# Análisis de temas emergentes a través de Twitter

*Analysis of emerging subjects in Twitter*

**José Luis ALONSO BERROCAL, Carlos G. FIGUEROLA, Ángel F. ZAZO RODRÍGUEZ**

Departamento de informática y automática, Universidad de Salamanca.
{berrocal | figue | zazo}@usal.es

**Resumen**

El análisis de temas emergentes en las redes sociales se aplica para conocer las opiniones que expresan usuarios individuales, para controlar actividades y actos de asociaciones, analizar las campañas de los políticos o estudiar el impacto de campañas publicitarias por parte de las empresas. Para la detección de dichos temas se aplicó el algoritmo Latent Dirichlett Allocation a un conjunto de perfiles del ámbito de la información y documentación, con el fin de conocer los temas que se tratan en dichos grupos y para evaluar si el sistema de detección es fiable. El sistema funciona correctamente y proporciona resultados fialbes.

**Palabras clave:** Twitter. Temas de actualidad. Detección de temas. Latent Dirichlett Allocation.

**Abstract**

Analysis of emerging issues in social networks applies to the views expressing individual users, to control activities and acts of associations, analyze political campaigns or study the impact of advertising campaigns by companies. For detection of these issues the algorithm Latent Dirichlett Allocation shall apply to a set of profiles in the field of information and documentation, in order to know the topics covered in these groups and to assess whether the detection system is reliable. The approach works correctly, and provides reliable results.

**Keywords**: Twitter. Trending topics. Topic detection. Latent Dirichlett Allocation.

## 1. Introducción

El crecimiento y diversificación de los recursos informativos, asociado a la popularización de la web ha propiciado un cambio en los hábitos de consumo de información por parte de los ciudadanos. En 2010 la OCDE publica el informe *The Evolution of News and the Internet* (Wunsch-Vincent y Vickery, 2010) en el que se documenta que más de la mitad de la población de los países occidentales se informa habitualmente a través de las ediciones digitales de los periódicos. Ese mismo año el Pew Research Center registra, por primera vez desde el surgimiento de internet, que el número de lectores de la prensa digital supera al de las ediciones en papel de los medios.

Por otra parte, debe considerarse que el cambio descrito ha estado necesariamente acompasado con la creciente implantación de las tecnologías de acceso (más del 68% de los hogares españoles cuenta con acceso a internet), así como con la difusión de nuevas aplicaciones, productos o servicios informativos, entre los que se incluyen desde hace algo más de una década las redes sociales (Pinholster y Ham, 2013).

El uso de las redes sociales se ha extendido entre aquellos agentes más directamente implicados en la difusión de la información. Así para aquellos actores sociales que habitualmente han ejercido el papel de fuente informativa, las redes se han convertido en herramientas útiles para comunicarse de manera directa y sin intermediarios con aquellos colectivos a los que se desea hacer llegar un mensaje. Redes como Twitter permiten a individuos y organizaciones saltarse los filtros tradicionales y acceder de manera directa a lectores, consumidores, fans, etc.

El estudio de los llamados “Trending Topics” (TT), término que hace referencia a los temas tendencia o a los temas que más se hablan en un determinado momento, se aplica para conocer las opiniones que expresan usuarios individuales, para controlar actividades y actos de asociaciones, analizar las campañas de los políticos o estudiar el impacto de campañas publicitarias por parte de las empresas. Tal es la magnitud que ha adoptado este concepto, que incluso ha traspasado las fronteras del propio mundo de Twitter, para pasar a ser una frase de uso común en los medios de comunicación, siendo noticia aquello que se vuelve tema del momento en esta red social. Esto ha propiciado, por ejemplo, que se creen secciones especializadas en periódicos online que se basen en aquello que es tendencia en Twitter.

Este trabajo está organizado como sigue: en el capítulo siguiente se hace una breve introducción

al estado del arte en este campo, para, a continuación, exponer los principales principios metodológicos seguidos en nuestra investigación. Después se expondrán los principales resultados obtenidos del análisis realizado. Finalmente se ofrecen unas conclusiones.

## 2. Trabajos previos

Topic Detection and Tracking (TDT) es un área que comenzó como track en las Text Retrieval Conferences (TREC) (Allan, 2002). Sin embargo, la aplicación de estas técnicas a Twitter es relativamente reciente. Algunos trabajos reseñables son los de (Sankaranarayanan et al., 2009), que aplicaron técnicas de clustering, o los de Petrovic (Petrović et al., 2010), que además constituyó una colección experimental de tweets que ha sido empleada en otros trabajos, el Edinburgh Twitter Corpus (Petrovic et al., 2010). Mathioudakis y Koudas (Mathioudakis and Koudas, 2010) propusieron un sistema para detectar trending topics a partir de un stream de tweets. También sobre la detección de trending topics han trabajado Shariffi, Hotton y Kalita (Sharifi et al., 2010), al igual que Cheong y Lee (Cheong y Lee, 2009), aunque el trabajo de éstos se centra en la evolución temporal de los trending topics.

La cuestión es, en consecuencia, cómo determinar la similitud entre todos los pares de tweets correspondientes a una entidad dada. La similitud entre dos documentos es uno de los problemas centrales de la Recuperación de Información y puede abordarse de diversas formas.

Podemos trabajar con algoritmos basados en modelos estadísticos. Su objetivo es la identificación de los términos relevantes o términos clave del texto, a través del cálculo de los pesos asociados a dichos términos. A este cálculo se le conoce como ponderación del término. Uno de las más conocidos es el de considerar cada documento (tweet) como un bag of words y aplicar un esquema clásico tf x idf (Salton y Buckley, 1988).

Otros algoritmos se basan en la minería de datos que es un campo de las ciencias de la computación que agrupa una serie de técnicas y tecnologías que tienen como misión la extracción de información a partir de patrones contemplados en grandes colecciones de datos estructurados, almacenados en bases de datos. Trata de detectar patrones repetitivos que expliquen el comportamiento de los datos en un determinado contexto. Por ello se trata de una buena técnica para la detección de temas, ya que es capaz de determinar qué elementos son propensos a coocurrir en un conjunto de datos o transacciones. En un contexto de redes sociales y concretamente en Twitter, un elemento w es cualquier término mencionado en un tweet (sin incluir signos de puntuación, palabras vacías, etc). De esta forma la transacción se corresponde con un tweet y el conjunto de transacciones o datos a analizar son todos los tweets que se producen en un intervalo de tiempo $T_j$. El número de veces que un conjunto de términos dado se produce en un intervalo de tiempo se llama soporte, y cualquier conjunto de elementos que cumple un soporte mínimo se llama patrón.

Por otro lado, y entrando en algoritmos de un mayor nivel de dificultad, tenemos los basados en métodos probabilísticos. Según el tipo de salida que produzcan, podemos distinguir varios grupos de algoritmos (Aiello et al., 2013): algoritmos basados en documentos y en funciones.

En el primer grupo destaca el algoritmo LDA. LDA (del inglés, “Latent Dirichlett allocation”) es uno de los modelos Bayesianos más conocidos y usados en la detección de TT. Se trata de un modelo de aprendizaje supervisado para caracterizar el contenido de los mensajes (Benhardus and Kalita, 2013). En él, cada documento se representa mediante una bolsa de términos, que son la única variable observada, a través de la cual se intenta extraer los temas tendencia. Es decir, LDA tiene como entrada un conjunto de términos correspondientes a la representación de cada documento de la colección, dando como salida los temas latentes de cada conjunto de palabras, que es lo mismo que los temas latentes de cada documento. Formalmente, un documento se asocia con una distribución multinomial de temas que a su vez son distribuciones multinomiales de palabras. De esta forma podemos definir la siguiente estructura (Blei et al., 2003):

- Palabra, es la unidad básica y discreta de datos. Se define como un conjunto de elemento de un vocabulario de tamaño V
- Documento (w = {w1,w2,…wN}), es una secuencia de N palabras
- Cuerpo (D= { w1,w2,…wM }), es una colección de M documentos

Al contrario que muchos modelos de agrupación, LDA permite que un documento pueda estar asociado con múltiples temas. Esto se debe a los tres niveles implicados en el algoritmo y en particular, al nodo tema que se muestrea repetidamente en un documento.

Los tweets, sin embargo, son documentos con una serie de características especiales que deberían tomarse en cuenta. Anta y colegas (Anta et al., 2013) mencionan varias de ellas. A las

cuestiones 'clásicas' de la utilización de unigramas o bi, trigramas, etc., o del stemming, en el caso de los tweets debemos añadir los derivados del uso de emoticones, abreviaturas, incluso de un argot propio de ese medio, así como de numerosas anomalías orto-tipográficas. La brevedad de los tweets es otra cuestión importante a considerar (Sriram et al., 2010).

El trabajo de (Mehrotra et al., 2013) utiliza modelos basados en LDA, para trabajar sobre tweets, permitiendo la reconstrucción de clusters y facilitando el análisis de la coherencia de los topics.

El trabajo de Ostrowski (Ostrowski , 2015) analiza diferentes modelos de detección de topics para concluir que el algoritmo LDA, empleado como mecanismo de clasificación, identifica los temas que son dignos de mención y así poder filtrar los mensajes de Twitter.

## 3. Metodología

El trabajo con twitter se ha llevado a cabo trabajando con la API que suministra la red social. Hemos trabajado con la REST API, que nos ofrece el acceso al core de los datos de Twitter. Podemos obtener aproximadamente los últimos 3200 tweets en la primera recogida y a partir de ese momento podemos recoger todo lo nuevo que emite un perfil. Para probar nuestro planteamiento creamos cinco grandes grupos de perfiles relacionados con el campo de la información y documentación, instituciones (26), bibliotecas (33), empresas (25), webs-blogs (14) y profesionales (20), dando un total de 118 perfiles (Tabla II, en apéndice). Hay que indicar que para determinar los perfiles que se iban a analizar se realizó una encuesta a los alumnos del máster en Sistemas de Información Digital, para que nos indicaran aquellos perfiles que consideraban como prioritarios a analizar, y los hemos mantenido tal cual. Con este mecanismo de selección de los perfiles pretendíamos trabajar con perfiles daos por alumnos, que además tienen una procedencia dispar, en cuanto a la titulación de la que proceden. Es evidente que en dicha lista pueden quedar algunos perfiles fueran, que los profesionales pueden echar en falta, pero la visión de los alumnos nos pareció interesante de evaluar.

La serie de resultados obtenida compone el total de los tweets emitidos durante el periodo de recogida para todos los usuarios seleccionados. Esos tweets emitidos incluyen tanto los tweets propios, los escritos directamente por esos usuarios, como sus retweets, tweets escritos por otros, pero difundidos por los usuarios de nuestra lista.

Twitter ofrece una información muy rica para cada tweet descargado. Buena parte de ella es descriptiva: el texto completo del mensaje o datos como el usuario emisor, la fecha y hora de envío, los usuarios mencionados, las direcciones de sitio web citados (url), las referencias a recursos multimedia externos o las etiquetas (hashtags) sirven para caracterizar cada mensaje. Otros campos pueden considerarse más bien metadatos, como los identificadores, las relaciones conversacionales que pueden existir entre los mensajes (respuesta, cita), estados atribuidos al mensaje o número de veces que se ha marcado como favorito, por mencionar algunos. El resto de la información es de tipo administrativo y tiene su papel en la gestión interna de Twitter: preferencias de usuario o atributos para el tratamiento de los mensajes. Toda esta información se recibe en ficheros de texto plano, estructurada según el formato JSON (JavaScript Object Notation). La sintaxis de este formato mediante pares claves-valor que están agrupados en objetos. Conceptualmente se corresponden con los campos y valores de las tablas de una base de datos, pero formalmente son más bien datos brutos que requieren algún tipo de procesado para poder abordar su análisis, bien mediante programas escritos ad hoc, bien mediante una conversión de los datos a un formato que permita llevar a cabo consultas normalizadas.

Es importante señalar que hemos trabajado solamente con los tweets originales emitidos por cada uno de los perfiles, prescindiendo de los retweets.

Los tweets se han procesado eliminando de los mismos las palabras vacías en español, inglés y francés. También se han eliminado los enlaces existentes, así como las menciones de usuario. Hemos mantenido los hashtags, pues pueden dar una buena idea temática o por donde van las tendencias en un determinado momento.

Para analizar la información recogida hemos trabajado con la información correspondiente a 2015, de manera que podemos representar la evolución de temas a lo largo del año completo.

Se han separado los tweets correspondientes a cada semana del año y sobre ellos se ha aplicado el algoritmo de detección LDA. Así hemos obtenido los temas de cada grupo de perfiles, para cada una de las semanas del año.

Este sistema nos permite ver la evolución temática semana a semana, y así podemos capturar los eventos que suceden en un periodo corto de tiempo, que si se analizaran en periodos más largos podían pasar desapercibidos.

## 4. Resultados

El número de tweets analizados en el año 2015 para cada uno de los perfiles fue:

| *Grupo* | *Nº de tweets* |
|---|---|
| Instituciones | 28.188 |
| Bibliotecas | 38.059 |
| Empresas | 14.252 |
| Webs-blogs | 14.328 |
| profesionales | 23.444 |

*Tabla I. Número de Tweets analizados para cada grupo*

Para representar la detección de temas, en el apéndice, hemos puesto las 4 semanas del mes de mayo para cada grupo de perfiles. Podemos así observar la evolución en los temas.

Es significativo cómo uno de los temas emergentes en la última semana de mayo es el hashtag #fesabid (Figuras 16-20), con menor repercusión en webs-blogs que aparece con poco peso, coincidiendo con las jornadas en la última semana del mes. Pero viendo la evolución del hashtag, podemos ver que las instituciones lo incluyen en la tercera semana, las bibliotecas solamente en la 4ª semana, y los profesionales empiezan desde la 2ª semana (Figura 10).

En la primera semana de mayo se celebró la 5ª Conferencia internacional sobre revistas de ciencias sociales y humanidades (CRECS) y tiene su reflejo esencialmente en el grupo webs-blogs y profesionales (Figuras 4-5).

En el grupo webs-blogs aparece todo el mes el hashtag #recbib (Figuras 9-19, con más fuerza), centrado en recursos bibliotecarios, que aparece con fuerza en la última semana del mes de marzo y se mantiene hasta la segunda semana de agosto, desaparece y vuelve a aparecer con fuerza en la primera semana de septiembre.

En bibliotecas se hace eco en la primera semana del centenario de la BID y en empresas las #encuestasneodoc, relacionadas con la #gestiondocumental.

La tercera semana de los profesionales (Figura 15) tiene como tema emergente #cibermetria, coincidiendo con la presentación del primer libro de la colección EPI Scholar *Cibermetría. Midiendo el espacio red.* También aparece el tema #jgic, que refleja las jornadas de gestión de la información científica, que tuvieron lugar el mes de mayo de 2015. Lo mismo sucede con las jornadas bibliosalud, cuyo primer encuentro tuvo lugar el mes de mayo.

Podemos ir viendo el reflejo de los temas y su evolución con el paso de las semanas, para cada grupo de perfiles.

### 4.1. Semana 1


*Figura 1. Instituciones*


*Figura 2. Bibliotecas*


*Figura 3. Empresas*


*Figura 4. Webs-blogs*


*Figura 5. Profesionales*

## 4.2. Semana 2


*Figura 6. Instituciones*

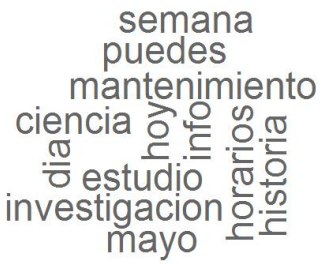


*Figura 7. Bibliotecas*


*Figura 8. Empresas*


*Figura 9. Webs-blogs*


*Figura 10. Profesionales*

## 4.3. Semana 3


*Figura 11. Instituciones*


*Figura 12. Bibliotecas*


*Figura 13. Empresas*


*Figura 14. Webs-blogs*


Figura 15. Profesionales

## 4.4. Semana 4


*Figura 16. Instituciones*


*Figura 17. Bibliotecas*


*Figura 18. Empresas*


*Figura 19. Webs-blogs*


*Figura 20. Profesionales*

## 5. Conclusiones

Planteamos un mecanismo de análisis de los temas emergentes, basado en el algoritmo LDA, que van apareciendo en un conjunto de perfiles y hemos podido ver cómo se refleja la actualidad y el interés en cada uno de dichos grupos.

El sistema funciona correctamente, y nos da unos resultados que creemos muy fiables de las temáticas que se tratan en los grupos de perfiles.

El sistema requiere mejorar los perfiles asignados a los grupos, y probar el empleo de técnicas de stemming, para comprobar si mejora o empeora el sistema.

Este sistema de detección pretendemos implantarlo en tiempo real, de forma que se vayan recogiendo los tweets de forma diaria y al mismo tiempo se vaya construyendo la red de temas emergentes, permitiendo conocer en tiempo real los temas que se están tratando.

## Referencias


Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Jaimes, A. (2013). Sensing trending topics in Twitter. // Multimedia, IEEE Transactions on. 15:6, 1268-1282.

Allan, J. (2002). Introduction to topic detection and tracking Topic detection and tracking (pp. 1-16): Springer.

Anta, A. F., Chiroque, L. N., Morere, P., & Santos, A. (2013). Sentiment analysis and topic detection of Spanish tweets: A comparative study of of NLP techniques. // Procesamiento del lenguaje natural. 50, 45-52.

Benhardus, J.; Kalita, J. (2013). Streaming trend detection in twitter. International Journal of Web Based Communities. 9:1, 122-139.

Blei, D. M.; Ng, A. Y.; Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research. 3, 993-1022.

Cheong, M.; Lee, V. (2009). Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. // Proceedings of the 2nd ACM workshop on Social web search and mining.

Mathioudakis, M.; Koudas, N. (2010). Twittermonitor: trend detection over the twitter stream. // Proceedings of the 2010 ACM SIGMOD International Conference on Management of data.

Mehrotra, R.; Sanner, S.; Buntine, W.; Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. // Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval.

Ostrowski, D. A. (2015). Using latent dirichlet allocation for topic modelling in twitter. // Semantic Computing (ICSC), 2015 IEEE International Conference on.

Petrovic, S.; Osborne, M.; Lavrenko, V. (2010). The edinburgh twitter corpus. // Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media.

Petrović, S.; Osborne, M.; Lavrenko, V. (2010). Streaming first story detection with application to twitter. Paper presented at the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Pinholster, G.; Ham, B. (2013). Science communication requires time, trust, and Twitter. // Science. 342 (6165), 1464-1464.

Salton, G.,; Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. // Information processing & management. 24:5, 513-523.

Sankaranarayanan, J.; Samet, H.; Teitler, B. E.; Lieberman, M. D.; Sperling, J. (2009). Twitterstand: news in tweets. // Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems.

Sharifi, B.; Hutton, M.-A.; Kalita, J. K. (2010). Experiments in microblog summarization. // Social Computing (SocialCom), 2010 IEEE Second International Conference on.

Sriram, B.; Fuhry, D.; Demir, E.; Ferhatosmanoglu, H.; Demirbas, M. (2010). Short text classification in twitter to improve information filtering. // Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.

Wunsch-Vincent, S., & Vickery, G. (2010). The evolution of news and the Internet: OECD.


Enviado: 2016-04-29. Segunda versión: 2016-09-19.
Aceptado: 2016-10-10.