Universidad Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Caracterización de usuarios y propagación de mensajes en Twitter en el entorno de temas sociales

Autora:
Mariluz Congosto Martínez

Director:
Luis Sánchez Fernández

DEPARTAMENTO DE INGENIERÍA TELEMÁTICA

Leganés, enero 2016

TESIS DOCTORAL

Caracterización de usuarios y propagación de mensajes en Twitter en el entorno de temas sociales

Autora: Mariluz Congosto Martínez

Director: Luis Sánchez Fernández

Firma del Tribunal Calificador:

Firma

Presidente: Rafael Rubio Núñez

Vocal: Daniel Gayo Avello

Secretario: Carlos García Rubio

Calificación:

Leganés, 28 de enero de 2016

Dichosa edad y siglos dichosos aquellos a quien los antiguos pusieron nombre de dorados, y no porque en ellos el oro, que en esta nuestra edad de hierro tanto se estima, se alcanzase en aquella venturosa sin fatiga alguna, sino porque entonces los que en ella vivían ignoraban estas dos palabras de *tuyo* y *mío*. Eran en aquella santa edad todas las cosas comunes.

El Quijote, capítulo XI

Miguel de Cervantes

AGRADECIMIENTOS

Al Dr. Luis Sánchez por su labor como director de tesis.

A la Dra. Damaris Fuentes-Lorenzo por su inestimable apoyo y por sus consejos.

Al Dr. Jesús Arias por el tiempo que me dedicó para mejorar el entorno de desarrollo con el que llevé a cabo los experimentos.

Al Dr. Pablo Basanta por las ideas que me aportó en los métodos de procesado de los experimentos.

Al Dr. José Antonio Pozas, el mejor jefe que tuve, por sus consejos y comentarios.

A Luis Marmisa, mi gurú técnico, por todos los buenos consejos que he recibido siempre de él.

A Jonathan Almodóvar por cuidar de los servidores en los que he ido recogiendo los datos durante estos años.

A mi familia por el cariño, comprensión y apoyo que me han dado en todo momento.

A todos mis amigos que me han animado para no desfallecer.

RESUMEN

La Web, que nació bajo el espíritu de la colaboración y la libertad de información frente al modelo vigente de competitividad y derechos de propiedad, ha ido evolucionando hasta nuestros días de una manera que nadie podría haber supuesto. Actualmente, la sociedad está fuertemente conectada y comparte una gran cantidad de información pero no lo hace de una forma distribuida, como se esperaba, sino centralizada desde plataformas conocidas como *redes sociales*. El lado positivo de esta concentración es la facilidad para obtener los datos de interacción social. De todas las redes sociales, Twitter se ha caracterizado por su carácter abierto tanto en sus contenidos como en el acceso a sus datos mediante APIs, y aunque el caudal completo de sus datos no está accesible de forma gratuita, es hoy por hoy la fuente más importante de datos sociales de la que disponen los investigadores en Internet.

Esta tesis aborda el análisis de la propagación de mensajes en Twitter en temas sociales y el papel que desempeñan las personas en la difusión. El enfoque se realiza desde un análisis empírico a través de un conjunto de casos de estudio con diferentes dimensiones, duraciones y contextos.

Para poder abordar esta investigación he diseñado la plataforma T-hoarder que captura los mensajes que publican los usuarios de Twitter, los analiza y visualiza los resultados, permitiendo detectar los momentos más virales y los usuarios más destacados. Esta plataforma dispone de mecanismos de procesado por partes y su posterior integración, gracias a los cuales ha funcionado continuamente durante más de cuatro años sin problemas de escalabilidad. Desde ella he podido observar más de cuarenta casos relacionados con los acontecimientos de impacto social, los movimientos sociales, las elecciones en España, las tendencias en Twitter y la relación entre Twitter y Televisión.

Basándome en las observaciones en sucesivos experimentos y mediante un proceso de refinamiento, he establecido la clasificación de usuarios que se presenta en esta tesis. Esta clasificación se valida con distintas métricas en las que la agrupación de los tipos de usuarios es coherente. Por otro lado, he definido los atributos de Alcance, Difusión, Participación, Incorporación y Automatismo y los he medido cada hora para cada uno de los casos. Las correlaciones encontradas para estos atributos, salvo el Automatismo, respecto al número de tuits publicados en cada intervalo de tiempo son muy altas en la mayoría de los casos.

Macroscópicamente he encontrado una burbuja de actividad en todos los casos en la que el 80% de los mensajes difundidos fueron publicados por una minoría y los causantes del 80% de la propagación formaron grupos reducidos de usuarios. Analizados año a año los casos de estudio de duración superior a los dos años he descubierto que cada año va aumentando el porcentaje de retransmisiones mientras que el tamaño de los grupos que las producen disminuye.

Un rasgo de meritocracia descubierto ha sido que la capacidad de propagación de mensajes de un usuario no depende la estructura de su red.

ABSTRACT

The Internet, which was born in the spirit of collaboration and freedom of information in the presence of the prevailing model of competitivity and copyright has evolved until now in a way which nobody could have imagined. Nowadays we are all very connected and we share a lot of information but we do not do it in a distributed way, as may have been imagined, but in a more centralised way from platforms we call Social Media. The positive part of this concentration is the ease with which social interaction data can be gathered. Of all the Social Media, Twitter stands out for its openness both in content and in accessibility of data through API's and although the complete flow of data is no available freely, it is today, the most important source of social data that researchers have on the Internet.

This thesis tackles the analysis of propagation of messages on Twitter about social issues and the role carried out by people in the spread. The focus was done by experiential analysis through eighteen case studies of different dimensions, durations and contexts. In order to carry out this research I have designed the T- hoarder platform which captures messages posted by Twitter users, analyses and visualises the results, allowing the detection of the most viral moments and the most prominent users. This platform has process mechanisms, and its later integration, thanks to which it has functioned continually for over four years without any scale problems. From it, over forty cases related to social impact, social movements, the elections in Spain, trends on Twitter and the relation between Twitter and television.

Based on observations in successive experiments and through a process of refinement, I have established the classification of users which is presented in this thesis. This classification was validated using distinctive metrics in which the grouping of types of user is coherent. On the other hand, I have defined the attributes of Reach, Diffusion, Participation, Incorporation and Automation and I have measured them every hour in each of the cases. The correlations found for these attributes, except for the Automation, with respect to the number of tweets posted at each interval of time are very high in the majority of cases.

Macroscopically I have found a bubble of activity in all the cases in which 80% of the messages spread were posted by a minority and the source of 80% of the spread formed small groups of users. Analyzing year on year the cases studied for a duration of over two years I have discovered that each year the percentage of retweets is increasing while the size of the groups producing them is falling.

A feature of meritocracy uncovered has been that the capacity of spread of messages of a user is not related to his/her network.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABLAS

PARTE I. Introducción y estado del arte

1 INTRODUCCIÓN

El nacimiento de la Web (Berners-Lee, Cailliau, Luotonen, Frystyk Nielsen, & Secret, 1994) a principios de los años 90 supuso una revolución en la forma de difundir contenidos. La Web nació con un espíritu de colaboración y libertad de información, frente al modelo vigente de competitividad y derechos de propiedad (Castells, 2002). La simplicidad del lenguaje HTML[1] y del protocolo HTTP[2] facilitó el crecimiento exponencial de páginas web, quedando accesibles a todas las personas conectadas a Internet. A pesar de que la tecnología estaba poco desarrollada, se crearon unas expectativas exageradas de negocio en las empresas vinculadas a Internet que dieron lugar a la burbuja "puntocom" ente los años 1997 y 2001.

Cuando se desinfló la burbuja desaparecieron muchas empresas, pero no la tecnología; diversas técnicas se fueron desarrollando durante esta etapa para hacer la Web mucho más dinámica e interactiva. Al margen de las especulaciones financieras de las empresas de Internet, otras personas estaban pensando en cómo sería la Web del futuro. En 1999, Darcy DiNucci mencionó por primera vez el término Web 2.0 (DiNucci, 1999) anticipando muchos de los cambios que surgirían años más tarde. Pero los que realmente impulsaron el concepto Web 2.0 fueron Dale Dougherty y Tim Reilly, que iniciaron las Conferencias sobre la Web 2.0 en octubre de 2004.

En torno al concepto Web 2.0 se agruparon un conjunto de características que Markus Angermeier (Angermeier, 2005) resumió en el mapa conceptual[3] de la Figura 1.

Figura 1 Mapa conceptual de la Web 2.0

[1] http://www.w3.org/html/
[2] http://www.w3.org/Protocols/
[3] http://kosmar.de/archives/2005/11/11/the-huge-cloud-lens-bubble-map-web20/

En la etapa inicial de la Web 2.0 surgieron comunidades de desarrolladores independientes que crearon servicios interactivos y abiertos que facilitaban la interacción entre personas. La Web se hizo más social, manteniendo los principios de colaboración y libertad de contenidos. El usuario se convirtió en *prosumidor,* es decir, en consumidor y productor de contenidos a la vez. Algunos de los frutos de estos trabajos fueron las plataformas de *blogs*, de *microblogging*, y las *redes sociales*.

En el inicio de la Web 2.0 las aplicaciones más populares fueron los *blogs*, que eran espacios personales donde se vertían opiniones, conocimientos, informaciones o experiencias y que recibían comentarios de sus lectores. Se establecían conexiones entre los *blogs* mediante citas o recomendaciones para su lectura. Esta red de *blogs* se denominó *blogosfera* y surgieron buscadores específicos para ella como Technorati[4]. Se crearon muchas plataformas de *blogs*, algunas de ellas funcionando actualmente, como Blogger, WordPress, o TypePad[5]. Fueron muy populares las plataformas españolas Blogia, La coctelera, Nireblog o Bitácoras.com[6].

Las plataformas de *microblogging* partían del mismo concepto que las de los *blogs*, pero con publicaciones más breves. Aunque algunas de estas plataformas existían desde el año 2006, no empezaron a ganar adeptos hasta el año 2010 (Perrin et al., 2015). La causa de su despliegue coincidió con la popularización de los teléfonos inteligentes y la aparición de las tarifas planas para el acceso a Internet. El acceso en movimiento a la información se realizaba de forma intermitente, aprovechando trayectos o tiempos muertos. La brevedad de los mensajes de *microblogging* encajaba perfectamente en este nuevo escenario. Entre las plataformas más famosas se encuentran Twitter, Tumblr, Weibo o Plurk[7].

Las *redes sociales* crearon espacios para que las personas contactasen y compartieran información. Las relaciones podían ser profesionales (LinkedIn[8]) o por amistad (Facebook, Orkut, MySpace o Tuenti)[9].

Las plataformas de *microblogging* y *redes sociales* permitían estar informado y compartir información con menos esfuerzo que los *blogs*. Por este motivo, muchos usuarios que no tenían *blog* se incorporaron a estas plataformas y los que sí lo tenían fueron bajando su ritmo de publicación a la vez que aumentaba su participación en *microblogging* y *redes sociales*.

El panorama actual nos muestra una concentración de plataformas, tanto en *blogs* como en *microblogging* y *redes sociales*. La cantidad de datos generados es tan ingente que ha dado lugar a la era del *big data* (Manyika, Chui, Brown, Bughin, & Dobbs, 2011), donde son necesarias nuevas tecnologías para tratar un enorme volumen de datos complejos en tiempo real. Nunca hasta la fecha se ha estado en mejor condición de poder analizar el comportamiento social con los datos generados por cientos de millones de personas en

[4] http://technorati.com/
[5]Blogs: https://www.blogger.com, https://es.wordpress.com/, http://www.typepad.com/,
[6]Blogs Españoles: https://www.blogia.com/, la coctelera y Nireblog cerraron y http://bitacoras.com/ se reconvirtió en una red social de blogs
[7]Microblogging: https://twitter.com/, https://www.tumblr.com/, http://www.we bo.com/ (*microblogging* Chino), http://www.plurk.com/
[8] Red social profesional https://www.linkedin.com/
[9]Redes sociales de amigos: https://www.facebook.com/, https://orkut.google.com/, https://myspace.com/, https://www.tuenti.com

todo el mundo, pero la mayoría de estos datos están en manos de unas pocas compañías privadas.

Compañías como Facebook o LinkedIn, por razones de privacidad, limitan el acceso a la información de sus usuarios en función de si están o no conectados. Pero las cláusulas que hay que aceptar cuando se abre un perfil en estas redes les permiten vender publicidad o servicios segmentados. Estas empresas utilizan la información de los perfiles e interacciones de las personas en su red con fines comerciales, mientras que el acceso de los investigadores a esta información está muy limitado y estrictamente controlado por estas compañías. En el caso de Twitter, la información que generan la mayoría de sus usuarios es pública, y se puede acceder a ella a través de su API; sin embargo, el caudal completo de mensajes generados no es accesible de forma gratuita, solo una parte. Aun así, hoy por hoy, Twitter es la fuente más importante de datos sociales de la que disponen los investigadores.

Twitter se caracteriza por limitar la longitud de sus mensajes, llamados tuits (*tweets*), a 140 caracteres, y por las relaciones asimétricas entre sus usuarios. Actualmente se publican 500 millones de tuits al día[10]. Desde el 2008 ha sido fuente de investigaciones (Huberman, Romero, & Wu, 2008) y se ha aplicado en el estudio de diferentes ámbitos como elecciones políticas (Conover, Gonc, Ratkiewicz, Flammini, & Menczer, 2011) (Gayo-Avello, 2011b) (Barberá & Rivero, 2012), movimientos sociales (Peña-lópez, Congosto, & Aragón, 2014), predicciones (Bollen, Mao, & Zeng, 2011) (Bollen & Pepe, 2009) (Asur & Huberman, 2010), influencia de usuarios (Cha & Gummadi, 2010), comportamiento (Dodds, Harris, Isabel, Bliss, & Danforth, 2011) o propagación de mensajes (De Domenico, Lima, Mougel, & Musolesi, 2013). Twitter es, por consiguiente, una fuente de incalculable valor para poder estudiar las dinámicas sociales de comunicación.

El objetivo de esta tesis es el estudio de la propagación de mensajes en Twitter en temas sociales y el papel que desempeñan las personas en la difusión. El enfoque se realiza desde un análisis empírico a través de un conjunto de casos de estudio. Para poder abordar esta investigación he diseñado T-hoarder[11], una plataforma que captura tuits, los analiza y visualiza los resultados, permitiendo detectar los momentos más virales. Con esta plataforma, que ha funcionado continuamente durante más de cuatro años, he podido observar más de cuarenta casos relacionados con los acontecimientos de impacto social, los movimientos sociales, las elecciones, las tendencias en Twitter y la relación entre Twitter y los programas de televisión. Basándome en estas observaciones he definido una clasificación de tipos de usuarios y unas métricas de propagación que he aplicado a dieciocho casos de estudio seleccionados. Por la larga duración del periodo analizado es posible obtener conclusiones sobre los cambios de las dinámicas de comunicación que se han producido en estos últimos años, pero no solo en sucesos aislados, también de una forma continua.

[10] About Twitter: https://web.archive.org/web/20151005104907/https://about.twitter.com/company
[11] T-hoarder: http://t-hoarder.com/

1.1 MOTIVACIÓN

Toda investigación, independientemente de su área de conocimiento, siempre tiene la finalidad de entender un fenómeno para mejorar la sociedad. En mi caso, aportar luz a los mecanismos de propagación y al papel que juegan las personas en su difusión me parece una contribución necesaria para comprender la propagación de ideas que dan lugar a los cambios sociales.

Manuel Castells, en su libro "Comunicación y Poder" (Castells, 2009), afirmaba que «*es mediante la comunicación como la mente humana interactúa con su entorno social y natural*». La comunicación es, consecuentemente, un factor esencial para la difusión de ideas y la creación de opinión. Las redes sociales han abierto una puerta para comunicar a las personas sin intermediarios y esto posibilita nuevos caminos para que la información fluya por otros cauces diferentes de los tradicionales. Aunque los medios de comunicación siguen marcando la agenda, la comunicación ya no está solo en sus manos y la transformación social incluye nuevos parámetros que aún no han sido estudiados en profundidad.

Vivimos unos tiempos de cambios acelerados y el estudio de las dinámicas de propagación de la información puede ayudar a comprender qué capta la atención de las personas, qué corrientes de opinión van ganando adeptos, cómo influyen unas personas sobre otras, cómo se polarizan las personas ante temas conflictivos e incluso cómo es de natural o forzada esta comunicación.

1.2 PROPÓSITO Y OBJETIVOS

El propósito principal de esta tesis es contribuir al conocimiento de los factores que facilitan la propagación de mensajes. Aunque Twitter tiene mucha actividad relacionada con temas de entretenimiento (TV, música, deportes, etc.) o en marketing de productos y servicios, el foco lo he dirigido hacia los datos sociales, políticos y socio-políticos. Las opiniones de los usuarios sobre temas relacionados con la política, los movimientos sociales o los sucesos que crean alarma social suponen la información base de este estudio.

El objetivo final de esta investigación es obtener, mediante métodos cuantitativos, conclusiones sobre cómo se difunden estos datos sociales, qué agentes actúan de catalizadores, cómo evolucionan con el tiempo y si existe alguna manipulación en esta difusión.

Las tareas parciales para conseguir el objetivo final son las siguientes:

> **Tarea 1**: *Definir roles de usuarios según su actividad en Twitter para analizar la propagación desde el punto de vista de las personas que participan*.
>
> ***Tarea 2**: Definir los parámetros para medir la propagación de los mensajes que permitan comparar distintos casos de estudio y ver su evolución en el tiempo.*
>
> **Tarea 3**: *Obtener de Twitter casos de estudio relacionados con temas sociales que cumplan los requisitos de persistencia (crisis, política, movimientos sociales,*

escándalos, etc.) o que sean recurrentes (campañas electorales, congresos de partidos políticos, etc.).

Tarea 4: *Medir para cada caso de estudio la participación de cada uno de los roles, los parámetros de propagación definidos y su evolución en el tiempo.*

1.3 METODOLOGÍA

Twitter produce una ingente cantidad datos que dificulta el análisis cualitativo precisando de la ayuda de métodos cuantitativos. En este contexto, los algoritmos cobran una vital importancia para estructurar información partiendo de unos datos fragmentados y desestructurados. Pero los algoritmos no siempre bastan, algunas veces es necesario dar forma a los datos mediante su visualización para poder detectar patrones de comportamiento y descubrir las partes que requieren un análisis más detallado. La metodología utilizada se basa en algoritmos de minería de textos con los que se extraen entidades, relaciones, propagaciones y localizaciones que cuando son visualizadas facilitan la comprensión de los datos y su contexto.

El caudal completo de información de Twitter no está disponible de forma gratuita. La única opción para acceder a él es contratarlo a través de la empresa Gnip[12], propiedad de Twitter. Anteriormente estos datos se podrían adquirir en DataSitf[13] o NTT Data[14], pero desde el 2015 ya no es posible. El coste del caudal total de los tuits está fuera del alcance de los investigadores. Afortunadamente, los datos sociales representan un subconjunto de ellos que puede ser obtenido mediante las APIs de Twitter.

Twitter fue pionera abriendo sus datos mediante APIs, estando disponibles desde el año 2008[15]. La API Search permitía realizar búsquedas de tuits que contuvieran palabras clave. La API REST daba acceso a toda la información de un usuario específico. Estas APIs funcionaban en modo síncrono, se solicitaba una información y la API la proporcionaba.

En el año 2009, Twitter comenzó a dar acceso, en tiempo real, a sus tuits. Primero permitió acceder al proverbial "*Firehose*" (el flujo total de tuits en tiempo real) pero al cabo de unos meses restringió su acceso por motivos técnicos. En su lugar creó la API Streaming[16], que proporcionaba sólo una parte de su caudal.

Las versiones de sus APIs han ido cambiando con el tiempo, modificando la forma de acceso, la cantidad de información por unidad de tiempo o integrando una API en otra. Hasta este momento han existido tres versiones:

- Twitter API V0.0: desde 2006 hasta junio de 2010.
- Twitter API V1.0: desde junio de 2010 hasta junio de 2013. Desaparece la autenticación básica (usuario/contraseña) en la API REST y se sustituye por OAuth.

[12] Gnip: https://gnip.com/
[13] Datasift: http://datasift.com/
[14] NTT data: http://www.nttdata.com/
[15] APIS de Twitter: https://web.archive.org/web/20090106032955/http://dev.twitter.com/
[16] API Streaming: http://hasin.me/2009/06/20/collecting-data-from-streaming-api-in-twitter

- Twitter API V1.1: desde junio de 2013 hasta la fecha. Integra la API Search en API REST, extiende la autenticación por OAuth a la API Streaming y modifica los límites de velocidad.

El acceso a cualquier API de Twitter se realiza por OAuth. Para ello es necesario dar de alta una aplicación en Twitter y obtener las claves de acceso para esa aplicación (*Consumer Key* y *Consumer Secret*)[17]. Por otro lado es necesario crear un conjunto de usuarios en Twitter y generar sus claves de acceso (*access token),* para esa aplicación. De esta forma, queda constancia en Twitter sobre qué usuario accede y desde qué aplicación.

Twitter proporciona desde la versión V1.1 dos APIs[18] para obtener información: la API REST y la API Streaming. La API REST permite realizar todo tipo de consultas a los datos de Twitter de una forma síncrona. La API Streaming establece un *socket* entre Twitter y un servidor por el que se recibe información de forma asíncrona. En ambas APIs hay que tener en cuenta las restricciones que impone Twitter en su uso:

- **Límite de velocidad**: restringe el número de consultas a la API REST durante quince minutos. El límite depende del tipo de consulta que se haya realizado. En el caso de la API Streaming existe una limitación de 50 tuits por segundo (TPS).
- **Límite temporal**: solo es posible recuperar los tuits desde siete días previos con el método GET /search/tweets de la API REST, los 3.200 últimos tuits de un usuario con el método GET /statuses/user_timeline de la API REST y ningún tuit previo con la API Streaming.
- **Muestra de tuits**: tan solo el método GET /statuses/user_timeline de la API REST proporciona la muestra completa de los últimos 3.200 mensajes del usuario. El método GET /search/tweets de la API REST y la API Streaming proporcionan un porcentaje de los tuits (85%-95%), incluso cuando la frecuencia de mensajes es inferior a 50 tuits por segundo (TPS).

En la Tabla 1 se resumen las características de los métodos para obtener tuits.

API	Búsqueda histórica	Tipo de búsqueda	Tuits servidos
Streaming	0	Palabras clave o usuarios o localizaciones	Si no supera los 50 tuits por segundo, se recupera entre el 80%-90% de los tuits.
REST (GET search)	1 semana atrás	Palabras clave o usuarios o localizaciones	Desconocido, ya que suministra los tuits que considera relevantes. La cantidad es inferior a la API Streaming[19].
REST (GET statuses)	3200 últimos tuits	Tuits de un usuario	Todos

Tabla 1 Diferencias entre las APIs de Twitter

[17] Apps en Twitter: https://apps.twitter.com/
[18] Documentación de las APIs: https://dev.twitter.com/overview/documentation
[19] API Streaming vs. API Search: https://wiki.digitalmethods.net/Dmi/BiasTwitterSearchStream

La más idónea de las APIs de Twitter para obtener capturas de tuits a largo plazo es la API Streaming. Esta API permite obtener los tuits mediante filtrado o por muestra aleatoria.

El filtrado se puede realizar por una de estas tres formas:

1. **Por palabras clave**: se obtienen los tuits que se ajustan a un patrón de búsqueda proporcionando una lista de palabras o expresiones separadas por comas. Las expresiones son palabras separadas por espacios que si se encuentran en el mensaje, independientemente del orden en que aparezcan, cumplen la condición. Las comas que separan las palabras o expresiones actúan como el operador lógico OR y en una expresión los espacios que separan las palabras se comportan como el conector lógico AND. Por ejemplo, la búsqueda "#metromadrid, metro Madrid" proporcionaría los tuits que contengan "#metromadrid" OR ("metro" AND "Madrid").
2. **Por usuarios**: se adquieren los tuits de los usuarios suministrados mediante una lista con sus identificadores separados por comas. También se obtienen todas las reacciones a estos mensajes, bien sean retransmisiones (RTs) automáticas o manuales, menciones, citas o respuestas.
3. **Por geolocalización**: se reciben los tuits publicados desde una o varias zonas geográficas rectangulares, cada una de ellas delimitada por una coordenada superior este y otra inferior oeste. Con esta opción sólo se recogen mensajes *geolocalizados.*

Estas tres opciones son excluyentes, no es factible buscar por más de una opción a la vez. Cuando es necesario monitorizar por usuarios y por palabras clave, la única solución es capturar dos flujos de datos en paralelo para luego fusionarlos eliminando los mensajes duplicados. Este caso es muy frecuente en el seguimiento de campañas electorales en las que es interesante conocer la conversación en torno a los candidatos (usuarios) y los lemas de la campaña (lista de palabras clave).

La API de Twitter suministra los datos codificados en UTF-8[20], en formato JSON (Crockford, 2006) e incluye abundante información de contexto para cada tuit.

Inicialmente las APIs de Twitter podían usarse sin limitaciones. Esto favoreció la creación de servicios Web que recopilaban los datos como Tweetbackup.com o TwapperKeeper[21]. En aquellos momentos se podía obtener un *dataset* utilizando estos servicios, pero en septiembre de 2012 la nueva normativa de uso de las APIs impidió compartir tuits[22]. Estos servicios cerraron o liberaron su código para que pudiera ser reutilizado.

Actualmente, las normas de uso de las APIs de Twitter permiten compartir colecciones de datos de una forma limitada: «*If you provide Content to third parties, including downloadable datasets of Content or an API that returns Content, you will only distribute or allow download of Tweet IDs and/or User IDs. You may, however, provide export via non-automated means (e.g., download of spreadsheets or PDF files, or use of a "save as"*

[20] UTF8: http://en.wikipedia.org/wiki/UTF-8
[21] TwapperKeeper: https://twapperkeeper.wordpress.com/2011/02/22/removal-of-export-and-download-api-capabilities/
[22] Normas de uso de Twitter: https://dev.twitter.com/archive/terms/api-terms/diff-20130702

button) of up to 50,000 public Tweets and/or User Objects per user of your Service, per day» [23].

En esas circunstancias era necesario disponer de una infraestructura de servidores para recopilar los tuits. Para mis investigaciones sobre propagación de mensajes necesitaba una plataforma que permitiera, de forma gratuita, disponer de varias colecciones de datos, recoger tuits durante largos periodos de tiempo (años), poder localizar sucesos significativos para analizar propagaciones e identificar las fuentes de la información obtenida.

Encontré varias plataformas como TwapperKeeper[24], Twitter-Tap[25] o twitterstream-to-mongodb[26] que utilizaban bases de datos SQL o mongoDB. El uso de base de datos implicaba ciertos problemas cuando el volumen de datos iba incrementándose. Se necesita más almacenamiento, más capacidad de proceso. Los respaldos no podían realizarse del total de los datos sino por partes, era complicado mover o fusionar tablas de datos, etc.

Para llevar a cabo la **Tarea 3** no hallé una plataforma que funcionase a largo plazo con los recursos de almacenamiento que disponía, por ese motivo decidí crearla. El resultado fue T-hoarder, una plataforma con muy baja sobrecarga. Elegí Unix como sistema operativo y Python como lenguaje de programación para su desarrollo. Para el almacenamiento de la información descarté el uso de las bases de datos y opté por almacenar la información mediante el sistema de ficheros de Unix.

Adicionalmente, T-hoarder realiza el procesado y la visualización de los datos en tres ejes: temporal, espacial y de relevancia. El eje temporal muestra la evolución en el tiempo de un conjunto de indicadores, el eje espacial ubica los tuits geográficamente y el eje de relevancia descubre los mensajes más difundidos.

La descripción de la plataforma T-hoarder se realiza de forma detallada en el punto 4.1. Su desarrollo ha consumido una gran parte de los esfuerzos de la investigación pero gracias a él dispongo de un gran número de casos de estudio sobre los que medir la propagación y la participación de los usuarios. Hasta el momento T-hoarder ha capturado los tuits relacionados con todas las elecciones en España, desde las catalanas del año 2010 hasta las municipales de abril del 2015. He recogido el inicio de los movimientos sociales en mayo de año 2011, más conocidos como el movimiento 15M, las manifestaciones ante el Congreso de los diputados (25S), las protestas contra los recortes del presupuesto, la abdicación del rey Juan Carlos y la proclamación de Felipe VI, la independencia de Cataluña, los escándalos políticos como Bankia, el caso “Bárcenas”, los EREs de Andalucía, el caso “Madrid-Arena”, el caso “Pujol” y la operación Púnica, la alerta social que provocó el ébola y lo que se habla en el mundo sobre sostenibilidad y medio ambiente.

La **Tarea 4** también se describe en el punto 4.2 debido a que es necesario conocer el funcionamiento de T-hoarder para explicar cómo se ha llevado a cabo.

[23] https://dev.twitter.com/overview/terms/policy
[24] https://github.com/540co/yourTwapperKeeper
[25] https://github.com/janezkranjc/twitter-tap
[26] https://github.com/gdelfresno/twitterstream-to-mongodb

1.4 ESTRUCTURA DEL DOCUMENTO

Los capítulos de este documento se agrupan en estas cinco partes:

PARTE I. Introducción y Estado del arte*: Esta parte comprende los capítulos 1 y 2. Después de la introducción realizada en este capítulo, el capítulo 2 introduce el estado del arte de las investigaciones realizadas con los datos de Twitter.*

PARTE II. Requisitos*: Esta parte contiene el capítulo 3, en el que se definen los tipos de usuario (tarea 1), los indicadores de propagación (tarea 2) y el factor de ruido (tarea 3).*

PARTE III. Desarrollo*: Esta parte contiene el capítulo 4, dedicado a la descripción del desarrollo de la tarea 4 (obtención de las colecciones de datos, su procesado y su visualización), de la tarea 5 (medidas de participación de roles e indicadores de propagación).*

PARTE IV. Resultados, contribuciones, conclusiones y trabajos futuros*: Esta parte contiene el capítulo 5 con los resultados de la investigación, el capítulo 6 con las conclusiones, las contribuciones y los trabajos futuros.*

PARTE V. Apéndices, referencias y anexo*:* Esta parte contiene el apéndice A con el glosario de acrónimos y las definiciones. El apéndice B con los datos detallados de las mediciones. El apéndice C con las publicaciones realizadas. Las referencias y el Anexo I con todas las gráficas de las medidas. En la versión digital de la tesis este anexo aparece en documento separado por razones del tamaño del fichero.

2 ESTADO DEL ARTE

Twitter es una plataforma de *microblogging* en la que prima la brevedad. Para crear un perfil, los usuarios sólo tienen que elegir una manera de llamarse en esta red, aunque pueden opcionalmente añadir su nombre real, de dónde son o dónde viven o describirse en menos de 157 caracteres. A la hora de publicar mensajes (tuits) nunca pueden sobrepasar los 140 caracteres. Para comunicarse entre ellos se mencionan por su nombre de perfil precediéndolo por el símbolo "@". Los usuarios deciden a quién seguir, es decir, en quién están interesados. Un usuario puede seguir a otro sin tener que ser correspondido, esto da lugar a relaciones asimétricas y a que todos los tuits sean públicos. Existe la opción de que los tuits sean privados, pero se usa minoritariamente. Cada usuario puede ver ordenados de forma secuencial los tuits que emiten los usuarios que sigue (*timeline*).

Twitter fue ideada en el año 2006 para que la gente pudiera dar a conocer a sus amigos y conocidos lo que estaban haciendo, pudiendo interaccionar entre ellos mediante conversaciones (*reply*) o menciones. Pero a veces ocurre que una herramienta creada para un fin se transforma por su uso, como ocurrió con Twitter. Con el tiempo, los usuarios descubrieron que era más útil comentar lo que pasaba a su alrededor que hablar de sí mismos, convirtiéndose en observadores de su entorno. Este cambio dio lugar a que en el año 2009 Twitter cambiara la pregunta inicial ¿qué estás haciendo? por ¿qué está pasando? (Stone, 2009b).

A las personas les gusta saber qué está ocurriendo y contarlo a sus conocidos, por lo que encontraron en Twitter algo similar a un patio virtual de vecinos en el que se podía escuchar, comentar y repetir lo que por allí circulaba. Pero este patio virtual no es global, cada usuario está encerrado en su red, escucha las conversaciones de los que sigue y puede hacerse escuchar por los que le están siguiendo. En realidad, Twitter es un conjunto de patios virtuales, donde los contactos son muchas veces de la misma localidad (Gonzalez, Cuevas, & Cuevas, 2011) o tienen similares intereses (Conover et al., 2011). Por tanto, los mensajes circulan en entornos cerrados, teniendo que levantar mucho la voz para que un mensaje pase de un patio a otro.

Los usuarios, tomando de nuevo la delantera, crearon dos mecanismos para que la información pudiera propagarse más allá de la red de contactos directa. El primero de ellos fue la retransmisión de mensajes que consistía en repetir el mensaje, indicando la fuente. Se realizaba anteponiendo al texto de un tuit la sigla RT (*retweet*) y el autor del mismo (Boyd, Golder, & Lotan, 2010). La segunda consistió en el etiquetado de mensajes, comúnmente llamado *hashtag*, para facilitar la agrupación de mensajes por temática. Las etiquetas comenzaban con el símbolo #.

La retransmisión facilitaba que un tuit originado en una red de contactos pudiera llegar a otras redes al ser duplicado por otro usuario que potencialmente podría tener nuevas

conexiones. El etiquetado facilitaba la búsqueda de tuits por temas. Ambos mecanismos ampliaban el alcance de los mensajes, haciéndolos visibles más allá de los contactos de primer nivel, difundiéndose entre distintas subredes conectadas.

Dada la gran aceptación de la retransmisión, Twitter creó en el año 2009 un botón RT para difundir mensajes automáticamente (Stone, 2009a). La forma de hacerlo distaba del método original, ya que se perdía la trazabilidad de la propagación. En la retransmisión manual aparecían las cadenas de RTs/usuario por las que iba pasando el mensaje, pudiendo conocer por qué caminos se había difundido. En la retransmisión automática solo quedaba constancia del autor original del tuit y de la persona que lo había retransmitido, independientemente de si ambos eran contactos directos o indirectos (Congosto, 2009b). El botón de RT disminuyó el esfuerzo para difundir mensajes, abriendo el camino hacia multiplicación de mensajes por la propagación, llegando hoy día a alcanzar los 500 millones de mensajes al día.

El fenómeno de la retransmisión en Twitter ha sido ampliamente estudiado desde muchos puntos de vista. Los investigadores se han preguntado por qué se difunden unos temas más que otros, cómo participan los usuarios en este proceso, qué relaciones hay entre ellos, cómo medir la influencia de unos sobre otros, por qué caminos se difunden los tuits o cuánto tardan en propagarse. Los trabajos relacionados con la propagación en Twitter se estructuran a continuación en dos partes: los que ponen el foco en el usuario y los que los hacen en la propagación.

2.1 CARACTERIZACIÓN DE USUARIOS

Los usuarios han sido objeto de estudio desde distintos puntos de vista. Por un lado, la pobreza de datos del perfil de Twitter ha llevado a inferir atributos demográficos para conocer mejor a quienes participan en Twitter. Por otra parte, se ha analizado la forma en que se conectan, bien sea por seguir a otros usuarios o por interactuar con ellos. También ha despertado mucho interés la influencia que unos usuarios ejercen sobre otros. Adicionalmente, se han realizado definiciones específicas de tipos de usuarios encaminadas a estudiar patrones de comportamientos.

2.1.1 DEMOGRAFÍAS

Como he comentado anteriormente, los usuarios en Twitter tienen un perfil con sólo un dato obligatorio, su nombre de usuario. Es opcional su nombre real, de dónde son, dónde viven o su descripción. Con estos datos tan escasos es muy difícil conocer el género, la edad, la localización o la profesión de los usuarios. A continuación se muestran algunos trabajos orientados a la deducción de atributos demográficos de los usuarios.

(Rao, Yarowsky, Shreevats, & Gupta, 2010) buscaron cómo deducir del texto de los tuits cuatro atributos: el género, la edad, la región y la polaridad política. Utilizaron un clasificador basado en tres modelos: el socio-lingüístico, por n-gramas y por entrenamiento. Para el género obtuvieron los mejores resultados con el modelo de entrenamiento (72,33% de fiabilidad). En cuanto a la edad, la clasificación fue muy limitada, distinguiendo solo entre mayores o menores de 30 años, consiguiendo una precisión del 74,11% con el modelo de entrenamiento. Respecto a la región, con el modelo

socio-lingüístico lograron una fiabilidad de 77,08%. Finalmente, la polaridad política la detectaron con una precisión del 82,84% con el modelo de n-gramas.

(Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011) desde un enfoque muy pragmático, buscaron la manera de deducir tres atributos del perfil de los usuarios: geografía, género y etnia. Encontraron que el 75,3% de los usuarios habían cumplimentado información de origen. Basándose en la API de Google Maps consiguieron la geolocalización. Respecto al género lo obtuvieron del nombre de pila, utilizando una lista de la seguridad social con los mil nombres más frecuentes clasificados por género. La etnia la dedujeron del apellido mediante los datos del censo que proporciona una distribución de etnias por apellidos.

(Gayo-Avello, 2011a) utilizó el algoritmo interactivo McC-Splat (Chakrabarti, Dom, & Indyk, 1998) que etiqueta nodos en un grafo parcialmente etiquetado. Mediante este método se pueden ir detectando atributos demográficos de un pequeño grupo de usuarios y mediante su red declarada se puede ir deduciendo los del resto de sus contactos. Obtuvo los atributos de género, edad, orientación política, religión, raza, etnia y orientación sexual.

2.1.2 RED DECLARADA VS. RED DINÁMICA EN TWITTER

Los usuarios de Twitter tienen una red declarada de conexiones formada por sus seguidores (los que le siguen a él) y sus seguidos (los que sigue él). Como no es necesaria la correspondencia, esta red es asimétrica. Estas conexiones se van realizando con distintos criterios y se van incrementando con el tiempo de forma que la inercia hace que se vayan acumulando. A través de la red declarada llegan los tuits que publican los usuarios seguidos, siendo la red dinámica la que surge de las interacciones entre ellos.

Tener muchos seguidores está sobrevalorado en Twitter. Muchos usuarios tienden a seguir a otros con la esperanza de ser correspondidos (*follow back*) y aumentar su número de seguidores. Estas relaciones no están basadas en el interés real por un usuario sino en la cortesía. Por otro lado, las personas populares llegan a sobrepasar el millón de seguidores y es poco probable que estén todos ellos atentos a sus publicaciones. Esto lleva a pensar que existe una burbuja en las redes declaradas en Twitter (Pravda, 2009).

La diferencia entre la red estática y dinámica la estudió (Huberman et al., 2008), que detectó que el 25,4% de los tuits contenían menciones a otros usuarios, es decir, que se interrelacionaban. Definió como *friends* a aquellos que interactuaban entre sí. Encontró que la mayoría de los usuarios tenían muchos menos *friends* que seguidores. Por tanto, las redes dinámicas son mucho más pequeñas y menos densas que las redes declaradas, siendo las que desvelan las verdaderas relaciones.

2.1.3 LA INFLUENCIA

No es fácil definir la influencia. En el ámbito de la sociología coexisten muchas teorías sobre cómo nos influye el entorno y a través de qué instituciones o personas somos más receptivos a cambiar de opinión o a tomar decisiones. En el ámbito que nos ocupa, aquí se recogen los trabajos más relevantes que han intentado medir esta influencia en Twitter de una manera empírica.

He agrupado estos trabajos en dos grupos: las medidas basadas en el grafo de la red declarada, es decir, en conocer quién sigue a quien para determinar quién influye a quién, y las que están basadas en la dinámica de publicación, en la que no se precisa conocer la red de contactos para efectuar la medida.

1.1.1.1 Medidas basadas en la red de contactos

Las primeras medidas se basaron en la red declarada de los usuarios. (Tunkelang, 2009) propuso el TunkRank, adaptando el algoritmo de PageRank (Page, Brin, Motwani, & Winograd, 1998) a Twitter con este criterio:

$$Influecia\ (X) = \sum_{Y\ \epsilon\ seguidores\ (X)} \frac{1 + p * Influencia\ (Y)}{|seguigos\ (Y)|}$$

El TrstRank mide la importancia basándose en la centralidad del PageRank de la red declarada de un usuario. Es posible conocer la puntuación de un perfil mediante una API de Infochimps[27] (Drew Conway, 2010).

(Weng, Lim, & Jiang, 2010) propusieron el TwitterRank, una extensión del algoritmo PageRank, que propone medir la influencia de los usuarios en Twitter tomando en cuenta la similitud de los temas que publican los usuarios y la estructura de vínculos de su red declarada.

Las siguientes propuestas tienen en cuenta la red declarada pero para analizar las interacciones entre usuarios.

(Bakshy, Hofman, Watts, & Mason, 2011) definieron la Influencia usuario como la media del tamaño de todas las propagaciones de las que es origen. Durante dos meses, trazaron la aparición de URLs en los tuits con el siguiente criterio: si la persona B está siguiendo a la persona A y la persona A publica la URL antes de B, y fue el único de los seguidos de B en publicar la URL, entonces se considera que la persona A influencia a la persona B para publicar la URL. En el caso de que B tenga más de un seguido que haya publicado con anterioridad la misma URL, se elige el que publicó la URL primero. De esta forma fueron construyendo árboles de influencia, que denominaron “cascadas”. Encontraron que el tamaño y la profundidad de las cascadas seguían una distribución de potencias, donde la mayoría de las cascadas eran muy cortas y una minoría largas. Definieron la Influencia usuario como la media del tamaño de todas las cascadas en las que fue origen. Para cada usuario calcularon la influencia durante el primer mes y la influencia en el periodo total. La mayor correlación la encontraron entre la influencia en el primer mes y el número de seguidores. La conclusión a la que llegaron fue que las personas que han sido influyentes en el pasado y que tienen muchos seguidores son más propensos a ser influyentes en el futuro; sin embargo, esto solo es cierto en promedio.

(Romero & Huberman, 2011) definieron la pasividad como una barrera para la propagación, por lo que diseñaron el algoritmo de Influencia-Pasividad (I-P) que asigna a cada usuario una puntuación a la influencia y otra a la pasividad, de una manera relativa respecto al total de la red. Suponen que la influencia de un usuario depende de la cantidad

[27] http://www.infochimps.com/

y la calidad de la audiencia que influye. La pasividad de un usuario es una medida de lo difícil que es para otros usuarios influirle. El modelo asume:

- La influencia de un usuario depende del número de personas que influye, así como su pasividad.
- La influencia de un usuario depende de la dedicación de la gente que influye, es decir, la cantidad de atención que un usuario presta a otro en comparación con todos los demás.
- La pasividad de un usuario depende de la influencia a la que está expuesto pero no es influenciado.
- La pasividad de un usuario depende de lo mucho que rechaza la influencia de otro usuario en comparación con todos los demás.

Utilizaron una muestra de 22 millones de tuits con enlaces, de la que extrajeron 15 millones de URLs válidas. Obtuvieron el grafo total de conexiones declaradas de todos los usuarios encontrados que publicaron URLs válidas. Para cada uno de los usuarios midieron los siguientes atributos: el número de seguidores, el número de RTs recibidos, su puntuación I-P, el PageRank y el h-index (Hirsch, 2005), según los RTs recibidos. Por otro lado, calcularon el número de clics de cada una de las URLs como una métrica de la atención prestada al usuario que publicó el enlace. Seguidamente, para cada URL, calcularon los valores medios de los atributos de los usuarios que habían publicado dicha URL cotejándolo con el número de clics que éstas recibieron. La mayor correlación la encontraron en la puntuación I-P. Adicionalmente, compararon el I-P de cada usuario y su número de seguidores y no encontraron una buena correlación, lo que refuerza que popularidad no es lo mismo que influencia.

Finalmente, una medida que se ha hecho muy popular y que se utiliza como un estándar de facto de influencia es Klout[28], que proporciona una puntuación de 0 a 100. El algoritmo no está publicado pero se basa en estos tres conceptos:

- **Alcance** (*Reach*): número de personas a las que llega el usuario con su contenido u opinión (seguidores, amigos, gente que comparte grupos).
- **Amplificación** (*Amplification*): mide el grado en el que sus contenidos son compartidos por otros usuarios y por lo tanto cuánto influye el usuario en su audiencia.
- **Red** (*Network*): mide la relevancia de los nodos de la red del usuario.

1.1.1.2 Medidas basadas en la dinámica de publicación

(Cha & Gummadi, 2010) analizaron la influencia de los usuarios desde tres puntos de vista: seguidores, RTs y menciones. Utilizaron el *Spearman's rank correlation* (Zar, 1998) para calcular la correlación entre las tres métricas. Compararon los resultados para tres grupos: el general, el del 10% y el del 1% del ranking. Encontraron que los que estaban en las posiciones superiores de ranking tenían una correlación más baja entre su red declarada y su red dinámica, pero tenían una correlación más fuerte entre RTs y Menciones, lo que viene a reforzar la idea de la burbuja de las redes declaradas. También analizaron cómo evolucionaba la influencia en el tiempo durante ocho meses para tres grupos del ranking: los 10 (mayoritariamente medios de comunicación), 11-100 (personas populares) y 101-233 (personas populares y líderes de opinión). El grupo que más osciló fue el de los 10 primeros del ranking, no teniendo tendencia creciente ninguno de los tres grupos. Asimismo, detectaron la figura del *topical influential,* consistente en perfiles

[28] https://klout.com

normales que de repente adquirían influencia sobre un tema pero ésta se mantenía en un corto periodo de tiempo y decaía.

(Gayo-Avello, Brenes, Fernández-Fernández, Fernández-Menéndez, & García-Suárez, 2010) propusieron una original medida de la influencia basada (metafóricamente) en la segunda ley de Newton de la física.

$$F = m.a$$

Equiparan la masa de un usuario a su número de seguidores. La fuerza aplicada a poner un usuario en movimiento era el número de menciones recibidas o de RTs. De esta forma, un perfil con muchos seguidores necesita más menciones o RTs para empezar a moverse que otro con menos. Esta ecuación asume fuerzas y aceleraciones instantáneas y tiempo continuo. Para la implementación del algoritmo se opera en intervalos de tiempo discretos.

$$v_t = v_{t-1} + \frac{F_a}{m} - \zeta$$

.

El modelo, denominado *velocidad*, asume que el tiempo se mide en intervalos de una hora, F_a es el número de menciones por intervalo de tiempo, m es el logaritmo del número de seguidores y ζel rozamiento implementado mediante una constante que decrementa la velocidad cuando no hay ninguna mención en el intervalo de tiempo.

Compararon la correlación entre los clics de las URLs con los rankings de *velocidad,* Followers, PageRank, TunkRank e Influencia I-P. Teniendo en consideración todo el periodo de tiempo de los datos analizados obtuvieron que el I-P < TunkRank < *velocidad* < PageRank < Followers. Estos resultados eran diferentes a los publicados por (Romero & Huberman, 2011), en el que el orden era *Followers* < PageRank < I-P. Esta discrepancia la achacan a que puede ser debida tanto a la forma de seleccionar los datos como a la manera de calcular las puntuaciones en ambos experimentos. Midiendo *velocidad* por semanas mejora sensiblemente su correlación con el número de clics de las URLs, siendo la medida con mejor comportamiento.

Adicionalmente, proponen *velocidad* como una medida en tiempo real para detectar tendencias de usuarios; es decir, aquellos usuarios que alcanzan altas velocidades y que pueden ser de interés para un público que todavía no los conoce. Al igual que en el cálculo de las tendencias de Twitter se favorecen los temas novedosos frente a los más recurrentes, proponen una manera de que se primen los casos de usuarios que empiezan a destacar frente a los que ya son muy populares, midiendo la aceleración relativa de los usuarios.

(González-Bailón, Borge-Holthoefer, & Moreno, 2013) definen el concepto de influencia como el ratio entre el número de menciones recibidas y el número de mensajes publicados. Relacionan la medida de influencia con el ratio de red (seguidos/seguidores) para determinar si esta influencia proviene de usuarios populares u otros menos conocidos.

(Morales, 2014) plantea la eficiencia de usuario como el ratio entre el número de retransmisiones recibidas y el número de mensajes publicados, midiendo los efectos que tienen los esfuerzos individuales sobre la reacción colectiva. Encontró que los que tenían más seguidores que seguidos tenían una mayor eficiencia, pero no en todos los casos.

Observó que la distribución de esta propiedad es ubicua en seis casos de estudio, sin importar sus dimensiones ni contextos. Por lo cual sugiere que existe universalidad en la relación entre esfuerzos individuales y reacciones colectivas en Twitter.

2.1.4 CLASIFICACIÓN DE USUARIOS

Klout realizó en el 2009 una clasificación de usuarios que ahora no está disponible en su web, pero que se explica detalladamente en un post del Wwwhats' new[29]. Esta clasificación se realizaba en cuatro ejes: *Sharing - creating*, *broad - focused*, *consistent - casual*, *listening - participation*, dando lugar a dieciséis tipos de usuarios (Figura 2).

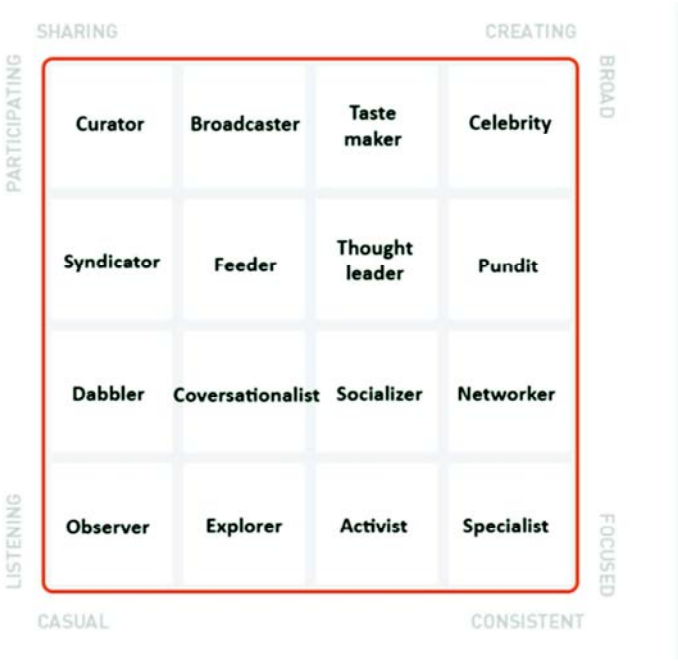

Figura 2 Clasificación de usuarios de Klout

De la misma manera que el algoritmo de Klout es desconocido, la forma de encasillar a estos tipos de usuarios también se ignora. Es una clasificación bastante compleja con algún caso interesante como el de *Networker.*

(Quercia, Ellis, Capra, & Crowcroft, 2011) analizan la forma de expresarse en Twitter para cinco tipos de usuarios a fin de determinar sus diferencias lingüísticas. La clasificación de usuarios que realizan es la siguiente:

- *Popular*: seguido por muchos (alto número de seguidores).
- *Influential*: puntuaciones altas en Klout y TrstRank.
- *Listener*: siguiendo a muchos (alto número de seguidos).
- *Star*: Sigue a pocos y es seguido por muchos (alto valor en el ratio $\frac{seguidores}{seguidos}$).
- *Highly-read*: Aparece en muchas listas.

Esta clasificación genera tipos no disjuntos, es decir, un usuario puede pertenecer a varias clases. Por ejemplo, el perfil del Barack Obama (@BarackObama) es Popular (65,7 millones de seguidores), es *Influential* (98 Klout), es *Star*, (64 Millones/632K) y *Highly-read* (213K listas).

[29] http://wwwhatsnew.com/2011/02/09/la-clasificacion-de-los-nuevos-lideres-segun-klout/

(Uddin, Imran, & Sajjad, 2014) definieron dieciocho características para clasificar a los usuarios en seis grupos según la finalidad de su presencia en Twitter:

- *Personal users*: los que lo usan para entretenerse, aprender o leer noticias.
- *Professional users*: los que comparten información especializada en la que son expertos.
- *Business users*: los que lo utilizan para promocionar su negocio.
- *Spam users*: los que tienen una finalizad maliciosa. Generalmente son programas automáticos (*bots*). Suelen utilizar el recurso de *follow-back* para adquirir seguidores pero con el tiempo los van perdiendo.
- *Feed/news*: son los que publican noticias de los medios de una forma automática. Suelen estar automatizados como los *Spam users* pero se distinguen de ellos en que sus URLs no son maliciosas y que con el tiempo van adquiriendo seguidores.
- *Viral/Marketing Services*: utilizan técnicas de marketing para posicionar marcas con la ayuda de la tecnología y el conocimiento de cómo funcionan las redes sociales. Utilizan *bots* inteligentes que propagan información e interactúan con su audiencia.

Entrenaron con las 18 características un SVM (*Support Vector Machine*) y obtuvieron una fiabilidad entre el 0,532 (*Spam users)* y el 0,895 (*Business users).* En esta tesis se implementa el usuario Automático que se basa en la definición de *Feed/news.* El usuario *Spam users,* al que los autores no han conseguido identificar con mucha fiabilidad, no se incluye en la tesis por su complejidad para detectarlo.

(Osteso, Claes, & Deltell, 2013) clasificaron los usuarios en instituciones, personas reales y personajes (*fakes* o parodias). Es una clasificación muy sencilla pero muy reveladora para analizar la capacidad de propagación de las personas frente a las organizaciones o ante las figuras cada vez más frecuentes, sobre todo en temas sociales, que bajo un anonimato utilizan el sarcasmo como medio de expresión. Aunque esta clasificación no se implementa en esta tesis porque requiere una clasificación manual, sí la he utilizado para medir la viralidad de estos tipos de usuarios en distintos casos de estudio (Congosto, 2014c) y (Congosto, 2015d).

(González-Bailón et al., 2013) clasificaron a los usuarios en *influentials*, *hidden influentials*, *broadcasters* y *common users* en función de cómo se distribuían en un diagrama de dispersión respecto al ratio de influencia (menciones/mensajes) y al de red (seguidos/seguidores). La clasificación es sencilla y clarificadora.

2.2 LA DIFUSIÓN

Esta tesis pone el foco en la difusión a través del tiempo y en los parámetros para evaluarla. Por ese motivo, a continuación se muestran una serie de trabajos que han influido para definir la forma y los atributos de medida de esta tesis.

(Kwak, Lee, Park, & Moon, 2010) realizaron el primer estudio cuantitativo los datos totales de Twitter durante casi un mes en el año 2009, obteniendo unos 41,7 millones de usuarios, 1,47 billones de relaciones, y 106 millones de tuits. Hoy día sería casi inabordable analizar todos los datos de un mes en Twitter porque tan solo en un día se publican 500 millones de tuits. Este estudio es uno de los más importantes que se han realizado y tiene el valor de contemplar el entorno completo de Twitter durante un periodo de tiempo. Midieron la

capacidad de un usuario de difundir más allá de su red de seguidores y encontraron que para aquellos que tenían más de mil seguidores, el promedio de los usuarios que no eran de su red no se veía afectado por el número de seguidores. Es decir, no importaba cuántos seguidores tuviera un usuario, el tuit era probable que llegase a un cierto número de audiencia una vez que se difundía mediante el RT. En las medidas temporales de difusión encontraron que la mitad de las difusiones se realizaron en la primera hora y el 75% en el día.

(Liere, 2010) encontró que en las retransmisiones, el 60% de los tuits se propagaban durante la primera hora, valor muy parecido a la medida anterior. Las propagaciones suelen ser rápidas y no halló ningún tuit que fuera retransmitido pasadas las 24 horas. Midió la distancia física en kilómetros, basada en el dato de localización del perfil, entre el autor original del tuit y los usuarios que le retransmitieron. Encontró que el valor medio de la distancia era 955 km. y la mediana 1.698 Km. Lo que refuerza que Twitter no es tan global como puede parecer.

(Sakamoto, 2015) Estudió la propagación en Twitter confirmando también que el 50% de los RTs ocurren en la primera hora y que existía una correlación entre el número de seguidores y el número de retransmisiones obtenidas. Encontró que la longitud de las cascadas de difusión oscilaba entre 1 y 12 para mensajes retransmitidos 100 veces. A mayor profundidad de las cascadas había menos seguidores directos en la propagación. Analizando conjuntamente todos los casos halló que, en la mayoría de las veces, solo un tercio de los usuarios que participaron en la retransmisión eran contactos directos del autor del tuit propagado. La media de seguidores directos que participaron en la propagación de mensajes de texto (9,16) era mayor que en la de mensajes con imágenes (8,82), lo que refuerza la idea de que los mensajes con imágenes se difunden mejor fuera de la red de contactos directos.

(De Domenico et al., 2013) realizaron un análisis espacio-temporal de las dinámicas de propagación en tres ventanas de tiempo: antes, durante y después de la noticia del descubrimiento de la partícula Higgs boson en el CERN. Definieron el *intertimes* como la demora temporal entre dos tuits consecutivos publicados e *interspaces* como la distancia espacial entre dos tuits geolocalizados consecutivos. Calcularon los *intertimes* e *interspaces* para las tres ventanas de tiempo. Encontraron que los *interspaces* se distribuían de forma similar en las tres ventanas mientras que el *intertime* variaba mucho. Desde un punto de vista global, los *intertimes* mostraban una distribución de potencias antes y después del evento principal, con un gran número de tuits enviados dentro en unos pocos segundos, y un pequeño número enviado a los pocos minutos. Por otra parte, la dinámica del proceso cambiaba drásticamente durante el evento principal, cuando esta frecuencia oscilaba entre menos de dos segundos y no más de seis segundos.
Profundizando en el análisis por usuarios, definieron para cada uno de ellos los *inter-arrival times* para las actividades de publicación de tuits, respuestas y retransmisiones como $\tau_u = t_u(i+1) - t_u(i)$, es decir, el tiempo que trascurre entre dos actividades. Encontraron que para los *inter-arrival times* de la publicación de tuits en la ventana pre y post al evento, los usuarios publican a ráfagas enviando varios tuits en pocos minutos seguidos de largos periodos sin actividad de hasta un día, dando lugar a una distribución de potencias. Durante el evento principal mostraban un comportamiento muy diferente, más compatible con una distribución normal que con una ley de potencias. En el caso de los *inter-arrival*

times de los RTs también seguían una distribución de potencias con exponente -0,8, un caso poco usual de exponente en las distribuciones de potencias (Newman, 2005).

(Yang et al., 2010) analizaron el comportamiento de los usuarios que hacen RT. Encontraron que la difusión era muy asimétrica, siguiendo una distribución de ley de potencias. La mayoría de los usuarios hacían muy pocos RTs mientras que un pequeño grupo realizaba muchos (*retweet-aholics*). El 3,13% de las retransmisiones las realizaron usuarios que hicieron RT más de 1.000 veces. Respecto a la relación entre número de tuits publicados y número de retransmisiones, los usuarios poco activos hacían pocos RTs mientras que los que publicaron más de 200 tuits eran más propensos a la retransmisión. También hallaron que los patrones temporales de publicación de los *retweet-aholics* eran diferentes a la de los usuarios normales.

(Grabowicz, Ramasco, Moro, Pujol, & Eguiluz, 2012), basándose en la teoría de *The strength of weak ties* (Granovetter, 1973), buscaron en Twitter esas conexiones débiles que, según la teoría, propiciaban la difusión de la información. Partieron de un conjunto de 2,5 millones de usuarios que obtuvieron por el procedimiento que se conoce como bola de nieve. Este procedimiento consiste en obtener la red desde un conjunto de usuarios iniciales (semillas) a los que se van explorando sus contactos y de estos a su vez los suyos y así sucesivamente hasta que se llega al límite deseado. Seguidamente obtuvieron los tuits de estos usuarios durante los meses de noviembre y diciembre del 2008. De los tuits obtenidos extrajeron aquellos que tenían interacciones (menciones y RTs). Crearon un grafo en la que los nodos eran los usuarios y las conexiones las interacciones. La agrupación de usuarios la obtuvieron aplicando al grafo el algoritmo de detección de comunidades Oslom (Fortunato, 2010). Estas comunidades no eran disjuntas y un usuario podía pertenecer a más de un grupo. Las conexiones de los grupos podían ser: internas al grupo, entre dos grupos que tenían usuarios en común y entre dos grupos disjuntos. Encontraron que las menciones eran más habituales dentro de un grupo y que los RTs solían pasar de un grupo a otro mediante los usuarios que tenían en común. Por lo tanto no eran los usuarios que tenían más conexiones dentro de un grupo los que más propagaban, sino aquellos que tenían mayor nivel de intermediación (pertenecía a dos o más grupos). Por tanto, la estructura de las redes de interacción dinámica favorece la propagación cuando los grupos que se forman no son cerrados y existen usuarios conectados a varios grupos.

2.3 CONSIDERACIONES SOBRE LA DIFUSIÓN

Siempre que se analiza un estado del arte se encuentran lagunas que no están cubiertas o temas que merecen un análisis más profundo o métodos que pueden ser mejorados. De todo lo anteriormente expuesto, paso a describir los aspectos que mueven a esta tesis a realizar sus contribuciones.

2.3.1 LA DIFUSIÓN MÁS ALLÁ DE LA RED DE CONTACTOS

Muchos estudios de propagación se basan en análisis de la red declarada para explicar este fenómeno y para crear modelos predictivos. Esto implica un enorme esfuerzo tanto para obtener los datos de red de usuarios desde Twitter como para procesar complejos algoritmos de redes. Todo ello se agrava con el crecimiento de usuarios y mensajes lo que implica trabajar cada vez con mayores volúmenes de datos, lo que supone un hándicap

investigar por esta vía. Por otro lado, las dinámicas de comunicación están cambiando y la difusión ya no solo se hace por la red de contactos.

En los últimos años hemos asistido a popularización de las tendencias de Twitter. Incluso usuarios que no tienen perfil en Twitter saben lo que significa porque se hacen eco los medios de comunicación. Aquel algoritmo que en el año 2010 creó Twitter para determinar de qué se estaba hablando con más frecuencia (Bowman, 2010) y que fue evolucionando para dar información cada vez más local y detallada de las conversaciones más frecuentes (Kamdar, 2010), es hoy día una especie de portada de lo que está ocurriendo en tiempo real.

Los usuarios consultan frecuentemente las tendencias en Twitter, tanto locales como globales, para conocer los temas de actualidad. Esta consulta abre la puerta de entrada a una avalancha de tuits que podrían no llegar por la red de contactos. Por este motivo las tendencias aceleran la difusión de mensajes de una manera diferente a la retransmisión, mucho más rápida pero también más concentrada en unos pocos temas.

Otro factor que influye en la difusión de mensajes son los medios de comunicación. Esta influencia es bidireccional; una tendencia de Twitter puede convertirse en noticia o viceversa. En España, el 55 por ciento de las tendencias tratan sobre eventos de los que también se hacen eco los medios de comunicación tradicionales (Carrascosa, Cuevas, Gonzalez, Azcorra, & García, 2015). El poder de difusión de los medios, sobre todo la televisión, actúa de una manera intensa mientras dura la emisión, tal y como he detectado en varios experimentos (Congosto, 2015f) (Congosto, 2014a).

La Figura 3 muestra de una manera visual cómo los mecanismos de difusión facilitan la propagación entre distintos patios virtuales.

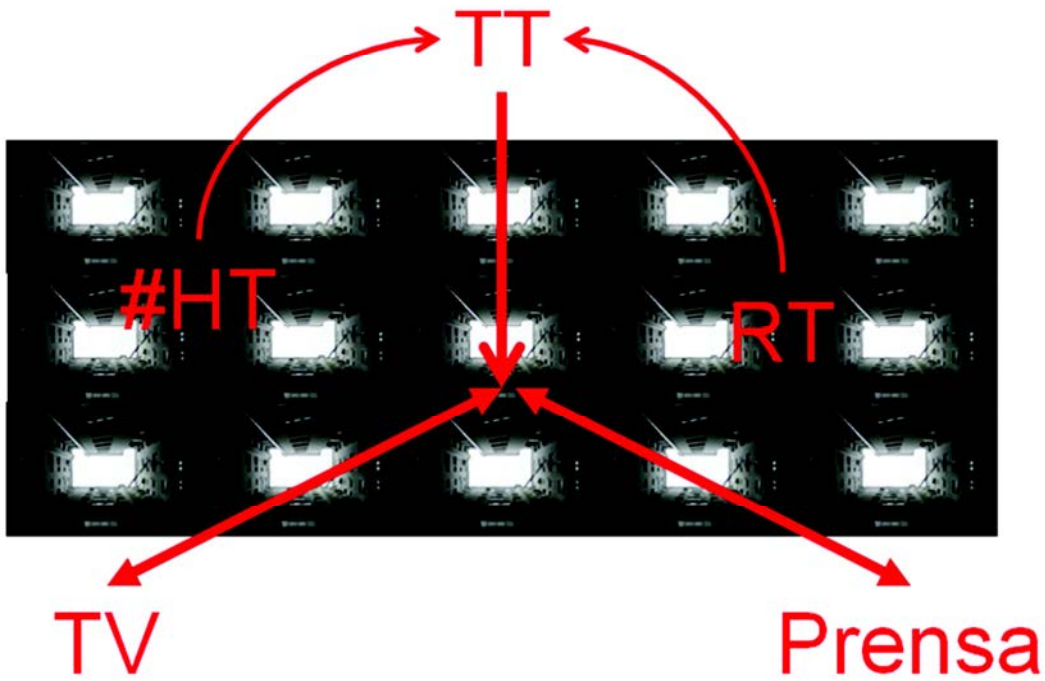

Figura 3 Mecanismos de difusión

La difusión puede iniciarse desde dentro de Twitter mediante los mecanismos de etiquetado que agrupan tuits por temas y las retransmisiones que los difunden multiplicando el número de mensajes. Ambos factores hacen que ciertos temas sean tendencia. Una vez que algo es tendencia es visible para todos los patios virtuales y la información fluye de arriba abajo. Adicionalmente la tendencia puede ser noticia en la prensa o en la televisión, lo que aumenta aún más su visibilidad. Por otro lado, cada vez son más frecuentes los programas de televisión que llevan a sus espectadores a Twitter para interactuar con ellos, creando nuevos hábitos de consumo de medios.

Las redes declaradas en Twitter siguen siendo muy importantes pero no es el único medio de propagación y eso hay que tenerlo en cuenta.

2.3.2 LA DIFUSIÓN EN DISTINTOS CONTEXTOS

Muchos estudios de propagación se realizan para un conjunto de datos de un periodo definido, por lo que miden atributos y propiedades sobre un único caso de estudio. Recientemente he encontrado trabajos en los que ya tienen en cuenta la medidas empíricas sobre distintos entornos (Morales, 2014) y (Val, Rebollo, & Botti, 2015), pero no son los casos más frecuentes.

Otro factor muy importante es la duración de casos de estudio. No he encontrado ninguno que analice un periodo superior a un año.

PARTE II. Requisitos

3 DEFINICIÓN DE LOS INDICADORES DE PROPAGACIÓN

La difusión en Twitter va mucho más allá de la propagación en red. Hoy día todo está interconectado, siendo muchas las vías por las que se conoce lo que ocurre en Twitter. Las tendencias de Twitter dan visibilidad a los temas candentes del día y en cuestión de minutos muchas personas pueden acceder e incluso participar en la conversación. Por otra parte, los medios de comunicación se hacen eco de lo que ocurre en Twitter o lo utilizan para interactuar con sus audiencias. Al ser muchas las fuentes a través de las cuales nos enteramos de los que sucede en Twitter, resulta muy complicado establecer cómo se produce la propagación. Por este motivo, propongo analizar los patrones de difusión en distintos entornos de una forma empírica, mediante la medida de un conjunto de indicadores a través de un conjunto de casos de estudio.

3.1 IDENTIFICACIÓN DE LOS INDICADORES

Para abordar el análisis a través de distintos casos de estudio hay que realizar las mismas medidas en cada uno de ellos de forma que sea posible contrastar los resultados y descubrir patrones recurrentes.

Estas medidas se realizan mediante un conjunto de indicadores predefinidos basados en tipos de usuarios y en métricas de propagación.

3.1.1 EL USUARIO COMO INDICADOR

Los usuarios de Twitter, a diferencia de Facebook o de Linkedin, tienen relaciones asimétricas, es decir, no se requiere consentimiento mutuo entre ambos para que exista relación. Por ende, es posible que alguien siga a otro sin que sea correspondido y el número de seguidores y seguidos puede ser diferente. Dependiendo de la proporción entre seguidos/seguidores se pueden agrupar en tres perfiles (Congosto, 2009a):

- Populares: cuando el número de seguidores es muy superior al de seguidos.
- Fan: si la cantidad de seguidos es muy superior a los seguidores.
- Amigables: en el caso de estar compensado el número de seguidores y seguidos.

Cada usuario va creando su propia red de contactos por la que puede tener acceso a las publicaciones de sus seguidos mientras que sus seguidores pueden acceder a las suyas. En consecuencia, el número de seguidores de un usuario determina el alcance que pueden tener sus mensajes, aunque las interacciones solo se producen con una pequeña parte de ellos. Por este motivo, el número de relaciones declaradas de los usuarios es siempre mucho mayor que la interactividad real que existe entre ellos (Huberman et al., 2008) (Morales, 2014).

La proporción de seguidos/seguidores es un dato que define al perfil del usuario pero que no siempre determina el grado de propagación de sus tuits (Cha & Gummadi, 2010). Son las dinámicas de comunicación las que afloran la influencia oculta de algunos usuarios.

La Figura 4 muestra la clasificación de los usuarios que participaron en la movilización social del 15M realizada por (González-Bailón et al., 2013).

> *«The scatterplot summarizes how users distribute in the network of followers and in the allocation of targeted messages. Both axes are expressed as ratios so that it is easier to identify outliers, that is, users who depart from symmetrical networks or from the volume of message exchange entailed by mere reciprocation. The vertical axis tracks the number of messages that users received over the number of messages they sent: the most visible users (those who were mentioned more often in protest messages) are above the dashed line. The horizontal axis tracks the number of other accounts a user is following over the number of followers they have: the most central and popular users in this communication network are on the left of the dashed line».*

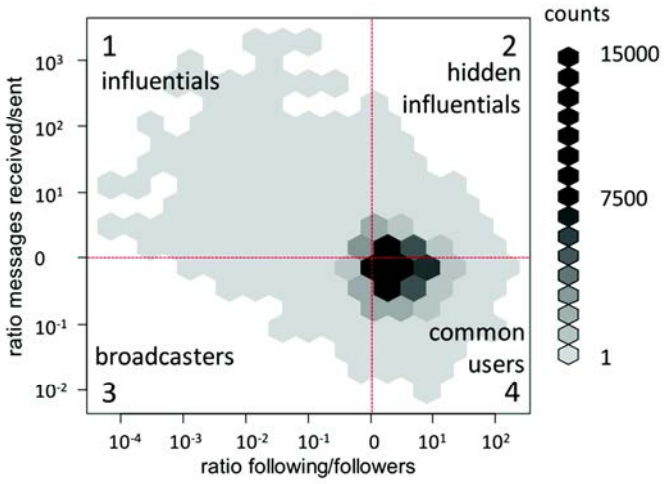

Figura 4 Influencia vs. Red (González-Bailón, Borge-Holthoefer, & Moreno, 2012)

Esta clasificación es sencilla y clarifica muy bien el tipo de usuarios en Twitter. Aparecen cuatro grupos en función de la capacidad de influencia y de la proporción de seguidos/seguidores: *influentials*, *hidden influentials*, *broadcasters* y *common users*. Sin embargo, caben algunas matizaciones en estos grandes grupos. Entre los *influentials* existen distintos grados según su capacidad de ser propagados. Los *broadcasters* pueden ser usuarios que publican muchos tuits que no son difundidos, que simplemente se dedican a retransmitir mensajes de otros o que publican tuits automáticamente mediante sindicación de contenidos.

Por este motivo he ampliado esta clasificación recogiendo nuevos roles según el tipo de actividad como se muestra en la Figura 5.

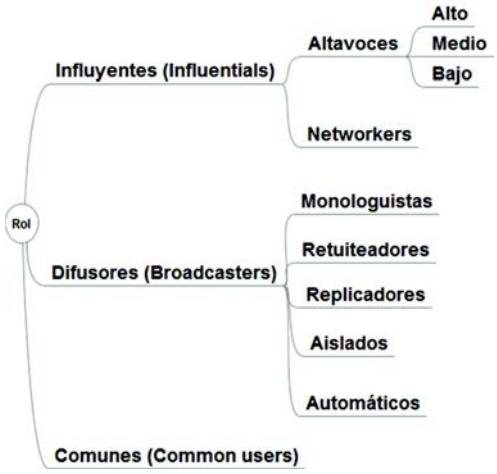

Figura 5 Clasificación de usuarios por actividad

Los roles quedan definidos de la siguiente manera:

- Altavoces: son los usuarios capaces de amplificar sus tuits. Pertenecen a este grupo aquellos cuyo número de retransmisiones recibidas ha sido tres veces mayor que el volumen de tuits emitidos. Hay tres niveles:
  - Alto: formado por los usuarios con más impacto que acapararon el 20% de las retransmisiones.
  - Medio: los siguientes más difundidos que obtuvieron un 50% de las retransmisiones.
  - Bajos: el resto.
- *Networkers*: mantienen una actividad alta, emitiendo tuits, difundiéndolos y siendo retransmitidos de una manera balanceada.
- Retuiteadores: muestran una frecuencia alta de publicación predominando las retransmisiones frente a los tuits propios.
- Monologuistas: manifiestan una participación alta con tuits propios que no reciben retransmisiones.
- Replicadores: La mayoría de sus tuits corresponden a respuestas de otros mensajes.
- Aislados: Los que no hacen ni reciben retransmisiones.
- Automáticos: la mayoría de sus tuits proviene de una fuente de sindicación, y son publicados con aplicaciones como Twitterfeed o Dlvr.it
- Comunes: el resto de usuarios que no tienen ninguna de las características descritas.

En el siguiente seudocódigo se detalla el algoritmo de clasificación por roles:

```
lista del top 20 = usuarios que han sido el origen del 20% de
los RTs
lista del top 50 = usuarios que han sido el origen del 50% de
los RTs
tuits_propios= número total de tuits – número de RTs enviados
k_rt = número de RTs recibidos / número de RTs enviados
k_out = número de RTs enviados / número total de tuits
k_in = número de RTs recibidos / tuits_propios
```

```
k_reply_out = número de respuestas enviadas / tuits_propios
k_auto = número de tuits automáticos / tuits_propios

if (k_in >= 3) y usuario en la lista del top 20:
    el rol es 'Altavoz alto'
elif (k_in >= 3) y usuario en la lista del top 50:
    el rol es 'Altavoz medio'
elif (k_in >= 3):
    el rol es 'Altavoz bajo'
elif número de tuits >= que la media y número de Rts
recibidos >= que la media y k_rt >= 0.50:
    el rol es 'Networker'
elif número de tuits >= que la media y (k_out >= 0.50):
    el rol es 'Retuiteador'
elif número de tuits >= que la media y (k_in <= 0.30):
    el rol es 'Monologuista'
elif (k_reply_out >= 0.60):
    el rol es 'Replicador'
elif k_auto >= 0.75
    el rol es 'Automático'
elif número de RTs recibidos = 0 y número de Rts enviados =
0:
    el rol es 'Aislado'
else:
    el rol es 'Común'
```

Esta clasificación detallada en distintos roles será un indicador del tipo de propagación debido a que según la proporción en la que aparezcan se puede saber si la difusión fue debida a la presencia de grandes Altavoces, al trabajo de los *Networkers*, a una campaña apoyada por Retuiteadores o por Monologuistas o por *spam*.

3.1.2 MÉTRICAS DE PROPAGACIÓN

Twitter es un espacio de interacción que se retroalimenta con sus propios contenidos. Los mensajes llegan por la redes de contactos pudiendo pasar desapercibidos o captando la atención para ser difundidos o respondidos, generando nuevos tuits. A continuación de definen una serie de indicadores que podrían impactar en la publicación de tuits.

3.1.2.1 El alcance

Como se ha comentado anteriormente, el número de seguidores de un usuario determina el alcance de sus tuits, es decir, el número de usuarios que potencialmente podrían leer un mensaje. Evidentemente no todos los usuarios están permanentemente conectados a Twitter y los que pudieran estarlo puede que no lo lean o en caso de hacerlo no le dediquen atención. Es importante la influencia que ejerza el autor del tuit o de la motivación que despierte su contenido para que otros usuarios reaccionen activamente (Romero & Huberman, 2011).

El alcance determina la capacidad potencial de hacer visible un mensaje y por tanto aumenta la probabilidad de que el mensaje se difunda o que dé lugar a nuevos tuits, aunque no es el único factor que influye en la propagación. En algunos casos los usuarios con un gran alcance difunden menos que otros con menos seguidores, siendo determinante el grado de identificación de los seguidores con el contenido (Congosto, 2014b) (Congosto, 2015c). Aun así, es un indicador importante para el análisis de las propagaciones.
La medida del alcance encierra en sí una inexactitud debido a que los usuarios pueden tener seguidores comunes. Eliminar los casos en que esto ocurre requeriría un esfuerzo de procesado inviable debido a las restricciones temporales que impone la API de Twitter para obtener la red de seguidores. Por este motivo, se asume este solapamiento de seguidores.

3.1.2.2 La retransmisión

El RT es el mecanismo de difusión utilizado por los usuarios de Twitter. La proporción entre el número de tuits totales frente a los retransmitidos puede indicar el volumen de información redundante que está circulando en un momento dado. Es una medida sin contexto que debe ser contrastada con otras medidas para mejorar la interpretación de la difusión. Es interesante medir también la retransmisión de los mensajes que contienen enlaces o etiquetas por estos motivos:

- La presencia de enlaces invita a acceder a información más detallada fuera del entorno Twitter que no es posible resumir en 140 caracteres. Es una medida de difusión de contenidos externos vía Twitter.
- La inclusión de etiquetas puede, con frecuencia, obedecer a un afán de clasificación y estructuración del contenido del tuit. Puede ser una métrica de difusión de acciones organizadas.

3.1.2.3 La participación

El número de tuits publicados en un momento dado hay que contrastarlo con la cantidad de usuarios diferentes que los han generado. Esta relación indicará si la publicación de tuits está inflada o se encuentra en parámetros normales. Un caso extremo de activismo, casi próximo al *spam,* fue el del usuario @isaranjuaez que llegó a publicar más de mil tuits en un día, a veces con una frecuencia de nueve mensajes por minuto (Congosto, 2015e).

También es importante medir los nuevos usuarios que se van incorporando a la conversación porque esto indica si se está ampliando el radio de difusión o si se ha saturado.

Para ilustrar cómo se produce esta saturación se muestra en la Figura 6 el caso del 15M donde se aprecia que después de que se alcanzó el máximo número de usuarios, las nuevas incorporaciones decayeron.

Figura 6 Incorporación de nuevos usuarios durante el 15M

Los usuarios únicos que participaron cada día (línea verde) se descomponían en dos grupos: los nuevos, es decir, los que participaron por primera vez (línea naranja) y los consolidados, que ya habían publicado antes algún tuit sobre este tema (línea azul).

3.1.2.4 La automatización

Hay un conjunto de usuarios cuyos tuits en su mayoría, y en algunos casos la totalidad, son automáticos, es decir, están publicados mediante aplicaciones que sindican contenidos. Es una práctica muy usada por el marketing de algunas marcas o servicios de atención al cliente. En el caso concreto de los datos sociales generalmente están asociados a la publicación de noticias de periódicos. Estos usuarios son fácilmente identificables por la aplicación que utilizan para publicar tuits (Twitterfeed, Dlvr.it) y porque en el instante en que se publica una noticia aparecen simultáneamente sus tuits con el titular y el enlace a esta información.

Estos usuarios publican por ráfagas (cuando hay noticias nuevas) y de forma casi sincronizada. Este ritmo perturba al número de tuits por unidad de tiempo y por tanto produce un ruido en la medida de los tuits generados por personas reales. Se medirán los tuits automáticos para conocer el grado de ruido de cada caso de estudio.

3.1.3 MEDIDA TEMPORAL

En Twitter, la información fluye en tiempo real y la propagación de un tuit es muy rápida al principio decayendo lentamente, pudiendo tener una latencia de casi un mes. En las medidas realizadas por (Kwak et al., 2010), el 50% de las retransmisiones ocurren durante la primera hora, el 75% durante el primer día y el 10% durante el primer mes. Un resultado similar obtuvo (Sakamoto, 2015) en un experimento realizado años más tarde.

La duración de los casos de estudio es en algunas veces superior a los dos años. Durante este tiempo hay periodos llanos con poca participación y momentos con picos de actividad. Por este motivo es posible medir la propagación en la dinámica rutinaria y en los casos excepcionales para ver en qué parámetros varía.

Respecto a la unidad temporal de medida he evaluado la propagación en tres periodos de tiempo: por minuto, por media hora y por día. La medida por minuto es demasiado corta y no recogería los efectos de la propagación y la de por día es demasiado larga y atenuaría mucho los picos. Midiendo cada hora se recogería el 50% de las retransmisiones aunque en los eventos de corta duración como los TT o los relacionados con la televisión se obtendrían pocas muestras para comparar correlaciones. Por ese motivo he optado por medir los indicadores de la propagación cada hora y cada cinco minutos. En los casos en que se obtengan pocas muestras se podrá contrastar la métrica realizada por hora con la obtenida cada cinco minutos.

3.1.4 INDICADORES DE PROPAGACIÓN

Atendiendo a lo antes expuesto, los indicadores que van a ser utilizados para evaluar la propagación en distintos casos de estudio se muestran en la Tabla 2.

Indicador	Valores	Medida temporal
Rol	Altavoces (altos, medios y bajos), Networkers, Retuiteadores, Monologuistas, Replicadores, Normales y automáticos	En global Cada hora y cada cinco minutos
Alcance	Número de seguidores de los usuarios	Cada hora y cada cinco minutos
Difusión global	Número de RTs	Cada hora y cada cinco minutos
Difusión enlace	Número de RTs con enlaces	Cada hora y cada cinco minutos
Difusión etiquetado	Numero de RTs con etiquetas	Cada hora y cada cinco minutos
Participación	Número de usuarios diferentes	Cada hora y cada cinco minutos
Incorporación	Número de usuarios que participan por primera vez	Cada hora y cada cinco minutos
Automatización	Número de tuits generados automáticamente	Cada hora y cada cinco minutos

Tabla 2 Indicadores de propagación

3.2 CASOS DE ESTUDIO

Esta tesis afronta el análisis de los indicadores de propagación a través de distintos casos de estudio. Al medir de la misma manera distintos entornos de conversación es posible compararlos y evaluar en qué puntos son coincidentes y en cuáles divergentes.

La selección de los casos de estudio ha sido cuidadosamente estudiada, agrupándose cada uno de ellos por afinidades. Cada grupo mide una característica y a su vez contiene distintas colecciones de datos para aportar más puntos de referencia. La estructuración de los casos es la siguiente:

- Barómetro social: corresponde a las opiniones de la ciudadanía sobre temas sociales relevantes que están presentes en el día a día. El periodo de recogida de datos se encuentra entre uno y tres años.
- Movimientos sociales: recoge momentos decisivos de movilización social que posteriormente han producido cambios sociales y políticos.
- Elecciones: agrupa las campañas electorales en España desde el 2010. Para unificar el criterio de selección de datos se han seleccionado las interacciones de los candidatos y partidos con los usuarios, descartando los tuits que no cumplen este requisito.
- Prensa: aporta el punto de vista de la propagación de noticias en Twitter para comprender cómo los periódicos influyen en las conversaciones.
- Casos internacionales: proporcionan un referente fuera de España para conocer las diferencias de uso en España y en otros países.
- Tendencias: suministra un punto de vista sobre la propagación en los momentos de máximo protagonismo de un tema.
- Twitter y televisión: contrasta el modelo de difusión en red con el de emisión de la TV para observar las diferencias.

3.2.1 BARÓMETRO SOCIAL

De las colecciones de datos que he ido recogiendo durante años relacionadas con los problemas sociales o los escándalos políticos he seleccionado dos casos significativos para mi investigación: los recortes presupuestarios y el caso Bankia. El resto de los casos tienen características muy similares como la larga duración, periodos de poca actividad y picos de mucha frecuencia de publicación, y están influenciados por las noticias de los medios. Por este motivo, incluir más casos no aportaría mucha información adicional.
En la Tabla 3 se describe con qué palabras clave se han obtenido los datos, durante cuánto tiempo, el volumen de tuits y los usuarios que han participado.

Caso	Palabras clave	Periodo	Participación
Recortes	#19F, #contrareforma, #reformalaboral, #19Ftomalacalle, #razones19F, #huelgageneral, #29F, #11M, #29M, #29NoaLaHuelgaGeneral, #noalahuelgageneral, #pideunpiquete, #piqueteempresarial, #huelgadespierta, recortes, reforma laboral, retallades,#15S,#14N	Del 17/02/2012 Hasta 02/09/2015 1.293 días	10.435.262 tuits de 2.067.064 usuarios
Bankia	#bankia, Rodrigo Rato, Blesa, tarjetas black, #tarjetasblack y Caja Madrid	Del 01/03/2013 Hasta 21/10/2015 964 días	4.697.599 tuits de 627.666 usuarios

Tabla 3 Casos del barómetro social

3.2.2 MOVIMIENTOS SOCIALES

El movimiento social más significativo corresponde a las movilizaciones de mayo del 2011 que llevaron a ocupar la Puerta de Sol de Madrid y otras plazas de ciudades de España. Es conocido con el nombre de 15M por iniciarse en una manifestación convocada por Democracia Real el 15 de Mayo. Un año después surgió otro movimiento para acudir ante el Congreso de los Diputados el 25 de septiembre, al que se denominó 25S, y que tuvo mucha repercusión tanto en las redes sociales como en la participación ciudadana.

En ambos casos subyace una organización, por lo que resulta muy interesante analizar qué tipos de usuarios participaron y cómo se propagaron los mensajes. También aporta una visión temporal debido a que entre ambos movimientos trascurrió más de un año y por tanto es posible detectar cambios en las dinámicas de comunicación en los movimientos sociales.

En la Tabla 4 se detalla el contenido de ambos casos. Como se puede observar, los casos están acotados en el tiempo desde unos días previos a la movilización a otros posteriores al punto álgido de la movilización. Se han descartado los tuits que no estaban en este intervalo para medir el momento de máxima difusión.

Caso	Palabras clave	Periodo	Número de tuits
15M	#15M, 15-M, #democraciarealya, #tomalacalle, #Nolesvotes, #spanishrevolution, #acampadasol, #acampadabcn, #indignados, #notenemosmiedo, #nonosvamos, #yeswecamp	Del 13/05/2011 Hasta 01/06/2011 19 días	1.443.131 tuits de 206.601 usuarios
Rodea el Congreso	25S, #voces25S, #25S,asalto al congreso, @ocupaelcongreso, #ocupaelcongreso, ocupa el congreso, #tomaelcongreso, toma el congreso, 29-S, 29S, #29S,#vamos29S	Del 24/08/2012 Hasta 01/10/2012 38 días	1.101.041 tuits de 253.854 usuarios

Tabla 4 Casos de movimientos sociales

3.2.3 POLÍTICA

Desde el año 2010 al 2015 se han celebrado en España diez elecciones: unas Generales (2011), unas Europeas (2014), tres autonómicas en Cataluña (2010, 2012 y 2015), dos en Andalucía (2012 y 2015), unas en Galicia y País Vasco (2012), dos autonómicas y municipales (2011 y 2015). De todas ellas he seleccionado las generales de 2011 y las europeas del 2014 por ser de ámbito nacional junto con las de Cataluña por ser las más recientes. Adicionalmente he incluido el caso de los Diputados, que contiene los tuits publicados por diputados del Congreso durante casi cuatro años y el impacto de éstos en los usuarios.

Los mensajes se canalizan en Twitter por tres vías durante las campañas electorales: mediante tuits que incluyen etiquetas con lemas electorales, por medio de las noticias que circulan sobre los candidatos y por la actividad que desarrollan los perfiles de los partidos y cabezas de lista. De estas tres vías he seleccionado la tercera, es decir, los mensajes que

están relacionados con las cuentas en Twitter de los candidatos. Esta elección corresponde a una manera más objetiva de medir ya que no es posible coleccionar todas las etiquetas de la campaña ni todas las noticias relacionadas con las elecciones, pero si está acotada la actividad de los usuarios en Twitter.

Con estos cuatro casos se pueden contrastar los cambios que se producen a lo largo del tiempo, del 2011 al 2015 y de las diferencias de comunicación en periodo electoral o fuera de él.

En la Tabla 5 se especifica cada uno de los casos indicando el volumen de tuits, la participación de usuarios y el periodo de recogida de datos.

Caso	Usuarios	Periodo	Número de tuits
Diputados	Diputados del Congresos que tienen perfil en Twitter	Del 29/12/2011 Hasta 21/10/2015 1.392 días	10.706.514 tuits de 859.963 usuarios
Elecciones generales 2011	Candidatos (cabezas de lista) y partidos de las elecciones generales del 2011 con perfil en Twitter	Del 08/10/2011 Hasta 22/11/2011 45 días	452.989 tuits de 98.232 usuarios
Elecciones europeas 2014	Candidatos (cabeza de lista) y partidos de las elecciones europeas del 2014	Del 04/05/2014 Hasta 28/05/2014 24 días	732.679 tuits de 152.328 usuarios
Elecciones catalanas 2015	Candidatos (cabeza de lista) y partidos de las elecciones catalanas de 2015	Del 07/09/2015 Hasta 30/09/2015 23 días	437.641 tuits de 57.012 usuarios

Tabla 5 Casos de política

3.2.4 PRENSA

Una parte importante de los tuits que circulan en Twitter son noticias. La agenda de los medios condiciona las conversaciones, comentarios y opiniones de los usuarios en Twitter. Para estudiar la propagación de las noticias he seleccionado cuatro casos correspondientes a dos periódicos que tienen edición en papel y digital, El País y el ABC, y dos periódicos digitales, uno con mayor polarización política que el otro, eldiario.es y 20 Minutos.

En la Tabla 6 se muestra cómo todos los casos tienen el mismo intervalo de captura de datos. Las palabras clave proporcionan todas las referencias a las cuentas en Twitter de estos periódicos y los enlaces a sus noticias al incluir total o parcialmente el nombre del dominio en Internet del medio.

Caso	Palabras clave	Periodo	Número de tuits
El País	@el_pais, elpais	Del 07/05/2014 Hasta 21/10/2015 532 días	14.340.517 tuits de 1.752.712 usuarios
ABC	@abc_es, abc.es	Del 07/05/2014 Hasta 21/10/2015 532 días	3.161.108 tuits de 592.317 usuarios
El diario.es	@eldiarioes, eldiario.es	Del 07/05/2014 Hasta 21/10/2015 532 días	1.794.186 tuits de 252.471 usuarios
20 Minutos	@20m, 20minutos	Del 07/05/2014 Hasta 21/10/2015 532 días	3.161.108 tuits de 592.317 usuarios

Tabla 6 Casos de prensa

3.2.5 CASOS INTERNACIONALES

Para tener un referente internacional que permita contrastar la propagación en España con la propagación global o de otros países he escogido dos casos internacionales. El primero de ellos es el del Ébola que reúne los tuits durante la etapa en que varios infectados por esta enfermedad fueron repatriados a Europa o a Estados Unidos. Durante este periodo la reacción en Twitter fue global y muy intensa. El segundo corresponde a los mensajes relacionados con la ecología en el entorno anglosajón que contienen algún término en inglés relacionado con el medioambiente y la sostenibilidad.

En la Tabla 7 se puede observar que ambos casos tienen más de cuarenta millones de tuits y casi diez millones de usuarios.

Caso	Palabras clave	Periodo	Número de tuits
Ébola	ebola,ébola,#ebola,#ébola,#giveusyourserum	Del 04/08/2014 Hasta 24/10/2014 81 días	44.665.469 tuits de.9.946.940 usuarios
Ecology	Green economy, Sustainable development, Green jobs, Social inclusion, Human welfare, Renewable energy, Energy efficiency, Sustainable energy, Energy consumption, Sustainable cities, Resilient cities, Food security, Sustainable agriculture, Hunger eradication, Poverty eradication, Crop diversity, Malnutrition, Deforestation, Land-use change, Water quality, Water consumption, Access to safe drinking water, Global water crisis, Eutrophication, Marine and coastal biodiversity, Marine and coastal resources, Marine fisheries, Ocean fisheries, Fisheries depletion, Fisheries overexploitation, Overfishing, Marine and	Del 09/06/2012 Hasta 20/10/2015 1.228 días	42.671.868 tuits de 9.262.642 usuarios

Caso	Palabras clave	Periodo	Número de tuits
	coastal pollution, Sea level rise, Ocean acidification, Loss of coastal habitats, Loss of biodiversity, Marine protected areas, Natural disasters, Resilience, Disaster-resilient Societies, Resilient ecosystems, Human population growth, Overpopulation, Ecosystem services, Global change, Loss of biodiversity, Climate change, Global warming		

Tabla 7 Casos internacionales

3.2.6 CASOS DE TENDENCIAS

Para analizar los casos de tendencias he elegido dos muy recientes. El primero corresponde a la etiqueta #YoSoyNaranjito que llegó ser tendencia mundial cuando el líder del partido político Ciudadanos contratacó ante un comentario de un dirigente del Partido Popular (Congosto, 2015a). El segundo, #ZapataDimisión, recoge un precedente de renuncia de un cargo por la presión en las redes sociales. (Congosto, 2015b).

En la Tabla 8 se detalla el volumen de ambos casos

Caso	Palabra clave	Periodo	Número de tuits
#YoSoyNaranjito	#YoSoyNaranjito	Desde 05/03/2015 10:55 Hasta 05/03/2015 22:05 11,17 horas	51.352 tuits de 17.267 usuarios
#ZapataDimisión	#ZapataDimisión	Desde 13/06/2015 15:20 Hasta 15/06/2015 12:40 45,33 horas	130.954 tuits de 37.979 usuarios

Tabla 8 Casos de tendencias (TT)

3.2.7 TWITTER Y TELEVISIÓN

Para analizar la relación entre Twitter y televisión he seleccionado dos casos recientes en los que ambos medios se complementaron y amplificaron. El primer caso recoge los tuits del debate en Telemadrid entre las candidatas Manuela Carmena y Esperanza Aguirre durante la campaña electoral de las municipales de mayo del 2015. El segundo corresponde al debate entre Albert Ribera y Pablo Iglesias en la cadena de televisión La sexta, en el programa de Salvados durante octubre de 2015.

El volumen y la duración de ambos casos se describen en la Tabla 9.

Caso	Descripción	Periodo	Número de tuits
Debate Aguirre-Carmena 2015	#DebateTM	Desde 19/05/2015 19:05 Hasta 20/05/2015 0:00 4,92 horas	39.519 tuits de 11.941 usuarios
Albert vs. Pablo en la Sexta	#A bertvsPablo	Desde 18/10/2015 19:05 Hasta 19/10/2015 0:00 4,92 horas	256.046 tuits de 61.622 usuarios

Tabla 9 Casos de Twitter y televisión

PARTE III. Desarrollo

4 PROCESOS DE MEDICIÓN

Una parte importante del esfuerzo dedicado en la realización de esta tesis ha consistido en el desarrollo de herramientas de apoyo al análisis cuantitativo. Las necesidades han ido cambiando conforme he ido profundizando en la investigación. Las herramientas al principio eran muy elementales y poco a poco he tenido que adaptarlas a las nuevas necesidades.

Al comenzar a obtener colecciones de datos encontré que las APIs de Twitter fallaban, con la consiguiente pérdida de información. Reforcé los mecanismos de comunicación pero, pasado un tiempo, los datos crecían mucho y el tiempo de proceso era muy elevado. Entonces, comencé a estructurar los datos en paquetes y a adaptar los algoritmos a un proceso por partes para luego integrar los resultados. Una vez que el tamaño y el proceso de los datos ya no era problema, diseñé una interfaz gráfica y comencé a procesar automáticamente todos los experimentos. El conjunto de herramientas que hicieron posible todo esto se denomina T-hoarder. Este nombre está compuesto de la T de Twitter y de la palabra *hoarder*, que significa acaparador, y también se utiliza como adjetivo para las personas que tienen el síndrome de Diógenes. Por tanto, T-hoarder es un "Diógenes digital" que acapara información de Twitter, pero lo hace de una manera ordenada y estructurada.

El resto de herramientas que he desarrollado han seguido el mismo mecanismo de procesado por partes e integración de resultados, lo que ha evitado que el tamaño de la información y el tiempo de ejecución fueran un problema.

4.1 PLATAFORMA T-HOARDER

T-hoarder fue concebido como un medio para almacenar tuits sobre ciertos temas candentes en los que se pudieran analizar los distintos tipos de propagación que podían tener los mensajes. Actualmente la plataforma es una fuente de información elaborada, que puede ser consultada por aquellas personas que muestran un interés por los acontecimientos sociales en España. Ha transcendido su uso más allá de la investigación para dar a conocer la evolución de un conjunto de eventos de interés social como movilizaciones ciudadanas, opiniones sobre la crisis económica, escándalos políticos, corrupción, etc.

La plataforma T-hoarder almacena tuits por líneas temáticas y los procesa automáticamente en tres ejes: temporal, espacial y de relevancia. El eje temporal permite ver tanto la evolución en el tiempo de un conjunto de indicadores como la proporción de mensajes retransmitidos, los usuarios más mencionados o más activos, los *hashtags* más populares, las palabras más frecuentes, etc. El eje espacial ubica los tuits geográficamente y la relevancia muestra los mensajes más difundidos. Esta plataforma está dotada de una interfaz gráfica interactiva que facilita la navegación por estos tres ejes.

4.1.1 ARQUITECTURA T-HOARDER

T-hoarder tiene una arquitectura sencilla que evita la dependencia de otros paquetes software. Utiliza Unix como sistema operativo y está desarrollado en Python. Para almacenar información usa ficheros Unix en vez de bases de datos por los siguientes motivos:

- Poder funcionar en entornos de desarrollo mínimos, como por ejemplo en una Raspberry PI[30].
- Facilitar el traslado de los datos de un servidor a otro. Cuando las bases de datos son muy grandes, los respaldos/restauraciones son problemáticos.
- No tener que definir a priori un modelo de datos desconociendo el tipo de información que se va a procesar.
- Permitir mezclar las distintas colecciones de datos fácilmente.
- No requerir el acceso a la información de forma aleatoria ya que el procesado de los datos es secuencial.

Su arquitectura se estructura en tres capas desacopladas para evitar que el tiempo de ejecución de una no interfiera sobre las otras. La comunicación entre estas capas es siempre mediante ficheros. La división funcional es la siguiente:

- ***Capa* 1**: Recolección y almacenamiento de datos
- ***Capa* 2**: Procesado de datos
- ***Capa* 3**: Visualización

En la Figura 7 se muestra esquemáticamente su arquitectura

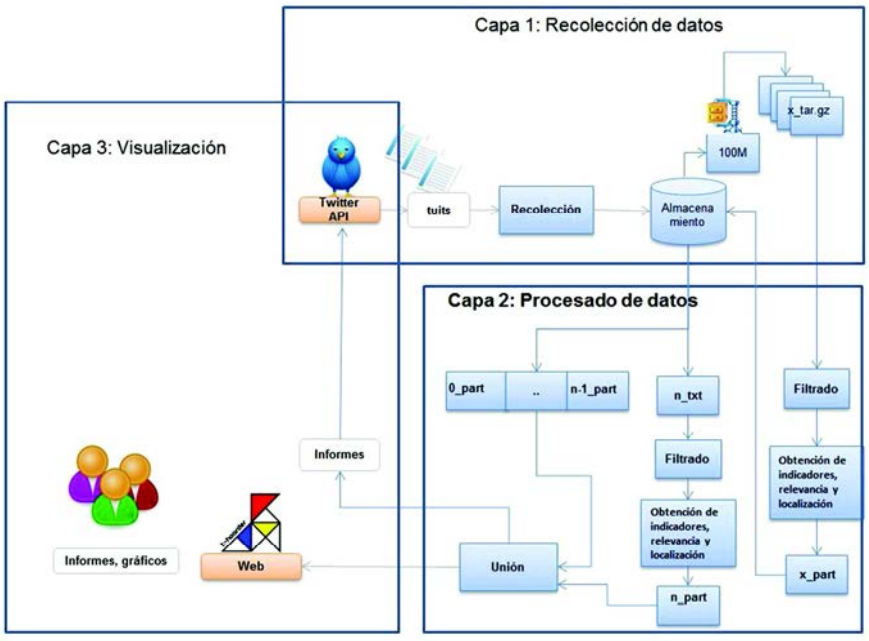

Figura 7 Arquitectura de T-hoarder

El código con el que está desarrollado T-hoarder se encuentra compartido en el repositorio Software GitHub[31]. Los componentes están programados en Python, estando formado el

[30] http://www.raspberrypi.org/
[31] https://github.com/congosto/t-hoarder

nombre de cada programa por el nombre del componente y la extensión “py”, siendo fácilmente localizables en el repositorio.

4.1.2 CAPA 1: RECOGIDA Y ALMACENAMIENTO DE DATOS

T-hoarder dispone de una aplicación llamada `Tweetdekc` y un conjunto de usuarios en Twitter. Para poder acceder a las APIs de Twitter mediante OAuth es necesario crear las claves de acceso para la aplicación y para los usuarios. Estas claves se generan en el componente `tweet_auth` y quedan almacenadas en ficheros para que los componentes puedan autenticarse automáticamente ante las APIs.

T-hoarder utiliza la API Streaming frente a la API REST por los siguientes motivos:

1. Porque es el método más adecuado para obtener información en tiempo real con la única limitación de los 50 TPS. La alternativa sería el uso del método GET /search/tweets de API REST, pero esto obligaría a hacer consultas periódicas para obtener los mensajes con la consiguiente complicación tanto en determinar la frecuencia de muestreo como evitar el límite de velocidad.
2. Porque cuando se inicia un experimento, generalmente, se recogen los mensajes desde ese mismo momento durante un periodo largo de tiempo. En otro tipo de experimentos de duración corta (días) y en los que es necesario remontarse a una fecha anterior, caso de los *trending topic* (TT) no es necesario el uso de T-hoarder. Es más razonable obtener los tuits con el método GET /search/tweets de la API REST, con los que se podrán recuperar hasta siete días previos y obtener el *dataset* completo.

4.1.2.1 Captura de datos

Dentro de las opciones de búsqueda de la API Streaming, T-hoarder usa preferentemente las palabras clave y los usuarios. La geolocalización es información que carece de contexto temático y representa una muestra muy pequeña de los tuits (1,5% en España).

Los flujos de datos se obtienen mediante el componente `tweet_streaming`. Se pueden extraer varios flujos simultáneos ejecutando el componente en paralelo, cada uno con un usuario diferente.

La API suministra los tuits solicitados en formato JSON. Los datos se transforman a texto plano separados por tabulaciones en una línea por mensaje. Esto permite leer los mensajes que se van recogiendo con facilidad e importarlos en una hoja de cálculo, si se desea.

De toda la información recibida se seleccionan los datos que son de utilidad para analizar el contexto del tuit:

- **id_tweet**: identificador del tuit. Es un número creciente que va asignando Twitter secuencialmente a cada mensaje.
- **timestamp**: fecha y hora GMT de tuit.
- **@autor**: nombre de usuario del autor del tuit.
- **texto**: texto del tuit.
- **app**: aplicación desde la que se ha publicado el tuit.

- **id_autor**: identificador del autor. Es un número creciente que va asignando Twitter a los usuarios conforme se van dando de alta.
- **seguidores**: número de seguidores en el momento de la publicación.
- **siguiendo**: número de usuarios seguidos en el momento de la publicación.
- **mensajes**: número de tuits publicados anteriormente.
- **localización**: localización declarada en el perfil de usuario.
- **url**: enlace si el tuit contiene una URL, en caso contrario se almacena el valor *None.*
- **geolocalización**: coordenadas si el tuit está geolocalizado, en caso contrario se almacena *None.*
- **nombre**: nombre proporcionado por el usuario.
- **bio**: descripción del usuario.
- **url_media**: URL si el tuit contiene información multimedia, en caso contrario se almacena el valor *None.*
- **tipo_media**: tipo de información multimedia (foto, video,.), en caso de no existir, su valor es *None.*
- **lenguaje**: idioma del tuit, si se ha podido detectar.

Algunos datos como **localización**, **nombre** y **bio** pueden contener saltos de línea o tabulaciones. Para evitar conflictos con los delimitadores se filtran las tabulaciones y los saltos de línea en estos datos.

Puede parecer poco eficiente almacenar información redundante como **localización**, **nombre** y **bio** pero se ha llegado a una solución de compromiso para que la información de los tuits esté auto-contenida, evitando la consulta de información exterior. Además, estos datos pueden cambiar con el tiempo, tanto que hasta existe la herramienta bioischanged[32] para conocer el historial de cambios de un usuario. Por este motivo, asociar esta información al momento en que se publicó el tuit es más riguroso.

4.1.2.2 Almacenamiento de los datos

Los datos se han organizado en una estructura de directorios predefinida y con una notación de prefijos y sufijos para facilitar la localización de la información almacenada. Desde una raíz inicial, denominada `$STORE`, se guardan los distintos flujos de datos de cada experimento y las claves de acceso a la API de Twitter (Figura 8).

[32] http://bioischanged.com/

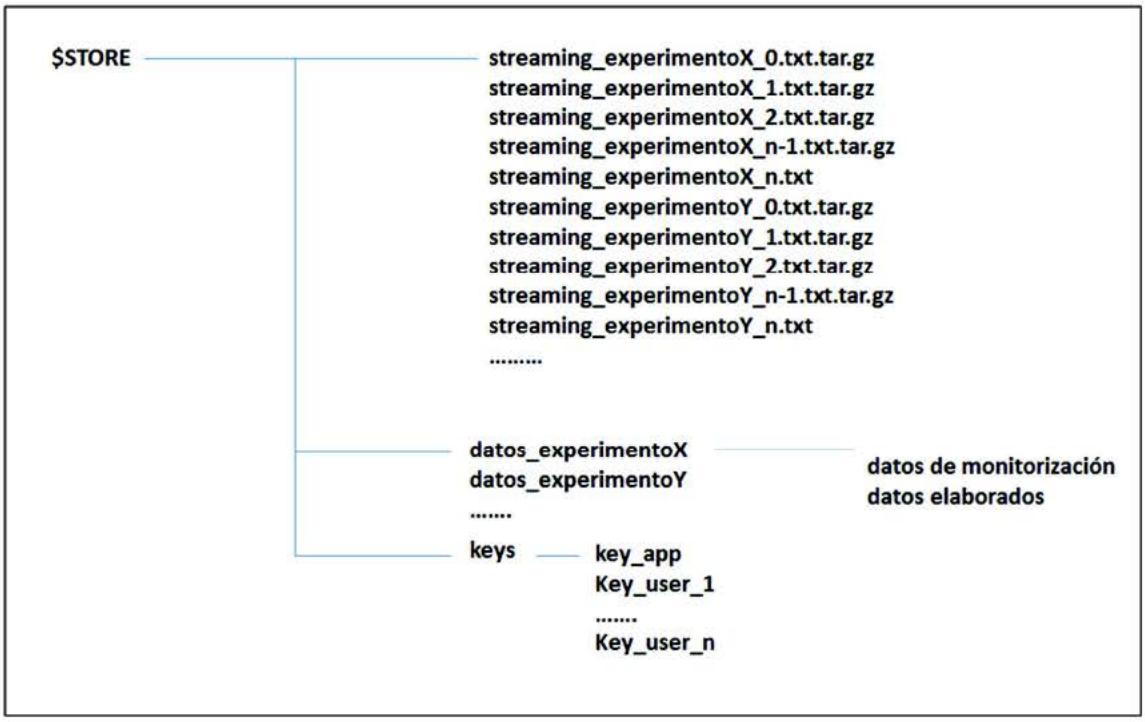

Figura 8 Estructura de directorios

Los flujos de datos de cada experimento se archivan en paquetes de ficheros. Los ficheros se van generando con el patrón de nombre `streaming_experimento_x.txt, (x: 0,n)`. El primer fichero se numera desde cero y cuando alcanza el tamaño de 100MB se comprime y se crea un nuevo fichero con una numeración creciente. Al ser ficheros de texto, la compresión es muy eficiente y el tamaño de los datos se reduce a la tercera parte.

Debajo de `$STORE` se crea un directorio para cada experimento, cuyo nombre será `datos_experimento_X`. Dentro de ese directorio se encuentra el fichero para seleccionar los tuits, bien por palabras clave, por usuarios o por localizaciones. También se utiliza este directorio para almacenar los datos elaborados.

Las claves de acceso de la aplicación y de los usuarios se archivan en el directorio `$STORE/keys`.

4.1.3 CAPA 2: PROCESADO DE DATOS

Esta capa se ejecuta de forma independiente a la captura y almacenamiento para evitar pérdida de tuits. Se utiliza un *cron* para ejecutar periódicamente los algoritmos de esta capa.

Los experimentos de larga duración podrían generar colecciones de datos muy grandes que resultarían muy costosas de procesar. Sin embargo, gracias al método de almacenamiento de los flujos de datos en paquetes de ficheros delimitados por tamaño, es viable realizarlo por partes para luego integrar los resultados. Este método tiene las siguientes ventajas:

1. Es factible procesar directamente los datos comprimidos con Python sin que se incremente el tiempo de ejecución.
2. Al tener los ficheros un tamaño manejable, no hay problemas de escalabilidad de los algoritmos.
3. Es posible procesar en paralelo los distintos paquetes de ficheros.
4. No es necesario volver a procesar un paquete ya procesado, tan solo los datos nuevos desde la última iteración.

En esta fase también se dispone de una estructura de directorios predeterminada (Figura 9). En el directorio `$RESOURCES` se almacenan los distintos recursos necesarios para procesar los datos, como por ejemplo: tablas de nombres por género, geolocalización de localidades, diccionarios para clasificar tuits, etc.

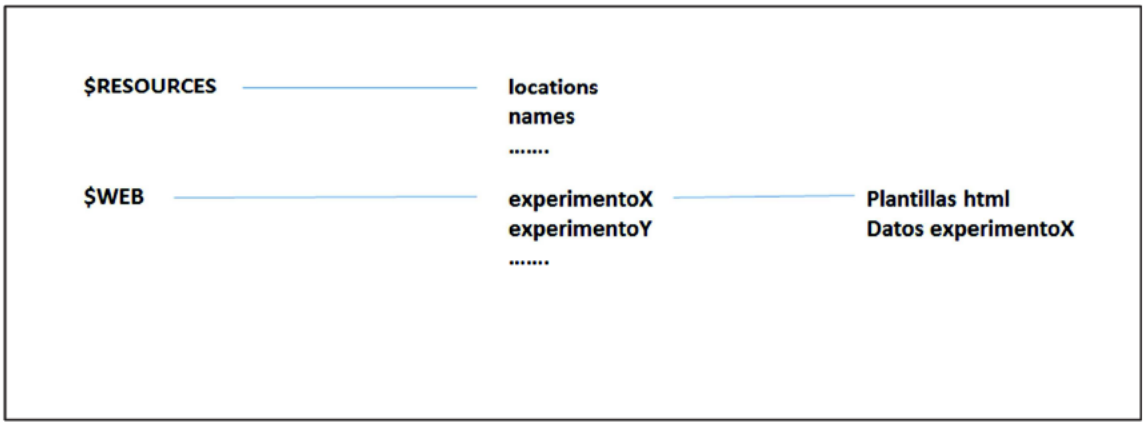

Figura 9 Entorno de proceso

En el directorio `$WEB` existirá un directorio para cada experimento en los que se almacenarán los datos elaborados para ser presentados en la interfaz gráfica web.

Para cada uno de los paquetes se realizan las siguientes operaciones:

- Filtrar los falsos positivos
- Extraer los indicadores
- Extraer relevancia
- Extraer localización
- Generar estado del paquete

4.1.3.1 Filtrado de falsos positivos

Algunas veces se capturan falsos positivos debido a que los términos de búsqueda contienen palabras ambiguas o las palabras de las expresiones no aparecen en el orden esperado. Por ejemplo, si recogemos tuits que contengan la expresión “metro de Madrid” la API nos proporcionará todos los tuits que contengan las palabras “metro” y “Madrid” independientemente del orden en que aparezcan. El resultado puede incluir tuits relacionados con el fútbol debido a que Metro es un canal de televisión que emite partidos del Real Madrid o del Atlético de Madrid. Los falsos positivos se detectan cuando al procesar los tuits aparecen mensajes que no son del contexto buscado. Esto implica que hay que descartar los mensajes no deseados y reprocesar el paquete. Para el filtrado se utiliza un fichero llamado `filter.txt` que contiene un conjunto de palabras o

expresiones que no corresponden al contexto y que permite desechar los mensajes que las contengan.

El filtrado se realiza mediante el componente `tweets_select_filter` que permite seleccionar o excluir tuits que contengan algunos términos o expresiones o que hayan sido publicados por ciertos usuarios.

4.1.3.2 Extraer los indicadores

Para observar la evolución de los datos almacenados se van calculando una serie de indicadores para cada día que serán expuestos más tarde en el eje temporal. Estos indicadores proporcionan una idea de la participación y modo de publicación de los mensajes:

- Número de tuits: cantidad de tuits recogidos.
- Número de RTs: cantidad de tuits que son difundidos mediante el mecanismo de retransmisión.
- Número de *replies*: cantidad de tuits que son respuestas a otro tuit.
- Número de menciones: cantidad de tuits que contienen menciones.
- Número de usuarios únicos: cantidad de usuarios diferentes que han tuiteado.
- Número de usuarios nuevos: cantidad de usuarios que tuitean por primera vez ese día.
- Top *hashtags*: para cada uno de los hashtags más mencionados, la cantidad de veces que aparece en los tuits.
- Top palabras: para cada una de las palabras más frecuentes (no se tienen en cuenta las *stop words*), la cantidad de veces que aparece en los tuits.
- Top usuarios mencionados: para cada uno de los usuarios más mencionados, la cantidad de veces que aparece en los tuits.
- Top usuarios activos: para cada uno de los usuarios más activos, la cantidad de tuits que han publicado.

Los indicadores se extraen mediante el componente `tweets_counter` en dos pasos. El primer paso va descomponiendo cada tuit en entidades que se van acumulando de forma global. Una vez finalizado, el otro paso obtiene las entidades de cada tipo más frecuentes y se contabiliza su aparición día a día.

```
Primer paso:
      Para cada tuit:
            Obtener autor y contabilizarlo
            Obtener menciones a usuarios y contabilizarlas
            Obtener el origen del tuit y contabilizarlo
            Obtener la localización declarada del autor y
            almacenarla
            Obtener las palabras del tuit que no sean stopwords y
            almacenarlas
            Obtener los hashtags del tuit y almacenarlos
Segundo paso:
      Obtener el top de autores, menciones, orígenes del tuit,
      localizaciones, palabras y hashtags
```

```
Para cada día:
      Para cada tuit de ese día:
            Contabilizar el top de autores
            Contabilizar el top menciones a usuarios
            Contabilizar el top de los orígenes del tuit
            Contabilizar el top de localizaciones
            Contabilizar el top de palabras
            Contabilizar el top de hashtags
```

4.1.3.3 Extraer relevancia

En T-hoarder, la relevancia se mide por la difusión de los mensajes. Los mensajes se difunden porque captan la atención de otros usuarios que a su vez quieren darle visibilidad en su entorno. En Twitter, la propagación de mensajes se realiza mediante el mecanismo de RT. El RT es una convención creada en los inicios de Twitter por los usuarios que querían compartir un tuit con sus seguidores y se realizaba mediante la publicación del mensaje de otro usuario anteponiéndole las siglas RT y el nombre del autor original. A partir del 2009 Twitter incluyó un botón de RT que hacía lo mismo pero automáticamente, lo que facilitó mucho la propagación de mensajes. Generalmente se difunden los tuits con los que se está de acuerdo, por lo que se puede contabilizar como un voto positivo al mensaje (Conover et al., 2011).

T-hoarder descarta usar el dato del número de RTs que suministra la API de Twitter por ser un dato dinámico que varía con el tiempo y que además, en el momento de la captura del tuit, en tiempo real, su valor sería cero o un valor muy bajo. En su lugar, detecta la difusión de los tuits comparando la similitud de mensajes y teniendo en cuenta la estructura del RT. Por lo tanto detecta RTs automáticos y RTs manuales.

Se considera que un tuit es la retransmisión de otro cuando comienza por "RT @usuario" y el resto del texto coincide más de un 90% con algún tuit original anterior (Figura 10). La coincidencia puede no ser del 100% porque al retransmitirse mensajes de casi 140 caracteres se truncan, como es el caso del ejemplo siguiente.

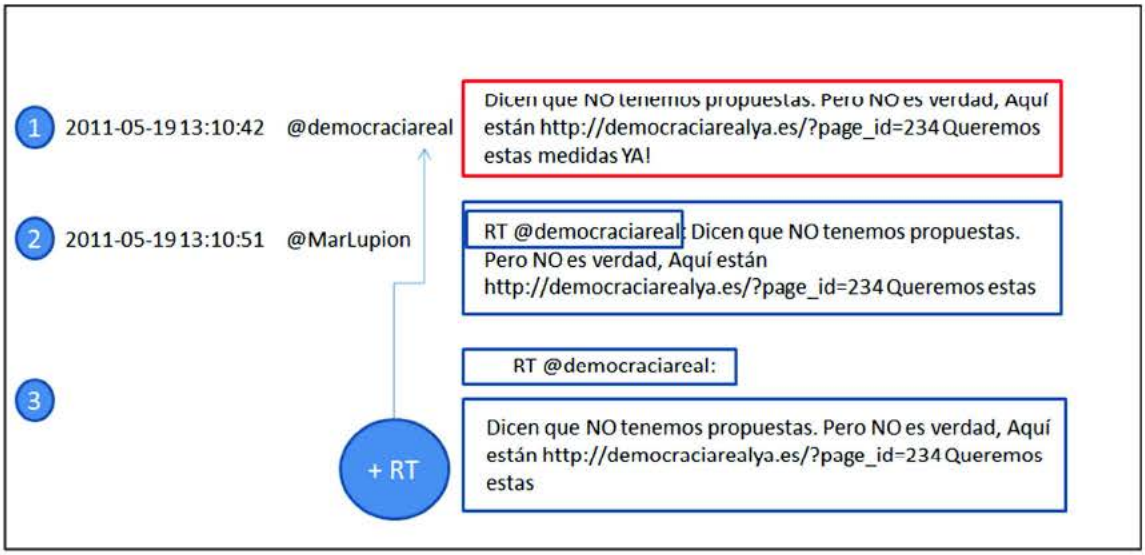

Figura 10 Detección de difusión de tuits por similitud de mensajes

La difusión de mensajes se calcula por día y para todo el período de captura de tuits. De esta forma se conoce lo más relevante de cada jornada y lo más destacado en global. Los datos que se almacenan de los tuits más difundidos son:

- Identificador del tuit.
- Fecha y hora del tuit.
- Autor del tuit.
- Texto del tuit.
- Número de veces que se ha difundido.

Los mensajes más difundidos se obtienen con el componente `tweets_talk`**.** Cada tuit es comparado con un *buffer* de tuits previos analizados. Si se detecta que es una retransmisión de algunos de ellos se incrementa el contador de RTs, en caso contrario se almacena en el *buffer* como nuevo mensaje. Cada hora o cada 15.000 tuits se salvan los 2.000 tuits más difundidos del *buffer* y el resto se descarta. De esta manera se evita que el número de comparaciones con tuits no difundidos ralenticen el proceso. Se mantiene un búfer global y otro del día.

```
Para cada tuit:
      ¿Es RT de algún tuit del búfer global?
            Sí:
                  Incrementar contador de RT del tuit del búfer
                  global
            No:
                  Almacenar el tuit en el búfer global
      ¿Es RT de algún tuit del búfer del día?
            Sí:
                  Incrementar contador de RT del tuit del búfer del
                  día
            No:
                  Almacenar el tuit en el buffer del día
      ¿Hay cambio de hora o el búfer tiene más de 15000 tuits?
            Sí:
                  Reducir el búfer global a los 2000 tuits con más
                  RTs
                  Reducir el búfer del día a los 2000 tuits con más
                  RTs
      ¿Hay cambio de día?
            Sí:
                  Almacenar el búfer del día
                  Vaciar el búfer del día
Almacenar el búfer global
```

4.1.3.4 Extraer localización

La ubicación de los tuits se puede conocer por dos caminos. El primero es por la localización declarada del perfil del usuario. Este dato puede no estar completado o contener el nombre de una ubicación ficticia por lo que no es posible ubicar todos los mensajes geográficamente. No obstante, es posible localizar un porcentaje elevado de

tuits (entre el 60% - 70%). La segunda opción la proporcionan los tuits geolocalizados de los usuarios que tienen activada la geolocalización en Twitter. En este caso el porcentaje es mucho más pequeño (en España el 1,5%).

Para la localización por perfil de usuario se utiliza un fichero con los municipios de España[33] a los que se les ha calculado previamente sus coordenadas (longitud y latitud) y se han clasificado por autonomía y provincia. Con estos datos se pueden situar los tuits en un mapa y también se pueden agregar por provincia o por autonomías. Para la geolocalización simplemente se extraen las coordenadas del tuit.

Para cada día se almacenan por un lado los tuits localizados por perfil del usuario y por otro los tuits geolocalizados. En ambos casos se guardan los mismos datos:

- Identificador del tuit
- Fecha y hora del tuit
- Autor
- Texto del tuit
- Coordenadas

Las localizaciones se obtienen con el componente `tweets_location` que analiza la localización declarada del autor del tuit y comprueba si coincide con algún municipio, provincia o autonomía de España.

```
Para cada tuit
      ¿Localización coincide con municipio?
            Sí:
                  Agregar las coordenadas del municipio
            No:
                  ¿Localización coincide con provincia?
                  Sí:
                        Agregar las coordenadas de la capital
                        provincia
                  No:
                        ¿Localización coincide con la autonomía?
                              Sí:
                                    Agregar las coordenadas de la
                                    capital de la autonomía
      ¿Está el tuit localizado?
            Sí:
                  Añadir al fichero de localizaciones
      ¿Está el tuit geolocalizado?
            Sí:
                  Añadir al fichero de geolocalizaciones
```

4.1.3.5 Generar estado del paquete

Cuando un paquete es procesado total o parcialmente se almacena, en un fichero denominado `experimento_x_status.txt`, una información de estado:

- Fecha inicial: fecha de tuit más antiguo.

[33] http://www.ine.es/jaxi/menu.do?type=pcaxis&path=/t20/e245/codmun&file=inebase

- Fecha final: fecha de tuit más reciente.
- Estado: estado del proceso del paquete. Puede tomar los valores: semi-procesado, procesado.
- Ultimo tuit procesado: identificador del último tuit procesado.
- Longitud del paquete en el momento de procesarlo.
- Número de tuits.
- Tiempo de ejecución: tiempo de ejecución del paquete.

4.1.3.6 Integración de resultados

Los resultados están calculados por día, por lo que la integración es algo tan sencillo como la concatenación de resultados. Solo hay que tener en cuenta el efecto "borde" que se produce al dividir las colecciones de datos en paquetes de 100K. La partición puede dejar un día en diferentes paquetes.

En el caso de los *tops* (palabras, *hashtags*, usuarios mencionados y usuarios activos) se reduce del *top* 1000 que se almacena al *top* 10 para su visualización. Por este motivo hay que recalcular cuáles han sido los *tops* en el conjunto total de la colección de datos.

Los resultados se depositan en el directorio de intercambio `$WEB` para que el servidor web pueda acceder a ellos. El formato es texto plano con separadores y adaptado a las herramientas de visualización. Los datos se integran con el componente `join_results`.

```
Para cada paquete de datos
      Almacenar contadores de entidades
      Almacenar top de entidades (teniendo en cuenta que un día
      puede estar en dos paquetes diferentes)
      Almacenar RTs globales
      Almacenar RTs por día (teniendo en cuenta que un día puede
      estar en dos paquetes diferentes)
      Almacenar localizaciones
Generar resultado final de contadores de entidades
Generar top de entidades reduciendo el top de 1000 a 10
Generar RTs globales
Generar RTs por día
Generar localizaciones
```

4.1.4 CAPA 3: VISUALIZACIÓN

Para conocer la evolución de la información recuperada, T-hoarder dispone de unos paneles web que permiten visualizar los datos procesados. Estos paneles están construidos mediante una estructura que alberga <*iframes*>.

El <*iframe*> es un recurso HTML que permite anidar documentos HTML. Es muy utilizado para incrustar pequeñas piezas HTML con una función específica dentro de una página web. El diseño con <*iframe*> permite construir las webs como si fueran un puzle.

Estos paneles se apoyan en los siguientes recursos:

- El *framework* boostrap[34] de HTML, estilos CSS y JavaScript que permiten dar una estructura, un estilo y una interactividad a los distintos elementos del panel.
- La librería dygraphs[35] para las gráficas temporales, que tiene opciones avanzadas para favorecer la interactividad, permitiendo hacer zoom y llamar a funciones desde un punto de la gráfica para contextualizar la información.
- Google Maps para la visualización de mapas (en un futuro en Cartodb[36]).

La creación de paneles se realiza con un conjunto de plantillas genéricas que se particularizan para cada caso mediante el comando `make_panel`.

4.1.4.1 Plantilla de la página principal

La plantilla `home.html` contiene la estructura del panel Web en la que se particulariza la descripción del experimento y el acceso a las distintas gráficas temporales o mapas. Para ello se sustituye el *token* "@experiment" por el nombre del experimento. La plantilla principal consta de cuatro partes (Figura 11):

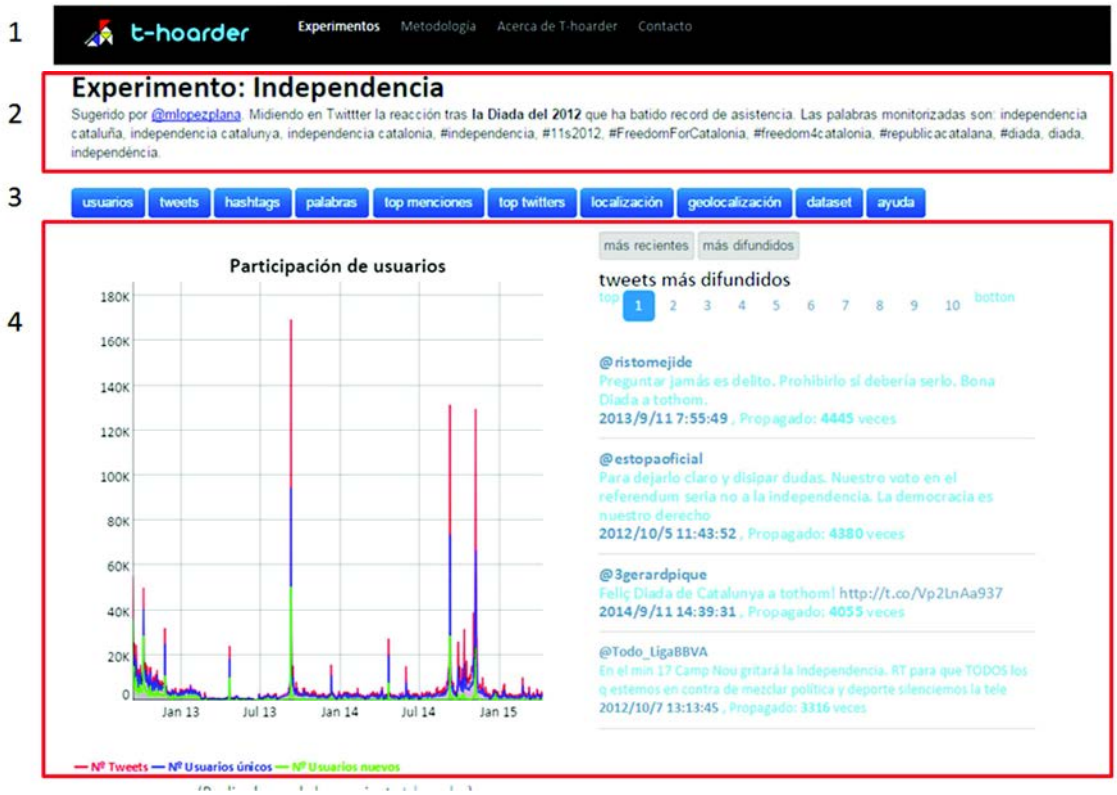

Figura 11 Panel T-hoarder

1. Barra de navegación desde la que se puede acceder a la información de la plataforma. Esta barra es común a todos los paneles.
2. Descripción del experimento con las entidades que están siendo monitorizadas. Es un *iframe* con información textual.

[34] http://getbootstrap.com/
[35] http://dygraphs.com/
[36] http://cartodb.com

3. Opciones de menú para seleccionar distintas vistas de la información: usuarios, tipo de tuits, *hashtags* y palabras más frecuentes, usuarios más mencionados y más activos, localización y geolocalización de los mensajes, información del *dataset* y la ayuda.
4. Gráficas interactivas con la información seleccionada en el menú. Es un *iframe* en el que se incrusta la página HTML que le corresponde a la opción del menú seleccionada. Existen dos tipos de gráficas, las temporales donde se muestra la evolución de los distintos indicadores y las geográficas que se representan mediante un mapa.

4.1.4.2 Plantilla para gráficas temporales

La plantilla `grafica_panel_cgi.html` contiene la estructura de la gráfica temporal que se particulariza para cada experimento. Para ello se sustituye el *token* "@experiment" por el nombre del *dataset* y el *token* "@data_file" por el nombre del fichero con los datos. Por lo tanto, esta plantilla se adapta a cada una de las gráficas temporales del experimento.

Las gráficas temporales constan de dos partes interrelacionadas (Figura 12):

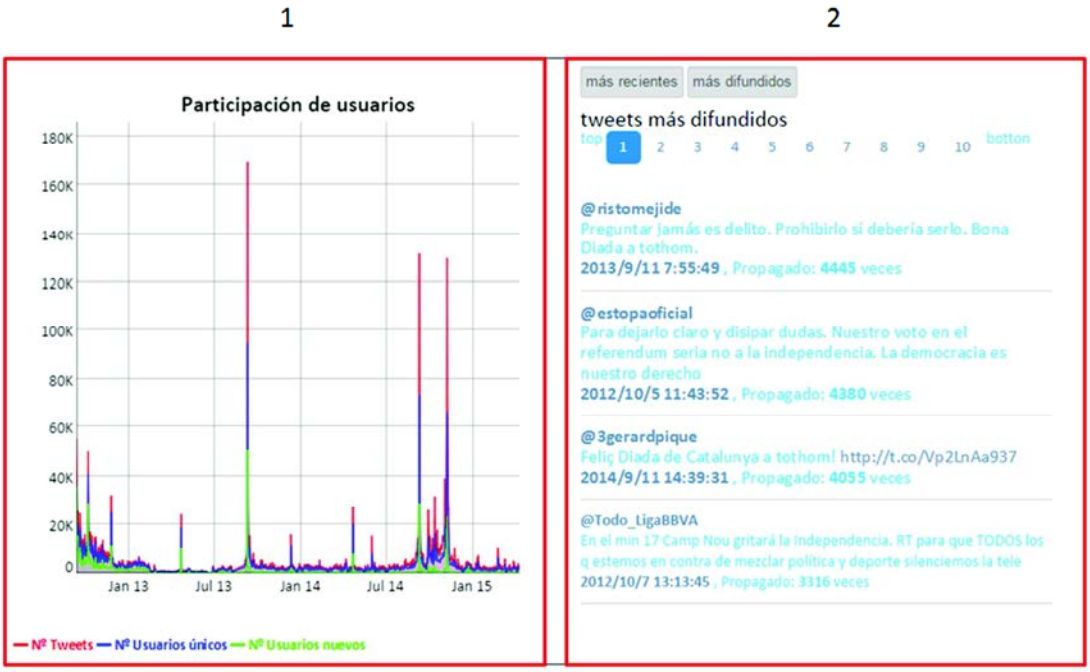

Figura 12 Panel temporal

1. La evolución de la entidad seleccionada, en la que se muestra gráficamente la variación de los distintos valores en el tiempo. Pasando el ratón por la gráfica se pueden ver los valores numéricos de los elementos de las leyendas. Para hacer *zoom*, se pulsa botón izquierdo del ratón y se arrastra. Para eliminar el *zoom*, se hace doble clic.
2. Los tuits más relevantes de la entidad. Por defecto se muestran los más difundidos recientemente. Es posible ver los más propagados durante toda la duración del experimento pulsando en el botón "más difundidos". Para descubrir los tuits más populares en un día concreto tan solo hay que hacer clic en la fecha de la gráfica

de la izquierda. En todos los casos, los mensajes están paginados de cuatro en cuatro y se pueden consultar hasta un máximo de cuarenta.

4.1.4.3 Plantilla para localización de tuits

La plantilla `grafica_location.html` contiene un plano de Google Maps en el que se representa la frecuencia de tuits por áreas mediante un mapa de calor, resaltando con un código de color las zonas más densas de tuits (Figura 13). Se particulariza sustituyendo el *token* "@data_file" por el nombre del fichero con los datos.

Figura 13 Origen de los tuits según la localización declarada en los perfiles de usuario

4.1.4.4 Plantilla para geolocalización de tuits

La plantilla `grafica_geolocation.html` contiene un mapa de Google Maps para situar los tuits geolocalizados mediante un puntero (Figura 14). Se particulariza sustituyendo el *token* "@data_file" por el nombre del fichero con los datos. Pasando el ratón por el puntero se puede leer el mensaje publicado en ese lugar.

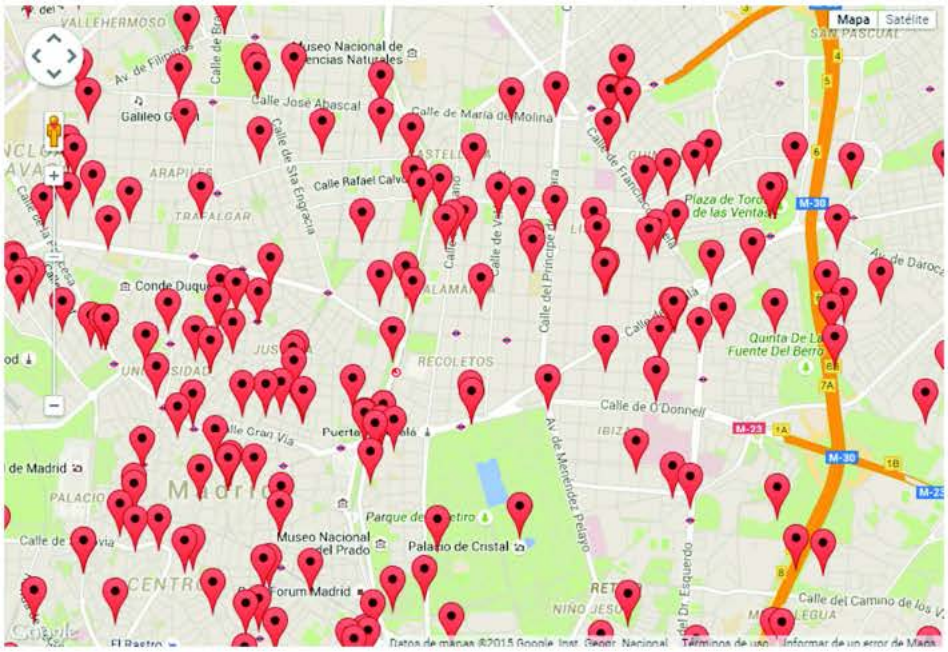

Figura 14 Tuits geolocalizados

4.1.5 EVALUACIÓN

Todos los componentes de T-hoarder calculan el tiempo de ejecución que han consumido. Esta medida permite planificar los recursos de computación y conocer la escalabilidad de los algoritmos. También sirve para poder compararlo con otras plataformas similares para evaluar la agilidad de su código.

El entorno de ejecución utilizado son dos servidores Unix con una CPU de 2GHz (8 *cores*) y 8 Gigabytes de memoria, con un sistema operativo Debian 3.2.65-1+deb7u1. Los componentes evaluados son `tweets_counter`, `tweets_talk`, `tweets_location` y `join_results`. Para determinar el comportamiento de los algoritmos en distintos tipos de experimentos, se han seleccionado cuatro de ellos atendiendo a dos características: frecuencia de publicación y duración del periodo de recogida de tuits. El número de tuits procesados oscila entre 10 millones y 44,5 millones. En la Tabla 10 se resume el entorno de la evaluación.

Infraestructura	Componentes	Experimentos
CPU de 2GHz (8 *cores*) 8 Gigabytes de memoria SO Debian 3.2.65-1+deb7u1	tweets_counter tweets_talk tweets_location join_results	10-4,54 millones de tuits 26-169 paquetes de 100 MB

Tabla 10 Entorno de evaluación

Los experimentos seleccionados son:

- `Diputados`: baja frecuencia y larga duración, contiene el impacto de los tuits de los diputados del parlamento español.

- `El País`: de frecuencia y duración media, recoge los mensajes que contienen una URL del periódico El País o mencionan a su perfil en Twitter.
- `Ecology`: frecuencia media y larga duración, almacena los tuits que mencionan palabras relacionadas con la sostenibilidad del entorno.
- `Ebola`: de frecuencia alta y duración media, contiene los tuits que hablan sobre el ébola.

En la Tabla 11 se muestran las características de los cuatro experimentos seleccionados en el momento de realizar la evaluación.

Experimento	Número de paquetes	Número de tuits	Tamaño	Media de tuits al día	Días	Desde	Hasta
Diputados	26	10.100.380	2,77 GB	8.177	1.235	11-12-29	15-05-17
El País	49	11.703.899	5,15 GB	27.052	433	14-05-07	15-07-14
Ecology	100	37.219.566	10,28 GB	32.940	1.130	12-06-09	15-07-14
Ebola	169	44.550.169	16,80 GB	555.241	80	14-08-04	14-10-24

Tabla 11 Características de las colecciones de datos

La evaluación consta de dos fases para cada una de las colecciones de datos. En la primera fase se procesan los distintos paquetes midiendo su tiempo de ejecución para los componentes: tweets_counter, tweets_talk y tweets_location

En la segunda fase se integran los resultados de cada uno de los paquetes generando el resultado total mediante el componente `join_results`.

4.1.5.1 Proceso de los paquetes

El tamaño de los paquetes es aproximadamente 100Mb. El número de mensajes de cada uno de ellos depende del formato de los tuits almacenados. En los experimentos más antiguos se almacenaba menos información y, conforme se fue necesitando, se incluyeron datos adicionales al final de cada registro a fin de que los componentes fueran compatibles hacia atrás. Por este motivo, en las colecciones `Diputados` y `Ecology` el número de tuits por paquete oscila entre 300.000 a 400.000, mientras que en las de `El País` y `Ebola`, más modernos, el rango se encuentra entre 210.000 y 300.000. La variación del número de tuits en las colecciones de datos del mismo formato depende del número de URLs y geolocalización que contengan y del tamaño de las biografías de los usuarios que participan.

En los cuatro experimentos seleccionados, el tiempo medio de ejecución por mensaje oscila entre 1,3-1,5 milisegundos, una diferencia muy pequeña. De estos datos se deduce que la frecuencia de tuits y la duración del experimento no influyen demasiado en el tiempo necesario para procesar un mensaje. La cantidad de tuits que T-hoarder es capaz de procesar en un segundo se muestran en la Tabla 12.

Experimento	tweets_counter	tweets_talk	tweets_location	En conjunto
Diputados	1.282	3.377	2.314	663
El País	1.392	4.481	2.184	714
Ecology	1.492	4.702	2.142	741
Ebola	1.485	3.312	2.098	689

Tabla 12 Número de tuits ejecutados por segundo

No todas las plataformas indican el tiempo de ejecución de sus componentes. De las que se han analizado en el estado del arte se ha encontrado esta información en la plataforma Trendminer (Preotiuc-Pietro & Samangooei, 2012). Esta plataforma analiza los tiempos de ejecución que consume cuando descompone un tuit en entidades y analiza el idioma. En un entorno local similar al utilizado por T-hoarder es capaz de analizar 510.000 tuits a la hora, es decir 141,66 tuits por segundo. Este dato indica que T-hoarder es más eficiente que Trendminer, ya que su análisis es más extenso y es capaz de procesar casi cinco veces más tuits.

Los resultados de la evaluación de cada una de las colecciones de datos se muestran en las Figura 15-Figura 18. En cada una de ellas se representa para cada paquete el tiempo de ejecución (eje y) y el número de tuits (eje x). En los cuatro casos el tiempo de ejecución es lineal en función del número de tuits. De los tres componentes analizados `tweets_counter` es el más costoso y `tweets_talk` el más liviano.

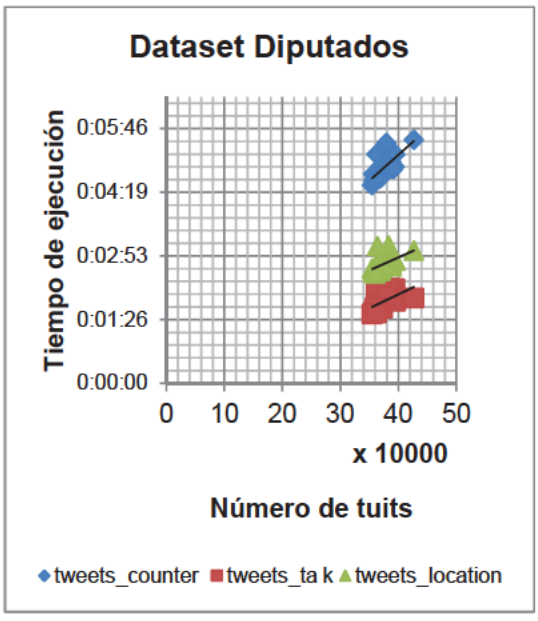

Figura 15 Tiempo de ejecución de Diputados

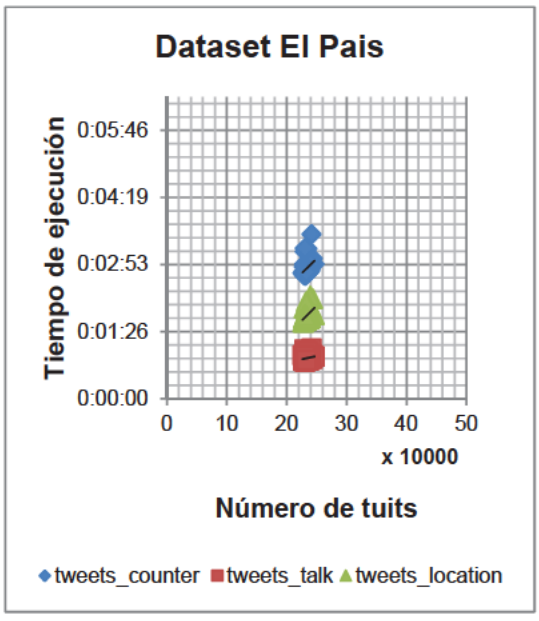

Figura 16 Tiempo de ejecución de El País

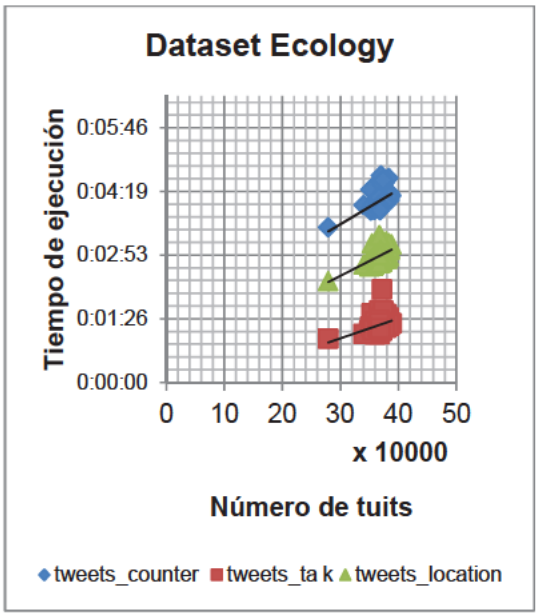

Figura 17 Tiempo de ejecución de Ecology

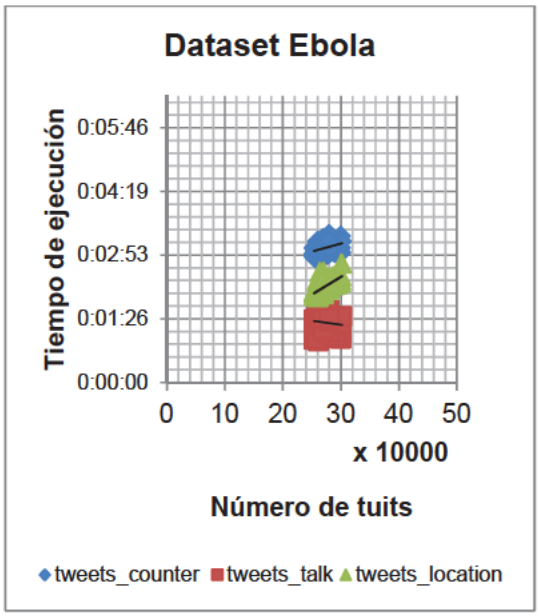

Figura 18 Tiempo de ejecución del Ébola

4.1.5.2 Proceso de integración de paquetes

El tiempo de ejecución de la integración de los paquetes depende de la duración del experimento. Esto es debido a que la información se procesa por días y por tanto aumenta el número de iteraciones del algoritmo. La Figura 19 muestra el tiempo utilizado para integrar los resultados de cada uno de los paquetes en un resultado final. Se puede apreciar cómo los contadores son los que precisan más tiempo y cómo las colecciones de datos de larga duración `Diputados` y `Ecology` consumen más tiempo a pesar de tener menos paquetes.

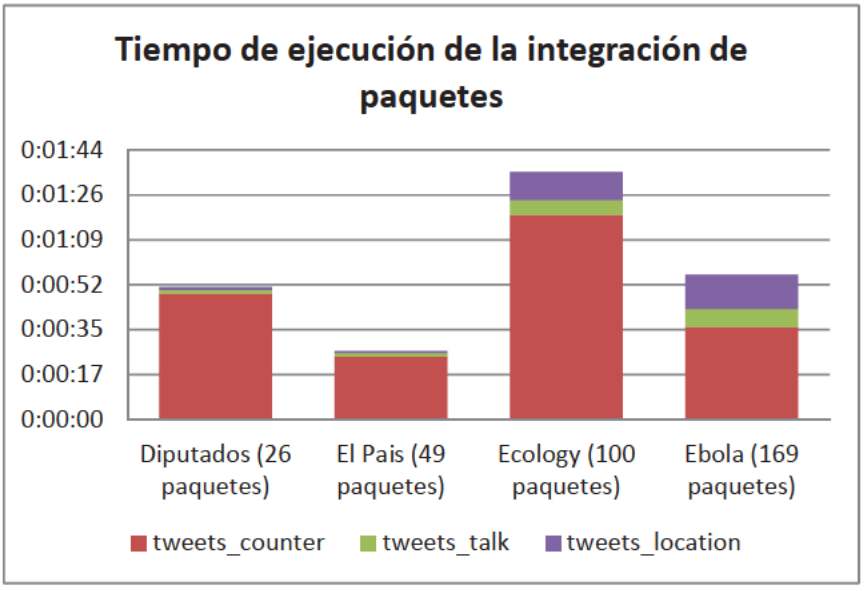

Figura 19 Tiempo de ejecución de la integración de los paquetes

Debido a que una vez procesado un paquete no es necesario volverlo a procesar, se puede medir la mejora en el tiempo de ejecución del modelo basado en paquetes sobre un modelo de un solo fichero. Para un experimento de 10 paquetes, el tiempo de ejecución de los datos totales es 1:04:43, más del ocho veces mayor que ejecutándolo por paquetes.

En la Figura 20 se puede observar la diferencia de los tiempos de ejecución de ambos modelos, en la esquina inferior izquierda el modelo de paquetes y en la superior derecha el de un solo fichero.

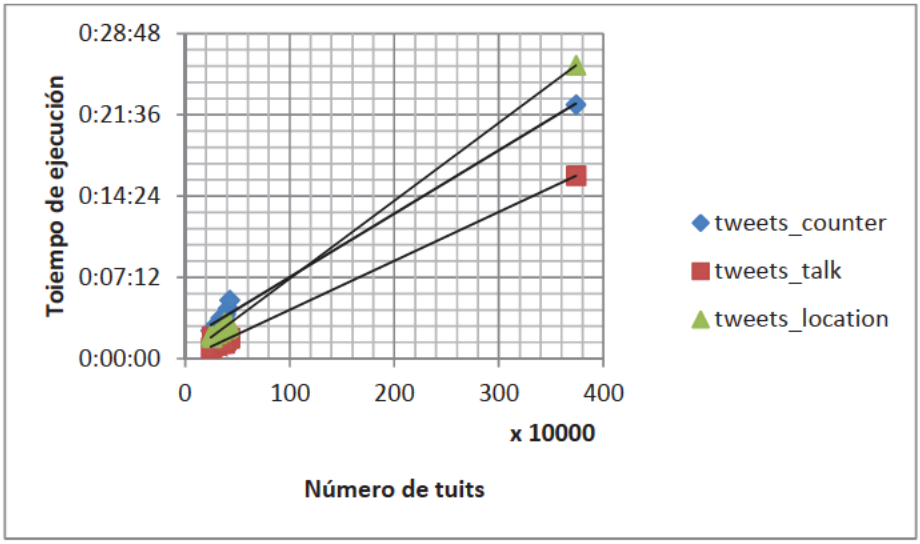

Figura 20 Correlación entre modelo de paquetes y el modelo compacto

4.1.6 REUTILIZACIÓN DE LA PLATAFORMA T-HOARDER

Para probar la capacidad de reutilización de T-hoarder, he creado el experimento Metroaverías, una prueba de concepto del ciudadano como sensor de la calidad de los servicios públicos (Congosto, Fuentes-Lorenzo, & Sánchez, 2015) que mide las quejas de los usuarios del metro de Madrid en Twitter y alerta de posibles averías. Los cambios realizados son mínimos y afectan a la capa de procesado y estéticamente a la capa de visualización.

Metroaverías incorpora un análisis semántico de los tuits para la detección de las quejas de los usuarios. El análisis se realiza por medio de un diccionario de quejas que incluye un conjunto de términos, entre ellos la jerga que se utiliza alegóricamente para protestar ("sardinas en lata", "envasado al vacío", "petado", etc.). Este diccionario se ha podido construir gracias a que T-hoarder es capaz de calcular la relevancia obteniendo los tuits más difundidos y contabilizar las palabras más frecuentemente usadas (Figura 21).

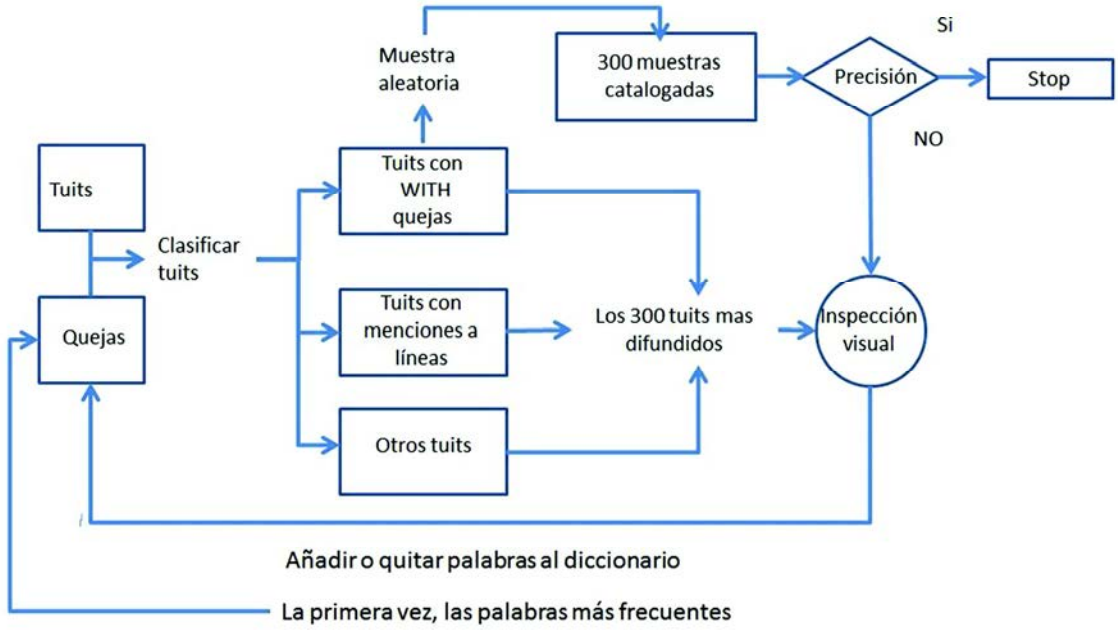

Figura 21 Metodología para generar el diccionario de quejas de Metroaverías

Los tuits se clasifican por quejas, por menciones de líneas y de estaciones. De esta manera se obtienen los problemas más frecuentes asociándolos al lugar dónde ocurren. Como las estaciones de metro están geolocalizadas es posible situar en un mapa las quejas que existen en cada una de ellas.

También incluye un sistema de alertas basadas en el algoritmo *Leaky bucket*[37]. Cada minuto examina si ha habido quejas de averías. En caso afirmativo, se incrementa el contador de averías con el número de quejas detectadas; si no las hubiera, se decrementa el contador si su valor es mayor que cero. Cuando el contador alcanza un umbral de diez se emite una alerta de avería. La detección de averías se ha contrastado con la información que proporciona @metro_madrid, la cuenta oficial del Metro Madrid, y se ha obtenido una precisión superior al 90%.

La capa de visualización solo ha cambiado estéticamente. Metroaverías dispone de nuevas plantillas para presentar las estadísticas de las quejas y utiliza el servicio de mapas de CartoDB en vez del de Google Maps.

37

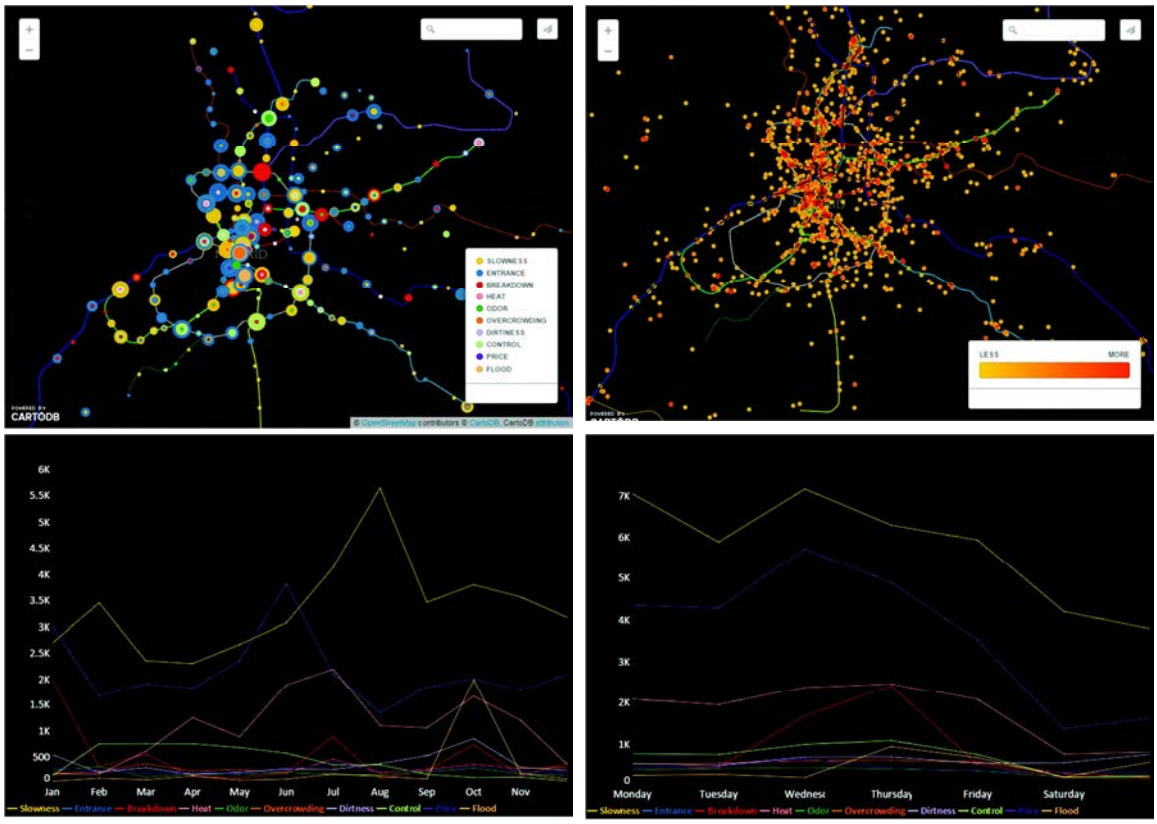

Figura 22 Mapa de quejas y estadísticas de Metroaverías

En la parte superior izquierda de la Figura 22 se encuentra el mapa del metro de Madrid en el que se ubican cada una de las quejas en sus estaciones mediante un círculo, cuyo tamaño es directamente proporcional al número de protestas y su color corresponde al tipo de reclamación. Cuando una estación tiene más de un tipo de queja, éstas quedan anidadas en círculos concéntricos. En la parte superior derecha se puede ver la geolocalización de los tuits que hablan del Metro de Madrid. En la parte inferior aparece cómo se distribuyen las quejas según los meses y los días de la semana.

4.2 MEDICIÓN DE ROLES Y PARÁMETROS DE PROPAGACIÓN

Gracias a la experiencia adquirida en T-hoarder con el procesado de paquetes y su posterior integración, ha sido posible calcular los parámetros de propagación de los distintos casos de estudio, algunos de ellos de más de 44,5 millones de tuits.

La obtención de los indicadores, por consiguiente, se realiza en dos fases. La primera consiste en procesar cada uno de los paquetes del caso de estudio y la segunda en integrar los resultados y generar las gráficas.

4.2.1 PROCESADO DE LOS PAQUETES

En esta fase se prepara la información para su posterior integración. Para cada paquete se calculan los siguientes resúmenes:

Resumen de intervalos (de 5 minutos y 1 hora):

- Alcance: número de seguidores de todos los usuarios que han publicado en el intervalo.
- RTs: número de retransmisiones que se han efectuado en el intervalo.
- RTs_enlaces: número de retransmisiones que incluían enlaces en ese intervalo.
- RTs etiquetas: número de retransmisiones que llevaban etiquetas en ese intervalo.
- Usuarios_unicos: número de usuarios únicos que han publicado en ese intervalo.
- Tuits_automaticos: número de tuits automáticos que se han publicado en ese intervalo.

Resumen de tuits:

- Fecha y hora: fecha y hora del tuit.
- Autor: autor del tuit.
- Ratio_red: ratio de seguidos/seguidores.
- App: aplicación con la que se publicó el tuit.
- RT: si es retransmisión el nombre del autor del tuit retransmitido.
- Reply: si es respuesta, el nombre del autor al que se responde.

4.2.2 INTEGRACIÓN DE RESULTADOS

La integración se ejecuta en cuatro pasos:

- Clasificación de usuarios: se realiza según el algoritmo descrito en apartado 3.1.1. y además se obtiene cómo se van incorporando los usuarios nuevos en cada intervalo. Esto no es posible hacerlo aisladamente en cada proceso de paquetes porque se necesita contexto global.
- Resumen de intervalos por roles: es similar al resumen de intervalos que se realizó en cada paquete, pero particularizado para cada rol. Esto permitirá ver en qué momentos han participado cada uno de ellos.
- Integración de resúmenes de intervalos: se agregan cada uno de los resúmenes de intervalos.
- Generación de gráficas: con la librería matplotlib se van generando las distintas gráficas.

PARTE IV. Resultados, conclusiones, contribuciones y trabajos futuros

5 RESULTADOS

5.1 VISIÓN MACROSCÓPICA DE LOS CASOS DE ESTUDIO

Como se indicó en el capítulo 3, los dieciocho casos de estudio han sido cuidadosamente seleccionados para medir la propagación en contextos distintos. Estos casos se diferencian en su duración, siendo algunos breves como las Tendencias en Twitter, de intervalos periódicos como las Elecciones o persistentes durante un largo periodo de tiempo como el Barómetro social. Existen también otros factores diferenciadores como la frecuencia de publicación, la intensidad de la propagación, el tipo de contenidos que se difunden y la renovación o estancamiento de los usuarios que participan. Por este motivo, antes de entrar en el detalle de los resultados, puede ser muy clarificador analizar los casos de estudio de una forma macroscópica.

La primera comparación corresponde a la proporción entre el número de usuarios que han participado y el número total de tuits que han generado. Mientras que un tema está de actualidad la incorporación de nuevos usuarios es creciente hasta alcanzar un punto de saturación en que empieza a decrecer (Congosto, 2011). Cuando este ratio es muy bajo denota que la participación está saturada o que algunos usuarios publican masivamente tuits. Cuando el valor se acerca a uno indica que hay una renovación continua de usuarios y que éstos participan de forma poco activa.

La segunda característica que se evalúa es el porcentaje de retransmisiones frente al total de tuits publicados. Adicionalmente se miden los porcentajes de retransmisiones que incluyen enlaces o etiquetas. Si el porcentaje de retransmisiones es muy alto nos encontramos ante un caso de información muy redundante con un número de mensajes originales pequeño. Cuando el porcentaje de retransmisiones con enlaces es elevado generalmente es debido a alguna noticia que necesita más de 140 caracteres y requiere de información adicional de un medio externo o incluye alguna imagen. En el caso de que el porcentaje de RTs con etiquetas sea muy alto, posiblemente sea debido a una campaña de comunicación organizada.

En tercer lugar se contrasta la distribución de los distintos roles en cada uno de los casos de estudio. La aparición de perfiles específicos como los Retuiteadores, Replicadores o Automáticos puede dar indicios del tipo de participación. Como se ha comentado anteriormente, la retransmisión actúa como un voto positivo al mensaje (Conover et al., 2011); por este motivo, una alta presencia de Retuiteadores puede revelar que en algunos casos sea una difusión organizada o un público muy receptivo a tema tratado. Las réplicas a los tuits, es decir, cuando un mensaje comienza por @usuario, suelen corresponder a tres supuestos: a una conversación entre dos usuarios, a llamar la atención a un usuario influyente (similar a escribir en el muro en Facebook) o a manifestar una opinión negativa. Por tanto, un porcentaje elevado de Replicadores podría corresponder a peticiones o a

opiniones discordantes. La presencia de usuarios Automáticos apunta a que la participación puede estar aumentada artificialmente.

5.1.1 PROPORCIÓN ENTRE USUARIOS Y TUITS

La Figura 23 muestra para cada uno de los casos de estudio la proporción entre usuarios que participaron y tuits totales que publicaron.

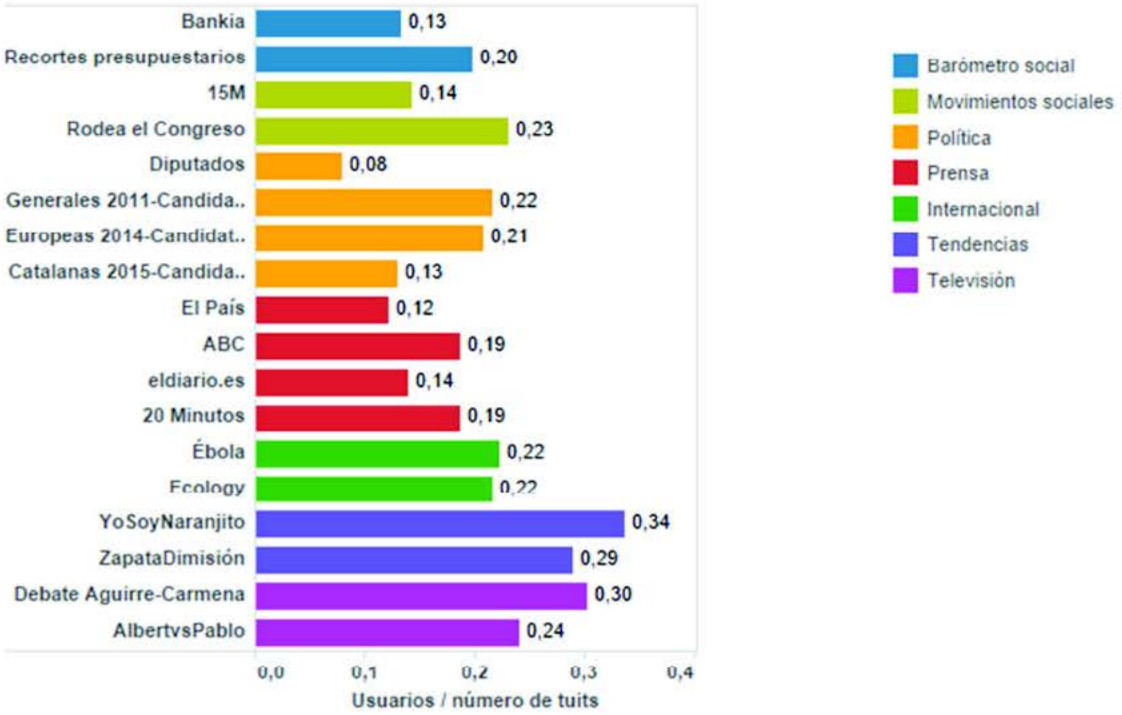

Figura 23 Proporción entre usuarios y tuits

Los valores más altos corresponden a los casos de Tendencias y Televisión, ambos de corta duración (horas o a lo sumo un día). En los casos de las campañas electorales, con una duración acotada a unos días, este ratio es superior al de los casos de periodos largos (meses o años). Parece que el tiempo es un factor determinante en la saturación del número de usuarios nuevos que se van incorporando. No obstante, en los casos internacionales el ratio es similar al de los casos de las elecciones. Esto puede ser debido a que, al ser temas globales, pueden seguir atrayendo a nuevos usuarios a pesar del transcurso del tiempo. Cabe destacar el valor del caso de los Diputados, muy por debajo del resto. Las causas se pueden deducir de la Figura 24, en la que se observa que la mayoría de los tuits proviene de menos del 15% de los usuarios.

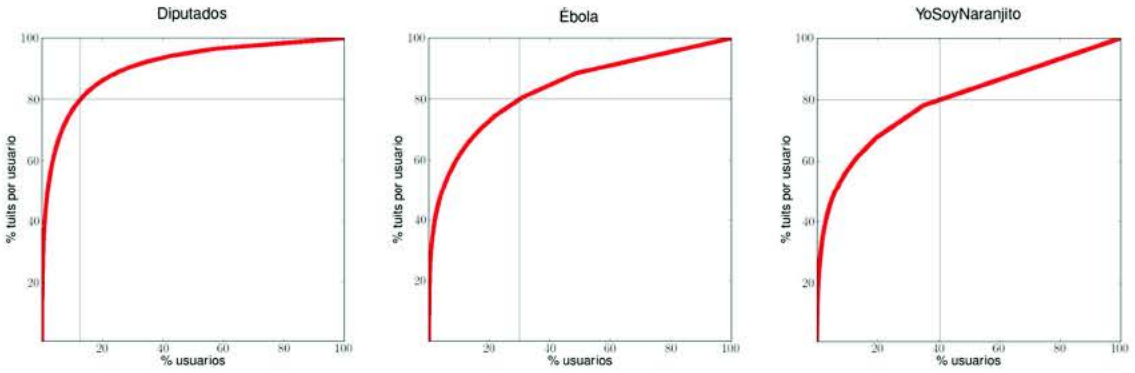

Figura 24 Distintos tipos de participación de usuarios

5.1.2 MENSAJES RETRANSMITIDOS

La Figura 25 recoge los porcentajes de propagación de cada uno de los casos de estudio. Muestra tres tipos de porcentajes: de retransmisiones, de difusión con enlaces y de propagación con etiquetas.

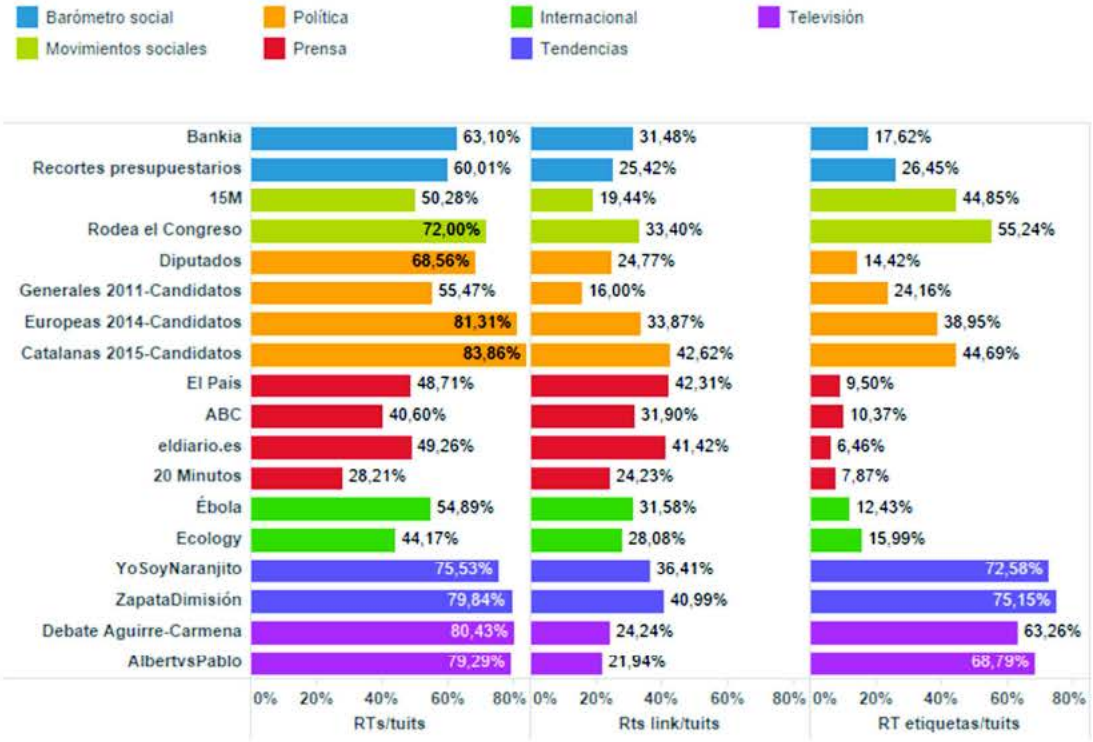

Figura 25 Porcentaje de mensajes retransmitidos

Los valores más altos de retransmisiones (75% - 83,86%) se encuentran en las dos últimas elecciones, las Tendencias y la Televisión, todos ellos de fechas recientes, de hace menos de un año. Las elecciones Generales del 2011 y el movimiento 15M tienen un porcentaje similar (50% - 55%). El movimiento Rodea al Congreso incrementa el número de RTs casi en un 22% respecto al 15M un año después. Las elecciones Europeas del

2014 y las Catalanas del 2015 aumentan más de un 25% el porcentaje de retransmisión respecto a las Generales del 2011. Entre los casos acotados temporalmente se aprecia una propensión a incrementar de la difusión de mensajes con el tiempo.

En los casos de duración más larga (años), los porcentajes de difusión son más bajos que los anteriores descritos. Los casos Bankia, Recortes presupuestarios y Diputados fluctúan entre el 60% y el 68%. El Ébola y Ecology oscilan entre el 44% y el 55%. La prensa se encuentra entre el 28% y el 49%.

Respecto a la difusión de mensajes con enlaces, los porcentajes mayores están en la prensa (24%-42%), lo cual es lógico dado que la mayoría de sus mensajes llevan enlaces a noticias. Los valores más bajos se encuentran en el 15M y las elecciones Generales del 2011 (16% -20%), posiblemente debido existencia de pocas noticias entre sus tuits. En el grupo de Televisión tiene sentido que el porcentaje de RTs sea bajo porque se comenta generalmente lo que se está viendo en directo de una forma más textual.

En cuanto a la propagación con etiquetas, los casos de Tendencias y Televisión son los que alcanzan mayores porcentajes porque casi todos los mensajes están etiquetados (63% -76%). En los casos de Movilización social también se utilizó el recurso del etiquetado para difundir mensajes (44% - 56%), lo que denota organización en la retransmisión. En menor medida se utilizó este recurso en las elecciones (25% - 45%). En los casos de la Prensa e Internacional el etiquetado fue testimonial (6% - 16%).

La publicación de tuits sigue una distribución de la ley de potencias, es decir, un pequeño grupo de usuarios tiene mucha actividad mientras que el resto es muy poco activo (Sysomos, 2009) (Bild, Liu, Dick, Mao, & Wallach, 2015). Lo mismo ocurre con el número de retransmisiones; la mayoría de los tuits difundidos provienen de un reducido número de usuarios mientras que los difusores también son minoría.

Para ilustrar cómo se distribuyen los RTs recibidos y los RTs realizados en la Figura 26 se muestra el caso del 15M, donde no llega al 4% el número de usuarios que acapararon el 80% de los RTs y en el que menos del 20% de los perfiles generaron el 80% de las retransmisiones.

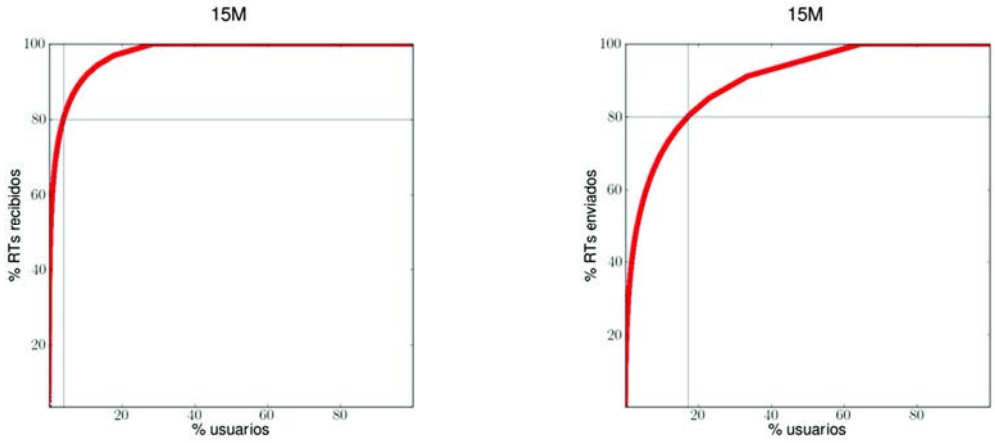

Figura 26 Distribución de los RTs recibidos y enviados en el 15M

Teniendo en cuenta que en algunos casos la retransmisión de mensajes puede superar el 80% de los tuits publicados nos encontramos ante un caso de burbuja de actividad. Un tema puede parecer que tiene mucha resonancia y una gran participación pero en realidad está estimulado por una minoría.

En la Figura 27 se muestra para cada caso de estudio cuántos usuarios obtuvieron el 80% de los RTs y cuántos realizaron el 80% de las retransmisiones.

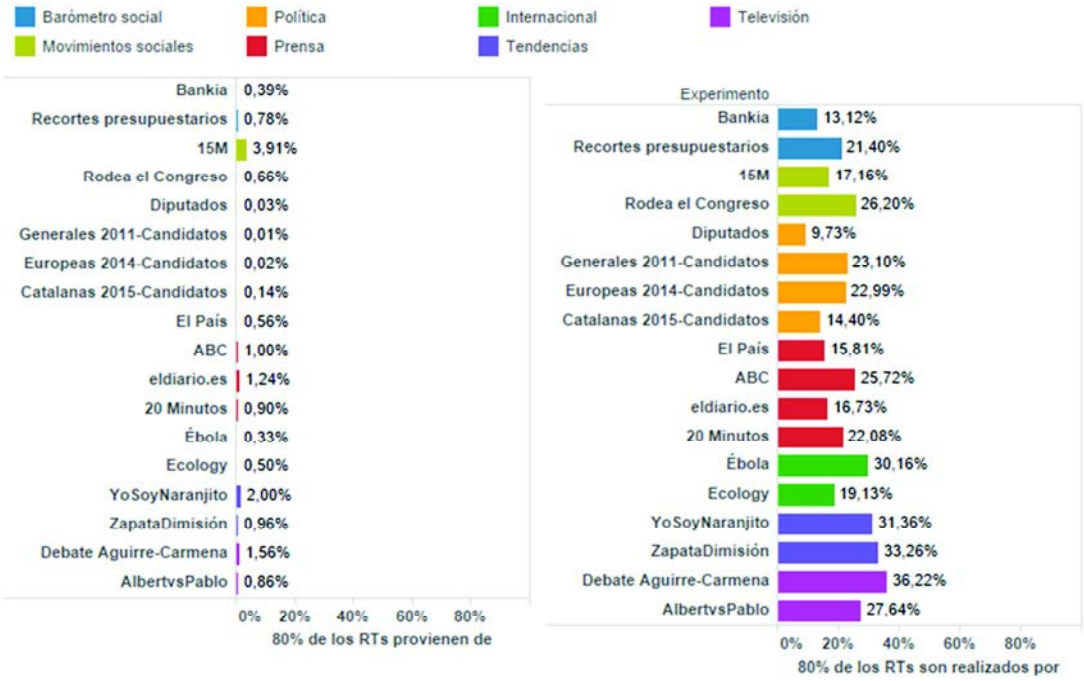

Figura 27 De dónde vienen las retransmisiones

En la parte de la izquierda de la figura se pueden ver los porcentajes de usuarios cuyos tuits han acaparado el 80% de las retransmisiones. Los valores más bajos corresponden al grupo de Política. Los casos de este grupo son especiales porque sus mensajes recogen la interacción de los políticos con los usuarios y por tanto los usuarios más difundidos son estos. El valor más alto y muy destacado del resto es el del 15M (3,91%), oscilando los demás casos entre el 0,33% y el 2%. Este porcentaje tan bajo de usuarios que han conseguido el 80% de los RTs indica un liderazgo muy centralizado, salvo en el caso del 15M que parece un poco más distribuido.

En la parte derecha de la imagen aparecen los porcentajes de los usuarios que han generado el 80% de las retransmisiones. Por debajo del 10% se encuentra el caso Diputados. Entre el 10% y el 20% se sitúan las elecciones catalanas del 2015, Bankia, el periódico El País, el diario.es, el 15M y Ecology. Entre el 20% y el 30% se ubican la mayoría de los casos. Entre el 30% y el 40% aparecen el Ébola, los casos de Tendencias y el Debate Aguirre-Carmena. Salvo algunas excepciones parece que en los casos de larga duración o periódicos tienen porcentajes más bajos que los más breves.

Con estos datos se constata que la difusión de mensajes en los temas sociales en Twitter está estimulada por menos del 40% de los participantes y que la mayoría de los contenidos son tuits duplicados provenientes como máximo del 4% de los usuarios.

5.1.3 DISTRIBUCIÓN DE LOS ROLES

La Figura 28 recoge la distribución los distintos roles para cada uno de los casos de estudio. Para hacerla más legible solo se han incluido los valores de los porcentajes más altos. No obstante, en la Tabla 13 y en la Tabla 14 del 0están detallados todos los valores.

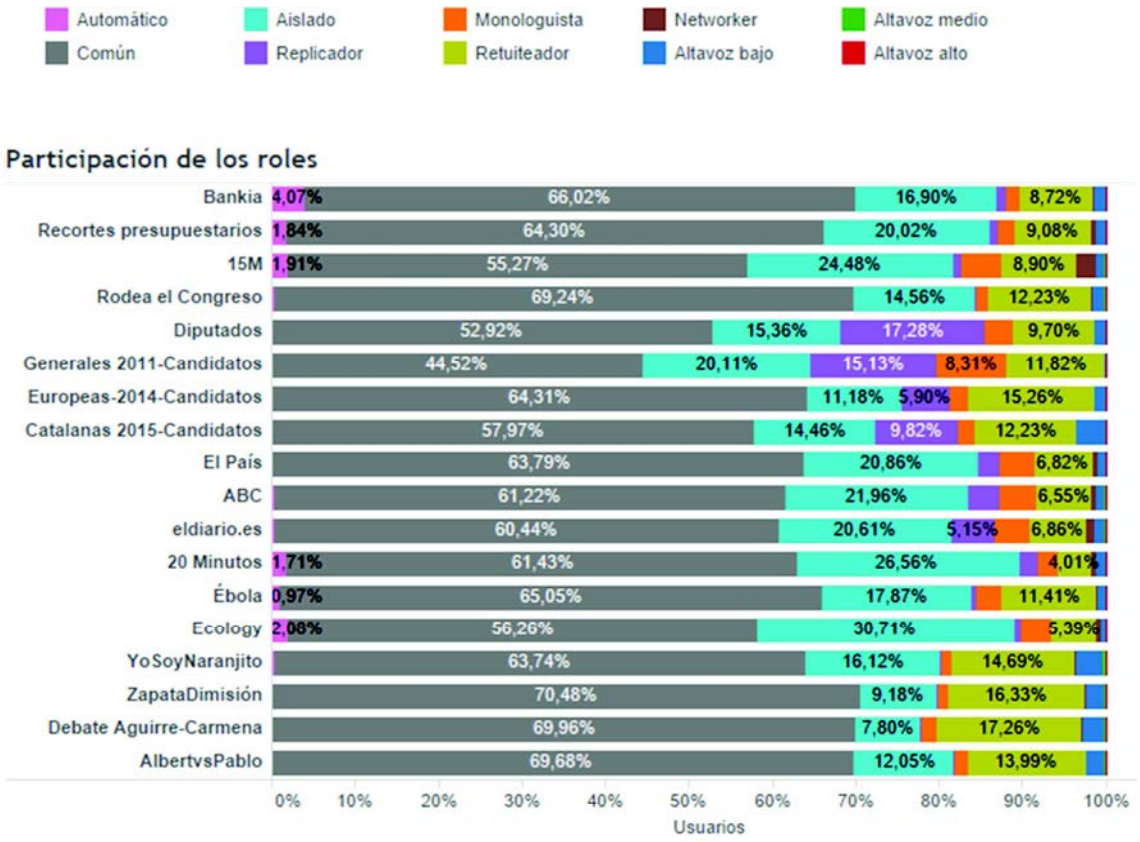

Figura 28 Distribución de los roles

Esta medida solo se refiere a los usuarios de cada tipo, no a la actividad que hayan desarrollado. Por ejemplo, podría haber un pequeño porcentaje de Retuiteadores muy activos que consiguieran generar más retransmisiones que un grupo mayor de éstos pero con menos actividad. Por tanto, los valores de esta gráfica orientan tan solo en la distribución de los roles en cada uno de los casos.

Los Automáticos, que publican tuits por medio de aplicaciones que agregan contenidos, oscilan entre 0 y 4,07%, destacando Bankia sobre los demás casos, seguido de 20 Minutos y los Recortes presupuestarios. Más adelante, cuando se analice el indicador Automatización, se podrá contrastar la actividad de estos usuarios.

Los usuarios Comunes es el grupo más numeroso en todos los casos. Se incluyen en este este conjunto los que no tienen un perfil específico de comportamiento. Su porcentaje oscila entre el 44,52% (Generales 2011) y el 70,48%.(ZapataDimisión).

Los Aislados, es decir, aquellos que no han retransmitido mensajes ni han sido difundidos, se encuentran en el rango de 7,80% al 30,71%. Los casos más recientes (Tendencias y Televisión y últimas elecciones) tienen porcentajes más bajos.

Los Replicadores, cuyos tuits son mayoritariamente respuestas, tienen valores más altos en el grupo de Política (5,90% - 17,28%). Una inspección visual de estos tuits ha encontrado a usuarios que intentan llamar la atención de los políticos algunas veces de forma positiva y otras increpándoles. Esta tesis no aborda el análisis de sentimiento por lo que no es posible proporcionar una medida de los mensajes favorables o contrarios hacia los políticos.

Los Monologuistas, que publican tuits que tienen poca o nula difusión, alcanzan el valor máximo en el Generales 2011-Candidatos con un 8,31%. En el resto de los casos tienen valores más bajos, sobre todo los casos más recientes.

Los Retuiteadores, el grupo que más practica la difusión, oscila entre el 4,01% y el 17,26%. Los valores más altos se encuentran en los casos más actuales.

Los Altavoces, cuyos mensajes son más difundidos, son una minoría pero ha aumentado su porcentaje en los casos más recientes.

Se podría concluir que desde el 2011 hasta el 2015 existe un aumento de las retransmisiones, que se traduce en un incremento del número de Altavoces y Retuiteadores, a la vez que disminuyen los Aislados y Monologuistas.

5.1.4 PARTICIPACIÓN DE LOS ROLES EN EL PROCESO DE DIFUSIÓN

La distribución de las retransmisiones generadas por cada por cada uno de los roles (Figura 29) muestra la asimetría en la propagación. El grupo minoritario de los Retuiteadores, que osciló entre el 4,01%-17,26%, generó entre el 45,57% al 63,53% de los RTs.

En esta distribución desaparecen los usuarios Aislados y Replicadores por no participar en el proceso de difusión.

Los Networkers mantuvieron una actividad alta de propagación en los casos de prensa y en aquellos más antiguos, sobre todo en el caso 15M en el que alcanzó su máxima participación.

Los Altavoces Bajos han ido ganando peso en la difusión en los eventos más recientes y en los relacionados con la política.

Da la impresión que el perfil Networker ha evolucionado en el tiempo al rol de Altavoz Bajo, participando activamente en el proceso de difusión y a la vez mejorando la efectividad de sus mensajes para ser difundidos.

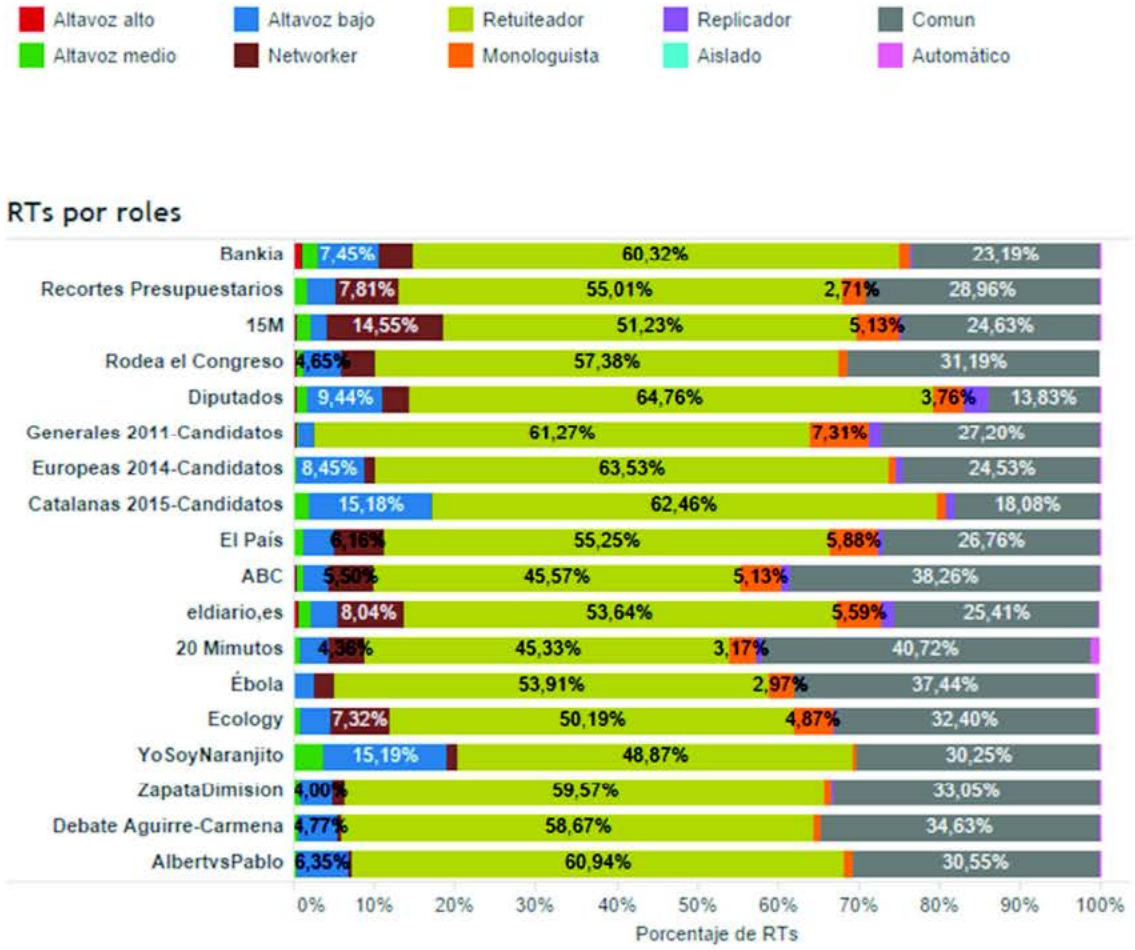

Figura 29 Porcentajes de RTs por role

5.2 BUSCANDO RELACIONES

Las distintas medidas de los indicadores van a ser comparadas unas con otras con el fin de conocer si existe alguna relación entre ellas. Para determinar el grado de relación entre los valores obtenidos de dos indicadores se utilizará el coeficiente de correlación de Pearson que se calcula como:

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \, \sigma_Y}$$

Donde σ_{XY} es la covarianza de (X,Y), σ_X es la desviación típica de X y σ_Y la desviación típica de Y. Este coeficiente mide si existe una correlación lineal entre dos variables. Esta relación puede ser positiva si los valores de las variables aumentan o disminuyen simultáneamente o negativa si una variable disminuye conforme la otra aumenta. Los valores que puede adquirir este coeficiente son:

- Si r= 1, existe una correlación positiva perfecta. Conociendo una variable se puede determinar el valor de la otra.
- Si 0 < r < 1, existe una correlación positiva. Cuanto más próximo está el valor a 1 más fuerte es la correlación.

- Si r= 0, no existe relación lineal. Pero esto no necesariamente implica que las variables sean independientes, pueden existir todavía relaciones no lineales entre ellas.
- Si -1 < r < 0, existe una correlación negativa. Cuanto más próximo está el valor a -1 más fuerte es la correlación.
- Si r= -1, existe una correlación negativa perfecta.

Adicionalmente se obtendrá el coeficiente de determinación (R^2), que determina la calidad de la relación. Se calcula como el cuadrado del coeficiente de Pearson, adquiriendo valores un rango entre 0 y 1. Cuanto más se acerca a 1 mayor es la calidad de la relación de las variables. Este coeficiente refuerza las correlaciones fuertes.

La correlación de dos variables se puede observar visualmente cuando se representan en un diagrama de dispersión. La correlación es muy alta si la disposición de la nube de puntos se ajusta a una recta, es más débil si los puntos adquieren una forma ovalada y prácticamente no existe si la forma es redonda.

Para mejor comprensión de los resultados he generado dos tipos de gráficas:

- Diagrama de dispersión – histograma: permite ver tanto la correlación de ambas variables como su distribución.
- Diagrama de dispersión – color: muestra la correlación de dos variables en la que los puntos de la nube están codificados por un código de color.

Las gráficas incluyen para los histogramas los valores de la media (μ), la mediana (*median*) y la desviación típica (σ) y para los diagramas de dispersión el coeficiente de Pearson (ρ) y el coeficiente de determinación (R^2). La Figura 30 muestra dos ejemplos de los diagramas de dispersión utilizados.

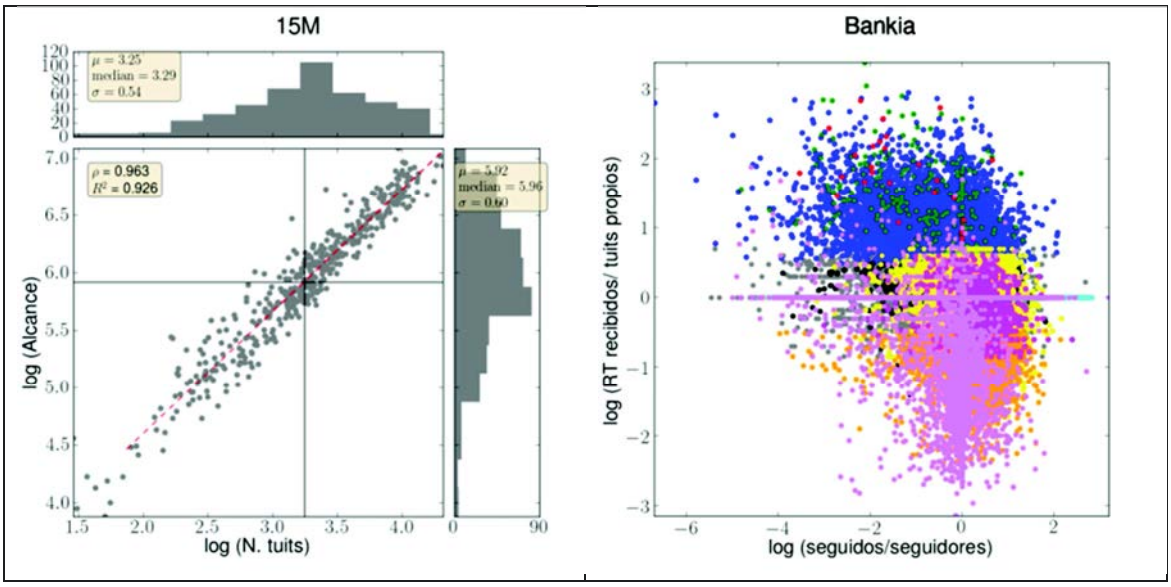

Figura 30 Ejemplos de diagramas de dispersión utilizados

El Anexo I recoge, estructurado por tipos de relaciones, las medidas de cada uno de los casos. La resolución de las imágenes es alta pero se ha adaptado su tamaño para facilitar la comparación entre los distintos entornos medidos, pudiéndo ser ampliadas sin

menoscabo de su calidad. El tamaño del Anexo I sobrepasa los límites de tamaño de las versiones digitales para las tesis por lo que aparecerá en un documento separado. No obstante se hará referencia a este Anexo desde la memoria de esta tesis.

Durante la exposición de los resultados se utilizarán algunos ejemplos significativos de estas gráficas pero para mejor comprensión de los resultados aconsejo una mirada detallada al Anexo I.

5.3 EL USUARIO COMO INDICADOR

Para analizar la propagación vista desde los perfiles de usuario, esta tesis se basa en el modelo de (González-Bailón et al., 2013), explicado en la Figura 4. Este modelo mide la influencia de los usuarios mediante dos parámetros, el ratio de red (seguidos/seguidores) y el ratio de mención (menciones recibidas/mensajes publicados). A priori se podría creer que las personas que tienen más seguidores que seguidos son las más influyentes, pero no es así, existe una influencia oculta formada por los usuarios con menos seguidores que seguidos que también consiguen ser mencionados.

Para las mediciones de los usuarios se ha modificado el ratio de mención (menciones recibidas/mensajes publicados) por el ratio de propagación, basado en el *coeficiente de eficiencia de usuario* definido por (Morales, 2014):

$$\eta_i = \frac{R_i}{A_i}$$

Siendo R_i el número de RTs recibidos por el usuario *i* y A_i el número de mensajes publicados por el usuario *i*.

El ratio de propagación realiza el siguiente ajuste del *coeficiente de eficiencia de usuario:*

$$rp_i = \frac{R\_in_i}{A_i - R\_out_i}$$

Siendo R_in_i es el número de RTs recibidos por el usuario i, A_i el número de mensajes publicados por el usuario i *y* R_out_i número de RTs enviados por el usuario i.
De esta manera, lo que se mide es la capacidad de propagación de los mensajes de un usuario de los que es autor.

Con este criterio se ha calculado para cada caso de estudio la relación entre el ratio de red y el ratio de propagación, obteniendo en todos los casos resultados similares a los de (González-Bailón et al., 2013).

5.3.1 RATIO DE PROPAGACIÓN VS. RATIO DE RED

Para determinar si existe alguna relación entre tipo de perfil de usuario y la atención prestada a sus mensajes se ha comparado el ratio de red (seguidores/seguidos) con el ratio de propagación. Esta relación se representa mediante un diagrama de dispersión - histograma. Todas las gráficas se encuentran en el Anexo I, punto 1.1.1. La Figura 31 contiene los diagramas de una selección de casos.

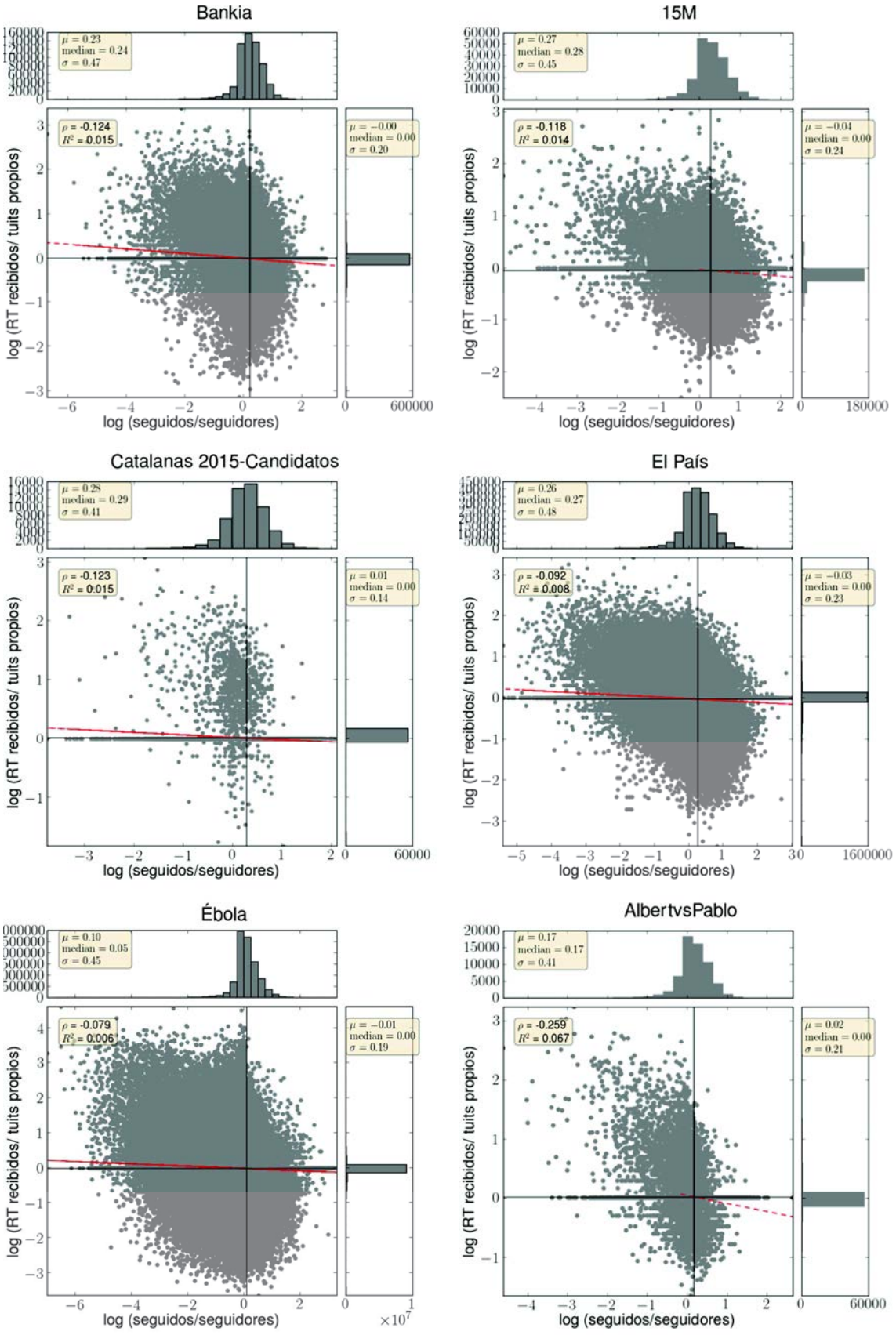

Figura 31 Ratio de propagación vs. ratio de red

En todos los casos, tengan o no muchos usuarios, sean de duración breve, periódica o larga, las gráficas revelan similitudes como baja correlación de los ratios y distribución similar del ratio de propagación. La distribución de los ratios de red no es siempre la misma pero es parecida en algunos casos entre sí.

Lo primero que llama la atención es que existe una correlación negativa muy baja entre el ratio de red y el ratio de propagación. A priori, el tipo de red no es un hándicap para propagar mensajes. Uno de los datos macroscópicos de los casos de estudio era la asimetría de las propagaciones en las que el 80% de los RTs provenían de menos del 4% usuarios. Esta medida puede inducir a pensar que la propagación está limitada a unas élites. Sin embargo, esta métrica más precisa que relaciona los ratios de red y de propagación desvela que las propagaciones provienen de todo tipo de usuarios.

La distribución del ratio de propagación es similar a todos los casos, estando concentrados la mayoría de los usuarios próximos al eje de las abscisas. Esto indica que los usuarios son generalmente poco o nada difundidos, lo que corrobora el dato macroscópico de la asimetría de las propagaciones.

En el ratio de red encontramos que para los usuarios con menos seguidores que seguidos la distribución muestra tramos escalonados, mientras que para los otros usuarios los tramos son más abruptos.

5.3.2 RATIO DE PROPAGACIÓN VS. RATIO DE RED POR ROLES

Al estar los usuarios clasificados por roles es posible representar cada uno de estos tipos en un diagrama de dispersión del ratio de propagación y el ratio de red mediante un código de colores:

- Altavoces: altos – rojo, medios – verde, bajos –azules.
- Networkers - negro.
- Retuiteadores - amarillo.
- Monologuistas - naranja.
- Replicadores - morado.
- Aislados - cian.
- Automáticos – violeta.
- Comunes - gris.

Para mayor claridad, las gráficas se han generado por capas, de forma que si dos roles tienen la misma posición en el diagrama de dispersión, la última capa se sobrepondrá a las anteriores. He evitado el recurso de la transparencia porque el resultado en la parte central del diagrama era mucho menos claro por la distorsión y por la superposición de colores. El orden de dibujo de las capas es el siguiente:

- Comunes: al ser el grupo más numeroso y por no tener características especiales se genera en la primera capa.
- Altavoces: bajos, medios y altos, en este orden, de forma que los casos menos frecuentes queden resaltados.
- Networkers, Retuiteadores, Monologuistas, Replicadores y Aislados.
- Automáticos: se han dejado para el final para destacar el ruido.

Todas las gráficas se encuentran en el Anexo I, punto 1.2.1. La Figura 32 contiene una selección de casos de estudio.

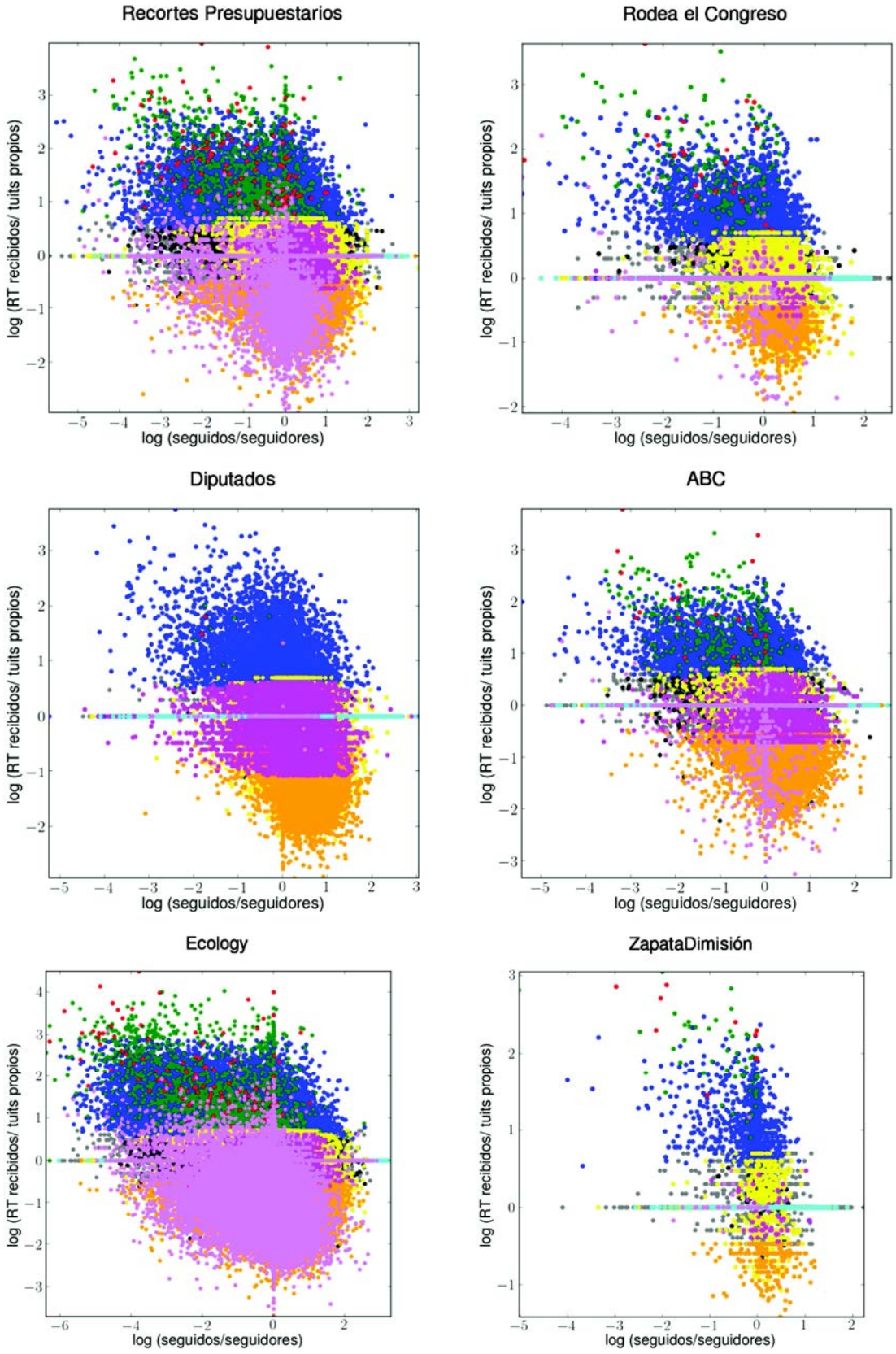

Figura 32 Ratio de propagación vs. ratio de red por roles

Los altavoces (rojos, verdes y azules) se encuentran en la parte superior del diagrama, tanto en el cuadrante izquierdo como el derecho respecto al origen de las coordenadas; es decir, hay altavoces con ratios de red superiores e inferiores a uno. Algunos Altavoces altos y medios aparecen con ratios de propagación inferiores a los de los Altavoces bajos. Esto es debido a la proporción entre RTs recibidos y tuits publicados. Por ejemplo, el perfil oficial en Twitter de un partido político podría ser un Altavoz alto porque el número de RTs que recibe está dentro del grupo del 20% más difundido, pero podría tener un ratio de difusión bajo porque los RTs provienen de bastantes mensajes. Es la diferencia entre un mensaje muy difundido o varios mensajes bastante difundidos. Comparándolo con el concepto de H-index académico (Hirsch, 2005), el valor de este índice sería mayor para el caso de los Altavoces que publican mucho y tienen una difusión constante de aquellos que con un solo tuit consiguen una gran difusión. La disposición de los Altavoces altos y medios en el diagrama proporciona información sobre la manera en que se ha conseguido la difusión.

Los Networkers (negros) y los Retuiteadores (amarillos) ocupan la parte central del diagrama. Aparece esporádicamente algún Networker con ratios muy bajos de difusión. Esto es debido a que se les clasifica por tener una actividad superior a la media y por tener equilibrado el número RTs enviados y recibidos pero en algunos casos pueden haber publicado muchos tuits sin propagación que le bajan el ratio de difusión.

Los Monologuistas (naranja) se sitúan en la parte inferior del diagrama con ratios de propagación inferiores a uno. Al igual que los altavoces aparecen tanto en el cuadrante izquierdo como el derecho respecto al origen de las coordenadas, lo que muestra que tener más seguidores que seguidos no garantiza difusión.

Los Replicadores (morados), muy presentes en los casos de Política y de Prensa, se distribuyen en la parte central, en una zona que generalmente ocupan los Retuiteadores, superponiéndose a ellos.

Los Aislados (cian) se encuentran en el eje de abscisas, tomando valores del ratio de red mayores y menores de uno.

Los Automáticos (violeta) se sitúan en la parte inferior del diagrama, en la zona de los Monologuistas, pero hay algunos casos en que se mezclan con Retuiteadores, Networkers e incluso Altavoces. Esto es debido a que hay dos tipos de Automáticos: las cuentas oficiales de medios de comunicación que publican automáticamente sus contenidos y los usuarios que sindican contenidos emitiendo tuits conforme se publican noticias. Los primeros son difundidos, los segundos no.

La clasificación empírica por roles que se realiza en esta tesis queda validada cuando se representa en estos diagramas de dispersión, ocupando posiciones coherentes con su definición.

5.3.3 ACTIVIDAD DE LOS ROLES PROPAGADORES

Los roles que más favorecen la propagación son los Retuiteadores y los Networkers, que difunden los tuits que publican los Altavoces. Analizar la participación de estos perfiles a lo largo del tiempo puede aportar un punto de vista de cómo se realizó la propagación. Por

este motivo he contabilizado la actividad de estos roles en intervalos de una hora y lo he representado gráficamente mediante un diagrama de dispersión. Para cada intervalo se sitúa en el eje X el número de tuits y en el eje Y el número de perfiles de cada tipo. Se utiliza una escala logarítmica para facilitar la representación.

De la misma manera que en el punto 5.3.2, utilizo un código de colores para cada rol y genero las gráficas por capas, de forma que si dos roles tienen la misma posición en el diagrama de dispersión, la última capa se sobrepondrá a las anteriores. El criterio de dibujo de cada rol está en función de que sea más o menos frecuente para que si se solapan quede el menos abundante resaltado. El orden de dibujo es el siguiente: Retuiteadores, Networkers, Altavoces bajos, medios y altos.

Todas las gráficas generadas se encuentran en el Anexo I, punto 1.1.3. La Figura 33 recoge cuatro ejemplos seleccionados, dos casos de larga duración (años) y dos más cortos (días).

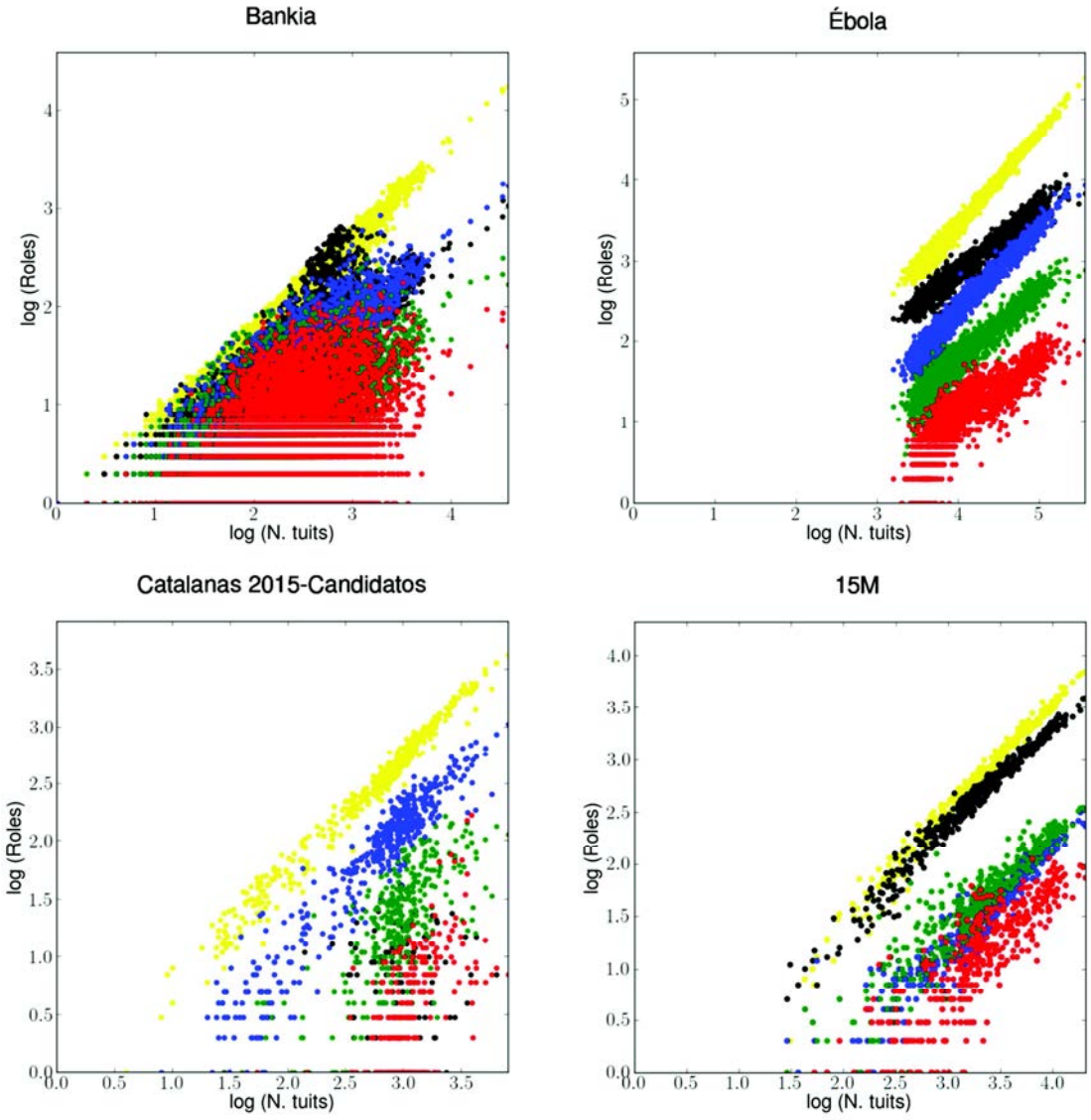

Figura 33 Ejemplos de actividad de los roles vs. número de tuits

En los casos de larga duración (años) existe más solapamiento entre roles, debido a que la forma de participar puede variar con el tiempo o las circunstancias. En los aquellos con duraciones más breves (días) queda más claro cómo fue la actividad de cada perfil. Para los que su duración fue de horas (Tendencias o Televisión) en el Anexo I, punto 1.1.4 se encuentra la misma métrica pero realizada cada cinco minutos, a fin de obtener más muestras. El resultado obtenido es similar al de los casos de duración por días.

Un patrón común en todos los casos es la actividad de los Retuiteadores (amarillo). Su distribución está próxima a la diagonal de la gráfica, mostrando una clara correlación con el número de tuits. Los Networkers (negros), los Altavoces bajos (azules), medios (verdes) y los Altavoces altos (rojos) aparecen en la parte inferior de la diagonal y tienen una presencia más dispersa.

En algunos casos la actividad de los Networkers se solapa con la de los Retuiteadores (El País, ABC, eldiario.es y 20 Minutos), otras con los Altavoces (Bankia y Diputados), y en otros casos aparecen claramente diferenciados del resto de los grupos y alineados respecto al número de tuits (Rodea el congreso, 15M, Ébola y Ecology).

Los Altavoces suelen aparecer superpuestos en casi todos los casos de larga duración salvo en los casos Internacionales (Ébola y Ecology), que aparecen claramente diferenciados. Los casos Internacionales tienen casi un orden más de magnitud superior de usuarios que los casos de larga duración.

La actividad de los roles en los casos del Barómetro social, Movimientos sociales, Política, Prensa e Internacional se parecen entre sí, con pequeñas variaciones en el solapamiento de perfiles. Da la impresión que en cada entorno los roles participan de forma similar.

5.4 INDICADORES DE PROPAGACIÓN EN EL CORTO PLAZO

Los indicadores de propagación se han medido en periodos de tiempo de una hora y de cinco minutos, como quedó explicado en el punto 3.1.3. Para cada intervalo temporal se calculó el valor de cada indicador y del número de tuits publicados en ese momento de forma que se pueda conocer su correlación. En algunos casos de corta duración, como pueden ser los debates televisados o las tendencias de Twitter, en el número de muestras puede ser muy pequeño. Aun así, he mantenido la comparación con la métrica por hora porque la de cinco minutos proporciona valores diferentes tanto en los casos de larga como de corta duración. Con todo, mostraré en los casos que se requiera la medida de los cinco minutos.

Las correlaciones se representan mediante un diagrama de dispersión – histograma explicado en el punto 5.2.

5.4.1 ALCANCE

Este indicador mide la correlación entre el número potencial de usuarios que pueden leer los tuits y la cantidad de tuits que se generan en un intervalo temporal. Adicionalmente se evalúa cómo se traduce en retransmisiones. Todas las gráficas generadas se encuentran en el Anexo I, puntos 1.2.1 y 1.2.2.

La Figura 34 muestra para cada uno de los casos la correlación entre el Alcance y el número de tuits publicados y la correlación con el número de RTs.

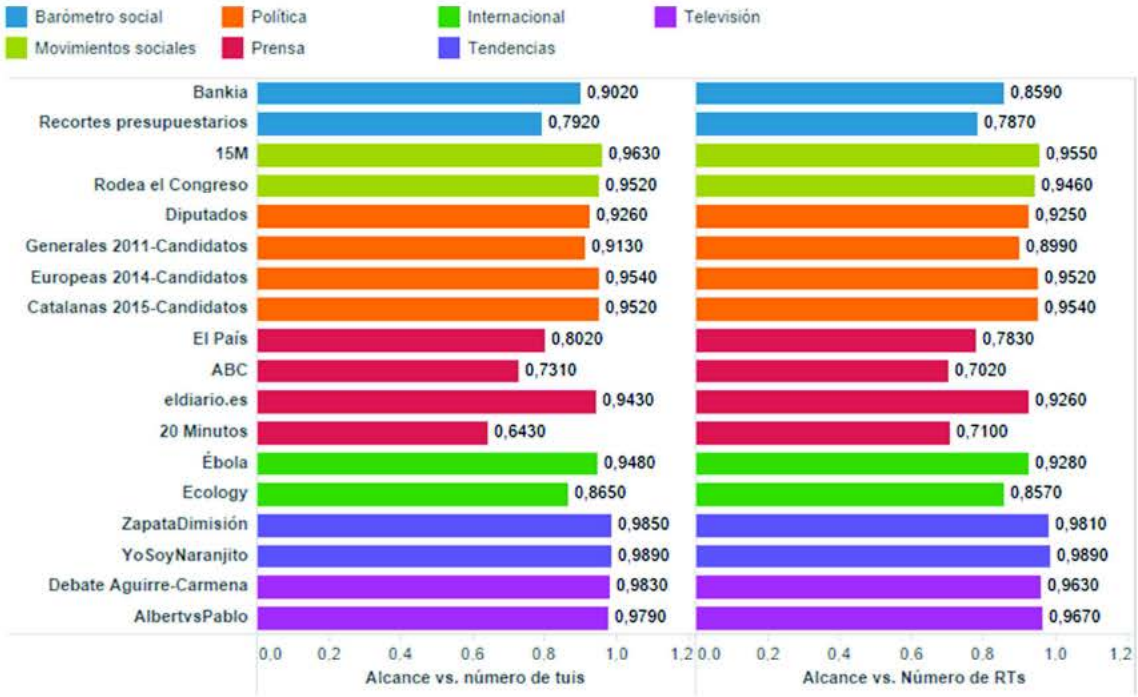

Figura 34 Alcance vs. número de tuits y número de RTs

En casi todos los casos existe una alta correlación entre el Alcance y el número de tuits. Es decir, cuando en un intervalo de tiempo la suma de los seguidores de los usuarios que han publicado aumenta, el número de tuis publicados también lo hace. Lo mismo ocurre con las retransmisiones; a mayor Alcance, más retransmisiones.

Los casos en los que estas correlaciones son mayores son los de duración corta (Tendencias y Televisión), aunque hay que matizar que el número de muestras es bajo. En los casos de las Elecciones, los Movimientos sociales y los Diputados también tienen correlaciones muy altas. El punto común entre ellos es que están estrechamente relacionados con la política. En los casos de larga duración, Barómetro social e Internacional las correlaciones siguen siendo bastante altas. En la Prensa estos valores bajan considerablemente en alguno de los casos, sobre todo en 20 Minutos, aunque se mantiene elevada correlación en eldiario.es, cuya tendencia política está más marcada.

La Figura 35 recoge una selección de cuatro casos, dos con alta y dos de más baja correlación.

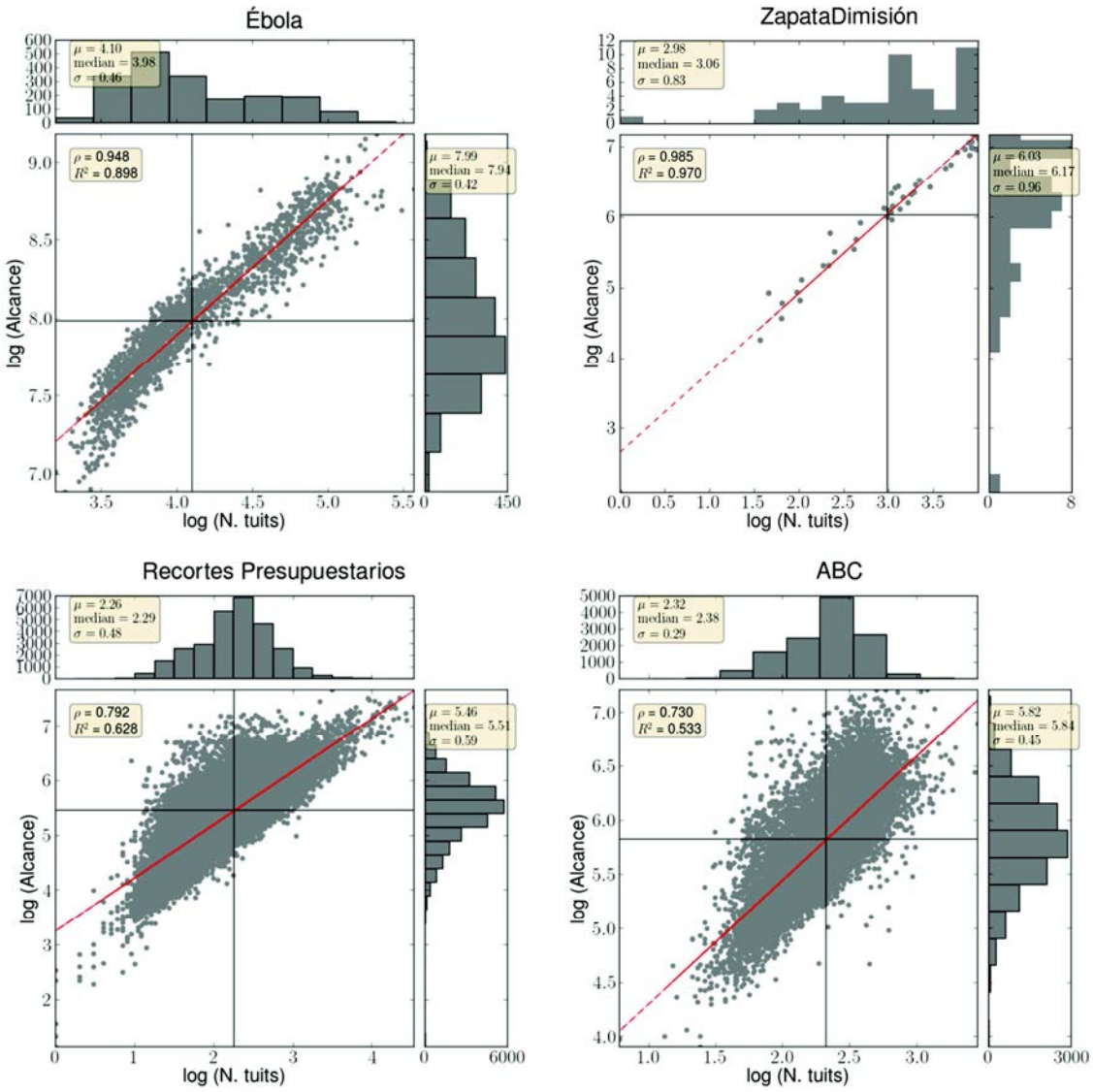

Figura 35 Ejemplos de correlaciones de Alcance vs. número de tuits

En Recortes Presupuestarios y ABC, con correlación más baja, existen momentos en el que un alto Alcance no logra generar un número proporcional de tuits.

En casi todos los casos la correlación entre el Alcance y el número de tuits es superior a la que existe con el número de RTs. La Figura 36 contiene algunos ejemplos de las correlaciones del Alcance respecto al número de tuits y al número de RTs de dos casos, uno con menor y otro con mayor correlación.

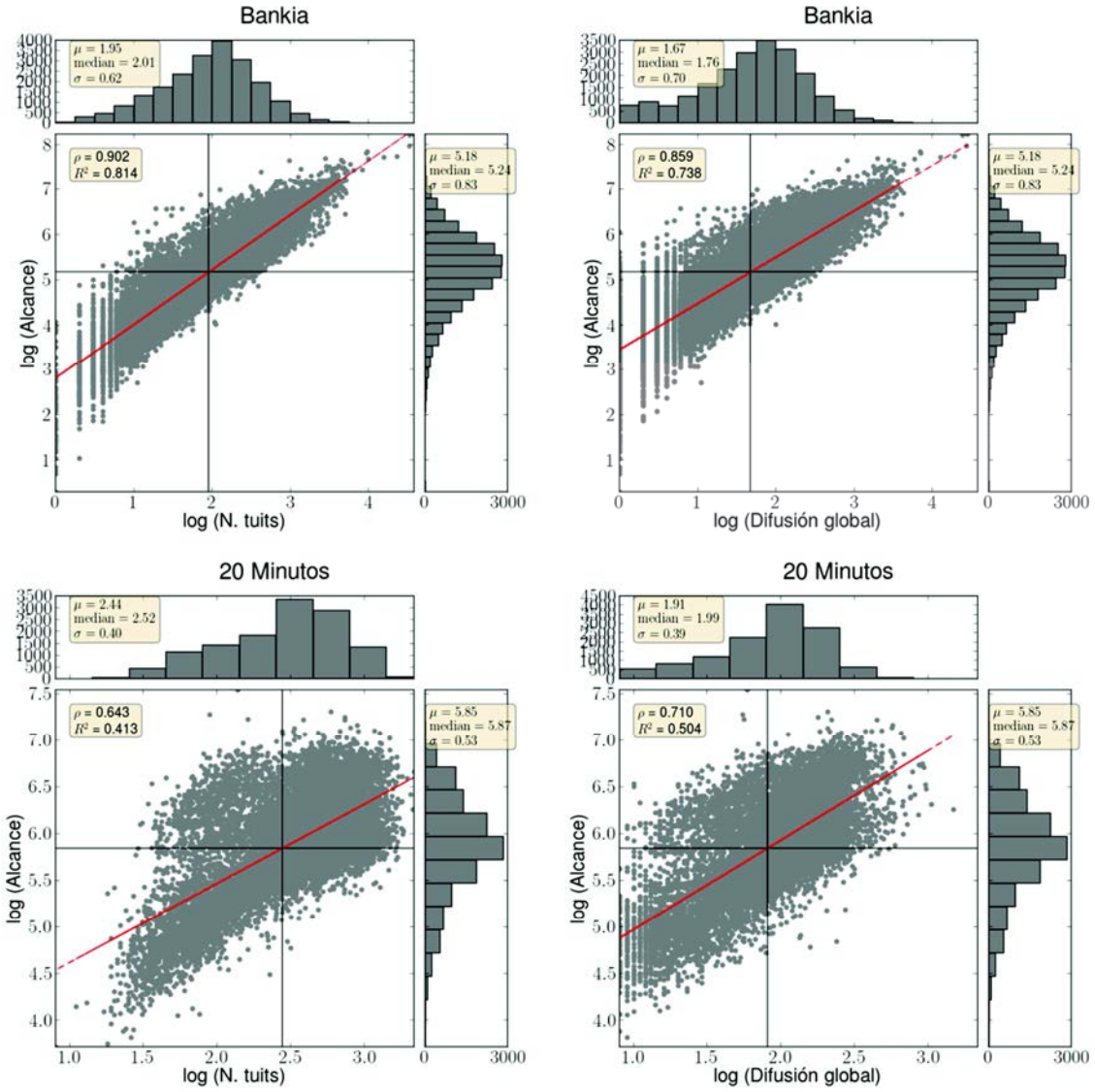

Figura 36 Ejemplos de correlaciones de Alcance vs. Propagación

El caso 20 minutos tiene un comportamiento muy diferente al resto de los casos respecto al indicador de Alcance. Más adelante se analizará si está relacionado con la participación sistemática de usuarios automáticos.

5.4.2 DIFUSIÓN

Este indicador mide la correlación entre el número de retransmisiones y la cantidad de tuits que se generan en un intervalo temporal. Todas las gráficas generadas se encuentran en el Anexo I, puntos 1.2.3, 1.2.4 y 1.2.5.

La Figura 37 muestra, para cada uno de los casos de estudio, las correlaciones entre la Difusión global, Difusión con enlace y Difusión con etiquetas respecto al número de tuits. La Difusión global corresponde al número de RTs totales, la Difusión por enlace al número de RTs con enlaces y la Difusión con etiquetas al número de RTs que incluyen etiquetas.

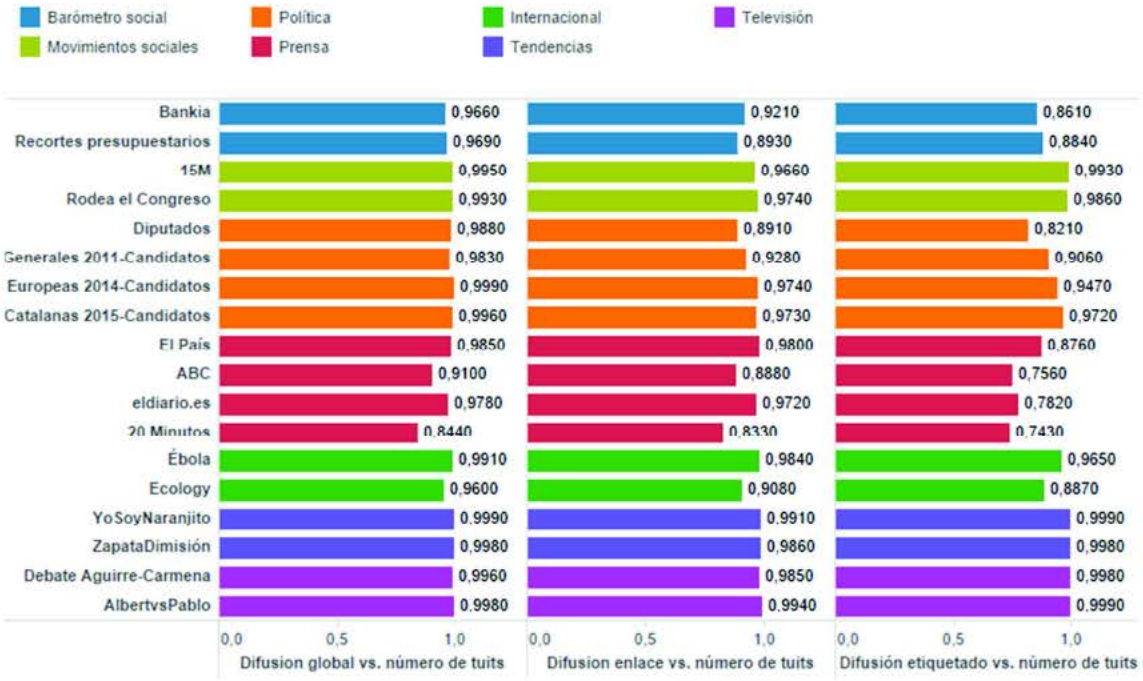

Figura 37 Difusión vs. número de tuits

En todos los casos se observa que la relación entre número de RTs y número de tuits en intervalos de una hora es proporcional. En menor medida ocurre en los casos en los que los mensajes retransmitidos contienen enlaces o etiquetas. Como se vio en el análisis macroscópico de los mensajes retransmitidos (punto 5.1.2) el porcentaje de RTs frente a número total de mensajes se encontraba en el rango entre el 83,86% (Elecciones catalanas 2015) y el 28,21% (20 Minutos). A pesar de esta diferencia tan grande, está muy ajustada la relación entre mensajes difundidos y mensajes publicados, incluso en el caso de 20 minutos. Parece que existe un coeficiente de propagación para cada caso.

En la Figura 38 se puede ver el caso de mayor correlación (Europeas-2014-Candidatos), el de Bankia y el del Ébola con elevados valores y el de 20 minutos que, teniendo una correlación alta, presenta una mayor dispersión.

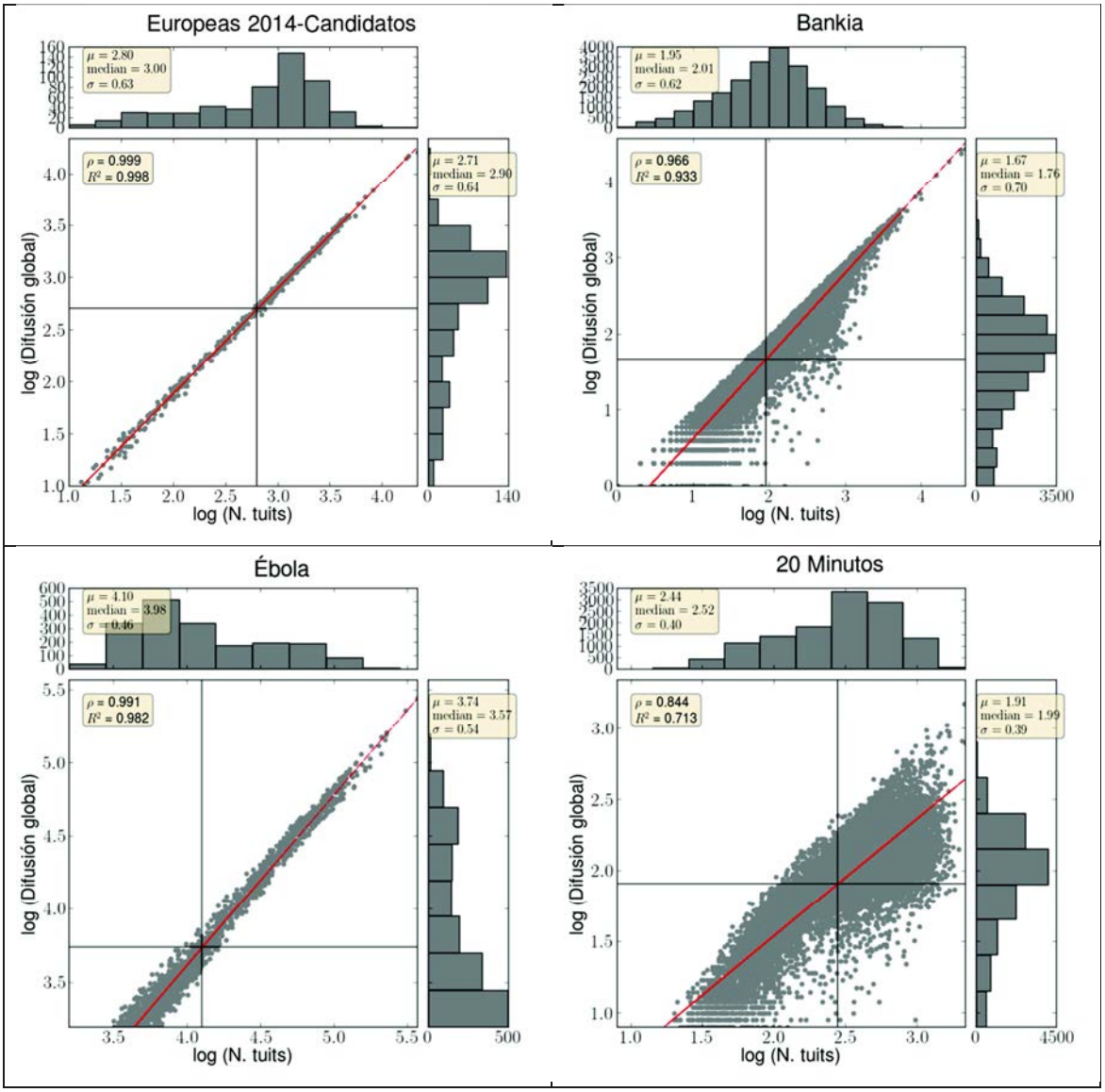

Figura 38 Ejemplos de correlaciones de difusión vs. número de tuits

5.4.3 PARTICIPACIÓN

Este indicador mide la correlación entre el número de usuarios y la cantidad de tuits que publican en un intervalo temporal. Todas las gráficas generadas se encuentran en el Anexo I, punto 1.2.6.

La Figura 39 muestra, para cada uno de los casos de estudio los resultados obtenidos.
Es el indicador con correlaciones más altas de todos los analizados, por encima del 0,98%.

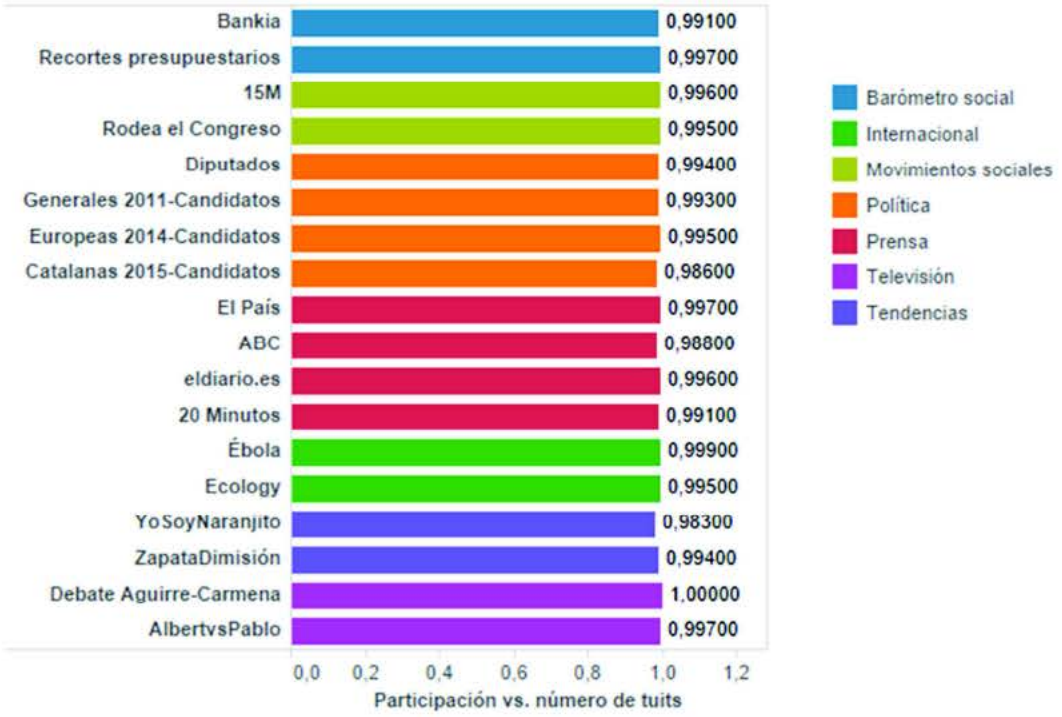

Figura 39 Participación vs. número de tuits

Como se vio en el análisis macroscópico, la proporción de usuarios y tuits (punto 5.1.1) oscilaba entre 0,08 (Diputados) y 0,34 (YoSoyNaranjito). Esta diferencia era debida a la frecuencia de publicación de los usuarios más activos. Los valores de correlación tan altos en este indicador desvelan que la distribución de usuarios en el tiempo es proporcional al volumen de tuits.

La Figura 40 contiene cuatro ejemplos de correlación. El caso del Debate Aguirre-Carmena tiene una correlación de uno, aunque con pocas muestras. El 15M tiene una dispersión muy baja. En los casos 20 Minutos y Bankia existen momentos en los que esta proporción no se mantiene y hay menos usuarios por intervalo.

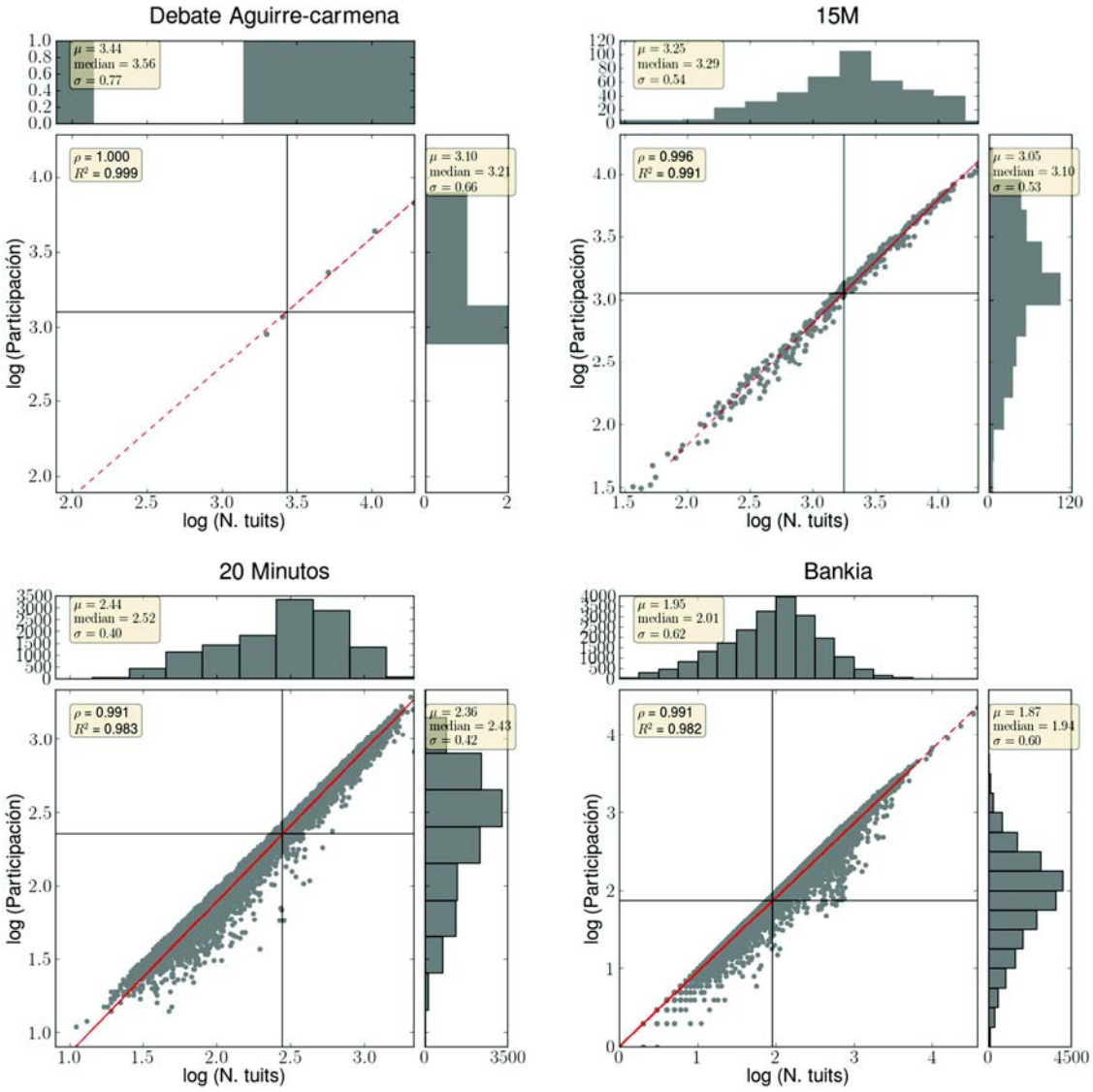

Figura 40 Ejemplos de correlaciones de Participación vs. número de tuits

Si en cada uno de los intervalos de tiempo el número de tuits fuera igual al número de los usuarios que los han publicado, los valores se situarían en la diagonal de la gráfica (las gráficas se han forzado para que el eje de las abscisas y el eje de las coordenadas sean iguales). En algunos de los casos se aproximan mucho a esta situación, arrancando la línea de regresión desde el punto 0,0 como es el caso de Bankia, Recortes presupuestarios, Rodea al Congreso y Ecology, aunque la pendiente de la línea de regresión es inferior a los 45º y no se ajusta completamente a la diagonal. En los demás casos esto no ocurre, porque siempre existen usuarios que publican más de un mensaje en la misma hora.

La participación en los eventos de Televisión es mucho más activa, es decir, el número de usuarios que repiten publicación en un intervalo de tiempo es superior al resto de los casos. Esto se aprecia en que la línea de regresión se encuentra mucho más abajo de la diagonal, incluso en los casos de las Tendencias, que son también casos de corta

duración. Esto puede denotar que la atención que se presta a los eventos televisados es superior a la que se da en las Tendencias en Twitter, siendo ambos casos de mucha visibilidad.

1.1.1 INCORPORACIÓN

Este indicador mide el número de usuarios nuevos que se han incorporado en un intervalo de tiempo respecto al número de tuits de dicho momento. Todas las gráficas generadas se encuentran en el Anexo I, punto 1.2.7.

La Figura 41 muestra para cada uno de los casos de estudio los resultados obtenidos.

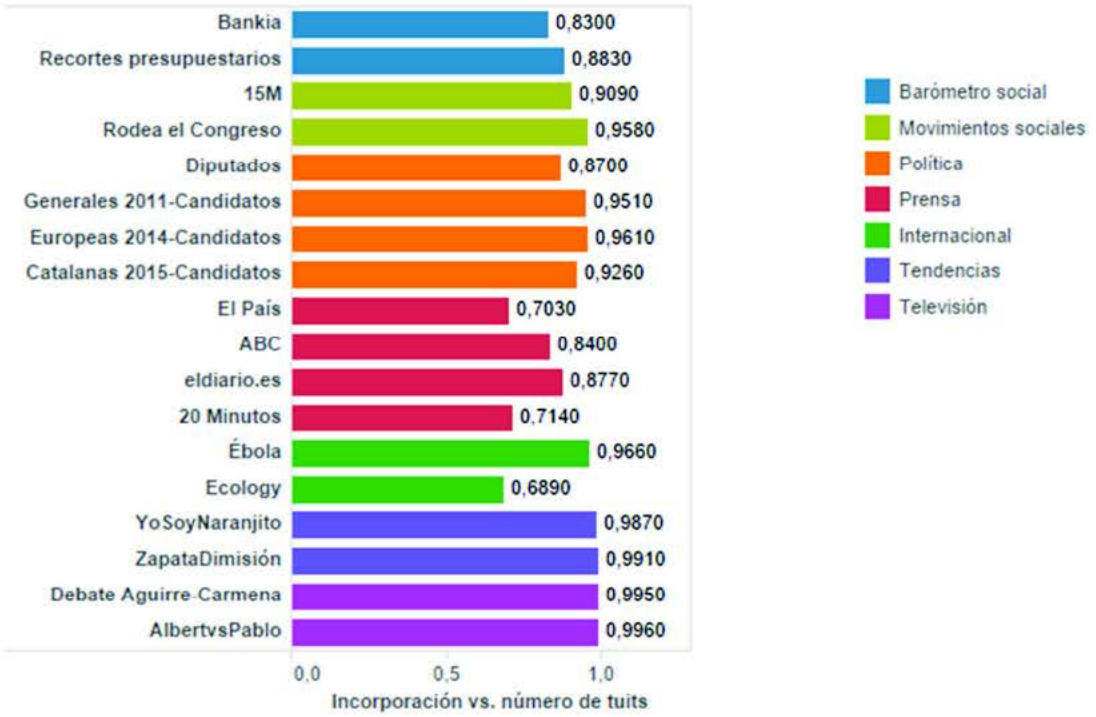

Figura 41 Incorporación vs. número de tuits

Los casos de estudio de larga duración (años), como barómetro Social, Diputados, Prensa y Ecology, tienen una correlación más baja, situándose entre el 0,689 y 0,833. Los otros casos más breves, de días u horas, como las Elecciones, los casos de las Tendencia o de la Televisión oscilaron entre el 0,951 y 0,996. Todo parece indicar que transcurrido un tiempo se alcanza el máximo radio interés por el tema y decrece el número usuarios nuevos que se incorporan, incluso si hay novedades. En los casos de duración breve esto no ocurre y la participación de nuevos usuarios es proporcional al número de tuits del momento.

El caso del Ébola es una excepción, aunque hay que tener en cuenta que la captura de estos tuits se realizó durante los tres meses en los que fue máxima actualidad.

La Figura 42 contiene una selección de casos de coeficientes altos y dos más bajos.

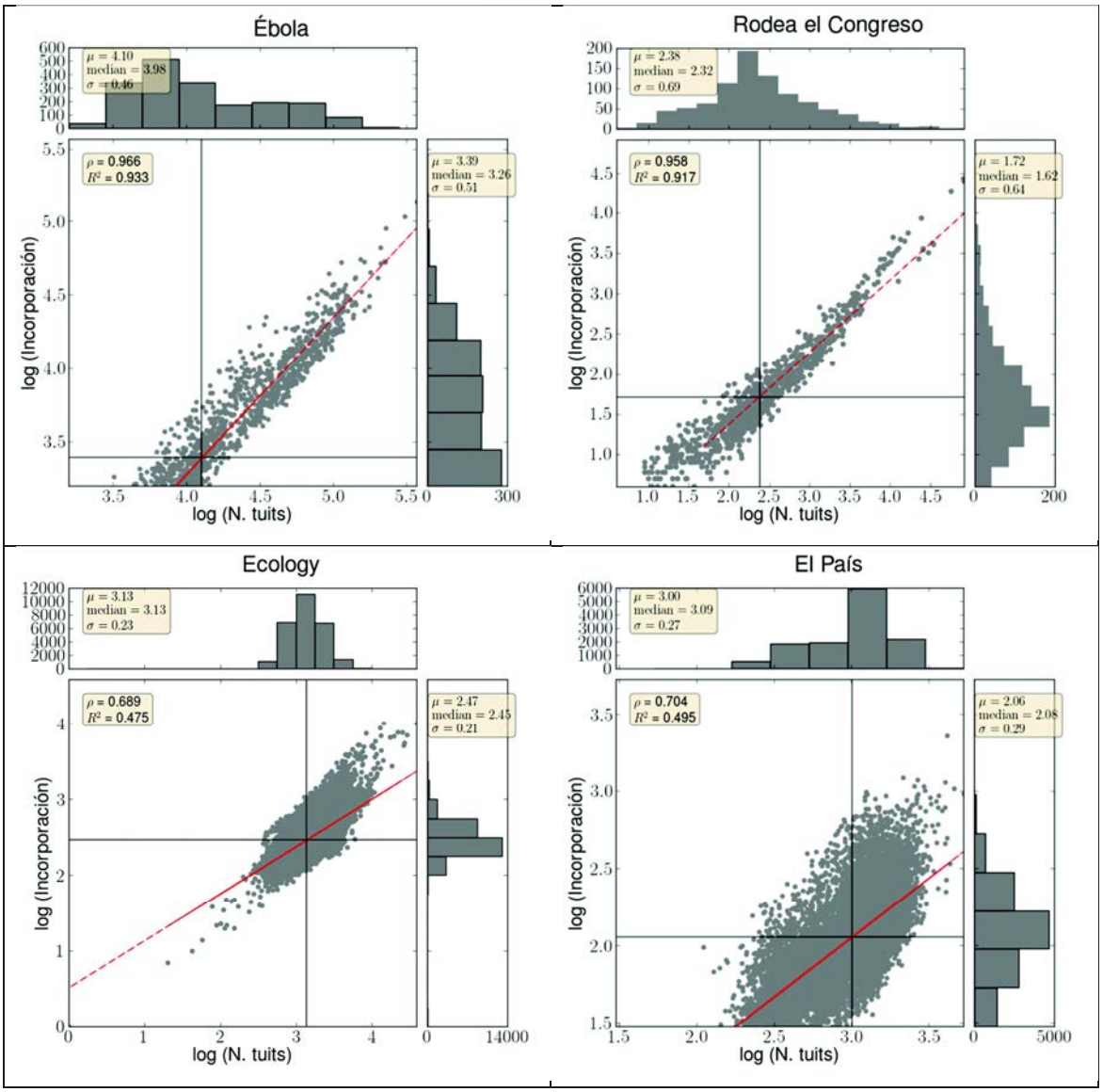

Figura 42 Ejemplos de correlaciones de Incorporación vs. número de tuits

El caso de Ébola tiene dos singularidades. La primera por la alta correlación del indicador de Incorporación siendo de duración media y por la distribución de cómo se realizan las incorporaciones. En casi todos los casos la incorporación de nuevos usuarios presenta una distribución casi gausiana (es ligeramente asimétrica), pero en el caso del Ébola la distribución es diferente; se concentran las nuevas incorporaciones en los intervalos con menos número de tuits.

5.4.4 AUTOMATISMO

Este indicador mide la correlación entre los tuits generados automáticamente en un intervalo respecto al volumen de tuits de ese momento. Como recoge el punto 3.1.2.4, los tuits automáticos están publicados mediante aplicaciones que sindican contenidos como

Twitterfeed o Dlvr.it. Todas las gráficas generadas se encuentran en el Anexo I, punto 1.2.8.

La Figura 43 muestra los resultados obtenidos para cada uno de los casos.

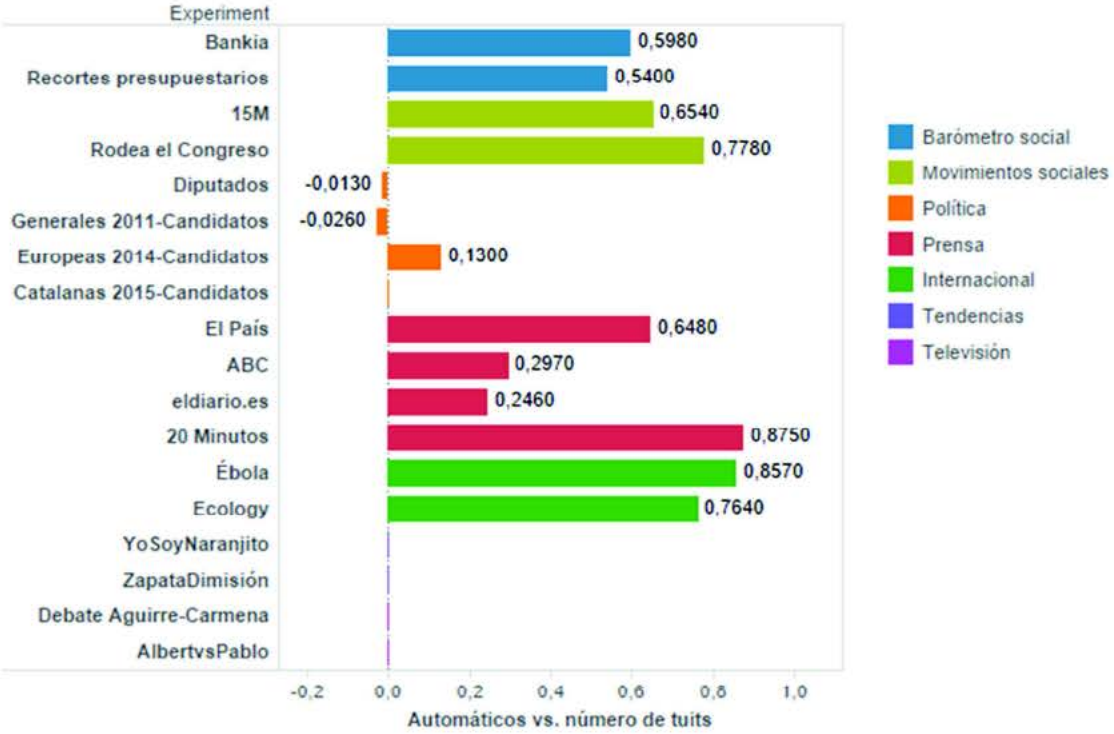

Figura 43 Automatismo vs. número de tuits

Este indicador es el que presenta valores más diferenciados para cada uno de los casos. Aparecen altas correlaciones en los casos de 20 Minutos y Ébola, siendo la correlación más moderada en los casos de Barómetro social y Movimientos sociales. Es prácticamente nula en la Política, las Tendencias y la Televisión.

La aparición de los tuits automáticos está muy vinculada a la publicación de noticias. Las noticias que generan los medios son el punto de arranque de las conversaciones que se encuentran en muchos de los casos analizados. El caso Bankia se nutre de los distintos sucesos que van ocurriendo o escándalos que se van desvelando. Los recortes presupuestarios siguen generando artículos de denuncia de recortes sociales. El Ébola fue portada de los periódicos durante meses. La ecología genera reportajes en secciones especializadas de los medios y en muchos blogs.

Esto no ocurre en los casos de Política porque los tuits corresponden a la interacción de los políticos con otros usuarios. Tampoco sucede en las Tendencias que siendo causadas por una noticia, la avalancha de tuits se genera por una polémica. En el caso de la Televisión, lo que se comenta generalmente es lo que se está viendo y no tienen cabida las noticias.

Casi todos los casos que incluyen noticias tienen una correlación grande, excepto el periódico ABC y eldiario.es. No obstante, una mirada detallada a los diagramas de dispersión aportará información sobre cuán abundante es la aparición de estos tuits automáticos.

Merece un análisis específico los casos de la prensa (Figura 44), que siendo entornos similares muestran diferencias significativas.

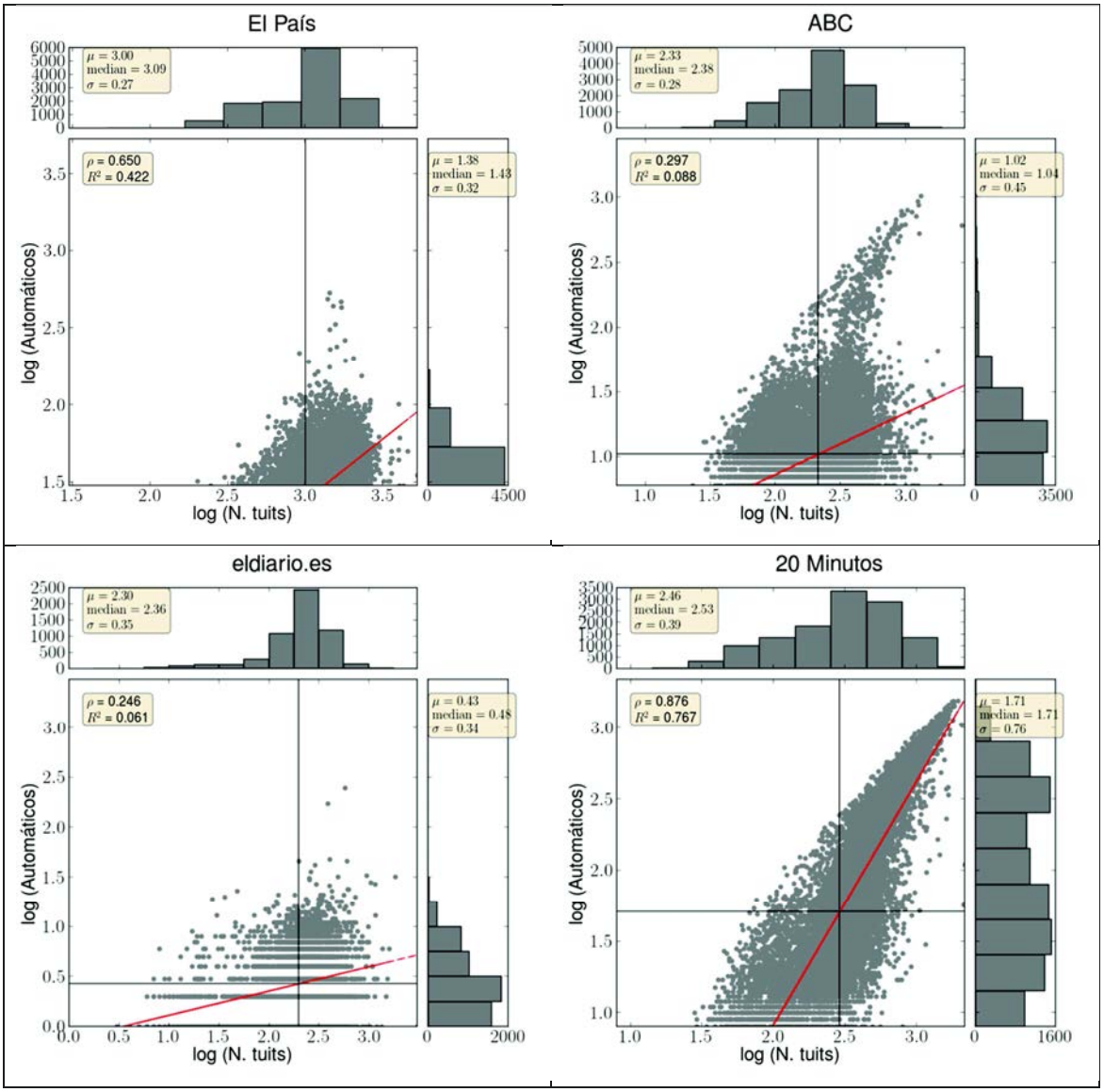

Figura 44 El automatismo en la prensa

En El País la mayoría de las veces el número de tuits automáticos por intervalo se encuentra entre 35 y 50 (mediana en 29), llegando muy esporádicamente a 500 tuits automáticos a la hora.

En el ABC el número más frecuente de aparición de estos tuits por intervalo se halla entre 3 y 50 (mediana en 11), llegando en algunos casos a detectarse entre 100 y 1.000.

En eldiario.es la presencia de estos tuits es menor, entre 1 y 5 (mediana en 3,2) llegando excepcionalmente a 300.

20 Minutos es un caso muy diferente. La distribución de estos tuits es casi uniforme, la correlación es muy alta y la mediana se sitúa en los 50 tuits por intervalo, superando en algunos momentos los 1.000 tuits.

Para ver el papel que juegan los usuarios que publican estos tuits recurro a la Figura 45 que muestra, para los casos de la Prensa, los diagramas de dispersión del ratio de propagación vs. el ratio de red para cada uno de los roles.

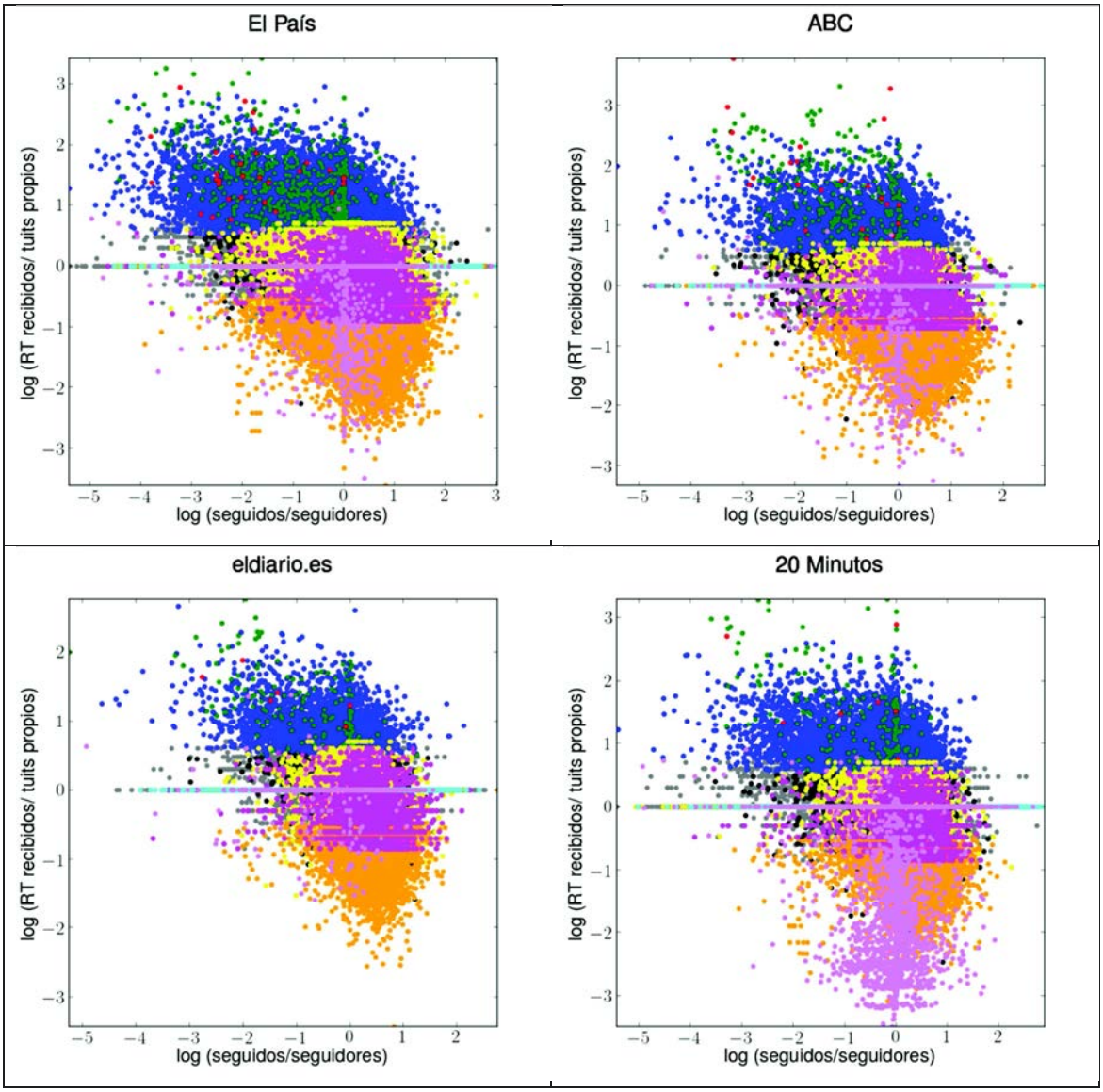

Figura 45 Perfiles automáticos en la prensa

En los diagramas de dispersión de El País y del ABC, los roles Automáticos (color violeta) se encuentran mayoritariamente ubicados en la parte inferior del diagrama próximos a los ejes de abscisas y de ordenadas. Esto significa que el ratio de red está cerca de 1 (número de seguidores = número de seguidos) y que sus mensajes se difunden poco. Respecto su

número, en El País son el 0,15% de los usuarios y en el ABC el 0,43%, como se vio en el análisis macroscópico de la participación de roles (punto 5.1.3).

En eldiario.es el porcentaje de Automáticos es el 0,44%, situándose cerca del eje de abscisas y en la parte inferior del diagrama de una manera más dispersa.

En 20 Minutos, el porcentaje de Automáticos se eleva a 1,71% distribuyéndose en el eje de abscisas y muy extensamente en la parte inferior del diagrama, en algunos casos con ratios de propagación inferiores a 1/1000. Estos ratios de propagación tan bajos son consecuencia de publicaciones masivas sin difusión. También se observan en estos usuarios ratios de red inferiores a cero (más seguidores que seguidos).

Todo apunta a que en 20 Minutos la publicación de los tuits está forzada artificialmente, mientras que en el País, el ABC y eldiario.es tan solo existe algo de ruido. Una pregunta que surge es si la inyección de tuits automáticos favorece a la publicación de tuits. Para ver cómo afecta he vuelto a calcular el índice de correlación del Alcance respecto al número de tuits, eliminando previamente los tuits automáticos. El resultado se muestra en la Figura 46.

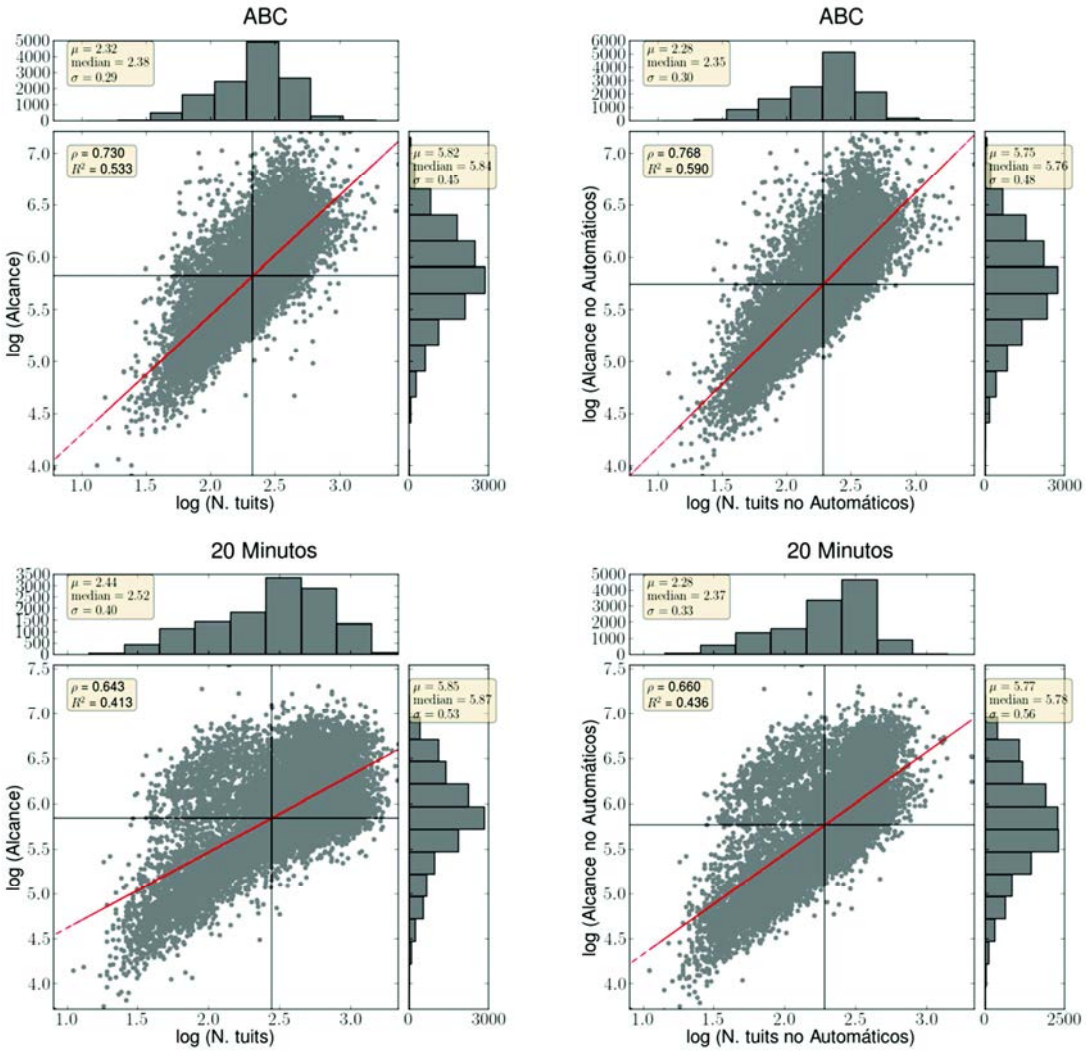

Figura 46 El Alcance con y sin Automáticos

He seleccionado los casos de ABC y 20 Minutos por ser los que tienen correlaciones más bajas de Alcance y más presencia de Automáticos.

En el caso de ABC la correlación aumenta ligeramente y se mantienen las distribuciones del número de tuits y Alcance en ambos casos.

En 20 Minutos la correlación de alcance baja ligeramente y en el caso de no Automáticos la distribución de tuits cambia, desapareciendo del histograma los momentos de mayor frecuencia de tuits.

De estos resultados se puede concluir que el ruido producido por las publicaciones automáticas no afecta a la difusión. El indicador de Alcance está más relacionado con la receptividad de la audiencia que con la inyección automática de tuits.

5.5 La evolución en el largo plazo

Una de las ventajas de disponer de casos de estudio de larga duración, entre dos y cuatro años, es la posibilidad de medir la evolución de la propagación de una manera continua. Esta medida sirve para fortalecer o debilitar las tendencias observadas en los casos de corta duración, como las campañas electorales o los movimientos sociales, que ocurrieron en distintos años.

5.5.1 Evolución de los mensajes retransmitidos

La primera tendencia observada en casos que sucedieron en años diferentes es el incremento del porcentaje de retransmisiones con el tiempo (Figura 25). Esta propensión se confirma en las medidas obtenidas para los casos de larga duración.

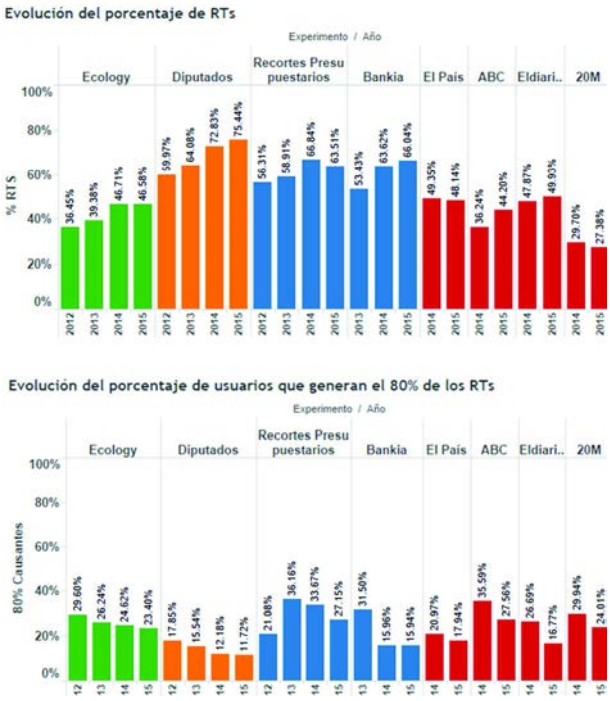

Figura 47 Evolución de la propagación

Salvo en el caso de los recortes presupuestarios, todos los casos muestran una tendencia creciente en el porcentaje de retransmisiones como se puede ver en la Figura 47. En la parte inferior de esta imagen se puede observar una tendencia decreciente en la evolución del porcentaje de usuarios que generan el 80% de las retransmisiones. Esta predisposición no coincide en el caso de los movimientos sociales pero sí en el las campañas electorales (Figura 27).

5.5.2 EVOLUCIÓN DE LA PARTICIPACIÓN DE LOS ROLES

La evolución de los distintos perfiles a través de los años puede indicar si están cambiando las dinámicas de comunicación. La Figura 48 muestra la evolución de los perfiles Altavoz alto y medio. Debido a los valores tan pequeños de estos porcentajes, la escala de la gráfica está acotada de 0 al 1%.

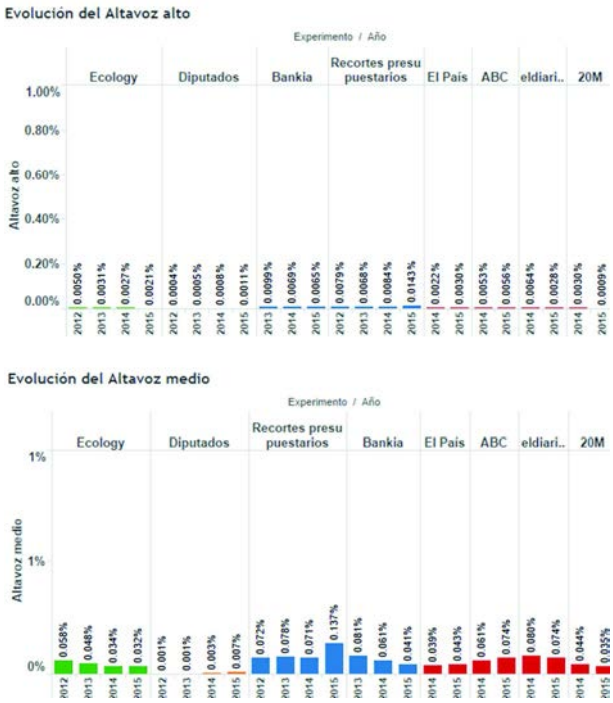

Figura 48 Evolución del Altavoz alto y medio

Como se puede apreciar la variación del Altavoz alto es mínima y la del Altavoz medio se incrementa en algunos casos y disminuye en otros. En los casos de duración corta, del 15M a Rodea el Congreso, existe una reducción considerable del 0,23% al 0,07% en el Altavoz medio. Los casos de las campañas electorales tienen valores tan bajos que no se puede distinguir la diferencia (Tabla 13). No se puede determinar ninguna tendencia para estos casos.

El perfil Altavoz bajo que aumentó en los casos de duración corta más recientes (Figura 28), también creció en los casos de duración larga, quedando la tendencia fortalecida. La Figura 49 muestra la evolución de este perfil.

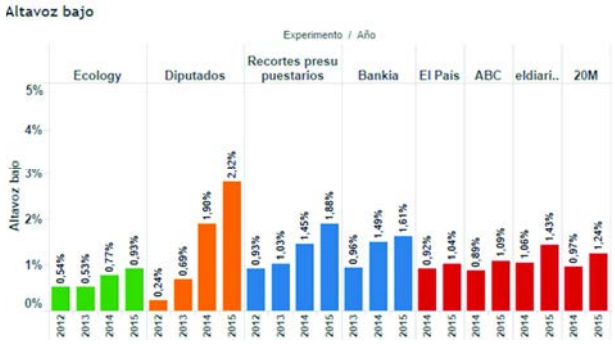

Figura 49 Evolución del Altavoz bajo

Los perfiles Aislado y Monologuista que aportan nula o poca difusión respectivamente, disminuyeron en los casos más recientes igual que ocurrió en los casos de larga duración, confirmando así la tendencia. La Figura 50 muestra la evolución de estos perfiles a lo largo de los últimos años.

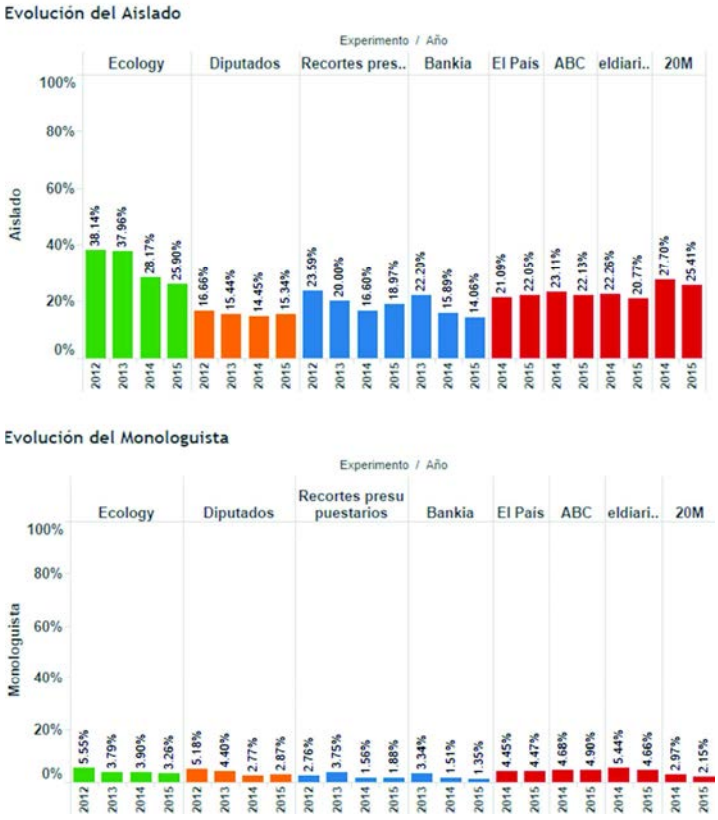

Figura 50 Evolución del Aislado y Monologuista

La Figura 51 recoge la evolución del perfil Retuiteador y Común. El Retuiteador se ha mantenido sin muchos cambios en los casos de larga duración. Podría pensarse que si han crecido las retransmisiones debería haber aumentado también este tipo de usuarios, pero no es así. El aumento de las difusiones debe ser causado por un incremento de los RTs de este perfil o de otros grupos. El perfil común aumentó en algunos casos.

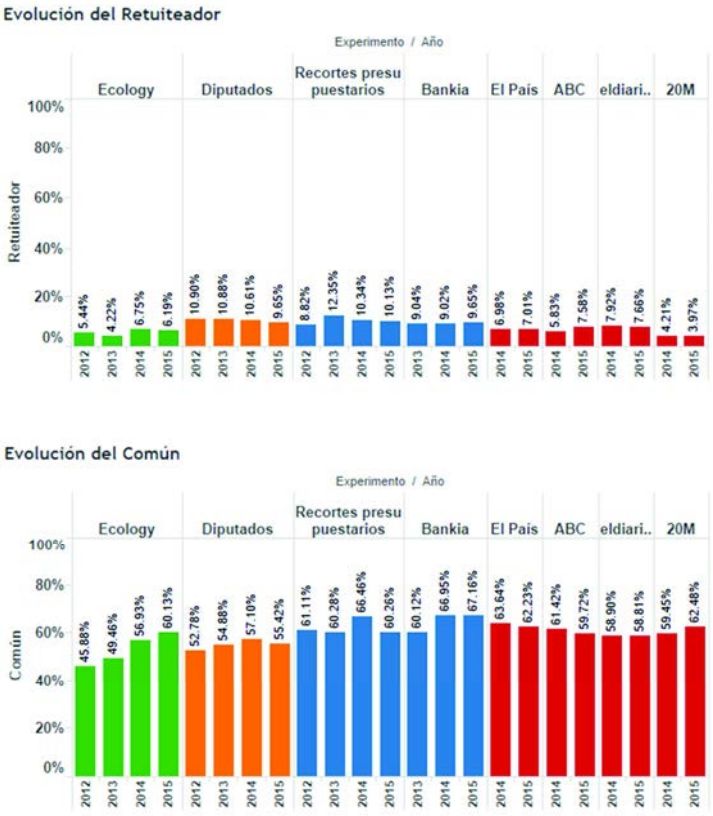

Figura 51 Evolución del Retuiteador y Común

La evolución de los perfiles en el tiempo confirman las tendencias encontradas en los casos de corta duración que sucedieron en distintos años. Los tipos de usuario que aportan menos a la difusión disminuyen, el Retuiteador apenas cambia y aumenta el Altavoz Bajo. El aumento de este último perfil indica se amplía el número de usuarios que publican tuits que interesan a otros usuarios.

6 CONCLUSIONES, CONTRIBUCIONES Y TRABAJOS FUTUROS

6.1 CONCLUSIONES

Esta Tesis aborda el estudio de la difusión en Twitter desde el punto de vista del usuario, desde el comportamiento global en el corto plazo (una hora) y desde la evolución en el largo plazo. Todo ello lo realiza a través de la observación de un conjunto de casos de estudio con diferentes dimensiones, duraciones y contextos. A continuación se exponen las principales conclusiones que se han obtenido tras este análisis.

La participación de los usuarios en Twitter en temas sociales se manifiesta de una forma asimétrica, unos pocos despliegan mucha actividad mientras que el resto participa de una forma pasiva. Salvo en los casos de la Prensa, más de la mitad de los mensajes publicados son retransmisiones y en temas relacionados con la política esta cifra puede llegar al 80%. Se podría decir que existe una burbuja de actividad en la que la información original que se publica es pequeña, siendo el impacto que se percibe grande.

Una élite de usuarios, entre el 0,1% y el 3,9% según los casos, son los autores de los tuits que acaparan el 80% de los RTs, mientras que son menos de un tercio los que realizan la mayoría de las retransmisiones. El resto de usuarios participa esporádicamente. Podría parecer en este escenario que la propagación está acaparada por unas élites de las que generalmente se asume que tienen más seguidores que seguidos. Sin embargo, se ha encontrado en todos los casos analizados un rasgo de meritocracia en el que usuarios con menos seguidores que seguidos han sido propagados. De hecho, la correlación entre el ratio de red (seguidores/seguidos) y los mensajes difundidos de un usuario es prácticamente nula. Induce a pensar que hay otros factores que influyen más en la propagación que la relevancia del autor en Twitter y que posiblemente uno de ellos sea la calidad del mensaje emitido. Es posible que la forma de consumir información en Twitter desde las tendencias de Twitter o desde programas de televisión que invitan a participar al espectador o siguiendo sucesos que despiertan la curiosidad del usuario por conocer la información más actual, propicien que los mensajes puedan ser leídos más allá de la red de seguidores, siendo probable que un mensaje ingenioso se difunda independientemente si se conoce o no al autor.

Se ha observado que la distribución de los tipos de usuarios es diferente en los casos de estudio, lo que indica distintas dinámicas de comunicación. Los casos con mayor porcentaje de Retuiteadores y Altavoces bajos son los que tienen un mayor porcentaje de RTs. Estos casos coinciden generalmente con los temas relacionados con la política, que parecen suscitar más interés entre los usuarios que en otros entornos.

Cuando se mide lo que ocurre a corto plazo se encuentra una relación fuerte entre el Alcance de ese momento (audiencia posible) y los tuits que se publican en ese intervalo (reacción). Es como si durante ese tiempo las posibles audiencias respondieran proporcionalmente. Las distintas medidas del Alcance en el tiempo siguen una distribución casi-normal, lo mismo que la publicación de tuits por intervalo. Todo parece indicar que la publicación de tuits se produce aleatoriamente y que las audiencias expuestas a estos tuits generan un número proporcional de tuits con un alcance también aleatorio. Cuando se mide en un plazo más largo, las distribuciones son muy distintas; los tuits publicados están concentrados en un grupo de usuarios y siguen una distribución de la ley de potencias. Lo mismo ocurre cuando se mide en unidades cortas de tiempo la proporción entre Alcance y RTs o RTs y mensajes publicados. Es como si esfuerzos individuales muy asimétricos confluyeran en un intervalo corto de tiempo con una frecuencia aleatoria. Da la impresión que la sociedad es como un ser independiente que emerge de los individuos que la componen y trasciende a la realidad de éstos

La participación de los usuarios a corto plazo respecto al número de tuits publicados es la relación más fuerte encontrada de todos los indicadores y para todos los casos. Es como si existiese un coeficiente de participación propio de cada contexto. Esta proporción es diferente en cada caso y se cumple cuando la actividad de los usuarios es alta o baja o cuando los casos de estudio son de corta o de larga duración. La relación entre la incorporación de nuevos usuarios a corto plazo y el número de tuits es menos fuerte. En este caso, el tiempo transcurrido o la pérdida de actualidad del tema estudiado saturan la aparición de nuevos usuarios.

Finalmente, analizados los casos de más de dos años de duración se ha encontrado que aumenta el número de retransmisiones mientras que disminuye el tamaño de los grupos que realizan los RTs. En los casos que son recurrentes, como las campañas electorales, también se observa la misma tendencia. Se podría decir que los usuarios activos son cada vez más activos. Respecto a los tipos de usuario, disminuyen los que propiciaban poco la difusión (Aislados y Monologuistas), los Retuiteadores apenas varían y aumentan los Altavoces bajos. Se podría concluir que está aumentando el número de usuarios que consiguen captar la atención de otros aunque su difusión es baja.

6.2 CONTRIBUCIONES

Una de las principales contribuciones de esta tesis ha sido la implementación de la plataforma T-hoarder que almacena, procesa y visualiza tuits según un criterio de búsqueda. Esta plataforma tiene una arquitectura sencilla, pocas dependencias de paquetes software, usa eficientemente los recursos de almacenamiento, es escalable para grandes volúmenes de datos y su código está liberado en un repositorio software público con una licencia *GNU general public license versión 3*. Es una herramienta que ha nacido bajo el espíritu 2.0, siendo concebida para ser compartida y reutilizada por otros investigadores en sus experimentos. Hasta el momento ha servido como herramienta cuantitativa para una tesis (Crespo, 2013) y una tesis de licenciatura en Cuba (Reyes, 2014).

T-hoarder es la base sobre la que se sustenta el experimento Metroaverías, una prueba de concepto del ciudadano como sensor de la calidad de los servicios públicos (Congosto et

al., 2015). Este ensayo que combina el análisis de la propagación junto con una capa semántica puede ser una evolución de T-hoarder hacia servicios de medida de satisfacción de clientes.

En esta tesis se define una clasificación original de usuarios basada en su comportamiento, habiendo sido validada por la coherencia en la que aparecen estos perfiles en las métricas realizadas. El enfoque de la propagación desde el punto de vista de los perfiles que participan en ella abre una nueva vía para el análisis de las dinámicas de comunicación. El algoritmo para clasificarlos está basado en las interacciones entre los usuarios de forma que será posible conocer en tiempo real cómo está siendo la difusión en función de los perfiles que participan.

El análisis de la propagación por medio de Indicadores ha permitido conocer de forma detallada cómo a las propagaciones les afecta el tiempo, el interés que despiertan los temas tratados y el ruido que producen las publicaciones automáticas.

Finalmente, esta tesis aporta una visión de los datos de larga duración. No he encontrado ningún otro trabajo que aborde el estudio de casos de más de dos años de duración en Twitter. Este enfoque ha permitido ver tendencias basadas en medidas continuas, no solo en análisis de sucesos aislados.

6.3 TRABAJOS FUTUROS

Los trabajos futuros podrían enfocarse por dos vías: la evolución de T-hoarder y la aplicación del modelo de tipos de usuario para el estudio de la propagación.

T-hoarder podría ser adaptado para ser usado por investigadores de las ciencias sociales, añadiéndole una interfaz gráfica que facilite la creación y gestión de experimentos de una manera automática, además de dotarle de un sistema de consulta de datos. Podría ser una valiosa herramienta cuantitativa que facilitase los análisis cualitativos.

Otra evolución del T-hoarder consistiría en la inclusión de los tipos de usuario entre sus indicadores de forma que se pudiera conocer en tiempo real qué tipos de usuario están participando de forma global o por días o incluso por horas. Esto permitiría hacer observaciones sobre las dinámicas de comunicación.

Una vez realizada la prueba de concepto de Metroaverías y con la experiencia adquirida se podría evolucionar la plataforma T-hoarder hacia un T-hoarder-S añadiéndole una capa semántica que incluyera clasificadores automáticos. Desde esta plataforma ampliada sería factible dar servicios similares al que se ofrece con Metroaverías pero aplicados a otros servicios públicos y por extensión a servicios comerciales de atención al cliente.

El análisis de las dinámicas de comunicación puede abordarse desde el enfoque de la participación de los distintos tipos de usuario. Una manera de realizarlo sería el estudio de las cadenas de difusión en la que se tuviera en cuenta la proporción de estos tipos y su distribución en el tiempo, siendo posible calcular la velocidad de propagación por tipo de usuario.

También sería interesante estudiar los patrones de propagación para los usuarios con un ratio de red superior o inferior a 1 para ver si el tipo de red determina la forma de las cadenas de difusión en el tiempo y en el tipo de usuarios que participan.

PARTE V. Apéndices y referencias

APÉNDICE A ACRÓNIMOS Y DEFINICIONES

ACRÓNIMOS

API

Application **P**rogramming **I**nterface (interfaz para programar aplicaciones).

DM

Direct **M**essage (mensaje privado entre usuarios de Twitter).

HTML

Hyper**T**ext **M**arkup **L**anguage (lenguaje de marcas de hipertexto con el que se crean las páginas web).

HTTP

Hyper**T**ext **T**ransfer **P**rotocol (protocolo utilizado para las transacciones con la Web).

JSON

Java**S**cript **O**bject **N**otation (formato ligero de transferencia de datos en el que la estructura está autocontenida y viaja con los datos).

OAuth

***O**pen **Auth**orization* (un protocolo abierto que permite la autorización segura a aplicaciones web, móviles y de escritorio que aceden a un servicio web).

RT

Re**T**weet (mensaje de Twitter retransmitido. Es un duplicado un mensaje en el que se indica el autor, el texto y quién lo duplica).

SVM

Support **V**ector **M**achine. Son algoritmos de aprendizaje supervisado que entrenados mediante un conjunto de ejemplos etiquetados son capaces de etiquetar automáticamente otras muestras.

TPS

Tweets **P**er **S**econd (Número de tuits por segundo).

URL

Uniform **R**esource **L**ocator (localizador de un recurso dentro de Internet).

UTF-8[38]

Unicode **T**ransformation **F**ormat **8**-bit (codificación de caracteres unicode que codifica símbolos de longitud variable).

Web

World **W**ide **W**eb (es un sistema de distribución de documentos de hipertexto interconectados y accesibles vía Internet).

[38] http://en.wikipedia.org/w ki/UTF-8

DEFINICIONES

Bases de datos

Repositorios que almacenan datos estructurados, categorizados y enlazados por relaciones y dotados de herramientas que permiten buscar fácilmente información.

Big data

Se refiere a los conjuntos de datos cuyo tamaño va más allá de la capacidad de las herramientas típicas de software de base de datos para capturar, almacenar, gestionar y analizar. Se le asocian tres magnitudes (las tres v): velocidad, variedad y volumen.

Blog

Término resultante de la contracción de las palabras inglesas ***web log*** (en español se denominó bitácora, aunque de forma minoritaria). Es un sitio web en el que el autor o autores publican temas de su interés con cierta frecuencia y que pueden ser comentados por sus lectores.

Bot

En este contexto, programa que publica tuits automáticos.

Cron

Planificador para ejecutar periódicamente tareas (comandos, scripts o programas). Utilizado en sistemas operativos del tipo Unix.

Distribución de potencias

Cuando la probabilidad de alguna variable se distribuye de acuerdo con una ley de potencia, su función de distribución se define como:

$$p(x) = Cx^{-\alpha}$$

Siendo $p(x)$ la probabilidad (frecuencia) de que la variable tome un valor de x; α el exponente de la distribución; x la variable que se quiere analizar y C una constante que depende del tipo de evento.

Distribución normal

Una variable aleatoria continua, x, sigue una distribución normal de media μ y desviación típica σ, y se designa por N(μ, σ), si se cumplen las siguientes condiciones:

- La variable puede tomar cualquier valor: $(-\infty, +\infty)$
- La función de densidad, es la expresión en términos de ecuación matemática de la curva de Gauss:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\frac{x-\mu}{2\sigma})^2}$$

Fake

En este contexto, perfil de Twitter de una persona que se escuda en un seudónimo. En algunos casos realizan parodias de un famoso y en otros crean personajes ficticios. En general utilizan el sarcasmo como el medio más habitual de expresión.

Firehose

El caudal completo de Twitter en tiempo real.

Follow back

En Twitter, acción de seguir a un usuario para corresponderle cuando éste nos sigue.

Hashtag

En Twitter, palabra precedida por el símbolo # que se utiliza para etiquetar un tuit.

Iframe

Es un recurso HTML que permite anidar documentos HTML. Es muy utilizado para incrustar pequeñas piezas HTML con una función específica dentro de una página web. El diseño mediante este recurso permite construir las webs como si fueran un puzle.

Listas

Las listas de Twitter permiten etiquetar a los usuarios. Cada usuario puede crear listas bajo su criterio (generalmente son temáticas) e incluir en ellas a usuarios. Las listas pueden ser públicas y otros usuarios se pueden suscribir a ellas. Sirven para:

- Organizar las conversaciones cuando se escucha a demasiados usuarios.
- Facilitar la segmentación de las conversaciones por grupos de interés.
- Agrupar a usuarios con similares intereses.

Mención

Forma de nombrar a un usuario en Twitter mediante la convención de anteponer el símbolo "@" al nombre de su perfil. Cuando un usuario es nombrado por otro, Twitter se lo notifica.

Microblogging

Es una variante del *blog* en el que las publicaciones son muy breves y en algunos casos está limitado a un número de caracteres.

None

Valor nulo de un dato.

Parodia

En este contexto, un caso especial de *Fake* que realiza parodias de un famoso. No intenta suplantar al famoso y suele advertir en su perfil que es una parodia.

Prosumidor

La palabra **prosumidor**, o también conocida como *prosumer*, es un acrónimo formado por la fusión original de las palabras en inglés *productor* (productor) y *consumer* (consumidor). En este contexto se utiliza como productor y consumidor de contenidos.

Python

Es un lenguaje de programación interpretado con una sintaxis que favorece la legibilidad del código. Soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional. Usa tipos dinámicos y es multiplataforma.

Reply

En Twitter, respuesta de un usuario a otro. El mensaje comienza mencionado al usuario al que se responde (@usuario).

Repositorio Software

Sitio centralizado, generalmente accesible desde Internet, donde se almacena, se mantiene y se comparte el Software.

Retweet-aholics

Usuario que retransmite mensajes de Twitter de una manera muy activa.

Socket

Es un mecanismo de comunicación para el intercambio de datos que establece la conexión extremo a extremo entre dos ordenadores remotos conectados mediante una red.

Software

Programas, procedimientos, reglas y cualquier documentación y datos asociados concernientes a la operación de un sistema de ordenador.

Spam

En Twitter, mensajes que son masivamente publicados cuyo contenido está repetido y que incluyen enlaces a publicidad comercial, a veces de productos dudosos o servicios cuasi-legales.

Spam users

Usuarios que publican mensajes de *spam*. Generalmente son *bots*.

Stopwords

Son palabras que por su alta frecuencia de aparición o por no poseer contenido semántico, carecen de valor de discriminación. Suelen ser artículos, pronombres, preposiciones, etc.

Timeline

En Twitter, corriente cronológica continua de tuits que emiten los usuarios seguidos por un perfil.

Token

En este contexto, componente léxico formado por una cadena de caracteres que tiene un significado coherente.

Tuit

Mensaje digital que se envía a través de la red social Twitter® y que no puede rebasar un número limitado de caracteres (definición de la RAE).

Tuit geolocalizado

Tuit que incluye las coordenadas geográficas desde las que fue publicado.

APÉNDICE B TABLAS DETALLADAS

I. VISIÓN MACROSCÓPICA DE LOS USUARIOS

Caso	Altavoz alto	Altavoz medio	Altavoz bajo	Networker	Retuiteador
Bankia	0,00%	0,03%	1,27%	0,29%	8,72%
Recortes presupuestarios	0,01%	0,05%	1,08%	0,62%	9,08%
15M	0,02%	0,23%	0,99%	2,41%	8,90%
Rodea el Congreso	0,01%	0,07%	1,40%	0,34%	12,23%
Diputados	0,00%	0,00%	1,24%	0,08%	9,70%
Nacionales 2011-Candidatos	0,00%	0,00%	0,10%	0,00%	11,82%
Europeas-2014-Candidatos	0,00%	0,00%	1,27%	0,03%	15,26%
Catalanas 2015-Candidatos	0,00%	0,01%	3,51%	0,00%	12,23%
El País	0,00%	0,03%	0,93%	0,63%	6,82%
ABC	0,00%	0,06%	0,99%	0,70%	6,55%
eldiario.es	0,00%	0,06%	1,28%	0,91%	6,86%
20 Minutos	0,00%	0,03%	1,09%	0,54%	4,01%
Ébola	0,00%	0,02%	0,86%	0,27%	11,41%
Ecology	0,00%	0,02%	0,63%	0,48%	5,39%
YoSoyNaranjito	0,02%	0,32%	3,16%	0,25%	14,69%

Caso	Altavoz alto	Altavoz medio	Altavoz bajo	Networker	Retuiteador
ZapataDimisión	0,03%	0,14%	2,10%	0,19%	16,33%
Debate Aguirre-Carmena	0,03%	0,18%	2,57%	0,22%	17,26%
AlbertvsPablo	0,01%	0,11%	2,16%	0,13%	13,99%

Tabla 13 Porcentaje de usuarios que propician la difusión

Caso	Monologuista	Replicador	Aislado	Común	Automático
Bankia	1,52%	1,17%	16,90%	66,02%	4,07%
Recortes presupuestarios	2,13%	0,88%	20,02%	64,30%	1,84%
15M	4,68%	1,13%	24,48%	55,27%	1,91%
Rodea el Congreso	1,48%	0,19%	14,56%	69,24%	0,48%
Diputados	3,40%	17,28%	15,36%	52,92%	0,01%
Nacionales 2011-Candidatos	8,31%	15,13%	20,11%	44,52%	0,01%
Europeas-2014-Candidatos	2,05%	5,90%	11,18%	64,31%	0,00%
Catalanas 2015-Candidatos	2,01%	9,82%	14,46%	57,97%	0,00%
El País	4,23%	2,56%	20,86%	63,79%	0,15%
ABC	4,41%	3,68%	21,96%	61,22%	0,43%
eldiario.es	4,25%	5,15%	20,61%	60,44%	0,44%
20 Minutos	2,39%	2,24%	26,56%	61,43%	1,71%
Ébola	2,99%	0,57%	17,87%	65,05%	0,97%
Ecology	3,54%	0,90%	30,71%	56,26%	2,08%
YoSoyNaranjito	1,24%	0,09%	16,12%	63,74%	0,38%

Caso	Monologuista	Replicador	Aislado	Común	Automático
ZapataDimisión	1,17%	0,27%	9,18%	70,48%	0,12%
Debate Aguirre-Carmena	1,74%	0,13%	7,80%	69,96%	0,11%
AlbertvsPablo	1,70%	0,13%	12,05%	69,68%	0,05%

Tabla 14 Porcentaje de usuarios que propician menos la difusión

APÉNDICE C PUBLICACIONES

La plataforma T-hoarder ha servido de base para el análisis de ciudadano como sensor del funcionamiento de los servicios públicos, en el caso concreto del Metro de Madrid.

- Congosto, M., Fuentes-Lorenzo, D., & Sánchez, L.,(2015). Microbloggers as Sensors for Public Transport Breakdowns. IEEE Internet Computing, *19*(6), 18 – 25.. Factor de impacto **1,713,** Cuartil Q1

La caracterización de usuarios se realizó por primera vez en la siguiente publicación

- Peña-López, I., Congosto, M., & Aragón, P. (2014). Spanish Indignados and the evolution of the 15M movement on Twitter: towards networked para-institutions. *Journal of Spanish Cultural Studies*,*15*(1-2), 189-216. Cuartil Q4

Por medio de T-hoarder he analizado todas las elecciones en España desde el 2011 aportando un nuevo punto de vista en el análisis de las campañas electorales en Twitter. Los trabajos publicados sobre este tema han sido los siguientes:

- Congosto, M. L. (2015). Elecciones Europeas 2014: Viralidad de los mensajes en Twitter. *Redes: revista hispana para el análisis de redes sociales*, *26*(1), 23-52.
- Triviño, A. I. B., & Congosto, M. (2014). Campaña electoral de las elecciones europeas: Medios de comunicación vs. viralidad de la Red. ALICE 2014
- Congosto, M. (2013). Twitter, una sonda permanente de opinión y una vía para canalizar acciones ciudadanas. GIGAPP 2013
- Congosto, M.L. & Aragón, P. (2012). Análisis de las elecciones 20N. ALICE 2012
- Congosto, M. L., and Pablo Aragón. "Twitter, del sondeo a la sonda: nuevos canales de opinión, nuevos métodos de análisis. *"Más poder local* 12 (2012): 50-56.
- Congosto, M. L., Fernández, M., & Moro, E. (2011). Twitter y política: Información, opinión y ¿Predicción? Cuadernos Évoca nº 4[39]

Aunque el foco de esta tesis esta puesto en el análisis de temas sociales, he explorado otros ámbitos como el entretenimiento con un doble fin: tener otro punto de vista de las dinámicas en Twitter y compartir metodologías con grupos multidisciplinares. Fruto de esta colaboración son las siguientes publicaciones:

- Claes, F., Deltell, L., & Congosto, M. L. (2015). Audiencia social ¿comunidad o enjambre? Caso de estudio: goyas 2014. *Ar@ cne: revista electrónica de recursos en internet sobre geografía y ciencias sociales*. 194. 1 de marzo de 2015 194. 1 de marzo de 2015
- Congosto, M. L., Escolar, L. D., Claes, F., & Osteso, J. M. (2013). Análisis de la audiencia social por medio de Twitter. Caso de estudio: los premios Goya 2013.*Icono14*, *11*(2), 4-30.

[39] http://neolabs.es/evoca/down/cuadernos4.pdf

Con el mismo grupo de investigación participé en un análisis de la figura de Hugo Chavez en el entorno de Twitter:

- Escolar, L. D., Congosto, M. L., Claes, F., & Osteso, J. M. (2013). Identificación y análisis de los líderes de opinión en Twitter en torno a Hugo Chávez. *Revista Latina de comunicación social*, (68), 31-23.

REFERENCIAS

Angermeier, M. (2005). The huge cloud lens bubble map web2.0. *kosmar*. Retrieved from http://kosmar.de/archives/2005/11/11/the-huge-cloud-lens-bubble-map-web20/

Asur, S., & Huberman, B. A. (2010). Predicting the Future With Social Media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference* (pp. 492–499).

Bakshy, E., Hofman, J. M., Watts, D. J., & Mason, W. A. (2011). Everyone ' s an Influencer : Quantifying Influence on Twitter Categories and Subject Descriptors. In *ACM international conference on Web search and data mining* (pp. 65–74).

Barberá, P., & Rivero, G. (2012). Desigualdad en la discusión política en Twitter. *Congreso ALICE*.

Berners-Lee, T., Cailliau, R., Luotonen, A., Frystyk Nielsen, H., & Secret, A. (1994). The World-Wide Web. *Comunications of ACM*, 76–82.

Bild, D. R., Liu, Y., Dick, R. P., Mao, Z. M., & Wallach, D. S. (2015). Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph. *ACM Transactions on Internet Technology (TOIT) (2015): 4.*, *15*(1), 17. doi:10.1145/2700060

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market . *Computer*, *2*(1), 1–8.

Bollen, J., & Pepe, A. (2009). *Modeling Public Mood and Emotion : Twitter Sentiment and Socio-Economic Phenomena*. *arXiv preprint arXiv:0911.1583*.

Bowman, D. (2010). Tweaking the Twitter homepage. *Blog Twitter*. Retrieved from https://blog.twitter.com/2010/tweaking-twitter-homepage

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the Annual Hawaii International Conference on System Sciences* (pp. 1–10). doi:10.1109/HICSS.2010.412

Carrascosa, J. M., Cuevas, R., Gonzalez, R., Azcorra, A., & García, D. (2015). Quantifying the Economic and Cultural Biases of Social Media through Trending Topics. *Plos One*, *10*(7), e0134407. doi:10.1371/journal.pone.0134407

Castells, M. (2002). *The Internet galaxy: Reflections on the Internet, business, and society. , 2002*. (Oxford University Press, Ed.).

Castells, M. (2009). *Comunicación y Poder*. (Alianza Editorial, Ed.). Alianza Editorial.

Cha, M., & Gummadi, K. P. (2010). Measuring User Influence in Twitter : The Million Follower Fallacy. In *CWSM, - aaai.org* (pp. 10–17).

Chakrabarti, S., Dom, B., & Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. *ACM SIGMOD Record*, *27*(2), 307–318. doi:10.1145/276305.276332

Congosto, M. (2009a). Definiendo Twitter en pocas palabras. *Barriblog*. Retrieved from

http://www.barriblog.com/2009/09/definiendo-twitter-en-pocas-palabras/

Congosto, M. (2009b). Por qué NO me gusta el RT (beta) de Twitterd-rollout. *Barriblog*. Retrieved from https://blog.twitter.com/2009/retweet-limited-rollout

Congosto, M. (2011). Evolución de la propagación del 15M en la plaza de Twitter. *Barriblog*. Retrieved from http://www.barriblog.com/2011/05/evolucion-de-la-propagacion-del-15m-en-la-plaza-de-twitter/

Congosto, M. (2014a). El Discurso del Rey 2014. *Barriblog*. Retrieved from http://www.barriblog.com/2014/12/el-discurso-del-rey-2014/

Congosto, M. (2014b). El poder de difusión de un buen tweet. *Barriblog*. Retrieved from http://www.barriblog.com/2014/03/el-poder-de-difusion-de-un-buen-tweet/

Congosto, M. (2014c). Twitter como fuente para conocer la opinión pública. In C. A. de C. / 64 CAC (Ed.), *Las nuevas tecnologías audiovisuales frente a los procesos tradicionales de comunicación* (pp. 117–142).

Congosto, M. (2015a). #YoSoyNaranjito o La ocasión la pintan calva. *barriblog*. Retrieved from http://www.barriblog.com/2015/03/yosoynaranjito-o-la-ocasion-la-pinta-calva/

Congosto, M. (2015b). El hashtag #ZapataDimisión en 360º. *barriblog*. Retrieved from http://www.barriblog.com/2015/06/el-hashtag-zapatadimision-en-360o/

Congosto, M. (2015c). El novio ninja. *Barriblog*. Retrieved from http://www.barriblog.com/2015/08/el-novio-ninja/

Congosto, M. (2015d). Elecciones Europeas 2014 : Viralidad de los mensajes en Twitter. *Revista Redes*, *26*, 23–52.

Congosto, M. (2015e). La delgada línea entre el activismo y el spam en Twitter. *Barriblog*. Retrieved from http://www.barriblog.com/2015/02/la-delgada-linea-entre-el-activismo-y-el-spam-en-twitter/

Congosto, M. (2015f). Propagación TV + Twitter: el caso de #leoncomegamba-leoncomegamba. *barriblog*. Retrieved from http://www.barriblog.com/2015/04/propagacion-tv-twitter-el-caso-de-leoncomegamba/

Congosto, M., Fuentes-Lorenzo, D., & Sánchez, L. (2015). Microbloggers as Sensors for Public Transport Breakdowns. *IEEE Internet Computing*, *19*(6), 18 – 25.

Conover, M. D., Gonc, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). Predicting the Political Alignment of Twitter Users. In *IEEE Third Inernational Conference on Social Computing (SocialCom)* (pp. 192–199).

Crespo, M. (2013). *Predicción Electoral Mediante Análisis de Redes Sociales*. Universidad Complutense de Madrid. Retrieved from http://eprints.ucm.es/22019/1/T34588.pdf

Crockford, D. (2006). The application/json media type for javascript object notation (JSON). Retrieved from https://tools.ietf.org/html/rfc4627

De Domenico, M., Lima, a, Mougel, P., & Musolesi, M. (2013). The anatomy of a scientific rumor. *Scientific Reports*, *3*, 2980. doi:10.1038/srep02980

DiNucci, D. (1999). Fragemented Future. *Http://www. Darcyd. Com/fragmented Future. Pdf*.

Dodds, P. S., Harris, K. D., Isabel, M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS One 6.12*,

6(12), e26752.

Drew Conway. (2010). Package " infochimps ." Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.192.487&rep=rep1&type=pdf

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3), 75–174. doi:10.1016/j.physrep.2009.11.002

Gayo-Avello, D. (2011a). All liaisons are dangerous when all your friends are known to us. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia* (pp. 171–180).

Gayo-Avello, D. (2011b). Don't turn social media into another "Literary Digest" poll. *Communications of the ACM*, *54*(10), 121–128. doi:10.1145/2001269.2001297

Gayo-Avello, D., Brenes, D. J., Fernández-Fernández, D., Fernández-Menéndez, M. E., & García-Suárez, R. (2010). De retibus socialibus et legibus momenti, 1–20. doi:10.1209/0295-5075/94/38001

Gonzalez, R., Cuevas, R., & Cuevas, A. (2011). *Where are my followers ? Understanding the Locality Effect in Twitter*. *arXiv preprint arXiv:1105.3682*.

González-Bailón, S., Borge-Holthoefer, J., & Moreno, Y. (2013). Broadcasters and Hidden Influentials in Online Protest Diffusion. *American Behavioral Scientist*, (0). doi:10.1177/0002764213479371

Grabowicz, P. a, Ramasco, J. J., Moro, E., Pujol, J. M., & Eguiluz, V. M. (2012). Social features of online networks: the strength of intermediary ties in online social media. *PloS One*, *7*(1), e29358. doi:10.1371/journal.pone.0029358

Granovetter, M. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380. doi:10.1086/225469

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proc Natl Acad Sci U S A*, *102*(46), 16569–16572. doi:10.1073/pnas.0507655102

Huberman, B. A., Romero, D. M., & Wu, F. (2008). Social networks that matter : Twitter under the microscope. *Available at SSRN: http://ssrn.com/abstract=1313405*.

Kamdar, J. (2010). More Cities, More Trends, More Understanding. *Blog Twitter*. Retrieved from https://blog.twitter.com/2010/more-cities-more-trends-more-understanding

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter , a Social Network or a News Media? In *The International World Wide Web Conference Committee (IW3C2)* (pp. 1–10). doi:10.1145/1772690.1772751

Liere, D. Van. (2010). How Far Does a Tweet Travel ? Information Brokers in the Twitterverse. In *Proceedings of the International Workshop on Modeling Social Media* (pp. 1–4).

Manyika, J., Chui, M., Brown, B., Bughin, J., & Dobbs, R. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. In *ICWSM* (pp. 554–557).

Morales, A. (2014). *Análisis y modelización de la dinámica emergente durante el proceso de difusión de información en las redes´ sociales de internet*. Doctoral dissertation, Agronomos.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Power Laws, Pareto Distributions and Zipf's Law. Contemporary Physics*, *46*(5), 323–351. doi:10.1016/j.cities.2012.03.001

Osteso, J., Claes, F., & Deltell, L. (2013). Teoría de la urdimbre comunicativa. Política, activismo y formación de líderes de opinión por medio de Twitter en España. In *Seminario internacional AISOC*.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web Internet And Web Information Systems*, *54*(1999-66), 1–17. doi:10.1.1.31.1768

Peña-lópez, I., Congosto, M., & Aragón, P. (2014). Journal of Spanish Cultural Studies Spanish Indignados and the evolution of the 15M movement on Twitter : towards networked para-institutions. *Journal of Spanish Cultural Studies*, *15*(1-2), 189–216.

Perrin, A., Duggan, M., Rainie, L., Smith, A., Greenwood, S., Porteus, M., & Page, D. (2015). *Social Media Usage: 2005-2015*. *Pew Research Center*. Retrieved from www.pewresearch.org

Pravda, K. (2009). The Million Followers Fallacy. *Adi Avnite*. Retrieved from http://web.archive.org/web/20090826115629/http://pravdam.com/2009/08/20/the-million-followers-fallacy-guest-post-by-adi-avnit/

Preotiuc-Pietro, D., & Samangooei, S. (2012). Trendminer: An architecture for real time analysis of social media text. *Sixth International AAAI Conference on Weblogs and Social Media*, 4–7. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/download/4739/5087

Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2011). In the mood being influential on twitter mood. In *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011* (pp. 307–314). doi:10.1109/PASSAT/SocialCom.2011.27

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents - SMUC '10* (pp. 37–44). doi:10.1145/1871985.1871993

Reyes, S. (2014). *Interacción social entre usuarios cubanos en Twitter*. Universidad de La Habana.

Romero, D. M., & Huberman, B. A. (2011). Influence and Passivity in Social Media. In *Machine learning and knowledge discovery in databases. Springer Berlin Heidelberg, 2011.* (pp. 18–33).

Sakamoto, T. (2015). *An analysis on information diffusion by retweets in Twitter*. *Doctoral dissertation, Massachusetts Institute of Technology*.

Stone, B. (2009a). Retweet Limited Rollout. *Blog Twitter*. Retrieved from https://blog.twitter.com/2009/retweet-limited-rollout

Stone, B. (2009b). What's Happening? *Blog Twitter*. Retrieved from https://blog.twitter.com/2009/whats-happening

Sysomos. (2009). Inside Twitter. *Sysomos*. Retrieved from https://sysomos.com/inside-twitter

Tunkelang, D. (2009). A Twitter Analog to PageRank. *The Noisy Channel*. Retrieved from http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank

Uddin, M. M., Imran, M., & Sajjad, H. (2014). *Understanding Types of Users on Twitter*. Retrieved from arXiv preprint arXiv:1406.1335

Val, E. del, Rebollo, M., & Botti, V. (2015). Does the Type of Event Influence How User Interactions Evolve on Twitter? *Plos One*, *10*(5), e0124049. doi:10.1371/journal.pone.0124049

Weng, J., Lim, E., & Jiang, J. (2010). TwitterRank : Finding Topic-sensitive Influential Twitterers. In *third ACM international conference on Web search and data mining* (Vol. Paper 504, pp. 261–270). doi:10.1145/1718487.1718520

Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L., & Su, Z. (2010). Understanding Retweeting Behaviors in Social Networks. In *ACM international conference on Information and knowledge management* (pp. 1633–1636). doi:10.1145/1871437.1871691

Zar, J. H. (1998). Spearman rank correlation. In *Encyclopedia of Biostatistics.*