Prototipo de análisis de datos ciudadanos relacionados a contaminación ambiental

Trabajo Terminal No. _ _ _ _-_ _ _

*Alumnos: *Hernández Clemente Samantha, Medina Flores Susana, Olivares Conchillos Leonel*

Directores: Zagal Flores Roberto Eswart

**e-mail: shernandezc1404@alumno.ipn.mx*

Resumen - Las redes sociales generan una gran cantidad de datos no estructurados que se pueden utilizar como una fuente de información, en este caso tomaremos en cuenta las publicaciones referentes a eventos contaminantes que afectan la calidad del aire en Ciudad de México. En el presente Trabajo Terminal se planea realizar un prototipo de análisis de datos ciudadanos relacionados a contaminación ambiental, como puede ser tráfico ó pirotecnia, extrayendo información de redes sociales con un robot de extracción e implementando un modelo de minería de datos para su análisis, y así mostrar los resultados en un dashboard.

Palabras clave - Contaminación del aire, Minería de datos, Redes sociales.

1.- Introducción

La contaminación se ha vuelto cada vez más presente en nuestro día a día. La Ciudad de México no es una excepción. En el año 2018, la Secretaría del Medio Ambiente de la Ciudad de México (SEDEMA), en el informe anual de calidad de aire publicó que el índice de la calidad del aire, en la mayoría del año se presentó una mala calidad del aire, con 276 días que superaron los 100 puntos, el valor máximo indicado por la Norma Oficial Mexicana de salud (NOM) y sólo 89 días donde ningún contaminante superó dicha cifra [1].

La Procuraduría Ambiental y del Ordenamiento Territorial de la Ciudad de México (PAOT) es un organismo público descentralizado de la Administración Pública que una de sus funciones es defender los derechos de la población referente al medio ambiente [2]. La ciudadanía puede presentar una denuncia referente a daños o incumplimiento de las normas al medio ambiente. En lo transcurrido de este año 2021, ha recibido más de 9,000 denuncias, donde apenas el 1.3% son acerca de las emisiones del aire [2].

Algunas de las fuentes más conocidas de compuestos orgánicos volátiles (COV) y óxidos de nitrógeno que afectan a la calidad del aire son las emisiones de vehículos e industrias, aerosoles, aromatizantes, entre otros [3]. Estas sustancias son comunes en la vida diaria, pero también hay días específicos en los cuales se generan una gran cantidad de ozono dañino: en los eventos sociales. Hay eventos o festividades que ocurren a lo largo del año en los cuales la contaminación se incrementa de forma alarmante, por ejemplo, el 15 de septiembre, en la Ciudad de México se celebra el grito de la Independencia, desde hace varios años se acostumbra utilizar grandes cantidades de pirotecnia que provoca la liberación de sustancias tóxicas al aire, incluso hoy en día, terminando en incendios forestales o daño a infraestructura.

Esta clase de eventos también aumentan considerablemente el tráfico por varias rutas, ocasionando una fuerte concentración de sustancias perjudiciales para la calidad del aire. Sucede de forma similar en algunas marchas, que se apoyan de pirotecnia o productos similares que aumentan el factor contaminante en el área. La mayoría de esta información no se anuncia de forma oficial o no se toma en cuenta, por lo que, los ciudadanos buscan otros medios para dar a conocer e informarse acerca de estos eventos que dañan al medio ambiente: las redes sociales.

La comunicación a través de las redes sociales se ha vuelto de vital importancia en los últimos años, no solo como forma de entretenimiento sino como una alternativa para transmitir información, que toma los elementos, recursos y características de los medios tradicionales pero que incorpora un nivel de interacción más grande [4].

Existen datos en redes sociales que pueden ayudar a identificar denuncias de carácter social relacionadas al medio ambiente, por ejemplo, existen ciudadanos que reportan a las autoridades o comunidades vecinales situaciones que afectan la calidad del aire y que afectan a la convivencia social; esta información no ha sido analizada para encontrar una caracterización de estos eventos contaminantes que pudiera ayudar a encontrar situaciones de riesgo al ambiente

u otros seres vivos. Esta información es una fuente alternativa de denuncias ciudadanas que al tratarse datos no estructurados, no se toman en cuenta en el monitoreo del impacto ambiental por parte de las autoridades correspondientes.

El reto técnico a realizar es obtener, limpiar y analizar una gran cantidad de información no estructurada sobre la calidad del aire que nos arroje un valor sobre la contaminación en la Ciudad de México y nos permita generar datos más certeros sobre la contaminación a partir de las denuncias ciudadanas en las redes sociales.

Los proyectos similares que se han desarrollado se muestran en la Tabla 1:

SOFTWARE	CARACTERÍSTICAS	PRECIO EN EL MERCADO
[5] Buscador geosocial para monitoreo de polución del aire urbano.	Un mecanismo geosocial de búsqueda basado en palabras clave para registrar patrones espaciales de denuncias a la calidad del aire a partir de publicaciones de Twitter.	No tiene un precio definido por el momento.

Tabla 1. Resumen de proyectos similares.

2.- Objetivo

2.1 Objetivo General

Desarrollar un prototipo de software que permita analizar descriptivamente publicaciones extraídas de Twitter relacionadas a situaciones de contaminación del medio ambiente en Ciudad de México a fin de obtener una caracterización social, espacial y temporal sobre eventos que generan contaminación del aire como es el caso de la pirotecnia y el tráfico.

2.2 Objetivos Específicos

- Construir un proceso de extracción, transformación y carga de publicaciones en Twitter.
- Generar análisis para categorizar los datos extraídos de Twitter.
- Generar un proceso de análisis de datos ciudadanos.
- Desarrollar un tablero de datos para presentar los resultados y datos extraídos.
- Generar componentes de software de análisis semiautomático para los datos extraídos.
- Generar un proceso semiautomático de análisis de datos sociales.
- Detectar y caracterizar publicaciones en redes sociales relacionado a eventos o situaciones de contaminación del ambiente
- Identificar, analizar y delimitar los eventos que describan situaciones que generen contaminación.

3.- Justificación

Un ejemplo previo acerca de la integración de redes sociales con sensores de datos para estudiar el impacto de la polución del aire ha sido desarrollado en un trabajo anterior [5]. Los autores desarrollaron una implementación de un mecanismo geosocial de búsqueda basado en palabras clave para registrar patrones espaciales de denuncias referentes a la calidad del aire en publicaciones de Twitter. Los resultados mostraron una significante correlación a lo largo del tiempo en una serie de ciudades en Francia, Brasil y China. Con la ayuda de un diccionario sobre términos de polución, las publicaciones relevantes fueron identificadas y categorizadas, después mapeadas en diferentes vecindarios urbanos y una comparativa sociocultural así como aparecen en el diseño de la ciudad. De estos patrones históricos, unas pocas predicciones fueron también generadas.

Es importante concientizar a las personas sobre la contaminación al medio ambiente ya que esto trae repercusiones en el cambio climático y el calentamiento global. En los últimos 50 años, el aumento en la formación de contaminantes y gases invernadero ha propiciado el alza de temperaturas y el cambio climático. Este aumento de temperatura ha sido acompañado con la degradación de la calidad del aire en distintas partes del mundo. En particular, la Organización Mundial de la Salud (OMS) reportó que mueren alrededor de 4.2 millones de personas al año debido a la contaminación en el exterior [6].

Como ya hemos observado, las personas están proporcionando de forma indirecta información del impacto real de eventos contaminantes los cuales no han sido recopilados, integrados, y analizados de forma específica en la Ciudad de México, de acuerdo a la revisión del estado del arte.

Por ello, nosotros vamos a desarrollar un prototipo de software que analiza de forma descriptiva publicaciones de las redes sociales referentes a eventos que dañan al medio ambiente como pueden ser el tráfico y el uso de pirotecnia para determinar el impacto del lado de la ciudadanía en la Ciudad de México.

Obtendremos los datos no estructurados de una de las redes sociales en México: Twitter. Esta red social no necesariamente es la más popular ya que Twitter ocupa el quinto lugar en preferencia [7], sin embargo esta red social nos permite obtener la información de diferentes hashtags clave para identificar denuncias sobre contaminación al medio ambiente. Por otro lado, las características de esta red social nos permite obtener de manera más sencilla la información, por ejemplo, la principal característica de Twitter es su sencillez y capacidad de sintetizar, pues el tope de escritura es de alrededor de 140 caracteres [8]. Además, de cada 10 usuarios de Internet (más de 80 millones), 3 están en Twitter. La mayor actividad en la twittósfera mexicana se realiza en la Ciudad de México (60%), en Monterrey (17%) y en Guadalajara (10%)[8].

3.- Productos o Resultados esperados

Tras el desarrollo del presente trabajo terminal esperamos obtener los siguientes productos:

1. Base de datos de publicaciones en redes sociales de la Ciudad de México.
2. Extractor de datos sociales para obtener publicaciones geolocalizadas.
3. Análisis de textos para tópicos del medio ambiente y polución del aire.
4. Social web dashboard para describir el discurso ambiental
5. Navegación de posts clasificados.
6. Prototipo de análisis de datos ciudadanos relacionados a contaminación ambiental.

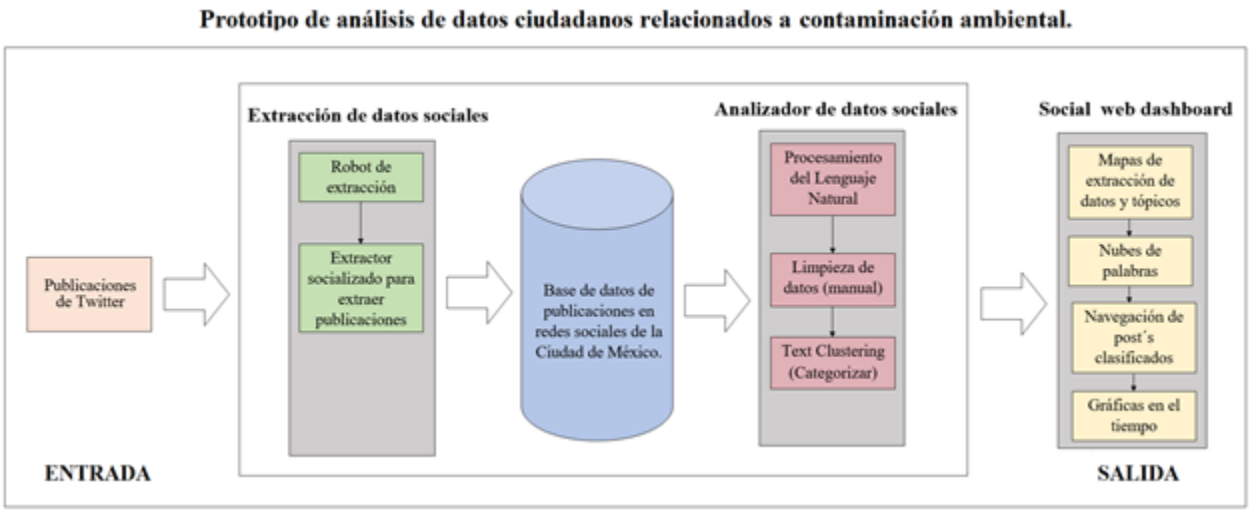


Figura 1. Arquitectura del sistema.

5.- Metodología

La metodología a usar será el modelo incremental. Este modelo de desarrollo nos permitirá construir el proyecto en etapas, que reciben el nombre de “incremento”, en donde cada una de ellas agrega una funcionalidad hasta llegar al sistema final. Este modelo nos permite reducir el riesgo de errores a través de la visibilidad de las nuevas versiones y obtener retroalimentación de las funcionalidades generadas en cada incremento. Nos evitará realizar un proyecto muy largo por la alta y cuidadosa planeación y nos permitirá generar valor al proyecto con cierta frecuencia.

El proyecto tiene la característica de poder ser dividido en diferentes subsistemas como lo pudimos definir en la sección de “Productos o Resultados esperados”. Cada incremento cuenta con etapas que nos permitirán definir los requisitos de cada uno de los subsistemas y evaluar el avance del proyecto.

En seguida se muestra la definición de procesos que se realizan en cada una de las etapas de cada incremento.

Etapa	**Proceso**
Comunicación	Se realiza una reunión con el equipo para establecer los requerimientos.
Planeación	Se plantea la iteración correspondiente.
Modelado	Se realiza el diseño del subsistema.
Construcción	Se construye el subsistema propuesto.
Despliegue	Se presenta el subsistema y se evalúa por los participantes.

Tabla 2. Etapas de cada incremento del modelo incremental.

Adoptaremos características de desarrollo de la metodología ágil SCRUM para poder trabajar de manera más eficaz y tener un panorama grande del trabajo que se hace. Realizaremos reuniones diarias (Daily SCRUM) para visualizar el avance de cada incremento desde la planeación hasta el despliegue y cada subsistema será visto como un Product Backlog y a su vez, poder crear sprints internos en cada incremento.

6.- Cronograma

CRONOGRAMA Nombre del alumno(a): Hernández Clemente Samantha TT No.:
Título del TT: Prototipo de análisis de datos ciudadanos relacionados a contaminación ambiental.

Actividad	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
Extracción de datos sociales (Twitter).											
Limpieza de datos sociales (manual).											
Crear la base de datos e ingresar los datos sociales.											
Revisión y pruebas.											
Evaluación de TT I.											
Gráficas en el tiempo (número de tweets en el tiempo).											
Mapas de extracción de datos y tópicos.											
Prototipo de clasificación de textos para tópicos del medio ambiente y polución del aire.											
Evaluación de TT II.											

CRONOGRAMA Nombre del alumno(a): Olivares Conchillos Leonel TT No.:
Título del TT: Prototipo de análisis de datos ciudadanos relacionados a contaminación ambiental.

Actividad	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
Extracción de datos sociales (Twitter).											
Limpieza de datos sociales (manual).											
Crear la base de datos e ingresar los datos sociales.											
Revisión y pruebas.											
Evaluación de TT I.											
Navegación de post clasificados.											
Prototipo de clasificación de textos para tópicos del medio ambiente y polución del aire.											
Evaluación de TT II.											

CRONOGRAMA Nombre del alumno(a): Medina Flores Susana TT No.:
Título del TT: Prototipo de análisis de datos ciudadanos relacionados a contaminación ambiental.

Actividad	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
Análisis y diseño del robot de extracción.											
Construcción del robot de extracción.											
Revisión y pruebas.											
Evaluación de TT I.											
Diseño del social web dashboard.											
Nubes de palabras para describir el discurso ambiental.											
Text clustering.											
Evaluación de TT II.											

7.- Referencias

[1] Secretaría del Medio Ambiente de la Ciudad de México. (2020). "Calidad del aire en la Ciudad de México, Informe 2018," Dirección General de Calidad del Aire, Dirección de Monitoreo de Calidad de Aire, Ciudad de México. [En línea]. Disponible en: http://www.aire.cdmx.gob.mx/descargas/publicaciones/informe-anual-calidad-del-aire-2018.pdf

[2] Procuraduría Ambiental y del Ordenamiento Territorial. (2002). Estadísticas Generales. [En línea]. Disponible en: http://www.paot.org.mx/contenidos_graficas/delegaciones/graficas_gral.php

[3] D. R. Andrés, "Cambio climático, la penalidad del ozono y la mortalidad asociada en la Ciudad de México," Tesina Licenciatura de Econ., C. de I. y Docencia Econ. A.C, CIDE, CDMX, 2020.

[4] C. Freire, "Las redes sociales trastocan los modelos de los medios de comunicación," Revista Latina de Comunicación Social, vol. 11, núm. 63, España, Canarias, 2008, pp. 287-293.

[5] M. Sammarco, R. Tse, G. Pau, G. Marfia, "Using geosocial search for urban air pollution monitoring," Pervasive Mob. Comput., pp. 15-31, 2017. [En línea]. Disponible en: https://doi.org/10.1016/j.pmcj.2016.07.001

[6] Camargo, R. (2020). Cambio climático, la penalidad del ozono y la mortalidad asociada en la Ciudad de México. [En línea]. Extraído de: http://revistas.sena.edu.co/index.php/rnt/article/view/3517/3953

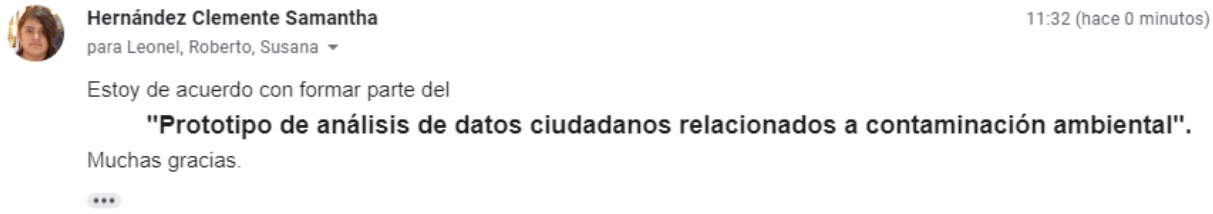
[7] Hootsuite. (2021). Informe Global sobre el entorno digital 2021 [En línea]. Disponible en: https://www.hootsuite.com/es/recursos/tendencias-digitales-2021

[8]B. Carlos. (2020). Solo 1.3 de cada 10 mexicanos tiene una cuenta de Twitter y .8 son usuarios activos [En línea]. Disponible en: https://www.abestudiodecomunicacion.com.mx/solo-1-3-de-cada-10-mexicanos-tiene-una-cuenta-en-twitter-y-8-son-usuarios-activo

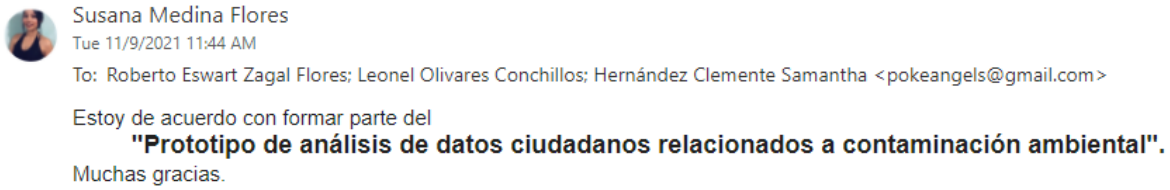
8.- Alumnos y Directores

CARÁCTER: Confidencial
FUNDAMENTO LEGAL: Artículo 11 Frace. V y Artículos 108, 113 y 117 de la Ley Federal de Transparencia y Acceso a la Información Pública.
PARTES CONFIDENCIALES: Número de boleta y teléfono.

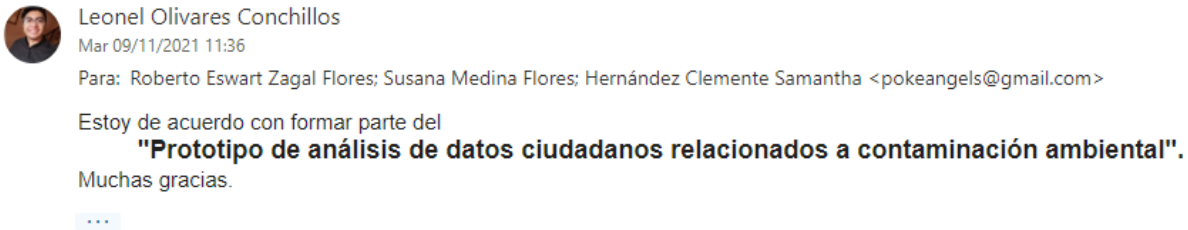
Hernández Clemente Samantha - Alumna de la carrera de
Ing. en Sistemas Computacionales en ESCOM,
Especialidad Sistemas, Boleta: 2015080602,
Tel. 5540936719, email: shernandezc1404@alumno.ipn.mx
Firma:



Medina Flores Susana - Alumna de la carrera de
Ing. en Sistemas Computacionales en ESCOM,
Especialidad Sistemas, Boleta: 2017630142,
Tel. 5611235465, email: Smedinaf1601@alumno.ipn.mx
Firma:



Olivares Conchillos Leonel - Alumno de la carrera de
Ing. en Sistemas Computacionales en ESCOM,
Especialidad Sistemas, Boleta: 2014071145,
Tel. 5561591655, email: lolivaresc1300@alumno.ipn.mx
Firma:



Dr. Roberto Zagal Flores. Es egresado de la Ingeniería en Sistemas Computacionales de la Escuela Superior de Cómputo del IPN, culminó sus estudios de Maestría en Ciencias de la Computación en el Centro de Investigación en Computación del IPN (No. De Cédula 11050111). Tiene un doctorado en tecnología avanzada de la Sección de Estudios de Posgrado de la UPIITA IPN. Actualmente es profesor de la escuela Superior de Cómputo y sus áreas de interés son: Data Mining, Spatial Data Mining, GIS, Web Semántica, Data Integration IoT y Arquitectura de Sistemas de Información. Ha trabajado en proyectos de tecnología en la iniciativa privada y en el sector público. Email: rzagalf@ipn.mx, Tel. 57296000, Ext. 52032.