Capítulo 1. Minería de datos de los medios sociales: herramientas para recopilar datos de Twitter

Social media data mining: tools for collecting Twitter data

BENÍTEZ-ANDRADES, José-Alberto (1)

(1) Grupo de Investigación Salud, Bienestar y Sostenibilidad Sociosanitaria (SALBIS), Departamento de Ingeniería Eléctrica y de Sistemas y Automática, Universidad de León, Campus de Vegazana s/n, 24071, León, España, jbena@unileon.es.

Resumen

En la actualidad la minería de datos de los medios sociales suele estar centrada en recopilar información procedente de Twitter. Uno de los principales problemas en el análisis de redes sociales en estos casos es la adquisición de los datos en forma de grafo. La posibilidad de hacer esta labor «a mano», es impensable cuando se tratan redes de cientos o miles de nodos en constante comunicación entre ellos. Twitter, además, es un ejemplo muy claro de esta problemática. De cualquier *trending topic* se pueden generan cientos de tweets en una hora. Es necesario herramientas de adquisición de todos esos datos de una forma automatizada utilizando las posibilidades que la propia red social ofrece. Un ejemplo de herramienta que ofrece una obtención de datos, con limitaciones, es la herramienta Netlytic. No es la única, pero es sencilla de utilizar. En este documento se muestra un caso de uso de un proyecto recoplado con Netlytic y graficado mediante Gephi.

Palabras clave: Minería de datos, Análisis de redes sociales, Twitter, Gephi, Netlytic.

Abstract

Currently, social media data mining is often focused on collecting information from Twitter. One of the main problems in social network analysis in these cases is the acquisition of data in the form of a graph. The possibility of doing this task "by hand" is unthinkable when dealing with networks of hundreds or thousands of nodes in constant communication with each other. Twitter, moreover, is a very clear example of this problem. Any trending topic can generate hundreds of tweets in an hour. Tools are needed to acquire all this data in an automated way using the possibilities offered by the social network itself. An example of a tool that offers data acquisition, with limitations, is the Netlytic tool. It is not the only one, but it is simple to use. This document shows a use case of a project collected with Netlytic and plotted using Gephi.

Keywords: Data mining, Social Network Analysis, Twitter, Gephi, Netlytic.

1.1. Introducción

Según (Zafarani et al., 2014), los datos procedentes de los medios sociales son significativamente distintos a los obtenidos mediante otras vías más tradicionales dentro del contexto de la minería de datos. Los datos obtenidos mediante este tipo de medios suelen tener un tamaño muy elevado y, además, es común obtener datos generados por los usuarios siendo estos desestructurados y teniendo un alto nivel de ruido. Este hecho provoca que se deban realizar distintas formas de análisis de datos en combinación con el análisis de redes sociales procedente de las teorías sociales.

Es posible combinar las teorías sociales con distintos métodos computacionales de forma que se estudie cómo los individuos interaccionan y cómo se forman las comunidades. De este tipo de análisis surge lo que se conoce como minería de los medios sociales. Se llama así al proceso de representar, analizar y extraer patrones de datos de los medios sociales. Esta área de investigación provoca la necesidad de crear un nuevo perfil de científico de datos el cual debe tener un conocimiento elevado en las teorías sociales, pero también en las teorías computacionales. De esta forma será posible que resuelva problemas haciendo uso tanto de las teorías sociales como de las herramientas computacionales dentro del mundo de los medios sociales.

Dentro de los medios sociales surgen los conceptos de átomos sociales (individuos, usuarios, etc.); entidades (contenidos, sitios, redes, etc.); interacciones entre átomos y entidades (entre átomo y átomo, o entre átomo y entidad). El proceso completo de minería de datos se compone de tres tareas principales: (1) recolectar información sobre átomos y entidades; (2) medir sus interacciones; (3) descubrir patrones para entender el comportamiento humano.

Para poder realizar una minería de datos sobre la red social Twitter, existen distintas herramientas que permiten realizar esta tarea sin necesidad de ser expertos en ingeniería informática, por ejemplo: Netlytic, T-Hoarder, Socioviz o Tweepy. A lo largo de este capítulo se expondrá un caso práctico haciendo uso de la herramienta Netlytic combinada con la herramienta Gephi para realizar un análisis de redes sociales completo.

1.2. Retos de la minería de medios sociales y teorías sociales

En esta sección se exponen distintos problemas existentes a la hora de realizar una minería de datos en los medios sociales y, también, se enumeran tres teorías sociales necesarias para realizar un buen análisis de redes sociales sobre estos datos.

1.2.1. Problemática de la minería de los medios sociales

Existen diversos problemas dentro del ámbito de la minería de los medios sociales que deben tenerse en cuenta siempre que se vaya a hacer uso de esta metodología de la investigación.

En primer lugar, hay que destacar que la cantidad de datos existente en los medios sociales es muy elevada. Sin embargo, a la hora de estudiar un individuo en cuestión, es posible que no se disponga de mucha información sobre él. Por ello se recomienda utilizar distintos medios sociales, dimensiones,

fuentes y procedencia de datos que añadan o completen con más información a cada individuo de estudio.

En muchas ocasiones las APIs de los medios sociales no ofrecen una total libertad a la hora de obtener la información que posteriormente queremos analizar. Esto puede provocar que la información obtenida no sea representativa de la información sobre cada individuo o átomo social.

Además, cuando se trabaja con grandes cantidades datos, se suelen realizar tareas de preprocesamiento y eliminación de ruido que, en algunos casos, pueden eliminar información que quizá sí era representativa para el estudio en cuestión. El ruido que se obtiene en un conjunto de datos dependerá de la tarea que se esté realizando.

Por todo ello se puede decir que, en comparación con la minería de datos tradicional, los datos de los medios sociales son numerosos, relacionados, ruidosos, altamente desestructurados e incompletos y, por todo ello, difieren de los datos de la minería de datos tradicional (Tang et al., 2014).

El científico de datos encargado de analizar y minar los datos de los medios sociales, debe ser consciente de la naturaleza de los datos que va a analizar para tener una perspectiva completa. Solo de esta forma podrá realizar un proceso de filtrado de datos idóneo, realizará buenas hipótesis y obtendrá unos patrones de comportamiento útiles para obtener conclusiones.

1.2.2. Teorías en los medios sociales

Para poder explicar los distintos tipos de fenómenos sociales se utilizan distintas teorías sociales. De entre todas ellas, hay tres que son las más utilizadas: (1) teoría de correlación social, (2) teoría del balance y (3) teoría del estatus.

La teoría de correlación social está basada en dos procesos que suceden de forma alternativa, la homofilia social y la influencia. La homofilia social es lo que se conoce como “el amor a lo mismo”. Se trata de la tendencia de los individuos a asociarse y relacionarse con otros similares. En este caso, los individuos suelen compartir características comunes (valores, educación, creencias, etc.). A este proceso también se le conoce como asortatividad. Por otro lado, la influencia es aquella que representa la acción de que los individuos suelen seguir el comportamiento de las personas más próximas (La Fond & Neville, 2010; Neville & Jensen, 2007).

Por otro lado, la teoría del balance o del equilibrio social se basa en el equilibrio o desequilibrio de la relación de confianza en las relaciones entre dos o tres personas. De esta teoría se obtiene que los miembros de un mismo grupo tienen afinidad entre ellos y los miembros de distintos grupos sienten aversión. Mediante esta teoría se suele enunciar que “los amigos de mis amigos, son mis amigos” o que “los enemigos de mis amigos son mis enemigos”.

Por último, la teoría del estatus es aquella mediante la cual es posible describir la posición social que ocupa un individuo dentro de una sociedad o en un grupo social de personas.

1.3. Recolección de datos mediante la herramienta Netlytic

Para extraer redes de Twitter, se va a utilizar la herramienta online Netlytic disponible en la web www.netlytic.org. Esta herramienta permite al usuario conectar su cuenta de Twitter y extraer hasta 1.000 tuits realizando la búsqueda que desee (por usuario, hasthag o cualquier término).

1.3.1. Creación del proyecto de importación de datos

El primer paso es registrarnos en la plataforma. Una vez nos hemos registrado, disponemos gratuitamente de tres importaciones de datos posibles desde Twitter.

Para obtener un nuevo conjunto de datos, hacemos click sobre el menú *New Dataset*, indicamos un nombre para nuestro conjunto de datos en el campo *Dataset Name* y escribimos el término o términos a buscar en el campo *Search Keywords.* Dentro de este campo es posible usar operadores booleanos como “AND” u “OR” para realizar una búsqueda avanzada. En la ilustración 1 se puede observar un ejemplo de esta pantalla.

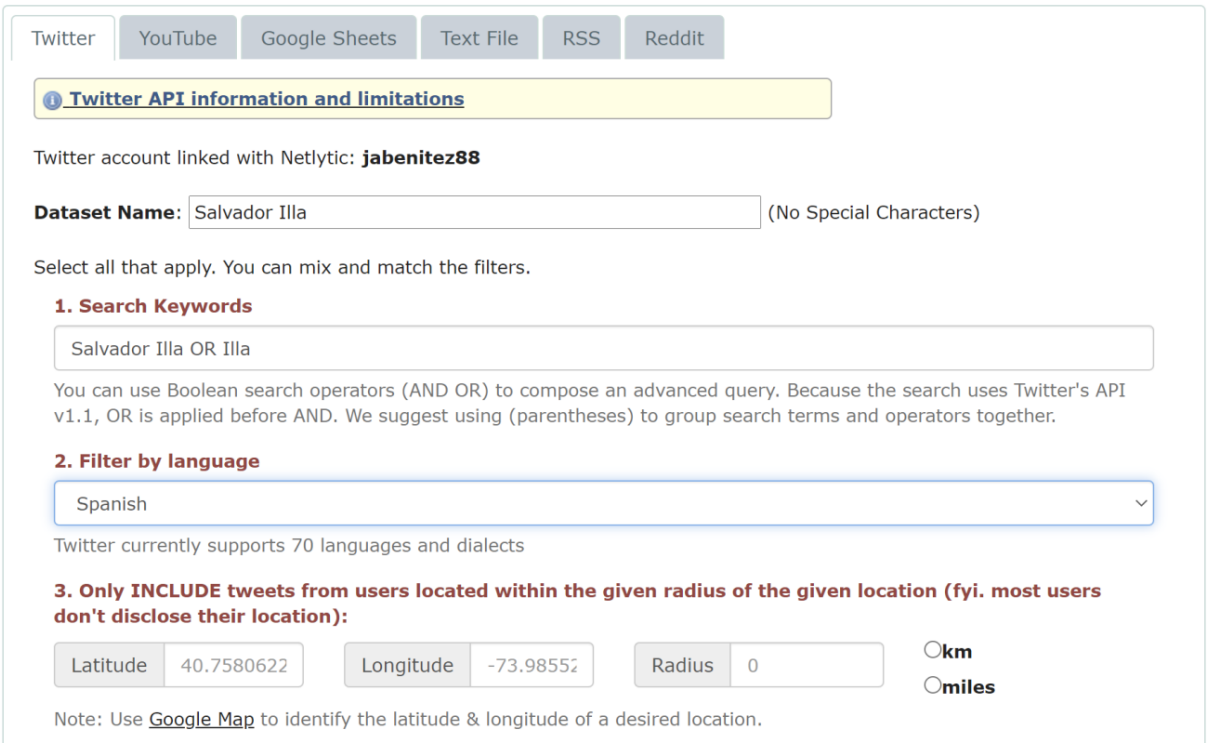

Ilustración 1. Pantalla inicial para crear un nuevo conjunto de datos desde Netlytic.org

Además de los dos parámetros indicados anteriormente, desde esta interfaz es posible personalizar los siguientes parámetros de búsqueda:

- Idioma: mediante este campo es posible seleccionar el idioma de los tuits que se recopilan.
- Tuits localizados según geolocalización: en este campo el usuario puede seleccionar una latitud, una longitud y un radio en Km o millas para que los datos recopilados se encuentren en la zona indicada.

- Criterios de inclusión: esta herramienta ofrece la oportunidad de incluir tuits que contengan retuits, respuestas, imágenes, vídeos, enlaces o noticias si se desea.
- Criterios de exclusión: también permite excluir por alguno de los criterios anteriormente mencionados.
- Número mínimo de retuits: se puede determinar una cifra mínima de retuits que deben tener los tuits que se recopilan.
- Número mínimo de “me gusta”: se puede determinar una cifra mínima de “me gusta” que deben tener los tuits que se recopilan.
- Tuits dirigidos a un usuario: es posible indicar que los tuits vayan dirigidos a un usuario determinado.
- Tuits realizados por un usuario: es posible indicar que los tuits estén escritos por un usuario determinado.

Una vez determinados todos los parámetros, se debe pulsar el botón “import” y el proyecto entrará en una cola de tareas. Al finalizar, el usuario podrá ver los resultados obtenidos mediante la búsqueda personalizada indicada.

1.3.2. Análisis de los datos recolectados

Tras finalizar el proceso de importación, se observa una búsqueda realizada tal y como se muestra en la ilustración 2.

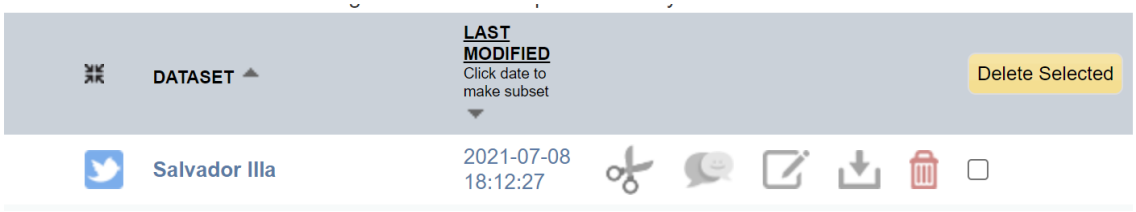

Ilustración 2. Pantalla para gestionar los conjuntos de datos creados con Netlytic.org

Desde esta pantalla es posible acceder a los datos haciendo click sobre el título del *dataset*. También es posible editar la búsqueda, eliminarla o descargar los datos directamente.

Tras hacer click en el título del *dataset* es posible ver una previsualización de los datos recopilados. También es posible descargar los datos en formato hoja de cálculo o CSV tal y como se muestra en la ilustración 3.

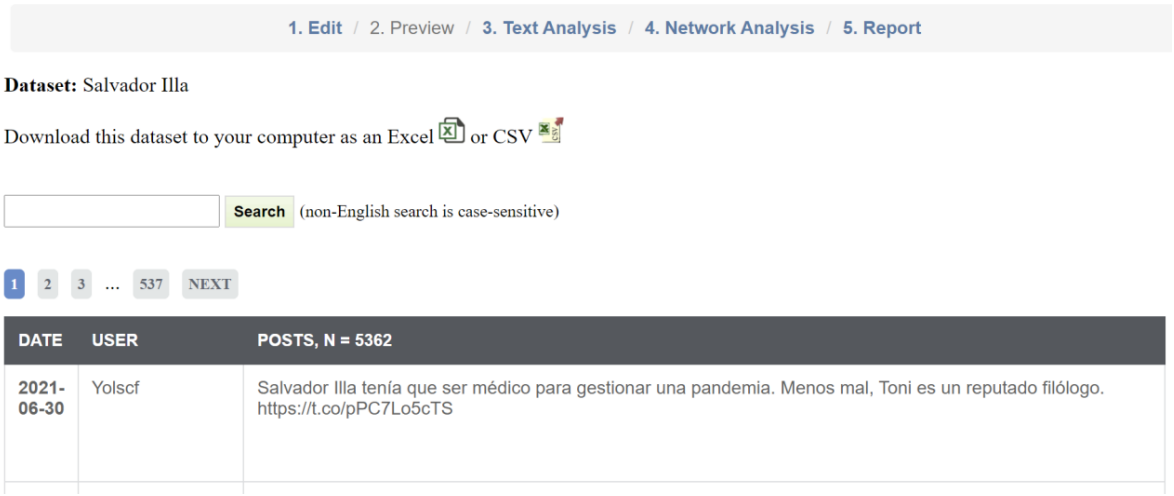

Ilustración 3. Pantalla de previsualización de datos importados mediante Netlytic.org

En el siguiente paso el usuario puede realizar un análisis de los textos como se aprecia en la ilustración 4.

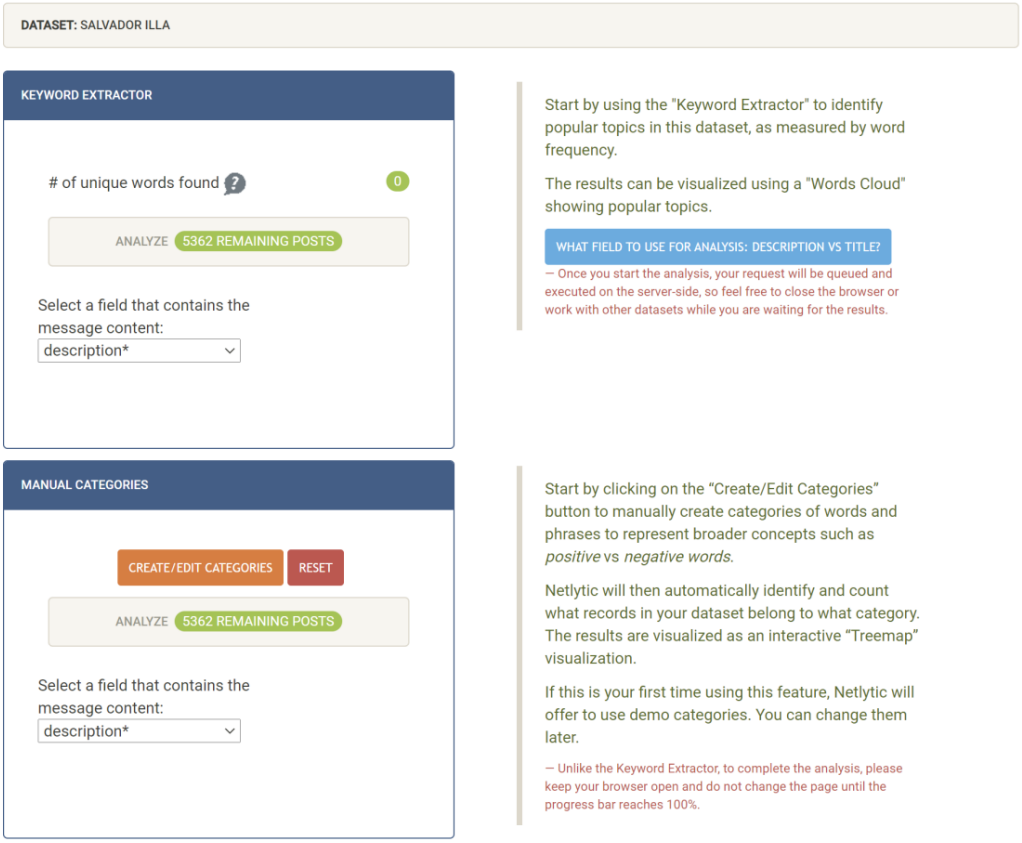

Ilustración 4. Pantalla de extracción de palabras clave y categorización de datos importados en Netlytic.org

Seguidamente, se pueden analizar las redes seleccionando qué consideramos que es un enlace de entre las siguientes opciones:

- Usuario A responde a usuario B.
- Usuario A cita a usuario B.
- Usuario A retuitea a usuario B.
- Usuario A menciona a usuario B.

Estas opciones se observan en la ilustración 5:

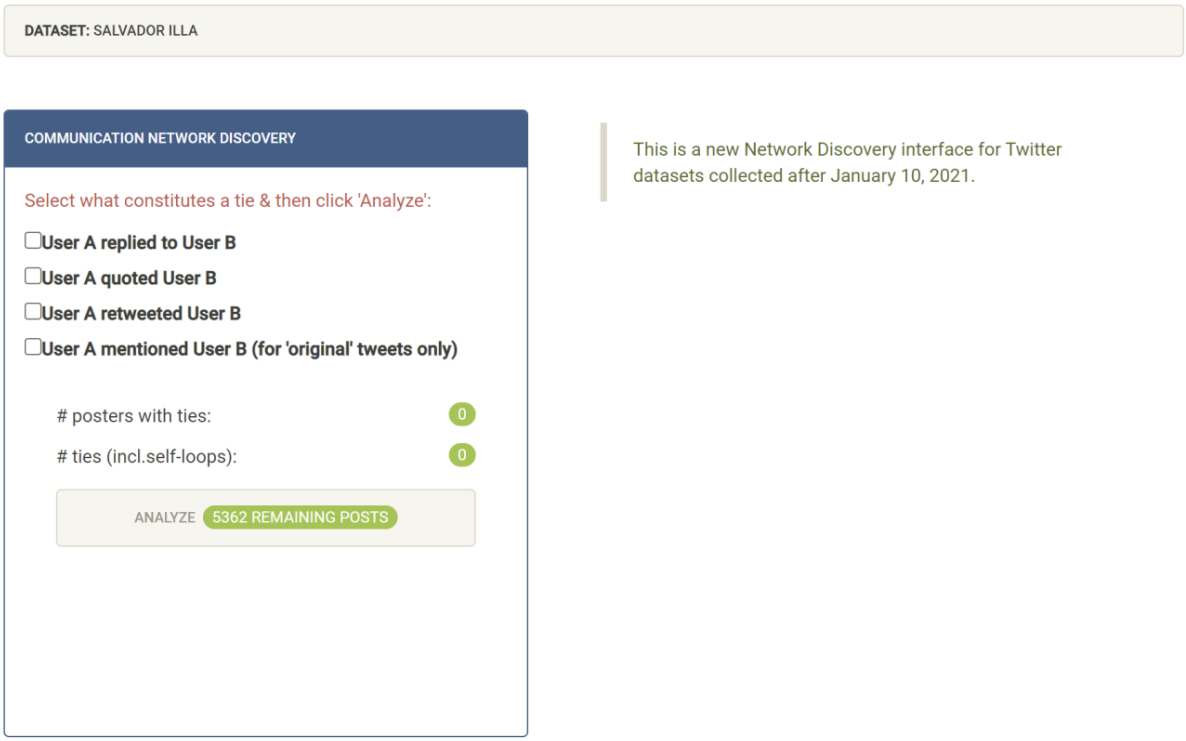

Ilustración 5. Pantalla para seleccionar qué tipo de relaciones generan una red en Netlytic.org

Tras haber seleccionado el patrón de enlace, la plataforma ofrece la opción de exportar la red en distintos formatos:

- Lista de aristas.
- Fichero en formato .gefx para abrir mediante Gephi.
- Fichero en formato GraphML.

Una vez descargada la red en formato .gefx es posible abrirla mediante Gephi y graficar haciendo uso de las distintas funcionalidades que ofrece Gephi.

1.4. Conclusiones finales e información complementaria

Gracias a la herramienta Netlytic es posible realizar un proceso de recolección de datos de forma sencilla a través de la plataforma Twitter configurando distintos datos para la misma como, por ejemplo, los términos de búsqueda, el idioma o los criterios de inclusión y exclusión basados en distintas métricas de Twitter. Tras esta recolección de datos, y de formas sencilla, es posible graficar la red y calcular las

distintas métricas de un análisis de redes sociales como, por ejemplo, la densidad, el diámetro o la centralidad de la red gracias a la herramienta Gephi y a la compatibilidad que ofrece Netlytic con esta. Para aprender a utilizar las distintas funcionalidades de Gephi se recomienda al lector que haga uso de la guía oficial de Gephi disponible en su sitio web https://gephi.org/users/.

1.5. Referencias

La Fond, T., & Neville, J. (2010). Randomization tests for distinguishing social influence and homophily effects. *Proceedings of the 19th international conference on World wide web*, 601-610. https://doi.org/10.1145/1772690.1772752

Neville, J., & Jensen, D. (2007). Relational Dependency Networks. *The Journal of Machine Learning Research*, *8*, 653-692.

Tang, J., Chang, Y., & Liu, H. (2014). Mining social media with social theories. *ACM SIGKDD Explorations Newsletter*, *15*(2), 20-29. https://doi.org/10.1145/2641190.2641195

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social Media Mining: An Introduction*. Cambridge University Press.