ISSN 1992-6510
e-ISSN 2520-9299

REALIDAD y reflexión

Reality and Reflection

47

Año 18, N° 47, San Salvador, El Salvador, Centroamérica. Revista Semestral Enero-Junio 2018

Year 18, N° 47, San Salvador, El Salvador, Central America. Semestral Journal January-June 2018

Pescando información en el océano de datos de Twitter

Fishing for information in the ocean of Twitter data

James Edward Humberstone Morales
Ingeniero en Ciencias de la Computación y Maestro en Informática Aplicada a Redes por la Universidad Francisco Gavidia
Docente de la Facultad de Ingeniería y Sistemas
Investigador del Laboratorio de Nanotecnología y Centro de Modelaje Matemático en la Universidad Francisco Gavidia
jhumberstone@ufg.edu.sv

Recibido: 19 de marzo de 2018
Aprobado: 7 de junio de 2018
DOI: http://dx.doi.org/10.5377/ryr.v0i47.6273

RESUMEN

Las redes sociales son portales que ofrecen plataformas donde las personas comparten libremente sus opiniones y sentimientos sobre los temas de interés que acontecen en una sociedad. En la presente investigación se presenta un procedimiento para pescar información del océano de datos de la red social Twitter. Para ello se analiza una muestra de 1,000 tuits generados por usuarios del municipio de San Salvador, El Salvador, acerca de un tema mediático.

Palabras clave: minería de texto, análisis de tuits, API de Twitter, El Salvador.

ABSTRACT

Social networks are portals that offer platforms where people freely share their opinions and emotions about the topic of interest that happen in a society. This research presents a procedure for fishing information from data ocean the social network Twitter. For this purpose, a sample of 1,000 user-generated tweets from San Salvador, El Salvador, about a media topic.

Keywords: *text mining, tweet analysis, Twitter API, El Salvador.*

Introducción

Nos encontramos en la era de la información donde los usuarios tienen un rol activo en las comunicaciones, colaboran en iniciativas, generan información cada segundo y la distribuyen rápidamente por las diferentes redes sociales.

Las redes sociales se han convertido en parte esencial de la rutina de las personas, ya que facilitan a los usuarios poder expresar sus opiniones y sentimientos. En nuestro país nueve de cada diez personas hacen uso de las redes sociales a diario (Analitika, 2015, p. 28).

Las empresas aprovechan las redes sociales para la difusión de noticias y publicidad, que es altamente aceptada por los salvadoreños (Analitika, 2015, p. 40). Gracias a los comentarios de las personas, las empresas pueden conocer el perfil de los clientes satisfechos e insatisfechos con los productos y/o servicios que ofrecen, así como las razones más habituales de su satisfacción o insatisfacción. Prácticamente, las empresas tienen una oportunidad para generar ventajas competitivas cuando obtienen y analizan las opiniones, quejas y sugerencias de sus usuarios.

En el ámbito político, Barack Obama, expresidente de los Estados Unidos, recurrió a las redes sociales para mejorar su posición respecto a la población durante su campaña (CulturaCRM, 2017). Contrató a un grupo de expertos en análisis de datos masivos con el objetivo de sondear las redes sociales. El resultado fue indiscutible para invertir en marketing, definir las horas a las que tendría mayor repercusión el mensaje y los canales más apropiados para hacerlo.

En el ámbito empresarial; en el 2013 Tim Burke y Stephen Hankinson vieron en Twitter la oportunidad de encontrar nuevos clientes para su negocio de aplicaciones (BBVA, 2018). Por eso se dedicaron a analizar durante un año los datos gráficos de Twitter para identificar y segmentar rápidamente grupos y eso hizo crecer su base de clientes. De ahí nació la compañía Affinio, que se dedica a planificar y ejecutar estrategias de marketing para diferentes marcas.

A partir de los casos anteriores, se puede decir que las redes sociales son océanos de datos alimentados en tiempo real por millones de usuarios. Twitter es una de las redes sociales más populares a nivel mundial. Según Analitika (2015, p. 6) en nuestro país es la tercera red más utilizada, por detrás de Facebook y YouTube, y el 33% de los usuarios salvadoreños utilizan dicha red para comentar la actualidad. Estas estadísticas reflejan que Twitter es un lugar donde se puede buscar y encontrar información relevante sobre temas de acontecer nacional.

Pero ¿cuál es el procedimiento para pescar información en el océano de datos que generan los usuarios de Twitter en El Salvador?

Método

El objetivo de este documento es exponer el procedimiento para pescar información en el océano de datos que generan los usuarios de Twitter. Para ello se aprovechó un evento mediático que generó una gran cantidad de tuits en el municipio de San Salvador. Se obtuvo una muestra de 1,000 tuits que luego se analizaron aplicando técnicas de minería de textos.[1]

En la pesca se deben tener en cuenta diferentes variables para la obtención de buenos resultados. Uno de los puntos que más influye a la hora de pescar es la elección donde se desarrollará la actividad, ya que si hay algo que altere el estado del mar y que influya en la pesca son las olas porque inciden directamente el lugar donde se encuentran los peces. En el caso del océano de Twitter los peces son los tuits y las olas son eventos que producen una gran cantidad de tuits y retuits. Una ola reciente es el error que presentó el sistema informático que la empresa SMARTMATIC utilizó para transmitir el escrutinio preliminar de las elecciones de diputados y alcaldes 2018 en El Salvador. Para fines de explicar el procedimiento de pesca se aprovechará la ola de tuits que generó dicho evento.

La metodología comienza con la preparación de los instrumentos para el desarrollo de la actividad de pesca. Los instrumentos utilizados fueron: API[2] REST[3] pública de Twitter, lenguaje de programación R versión 3.4.3 de 64 bits, Rstudio versión 1.1.423 como Interfaz de Entorno de Desarrollo (IDE). Luego se lanza la red, script[4], al océano de datos de Twitter para pescar tuits; se continúa con la limpieza de los pescados, tuits, y finalmente se analizan para generar un producto; en este caso, nueva información.

Preparando la red para adquirir tuits

1. Crear las credenciales de acceso a la API pública de Twitter. Para hacer uso de dicha API es necesario registrarse como desarrollador (https://developer.twitter.com/) y luego se debe crear una aplicación en el sitio web https://apps.twitter.com/app/new

2. Una vez completado el paso anterior, se debe tomar nota de los valores de las siguientes variables: *Consumer Key, Consumer Secret, Access Token y Access Token Secret*, ya que son las credenciales que se deben configurar en el script para la adquisición de tuits.

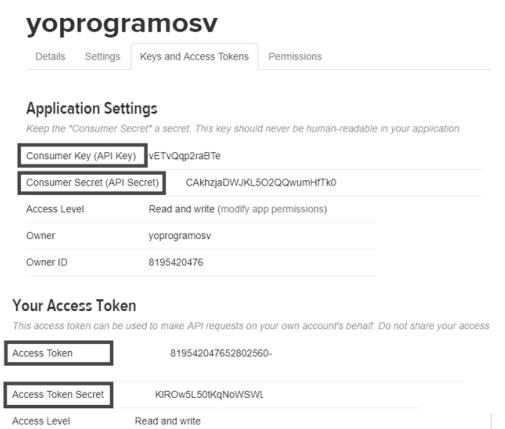

Figura n.° 1: Llaves de acceso a la *API pública de Twitter.* Las variables *(Consumer Key, Consumer Secret, Access Token y AccessToken Secret)* enmarcadas con recuadros se encuentran en la pestaña *"Key and Access Token"* de las opciones de configuración de la aplicación y son indispensables para poder hacer uso de la API REST de Twitter. Fuente: Elaboración propia.

3. Finalmente se deben instalar las librerías de R que son de utilidad para la adquisición, limpieza y análisis de los datos.

1 En ingles Text Minig. Consiste en el proceso de generar nueva información de los textos.

2 Interfaz de programación de aplicaciones (API), conjunto de funciones que se ponen a disposición para ser utilizadas por otro software.

3 Transferencias de estado representacional (REST): Interfaz entre sistemas que utilizan el protocolo HTTP para compartir datos o generar operaciones sobre los datos.

4 Conjunto de instrucciones o comandos escritos en determinado lenguaje de programación que conforman una rutina o programa.

Tabla n.° 1

Librerías que se utilizaron para el desarrollo del experimento

Librería	Utilidad
twitteR	Librería que permite hacer uso de la API de Twitter. Se utilizó para realizar la búsqueda y adquisición de tuits.
Stringi	Librería que permite codificar los caracteres en cualquier alfabeto. Se utilizó para codificar los tuits en el alfabeto latino.
Dplyr	Librería que permite manipular gramaticalmente las variables de tipo texto. Se utilizó para la limpieza y tratamiento de los datos.
Tidyr	Librería para clasificación y ordenamiento de los datos.
Widyr	Librería para tratamiento de los datos en grandes matrices. Se utilizó para realizar operaciones tales como correlaciones y conteos de ocurrencia.
tm	Librería de Minería de Texto. Se utilizó para el procesamiento de los datos, administración de la información de los datos (metadata), creación de diccionarios y matrices de términos relevantes.
Tidytext	Librería de Minería de Texto para el proceso de palabras y análisis de semántico.
wordcloud	Librería para generar gráficos de nubes de palabras.
ggplot2	Librería para la generación de gráficos de barra, círculo, histogramas, etc.
Igraph	Librería para la generación de gráficos simples y análisis de redes. Se utilizó para elaborar grafos[5] de los términos relevantes.
Ggraph	Extensión de la librería ggplot2. Se utilizó para la creación de gráficos de redes.
graphTweets	Librería para construir gráfica de grafos a partir de los tuits. Se utilizó para generar los gráficos de interacción entre los usuarios.

Lanzando la red al océano de datos de Twitter

Una vez preparada la red:

1. Para la adquisición de los tuits, se preparó un script de R y se utilizó la librería twitteR para conectarse a la API REST pública de Twitter por medio de la función setup_twitter_oauth y los parámetros Consumer Key, Consumer Secret, Access Token y AccessToken Secret obtenidos al crear la aplicación.

2. Para la obtención de los datos se utilizó la función searchTwitter que provee la librería twitteR y se configuró con los siguientes parámetros. Ver Tabla n.° 2.

3. Convertir la lista de tuits recuperados en un objeto DataFrame[6] por medio de la función twListToDF.

Limpiando los pescados, tuits recolectados

En la minería de texto el proceso de limpieza consiste en eliminar del texto todo aquello que no aporte información sobre la temática de interés, por lo que se procede a eliminar: patrones no informativos (direcciones web), signos de puntuación, caracteres sueltos, espacios adiciones.

5 Tipo abstracto de datos que consiste en un conjunto de nodos y un conjunto de aristas que establecen la conexión entre los nodos.

6 Estructura de datos similar a una matriz con la diferencia que cada columna puede ser de diferente tipo de dato.

Tabla n.° 2

Parámetros de búsqueda

Parámetro	Descripción	Valor
searchTerm	Termino de búsqueda puede ser: una palabra, un hashtag (#smartmatic), un usuario (@tseelsalvador)	smartmatic
geocode	Referencia geográfica de donde se quiere obtener la información. Incluye a su vez tres datos: Latitud, Longitud y millas a la redonda	13.6893500,-89.1871800,50mi
resultType	Tipos de tuits pueden ser: recientes o destacados.	recent
N	Cantidad de tuits que recuperar en la petición, el valor máximo son 1,500.	1,000

Tabla n.° 3

Comandos utilizados para eliminar información que no aporta a la temática

Comando	Acción en los tuits
tweets.text=gsub("@\\w+", "", tweets.text)	Eliminar las menciones.
tweets.text=gsub("[[:punct:]]","",tweets.text)	Eliminar los signos de puntuación
tweets.text=gsub("http\\w+", "", tweets.text)	Eliminar las direcciones web
tweets.text=gsub("[|\t]{2,}", "", tweets.text)	Eliminar los tabuladores
tweets.text=gsub("^ ", "", tweets.text)	Eliminar los espacios adicionales al principio de las palabras.
tweets.text <- gsub(" $", "", tweets.text)	Eliminar los espacios adicionales al final de las palabras.

Luego, se debe separar el texto en palabras ya que es el elemento más sencillo con significado propio para el posterior análisis.

Sucede con frecuencia que el lenguaje utilizado por los usuarios no siempre es correcto (Montecinos García, 2014). Muchas veces no ocupan tildes y escriben palabras que no aparecen en los diccionarios. Por ejemplo "porfa" en vez de "por favor", "xq" en vez de "porque". También es común que dupliquen vocales en las palabras y cambien la letra "q" por la letra "k", todas estas deformaciones del lenguaje dificultan el análisis. Después de separar el texto en palabras se debe eliminar aquellas que no tienen significado propio: artículos, preposiciones, conjunciones, deformaciones del lenguaje, etc. La mejor forma de realizarlo es utilizando un diccionario que incluya esas palabras. Para el desarrollo de esta investigación se utilizó los diccionarios que proveen las librerías *tidyText* y *tm.*

```
stop_words <- get_stopwords(language = "es", source = "snowball")
stop_words <- bind_rows(stop_words,
                        data_frame(word = tm::stopwords("spanish"),
                        lexicon = "custom"))
```

Figura n. ° 2: Fragmento de código para crear el diccionario de palabras a suprimir. La primera línea del código define un variable de nombre stop_words de tipo lista con palabras que incluyen artículos, preposiciones, conjunciones, etc., definidas en la librería tidyText. La segunda línea de código agrega más palabras a la variable stop_words definidas en la librería tm. Fuente: elaboración propia.

Procesando y empacando para generar el producto

Después de limpiar los tuits se debe hacer el análisis según el interés de la investigación. En este caso se quiere conocer la opinión de los usuarios acerca del error que presentó el sistema de transmisión de votos de la empresa SMARTMATIC en el escrutinio preliminar de las elecciones de alcaldes y diputados 2018, para ello se hizo un análisis exploratorio de la siguiente manera:

1. Determinar las palabras más utilizadas en los tuits capturados por medio de la construcción de una tabla de frecuencia, gráfico de barras con las diez palabras más utilizadas, gráfico tipo nube de palabras para conocer los términos con una frecuencia mayor a quince.

2. Determinar la relación que tienen estas palabras por medio de la construcción de grafos donde los nodos corresponden a las palabras y las aristas son la frecuencia en que esas palabras aparecen en un mismo tuit.

3. Conocer la dinámica de conversación entre los usuarios de los tuits capturados por medio de la construcción de un grafo donde los nodos representan a los usuarios y las aristas representan las menciones que hacen a otros usuarios.

Resultados

El análisis exploratorio de los datos se inició graficando las palabras más utilizadas por los usuarios de Twitter en el municipio de San Salvador respecto al término de búsqueda "smartmatic". (Ver Figura n.° 3)

Utilizando la librería wordcloud (CRAN, 2015) de r se generó una nube de palabras para una mejor exploración de los términos con mayor frecuencia. (Ver Figura n.° 4)

Una vez graficados las palabras con mayor frecuencia es de interés conocer si existe relación entre ellas. La relación implica con qué frecuencia las palabras aparecen en el mismo mensaje. Sin embargo, no implica que las palabras deben aparecer una después de otra.

Figura n.° 4: Nube de Palabras de los términos con mayor frecuencia utilizado por los usuarios. La nube muestra de forma cualitativa los términos con mayor frecuencia usados en los tuits capturados. La frecuencia del término está dada por el tamaño de la letra. Fuente: elaboración propia.

Una vez graficados las palabras con mayor frecuencia es de interés conocer si existe relación entre ellas. La relación implica con qué frecuencia las palabras aparecen en el mismo mensaje. Sin embargo, no implica que las palabras deben aparecer una después de otra.

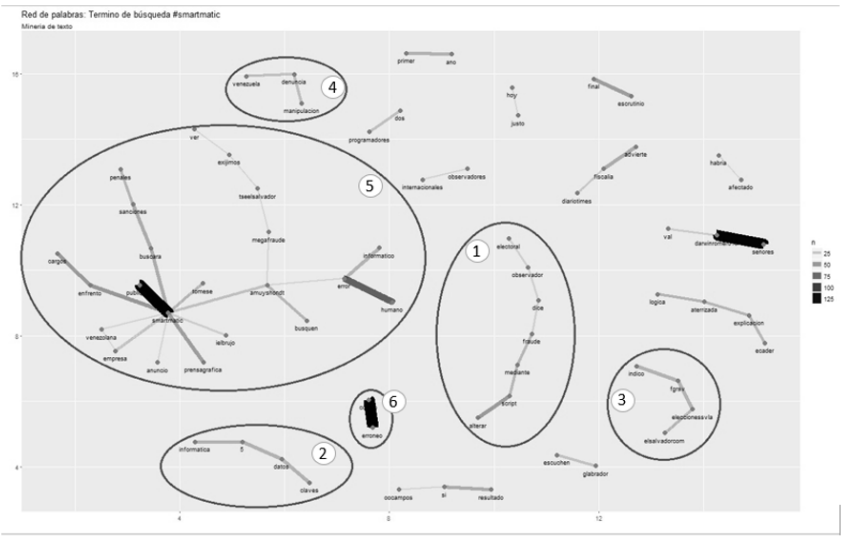

Figura n.° 5: Grafo de relaciones entre los términos con mayor frecuencia. El gráfico muestra un grafo no conexo de los términos, encerrados con elipses se encuentran los subgrafos de términos relaciones con una frecuencia mayor a 25. Fuente: elaboración propia.

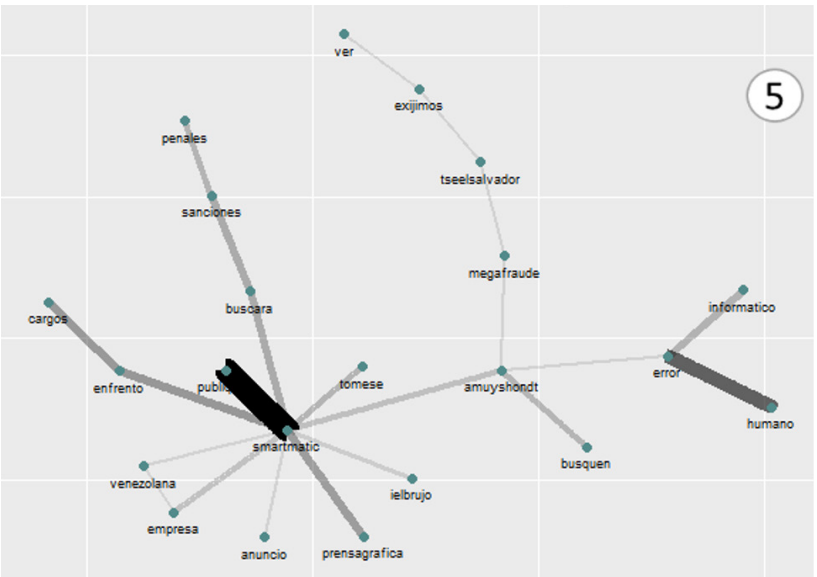

Figura n.° 6: Ampliación del subgrafo 5 de la Figura n.° 5. El subgrafo muestra la frecuencia de la relación entre las palabras más utilizadas. La frecuencia está dada por el grosor y color de la arista, la línea que los une. La línea negra indica una frecuencia de 125 (límite superior). La línea gris más delgada indica una frecuencia de 25 (límite inferior). Fuente: elaboración propia.

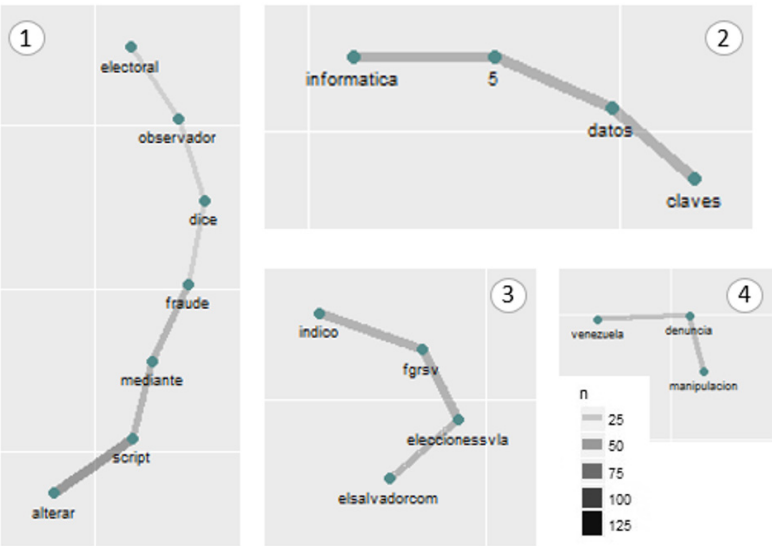

Figura n.° 7: Ampliación de los subgrafos del 1 al 4 de la Figura n.° 5. Cerca del subgrafo 4 se encuentra la escala de frecuencia para las relaciones entre palabras. Fuente: elaboración propia.

En los tuits además del mensaje se encuentran las siguientes entidades: hashtag, palabras o frases que inician con el carácter de numeral (#); menciones, nombres de usuarios de Twitter que inician con el carácter de arroba (@); y direcciones web que podemos identificar porque inician con http o https.

En la muestra de tuits obtenida, el 85% de los tuits contiene menciones, por tanto, se exploró la relación que existe entre los usuarios mencionados para determinar a los posibles generadores de opinión. (Ver Figura n.° 8)

Discusión

Esta investigación tuvo como propósito explicar el procedimiento para pescar información en el océano de datos de Twitter. Para ello se experimentó adquirir una muestra 1,000 tuits acerca del error ocurrido en el sistema de transmisión de datos de la empresa SMARTMATIC durante el escrutinio preliminar de las elecciones de alcalde y diputados 2018.

Durante la etapa de adquisición de los tuits se experimentó pescar las tendencias, tópicos que se están hablando, de los usuarios de la red social en San Salvador por medio de la API REST publica de Twitter y el lenguaje R. Se utilizó la función *getTrends* de la librería *twitteR* para recuperar las tendencias según la documentación que ofrece *The Comprehensive R Archive Network* (CRAN, 2016, pág. 8) pero no se logró obtener las tendencias debido a que la función requiere el parámetro woeid que indica el código describe a una ubicación según el servicio de *Yahoo! Where On Earth ID* y dicho código no se encuentra definido para el país. Entonces para determinar las tendencias en el país se puede utilizar la función SearchTerm, explicada en el apartado de método, para

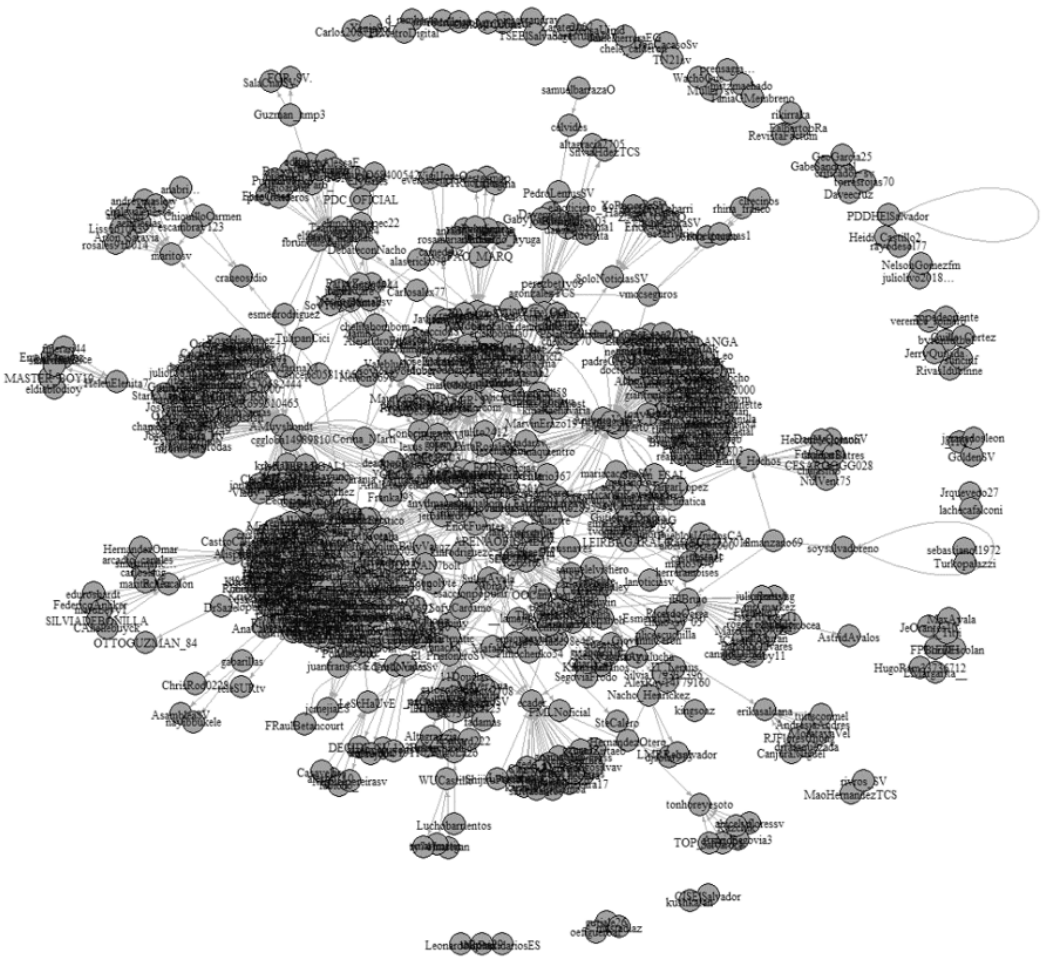

Figura n.° 8: Grafo de relaciones entre las menciones. El resultado es un multígrafo donde los nodos representan a los usuarios mencionados en los Tweets y la arista indica la dirección de la comunicación. Fuente: elaboración propia.

recuperar los Tweets y luego identificar los hashtags y/o palabras más frecuentes.

De los resultados obtenidos se puede afirmar que es posible pescar información del océano de datos de Twitter. Si dicha información se acompaña con experiencia se transformará en nuevo conocimiento. Ahora bien, la calidad de la información depende del tratamiento (limpieza y transformación) de los datos capturados. Como parte de la limpieza de los datos es necesario eliminar las palabras que no tienen significado propio, para ello se requiere construir un diccionario que incluya dichas palabras.

La nube de palabras y el grafo de relación entre palabras permiten explorar los términos de mayor uso (tendencias) en una ola de Tweets, sin embargo, no indican si las palabras transmiten un sentimiento positivo o negativo.

El grafo de relaciones entre menciones (usuarios) muestra un multígrafo con muchos nodos y aristas, pero la presentación de dicho grafo por medio de una imagen no permite explorar con claridad la información que contiene. En la imagen resultante es complicado determinar que menciones son los generadores de opinión y quienes son los seguidores. Para poder realizar un mejor análisis es necesario preparar los datos y exportarlos a un software de análisis de redes de datos como Gephi.

Bibliografía

Álvarez, A. (2018). Deporte y tiempo libre. *www.a-alvarez.com*. Recuperado de: https://www.a-alvarez.com/blog/pesca/la-pesca/leer-el-mar-trucos-para-saber-donde-estan-los-peces/4502

Analitika Market Research. (2015). *Estudio de redes sociales en El Salvador*. Recuperado de: https://www.gestiopolis.com/estudio-de-redes-sociales-en-el-salvador-2015/

Banco Bilbao Vizcaya Argentaria. (2018). Case study: Affinio y las APIs de Twitter para definir campañas publicitarias. *BBVA*. Recuperado de: https://bbvaopen4u.com/es/actualidad/case-study-affinio-uso-las-apis-de-twitter-para-definir-campanas-publicitarias

CulturaCRM. (2017). Data mining: casos de éxito. *CulturaCRM*. Recuperado de CulturaCRM: https://culturacrm.com/data-mining/data-mining-casos-exito/

Montecinos García, L. (Agosto de 2014). *Análisis de sentimientos y predicción de eventos en Twitter*. (Memoria para optar al título de Ingeniero Civil Eléctrico). Universidad de Chile. Recuperado de: http://repositorio.uchile.cl/bitstream/handle/2250/130479/cf-montesinos_lg.pdf?sequence=1

The Comprehensive R Archive Network. (20 de febrero de 2015). *CRAN*. Recuperado de https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf

The Comprehensive R Archive Network. (29 de agosto de 2016). *CRAN*. Recuperado de Package 'twitteR': https://cran.r-project.org/web/packages/twitteR/twitteR.pdf