UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE QUITO

CARRERA DE COMPUTACIÓN

ANÁLISIS DE TWEETS COMO FUENTE DE INFORMACIÓN PARA MEJORAR LAS ESTRATEGIAS DE PREVENCIÓN EN SINIESTROS DE TRÁNSITO EN QUITO

Trabajo de titulación previo a la obtención del
Título de Ingeniero en Ciencias de la Computación

AUTOR: JEAN PIERRE CASA LUMBI
TUTOR: JULIO RICARDO PROAÑO ORELLANA

Quito - Ecuador
2024

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Yo, Jean Pierre Casa Lumbi con documento de identificación N.º 2200207476; manifiesto que:

Soy autor y responsable del presente trabajo; y, autorizo a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Quito, 6 de febrero del 2024

Atentamente,

Jean Pierre Casa Lumbi
2200207476

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Yo, Jean Pierre Casa Lumbi con documento de identificación No. 2200207476, expreso mi voluntad y por medio del presente documento cedo a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que soy autor del Artículo Académico: "Análisis de tweets como fuente de información para mejorar las estrategias de prevención en siniestros de tránsito en Quito", el cual ha sido desarrollado para optar por el título de: Ingeniero en Ciencias de la Computación, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribo este documento en el momento que hago la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Quito, 6 de febrero del 2024

Atentamente,

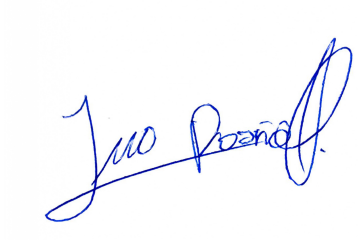
Jean Pierre Casa Lumbi
2200207476

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Julio Ricardo Proaño Orellana con documento de identificación N° 0103909412, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: ANÁLISIS DE TWEETS COMO FUENTE DE INFORMACIÓN PARA MEJORAR LAS ESTRATEGIAS DE PREVENCIÓN EN SINIESTROS DE TRÁNSITO EN QUITO, realizado por Jean Pierre Casa Lumbi, con documento de identificación N.º 2200207476, obteniendo como resultado final el trabajo de titulación bajo la opción Artículo Académico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Quito, 6 de febrero del 2024

Atentamente,



Ing. Julio Ricardo Proaño Orellana, MSc.
0103909412

Análisis de tweets como fuente de información para mejorar las estrategias de prevención en siniestros de tránsito en Quito

1st Jean Pierre Casa Lumbi
Ingeniería en las Ciencias de la Computación
Universidad Politécnica Salesiana
Quito, Ecuador
jcasal@est.ups.edu.ec

2nd Julio Ricardo Proaño Orellana
Ingeniería en las Ciencias de la Computación
Universidad Politécnica Salesiana
Quito, Ecuador
jproanoo@ups.edu.ec

***Resumen*—Los siniestros de tránsito son un problema grave que causa lesiones y discapacidades permanentes, así como la muerte de millones de personas en todo el mundo cada año. Estudios han demostrado que las redes sociales son una fuente de información en especial Twitter. Mediante la extracción de datos que proporcionan los tweets como el usuario, el lugar, la fecha, números de lesionados y muertos. En el presente trabajo se propuso realizar un análisis de tweets como fuente de información para mejorar las estrategias de prevención en siniestros de tránsito en Quito, aplicando dos metodologías, observacional para recopilar y analizar información ya existente en los tweets; y descriptivo en la cual servirá para describir y analizar patrones, tendencias y ubicaciones de los siniestros de tránsito. Se descubre que la mayor incidencia de siniestros de tránsito ocurre en la Avenida Simón Bolívar, siendo esta la calle principal con más lesionados y muertos reportados en la plataforma Twitter. Este hallazgo resalta la importancia de implementar medidas de seguridad vial en esta vía con el fin de reducir el número de siniestros de tránsito en Quito.**

***Palabras Clave*—Twitter, API, Análisis de tweets, Siniestros de tránsito, Tweets, Patrones y tendencias, Observación y análisis, Extracción de datos**

***Abstract*—Road crashes are a serious problem that causes injuries and permanent disabilities, as well as the death of millions of people around the world every year. Studies have shown that social networks are a source of information, especially Twitter. By extracting data provided by tweets such as user, location, date, numbers of injured and dead. In this work we proposed to analyze tweets as a source of information to improve prevention strategies in traffic crashes in Quito, applying two methodologies, observational to collect and analyze existing information in tweets; and descriptive in which will serve to describe and analyze patterns, trends and locations of traffic crashes. It is found that the highest incidence of traffic accidents occurs on Avenida Simón Bolívar, being this the main street with more injuries and deaths reported on the Twitter platform. This finding highlights the importance of implementing road safety measures on this road in order to reduce the number of traffic accidents in Quito.**

***Keywords*—Twitter, API, Tweet analytics, Traffic fatalities, Tweets, Patterns and trends, Observation and analysis, Data mining**

I. Introducción

La llegada del Internet ha revolucionado la comunicación y ha tenido un impacto significativo en la sociedad. Antes, la gente dependía de los periódicos y las noticias para obtener información, pero con el Internet se ha facilitado el intercambio de información [1]. Una de las redes sociales más destacadas es Twitter, que fue creada en 2006. Twitter permitió compartir información de forma breve y precisa, limitando los mensajes a 280 caracteres llamados tweets [2]. Twitter se ha convertido en una fuente inagotable de información y puede utilizarse como material científico en donde se han realizado diversos estudios que han demostrado la utilidad de analizar tweets para comprender y predecir diferentes fenómenos [3]. La extracción de datos se refiere al proceso de obtener información tanto estructurada como no estructurada. Esta información pasa por un proceso de transformación y limpieza de datos mediante diversas técnicas que, en última instancia, permiten garantizar la calidad y coherencia de los datos antes de su análisis [4]. Los siniestros de tránsito son sucesos imprevistos que involucran vehículos en movimiento que producen daños materiales, heridos o fallecidos. Pueden ser causados por exceso de velocidad, incumplimiento de normas, estado del conductor o condiciones peligrosas de la vía [5].

Mundialmente, segun informes de Centers for Disease Control and Prevention (2020). aproximadamente 1,35 millones de vidas se pierden anualmente en accidentes viales. Diariamente, alrededor de 3700 personas fallecen en colisiones que involucran diversos medios de transporte, peatones, bicicletas, automóviles, camiones, motocicletas o autobuses. La mayoría de las víctimas son ciclistas, peatones y ciclistas. Se estima que los accidentes de tráfico ocupan el octavo puesto entre las principales razones de muerte en todas las franjas de edad, siendo la causa primordial de mortalidad en niños y jóvenes de 5 a 29 años. Hoy en día fallecen más personas por accidentes que a causa del VIH/SIDA. Además, se prevé que entre 2015 y 2030, las lesiones mortales y no mortales por accidentes representarán un costo económico aproximado de $1,8 billones

de dólares en la economía mundial, equivalente a un impuesto anual del 0,12% sobre el Producto Interno Bruto global [6].

Dentro del cantón Quito,según informes del El Instituto Nacional de Estadística y Censos – INEC, para el primer y segundo trimestre del 2022, se registró un total de 1014 lesionados y 137 fallecidos, víctimas de accidentes de tránsito, lo cual corresponde a un 11,24% de lesionados y un 12,41% de fallecidos a nivel nacional, siendo Quito el segundo cantón del país, solo después de Guayaquil, con mayor incidencia de accidentes de tránsito registrados [7].

Partiendo de lo antes mencionado el presente proyecto propone resolver el problema: ¿Cómo se ha utilizado el análisis de Twitter como fuente de información para informar y mejorar las estrategias de prevención en siniestros de tránsito en Quito? La elección del uso de tweets como base para la recopilación de datos es debido a la masiva cantidad de estos y su creciente popularidad en la última década. La extracción de datos de tweets relacionados con siniestros de tránsito en Quito proporcionará información valiosa sobre las circunstancias, ubicaciones, causas y efectos de estos eventos. Este análisis permite identificar patrones, tendencias y factores de riesgo asociados con los siniestros de tránsito en la ciudad. Además, se pueden detectar áreas problemáticas específicas y evaluar la eficacia de las medidas de prevención implementadas [4] [3].

De acuerdo con lo anterior, se utilizaron las metodologías descriptivo y observacional con la finalidad de realizar un análisis con los datos obtenidos vinculados a esta problemática [8] [9]. Posteriormente, se presentaron gráficamente estos hallazgos utilizando las herramientas adecuadas.Finalmente, toda la información es almacenada en una base de datos documental para facilitar su uso en futuros proyectos relacionados con la mejora de estrategias de prevención en siniestros de tránsito en Quito.

II. Metodología

Para el desarrollo de este artículo se utilizó la metodología de análisis descriptivo y análisis observacional. Los cuales son bastante conocidas dentro de análisis de datos. En referencia al método observacional esta investigación se centra en el análisis de Twitter como fuente de información. En este caso se recopilan y analizan información ya existente en los tweets. De igual forma el método descriptivo servirá para describir y analizar patrones, tendencias y ubicaciones de los siniestros de tránsito reportados en Quito a partir de la información recopilada en los tweets, solo se quiere obtener una comprensión detallada de los datos recopilados [8] [9].

Con el fin de lograr esto, se llevan a cabo los siguientes pasos: i) Recopilación de datos: Utilizando la API de Twitter se realiza una selección de datos sobre siniestros de tránsitos mediante los siguientes criterios de selección, hashtag, ubicación (dentro de Quito), fecha y hora. Se almacena en una base de datos documental, ii) Selección, Limpieza, transformación: Una vez seleccionados los datos, se aplicarán técnicas de transformación y limpieza de los datos para extraer la información más relevante, iii) Interpretación y visualización: En esta etapa, se llevan a cabo análisis descriptivos y visualizaciones para identificar patrones, tendencias y ubicaciones de los siniestros de tránsito. Mediante herramientas gráficas.

A. Recolección de datos

Para la recopilación de datos se utilizó la API Twitter Scraper, este programa posibilita la extracción ilimitada de tweets públicos a gran escala, sin imposiciones ni restricciones. Extrae y transforma los tweets públicos en una API JSON lista para su empleo. Tiene la capacidad de recopilar información de tweets desde el 21 de marzo de 2006 (fecha del primer tweet en Internet) hasta 23 de octubre de 2023, prescindiendo de la obligación de contar con una cuenta de desarrollador de Twitter [10].

En la Fig. 1 se muestra el diagrama de flujo del proceso de recopilación de datos realizado en este estudio. Se inició importando las librerías necesarias como pickle, request y json (paso 1). Se crearon dos listas vacías tweet y user. Abriendo el archivo tweet.pkl en modo escritura binaria, se serializó y almacenó la lista tweet mediante pickle.dump. De igual forma, abriendo el archivo user.pkl se realizó lo mismo (paso 2). Se creó un parámetro llamado término en el cual se colocaron palabras clave que se deseaba buscar, incluyendo usuarios, como se muestra en la Tabla I. Después para obtener la información utilizamos la API Twitter Scraper, se construyó un querystring con los siguientes parámetros: i) ”searchTerms”: término, ii) ”maxTweets”: número de tweets, iii) ”near”: ubicación, iv) ”lang”: idioma, v) ”fromDate”: fecha de inicio y vi) ”toDate”: fecha final y se definieron los headers con la API key y host. Se hizo una solicitud GET pasando los headers y el querystring donde finalmente, se imprimió como JSON el resultado devuelto por la API (paso 3). Posteriormente, se guardan las listas en los archivos existentes tweet.pkl y user.pkl que se cargaron las listas tweet y user previamente almacenadas en donde se itera cada vez que obtengamos información que se menciona en el paso 3, donde finalmente, se abren nuevamente los archivos .pkl en modo escritura binaria y se serializan las listas actualizadas con pickle.dump (paso 4). Finalmente para guardar en un archivo Excel, se abrieron los archivos .pkl en modo lectura binaria. Se convirtieron las listas a DataFrames de Pandas utilizando pd.DataFrame.from_dict . Se exportaron los DataFrames a archivos Excel usando el método .to_excel(), generando tweets.xlsx y usuarios.xlsx (paso 5).

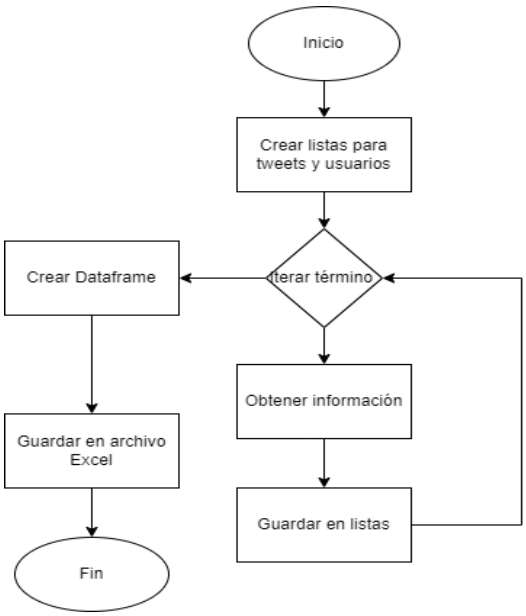


Fig. 1: Recolección de datos

Tabla I: TÉRMINOS

Usuarios	Términos Usados
@BomberosQuito	Av. Simón Bolívar
@radio_pichincha	accidente de transito
@elcomerciocom @QuitoNoticias1 @radioquitoec @TiempoRealEC @CDL_NOTICIAS @RadaresEC @AguantaMijin @EcuavisaInforma @RTSEcuador @tctelevision @el_telegrafo @AMT_Quito @ANT_ECUADOR	accidente quito

B. Pre-procesamiento de datos

En la Fig. 2 se muestra el diagrama de flujo que describe el proceso de curación de datos realizado en este estudio.

Se inició importando las librerías necesarias como spanish-nlp, unidecode y pandas (paso 1). Primero accedemos a los datasets tweets.xlsx y usuarios.xlsx utilizando pandas la función read_excel (paso 2). Se crea un objeto llamado tweets que tendrá la columna *full_text* del dataset tweets.xlsx, para transformar de texto a listas (paso 3). Como se mencionó anteriormente se utilizó la librería spanish-nlp lo cual realiza el proceso de nlp de forma automática pero, debemos configurar correctamente los parámetros del objeto de la clase llamada SpanishPreprocess que se muestra en la Tabla II a la forma que deseamos que será almacenada en el objeto sp y que posteriormente se aplicará en un bucle for que recorra cada tweet en la lista de tweets aplicado el preprocesamiento especificado que configuramos para ser almacenada en el objeto tweets las modificaciones de la columna *full_text* (paso 4). Para transformar el dataset, se crea un objeto llamado dataset que tendrá las columnas *created_at, full_text, retweet_count* para posteriormente anexar con la columna screen_name extraído del dataset usuarios.xlsx, guardando las modificaciones en un objeto llamado result (paso 5). Con el objeto result, procedemos a realizar la eliminación de duplicados con la función drop_duplicates (paso 6). Finalmente guardamos el dataframe final con los tweets limpios y transformados a un archivo Excel llamado tweets_curado.xlsx.

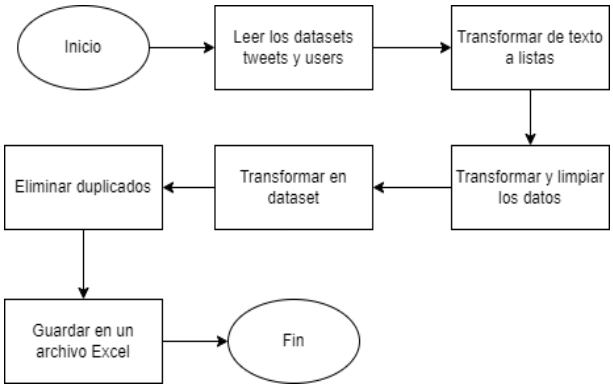


Fig. 2: Pre-procesamiento de datos

Tabla II: Parámetros de SpanishPreprocess

Parámetros	**Uso**
lower=True	Convierte el texto a minúsculas
remove_url=True	Elimina enlaces (URLs) del texto
remove_hashtags=True	Suprime los hashtags del texto
normalize_breaklines=True	Normaliza los saltos de línea en el texto
split_hashtags=False	No divide las palabras en los hashtags
remove_emoticons=True	Elimina los emoticonos del texto
remove_emojis=True	Remueve los emojis del texto
convert_emoticons=False	No convierte los emoticonos a su forma textual
convert_emojis=False	No convierte los emojis a su forma textual
normalize_inclusive_language=True	Normaliza el lenguaje hacia formas más inclusivas
remove_vowels_accents=True	Elimina acentos y diacríticos de las vocales en las palabras
remove_multiple_spaces=True	Elimina espacios en blanco consecutivos
remove_punctuation=True	Suprime signos de puntuación del texto
remove_unprintable=True	Elimina caracteres no imprimibles o especiales
remove_stopwords=False	No elimina las palabras comunes (stopwords) del texto
remove_numbers=False	No elimina los números del texto
lemmatize=True	Realiza la lematización de las palabras
stem=False	No realiza la reducción de palabras a sus raíces mediante el proceso de stemming
remove_html_tags=True	Elimina etiquetas HTML del texto

C. Extración de datos

En este procedimiento, se llevaron a cabo las siguientes etapas: i) aplicación del proceso de stemming, ii) recopilación de información sobre heridos y fallecidos, iii) obtención de datos de ubicación, y iv) inclusión en un conjunto de datos final.

*1) **Proceso de stemming**:* En la Fig. 3 se muestra el diagrama de flujo que describe el proceso de stemm realizado en este estudio. Se inició importando las librerías necesarias como pandas, nltk, stopwords, SnowballStemmer y word_tokenize (paso 1). Se lee el archivo Excel llamado tweets_curado.xlsx. con la función read_excel, se crea un objeto llamado text_list que tendrá la columna full_text transformado en lista, cabe recalcar que la columna full_text son los tweets (paso 2). Se crea una función llamada stem_documents que recibirá como parámetro un objeto llamado documents, se define el stemmer para el idioma español, se inicializa una lista vacía para documentos stemmeados llamada stemmed_documents, se crea un for que recorra el objeto documents el cual almacena los tweets para luego realizar un proceso de tokenización al documento y posteriormente stemmear cada proceso del documento finalizando con todos los tweets stemeados. Luego se almacenan en la lista vacía previamente creada llamada stemmed_documents (paso 3). Finalmente se guarda el objeto para realizar el siguiente otros procesamientos (paso 4).

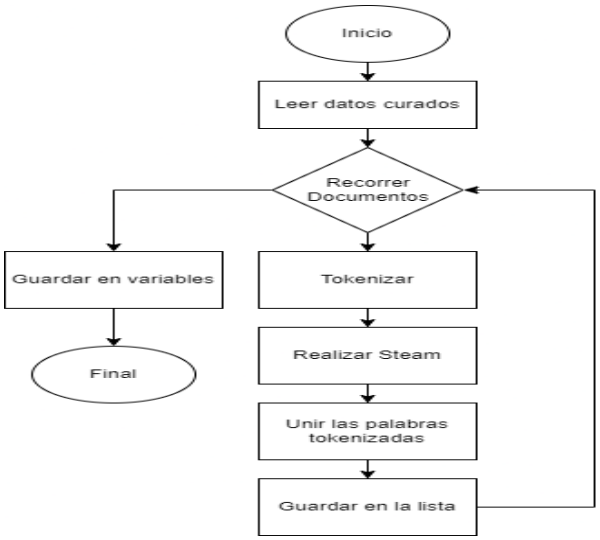


Fig. 3: Proceso de stemming

*2) **Extracción de datos sobre heridos y fallecidos**:* Para este proceso se utilizó la librería Word2number-es, permite convertir palabras en español a números y también proporciona funciones para transcribir expresiones numéricas escritas en texto a su representación decimal [11].

A continuación, en la Fig. 4 se muestra el diagrama de flujo que describe el proceso de extracción de datos sobre heridos y fallecidos realizado en este estudio. Se inició importando las librerías necesarias como re, word2number_es y numpy (paso 1). Después se cargan las diferentes listas de datos, en este caso texto de los tweets original y stemmed ya que necesitamos observar si existe palabras relacionadas con números o números en sí, luego se ubica el término de búsqueda con proceso de stemmed ya que se quiere buscar la información relevante en el texto en este caso numero de cierto término(paso 2). Luego recorremos los distintos textos y realizamos un proceso de split por cada texto para luego procesar dentro de una función con el término de búsqueda si existe digitos numericos o en cuyo caso observar si la palabra transformada es un dígito usando la librería word2number-es (paso 3). Este proceso realizarlo contando 3 palabras ala izquierda del texto primero y luego ala derecha esto con el fin que si en alguno de los 2 bucles de busqueda encuentra informacion guarde ese numero y lo devuelva , y si no existen datos de resultado 0 numeros. Finalmente, una vez que se obtiene el número de muertos en lista, fallecidos, heridos se suma los sinónimos de términos en este caso muertos y fallecidos, se obtiene el número total de muertos, se crea el dataset con los tweets , lista número de muertos , numero de heridos, se guarda la información en un excel (paso 4).

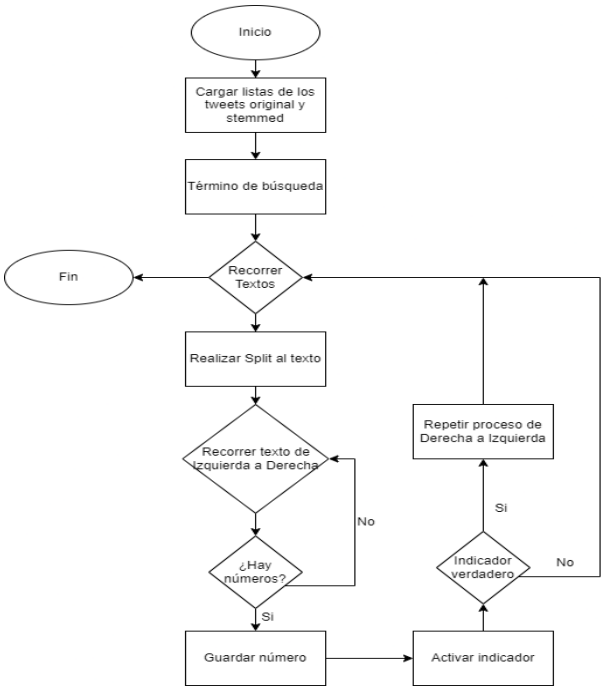


Fig. 4: Extración de heridos y muertos

*3) **Extracción de la ubicación**:* Para esta parte se utilizó la "API de Geocoding", permite convertir direcciones postales en coordenadas de latitud y longitud. Facilita ubicar puntos geográficos a partir de descripciones de lugares como ciudades, calles o códigos postales [12].

A continuación, en la Fig. 5 se muestra el diagrama de flujo que describe el proceso de extracción de ubicación realizado en este estudio. Se inició importando las librerías necesarias como time, requests, pandas, preprocess, nltk y stopwords (paso 1). Se realiza una lista de sinónimos relacionado a calles esto con el fin de optimizar la búsqueda de los tweets,

estos sinónimos se pueden visualizar en la Tabla III (paso 2). Se carga la información sobre los tweets (paso 3). Se hace un split al texto, ingresa el texto a un bucle for el cual recorre cada palabra y mediante una condición que verifica la existencia de las palabras en la lista de sinónimos, en el caso de ser verdadero se toma la posición de esa palabra en coincidencia hasta 5 posiciones delante de esa palabra ya que nos interesa solo la información relevante para finalmente retornar el texto con la calle (paso 4). Luego con la API Geocoding, se especifica los parámetros de búsqueda , por dirección , el api key , el resultado que sea de tipo ROOFTOP es decir el más preciso , luego pasamos los parámetros y esto nos devolverá un JSON para finalmente obtener los datos relacionados a latitud , longitud y dirección, si no hay datos sobre la calle o localidad mostrará que no existe datos (paso 5). Finalmente, guardamos en listas diferentes los datos para luego ser concatenados en un solo dataframe (paso 6).

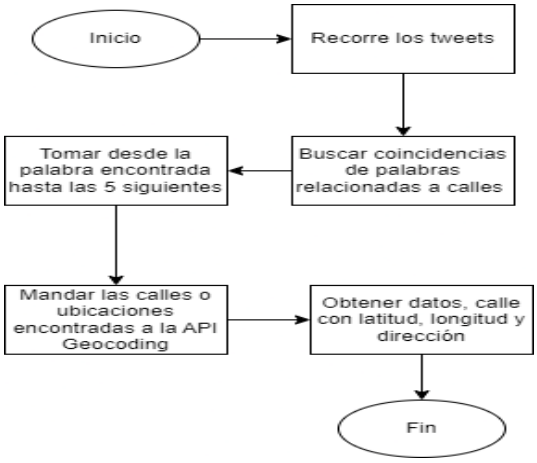


Fig. 5: Extración de ubicación

Tabla III: Sinónimos de calles usadas

Sinónimos usados
av, avenida, calle, via, sector, barrio, ubicacion, bajada, subida, paso, pasaje, peatonal, ronda, redondel, sector, bulevar, travesia, zona y lugar

D. Visualización de datos

La exhibición de gráficos desempeña un papel crucial en cualquier investigación, facilitando una representación más nítida y precisa de los resultados obtenidos. Para lograrlo, es necesario llevar a cabo las siguientes etapas.

1) Se importaron las librerías pandas, numpy, duckdb, folium, matplotlib.pyplot y cartopy para manipular datos y crear visualizaciones gráficas, respectivamente.
2) Se crearon dos mapas geográficos utilizando folium, cada uno mostrando la ubicación dentro de la ciudad de Quito y la ubicación con más frecuencias de siniestros de tránsito.
3) Se crearon gráficos utilizando matplotlib.pyplot para representar:
   a) La frecuencia de calles con más siniestros de tránsito.
   b) La cantidad de heridos y muertos.
   c) La cantidad de muertos con respecto a la calle.
   d) La cantidad de heridos con respecto a la calle.
   e) El ratio de muertos/heridos por calle.
   f) Palabras más usadas.
   g) Mapa de calor de la latitud y longitud.

En la Tabla IV se muestran las librerías y funciones que se utilizaron para obtener gráficas tipo pastel, barra, barras horizontales, mapa de calor para el respectivo análisis.

Tabla IV: Librerías y funciones

Librería	Función	Utilidad
pandas	read_excel()	Acceder a los archivos excel y crear dataframe
	drop()	Elimina columnas, filas específicos
	to_list()	Convertir en una serie o dataframe
	index()	Accedere, manipular o obtener el índice de un objeto
folium	map()	Crear mapa básico, visualización dinámica
	Marker()	Agrega puntos específicos en el mapa
	save()	Guarda un objeto de mapa en un archivo HTML
duckdb	query()	Realiza operaciones de extracción y manipulación de datos
numpy	random()	Genera números aleatorios
matplotlib.pyplot	pie()	Crea gráficos de tipo pastel, representan la distribución de diferentes categorías en un conjunto de datos
	bar()	Crea gráficos de tipo barras, representan la comparación de datos en diferentes categorías
	barh()	Crea gráficos de barras horizontales, representan la comparación de datos en diferentes categorías
	figure()	Crea un lienzo en el que se puede trazar gráficos
	scatter()	Crea gráficos de dispersión

III. Resultados y Análisis

A continuación, se exponen los principales resultados obtenidos del estudio realizado.

A. Principales calles con mayor frecuencia de siniestros de tránsito

Como se observa en la Fig. 6, la distribución porcentual de siniestros de tránsito en las principales calles en Quito, evidencia una marcada concentración en algunos puntos específicos. La calle Avenida Simón Bolívar presenta la mayor frecuencia de siniestros de tránsito, representando el 69% del total de casos registrados en el periodo de estudio. Le siguen en orden descendente las calles Avenida Mariscal Sucre con 10.5%, Avenida Panamericana Nte. con 6.2%, Avenida 6 de

Diciembre con 5.6%, Avenida Oswaldo Guayasamín con 5%, y Avenida Pedro Vicente Maldonado con 3.7%.

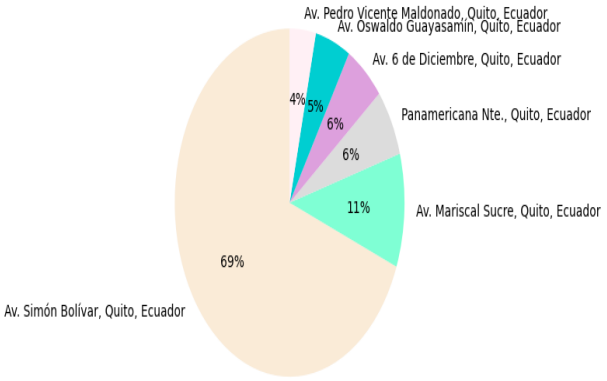


Fig. 6: Calles con mayor frecuencia de siniestros de tránsito

B. Tasas de mortalidad y morbilidad

En la Fig. 7 se muestran las menciones en Twitter sobre heridos y muertos por siniestros de tránsito en Quito, recolectadas mediante la API gratuita Twitter Scraper2. Tras aplicar técnicas de extracción de datos para obtener información relevante, se identificó un total de 7088 datos en los tweets. De estos, 6259 datos corresponden a lesionados y 839 datos a fallecidos. Se observa así una preponderancia de menciones a heridos comparado con fallecidos, siendo los reportes sobre lesionados los de mayor frecuencia en los tweets analizados.

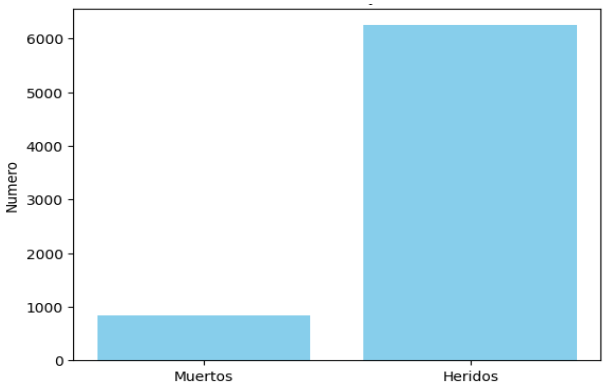


Fig. 7: Números de mortalidad y morbilidad

C. Tasas de mortalidad por principales calles

Como se observa en la Fig. 8, se muestran los datos de mortalidad con respecto a las intersecciones con mayor frecuencia de siniestros de tránsito identificadas en la Fig. 6. La Avenida Simón Bolívar presenta la tasa más alta de mortalidad, con 81 datos mencionados. Le siguen en orden descendente la Avenida Panamericana con 12 datos, la Avenida Mariscal Sucre con 6 datos, la Avenida Pedro Vicente Maldonado también con 6 datos, la Avenida 6 de Diciembre con 1 dato y la Avenida Oswaldo Guayasamín sin menciones. Cabe recalcar que estos números de datos son bajos debido a que se seleccionaron únicamente las intersecciones con mayor frecuencia de reportes en los tweets analizados.

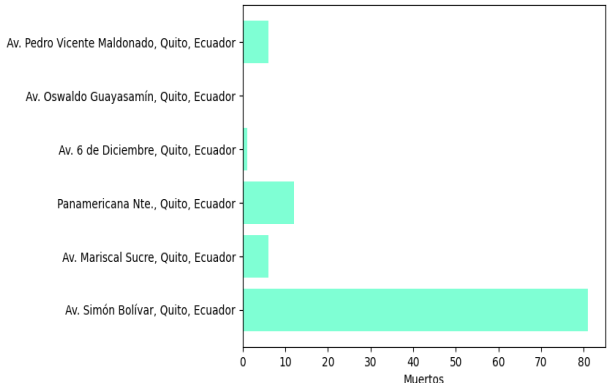


Fig. 8: Números de muertos por principales calles

D. Tasas de morbilidad por principales calles

Para este análisis, cabe mencionar que se utilizó la Fig. 6, ya que solo se centrará en cuantificar los datos de morbilidad mencionados en las intersecciones con mayor frecuencia de siniestros de tránsito en Quito. En la Fig. 9, se puede observar que la Avenida Simón Bolívar presenta la tasa más alta de morbilidad, con 168 datos mencionados. Le siguen en orden descendente la Avenida Panamericana Norte con 56 datos, la Avenida Pedro Vicente Maldonado con 40 datos, la Avenida Mariscal Sucre con 31 datos, la Avenida Oswaldo Guayasamín con 2 datos y la Avenida 6 de Diciembre con 1 dato mencionado.

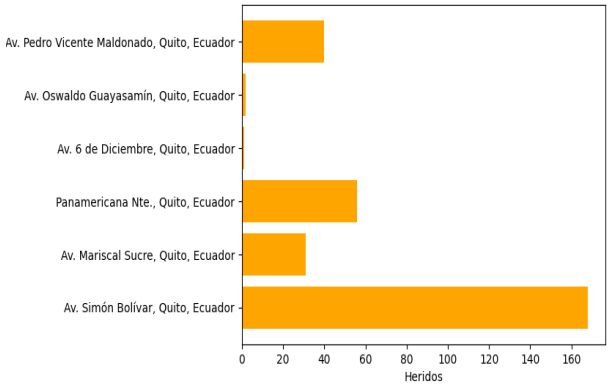


Fig. 9: Números de heridos por principales calles

E. Ratio de fallecidos/heridos por mayor frecuencia de intersecciones

En la Fig. 10 se presenta la proporción de muertos y heridos en diferentes avenidas. Se puede afirmar que si las proporciones son iguales, el ratio es del 100%. Si la barra es pequeña, indica que hay más heridos que muertos; y si es grande, señala más muertos que heridos. Por lo tanto, se distingue que en la Avenida Pedro Vicente Maldonado con 15%, la Avenida Oswaldo Guayasamín con 0%, la Avenida Panamericana con 21%, la Avenida Mariscal Sucre con 19% y la Avenida Simón Bolívar con 48% hay mayor cantidad de heridos que fallecidos. En cambio, en la Avenida 6 de Diciembre con 100% se presenta la misma proporción de muertos y lesionados, pero dado que se presentan pocos datos

mencionados en los muertos y heridos que se extrajo de los tweets es indeterminado. Por lo tanto, los datos analizados en este estudio evidencian que en los siniestros de tránsito en Quito suele haber mayor cantidad de heridos que de muertos.

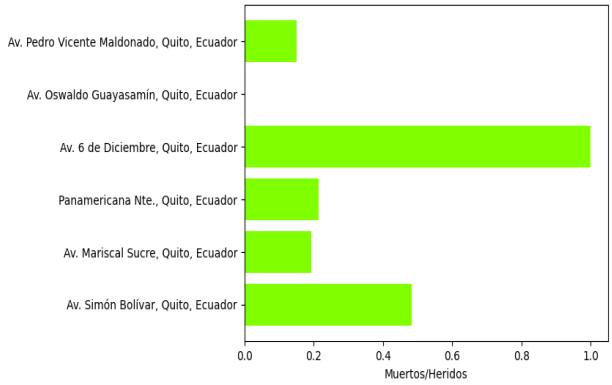


Fig. 10: Ratio de muertos y heridos por mayor frecuencia de calles

F. Frecuencia de palabras clave en los tweets

Para el análisis de la frecuencia de palabras, se extrajo las 30 palabras claves más usadas en los tweets, se recopilaron y procesaron un total de 25392 palabras claves relacionados con siniestros de tránsito en Quito. En la Fig. 11 se muestra la distribución de las palabras claves más mencionadas en los tweets. Se observa que las palabras clave "accidente", "transito", "amt_quito" fueron las más frecuentes, lo que sugiere que son un fuerte énfasis dentro de la discusión en Twitter sobre el tema de estudio. Estas frecuencias de palabras claves proporcionan una visión general de las tendencias que se discuten en la plataforma de Twitter.

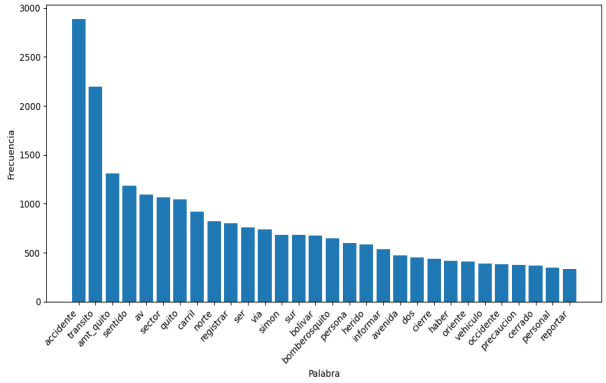


Fig. 11: Top 30 palabras claves usadas en los tweets

G. Concentración de Siniestros de Tránsito según Latitud y Longitud

En la Fig. 12, se muestran las distintas coordenadas en el mapa de calor. Las coordenadas más cercanas al punto representan la ciudad de Quito, mientras que las más alejadas son el resultado de errores del API de Geocoding, que incluyó calles de otros países o lugares. Este gráfico resulta útil para analizar la precisión del API al buscar calles y para identificar posibles errores dentro del conjunto de datos. Como se observa, la mayoría de los puntos se concentran en el centro, siendo pocos los puntos alejados.

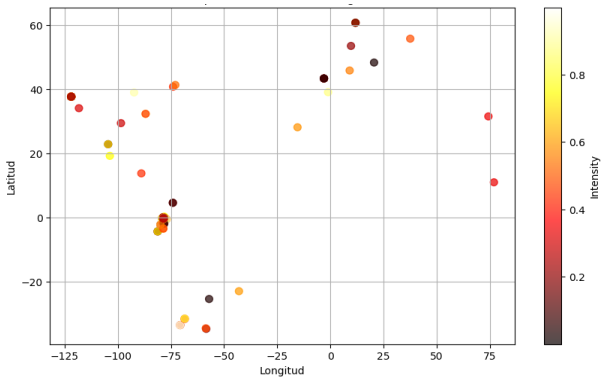


Fig. 12: Concentración de Siniestros de Tránsito según Latitud y Longitud

IV. Conclusiones

Como se puede observar, la gran tendencia de las redes sociales han proporcionado el intercambio de información de manera eficaz siempre siguiendo fuentes confiables, donde se destaca la red social Twitter que la mayoría de usuarios publican noticias relevantes como fue el caso de estudio, siniestros de tránsito en Quito. Para los analistas de datos,Twitter representa una fuente inagotable de información que mediante procesos de extracción de datos y pre-procesamiento se puede obtener datos de la forma que desea analizar. El objetivo principal de este trabajo es realizar un análisis de tweets como fuente de información para mejorar las estrategias de prevención en siniestros de tránsito en Quito. En este sentido, se destacan aspectos importantes sobre análisis de tweets y como este proporciona información tras pasar procesos de extracción de datos para posteriormente dar una estrategia de prevención para reducir los siniestros de tránsito que afectan en la ciudad de Quito.

En el periodo de este trabajo se obtuvo que la calle con mayor índice de siniestros de tránsito es la Avenida Simón Bolívar con 65%, los resultados sobre las tasas de mortalidad y morbilidad por mayor indice de calles radica la atención en la Avenida Simón Bolívar, el ratio sobre muertos/heridos se obtuvo que las Avenidas Simón Bolívar, Mariscal Sucre y Panamericana tienden a tener más heridos que muertos durante un siniestro de tránsito. Por lo tanto para reducir los siniestros de tránsito en Quito, en general se debe iniciar con educar a los conductores y peatones sobre la seguridad vial, esto se puede realizar a través de campañas que fomenten la concientización sobre la seguridad vial. También renovar e implementar nuevos moderadores de tráfico como señalización, resaltar rompevelocidades, colocar nuevos radares, nuevos centros de comisarías en puntos estratégicos y también colocar puestos de enfermería. Estas estrategias de prevención mencionadas podrán reducir los siniestros de tránsitos en Quito.

La red social Twitter se ha convertido en una poderosa herramienta para difundir información sobre incidentes que están ocurriendo en la ciudad de Quito como siniestros de tránsito. Muchos organismos de seguridad, policia de transito, bomberos y servicios de emergencia utilizan esta red social para informar sobre siniestros viales, rutas alternas y dar recomendaciones ya que al provenir de fuentes oficiales y confiables, los usuarios pueden estar al tanto de la situación y tomar decisiones para movilizarse de manera segura y así evitar mayores congestiones. Asimismo, las autoridades pueden utilizar Twitter como medio de difusión de campañas preventivas y consejos a los conductores sobre límites de velocidad, estado de ebriedad, cambios de normativas de tránsitos o nuevas tecnologías implementadas. Este es un recurso que bien utilizado puede ayudar a reducir los siniestros de tránsito en Quito.

En cuanto al uso de los métodos empleados, se destaca que el uso de API Twitter Scraper resultó muy valioso para extraer fuentes de información de esta red social, permitiendo filtrar la búsqueda por ubicación, hashtag, fecha y idioma. El uso de estas herramientas de acceso libre agiliza enormemente la recolección de datos relevantes. A si mismo, la aplicación de técnicas de transformación y limpieza de datos resultó fundamental para poder extraer la información más relevante para nuestro análisis, destacando la liberia spanish-nlp donde agilizo la tarea de eliminar los datos faltantes, eliminación de datos duplicados, eliminar caracteres especiales, emoticonos, hashtags, URLS. En definitiva, la dedicación a estas técnicas de transformación y limpieza de datos fue un paso clave para extraer la información necesaria, como ocurrió en este caso con respecto a los siniestros de tránsito en Quito.

REFERENCES

[1] A. M. Astobiza, "Internet y cognición social," /, vol. 33, p. 16, 2018.

[2] H. M. A. Luis Alberto Ortíz Palma, "Herramienta para la extracción y análisis de información obtenida de la red social twitter, como apoyo a los procedimientos: nuevo registro calificado y renovación de registros," /, vol. 18, p. 51–63, 2020.

[3] W. G. Yan Wang, Shangde Gao, "Investigating dynamic relations between factual information and misinformation: Empirical studies of tweets related to prevention measures during covid-19," *Journal of Contingencies and Crisis Management*, vol. 30, p. 427–439, 2022.

[4] S. Mishra, R. Rezapour, and J. Diesner, "Information extraction from social media: A hands-on tutorial on tasks, data, and open source tools," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, no. 4. Association for Computing Machinery, 2022, p. 5148–5151.

[5] J. B. B. L. P. P. J. D. Ana E., Congacha, "Caracterización de los siniestros viales en el ecuador," vol. 2, pp. 17–29, 2019.

[6] Erin Sauber-Schatz, Erin Parker, Michael Ballesteros. (2020) Road Traffic Safety. CDC Health Information for International Travel. [Online]. Available: https://wwwnc.cdc.gov/travel/yellowbook/2024/air-land-sea/road-and-traffic-safety

[7] Instituto Nacional de Estadística y Censos – INEC. (2022) Estadísticas de Transporte Siniestros de Tránsito Trimestral I y II Trimestre, 2022. Informe estadístico en formato PDF. [Online]. Available: https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Economicas/Estadistica%20de%20Transporte/ESTRA_2021/2022_ESTRA_SINIESTROS.pdf

[8] L. K. R. P. M. H. D. A. H. E. C. P. Peter Miksza, Julia T. Shaw, "Descriptive statistics," *online edn*, vol. /, p. 325–346, 2023.

[9] M. T. Anguera, A. Blanco-Villaseñor, J. L. Losada, and M. Portell, "Pautas para elaborar trabajos que utilizan la metodología observacional," *Anuario de Psicología*, vol. 48, no. 1, pp. 9–17, 2018.

[10] "Twitter Scraper," https://rapidapi.com/microworlds/api/twitter-scraper2/, 2022.

[11] " Word to Number ," https://pypi.org/project/word2number-es/.

[12] "API Geocogin," https://developers.google.com/maps/documentation/geocoding/?hl=es_419, /.