Maestría en Comunicación

Departamento de Comunicación Social

Universidad del Norte

DIFUSIÓN NOTICIOSA Y MEDIOS SOCIALES: UN ANÁLISIS DE CONTENIDO A LA DIFUSIÓN DE TEMAS NOTICIOSOS DEL PERIÓDICO EL TIEMPO EN TWITTER DURANTE EL AÑO 2013

Tesis de Maestría presentada por

Eduar Barbosa Caro

Bajo la dirección del Doctor

Carlos Arcila Calderón

Barranquilla, diciembre de 2014.

A Dios.

A mis hermanos en Cristo.

A mi familia y a Johanna, gracias por tanto amor.

AGRADECIMIENTOS

Agradezco a Dios por darme la oportunidad de cursar esta Maestría en Comunicación. La fuerza y la provisión para afrontar este reto vinieron de Él.

Extiendo un agradecimiento especial a la Universidad del Norte por su apoyo durante estos años de estudio. Gracias a la Dra. Pamela Flores por su disposición para acompañar este proceso, su dedicación ha sido invaluable.

A mi director de tesis, el Dr. Carlos Arcila, por la rigurosidad y la motivación para continuar aprendiendo cada día. Su experiencia académica y profesional ha enriquecido este trabajo indiscutiblemente. Además, agradezco a la Universidad del Rosario por el espacio y los recursos facilitados para llevar a cabo mi estancia de investigación.

Doy gracias a mis padres y hermano por sus palabras de aliento y atenciones constantes. Mi corazón está con ustedes aun (y más) en la distancia. A Johanna, por sus consejos, su paciencia y nuestras largas conversaciones. Esta ha sido también una prueba de amor. A Esmeralda, Antonio, Piero y Pilly, por abrirme las puertas de su casa, y a la familia Ramírez Suavita, por su cariño e inmensa generosidad. Al profesor José Otero, por haber estado ahí durante tanto tiempo. A Luis Alberto y Katherine, porque gracias al Señor hemos podido compartir y fortalecernos mutuamente en la fe.

Al Pastor Orosmán Rozo, su familia y los hermanos de la Iglesia Bautista del Norte en Bogotá. Gracias por sus oraciones, su amabilidad y servicio.

Tabla de contenido

1. Introducción

La producción periodística [que para efectos de este estudio, se entiende como "cualquier texto en formato escrito, de audio o video que afirma ser (o que se presenta a la audiencia como) una declaración veraz sobre, o un registro de, una hasta ahora desconocida (nueva) característica del mundo social actual" (McNair, 1998, p. 4)] ha ido cambiando de escenario conforme avanzan las décadas. Es por esto que el presente estudio busca aportar, más que un recorrido teórico, herramientas metodológicas para abordar la relación actual del *social media* y los medios masivos que, en palabras de McQuail (2010), son "un canal de representación y expresión cultural, y una fuente primaria de imágenes de la realidad social y de materiales para formar y mantener una identidad social". Por tal motivo, este trabajo de corte cuantitativo se enfoca en el análisis de cómo se ha dado ese salto a los entornos digitales por parte de los medios masivos en Colombia, haciendo énfasis en el comportamiento de las temáticas subyacentes evidenciadas en los contenidos del periódico @ELTiempo en Twitter durante el año 2013.

El documento está dividido en cuatro grandes bloques, que han sido construidos en consonancia con el planteamiento del problema, la justificación, los objetivos y las preguntas de investigación formuladas.

En el primero de ellos está el estado del arte, donde se hace una revisión de resultados de estudios previos que sirven de referencia a la presente investigación. En segundo lugar se han construido los fundamentos conceptuales con los cuales se engloban los elementos fundamentales de la *Teoría de la Difusión de Innovaciones* que son base para

el estudio de la difusión de temas noticiosos, y las nociones principales sobre *Big Data*, cambios en el paradigma de los datos, el lugar de la teoría en la actualidad, análisis automatizado de contenido, *data mining* y *machine learning*, modelamiento de temas y análisis de sentimiento automatizado.

Un tercer bloque explica la metodología que se siguió para la recolección, procesamiento y análisis de los datos, para posteriormente pasar a un cuarto y último bloque compuesto por los resultados, la discusión y las conclusiones del estudio donde se exponen los principales hallazgos e inferencias que surgen del previo análisis de contenido.

2. Planteamiento del problema

Como en muchas otras disciplinas, la producción periodística ha sido partícipe de la evolución vertiginosa e innovadora de los entornos digitales en la última década. Debido a los cambios en la forma de pensar, gerenciar y distribuir la información, los canales de comunicación se han ido transformando a la par de las necesidades de los destinatarios finales, sean estos lectores, televidentes u oyentes.

Uno de los aspectos más relevantes de este cambio ha sido la implementación de herramientas web para extender las redes comunicacionales y así ampliar el *target* mediático. De ahí se desprende que los medios masivos de comunicación alrededor del mundo —incluyendo los de América Latina— le apuesten con más seguridad al uso de estas herramientas interactivas y los medios sociales para publicar, informar e interactuar con sus lectores y audiencias (Said-Hung et al., 2013; García de Torres et al., 2008; Lasorsa, Lewis & Holton, 2012; Caballero, 2000) .

Teniendo en cuenta lo anterior, se hace necesario el análisis de cómo se ha dado ese salto a los entornos digitales por parte de los medios masivos de comunicación en Colombia, haciendo énfasis, en este caso, en el comportamiento de las temáticas subyacentes evidenciadas en sus contenidos.

Como afirman Micó et al. (2008), este auge ha incrementado los cuestionamientos sobre el uso de la red en la producción periodística y la credibilidad de los que ahora son,

además, cibermedios. Por esta razón, el estudio y sistematización de los contenidos generados por los medios masivos colombianos en los medios sociales busca ampliar el conocimiento sobre este fenómeno en particular que toca la esfera de lo público desde lo digital.

Esta nueva forma de interactuar con sus públicos objetivos, le ha permitido a los medios masivos de comunicación aprender a moverse en estos nuevos medios y adaptar o transformar su lenguaje y su discurso. Dicha alfabetización mediática, útil en nuestras experiencias diarias, es definida por la Comisión Europea (2007) como "la capacidad de acceder a los medios de comunicación, comprender y evaluar con sentido crítico diversos aspectos de los mismos y de sus contenidos, así como de establecer formas de comunicación en diversos contextos". Por esto, es menester investigar y describir la manera en que, además de transmitir información, los medios masivos introducen variadas temáticas en su agenda noticiosa.

Por tal motivo, los medios sociales se convierten entonces en la base para el presente estudio, debido a su papel fundamental como mediadores entre las personas que se nutren a diario de múltiples informaciones y los medios que las producen.

En América Latina, los usuarios pasan en promedio 24,6 horas al mes en Internet y 9,4 horas en medios sociales (ComScore, 2013). Estos porcentajes, que incluso superan a los de Medio Oriente y Asia Pacífico, auguran un crecimiento exponencial en el consumo de medios sociales por parte de los latinoamericanos.

Para el caso de Colombia, el uso de estas plataformas se distribuye de la siguiente manera:

> Mantenerse en contacto con su familia y amigos (97%), enviar videos o fotos a sus contactos (89%), para expresar sus opiniones sobre temas de actualidad (83%), compartir información noticiosa con sus contactos (83%) y movilizar a sus contactos en torno a causas sociales o políticas (48%). Si tomamos como referencia estos valores, podremos observar que las posibilidades de estar expuestos a dichos mensajes generados desde plataformas de interacción digital aumentan considerablemente (Centro de investigación en comunicación y política – Universidad Externado de Colombia, 2012).

Nos proponemos, por las anteriores razones, evaluar desde los contenidos publicados en la cuenta de Twitter del periódico *El Tiempo* aquellas variables que generen un conocimiento más profundo de la difusión de temas noticiosos en los medios sociales de los medios masivos colombianos, acercándonos a la tendencia global de la generación de mensajes inmediatos, replicables y de alto impacto en la esfera digital.

2.1 Medios sociales y datos masivos

Al enfrentarnos al crecimiento exponencial de la capacidad de captura, procesamiento y análisis de datos, los medios sociales se han convertido en una de las fuentes que pueden proveer gran cantidad de data a un costo notablemente reducido, y además, "no nos ofrecen meramente una forma de localizar y mantener el contacto con amigos y

colegas: también toman elementos intangibles de nuestra vida diaria y los transforman en datos que pueden usarse para hacer cosas nuevas" (Mayer-Schönberger & Cukier, 2013, p. 116). Pero esto, a su vez, hace que emerjan nuevos retos desde distintas perspectivas, como lo enuncian Mayer-Schönberger & Cukier (2013) cuando reflexionan sobre la necesidad de dejar hablar a los datos, pues así "podemos establecer conexiones que nunca hubiésemos sospechado".

Para esto es necesario no tomar, como afirman los autores, "una muestra en lugar de un todo", pues "el conjunto de datos carece de la extensibilidad o maleabilidad que serían necesarias para que los mismos datos pudieran ser analizados otra vez con un propósito enteramente distinto de aquel para el que fueron recopilados en origen" (Mayer-Schönberger & Cukier, 2013, p. 40). Esta, entre otras, es una de las principales diferencias que suponen este nuevo modo de pensar y hacer la investigación.

Es este enfoque de datos masivos el que se abordará en la presente investigación, pues, siguiendo con el argumento de los autores, dicha perspectiva "puede no suponer el <<fin de la teoría>>, pero sí que transforma radicalmente nuestra forma de explicar el mundo" (Mayer-Schönberger & Cukier, 2013, p. 94), de tal manera que hoy podemos tener *muestras* más grandes, análisis más robustos y descripciones amplias de un fenómeno como lo es, en este caso, la relación del *social media* y los medios masivos.

3. Formulación del problema

¿Cómo fue la difusión de temas noticiosos en el canal de Twitter del periódico *El Tiempo* durante el año 2013?

4. Justificación

Esta investigación se ha planteado con el fin de contribuir al conocimiento científico y la reflexión sobre la frontera de dos campos de estudio que, en los últimos años, han observado sus propias convergencias desde variadas perspectivas: las ciencias de la computación y la comunicación.

El presente proyecto enmarcado en la línea de investigación titulada *Nuevos Medios* del Grupo de Investigación en Comunicación y Cultura (PBX), tiene como una de sus metas aportar en términos metodológicos y de resultados al análisis de grandes cantidades de mensajes producidos a través de medios sociales. Esto, debido a que como enuncia Rogers (2003), "los canales de medios masivos son usualmente los más rápidos y más eficientes para informar a una audiencia (…), para crear conciencia y conocimiento" (p. 18).

Hoy en día, se puede decir que "la comunidad de *digital humanities* fue una de las primeras adoptantes de los medios sociales, utilizándolos para la comunicación académica, la colaboración y diseminación" (Ross, 2012, p. 23) de ideas, trabajos y recursos, razón por la cual consideramos pertinente revisar, desde este mismo ámbito académico, el uso que los medios masivos hacen de estos.

Para efectos del análisis, se seleccionó el perfil de Twitter del periódico *El Tiempo* (@ElTiempo) puesto que dicho diario ocupa el primer lugar en el ranking de Alexa para

Colombia (www.alexa.com) entre todos los medios masivos digitales del país. Además, se ha considerado para su escogencia como caso de estudio la influencia que tiene como *cibermedio*, la cantidad de seguidores en su perfil de Twitter y el número de mensajes que publican diariamente en dicha cuenta. Así, como referente del país para América Latina, es importante registrar cómo realizan su difusión noticiosa en este entorno para poder generar una base de comparación frente a los perfiles de otros medios de la región.

La recopilación de los datos se llevó a cabo entre el 02/01/2013 y el 02/01/2014, obteniendo así mensajes publicados a lo largo de todo un año y no solo una muestra reducida de este. Se seleccionó este periodo de tiempo puesto que Twitter solo permite acceder, aproximadamente, a los últimos 3.200 tuits de cualquier perfil, razón por la cual la disponibilidad de los mensajes fue uno de los principales factores que contribuyeron a demarcar el lapso.

Se espera que esta investigación sirva como referencia a futuras indagaciones cuantitativas a partir de métodos inductivos que ya se vienen realizando en otras áreas del saber, y que desde hace pocos años han incluido a la comunicación y los análisis de contenido mediático[1]. Con estas nuevas formas de hacer ciencia, se abarca una mayor cantidad de información que nos permite profundizar sin sacrificar buena parte de los datos, reforzando así el hecho de que "a través del auge de los medios sociales y la creciente digitalización de la industria noticiosa, los investigadores en Ciencias Sociales pueden aprovechar un gran número de fuentes de conocimiento" (Verbeke et al., 2014).

[1] Como ejemplo se puede destacar el número especial del *Journal of Communication* titulado *Big Data in Communication Research* (http://onlinelibrary.wiley.com/doi/10.1111/jcom.2014.64.issue-2/issuetoc).

5. Objetivos

Objetivo general:

Caracterizar el proceso de difusión de temas noticiosos en el canal de Twitter del periódico *El Tiempo* durante el año 2013.

Objetivos específicos:

- Determinar las propiedades del contenido de la innovación (temas noticiosos) difundidas a través de la cuenta de Twitter del periódico *El Tiempo*.
- Describir el tiempo y los canales en el cual se enmarca la difusión de temas noticiosos publicados a través de la cuenta de Twitter del periódico *El Tiempo*.
- Determinar las variables que influyen en la difusión de temas noticiosos.

6. Preguntas de investigación

PI 1: ¿Cómo fue la difusión de innovaciones (temas noticiosos) en el canal de Twitter del periódico El Tiempo durante el año 2013?

> **PI 1.1** ¿Cuál es el contenido de las *innovaciones* (*temas noticiosos*) que emerge de los tuits publicados por el periódico @ElTiempo durante el 2013?
>
> **PI 1.2** ¿De qué manera se evidenciaron las *innovaciones* (*propiedades innovadoras*) en los temas noticiosos difundidos por la cuenta @ElTiempo?
>
> **PI 1.3** ¿Cómo se difundieron en el *tiempo* (momento de producción) los temas noticiosos a través del perfil @ElTiempo en el 2013?
>
> **PI 1.4** ¿Qué uso se le dio al *canal* de Twitter desde las fuentes (autoría del mensaje - @ElTiempo o RT) en los temas noticiosos difundidos por la cuenta @ElTiempo durante el 2013?

PI 2: ¿Cuáles son los factores que influyen en la difusión de innovaciones (temas noticiosos) en la cuenta @ElTiempo?

> **PI 2.1** ¿Influyen las características de la innovación (*propiedades innovadoras)* en la difusión de temas noticiosos en los tuits publicados por @ElTiempo durante el 2013?
>
> **PI 2.2** ¿Cómo se dio la relación entre el *tiempo* (momento de producción) de un tuit de la cuenta @ElTiempo y las *innovaciones* (temas noticiosos) que emergieron de los mensajes?
>
> **PI 2.3** ¿En qué forma afecta el *canal* (autoría del mensaje) a las *innovaciones* (temas noticiosos) que se observan en los tuits difundidos por el perfil @ElTiempo durante el 2013)?

6.1 Operacionalización de variables

Objetivo	Categorías	Subcategorías	Indicador
Determinar las propiedades del contenido de la innovación (temas noticiosos) difundidas a través de la cuenta de Twitter del periódico El Tiempo.	Innovación (TEMA NOTICIOSO)	Contenido de la innovación	Tema
		Propiedades innovadoras	Número de caracteres
			Número de palabras
			Número de enlaces
			Número de menciones por tuit
		Sentimiento	Tono
Describir el tiempo, el canales y el sistema social en el cual se enmerca la difusión de temas noticiosos difundidos a través de la cuenta de Twitter del periódico El Tiempo.	Tiempo	Momento de producción	Día codificado
			Hora
	Canal	Autoría del mensaje	RT o no
Determinar las variables que influyen en la difusión de temas noticiosos.	Innovación, tiempo, canal y sistema social	-	Modelo de regresión lineal múltiple para cada tema

Tabla 1. Operacionalización de variables.

6.2 Modelo de relación de subcategorías y las preguntas de investigación (PI)

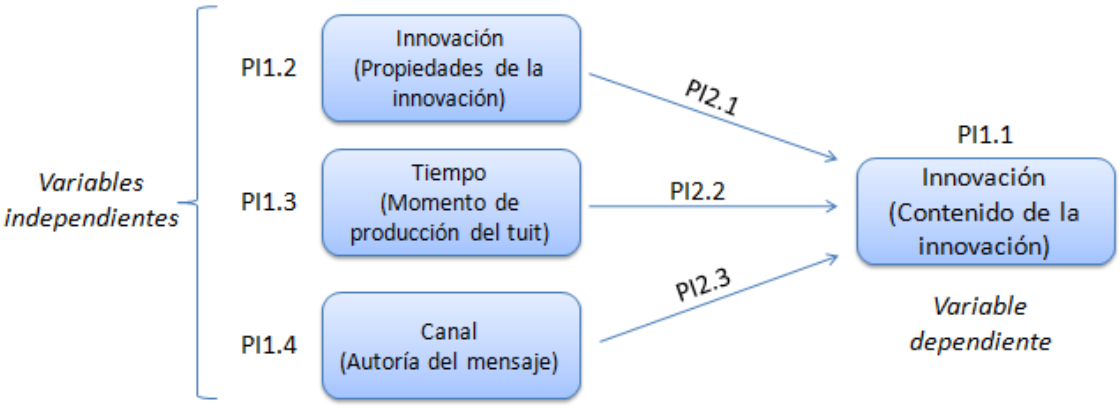


Diagrama 1. Modelo de relación de subcategorías.

7. Estado del arte

Los medios sociales han transformado la manera de hacer y consumir los productos periodísticos. Stassen (2010) afirma que esta actividad particular "se limitaba a leer el periódico, escuchar un boletín en radio o verlo por televisión" (p. 2), pero hoy, con la entrada al escenario comunicacional de los medios sociales digitales, vemos cómo los medios tradicionales han tenido que adaptarse a las demandas de estos nuevos espacios de interacción.

Este vínculo, que se teje actualmente entre los medios masivos de difusión noticiosa y los medios sociales, es cada vez más fuerte. Estos últimos permiten compartir mensajes de diferente longitud, acompañados de contenidos multimedia y enlaces a otros sitios, lo que aumenta notablemente el universo de información al que nos vemos expuestos diariamente. Así, se configura la *Big Data* desde esta perspectiva como vastos conjuntos de datos (Richards & King 2014) con diversas metodologías de análisis que permiten entender mejor la esfera pública electrónica en la que nos encontramos inmersos (Neuman et al., 2014).

Como afirman Said, Arcila, & Méndez (2011), "los cibermedios están experimentando constantes cambios en sus formatos y contenidos con el fin de proveer a sus visitantes de servicios informativos de calidad" (p. 48). Así, los periódicos –que conforman buena parte del espectro mediático– a los que antes solo se podía acceder en papel, ahora brindan la posibilidad a sus lectores de consumir e interactuar con sus noticias mediante Internet, más específicamente, desde sus portales web y perfiles en medios sociales.

7.1 Producción periodística y medios sociales

Ure & Parselis (2013) observaron el comportamiento de 22 perfiles en medios sociales entre los que se encontraban, precisamente, medios de comunicación (corporativos) y periodistas (personales). En dicho análisis se encontró que, para el caso de Argentina, en un período de 12 semanas (con una recolección de data por semana) los medios masivos recibieron en promedio una respuesta por cada 322 tuits (0,31%), contrario a lo que sucedía con los perfiles personales (una por cada 3 tuits, es decir, 31,76%). Esto deja ver que, como bien esbozan los autores, el nivel de reciprocidad para los medios estudiados se acerca a cero, mientras que el de los periodistas es considerablemente mayor.

En otra experiencia, Gil de Zúñiga (2012) confirma la relación entre el uso de sitios de medios sociales para mantenerse informado sobre política, asuntos públicos, la obtención de información sobre la comunidad y el capital social (entendido como un antecedente de comportamiento relacionado con el ámbito público y el bienestar colectivo, ya sea en el campo político o comunitario). Este trabajo arrojó como resultado, entre otros, una relación positiva con significancia estadística ($\beta = .153$, $p. < .001$) entre las variables mencionadas en el párrafo anterior, lo que nos invitaría a pensar en la importancia de los medios masivos al momento de entregar información relevante, inmediata y confiable a través de sus perfiles en medios sociales. Esto, debido a que los contenidos podrían influir en la forma como las personas piensan o actúan frente a la realidad circundante.

También se han conducido estudios desde otros puntos de vista, como por ejemplo, los realizados por Schultz (2012) y Wasike (2013) en Estados Unidos. El primero, un estudio sobre *social media* y *branding* realizado con periodistas y basado en la teoría de Rogers (2003), descubrió lo siguiente:

> [Aquellos] con más de 20 años de experiencia tienen las respuestas más bajas con respecto al *social media* y asuntos de *branding* (…) y que comparado con otros grupos, los reporteros de periódicos estaban significativamente menos inclinados a creer que el reportaje a través de *social media* era importante ($t = -3.26$, $df = 198$, $p < .001$) (Schultz, 2012, pp. 102, 104).

Por su parte, Waskie (2013) expone una descripción detallada de las publicaciones hechas en las cuentas personales de 8 *Social Media Editors (SME)*. En este caso particular, 950 tuits fueron analizados para observar los temas y el tipo de contenidos que se publican desde estos perfiles, emergiendo así relaciones y patrones entre estos mensajes.

La anterior publicación concluye que hay una relación significativa entre el formato del medio y la utilización de ciertos marcos temáticos específicos en las publicaciones [$X^2(8, N=423) = 16.07$, $p < 0.05$]. Además, se pudo evidenciar que los *SME* de medios impresos "tienden a publicar más artículos sobre ciencia y tecnología junto a los de guerra y terrorismo, mientras que los que pertenecen al medio televisivo enlazaron más artículos relacionados con entretenimiento" (Wasike, 2013).

Aunque han surgido novedosas formas de llevar a cabo el ejercicio periodístico en la red (replicando comentarios y noticias desde otras páginas, cuentas o perfiles, por ejemplo), algunos medios mantienen las fuentes oficiales como base fundamental de sus contenidos. Este es el caso expuesto en el trabajo de Harlow & Johnson (2011), donde se observa cómo en los artículos del New York Times (92%) fue mucho más probable que se utilizaran fuentes oficiales no-ciudadanas para relatar lo ocurrido en Egipto [$X^2(2)= 85.27$, $p < .001$], a diferencia de lo que se vio en Twitter, donde las fuentes oficiales no-ciudadanas solo fueron utilizadas en un 18%. Este contraste, como argumentan los autores, podría ser una falencia al momento de cubrir eventos importantes como este, pues se "privilegian las fuentes oficiales sobre las ciudadanas" (Harlow & Johnson, 2011).

Lotan et al. (2011) aseguran que tanto en Egipto como en Túnez, los medios masivos primaron en la entrega de informaciones noticiosas a través de Twitter (70,7% y 69%, respectivamente), lo que se corresponde, en cierto modo, con los resultados del estudio enunciado previamente.

Aplicando el modelo de Difusión de Innovaciones de Rogers (2003) en una universidad de los Estados Unidos, los autores Peslak, Ceccucci & Sendall (2010) hallaron que este uso general de los medios sociales se debe, básicamente, a cuatro condiciones puntuales: 1. Compatibilidad con el modo de vida del usuario; 2. Baja complejidad o simplicidad en su uso; 3. La habilidad de probar la tecnología de manera fácil; 4. Los beneficios y ventajas de su uso. Todos elementos que, partiendo de la teoría de Rogers (2003), pueden observarse en la difusión de una innovación tecnológica, idea o información.

A partir de lo descrito, se explica que herramientas como Twitter pueden funcionar incluso como actores y catalizadores para el cambio (Harlow & Johnson, 2011), aprovechando su potencial de influencia y la libertad que se tiene en la plataforma frente a la llamada *imparcialidad periodística*. En contraste con la marcada línea institucional que se puede manejar desde una página web corporativa, los medios sociales brindan la oportunidad de reinventar las maneras de contar estas historias y darle voz a quienes, desde otras perspectivas, también construyen los hechos noticiosos (con respuestas a los mensajes, retuits, etcétera).

7.2 Difusión de información en medios

Los estudios que abordan esta difusión noticiosa pueden rastrearse incluso décadas atrás. Estos atendían a necesidades de comprensión de fenómenos particulares, noticias importantes y eventos de talla mundial. Deutschmann & Danielson (1960) encontraron, por ejemplo, en la década de los 60, que una noticia podía tardar entre uno y dos días en completar el proceso de difusión, incluso teniendo gran despliegue en los medios tradicionales. Esto, sin duda alguna, ha cambiado drásticamente en nuestro tiempo.

Por otra parte, en su estudio sobre la difusión noticiosa del asesinato de Kennedy, Greenberg (1964) reporta que casi 9 de cada 10 se enteraron de la noticia entre los primeros 60 minutos posteriores al primer anuncio, y que además la mitad (50%) conoció el evento por medio de otra persona. Aquí cabe resaltar que su primera reacción fue *usar luego la radio y la televisión* con el fin de corroborar los datos.

Un estudio similar, esta vez enfocado en la difusión de noticias referentes a la muerte de la princesa Diana, fue llevado a cabo por Henningham (2000). En él se encontró que aquellos más afectados emocionalmente (en mayor o menor medida) por la muerte de la princesa Diana, tenían más probabilidades de contarle a otro sobre el suceso. Además, los principales canales a través de los cuales se enteraron de la noticia fueron la televisión (40%), otra persona (29%) y la radio (28%).

Como vemos, estos trabajos van dando indicios de lo que vendría a ser un campo de estudio con numerosos trabajos de investigación desde la Comunicación. Sin embargo, por las fechas en que fueron realizados estos y otros estudios, los medios sociales no eran un elemento fundamental en el proceso. Nos sirven, pues, como precedente de análisis en un enfoque influido por la teoría de Rogers (2003), y desde los cuales derivan otros intereses y objetos de estudio como el que ocupa esta investigación.

Actualmente, trabajos transculturales como el de Golder et al. (2011) y Ritter, Preston & Hernández (2013) han puesto énfasis en otros aspectos de la difusión de información que pueden ser explorados a través de los medios sociales. Utilizando el *software* para análisis textual *LIWC* y un corpus de 509 millones de tuits –que, según Parks (2014), calificaría como *Big Data*–, Golder et al. (2011) encontraron que los sentimientos positivos en Twitter se manifiestan más durante las primeras horas de la mañana y cerca a la media noche (Golder et al., 2011). Además, evidenciaron que los días en que se publican los mensajes más positivos son los sábados y domingos (M = 0.058), teniendo en cuenta que durante la semana los horarios donde hay mayor volumen de

publicaciones va de 9:00 A.M. a 10:00P.M., como se observa en el Gráfico 1 (Golder et al., 2011).

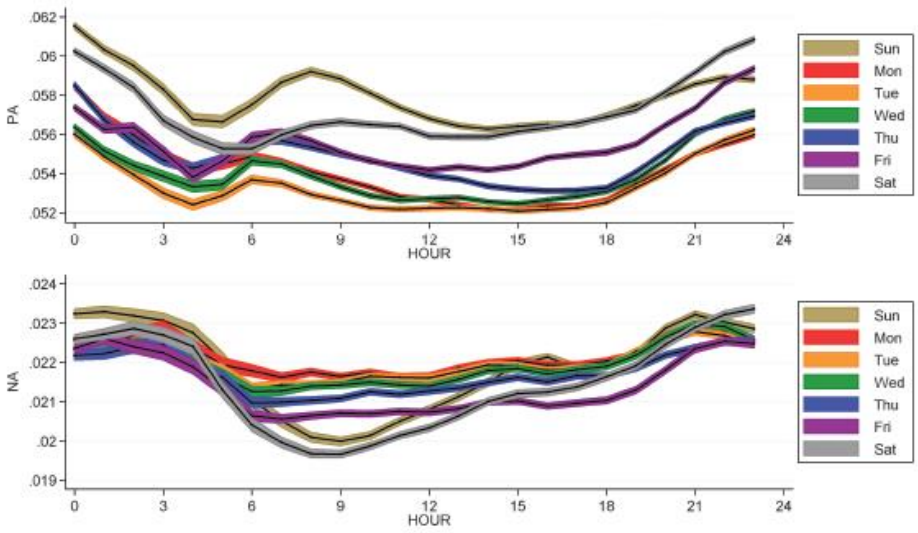


Gráfico 1. Cambios en los afectos de los individuos por horas y días de la semana. PA: *Positive affect* / NA: *Negative affect*. Fuente: Tomado de Golder et al. (2011).

Otras investigaciones como las de Gruzd & Roy (2014), Conover et al. (2011), Jungherr (2014) y Mejova & Srinivasan (2012) proveen acercamientos a temas políticos desde Twitter, ya sea en términos de filiación o sentimientos. Entre los resultados obtenidos por Gruzd & Roy (2014), podemos destacar que un 40% de los mensajes publicados por personas cuya orientación política se inclina hacia la izquierda tienden a ser negativos u hostiles en su naturaleza (Gruzd & Roy, 2014). El Gráfico 2 expone una visualización de los vínculos entre perfiles a partir de su filiación política:

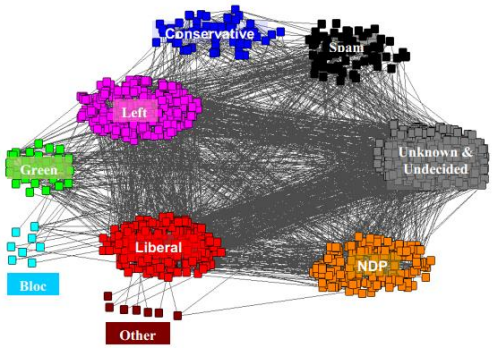


Gráfico 2. Red de menciones en Twitter. Relaciones entre cuentas de Twitter a partir de su filiación política. Fuente: Tomado de Gruzd & Roy (2014).

Mejova & Srinivasan (2012), en otro escenario político donde contrastaron comentarios de YouTube y publicaciones en Twitter, encontraron que un 40% de los contenidos en la red de microblogs carecen de sentimiento, y que solo el 17,5% de los mensajes son positivos. El restante 42,4% se divide entre los mensajes negativos (40,6%) y aquellos que tienen un sentimiento *mezclado* (1,8%). Por su parte, Jungherr (2014) encontró que es notable el aumento de mensajes políticos en Twitter cuando se acerca un evento altamente publicitado, como por ejemplo un debate televisivo que involucra a los candidatos que puntean las encuestas. En estos periodos de tiempo, los *RTs* y menciones suelen disminuir (Jungherr, 2014).

Con respecto a otros estudios en el área, vale la pena destacar a continuación algunos de los que han volcado su atención a la extracción de temas, el manejo del lenguaje y la manera en que se ha dado la adopción y uso de estas nuevas plataformas por parte de los medios masivos tradicionales, personas y organizaciones.

En general, hemos presenciado que la preocupación académica por abordar los medios sociales se ha incrementado, de tal forma que se han podido integrar técnicas de las

Ciencias de la Información —como el *Machine learning* y el *Data mining*, entre otras— a los procesos analíticos de otros ámbitos, proveyéndose así de nuevas herramientas que respondan a los cambios en los paradigmas de investigación.

Aplicaciones como la de Ramage, Dumais & Liebling (2010), donde se implementó un modelo de *LDA (Latent Dirichlet Allocation)* con el objetivo de develar estructuras latentes en una colección de documentos (8'214.019 publicaciones de Twitter), han mostrado las posibilidades que se abren ante esta combinación de técnicas. En este caso, con el uso del modelo se pudieron construir cuatro categorías en las que los autores dividen los mensajes de Twitter: 1. *Sustancia*, que corresponde a los eventos, ideas, cosas o personas; 2. *Social*, que tienen una finalidad comunicativa; 3. *Estado*, que corresponde a las actualizaciones de *status* personales; 4. *Estilo*, que indica un uso elaborado de la lengua (Ramage, Dumais & Liebling 2010).

En consonancia con este enfoque, Kang, O'Donovan & Höllerer (2012) desarrollaron varios modelos computacionales para medir la *credibilidad* en Twitter. Así, se toparon con ciertos hallazgos importantes en términos de caracterización de los mensajes, como por ejemplo que "la presencia de enlaces en los tuits se relaciona con el sentimiento en una forma interesante: si la métrica de sentimiento está polarizada, ya sea de manera negativa o positiva, los links son más comunes" (Kang, O'Donovan & Höllerer, 2012, p. 186). Además de este resultado, el estudio arrojó que aquellos tuits en los cuales había URLs insertas fueron más retuiteados que aquellos que no los tenían, probando en cierta medida que existe una relación entre estos indicadores.

Dentro de este espectro de investigaciones también se encuentran algunos proyectos que han puesto su mirada en otras plataformas como Facebook y su relación con los cibermedios (Arcila & Said, 2012; Noguera, 2010), la proximidad de los contenidos (Schaal, O'Donovan & Smyth, 2012) o el uso de los idiomas español e inglés en Twitter (Argüelles & Muñoz, 2012).

La investigación planteada por Noguera (2010) acogió dentro de su muestra periódicos nacionales con referentes en papel, versiones digitales de medios regionales, medios que solo se encuentran en digital e incluso diarios digitales locales. En ella se pudo establecer, entre otros resultados, que el 69,3% de los medios estudiados no establecen conversaciones ni responden a los comentarios de quienes generan interacciones en sus perfiles.

Por su parte, Argüelles & Muñoz (2012) abordaron dos cuerpos de mensajes (uno en inglés y otro en español) con más de 4 millones de palabras cada uno. En los conteos caracteres, Argüelles & Muñoz (2012) encontraron que en promedio solo se utiliza un poco más de la mitad del espacio (51,07%) que provee Twitter para publicar (140 caracteres), pero que en inglés este número es más elevado (75,63%) que en español (67,37%). Además, teniendo como referencia los emoticones, Argüelles & Muñoz (2012) detectaron que son más los tuits que indican felicidad (21.726) que aquellos que indican sentimientos cercanos a la tristeza (3.190). También develaron que el día más mencionado en ambos corpus fue el viernes (*Friday* = 1.013; Viernes = 1.156), y para el caso de los enlaces, los porcentajes de uso fueron bajos tanto en español (2,34%) como en inglés (2,24%). Aunque para las menciones los indicadores son un poco más altos,

siguen estando por debajo del 10% (español = 4,98%; inglés = 6,20%) (Argüelles & Muñoz, 2012).

Con el fin de observar el uso de la Web 2.0 por parte de los cibermedios, en América Latina y España se han hecho aportes teórico-prácticos en la investigación del binomio cibermedios-medios sociales, donde se pueden resaltar los trabajos de Said & Arcila (2011) y Rodríguez-Martínez, Codina & Pedraza-Jiménez (2010). Estos últimos construyeron un estudio comparativo con cinco diarios españoles, uno de Inglaterra, uno de Francia y otro de Estados Unidos. Allí se constató una tendencia por parte de los diarios a vincularse a medios sociales como Facebook, Twitter y YouTube (Rodríguez-Martínez, Codina & Pedraza-Jiménez, 2010), respaldado por el hecho de que el 100% de los medios tenían perfiles activos en las dos primeras plataformas sociales.

García et al. (2011) combinaron un análisis del uso de los medios sociales (Twitter y Facebook) por parte de medios informativos iberoamericanos con la aplicación de entrevistas semi-estructuradas a las personas encargadas de coordinar estos medios sociales en cada caso de estudio. Como conclusiones, la investigación arrojó que en Twitter "el porcentaje de mensajes conversacionales asciende a 25,8% y los mensajes constituidos por titulares con enlace tienen algo menor peso (63,5%) pero son, también, mayoritarios" (García et al., 2011). Sumado a lo anterior, se observó a través de la medición que en el 68,1% de los casos el titular y el enlace que publicaron los medios en Facebook sobre sus noticias es idéntico al publicado en su sitio web.

Este tipo de análisis se han replicado en otras latitudes. Uno de similares características fue llevado a cabo en periódicos de Suecia. En él, Telja (2011) mostró que todos los

medios estudiados dan la posibilidad de compartir sus contenidos en Facebook y Twitter, a excepción de uno de ellos que no cuenta con este último medio social. En India, Guul & Islam (2013) estudiaron 21 periódicos de los 58 listados por el *CNS Directory*. Entre estos medios, el 42,85% hace uso de Twitter y el 38,9% de Facebook (ver Gráfico 3). A continuación encontramos una gráfica que muestra el panorama general de adopción:

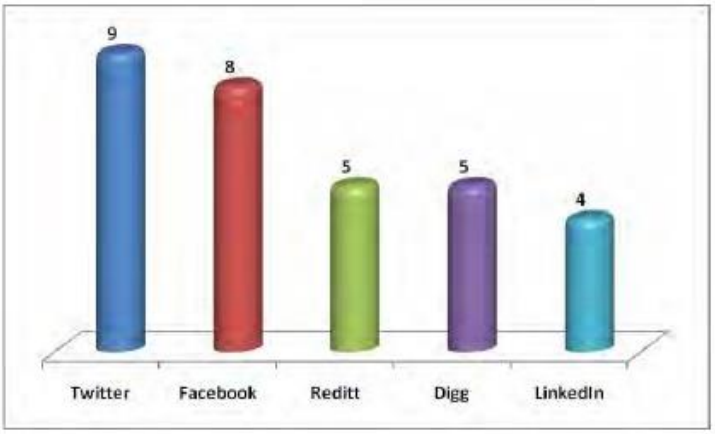


Gráfico 3. Adopción de *social media* por parte de periódicos *online* de Kashmir (India). Fuente: Tomado de Guul & Islam (2013).

Considerando el momento de reconfiguración por el que atraviesan actualmente de las Humanidades (Romero, 2014) y los nuevos recursos tecnológicos al servicio de la investigación y la colaboración científica (Arcila et al. 2014), es menester revisar estudios que se muevan en esta hibridación teórico-metodológica, evidenciando contribuciones conceptuales y prácticas en el análisis de los medios sociales y su relación con los medios masivos.

Antes de examinar aquellos trabajos que se enfocan en la difusión noticiosa, vale la pena traer a colación aplicaciones para técnicas como el *Topic modeling* y el *Sentiment*

analysis que, si bien no se topan con lo noticiosos, ven en el *social media* su fuente primaria de datos.

Se han evaluado distintas formas de extraer temas de microblogs (Rajani, 2014; Yang et al., 2013; Lim, Chen & Buntine, 2013), lo que supone de igual forma estudios a distintos niveles. Paul & Dredze (2014), por ejemplo, utilizaron el modelamiento de temas con un objetivo primordial: descubrir temas de salud teniendo como base un corpus de tuits. Esto les permitió identificar dolencias a partir de grupos de mensajes (como se observa en el Gráfico 4), en donde emergieron algunas tales como alergias, depresión, tos y enfermedades respiratorias, ansiedad, lesiones deportivas, entre otras, que luego pudieron ser comparadas en el tiempo (ver ejemplo de la *influenza* en el Gráfico 5) con datos reales de los Estados Unidos (Paul & Dredze, 2014).

	Ailments					
	Influenza-like Illness	**Insomnia & Sleep Issues**	**Diet & Exercise**	**Cancer & Serious Illness**	**Injuries & Pain**	**Dental Health**
General Words	better hope ill soon feel feeling day flu thanks xx	night bed body ill tired work day hours asleep morning	body pounds gym weight lost workout lose days legs week	cancer help pray awareness diagnosed prayers died family friend shes	hurts knee ankle hurt neck ouch leg arm fell left	dentist appointment doctors tooth teeth appt wisdom eye going went
Symptoms	sick sore throat fever cough	sleep headache fall insomnia sleeping	sore throat pain aching stomach	cancer breast lung prostate sad	pain sore head foot feet	infection pain mouth ear sinus
Treatments	hospital surgery antibiotics fluids paracetamol	sleeping pills caffeine pill tylenol	exercise diet dieting exercises protein	surgery hospital treatment heart transplant	massage brace physical therapy crutches	surgery braces antibiotics eye hospital

Gráfico 4. Palabras relevantes asociadas con los temas. Fuente: Tomado de Paul & Dredze (2014).

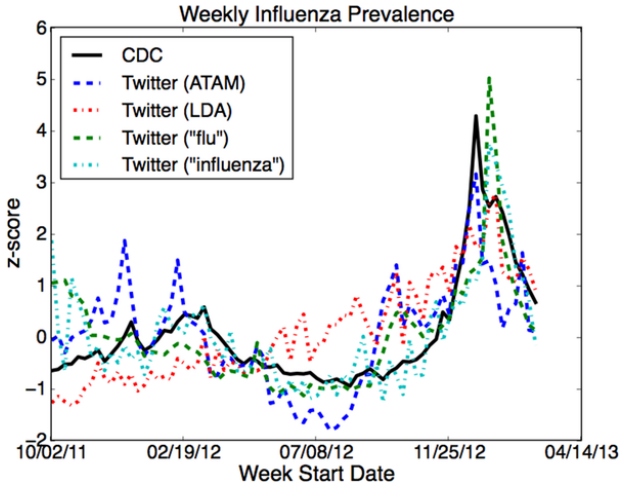


Gráfico 5. *Influenza* a través del tiempo. Fuente: Tomado de Paul & Dredze (2014).

Otras aproximaciones ilustrativas han mapeado, a partir de tuits y geolocalización, contenidos ciudades y países completos (Gerber, 2014; Ferrari et al., 2011; Hong et al., 2012; Ghosh & Guha, 2013).

En el estudio de Ghosh & Guha (2013), en el que buscaban rastrear terminología referente a la obesidad, se configuraron tres temas base sobre el tema, a saber, 1. *Childhood obesity and schools* (21.887 tuits); 2. *Obesity prevention* (32.000 tuits aprox.); 3. *Obesity and food habits* (44.230 tuits) (Ghosh & Guha, 2013). Habiendo consolidado estos resultados y la procedencia de los tuits (455.981 geolocalizados), se pudieron hacer comparaciones entre, por ejemplo, los tuits que pertenecían a la categoría de *obesity and food habits* (3er tema) y los restaurantes de McDonald's. El Gráfico 6 muestra las visualizaciones resultantes donde, cabe aclarar, la distancia promedio entre los tuits y los restaurantes de McDonald's es de menos de una milla:

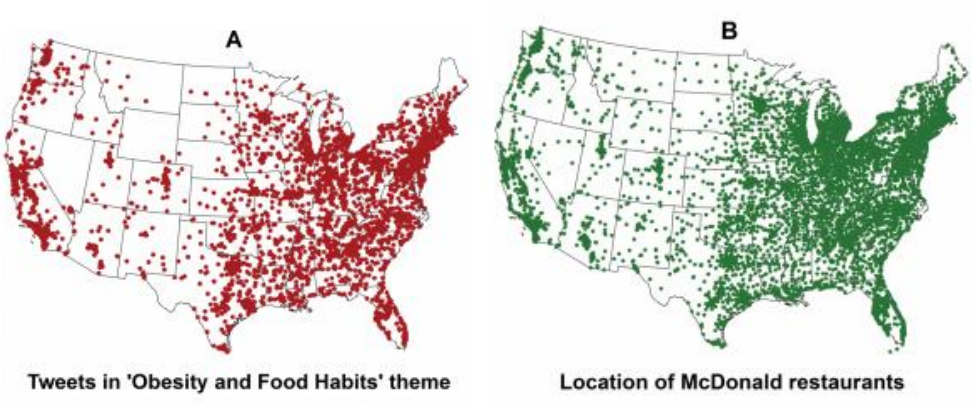


Gráfico 6. Ubicación de los tuits del tema *obesity and food habits* (A) y los restaurantes de McDonald's (B). Fuente: Tomado de Ghosh & Guha (2013).

Algunas experiencias, por otra parte, han estudiado el comportamiento de los usuarios (Bogdanov et al., 2013) y los temas a los que aluden frecuentemente (Asfari et al., 2013; Michelson & Macskassy, 2010), pero se mantienen por fuera del contexto específicamente noticioso. Por tal motivo, y habiendo visto ejemplos en otras disciplinas, a continuación ampliaremos estudios que, precisamente, emprenden estas tareas de análisis de grandes cantidades de datos con computación avanzada.

7.3 Noticias, medios sociales y análisis de grandes cantidades de datos

A continuación se explorará lo que ha sucedido, específicamente, con el *social media*, los medios masivos y el tratamiento de grandes cantidades de datos con computación avanzada.

Actualmente, podemos afirmar que los medios sociales han creado un ecosistema más complejo en términos de la creación y distribución de los productos noticiosos

(Newman, Dutton & Blank, 2012), y de la misma forma han modificado las prácticas científicas alrededor de diversos temas de investigación, como lo son, por ejemplo, las respuestas emocionales en función de la difusión noticiosa (Ibrahim, Ye & Hoffner, 2008), los temas y entidades expuestas en artículos de noticias (Newman, et al. 2006), la dinámica temporal de mensajes con respecto a eventos específicos (Jungherr, 2014) y la aceptación o rechazo de información (Emery et al., 2014).

Por tal motivo, vemos la necesidad de que desde las exploraciones académicas se aporte a la construcción de conocimiento sobre este fenómeno, sus elementos y dinámicas. Para Kim & Oh (2011), el problema de entender estas grandes cantidades de datos noticiosos a través de un largo periodo de tiempo radica en las diversas etapas del proceso, que incluyen el descubrimiento de los temas, hallar aquellos que son similares y agruparlos, encontrar asuntos de corta duración en los temas e identificar cómo estos cambian en el tiempo (Kim & Oh, 2011).

En su trabajo sobre cadenas de temas en artículos noticiosos, Kim & Oh (2011) se enfocaron en "temas y detección y rastreo de eventos, modelamiento de temas probabilístico y *temporal news mining*" (p. 2). Allí, habiendo dividido un corpus de 130.000 documentos de noticias coreanas en intervalos de tiempo (debido a que la secuencialidad de las noticias lo permite), establecieron que la mayoría de los temas de larga duración encontrados en el corpus "pueden ser rotulados como *política*, *negocios* o *deportes*, y que los temas en esas cadenas reflejan a su vez una amplia variedad de temas en esas categorías generales" (Kim & Oh, 2011, p. 12).

También confirmaron que algunos *temas únicos* o de corta duración que aparecieron en el *topic modeling* son incoherentes, sin embargo, buena parte de ellos podrían representar eventos relevantes como la muerte de alguien famoso o el aumento en la seguridad aérea (Kim & Oh, 2011). Por esta razón, dichos *asuntos temporales* pueden ser de interés en investigaciones futuras.

Para Ha-Thuc et al. (2009), la escalabilidad y la inhabilidad para desechar segmentos de texto de poca relevancia son dos de las limitaciones que motivaron, en parte, su estudio sobre el seguimiento de eventos noticiosos a partir de *topic modeling*. En su investigación, Ha-Thuc et al. (2009) utilizaron 1 millón de *blogs* provistos por el servicio de indexación *Spinn3r* para el análisis, y seleccionaron 10 eventos noticiosos (de la lista de los eventos más populares de *Wikipedia*) para seguirlos por espacio de dos meses.

Con el modelo plantado por Ha-Thuc et al. (2009), se observó la intensidad de los eventos en el tiempo, pero además se hizo un *sub-event tracking* donde se desglosó uno de estos eventos en particular, a saber, las elecciones presidenciales en los EEUU. Los resultados de visualización de estos análisis basados en la temporalidad y la distribución de temas se muestran en los Gráficos 7 y 8:

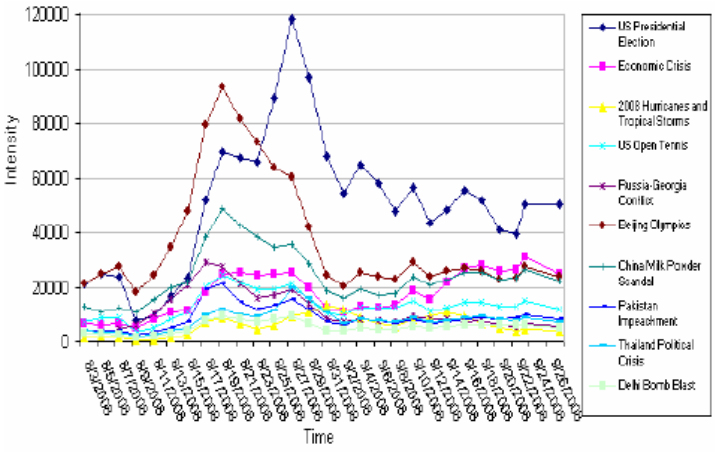


Gráfico 7. Intensidad temporal de los eventos. Fuente: Tomado de Ha-Thuc et al. (2009).

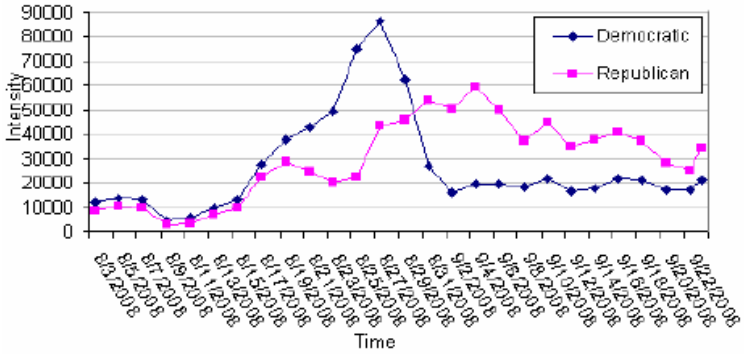


Gráfico 8. Intensidad temporal de los sub-eventos. Fuente: Tomado de Ha-Thuc et al. (2009).

En otra perspectiva, Zhao et al. (2011) abordaron la cuestión de si en realidad Twitter no es más que un *news feed* más rápido que los medios noticiosos tradicionales comparando contenidos de este medio social (1'225.851 documentos) con los del *New York Times* (NYT) (11.924 documentos), cubriendo un lapso de tiempo que iba desde el 11 de noviembre de 2009 hasta el 1^{ero} de febrero de 2010.

Para ejecutar esta tarea, Zhao et al. (2011) recurrieron al *unsupervised topic modeling*, una forma de extraer temas (temas semánticos subyacentes) usando únicamente las palabras que se encuentran en un conjunto de documentos (Blei & McAuliffe, 2007).

Así, categorizaron los artículos del NYT y los tuits basándose tanto en los estándares del medio (NYT) como en una tipología tópica en la que los *temas semánticos* se clasificaron como *event-oriented topics*, *entity-oriented topics* y *long-standing topics* (temas orientados a eventos, temas orientados a entidades y temas de larga duración); de esta manera, por ejemplo, cada tuit fue asociado a un tema, y cada tema a una categoría específica (Zhao et al., 2011).

En sus resultados, Zhao et al. (2011) muestran que los *event-oriented topics* están más presentes en el NYT que en Twitter, pues en este último predominan los *entity-oriented topics*, y añaden que los grados en que se presentan cada una de las categorías tópicas son bastante diferentes (Zhao et al., 2011), tal como se muestra en los Gráficos 9, 10 y 11:

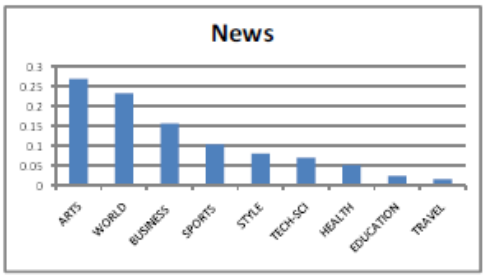


Gráfico 9. Distribución de categorías en el NYT.

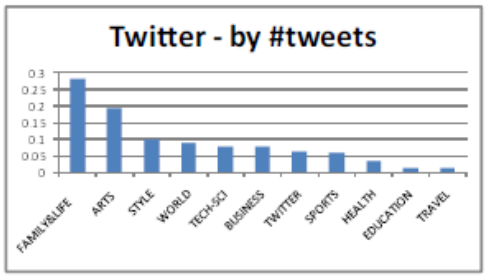


Gráfico 10. Distribución de categorías en Twitter (# tuits).

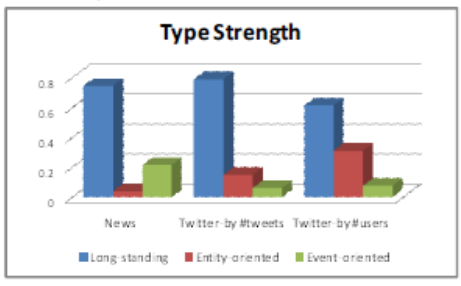


Gráfico 11. Distribución de las tipologías de temas en los *data sets*. Fuente: Tomados de Zhao et al. (2011).

En consonancia, Zhu & Li (2013) emprendieron una investigación sobre temas comunes/no comunes entre *social media* y medios masivos, y para esto seleccionaron dos casos de estudio (*Occupy Wall Street* y *Kim Jong-il's death*) y dos eventos específicos (*Tianagong-1's Launch* y *Yueyue Accident*) como base para comparar estas diferencias.

Con respecto al caso de estudio *Occupy Wall Street*, se hallaron cinco (5) temas comunes entre ambos corpus (*Background words*, *US Finance Economy*, *Demonstrations*, *A journal* y *Jobs*), y para el caso de *Kim Jong-il's death*, otros seis (6) temas más (*Background words*, *States said…*, *Funeral Silence Memorial*, *Die at 69*, *Kim's* y *destroy*), lo que para los autores ratifica que el número de temas varía según el evento al que se encuentren relacionados (Zhu & Li, 2013).

Como aporte fundamental del estudio de Zhu & Li (2013), es importante destacar la comparación realizada entre la presencia de los mismos temas noticiosos tanto en medios sociales como en los medios masivos, y el hecho de que concluyeran que más investigación es necesaria para observar la *esperanza de vida* (Zhu & Li, 2013) de temas en *microblogs* y medios masivos.

Al tener data de ambas plataformas (con los eventos de *Tianagong-1's Launch* y *Yueyue Accident*), Zhu & Li (2013) cotejaron estas variaciones, resultando en la Tabla 2, que se encuentra a continuación:

Event	Discovered Top Words (Manual Label)	Microblog	News
Tiangong-1's Launch (K = 12)	minute, hour, launch, 21, day (Launch)	0.078	0.171
	China, space station, outer space, space, dock (Chinese Space Station)	0.072	0.202
	no, watch, all, Tiangong, say (User Activities)	0.375	-
	mission, control, astronaut, rendezvous, flight control (Mission)	-	0.588
Yueyue Accident (K = 12)	moral, social, Heroic Acts, law, not, protect (Moral and Law)	0.191	0.123
	Chen Xianmei, rubbish scavenge, journalist, hospital, granny (Chen Xianmei)	0.02	0.347
	watch, Yue, see, no, say, video (Watch the Video)	0.224	-
	Hu, express, deputy director, introduce, functionality, brain death (Long Reports)	-	0.288

Tabla 2. Temas de Microblog vs Temas de medios masivos. Fuente: Tomado de Zhu & Li (2013).

Consideraciones finales

Los estudios descritos en el apartado anterior demuestran que la popularidad de Twitter va en crecimiento y sus usuarios se encuentran en todos los continentes (Java et al., 2007), y reafirma la necesidad de observar la plataforma desde perspectivas mediáticas y noticiosas, pues se han realizado estudios a nivel global en las ciencias de la salud (Paul & Dredze, 2014), con componentes de geolocalización (Gerber, 2014; Ghosh & Guha, 2013; Ferrari et al., 2011), temas de interés (Asfari et al., 2013; Michelson & Macskassy, 2010) y análisis de comportamiento de usuarios (Bogdanov et al., 2013) pero el abordaje del hecho noticioso en medios sociales es aún incipiente en nuestro contexto de América Latina.

Debido a la insuficiencia de abordajes de este tipo (en envergadura y técnicas de análisis) en el ámbito colombiano, los proyectos recogidos en este estado del arte suscitan la pregunta de investigación que guía este estudio: ¿Cómo fue la difusión de temas noticiosos en el canal de Twitter del periódico *El Tiempo* durante el año 2013? Esto, con el fin de aportar a la construcción de avances en la distribución temporal de temas (ej. Zhu & Li, 2013; Kim & Oh, 2011), el análisis de los contenidos noticiosos (como el caso de Zhao et al., 2011), el análisis de sentimiento (ver Golder et al., 2011; Mejova & Srinivasan, 2012) y la aplicación de estos métodos que en otro momento eran más difíciles de ejecutar debido a las restricciones de accesibilidad a la tecnología, lo reciente de las nuevas plataformas y la capacidad de procesamiento de grandes cantidades datos.

8. Fundamentos conceptuales

8.1 Difusión de innovaciones

La teoría de la Difusión de Innovaciones, en el área de las Humanidades, ha sido desarrollada desde los años 60 por el sociólogo Everett Rogers y otros académicos, tanto dentro como fuera de los EEUU. Esta perspectiva, que se ha ido extendiendo a otras geografías, muestra grandes repercusiones en la forma de abordar la realidad social y sus cambios constantes.

La teoría *Difusión de Innovaciones* fue aplicada en principio al campo de la sociología (rural), pues aún hoy, en palabras de Rogers (2003), "explica el cambio social, uno de los procesos más fundamentales del ser humano" (p. xviii). Esto, debido a que evalúa cómo se propagan e introducen en la sociedad las diferentes tecnologías y mensajes (innovaciones) que modifican, mejoran o distorsionan nuestra manera de hacer las cosas.

Esta propagación de tecnologías y mensajes, llamada *difusión*, está definida por Everett Rogers en su libro *Diffusion of innovations* (5ª. ed.) de la siguiente forma:

> *Difusión* es el proceso en el cual una innovación es comunicada a través de ciertos canales en el tiempo entre los miembros de un sistema social. Es un tipo especial de comunicación, en el que los mensajes están cargados de nuevas ideas. (Rogers, 2003, p. 5, traducción propia)

Este primer acercamiento nos permite identificar elementos clave que emergen del estudio de la difusión de innovaciones, a saber: "(1) una *innovación* (2) es *comunicada* a través de ciertos *canales* (3) en un *tiempo* definido (4) entre miembros de un *sistema social*". (Rogers, 2003. p. 11)

Aquí, es preciso explicar brevemente cada uno de los ítems enunciados en el párrafo inmediatamente anterior, pues ellos nos servirán como herramientas de análisis posteriormente.

En primer lugar encontramos una innovación, que bien puede ser una idea, una práctica o un objeto que es apreciado como nuevo por un sujeto u otra unidad de adopción (Rogers, 2003). En este punto se aclara que, para el presente estudio, se acepta la noción de que cada uno de los temas publicados desde la cuenta de Twitter escogida para el análisis es una innovación, puesto que, como afirma Rogers (2003), el proceso de difusión incluye la "propagación de nuevas ideas, las planeadas y las espontáneas" (p. 6).

Luego, encontramos que para que dicha innovación llegue a sus destinatarios, esta debe utilizar ciertos canales de comunicación, como el del caso particular que nos ocupa: una plataforma de comunicación virtual y masiva de mensajes en 140 caracteres con múltiples posibilidades de interacción, a saber: retuits, publicación de imágenes y videos, favoritos y respuestas a los contenidos. Los retuits, por ejemplo, indican si un mensaje es propio o proviene de otra fuente, lo que permite compartir contenidos de otros *canales* haciendo referencia a su autor original. Aquí, la siguiente aclaración de Rogers (2003) puede arrojar más luz al respecto: "En adición a los medios masivos de

comunicación y la comunicación interpersonal, la comunicación interactiva vía Internet se ha vuelto más importante para la difusión de ciertas innovaciones en décadas recientes" (p. 18). De esta manera, se plantea desde la teoría que la comunicación de uno hacia muchos, en el ámbito digital, supone más retos en cuanto a creatividad y alcance de nuevas ideas o temáticas.

El tercero de los elementos es el factor *tiempo*, a partir del cual se demarca un espacio temporal donde se observa cómo la innovación es recibida, adoptada (o no), modificada o reemplazada por parte del público al que está dirigida. Esta dimensión temporal está involucrada en el proceso de difusión durante tres etapas: un primer momento en donde el individuo o unidad de adopción decide si adopta o rechaza la innovación, una segunda instancia donde se evalúa la precocidad/retraso en la adopción de la innovación, y por último, en la cuantificación del número de miembros de un sistema que adoptaron dicha innovación en un período de tiempo determinado (Rogers, 2003).

Para el caso de estudio, la dimensión *tiempo* será fundamental para entender el comportamiento de las temáticas observadas en las publicaciones. Dicho análisis permitirá establecer un mapa de temas a lo largo de un período de tiempo previamente establecido que ayudará a evidenciar una clara diferencia con otras innovaciones: "que los eventos noticiosos se difunden mucho más rápido" (Rogers, 2003, p. 75). En la presente investigación, el día y la hora es una adaptación del concepto de Rogers (2003), puesto que se toma como indicador el día de la semana y la hora en que fue publicado el tuit (es decir, el momento específico en que aparece en la línea de tiempo de la plataforma), observando el comportamiento temporal de los temas (su difusión) y no el rechazo o adopción de los mismos.

El cuarto elemento presente en la difusión de una innovación es el *sistema social*, definido por Rogers (2003) como "un conjunto de unidades interrelacionadas que se involucran en la solución de problemas para llegar a objetivos comunes" (p. 23).

Estos sistemas sociales son los encargados de reproducir, masificar o frenar el ascenso de una innovación en la curva de adopción. En ellos se encuentran las relaciones de sujetos que, en mayor o menor medida, impulsarán o rechazarán una *tecnología*. Este término, usualmente asociado con *hardware* o asuntos informáticos, también puede ser asumido en términos de comunicación y lenguaje: "como ejemplos tenemos una política filosófica como el Marxismo, una idea religiosa como el Cristianismo, un evento noticioso o una política municipal como la ordenanza de no fumar" (Rogers, 2003. p. 13).

Rogers (2003) estructura distintas categorías de adoptantes en un proceso de innovación dependiendo del punto en donde se ubique el sujeto en el área bajo la curva de adopción (Gráfico 12). Así, estos se enmarcan dentro de cinco clasificaciones:

I. Innovadores: Su interés en nuevas ideas los lleva fuera de su círculo local de redes con pares hacia relaciones sociales cosmopolitas. Necesitan habilidades para entender y aplicar conocimientos técnicos complejos, y deben ser capaces de hacer frente a un alto grado de incertidumbre sobre una innovación al momento en que la adoptan.
II. Primeros adoptantes: Están más integrados a su sistema social local que los innovadores. Mientras los innovadores son cosmopolitas, los primeros adoptantes son locales. Tienen el más alto nivel de liderazgo de opinión en la mayoría de los

sistemas. En cierto modo, los primeros adoptantes ponen su sello de aprobación a una nueva idea cuando la adoptan.

III. Mayoría precoz: Adoptan nuevas ideas justo antes que el miembro promedio del sistema. Interactúan frecuentemente con sus pares pero raramente tienen posiciones de liderazgo de opinión. Su proceso de decisión frente a una innovación es relativamente más largo que el de los innovadores y primeros adoptantes.

IV. Mayoría tardía: La adopción para este grupo puede ser tanto por una necesidad económica como el resultado de la presión de sus pares. Las innovaciones son abordadas escépticamente y con un aire cauteloso, y no adoptan hasta que la mayoría de los otros en su sistema lo han hecho.

V. Rezagados: Son los últimos en un sistema social en adoptar una innovación. El punto de referencia de los rezagados es el pasado, y sus decisiones suelen ser tomadas en términos de lo que se ha hecho previamente, interactuando primariamente con otros que también tienen valores relativamente tradicionales. (Rogers, 2003)

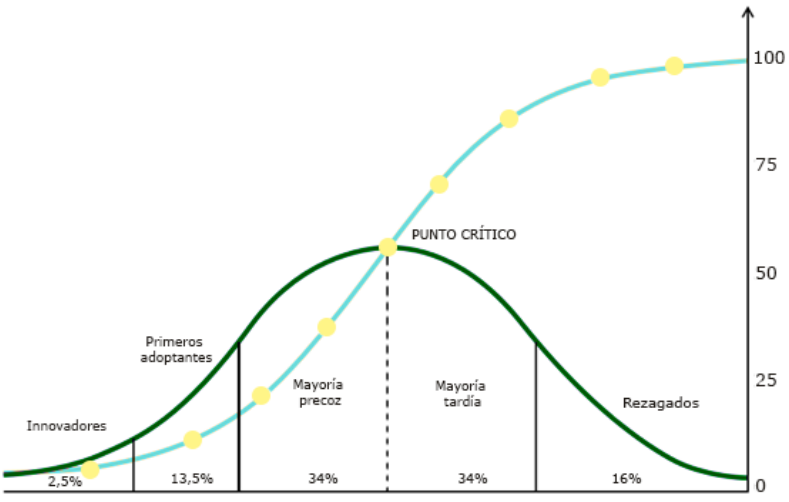


Gráfico 12. Tipología de adoptantes según Rogers (2003). Fuente: Elaboración propia.

8.1.1 Difusión de noticias

Entre los planteamientos de Rogers (2003) también podemos encontrar construcciones conceptuales específicas alrededor de los medios masivos, razón por la cual se toma su teoría como referente para el desarrollo teórico y metodológico de esta investigación.

En primer lugar, para Rogers (2003) las noticias pueden ser vistas desde el punto de vista de la *prominencia*, concepto que expresa el "grado en el que un evento noticioso es considerado como importante por los individuos" (p. 77). Por tal motivo, se considera que de los canales de comunicación esparcidos en todo el espectro mediático, el público escoge y estructura las noticias que consumen, llevándolos a "buscar más información al respecto y contarle a personas, usualmente extraños, sobre las noticias" (Rogers, 2003:78).

Rogers & Seidel (2002) aplican el concepto de *grado de difusión* con respecto a noticias en su estudio sobre el 11 de septiembre de 2011. Aquí, establecen que dicha medición implica el "porcentaje acumulado de individuos a los cuales ha llegado el evento noticioso a lo largo del tiempo" (Rogers & Seidel, 2002). Este lapso temporal que va desde que se da el suceso hasta su descubrimiento está, según Mayer et al. (1990), "correlacionado con cuántas personas se enteran del evento" (p. 114).

El estudio de la difusión de noticias, como bien explica Rogers (2000), "arroja luces sobre el complejo proceso mediante el cual los medios masivos transmiten nuevas historias a los individuos de la audiencia" (p. 562), lo que nos lleva a pensar en un campo de estudio que se desarrolla vertiginosamente a medida que emergen nuevas

formas de informar y comunicar. Dicho nivel de complejidad en esta área viene dado, además, por la "naturaleza impredecible de la ocurrencia de un evento noticioso, combinado con su rápida difusión" (Rogers, 2000, p. 563).

Todos los conceptos presentados en este apartado, aportan a una visión más clara desde la cual puede haber una aproximación analítica para describir cómo, a través de estas nuevas plataformas de comunicación, se están generando agendas noticiosas que integran nuevos conceptos y temas que, de una u otra forma, terminarán influyendo en la opinión pública.

8.2 Grandes cantidades de datos

El volumen de información que se maneja hoy en día alcanza dimensiones astronómicas. Como ejemplifican Cukier & Mayer-Schöenberger (2013), "si pusiéramos toda la información disponible hoy [unos 1,200 exabytes] en CDs, y los apiláramos, los CDs formarían cinco pilas separadas y todas llegarían a la luna".

Esta ilustración nos lleva a imaginarnos la cantidad de datos que hay alrededor nuestro, todos en variadas fuentes y formatos. En el área de las Humanidades, este crecimiento informático se ha visto reflejado en lo que hoy se conoce como *Digital Humanities*, pues, como asegura Romero (2014), estas "viven un tiempo de redefinición, cuestionamiento y puesta en valor en un contexto social muy mediatizado por las tecnologías digitales" (p. 19).

Según Burdick et al. (2012), "las primeras olas del acercamiento de las humanidades con las redes y la computación abrazaron los trabajos pioneros de 1940 y los modelos que inspiraron los proyectos de archivo en Oxford a comienzos de 1970" (p. 8). Posteriormente, las Humanidades siguieron abriéndose paso hacia lo digital como una manera de reforzar sus métodos tradicionales de trabajo y, en últimas, llegar a una mayor cantidad de público (Burdick et al., 2012).

Ejemplos puntuales de estas primeras *olas* en el ámbito fueron los "análisis textuales y la catalogación, el estudio de características lingüísticas, el énfasis en los apoyos pedagógicos y entornos de aprendizaje, y las preguntas de investigación dirigidas por el análisis de datos estructurados" (Burdick et al., 2012, p. 8). Luego, hemos asistido a la llegada de nuevas técnicas, tamaños muestrales y tecnologías para el estudio de la realidad. Aquí se incluye a la actual *big data*, con la cual ha sido posible una "expansión del conocimiento revelando y abriendo nuevas posibilidades, ideas, hechos y acciones que, previamente, estaban ocultas o eran inaccesibles por, en parte, el pequeño tamaño de los datos" (Kosciejew, 2013, p. 52).

Las aproximaciones a este fenómeno se han realizado desde distintas posiciones, por lo que para comprender mejor el contexto histórico, las ventajas y limitaciones de la *big data* es menester traer a colación variados planteamientos sobre el tema.

Para Zikopoulos (2012), "el impulso de Wallmart para usar Identificación por Radiofrecuencia [RFID, por sus siglas en inglés] para optimizar su cadena de suministro es una gran historia que ilustra el amanecer de la era *Big data*" (p. 4), pues esta

comprende información generada por computadores que a su vez puede ser recolectada y analizada (Zikopoulos, 2012).

De acuerdo con Verbeke et al. (2014), se podría definir *big data* como "volúmenes masivos de información tanto estructurada como no estructurada que es dinámica por naturaleza, que es recogida durante cierto periodo de tiempo y que requiere de métodos computacionales para extraer conocimiento de ella" (p. 2). Aquí podemos distinguir tres elementos claves de la *big data*: su tamaño, su momento de producción y su forma de recolección. Zikopoulos (2012) resumiría estas características en cuatro *V*s: "volumen, variedad, velocidad y veracidad" (p. 9), lo que permite, desde esta corriente teórica, establecer ciertas nociones sobre los componentes fundamentales de un corpus para ser rotulado como *big data*.

Para ejemplificar el concepto de *volumen* de la *big data*, el autor utiliza el siguiente enunciado: "Seis o siete años a partir de ahora, el número [de *zettabytes* en el mundo] se estima que estará alrededor de 35ZB, lo que equivale a cuatro trillones de iPods de 8GB" (Zikopoulos, 2012, p. 9). Ante esta realidad, es claro que, aunque la data que se analice en otros estudios no tenga estas dimensiones descomunales, la proporción en cuanto al fenómeno estudiado en casos particulares nos puede poner frente a un conjunto de datos que pertenezca a esta categoría.

La *variedad*, por otra parte, se refiere a la captura de "toda la data perteneciente a nuestro proceso de toma de decisiones" (Zikopoulos, 2012, p. 9), con el fin de entender mejor los datos y realizar un análisis que pueda abarcar un abanico de características

más amplio. Junto a la *velocidad*, estos dos componentes apuntan a recopilar información de forma rápida e integral, enriqueciendo de esta forma el análisis.

La *velocidad* tiene a su vez dos elementos importantes: el ritmo de llegada y la velocidad de procesamiento (Zikopoulos, 2012), lo cual deja una pregunta central en el asunto: "¿Cuánto tiempo le toma hacer algo al respecto [con esa data]?" (Zikopoulos, 2012, p. 10). Esta reflexión pone en evidencia una de las principales diferencias entre trabajar con *big data* y otros tamaños de datos, pues la cantidad de recursos humanos y tecnológicos suele ser menor en la primera (aunque más compleja).

Por último, encontramos la *veracidad*, cuya acepción "se refiere a la calidad e integridad de los datos" (Zikopoulos, 2012, p. 14). Esto incluye la cantidad de ruido, datos incompletos y el porcentaje de información que es útil para el análisis que se ha de proponer.

Otro concepto sobre *big data*, acuñado por Murphy & Barton (2014) es el siguiente: "*Big data* implica dibujar datos desde una potencial amplia variedad de conjuntos de datos que, históricamente, nunca tuvieron el propósito de ser combinados" (p. 8). Esta definición, que bien puede remitirnos a la *variedad* de la que nos habla Zikopoulos (2012), trae implícitamente dos de los factores más relevantes de las grandes cantidades de datos: su intencionalidad y utilidad.

Cuando se reúnen grandes cantidades de datos para una finalidad específica, estos suelen perderse en un mar de información sin pensar en usos secundarios. Mayer-Schönberger & Cukier (2013) conceptualizan este comportamiento en lo que ellos

llaman “tres vías poderosas para desencadenar el valor de opción de los datos: la reutilización básica, la fusión de conjuntos de datos y el hallazgo de combinaciones <<dos por uno>>” (p. 132), donde hasta los desechos digitales pueden ser objeto de estudio y beneficio, tanto en términos de conocimiento como económicos.

Nunan & Di Domenico (2013) enmarcan el concepto de *big data* en tres miradas que se asemejan a lo que hemos visto anteriormente, por lo que nos sirve a manera de resumen:

> La primera es una respuesta a problemas tecnológicos asociados al almacenamiento, seguridad y análisis de los siempre crecientes volúmenes de datos. (…) La segunda perspectiva se enfoca en el valor comercial que puede ser añadido a las organizaciones a través de la generación de *insights* más efectivos a partir de su data. (…) La tercera considera los amplios impactos sociales de la *big data*, particularmente las implicaciones para la privacidad personal (p. 3).

Gobble (2013) nos habla, por otra parte, de la gran promesa de la *big data*: “remodelar todo, desde el Gobierno y el desarrollo internacional, hasta cómo hacemos ciencias básicas” (p. 64). Dichas pretensiones auguran esfuerzos desde muchas disciplinas, incluyendo las Humanidades y la comunicación, para hacer el mejor uso posible de estas herramientas. Uno de estos, que podría ser visto en principio como una limitación emergida de las grandes cantidades de datos, es su almacenamiento. “En la era de la *big data*, las empresas están reuniendo información a la velocidad de la luz y las estrategias de almacenamiento tradicionales no son suficientes” (Collett, 2013, p. 16). Por este motivo, más empresas están buscando alternativas para contener sus datos (como la

nube), lo que ha generado incluso que se subcontraten agentes externos para manejar esta vasta cantidad de información.

Esto se debe principalmente a que, como reflexionan Marozzo et al. (2013), "los repositorios digitales de datos son más y más masivos, complejos y ubicuos. Por lo tanto, necesitamos técnicas inteligentes de análisis de datos y arquitecturas escalables para extraer información útil eficientemente" (p. 182). Así, no se trata solo de saber qué datos recopilar para qué propósito o cómo obtenerlos, sino también cómo almacenarlos y procesarlos para sacar el mayor provecho posible.

Volviendo sobre Gobble (2013), es importante resaltar que "la data es *big data* cuando es muy grande para ser manejada por sistemas convencionales" (p. 64), lo que implica que usarla, "ya sea para dirigir una innovación o reformar el proceso de innovación, no será fácil" (Gobble, 2013, p. 65). A esto se suma la cuestión de la privacidad, otro de los puntos claves en la administración de la *big data*.

Como hemos visto en este apartado, son muchas las ventajas cuando hablamos de profundidad y amplitud al utilizar *big data*, pero al mismo tiempo este nuevo acontecer de la ciencia trae limitaciones en su interior, ya sea por procesamiento, almacenamiento o privacidad. Con los desarrollos tecnológicos que han surgido en las últimas décadas, estos inconvenientes han ido resolviéndose eficientemente, con lo cual hoy podemos depositar un buen grado de confiabilidad en los resultados obtenidos a partir de estas diversas técnicas de análisis.

Para concluir, un pensamiento sobre el futuro que sirve como abrebocas para el siguiente punto en nuestro recorrido teórico: "Hoy por hoy, en las primeras fases de la era de los datos masivos, las ideas y las capacidades parecen atesorar todo el valor, pero a la larga lo más valioso serán los propios datos: podremos hacer más cosas con la información" (Mayer-Schönberger & Cukier, 2013, p. 168).

8.2.1 Tres cambios de mentalidad: rupturas en el paradigma según Mayer-Schönberger & Cukier

Cuando nos aproximamos desde la academia al término *big data*, encontramos posiciones de todo tipo frente al tema: unas que ven más sus ventajas y otras que hacen hincapié en los riesgos que implica *confiar* en las máquinas.

En este inciso resaltaremos brevemente tres cambios importantes de mentalidad que hacen referencia al pensamiento de datos masivos, cuya influencia ha traspasado ya disciplinas y campos de estudio. Como afirman Boyd & Crawford (2011), son muchas las aristas del asunto: "Es indispensable que comencemos a hacernos preguntas críticas sobre lo que significa la *big data*, quién tiene acceso a ella, cómo se despliega y con qué fines" (p. 2).

Para Mayer-Schönberger & Cukier (2013) estos tres giros en el pensamiento, "al estar interrelacionados se refuerzan entre sí. El primero es la capacidad de analizar enormes cantidades de datos de información sobre un tema dado, en lugar de verse uno forzado a conformarse con conjuntos más pequeños" (p. 33). En este primer escenario se

encuentran las empresas que describen los autores como aquellas para las que "primero están los datos. Se trata de las compañías que disponen de los datos o, cuando menos, del acceso a ellos" (Mayer-Schönberger & Cukier, 2013, p. 157). Este estadio integra a todas las organizaciones, empresas y entidades que pueden tener acceso a los datos pero que, en pocas palabras, no tienen el conocimiento ni la capacidad para obtener lo que esperan y *usarlo*.

En segundo lugar, continuando con las transformaciones del paradigma, encontramos la "disposición a aceptar la imprecisión y el desorden –muy del mundo real– de los datos, en lugar de anhelar la exactitud" (Mayer-Schönberger & Cukier, 2013, p. 33). Pero no debemos asustarnos. Como ha aumentado la cantidad de datos, este margen de error (presente también en los trabajos con muestras pequeñas) puede ser minimizado. De aquí que "cualquier medición aislada puede ser incorrecta, pero la agregación de tantas mediciones ofrecerá una imagen mucho más exhaustiva, porque este conjunto, al consistir en más puntos de datos, es más valioso" (Mayer-Schönberger & Cukier, 2013, p. 51).

Por tratarse de enormes cantidades de datos en algunos casos, estos puntos de los que hablan los autores vienen a configurar relaciones que nos permiten observar más y mejor fenómenos subyacentes. Es en este momento donde irrumpen las *correlaciones*. Citando a Mayer-Schönberger & Cukier (2013), "el tercer cambio pasa por empezar a respetar las correlaciones, en vez de buscar constantemente la elusiva causalidad" (p. 33). Aunque parte importante de muchos estudios en la actualidad, la causalidad no debe ser el fin último de toda observación (si bien es uno de los más relevantes), pues una correlación puede cuantificar "la relación estadística entre dos valores de datos. Una

correlación fuerte significa que, cuando cambia uno de los valores de datos, es altamente probable que cambie también el otro" (Mayer-Schönberger & Cukier, 2013, p. 72).

Para finalizar, es importante hacer la siguiente anotación: la causalidad no lo es todo. Mayer-Schönberger & Cukier (2013) lo explicitan cuando le quitan el aura de misticismo a esta idea:

> Al contrario de la sabiduría convencional, esta intuición humana de la causalidad no acrecienta nuestra comprensión del mundo. En muchos casos, es poco más que un atajo cognitivo que nos depara una ilusión de percepción, cuando en realidad nos deja la inopia respecto al mundo que nos rodea (p. 85).

8.3 El lugar de la teoría: una discusión abierta

La discusión sobre la necesidad o no de una teoría para aprehender lo que sucede a nuestro alrededor se ha tornado cada vez más polarizada. A pesar de que esta disyuntiva se remonta tiempo atrás, hubo un momento crucial en los años recientes que reactivó dicho enfrentamiento: la publicación del artículo titulado *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete* por Chris Anderson en el año 2008.

En el texto divulgado por la revista *Wired*, Anderson (2008) argumenta que hace décadas solo teníamos modelos, "desde ecuaciones cosmológicas hasta teorías del comportamiento humano, capaces de (…), imperfectamente, explicar el mundo

alrededor nuestro. Hasta ahora". Partiendo de esto, describe cómo a lo largo de los años la Internet y los computadores han facilitado tanto las tareas, que hoy vemos cómo "Google y compañías similares se tamizan a través de la época más medida de la historia, tratando sus *corpus* masivos como un laboratorio de la condición humana. Son los hijos de la *Era Petabyte*" (Anderson, 2008).

Una de las declaraciones más atrevidas según la crítica (a la que retomaremos más adelante) fue la siguiente: "Las correlaciones reemplazan a la causalidad, y la ciencia puede avanzar incluso sin modelos coherentes, teorías unificadas o cualquier mecanismo de explicación" (Anderson, 2008). Con esto, el autor enfatiza en que hemos entrado en un nuevo momento histórico para el análisis de datos, inclusive para la humanidad.

Antes de continuar exponiendo otras posiciones, volvamos sobre nuestros pasos hasta la forma clásica de investigación, con el fin de entender cómo es que muchos autores alrededor del mundo han ubicado la teoría en un segundo plano. El Gráfico 13 de Wanamaker & Bean (2013) expone de manera clara la estructura tradicional:

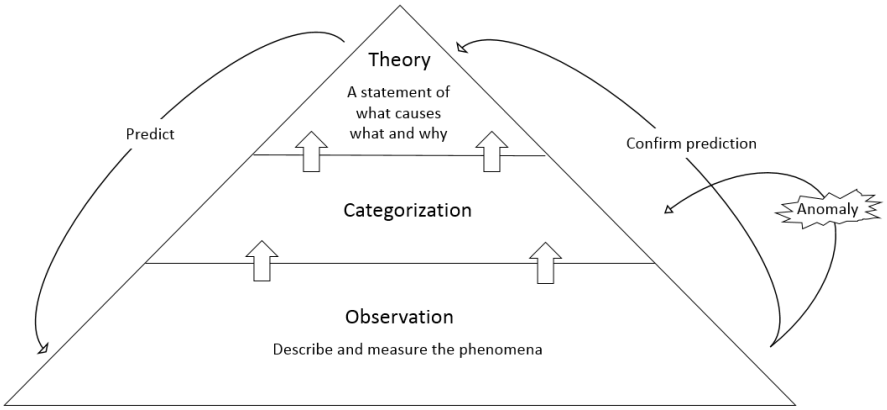


Gráfico 13. Modelo clásico del desarrollo de teorías. Fuente: Wanamaker & Bean (2003).

Este gráfico nos muestra la forma en que los investigadores han abordado, durante años, los fenómenos. De manera más puntual, esto es lo que dicen los autores al respecto:

> Primero ellos observan, describiendo y midiendo el fenómeno cuidadosamente. Luego, agrupan estas observaciones en distintas categorías, generalmente buscando similitudes y diferencias entre atributos. Por último, los investigadores desarrollan una teoría que explica cómo cierto conjunto de atributos lleva a cierto resultado (Wanamaker & Bean, 2003).

En este modo de ver las cosas, la relación *causa-efecto* cumple un papel fundamental, pues guía los esfuerzos científicos hacia ese descubrimiento. Pero, tal y como se pregunta White (2013), "¿qué pasaría si un mundo que se basa en causa y efecto se deshace de las causas?". Este cuestionamiento es, básicamente, lo que ha generado luchas entre académicos y científicos por una u otra acepción.

Graham (2012) ajusta este pensamiento y se pone de alguna manera en el centro de la discusión –inclinándose más hacia la tradición científica–, al proponer que "la *big data* es indudablemente útil para dirigir y superar muchos asuntos importantes que enfrenta la sociedad. Pero debemos asegurarnos de que no estamos siendo seducidos por las promesas de la *big data* de hacer la teoría innecesaria". El autor además agrega que "podríamos llegar al punto en que las suficientes cantidades de *big data* puedan ser cosechadas para responder todas las preguntas sociales que más nos preocupan, aunque [esto] me genera dudas" (Graham, 2012).

Enfoques como el anterior van de un lado a otro en el espectro científico. Mientras algunos arguyen que la teoría es siempre necesaria, otros señalan que solo en algunos casos debe usarse o, incluso, que ya no es necesaria. Felten (2008) aclara, frente a esta situación, que "en un mundo con más y más datos, con mejores y mejores herramientas para hallar correlaciones, necesitamos el método científico más que nunca". En consonancia, el autor expone el ejemplo de la teoría en física, que sería "más útil si tuvieran más datos. Y lo mismo es verdad para teoría científica en general: teoría y experimentación avanzan en tándem, con adelantos en uno se crean oportunidades para el otro" (Felten, 2008).

Quienes sugieren que la teoría es innecesaria explican que "los motores de predicción impulsados por datos generarán predicciones que obviarán la necesidad de un análisis guiado por la teoría" (Wanamaker & Bean, 2013), pero al mismo tiempo aquellos que no rechazan ni lo uno ni lo otro, buscan la complementariedad de los métodos.

Knapp & Michaels (1982) ya habían discutido en su época sobre esta cuestión, cuando aseguraron que la *teoría* es "el nombre de todas esas formas en que las personas han intentado pararse por fuera de la práctica con el fin de gobernarla desde el exterior" (p. 742). Acentuando ese detalle, los autores terminan su diatriba con una tesis que, para muchos, podría sonar extremista: "Nadie puede llegar a una posición fuera de la práctica, los teóricos deberían dejar de tratar y la empresa teórica debería llegar a su fin" (Knapp & Michaels, 1982, p. 742).

Con respecto al párrafo anterior, se puede decir entonces que estamos viviendo un tiempo esencial para el desarrollo de nuevas formas de hacer ciencia. Wanamaker &

Bean (2013) nos dejan, para culminar, con esta idea: “el uso inadecuado de la *big data* nos podría llevar a conclusiones incorrectas y oportunidades perdidas. Usada correctamente, en efecto tiene tremendo potencial para revolucionar (…) haciendo los análisis más asequibles y accesibles”. Queda entonces abierta la pregunta: ¿hasta dónde llegará esta dualidad?

8.4 Métodos computacionales para el análisis de las grandes cantidades de datos

Los métodos computacionales son herramientas indispensables para avanzar en el conocimiento de nuevos objetos de estudio y tamaños de muestra que, incluso, pueden acercarse al *todo*, y no solo una parte de la data (Mayer-Schönberger & Cukier, 2013). Esta vasta cantidad de información puesta en forma de datos nos lleva a explorar otras formas de hacer ciencia y, por ende, de optimizar los procesos que anteriormente requerían de mayores esfuerzos humanos y económicos. Por tal motivo, en este apartado se detallan algunos de los más importantes métodos computacionales para el análisis de grandes cantidades de datos, a saber: análisis automatizado de contenido, *Data mining*, *Machine learning*, Modelamiento de temas o *topic modeling* y el análisis de sentimiento automatizado.

8.4.1 Análisis automatizado de contenido

Para conceptualizar alrededor de este tema, en primer lugar, consideraremos la siguiente noción de análisis de contenido esbozada por Leetaru (2012): “es cualquier tipo de

análisis que pretende derivar nuevos significados del contenido existente" (p. 2). Aquí podemos destacar dos componentes principales: los significados y el contenido como tal, que se vinculan entre sí a manera de proceso para constituir un análisis de contenido eficaz.

En ese sentido, Krippendorff (2004) puntualiza que "en el análisis de contenido los investigadores examinan artefactos de la comunicación social. Típicamente, estos son documentos escritos o transcripciones de comunicación verbal grabada" (p. 240), por lo que esta técnica ha sido utilizada desde hace décadas para tratar con textos de procedencias diferentes.

A lo largo de su historia, el análisis de contenido se ha servido de otras técnicas que mejoran su alcance. Además, como explican White & Marsh (2006), este se ha venido aplicando en "marcos de investigación cuantitativos, cualitativos y mixtos" (p. 2), mientras "emplea un amplio rango de técnicas analíticas para generar descubrimientos y ponerlos en contexto" (White & Marsh, 2006, p. 22). Los avances tecnológicos para el procesamiento de datos mencionados en el ítem sobre *big data*, pertenecen también –y fundamentalmente– a las herramientas que coadyuvan a la extracción de conocimiento a través del análisis de contenido, motivo que nos lleva a bosquejar, acto seguido, sus beneficios y aportes al campo.

Según Lewis, Zamith & Hermida (2013), "los métodos computacionales, en teoría, ofrecen el potencial de superar algunas limitaciones de muestreo y codificación de los análisis de contenido tradicionales" (p. 38), planteamiento que se complementaría con tres elementos que, para Leetaru (2012), son el núcleo ventajoso frente a los métodos

"humanos", a saber: confiabilidad, reproductibilidad y escala (p. 2). Aquí vale la pena hacer énfasis en que lo estricto de las normas en las máquinas, lejos de ser una limitación "puede, en realidad, favorecer el análisis de contenido, en el cual los codificadores humanos son muchas veces incapaces de dejar a un lado las creencias personales que pueden sesgar su interpretación de un texto" (Leetaru, 2012, p. 70).

Este tipo de consideraciones generan reflexiones, tanto positivas como negativas, con respecto a la fiabilidad y validez de dichos análisis computarizados. Refiriéndose a estas inquietudes, Hardwood & Garry (2003) recalcan que en el caso de "no ser satisfechas, pueden suscitar dudas en cuanto a la generalización de los resultados" (p. 493).

En contraposición a la anotación anterior, West (2001) defiende la utilización de métodos computacionales como sigue:

> El cambio más prometedor en el análisis de contenido es la habilidad de buscar cantidades masivas de material instantáneamente. Mientras esto podría reducir la profundidad del análisis, incrementa dramáticamente la amplitud de un estudio. Por sí mismo, esto es suficiente para alabar el valor del computador (…) (p. 5).

Con este *valor* computacional mencionado anteriormente, se provee de *extremidades* más amplias al científico de datos, lo que impulsa trabajos de investigación cada vez más diversos. Por ejemplo, Cheng et. al. (2008) destacan, entre otros, el incremento en el uso de "la lingüística computacional (…) aplicada a dominios como los de la captura de datos de inteligencia, traducción con máquinas, análisis de contenido automatizado y la indexación y recuperación de bases de datos completas" (p. 2).

Actualmente, hemos visto cómo estas técnicas se han consolidado como una herramienta fundamental para estudiar, entre otras cosas, miles de mensajes publicados por medios masivos de difusión de noticias, publicaciones a través de medios sociales y todo tipo de data a nuestro alcance. Somos testigos de la abundancia y, por qué no, de una *inundación de datos* que cubre todas las esferas de nuestras vidas. Esto ha sido aprovechado, por ejemplo, para desarrollar aproximaciones como las de Fernández et al. (2012), Rich (2012) y Hanna (2013), en las cuales el análisis de contenido automatizado ha facilitado (cuando no *permitido*) tareas de tales magnitudes.

Nuevas formas de pensar el mundo se vislumbran, así, desde la ventana del quehacer científico. Es pasar de *contemplar* aquellas ideas que se veían como irrealizables para *ejecutar* lo que, en algún punto, fue inconcebible. Wing (2006) lo sintetiza así: "Los métodos computacionales nos dan el coraje para resolver problemas y diseñar sistemas que ninguno de nosotros hubiese sido capaz de abordar por sí solo" (p. 33).

Con el paso del tiempo seguiremos descubriendo nuevos sentidos y formas al interior de los textos que, de otra manera, quedarían escondidos bajo la superficie. Cada autor deja huellas en ellos, y a fin de poder encontrarlas, es necesario aumentar nuestra capacidad de búsqueda y recuperación masiva de dichos elementos.

8.4.2 *Data mining* y *machine learning*

8.4.2.1 *Data mining*

En principio, examinaremos algunas acepciones sobre *data mining*. Dhar (2013) argumenta que el "campo del *data mining* floreció a principios de 1990 a medida que la tecnología relacional de bases de datos maduró y los procesos de negocio crecieron en automatización" (p. 67), fomentando así la creación de *software* orientado a aprovechar los datos sobre comportamiento y transacciones, con el fin de predecir y planear de manera más acertada (Dhar, 2013).

Han & Kamber (2006) construyeron una definición de *data mining* (o en español, *minería de datos*) que engloba lo esencial del asunto: "*data mining* es extraer o hacer *minería* de conocimiento de grandes cantidades de datos" (p. 5). Como veremos también en el *machine learning*, este subcampo de las ciencias computacionales cuenta con varias técnicas y fases primordiales, aplicables a estudios en todo tipo de experiencias.

Siguiendo la línea de Han & Kamber (2006), lo que ellos referencian como *knowlegde discovery from data* (término que se ha usado a la par de *data mining*) se puede dividir en siete momentos, a saber:

1. Limpieza de datos: remover *ruido* y datos inconsistentes.
2. Integración de los datos: donde múltiples fuentes de datos pueden ser combinadas.

3. Selección de datos: los datos relevantes para la tarea de análisis son recuperados de la base de datos.
4. Transformación de los datos: la data es transformada o consolidada en formas apropiadas para el *data mining*, realizando operaciones de resumen o agregación, por ejemplo.
5. Minería de datos: proceso esencial donde métodos de inteligencia son aplicados para extraer patrones de los datos.
6. Evaluación de patrones: identificar patrones verdaderamente interesantes sobre la base de ciertas medidas de interés.
7. Presentación del conocimiento: las técnicas de visualización y presentación del conocimiento son usadas para presentar el conocimiento *minado* al usuario (Han & Kamber, 2006, p. 7).

Cabe resaltar que desde esta perspectiva se identifica al *data mining* como solo uno de varios *momentos* –si bien de suma importancia– para el conocimiento a partir de los datos, lo que no resta trascendencia a su posición como un instrumento de análisis eficiente.

Autores como Kaur & Singh (2011) señalan que, "básicamente, la *minería de datos* es el análisis de conjuntos de datos observacionales para encontrar asociaciones inesperadas y para resumir los datos en nuevas formas que sean claras y útiles para el dueño de la data" (p. 336). En esta definición encontramos, por una parte, la referencia a datos *observacionales*, que para Hand, Mannila & Smyth (2001) se relaciona con el hecho de que la "*minería de datos* típicamente trata con datos que ya han sido recopilados para algún propósito distinto al del análisis de *minería de datos*" (p. 6). Por otro lado, están las asociaciones *inesperadas* que, en estos tiempos, han despertado el

interés de estudios con significativas aplicaciones prácticas, ya sea en la academia, lo público o la empresa privada [ver, por ejemplo, los casos expuestos por Mayer-Schönberger & Cukier (2013) en el capítulo titulado *Correlaciones*]. Se han desarrollado, además, otros ejercicios con enfoques noticiosos (Murata, 2008), de género, edad y *sentimiento* (Thelwall, Wilkinson & Uppal, 2010) y lingüístico-geográficos (Mocanu et al., 2013) que aportan al avance en este campo.

Hand, Mannila & Smyth (2001) comentan que "este proceso de adquisición de datos digitales y tecnologías de almacenamiento han resultado en el crecimiento de grandes bases de datos" (p. 5). Además, sugieren que se ha expandido de este modo:

> En todas las áreas del quehacer humano, desde las *mundanas* (como los datos de transacciones en el supermercado, registros de uso de tarjetas de crédito, detalles de llamadas telefónicas y estadísticas gubernamentales) hasta las más *exóticas* (imágenes de cuerpos astronómicos, bases de datos moleculares y registros médicos) (Hand, Mannila & Smyth, 2001, pp. 5-6).

Para Kalina (2013), una de las principales diferencias entre la estadística clásica y la *minería de datos* es que esta última "no tiene la ambición de generalizar sus resultados más allá de la data resumida en una base de datos" (p. 10). Lo anterior aportaría, según la autora, una perspectiva desde donde los datos hablan de ese caso particular, y donde "no se asume usualmente una muestra de cierta población y los datos se analizan e interpretan como si constituyeran el total de la población" (Kalina, 2013, p. 10). Esto es, ir más profundo –y abarcar más– sin que nuestra principal motivación sea la generalización total de los descubrimientos.

Pero debemos tener cuidado y evitar pensar que este tipo de técnicas reemplazan totalmente nuestra labor como investigadores. He aquí, a modo de cierre, una apreciación provocadora de Two Crows Corporation (1999):

> D*ata mining* es una herramienta, no una varita mágica. No va a sentarse frente a su base de datos viendo qué sucede y le va a mandar un *e-mail* para atraer su atención cuando vea un patrón interesante. No elimina la necesidad de conocer su negocio, de comprender sus datos o entender los métodos analíticos (p. 1).

8.4.2.2 *Machine learning*

En palabras de Blum (2003), "el área de *machine learning* trata con el diseño de programas que pueden aprender reglas a partir de datos, adaptarse a cambios y mejorar el rendimiento con la experiencia" (p. 1). Con dicha técnica se reducen tiempos y costos en los procesos, y también se obtienen resultados fiables a través del *aprendizaje* que realiza la máquina al agregársele parámetros y configuraciones específicas para cada estudio.

De acuerdo con Mitchell (1997), el *machine learning* "es inherentemente un campo multidisciplinario. Se basa en resultados de inteligencia artificial, probabilidad y estadística, teoría de la complejidad computacional, teoría del control, teoría de la información, filosofía, psicología, neurobiología y otros campos" (p. 2). Al conocer los

campos de estudio implicados en el *maching learning*, permanecen dos preguntas concisas pero clave: a) ¿Cómo funciona? b) ¿De qué pasos y elementos se compone?

El gráfico que se encuentra a continuación ilustra las distintas etapas de esta forma de trabajo. En la conceptualización de Lula & Wójcik (2011), el flujo del proceso se representaría como aparece en el Gráfico 14:

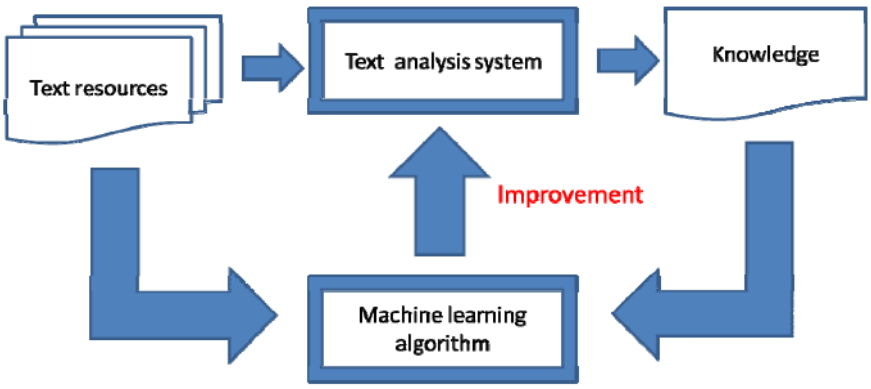


Gráfico 14. Proceso del *machine learning*. Fuente: Lula & Wójcik (2011).

Desglosando el gráfico, tenemos en primera instancia unos *recursos de texto* que son cargados al *sistema de análisis*, ya sea a través de una interfaz o una línea de códigos. Este sistema permite a su vez generar un *conocimiento* basado en el corpus introducido previamente, lo que se convierte en un *algoritmo* con el cual la máquina *aprende* las reglas subyacentes en dichos documentos. Luego de este paso, estas reglas o patrones son ingresados nuevamente al *sistema de análisis* para mejorar progresivamente los resultados, forjando un análisis cada vez más robusto.

Domingos (2012) nos da ejemplos de los ámbitos en donde esta técnica se está usando actualmente: "búsquedas en la Web, filtros de *spam*, sistemas de recomendaciones, ubicación de anuncios publicitarios, puntuaciones para créditos, detección de fraude,

comercio de acciones, diseño de fármacos y muchas otras" (p. 78). A esto se puede agregar uno de los objetivos principales que Domingos (2012) asigna al *machine learning*: "generalizar más allá de los ejemplos que se utilizan para entrenar [a la máquina]" (p. 80).

Dietterich (2003) plantea con respecto a este uso del *machine learning*, que existen cuatro tipos de situaciones que estimulan al uso de esta técnica. El primero emerge cuando "hay problemas para los cuales no hay expertos [en resolverlos]" (Dietterich, 2003, p. 1), y por tales razones se utiliza *machine learning* con el objetivo de estudiar la data existente y que así la máquina pueda "aprender normas de predicción" (Dietterich, 2003, p. 1).

La segunda de las dificultades descritas se da cuando los seres humanos no son capaces de explicar de qué manera realizan un proceso o tarea de análisis de acuerdo a su experticia (Dietterich, 2003). Este es el caso, según Dieterrich (2003), "de muchas tareas de percepción, como reconocimiento de voz, reconocimiento de escritura a mano y comprensión del lenguaje" (p. 1). Para resolverlo, el *machine learning* puede proveer, a través del aprendizaje, algoritmos que asignen de manera adecuada cada *input* al correspondiente *output* (Dietterich, 2003).

En tercer lugar encontramos una de las características más importantes de la vida en general, y del mundo tecnológico y científico en particular: cambia constantemente. En el caso financiero, por ejemplo, los "comportamientos cambian frecuentemente, así que incluso si un programador construye un buen programa predictivo, este tendría que ser reescrito también frecuentemente" (Dietterich, 2003, p. 1). A través de las *reglas de*

predicción mencionadas anteriormente, un programa de *machine learning* "puede aliviar al programador de esta carga" (p.1).

Por último, tenemos una tarea que tiene que ver más con la personalización y la experiencia de usuario. Para algunos trabajos no tenemos un algoritmo único, como alguno que permita "diferenciar *spam e-mails* de correos electrónicos legítimos" (Alpaydin, 2010, p. 1), pero lo que "nos falta en conocimiento, lo compensamos con datos" (Alpaydin, 2010, p. 1). Precisamente, Dietterich (2003) sugiere el ejemplo de los filtros para mensajes de correo electrónico no deseados dado que, como cada uno de los usuarios tiene una dirección de correo distinta, sería una labor monumental configurarlas una a una. Con la ayuda del *machine learning* se puede "aprender cuáles mensajes rechaza el usuario y mantener las reglas de filtrado automáticamente" (Dietterich, 2003, p. 1).

A la par de estas observaciones, Murphy (2012) divide el *machine learning* en dos grandes grupos. En primer lugar, se encuentra el "*predictivo* o de *aprendizaje supervisado*, donde el objetivo es *aprender* un mapeo de los *Inputs X* a los *Outputs Y*" (p. 2). Por otra parte, se encuentra el "*descriptivo* o *aprendizaje sin supervisión* (…), donde solo se dan los *Inputs* y se tiene como objetivo encontrar *patrones interesantes* en los datos" (p. 2). Con base en la anterior clasificación, podríamos decir que la ciencia hoy en día toma distintas posiciones frente al *machine learning*, ajustando sus herramientas, *software* y metodologías a los resultados esperados (o inesperados) de sus investigaciones, dependiendo de su interés.

Entre los estudios que utilizan esta técnica pueden nombrarse investigaciones como las de Pennacchiotti & Popescu (2011), quienes trabajan en las inmediaciones del *social media* y el *machine learning* para detectar atributos como la inclinación política, la etnia o afinidades de negocio, o como el de Téllez, Montes & Villasenor (2009) quienes aportan conocimientos metodológicos para recopilar y analizar datos sobre reportes noticiosos a través del *machine learning*.

Rätsch (2004) va un poco más allá y expone que "parece gratificante y casi obligatorio para los científicos computacionales e ingenieros aprender cómo y dónde el *machine learning* puede ayudarlos a automatizar tareas o proveer predicciones donde los humanos tienen dificultades para comprender grandes cantidades de datos" (pp. 4-5). Teniendo en el horizonte este pensamiento, se anota una interrogante para futuras (y más específicas) indagaciones: ¿dónde están situados, en todo este proceso, los científicos sociales?

8.4.2.2.1 Modelamiento de temas o *topic modeling*

El modelamiento de temas o *topic modeling* se desprende de lo que en el apartado anterior se ha puntualizado como *machine learning*. Para definirlo, Arora et al. (2013) precisan que es un "método que aprende estructuras temáticas de grandes colecciones de documentos sin supervisión humana" (p. 1). Esta *habilidad computacional* permite abarcar desde unos cuantos cientos hasta millones de documentos sin tener que ir uno por uno para clasificarlos, asignándoseles temas a los cuales –muy probablemente– pertenecen. Blei (2012) centra su atención en que "a medida que más y más textos están

disponibles en línea, sencillamente no tenemos el poder humano para leer y estudiarlos" (p. 77), lo que nos empuja a encontrar nuevas maneras de procesarlos en este entorno de constante crecimiento.

Para aclarar un poco más este concepto, se toma en cuenta la siguiente definición de Blei (2012): "Los algoritmos de *modelamiento de temas* son métodos estadísticos que analizan palabras de textos originales para descubrir los temas que viajan a través de ellos, cómo esos temas están conectados entre sí, y cómo cambian a través del tiempo" (p. 77). Un ejemplo claro, expuesto por el mismo Blei (2012), es el Gráfico 15, un análisis realizado al *Yale Law Journal*:

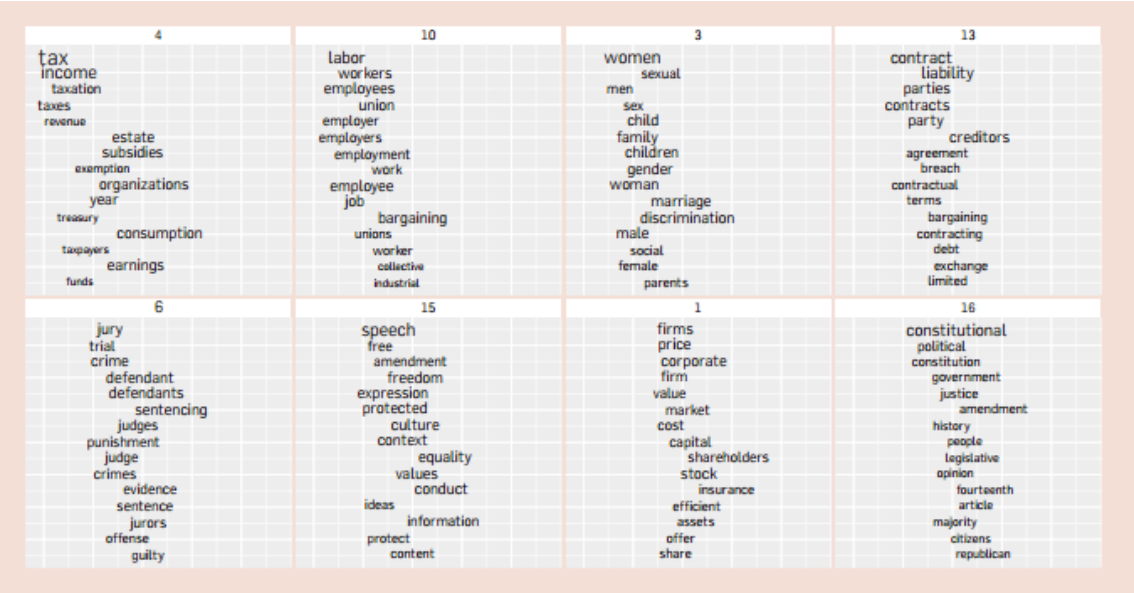


Gráfico 15. Ejemplo de modelamiento de temas. Fuente: Blei (2012).

Esta metodología de corte *inductivo* parte de la misma data para obtener los temas (nombrados a posteriori por el investigador) en los que luego serán agrupados los documentos (o colecciones de estos). Por esta razón, lo que vemos en el Gráfico 15

corresponde a varios listados de palabras frecuentes agrupados en cuadros numerados, cada uno de representando un tema encontrado en el análisis. En este ejemplo en particular, las palabras que más específicas dentro de cada tema son aquellas que más se acercan al eje X del centro de los cuadros (Blei, 2012).

Para llevar a cabo esta tarea, se seleccionan automáticamente palabras del corpus que aparecen frecuentemente, lo que indicaría que podrían pertenecer o no a cierto tema y, *observando* la presencia de estas en los documentos, podemos clasificarlos sin intervención humana.

"Por otra parte, la inferencia posterior a nivel de los documentos es útil para la recuperación de información, clasificación y búsqueda por temas" (Blei & Lafferty, 2006, p. 1), lo que diversifica el uso que puede dársele a esta técnica y, por ende, los resultados arrojados por ella.

Un ejemplo (no digital) de cómo funciona el *topic modeling* puede ser trabajar sobre un texto y resaltar palabras claves con diferentes colores de acuerdo al tema: al final, si reunimos esos listados de palabras por colores, cada uno representaría un *tema* distinto de los que se encuentran en dicho texto (Brett, 2012).

Con la finalidad de sintetizar los elementos constituyentes del *modelamiento de temas*, Brett (2012) los agrupa en tres necesidades básicas para llevarlo a cabo: en primer lugar, "un corpus, preferiblemente grande", añadiendo que considera pertinente "como mínimo estar en los *cientos* sino en un mínimo de 1.000 documentos" (Brett, 2012) para poder realizar la tarea adecuadamente. Esto podría incluir "cambiar el texto de frases

legibles para humanos a una cadena de palabras quitando la puntuación y las mayúsculas" (Brett, 2012), o en otras palabras, *limpiando* el corpus de ciertas marcas o caracteres que podrían entorpecer el análisis.

En segundo lugar, Brett (2012) señala como útil el conocimiento previo del corpus –al menos en una manera superficial– que permita saber hasta cierto punto qué deberíamos tener en él. La autora recalca que "el *topic modeling* no es, en ninguna circunstancia, una ciencia exacta" (Brett, 2012), por lo que se recomienda estar familiarizados con la data que pretendemos analizar.

Por último, pero no menos importante, Brett (2012) destaca la necesidad de "una herramienta para realizar el *modelamiento de temas*". De acuerdo a las particularidades de cada estudio, se enuncian herramientas útiles que se acomoden de la mejor manera al trabajo. Por un lado se encuentra MALLET (uno de los más difundidos, que puede descargase gratuitamente desde http://mallet.cs.umass.edu/), que como lo describen sus creadores, es "un paquete basado en Java para el procesamiento estadístico de lenguaje natural, clasificación de documentos, *clustering*, modelamiento de temas, extracción de información y otras aplicaciones de *machine learning* al texto" (McCallum, 2002).

Para Blei (2012), el modelo más simple de *topic modeling* es el *Latent Dirichlet Allocation* (LDA, por sus iniciales en inglés) que tiene dos principios básicos: "1. Hay un número fijo de patrones en el uso de palabras, grupos de términos que tienden a aparecer juntos en los documentos. Llamémoslos *temas*. 2. Cada documento del corpus exhibe los temas en mayor o menor grado" (Blei, 2012, p. 9), y que responde a una

pregunta sencilla pero ilustrativa del proceso: "¿Cuál es la estructura de temas oculta más probable que generó los documentos que observo?" (Blei, 2012, p. 9).

También existen interfaces como el Standford Topic Modeling Toolbox (http://nlp.stanford.edu/software/tmt/tmt-0.4/) o el Topic Modeling Tool (https://code.google.com/p/topic-modeling-tool/), instrumentos con los cuales se puede realizar *modelamiento de temas* y que, a diferencia de MALLET, proveen de un entorno gráfico más *amigable* para llevar a cabo los procesos sin necesidad de conocer la materia en su totalidad.

Estudios como los de Zhao et al. (2001) y Panaccione & Foltz (2009) muestran acercamientos al *topic modeling* desde los medios y otras plataformas de comunicación, indicando unos caminos (entre muchos otros) que se han tomado con el modelamiento de temas.

A pesar de todas las ventajas y avances enunciados en párrafos anteriores, aún se discute sobre la necesidad de ir más allá en la comprobación y estandarización de los métodos, tema que queda en el tintero para continuar en su desarrollo. Por este motivo, Wallach et. al (2009) añaden con firmeza que existe una "necesidad de un método universal que mida la capacidad de generalización del modelamiento de temas en una manera exacta, computacionalmente eficiente e independiente de cualquier aplicación específica" (p. 1).

8.4.3 Análisis de sentimiento automatizado

Antes de comenzar a hablar de *análisis de sentimiento automatizado*, vale la pena distinguir el siguiente pensamiento de Pang & Lee (2008): "*lo que otras personas piensan* siempre ha sido una parte importante para la mayoría de nosotros" (p. 1). En ese sentido, la curiosidad que despiertan las opiniones y emociones de otros individuos puede considerarse como punto de referencia (si bien en principio no *científico*) que se ha venido transformado a lo largo del tiempo para convertirse en un interés académico estudiado desde la investigación.

El análisis de sentimiento automatizado es una de las técnicas que se ha perfeccionado a la par de otras, como la clasificación de textos, con sus respectivas ventajas y limitaciones. Leetaru (2012) lo define de la siguiente manera: "es una forma de análisis de vocabulario que utiliza *lexicons* para medir la carga emocional latente en un cuerpo de texto" (p. 70), y así poder, por ejemplo, saber si los mensajes contienen emociones positivas, negativas o neutras en su estructura.

Desde una perspectiva similar, Feldman (2013) asegura que existen variados tipos de *sentiment analysis*, entre los que se encuentran: a) *document-level sentiment analysis* (*análisis de sentimiento a nivel de documento*, que asume que en el documento observado existe una opinión o posición del autor frente a algo), b) *sentence-level sentiment analysis* (análisis de sentimiento a nivel de frases) que se utiliza cuando "queremos una visión de *grano fino* acerca de las diferentes opiniones expresadas en un documento" (Feldman, 2013, p. 84), c) *aspect-based sentiment analysis* (análisis de sentimiento basado en atributos) en el cual el análisis se enfoca a los *atributos* que

poseen las entidades sobre las cuales se habla y el d) *comparative sentiment analysis* (análisis de sentimiento comparativo), que busca extraer este *sentimiento* en función de las comparaciones hechas dentro del mismo texto.

Para Kechaou, Ben Ammar & Alimi (2013), "la meta detrás de la aplicación de *opinión mining*, también llamado *sentiment analysis*, es hacer a la computadora capaz de procesar, reconocer y evaluar emociones o sentimientos" (p. 2), y cuyo crecimiento responde también al aumento *notable* de la cantidad de contenido generado por los mismos usuarios de la Red, ya sea en blogs, medios sociales u otras (Kechaou, Ben Ammar & Alimi, 2013).

A diferencia de Kechaou, Ben Ammar & Alimi (2013), hay quienes argumentan que *opinion mining* y *sentiment analysis* son dos cuestiones distintas. Por ejemplo, Cambria et al. (2013) arguyen que el "*opinion mining* y el *sentiment analysis* en realidad se enfocan en la detección de polaridad y el reconocimiento de emociones, respectivamente" (p. 15). En este caso, la polaridad se refiere a cuando se clasifica una opinión con respecto a dos sentimientos opuestos (positivo o negativo, por ejemplo) (Cambria et al., 2013).

Los autores, sin embargo, esclarecen que "debido a que la identificación de sentimientos es a menudo explotada para la detección de la polaridad (…) los dos campos se combinan generalmente bajo la misma sombrilla o incluso se utilizan como sinónimos" (Cambria et al., 2013, p. 15). Aún así, Liu (2010) ratifica que "nuestra comprensión y conocimiento del problema y su solución son todavía limitadas. La razón principal es que es una tarea de procesamiento de lenguaje natural, y el lenguaje natural no tiene

problemas fáciles" (p. 32). De esta manera nos encontramos con uno de los principales retos del *sentiment analysis*, pues al tratarse de una tarea en la que confluyen tantas *reglas* o emociones subyacentes, esta requiere que el nivel y la calidad del procesamiento hecho por la máquina sea óptimo.

En la construcción que hacen Vinodhini & Chandrasekaran (2012), destacan cuatro tipologías de fuentes de contenido para ejecutar los análisis de sentimiento que tienen como base nuevas tecnologías: a) *blogs*, por su capacidad para presentar las opiniones personales, b) *Review sites*, pues estos guían las decisiones de compra de los usuarios, c) *Datasets*, donde se incluyen *reviews* de productos o películas y d) *micro-blogging*, cuya plataforma más representativa es Twitter.

Yendo específicamente hacia el campo de los nuevos medios, ¿qué tanto se han utilizado estas técnicas? Un ejemplo es el estudio realizado Stieglitz & Dang-Xuan (2013), donde además los autores señalan que "pocas investigaciones le han prestado atención a las emociones como potenciales conductoras de difusión de información en el *social media*, en particular con respecto al comportamiento de compartir información de usuario" (p. 218).

En principio podemos resaltar estudios como el de Turney (2002), que aplica el análisis de sentimiento con respecto a *reviews* y de esta manera poder clasificarlas como *recomendadas* o *no recomendadas*, o trabajos como los de Meena & Prabhakar (2007) que se enfocan en extraer sentimiento de frases u oraciones. Por su parte, otros como Cai et al. (2010) llegan incluso a combinar las técnicas del *sentiment analysis* y el *topic*

modeling para extraer resultados más concretos sobre estas características (sentimiento y tema) y sus relaciones.

De cara al futuro, las investigaciones en el tema han devenido en nuevas líneas de investigación que van tomando forma a medida que el ejercicio científico descubre técnicas cada vez más potentes. Una de ellas es el *Affective computing*, cuyos exponentes del MIT *Affective computing* definen como "computación que está relacionada *con*, surge *de*, o deliberadamente influye *en* las emociones u otro fenómeno afectivo" (ver más en: http://affect.media.mit.edu/index.php).

Por todo lo anterior, podríamos decir que el horizonte de exploración en torno a las emociones es amplio, pero, ¿cuánta más información podremos abarcar, extraer y procesar en las próximas décadas?

9. Metodología

Para este estudio de corte cuantitativo se realizó un diseño no experimental que va del nivel descriptivo al correlacional, enfocándonos en las relaciones y posibles explicaciones de los fenómenos que se estudiarán con respecto a los tuits.

Sumado a los fundamentos conceptuales explorados en etapas previas de la investigación, se planteó un enfoque inductivo que orientara el proceso analítico, con el fin de obtener clasificaciones y temas a partir de los mismos datos. Así, este componente aportó herramientas metodológicas para el abordaje del objeto de estudio.

Específicamente, se llevó a cabo una investigación de corte cuantitativo en la que se tomaron 54.878 tuits recopilados desde el 02/01/2013 hasta el 02/01/2014 provenientes de la cuenta de Twitter del periódico *El Tiempo* @ElTiempo. Estas publicaciones fueron obtenidas a través de la aplicación alojada en el dominio www.allmytweets.net, cuya estructura permite capturar los últimos 3.200 mensajes emitidos por cualquier perfil de Twitter.

Se utilizó la técnica del análisis automatizado de contenido (AAC), acompañada de herramientas computacionales de minería de datos como el *machine learning*, *topic modeling* y *sentiment analysis*. Dicha tecnología permitió abarcar la totalidad de los tuits recopilados y, además, ejecutar de manera efectiva los procedimientos esperados. Debido a la envergadura del proyecto, estas herramientas (explicadas anteriormente) fueron esenciales para procesar las grandes cantidades de datos con los menores

inconvenientes posibles, a lo que se agregaron modelos multivariable para ahondar en las relaciones entre las categorías estudiadas.

Con el fin de llevar a cabo este análisis, se realizó la siguiente operacionalización de variables (ver Tabla 3), cuyo objetivo principal era dar un orden lógico a la estructura planteada para la recopilación, análisis e interpretación de los tuits:

Objetivo	Categorías	Subcategorías	Indicador
Determinar las propiedades del contenido de la innovación (temas noticiosos) difundidas a través de la cuenta de Twitter del periódico El Tiempo.	Innovación (TEMA NOTICIOSO)	Contenido de la innovación	Tema
		Propiedades innovadoras	Número de caracteres
			Número de palabras
			Número de enlaces
			Número de menciones por tuit
		Sentimiento	Tono
Describir el tiempo, el canales y el sistema social en el cual se enmerca la difusión de temas noticiosos difundidos a través de la cuenta de Twitter del periódico El Tiempo.	Tiempo	Momento de producción	Día codificado
			Hora
	Canal	Autoría del mensaje	RT o no
Determinar las variables que influyen en la difusión de temas noticiosos.	Innovación, tiempo, canal y sistema social	-	Modelo de regresión lineal múltiple para cada tema

*Las primeras tres categorías corresponden a los elementos fundamentales de la difusión de innovaciones descritos por Everett Rogers en "Difusión de Innovaciones" 5ª edición (2003) y que han sido adaptados para el presente estudio.

Tabla 3. Operacionalización de variables.

Para clarificar ciertas acepciones de Rogers (2003) que se han adaptado para el presente estudio, se propone la siguiente precisión conceptual:

- Innovación: Se toma la noción de Rogers (2003) de *la idea nueva como innovación*, para extrapolarla al *tema como innovación* que es percibido como nuevo por los *individuos* o *nodos* expuestos a los mensajes.
- Tiempo (Momento de producción): Para Rogers (2003) el tiempo se mide en términos del sujeto que adopta una innovación. En este caso, se asume el *tiempo* (momento de producción) como el día y la hora en que fue publicado el mensaje.
- Canal: En este estudio se asume el canal como la fuente de donde proviene el mensaje (autoría), es decir, si es de @ElTiempo u otro perfil de la plataforma.

Con estas aclaraciones, se procede a relacionar en los siguientes renglones los datos de la ficha para el análisis, con los siguientes indicadores:

- Hora.
- Día codificado.
- Si es retuit o no (RT).
- Número de hashtags.
- Número de menciones por tuit.
- Número de enlaces.
- Número de caracteres.
- Número de palabras.
- Tono (positivo, negativo o neutro).
- Tema al que pertenece.
- Modelo de regresión lineal múltiple para cada uno de los temas.

- Modelo de regresión logística múltiple para la variable *Innovación*.

Los *software* que se utilizaron para extraer la información de los datos se enuncian a continuación. Cada uno se ajustó con los parámetros necesarios para realizar las tareas pertinentes en cada fase de la investigación:

- Topic Modeling Tool: Esta herramienta es una interfaz que permite realizar modelamiento de temas de acuerdo a las necesidades del usuario. Al ajustar los parámetros (que incluyen el número de temas y el número de palabras por tema) se pueden obtener, a partir de la propia data, diversos grupos de términos asociados cada uno a un tema específico.
- Textalytics for Excel: Su principal función es realizar análisis de sentimiento automatizado a través del uso de una *API*. Está disponible para varios idiomas distintos, entre ellos el español y el francés.
- Microsoft Excel: A partir de una matriz realizada en Microsoft Excel, el conteo de frecuencias y otros ejercicios estadísticos pudieron ser llevados a cabo automáticamente. Se tomaron las fórmulas de la Tabla 4, ajustadas por el autor, para gestionar esta parte del análisis:

Descripción	Fórmula
Hallar los caracteres RT (Retuit – no incluido como parte de palabras) en el texto y arrojar 1=Sí o 2=No.	=SI.ERROR(ENCONTRAR("RT",**CELDA**),2)
Contar el número de caracteres contenidos en un tuit.	=LARGO(**CELDA**)
Contar el número de palabras contenidas en un tuit, incluyendo los enlaces.	=SI(LARGO(ESPACIOS(A1))=0,0,LARGO(ESPACIOS(**CELDA**))-LARGO(SUSTITUIR(**CELDA**," ",""))+1)
Contar el número de menciones contenidas en un tuit.	=LARGO(**CELDA**)-LARGO(SUSTITUIR(**CELDA**,"@",""))
Contar el número de enlaces en un tuit (posteriormente deben recodificarse los resultados reemplazando 4 por 1, 8 por 2, y así sucesivamente).	=LARGO(**CELDA**)-LARGO(SUSTITUIR(**CELDA**, "http",""))
Contar el número de hashtags.	=LARGO(**CELDA**)-LARGO(SUSTITUIR(**CELDA**, "#",""))

Tabla 4. Fórmulas para codificación en Excel.

- SPSS Statistics: Con este paquete se realizaron cruces de variables y otras operaciones estadísticas, con el fin de encontrar posibles correlaciones.
- QDA Miner (http://provalisresearch.com/es/productos/software-de-analisis-cualitativo/): Este paquete permitió obtener las frecuencias de palabras claves en el corpus, *hashtags* más usados, frases más usadas y las palabras clave en contexto.

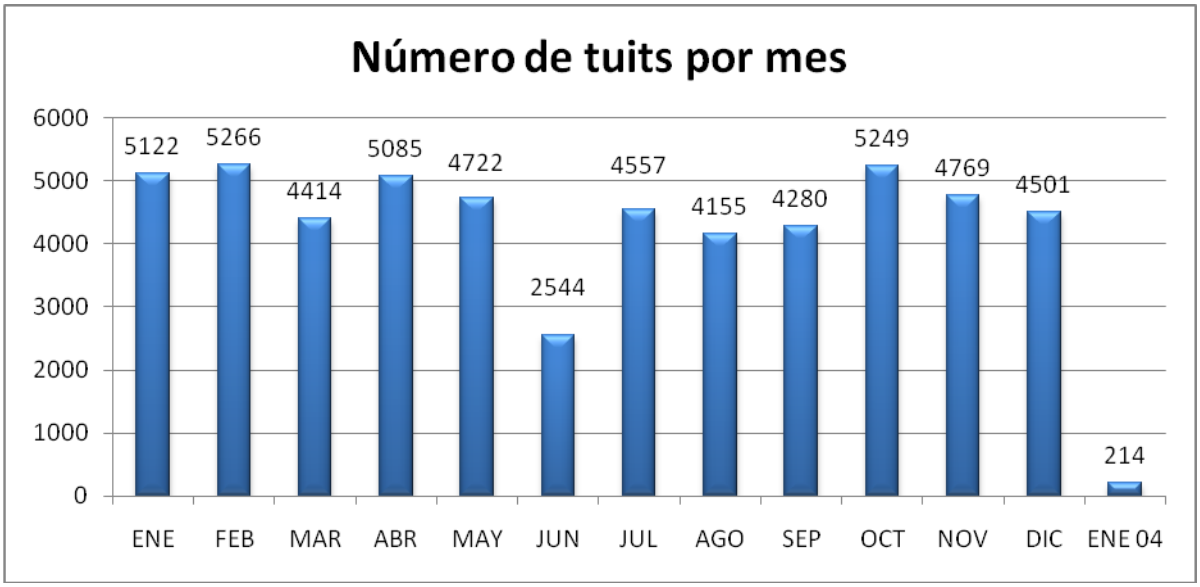
10. Resultados

10.1 Descripción del corpus

El corpus analizado está compuesto por 54.878 tuits provenientes de la cuenta @ElTiempo que fueron recopilados entre el 02/01/2013 y el 02/01/2014. Entre ellos se encuentran mensajes propios del medio de difusión noticiosa y también *RTs*, es decir, tuits replicados por @ElTiempo que fueron generados desde otras cuentas.

10.2 Tuits por mes, año 2013

Para responder a la pregunta de cómo se difundieron los temas noticiosos en el tiempo (PI 1.3), podemos decir en primera instancia que el número de tuits emitidos por el periódico *El Tiempo* en Twitter durante el año 2013 tuvo su punto más bajo en los meses de junio* y agosto (2.544 y 4.155 mensajes, respectivamente) y el más alto en el mes de febrero (5.266 mensajes). Es necesario aclarar que del mes de enero de 2014 solo se tomaron dos días (01/01/2014 y 01/02/2014) para completar un año de datos. Lo anterior se expone en el Gráfico 16:



*Del 9 al 20 de junio de 2013 no se tiene registro de tuits por una falla técnica de la aplicación con la cual se capturaron los mensajes de la cuenta @ElTiempo.

Gráfico 16. Número de tuits por mes. Perfil @ElTiempo durante 2013.

10.3 Día

Los tres días de la semana con mayor cantidad de tuits (emitidos por @ElTiempo) durante el 2013 fueron, en orden decreciente: jueves (9.155), martes (9.114) y miércoles (8.675). El compendio de frecuencias para cada uno de los siete días evaluados, que ayuda a responder la PI 1.3 sobre la difusión de temas noticiosos en el tiempo, se presenta en el Gráfico 17:

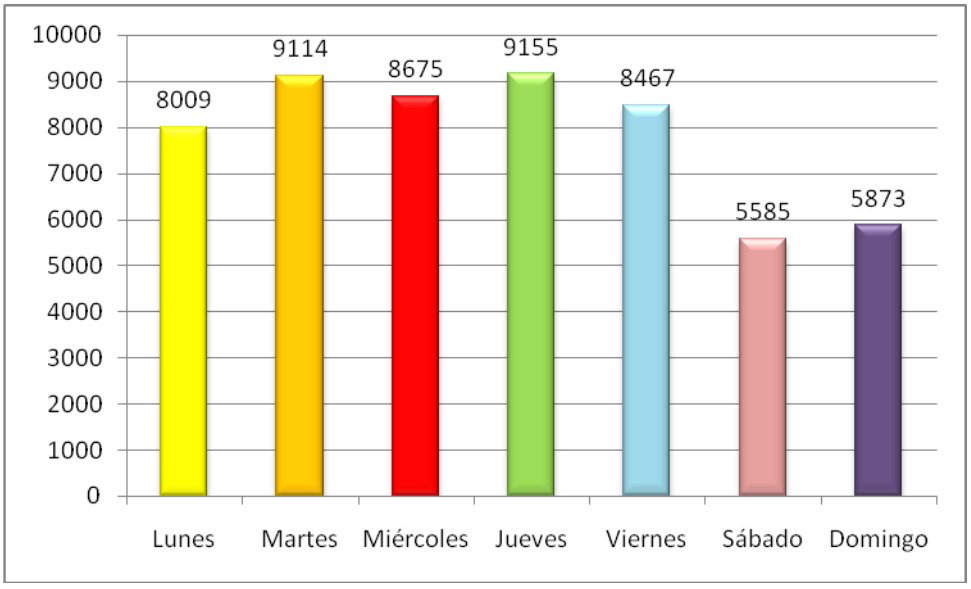


Gráfico 17. Número de tuits por día de la semana durante 2013. Perfil @ElTiempo.

10.4 Hora

Para complementar la respuesta a la pregunta de cómo se difundieron los temas noticiosos en el tiempo durante 2013 (PI 1.3), se exponen a continuación los totales de tuits por hora en cada día de la semana (durante todo el año). El Gráfico 18 agrupa los siete días de la semana, mientras que los Gráficos 19, 20, 21, 22, 23, 24 y 25 presentan una curva que permite observar la tendencia de publicación durante las distintas horas por día específico, teniendo en cuenta que el lunes empieza en la hora 0 y el domingo termina con la 168.

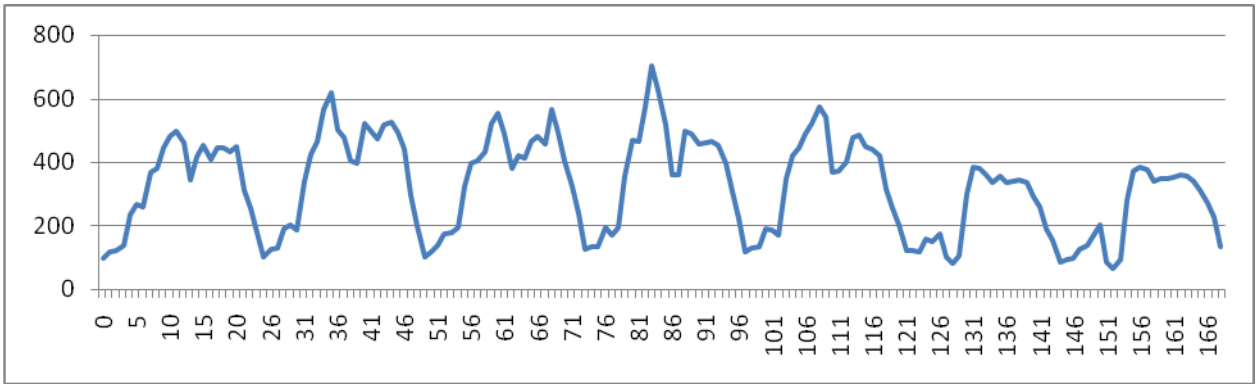


Gráfico 18. Total de tuits por hora durante la semana. Acumulado 2013. Perfil @ElTiempo.

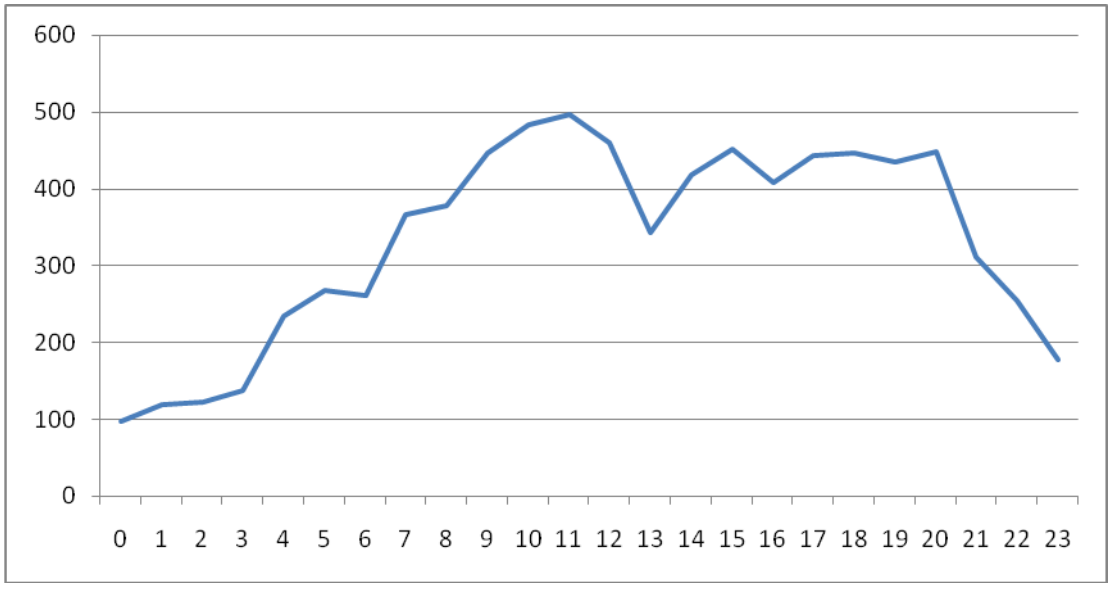


Gráfico 19. Total de tuits día lunes. Acumulado 2013. Perfil @ElTiempo.

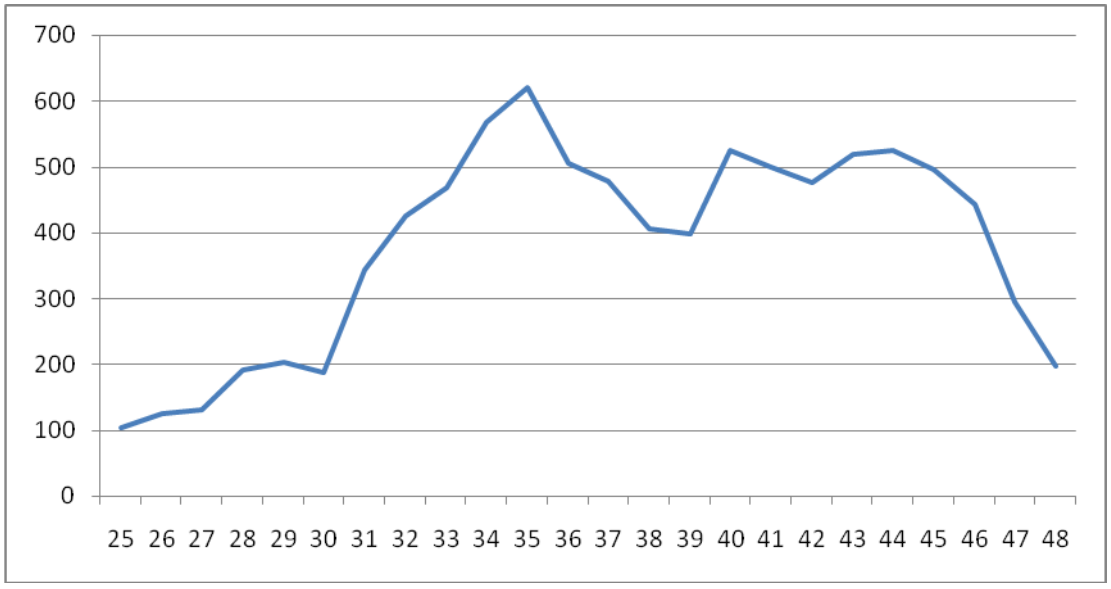


Gráfico 20. Total de tuits día martes. Acumulado 2013. Perfil @ElTiempo.

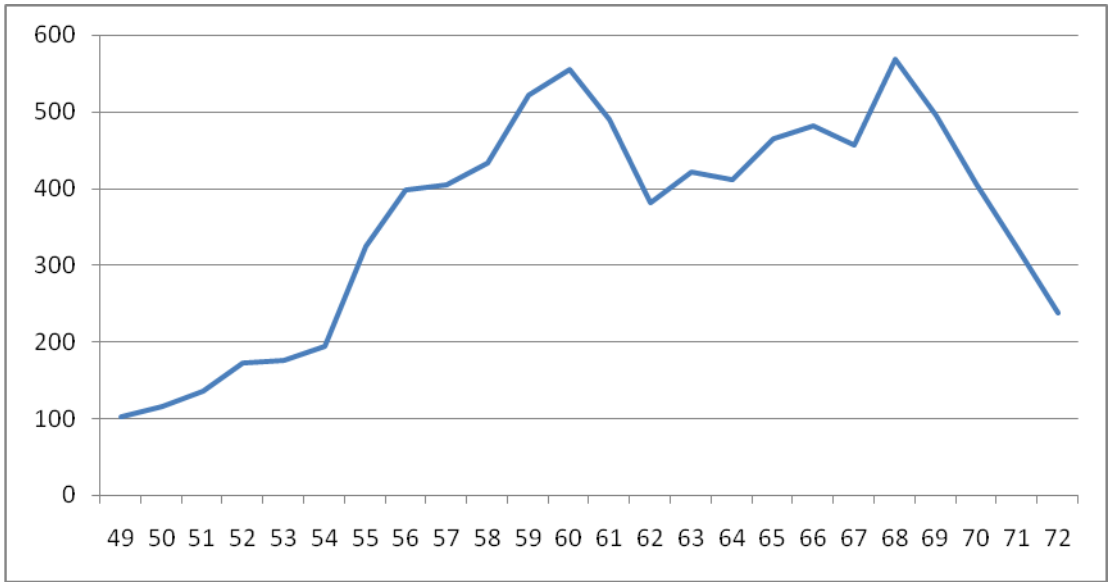


Gráfico 21. Total de tuits día miércoles. Acumulado 2013. Perfil @ElTiempo.

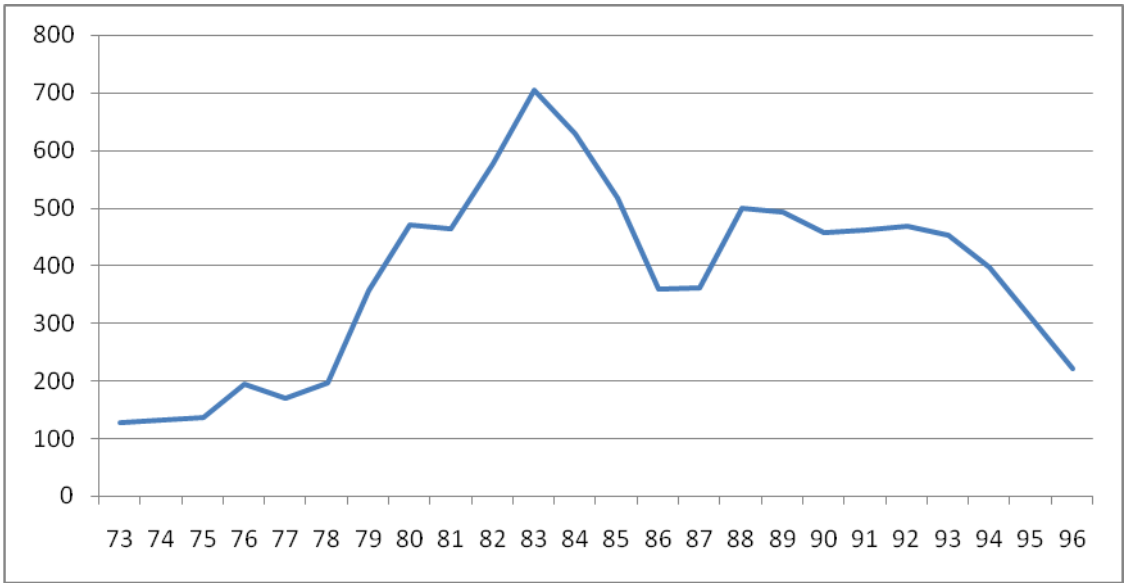


Gráfico 22. Total de tuits día jueves. Acumulado 2013. Perfil @ElTiempo.

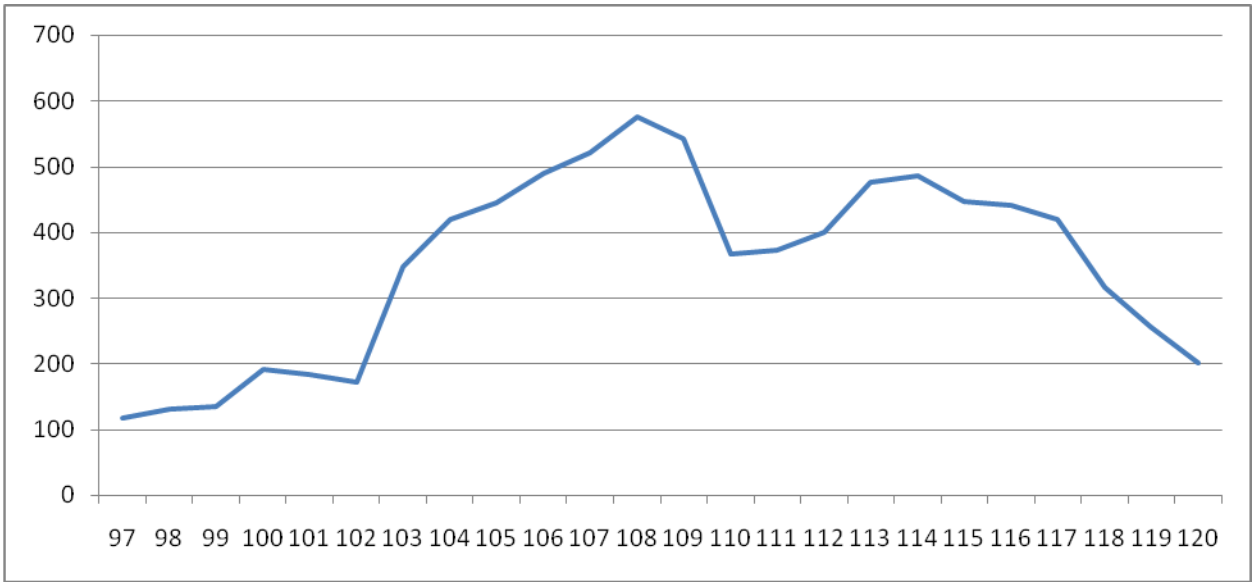


Gráfico 23. Total de tuits día viernes. Acumulado 2013. Perfil @ElTiempo.

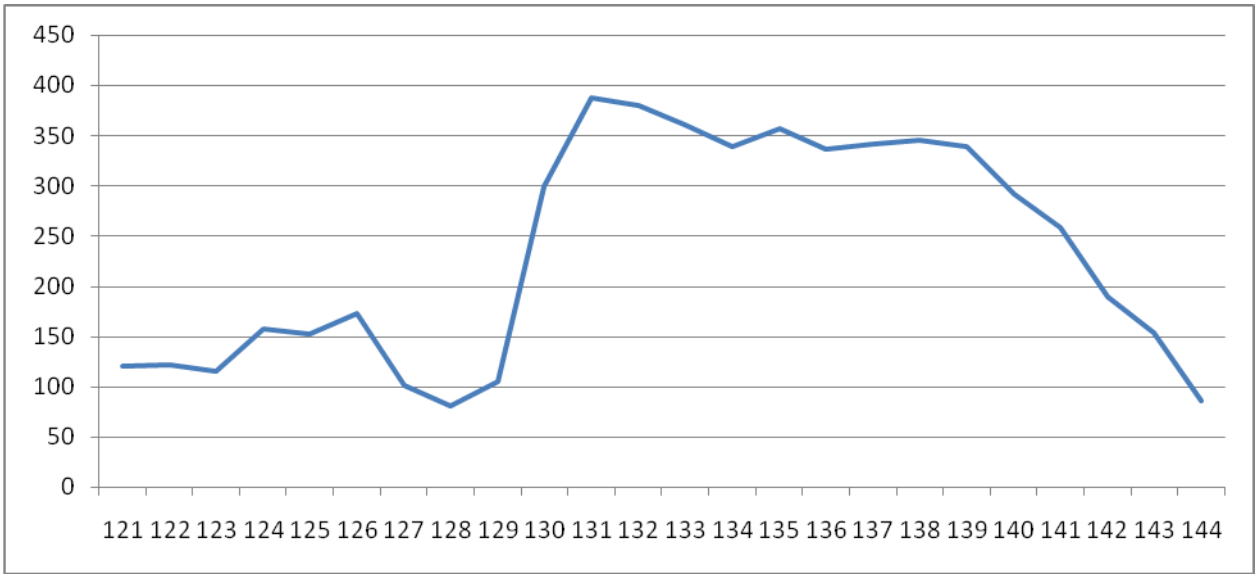


Gráfico 24. Total de tuits día sábado. Acumulado 2013. Perfil @ElTiempo.

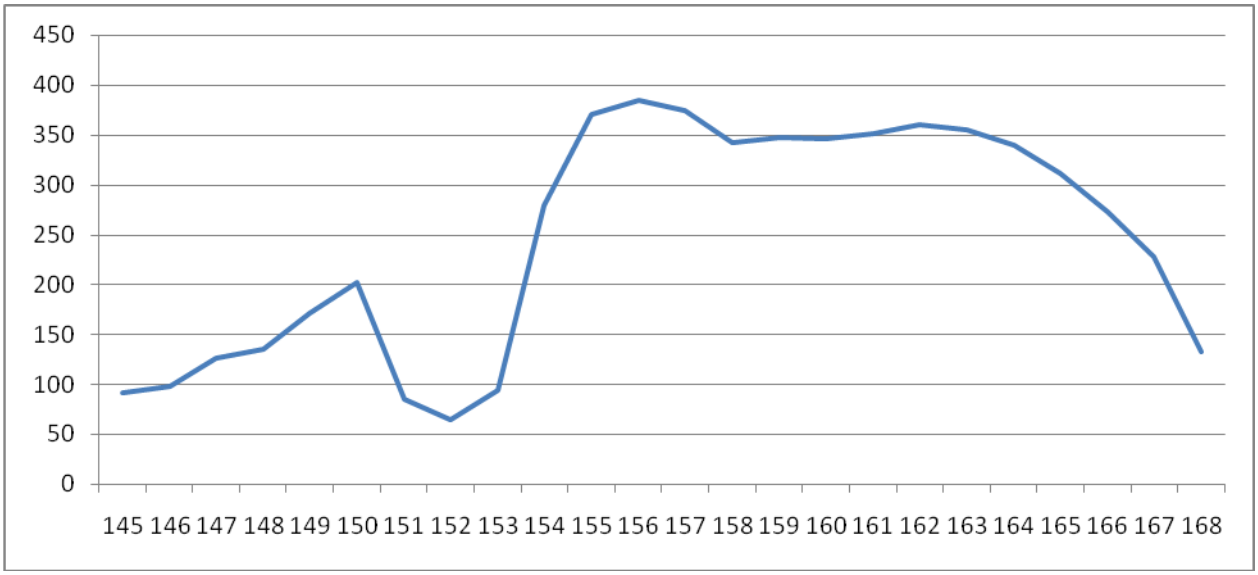


Gráfico 25. Total de tuits día domingo. Acumulado 2013. Perfil @ElTiempo.

10.5 Tema [TemaPRINCIPAL]

Posterior al *topic modeling* y para responder a la cuestión de cuáles fueron los temas noticiosos que emergieron del canal de Twitter de @ElTiempo durante 2013 y cómo se dio su difusión (PI 1.1, PI 1.3), se extrajo el porcentaje de presencia de cada uno de ellos en el corpus. Se encontraron cuatro (4) grandes temas que se relacionan a continuación:

- **Venezuela/Internacional.** Este tema abarca noticias enfocadas en asuntos de orden internacional y, además, aquellos tuits que se refieren a Venezuela (política y militarmente) ya sea directa o indirectamente. Poseen palabras clave 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*.
- **Deportes/Entretenimiento/General.** En esta clasificación entran los tuits que abordan temáticas deportivas y del mundo del entretenimiento. También hay noticias de menor envergadura, aunque con mucha menor presencia. Algunas de las palabras claves que pueden resaltarse son *mundial*, *Colombia*, *liga*, *final*, *copa*, *tiempo*, *partido*, *gol*, *Falcao*, *Medellín* y *colombiano*. Un tuit que ejemplifica este tema es: *Todos al acecho de Santa Fe en el grupo A de la Liga http://bit.ly/11S86bl*.
- **Temas residuales.** En este tema entran los tuits que no pudieron ser asociados claramente con alguno de los otros tres grupos.
- **Política/Interés Nacional/Conflicto.** Este último tema agrupa todos aquellos tuits que hacen referencia a situaciones de orden político, de guerra o conflicto

armado. Su delimitación deriva de palabras claves como *Gobierno*, *FARC*, *paro*, *política*, *paz*, *proceso*, *ataque*, *libertad* y *Santos*. Un tuit representativo de este conjunto es: *El balance de las Farc y el Gobierno de siete meses de #DiálogosDePaz http://bit.ly/11SupPe*

Frente a lo anterior, se obtuvo que un poco menos de la mitad (39,1%) de los tuits corresponden a la categoría *Política/InterésNal/Conflicto*, cuya composición, recordemos, abarca temáticas como el conflicto armado, violencia, hechos nacionales de amplia envergadura y sucesos correspondientes a personajes de la política o instituciones gubernamentales. En segundo lugar se encuentra un 27,7% que se inscribe en el tema *Deportes/Entret/General*. El Gráfico 26 muestra los porcentajes hallados para cada tema:

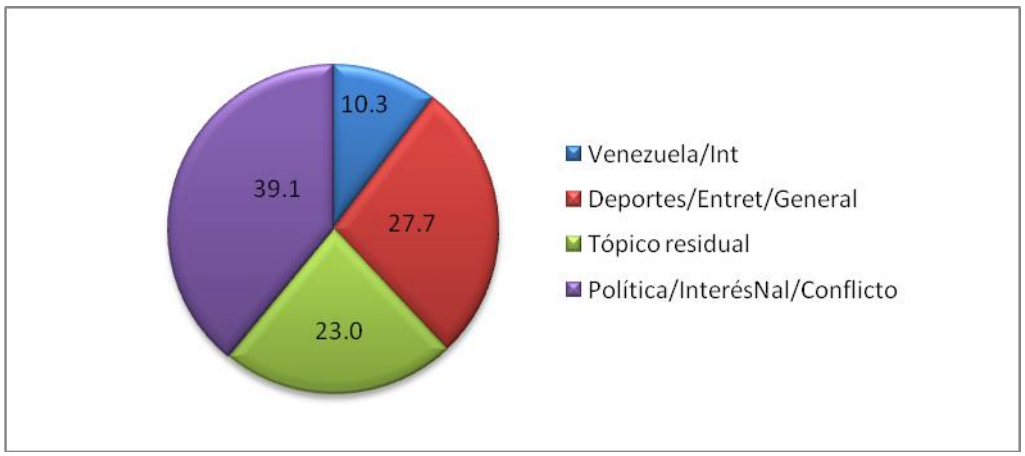


Gráfico 26. Porcentaje por tema en el corpus de tuits de @ElTiempo durante 2013.

Para ver cómo se comportaron cada uno de estos temas durante el año 2013, se pueden observar los Gráficos 27 al 30:

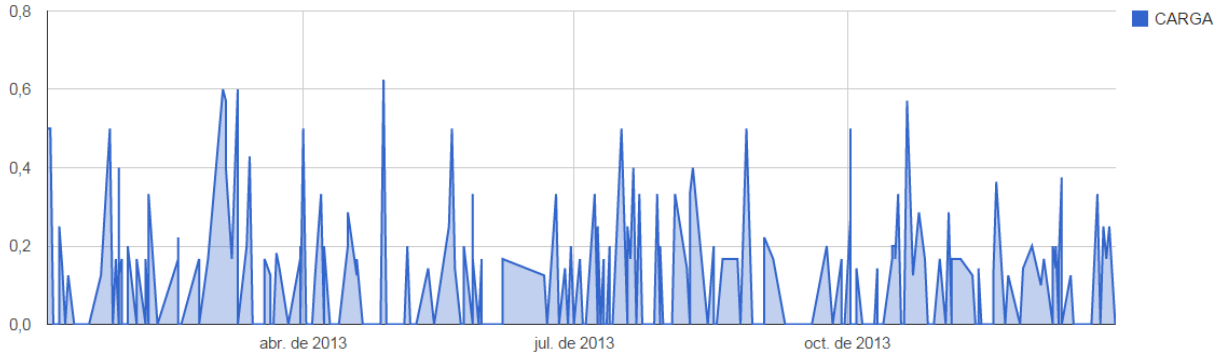


Gráfico 27. Evolución temporal del tema 1 (Venezuela/Internacional) durante el año 2013.

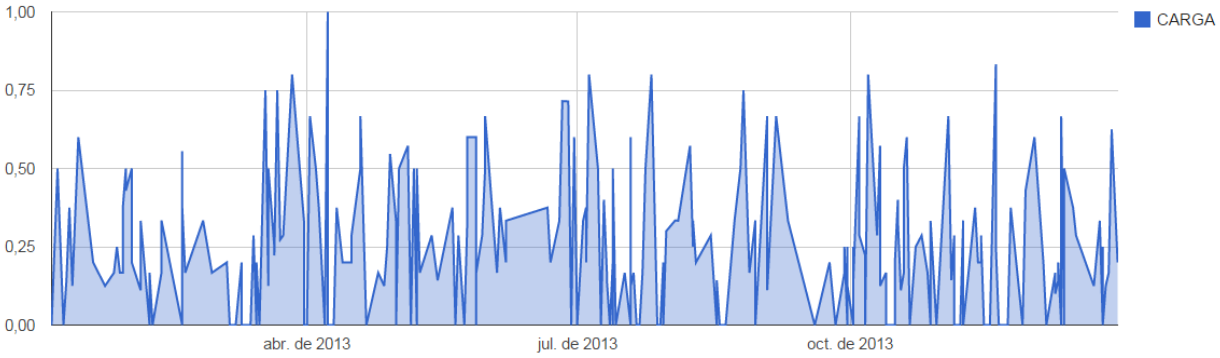


Gráfico 28. Evolución temporal del tema 2 (Deportes/Entret/General) durante el año 2013.

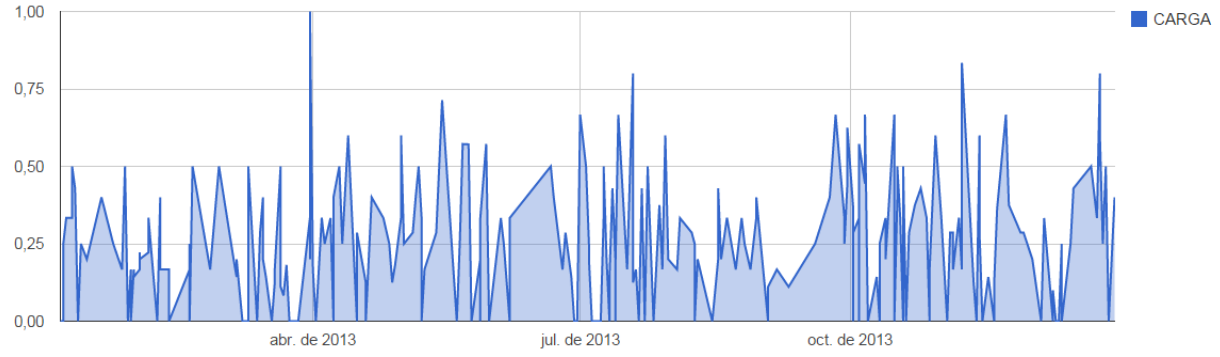


Gráfico 29. Evolución temporal del tema 3 (Temas residuales) durante el año 2013.

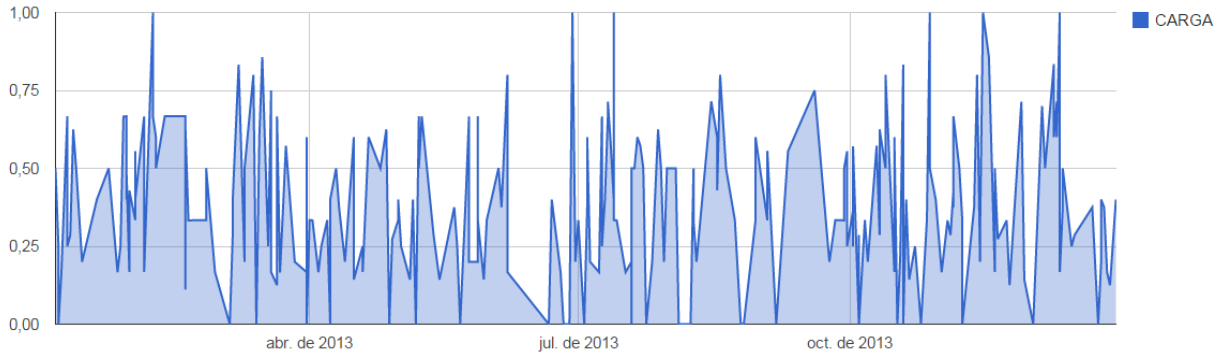


Gráfico 30. Evolución temporal del tema 4 (Política/Interés Nacional/Conflicto).

10.6 Tono [TonoETI]

Con el fin de mostrar de qué manera se evidenciaron las propiedades innovadores en los temas noticiosos difundidos a través de @ElTiempo durante 2013 (PI 1.2), se abordó el tono de los tuits publicados por la cuenta. Este está dividido en una escala que contiene seis (6) etiquetas, a saber: *P+*, *P*, *NEU*, *N*, *N+ y No tiene*. En esta clasificación, *P+* indica *positivo fuerte*, *P* es igual a *positivo*, *NEU* corresponde a *neutral*, *N* significa *negativo* y *N+* señala que el mensaje es *negativo fuerte*. Cuando el mensaje está etiquetado con *No tiene*, quiere decir que no hay marcas de subjetividad en el texto que compone el tuit.

Habiendo aclarado lo anterior, el Gráfico 31 presenta los porcentajes para esta propiedad innovadora:

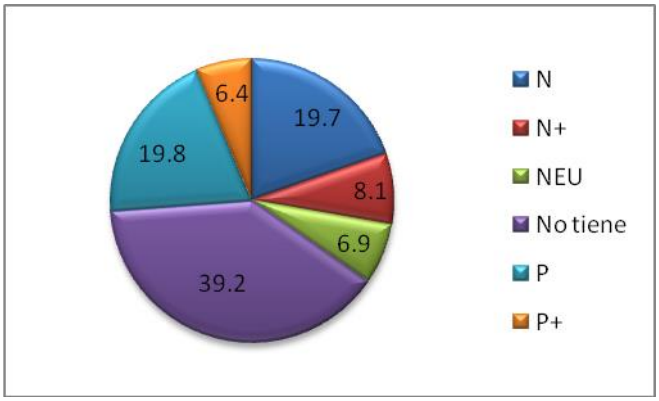


Gráfico 31. Porcentajes y clasificación de tono en los tuits de @ElTiempo durante 2013.

En este ítem se pudo identificar que el mayor porcentaje de los tuits emitidos por la cuenta @ElTiempo durante 2013 no posee marcas de subjetividad (39,2%), pero existen cantidades muy similares de tuits que pueden categorizarse como positivos (19,8%) y negativos (19,7%). En menor proporción encontramos los mensajes muy negativos (8,1%) y aquellos muy positivos (6,4%).

A fin de aclarar esta clasificación, se exponen a continuación ejemplos de mensajes calificados como *P+*, *N+* y *NEU* en el análisis:

- ***P+*: *España debutó con triunfo 4-1 frente a EE.UU. en Mundial Sub-20 http://bit.ly/14c5nwd***
- ***N+*: *No aparece maletín de agente de DEA asesinado http://bit.ly/13Zrq3k***
- ***NEU*: *"Mujeres militares en zonas de guerra y una cuestión de igualdad", por Sergio Muñoz Bata http://t.co/WHJEB9yP***

10.7 Palabras clave

Las palabras más utilizadas por @ElTiempo en Twitter durante 2013 se presentan a continuación. Entre ellas, hay un claro predominio de las palabras *Colombia* (2.971 casos) y *Bogotá* (2.391 casos) en el corpus de tuits, que se encuentran listadas junto a aquellas referentes al conflicto, hechos delictivos o violencia (*paz*, *gobierno*, *FARC*, *presidente*, *policía*, *Santos*). Cabe resaltar la presencia de las palabras *mundial* (723 casos) y (*mujer*, 482 casos) dentro de las más utilizadas, pues si bien no están relacionadas directamente con las anteriores, sí se presentaron de manera significativa en los mensajes del periódico @ElTiempo durante el año 2013 (ver Gráfico 32).

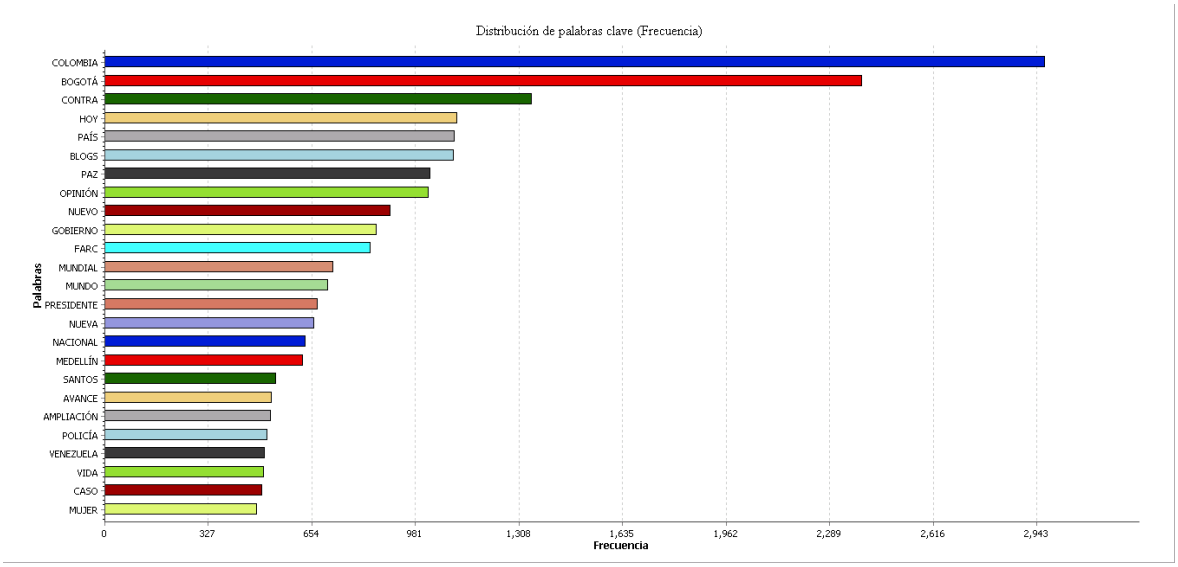


Gráfico 32. Palabras más usadas durante 2013 por la cuenta @ElTiempo.

10.8 RT o no

Para responder la PI 1.4, que indaga por el uso que se le dio al canal de Twitter desde las fuentes, se realizó el análisis de los retuits hechos desde el perfil @ElTiempo. Así, se encontró que el 18,6% de los tuits emitidos durante el año 2013 provenían de otras fuentes, es decir, eran *retuits* (*RT*). El porcentaje de tuits propios asciende a 81,4%, tal como se resume en el Gráfico 33:

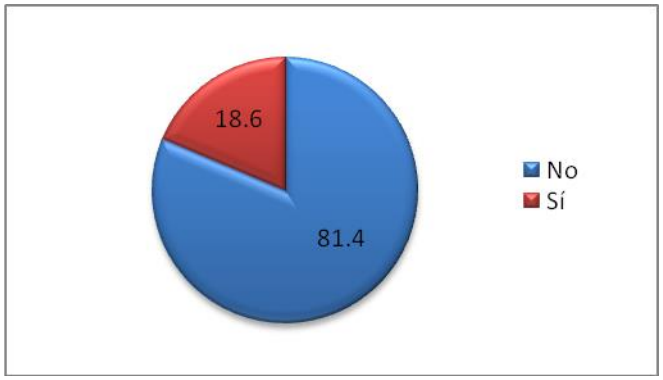


Gráfico 33. Porcentaje de *RTs* en los tuits de @ElTiempo durante 2013.

En el Gráfico 34 se muestra la cantidad de RTs por día de la semana:

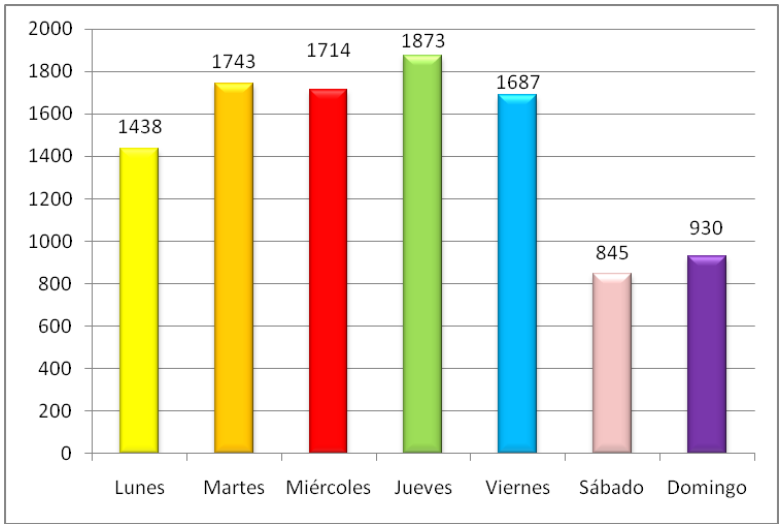


Gráfico 34. Total de *RTs* por día en los tuits de @ElTiempo durante 2013.

Los indicadores del 10.9 al 10.12 abordan la PI 1.2, que se enfoca en cómo se evidenciaron las innovaciones (*propiedades innovadoras*) en los temas noticiosos difundidos por la cuenta @ElTiempo durante el año 2013.

10.9 Número de caracteres [NUMcaracteres]

El número total de caracteres en el corpus es de 5'250.447. Para los 54.878 tuits, el número promedio de caracteres utilizados fueron 95,67 (M=95,67, SD= 21.350), siendo 91 la cantidad de caracteres con mayor frecuencia (M_o=140). Por último, se obtuvo M_e=92. La cantidad mínima de caracteres utilizada por @ElTiempo en sus tuits fue de 19, y la máxima de 140.

10.10 Número de palabras [NUMpalabrasSIN]

La suma del número total de palabras dio como resultado 639.264. Cada tuit utilizó, en promedio, casi 12 palabras (M=11,65, SD= 3.980). La mediana para este caso corresponde a M_e=11, teniendo como valor que más se repite en esta medición M_o=10 (7.404 casos). El máximo número de palabras encontradas en un tuit fue 30, y el mínimo 0 (que corresponde a un tuit que solo tenía un enlace).

10.11 Número de menciones [NUMmenciones]

Este indicador, que buscaba medir el número de menciones (*@usuario*) en los tuits del corpus, arrojó un total de 19.184. En primer lugar, se evidencia que más de la mitad de los tuits (71,4%) no tenía menciones. Un porcentaje considerable (23,6%) poseía una referencia a otra cuenta, ya sea a manera de respuesta directa, por estar incluida en un RT o por ser fuente/víctima/implicado en un hecho noticioso. La media para este ítem en particular fue de M=.35 (SD= .628), y la mediana y la moda fueron 0 (M_e= .00, M_o=0).

El número de menciones varió entre 0 y 9, y sus respectivas frecuencias se presentan en el Gráfico 35:

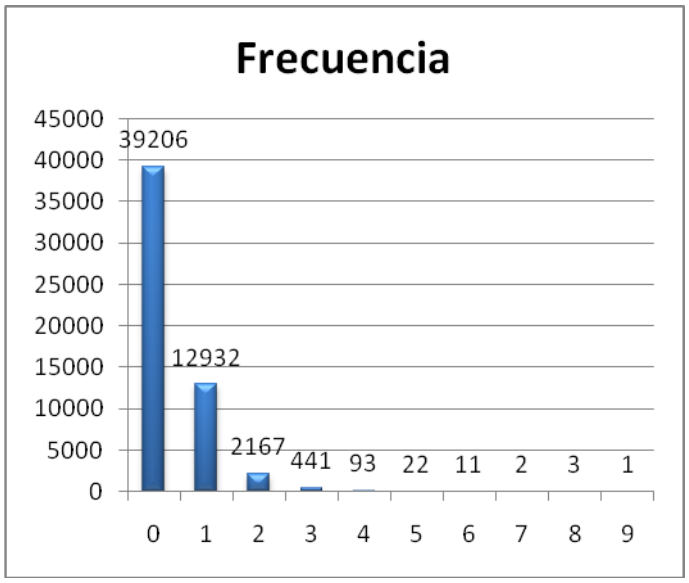


Gráfico 35. Número de menciones en el corpus y su frecuencia. Perfil @ElTiempo durante 2013.

Además de lo expuesto en el Gráfico 35, se añaden en el Gráfico 36 las cuentas más mencionadas por el perfil @ElTiempo durante el año 2013, destacándose, en orden descendente, @Portafolioco, @ElTiempo, @FUTBOLRED y la cuenta del presidente @JUANMANSANTOS.

Gráfico 36. Cuentas más mencionadas por el periódico @ElTiempo durante el año 2013.

En el Gráfico 37 se muestra la cantidad de menciones por día de la semana:

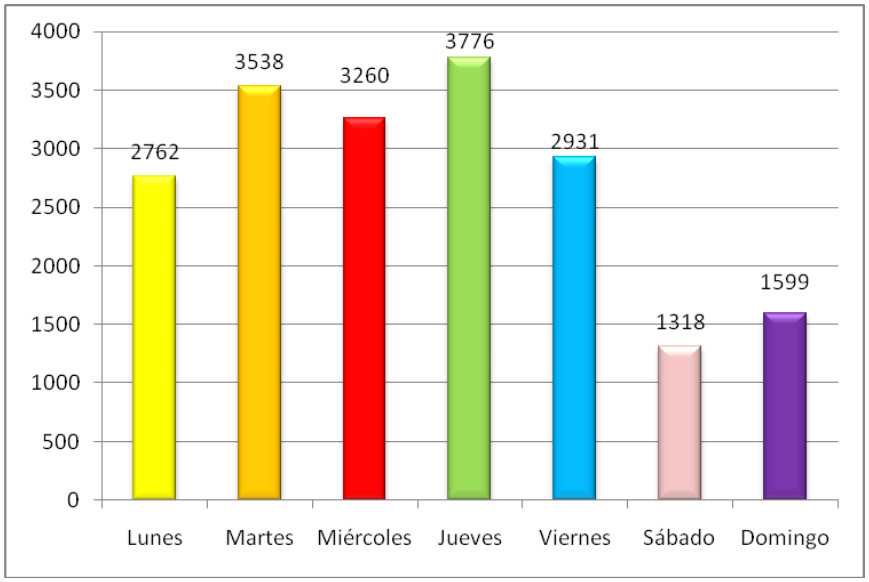


Gráfico 37. Total de menciones por día en los tuits de @ElTiempo durante 2013.

10.12 Número de enlaces [NUMenlacesRECOD]

Un gran porcentaje de los tuits analizados tienen un enlace dentro de su estructura (86,5%). También cabe resaltar que aquellos tuits que no tienen enlaces configuran el segundo porcentaje más alto (12,8%), seguidos por aquellos que tienen dos (0,8%). La media para este ítem fue de M=.88 (SD= .348), mientras que la mediana M_e= 1 y M_o=1.

Para visualizar estos porcentajes y los restantes (por número de enlaces) se elaboró el Gráfico 38:

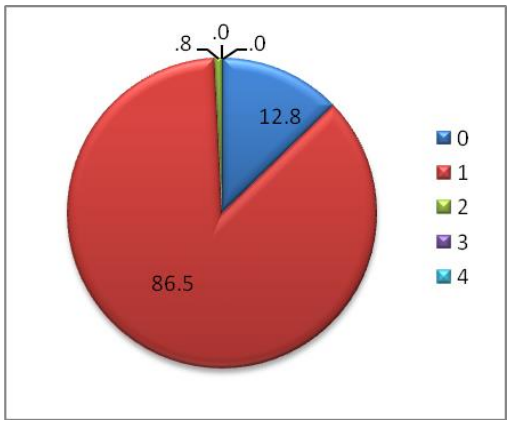


Gráfico 38. Porcentajes por cantidad de enlaces. Perfil @ElTiempo durante 2013.

En el Gráfico 39 se muestra el total de enlaces por día de la semana:

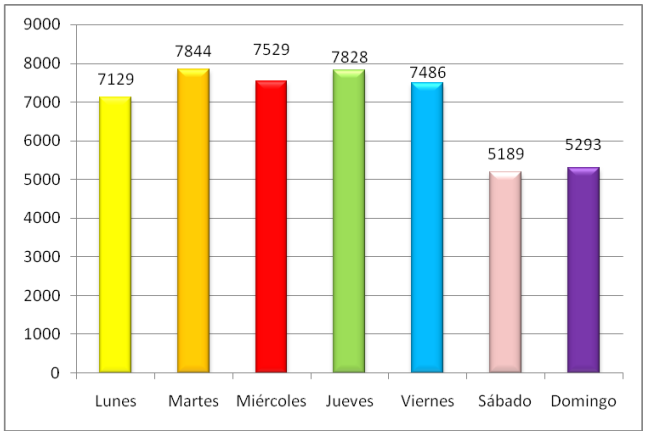


Gráfico 39. Total de enlaces por día en los tuits de @ElTiempo durante 2013.

10.13 Número de etiquetas [NUMhashtags]

Con respecto al número de etiquetas (palabras o expresiones que comienzan por el símbolo #) presentes en los tuits publicados por @ElTiempo durante 2013, se puede observar que más de la mitad de los mensajes (70%) en el corpus no tienen ningún tipo de etiqueta (*hashtag*), y los que tienen uno solo llegan al 27% (ver Gráfico 40). De ahí en adelante, el siguiente porcentaje corresponde a aquella cantidad de tuits con dos etiquetas (2,9%).

En este orden de ideas, podemos ver que M_o=0 y M_e= .00. Para este ítem, M= .33 (SD= .540), lo que indica que el promedio de *hashtags* se encuentra por debajo de 1.

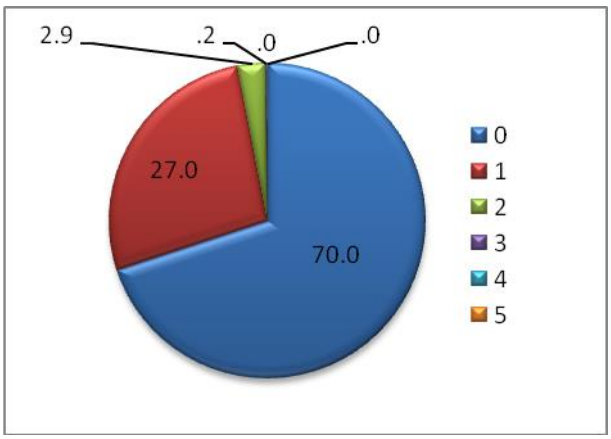


Gráfico 40. Porcentajes por número de etiquetas. Perfil @ElTiempo durante 2013.

En el Gráfico 41 se muestra el total de etiquetas por día de la semana:

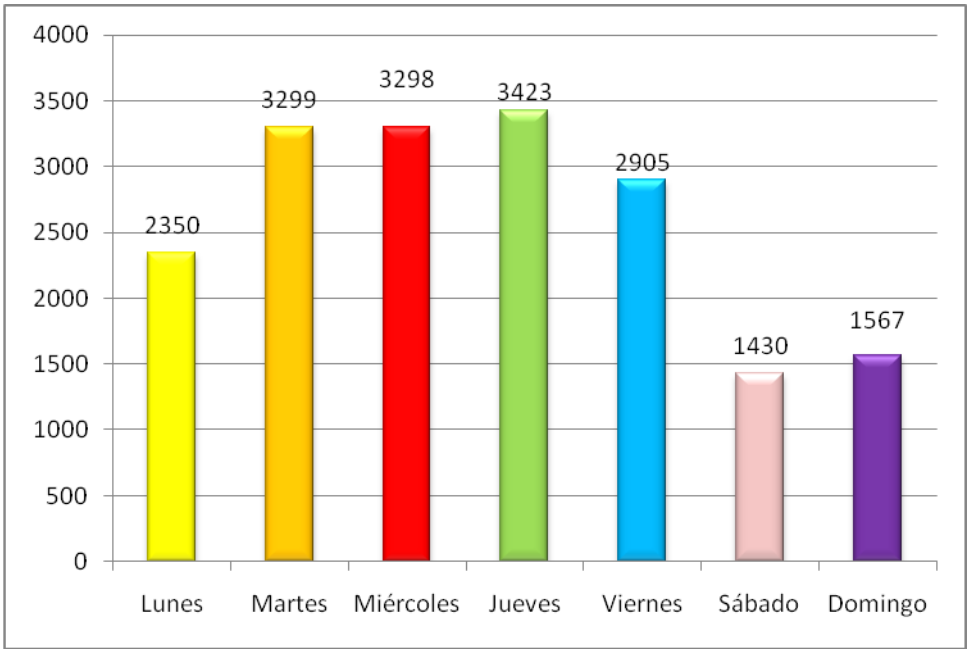


Gráfico 41. Total de etiquetas por día en los tuits de @ElTiempo durante 2013.

La Tabla 5 resume las tendencias centrales para los anteriores cinco (5) ítems:

	Media (M)	SD	Mediana	Moda
No. de caracteres	93.91	21.350	90	91
No. de palabras	11.65	3.980	11	10
No. de menciones	0.35	.628	0	0
No. de enlaces	0.88	.348	1	1
No. de etiquetas	0.33	.540	0	0

Tabla 5. Resumen de tendencias centrales. Perfil @ElTiempo durante 2013.

10.4 Análisis factorial y regresiones lineales múltiples

Con el objetivo de estimar el comportamiento de la variable dependiente *innovación* (*contenido de la innovación - tema noticioso*), se llevó a cabo un análisis factorial exploratorio (ver Tabla 6) para luego ejecutar regresiones lineales múltiples para cada uno de los temas noticiosos encontrados. El análisis factorial exploratorio pretendía encontrar grupos de variables entre las analizadas, es decir, los factores subyacentes en nuestros ítems. Los resultados del análisis, en el cual la medida de adecuación muestral KMO (Kaiser-Meyer-Olkin) indica que es válido para la muestra (KMO = 0,5), se exponen a continuación:

Ítem	1	2
Puntuación Z: Número de caracteres	.874	
Puntuación Z: Número de palabras (sin enlaces)	.911	
Puntuación Z: Número de menciones	.645	
Puntuación Z: Número de enlaces recodificado	-.606	
Puntuación Z(diarecod)		.649
Puntuación Z: Hora		.745

Tabla 6. Análisis factorial. Matriz de componentes rotados.

Como podemos observar en la Tabla 6, se incluyeron en el análisis factorial las puntuaciones estandarizadas (*Zscores*) de seis (6) ítems: número de caracteres, número de palabras sin enlaces, día de publicación, hora de publicación, número de menciones por tuit y número de enlaces por tuit. Esto dio como resultado dos factores claros: un primer componente al que hemos denominado *propiedades innovadoras* (que agrupa los ítems anteriores a excepción del día y la hora) y otro, al que se ha rotulado como *tiempo*

(día y hora). Se excluyeron del análisis factorial el tema noticioso, el tono del mensaje y si es RT o no por ser indicadores que no están agrupados junto a otros (ver Tabla 1).

Seguidamente se realizó una regresión lineal múltiple para cada uno de los cuatro (4) temas encontrados (variables dependientes), teniendo en cuenta los componentes hallados en el análisis anterior (factores). Los resultados se enuncian a continuación:

- **Tema 1:** Con el fin de conocer los factores predictores del *Tema 1 – Venezuela/Internacional* se realizó un modelo de regresión lineal múltiple. Para este caso, se corroboró la ausencia de multicolinealidad (con los valores de tolerancia cercanos a 1 y los FIV por debajo de 5) y un modelo significativo [$F(4, 54872)=2.417$, $p < .05$], aunque este no explica la varianza en la presencia del *Tema 1* (R^2=.000).
- **Tema 2:** En el caso del *Tema 2 – Deportes/Entretenimiento/General*, se cumplen asimismo los supuestos de multicolinealidad (valores de tolerancia cercanos a 1 y los FIV por debajo de 5), y el modelo es significativo [$F(4, 54872)=84.483$, $p < .000$]. Se encontró que las propiedades innovadoras (β=-0.021, $p < .000$), el tiempo (β=0.061, $p < .000$) y el tono del mensaje (β=0.043, $p < .000$) son predictores significativos, aunque el resumen del modelo indica que solo explican en un 0,6% (R^2=0.006) la varianza del *Tema 2*. Cabe resaltar, con respecto a las propiedades innovadoras, que este factor es el único que presenta una dirección de relación *negativa*.
- **Tema 3:** En la regresión lineal múltiple para el *Tema 3 – Residuales* se cumplieron los supuestos de multicolinealidad (valores de tolerancia cercanos a 1 y los FIV por debajo de 5). El modelo fue significativo [$F(4, 54872)=11.731$, p

< .000] y se encontró que las propiedades innovadoras (β=0.017, p < .000), el tiempo (β=-0.015, p < .000) y si es RT o no (β=-0.021, p < .000) son predictores significativos al nivel de p < .000, y el tono del mensaje a un nivel de p < .01 (β=-0.012, p < .01). Este modelo de regresión lineal múltiple explica el 0,1% de la varianza (R^2=0.001) en el *Tema 3*.

- **Tema 4:** Para el *Tema 4 – Política/Interés Nacional/Conflicto*, se cumplieron también los supuestos de multicolinealidad (valores de tolerancia cercanos a 1 y los FIV por debajo de 5) en el modelo. Este fue significativo [F(4, 54872)=44.187, p < .000], y a excepción de las propiedades innovadoras (β=0.007, p > .05), todos los demás son predictores significativos: tiempo (β=-0.040, p < .000), RT o no (β=0.019, p < .000) y tono del mensaje (β=-0.034, p < .000). El resumen del modelo refleja un R^2=0.003, lo que indica que explica la varianza del *Tema 4* en un 0,3%.

La Tabla 7 resume los resultados de las cuatro regresiones lineales presentadas anteriormente:

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

*p < .05; **p < .01; ***p < .001.

Tabla 7. Resumen de regresiones lineales por tema noticioso.

Con lo anterior, podemos afirmar que, para el caso de la pregunta PI 2.1 (¿Influyen las propiedades innovadoras en la difusión de temas noticiosos?), los resultados de las regresiones demuestran que en general las propiedades innovadoras, si bien tienen una influencia baja (en su conjunto explican muy poca de la varianza) en el tema noticioso, son predictoras significativas del mismo (ver regresiones de los temas 2 y 3).

Con respecto a la pregunta PI 2.2 (¿Cómo se dio la relación entre el tiempo de un tuit de la cuenta @ElTiempo y los temas noticiosos), se puede resaltar que para todos los temas encontrados el *tiempo* fue un predictor significativo, aunque por su peso bajo explica en menor medida la varianza del tema noticioso.

Para dar respuesta a la PI 2.3 (¿En qué forma afecta el canal a los temas noticiosos?), se encontró que el canal (*RT o no*) fue un predictor significativo para el tema 4 (β=0.019, $p < .000$). En el tema 2 (*Deportes/Entretenimiento/General*), el canal no aparece como un ítem significativo en la varianza (β=0.005, $p > .05$).

Por último, y a fin de constatar dos de las correlaciones encontradas por Kang, O' Donovan & Höllerer (2012) [*tono/enlaces* y *enlaces/RT*, que para efectos de este estudio han sido evaluadas como propiedades innovadoras de los temas noticiosos (PI 1.2)], se presentan a continuación los resultados que corresponden al cruce de las variables independientes *Tono/Número de enlaces* y *Número de enlaces/RT o no*:

- La relación entre el tono y el número de enlaces es significativa pero muy baja ($r = -0.71$, $p = .000$), siendo esta inversamente proporcional. Esto quiere decir que

existe una probabilidad de que en presencia de un tono más positivo, el número de enlaces disminuya, y viceversa.

- La relación entre el número de enlaces y los RT es significativa pero baja (r = -.242, p = .000), y de igual forma inversamente proporcional. Lo anterior sugiere que a mayor número de enlaces en un mensaje, menos probabilidades tiene de ser un RT.

11. Discusión

La pregunta central que guió esta investigación indagaba sobre cómo fue la difusión de temas noticiosos en el canal de Twitter del periódico *El Tiempo* durante el año 2013. Así, y antes de emprender la tarea de recolección y análisis de los datos, se plantearon dos macro-preguntas de investigación, a saber: ¿Cómo fue la difusión de innovaciones (temas noticiosos) en el canal de Twitter del periódico El Tiempo durante el año 2013? (PI 1) y ¿cuáles son los factores que influyen en la difusión de innovaciones (temas noticiosos) en la cuenta @ElTiempo? (PI 2). De ellas se desprendieron las preguntas (y, por ende, los resultados) que se discuten a continuación.

Con respecto a la pregunta PI 1.1, [¿Cuál es el contenido de las *innovaciones* (*temas noticiosos*) que emerge de los tuits publicados por el periódico @ElTiempo durante el 2013?], podemos resaltar que los tres temas principales presentes durante 2013 en la cuenta @ElTiempo fueron *Venezuela/Internacional*, *Deportes/Entretenimiento/General* y *Política/Interés Nacional/Conflicto*.

Tal como se ve reflejado en el Gráfico 26, el tema *Política/Interés Nacional/Conflicto* ocupó la mayor parte de los mensajes publicados por la cuenta @ElTiempo durante 2013 (39%), seguido por *Deportes/Entretenimiento/General* (27,7%). Cabe aclarar que en el primero se incluyen tuits que abarcan temáticas como el conflicto armado, violencia, hechos nacionales de amplia envergadura y sucesos correspondientes a personajes de la política o instituciones gubernamentales.

Si observamos algunos de los hechos noticiosos más importantes para Colombia durante 2013 (como el comienzo de los diálogos de paz entre el Gobierno Santos y las Fuerzas Armadas Revolucionarias de Colombia FARC – EP a finales de 2012 y las eliminatorias para la Copa Mundial de la FIFA Brasil 2014), vemos que se corresponden, además de los temas encontrados, con las palabras clave (*paz*, *gobierno*, *FARC*, *presidente*, *policía*, *Santos*, *contra*, *mundo* y *mundial*, entre otras). Hay un claro énfasis del periódico en estos dos temas que se evidencia en los Gráficos 28 y 30.

Utilizando también *topic modeling*, Kim & Oh (2011) observaron que los temas de larga duración identificados en las noticias podrían rotularse como *política*, *negocios* o *deportes*. Teniendo en cuenta los resultados de Kim & Oh (2011), y al contrastarlos con lo obtenido en la pregunta PI 1.1 [¿Cuál es el contenido de las *innovaciones* (*temas noticiosos*) que emerge de los tuits publicados por el periódico @ElTiempo durante el 2013?] de este estudio, podríamos decir que existe cierta correspondencia en dos de los tres temas generales: *política* y *deportes*.

Por otra parte, la pregunta PI 1.2 del presente estudio buscaba observar cómo se evidenciaron las propiedades innovadoras en los temas noticiosos difundidos por la cuenta @ElTiempo durante 2013. Al analizar el número de caracteres como primera propiedad innovadora, es importante destacar que los tuits que ocupan la totalidad del espacio proveído por la plataforma (140 caracteres) son el 4,5%, es decir, 2.495 casos. Aunque es un porcentaje relativamente bajo, es el más alto con respecto a otras longitudes de mensaje. Esto puede sugerir que la tendencia del medio es a publicar mensajes relativamente cortos, resumiendo informaciones en pocas palabras y dando respuestas concisas a otros usuarios de la plataforma (ver Tabla 5).

El número de enlaces y de palabras son otras de las propiedades innovadoras que se analizaron en esta investigación. Aquí podemos resaltar brevemente el hecho de que el 71,4% de los tuits no tenían menciones y el 86,5% tenían enlaces. Lo primero podría sugerir una baja interacción con otros usuarios de la plataforma o la posibilidad de que las personas/entidades nombradas en los mensajes no posean una cuenta de Twitter, y lo segundo una alta tasa de redireccionamiento desde esta plataforma.

Con respecto a nuestros resultados para contrastar con las correlaciones *tono/enlaces* y *enlaces/RT* halladas por Kang, O' Donovan & Höllerer (2012), podemos decir que no ratifican totalmente lo encontrado por estos autores (que cuando el sentimiento está polarizado, es decir, es negativo o positivo los enlaces son más comunes, y que cuando hay presencia de enlaces es más posible que los tuits sean RT). Sin embargo, sí se encontró que, al igual que en el estudio de Kang, O' Donovan & Höllerer (2012), hay una correlación estadísticamente significativa entre estas tres *propiedades innovadoras*.

Sobre cómo se difundieron los temas noticiosos en el tiempo (PI 1.3), podemos argumentar que aunque en algunos meses la emisión de tuits disminuyó (como es el caso de marzo y septiembre), se puede inferir que la tendencia de publicación fue constante. El rango de tuits publicados varió entre 4.000 y 5.500 aproximadamente (por mes), evidenciándose una notable disminución de la producción de tuits durante los días sábado (5.585) y domingo (5.873) con respecto a los otros días de la semana. Esta situación cambia con el aumento presentado los días lunes (se registraron en este día 2.136 tuits más que el domingo), lo que podría obedecer a dinámicas periodísticas y el volumen de información generado durante los fines de semana.

Argüelles & Muñoz (2012) hallaron que el día más mencionado en su corpus fue el viernes (*Friday* = 1.156; Viernes = 1.013). No obstante, al analizar esta tendencia en los tuits de @ElTiempo durante 2013 se encontró que, a diferencia del estudio referenciado anteriormente, el día más mencionado fue el domingo (167 veces). A este le siguieron el lunes (156 veces), martes (147), jueves (137) y viernes (120). Como argumentan Argüelles & Muñoz (2012) en consonancia con Stubbs (2001), esto puede ocurrir debido a que factores culturales influyen en la aparición o no de los días de la semana en el corpus. Lo anterior, si se entienden estos lapsos como constructos sociales y mentales que se refieren indirectamente (a través de representaciones cognitivas) a una realidad que han ayudado a crear (Stubbs, 2001), teniendo un efecto real en el comportamiento de las personas. Así, los días de la semana pueden asociarse con ciertos tipos de mensajes o emociones que se quieren transmitir, ya sea por parte de una persona o, en este caso, un medio de difusión de noticias (el *descanso* o el *deporte* asociados a los domingos, la diversión con otros días del fin de semana, etc.).

Al examinar el uso que se le dio al canal desde el cual se emitieron los mensajes (PI 1.4), vemos que el porcentaje de tuits propios asciende a 81,4%. La poca frecuencia de RT (18,6%) indica bajos niveles de replicación de contenidos, es decir, que el periódico genera escasos momentos de visibilidad para mensajes de otras cuentas. Manteniendo este alto porcentaje de originalidad en sus publicaciones, la cuenta @ElTiempo redirecciona a los usuarios a su página en Internet u otros recursos *online*, convirtiendo a los sitios web externos en un marco de referencia para la lectura de sus tuits. Así, aunque los mensajes propios sean cortos (ver Tabla 5), proveen mayor información que se aloja en otros dominios. A pesar de esto, se puede destacar que el 70% de los tuits no

tienen *hashtags*. Esto limitaría de cierto modo la inclusión de los mensajes en conversaciones diferentes a las generadas por el medio de difusión noticiosa, teniendo un impacto en la viralidad de sus contenidos.

A pesar de que el modelo planteado solo explica el 1% de la varianza en el tema noticioso, se puede resaltar que las *propiedades innovadoras* aparecen como predictoras significativas de este en todos los casos, lo que no sucede con el canal, por ejemplo. Este último se destaca como un predictor significativo solo en el tema 4, que abarca asuntos políticos y de conflicto. Esta asociación podría deberse en alguna medida al hecho de que, generalmente, para reportar sucesos políticos o de conflicto las fuentes (tomadas como *canal* en este estudio) suelen ser citadas textualmente para apoyar la información. Lo anterior invitaría a pensar en la utilización de mensajes de otros perfiles (p. ej. personalidades de la política) para comunicar este tipo de eventos noticiosos. Sin embargo, se considera que otros análisis complementarios son necesarios para valorar mejor la afectación de esta variable dependiente (*tema noticioso*) en función de las *propiedades innovadoras* que presenta Twitter.

Con todo lo anterior, se puede argüir que estos resultados dan luces, a partir de la utilización de la teoría Rogers (2003) y de los métodos de computación avanzada, sobre los elementos característicos de la difusión noticiosa en medios sociales. Debido a que los medios masivos de comunicación alrededor del mundo —incluyendo los de América Latina— le apuestan cada vez más al uso de los medios sociales para publicar, informar e interactuar con sus lectores y audiencias (Said-Hung et al., 2013; García de Torres et al., 2008; Lasorsa, Lewis & Holton, 2012; Caballero, 2000), este abordaje se permitió abarcar todo un año de publicaciones sin dejar de lado el análisis a profundidad de

características, relaciones y posibles causas que llevaran a comprender mejor la difusión de temas noticiosos en Twitter.

Las metodologías de análisis de contenido automatizado, *topic modeling* y análisis automatizado de sentimiento, entre otras, demostraron ser convenientes para este tipo de trabajos, resultando así en un apoyo teórico-práctico tanto para el estudio de las dinámicas periodísticas (en específico) como el de la comunicación (en general). Se destacan asimismo la disminución del tiempo de limpieza y procesamiento de datos y el aumento de la capacidad de análisis que, junto a la alta confiabilidad de los algoritmos y fórmulas, acrecentan las posibilidades y objetos de estudio.

12. Conclusiones

Atendiendo a la necesidad de comprender mejor la difusión noticiosa en Twitter, el presente trabajo tenía como objetivo general precisamente caracterizar el proceso de difusión de temas noticiosos en el canal de Twitter del periódico colombiano *El Tiempo* durante el año 2013. Además, desde sus objetivos específicos, se pretendió determinar las propiedades de los temas noticiosos, describir su difusión en función del tiempo y el canal y determinar las variables que influyen en el proceso.

Considerando que los análisis de medios se han enfocado en momentos cortos de tiempo y eventos específicos (Deutschmann & Danielson, 1960; Greenberg, 1964; Henningham, 2000), esta investigación tomó elementos del tratamiento de la *Big Data* para contrarrestar, de algún modo, el fenómeno que Mayer-Schönberger & Cukier (2013) llaman "una muestra en lugar de un todo" (p. 40).

Se puede concluir, en primer lugar, que nuestro modelo basado en la Teoría de Difusión de Innovaciones de Rogers (2003), si bien permite caracterizar cómo se da la difusión de mensajes en Twitter, no resulta ampliamente explicativo al momento de estimar el comportamiento del *tema noticioso* cuando este es tratado como *innovación*. Por tal motivo, se considera necesaria una investigación más extensa (tal vez desde otros acercamientos teóricos y enfoques metodológicos) para hallar los factores que influyen en esta variable en particular.

Al evaluar cómo se evidenciaron las *propiedades innovadoras* (PI 1.2), cómo fue la difusión en el *tiempo* de los temas noticiosos (PI 1.3) y el uso del *canal* de Twitter por

parte del medio (PI 1.4), podríamos afirmar que @ElTiempo, específicamente, tiene un bajo índice de contenidos replicados en su perfil (con mensajes propios relativamente cortos), un promedio mensual de tuits publicados que varía entre 4.000 y 5.500 y gran cantidad de enlaces que dirigen a los usuarios a otros sitios web. Se da, además, poca inserción de estos mensajes en conversaciones a través de etiquetas.

En el corpus se encontraron, a través del *topic modeling*, tres temas principales que responden a la PI 1.1 [¿Cuál es el contenido de las innovaciones (temas noticiosos) que emerge de los tuits publicados por el periódico @ElTiempo durante el 2013?], a saber: *Venezuela/Internacional*, *Deportes/Entretenimiento/General* y *Política/Interés Nacional/Conflicto*.

Con las cuatro regresiones lineales múltiples llevadas a cabo se estableció que las *propiedades innovadoras* son predictoras significativas de los temas al igual que el tiempo (aunque este en menor medida). También se pudo observar que el canal (*RT o no*) fue un predictor significativo para el tema 4 (β=0.019, $p < .000$).

Cuando se observan los perfiles que más mencionó el periódico El Tiempo durante 2013 desde su cuenta de Twitter, se puede concluir que tanto en temas sociales, políticos y económicos como en los asuntos deportivos, buena parte de los perfiles más nombrados (@Portafolioco, @ElTiempo, @FUTBOLRED, @CityTV) pertenecen a su misma casa editorial (Casa Editorial El Tiempo). La cuenta del presidente @JUANMANSANTOS es la personalidad pública más nombrada, seguida de @PETROGUSTAVO, @FALCAO y @BARACKOBAMA. Los datos sugieren que existe una relación entre los perfiles más mencionados, las palabras claves y los temas

subyacentes encontrados en el *topic modeling*, puesto que los resultados de todos estos ítems encajan en las categorías de *política* o *deportes*.

Una de las limitaciones que presentó el estudio fue la falta de datos del 9 al 20 de junio de 2013, causada por un problema técnico de la aplicación que se utilizó para recopilar los tuits. Esto pudo haber ocasionado un ligero bache al observar la data de este mes, teniendo en cuenta que no hay forma de recuperar aquellos mensajes que desaparecieron de la línea de tiempo a menos de que la fuente (en este caso, el periódico El Tiempo) los suministre. Otra limitación de este estudio consistió en la escasez de investigaciones previas en América Latina que tuvieran características similares, tanto en el tamaño de la muestra como en las técnicas utilizadas. Por tal razón, la comparación de datos empíricos con otras exploraciones de países vecinos no fue posible.

Se puede agregar, además de lo anterior, que por ser Twitter una plataforma en la que los mensajes tienen una *vida corta* (ya que van *desapareciendo* los más antiguos a medida que se publican nuevos tuits), otra de las limitaciones fue la recolección de un archivo que abarcara un lapso más amplio, pues esto implicaría tener acceso abierto a los contenidos almacenados por los propietarios de los perfiles. Por último, y con respecto a la adaptación de la teoría de Rogers (2003) en la investigación, surgió la limitante de no poder incluir el sistema social como una categoría de análisis, debido a las características de Twitter que se pudieron abordar desde esta perspectiva.

Con todo lo anterior, queda abierto un espacio de investigación donde los métodos computacionales sirvan de soporte para estudios más amplios y profundos sobre difusión noticiosa en medios sociales, específicamente en cuanto al comportamiento de

los *temas noticiosos*. Se hace hincapié en la necesidad de investigaciones de mayor envergadura en los medios masivos de América Latina y sus perfiles en medios sociales, donde se contrasten los resultados de varios casos específicos. Este ejercicio en particular permite vincular saberes y herramientas de otras disciplinas al análisis de noticias, lo que significa un incremento potencial de la capacidad de comprensión de fenómenos noticiosos en entornos digitales.

Así, sería relevante preguntarse en futuras indagaciones sobre cómo se da el proceso de adopción de estos temas noticiosos en medios sociales desde la perspectiva del sujeto, cómo aparecen y desaparecen temas noticiosos de corta duración a lo largo del tiempo y qué modelos pueden plantearse para intentar explicar la varianza de otras características de los mensajes en plataformas como Twitter.

13. Referencias

Alpaydin, E. (2010). *Introduction to Machine learning*. Cambridge/London: The MIT Press.

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired. Recuperado de http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

Arcila, C. et al. (2014). *e-Investigación Social en América Latina*. En Romero, F. & Sánchez, M. (eds.) *Ciencias Sociales y Humanidades Digitales Técnicas, herramientas y experiencias de e-Research e investigación en colaboración*. La Laguna: Sociedad Latina de Comunicación Social.

Arcila, C. & Said, E. (2012). Factores que inciden en la variación de seguidores en los usuarios Top 20 más vistos en Twitter en América Latina y Medio Oriente. *Interciencia*, *37*(12), pp. 875-882.

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*, pp. 37–50.

Arora, S. et al. (2013). *A Practical Algorithm for Topic Modeling with Provable Guarantees*. Paper presentado en la 30th International Conference on Machine Learning (ICML). Atlanta, EEUU. Recuperado de http://jmlr.org/proceedings/papers/v28/arora13.html

Asfari, O. et al. (2013). Ontological Topic Modeling to Extract Twitter users' topics of interest. Paper presentado en *The 8th International Conference on Information Technology and Applications (ICITA 2013)*, Sydney, Australia.

Blei, D. & Lafferty, J. (2006). *Dynamic topic models. En Cohen*, W. & Moore, A. (Eds.) *Proceedings of the 23nd International Machine Learning Conference*. Corvallis: Omni Press.

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

Blei, D. (2012b). Introduction to probabilistic topic models. Communications of the ACM, pp. 77–84.

Blei, D. & McAuliffe, J. (2007). Supervised topic models. *Neural Information Processing Systems*, 21, pp. 1-8.

Blum, A. (2003). Machine learning theory. FOCS 2003 The 44th Annual IEEE Symposium on Foundations of Computer Science.

Bogdanov, P. et al. (2013). The Social Media Genome: Modeling Individual Topic-Specific Behavior in Social Media. Paper presentado en *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining ASONAM*, Niagara Falls, Canada.

Boyd, D. & Crawford, K. (2011). Six provocations for big data, ponencia presentada en el simposio "A decade in internet time: symposium on the dynamics of the internet and society" del Oxford Internet Institute, 21 de septiembre de 2011.
http://softwarestudies.com/cultural_analytics/Six_Provocations_for_Big_Data.pdf

Boyd, D. & Ellison, N. (2008). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*(1), pp. 210-230.

Boyd, D., Golder, S. & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *HICSS-43. IEEE: Kauai, HI*.

Brett, M. (2012). Topic Modeling: A Basic Introduction. *Journal of Digital Humanities*, *2*(1). Recuperado de http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/

Burdick, A. (2012). *Digital Humanities*. Cambridge: MIT Press.

Caballero, U. (2000). "Periódicos mexicanos en internet", *Revista Universidad de Guadalajara*. Recuperado de: http://www.cge.udg.mx/revistaudg/rug22/rug22dossier3.html, 10-06-2012.

Cai, K. et al. (2010). Leveraging sentiment analysis for topic detection. *Web Intelligence and Agent Systems: An International Journal*, 8, pp. 291–302.

Cambria, E. et al. (2013). *Knowledge-Based Approaches to Concept-Level Sentiment Analysis*. *Intelligent Systems, IEEE, 28*(2), pp. 15-21.

Collett, S. (2013). Big data, big storage: the era of big data requires new storage strategies. *Computerworld*, *47*(17), pp. 14-18.

Comisión Europea (2007). Comunicación de la comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. Un planteamiento europeo de la alfabetización mediática en el entorno digital. Recuperado de http://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:52007DC0833&from=ES, 27-03-2014.

ComScore (2013). Futuro Digital Colombia 2013. Recuperado de https://www.comscore.com/lat/Insights/Presentations-and-Whitepapers/2013/2013-Colombia-Digital-Future-in-Focus, 01-04-2014.

Conover, M. et al. (2011). Political polarization on Twitter. En N. Nicolov & Shanahan, J. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, España.

Cheng, A. et al. (2008). Advancing social science research by applying computational linguistics. *Proceedings of the American Society for Information Science and Technology*, *45*(1), pp. 1-12.

Cukier, K. & Mayer-Schöenberger, V. (2013). The Rise of Big Data. *Foreign Affairs*, *92*(3).

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

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

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), pp. 64-73.

Dietterich, T. G. (2003). *Machine Learning*. En *Nature Encyclopedia of Cognitive Science*. London: Macmillan.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), p78-87.

Emery, S. et al. (2014). Are you scared yet? Evaluating fear appeal messages in tweets about the Tips Campaign. *Journal of Communication*, 64, pp. 278-295.

Escobar, J. & Cuervo, Á. (2008). Validez de contenido y juicio de expertos: una aproximación a su utilización. *Avances en medición*, 6, pp. 27-36.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), pp. 82-89.

Felten, E. (2008, junio 26). The End of Theory? Not Likely. Freedom to tinker. Recuperado de https://freedom-to-tinker.com/blog/felten/end-theory-not-likely/, 10/04/2014.

Ferrari, L. et al. (2011). Extracting urban patterns from location-based social networks. En I. Cruz & D. Agrawal. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social* Network, pp. 9–16.

Fernández, I. et al. (2012). Language use depending on news frame and immigrant origin. *International Journal of Psychology*, DOI:10.1080/00207594.2012.723803.

García, E. et al. (2011). Uso de Twitter y Facebook por los medios Iberoamericanos. *El profesional de la información*, *20*(6), pp. 611–620.

García de Torres, E. et al. (2008). Las herramientas 2.0 en los diarios españoles 2006-2008: tendencias. *PRISMA.COM*, 7, pp. 193-222.

Gerber, M. (2014). Predicting Crime Using Twitter and Kernel Density Estimation. *Decision support systems*, *61*, pp. 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), pp. 90–102.

Gil de Zúñiga, H. (2012). Social Media Use for News and Individuals' Social Capital, Civic Engagement and Political Participation. *Journal of Computer-Mediated Communication*, *17*(3), pp. 319-336.

Gobble, M. (2013). Big data: the next big thing in innovation. *Research Technology Management*, *56*(1), pp. 64-66.

Golder, S. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, *333*, pp. 1878 – 1881.

Graham, M. (2012). Big data and the end of theory? DATABLOG. Recuperado de http://www.theguardian.com/news/datablog/2012/mar/09/big-data-theory, 07/04/2014.

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

Gruzd, A. & Roy, J. (2014). Investigating Political Polarization on Twitter: A Canadian Perspective. *Policy & Internet*, *6*(1), pp. 28–45.

Guul, S. & Islam, S. (2013). Adoption of social media by online newspapers of Kashmir. *Annals of Library and Information Studies*, *60*, pp. 56 – 63.

Han, J. & Kamber, M. (2006). *Data mining. Concepts and techniques*. San Francisco: Morgan Kauffman Publishers.

Hand, D., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining*. Cambridge: MIT Press.

Hanna, A. (2013). Computer-aided content analysis of Digitally enabled movements. *Mobilization: An International Quarterly*, *18*(4), pp. 367-388.

Hardwood, T. & Garry, T. (2003). An overview of content analysis. *The Marketing Review*, *3*, pp. 479-498.

Harlow, S. & Johnson, T. (2011). Overthrowing the Protest Paradigm? How The New York Times, Global Voices and Twitter Covered the Egyptian Revolution. *International Journal of Communication*, *5*, pp. 1359–1374.

Ha-Thuc, V. et al. (2009). A Relevance-based Topic Model for News Event Tracking. En J. Allan & J. Aslam. *SIGIR '09 Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 764-765.

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

Hong, L. et al. (2012). Discovering Geographical Topics In The Twitter Stream. Paper presentado en *WWW 2012*, Lyon, Francia.

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), pp. 91-101.

Java, A. et al. (2007). Why We Twitter: Understanding Microblogging Usage and Communities. En H. Zhang et al. *Procedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007*, pp. 56 – 65.

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

Kalina, J. (2013). Highly robust methods in Data Mining. *Serbian Journal of Management*, *8*(1), pp. 9-24.

Kang, B., O'Donovan, J. & Höllerer, T. (2012). Modeling Topic Specific Credibility in Twitter. En C. Duarte & L. Carriço. *IUI '12 Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pp. 179-188.

Kaur, G. & Singh, L. (2011). Data mining: An Overview. *IJCST*, *2*(2), pp. 336-339.

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

Kim, D. & Oh, A. (2011). Topic chains for understanding a news corpus. En A. Gelbukh. *CICLing'11 Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*, pp. 163-176.

Knapp, S. & Michaels, W. (1982). Against Theory. *Critical Inquiry*, *8*(4), pp. 723-742.

Kosciejew, M. (2013). The Era of Big Data. *Feliciter*, *59*(4), pp. 52-55.

Krippendorff, K. (2004). *Content analysis. An introduction to its methodology*. Los Angeles: Sage Publications.

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

Leetaru, K. (2012). *Data mining methods for the content analyst. An introduction to the Computational Analysis of Content*. New York: Routledge.

Lewis, S., Zamith, R. & Hermida, A. (2013). Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *Journal of Broadcasting & Electronic Media*, *57*(1), pp. 34-52.

Lim, K., Chen, C. & Buntine, W. (2013). Twitter-Network Topic Model: A Full Bayesian Treatment for Social Network and Text Modeling. Paper presentado en *NIPS 2013*, Nevada, Estados Unidos.

Liu, B. (2010). *Sentiment Analysis and Subjectivity*. En Indurkhya, N. & Damerau, F. (Eds.) *Handbook of Natural Language Processing, Second Edition*, Boca Raton: Chapman and Hall/CRC.

Lotan, G. et al. (2011) The Revolutions Were Tweeted: Information Flows During the 2011 Tunisian and Egyptian Revolutions. *International Journal of Communication, 5*, pp. 1375–1405.

Lula, P. & Wójcik, K. (2011). Sentiment analysis of consumer opinions written in polish. *Economics & Management*, 16, pp. 1286-1291.

McNair, B. (1998). *The sociology of journalism*. London/New Delhi/New York/Sydney: Bloomsbury Academic.

McQuail, D. (2010). *McQuail's Mass Communication Theory*. Los Angeles/London/New Delhi/Singapore/Washington D.C.: SAGE Publications.

Marozzo, F. et al. (2013). A Cloud Framework for Big Data Analytics Workflows on Azure. *Cloud Computing and Big Data*. DOI: 10.3233/978-1-61499-322-3-182.

Mayer-Schönberger, V. & Cukier, K. (2013). *Big Data. La revolución de los datos masivos*. Madrid: Turnerr Publicaciones S.L.

McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit. Recuperado de http://mallet.cs.umass.edu

Meena, A. & Prabhakar, T. (2007). Sentence Level Sentiment Analysis in the Presence of Conjuncts Using Linguistic Analysis. En Amati, G., Carpineto, C. & Romano, G. (Eds.): ECIR 2007, LNCS 4425.

Mejova, Y. & Srinivasan, P. (2012). Political Speech in Social Media Streams: YouTube Comments and Twitter Posts. Paper presentado en *WebSci 2012*, Evanston, Illinois, USA.

Micó, J. et al. (2008). La ética en el ejercicio del periodismo: credibilidad y autorregulación en la era del periodismo en Internet. *Estudos em Comunicação*, 4, pp. 15-39.

Michelson, M. & Macskassy, S. (2010). Discovering users' topics of interest on Twitter: a first look. Paper presentado en *AND'10*, Toronto, Ontario, Canada.

Mitchell, T. (1997). *Machine learning*. New York: McGraw-Hill.

Mocanu, D. et al. (2013). The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *PLOS ONE*, *8*(4), pp. 1-9.

Murata, T. (2008). Detection of breaking news from online web search queries. *New Generation Computing*, 26, pp. 63-73.

Murphy, K. (2012). *Machine Learning. A Probabilistic Perspective*. Cambridge/ London: The MIT Press.

Murphy, M. & Barton, J. (2014). From a Sea of Data to Actionable Insights: Big Data and What It Means for Lawyers. *Intellectual Property & Technology Law Journal*, *26*(3), pp. 8-17.

Neuman, W. et al. (2014). The dynamics of public attention: Agenda-setting Theory meets Big Data. *Journal of Communication*, 64, pp. 193-214.

Newman, D. et al. (2006). Analyzing entities and topics in news articles using statistical topic models. En S. Mehrotra et al. *ISI'06 Proceedings of the 4th*

IEEE international conference on Intelligence and Security Informatics, pp. 93-104.

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), pp. 6-22.

Noguera, J. (2010). Redes sociales como paradigma periodístico. Medios españoles en Facebook. *Revista Latina de Comunicación Social*, 65, pp. 176–186.

Nunan, D. & Di Domenico, M. (2013). Market research and the ethics of big data. *International Journal of Market Research*, *55*(4), pp. 2-13.

Panaccione, C. & Foltz, P. (2009). Implicit communication detection using topics model on asynchronous communication data. En *Proceedings of the NIPS workshop on Topic Models: Text and Beyond*. Recuperado de http://www.umiacs.umd.edu/~jbg/nips_tm_workshop/21.pdf.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), pp. 1–135.

Parks, M. (2014). Big Data in Communication Research: Its contents and discontents. *Journal of Communication*, 64, pp. 335-360.

Paul, M. & Dredze, M. (2014). Discovering Health Topics in Social Media Using Topic Models. *PLoS ONE*, *9*(8): e103408.

Pennacchiotti, M. & Popescu, A. (2011). *A Machine Learning Approach to Twitter User Classification*. En *Proceedings of the Fifth International Conference on Weblogs and Social Media*. Menlo Par: The AAAI Press.

Peslak, A., Ceccucci, W. & Sendall, P. (2010). An empirical study of social networking behavior using Diffusion of Innovation Theory. En *Conference on Information Systems Applied Research 2010 CONISAR Proceedings*, Nashville Tennessee, USA.

Rajani, N. (2014). Topic discovery in microblogs. En S. Geva & A. Trotman. *SIGIR 2014. The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Gold Coast, QLD, Australia.

Ramage, D., Dumais, S. & Liebling, D. (2010). Characterizing microblogs with topic models. En M. Hearst. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Washington, DC, Estados Unidos.

Rätsch, G. (2004). *A Brief Introduction into Machine Learning*. Ponencia presentada en el XXI Chaos Communication Congress, Berlin, Alemania. Recuperado de http://events.ccc.de/congress/2004/fahrplan/files/105-machine-learning-paper.pdf

Recuero, R., Araujo, R. & Zago, G. (2011). How does social capital affects retweets? *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, ICWSM 11*, Barcelona.

Rich, T. (2012). Deciphering North Korea's nuclear rhetoric: An automated content analysis of kcna news. *Asian Affairs*, *39*(2), pp. 73-89.

Richards, N. & King, J. (2014). Big Data Ethics. *Wake Forest Law Review*, 2014. Disponible en: http://ssrn.com/abstract=2384174

Ritter, R., Preston, J. & Hernández, I. (2013). Happy Tweets: Christians Are Happier, More Socially Connected, and Less Analytical Than Atheists on Twitter. *Social Psychological and Personality Science*, *00*(0), pp. 1–7.

Rodríguez-Martínez, R., Codina, Ll. & Pedraza-Jiménez, R. (2010). Cibermedios y web 2.0: modelo de análisis y resultados de aplicación. *El profesional de la información*, *19*(1), pp. 35–44.

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

Rogers, E. & Seidel, N. (2002). Diffusion of News of the Terrorist Attacks of September 11, 2001. *Prometheus*, *20*(3), pp. 209-219.

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

Rojas, H (2012). Encuesta Comunicación Colombia 2012. Nuevas tecnologías de comunicación. *Boletín de prensa del Centro de investigación en Comunicación política – Universidad Externado de Colombia*, pp. 1-18.

Romero, E. (2014). *Ciencias Sociales y Humanidades Digitales: una visión introductoria*. En Romero, F. & Sánchez, M. (eds.) *Ciencias Sociales y Humanidades Digitales Técnicas, herramientas y experiencias de e-Research e investigación en colaboración*. La Laguna: Sociedad Latina de Comunicación Social.

Ross, C. (2012). *Social media for digital humanities and community engagement*. En Warwick, C., Terras, M. & Nyhan, J. (Eds.) *Digital humanities in practice*. Londres: Facet Publishing.

Said, E. et al. (2013). Ibero-American Online News Managers' Goals and Handicaps in Managing Social Media. *Television and New Media*, *4*(2).

Said, E. & Arcila, C. (2011). Los cibermedios en América Latina y la Web 2.0. *Comunicar*, 37, pp. 125–131.

Said, E.; Arcila, C. & Méndez, J. (2011). Desarrollo de los cibermedios en Colombia. *El profesional de la información*, *20*(1), pp. 47–53.

Schaal, M., O'Donovan, J. & Smyth, B. (2012). An Analysis of Topical Proximity in the Twitter Social Graph. En K. Aberer et al. (Ed.) *Social Informatics. 4th International Conference, SocInfo 2012*, Lausanne, Suiza.

Schultz, B. (2012). New brand: The rise of the independent reporter through Social Media. *Online Journal of Communication and Media Technologies*, *2*(3), pp. 93–112.

Stassen, W. (2010). Your news in 140 characters: exploring the role of social media in journalism. *Global Media Journal: African Edition*, *4*(1), pp. 1-16.

Stubbs, M. (2001). *Words and phrases: Corpus Studies of Lexical Semantics.* Blackwell: Oxford.

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), pp. 217-247.

Telja, C. (2011). The uses of social media in the swedish online newspaper Aftonbladet - a case study. En *Media in Transition 7 Conference, MiT7 2011. Massachusetts Institute of Technology*, Boston, USA.

Téllez, A., Montes, M. & Villasenor, L. (2009). Using Machine Learning for Extracting Information from Natural Disaster News Reports. *Computación y Sistemas*, *13*(1), pp. 33-44.

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)*, Philadelphia, July 2002, pp. 417-424.

Two Crows Corporation. (1999). *Introduction to Data Mining and Knowledge Discovery*. Potomac: Two Crows Corporation.

Thelwall, M., Wilkinson, D. & Uppal, S. (2010). Data Mining Emotion in Social Network Communication: Gender Differences in MySpace. Journal of the American Society for Information Science and Technology, *61*(1), pp. 190–199.

Ure, M. & Parselis, M. (2013). Argentine Media and Journalists Enhancing and Polluting of Communication on Twitter. *International Journal of Communication*, *7*, pp. 1784–1800.

Verbeke, M. et al. (2014). When Two Disciplines Meet, Data Mining for Communication Science. Ponencia presentada en el 64th Annual ICA Conference, Seattle.

Vinodhini, G. & Chandrasekaran, R. (2012). Sentiment Analysis and Opinion Mining: A Survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, *2*(6), pp. 282-292.

Wallach, H, et al. (2009). *Evaluation Methods for Topic Models*. *Proceeding ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning*. New York: ACM.

Wanamaker, B. & Bean, D. (2013). Big Data: The end of theory in healthcare? Clayton Christensen Institute. Recuperado de http://www.christenseninstitute.org/big-data-the-end-of-theory-in-healthcare/, 15/05/2014.

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), pp. 5-23.

West, M. (2001). *Theory, Method, and Practice in Computer Content Analysis.* Progress in Communication Sciences. Westport: Ablex Pub.

Wing, J. (2006). Computational thinking. *Communications of the ACM*, *49*(3), pp. 33-35.

White, M. & Marsh, E. (2006). Content Analysis: A Flexible Methodology. *Library trends*, *55*(1), pp. 22-45.

White, M. (2013, noviembre 8). How Big Data Is Changing Science (and Society). Pacific Standard. Recuperado de http://www.psmag.com/navigation/nature-and-technology/big-data-changing-science-society-69650/, 07/04/2014.

Yang, X. et al. (2013). A biterm topic model for short texts. En D. Schwabe, V. Almeida & H. Glaser. *WWW '13 Proceedings of the 22nd international conference on World Wide Web*, Rio de Janeiro, Brasil.

Zhao et al. (2011). *Comparing Twitter and Traditional Media Using Topic Models*. En *Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR*, 2011, Dublin, Ireland, April 18-21, 2011. Proceedings. Berlin: Springer.

Zhu, Q. & Li, F. (2013). Topic Detection from Social Media and News Media. Paper presentado en *Workshop on understanding the positive and negative sides of social media*. Beijing, China.

Zikopoulos, P. et al. (2012). *Harness the power of Big Data*. New York: McGraw-Hill.