

Prototipo para la identificación de efectos de vacunación contra el COVID-19 en la Ciudad de México

Trabajo Terminal No.

Alumnos: Arango León Moisés, *Quintana Martínez Erick

Directores: Dr. Zagal Flores Roberto Eswart, Dr. Mata Rivera Miguel Félix

*e-mail: equintanam1500@alumno.ipn.mx

Resumen – Durante los primeros dos años de la pandemia de COVID-19 en México, no se usó algún instrumento que permita recabar datos sobre los efectos en adultos de las campañas de vacunación. En este sentido las redes sociales han sido un medio para que los ciudadanos expresen los malestares físicos ocasionados por la aplicación de la vacuna. Recabar esta información permitirá visualizar el impacto temporal en la salud de los ciudadanos en diferentes regiones de la Ciudad de México. El reto es cómo extraer esta información en tiempo y en espacio, comprender la dinámica de los términos que describen efectos de vacunación en redes sociales, y posteriormente analizarla e identificar los efectos conocidos de las diferentes vacunas de COVID-19. En este trabajo se propone el uso de extractores de narrativas de salud de Twitter que procesaran con técnicas de Lenguaje Natural y Minería de Datos. Este estudio permitiría complementar los datos de las campañas de vacunación y definir estrategias o políticas de salud en el futuro.

Palabras clave – efectos secundarios vacunación COVID-19, Minería de datos, Social listening, Social data

1. Introducción

A principios del año 2020, en el mundo se registraron los primeros casos del virus COVID-19, ante su aumento los gobiernos tomaron medidas de confinamiento que causaron el aislamiento social, sin embargo, a pesar de ello la transmisión del virus perpetuaba, por lo que empresas farmacéuticas empezaron a diseñar vacunas de emergencia para combatir la enfermedad, y fue hasta finales del mismo año que diversos países aprobaron y comenzaron a distribuir y a aplicar dichas vacunas.

En México, las campañas de vacunación iniciaron a finales del año 2020 con el personal del sector salud, para posteriormente continuar con el resto de la población mexicana, es así como durante todo el año siguiente los mexicanos recibieron las dosis necesarias de las vacunas (Sputnik V, AstraZeneca, Moderna, Pfizer, Johnson & Johnson, Cansino y Sinovac). La mayoría de las personas presentaron reacciones normales durante las primeras 24 horas como dolor en el brazo, cansancio, dolor de cabeza, dolor muscular, escalofríos, fiebre, náuseas, etc., no obstante, algunas que presentaron efectos secundarios graves, días posteriores tales como cansancio, dolor en las articulaciones, fiebre, náuseas, reacciones alérgicas como sarpullido, picazón o hinchazón de la cara, entre otras. [1]

En algunos países crearon diversas herramientas para registrar toda reacción secundaria que se presentara y contar con estadísticas con el fin de proporcionar a las farmacéuticas este tipo de información para que pudieran realizar evaluaciones sobre los efectos secundarios y de la efectividad de sus vacunas [2] y [3]. No obstante, en México no se ha llevado acción alguna para conocer las consecuencias del uso de esta vacuna a largo plazo, por lo que no ha generado algún tipo de información pública al respecto que contribuya para el desarrollo de vacunas cada vez más eficaces y que aumenten la confianza de la ciudadanía hacia las mismas.

Ante este contexto en los últimos años, el monitorear enfermedades tradicionalmente, realizado a través de agencias de salud pública que retoman informes de médicos, clínicas y hospitales para recopilar datos, se ha visto complementado por las redes sociales que también han servido para la vigilancia de enfermedades, pues a través de la minería de datos se han podido rastrear, predecir y vigilar el desarrollo de brotes de enfermedades infecciosas. En 2008, Google Flu Trends comenzó a proporcionar datos en tiempo real al público cuando observó brotes de gripe en todo el mundo, según los términos relacionados con la gripe que las personas buscaban en Internet [4], pudo rastrear enfermedades infecciosas como la tuberculosis y la influenza.

Por tanto, mediante estos nuevos canales, se han podido detectar puntos críticos de brotes antes que la vigilancia tradicional lo haga, además, a través de las redes sociales, la vigilancia de enfermedades puede contribuir potencialmente a un análisis más preciso y completo de la estimación de brotes, porque identifica casos de enfermedades incluso cuando los individuos no buscan atención médica.

Aunado a ello, las redes sociales son de gran apoyo para conocer las opiniones de las personas en relación con algún determinado tema. Específicamente la red social Twitter se ha consolidado como un canal de denuncia alternativo al que ofrecen los gobiernos y empresas farmacéuticas, lo que brinda una oportunidad para la vigilancia de los efectos secundarios de la vacuna debido a que los usuarios de la

red social hablan de sus síntomas.

En meses recientes, se han generado diversas investigaciones para conocer la percepción de la población referente a la vacuna COVID-19, como, por ejemplo, en University of Texas Southwestern Medical Center hicieron un análisis de sentimientos y emociones e inferencia demográfica de los usuarios en los tweets relacionados con la vacuna COVID-19 para proporcionar información sobre la evolución de las actitudes públicas [5]. También el estudio realizado por investigadores de la Universidad de Guelph permitió comprender los sentimientos y las opiniones sobre la vacunación a través de Twitter. Ambas investigaciones pueden ayudar a las agencias de salud a aumentar los mensajes positivos y eliminar los mensajes opuestos, con el propósito de mejorar la captación de vacunas. [6]

Incluso, hay un aporte referente al monitoreo de los efectos adversos utilizando datos de Twitter. Los investigadores Andrew T. Lian, Jingcheng Du y Lu Tang desarrollaron un enfoque de procesamiento de lenguaje natural y aprendizaje automático para identificar Eventos Adversos de Vacunas (Vaccine Adverse Events, VAE) COVID-19, en el que descubrieron que los cuatro estados más poblados (California, Texas, Florida y Nueva York) en los EE. UU. fueron testigo de la mayoría de las discusiones de VAE en Twitter. En su artículo mencionan que los hallazgos demostraron la viabilidad de usar datos de redes sociales para monitorear VAE, puesto que es un excelente complemento para los sistemas de farmacovigilancia de vacunas existentes. [4]

1.1 Planteamiento del problema

Al momento en el que se elabora este protocolo, no existe un mecanismo o sistema oficial que permita recabar los síntomas o efectos secundarios de la vacunación de COVID-19 en la Ciudad de México. En la red social Twitter existen narrativas que reportan estos efectos, sin embargo, el poder extraerla y analizar estos datos implica procesar información dispersa en el tiempo y en el espacio, además lingüísticamente la forma de expresar un malestar físico puede variar.

2. Objetivos

Objetivo general

Desarrollar un prototipo para la identificación de efectos de la vacunación contra el COVID-19 en la Ciudad de México procesando y analizando datos de redes sociales y minería de datos.

Objetivos específicos

- Desarrollar una conceptualización de los principales efectos de la vacunación por COVID-19 de acuerdo a la literatura científica
- Definir zonas y parámetros para la extracción de datos sociales.
- Desarrollar un extractor datos sociales aplicado a la vacunación de COVID-19
- Investigar, analizar e integrar técnicas de procesamiento de datos sociales sobre campañas de vacunación
- Definir una arquitectura de data mining para la extracción, procesamiento y visualización de datos sociales referentes a campañas de vacunación

3. Justificación

La Secretaría de Salud de la Ciudad de México no tiene mecanismos ni lleva un registro para medir los efectos secundarios de la vacuna COVID-19 en los habitantes, aunque, el gobierno mexicano cuenta con un programa de farmacovigilancia con la Comisión Federal para la Protección contra Riesgo Sanitario (COFEPRIS), que recopila los efectos adversos de un medicamento o vacuna, este no se ha llevado a cabo con las vacunas para el COVID-19, pues no han realizado un estudio a profundidad que aclare las consecuencias de su uso a largo plazo, por lo que no han generado algún tipo de información pública al respecto que contribuya para el desarrollo de vacunas cada vez más eficaces y que aumenten la confianza de la ciudadanía hacia las mismas.

En contraste con Israel que al tener un sistema de salud universal y un registro de salud digitalizado, le ha sido posible recolectar datos de número de casos de COVID-19, número de muertes, hospitalizaciones, el número de pacientes que necesitan respiradores y si hay cualquier efecto de la vacuna [2]; mientras que Argentina y España han publicado constantemente informes sobre los efectos secundarios presentados en los últimos meses en su población. [3] y [7]

Dicho lo anterior, en México se identifica la necesidad de desarrollar herramientas capaces de poder obtener y recolectar todo registro que ayude a ampliar el panorama que se tiene dentro de la población mexicana relacionado a efectos que sean causados por la aplicación de la vacuna. Dado que, la falta de registro o información, se ha constituido como un problema, ya que muchas personas al no tener un

comunicado oficial (con cifras y experiencias reales), para consultar sobre los efectos secundarios, deciden dejar de aplicarse las dosis siguientes de la vacuna que le sea asignada, provocando una gran alza en esquemas de vacunación incompletos o sin iniciar, dejando expuestas a estas personas a contraer el virus y que la propagación nunca acabe o se logre controlar en la población mexicana. [8]

El proyecto consistirá en recolectar tweets para encontrar referencias sobre los efectos secundarios, los cuales podrán ser entregados a las farmacéuticas, que distribuyen las vacunas en la Ciudad de México, que les ayudará a estudiar con más detalle sus causas, de esta forma pueden prevenir y reducir los efectos adversos de sus vacunas.

Cabe mencionar que, Twitter tiene un alcance que llega a millones de mexicanos, los cuales tienen una cuenta asociada en esta red social, así pues, se puede afirmar que el campo de investigación es muy amplio, de ahí que los resultados obtenidos ayudarán de gran manera a las investigaciones de los epidemiólogos.

En conclusión, el desarrollo de este proyecto puede ser una buena alternativa para identificar efectos secundarios de la vacuna COVID-19, con ello los tomadores de decisiones en los diferentes institutos de salud pública de la Ciudad de México podrán tener información sobre el impacto de los efectos de vacunación y desarrollar políticas sociales o de salud para las campañas de información.

4. Productos o resultados esperados

Como podemos ver en nuestro bloque de flujos describe el funcionamiento que queremos obtener para nuestro prototipo y que resultados nos va a brindar. En la figura 1 se muestran las fases y componentes claves para desarrollar este trabajo. (1) Extracción de datos: Se desarrollará un extractor de tweets y estrategias para localizar datos sociales útiles para este estudio (2) Realizamos un Pre- procesamiento de datos: Se requieren técnicas de tratamiento de lenguaje natural para detectar tópicos que sean tendencias en los tweets recabados, esto requiere una fase de limpieza de los datos, (3) Análisis de tendencias con Minería de Datos: Es un análisis estadístico para detectar tendencias en los datos recabados haciendo uso del topic model y data cube, (4) Visualización de datos; es la fase para desplegar e interpretar los tópicos descubiertos y estadísticas asociadas.

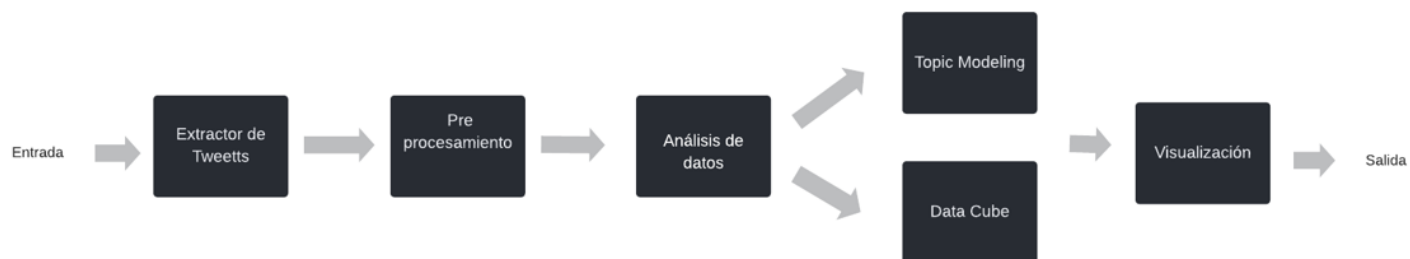


Figura 1. Arquitectura del prototipo

Alcances:

1. Nuestro prototipo tendrá la capacidad de diferenciar distintos términos de los efectos secundarios, los cuales son de vital importancia para poder llevarlos a las categorías adecuadas.
2. La zona a analizar será la CDMX, pero se delimitará a cada alcaldía para extraer los datos que necesitamos. Todos los tweets recolectados pasarán por el proceso de limpieza para su mejor análisis.
3. Aplicando las técnicas de modelado adecuadas se espera que el prototipo sea capaz de identificar los efectos secundarios, con ello se podrá disponer de representaciones gráficas en relación tiempo y espacio con todos los datos registrados, para saber cómo han impactado los efectos. Cabe mencionar que el tiempo no será de la fecha en la que se recuperaron los tweets, sino más bien es la fecha en la que los usuarios de la CDMX mencionaron los efectos que tuvieron después de su aplicación.

5. Metodología

Hay diversas metodologías para trabajar con la minería de datos, pero la que se empleará en el desarrollo de este proyecto es CRISP-DM (Cross Industry Standard Process for Data Mining) dado que es intuitiva, flexible y simple.

CRISP-DM tiene seis etapas, algunas de las cuales son bidireccionales y que pueden tener iteraciones de acuerdo con las necesidades. [9]

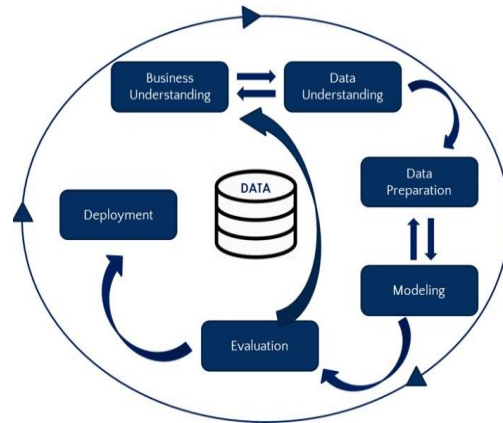


Figura 2. Etapas CRISP-DM

La primera etapa es entendimiento del negocio, sabemos que el principal problema es que no existe un proceso o sistema oficial que permita recabar los síntomas o efectos de la vacunación de COVID-19 en la Ciudad de México, por lo cual el objetivo del proyecto será en proporcionar un prototipo a la Secretaría de Salud de la Ciudad de México que servirá para llevar registros de los efectos secundarios de las campañas de vacunación.

Para lograr lo anterior, la herramienta esencial será el lenguaje de programación Python, dado que ofrece muchas librerías para el análisis de datos. Para la extracción de los tweets se hará uso de snsrape que es un scraper para las redes sociales, permite obtener tweets de tiempo atrás además de especificar las coordenadas geográficas. En la modelación se usarán las librerías scikit-learn, NLTK y gensim para el procesamiento de los datos. Y para implementar datos multidimensionales se ocupará SQL Server Analysis Services (SSAS) que nos permitirá hacer el análisis estadístico en dimensiones de espacio y tiempo.

La siguiente etapa es el entendimiento de los datos, por lo que los datos a extraer serán de la red social Twitter para ello se debe de identificar palabras clave (como dolor cabeza, dolor garganta, dolor vacuna, efecto vacuna, reacción vacuna, vacuna, vacuna #AstraZeneca, vacuna síntoma, etc.) además será necesario definir coordenadas geográficas (demarcaciones territoriales, sedes de vacunación, etc.) para rastrear los posibles efectos secundarios en distintas zonas de la Ciudad de México.

Los tweets recolectados se guardarán en un archivo csv (valores separados por comas) para su mejor manipulación, que tendrá como columnas principales el id del tweet, consulta de búsqueda, la publicación, la fecha de extracción, la cantidad de reacciones, la latitud y longitud y el radio de distancia.

La tercera etapa es la preparación de los datos. Entre todas las columnas la que se seleccionará es la de publicación dado que es la que contiene los términos buscados, después se hará una limpieza de los tweets eliminando palabras vacías (artículos, pronombres, preposiciones, etc.), cuentas de usuarios de noticias con la finalidad de tener solo el contenido relacionado con experiencias personales, etc.

Para la etapa del modelado se seleccionará el método de agrupamiento K-medias (k-means clustering) para asociar ciertas palabras que describen ciertos grupos de tweets, los datos para el entrenamiento serán del 80% mientras que el resto que es del 20% serán utilizados como datos de prueba. Las palabras podrán ser extraídas por medio de la técnica modelado de temas que servirá para saber las frases más representativas de los grupos de tweets y podrán ser representadas por medio de la técnica de words clouds, después cada tweet se le aplicará análisis de sentimientos para conocer el tono (negativo, positivo, neutro) usando el clasificador Naive Bayes.

En la etapa de evaluación se harán las pruebas necesarias para los modelos generados. Una vez realizada la evaluación, se debe decidir si los objetivos han sido cumplidos y de ser así se puede avanzar a la fase de implantación.

6. Cronograma

Título del Trabajo Terminal: Prototipo para la identificación de efectos de vacunación contra el COVID-19 en la Ciudad de México

[illegible]

Nombre del alumno: Quintana Martínez Erick

Título del Trabajo Terminal: Prototipo para la identificación de efectos de vacunación contra el COVID-19 en la Ciudad de México

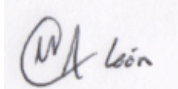
Actividad	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Enero	Febrero	Marzo	Abril	Mayo	Junio
Entendimiento de los efectos secundarios											
Diseño de la arquitectura de Minería de datos											
Diseño de la visualización de los datos											
Generación del reporte técnico											
Evaluación TT I											
Implementación de la arquitectura de minería de datos											
Extracción de datos											
Análisis de los datos											
Pruebas y experimentos											
Evaluación TT II											

7. Referencias

- [1] Centros para el Control y la Prevención de enfermedades CDC, (2022, ene. 12). “Posibles efectos secundarios después de vacunarse contra el COVID-19”. [Internet]. Disponible en <https://espanol.cdc.gov/coronavirus/2019-ncov/vaccines/expect/after.html>
- [2] I. Encabo, (2021, ene. 21). “El precio del "milagro" israelí con la vacuna: pagar más y dar datos a Pfizer”. [Internet]. Disponible en <https://www.elindependiente.com/vida-sana/salud/2021/01/21/el-precio-del-milagro-israeli-con-la-vacuna-pagar-mas-y-dar-datos-a-pfizer/>
- [3] Heraldo Salud, (2021, sep. 08). “Nuevos efectos secundarios de las vacunas de Pfizer, AstraZeneca, Moderna y Janssen”. Disponible en <https://www.heraldo.es/noticias/salud/2021/03/10/nuevos-efectos-secundarios-vacuna-pfizer-moderna-1476559.html?fbclid=IwAR3gMiJDldTNXBQ3lUg80Mji2lt63Q2pd7P95XsMhIWAFUP699QXFFsvfzk>
- [4] A. T. Lian, J. Du, and L. Tang, “Using a Machine Learning Approach to Monitor COVID-19 Vaccine Adverse Events (VAE) from Twitter Data”, Vaccines, vol. 10, no. 1, p. 103, ene. 2022, doi: 10.3390/vaccines10010103
- [5] S. N. Saleh, A. McDonald, A. Basit, S. Kumar, R. J. Arasaratnam, C. U. Lehmann y R. J. Medfor, “Public Perception of COVID-19 Vaccines through Analysis of Twitter Content and Users”, medRxiv, s. v, s.n, s.p, abr. 2021, doi: <https://doi.org/10.1101/2021.04.19.21255701>
- [6] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos y S. Sharif, “An analysis of COVID-19 vaccine sentiments and opinions on Twitter”, International Journal of Infectious Diseases, vol. 108, s.n, p. 256-262, jul. 2022, doi: <https://doi.org/10.1016/j.ijid.2021.05.059>
- [7] P. Vega, (2021, abr. 23) "Por qué cada vez más países recurren a la vacuna Sputnik V: una efectividad del 97,6% y no produce trombos". Disponible en <https://www.eleconomista.es/sanidad/noticias/11175777/04/21/Por-que-cada-vez-mas-paises-recurren-a-la-vacuna-Sputnik-V-una-efectividad-del-976-y-no-produce-trombos.html>
- [8] Instituto Mexicano del Seguro Social, (2022, abr. 17). “Vacuna contra COVID-19 es la medida más efectiva para prevenir cuadros graves de la enfermedad”. Disponible en https://www.gob.mx/imss/prensa/vacuna-contra-covid-19-es-la-medida-mas-efectiva-para-prevenir-cuadros-graves-de-la-enfermedad?state=published&fbclid=IwAR25UfXpEFWoomoN_vRsCdiyI8YsODHVh-CHslULhtPnJZ62k70ndB6kj4
- [9] V. Galán Cortina, “APLICACIÓN DE LA METODOLOGÍA CRISP-DM A UN PROYECTO DE MINERÍA DE DATOS EN EL ENTORNO UNIVERSITARIO”, Proyecto fin de carrera, Escuela Politécnica Superior Ing. en Inform., UC3M., Getafe, España, 2015. p. 21. Disponible en https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf

8. Alumnos y directores

Moisés Arango León. - Alumno de la carrera de Ingeniería en Sistemas Computacionales en ESCOM. Boleta 2019630170, teléfono: 5583560340, correo electrónico: marangol1500@alumno.ipn.mx



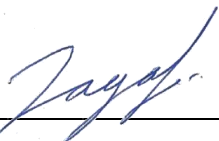
Firma: _____

Erick Quintana Martínez. - Alumno de la carrera de Ingeniería en Sistemas Computacionales en ESCOM. Boleta 2019630154, teléfono: 5516409955, correo electrónico: equintanam1500@alumno.ipn.mx



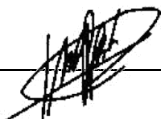
Firma: _____

Dr. Roberto Eswart 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. Cedula 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 Arquitecturas de Sistemas de Información. Ha trabajado en proyectos de tecnología en la iniciativa privada y en el sector público. Email: zagalmmx@gmail.com, Tel. 57296000, Ext. 52032.



Firma: _____

Dr. Miguel Félix Mata Rivera. - Es Egresado de la Ingeniería en Computación de la Escuela Superior de Ingeniería Mecánica y Eléctrica Unidad Culhuacán del IPN, culminó sus estudios de Maestría en Ciencias de la Computación, así como su Doctorado en Ciencias de la Computación en el Centro de Investigación en Computación del IPN. Actualmente es profesor de la Unidad Profesional Interdisciplinaria en Ingeniería y Tecnología Avanzada y sus áreas de interés son Geographic Information Retrieval, Spatial Semantic Web, Web-Mapping, Programación de Sistemas y Desarrollo de Aplicaciones Web y Móviles, Sistemas multimedia, Soluciones Móviles, Cómputo Ubicuo. Email: migfel@gmail.com Teléfono: 5796000 Extensión: 56853.



Firma: _____