RESEARCH ARTICLE

VOL. 47 | **SPECIAL ISSUE SI TEC 2026**| PP 1-25

Revista Mexicana de Ingeniería Biomédica

https://dx.doi.org/10.17488/RMIB.47.SI-TAIH.1529

E-LOCATION ID: e**1529**

Monitoreo y Predicción de Enfermedades Infecciosas a través del Análisis de Redes Sociales

Monitoring and Prediction of Infectious Diseases through Social Network Analysis

Pedro Wences[1], *Alicia Martínez*[1], *Hugo Estrada*[1], *Sabino Miranda*[2]

[1]Centro Nacional de Investigación y Desarrollo Tecnológico, Morelos - México

[2]Universidad Autónoma de la Ciudad de México

RESUMEN

Este artículo presenta un enfoque integral para el monitoreo y predicción de enfermedades infecciosas mediante el análisis de redes sociales, enfocado en la pandemia de COVID-19. El objetivo es identificar afirmaciones de contagio en tiempo real y estimar su evolución como complemento a la vigilancia epidemiológica tradicional. Se desarrolló un sistema que combina técnicas de procesamiento de lenguaje natural, usando el modelo BERT para clasificar afirmaciones de contagio en X, y la función de *Gompertz* para proyectar casos a corto plazo. La metodología también incluye análisis de publicaciones georreferenciadas, predicciones mediante ventanas móviles y representación espacial de zonas de riesgo mediante mapas de calor. Los resultados muestran una correlación significativa entre las menciones en X y los reportes oficiales, sugiriendo una sincronicidad temporal entre ambas fuentes. Se reconocen limitaciones importantes como el sesgo urbano en la muestra de usuarios y la escasa representación rural. Finalmente, se concluye que las redes sociales representan un recurso potencialmente valioso como fuente complementaria para generar alertas epidemiológicas oportunas, fortaleciendo así la toma de decisiones en salud pública.

PALABRAS CLAVE: COVID-19, redes sociales, BERT, *Gompertz*, vigilancia epidemiológica

ABSTRACT

This article presents a comprehensive approach for monitoring and predicting infectious diseases through social media analysis, focusing on the COVID-19 pandemic. The objective is to identify real-time contagion reports and estimate disease trends as a complement to traditional epidemiological surveillance. We developed a system integrating natural language processing techniques, employing the BERT model to classify contagion statements on X, and using the *Gompertz* function to forecast short-term case growth. The methodology also incorporates analysis of georeferenced posts, predictions via rolling windows, and spatial representation of risk areas through heat maps. Results indicate a significant correlation between X mentions and official health reports, suggesting temporal synchronicity between both data sources. Important limitations are acknowledged, such as the urban bias in X user samples and the underrepresentation of rural populations. Finally, it is concluded that social media represent a potentially valuable resource as a complementary source for generating timely epidemiological alerts, thereby strengthening public health decision-making.

KEYWORDS: COVID-19, social media, BERT, *Gompertz*, epidemiological surveillance

Autor de correspondencia

PARA: **ALICIA MARTÍNEZ REBOLLAR**
INSTITUCIÓN: **TECNOLÓGICO NACIONAL DE MÉXICO / CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO**
DIRECCIÓN: **INTERIOR INTERNADO PALMIRA S/N, COL. PALMIRA, CUERNAVACA, MORELOS, C.P. 62490, MÉXICO.**
CORREO ELECTRÓNICO: **alicia.mr@cenidet.tecnm.mx**

Recibido:

16 Febrero 2025

Aceptado:

5 Septiembre 2025

Publicado:

30 Enero 2026

INTRODUCTION

Las enfermedades infecciosas representan un desafío persistente y crítico para la salud pública global, lo que exige estrategias de monitoreo capaces de detectar y seguir en tiempo real las zonas de contagio, permitiendo una respuesta oportuna ante su propagación[1]. La pandemia de COVID-19 ha evidenciado la urgencia de desarrollar métodos más eficientes de vigilancia epidemiológica, dado su alto índice de transmisión y el impacto global sin precedentes[2]. No obstante, la dependencia exclusiva de datos provenientes de centros de salud y hospitales limita la capacidad de respuesta de las instituciones sanitarias, debido a los retrasos inherentes en la recopilación, procesamiento y análisis de esta información.

En las últimas décadas, las redes sociales han emergido como un recurso invaluable para la vigilancia de fenómenos sociales y de salud. Entre estas plataformas, Twitter, actualmente denominado X, se ha consolidado como una fuente de datos en tiempo real, que refleja no solo la percepción pública sobre temas de salud, sino también indicios de posibles brotes infecciosos mediante menciones de síntomas y la ubicación de los usuarios. La riqueza y rapidez de esta información presenta una oportunidad única para complementar los sistemas de monitoreo tradicionales mediante un análisis que capture eventos y tendencias de salud en el momento en que ocurren[3]. Sin embargo, el aprovechamiento de esta información requiere abordar desafíos como la gestión del ruido en los datos, la identificación precisa de menciones relevantes y la integración de estos hallazgos con modelos epidemiológicos convencionales.

Este artículo propone un enfoque innovador para el monitoreo y la predicción de enfermedades infecciosas a través del análisis de redes sociales, utilizando datos de X como insumo principal para identificar y proyectar áreas de contagio. Mediante la integración de técnicas de procesamiento de lenguaje natural, el modelo BERT, redes neuronales convolucionales y la función de *Gompertz* para proyecciones epidemiológicas, se busca proporcionar una herramienta moderna que complemente los métodos convencionales de vigilancia. En este estudio, el enfoque se implementó para el monitoreo del COVID-19 en México, aunque su aplicabilidad se extiende a diversas enfermedades infecciosas. Los resultados obtenidos no solo destacan la viabilidad de estas tecnologías, sino que también ofrecen herramientas concretas para fortalecer la respuesta epidemiológica y mejorar la toma de decisiones en salud pública, sentando las bases para una vigilancia más proactiva y eficaz ante futuras pandemias.

Antecedentes

El monitoreo epidemiológico ha incorporado enfoques innovadores con el auge de las redes sociales, consolidándose X como una fuente valiosa de datos no estructurados para identificar patrones y tendencias en salud pública. Diversas investigaciones han explorado su potencial para complementar los sistemas tradicionales de vigilancia epidemiológica mediante métodos avanzados de análisis de datos en tiempo real.

El uso de X para monitorear la pandemia de COVID-19 ha sido ampliamente estudiado. En España, Arjona[4] analizó la distribución espacial y temporal de términos relacionados con el virus, generando mapas de calor que facilitaron la comprensión de su propagación y resaltaron la utilidad de los datos sociales como complemento de los informes oficiales. De manera similar, en la investigación de Hoque[5] se analizaron millones de mensajes publicados durante las primeras etapas de la pandemia, revelando disparidades en las respuestas emocionales y conductuales a las medidas de salud pública entre distintas regiones.

Más allá del COVID-19, X ha sido empleado para mapear la epidemiología de otras enfermedades. En la investigación de Tulloch[6] se empleó en la vigilancia de la enfermedad de Lyme en el Reino Unido e Irlanda,

encontrando una fuerte correlación entre los datos obtenidos en la plataforma y las cifras oficiales, lo que subraya la capacidad de las redes sociales para reflejar con precisión la incidencia y distribución de enfermedades. Por otro lado, en la investigación de Chae[7] se combinaron datos de X, motores de búsqueda y registros oficiales para predecir la aparición de enfermedades infecciosas mediante modelos de aprendizaje profundo, demostrando el valor de integrar múltiples fuentes digitales para mejorar la precisión en las proyecciones epidemiológicas.

En otros ámbitos, como la salud mental y la gestión de desastres, las redes sociales han mostrado su utilidad como herramientas de monitoreo. Por ejemplo, en la investigación de Birjali[8] se implementaron algoritmos de aprendizaje automático para detectar patrones de ideación suicida en las publicaciones, mientras que Reynard y Shirgaokar[9] destacaron el papel de X en la toma de decisiones durante emergencias, utilizando datos geolocalizados para evaluar daños y asignar recursos de manera eficiente. Además, el análisis geoespacial ha cobrado relevancia en la interpretación de datos sociales, con estudios como el de Gu[10], quienes propusieron un modelo de confianza social para inferir la ubicación de los usuarios, y Redondo[11], que aplicaron técnicas de entropía y agrupamiento para identificar actividades inusuales en entornos urbanos.

Asimismo, investigaciones como las de Alonso[12] y Xiong[13] han explorado el uso de redes sociales basadas en la ubicación para recomendar actividades o puntos de interés, aprovechando datos espacio-temporales para mejorar la precisión de los resultados. Estos enfoques demuestran la versatilidad del análisis de datos en redes sociales para inferir patrones espaciales y comportamentales en distintos contextos.

No obstante, el uso de X como fuente de datos para estudios epidemiológicos presenta limitaciones importantes que deben abordarse con cuidado y procesarse adecuadamente. En primer lugar, diferenciar publicaciones provenientes de personas reales y cuentas automatizadas (*bots*) continúa siendo un desafío. Estudios recientes estiman que aproximadamente el 9 % de las publicaciones relacionadas con COVID-19 podrían haber sido generados por *bots*[14]. Estos *bots* frecuentemente se enfocan en contenido político, críticas hacia medidas sanitarias o difusión de mensajes negativos, lo que distorsiona la percepción general obtenida de la plataforma[15].

En segundo lugar, los datos generados por los usuarios en X reflejan fuertes sesgos cognitivos, ideológicos y emocionales. Por ejemplo, en la investigación de Jiang[16] se muestra que durante la pandemia de COVID-19 hubo una notable politización del discurso, con comunidades divididas que operaban en “cámaras de eco”, reforzando visiones polarizadas sobre temas clave como el uso de mascarillas o las vacunas. Asimismo, en la investigación de Xue[17] se identifica que emociones intensas como miedo, ira y ansiedad, comunes durante brotes epidémicos, pueden amplificar contenidos alarmistas y limitar la diversidad de la información compartida, sesgando así el análisis automatizado de contenido.

En tercer lugar, debe considerarse la limitada veracidad de las afirmaciones personales de contagio publicadas en redes sociales, ya que no existe un mecanismo formal de verificación clínica en estas plataformas. La naturaleza autorreportada de estas publicaciones implica que algunas personas pueden comunicar información imprecisa, exagerada o incluso falsa, ya sea intencionadamente o por error. Para mitigar este riesgo, algunos estudios han propuesto estrategias de filtrado más estrictas que incluyen únicamente aquellas publicaciones donde los usuarios declaran explícitamente haber recibido un diagnóstico confirmado, ya sea mediante prueba PCR o valoración médica directa[18][19]. Estas investigaciones han demostrado que, al aplicar estos criterios, es posible construir conjuntos de datos más confiables y útiles para el monitoreo epidemiológico. Aun así, incluso con estos filtros, persiste un grado de incertidumbre, por lo que estos datos deben interpretarse con cautela y preferentemente a nivel agregado.

En cuarto lugar, la presencia de desinformación y teorías conspirativas en X representó una limitación significativa durante la pandemia. Desde etapas tempranas circularon narrativas infundadas, como la supuesta relación entre el virus y la tecnología 5G o la idea de que las vacunas tenían propósitos ocultos. Investigaciones encontraron evidencia de que estos contenidos fueron amplificados por cuentas automatizadas y pequeños grupos altamente activos, lo que generó una falsa percepción de consenso en la plataforma[20][21]. Este fenómeno no solo afectó la calidad informativa del entorno digital, sino que también introdujo ruido en los datos, dificultando el análisis fiable del comportamiento y la percepción social frente a la pandemia.

Finalmente, la representatividad geográfica y sociodemográfica de los usuarios de X constituye otra limitante importante. Estudios previos indican que los usuarios de esta plataforma tienden a ser más jóvenes, urbanos, con mayor nivel educativo y políticamente más inclinados hacia posiciones liberales o progresistas, en comparación con la población general[22][23]. Esta subrepresentación de poblaciones rurales, mayores o con limitado acceso digital implica que los datos recopilados deben interpretarse con cautela al extrapolarlos a contextos poblacionales más amplios.

Estas limitaciones, aunque significativas, no anulan el valor potencial de los datos de redes sociales para complementar la vigilancia epidemiológica, especialmente cuando se interpretan correctamente y se combinan con métodos tradicionales de monitoreo. En conjunto, los antecedentes revisados subrayan tanto el potencial como las restricciones de X y otras redes sociales como herramientas para el monitoreo y predicción epidemiológica. En este trabajo, se propone un enfoque complementario de monitoreo que combina datos sociales con técnicas de procesamiento de lenguaje natural, modelado epidemiológico y aprendizaje profundo, con el objetivo de fortalecer los sistemas tradicionales al capturar señales tempranas de contagio que podrían pasar desapercibidas mediante métodos convencionales. Este enfoque busca enriquecer las herramientas existentes, ofreciendo una perspectiva en tiempo real basada en la actividad social en línea durante pandemias como la de COVID-19 en México.

MATERIALES Y MÉTODOS

En este apartado, se describen los conjuntos de datos, técnicas y herramientas utilizadas en el desarrollo y validación del enfoque propuesto para el monitoreo y predicción de enfermedades infecciosas mediante X. La metodología seguida en este estudio se compone de una serie de fases diseñadas para optimizar la identificación y el análisis de información relevante en redes sociales, con el objetivo de mejorar la vigilancia epidemiológica.

Este enfoque busca superar las limitaciones de los métodos convencionales en términos de inmediatez, proporcionando información en tiempo real que puede ser valiosa para la toma de decisiones en salud pública. La Figura 1 presenta un diagrama de flujo de la metodología empleada en esta investigación.

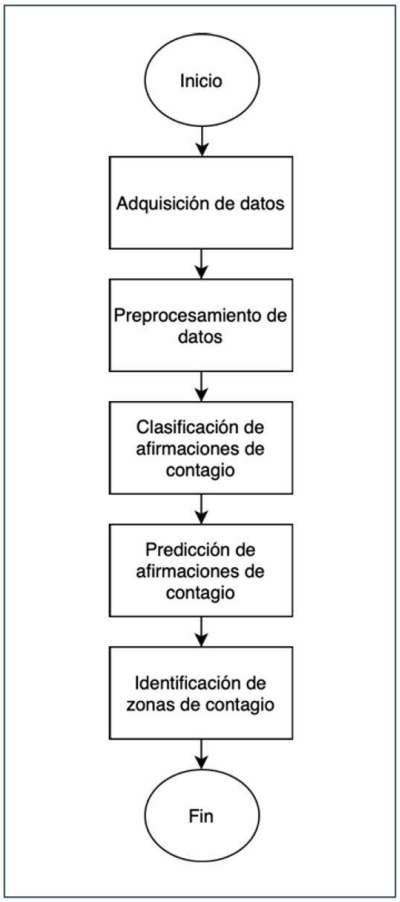

FIGURA 1. **Diagrama de flujo de la metodología para el monitoreo y predicción de enfermedades infecciosas utilizando redes sociales.**

La metodología integra técnicas de procesamiento de lenguaje natural, aprendizaje automático y modelado matemático para analizar menciones en X y proyectar posibles tendencias de contagio. Se compone de cinco fases principales: adquisición de datos, preprocesamiento de datos, clasificación de afirmaciones de contagio, predicción de afirmaciones de contagio e identificación de zonas de contagio. A continuación, se describe en detalle cada una de estas fases.

Adquisición de datos

La adquisición de datos es la fase encargada de obtener y almacenar las publicaciones de la plataforma X (anteriormente conocidas como tweets cuando se denominaba Twitter), con el objetivo de recopilar información relevante para el análisis. Estas publicaciones (también llamadas posts en su terminología actual) pueden contener texto, enlaces, imágenes, videos y metadatos adicionales, como la fecha y hora de publicación, información del usuario y, en algunos casos, datos de ubicación. Cuando el usuario así lo permite, estas publicaciones incluyen información de georreferenciación, es decir, coordenadas geográficas que indican el lugar aproximado desde donde fueron publicadas.

La API de X permite obtener publicaciones georreferenciadas si se configura un filtro por *locations*, el cual acepta uno o varios *bounding boxes* definidos por dos coordenadas diagonales opuestas. En este estudio, se delimitó la República Mexicana a través de 24 *bounding boxes*. Cada uno fue definido manualmente utilizando la herramienta *boundingbox.klokantech.com*, mediante trazos visuales sobre el mapa y copiando las coordenadas correspondientes al rectángulo generado.

La elección de estos 24 *bounding boxes* respondió a un criterio de equilibrio entre precisión geográfica y factibilidad operativa. Un número menor de *bounding boxes* habría implicado áreas demasiado amplias, con mayor probabilidad de capturar datos provenientes de países vecinos. Por otro lado, utilizar un número considerablemente mayor puede ofrecer una delimitación más precisa, aunque también incrementa la complejidad en la configuración. En este sentido, la cantidad de *bounding boxes* utilizada puede variar según los objetivos y recursos de cada investigador. En este estudio, cada *bounding box* fue definido cuidadosamente para cubrir el territorio de la República Mexicana, procurando que la superposición con zonas limítrofes de Guatemala, Belice o Estados Unidos fuera la menor posible. La Figura 2 muestra la distribución geográfica de los *bounding boxes* utilizados.

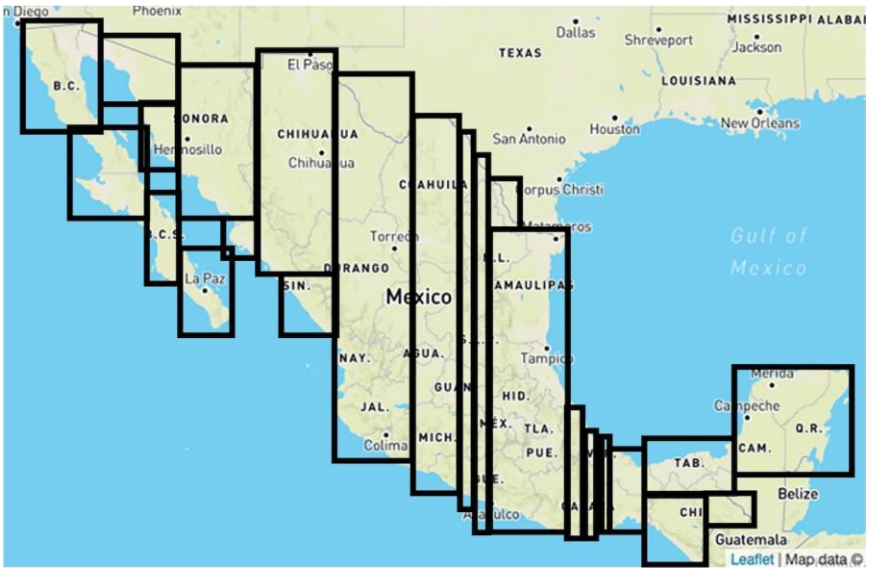

FIGURA 2. **Distribución geográfica de los 24 *bounding boxes* utilizados para delimitar el territorio de la República Mexicana durante el proceso de adquisición de datos en X**

La adquisición de publicaciones georreferenciadas se llevó a cabo mediante un proceso automatizado en Python que utilizó la biblioteca *Tweepy*[24], permitiendo interactuar con la API de X de manera eficiente. El acceso a la API de X requirió el registro en la plataforma para desarrolladores, donde se aprobaron y asignaron *tokens* de autenticación. Estos *tokens* fueron esenciales para la recopilación automatizada de publicaciones en tiempo real dentro de la región delimitada por los *bounding boxes*.

El proceso de adquisición de datos se llevó a cabo del 1 de octubre de 2021 al 13 de marzo de 2023. Este periodo comenzó tras finalizar las pruebas del sistema de recolección y concluyó debido a los cambios en las políticas de acceso a la API de X. Durante este período, se almacenaron en tiempo real todas las publicaciones georreferenciadas en México, segundos después de ser publicados en X. Como resultado, se recopilaron 20,008,585 registros, cada uno con 15 atributos, los cuales fueron almacenados en una base de datos relacional para facilitar su gestión y análisis. La Tabla 1 presenta la descripción de los atributos.

TABLA 1. **Descripción de los atributos de un registro.**

Atributo	Valor de ejemplo
tweet.id	123456789012345678
tweet.text	"Este es una publicación de ejemplo"
tweet.truncated	False
tweet.extended_tweet["full_text"]	"Este es el texto completo de una publicación extendida..."
tweet.user.id	987654321
tweet.user.name	"Monica Pérez"
tweet.user.screen_name	"MoniP123"
tweet.user.description	"Apasionada de la tecnología y la programación"
tweet.user.location	"Ciudad de México"
tweet.created_at	2025-02-15 12:34:56
tweet.place.id	"07d9cd2eef4f207f"
tweet.place.full_name	"Ciudad de México, México"
tweet.place.country	"México"
tweet.place.place_type	"city"
tweet.place.bounding_box.coordinates	[[[-99.3647, 19.0112], [-99.3647, 19.6399], [-98.9403, 19.6399], [-98.9403, 19.0112]]]

Preprocesamiento de datos

El preprocesamiento de datos es la fase en la que se organizan, filtran y transforman los datos adquiridos para que sean adecuados para el análisis. Su objetivo principal es mejorar la calidad, estructura y utilidad de la información mediante la eliminación de datos irrelevantes, la organización en una base de datos estructurada y la normalización del contenido textual. Esto asegura que los datos sean representativos y confiables para su posterior clasificación y modelado. En este estudio, el preprocesamiento se realizó en dos niveles principales:

Estructuración y almacenamiento de datos

Una vez obtenidos los datos en bruto desde X, estos se almacenaron en una base de datos relacional organizada en diversas tablas. Se creó una estructura que separa la información en tablas específicas, tales como:

- Usuarios: Contiene información del usuario que realizó la publicación, como su identificador único, nombre de usuario y cantidad de seguidores.

- Ubicaciones: Almacena información sobre la georreferenciación cuando está disponible, incluyendo coordenadas y *bounding boxes* asociados.

- Publicaciones: Contiene el texto de la publicación, su identificador único, la fecha y hora de publicación, y referencias a las tablas de usuarios y ubicaciones.

Esta organización permite optimizar la gestión de los datos, evitando redundancias y facilitando su recuperación eficiente para el análisis posterior.

Preprocesamiento de texto

Una vez almacenados, los textos de las publicaciones fueron sometidos a procesos de filtrado y transformación para mejorar la calidad y el rendimiento del modelo clasificador de afirmaciones de contagio.

- Filtrado de textos de *bots*, medios de comunicación y figuras públicas:

Para asegurar que los textos analizados reflejaran experiencias personales de contagio, se descartaron publicaciones generadas por bots, medios de comunicación y figuras públicas, debido a su tendencia a publicar

contenido institucional, repetitivo o general sobre la pandemia. Esta exclusión se realizó mediante una lista negra elaborada durante el preprocesamiento, en la cual se registraron manualmente los identificadores únicos (ID's) de los usuarios considerados irrelevantes.

- La identificación de estos usuarios se realizó mediante criterios heurísticos específicos:
- Cuentas verificadas (generalmente asociadas a celebridades, figuras públicas o periodistas reconocidos).
- Usuarios con gran número de seguidores (>10,000), lo cual habitualmente caracteriza a figuras públicas y medios de comunicación.
- Perfiles con actividad excesiva (más de 200 publicaciones diarias), comportamiento típico de bots automatizados.
- Usuarios con alta proporción de contenido duplicado o repetitivo.

El proceso de filtrado fue parcialmente automatizado. Al finalizar cada día, se ejecutaba un script que contabilizaba publicaciones por usuario, detectaba textos repetidos y registraba el número de seguidores. Con base en estos resultados, los colaboradores del proyecto revisaban y validaban manualmente cuáles usuarios debían incluirse en la lista negra, asegurando así que las publicaciones analizadas fueran relevantes y auténticas para el estudio epidemiológico propuesto.

- Filtrado de textos por palabras clave: Se seleccionaron únicamente los textos que contenían términos relacionados con la enfermedad, de acuerdo con un conjunto de palabras clave identificadas manualmente. En este caso de estudio, las palabras clave fueron las diversas formas en las que las personas mencionaron al COVID-19, tales como "covicho", "cobicho", "covid", "cobid", "coronavirus", "sars-cov2" y "coronabirus". Este filtrado permitió enfocar el análisis en publicaciones potencialmente relevantes para la detección de contagios.
- Transformaciones aplicadas a los textos seleccionados: Tras el filtrado, los textos fueron sometidos a una serie de modificaciones para estandarizar su formato y reducir el ruido. Las transformaciones aplicadas a los textos se visualizan en la Tabla 2.

TABLA 2. Técnicas empleadas en el preprocesamiento de los datos.

Técnica de limpieza	Ejemplos
Conversión de hashtags relacionados con COVID-19 a texto	**Entrada**: Amigos me dio el #COVID me siento mal **Salida**: Amigos me dio el COVID me siento mal
Eliminación de URL's	**Entrada**: Si tienen covid pueden enviar registrarse en http://tengocovid.com **Salida**: Si tienen covid pueden enviar registrarse en
Eliminación de caracteres especiales y emojis	**Entrada**: ¡No puedo creerlo! 😭😢 #COVID19 está afectando a mucha gente!!! 💔💔💔 **Salida**: No puedo creerlo COVID19 está afectando a mucha gente
Manejo de caracteres especiales	Caracteres como la 'Ñ' no se eliminaron **Entrada**: Soñe o tengo covid **Salida**: Soñe o tengo covid

Clasificación de afirmaciones de contagio

En esta fase se desarrolló un clasificador binario al que se le ha denominado ConBiBER (BERT + CNN), diseñado para identificar automáticamente textos que contienen afirmaciones positivas de contagio. Este clasificador combina BERT[25][26], encargado de extraer representaciones semánticas del lenguaje, con redes convolucionales (CNN)[27], que procesan los *embeddings* generados y detectan patrones relevantes en las secuencias de texto.

El modelo BERT se seleccionó para abordar la complejidad lingüística propia de las publicaciones en redes sociales, en las que las afirmaciones de contagio pueden expresarse de manera ambigua, informal o mediante lenguaje figurado. A diferencia de los filtros basados en palabras clave, que resultaron limitados para detectar frases no literales como “ya me tocó el cobicho” o “hoy amanecí positivo #covicho”, BERT permite interpretar el significado contextual de las expresiones. Esta capacidad resulta clave para identificar menciones de contagio redactadas de forma no convencional, ampliando el alcance del clasificador más allá de las afirmaciones explícitas.

Además, BERT fue seleccionado con la intención de facilitar futuras aplicaciones en el monitoreo de otras enfermedades infecciosas, permitiendo entrenar nuevos clasificadores de manera ágil y eficiente a partir de datos etiquetados, sin necesidad de diseñar manualmente reglas o filtros específicos para cada caso.

La arquitectura de ConBiBER incluye una capa de embeddings basada en BERT, seguida de capas convolucionales con filtros de bigrama, trigrama y cuatrograma para capturar relaciones contextuales[28]. La reducción de dimensionalidad se efectuó con *GlobalMaxPooling*, cuya salida alimentó una capa densa de 512 neuronas (ReLU), y finalmente una neurona sigmoide para la clasificación binaria.

Los hiperparámetros y ajustes específicos se resumen a continuación:

- Modelo base: BERT multi-cased L-12_H-768_A-12
- Tokenización: *WordPiece*, longitud máxima = 128 *tokens*
- Capas convolucionales paralelas:
- Filtros 2-gram, 3-gram y 4-gram (100 filtros cada uno, ReLU)
- *Pooling*: GlobalMaxPooling1D
- Capa densa oculta: 256 unidades, ReLU, regularización L1/L2 = 0.001
- *Dropout*: 0.5
- Capa de salida: 1 neurona, sigmoide, L1 = 0.01
- *Batch size*: 32 Épocas: 40
- Función de pérdida: binary_crossentropy
- Optimizador: Adam
- Pesos de clase: {0:1.0, 1:8.0}

El clasificador se entrenó con un conjunto de datos desbalanceado de 4,500 textos: 3,900 con menciones generales sobre COVID-19 y 600 con afirmaciones de contagio. Esta decisión buscó simular condiciones similares a las observadas en X. Para mitigar el desbalance entre clases, se empleó una técnica basada en asignar pesos diferenciados a cada clase durante el entrenamiento.

El conjunto de 4,500 textos utilizados para el entrenamiento fue etiquetado manualmente por un colaborador del proyecto. Las publicaciones fueron previamente filtradas mediante palabras clave asociadas al COVID-19 ("sars-cov2", "covid", "covicho", "coronavirus", "cobi", "covi"), y luego clasificados en dos categorías: afirmaciones positivas,

cuando el contenido expresaba un contagio personal, y menciones generales, cuando se aludía al virus sin confirmar un diagnóstico positivo. Se priorizaron textos redactados en primera persona y con indicios claros de contagio, aislamiento o síntomas compatibles, con el fin de asegurar una base confiable para el entrenamiento del clasificador.

La Tabla 3 presenta ejemplos del conjunto de datos de entrenamiento; el conjunto completo puede consultarse en[28]:

TABLA 3. Conjunto de datos de entrenamiento del clasificador ConBiBER.

Id	Texto	Clase
1	amigos me dio el covid	1
2	raza ya di positivo al cobid	1
...	...	...
600	oren por mi me dio el covicho	1
601	la covid-19 ha aumentado en mexico	0
602	gracias a dios fui negativo al covid	0
...	...	...
4500	sigo invicto al covicho	0

Predicción de afirmaciones de contagios

Los textos clasificados por *ConBiBER* se emplearon para predecir afirmaciones de contagio utilizando el modelo matemático de *Gompertz*, conocido por su eficacia en la modelización de fenómenos biológicos y epidemiológicos[29][30][31]. Este modelo es adecuado para describir la progresión de epidemias, ya que capta tanto la fase inicial de crecimiento exponencial como la fase de estabilización.

La estructura general del modelo *Gompertz* se representa matemáticamente mediante la función (1):

$$f(t) = ae^{-be^{-ct}} \tag{1}$$

donde:

- **f(t)** representa el número acumulado de casos en el tiempo t.
- **a** simboliza el límite asintótico, o el número máximo proyectado de casos.
- **b** es un parámetro que controla la posición de la curva en el eje temporal.
- **c** define la tasa de crecimiento, relacionada directamente con la velocidad de transmisión del virus.

El modelo de *Gompertz* fue utilizado para predecir la evolución de las afirmaciones positivas de contagio identificadas en X. Para ello, se empleó una estrategia de ventanas móviles de 7, 15 y 30 días, que permitió generar series de tiempo localizadas para distintos periodos. Sobre cada una de estas ventanas, se ajustó una curva del modelo, es decir, se calcularon los valores óptimos de los parámetros *a*, *b* y *c* que permiten que la curva generada se asemeje lo más posible a la evolución acumulada de afirmaciones observadas en dicho periodo. Esta estimación se realizó mediante regresión no lineal por mínimos cuadrados, utilizando el algoritmo de *Levenberg-Marquardt*.

Para realizar este ajuste, se generó una serie de tiempo acumulada con las afirmaciones positivas clasificadas por ConBiBER dentro de cada ventana. El eje temporal *t* representa los días transcurridos desde el inicio de la ventana, y *f(t)* corresponde al número acumulado de menciones afirmativas en ese intervalo. La estimación de los parámetros se realizó mediante regresión no lineal por mínimos cuadrados, utilizando el algoritmo de *Levenberg-Marquardt*. El objetivo fue minimizar la suma del error cuadrático (diferencias elevadas al cuadrado) entre los valores predichos por la curva de *Gompertz* y los datos reales. Este procedimiento se repitió iterativamente para cada ventana, generando predicciones personalizadas a corto plazo.

La Figura 3 ilustra este procedimiento. La sección en amarillo representa los datos históricos utilizados para el ajuste del modelo en cada ventana móvil, mientras que la sección en verde muestra el horizonte de predicción generado. En la parte superior, se observa una ventana móvil de 7 días que avanza progresivamente en la serie temporal, calculando predicciones de hasta 5 días. En la parte inferior, una ventana de 15 días ofrece una base más amplia para la predicción. Este enfoque garantiza que las proyecciones se realicen con base en la información más reciente disponible.

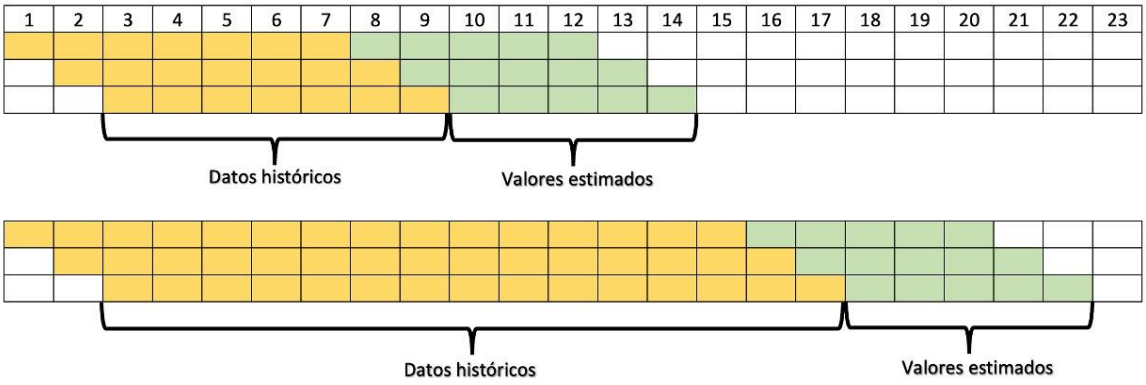

FIGURA 3. **Esquema del uso de ventanas móviles para la predicción de afirmaciones de contagio. Las ventanas de entrada (amarillo) se desplazan día a día y generan predicciones acumuladas (verde) mediante ajuste del modelo *Gompertz*.**

Este procedimiento tuvo como objetivo generar curvas que describieran de manera precisa la evolución acumulada de las afirmaciones de contagio a corto plazo. Las curvas resultantes capturaron adecuadamente tanto la fase de crecimiento exponencial como la de estabilización, lo que valida su utilidad como herramienta predictiva en contextos epidemiológicos. Las predicciones se generaron de forma iterativa, actualizando diariamente la ventana de entrada. Las métricas de error calculadas (RMSE y MAE) para horizontes de 1 y 5 días mostraron un buen ajuste, respaldando que es posible aproximar la evolución de los contagios acumulados a partir de los datos obtenidos en X.

Identificación de zonas de contagio

La identificación de zonas de contagio es un proceso clave en el monitoreo epidemiológico, ya que permite detectar áreas con alta concentración de casos positivos de contagio y analizar su evolución en el tiempo. Para ello, se empleó un enfoque geoespacial basado en datos obtenidos de X, utilizando técnicas de georreferenciación y modelado de riesgo. Este análisis se llevó a cabo mediante la integración de datos espaciales en un mapa de calor, el cual representa la distribución del riesgo de contagio en los estados de la República Mexicana. La actualización diaria de este mapa permitió identificar regiones con alta incidencia de afirmaciones de contagio, facilitando la toma de decisiones en salud pública y la detección de patrones en la propagación del virus.

La identificación de estas zonas se realizó a través de dos fuentes principales de información geográfica:

- *Bounding boxes*: X proporciona coordenadas georreferenciadas en formato de *bounding boxes*, los cuales delimitan una región aproximada de donde se emitió la publicación. Para asignar cada publicación a un estado específico, se tomó el centro del *bounding box* y se comparó con los polígonos geográficos de los 32 estados de México.

- Metadatos de la publicación: En algunos casos, X incluye el nombre del estado de la república de donde se emitió la publicación. Esta información fue utilizada cuando estaba disponible para mejorar la precisión de la asignación geográfica.

Una vez asignadas las publicaciones afirmativas de contagio a su ubicación correspondiente, se calculó un índice de riesgo para cada estado. Este índice se determinó dividiendo el número de casos activos por el número máximo esperado de casos en la región.

- Casos activos: Se definieron como las publicaciones afirmativas de contagio publicados en los últimos siete días, incluyendo la fecha de cálculo.

- Número máximo de casos esperados: Se estableció heurísticamente en función de un porcentaje de la cantidad total de publicaciones recopiladas en cada estado.

Con base en el índice de riesgo, se asignaron colores en el mapa de calor, permitiendo visualizar de manera intuitiva las áreas con mayor concentración de contagios. Los umbrales de color fueron definidos de acuerdo con la densidad de casos positivos, asegurando una representación clara del nivel de riesgo en cada estado. El índice de riesgo y la distribución de las zonas de contagio se recalcularon diariamente, generando una visualización dinámica que reflejaba la evolución de la pandemia en distintas regiones del país. Este enfoque permitió detectar zonas críticas de propagación, contribuyendo a mejorar las estrategias de vigilancia epidemiológica.

RESULTADOS Y DISCUSIÓN

Los hallazgos obtenidos a partir de la aplicación del enfoque propuesto en esta investigación se centraron en el monitoreo y predicción del COVID-19 en X. Estos resultados se presentan en tres apartados principales: análisis de afirmaciones COVID-19, evaluación de la predicción de menciones de contagios y visualización geoespacial en mapa de calor.

Análisis de afirmaciones COVID-19

El modelo de clasificación ConBiBER fue evaluado utilizando un conjunto de datos de prueba, compuesto por 650 textos relacionados con COVID-19 y 100 textos que afirmaban un diagnóstico positivo de la enfermedad. En la Figura 4 se muestra la matriz de confusión de la evaluación del conjunto de datos de prueba.

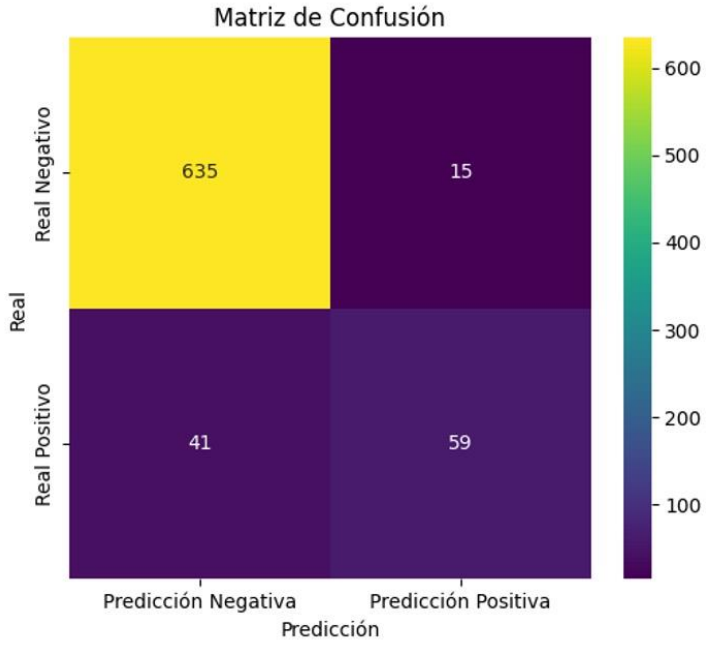

FIGURA 4. **Matriz de confusión resultante de la evaluación del modelo ConBiBER sobre un conjunto de prueba compuesto por 650 textos relacionados y 100 afirmaciones positivas de contagio.**

Las métricas de desempeño obtenidas del conjunto de datos de prueba se resumen en la Tabla 4. En particular, el modelo mostró un alto valor de Sensibilidad, lo que indica que es capaz de etiquetar correctamente la mayoría de los textos afirmativos, una característica esencial para no pasar por alto menciones de contagio en el contexto del monitoreo epidemiológico. La precisión del 73.9% refleja que el modelo incurre en una proporción relevante de falsos positivos, es decir, clasifica algunos textos no afirmativos como afirmativos. Este comportamiento está influenciado por el desbalance del conjunto de prueba, donde la clase negativa es mayoritaria. Por ello, la precisión debe interpretarse con cautela y complementarse con otras métricas más robustas ante el desbalance.

TABLA 4. **Métricas del conjunto de datos de prueba.**

Métrica	Valor
Exactitud	94.5%
Precisión	73.9%
Sensibilidad	91.0%
Especificidad	95.0%
F1-Score	81.6%

Es importante destacar que el proceso de filtrado aplicado durante el preprocesamiento contribuyó a reducir el ruido en los datos de entrada. Al eliminar publicaciones institucionales, duplicadas o no relacionadas con experiencias personales, se mejoró la coherencia semántica del conjunto de entrenamiento, lo que favoreció el rendimiento del clasificador ConBiBER. Esta depuración permitió que el modelo se enfocara en identificar patrones lingüísticos propios de declaraciones personales de contagio, lo cual se refleja en sus métricas de evaluación.

Posteriormente, el clasificador ConBiBER fue integrado en el flujo de datos de X para etiquetar publicaciones en tiempo real. Esta integración permitió clasificar casi en tiempo real publicaciones desde el 1 de noviembre de 2021 hasta el 13 de marzo de 2023.

Los resultados del etiquetado en tiempo real mostraron una correlación significativa entre las menciones en X y los contagios reportados oficialmente, con una correlación de Pearson de 0.83. Esta correlación no implica una relación causal, sino que sugiere una posible sincronicidad temporal entre ambas fuentes. Se presenta como un análisis exploratorio que evidencia cierta similitud en la forma de las curvas temporales. En este sentido, los datos extraídos de redes sociales podrían considerarse como un insumo complementario dentro de los sistemas de monitoreo epidemiológico, especialmente al capturar manifestaciones de contagio que pueden no ser registradas por los métodos tradicionales de salud pública.

Los conjuntos de datos de X y los reportes oficiales de contagios de la Secretaría de Salud de México fueron normalizados para facilitar la comparación, ajustando las series diarias a una escala común. Los datos oficiales fueron descargados del portal del Gobierno de México administrado por CONACYT[32]. La Figura 5 muestra cómo los datos de ambas fuentes presentan un comportamiento similar a lo largo del tiempo, reflejando picos y tendencias de manera alineada.

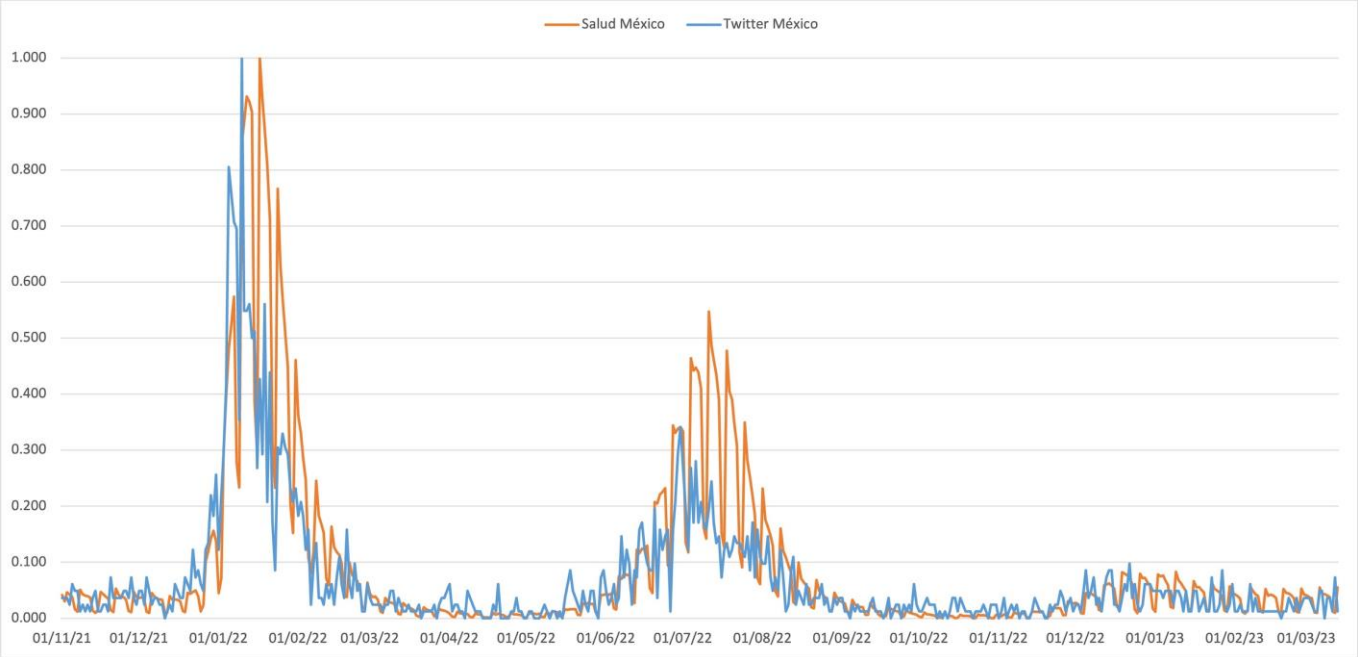

FIGURA 5. **Comparación temporal entre los casos oficiales de COVID-19 reportados por la Secretaría de Salud de México y las afirmaciones positivas clasificadas en X. Las series fueron normalizadas para facilitar la comparación. La correlación observada (r = 0.83) sugiere una sincronicidad temporal entre ambas fuentes.**

Al analizar la Figura 5, se observa que los picos de afirmaciones positivas de contagio en X tienden a estar ligeramente desplazados hacia la izquierda respecto a los picos de contagios reportados por la Secretaría de Salud. Esta diferencia temporal sugiere que las publicaciones en redes sociales podrían anticipar los brotes o incrementos en los contagios oficiales. Aunque esta hipótesis no puede ser validada directamente en este estudio, plantea la posibilidad de que las redes sociales funcionen como un sistema de alerta temprana.

Dos explicaciones plausibles sustentan esta observación. La primera es que los usuarios tienden a compartir sus síntomas o diagnósticos de manera inmediata en sus cuentas personales, antes de acudir a centros de salud o ser contabilizados en los sistemas oficiales. La segunda posibilidad es que, en algunos casos, las personas con síntomas o sospecha de contagio simplemente no son registradas oficialmente, ya sea por limitaciones de acceso, desinterés o falta de pruebas. Validar esta teoría requeriría estudios de campo con pruebas rápidas en zonas geográficas

específicas donde se detecten estas menciones, para confirmar si efectivamente anticipan un aumento real de contagios.

En este sentido, la comparación entre ambas curvas no solo permite evaluar la sincronía entre fuentes sociales y oficiales, sino que también abre la puerta al uso de datos de redes sociales como complemento temprano en estrategias de vigilancia epidemiológica.

Estos resultados pueden compararse con el trabajo de Osorio[4], cuyos autores también utilizaron Twitter para detectar posibles contagios en España. Sin embargo, en contraste con nuestro enfoque basado en clasificación supervisada con BERT (ConBiBER), en [4] emplearon un método basado en términos clave sin clasificación semántica avanzada. Esto posiciona nuestro método con mayor capacidad para interpretar afirmaciones ambiguas o no literales, mejorando así la precisión en la identificación de contagios reportados informalmente en redes sociales.

Evaluación de la predicción de menciones de contagio

El modelo predictivo de menciones de contagio se evaluó durante el período del 1 de noviembre de 2021 al 13 de marzo de 2023. En este análisis, se emplearon ventanas móviles de 7, 15 y 30 días con el objetivo de generar predicciones a 1 y 5 días a futuro.

La elección de estos tamaños de ventana se fundamentó en un equilibrio entre sensibilidad y estabilidad: las ventanas cortas (7 días) permiten capturar con mayor rapidez cambios recientes en la dinámica de contagios, mientras que las ventanas más amplias (15 y 30 días) suavizan el efecto de anomalías o fluctuaciones abruptas, favoreciendo predicciones más estables. Esta combinación permitió evaluar el desempeño del modelo en diferentes condiciones de variabilidad epidemiológica.

Las predicciones se validaron utilizando la raíz del error cuadrático medio (RMSE) y el error absoluto medio (MAE), lo que permitió un análisis detallado de la precisión en las estimaciones. Estas métricas son fundamentales para cuantificar el grado de ajuste entre los valores predichos y observados: cuanto más bajos son el RMSE y el MAE, mayor es la exactitud del modelo.

En este contexto, las menciones observadas corresponden al número acumulado diario de publicaciones clasificadas como afirmaciones positivas de contagio mediante el modelo ConBiBER. Las menciones predichas, por su parte, son las estimaciones generadas por el modelo de *Gompertz*, calculadas sobre las series más recientes mediante ventanas móviles. Esta comparación permite evaluar la capacidad del sistema para anticipar la evolución a corto plazo de los patrones de contagio expresados en redes sociales.

La Figura 6 presenta la comparación entre los valores observados y los valores predichos del modelo entrenado con un segmento de datos de 7 días previos y predicción a 1 día.

FIGURA 6. **Comparación entre las menciones afirmativas observadas y las predichas por el modelo de *Gompertz* entrenado con una ventana móvil de 7 días, con horizonte de predicción de 1 día.**

Las medidas de evaluación para el modelo predictivo entrenado con un segmento de datos de 7 días previos y cálculos a 1 día fueron RMSE = ±5.34 y MAE = 2.83.

La Figura 7 presenta la comparación entre los valores observados y los valores predichos del modelo entrenado con un segmento de datos de 7 días previos y cálculos a 5 días a futuro.

FIGURA 7. **Comparación entre las menciones afirmativas observadas y las predichas por el modelo de *Gompertz* entrenado con una ventana móvil de 7 días, con horizonte de predicción de 5 días.**

Las medidas de evaluación para el modelo predictivo entrenado con un segmento de datos de 7 días previos y cálculos a 5 días fueron RMSE = ± 11.47 y MAE = 5.18.

La Figura 8 presenta la comparación entre los valores observados y los valores predichos del modelo entrenado con un segmento de datos de 15 días previos y cálculos a 1 día a futuro.

FIGURA 8. **Comparación entre las menciones afirmativas observadas y las predichas por el modelo de *Gompertz* entrenado con una ventana móvil de 15 días, con horizonte de predicción de 1 día.**

Las medidas de evaluación para el modelo predictivo entrenado con un segmento de datos de 15 días previos y cálculos a 1 día fueron RMSE = ± 5.14 y MAE = 2.61.

La Figura 9 presenta la comparación entre los valores observados y los valores predichos del modelo entrenado con un segmento de datos de 15 días previos y cálculos a 5 días a futuro.

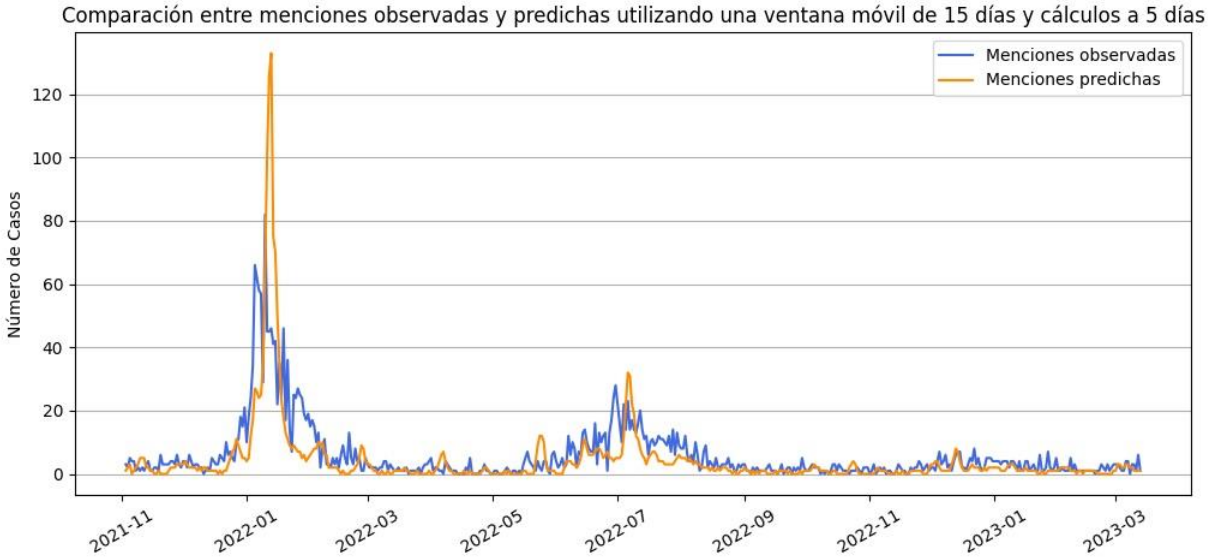

FIGURA 9. **Comparación entre las menciones afirmativas observadas y las predichas por el modelo de *Gompertz* entrenado con una ventana móvil de 15 días, con horizonte de predicción de 5 días.**

Las medidas de evaluación para el modelo predictivo entrenado con un segmento de datos de 15 días previos y cálculos a 5 días fueron RMSE = ± 8.57 y MAE = 3.73

La Figura 10 presenta la comparación entre los valores observados y los valores predichos del modelo entrenado con un segmento de datos de 30 días previos y cálculos a 1 día a futuro.

FIGURA 10. **Comparación entre las menciones afirmativas observadas y las predichas por el modelo de *Gompertz* entrenado con una ventana móvil de 30 días, con horizonte de predicción de 1 día.**

Las medidas de evaluación para el modelo predictivo entrenado con un segmento de datos de 30 días previos y cálculos a 1 día fueron RMSE = ± 5.98 y MAE = 2.96.

La Figura 11 presenta la comparación entre los valores observados y los valores predichos del modelo entrenado con un segmento de datos de 30 días previos y cálculos a 5 días a futuro.

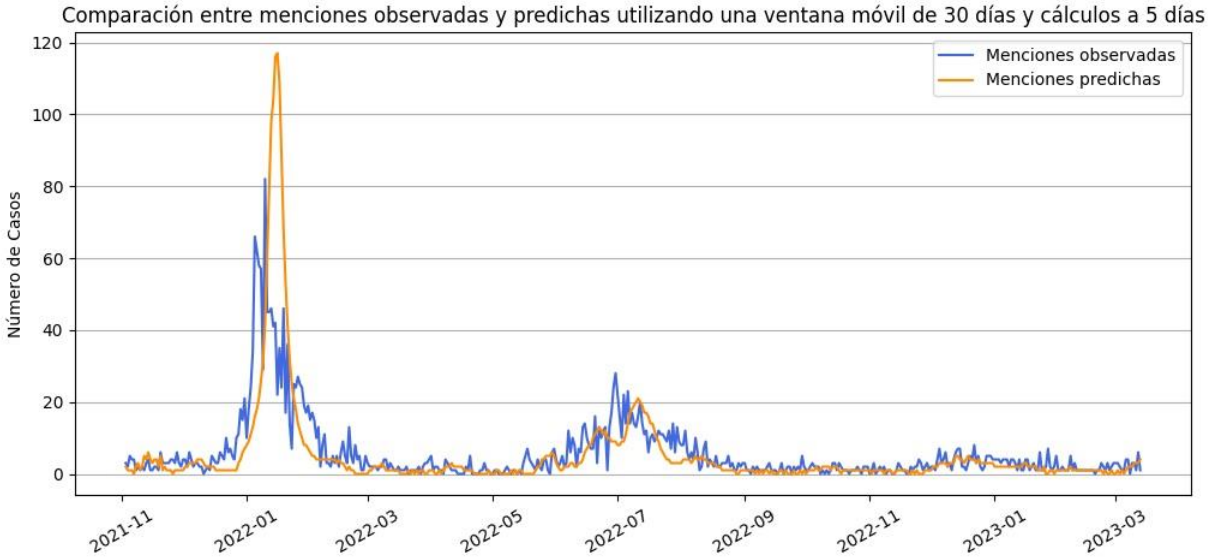

FIGURA 11. **Comparación entre las menciones afirmativas observadas y las predichas por el modelo de *Gompertz* entrenado con una ventana móvil de 30 días, con horizonte de predicción de 5 días.**

Las medidas de evaluación para el modelo predictivo entrenado con un segmento de datos de 30 días previos y cálculos a 5 días fueron RMSE = ± 10.08 y MAE = 3.89.

Las gráficas y métricas de error muestran que el modelo logra capturar adecuadamente las tendencias generales en la evolución de las afirmaciones de contagio. En general, el modelo tiende a comportarse de manera similar a las menciones observadas, con una ligera subestimación incluso durante picos moderados. No obstante, en eventos con cambios muy abruptos (como el registrado en enero de 2022) se observa una tendencia a la sobreestimación.

Se identificó que el modelo ofrece mejor precisión al realizar predicciones con un día de anticipación en comparación con las realizadas a cinco días. Asimismo, el modelo entrenado con una ventana móvil de 15 días fue el que mostró mejor desempeño tanto en el horizonte de 1 día como en el de 5 días, logrando un equilibrio entre estabilidad y sensibilidad ante cambios recientes. Estos hallazgos permiten orientar futuros ajustes del sistema de predicción hacia combinaciones óptimas de ventana y horizonte de pronóstico.

Varios estudios recientes han propuesto modelos predictivos aplicados a datos provenientes de redes sociales para anticipar el comportamiento de enfermedades infecciosas. Por ejemplo, investigaciones como las de [33][34] han empleado redes neuronales recurrentes (LSTM y variantes RNN) para proyectar la evolución del COVID-19. Si bien estos modelos ofrecen buena capacidad de predicción, su naturaleza de caja negra limita la interpretabilidad de los resultados en contextos epidemiológicos.

Otros trabajos, como [35] y [36], exploran comparativamente modelos como Prophet, ARIMA y redes neuronales, destacando sus ventajas en ciertas métricas, pero sin incorporar señales semánticas directas de afirmaciones personales.

Por su parte, enfoques basados en modelos compartimentales, como los SEIR modificados (por ejemplo, [37] y [38]), permiten incorporar parámetros clínicos y poblacionales, pero no aprovechan el contenido generado en redes sociales como fuente directa de detección temprana.

En contraste, el enfoque propuesto en este trabajo integra señales sociales explícitas (afirmaciones positivas de contagio clasificadas mediante ConBiBER) con un modelo interpretable de crecimiento epidémico (Gompertz), lo cual ofrece una combinación de sensibilidad semántica, capacidad de adaptación y claridad en la evolución temporal. Esta arquitectura permite realizar predicciones a corto plazo ajustadas dinámicamente, con mayor trazabilidad que los modelos tipo LSTM, y mayor especificidad textual que los enfoques basados exclusivamente en frecuencia de términos o movilidad.

Visualización geoespacial en mapa de calor

Los casos activos de COVID-19 y el número máximo esperado de casos positivos, se utilizaron para calcular un índice de riesgo que se calcula diariamente. A partir del índice de riesgo, se asignaron los colores en el mapa de calor que muestra la distribución geográfica del riesgo de contagio en los estados de la República Mexicana.

Con el fin de contrastar visualmente el riesgo estimado por el sistema propuesto con el semáforo epidemiológico de la Secretaría de Salud, se realizó una comparación cualitativa para el periodo del 1 al 14 de noviembre de 2021. La Figura 12 presenta dicha comparación: el panel A muestra el semáforo epidemiológico oficial; los paneles B, C y D corresponden a los mapas generados por este estudio para los días 1, 7 y 14 de noviembre, respectivamente.

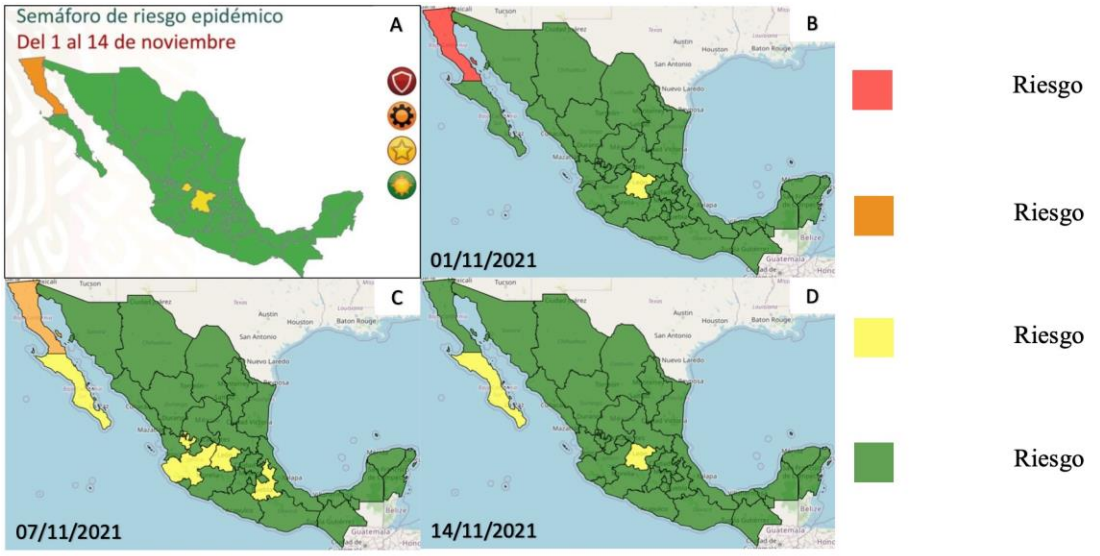

FIGURA 12. **Comparación visual entre el semáforo epidemiológico oficial (panel A) y el índice de riesgo estimado a partir de afirmaciones clasificadas (paneles B–D). Fechas representadas: 01/11/2021 (B), 07/11/2021 (C), 14/11/2021 (D).**

De manera similar, se realizó una segunda comparación cualitativa para el periodo del 29 de noviembre al 12 de diciembre de 2021. La Figura 13 presenta los resultados: el panel A muestra el semáforo epidemiológico oficial para dicho intervalo, mientras que los paneles B, C y D corresponden a los mapas generados por este estudio para los días 29 de noviembre, 6 de diciembre y 12 de diciembre, respectivamente. La escala de colores empleada se detalla en la leyenda.

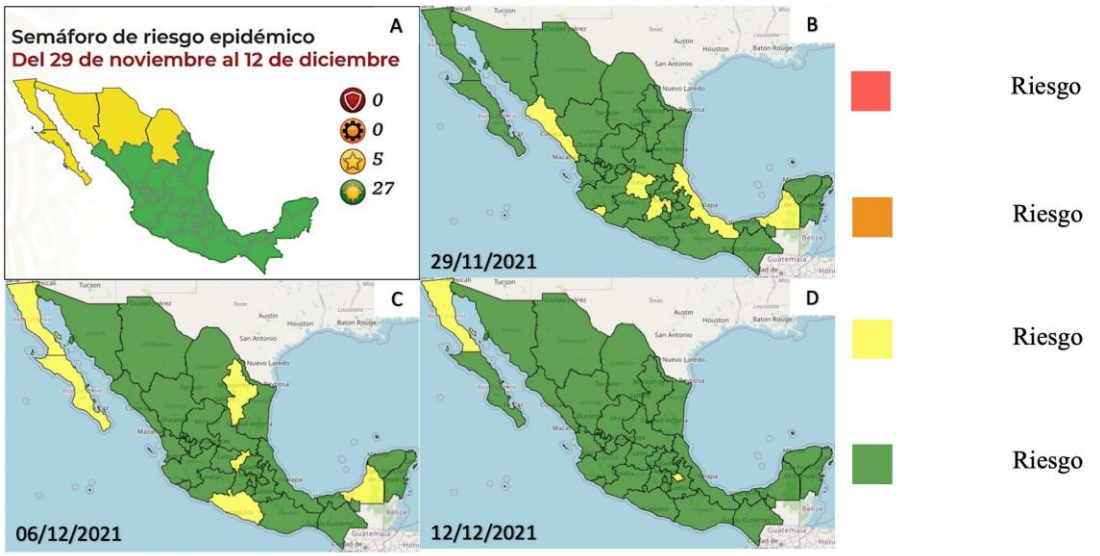

FIGURA 13. **Comparación visual entre el semáforo epidemiológico oficial (panel A) y el índice de riesgo estimado a partir de afirmaciones clasificadas (paneles B–D). Fechas representadas: 29/11/2021 (B), 06/12/2021 (C), 12/12/2021 (D).**

De manera similar, se llevó a cabo una tercera comparación cualitativa correspondiente al periodo del 10 de enero al 23 de enero de 2022. En la Figura 14 se muestran los resultados: el panel A representa el semáforo de la Secretaría de Salud para dicho intervalo, mientras que los paneles B, C y D corresponden a los mapas generados por este estudio para los días 10, 16 y 23 de enero, respectivamente.

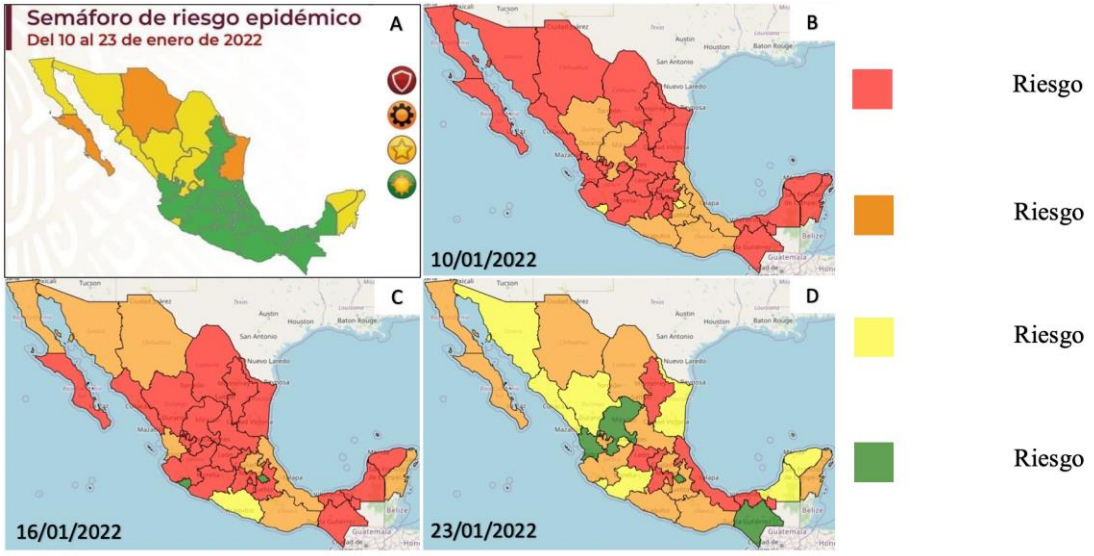

FIGURA 14. **Comparación visual entre el semáforo epidemiológico oficial (panel A) y el índice de riesgo estimado a partir de afirmaciones clasificadas (paneles B–D). Fechas representadas: 10/01/2022 (B), 16/01/2022 (C), 23/01/2022 (D).**

Aunque el índice de riesgo desarrollado en este estudio se basa en afirmaciones positivas extraídas de X y no en indicadores clínicos oficiales, la comparación visual realizada con el semáforo epidemiológico de la Secretaría de Salud permite identificar zonas de coincidencia y divergencia relevantes. Esta validación visual respalda la utilidad del enfoque como herramienta complementaria para la vigilancia geoespacial.

A diferencia del semáforo de la Secretaría de Salud, que se actualizaba cada 14 días con base en múltiples métricas (por ejemplo, ocupación hospitalaria, tasa de reproducción, número de contagios, disponibilidad de pruebas, entre otras), el índice propuesto en este estudio se actualiza diariamente. Esto representa una ventaja significativa, al permitir una detección más oportuna de cambios recientes en la dinámica epidemiológica. Esta diferencia en la frecuencia de actualización implica que el sistema basado en redes sociales puede anticipar variaciones de riesgo con horas o días de antelación, frente a esquemas oficiales que dependen de procesos institucionales más lentos para consolidar y validar datos.

Además, al emplear exclusivamente publicaciones clasificadas como afirmaciones positivas (en lugar de menciones generales sobre COVID-19, como en el caso de [4]), este enfoque ofrece una representación más específica de zonas potenciales de contagio activo, priorizando la especificidad de la señal sobre el volumen total de publicaciones.

Otros estudios, como [19] y [18], se enfocan principalmente en el análisis textual para identificar síntomas o diagnósticos auto-reportados a partir de publicaciones en redes sociales. Sin embargo, dichos enfoques no incorporan componentes de análisis geoespacial, lo que limita su utilidad para identificar zonas de riesgo localizadas. A diferencia de estos trabajos, la propuesta presentada en este estudio combina la detección semántica de contagios con proyección temporal y espacial, ofreciendo una solución integral y dinámica que puede adaptarse a distintas regiones y servir como marco replicable para futuras enfermedades emergentes.

CONCLUSIONES

El análisis de menciones relacionadas con COVID-19 en X, combinado con la modelización mediante BERT y la función de *Gompertz*, constituye una herramienta de monitoreo que puede complementar los métodos convencionales de vigilancia epidemiológica. La correlación observada entre los picos de menciones en X y los datos oficiales de contagio no implica una relación causal, pero sugiere una posible sincronicidad temporal. Este hallazgo exploratorio pone de manifiesto que las redes sociales pueden captar señales tempranas de brotes, especialmente en contextos donde los sistemas formales presentan rezagos en la recolección y consolidación de información.

A través de la metodología empleada, se ha logrado una caracterización predictiva del comportamiento del virus, donde el uso de ventanas móviles aporta dinamismo al modelo, facilitando así una estimación ajustada y continua de la situación epidemiológica.

La representación geoespacial del riesgo mediante mapas de calor refuerza la utilidad de este enfoque, proporcionando una visualización estratégica que permite identificar áreas de riesgo elevado en la República Mexicana, favoreciendo la toma de decisiones informadas en salud pública.

X ofrece una fuente vasta de datos en tiempo real que puede aprovecharse para el monitoreo epidemiológico. No obstante, es importante tener en cuenta que X no representa equitativamente a toda la población, ya que ciertos grupos demográficos (como adultos mayores, personas con menor acceso a tecnología y poblaciones rurales) tienden a estar subrepresentados en el uso de redes sociales. Por lo tanto, el monitoreo epidemiológico basado en X debe usarse con cautela y como un complemento de la vigilancia epidemiológica tradicional.

Finalmente, este enfoque de monitoreo requiere ajustes específicos según el caso de estudio, ya sea en el seguimiento de diferentes enfermedades o en la aplicación en otros países.

Como parte del trabajo futuro, se plantea mejorar el etiquetado automático de textos afirmativos relacionados con COVID-19 y extender esta metodología a otras enfermedades infecciosas. Para ello, se prevé explorar modelos avanzados, como GPT y otros basados en atención (*transformers*).

Además, se contempla combinar el modelo *Gompertz* con el modelo autorregresivo integrado de media móvil (ARIMA) o con redes neuronales recurrentes (RNN) para mejorar la precisión del modelo predictivo en proyecciones mayores a 5 días. De tal manera que el modelo *Gompertz* capture la fase inicial de crecimiento, y el modelo ARIMA o las redes neuronales recurrentes, se encarguen de las fases posteriores.

Asimismo, se propone desarrollar un índice de riesgo más robusto mediante la incorporación de variables adicionales, tales como factores socioeconómicos, comportamiento de movilidad y patrones demográficos. Un índice de riesgo más robusto se ajustará mejor a las realidades locales, lo cual contribuirá a una mejor comprensión de la dinámica de contagio.

CONFLICTOS DE INTERÉS

Los autores declaran que no tienen conflictos de intereses.

REFERENCIAS

[1] SAS, "Solutions for real-time disease surveillance and data-driven decision making." Accessed: Oct. 31, 2024. [Online]. Available: https://www.sas.com/es_mx/insights/articles/analytics/situational-awareness-guides-our-responses---routine-to-crisis.html

[2] Sasha Walek, "New COVID Local Risk Index Identifies Neighborhoods at Highest Risk of Pandemic's Impact." Accessed: Oct. 31, 2024. [Online]. Available: https://nyulangone.org/news/new-covid-local-risk-index-identifies-neighborhoods-highest-risk-pandemics-impact

[3] S. Masri et al., "Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic," BMC Public Health, vol. 19, no. 1, Jun. 2019, doi: https://doi.org/10.1186/s12889-019-7103-8

[4] J. O. Arjona, "Geolocated Social Networks And Covid-19: Analysis Of The Temporal And Spatial Activity Of Twitter Users In Spain During The Pandemic," GeoFocus, vol. 2022, no. 30, pp. 25–47, Dec. 2022, doi: https://doi.org/10.21138/GF.789.

[5] M. U. Hoque et al., "Analyzing Tweeting Patterns and Public Engagement on Twitter During the Recognition Period of the COVID-19 Pandemic: A Study of Two U.S. States," IEEE Access, vol. 10, pp. 72879–72894, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3189670.

[6] J. S. P. Tulloch, R. Vivancos, R. M. Christley, A. D. Radford, and J. C. Warner, "Mapping tweets to a known disease epidemiology; a case study of Lyme disease in the United Kingdom and Republic of Ireland," J Biomed Inform X, vol. 4, Dec. 2019, doi: https://doi.org/10.1016/j.yjbinx.2019.100060.

[7] S. Chae, S. Kwon, and D. Lee, "Predicting infectious disease using deep learning and big data," Int J Environ Res Public Health, vol. 15, no. 8, Aug. 2018, doi: https://doi.org/10.3390/ijerph15081596.

[8] M. Birjali, A. Beni-Hssane, and M. Erritali, "Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks," in Procedia Computer Science, Elsevier B.V., 2017, pp. 65–72. doi: https://doi.org/10.1016/j.procs.2017.08.290.

[9] D. Reynard and M. Shirgaokar, "Harnessing the power of machine learning: Can Twitter data be useful in guiding resource allocation decisions during a natural disaster?," Transp Res D Transp Environ, vol. 77, pp. 449–463, Dec. 2019, doi: https://doi.org/10.1016/j.trd.2019.03.002.

[10] Y. Gu, Y. Yao, W. Liu, and J. Song, "We know where you are: Home location identification in location-based social networks," in 2016 25th International Conference on Computer Communications and Networks, ICCCN 2016, Institute of Electrical and Electronics Engineers Inc., Sep. 2016. doi: https://doi.org/10.1109/ICCCN.2016.7568598.

[11] R. P. D. Redondo, C. Garcia-Rubio, A. F. Vilas, C. Campo, and A. Rodriguez-Carrion, "A hybrid analysis of LBSN data to early detect anomalies in crowd dynamics," Future Generation Computer Systems, vol. 109, pp. 83–94, 2020, doi: https://doi.org/10.1016/j.future.2020.03.038.

[12] O. Alonso, V. Kandylas, S. E. Tremblay, and S. Whiting, "Answering recreational web searches with relevant things to do results," Inf Process Manag, vol. 57, no. 2, p. 102184, 2020, doi: https://doi.org/10.1016/j.ipm.2019.102184.

[13] X. Xiong, S. Qiao, Y. Li, N. Han, G. Yuan, and Y. Zhang, "A point-of-interest suggestion algorithm in Multi-source geo-social networks," Eng Appl Artif Intell, vol. 88, no. September 2019, p. 103374, 2020, doi: https://doi.org/10.1016/j.engappai.2019.103374.

[14] W. Shi, D. Liu, J. Yang, J. Zhang, S. Wen, and J. Su, "Social bots' sentiment engagement in health emergencies: A topic-based analysis of the covid-19 pandemic discussions on twitter," Int J Environ Res Public Health, vol. 17, no. 22, pp. 1–19, Nov. 2020, doi: https://doi.org/10.3390/ijerph17228701.

[15] V. Suarez-Lledo and J. Alvarez-Galvez, "Assessing the Role of Social Bots During the COVID-19 Pandemic: Infodemic, Disagreement, and Criticism," J Med Internet Res, vol. 24, no. 8, Aug. 2022, doi: https://doi.org/10.2196/36085.

[16] J. Jiang, X. Ren, and E. Ferrara, "Social Media Polarization and Echo Chambers in the Context of COVID-19: Case Study," JMIRx Med, vol. 2, no. 3, p. e29570, Aug. 2021, doi: https://doi.org/10.2196/29570.

[17] J. Xue et al., "Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach," J Med Internet Res, vol. 22, no. 11, Nov. 2020, doi: https://doi.org/10.2196/20550.

[18] A. Z. Klein, S. Kunatharaju, K. O'Connor, and G. Gonzalez-Hernandez, "Automatically Identifying Self-Reports of COVID-19 Diagnosis on Twitter: An Annotated Data Set, Deep Neural Network Classifiers, and a Large-Scale Cohort," J Med Internet Res, vol. 25, 2023, doi: https://doi.org/10.2196/46484.

[19] Sarker, S. Lakamana, W. Hogg-Bremer, A. Xie, M. A. Al-Garadi, and Y.-C. Yang, "Self-reported COVID-19 symptoms on Twitter: An analysis and a research resource," Apr. 22, 2020. doi: https://doi.org/10.1101/2020.04.16.20067421.

[20] E. Ferrara, "What types of COVID-19 conspiracies are populated by Twitter bots?," First Monday, May 2020, doi: https://doi.org/10.5210/fm.v25i6.10633.

[21] P. Castioni, G. Andrighetto, R. Gallotti, E. Polizzi, and M. De Domenico, "The voice of few, the opinions of many: Evidence of social biases in Twitter COVID-19 fake news sharing," R Soc Open Sci, vol. 9, no. 10, Oct. 2022, doi: https://doi.org/10.1098/rsos.220716.

[22] J. Mellon and C. Prosser, "Twitter and Facebook are not representative of the general population: Political attitudes and demographics of british social media users," Research and Politics, vol. 4, no. 3, Jul. 2017, doi: https://doi.org/10.1177/2053168017720008.

[23] S. Wojcik and A. Hughes, "Sizing Up Twitter Users For Media Or Other Inquiries," 2019. [Online]. Available: www.pewresearch.org.

[24] Harmon, "Tweepy," Twitter for Python. Accessed: Apr. 08, 2025. [Online]. Available: https://www.tweepy.org

[25] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." [Online]. Available: https://github.com/tensorflow/tensor2tensor

[26] A. Rogers, O. Kovaleva, and A. Rumshisky, "A Primer in BERTology: What We Know About How BERT Works", doi: https://doi.org/10.1162/tacl_a_00349

[27] K. Kaur and P. Kaur, "BERT-CNN: Improving BERT for Requirements Classification using CNN," in Procedia Computer Science, Elsevier B.V., 2022, pp. 2604–2611. doi: https://doi.org/10.1016/j.procs.2023.01.234.

[28] Pedro Wences, "Monitoreo epidemiológico de COVID-19 con datos de X," CENIDET. Accessed: May 11, 2025. [Online]. Available: https://github.com/wopjk/MonitoreoCovidTwitterX

[29] P. Winsor, "The Gompertz Curve As A Growth Curve," 1932. [Online]. Available: https://doi.org/10.1073/pnas.18.1.1

[30] K. M. C. Tjørve and E. Tjørve, "The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family," PLoS One, vol. 12, no. 6, Jun. 2017, doi: https://doi.org/10.1371/journal.pone.0178691.

[31] E. Pelinovsky et al., "Gompertz model in COVID-19 spreading simulation," Chaos Solitons Fractals, vol. 154, Jan. 2022, doi: https://doi.org/10.1016/j.chaos.2021.111699.

[32] CONAHCYT, "Covid-19 México," CONAHCYT. Accessed: May 11, 2025. [Online]. Available: https://datos.covid-19.conacyt.mx

[33] A. Tarafdar, J. Mahato, R. K. Upadhyay, and P. Bhattacharya, "Exploring the synergy of media awareness and quarantine classes in SiSAQEIH model for pandemic control: A Deep LSTM-RNN predictions," Physica D, vol. 474, p. 134563, Apr. 2025, doi: https://doi.org/10.1016/j.physd.2025.134563.

[34] W. Yang and X. Chang, "Time series analysis and prediction of the trends of COVID-19 epidemic in Singapore based on machine learning," Computer Methods and Programs in Biomedicine Update, vol. 7, Jan. 2025, doi: https://doi.org/10.1016/j.cmpbup.2025.100190.

[35] A. Chhabra et al., "Sustainable and intelligent time-series models for epidemic disease forecasting and analysis," Sustainable Technology and Entrepreneurship, vol. 3, no. 2, May 2024, doi: https://doi.org/10.1016/j.stae.2023.100064.

[36] S. Y. Ilu and R. Prasad, "Improved autoregressive integrated moving average model for COVID-19 prediction by using statistical significance and clustering techniques," Heliyon, vol. 9, no. 2, Feb. 2023, doi: https://doi.org/10.1016/j.heliyon.2023.e13483.

[37] W. Angulo, J. M. Ramírez, D. De Cecchis, J. Primera, H. Pacheco, and E. Rodríguez-Román, "A modified SEIR model to predict the behavior of the early stage in coronavirus and coronavirus-like outbreaks," Sci Rep, vol. 11, no. 1, Dec. 2021, doi: https://doi.org/10.1038/s41598-021-95785-y.

[38] R. Ramalingam, A. J. Gnanaprakasam, and S. Boulaaras, "Stability and control analysis of COVID-19 spread in India using SEIR model," Sci Rep, vol. 15, no. 1, Dec. 2025, doi: https://doi.org/10.1038/s41598-025-93994-3.