
Identificación de incidentes de tráfico en Panamá por medio del análisis de datos de redes sociales

Liu, Lucia

Facultad de Ingeniería de Sistemas Computacionales, Universidad Tecnológica de Panamá
Ciudad de Panamá, Panamá
lucia.liu@utp.ac.pa

Guevara, Ameth

Dirección de Tránsito y Seguridad Vial, Autoridad del Tránsito y Transporte Terrestre
Ciudad de Panamá, Panamá
aguevara@transito.gob.pa

Sanchez-Galan, Javier E.

Facultad de Ingeniería de Sistemas Computacionales, Universidad Tecnológica de Panamá
Ciudad de Panamá, Panamá
javier.sanchezgalan@utp.ac.pa

Abstract

In Panama, the large number of cars on the roads and random traffic events produce constant and extensive traffic jams. These issues are usually not solved even when the authorities plan and move forward with the construction of more traffic lanes. This work proposes the development of a system that allows the visualization of information published on social media about traffic incidents. For this task, different Python libraries are used for social network information extraction (Selenium and Tweepy), analysis (sklearn, Pandas, fuzzy wuzzy, etc), data visualization (Plotly, Dash and WordCloud) and a geocoding method to obtain the approximate location of obtained posts. Results show that approximately 180 thousand tweets have been obtained since 2014, the preparation of data for the classification models (detection of tweets corresponding to traffic incidents) has started and the graphical interface has been developed. This system has advantages such as speeding up the detection and visualization of traffic incidents, which can be of great help to the country's traffic authorities.

Keywords: Incidents, traffic, social networks, classification, geocoding.

RESUMEN

En Panamá, la gran cantidad de carros en las vías y eventos aleatorios de tráfico producen embotellamientos constantes y extensos. Estos problemas no son solventados aun cuando las autoridades planifican y avanzan con la construcción de más carriles para el flujo vehicular. Se propone desarrollar un sistema que permita visualizar información publicada en las redes sociales acerca de incidentes de tráfico. Para esta tarea, diferentes librerías de Python son utilizadas para la extracción de información de redes sociales (Selenium y Tweepy), análisis (sklearn, Pandas, fuzzy wuzzy, etc), visualización de datos (Plotly, Dash y WordCloud) y un método de geocodificación para obtener la localización aproximada de las publicaciones obtenidas. Los resultados muestran que se han obtenido aproximadamente 180 mil tweets desde 2014. Además, se ha iniciado con la preparación de datos para los modelos de clasificación (detección de tweets correspondientes a incidentes de tráfico) y se ha desarrollado la interfaz gráfica. Este sistema presenta ventajas como la agilización y la rapidez de detección y visualización de incidentes de tráfico, que pueden ser de gran ayuda para las autoridades de tránsito del país.

Palabras claves: Incidentes, tráfico, redes sociales, clasificación, geocodificación.

1. INTRODUCCIÓN

Los incidentes de tráfico representan cualquier situación inusual en las vías de tránsito como las colisiones, la obstrucción de las vías o eventos que requieran el uso de las calles. Las congestiones en el tráfico de cualquier ciudad se vuelve un problema social, económico y hasta ambiental a largo plazo [1]. Muchos de estos eventos de tránsito, no son reportados o presentan largos tiempos antes de poder ser detectados por las autoridades encargadas del tránsito.

Por otro lado, en casos en partes de la ciudad donde no existe la estructura de sensores para el monitoreo del tráfico, las redes sociales basadas en localización tienen un gran impacto, ya que pueden aportar reportes en tiempo real [2]. Las redes sociales son un canal activo que son de mucha utilidad para el intercambio de información de incidentes de tráfico y hechos que afectan la movilidad, y por esta razón, la ATTT monitorea constantemente diferentes cuentas en las redes sociales dedicadas a compartir contenido sobre el estado del tráfico que reciben de los usuarios que se encuentran en vía y en tiempo real. De esta manera, las redes sociales significan un complemento a la limitada red de sensores y semáforos inteligentes que no solo existen en la ciudad de Panamá, sino en todo el país.

A. Propuesta

El objetivo de esta propuesta es desarrollar un sistema que permita visualizar información publicada en las redes sociales acerca de incidentes de tráfico en la ciudad de Panamá y áreas de alto tránsito mediante distintas librerías de Python para la extracción, análisis y visualización de datos.

Se propone el desarrollo un *backend* mediante herramientas como Selenium y Tweepy para obtener las publicaciones históricas de Twitter sobre incidentes de tráfico y también en tiempo real; el almacenamiento de datos en una base de datos PostgreSQL; la limpieza y preprocesado de texto con Pandas [3] y re; sklearn [4] para el entrenamiento y evaluación de modelos de clasificación; GeoPandas [5] y fuzzy wuzzy para aplicar un proceso de “geocoding” para obtener una posición geográfica aproximada del incidente y guardarla como referencia espacial. Además, se propone el desarrollo de un *frontend*, el cual permite visualizar la información extraída de los incidentes en un tablero (dashboard) creado con la librería Dash de Python y las gráficas generadas con Plotly. En la Figura 1, se muestra un diagrama esquemático del sistema propuesto.

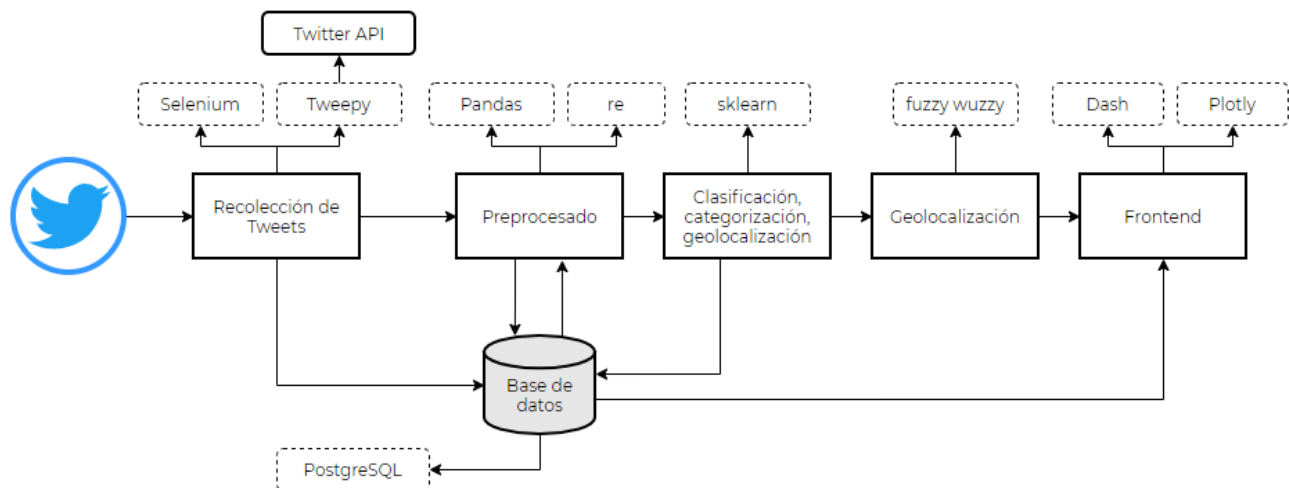


Figura 1. Diagrama del sistema.

El sistema que se propone se orienta a la sección de tránsito en Panamá, y por ende, esto crearía un impacto en la calidad de vida de los panameños a través de beneficios como mayor fluidez de tráfico, menores tiempos en embotellamientos y tiempos de tránsito, especialmente en la sección urbana. También, una rápida reacción a los incidentes de tránsito puede salvar vidas al ocurrir accidentes involucrando conductores y/o peatones porque la detección de estos incidentes ocurriría en un menor tiempo y por ende, una rapidez en el tiempo de atención.

2. MÉTODO

A. Exploración inicial

El estudio inicial del proyecto se basa en explorar las publicaciones sobre incidentes de tráfico en Twitter con relación a la cuenta de *@TraficoCPanama* y distintos noticieros o cuentas dedicadas a contenido sobre el tráfico en Panamá.

B. Recolección de datos

La obtención de los datos se divide en dos partes: la recolección de los datos históricos sobre incidentes de tráfico y la transmisión de datos en tiempo real. Se utiliza la librería Selenium que es una herramienta de extracción de datos mediante automatización, ya que mediante la API de Twitter solo permite obtener los tweets de los últimos 7 días y este se utiliza entonces para la transmisión de los tweets en tiempo real junto a la librería Tweepy.

Respecto a la conexión al API de Twitter para la transmisión en tiempo real, tweepy fue creada para comunicarse efectivamente con el API de Twitter; establece la autenticación correctamente y contiene los métodos adecuados para la obtención de recursos de Twitter, además de su extensiva documentación y contribución abierta de gran cantidad de usuarios.

C. Preprocesamiento de datos

El objetivo principal del preprocesamiento de los datos es convertir el texto obtenido de los tweets en una estructura léxica propia para la clasificación del incidente. Esto incluye la eliminación de: saltos de línea, enlaces, caracteres especiales, menciones (@), hashtags (#), espacios múltiples, palabras vacías y la conversión del texto a minúsculas.

D. Clasificación y Categorización

El propósito de la clasificación es determinar si el tweet corresponde a un incidente y la categorización, a qué tipo de incidente corresponde mediante diferentes palabras claves y técnicas de expresiones regulares. Para el modelo de predicción de incidentes, se probarán algunos clasificadores que han sido utilizados en diversos modelos de procesamiento de texto [6]–[8]: Support Vector Machines (SVM), Clasificador Naive Bayes, Random Forest y Modelos basados en el aumento de gradiente extremo (XGB).

E. Geocoding

Para el geocoding se planea obtener un punto de geolocalización para cada tweet mediante procedimientos de comparaciones difusas de texto (fuzzy matching) entre el contenido del tweet y los nombres de los lugares en la base de datos espacial. En la Figura 2 se muestra un flujograma esquemático que explica el procedimiento de geocoding propuesto.

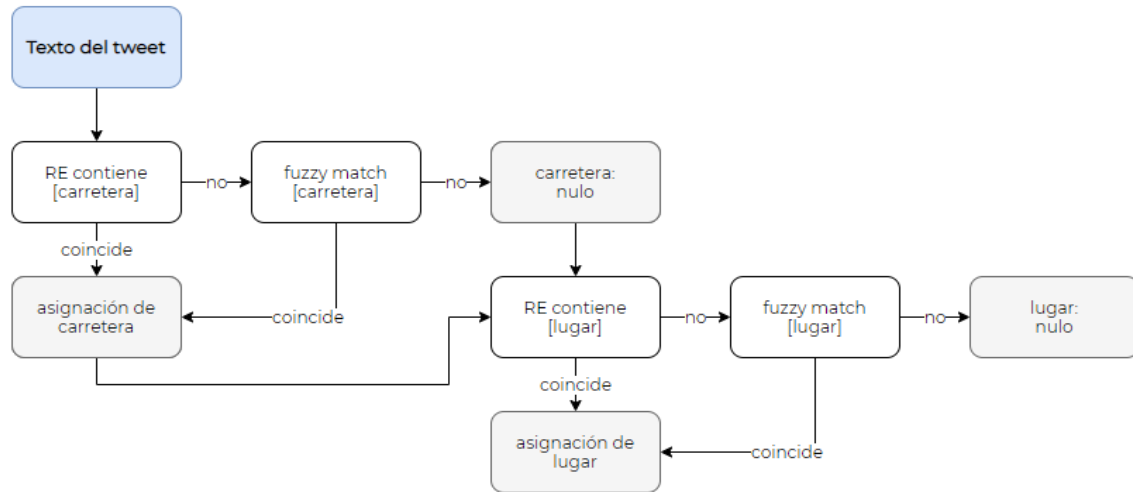


Figura 2. Procedimiento de geocoding.

En este proceso resultan dos campos: carretera y lugar, para luego realizar un proceso de geolocalización (intersecciones entre las áreas o puntos intermedios) para obtener un punto aproximado para la ubicación del tweet.

F. Desarrollo del Frontend

El *frontend* implica la creación de la interfaz en la que se visualizarán los datos obtenidos y procesados en los módulos del backend mediante la librería Dash de Python, que permite desarrollar dashboards con controles interactivos y elementos HTML.

En el dashboard se desplegarían los tweets más recientes obtenidos en tiempo real, gráficas de frecuencias de palabras (con la librería WordCloud), reportes por fecha y tipo de incidente y un mapa con los puntos de geolocalización de los tweets (graficado con Plotly).

3. RESULTADOS

Con la extracción de tweets realizado con Selenium, se lograron obtener alrededor de 180 mil tweets hasta el año 2021. En la Figura 3, se muestra la cantidad de tweets obtenidos de la búsqueda para los años comprendidos entre 2014 y lo que va del 2021. Los diferentes meses de cada año se representan en colores distintos.

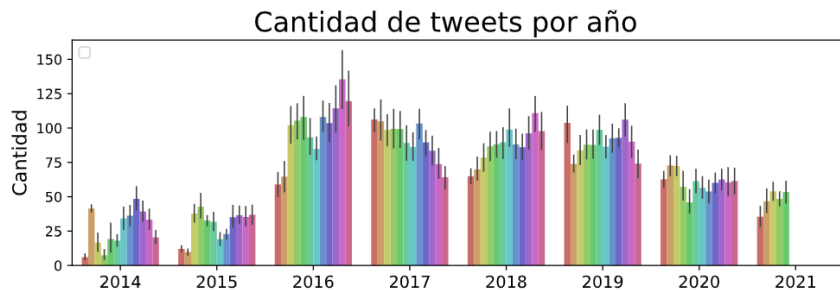


Figura 3. Cantidad de tweets obtenidos por año y mes.

En la interfaz que se desarrolló, se desplegó el mapa con las ubicaciones obtenidas en el proceso de geocoding, se muestran los tweets obtenidos en tiempo real y se crearon algunas gráficas adicionales para el análisis de los datos de tráfico.

Visualización de incidentes de tráfico

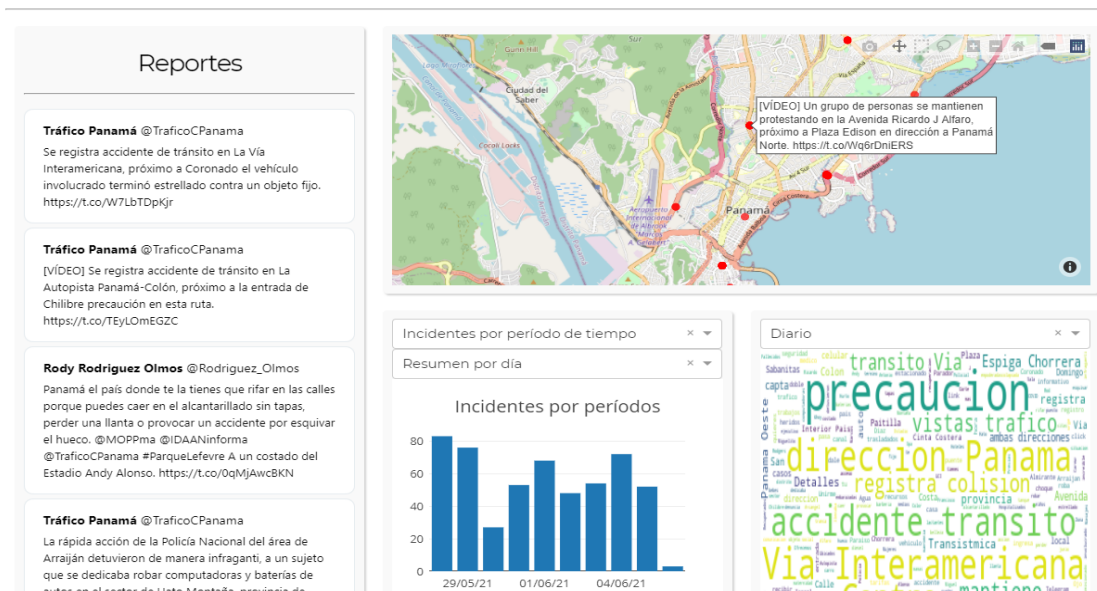


Figura 4. Interfaz de visualización de incidentes de tráfico.

En el proceso de geocoding, se realizaron pruebas preliminares con la librería fuzzy wuzzy para hacer comparaciones difusas entre 1000 tweets de prueba y las carreteras de la ciudad de Panamá. Se utilizaron los centroides para cada carretera para ubicar un punto específico en el mapa en lugar de un área (polígono), como se puede observar en la Figura 4. Por esta razón, hay puntos específicos que se superponen, sin embargo, esto se planea mejorar en los trabajos consiguientes.

4. CONCLUSIONES

Esta propuesta se considera de alta importancia ya que permite detectar los incidentes de tráfico de manera centralizada y en tiempo real, con esto, ayudando al personal de la ATTT a resolver y hasta evitar problemas en el tránsito con una solución tecnológica de bajo costo. Otro beneficio crucial de la implementación de esta propuesta es que la detección de accidentes de tránsito en un menor tiempo contribuirá a una rápida atención a incidentes que involucren la salud y la vida de las personas.

Dentro de las actividades futuras en el desarrollo del proyecto aún se debe realizar la clasificación y categorización mediante la limpieza de los datos obtenidos hasta el momento y el uso de expresiones regulares. Se espera evaluar los diferentes modelos de clasificación propuestos y cumplir con al menos un 90% de precisión en la detección de tweets con incidentes de tráfico.

También se ha desarrollado en un método de geocoding aproximado, sin embargo, futuramente se desea mejorarlo para obtener la geolocalización más exacta de los tweets. Además, también se desean crear más reportes incluidos tanto en las gráficas como en el mapa, como reportes por área, las ubicaciones que requieren más atención, los filtros según tipo de incidente o palabras claves.

Referencias

- [1] A. Salas, P. Georgakis, and Y. Petalas, "Incident Detection Using Data from Social Media," in *IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 745–749.
- [2] T. H. Silva *et al.*, "Urban Computing Leveraging Location-Based Social Network Data: a Survey," Nov. 2018. Accessed: Jan. 07, 2021. [Online]. Available: <https://pessoal.dainf.ct.utfpr.edu.br/thiagoh/papers/urbcompSurveyACM-silva.pdf>.
- [3] W. McKinney, "Data Structures for Statistical Computing in Python," *Proc. 9th Python Sci. Conf.*, pp. 56–61, 2010, doi: 10.25080/MAJORA-92BF1922-00A.
- [4] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.
- [5] K. Jordahl *et al.*, "geopandas/geopandas: v0.8.1," Jul. 2020, doi: 10.5281/ZENODO.3946761.
- [6] U. Malik, "Text Classification with Python and Scikit-Learn," Feb. 02, 2019. <https://stackabuse.com/text-classification-with-python-and-scikit-learn> (accessed Jun. 08, 2021).
- [7] M. F. Zafra, "Text Classification in Python," *Towards Data Science*, Jun. 15, 2019. <https://towardsdatascience.com/text-classification-in-python-dd95d264c802> (accessed Jun. 08, 2021).
- [8] S. Bansal, "A Comprehensive Guide to Understand and Implement Text Classification in Python," Apr. 23, 2018. <https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/> (accessed Feb. 17, 2021).

Autorización y Licencia CC

Los autores autorizan a APANAC XVIII a publicar el artículo en las actas de la conferencia en Acceso Abierto (Open Access) en diversos formatos digitales (PDF, HTML, EPUB) e integrarlos en diversas plataformas online como repositorios y bases de datos bajo la licencia CC:

Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

Ni APANAC XVIII ni los editores son responsables ni del contenido ni de las implicaciones de lo expresado en el artículo.