Minería de datos y aprendizaje profundo (DeepLearning) en Twitter con Gephi

M. en I.S.C. Leonardo M. MorenoV.*, Dr. F. Jacob Ávila C.*, Dr. Adolfo Meléndez Ramírez*

Resumen

El propósito de este artículo, es mostrar una perspectiva que permita visualizar de forma grafica la importancia que tiene la minería de datos mediante un software para el análisis de grafos como lo es Gephi; mediante un ejemplo, se puede observar el análisis de sentimientos y las reacciones que tienen los usuarios de la red social Twitter ante un comentario del presidente Enrique Peña Nieto.

Acerca de los autores...

* Profesor-investigador adscrito a la División de Informática del Tecnológico de Estudios Superiores de Ecatepec. lmoreno@tese.edu.mx, fjacobavila@tese.edu.mx, Adolfo_melendez@tese.edu.mx.

Abstract

The purpose of this article is to show a perspective that allows to visualize graphically the importance of data mining through software for the analysis of graphs such as Gephi; by means of an example, you can see the analysis of feelings and the reactions that users of the social network have Twitter to a comment by President Enrique Peña Nieto.

IndexTerms—Data Mining, Twitter, Gephi, Análisis de Sentimientos, Análisis de Grafos.

Introducción

El gran avance tecnológico en los últimos años ha permitido la incursión del Internet a nuestras vidas, lo cual ha traído consigo grandes beneficios para la humanidad, ya que facilita la comunicación alrededor del mundo, el acceso a la información, la forma de los negocios como las compras en línea, pagos electrónicos, las redes sociales, entre muchas otras actividades que cotidianamente utilizamos y dependen totalmente del internet. En consecuencia, la generación de datos se ha incrementado inimaginablemente, por ejemplo, según Chen, Shiwen, & Yunhao (2014), compañías como Google procesa miles de peta bytes (PB) de datos; Facebook genera más de 10 (PB) cada mes, y compañías chinas como Tabao y Alibaba generan y procesan datos en un volumen de 10 (PB), es decir, producen datos de decenas de terabytes, debido a un factor importante, el cual está relacionado con los últimos avances de las (TIC), las tecnologías emergentes como el Internet de las cosas, la nube, los smartphones; el fácil acceso a ellos aumenta la generación de datos, que pueden ser representados en muchas formas, no solo letras, sino en audio, video, sensores, sistemas GPS, medidores de temperatura, etcétera, por lo que todo aquello que se conecte a Internet y comparta una base de datos, aportará información en dimensión Big Data.

Como referencia, según Leboeuf (2016), en el año 2016 cada minuto se reprodujeron 2.78 millones de videos en YouTube, se realizaron 701,389 millones de inicios de sesión en Facebook, se enviaron 20.8 millones de mensajes en WhatsApp, se realizaron 2.4 millones de búsquedas en Google y se publicaron 347,222 nuevos tweets, esto solo para dar una perspectiva de la cantidad de datos generada, por lo que cada vez se necesita una mejor infraestructura de TI para poder soportar, almacenar y procesar esta información en tiempo real, razón por la cual se ha adoptado el término de Big Data, principalmente para describir el conjunto masivo de datos, pero difiriendo de almacenes de datos comunes, por su enorme cantidad de información y su difícil proceso de análisis.

Big Data es un concepto abstracto que es conceptualizado de distinta manera. En 2010, Apache Hadoop definió Big Data como “conjuntos de datos que no pueden ser capturados, administrados y producidos por ordenadores generales”. En mayo de 2011, McKinsey & Company, una agencia de consultoría global anunció el Big Data como la próxima frontera para la innovación, la competencia y la productividad. Big Data se entiende como los conjuntos de datos que no pudieron ser adquiridos, almacenados, y manejados por el software clásico de una base de datos. Esta definición incluye dos connotaciones: primero, los volúmenes de conjunto de datos, que se

ajustan al estándar de que los grandes datos están cambiando, y puede crecer con el tiempo o con los avances tecnológicos. Segundo, los volúmenes de conjuntos de datos que se ajustan al estándar de datos en diferentes aplicaciones en general, los datos grandes generalmente varían por la cantidad de Terabytes o Petabytes (Chen, Shiwen, & Yunhao , 2014).

Ejemplo de aplicación con Gephi

Recolección de Datos

Siendo Twitter una de las plataformas sociales que más datos genera por minuto, estimada en alrededor de 350,000 nuevos tweets cada minuto, es muy importante conocer algunas de las formas de minería de datos y herramientas de ciencia de datos que pueden ser aplicadas para poder explotar esta enorme fuente de información, la cual podría ser empleada, por ejemplo, para predecir los comportamientos de mercados financieros, como en el artículo publicado por Bollen, Mao, & Zeng (2010); la perfilación y predicción de la personalidad a través de los tweets publicados por un individuo (Quercia, 2011), y hasta la detección de epidemias de influenza mediante los mensajes y posts en Twitter (Culotta, 2010), lo que brinda un amplio panorama de todo lo que podemos obtener y realizar con solo analizar la información que esta red social nos aporta.

Para obtener los datos generados por Twitter, esta red social nos ofrece tres posibilidades: APIs Streaming API, REST API y Search API, aplicables a necesidades diferentes. El Streaming API proporciona un conjunto de tweets en casi tiempo real. Permite establecer una conexión permanente del usuario con los servidores de Twitter y mediante una petición http, se recibe un flujo continuo de tweets en formato json. Se puede obtener una muestra aleatoria (statuses/sample), un filtrado (statuses/filter) por palabrasclave o por usuarios.

El Search API permite obtener los tweets con determinada profundidad en un tiempo de siete días, que se ajustan a la query solicitada por el usuario. También es posible filtrar por cliente utilizado, lenguaje y localización. No requiere autenticación y los tweets se obtienen en formato json o atom.ede

El REST API ofrece a los desarrolladores el acceso al core de los datos de Twitter. Todas las operaciones que se pueden hacer vía web es posible realizarlas desde el API. Dependiendo de la operación requiere o no autenticación, con el mismo criterio que en el acceso web. Soporta los formatos: xml, json, rss, atom.

También es factible acceder a los datos de Twitter mediante proveedores oficiales de datos (Cured Data); dentro de estos proveedores tenemos a Datasifthttp://datasift.com/,GNIP (comprada por Twitter en abril del 2014), y Topsy (adquirida por Apple en diciembre del 2013); aunque estas dos herramientas se encuentran descontinuadas actualmente, se puede acceder a los datos de Twitter mediante otras herramientas que nos permitan realizar análisis de la información (Curated data + analytics), dentro de éstos destacan: Brandwatch, Hootsuite, y de igual formase puede verificar a los diferentes partners de twitter en: https://partners.twitter.com/.

Para ejemplificar el primer elemento, nos registramos con nuestro perfil en https://developer.twitter.com/ y una aplicación en https://apps.twitter.com/ esto para tener un control de las aplicaciones que están accediendo a la plataforma y un mejor manejo de la información mediante Gephi, una plataforma interactiva de código abierto (open source) para la visualización y exploración de todo tipo de redes y sistemas complejos con gráficos dinámicos y jerárquicos. En la Figura 1, podemos observar el registro de una aplicación en Twitter, en donde la red social nos ofrece los elementos esenciales

para trabajar con la información, como la Consumer Key (API Key), el Consumer Secret (API Secret), su Access Token y por último el Access Token Secret, sin los cuales no podríamos acceder a los tweets en la respectiva búsqueda.

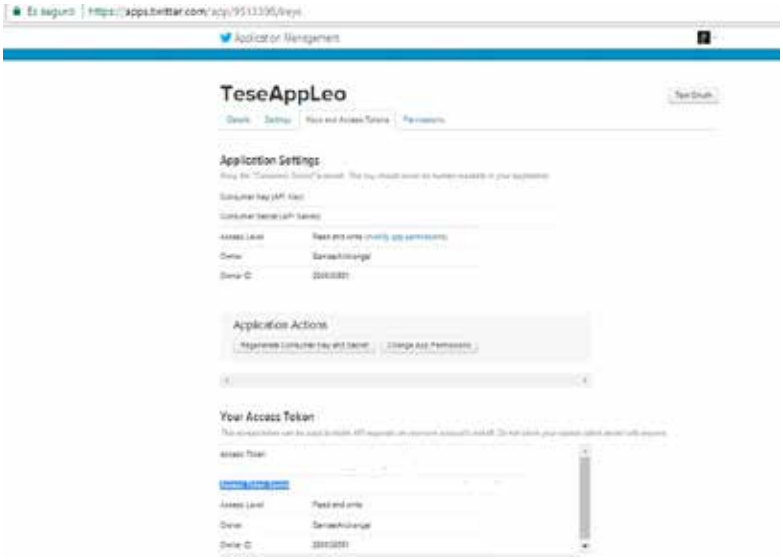

Figura 1

Application Management en Twitter. Fuente: Elaboración propia

Para esta aplicación registrada, utilizaremos Gephi, por lo que ocuparemos un plugin de la misma plataforma. En la Figura 2 podemos ver claramente la pantalla de inicio de Gephi, y la opción para descargar los plugin disponibles de esta herramienta. En este caso utilizaremos Twitter Streaming Importer, el cual permite conectarnos mediante las credenciales de nuestra aplicación generada anteriormente a los datos de Twitter para ser analizados en Gephi.

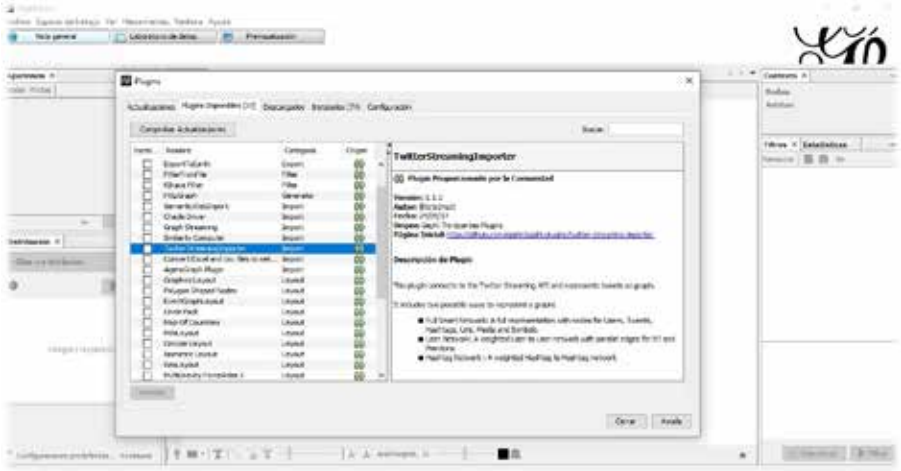

Figura 2

En Gephi descarga del plugin Twitter Streaming Importer. Fuente: Elaboración propia

Una vez realizada la instalación del plugin correspondiente, debemos configurarlo proporcionando las credenciales de la app de Twitter que previamente configuramos; esto puede verse claramente representado en la Figura 3, donde una vez accesados los datos al plugin, procedemos a realizar la conexión del mismo a la fuente de datos de Twitter, a la cual se puede acceder de dos formas diferentes:

1. Mediante la búsqueda de palabras o hashtags dentro de Twitter.

2. Siguiendo el comportamiento de un usuario.

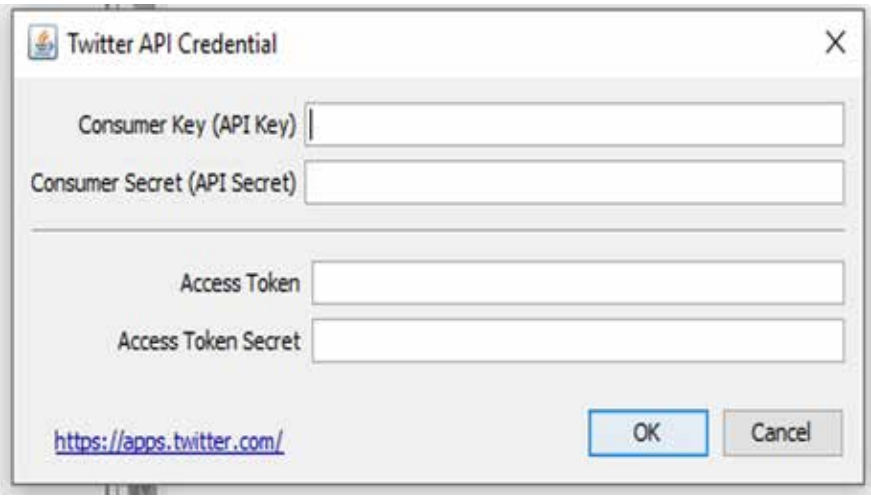

Figura 3

Acceso de credenciales de Twitter en Gephi. Fuente: Elaboración propia.

Es importante señalar que para realizar estas búsquedas existen tres lógicas de aplicación en Gephi.

1. Full Twitter Network. Nos permite aplicar nuestra lógica de búsqueda en toda la red de Twitter (Usuarios, Tweets, Hastags, URL, Media, Símbolos, etcétera)

2. Hashtag Network. Aplica nuestra lógica de búsqueda en la red de hashtags de Twitter.

3. User Network. Usa la lógica de búsqueda para solo acceder a los datos de nuestro panel usuario menciones, interacciones, etcétera.

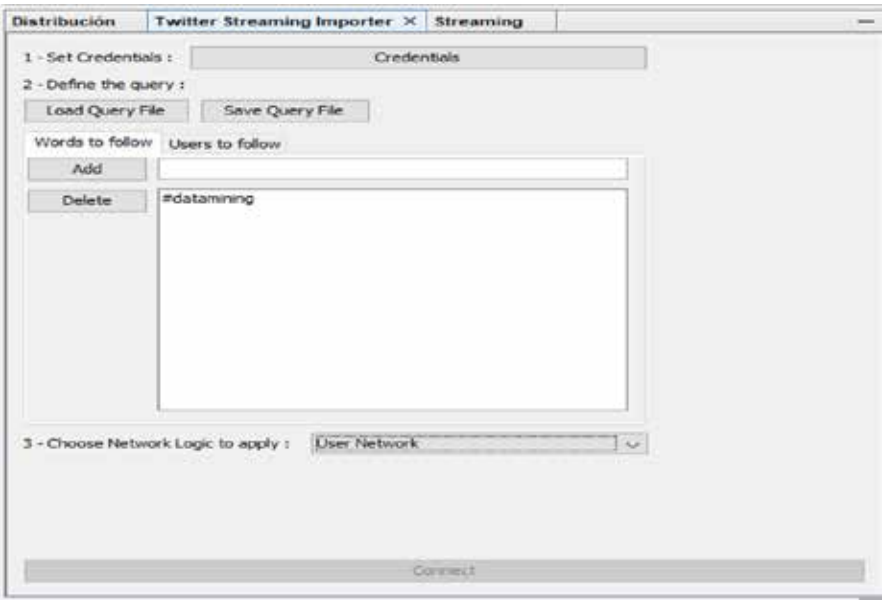

Figura 4

Lógica de búsqueda de Twitter Streaming Importer. Fuente: Elaboración propia.

Los botones de Load Query File y saber Query File, nos permiten cargar o guardar nuestros querys de búsqueda, respectivamente. Una vez que aceptamos la configuración, podremos visualizar en nuestro panel del grafo, la construcción en tiempo real de los datos que se van obteniendo de nuestra lógica de búsqueda en Twitter. Esto puede ser visualizado en la Figura 5 que se muestra a continuación.

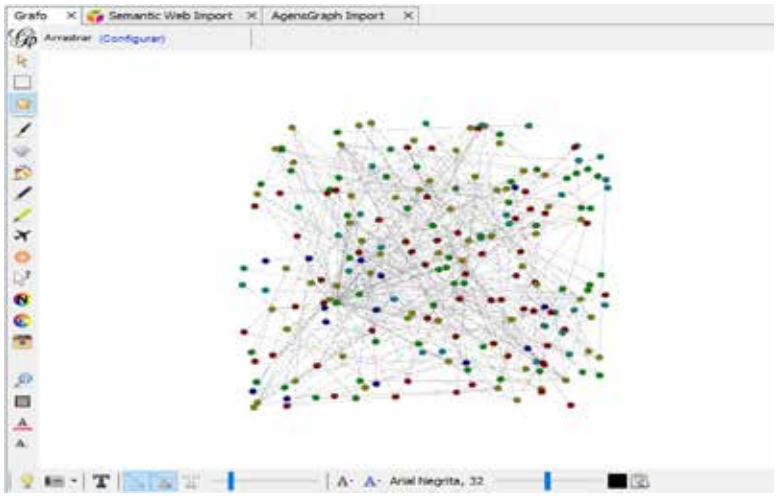

Figura 5

Grafo resultante de la búsqueda aplicando la lógica establecida por el usuario. Fuente: Elaboración propia.

Como podemos observar, el grafo resultante no representa sentido alguno para nosotros; es importante indicar que en los grafos resultantes, todos están integrados por un nodo (nodes) y las conexiones entre éstos (edges); cada nodo tiene un grado, que no es más que el número de aristas que inciden en él. Si bien entender esto es uno de los elementos básicos para Gephi, también lo son los algoritmos de distribución, que no son más que la forma como se van a visualizar los datos y van a ser distribuidos en el grafo. Gephi nos proporciona diferentes algoritmos para la visualización de la información, pero ¿cómo escoger cuál es el más adecuado? Todo va a depender del énfasis que se le dé a lo que deseamos visualizar.

Figura 6

Tipos de Layouts. Fuente: http://gephi.org

La tabla de datos obtenida por la lógica de búsqueda que establecimos, se encuentra almacenada en la pestaña denominada *Laboratorio de datos*. Como se muestra en la Figura 7, este laboratorio de datos nos permite trabajar con los datos crudos obtenidos de Twitter mediante nuestra lógica de búsqueda, y también nos proporciona una herramienta para descargar toda nuestra tabla en un archivo csv, para poder utilizar tales datos en cualquier otra herramienta de análisis de datos.

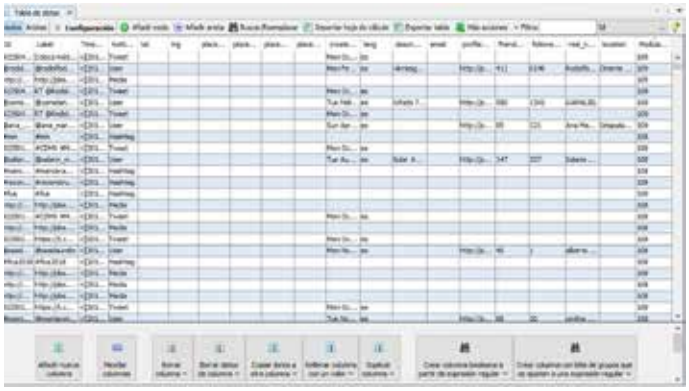

Figura 7

Laboratorio de datos (tabla de datos) de la lógica de búsqueda en Twitter. Fuente: Elaboración propia.

Análisis

A manera de ejemplo, realizamos una búsqueda por usuario, en este caso, a la persona identificada como EPN, a fin de obtener a todos los usuarios que están haciendo mención de él y poder efectuar un análisis para su grafo generado en esta consulta.

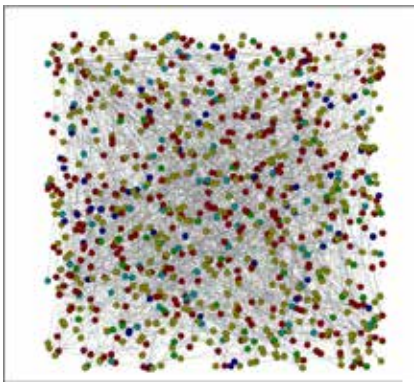

Figura 8

Grafo generado por la consulta al usuario EPN. Fuente: Elaboración propia.

Es importante destacar que se debe aplicar la modularidad, la cual es una medida de la estructura de los grafos, midiendo la fuerza de la división de una red en comunidades o módulos, que pueden definirse como conjuntos de nodos más densamente conectadosentre ellos, que con el resto de la red. Comúnmente se utiliza en métodos de optimización para detectar la estructura de una comunidad de nodos en una red o grafo; sin embargo, no es recomendado para la detección de módulos o comunidades pequeñas, debido a que con frecuencia presenta problemas llamados resolución. La modularidad está definida por la siguiente fórmula:

Sea un grafo con **n** nodos y **m** aristas, de tal manera que el cálculo de la modularidad **Q** propuesta por (Newman, 2006) sería de la siguiente manera:

$$Q = \sum_{ij}\left(\frac{A_{ij}}{2m} - \frac{k_i * k_j}{(2m)(2m)}\right)\delta(c_i, x_j)$$

Donde:

A_{ij}= Peso de la arista o Edge "i" al nodo destino "j"

K_i= Grado del nodo "i"

K_j= Grado del nodo "j"

$ln_= \sum_i^n K_i$= 2m número total de aristas o edges de entrada o salida

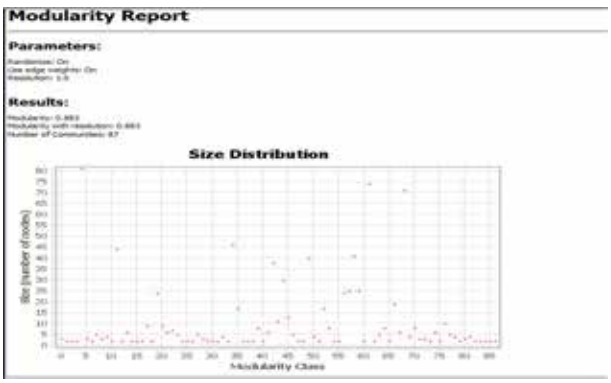

Figura 9

Reporte de modularidad en Gephi. Fuente: Elaboración propia.

Se podría establecer la posición en esos rangos, para lo cual se usa Page Rank, aplicado por lo general para formar las posiciones de las páginas de los buscadores, pero puede ser usado para medir su influencia en las redes sociales, en otras palabras, para encontrar los nodos importantes de cualquier grafo.

Si bien Page Rank fue creado para asignar un número de influencia a cada página web en un motor de búsqueda, es posible emplearlo en cualquier grafo dirigido a establecer la influencia de cada nodo en él.

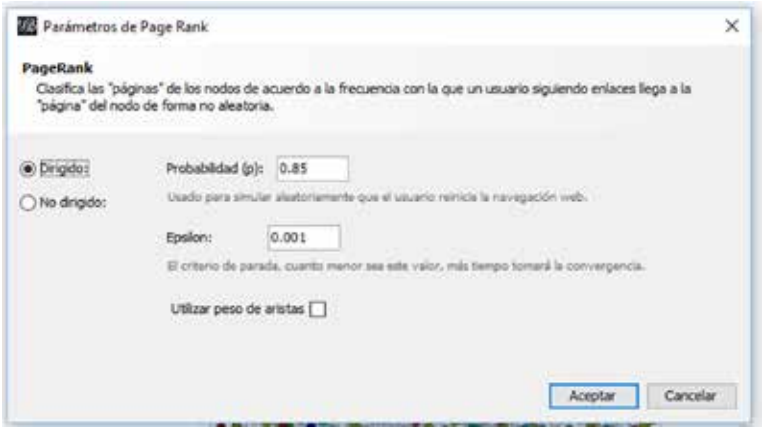

Figura 10

Aplicación del algoritmo Page Rank para el grafo generado.
Fuente: Elaboración propia.

Figura 11

Reporte de la aplicación del algoritmo Page Rank.

A continuación usaremos el algoritmo de distribución Force Atlas 2, el cual se tiene como característica contar con un modelo lineal-lineal, lo cual significa que la atracción y la repulsión es proporcional a la distancia entre los nodos, el cual será utilizado para poder dispersar los nodos y grupos, a fin de dar espacio a los nodos de mayor tamaño. A continuación se presenta la fórmula correspondiente al algoritmo de distribución(Jacomy, Venturini, Heymann, & Bastian, 2014):

$$F_r(n_1 n_2) = k_r \frac{(\deg(n_1)+1)(\deg(n_2)+1)}{d(n_1, n_2)}$$

Donde:

F_r = Fuerza de repulsión, que es proporcional al producto de los grados más 1 de los 2 nodos.

K_r= Es definido por la configuración.

Aplicando el algoritmo de distribución Force Atlas 2, nos arroja el siguiente Grafo:

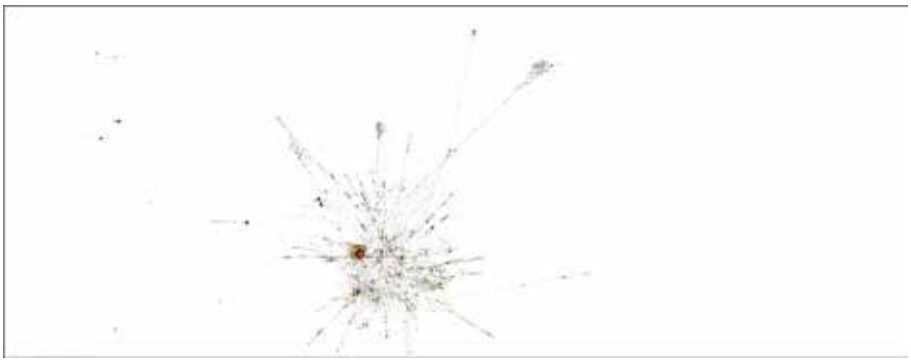

Figura 12

Grafo generado aplicando Force Atlas 2.
Fuente: Elaboración propia.

Observando cuidadosamente el grafo generado, al cual se le aplicó modularidad y el algoritmo Page Rank para establecer con base en estos algoritmos el tamaño de los nodo, se puede ver que en la imagen existen diferentes grupos de personas que mencionan al usuario EPN, dentro de ellos destaca uno con bastante influencia, que para esta consulta corresponde a una declaración realizada por el usuario que generó una reacción negativa en Twitter, como se aprecia en la Figura 13.

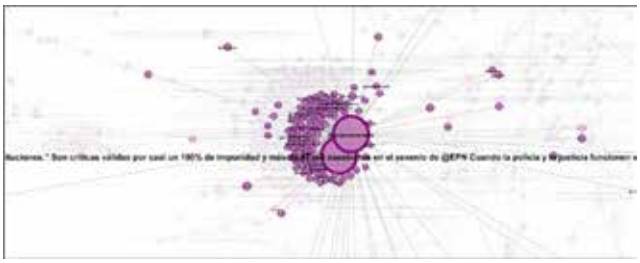

Figura 13

Nodo correspondiente a un tweet de un periodista sobre el comentario del usuario EPN.

Si ponemos atención en el grafo, el mismo generó comunidades, que corresponden a los tweets de periodistas famosos sobre el usuario EPN, y los usuarios que ha retweeteado a estos periodistas o los han mencionado junto con el usuario EPN, por ejemplo tenemos a Jorge Ramos (el que más influencia tiene en este grafo), Denisse Dresser, Julio Astillero, y Genaro Villamil, pero se puede observar el gran descontento que existe y las reacciones de los usuarios con base en un comentario realizado por el usuario EPN así como el gran impacto e influencia que ejercen los periodistas en este país.

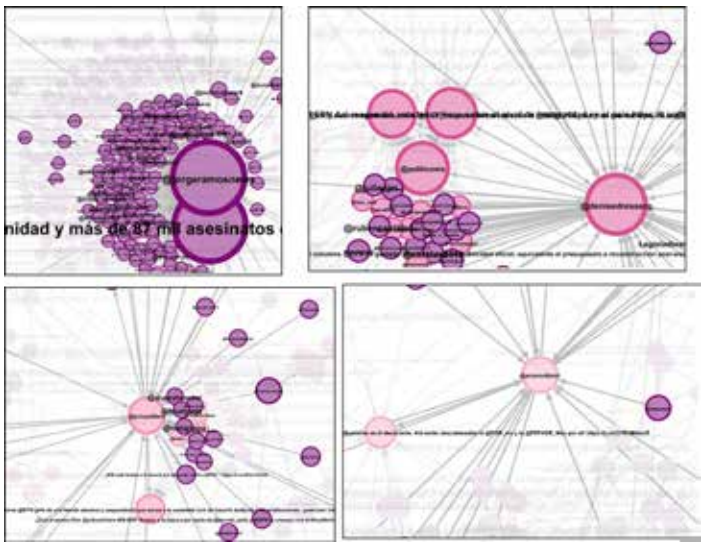

Figura 14

Periodistas y comunidades de usuarios que los retweetean.
Fuente: Elaboración propia.

Conclusiones

Siendo las redes sociales, en especial Twitter, una de las fuentes de información más grandes que existen en la actualidad (Big Data), es de suma importancia entender y desarrollar elementos que nos permitan analizar y comprender cómo fluye y se maneja toda la información generada casi de forma inmediata, ello con el fin de observar mediante herramientas de análisis, como lo es Gephi, las diferentes comunidades de usuarios de esta plataforma, la interacción que existe entre ellos, cómo fluye la información y cómo es usada por los diferentes medios de información, mencionando que el estudio de caso fue hecho para un análisis de sentimientos en reacción a un comentario que realizó EPN en días anteriores, donde se pudo observar cómo existen comunidades que giran en torno a periodistas que generan sentimientos negativos, ejemplificando claramente la importancia del uso de herramientas que nos permitan analizar los sentimientos en las redes sociales, no solo para detectar oportunidades de negocio, sino también para conocer el impacto que tiene un comentario en la imagen pública de una persona.

Referencias

Bollen, J., Mao, H., &Zeng, X.J. (2010). Twitter Mood Predicts the Stock Market. ArXiv.Org, cs. CE(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. Retrieved October 18, 2017, from http://snap.stanford.edu/soma2010/papers/soma2010_16.pdf

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoSONE, 9(6), e98679. https://doi.org/10.1371/journal.pone.0098679

Leboeuf, K. (2016). 2016 Update: What Happens in One Internet Minute? - Excelacom, Inc. Retrieved October 18, 2017, from http://www.excelacom.com/resources/blog/2016-update-what-happens-in-one-internet-minute

Millman, K.J., & Aivazis, M. (2011). Python for scientists and engineers. *Computing in Science and Engineering*. https://doi.org/10.1109/MCSE.2011.36

Newman, M.E.J. (2006). Modularity and community structure in networks. Proc. Natl. Acad. Sci. of the U.S.A., 103(23), 8577–8585. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1482622/pdf/zpq8577.pdf

Quercia, D. (2011). Our Twitter Profiles, Our Selves: Predicting Personality from Twitter, 180–185. https://doi.org/10.1109/PASSAT/SocialCom.2011.26