Revista Iberoamericana de Argumentación
ἐπει δὲ ταύτην ἐπιστήμην ζητοῦμεν

Segunda Época
RIA 15 (2017): 16-23

Directores: Luis Vega y Hubert Marraud **Secretaria:** Paula Olmos
ISSN 2172-8801 / doi 10.15366/ria / https://revistas.uam.es/ria

Cuando la ciudad toma la palabra. Geolocalización de la información producida en Twitter: el Proyecto Gnosis

When the city takes the floor. Geolocation of Twitter contents: Project Gnosis

Enrique Alonso
Departamento de Lingüística general, Lenguas modernas, Lógica y filosofía de la ciencia, y Teoría de la literatura y literatura comparada
Universidad Autónoma de Madrid. Facultad de Filosofía y Letras
Ciudad Universitaria de Cantoblanco. 28049 Madrid
enrique.alonso@uam.es

Artículo recibido: 19-10-2017
Artículo aceptado: 30-10-2017

RESUMEN
En este texto resumimos el trabajo realizado durante los últimos años en torno al seguimiento de la información producida en Twitter. Nos centramos de forma destacada en el uso de las herramientas de geolocalizacion de tweets obteniendo diversas conclusiones acerca del valor de los datos obtenidos.

PALABRAS CLAVE: Redes Sociales, Twittter, Big-Data.

ABSTRACT
In this paper I summarize the work carried out during the last years analyzing information produced on Twitter. I focus prominently on the use of geolocation tools for Twitter activity, drawing various conclusions about the value of the data obtained.

KEYWORDS: Social Networks, Twitter, Big-Data

Servicio de Publicaciones de la Universidad Autónoma de Madrid

1. ANTECEDENTES

Hace años que vengo estudiando los flujos de información en Twitter dedicando especial atención a comportamientos como la generación y persistencia de los *hashtgs*, la conducta discursiva en el entorno de los partidos políticos o la información generada por ciertos personajes públicos e instituciones de especial relevancia. En todos estos casos, los objetivos del estudio quedan fijados de antemano y son objeto de un seguimiento a lo largo del tiempo que se restringe a los perfiles analizados y a aquellos que interactúan con ellos en primera instancia. Si, por ejemplo, estamos interesados en analizar la recurrencia de ciertos términos en un partido o perfil determinado,[1] podremos hacerlo sin problemas, siempre y cuando no pretendamos extenderlo a perfiles cuyos resultados no están siendo capturados de antemano. Cuando la muestra de la que parte uno de nuestros análisis se centra en un cierto número de perfiles, la información disponible formará una especie de nube o mancha fuertemente centrada en esa selección previa.

Twitter ofrece, por su parte, la posibilidad de estudiar términos a lo largo de toda su red a través de un sistema de búsqueda avanzada disponible en la interfaz oficial de esta aplicación. Existen además múltiples filtros disponibles que pueden combinarse para acceder a un volumen de datos nada despreciable, aunque limitados por razones obvias. Este hecho junto con el formato en que Twitter ofrece los resultados de estas búsquedas hace que la interfaz oficial solo pueda ser considerada como una ayuda muy general para este tipo de análisis.

La metodología más comúnmente empleada consiste por ello en marcar objetivos fijos cuyos tweets se requieren, junto quizá a otros datos relevantes como menciones, retweets, etc. y se almacenan para su posterior análisis. No existe un límite al volumen de información almacenada distinto de aquel que marquen los recursos de cada equipo de investigación. El formato en que se recogen esos contenidos[2] permite un rápido análisis de la información disponible pudiendo explotar la totalidad de los datos que Twitter ofrece a través de su API (Application Programming Interface).

Cuando los objetivos se fijan en torno a perfiles o términos previamente determinados, hablaremos de búsquedas fuertemente dirigidas. Existen sin embargo otras opciones que hasta ahora no habían sido consideradas seriamente, ni por Twitter,

[1] Este trabajo lo hemos realizado con pares de términos con sesgo ideológico –España/nación, ciudadanos/trabajadores, etc.– dando resultados muy significativos.
[2] Se trata de archivos codificados en *json*.

ni por los equipos de investigación que trabajan con sus datos. Se trata de estudiar la información generada en torno a una determinada ubicación geográfica. A este tipo de análisis lo calificaremos como "débilmente dirigido".

A falta de una terminología más ajustada, lo que se quiere dar a entender con esta distinción es la concentración de intencionalidad que puede mostrar la información recogida en cada caso. Es evidente que los tweets producidos por un perfil, ya se trate de un partido político o un personaje más o menos popular, muestran una intención muy claramente reconocible y además supuestamente coherente y uniforme, a saber, la de sus autores por mostrar sus opiniones en torno a los asuntos de su interés. Aquellos que se recogen debido a la presencia de un cierto término fijado como objetivo exhiben una intencionalidad más difusa, la de múltiples perfiles que sin embargo emplean esa expresión de forma significativa en sus tweets. Por último, tendríamos colecciones de datos cuya característica común es haber sido producidos en una determinada ubicación. Como es obvio, la intencionalidad aparece en este caso completamente diluida ya que ni existe un tema común, ni un perfil reconocible como origen de la información. Podría decirse que este tipo de muestras recogen el ruido de fondo de la ciudad, el flujo de información que constituye la huella característica de cada ubicación a lo largo del tiempo. Mi interés es analizar el comportamiento de esa especie de murmullo urbano y aprender a manejarlo para obtener información valiosa para los estudiosos de la era digital.

2. DISEÑO DE UN EXPERIMIENTO

Hay muchas formas de elegir lugares de observación en proyectos de este tipo y, como cabe imaginar, cada posibilidad implica unos sesgos que deben ser tenidos en cuenta. El proyecto Gnosis[3] viene recopilando información desde agosto de 2016 a través de la selección de un total de 16 ubicaciones[4] de la almendra central de la ciudad de Madrid. El radio que separa estos puntos garantiza que no se produzcan solapamientos aunque al precio más que probable de generar espacios vacíos.

El sistema seguido para recoger datos consiste en generar una conexión cada hora con la API de Twitter solicitando la información producida en cada una de esas ubicaciones desde la última conexión. De esta forma, se evitan solapamientos horarios evitando al mismo tiempo alcanzar las tasas máximas de transacción de

[3] http://zeus.lllf.uam.es/Gnosis
[4] La descripción precisa puede ser encontrada en la página principal de la aplicación.

información que impone Twitter. Una frecuencia mucho más baja podría llevar a perder tweets que no entraran dentro de los cupos máximos que Twitter aporta en sus respuestas. La proporción elegida, una consulta a la hora, queda sobradamente dentro de tales límites. Las zonas seleccionadas no responden, en esta primera versión de la herramienta, a un criterio único. No son, por ejemplo, ubicaciones asociadas a medios de transporte colectivo –la herramienta Locus, ahora en desarrollo, sí obedece a dicho criterio– como pudieran ser estaciones centrales de ferrocarril y autobús, intercambiadores o aeropuertos. Pero tampoco se trata de espacios de reunión, centros deportivos, edificios administrativos o culturales. En realidad se ha seguido un criterio mixto confiando en aprender a reconocer los rasgos distintivos de cada uno de estos tipos de ámbitos urbanos.

La cuestión que hay que plantearse y que siempre surge en este tipo de estudios es la medida en que los resultados pueden ser representativos de alguna conducta relevante más allá de los límites propios de la plataforma digital en cuestión, Twitter en este caso. Hay que tener en cuenta que los análisis de los flujos de información en las redes sociales solo representan el comportamiento de sus usuarios, nada más. Nunca se pueden considerar sin las debidas precauciones como indicios de comportamientos generales que puedan hacerse extensivos al resto de la población. Para lograr una cierta eficacia en la interpretación de este tipo de datos debe operarse siempre teniendo en cuenta el propio comportamiento de la plataforma analizada en situaciones similares que puedan servir como base para la apreciación de comportamientos relevantes. Imagínese, por ejemplo, que estamos analizando la información producida un viernes a la tarde en una ubicación determinada. Solo podemos sacar conclusiones en aquellos casos en los que se aprecien desviaciones sustanciales respecto a los valores medios registrados en esa misma ubicación en periodos similares. Cuando esas desviaciones tienen lugar es cuando se puede pasar a estudiar el tipo de evento que está detrás del comportamiento anómalo intentado incluso descubrir su huella característica o su anticipación en el tiempo.

Otro de los datos relevantes que tienen que ver con el análisis de la información geolocalizada en Twitter es la proporción en que esta tiene lugar. Para que un tweet sea asociado a un lugar determinado tiene que darse en estos momentos una serie de características singulares que no son del todo frecuentes. Ese contenido tiene que haber sido generado activando alguna de las opciones de geolocalización que ofrece Twitter o alguna de las interfaces que emplean su API.

En la actualidad, esa localización puede ser de dos tipos. Una *precisa* que se expresa mediante coordenadas que se obtienen al activar los recursos de que disponen los celulares hoy en día y otra *menos exacta* en la que se hace referencia a una ubicación preseleccionada con anterioridad. Esta posibilidad permite su uso en dispositivos fijos pero puede ser eventualmente confundidora.

Dadas estas condiciones, parece razonable no esperar una incidencia demasiado alta de los tweets geolocalizados en la actividad total de Twitter y, en efecto, así es. Estimaciones aproximadas indican que este porcentaje puede oscilar entre el 2% y el 5% de los tweets generados a diario en la plataforma. Sin embargo sí es cierto que se aprecia un suave incremento de esta proporción en la medida en que los usuarios van aceptando como algo normal la pérdida de privacidad asociada a la geolocalización de su actividad digital. Tiempo atrás se tendía a considerar como una técnica intrusiva por parte de compañías y plataformas el intento de asociar una ubicación al material producido por sus usuarios. Estos aprendían rápidamente a bloquear dicha opción obteniendo así lo que consideraban una mayor privacidad en su vida cotidiana. La popularización de aplicaciones que requerían servicios activos de geolocalización en el celular como callejeros, buscadores de locales de ocio, etc. exigía cada vez más una conducta más permisiva con este tipo de opciones, lo que ha influido en el aumento de contenidos geolocalizados. La irrupción de aplicaciones que, como Instragram, optan por pedir a sus usuarios la localización de sus materiales, junto con la posibilidad de usar cuentas de Twitter vinculadas explica que buena parte del contenido geolocalizado en Twitter sea procedente de esta plataforma de contenidos gráficos. Esta situación, lejos de considerarse como una limitación debe ser vista como una oportunidad para explotar datos con un alto contenido informativo utilizable en direcciones imprevistas.

3. CONCLUSIONES DEL ESTUDIO

Uno de los resultados mejor establecido por los datos se refiere a la existencia de un perfil característico asociado a cada uno de los espacios urbanos analizados. Las tipologías, que no describiremos aquí por falta de espacio, se obtienen al combinar el volumen de información producida en cada zona con la distribución característica. Hay ubicaciones urbanas que son especialmente intensas en cuanto al número de tweets producidos por unidad de tiempo y otras típicamente mortecinas. Entre las primeras suelen encontrarse los nudos de comunicación que implican tiempos de espera

considerables, la Estación de Cercanías y AVE de Atocha, por ejemplo, pero también espacios turísticos donde son frecuentes los *selfies*. Las zonas menos productivas suelen estar asociadas a zonas con un nivel socieconómico menor o nudos de comunicación con cortos tiempos de espera, como los intercambiadores de transporte. También es posible reconocer tipologías basadas en la frecuencia horaria en que se genera la información en cada espacio. Hay espacios más propensos a producir mayor información en días no laborables mientras que otros operan justo al contrario.

Estas conclusiones se obtienen a partir de medias estadísticas características que permiten trazar un perfil semanal y horario típico de cada zona. Estos perfiles son útiles además por otras razones. Es típico encontrar en cada ubicación desviaciones sobre esa media que suelen abarcar el espacio de unas horas a lo largo de uno o varios días. Esas pautas, digamos anómalas, de producción de información se pueden investigar hasta identificar una causa que ofrezca una explicación de las mismas. Así se han podido identificar patrones característicos asociados a ciertos tipos de eventos entre los que destacan los siguientes: eventos deportivos locales, eventos deportivos internacionales, fiestas populares, manifestaciones públicas y conciertos musicales de estrellas públicas. Esto no quiere decir que no existan más modelos, sino que hemos sido capaces de extraer al menos los patrones característicos de estos cinco tipos de eventos. Los usos de estas pautas en la identificación temprana de posibles situaciones de estrés en los transportes o comunicaciones urbanas están por ver, pero las perspectivas son prometedoras.

Un análisis mucho menos desarrollado que el anterior es el que se refiere a lo que podríamos denominar el tono vital de cada uno de los espacios cartografiados. Medir si la información producida en un tweet contiene un tono vital optimista o uno pesimista no es fácil en principio. Las objeciones son muchas y en general tienen que ver con la complejidad de la comunicación humana, muy superior a la que las herramientas que podemos diseñar en términos puramente computacionales es capaz de manejar. Pero no se pretende aquí generar aplicaciones eficientes dentro del terreno de la AI –*Artificial Intelligence*– capaces de captar el tono vital de un tweet, en realidad nos conformamos con algo mucho más modesto. Bastará, en realidad, con hacer una apuesta simple basada en la ocurrencia de ciertos términos característicos y el uso recurrente de ciertos emoticonos. Es obvio que una técnica tan primitiva ha de producir un alto porcentaje de falsos positivos así como un número tampoco despreciable de falsos negativos. Las pruebas realizadas usando la técnica indicada muestran, sin embargo, que la incidencia de errores de ese tipo no es tan alta como en principio cabría

esperar. Por desgracia, la comunicación humana, y mucho más cuando está tan fuertemente prototipada como en un tweet, no es tan novedosa y creativa como solemos imaginar. La recurrencia de patrones comunes es frecuente y puede ser reconocida por técnicas relativamente ingenuas como las propuestas tras un cierto análisis experto.

Esta técnica ha arrojado de momento la identificación de ciertas zonas con una intensidad emocional, tanto positiva como negativa, más marcada que otras en las que la información muestra un perfil más descriptivo. Destaca, por ejemplo, la intensidad emocional entorno a la Estación ferroviaria de Atocha, algo explicable por el número de encuentros y despedidas que se producen en ese espacio a lo largo del tiempo. La zona aledaña a los Juzgados de Instrucción situados en Plaza de Castilla ofrece un perfil emotivo también alto, pero negativo esta vez, debido, como cabe imaginar, a los comentarios de los fallos judiciales que muchos ciudadanos hacen en su entorno.

Aún mucho menos desarrollado se encuentra el estudio de los componentes morales de la información expresada en un tweet. Este análisis debería seguir unas pautas no muy distintas a las que funcionan en el análisis de los contenidos emocionales, pero quizá resulte optimista confiar en obtener algo de provecho. Por esta razón, se ha optado por elaborar mapas de ciertos términos previamente seleccionados para analizar la ocurrencia de tales expresiones en periodos de tiempo que abarcan la totalidad de la muestra. Esta técnica se ha empleado, por ejemplo, para estudiar el lenguaje homófobo en la ciudad de Madrid pero sin resultados que aún puedan darse por definitivos.

El último estudio que nos hemos planteado a raíz de la explotación de la información geolocalizada en Twitter introduce un componente dinámico que no está presente en los análisis precedentes. La idea guarda una cierta relación con la motivación que inspira la Memética de Dawkins o Blackmore propuesta durante el pasado siglo. Para la Memética los seres humanos pasamos a ser vistos como el medio que ciertas ideas tipo o memes tienen para propagarse, reproducirse y evolucionar a formas más complejas o simplemente más eficientes. Se produce una especie de inversión entre el agente y el producto que es la que nos llamó la atención en este punto de nuestro trabajo. Los seres humanos somos, como resulta evidente, los únicos productores de información verbal en nuestro hábitat. Somos nosotros quienes producimos las palabras que forman ciertas unidades de información con las que nos comunicamos. Al desplazarnos a lo largo de un espacio urbano vamos diseminando esas piezas de información y con ellas las palabras que las forman a lo largo del tiempo. El uso de Twitter permite medir la forma en que las palabras viajan por la ciudad a lo

largo del día empleándonos como medio de transporte. Este es el tipo de inversión entre agente y producto que asimila este estudio al estilo propio de la Memética. Somos conscientes de que se trata de una idea fuertemente especulativa que posiblemente no lleve a conclusión alguna. Las ciudades, sus barrios, no son entidades a las que quepa atribuir conductas intencionales como sería en este caso la producción de información. Pero también creo que debería tenerse en cuenta que los seres humanos no producimos nuestras piezas de información en un espacio ideal totalmente neutro con respeto a su capacidad para influir en nuestras emociones o incluso en nuestras ideas. Ahora disponemos de un recurso que nos permitiría comprobar ya sea el ruido de la ciudad o, lo que sería harto sorprendente, una cierta capacidad para hacer viajar ideas expresadas en palabras a lo largo de sus complejas redes de comunicación. Especulativo, sin duda, ciertamente improbable, pero apasionante en cualquier caso.

AGRADECIMIENTOS: Este artículo forma parte del proyecto «La construcción de agentes argumentativos en las prácticas del discurso público» financiado por la Secretaria de Estado de Investigación, Desarrollo e Innovación (MINECO), FFI2014-53164-P.

E. ALONSO: es Profesor Titular de Lógica y Filosofía de la Ciencia en la Facultad de Filosofía y Letras de la Universidad Autónoma de Madrid. Allí es responsable de la docencia de Lógica Formal y de distintas asignaturas relacionadas con el tema de la Sociedad de la Información. Su investigación se ha desarrollado en torno a la obra de Gödel y los avances en teoría de la computación y recientemente ha colaborado con María Manzano en diversos trabajos relacionados con la completitud de los sistemas formales y la obra de Leon Henkin. En 2007 creó el grupo “Sociedad Digital y Conocimiento” en la Universidad Autónoma de Madrid en el que ha dirigido trabajos relacionados con la Web semántica y más recientemente con la monitorización de las Redes Sociales. Ha publicado recientemente varios libros sobre esta temática: *La quimera del usuario* (Madrid, UAM/Abada, 2014), *El nuevo Leviatán. Una historia política de la red* (Madrid, Díaz Pons, 2015).