Memorias del 2022 Congreso Nacional de Control Automático
12-14 de Octubre, 2022. Tuxtla Gutiérrez, México.

Text Mining Methodology for Building Water Leak Maps from Tweets

Lizeth Torres * **Roberto G. Ramírez-Chavarría** *
Martín R. Jiménez-Magaña ** **Lucero F. García-Franco** ***

* *Instituto de Ingeniería, Universidad Nacional Autónoma de México, Alcaldía Coyoacán, Ciudad de México, México.*
** *FES-Aragón, Universidad Nacional Autónoma de México, Ciudad de México, México.*
*** *Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Ciudad de México, México.*

Abstract: This article presents a text mining methodology that is used in the framework of a project called *Fugometría*. The objective of the proposed methodology is the construction of maps with markers that indicate the location of water leaks. To achieve this purpose, the initial step is the collection of *tweets* issued by citizens to report the existence of leaks in drinking water distribution networks. Once these *tweets* are collected, they are parsed for an address in the body of the message. If an address is detected, it is converted into a GPS coordinate, which is in turn used to build a map with markers indicating the location of water leaks. In order to show the applicability of the methodology, some preliminary results on the construction of a leak map of Mexico City are shown.

Keywords: Leak diagnosis; Text mining; Twitter; Social media; Content analysis

1. INTRODUCCIÓN

Una de las consecuencias del calentamiento global y la industrialización de los países en desarrollo es la escasez de agua. Un problema que, si no se aborda en el presente, provocará malestar social, aumento de enfermedades y mortalidad en el futuro.

En México, la distribución geográfica del agua no coincide con la distribución geográfica de la población. Existen diferencias sustanciales entre el sur y el norte del territorio. En el norte llueve menos que en el resto del país, mientras que en el sur llueve tanto que, en ciertas épocas del año, se producen inundaciones desastrosas. Además del clima, hay otros aspectos que afectan la justa distribución del agua en México, por ejemplo, acuíferos sobreexplotados y contaminados, la falta de monitoreo en tiempo real, el desperdicio y mal uso del agua en algunas actividades, el tratamiento insuficiente de las aguas residuales, el envejecimiento de la infraestructura de distribución, la presencia excesiva de fugas, entre otros. Es por eso que todas las posible soluciones para mejorar la distribución del agua potable son bienvenidas.

En este artículo se presenta una metodología de minería de textos que es utilizada en el marco de un proyecto nombrado *Fugometría*.[2] Por un lado, el objetivo de este proyecto es integrar diversas herramientas de ciencia de datos, geolocalización y estadística con la finalidad de cuantificar variables asociadas a la presencia de fugas en ciudades mexicanas. Por otro lado, el propósito de la metodología a presentar es la creación de mapas de fugas de agua a partir de reportes emitidos por usuarios de `Twitter`.

Un mapa de fugas puede tener diversas utilidades, por ejemplo, para determinar la densidad de fugas en una región en particular, para ayudar a contabilizar la reincidencia de ciertas fugas, o incluso, para calcular la probabilidad de fugas en un región. También puede convertirse en una herramienta para que usuarios e instituciones conozcan las coordenadas exactas de una fuga y, a partir de ello, suministren más información sobre las posibles causas.

En la Sección 2 se presenta un estado del arte resumido sobre el uso de *tweets* para aplicaciones científicas e ingenieriles. En la Sección 3 se describe la metodología propuesta para minar direcciones postales de fuga suministradas por usuarios de `Twitter`. En la Sección 4 se presentan resultados preliminares de la construcción de un mapa de fugas para la Ciudad de México. Finalmente, en la Sección 5, se presentan algunas conclusiones derivadas de los resultados preliminares de la metodología propuesta.

2. TRABAJO RELACIONADO

`Twitter` es una de las redes sociales más populares en el mundo. Se estima que tiene más de 300 millones de usuarios que generan 65 millones de *tweets* al día a través

[1] Corresponding author: forreso@iingen.unam.mx
[2] Nombre conformado por la palabra en español fuga y por la palabra griega *metrón*, que significa medida.

de sus computadoras y teléfonos móviles. En estos *tweets* se comparte información de índole diversa, por ejemplo, opiniones políticas, conversaciones culturales, propaganda empresarial o gubernamental, quejas, bromas, por lo que tener acceso a esta información es una oportunidad para comprender tendencias y patrones sobre pandemias, desastres naturales, manifestaciones, accidentes.

Una característica importante de la información a través `Twitter` es su evolución en tiempo real debida a la restricción impuesta por esta red social en la longitud de los mensajes de texto: situación que facilita su rápida lectura y retransmisión. Esta restricción también alienta a los usuarios de `Twitter` a escribir *tweets* varias veces al día, posibilitando que otros usuarios sepan qué están pensando o cómo la están pasando lo demás en el momento.

La gran cantidad de *tweets* da como resultado la emisión de numerosos informes de eventos sociales como fiestas, partidos de fútbol, campañas presidenciales; o de eventos funestos como tormentas, incendios, atascos de tráfico, disturbios, lluvias torrenciales, terremotos y huracanes. Como explican Sakaki et al. (2010), cada usuario de `Twitter` puede concebirse como un sensor social capaz de proporcionar información espacio temporal de noticias importantes. Es por ello que estos sensores sociales han sido utilizados para realizar investigaciones científicas con pertinencia social.

Sakaki et al. (2010) propusieron un sistema de notificación de sismos que monitorea *tweets* y envía notificaciones rápidamente a los usuarios registrados. Este sistema está basado en un Filtro de Kalman y en un Filtro de Partículas. Para detectar un evento sísmico, los autores diseñan un clasificador de *tweets* basado en una máquina de vectores de soporte. Este clasificador determina si un *tweet* es de clase positiva, i.e., que está reportando un sismo verdadero, o de clase negativa. Las características que se utilizan para la alimentar el clasificador son palabras clave relacionadas con un sismo, el número de palabras y su contexto. Tanto los *tweets* positivos, como los negativos, alimentan un modelo temporal que proporciona como salida la probabilidad de que realmente exista un sismo. Si se determina que existe el sismo, se obtienen las coordenadas de los *tweets* emitidos y, utilizando un filtro de Kalman o un filtro de partículas, se calcula el epicentro del sismo. Los autores afirman que su sistema de monitoreo puede detectar el 96 % de los terremotos detectados por la Agencia Meteorológica de Japón (JMA) y que su sistema entrega las notificaciones mucho más rápido que la JMA.

Gerber (2014) propuso una metodología para predecir crímenes en Chicago combinando registros históricos oficiales con información espaciotemporal integrada en *tweets* emitidos dentro del área geográfica de interés.

Sarker et al. (2016) diseñaron una técnica de clasificación automática supervisada para identificar *tweets* que contenían señales de abuso de medicamentos y evaluaron la utilidad de Twitter para investigar patrones de abuso a lo largo del tiempo. Para ello, recopilaron *tweets* asociados con tres medicamentos de los que se abusa con frecuencia (Adderall, oxicodona y quetiapina). Posteriormente, anotaron manualmente 6400 *tweets* que mencionaban estos tres medicamentos y un medicamento de control (metformina), que no es objeto de abuso. Finalmente, realizaron análisis cuantitativos y cualitativos para determinar si las publicaciones en `Twitter` contenían señales de abuso de medicamentos recetados.

de Bruijn et al. (2019) presentaron un algoritmo para detectar y localizar inundaciones en tiempo real a escala mundial utilizando información de `Twitter`. El algoritmo se desarrolló utilizando 88 millones de *tweets*, de los cuales se identificaron más de 10 000 eventos de inundación en 176 países en 11 idiomas en poco más de cuatro años. De acuerdo a los autores, este algoritmo está corriendo en tiempo real, y los datos que se obtienen con él están publicados en una plataforma de acceso libre: `https://www.globalfloodmonitor.org/`.

Recientemente, Jiang et al. (2022) propusieron un sistema robusto y eficaz para monitorear la enfermedad COVID-19. Este sistema está basado en la minería de *tweets* y en técnicas de aprendizaje profundo. Los autores claman que este sistema puede predecir la presencia de nuevos casos, así como tasas de mortalidad.

Como queda en evidencia de la revisión del estado del arte, un gran número de algoritmos, métodos y sistemas han sido propuestos para utilizar la información contenida en *tweets* con fines de monitoreo, estadísticos y de predicción. En particular, para supervisar desastres naturales, enfermedades, campañas electorales o el comportamiento del mercado. Sin embargo, en esta contribución es la primera vez que se plantea un método de minería de *tweets* para extraer información sobre fugas de agua en ciudades con un sistema de distribución.

La idea subyacente de esta propuesta es utilizar los 14 millones de usuarios de `Twitter` en México (Statista, 2022) como sensores sociales para atacar uno de los problemas más graves de este país: el desperdicio de agua.

3. METODOLOGÍA

El objetivo de la metodología que a continuación se presenta es la generación de mapas con marcadores indicando las coordenadas geográficas de fugas de agua reportadas por usuarios de `Twitter`. Esta metodología está inspirada en la serie de pasos propuesta por Yoon et al. (2013) para minar contenido de *tweets*.Los seis pasos que conforman esta metodología se ilustran en la Figura 1 y se describen a continuación.

1. **Recolección de Tweets**. El primer paso es la recolección de *tweets* de acuerdo a ciertas reglas y restricciones de búsqueda. La recolección se compone de dos etapas: la búsqueda y el almacenamiento de *tweets*, tareas que pueden llevarse a cabo utilizando una aplicación diseñada para ello, ya sea comercial o de autoría propia.
2. **Preprocesamiento del Cuerpo del Mensaje.** En este paso se limpian (o filtran de ruido) los *tweets*, es decir, se eliminan los caracteres, palabras, *emojis*

Figura 1. Metodología para minar direcciones postales de `Twitter`

que no aportan información. Para ello, se aplican una serie de filtros como los que proponen Yoon et al. (2013) y Ralston et al. (2014).

3. **Transformación de la Información.** En este paso la información se convierte en valores numéricos con un significado en el contexto en el que será utilizada.
4. **Identificación de la Información.** En este paso se extrae la información deseada, por ejemplo, nombres de personas, libros o películas, direcciones postales.
5. **Despliegue de la información.** En este paso la tarea es representar la información de una manera significativa y visual para su interpretación y comprensión. La visualización de información es una forma eficaz de compartir conocimientos en un formato digerible que ayude a hacer el mejor uso de ella.
6. **Uso de la información.** La información minada se puede usar para planear, tomar decisiones, programar actividades, crear estrategias. En conclusión, el cumplimiento de este paso es la meta final de minar *tweets*.

A continuación se presenta cómo se aplica la metodología para obtener las coordenadas en donde, según los reportes de usuarios de la Ciudad de México, existe una fuga.

3.1 Recolección de Tweets

La tarea a realizar en este paso es la recolección de *tweets* en los que se reportan fugas en la Ciudad de México. Para ello, se buscan *tweets* en español que incluyan en el cuerpo del mensaje alguna de las siguientes palabras: {fuga}, {fugando} o {SACMEX}[3]. Dado que también se reportan fugas de gas, de combustible, de información o de delincuentes, es necesario excluir los *tweets* que contengan la palabra {gas, combustible} o frases como {se dió a la fuga, fuga de divisas}. De esta manera se tienen 3 conjuntos de palabras: (1) las que se buscan en el *tweet*, (2) las que no debe contener el *tweet* y (3) las que, asociadas con cierta estructura, tampoco deben estar en el *tweet*.

[3] Organismo que gestiona el agua en la Ciudad de México

En la Figura 2 se muestran algunos reportes de fuga encontrados a partir de la ejecución de las reglas de búsqueda.

Figura 2. Capturas de pantalla de *tweets* de ciudadanos reportando fugas

Los *tweets* que cumplen las reglas de búsqueda se descargan en hojas de cálculo con los metadatos del *tweet* y la información del perfil del usuario, por ejemplo: fecha de emisión del *tweet*, nombre del usuario que lo emitió, el ID del *tweet*, los archivos multimedia, como fotos o vídeos, que se adjuntaron al *tweet*, entre otros.

En la Figura 3 se muestran algunas fotos adjuntadas por usuarios de `Twitter` en sus reportes.

Figura 3. Fotos compartidas por ciudadanos como evidencia de las fugas que reportan

3.2 Preprocesamiento del Cuerpo del Mensaje

En este paso se eliminan las palabras vacías[4], la puntuación, los nombres de usuario después de una @, los hipervínculos después *http*, *emojis*. No se elimina la conjunción (y), ya que a menudo se utiliza para indicar la unión de dos calles.

[4] Nombre que reciben las palabras sin significado como artículos, pronombres, preposiciones, etc. que son filtradas antes o después del procesamiento de datos en lenguaje natural (texto)

3.3 Identificación de la Información

La tarea a cumplir en este paso es la identificación de direcciones postales en el texto de un *tweet*. Para ello, se elige o desarrolla una aplicación que utilice una estrategia particular de identificación. En el caso de direcciones postales, el proceso de identificación varía de país en país (Khanwalkar et al., 2013).

En México, hay palabras clave que permiten identificar una dirección postal por ejemplo: Avenida o su abreviación Av., Calle o su abreviación C., la abreviatura de número (No.), el símbolo #, esquina y su abreviatura Esq. Para el proceso de identificación también son buscados arreglos de números de 1 a 4 dígitos.

3.4 Transformación de la Información

La tarea a ejecutar en este paso es la geocodificación de las direcciones postales, es decir, la transformación de direcciones postales en coordenadas de latitud y longitud. Esta tarea se realiza manualmente o de manera automatizada con ayuda de una aplicación. Para realizar esta tarea, desarrollamos un aplicación en la plataforma `Google Apps Script` basada en el código propuesto por Aleks (2021).

3.5 Despliegue de la Información

Para desplegar la información de manera eficaz, es decir, las coordenadas de las fugas, se utilizó el servicio `Google My Maps`. Para ello se desarrolló una aplicación en la plataforma `Google Apps Script`, que coloca un marcador en cada coordenada donde se reportó una fuga de agua.

Con la finalidad de facilitar su compresión, la metodología se presenta en forma de algoritmo a continuación.

Algoritmo 1 Algoritmo para construir mapa de fugas

1. Establecer un conjunto de palabras de búsqueda denotado por Q.

2. Buscar en `Twitter` cada t minutos palabras incluidas en el conjunto de palabras Q y obtener el conjunto de *tweets* T que las contengan.

3. Dado un conjunto de palabras no deseadas E, descartar los *tweets* del conjunto T que contengan una o varias palabras del conjunto E.

4. Dado un conjunto de palabras deseadas D, descartar los *tweets* del conjunto T que no contengan alguna palabra del conjunto D.

5. Dado un conjunto de palabras innecesarias denotado por F, eliminar dicho conjunto del conjunto de *tweets* T.

6. Para cada *tweet* $w \in T$ obtener sus coordenadas geográficas (latitud, longitud).

7. Colocar un indicador sobre un mapa en cada coordenada obtenida en el paso anterior.

4. DISCUSIÓN DE RESULTADOS

Los datos que a continuación se presentan se obtuvieron del 1 de mayo de 2022 al 6 de julio de 2022. Durante este periodo se encontraron 59 *tweets* que satisficieron las condiciones de búsqueda.

En la Tabla 1 se lista la cantidad de reportes de fuga por alcaldía en el periodo indicado. La alcaldía con mayor número de reportes fue la alcaldía de Tlalpan. Se teoriza que la situación se debe a diversos factores, por ejemplo, a la extensión territorial, a la condición de la infraestructura, al número de habitantes, entre otros. Sin embargo, para determinar la razón se requiere un estudio con más elementos de información.

En la Figura 4 se muestra un mapa donde están indicadas las posiciones de las fugas que se reportaron en `Twitter`.

En la Figura 5 se muestra un mapa con las posiciones de fuga y los límites geográficos de las alcaldías. Este mapa puede ayudar a visualizar la densidad de fugas por alcaldía. Mientras que la información utilizada para construir este mapa puede ayudar a calcular tal densidad.

Tabla 1. Reporte de Fugas

Alcaldía	Número de Reportes
Azcapotzalco	2
Coyoacán	9
Cuajimalpa de Morelos	0
Gustavo A. Madero	6
Iztacalco	2
Iztapalapa	8
La Magdalena Contreras	1
Milpa Alta	0
Álvaro Obregón	5
Tlahuac	2
Tlalpan	11
Xochimilco	1
Benito Juárez	6
Cuauhtémoc	1
Miguel Hidalgo	5
Venustiano Carranza	0

5. CONCLUSIONES

En este artículo se presentó una metodología para minar información de importancia contenida en *tweets* emitidos por usuarios de esta red social y de alguna red de distribución de agua potable en particular. La metodología fue probada y aplicada para extraer información relacionada con la ubicación geográfica de fugas en las diferentes alcaldías de la Ciudad de México. La información minada puede ser utilizada para alimentar modelos de pronóstico y sistemas de alerta, para la creación de registros históricos, para análisis estadísticos y evaluación de riesgos, pero sobre todo, para visibilizar eventos de fugas y su atención inmediata.

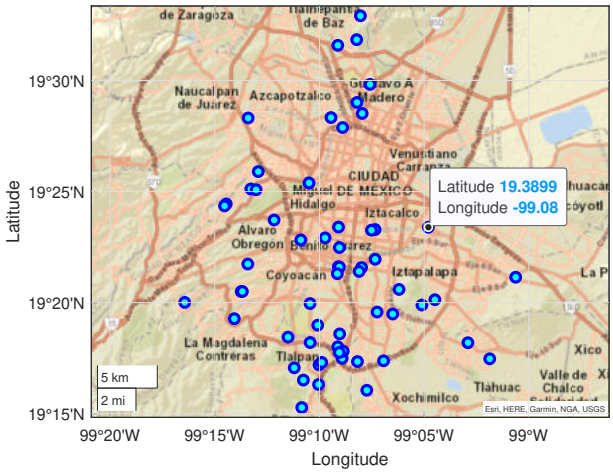

Figura 4. Mapa de fugas de la Ciudad de México

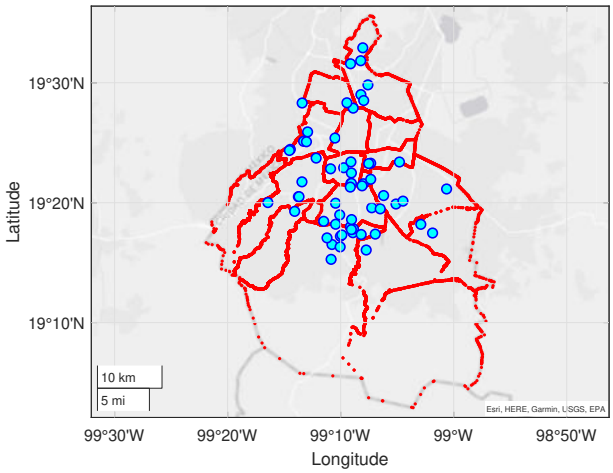

Figura 5. Mapa de fugas en las alcaldías de la Ciudad de México

6. AGRADECIMIENTOS

Este trabajo se llevó a cabo gracias al apoyo otorgado por CONACYT (Atención a Problemas Nacionales, Convocatoria 2017, Proyecto 4730: Estaciones de Diagnóstico y Monitoreo para Redes de distribución de Agua con Interconexicón a Internet).

REFERENCIAS

Aleks (2021). Convertir direcciones a coordenadas de latitud y longitud con google sheets y google maps (geocoding). URL `https://bit.ly/3RRa0u0`.

de Bruijn, J.A., de Moel, H., Jongman, B., de Ruiter, M.C., Wagemaker, J., y Aerts, J.C. (2019). A global database of historic and real-time flood events based on social media. *Scientific data*, 6(1), 1–12.

Gerber, M.S. (2014). Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61, 115–125.

Jiang, J.Y., Zhou, Y., Chen, X., Jhou, Y.R., Zhao, L., Liu, S., Yang, P.C., Ahmar, J., y Wang, W. (2022). Covid-19 surveiller: toward a robust and effective pandemic surveillance system based on social media mining. *Philosophical Transactions of the Royal Society A*, 380(2214), 20210125.

Khanwalkar, S., Seldin, M., Srivastava, A., Kumar, A., y Colbath, S. (2013). Content-based geo-location detection for placing tweets pertaining to trending news on map. En *The Fourth International Workshop on Mining Ubiquitous and Social Environments*, 37. Citeseer.

Ralston, M.R., O'Neill, S., Wigmore, S.J., y Harrison, E.M. (2014). An exploration of the use of social media by surgical colleges. *International Journal of Surgery*, 12(12), 1420–1427.

Sakaki, T., Okazaki, M., y Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. En *Proceedings of the 19th international conference on World wide web*, 851–860.

Sarker, A., O'connor, K., Ginn, R., Scotch, M., Smith, K., Malone, D., y Gonzalez, G. (2016). Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from twitter. *Drug safety*, 39(3), 231–240.

Statista (2022). Número de usuarios de twitter en algunos países de américa latina en enero de 2022. URL `https://es.statista.com/`.

Yoon, S., Elhadad, N., y Bakken, S. (2013). A practical approach for content mining of tweets. *American journal of preventive medicine*, 45(1), 122–129.