EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Doctorado

Predicción de dispersión de contagios de enfermedades infecciosas utilizando redes sociales

presentada por
M.C. Pedro Wences Olguín

como requisito para la obtención del grado de
Doctor en Ciencias de la Computación

Director de tesis
Dra. Alicia Martínez Rebollar

Codirector de tesis
Dr. Hugo Estrada Esquivel

Cuernavaca, Morelos, México. Noviembre de 2025.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Mor., a 19 de agosto de 2025

DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
P R E S E N T E

AT'n: DR. JUAN GABRIEL GONZÁLEZ SERNA
PRESIDENTE DEL CLAUSTRO DOCTORAL DEL DEPARTAMENTO DE CIENCIAS COMPUTACIONALES

Los abajo firmantes, miembros del Comité Tutorial del estudiante **M.C. PEDRO WENCES OLGUÍN** manifiestan que después de haber revisado el documento de tesis titulado **"Predicción de dispersión de contagios de enfermedades infecciosas utilizando redes sociales"**, realizado bajo la dirección de la **Dra. Alicia Martínez Rebollar** y codirección del **Dr. Hugo Estrada Esquivel,** el trabajo se ACEPTA para proceder a su impresión.

A T E N T A M E N T E

Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al Servicio de México"

DRA. ALICIA MARTÍNEZ REBOLLAR
TECNM/CENIDET

DR. HUGO ESTRADA ESQUIVEL
TECNM/CENIDET

DR. FRANCISCO JAVIER RUIZ ORTEGA
TECNM/I.T. TORREÓN

DR. JAVIER ORTIZ HERNÁNDEZ
TECNM/CENIDET

DRA. MARÍA YASMÍN HERNÁNDEZ PÉREZ
TECNM/CENIDET

DR. SABINO MIRANDA JIMÉNEZ
INFOTEC-AGUASCALIENTES

c.c.p: C Verónica Sotelo Boyas/ Jefa del Departamento de Servicios Escolares
c.c.p: C Nóe Alejandro Castro Sánchez / Jefe del Departamento de Ciencias Computacionales
c.c.p: Expediente

Centro Nacional de Investigación y Desarrollo tecnológico
Subdirección Académica

Cuernavaca Mor, 25/agosto/2025
Oficio No. SAC/211/2025
Asunto: Autorización de impresión de tesis

PEDRO WENCES OLGUÍN
CANDIDATO AL GRADO DE DOCTOR
EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Predicción de dispersión de contagios de enfermedades infecciosas utilizando redes sociales"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al Servicio de México"

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

c.c.p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 4104,
e-mail: acad_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

A Samara Wences y a mi bebé que viene en camino.
Llegaron en el momento perfecto.

Agradecimientos

A Dios por la vida, por su guía, protección, amor y salvación.

A mi esposa Ana Patricia por su amor, compañía, comprensión y paciencia en este doctorado. Te amo.

A mis padres, por ser un ejemplo de perseverancia, paciencia, humildad, sabiduría, fe y bondad. Ustedes siempre han estado a mi lado apoyándome en todo. Los amo. Gracias mamá, tus oraciones siempre me alcanzan.

A mis hermanos Eder, Edson y Febe por sus consejos y apoyo económico, moral y espiritual. Los amo.

Al Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) por el apoyo económico que me brindó para realizar mis estudios de doctorado.

Al Tecnológico Nacional de México campus Centro Nacional de Investigación y Desarrollo Tecnológico (TecNM/CENIDET) por haberme brindado la oportunidad de realizar mis estudios de posgrado en sus instalaciones.

A mi directora de tesis, Dra. Alicia Martínez Rebollar, por sus valiosos consejos, observaciones y paciencia para guiar mi trabajo de investigación. Le agradezco profundamente su apoyo e insistencia para que concluyera mi doctorado.

A mi codirector de tesis, Dr. Hugo Estrada Esquivel, por sus consejos, dinamismo y confianza.

A los miembros del comité tutorial: Dr. Joaquín Pérez Ortega, Dr. Javier Ortiz Hernández, Dra. Yasmín Hernández Pérez, Dr. Francisco Javier Ruiz Ortega y Dr. Sabino Miranda Jiménez, quienes me acompañaron a lo largo del doctorado y me aportaron sus conocimientos y comentarios que permitieron enriquecer mi formación doctoral.

A mis compañeros y amigos, Irving Rosas, Julio V, Armando Granillo, Gerardo Galindo, Rodrigo, Francisco Javier, Julia, Keila, Jorge, Samuel, Karen Gama, Itzel Blanco, que me acompañaron dentro y fuera del CENIDET.

Gracias a todos los que creyeron en mí.

Resumen

La identificación temprana y el monitoreo efectivo de enfermedades infecciosas son esenciales para la gestión de pandemias, como la del COVID-19. Estos procesos permiten implementar medidas preventivas y de contención de manera rápida y precisa, reduciendo la propagación del virus y salvando vidas. Además, facilitan la asignación adecuada de recursos médicos y logísticos, y mejoran la comunicación pública.

Las instituciones de salud suelen enfrentar demoras en la actualización y reporte de casos confirmados, lo que resulta en información desactualizada. La información desactualizada tiene como consecuencia, que las personas tomen decisiones no consientes de las áreas de mayor contagio y, por ende, aumentar el riesgo de contagio.

En este sentido, está tesis presenta un enfoque innovador de identificación temprana y monitoreo efectivo de enfermedades infecciosas como alternativa a los métodos tradicionales. Este enfoque, hace uso de datos generados en tiempo real a través de la red social X. Utilizando técnicas de procesamiento de lenguaje natural y análisis de datos, ofreciendo una solución innovadora para la gestión oportuna de pandemias.

El enfoque presentado en esta tesis utiliza la información proporcionada por X para monitorear la propagación del COVID-19. El enfoque se compone de tres fases principales. La primera fase, emplea un clasificador binario basado en BERT y redes convolucionales para etiquetar automáticamente las publicaciones en las que los usuarios afirman estar enfermos de COVID-19.

La segunda fase, se encarga de predecir el número de casos positivos de COVID-19 utilizando la función Gompertz, lo que permite realizar predicciones hasta 7 días en el futuro basadas en los casos afirmativos registrados diariamente.

La tercera fase, identifica las zonas de contagio, calcula un índice de riesgo y asigna un nivel de riesgo (alto, medio-alto, medio o bajo) a cada zona.

El enfoque desarrollado en esta tesis fue validado mediante pruebas experimentales, demostrando que la información generada en redes sociales puede ser una herramienta valiosa para la detección temprana y el monitoreo de enfermedades infecciosas. Los resultados obtenidos indican que el uso de X como fuente de datos en tiempo real mejora significativamente la gestión de la pandemia, proporcionando información actualizada y precisa sobre la propagación del COVID-19.

La investigación demuestra que las redes sociales basadas en la ubicación (*Location-Based Social Networking*, LBSN) son una fuente rica de datos que puede ser aprovechada para mejorar la detección y monitoreo de enfermedades infecciosas, ofreciendo una alternativa eficaz a las demoras en la actualización de datos por parte de las instituciones de salud.

Abstract

Early Identification and Effective Monitoring of Infectious Diseases are Essential for Pandemic Management, Such as COVID-19. These processes enable the implementation of preventive and containment measures more quickly and accurately, reducing the spread of the virus and saving lives. Additionally, they facilitate the proper allocation of medical and logistical resources and improve public communication.

Health institutions often face delays in updating and reporting confirmed cases, resulting in outdated information. Outdated information leads to individuals making uninformed decisions about high-contagion areas, thereby increasing the risk of infection.

In this context, this thesis presents an approach for early identification and effective monitoring as an alternative to traditional methods. This approach uses real-time data generated through the social network X. Utilizing natural language processing and data analysis techniques, it offers an innovative solution for pandemic management.

The approach presented in this thesis uses information provided by X to monitor the spread of COVID-19. The approach consists of three main phases. The first phase employs a binary classifier based on BERT and convolutional networks to automatically label posts in which users claim to be sick with COVID-19.

The second phase is responsible for predicting the number of positive COVID-19 cases using the Gompertz function, allowing predictions up to 7 days into the future based on daily recorded positive cases.

The third phase identifies contagion zones, calculates a risk index, and assigns a risk level (high, medium-high, medium, or low) to each zone.

The approach developed in this thesis was validated through experimental tests, demonstrating that information generated on social media can be a valuable tool for early detection and monitoring of infectious diseases. The results obtained indicate that using X as a real-time data source significantly improves pandemic management, providing updated and accurate information on the spread of COVID-19.

The research shows that Location-Based Social Networking (LBSN) is a rich data source that can be leveraged to improve the detection and monitoring of infectious

diseases, offering an effective alternative to delays in data updates by health institutions.

Tabla de contenido

Lista de figuras

Lista de tablas

1

Introducción

En este capítulo se presenta el contexto general de la investigación, abordando los elementos que motivaron su desarrollo y los objetivos que se persiguen. Se expone el planteamiento del problema, la justificación y motivación del estudio, así como los objetivos generales y específicos que orientan la propuesta metodológica. Asimismo, se enuncian la hipótesis de trabajo y la organización del documento, con el fin de ofrecer al lector una visión clara de la estructura y el alcance de la tesis.

1.1 Contexto de la investigación

La actualización de las zonas de contagio de enfermedades infecciosas depende en gran medida de los datos proporcionados por las instituciones del sistema de salud. Sin embargo, estas entidades a menudo enfrentan retrasos en la recolección, validación y difusión de los casos confirmados, lo cual limita la oportunidad de respuesta y reduce la visibilidad de los focos de contagio recientes. En consecuencia, la información epidemiológica disponible tiende a estar desactualizada o incompleta, dificultando la toma de decisiones en tiempo real [1], [2].

En contraste, el auge de las redes sociales ha generado nuevas oportunidades para la recolección de información pública en tiempo real. En particular, las redes sociales basadas en ubicación (*Location-Based Social Networking*, LBSN, por sus siglas en inglés) permiten asociar contenido generado por los usuarios como: publicaciones, imágenes o videos con ubicaciones geográficas específicas. Esta funcionalidad ha sido

ampliamente explorada en investigaciones sobre recomendaciones contextuales, movilidad urbana, comportamiento de usuarios y, más recientemente, en el monitoreo de fenómenos de salud pública, como la propagación de enfermedades infecciosas.

Una de las principales ventajas de estas plataformas es su carácter dinámico y continuo: los usuarios publican contenidos de manera espontánea y masiva, lo que permite capturar información sobre eventos en el momento en que ocurren. Además, muchas de estas plataformas, como la red social X (anteriormente Twitter), ofrecen interfaces de programación de aplicaciones (*Application Programming Interface*, API, por sus siglas en inglés) que permiten acceder a sus flujos de datos en tiempo real, incluyendo texto, metadatos y, en ciertos casos, datos de geolocalización.

La API de X constituye una fuente valiosa para el análisis automatizado de contenido textual asociado a eventos sanitarios. En este contexto, se propone un nuevo enfoque para el monitoreo de enfermedades infecciosas, basado en el análisis de la información generada en esta red social, con un estudio de caso centrado en el COVID-19. El enfoque desarrollado se compone de tres módulos fundamentales:

- Clasificación de afirmaciones de contagio: Un clasificador binario, construido mediante modelos BERT y redes neuronales convolucionales, identifica automáticamente publicaciones de usuarios que manifiestan estar enfermos de COVID-19, permitiendo la extracción y estructuración de afirmaciones relevantes.
- Predicción epidemiológica: Se emplea la función de crecimiento Gompertz para estimar, con base en los datos extraídos, el número de casos positivos en un horizonte de cinco días, generando una proyección dinámica del comportamiento del brote.
- Delimitación geográfica de riesgo: Se desarrolla un algoritmo heurístico que considera factores como la ubicación de las afirmaciones positivas, la movilidad de los usuarios, el tiempo de incubación y recuperación, la densidad poblacional y las predicciones generadas, para identificar zonas de contagio y asignarles un nivel de riesgo (alto, medio-alto, medio o bajo).

Esta investigación surge de la necesidad de diseñar herramientas computacionales capaces de complementar los sistemas tradicionales de vigilancia epidemiológica, mediante el uso de fuentes alternativas de datos sociales, con el objetivo de mejorar la detección temprana, el monitoreo continuo y la toma de decisiones informadas en contextos de emergencia sanitaria.

1.2 Planteamiento del problema

La propagación acelerada de enfermedades infecciosas, como la ocasionada por el virus SARS-CoV-2 (COVID-19), ha puesto en evidencia las limitaciones de los sistemas tradicionales de vigilancia epidemiológica, los cuales dependen en gran medida de reportes clínicos emitidos por instituciones de salud. Esta dependencia genera rezagos significativos en la obtención de datos actualizados y limita la capacidad de las autoridades sanitarias para implementar acciones preventivas o de contención en tiempo real. Aun cuando estos sistemas cumplen una función crítica, su naturaleza reactiva impide una respuesta anticipada ante brotes emergentes [1], [2].

En este contexto, las plataformas digitales, en particular las redes sociales, han demostrado ser fuentes potencialmente complementarias para la obtención de información pública en tiempo real. X se distingue por su gran volumen de publicaciones espontáneas, geolocalizadas y semiestructuradas, que permiten detectar patrones de comportamiento, reportes auto-referidos de síntomas y menciones explícitas de enfermedades. La riqueza y velocidad de difusión de esta información representa una oportunidad estratégica para complementar los sistemas tradicionales de monitoreo [3].

No obstante, el aprovechamiento efectivo de esta información exige abordar desafíos técnicos significativos, como la extracción automática de datos relevantes, la desambiguación semántica, la reducción del ruido textual y la localización espacial confiable de los eventos mencionados. Actualmente, no se cuenta con mecanismos estandarizados que integren, de manera sistemática, el análisis lingüístico de contenidos generados en redes sociales con modelos de predicción epidemiológica basados en aprendizaje automático. Esto limita el potencial de dichas plataformas como herramientas de apoyo en la toma de decisiones en salud pública.

En consecuencia, se hace necesaria la formulación de un enfoque metodológico computacional que permita el monitoreo continuo y la predicción dinámica de zonas de contagio, a partir del procesamiento de información textual generada por los usuarios de X. Por tanto, el problema central que se aborda en esta investigación es la ausencia de un modelo computacional integral que permita, mediante técnicas de procesamiento de lenguaje natural (PLN), aprendizaje profundo y análisis geoespacial, identificar afirmaciones relevantes sobre contagios de COVID-19, clasificarlas y asociarlas con ubicaciones específicas, a fin de estimar y actualizar de forma dinámica las zonas de mayor riesgo de transmisión.

1.3 Motivación

La gestión y monitoreo eficaz de enfermedades infecciosas tradicionalmente dependen de la información suministrada por entidades de salud pública. Este sistema de monitoreo se basa en el registro de casos confirmados de contagio, reportados por instituciones sanitarias. No obstante, diversos factores inciden en que esta información a menudo resulte insuficiente o desactualizada. En particular, la falta de disposición de los individuos a reportar síntomas por miedo a la crítica social, o simplemente por negligencia o la decisión de auto-medicarse, contribuye a un panorama incompleto de la situación epidemiológica actual [4], [5].

Por otro lado, frente a estas limitaciones, las redes sociales han emergido como una fuente alternativa y prometedora de datos en tiempo real. En especial, aquellas plataformas que incorporan información geoespacial (conocidas como redes sociales basadas en la ubicación) permiten observar dinámicas de comportamiento y movilidad que pueden reflejar, indirectamente, el estado de salud de los usuarios. La continua generación de contenido por parte de los propios individuos habilita una actualización constante del estado de las comunidades, característica clave para monitorear brotes infecciosos y anticipar su evolución.

Recientemente, se ha destacado el valor de las redes sociales como fuentes de datos geoespaciales de alta resolución y en tiempo real, capaces de captar dimensiones humanas relevantes para comprender la introducción y propagación de enfermedades emergentes. Sin embargo, este potencial ha sido explorado solo parcialmente en estudios epidemiológicos contemporáneos [6], [7].

Ante esta realidad, surge la motivación para desarrollar un nuevo paradigma que complemente los enfoques tradicionales. La propuesta de esta tesis consiste en diseñar una metodología capaz de extraer, identificar y localizar afirmaciones personales de contagio en publicaciones emitidas en la red social X, utilizando técnicas avanzadas de procesamiento de lenguaje natural, modelos de aprendizaje profundo y análisis geoespacial. Se espera que esta aproximación ofrezca una herramienta ágil, precisa y en tiempo real que fortalezca las capacidades de monitoreo epidemiológico y apoye la toma de decisiones sanitarias en contextos de emergencia.

1.4 Objetivos

1.4.1 Objetivo general

Desarrollar un modelo computacional para el monitoreo y predicción dinámica de zonas de contagio de una enfermedad epidemiológica, con el objetivo de mejorar su

detección temprana y optimizar la respuesta sanitaria, mediante el análisis de datos generados en la red social X, utilizando técnicas de procesamiento de lenguaje natural, aprendizaje profundo y análisis geoespacial.

1.4.2 Objetivos específicos

- Diseñar un mecanismo computacional para la extracción y preprocesamiento de datos generados en la red social X, con el fin de obtener información textual relevante sobre posibles contagios, mediante técnicas de procesamiento de lenguaje natural.
- Clasificar afirmaciones publicadas en la red social X que estén asociadas a posibles contagios, utilizando modelos de lenguaje preentrenados y redes neuronales profundas para lograr una detección semánticamente precisa.
- Desarrollar un modelo predictivo para estimar la evolución temporal y espacial de zonas de contagio, utilizando patrones derivados del análisis histórico de publicaciones en la red social X clasificadas como indicativas de contagio utilizando aprendizaje profundo.
- Diseñar un algoritmo heurístico de localización dinámica que permita establecer y actualizar las zonas de contagio identificadas basado en la información procesada y las predicciones realizadas.

1.5 Hipótesis

La información generada en la red social X puede ser utilizada como fuente alternativa para la detección temprana, el monitoreo activo y la predicción dinámica de zonas de contagio de alguna enfermedad epidemiológica, mediante la aplicación de técnicas avanzadas de procesamiento de lenguaje natural, aprendizaje profundo y modelado matemático.

1.6 Organización del documento

El resto de este documento está organizado de la siguiente manera:

El **Capítulo 2** presenta el marco teórico que fundamenta esta investigación, abarcando los conceptos clave relacionados con el análisis de redes sociales, el procesamiento de lenguaje natural y la predicción de enfermedades infecciosas.

El **Capítulo 3** expone los trabajos relacionados, con énfasis en estudios previos sobre clasificación de textos, modelos predictivos y seguimiento geoespacial a partir de

redes sociales. Se destacan las principales contribuciones, metodologías utilizadas y las limitaciones identificadas en dichas investigaciones.

El **Capítulo 4** describe la metodología propuesta para el monitoreo y la predicción de zonas de contagio, detallando el diseño del clasificador binario, la incorporación de datos de localización y la aplicación de modelos de predicción.

El **Capítulo 5** presenta la evaluación experimental, describiendo la implementación del enfoque desarrollado, los resultados obtenidos y un análisis de la efectividad del modelo en la identificación y predicción de zonas de riesgo.

Finalmente, el **Capítulo 6** expone las conclusiones generales del estudio, así como posibles líneas de trabajo futuro orientadas al perfeccionamiento y ampliación del modelo propuesto.

2

Marco teórico

Este capítulo presenta una revisión de los fundamentos conceptuales y metodológicos que sustentan la presente investigación. Se abordan tres áreas clave: el análisis de redes sociales como fuente alternativa de información en tiempo real, las técnicas de procesamiento de lenguaje natural (PLN) orientadas a la extracción y clasificación de contenido textual, y los enfoques computacionales para la predicción de enfermedades infecciosas. Estos campos de estudio proporcionan el marco teórico necesario para el diseño e implementación del modelo propuesto de monitoreo y predicción de zonas de contagio, basado en datos generados en la red social X. La integración de estas disciplinas permite construir un enfoque interdisciplinario que combina minería de datos sociales, inteligencia artificial y epidemiología computacional.

2.1 BERT

BERT (*Bidirectional Encoder Representations from Transformers*) es un modelo de representación del lenguaje desarrollado por Google AI. BERT se distingue por preentrenar representaciones bidireccionales profundas a partir de texto no etiquetado, condicionando conjuntamente en el contexto tanto a la izquierda como a la derecha en todas las capas. Este enfoque permite que el modelo preentrenado BERT se ajuste con una sola capa de salida adicional para crear modelos de vanguardia para

una amplia gama de tareas de procesamiento del lenguaje natural, sin necesidad de modificaciones sustanciales en la arquitectura específica de la tarea [8].

En cuanto a su fundamento técnico, BERT aprovecha la arquitectura de Transformers, que emplea mecanismos de auto-atención para procesar secuencias de palabras de manera paralela. A diferencia de los modelos recurrentes, BERT puede captar relaciones y dependencias contextuales profundas en todo el texto gracias a que cada token "atiende" de forma directa a los demás [9]. Esto posibilita generar representaciones contextuales ricas y robustas a distintos rangos de dependencia dentro de la oración.

Un aspecto crucial de BERT para manejar diversos tipos de tareas es su flexibilidad en la representación de entrada, capaz de codificar tanto una sola oración como un par de oraciones en un único bloque de *tokens*. Cabe señalar que el término oración en este modelo engloba cualquier sección contigua de texto, no sólo unidades gramaticales completas. Para ello, BERT implementa WordPiece embeddings con un vocabulario de 30,000 *tokens*, además de tres tipos de embeddings:

- Token embeddings: codifican el significado básico de cada palabra o subpalabra;
- Segment embeddings: señalan si un token corresponde a la primera o segunda oración (en escenarios donde existen dos oraciones);
- Position embeddings: añaden información de la posición secuencial de cada token.

Asimismo, en la secuencia de entrada se ubica el token especial [CLS] al inicio, cuya representación final se utiliza para tareas de clasificación, y se emplea el token [SEP] para delimitar oraciones o fragmentos de texto separados. Para distinguir claramente dos oraciones, BERT no sólo coloca el token [SEP] entre ellas, sino que asigna embeddings de segmento diferentes que identifican si se trata de la oración A o la oración B. En la Figura 2.1 (Fuente: [8]), se ilustra cómo estos tres componentes (token, segmento y posición) se combinan para conformar la representación final de cada token, lo que permite a BERT capturar relaciones contextuales completas y optimizar el rendimiento en múltiples tareas de PLN [8].

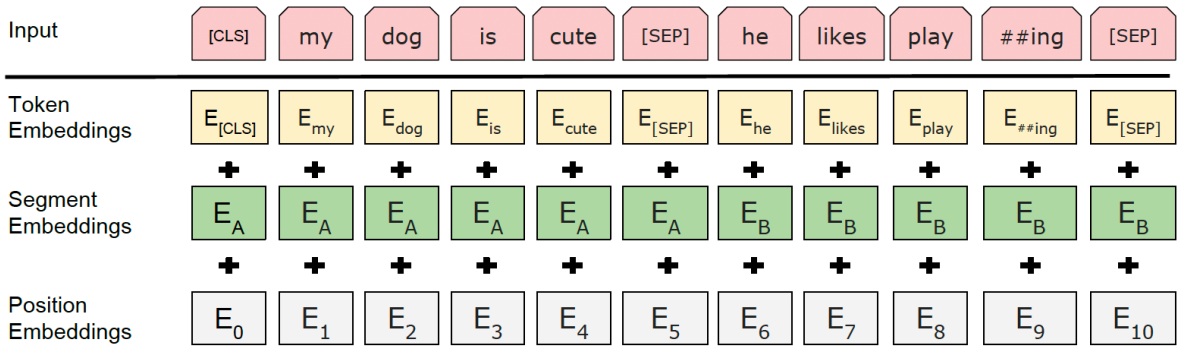

Figura 2.1 Representación de entrada en BERT, compuesta por *embeddings* de *tokens*, segmentos y posiciones.

2.2 Redes neuronales convolucionales

Las redes neuronales convolucionales (CNN) son un tipo de redes neuronales profundas que se especializan en procesar datos con una topología de cuadrícula, como imágenes digitales [10]. Estas redes están compuestas por múltiples capas que transforman la entrada (imágenes, sonido, texto) a través de operaciones como la convolución, la agrupación (*pooling*) y la aplicación de funciones de activación no lineales [11].
Las CNN están diseñadas para emular la manera en que el cerebro humano procesa la información visual, usando una serie de capas que simulan una respuesta neuronal a subregiones de la entrada, conocidas como campos receptivos. Estas capas incluyen:

- **Capas de Convolución:** Utilizan filtros para realizar convoluciones sobre la entrada, extrayendo características esenciales como bordes, texturas o formas. Cada filtro produce un mapa de características que representa una característica específica en diferentes ubicaciones de la entrada [12].

- **Capas de Pooling:** Simplifican la información de los mapas de características reduciendo su dimensión a través de operaciones como máximo (*max pooling*) o promedio (average pooling). Esto ayuda a hacer el modelo más eficiente y robusto a las variaciones de la entrada [13].

- **Capas Totalmente Conectadas:** Después de varias capas de convolución y *pooling*, las redes suelen terminar con una o más capas totalmente conectadas que integran las características extraídas para realizar tareas como clasificación o regresión [10], [13].

Las redes neuronales convolucionales utilizan funciones de activación para introducir no linealidades en el modelo. Estas funciones transforman la salida lineal en una forma más útil y no lineal. Esta transformación es crucial porque la mayoría de los problemas reales a modelar con estas redes son no lineales.

En las CNN, cada capa convolucional está seguida por una función de activación. Las funciones de activación ayudan a determinar la salida final de la red, como clasificaciones binarias (sí/no) o múltiples categorías [10], [11].

Las funciones no lineales más utilizadas son:

1. **ReLU (Unidad Lineal Rectificada)**: La más utilizada en CNNs, definida como $f(x) = \max(0, x)$

   ReLU ayuda a la red a converger rápidamente durante el entrenamiento.
2. **Sigmoid**: Transforma los valores a un rango entre 0 y 1, modelando la probabilidad y usada típicamente en la capa de salida para clasificación binaria.
3. **Tanh (Tangente Hiperbólica)**: Similar a la función sigmoid pero transforma los valores en un rango de -1 a 1, centrando los datos y a menudo llevando a un mejor rendimiento en ciertas aplicaciones.

Las funciones de activación ayudan a controlar el flujo del gradiente durante el entrenamiento, lo que es esencial para actualizar los pesos de la red mediante retropropagación.

2.3 Redes neuronales recurrentes

Las redes neuronales recurrentes (RNN) son un tipo de redes neuronales artificiales avanzadas, diseñadas para modelar datos secuenciales o series temporales. Estas redes son capaces de utilizar su estado interno, también conocido como memoria, para procesar secuencias de entradas, lo que las hace especialmente útiles en tareas como el reconocimiento de voz, la predicción de secuencias y la comprensión del lenguaje natural [14], [15].

Las RNN se diferencian de las redes neuronales *feedforward* por su capacidad de mantener información temporal a través de conexiones cíclicas entre sus unidades.

En lugar de tener una estructura de flujo de datos estrictamente unidireccional, las RNN incluyen conexiones recurrentes que permiten que la información se preserve y se pase a lo largo del tiempo [16]. Este diseño les permite manejar datos secuenciales de manera más eficiente al recordar y utilizar información pasada en la generación de salidas actuales.

Un nodo en una RNN recibe una entrada en el tiempo t, denotada como x_t, y genera una salida mientras actualiza su estado oculto. El estado oculto h_t en el tiempo t se calcula utilizando la entrada actual y el estado oculto del tiempo anterior, como se muestra en la siguiente ecuación [16]:

$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + B_h)$$

donde:

- x_t: vector de entrada en el tiempo t.
- h_t: vector de estado oculto en el tiempo t.
- h_{t-1}: vector de estado oculto en el tiempo anterior.
- W_{hx}: matriz de pesos que conecta la entrada con la capa oculta.
- W_{hh}: matriz de pesos que conecta el estado oculto previo con el actual (recurrente).
- B_h: vector de sesgo asociado a la capa oculta.
- $f(\ )$: función de activación (por ejemplo, tanh o ReLU).

Concluyentemente, la salida y_t se genera secuencialmente en cada nodo con:

$$y_t = g(W_y h_t + B_y)$$

donde:

- y_t: vector de salida en el tiempo t.
- W_y: matriz de pesos que conecta el estado oculto con la salida.
- B_y: vector de sesgo asociado a la salida.
- $g(\ )$: función de activación de la salida [14] (por ejemplo, softmax o sigmoid).

Las redes neuronales clásicas asumen que la entrada y la salida son directamente independientes entre sí, lo que no siempre es cierto en datos secuenciales [14]. Las RNN abordan esta limitación al permitir que la salida de cada nodo sea utilizada como entrada para nodos sucesivos, permitiendo así que la información pasada influencie las decisiones futuras. La Figura 2.2 (Fuente: [14]) muestra la arquitectura básica de una red recurrente.

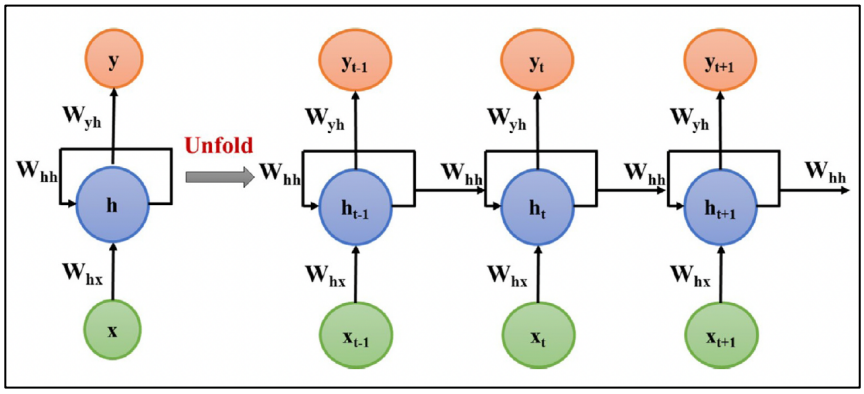

Figura 2.2 Arquitectura básica de una red neuronal recurrente.

2.4 Modelo Gompertz

El modelo de Gompertz es un modelo matemático utilizado para describir el crecimiento de poblaciones biológicas que crecen rápidamente y se desaceleran con el tiempo. Este modelo fue presentado por primera vez por Benjamín Gompertz en 1825, este modelo fue formulado para analizar tasas de mortalidad humana, pero rápidamente se adaptó a otras disciplinas, convirtiéndose en una herramienta esencial en áreas como la biología, la epidemiología y la dinámica de poblaciones [17], [18].

El modelo de Gompertz se caracteriza por su flexibilidad para ajustarse a datos de crecimiento que exhiben una rápida aceleración inicial seguida de una desaceleración progresiva hacia un valor asintótico. Su ecuación general incluye parámetros que representan la tasa de crecimiento relativa, el valor inicial y la asintota superior, proporcionando una interpretación biológica directa. Los investigadores han ajustado el modelo de Gompertz al crecimiento de plantas, crecimiento de aves, crecimiento de peces y crecimiento de otros animales, hasta el crecimiento de tumores y el crecimiento bacteriano [17].

La ecuación que describe al modelo Gompertz se presenta a continuación:

$$y = ae^{-be^{-ct}}$$

donde:

- $y(t)$ es el valor de la variable dependiente en el tiempo t.
- a es el valor asintótico máximo de $y(t)$.
- b determina la posición inicial de la curva en el tiempo t. Es una medida del desplazamiento de la curva a lo largo del eje del tiempo.
- c controla la tasa de crecimiento. A mayores valores de c, la tasa de crecimiento inicial es mayor, pero la curva se estabiliza más rápidamente.

2.5 Redes sociales basadas en la ubicación

Las redes sociales basadas en la ubicación (LBSN, por sus siglas en inglés) representan una evolución de las redes sociales tradicionales, integrando la dimensión espacial en las interacciones sociales. A través de servicios como Foursquare, Instagram, Yelp, X, entre otros. Los usuarios pueden compartir su ubicación y experiencias, lo que permite un análisis profundo de los patrones de comportamiento y las relaciones sociales influenciadas por la geografía [19].
Las redes sociales basadas en la ubicación usan sistemas de información geográfica (GIS) para abordar el comportamiento socioespacial de los usuarios, particularmente mediante datos de *check-in* (X) o reseñas vinculadas a lugares geolocalizados (Foursquare, Tripadvisor o Yelp) [20].

La disponibilidad de información georreferenciada ofrece una variedad de posibilidades en el análisis del comportamiento y patrones socioespaciales en el entorno urbano, como el mapeo de variables segmentadas e identificación de puntos de interés (POI) o áreas de interés (AOI).

La detección de POI y AOI, como la identificación de lugares que generan interés, afluencia de visitantes o puntos calientes (*hotspots*), ha sido uno de los objetivos de las ciencias de la computación.

El mapeo de agrupaciones de POI y AOI es una aplicación común de la extracción de datos georreferenciados de LBSN y facilita la visualización de grandes cantidades de información compleja. Adicionalmente, la información contextual como intereses, opiniones, distribución temporal de visitas, permite un enfoque alternativo a la distribución geográfica [21].

2.6 Zonas de contagio

Las zonas de contagio son áreas geográficas identificadas que determinan una probabilidad de contagio. Esta probabilidad se define de acuerdo con diversos factores que influyen en el aumento o disminución del contagio de enfermedades infecciosas. El conocimiento de las zonas de contagio es fundamental para que la población este consciente de los riesgos a los que está expuesta y, en consecuencia, tome las acciones necesarias para su seguridad [22].

El mapeo y visualización de zonas de contagio utiliza sistemas de información geográfica (SIG) y herramientas de análisis espacial.

La identificación temprana y la intervención en zonas de contagio son esenciales para contener la propagación de enfermedades infecciosas. Las estrategias de intervención incluyen:

- Implementación de cuarentenas y restricciones de movilidad.

- Campañas de vacunación y distribución de medicamentos.
- Mejora de las condiciones sanitarias y de infraestructura.
- Educación y concienciación de la comunidad sobre medidas preventivas.

En el marco de esta tesis, una zona de contagio se define como un área geográfica en la cual se concentra un número significativo de publicaciones en la red social X que contienen afirmaciones personales de infección por COVID-19, identificadas mediante el clasificador ConBiBER. Estas zonas se determinan a partir de los datos georreferenciados extraídos de la red social y permiten estimar, de manera relativa, el nivel de riesgo epidemiológico en cada región.

La relevancia de estudiar las zonas de contagio radica en la posibilidad de reducir la propagación de enfermedades y proteger la salud pública [23].

2.7 Índice de riesgo

Un índice de riesgo (*risk index*) es una herramienta que resume y cuantifica el riesgo de un evento o situación mediante valores numéricos o categóricos, con el propósito de facilitar su comprensión, comparación y comunicación. Este índice se construye a partir de una medida de riesgo, que es un resumen numérico de una distribución de probabilidad asociada al riesgo, y luego se transforma en un formato más comprensible, como una escala numérica, palabras, letras o colores.

El índice de riesgo evalúa la probabilidad y el impacto potencial de eventos adversos en un área específica, combinando factores como exposición, vulnerabilidad y capacidad de respuesta. El índice de riesgo ayuda a priorizar áreas que requieren mayor atención para la reducción del riesgo de desastres, proporcionando una puntuación comparativa que ilustra el nivel de riesgo. La construcción de un índice de riesgo puede involucrar diversas fuentes de datos, lo que lo convierte en una herramienta valiosa para la planificación y toma de decisiones efectivas [24].

Este tipo de índices suele formularse de manera similar a los índices compuestos como el Índice de Desarrollo Humano (IDH). En este enfoque, se seleccionan varios factores F_i relevantes para el fenómeno de estudio, se normalizan a un rango común [0,1] y posteriormente se integran en una sola medida sintética. Una forma general de cálculo es:

$$IR = \frac{1}{n}\sum_{i=1}^{n} N(F_i)$$

donde $N(F_i)$ es el valor normalizado del factor i y n es el número total de factores.

Cuando se desea dar más importancia a ciertos factores, se pueden aplicar ponderaciones w_i, cumpliendo que $\sum_{i=1}^{n} w_i = 1$:

$$IR = \sum_{i=1}^{n} w_i N(F_i)$$

Por ejemplo, si se consideran tres fatores principales "prevalencia de casos *(F_1)*, movilidad poblacional *(F_2)* y densidad demográfica *(F_3)*" el índice de riesgo se expresaría como:

$$IR = \frac{N(F_1) + N(F_2) + N(F_3)}{3}$$

De este modo, el índice integra información heterogénea en una única medida que permite comparar el nivel de riesgo entre regiones o periodos de tiempo, facilitando su interpretación y uso práctico en contextos epidemiológicos.

2.8 COVID-19

La COVID-19 es una enfermedad infecciosa causada por el coronavirus SARS-CoV-2, identificado por primera vez en diciembre de 2019 en Wuhan, China. Este virus pertenece a la familia de los coronavirus, conocidos por afectar tanto a animales como a humanos, y puede provocar enfermedades que van desde resfriados comunes hasta afecciones más graves como el síndrome respiratorio agudo severo (SARS) y el síndrome respiratorio de Oriente Medio (MERS).

La transmisión de la COVID-19 ocurre principalmente a través de gotículas respiratorias expulsadas cuando una persona infectada tose, estornuda, habla o respira. Estas partículas pueden ser inhaladas por personas cercanas o depositarse en superficies, donde el virus puede sobrevivir por períodos variables, facilitando la transmisión al tocarse la cara después de contactar con dichas superficies [25].

Los síntomas más comunes incluyen fiebre, tos seca y dificultad para respirar. Algunos pacientes también pueden experimentar fatiga, dolores musculares, pérdida del gusto o el olfato, congestión nasal, dolor de garganta o diarrea. La mayoría de las personas desarrollan síntomas leves a moderados y se recuperan sin necesidad de tratamiento hospitalario. Sin embargo, ciertos individuos, especialmente los mayores de 60 años o con condiciones médicas subyacentes como hipertensión, diabetes, enfermedades cardíacas o pulmonares, tienen un mayor riesgo de desarrollar formas graves de la enfermedad, que pueden llevar a complicaciones como neumonía, síndrome de dificultad respiratoria aguda, insuficiencia multiorgánica e incluso la muerte [26].

La prevención de la COVID-19 se centra en medidas como la vacunación, el uso de mascarillas, el distanciamiento físico, la higiene de manos y la ventilación adecuada

de espacios cerrados. Las vacunas desarrolladas han demostrado ser efectivas para reducir la incidencia de enfermedades graves y la mortalidad asociada al virus. No obstante, la aparición de variantes del SARS-CoV-2 ha planteado desafíos adicionales, subrayando la importancia de mantener las medidas de prevención y continuar con la vigilancia epidemiológica [25].

2.9 Red social X

X es una plataforma de microblogging lanzada en octubre de 2006, se ha consolidado como un espacio clave para la interacción social y el análisis de datos a gran escala. La plataforma ha permitido organizar discursos globales y facilitar la comunicación transnacional. Además, su apertura en términos de acceso a datos mediante interfaces de programación de aplicaciones (API), la posiciona como una herramienta invaluable para investigadores interesados en fenómenos sociales, políticos y culturales [27].

A pesar de su utilidad, el uso de X en la investigación presenta desafíos metodológicos significativos. Entre ellos se encuentran el sesgo de representación, ya que la población activa en la plataforma no es demográficamente representativa del mundo en *offline*; el desafío idiomático, debido a la diversidad lingüística de los usuarios, y el sesgo en los datos, derivado de las limitaciones técnicas en la recolección a través de las API públicas [28].

La investigación sociológica ha aprovechado X como una herramienta para entender dinámicas sociales, desde movimientos políticos hasta análisis de emociones y comportamientos. Ejemplos destacados incluyen el uso de análisis computacional para explorar redes de interacciones y el impacto de eventos globales en la percepción pública. A pesar de los cambios recientes, como la reestructuración de la plataforma bajo el nombre de X (anteriormente conocida como Twitter), X es un recurso crucial para explorar preguntas sociológicas clásicas y contemporáneas [27].

Esta combinación de ventajas y desafíos metodológicos destaca la importancia de X como una ventana hacia las interacciones humanas en el contexto digital, ofreciendo a los investigadores oportunidades únicas para capturar y analizar fenómenos sociales en tiempo real.

2.10 Ventanas móviles

Las ventanas móviles (*sliding windows*) constituyen un método ampliamente utilizado en el análisis de series temporales, al dividir los datos en segmentos de tamaño fijo que se desplazan secuencialmente a lo largo de la serie [29]. Esta reformulación

permite abordar la predicción de manera supervisada: los valores anteriores (contenidos en cada ventana) se emplean como insumos para estimar valores futuros. Dicho enfoque se aplica exitosamente en disciplinas como la economía, la climatología y el procesamiento de señales, donde la conservación de la estructura temporal de los datos resulta esencial [29], [30].

En la literatura, "el uso de observaciones previas para predecir la siguiente" se conoce como método de ventana deslizante [31]. Si se elige una ventana de tamaño 1, el modelo usará el dato de tiempo t–1 para predecir el valor en t; con una ventana de tamaño 2, usará los datos de t–2 y t–1 para predecir t, y así sucesivamente [29]. De este modo, la técnica reestructura los datos temporales en pares entrada-salida (características → valor futuro), habilitando el uso de algoritmos de predicción estándar sobre series cronológicas [29].

En la Figura 2.3, se ilustra este proceso, mostrando cómo la estructura original de los datos se transforma en instancias de entrenamiento sucesivas [29].

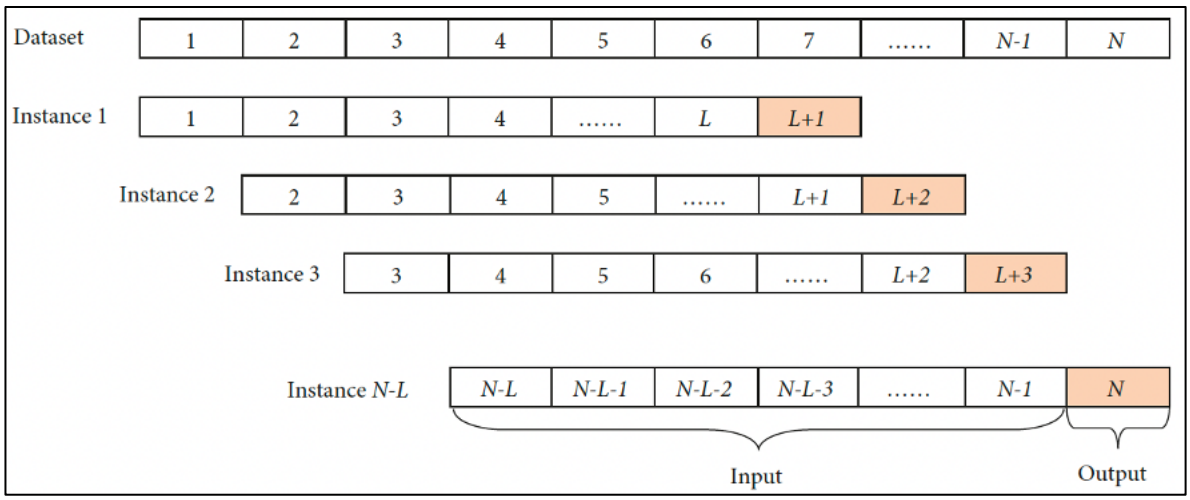

Figura 2.3 Proceso de ventana móvil de datos para crear series de datos temporales.

Al aplicar ventanas móviles en la predicción, existen dos enfoques principales para deslizar la ventana a lo largo del tiempo [32]:

- Ventana Móvil Fija (*Rolling Window*).

  Mantiene un tamaño de ventana constante a lo largo de todo el proceso de pronóstico. El modelo se entrena siempre con la misma cantidad de datos más recientes y, a medida que se avanza en el tiempo, la ventana "roda" descartando el dato más antiguo y añadiendo el dato nuevo. Por ejemplo, con una ventana fija de 12 pasos, primero se usan los datos 1–12 para predecir el día 13, luego 2–13 para predecir el 14, y así sucesivamente. Se recomienda

cuando la serie permanece relativamente estable en el tiempo, ya que el horizonte de entrenamiento permanece fijo [32].

- Ventana Móvil Incremental (*Expanding Window*).

  El tamaño de la ventana aumenta de forma progresiva a medida que avanza el pronóstico. El modelo se reentrena cada vez con un conjunto de datos mayor: por ejemplo, primero con 12 datos, luego con 24 (incorporando los 12 nuevos), y así sucesivamente hasta incluir todo el historial. Este método es útil cuando la serie presenta tendencias o cambios estructurales a lo largo del tiempo, permitiendo al modelo aprender de un histórico cada vez más amplio.

Ambas estrategias suelen usarse para realizar pronósticos paso a paso (*step-by-step forecasting*), reentrenando el modelo en cada paso para actualizar la información con los datos más recientes. A diferencia del enfoque tradicional "estático" (entrenar una sola vez y proyectar múltiples pasos), las ventanas móviles proporcionan un enfoque más realista y preciso del rendimiento predictivo, ya que cada predicción solo toma en cuenta la información disponible en ese momento futuro. Sin embargo, esto conlleva un mayor costo computacional por la necesidad de reentrenar repetidamente [32].

No obstante, la efectividad de la técnica depende en gran medida del tamaño de la ventana: si es muy pequeño, puede no capturar tendencias relevantes; si es demasiado grande, existe el riesgo de incluir ruido y sobreajustar el modelo [32].

Las ventanas móviles representan una estrategia esencial para abordar la predicción en series temporales. Al convertir el problema en un formato supervisado, se facilita la aplicación de algoritmos de aprendizaje y se preserva la naturaleza secuencial de los datos. Tanto la ventana móvil fija como la incremental ofrecen ventajas particulares según la estabilidad o el cambio estructural de la serie. Por su versatilidad y eficacia, las ventanas móviles mantienen su relevancia en múltiples ámbitos científicos y prácticos.

3

Trabajos relacionados

Este capítulo presenta una revisión detallada de los trabajos relacionados con los principales ejes temáticos que sustentan esta investigación. Para ello, se analizaron estudios representativos en las áreas de monitoreo epidemiológico, minería de datos en redes sociales, técnicas de aprendizaje profundo aplicadas a la clasificación de textos, modelos de predicción de brotes infecciosos y metodologías para la delimitación geoespacial de zonas de contagio y la evaluación de riesgo sanitario.

Cabe destacar que estos campos no son independientes ni excluyentes entre sí; por el contrario, diversas investigaciones integran componentes de múltiples áreas con el propósito de desarrollar enfoques interdisciplinarios más robustos para el análisis y pronóstico epidemiológico. El criterio seguido para la selección y organización de los trabajos revisados se basó en su relevancia, actualidad y aportación metodológica al desarrollo de soluciones computacionales aplicadas a la gestión de enfermedades infecciosas.

3.1 Monitoreo epidemiológico

El monitoreo epidemiológico desempeña un papel fundamental en la detección temprana y el control de enfermedades infecciosas. Diversos estudios han abordado este tema desde enfoques tradicionales e innovadores, utilizando tanto fuentes oficiales como trazas digitales para mejorar la vigilancia sanitaria.

Un estudio realizado en España [33], exploró el uso de X como fuente alternativa de información para monitorear la pandemia de COVID-19. Mediante el análisis de la distribución espacial y temporal de usuarios y términos clave durante 2019 y 2020, se generaron mapas de calor que facilitaron la visualización de focos de actividad vinculados con la enfermedad, permitiendo la identificación de patrones espacio-temporales significativos en la propagación del virus.

De manera similar, en el Reino Unido e Irlanda [34], se utilizó X para mapear la epidemiología de la enfermedad de Lyme. A través de técnicas de análisis geográfico y temporal, se identificaron patrones de incidencia mediante la detección de publicaciones que contenían el término "Lyme". Los resultados mostraron una fuerte correlación entre los datos obtenidos de X y los informes oficiales, lo que evidencia la efectividad de las redes sociales como herramienta de vigilancia en tiempo real.

Por su parte, la Red Nacional de Vigilancia Epidemiológica (RENAVE) en España [35], mantuvo una coordinación estrecha con organismos locales e internacionales durante la pandemia de COVID-19. A través del sistema SiViEs, se implementó un protocolo de notificación de casos confirmados que permitió la recolección diaria de indicadores clave como hospitalizaciones, ingresos a cuidados intensivos, recuperaciones y defunciones.

En otro enfoque más amplio, en la investigación [36], se diseñó y evaluó un sistema de alerta temprana para monitorear en tiempo casi real la actividad del COVID-19 en Estados Unidos, utilizando múltiples trazas digitales. El estudio abarcó el periodo de marzo a septiembre de 2020, e integró seis fuentes de datos: tendencias de búsqueda en Google, publicaciones en X, consultas clínicas en UpToDate, estimaciones del modelo GLEAM, movilidad humana derivada de datos de smartphones (Apple y Cuebiq), y mediciones de termómetros inteligentes Kinsa. Estas señales se compararon con tres “estándares de oro” epidemiológicos: casos confirmados, muertes y enfermedades tipo influenza (ILI). Mediante un modelo bayesiano, se identificaron eventos de crecimiento exponencial (*uptrends*) y de decrecimiento (*downtrends*) en las series temporales, evaluando su anticipación con respecto a los datos oficiales.

Los resultados mostraron que varias señales digitales (particularmente las búsquedas en Google y la actividad en X) anticipaban aumentos de casos confirmados y muertes entre 2 y 3 semanas antes. Asimismo, se observó que reducciones en la

movilidad precedían descensos en la incidencia de fiebre y casos de COVID-19, con un desfase de 3 a 4 semanas. Se propuso una métrica combinada basada en el valor P armónico de todos los indicadores, que alcanzó una sensibilidad de hasta 0.78 para predecir eventos de crecimiento en las métricas epidemiológicas. Este enfoque integrado demostró ser más robusto que el uso de señales individuales.

Además, el estudio introdujo un marco probabilístico para estimar el tiempo hasta un evento epidémico (crecimiento o descenso), lo que permitió predecir con éxito el 75% de los eventos una semana antes de su ocurrencia. Aunque cada fuente de datos presenta limitaciones la combinación de múltiples trazas digitales ofrece un enfoque prometedor para la vigilancia temprana de brotes y la planificación de intervenciones no farmacológicas.

Complementariamente, se han desarrollado revisiones sistemáticas sobre sistemas de vigilancia sanitaria que analizan el papel de la epidemiología digital. En [37], se destaca el uso de datos masivos provenientes de motores de búsqueda, redes sociales, registros electrónicos de salud, dispositivos portátiles y aplicaciones móviles para transformar los sistemas tradicionales de vigilancia. Se discuten ejemplos como Google Flu Trends, el seguimiento geográfico mediante X y el uso de datos de dispositivos como Fitbit para predecir patrones gripales. Además, se subraya el uso de PLN y algoritmos de machine learning para modelar brotes, junto con SIG para representar su propagación espacial.

En la revisión [38], el estudio se enfoca en identificar las metodologías más utilizadas para analizar datos de redes sociales con fines de vigilancia epidemiológica, abordando sus aplicaciones, fuentes de datos, algoritmos de clasificación de texto y desafíos asociados.

A partir del análisis se identificó que X es la fuente de datos más utilizada y que las enfermedades más monitoreadas son la influenza o enfermedades tipo influenza. El algoritmo más empleado fue la Máquina de Vectores de Soporte (SVM), seguido por Naive Bayes y sus variantes.

Entre las aplicaciones clave de estos sistemas se encuentran la predicción temprana de brotes, la vigilancia sindrómica, el análisis del impacto social de eventos sanitarios, la identificación de desinformación, y el estudio de la percepción pública durante crisis epidemiológicas. Sin embargo, también se destacan desafíos significativos como el ruido en los datos, la validación de la información, el sesgo demográfico, y problemas de privacidad y semántica en los textos analizados.

Este trabajo ofrece una visión amplia del estado del arte y propone futuras líneas de investigación, como la combinación de datos sociales con variables contextuales (clima, ubicación), el análisis multimodal (texto, imágenes, video), y el desarrollo de modelos predictivos integrados entre plataformas.

Finalmente, en [39], los autores presentan una revisión exhaustiva sobre el uso de datos de redes sociales e Internet para la vigilancia en salud pública. Se identifican tres enfoques principales: el uso de datos agregados por búsqueda, publicaciones sociales y plataformas de vigilancia participativa. Se discuten ejemplos como Google Flu Trends, Yelp, Flu Near You e Influenzanet. Los autores concluyen que los sistemas híbridos, que combinan fuentes digitales con datos clínicos tradicionales, pueden mejorar significativamente la capacidad de respuesta frente a emergencias sanitarias, aunque deben considerarse cuidadosamente los riesgos asociados a la privacidad, la deriva conceptual (*concept drift*), es decir, la variación en la relación entre las variables de entrada y las salidas esperadas a lo largo del tiempo, lo que puede afectar la validez de los modelos predictivos, así como la falta de validación clínica[1].

No obstante, es importante reconocer que, a partir de los cambios ocurridos entre 2022 y 2023, con la adquisición de Twitter, su transformación en X y las nuevas restricciones de acceso a la API, la dinámica de publicación y el volumen de datos disponibles han cambiado de forma sustancial, reduciendo parcialmente la viabilidad de replicar los estudios que dependían del flujo abierto de datos personales y georreferenciados.

En este contexto, la presente tesis se posiciona como una extensión y adaptación de esos enfoques, al proponer un sistema metodológico que integra la clasificación semántica de afirmaciones personales, la modelación temporal con la función de Gompertz y la identificación geoespacial dinámica de zonas de riesgo. A diferencia de los estudios revisados, que se enfocan principalmente en correlaciones o vigilancia descriptiva, esta investigación plantea un marco predictivo y operativo orientado al monitoreo diario de enfermedades infecciosas en el contexto mexicano, contribuyendo al desarrollo de una epidemiología digital más contextualizada y reproducible.

3.2 Minería de datos en redes sociales

La minería de datos en redes sociales ha emergido como una herramienta poderosa para el análisis de fenómenos sociales y de salud pública, al permitir la extracción de patrones relevantes a partir de grandes volúmenes de datos generados por los usuarios. Este enfoque ha sido aplicado para comprender el comportamiento,

[1] Un ejemplo de deriva conceptual puede observarse en modelos entrenados con datos de contagios de COVID-19 en 2020, cuando los patrones de movilidad y comportamiento social eran distintos; si se aplican esos mismos modelos en 2022, sus predicciones pueden perder precisión debido a los cambios en dichos patrones.

las emociones y las preocupaciones de las personas en tiempo real, lo cual resulta especialmente útil en el monitoreo de temas sensibles y de alto impacto.

Entre los enfoques que emplean la minería de datos en redes sociales se encuentra la investigación [40] en la cual los autores utilizaron X para detectar señales de ideación suicida mediante análisis de sentimientos. Utilizando un vocabulario específico relacionado con el suicidio y un algoritmo de aprendizaje automático basado en WordNet, el estudio identificó con éxito patrones emocionales y contextuales en las publicaciones, revelando la capacidad del análisis semántico de sentimientos para identificar y monitorear temas críticos de salud mental en redes sociales. Los resultados obtenidos mostraron que, para las publicaciones con riesgo de suicidio, el algoritmo SMO alcanzó la mayor precisión (89.5%) y F-score (89.3%), seguido de Naïve Bayes con un F-score de 82.9%. En contraste, para las publicaciones sin riesgo de suicidio, el algoritmo J48 obtuvo el mejor rendimiento con una precisión del 75.4% y un F-score de 63.8%. Estos hallazgos destacan la efectividad de los clasificadores basados en aprendizaje automático en contextos de salud pública digital.

En el contexto de enfermedades infecciosas, el estudio [3], analizó el uso de X como herramienta para mejorar la vigilancia del virus del Zika en Estados Unidos durante la epidemia de 2016. Mediante el análisis de publicaciones geoetiquetadas y modelos autorregresivos, el estudio evaluó la relación entre la cantidad de casos de Zika reportados semanalmente y la frecuencia de menciones de términos relacionados en X. Los resultados mostraron una correlación significativa entre los datos de X y los reportes oficiales de casos, con un coeficiente de determinación (R^2) de 0.74 para el modelo de Florida y 0.70 para el modelo nacional. Además, se observó que las publicaciones sobre Zika tendían a preceder el incremento de casos en aproximadamente una semana, lo que sugiere que X puede ser una fuente útil para la vigilancia epidemiológica en tiempo real.

Complementariamente, en [41], se analizó el uso de X para rastrear la desinformación sobre el virus del Zika durante la epidemia de 2016. El estudio recopiló más de 13 millones de publicaciones y empleó técnicas de *crowdsourcing* y aprendizaje automático para identificar rumores y campañas de clarificación promovidas por organizaciones de salud. Se examinaron seis rumores clave identificados por la Organización Mundial de la Salud (OMS) y el sitio de verificación Snopes, utilizando herramientas avanzadas de minería de datos y clasificación automatizada. Los resultados mostraron que el desempeño del clasificador Random Tree no fue uniforme: los rumores sobre "Zika vinculado a transgénicos" y "los estadounidenses son inmunes al Zika" presentaron las peores precisiones (29.6% y 10.1%, respectivamente), mientras que otros como "síntomas similares a la influenza" alcanzaron una precisión de 74.6%. En términos de recall, los rumores "vacunas causan microcefalia" y "insecticidas causan microcefalia" estuvieron por debajo de 0.50, mientras que el rumor sobre "Zika

vinculado a transgénicos" logró un recall de 0.869. El F-measure más alto fue de 0.688 (para el rumor "café como repelente de mosquitos"), lo que evidencia que, aunque es posible identificar automáticamente rumores en X, el rendimiento depende fuertemente del tipo de rumor analizado.

En la investigación [42], se exploró el impacto de las redes sociales en la alfabetización digital en salud en la población general española. Mediante un estudio transversal con 1,307 participantes, se encontró que el 72.1% utilizó activamente las redes sociales para buscar información sobre salud. Los resultados mostraron una correlación significativa entre el uso activo de estas plataformas y una mayor puntuación en el cuestionario eHEALS de alfabetización digital en salud. En concreto, el análisis inferencial reveló diferencias estadísticamente significativas entre los grupos de usuarios activos y no activos ($p=0.0001$), lo que indica que el uso activo de redes sociales se asocia con un mayor nivel de alfabetización digital en salud.

Durante las primeras etapas de la pandemia de COVID-19, en [43]
se analizaron 7,004,155 publicaciones, de las cuales 571,696 estuvieron relacionadas con la pandemia en los estados de Washington y Louisiana. La actividad en X mostró una correlación positiva con la densidad poblacional, significativa al nivel de 0.01. En el análisis de sentimientos, el promedio diario de las publicaciones generales se mantuvo en el rango positivo, mientras que las relacionadas con COVID-19 se ubicaron en el rango neutro o negativo; además, Washington presentó valores de sentimiento consistentemente más altos (menos negativos) que Louisiana. Finalmente, el análisis de las cuentas gubernamentales evidenció que, aunque predominaban los mensajes informativos, las publicaciones con llamados a la acción (por ejemplo, lavarse las manos) generaron los mayores niveles de interacción.

En el ámbito de la gestión de emergencias, en [44], se examinó el uso de X en la gestión de desastres naturales, proponiendo un enfoque que aprovecha las redes sociales basadas en la ubicación (LBSN) para la asignación de recursos en emergencias. Mediante el análisis de publicaciones con información sobre daños, problemas de transporte y sentimientos, este estudio destaca la relevancia de los datos de redes sociales en tiempo real para la toma de decisiones en situaciones de crisis.

En términos de localización geográfica, el estudio [45], desarrolló un modelo basado en la confianza social para identificar la ubicación de los hogares de los usuarios en redes sociales, mientras que en [46] combinaron análisis de entropía con técnicas de agrupamiento para detectar actividad inusual en áreas urbanas mediante datos geolocalizados, mostrando el potencial de estas metodologías en la detección de anomalías en los patrones de comportamiento.

Adicionalmente, los estudios [47] y [48] exploraron el uso de redes sociales basadas en la ubicación para la recomendación de actividades y puntos de interés, integrando información espacio-temporal y de contexto. En [47] se evidenció que la

incorporación de dichos factores mejora la precisión de los resultados de búsqueda recreativa frente a métodos tradicionales, incrementando la relevancia de las sugerencias. Por su parte, [48] presentó el modelo MTAS, que combina datos de redes sociales basadas en localización y microblogs, alcanzando mejoras de entre 0.6% y 23.8% en precisión respecto a enfoques de referencia, y demostrando que factores como la estructura de comunidad y la influencia temporal son determinantes para optimizar la calidad de las recomendaciones.

En conjunto, estos trabajos destacan el potencial de las redes sociales, en particular X, como fuente complementaria para el monitoreo, predicción y respuesta ante situaciones de salud pública. Su capacidad para reflejar patrones conductuales, emocionales, espaciales y comunicativos en tiempo real establece una base sólida para el desarrollo de modelos integrados de vigilancia epidemiológica, como el que se propone en esta investigación.

No obstante, es importante señalar que, tras los cambios estructurales ocurridos entre 2022 y 2023, incluyendo la adquisición de la plataforma por Elon Musk, su cambio de nombre (De Twitter) a X y la implementación de nuevas políticas de acceso a la API, el potencial de esta red social como fuente de información personal y georreferenciada se ha visto parcialmente reducido. Aun así, los principios metodológicos derivados de estas investigaciones siguen siendo aplicables a otras plataformas con dinámicas comunicativas y de datos similares.

A diferencia de los trabajos previos, que se centraron en la detección de correlaciones, análisis de sentimientos o seguimiento de rumores, la presente tesis amplía el alcance metodológico al integrar tres dimensiones complementarias: la clasificación semántica de afirmaciones personales de contagio mediante un modelo híbrido (BERT + CNN), la predicción temporal de menciones a través de la función Gompertz con ventanas móviles y la identificación geoespacial diaria de zonas de riesgo. Esta combinación aporta una visión integral y dinámica del fenómeno, orientada a fortalecer la vigilancia epidemiológica en tiempo real.

3.3 Modelos de predicción

La predicción de enfermedades mediante herramientas computacionales se ha consolidado como una estrategia clave para anticipar brotes y fortalecer las respuestas del sistema de salud. En los últimos años, los modelos predictivos han incorporado tanto fuentes tradicionales (como registros clínicos y reportes oficiales) como datos no convencionales, incluyendo publicaciones en redes sociales y consultas en motores de búsqueda. Esta integración de datos heterogéneos ha permitido generar sistemas de

alerta temprana más sensibles y adaptativos, capaces de reflejar en tiempo real la evolución de fenómenos epidemiológicos.

En este sentido, la investigación [49] desarrolló un modelo para predecir enfermedades infecciosas utilizando una combinación de datos de redes sociales (X), consultas en motores de búsqueda, variables meteorológicas y registros oficiales de salud en Corea del Sur. Los resultados evidenciaron que los modelos de aprendizaje profundo (DNN y LSTM) superaron a los enfoques tradicionales (OLS y ARIMA): en el caso de la varicela, los DNN mejoraron el rendimiento en un 24.45% y los LSTM en un 18.78% respecto a ARIMA; para la escarlatina, las mejoras fueron de 23.28% y 17.97%, respectivamente. Aunque la malaria presentó menor capacidad predictiva debido al reducido número de casos, los modelos profundos lograron captar mejor la tendencia general que los métodos clásicos. En conjunto, estos hallazgos muestran que la integración de datos digitales heterogéneos con algoritmos de aprendizaje profundo permite anticipar brotes con mayor precisión y complementar eficazmente la vigilancia epidemiológica convencional.

Por su parte, el estudio [50] presentó el sistema DEFENDER, una plataforma basada en análisis de datos para la detección y predicción de epidemias mediante el uso de redes sociales y medios de comunicación. El sistema integra algoritmos para la detección temprana de brotes, evaluación situacional y pronóstico de enfermedades. En la fase de *nowcasting*, el modelo ARIMA enriquecido con datos de Twitter (MAE = 8.20) superó al ARIMA tradicional (MAE = 13.05), lo que representó una mejora relativa del 37% y un desempeño claramente superior a controles ingenuos como la media (MAE = 31.50) o el *random walk* (MAE = 19.06). En el pronóstico de casos futuros (*forecasting*), incorporar patrones de movilidad de usuarios en Twitter permitió reducir el error medio de 1.43 a 1.35, lo que se tradujo en una mejora del 5∼6% en comparación con modelos tradicionales de series temporales. Estos resultados demuestran la capacidad de DEFENDER para complementar eficazmente los sistemas convencionales de vigilancia epidemiológica.

De manera complementaria, 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 [51] y [52] han empleado redes neuronales recurrentes (LSTM y variantes RNN) para proyectar la evolución del COVID-19. Estos modelos ofrecen buena capacidad de predicción; sin embargo, su naturaleza de caja negra dificulta la interpretación de los resultados en contextos epidemiológicos.

Otros trabajos, como [53] y [54], exploran comparativamente modelos como Prophet, ARIMA y redes neuronales, destacando sus ventajas en ciertas métricas, pero sin incorporar señales semánticas directas de afirmaciones personales.

Finalmente, enfoques basados en modelos compartimentales, como los SEIR modificados (por ejemplo, [55] y [56]), permiten incorporar parámetros clínicos y poblacionales, pero no aprovechan el contenido generado en redes sociales como fuente directa de detección temprana.

En conjunto, estos enfoques muestran la diversidad de aproximaciones disponibles para el pronóstico epidemiológico; no obstante, la mayoría carece de una integración explícita de señales semánticas derivadas de afirmaciones personales, lo que constituye el aporte diferencial de la metodología propuesta en este trabajo.

Aunque la plataforma X ha cambiado su dinámica y sus políticas de acceso en años recientes, los enfoques metodológicos revisados, particularmente los basados en aprendizaje profundo y fusión de datos heterogéneos, siguen siendo vigentes y adaptables a otras fuentes digitales de información.

En este contexto, la presente tesis amplía los trabajos anteriores al incorporar explícitamente la información semántica contenida en afirmaciones personales sobre contagio, empleando un modelo híbrido de aprendizaje profundo que se articula con modelos de predicción temporal (función de Gompertz) y un componente geoespacial heurístico para identificar zonas de riesgo. Este diseño metodológico, propone una estructura integral para el monitoreo y predicción de enfermedades infecciosas a partir de redes sociales, fortaleciendo la capacidad de alerta temprana y la toma de decisiones en salud pública.

3.4 Zonas de contagio y evaluación de riesgo

La identificación de zonas de contagio y la evaluación del riesgo epidemiológico en contextos geográficos específicos se han convertido en elementos fundamentales para la toma de decisiones en salud pública. Durante la pandemia de COVID-19, diversos estudios buscaron desarrollar metodologías capaces de estimar el riesgo de propagación en función de variables espaciales, demográficas y sociales. Estos enfoques, comúnmente apoyados en sistemas de información geográfica (SIG), permiten visualizar de forma precisa las áreas más vulnerables, facilitando la planificación de intervenciones focalizadas y la optimización de recursos sanitarios.

En la investigación [23], se propuso un método para evaluar el riesgo de contagio de COVID-19 a escala urbana tomando como unidad mínima de análisis los edificios. El estudio se apoyó en la ecuación general de riesgo $R = H \times E \times V$, que integra tres componentes fundamentales: el peligro (zonas de tránsito de personas infectadas), la exposición (densidad de población) y la vulnerabilidad (edad y concentración de habitantes por edificio). A partir de estas variables se generaron mapas de riesgo

mediante sistemas de información geográfica (SIG), clasificando los resultados en cinco niveles (R1–R5, de muy bajo a muy alto).

Los resultados se expresaron principalmente de forma cartográfica y categórica: más de la mitad del área urbana de Málaga se ubicó en los niveles alto (R4) o muy alto (R5) de riesgo. En particular, se identificaron *hotspots* en barrios densamente poblados como Carretera de Cádiz, Miraflores-Trinidad, El Ejido-Olletas y sectores de El Palo y Ciudad Jardín. En contraste, áreas periféricas como Puerto de la Torre o El Candado presentaron niveles bajos de riesgo (R1–R2). Si bien la densidad poblacional mostró una correlación general con el riesgo de contagio, se observaron excepciones significativas: zonas con densidad intermedia que fueron clasificadas con riesgo alto, atribuibles a la configuración urbana y a la conectividad espacial de los edificios.

En conjunto, este enfoque no entrega métricas estadísticas tradicionales (precisión, error), sino que produce indicadores categóricos de riesgo espacial que permiten visualizar, cuantificar y comparar la distribución territorial del contagio. Ello convierte la propuesta en una herramienta cartográfica valiosa para la planificación de intervenciones focalizadas, la gestión de pandemias y la implementación de medidas preventivas en áreas críticas.

Aunque la mayoría de los estudios de este tipo se basan exclusivamente en datos geoespaciales y demográficos, sin aprovechar la información generada en redes sociales, su integración con fuentes digitales permite enriquecer la evaluación del riesgo al incorporar indicadores conductuales y semánticos en tiempo real. En este sentido, la metodología propuesta en esta tesis amplía la perspectiva tradicional al combinar modelos predictivos con un índice heurístico de riesgo derivado de publicaciones georreferenciadas, ofreciendo una visión dinámica y complementaria de la propagación epidemiológica.

4

Metodología de solución

En este capítulo se describe detalladamente la metodología propuesta para el monitoreo y la predicción de zonas de contagio, a partir del análisis de publicaciones generadas en la red social X. El diseño metodológico se estructura en dos niveles de abstracción que permiten una implementación ordenada, flexible y replicable:

- El primero corresponde a cinco fases fundamentales del proceso: adquisición de datos, preprocesamiento, clasificación de afirmaciones de contagio, predicción de afirmaciones, e identificación de zonas de contagio;
- El segundo nivel contempla un conjunto de tareas generales que se ejecutan dentro de cada fase, las cuales permiten la implementación gradual y estructurada del sistema propuesto.

Aunque la metodología fue validada mediante un estudio de caso concreto, su descripción en este capítulo se presenta de manera general, con el propósito de facilitar su adaptación a diversos contextos de vigilancia epidemiológica basados en redes sociales.

En la Figura 4.1 (Fuente: elaboración propia), se presenta un esquema visual de las fases metodológicas y el flujo de información entre ellas, desde la recolección inicial hasta la generación final de mapas de riesgo.

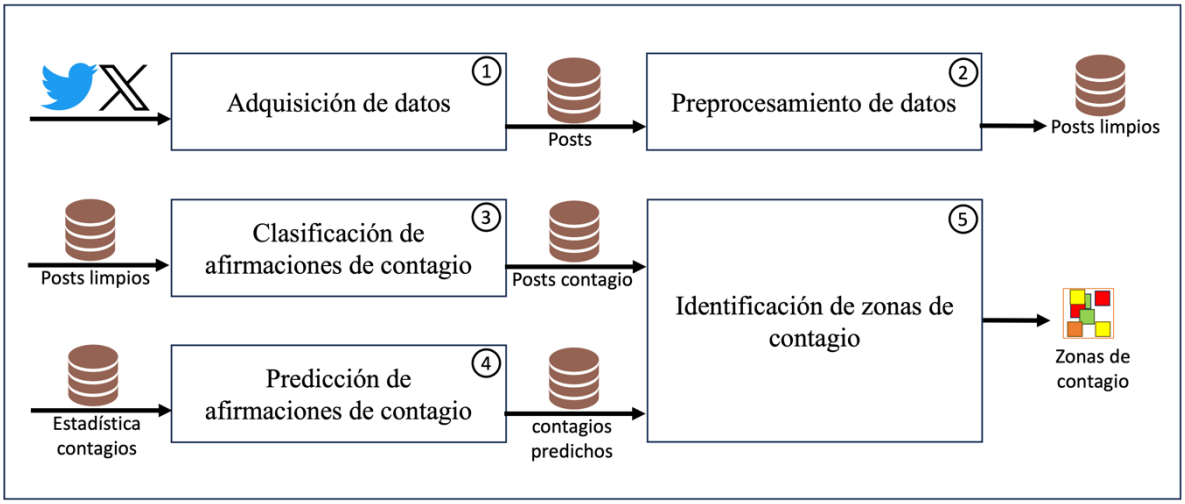

Figura 4.1 Metodología para el monitoreo y predicción de enfermedades infecciosas.

4.1 Adquisición de datos

La fase 1 de la metodología corresponde a la adquisición de datos, cuya finalidad es obtener y almacenar publicaciones generadas en la red social X, con el propósito de recopilar información textual y contextual relevante para su posterior análisis. Estas publicaciones, anteriormente conocidas como *tweets*, representan una valiosa fuente de datos en tiempo real, ya que contienen contenido breve, espontáneo y, en ciertos casos, están acompañadas de metadatos geoespaciales.

X fue seleccionada como fuente primaria debido a diversas características que la hacen adecuada para tareas de monitoreo epidemiológico. A diferencia de otras plataformas como Facebook o Instagram, X ofrece mecanismos públicos y estandarizados de acceso a su contenido mediante una interfaz de programación de aplicaciones (*Application Programming Interface*, API). Esta API permite a los desarrolladores recuperar publicaciones en tiempo real o de forma retrospectiva, incluyendo información estructurada como el texto del mensaje, la fecha y hora de publicación, datos del usuario, y cuando está disponible, información de ubicación.

Esta apertura en el acceso, junto con la estructura breve y semánticamente rica de los textos, posiciona a X como una opción viable y eficaz para la recolección masiva de datos con fines analíticos.

La recolección de datos puede aplicarse en cualquier país, región o conjunto de áreas geográficas. El periodo de recolección es igualmente flexible, pudiendo iniciarse o finalizarse conforme a los criterios establecidos por los investigadores. Esta fase puede adaptarse a contextos específicos, como situaciones de emergencia sanitaria, brotes localizados, o procesos de vigilancia continua, y su configuración dependerá del estudio de caso al que se aplique la metodología.

Las publicaciones recopiladas deben contar con información georreferenciada, por lo que es necesario delimitar previamente la región de estudio mediante el uso de *bounding boxes*. Una vez recolectadas, las publicaciones deben almacenarse en una base de datos relacional, conservando atributos esenciales como el contenido textual, los identificadores únicos, la ubicación aproximada y las marcas temporales correspondientes.

Los datos recolectados en esta fase constituyen la base operativa de las siguientes etapas, donde serán limpiados, transformados y analizados con el fin de identificar afirmaciones personales de contagio, generar predicciones temporales y construir mapas de zonas de riesgo. Las tareas específicas realizadas durante esta fase se describen en las siguientes secciones.

Esta fase, por tanto, no solo habilita el acceso a datos en bruto, sino que también condiciona la calidad general del sistema. La selección de una fuente confiable, la configuración precisa de los filtros y la persistencia estructurada de los registros constituyen el fundamento técnico para todas las fases posteriores del modelo propuesto.

4.1.1 Registro y configuración de acceso a la API de X

La red social X proporciona acceso programático a sus datos mediante una interfaz de programación de aplicaciones (API), la cual permite recuperar información pública en tiempo real o histórica. Esta API constituye una herramienta clave para investigadores, ya que permite acceder a publicaciones, metadatos de usuario, coordenadas de ubicación (cuando están disponibles), y otros elementos estructurados útiles para análisis computacional.

Para utilizar la API, es necesario registrarse como desarrollador en el portal oficial de X [57] y crear una aplicación dentro del proyecto correspondiente. Este proceso habilita la generación de credenciales de autenticación, las cuales incluyen: API Key, API Secret, Access Token y Access Token Secret. Estas credenciales son esenciales para establecer una conexión segura y autorizada entre la aplicación desarrollada y los servidores de X.

En la presente investigación, se gestionó el acceso con privilegios al *streaming* público, modalidad que permite capturar publicaciones en tiempo real conforme se generan, sin necesidad de consultas periódicas. Este tipo de acceso resultó especialmente adecuado para los objetivos del proyecto, ya que permitió una recolección continua y automatizada de publicaciones dentro del territorio mexicano.

Antes del año 2023 X ofrecía beneficios especiales para investigadores académicos, permitiendo el uso gratuito de la API con acceso extendido. No obstante,

a partir de 2023, estas políticas fueron modificadas, instaurando un modelo de acceso por suscripción, lo cual representa una barrera significativa para estudios futuros que requieran un volumen considerable de datos.

Además de los aspectos administrativos, el uso de la API de X impone restricciones técnicas importantes. Entre ellas se encuentran: límites en el número de conexiones simultáneas, cuotas de transmisión de datos por segundo, y eventos de desconexión no previstos. Estas condiciones obligan a implementar estrategias robustas de tolerancia a fallos y reconexión automática, a fin de garantizar la estabilidad del sistema de adquisición.

El procedimiento completo para el registro, configuración y autenticación de la API de X se encuentra documentado en [58] con el fin de facilitar su replicabilidad y adaptación en investigaciones similares.

4.1.2 Desarrollo del proceso de adquisición de datos en Python

Una vez obtenidas las credenciales de autenticación, se procedió al desarrollo de un sistema automatizado para la recolección de publicaciones utilizando el lenguaje de programación Python. Esta elección respondió a su robustez y flexibilidad para el procesamiento de datos en tiempo real, así como a la amplia disponibilidad de bibliotecas especializadas en análisis textual y conexión con redes sociales.

Para interactuar con la API de X, se empleó la biblioteca Tweepy [59], ampliamente reconocida por su estabilidad y facilidad de uso. Tweepy [59] proporciona una interfaz de alto nivel que simplifica la autenticación, el establecimiento de conexión con los *endpoints* de la API, y la captura eficiente de eventos en tiempo real. En particular, se utilizó el *endpoint* de *streaming*, que permite mantener una conexión persistente con los servidores de X, recolectando publicaciones conforme son emitidas y filtradas según criterios predefinidos.

El sistema fue diseñado con un enfoque modular y configurable, lo que permitió definir y ajustar parámetros como el idioma de las publicaciones, las palabras clave de interés y las regiones geográficas de monitoreo, a través de *bounding boxes*. Esta configuración adaptable facilita la personalización del sistema para distintos escenarios de análisis o para su réplica en nuevos estudios.

Cada publicación obtenida fue almacenada en su formato original, es decir, conservando la estructura jerárquica de los objetos JSON retornados por la API. Esta información se guardó dentro de una tabla denominada datosX, ubicada en una base de datos relacional MySQL. Esta decisión permitió preservar todos los atributos originales (como texto del mensaje, fecha de publicación, identificadores del usuario, metadatos de geolocalización y métricas de interacción), al tiempo que se facilitaba su

análisis posterior mediante consultas estructuradas. En la Figura 4.2, se muestra un ejemplo representativo de una publicación almacenada, ilustrando los campos clave capturados durante el proceso.


```
{
  "id": "123456789012345678",
  "created_at": "2025-02-15T12:34:56Z",
  "text": "Este es una publicación de ejemplo",
  "user": {
    "id": "987654321",
    "name": "Mónica Pérez",
    "screen_name": "MonicP123",
    "description": "Apasionada de la tecnología y la programación",
    "location": "Ciudad de México",
    "followers_count": 205
  },
  "place": {
    "id": "07d9cd2eef4f207f",
    "full_name": "Ciudad de México, México",
    "country": "México",
    "place_type": "city",
    "bounding_box": {
      "coordinates": [
        [-99.3647, 19.0112],
        [-99.3647, 19.6399],
        [-98.9403, 19.6399],
        [-98.9403, 19.0112]
      ]
    }
  }
}
```


Figura 4.2 Fragmento representativo de un objeto JSON correspondiente a una publicación recolectada mediante la API de X. Se incluyen campos clave como el texto, fecha, usuario, metadatos geográficos y número de seguidores.

Con el objetivo de garantizar la integridad de las publicaciones recolectadas, se implementó un mecanismo complementario de recuperación manual, destinado a cubrir vacíos ocasionados por fallos en la conexión o errores críticos en el sistema automático. Este mecanismo utilizó un *endpoint* distinto de la API, especializado en la recuperación histórica de publicaciones. Al recibir como entrada un rango de fechas determinado, el sistema completaba las publicaciones faltantes, asegurando la continuidad temporal de la base de datos y minimizando la pérdida de información valiosa.

La combinación de estas estrategias (automatización en tiempo real y recuperación manual asistida) permitió alcanzar una cobertura amplia y precisa de la actividad en X durante el periodo de recolección. Esta robustez en la adquisición fortalece las etapas posteriores de clasificación, predicción y análisis geoespacial de los datos.

4.1.3 Filtro de publicaciones

El proceso automatizado de recolección de publicaciones se implementó mediante la función *filter()* de la biblioteca Tweepy [53], la cual permite establecer

múltiples criterios para capturar publicaciones emitidas en tiempo real desde la red social X. Esta función resulta esencial para limitar el flujo de datos a aquellos que realmente son pertinentes para los objetivos de la investigación, evitando la captura masiva de publicaciones irrelevantes.

La función *filter()* admite tres parámetros principales: *languages*, *locations* y *track*, los cuales permiten definir filtros por idioma, región geográfica y contenido textual, respectivamente.

Parámetro *languages*. Este parámetro permite especificar el idioma de las publicaciones deseadas, utilizando el estándar ISO 639-1. Por ejemplo, `languages=['es']` restringe la captura a publicaciones escritas en español, mientras que `languages=['es', 'en']` permite incluir tanto español como inglés. Esta configuración garantiza que el corpus recolectado se alinee lingüísticamente con el modelo de análisis desarrollado posteriormente.

Parámetro *locations*. Permite delimitar áreas geográficas específicas mediante el uso de *bounding boxes*, que son rectángulos definidos por dos coordenadas geográficas: el vértice suroeste (longitud mínima, latitud mínima) y el vértice noreste (longitud máxima, latitud máxima). Esta configuración restringe la captura únicamente a publicaciones generadas dentro de la zona delimitada.

En la Figura 4.3 (Fuente: elaboración propia), se ilustra un ejemplo de *bounding box* aplicado a la República Mexicana. Los puntos morados indican las coordenadas extremas utilizadas para construir dicho rectángulo geográfico, el cual cubre la totalidad del país. Estas coordenadas son:

-117.5668958243, 14.3949683695, -86.3966790096, 32.9535735494

Figura 4.3 Ejemplo de un *bounding box* aplicado a la República Mexicana.

Cabe señalar que, a diferencia de los mapas tradicionales, la API de X utiliza el formato [longitud, latitud], lo cual debe respetarse estrictamente al definir cada coordenada.

El *bounding box* de la Figura 4.3 cubre la totalidad del territorio nacional, sin embargo, intercepta parcialmente zonas limítrofes de países vecinos como Estados Unidos, Belice, Guatemala, El Salvador y Honduras lo que da como resultado la obtención de publicaciones de esos países.

Para mejorar la especificidad geográfica y reducir el ruido, es recomendable definir múltiples *bounding boxes* más precisos, en lugar de utilizar una única región amplia que podría interceptar zonas no deseadas. Esta estrategia se adapta mejor a estudios donde se requiere un control más fino sobre la cobertura espacial del análisis. Dicha estrategia se detalla en la sección 5.1

Parámetro *track*. Este parámetro se utiliza para filtrar publicaciones en función del contenido textual. Permite definir palabras clave que deben estar presentes en el texto para que una publicación sea recolectada. La configuración admite dos formas de combinación:

- Si se utilizan espacios: `track=['palabra1 palabra2']`, se aplica una condición lógica AND, capturando únicamente publicaciones que contengan ambas palabras.

- Si se utilizan comas: `track=['palabra1, palabra2']`, se aplica una condición lógica OR, capturando publicaciones que contengan al menos una de las palabras.

Esta distinción es crucial para ajustar la especificidad del filtro. El uso de condiciones AND produce un conjunto de datos más preciso pero limitado, mientras que OR permite capturar un volumen mayor, aunque con menor relevancia semántica por publicación. La elección entre ambos enfoques debe basarse en los objetivos del análisis y la etapa metodológica correspondiente.

Es importante tener en cuenta que el parámetro *languages* es compatible con ambos filtros (*track* y *locations*); sin embargo, *track* y *locations* no pueden usarse simultáneamente. Si se configuran al mismo tiempo, la API dará prioridad a *track* y descartará cualquier restricción geográfica definida por *locations*. Esta limitación técnica debe ser considerada cuidadosamente durante el diseño del proceso de adquisición.

En resumen, el uso combinado y estratégico de los parámetros *languages*, *locations* y *track* permite optimizar la calidad de las publicaciones recolectadas, alineándolas con los criterios lingüísticos, geográficos y semánticos definidos por cada investigación. Esta etapa resulta fundamental para garantizar la coherencia y utilidad del conjunto de datos que alimentará las fases posteriores de preprocesamiento, clasificación y análisis predictivo.

4.2 Preprocesamiento de datos

La fase 2 de la metodología corresponde al preprocesamiento de los datos, en la cual se organizan, depuran y transforman los datos originales obtenidos desde la red social X. El propósito principal de esta fase es garantizar que la información recolectada posea un nivel adecuado de calidad, coherencia estructural y relevancia semántica, a fin de alimentar con eficacia los modelos de clasificación y predicción desarrollados en etapas posteriores.

El preprocesamiento se lleva a cabo en distintos niveles, abarcando tanto la estructura de almacenamiento como el contenido textual. Para ello, se definieron cuatro tareas interrelacionadas: (1) la estructuración y diseño del esquema de datos, (2) la limpieza y filtrado de registros, (3) el preprocesamiento y normalización del texto, y (4) la selección de atributos y su representación. Estas operaciones aseguran que el conjunto de datos resultante sea representativo, consistente y adecuado para los modelos de aprendizaje automático empleados en esta investigación.

4.2.1 Normalización y diseño del esquema de datos

Una vez completado el proceso de adquisición, los datos fueron organizados en una base de datos relacional con el propósito de optimizar su almacenamiento, acceso y procesamiento. Esta estructuración fue necesaria para manejar de forma eficiente grandes volúmenes de información textual y metadatos asociados, así como para facilitar la aplicación de operaciones complejas como filtrado, agregación y análisis contextual.

El diseño del esquema se orientó a eliminar redundancias, mejorar la integridad referencial y permitir una representación estructurada del conjunto de publicaciones, usuarios y ubicaciones. Para lograrlo, se normalizaron los datos en tres tablas principales:

- Usuarios: contiene atributos relacionados con la cuenta emisora de la publicación, tales como el identificador único del usuario, el nombre o alias, y el número de seguidores. Esta tabla permite vincular cada publicación con su emisor y evaluar dimensiones contextuales como el alcance potencial del mensaje.
- Ubicaciones: almacena información geoespacial vinculada a las publicaciones. Incluye el identificador del lugar, su nombre completo, el país, el tipo de lugar (ciudad, región, etc.), y las coordenadas del *bounding box* asociado. Esta tabla resulta fundamental para los análisis espaciales de zonas de contagio.
- Tweets: registra el contenido textual, el identificador único de la publicación, la fecha y hora de publicación, así como claves foráneas que vinculan cada publicación con su autor y ubicación geográfica. También conserva metadatos adicionales como el tipo de publicación (original, retweet, respuesta) y las métricas de interacción (favoritos, retweets).

Esta estructura relacional fue implementada en una instancia de base de datos MySQL, elegida por su rendimiento, compatibilidad con bibliotecas de análisis en Python y facilidad de integración con los módulos de procesamiento desarrollados posteriormente. En la Figura 4.4 (Fuente: elaboración propia), se ilustra el esquema general de la base de datos relacional propuesta. Esta organización no solo mejora la eficiencia en la consulta y manipulación de los datos, sino también permite una mayor flexibilidad al incorporar nuevas variables o integrar datos provenientes de futuras etapas de análisis.

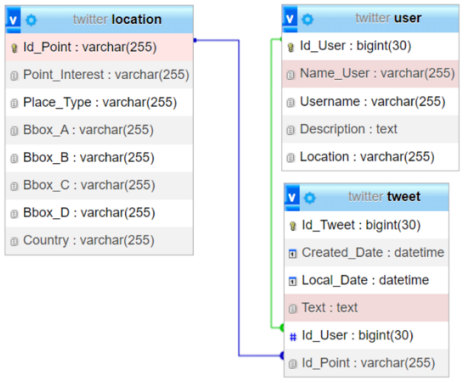

Figura 4.4 Diseño relacional de la base de datos.

4.2.2 Limpieza y filtrado de registros

Una vez estructurados los datos en una base relacional, se procedió a realizar una limpieza inicial del conjunto de registros con el objetivo de reducir el ruido y asegurar que la información utilizada reflejara interacciones auténticas de usuarios individuales. Esta etapa fue fundamental para eliminar publicaciones irrelevantes, redundantes o potencialmente distorsionantes para las fases posteriores de análisis y modelado.

1. Eliminación de duplicados.

A pesar de que la API de X permite configurar filtros que excluyen *retweets*, se detectaron publicaciones idénticas entre diferentes usuarios o emitidas por el mismo usuario en distintos momentos. Las duplicaciones fueron eliminadas mediante comparaciones exactas del contenido textual, la fecha y el identificador del usuario. Este paso fue clave para evitar sesgos estadísticos provocados por la repetición de mensajes con el mismo contenido.

2. Filtrado de cuentas no representativas.

Se excluyeron publicaciones generadas por bots, medios de comunicación y figuras públicas, ya que este tipo de usuarios suelen emitir contenido automatizado, institucional o masivo, que no representa experiencias personales vinculadas a contagios. Este filtrado se realizó a través de criterios heurísticos como:

- Presencia de cuenta verificada.
- Número de seguidores superior a 10,000.
- Más de 200 publicaciones diarias.

- Alta proporción de publicaciones repetidas o idénticas.

Adicionalmente, se construyó y mantuvo una lista negra dinámica de identificadores de usuario (user IDs) que cumplían estos criterios. Esta lista fue alimentada por un script que procesaba métricas diarias y era validada manualmente por el equipo de investigación.

3. Manejo de valores faltantes

Se identificaron registros con campos esenciales incompletos, como la ausencia de texto, fecha u hora de publicación. En los casos donde fue posible, se imputaron valores faltantes mediante reglas lógicas (por ejemplo, usando la hora del servidor como proxy de publicación). En los casos en que esta imputación no era fiable, se optó por la eliminación definitiva del registro.

Gracias a estas acciones, el corpus final quedó reducido a publicaciones con texto original, emitidas por usuarios individuales, dentro del periodo y región de estudio definidos. Esta depuración mejoró considerablemente la calidad del conjunto de datos y su capacidad de representar conversaciones reales y relevantes sobre contagios, lo cual es esencial para la clasificación de afirmaciones y la predicción de tendencias.

4.2.3 Preprocesamiento y normalización de texto

Una vez filtrada la colección de publicaciones para conservar únicamente aquellas relevantes y auténticas, se procedió al preprocesamiento textual, con el fin de eliminar ruido semántico y preparar los datos para su posterior vectorización y clasificación. Esta etapa consistió en una serie de transformaciones aplicadas sobre el contenido textual de cada publicación, orientadas a estandarizar su formato, preservar su valor informativo y reducir la ambigüedad o redundancia léxica.
Las operaciones realizadas se agrupan en las siguientes categorías:

1. Normalización básica del texto.

Se aplicó normalización de caracteres acentuados y manejo de caracteres especiales como la letra “ñ”, que fueron preservados para no alterar la integridad lingüística del español.

2. Eliminación de elementos no informativos.

Se removieron elementos que no aportan valor semántico directo para el análisis, tales como:

- URLs, que frecuentemente redirigen a fuentes externas.
- Menciones a usuarios (por ejemplo, @nombre), las cuales generalmente no tienen relevancia contextual.
- Emojis y símbolos que no tienen interpretación textual clara.
- Signos de puntuación repetidos, caracteres no alfabéticos y números aislados.

3. Sustitución semántica y estandarización de patrones.

Algunos elementos fueron transformados en lugar de eliminados para conservar su carga semántica:

- Hashtags temáticamente relevantes fueron convertidos a su forma textual equivalente.
- Emojis con significado claro fueron reemplazados por etiquetas descriptivas (por ejemplo, 😷 → "[enfermo]").
- Se aplicaron reglas de unificación ortográfica básica para reducir variantes comunes de escritura en redes sociales (por ejemplo, errores deliberados o creativos).

4. Filtrado de palabras clave específicas.

Los textos se filtraron por palabras clave relacionadas con la enfermedad de estudio. Este filtrado semántico inicial fue esencial para centrar el análisis en publicaciones con alta probabilidad de referirse a experiencias personales de enfermedad.

5. Limpieza final y tokenización.

En esta etapa se eliminaron los espacios innecesarios generados en los procesos previos, y los textos fueron segmentados en *tokens* mediante expresiones regulares. Se descartaron *tokens* de longitud muy corta (por ejemplo, menores a 2 caracteres) y números aislados sin relevancia semántica. Estas transformaciones se implementaron en Python utilizando las bibliotecas re [60] y nltk [61], con el objetivo de preservar el significado original de los mensajes sin introducir distorsiones. El resultado fue un corpus limpio, estructurado y adecuado para su representación mediante *embeddings* y posterior procesamiento por modelos de aprendizaje automático, como BERT, en las siguientes fases de la metodología.

4.2.4 Selección de atributos y representación del texto

Tras la limpieza y normalización de los registros almacenados, se realizó la selección de atributos con el objetivo de conservar únicamente aquellas variables

relevantes para las fases de clasificación de afirmaciones y predicción de contagios. Esta etapa permitió organizar el conjunto de datos de forma que cada publicación quedara representada por un conjunto de características clave, agrupadas en tres dimensiones principales: temporales, espaciales y contextuales.

1. Atributos temporales.

Se conservaron la fecha y la hora exacta de publicación de cada publicación, lo cual fue esencial para la construcción de series cronológicas. Esta dimensión permitió ordenar el corpus en el tiempo, aplicar ventanas móviles para el análisis y desarrollar modelos que proyectaran la evolución de las menciones afirmativas de contagio. Asimismo, facilitó la identificación de patrones estacionales, picos de actividad y periodos críticos en la conversación digital sobre la enfermedad de estudio.

2. Atributos espaciales.

Para cada publicación, se conservaron las coordenadas geográficas proporcionadas por el campo place.bounding_box. Esta dimensión espacial fue fundamental para geolocalizar las afirmaciones y posteriormente identificar zonas con alta densidad de menciones, como base para la generación de mapas de riesgo.

3. Atributos contextuales.

Se incluyeron variables asociadas tanto al autor como al comportamiento de la publicación dentro de la red social. Entre estos atributos destacan:

- El número de seguidores del usuario emisor, como indicador del alcance potencial del mensaje.
- El número de *retweets* y favoritos, que reflejan el grado de interacción o resonancia del contenido en la comunidad.

Estos atributos permitieron enriquecer el análisis con información sobre la visibilidad y el impacto de cada publicación, así como evaluar su relevancia en la propagación de información relacionada con contagios.

La combinación de estas tres dimensiones permitió construir una base de datos estructurada, con características clave que fueron utilizadas para alimentar los modelos de clasificación y predicción descritos en los capítulos posteriores. Esta representación también facilitó el análisis exploratorio y la visualización de patrones temporales y geoespaciales del fenómeno estudiado.

4.3 Clasificación de afirmaciones de contagio

Una de las principales tareas dentro de la metodología propuesta fue la identificación automática de publicaciones que contienen afirmaciones personales de contagio relacionadas con la enfermedad de estudio. Esta fase tuvo como finalidad filtrar, dentro del gran volumen de datos recolectados, aquellos mensajes que reflejan experiencias individuales de infección, excluyendo menciones generales, noticias o comentarios impersonales.

Para ello, se desarrolló un modelo de clasificación binaria, diseñado para distinguir entre publicaciones que expresan afirmaciones directas de contagio y aquellas que únicamente hacen referencia general a la enfermedad. Este modelo permitió automatizar una tarea que, en otros estudios, ha dependido de listas de palabras clave o de revisión manual, lo cual limita su escalabilidad y precisión semántica.

El modelo debe ser entrenado con un conjunto de datos previamente etiquetado, el cual, en la mayoría de los casos, presenta un desbalance natural: existen menos textos con afirmaciones personales de contagio que publicaciones generales sobre la enfermedad. Por ello, es necesario establecer criterios lingüísticos claros que guíen el etiquetado, como el uso de verbos en primera persona ("me contagié", "di positivo") y el contexto explícito del mensaje.

La clasificación semántica de las publicaciones permite generar una base de datos precisa de afirmaciones individuales, sobre la cual se construyen series temporales y modelos de predicción. Esta fase es crucial, ya que define el subconjunto informativo sobre el que se calcularán tendencias y se identificarán zonas de riesgo geográfico.
La implementación técnica del modelo, incluyendo su arquitectura, criterios de entrenamiento y evaluación, se describe detalladamente en el Capítulo 5.

4.4 Predicción de afirmaciones de contagio

Una vez identificadas las publicaciones con afirmaciones personales de contagio, se procedió a analizar su comportamiento temporal con el objetivo de anticipar posibles aumentos o disminuciones en la conversación digital vinculada a la enfermedad de estudio.

Para ello, se propuso el uso del modelo matemático de Gompertz, ampliamente reconocido por su capacidad para describir fenómenos biológicos y epidemiológicos. Este modelo es particularmente útil para representar la dinámica de brotes, al capturar tanto la fase inicial de crecimiento acelerado como la eventual estabilización de los casos reportados. Su naturaleza asintótica lo convierte en una herramienta adecuada para proyectar tendencias basadas en la evolución observada.

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

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

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 análisis temporal se estructuró mediante una estrategia de ventanas móviles, que consiste en dividir la serie cronológica en segmentos consecutivos de duración fija (por ejemplo, 7, 15 o 30 días). Estas ventanas se desplazan progresivamente sobre la secuencia diaria de afirmaciones clasificadas, permitiendo generar predicciones ajustadas a los datos más recientes.

Esta metodología ofrece flexibilidad para abordar distintos horizontes de predicción. Ventanas cortas proporcionan una mayor sensibilidad ante cambios abruptos en el comportamiento de las publicaciones, mientras que ventanas más amplias permiten suavizar fluctuaciones y detectar tendencias consolidadas.

En este marco, el modelo de Gompertz fue empleado para estimar la evolución futura de las afirmaciones de contagio, a partir de su comportamiento histórico en cada ventana temporal. Aunque la implementación detallada de este modelo se presenta en el capítulo siguiente, cabe destacar que su aplicación permitió generar estimaciones dinámicas de corto plazo, alineadas con los cambios en la actividad observada en la red social X.

En la Figura 4.5 (Fuente: elaboración propia), se ilustra este proceso, donde la sección en amarillo representa los datos históricos, correspondientes a los valores observados en un periodo determinado. A partir de estos datos, el modelo genera valores predichos (sección en verde), proyectando el comportamiento esperado en los días siguientes.

En la parte superior, se observa una ventana móvil de datos históricos de 7 días que avanza progresivamente en la serie temporal, calculando predicciones de hasta 5 días.

En la parte inferior, se muestra el caso de una ventana que abarca 15 días, proporcionando una base más amplia para la predicción. Este enfoque asegura que la predicción se base en los datos más recientes disponibles.

Figura 4.5. Esquema del uso de ventanas móviles en la predicción de menciones de contagio.

4.5 Identificación de zonas de contagio

La identificación de zonas de contagio constituye una fase crítica dentro del monitoreo epidemiológico digital, ya que permite detectar áreas con alta concentración de afirmaciones de contagio y analizar su evolución en el tiempo. Para ello, se emplea un enfoque geoespacial basado en datos obtenidos de la red social X, utilizando técnicas de georreferenciación y modelado de riesgo.

Este análisis permite integrar datos espaciales en un mapa de calor dinámico, el cual representa la distribución relativa del riesgo de contagio en las distintas regiones del territorio bajo estudio. La actualización periódica de este mapa facilita la identificación de zonas con alta incidencia de afirmaciones, apoyando la toma de decisiones en salud pública y el análisis de patrones en la propagación digital de la enfermedad.

La asignación geográfica de las publicaciones se realiza a partir de dos fuentes principales:

- *Bounding boxes*: Cuando la publicación contiene coordenadas aproximadas, se utiliza el centroide del *bounding box* proporcionado por la plataforma para asignarla a una región. Esta técnica resulta útil en ausencia de georreferenciación precisa.
- Metadatos de la publicación: En los casos en que la publicación incluye explícitamente el nombre de la ubicación, dicha información se aprovecha para refinar la geolocalización, comparándola con una base de polígonos administrativos previamente definida.

Una vez ubicadas las publicaciones afirmativas en sus respectivas regiones, se calcula un índice de riesgo relativo, definido como la proporción entre el número de casos activos y un umbral máximo estimado de menciones:

- Casos activos: Se consideran aquellas publicaciones afirmativas emitidas durante el periodo estimado de transmisibilidad de la enfermedad de estudio,

contado a partir de la fecha de la afirmación. Este periodo puede variar según las características clínicas y epidemiológicas de la enfermedad, pero para efectos del presente enfoque, se asume un rango de días representativo en los que la persona podría contagiar a otros tras haber reportado síntomas o un diagnóstico positivo.

- Máximo estimado: Se define como un umbral de referencia calculado heurísticamente a partir del volumen histórico de publicaciones en cada región. Este valor representa la cantidad máxima esperada de publicaciones afirmativas, y se utiliza como denominador para estimar el nivel relativo de riesgo en función del comportamiento típico observado en la zona.

Con base en este índice, se asignan rangos de color al mapa de calor, permitiendo visualizar intuitivamente las regiones con mayor concentración de afirmaciones. Los umbrales de color se definen en función de la densidad relativa de casos, asegurando una representación clara y diferenciada de los niveles de riesgo.

El índice de riesgo y la distribución espacial se recalculan de forma periódica, generando una visualización dinámica que refleja la evolución de la conversación digital vinculada al contagio. Este enfoque permite detectar zonas críticas de propagación, contribuyendo al fortalecimiento de las estrategias de vigilancia epidemiológica.

5

Estudio de caso: COVID-19

Este capítulo describe la implementación de la metodología propuesta para el monitoreo y predicción de zonas de contagio, aplicada al caso específico del COVID-19 en México. El estudio se basó en el análisis de publicaciones emitidas en X, con el fin de identificar afirmaciones personales de contagio, estimar su comportamiento temporal y localizar posibles zonas de riesgo.

La implementación se estructuró en cinco fases, adquisición de datos, preprocesamiento de datos, clasificación de afrimaciones de contagio, predicción de afirmaciones de contagio e identificación de zonas de contagio. Cada una de estas etapas se desarrolló utilizando herramientas de programación, técnicas de procesamiento de lenguaje natural y modelos predictivos, con el objetivo de transformar publicaciones digitales en indicadores útiles para la vigilancia epidemiológica.

5.1 Adquisición de datos

Durante el periodo comprendido entre el 1 de octubre de 2021 y el 13 de marzo de 2023, se llevó a cabo la adquisición automatizada de datos generados en la red social X. El objetivo central fue obtener los datos correspondientes a las publicaciones georreferenciadas emitidas por los usuarios dentro del territorio mexicano.

Para delimitar geográficamente la adquisición, se definieron múltiples áreas geográficas mediante cajas delimitadoras o *bounding boxes*. En total, se establecieron

24 *bounding boxes*, seleccionadas cuidadosamente con base en el criterio del investigador, buscando cubrir completamente las 32 entidades federativas de México. La configuración de estas cajas delimitadoras se realizó manualmente utilizando la herramienta boundingbox.klokantech.com, asegurando una cobertura geográfica nacional exhaustiva. Cada *bounding box* fue establecida con precisión, procurando minimizar la captura de publicaciones provenientes de países vecinos como Estados Unidos, Guatemala y Belice, al evitar que dichas áreas excedieran significativamente los límites territoriales nacionales, ya que esto podía generar la obtención de publicaciones ajenas al área de estudio y provocar un consumo innecesario de *tokens* disponibles en la API de X, los cuales estaban sujetos a un límite diario de uso.

La distribución precisa de estos *bounding boxes* se presenta gráficamente en la Figura 5.1 (Fuente: elaboración propia), permitiendo visualizar claramente cómo fue segmentado el territorio nacional para fines de adquisición de datos.

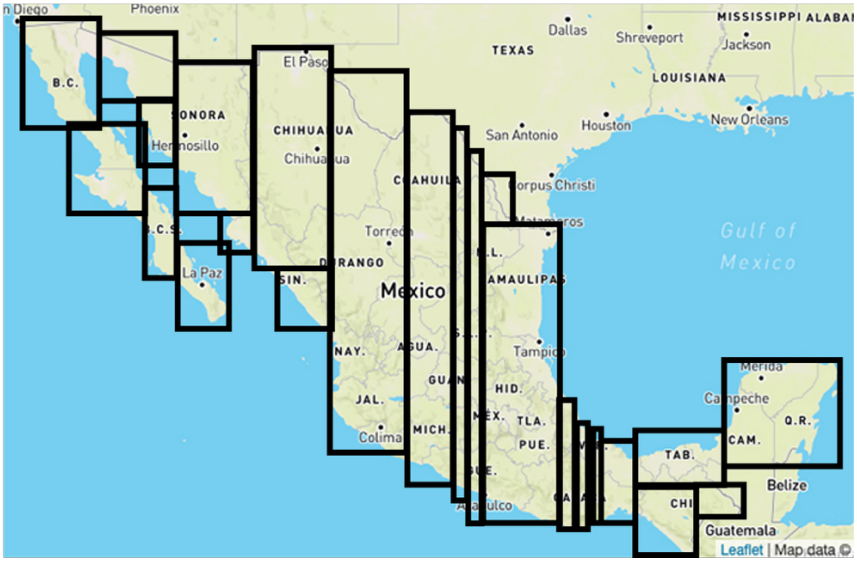

Figura 5.1 Distribución geográfica de las 24 *bounding boxes* utilizadas para la adquisición automatizada de datos georreferenciados en la red social X, correspondientes al territorio mexicano.

La implementación técnica del sistema se realizó utilizando el lenguaje Python y la biblioteca Tweepy [59], que facilitó la interacción eficiente con la API de X. El sistema se configuró para mantener una conexión persistente a través del *endpoint 'streaming'*, permitiendo así capturar los datos de las publicaciones en tiempo real, prácticamente segundos después de su emisión. Durante todo el periodo de recolección, se almacenaron de manera automática e ininterrumpida todos los datos relacionados con las publicaciones georreferenciadas dentro de México, incluyendo sus atributos originales y metadatos esenciales como identificadores únicos, texto completo, usuario emisor, fecha y hora de publicación, ubicación aproximada y métricas de interacción.

A partir de cada objeto JSON obtenido, se extrajeron todos sus atributos relevantes y se almacenaron en formato estructurado dentro de una tabla denominada

datosX, ubicada en una base de datos relacional MySQL. Esta decisión permitió conservar la totalidad de la información original, pero en un esquema tabular que facilita su consulta, análisis y explotación posterior mediante herramientas convencionales de análisis de datos.

No obstante, debido a la posibilidad de errores técnicos o interrupciones críticas en la conexión con la API, se implementaron mecanismos complementarios de control y recuperación de datos. Específicamente, en aquellos casos donde se detectaron desconexiones prolongadas o errores críticos que interrumpieron la adquisición automática, se recurrió a un segundo *endpoint*, llamado "*search30*", proporcionado también por la API de X. Este *endpoint* especializado permitió realizar extracciones manuales y retrospectivas de los datos faltantes, garantizando así la integridad y continuidad temporal del conjunto de datos.

Cabe mencionar que el periodo de adquisición fue seleccionado luego de una etapa previa de pruebas técnicas y operativas del sistema, confirmando así su estabilidad y robustez antes de comenzar la recolección formal. Además, el proceso concluyó en marzo de 2023 debido a cambios en las políticas de acceso a la API de X, las cuales impusieron restricciones importantes en el uso gratuito para fines académicos, estableciendo un modelo de suscripción. Este cambio representó una barrera significativa para la continuidad de investigaciones futuras que dependen de la obtención masiva de datos sociales.

Como resultado global del proceso descrito, se logró recopilar un total de 20,008,585 publicaciones georreferenciadas, almacenadas en una base de datos relacional. La distribución de publicaciones recolectadas por entidad federativa se presenta en la Tabla 5.1. La estructura del almacenamiento preservó meticulosamente los atributos originales de cada publicación, permitiendo realizar análisis posteriores precisos y detallados. La descripción completa y ejemplos de estos atributos se presentan en la Tabla 5.2.

Tabla 5.1 Distribución de publicaciones recolectadas por entidad federativa.

Estado	No. publicaciones	Porcentaje
Aguascalientes	190082	0.95%
Baja California	450193	2.25%
Baja California Sur	180077	0.90%
Campeche	134058	0.67%
Chiapas	310133	1.55%
Chihuahua	298128	1.49%
Coahuila	406174	2.03%
Colima	94040	0.47%
CDMX	4437904	22.18%
Durango	178076	0.89%
Estado de México	1854796	9.27%
Guanajuato	654281	3.27%
Guerrero	314135	1.57%
Hidalgo	328141	1.64%
Jalisco	1538660	7.69%
Michoacán	324139	1.62%
Morelos	342147	1.71%
Nayarit	148064	0.74%
Nuevo León	1560670	7.80%
Oaxaca	360155	1.80%
Puebla	884379	4.42%
Querétaro	574246	2.87%
Quintana Roo	754324	3.77%
San Luis Potosí	328141	1.64%
Sinaloa	382164	1.91%
Sonora	442190	2.21%
Tabasco	306130	1.53%
Tamaulipas	452194	2.26%
Tlaxcala	142061	0.71%
Veracruz	848364	4.24%
Yucatán	526226	2.63%
Zacatecas	264113	1.32%
Total	**20,008,585**	**100.00%**

Tabla 5.2 Descripción de los atributos almacenados en la base de datos datosX, derivados de un total de 20,008,585 publicaciones georreferenciadas recolectadas en México a través de la API de X.

Atributo	Valor de ejemplo
X.id	1346889436626259968
X.text	Learn how to use the user Tweet timeline and user mention timeline endpoints in the X API v2 to explore Tweet…
X.created_at	Wed Jan 06 18:40:40 +0000 2021
X.username	XDevelopers
X.author_id	2244994945
X.user.id	2244994945
X.user.name	X Dev
X.user.username	TwitterDev
X.user.protected	FALSE
X.user.created_at	2013-12-14T04:35:55Z
X.place.id	f7eb2fa2fea288b1
X.place.full_name	Lakewood, CO
X.place.country	United States
X.place.country_code	US
X.place.place_type	city
X.place.bounding_box.bbox	[-105.193475, 39.60973, -105.053164, 39.761974]
X.place.geo.geometry.coordinates	[-105.18816086351444, 40.247749999999996]
X.poll.id	1365059861688410112
X.poll.duration_minutes	5042
X.poll.end_datetime	2023-11-07T05:31:56Z
X.poll.voting_status	open
X.topic.name	Technology
X.topic.description	All about technology
X.matching_rule.id	120897978112909812
X.matching_rule.tag	Non-retweeted coffee Posts
X.truncated	FALSE

5.2 Flujo operativo del sistema

La metodología se implementó en un sistema de vigilancia epidemiológica el cual seguía un flujo operativo diario que garantizara la actualización continua de los

resultados. Este flujo permitió realizar análisis automatizados sobre los datos recolectados en tiempo real, generando información oportuna para la toma de decisiones en salud pública y la identificación temprana de zonas de riesgo.

Al concluir cada día (a las 23:59 horas), el sistema realizaba automáticamente un conjunto ordenado de tareas:

1. Cierre diario y almacenamiento de datos:

Se cerraba el periodo de adquisición del día, almacenando todas las publicaciones recolectadas en las últimas 24 horas en la base de datos relacional previamente estructurada (Sección 5.1).

2. Preprocesamiento automatizado:

Las publicaciones recolectadas eran inmediatamente sometidas al proceso automático de preprocesamiento (detallado en la sección 5.3), incluyendo limpieza textual, filtrado de publicaciones irrelevantes y normalización semántica. Este paso garantizó que el conjunto de datos estuviera limpio y estructurado para su análisis inmediato.

3. Clasificación automática de afirmaciones:

El modelo ConBiBER (BERT + CNN) procesaba automáticamente los datos del día para identificar aquellas publicaciones con afirmaciones personales de contagio. Esta clasificación generaba un conjunto diario de publicaciones afirmativas que era acumulado para análisis temporales.

4. Predicción temporal diaria:

Con base en las afirmaciones clasificadas hasta la fecha actual, el sistema aplicaba diariamente el modelo de Gompertz mediante ventanas móviles de 7, 15 y 30 días, con el fin de generar predicciones a corto, mediano y largo plazo. La selección de estas longitudes se sustentó en escalas de vigilancia epidemiológica comúnmente utilizadas: el ciclo semanal (7 días), ampliamente empleado en la notificación de enfermedades infecciosas [62]; la actualización quincenal (15 días) aplicada en el semáforo epidemiológico de México durante la pandemia; y los reportes mensuales (30 días), que constituyen un intervalo práctico y frecuente en sistemas nacionales de vigilancia [63]. Estas ventanas, además de estar alineadas con prácticas epidemiológicas, ofrecieron un equilibrio metodológico entre la detección de cambios súbitos y la observación de tendencias más estables. Las predicciones eran recalculadas diariamente, incorporando continuamente los datos más recientes disponibles.

5. Actualización diaria del mapa geoespacial:

Finalmente, el sistema actualizaba automáticamente el índice de riesgo relativo para cada región geográfica, recalculándolo diariamente a partir del número de casos activos. Estos resultados se representaban en un mapa de calor dinámico que se actualizaba al término de cada jornada, mostrando visualmente los cambios en las zonas de mayor riesgo.

Este flujo operativo garantizó que el sistema mantuviera un carácter preventivo y dinámico, proporcionando información en tiempo casi real sobre la propagación digital del contagio, lo cual resultó fundamental para apoyar la implementación oportuna de medidas de vigilancia y prevención epidemiológica.

5.3 Preprocesamiento de datos

En esta fase se llevó a cabo el preprocesamiento de los datos con el objetivo de estructurarlos, limpiarlos y prepararlos para su posterior análisis automatizado. Este paso fue fundamental para garantizar la calidad y coherencia del conjunto de datos que se emplearía en las etapas de clasificación y predicción. El proceso se organizó en cinco etapas claramente definidas: selección de atributos, estructuración en base de datos, limpieza de registros, filtrado semántico y normalización textual.

Se seleccionaron y conservaron únicamente aquellos atributos esenciales para las tareas posteriores. Estos atributos se agruparon en tres dimensiones principales:

- Temporales: Fecha y hora de publicación, necesarias para el análisis cronológico y la aplicación de ventanas móviles.
- Espaciales: Coordenadas del *bounding box* asociadas a cada publicación, indispensables para la identificación geográfica de zonas de contagio.
- Contextuales: Información adicional como el número de seguidores del autor o métricas de interacción (*retweets*, favoritos), útiles para evaluar el impacto y la propagación del contenido.

Una vez definidos los atributos relevantes, se procedió a diseñar una base de datos relacional en la cual fueron almacenadas las publicaciones obtenidas de la red social X. El esquema resultante constó de tres tablas principales:

- Usuarios: atributos relacionados con la cuenta emisora (identificador único, nombre, seguidores).
- Ubicaciones: información geográfica vinculada a cada publicación (*bounding box*, nombre del lugar, país).

- Publicaciones: contenido textual, fecha y hora de publicación, tipo de publicación y métricas de interacción.

Esta organización relacional garantizó la integridad referencial de los datos, facilitando consultas rápidas y precisas en etapas analíticas posteriores.

Posteriormente, los datos fueron sometidos a una limpieza exhaustiva con el fin de eliminar ruido informativo y asegurar la autenticidad de las publicaciones recolectadas. Se eliminaron duplicados exactos y se excluyeron cuentas consideradas no representativas (bots, medios de comunicación, figuras públicas). Este filtrado se realizó a través de reglas heurísticas definidas, tales como número excesivo de seguidores (>10,000), actividad excesiva (más de 200 publicaciones diarias) o contenido repetitivo. Asimismo, se gestionaron valores faltantes en campos esenciales como el texto o la marca temporal, descartándose aquellos registros que no podían imputarse de manera confiable.

En esta fase también se aplicó un filtrado semántico, seleccionando exclusivamente las publicaciones que contenían términos relacionados con COVID-19, se utilizaron palabras clave previamente identificadas de forma manual, incluyendo variantes lingüísticas comunes empleadas por los usuarios: "covi", "cobi", "covicho", "cobicho", "covid", "cobid", "coronavirus", "sars-cov2", "coronabirus", entre otras. La lista completa de los términos empleados se presenta en el **Apéndice A**. Este filtrado permitió enfocar el análisis exclusivamente en aquellas publicaciones con alta probabilidad de contener afirmaciones personales de contagio.

Finalmente, el contenido textual de las publicaciones fue sometido a una normalización semántica utilizando scripts en Python con las bibliotecas re y nltk. Este proceso incluyó:

- Preservación de caracteres lingüísticos especiales del español (acentos, “ñ”).
- Eliminación de elementos no informativos como URLs, menciones a usuarios, signos de puntuación repetidos, emojis sin valor semántico y números aislados.
- Transformación semántica de elementos con alta relevancia contextual. Con el propósito de preservar la información semántica implícita en ciertos símbolos, algunos emojis fueron sustituidos por etiquetas textuales descriptivas, de modo que su significado pudiera integrarse de manera uniforme en el análisis. La Tabla 5.3 presenta los emojis considerados y la etiqueta asignada en cada caso.

Tabla 5.3 Emojis y su sustitución semántica.

Emoji	Etiqueta asignada
🤒	[enfermo]
🤧	[fiebre]

🤧	[tos], [gripa]
🦠	[virus]
💉	[vacuna]
🏥	[hospital]
🛌	[reposo], [cama]
⚰️	[muerte]
💀	[muerte]
😢	[tristeza]
🙏	[oración]

De igual forma, los hashtags que incluían alguna de las palabras clave listadas en el **Apéndice A** fueron estandarizados a su forma base, con el fin de asegurar consistencia y comparabilidad durante el procesamiento de los datos.

- Tokenización especializada del texto y eliminación de *tokens* irrelevantes o demasiado cortos (<2 caracteres).

El resultado fue un conjunto de datos limpio, estructurado semánticamente y listo para ser representado computacionalmente en modelos posteriores, específicamente el clasificador ConBiBER y el modelo predictivo Gompertz. Esta estructura facilitó también tareas de análisis exploratorio, visualización interactiva y evaluación estadística en etapas posteriores.

5.4 Clasificación de afirmaciones de contagio

Una vez preprocesadas las publicaciones recolectadas, se procedió a la clasificación automática de aquellas que contenían afirmaciones personales de contagio. Para ello, se desarrolló un modelo supervisado denominado ConBiBER (*Convolucional Binary Classifier with BERT*), diseñado para identificar afirmaciones de contagio en textos coloquiales, como los que se publican en X.

El modelo fue entrenado con un conjunto de 4,500 textos etiquetados manualmente por el equipo de investigación, diferenciados en dos clases: afirmaciones de contagio (clase 1) y comentarios generales sobre la enfermedad (clase 0). Este conjunto no provino directamente del total de 20,008,585 publicaciones recolectadas, sino que fue construido a partir de una muestra representativa obtenida durante la primera etapa del sistema automatizado (octubre de 2021). Dado que en los datos georreferenciados de ese periodo no se identificó una cantidad suficiente de afirmaciones de contagio, se complementó la muestra con publicaciones no georreferenciadas que cumplían los criterios semánticos de la clase 1. De esta manera, se equilibró el conjunto

de entrenamiento con 600 ejemplos positivos (clase 1) y 3,900 negativos (clase 0), garantizando la adecuada representación de ambas categorías para el aprendizaje supervisado del modelo:

- Clase 1 (afirmaciones de contagio): 600 textos en los que los usuarios afirmaban haber contraído COVID-19 o reportaban síntomas compatibles.
- Clase 0 (comentarios generales): 3,900 textos relacionados con la enfermedad, pero sin indicios de contagio personal.

El proceso de etiquetado manual para el conjunto de entrenamiento, se basó en criterios lingüísticos definidos previamente, como el uso de verbos en primera persona (p. ej., "me contagié", "di positivo") y el contexto semántico del mensaje. Se consideraron también expresiones informales como "me pegó el covicho", las cuales fueron incluidas con el fin de entrenar al modelo para reconocer lenguaje coloquial característico de redes sociales. En la Tabla 5.4 se muestran ejemplos representativos de cada clase.

Tabla 5.4. 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

ConBiBER se construyó sobre las representaciones semánticas generadas por BERT multilingüe (*base cased*), que fueron procesadas por una red neuronal convolucional (CNN) con filtros de tamaño 2, 3 y 4, seguidos por una capa global *max pooling* y una capa densa de 512 neuronas con activación ReLU. La salida del modelo fue una neurona con activación sigmoide, representando la probabilidad de afirmación de contagio.

El modelo fue entrenado con la función de pérdida *binary_crossentropy*, el optimizador Adam, un tamaño de lote de 32 y un total de 40 épocas. Para mitigar el desbalance de clases, se emplearon pesos de clase, mejorando así la sensibilidad del modelo frente a la clase minoritaria. Los hiperparámetros y ajustes específicos se resumen en la Tabla 5.5.

Tabla 5.5 Hiperparámetros y ajustes del modelo ConBiBER.

Categoría	Parámetro	Valor utilizado
Framework	TensorFlow	2.19.0
	Keras API	Integrada en TensorFlow
	TensorFlow Hub	bert_multi_cased_L-12_H-768_A-12
Embeddings	Dimensión de embedding	768
	Tokenización	WordPiece (sensible a mayúsculas/minúsculas)
Optimizador	Algoritmo	Adam
	Learning rate	2×10^{-5}
	β_1	0.9
	β_2	0.999
	ε	1e−07
Entrenamiento	Batch size	32
	Número de épocas	10
	Función de pérdida	Binary Crossentropy
	Métricas	Accuracy, Precision, Recall
Regularización	Dropout	0.2
	L1/L2 en capa oculta	0.001 / 0.001
	L2 en capa de salida	0.001
Callbacks	EarlyStopping	Paciencia = 5 (monitoreando *val_recall*)
	Checkpoints	Guardado de mejor modelo según *val_recall*
Pesos de clase	Pesos de clase	{0:1.0, 1:8.0}

El modelo ConBiBER se evaluó mediante validación cruzada estratificada con k=5 sobre el conjunto de entrenamiento (4,500 textos). En cada iteración, 3,600 instancias se emplearon para entrenamiento y 900 para validación, preservando la proporción original de clases. La Tabla 5.6 muestra los resultados promedio y desviación estándar de la validación cruzada, evidenciando una alta exactitud y

especificidad, con variaciones moderadas en precisión y F1-Score atribuibles al desbalance de clases.

Tabla 5.6 Resultados promedio de la validación cruzada (k=5) del clasificador ConBiBER.

Métrica	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Promedio ± σ
Exactitud	92.362	93.281	94.200	95.119	96.038	94.2% ± 1.3
Precisión	70.564	72.332	74.100	75.868	77.636	74.1% ± 2.5
Sensibilidad	88.154	89.427	90.700	91.973	93.246	90.7% ± 1.8
Especificidad	93.020	94.010	95.000	95.990	96.980	95.0% ± 1.4
F1-Score	78.530	80.015	81.500	82.985	84.470	81.5% ± 2.1

Posteriormente, el modelo se reentrenó con la totalidad del conjunto de entrenamiento (4,500 textos) y se evaluó en un conjunto de prueba independiente compuesto por 750 textos (100 afirmaciones positivas y 650 comentarios generales). Este conjunto no provino del flujo operativo de publicaciones recolectadas entre noviembre de 2021 y marzo de 2023, sino que fue construido a partir de datos recopilados durante la etapa inicial del sistema (octubre de 2021). Para su conformación, se incluyeron publicaciones no georreferenciadas que cumplían los criterios semánticos de las clases 1 y 0, además de ejemplos elaborados para reforzar la representación de ambas categorías. En la Figura 5.2, se muestra la matriz de confusión obtenida en esta evaluación final, y las métricas asociadas se resumen en la Tabla 5.7, confirmando la consistencia de los resultados con los obtenidos en la validación cruzada.

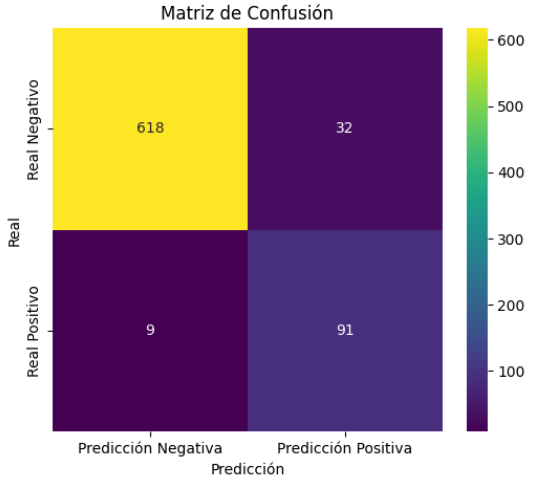

Figura 5.2 Matriz de confusión de la evaluación del clasificador ConBiBER sobre el conjunto de prueba independiente.

Tabla 5.7 Métricas de la evaluación en el conjunto de prueba independiente del clasificador ConBiBER.

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

Cabe destacar que la alta sensibilidad del modelo es especialmente relevante para el monitoreo epidemiológico, ya que minimiza la omisión de publicaciones que puedan representar contagios reales. Aunque se observaron falsos positivos, esto se atribuye al desbalance del conjunto de prueba.

Adicionalmente, se calculó el área bajo la curva ROC (AUC-ROC), cuyo valor fue de 0.925, lo que demuestra una excelente capacidad discriminativa del clasificador ConBiBER para distinguir entre afirmaciones personales de contagio y comentarios generales sobre la pandemia. Este resultado confirma la coherencia con las métricas previamente reportadas de sensibilidad (91%) y especificidad (95%), y respalda la solidez del modelo como herramienta de detección automatizada de contagios en publicaciones de redes sociales.

En la Figura 5.3, se visualiza la curva ROC generada a partir de las probabilidades de predicción del modelo. La proximidad de la curva al vértice superior izquierdo y el valor del área bajo la curva (AUC = 0.925) reflejan un equilibrio óptimo entre sensibilidad y especificidad, lo que confirma la fiabilidad del clasificador en la detección automatizada de casos potenciales.

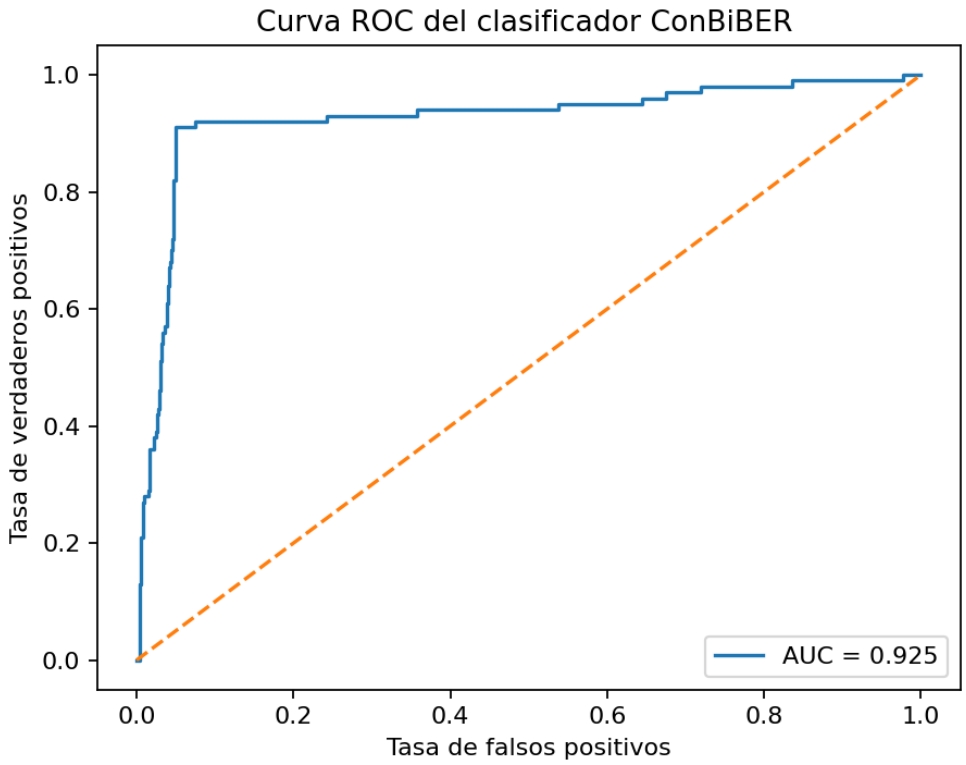

Figura 5.3 Curva ROC del clasificador ConBiBER.

Además, se observó una correlación de Pearson de 0.83 entre las afirmaciones clasificadas y los casos oficiales reportados por la Secretaría de Salud. Esta sincronicidad temporal, mostrada en la Figura 5.4, sugiere que las publicaciones en X podrían servir como complemento temprano en la vigilancia de brotes epidémicos.

Es necesario reconocer que los datos recolectados desde la plataforma X están sujetos a sesgos, principalmente de distribución geográfica, volumen y características demográficas de los usuarios. Aun con estas limitaciones, la fuerte correlación identificada sugiere que la dinámica temporal de las publicaciones refleja de manera consistente la evolución de la epidemia. En particular, frente al sesgo geográfico, resulta relevante considerar que los brotes epidémicos suelen concentrarse en zonas urbanas de alta densidad poblacional, que coinciden con las regiones con mayor concentración de usuarios activos de X. Así, el sistema opera como un "centinela digital" en los principales focos de transmisión, ofreciendo señales de alerta temprana de gran valor para la vigilancia epidemiológica. Este hallazgo es coherente con estudios previos [64], [65], que destacan que los datos de redes sociales, aun con sesgos de cobertura, permiten capturar patrones epidemiológicos relevantes cuando se analizan en series temporales agregadas. En consecuencia, la fortaleza del sistema no radica en la representatividad exhaustiva de la muestra, sino en su capacidad de complementar la información oficial y anticipar oportunamente cambios en la tendencia de los contagios a nivel nacional.

A partir de esta validación, ConBiBER fue integrado al sistema automatizado descrito en la sección 5.2, operando en tiempo real desde el 1 de noviembre de 2021 hasta el 13 de marzo de 2023. Durante dicho periodo, el modelo procesó más de veinte millones de publicaciones y clasificó 2,884 textos como afirmaciones positivas de contagio, los cuales se emplearon para la generación series temporales que sirvieron de base para las predicciones temporales (Capítulo 5.5) y la identificación geográfica de zonas de contagio (Capítulo 5.6).

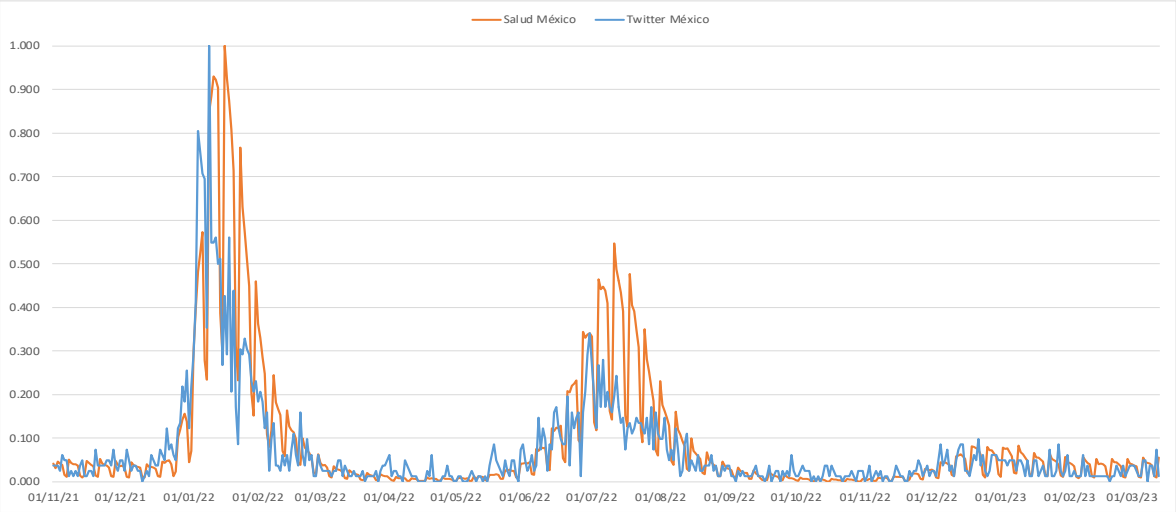

Figura 5.4 Correlación temporal entre las afirmaciones clasificadas como contagio por COVID-19 en X y los casos oficiales reportados por la Secretaría de Salud.

Al analizar la Figura 5.4, 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. Cabe aclarar que la Figura 5.4 muestra la serie temporal nacional de afirmaciones positivas, integrando publicaciones de todas las entidades federativas

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.

Por otro lado, durante la etapa de clasificación se consideró una clase 0, compuesta por publicaciones generales sobre la pandemia que no implicaban contagio personal. Esta categoría permitió al modelo ConBiBER diferenciar menciones contextuales de afirmaciones personales de infección, aumentando la precisión y reduciendo falsos positivos en el monitoreo automatizado.

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.

5.4.1 Resultados del clasificador ConBiBER

El modelo ConBiBER alcanzó un buen desempeño en la identificación automática de afirmaciones de contagio.

Los principales resultados sobre el conjunto de prueba independiente (750 textos) se visualizan en la **Error! Reference source not found.**:

Tabla 5.8 Resumen de resultados del clasificador ConBiBER.

Métrica	Valor
Exactitud	94.5 %
Precisión	73.9 %
Sensibilidad	91.0 %
Especificidad	95.0 %
F1-Score	81.6 %
AUC-ROC	0.925

Además, se observó una correlación de Pearson de **0.83** con los casos oficiales de la Secretaría de Salud, lo que sugiere que las publicaciones en X pueden anticipar incrementos reales de contagios.

5.5 Predicción de afirmaciones de contagio

Una vez clasificadas las publicaciones afirmativas de contagio mediante el modelo ConBiBER, se procedió a su análisis temporal con el fin de anticipar su evolución. Esta predicción se llevó a cabo utilizando la función de crecimiento de Gompertz, reconocida por su capacidad para modelar fenómenos epidemiológicos que exhiben una fase de crecimiento exponencial seguida por una desaceleración.

El modelo predictivo de menciones de contagio se evaluó durante el período del 1 de noviembre de 2021 al 13 de marzo de 2023, considerando la serie temporal

nacional, construida a partir del total de publicaciones clasificadas en las 32 entidades federativas de la República Mexicana.

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.

En la Figura 5.5, se presenta la comparación entre los valores observados y predichos del modelo entrenado con una ventana móvil de 7 días y horizonte de predicción de 1 día.

Figura 5.5 Comparación entre los valores observados y predichos de afirmaciones de contagio utilizando el modelo de Gompertz. Se muestra el desempeño del modelo con una ventana móvil de 7 días y un 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.

En la Figura 5.6, se ilustra el desempeño del modelo con la misma ventana de 7 días, pero horizonte de predicción a 5 días.

Figura 5.6 Comparación entre los valores observados y predichos de afirmaciones de contagio utilizando el modelo de Gompertz. Se muestra el desempeño del modelo con una ventana móvil de 7 días y un 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. En la Figura 5.7, se ilustra la sobreposición de las predicciones a 1 y 5 días, ambas con una ventana móvil de 7 días, para comparar su respectivo desempeño frente a los datos observados.

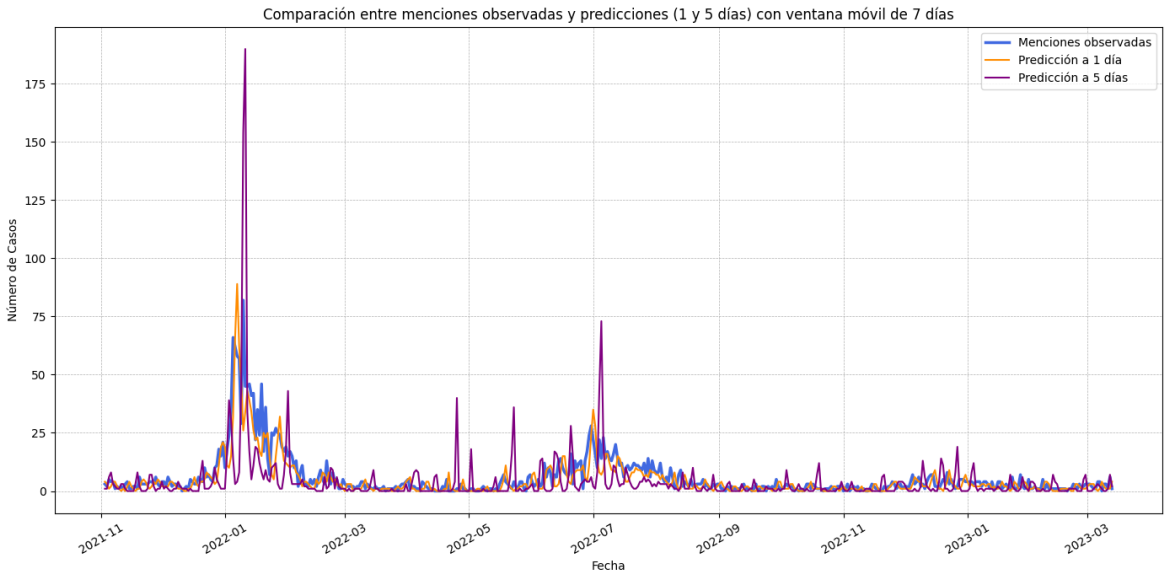

Figura 5.7 Comparación superpuesta de las predicciones a 1 y 5 días, utilizando una ventana móvil de 7 días.

Análogamente, en la Figura 5.8 se presentan los resultados para una ventana móvil de 15 días con predicciones a 1 día

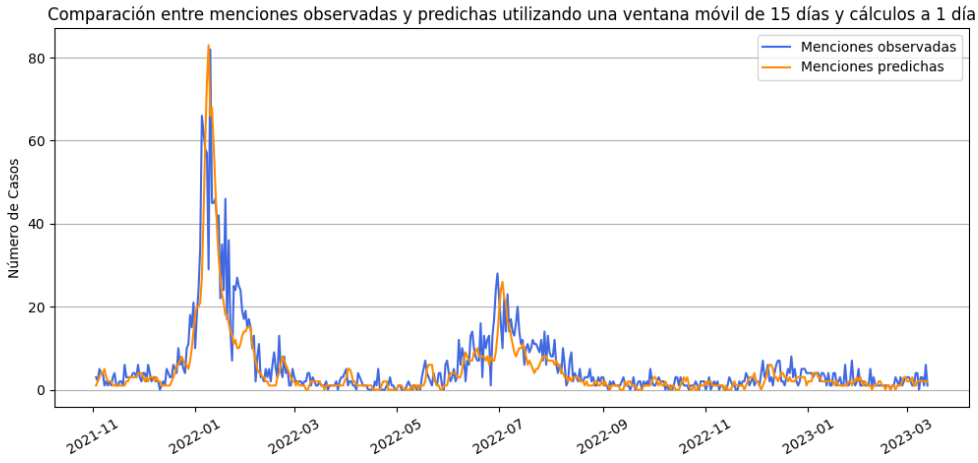

Figura 5.8 Comparación entre los valores observados y predichos de afirmaciones de contagio utilizando el modelo de Gompertz. Se muestra el desempeño del modelo con una ventana móvil de 15 días y un horizonte de predicción de 1 día.

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.

En la Figura 5.9, se presentan los resultados correspondientes a la ventana móvil de 15 días y predicciones a 5 días

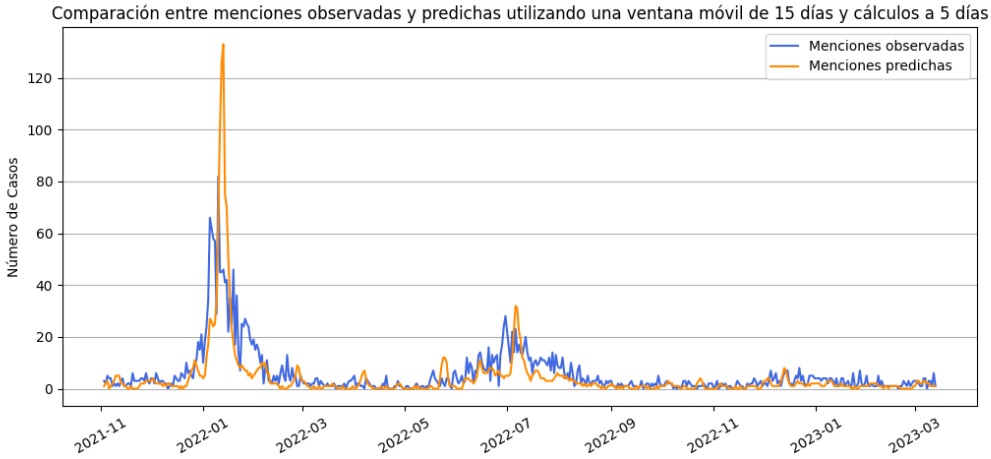

Figura 5.9 Comparación entre los valores observados y predichos de afirmaciones de contagio utilizando el modelo de Gompertz. Se muestra el desempeño del modelo con una ventana móvil de 15 días y un horizonte de predicción de 5 días.

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.

En la Figura 5.10, se ilustra la sobreposición de las predicciones a 1 y 5 días, ambas con una ventana móvil de 15 días, para comparar su respectivo desempeño frente a los datos observados.

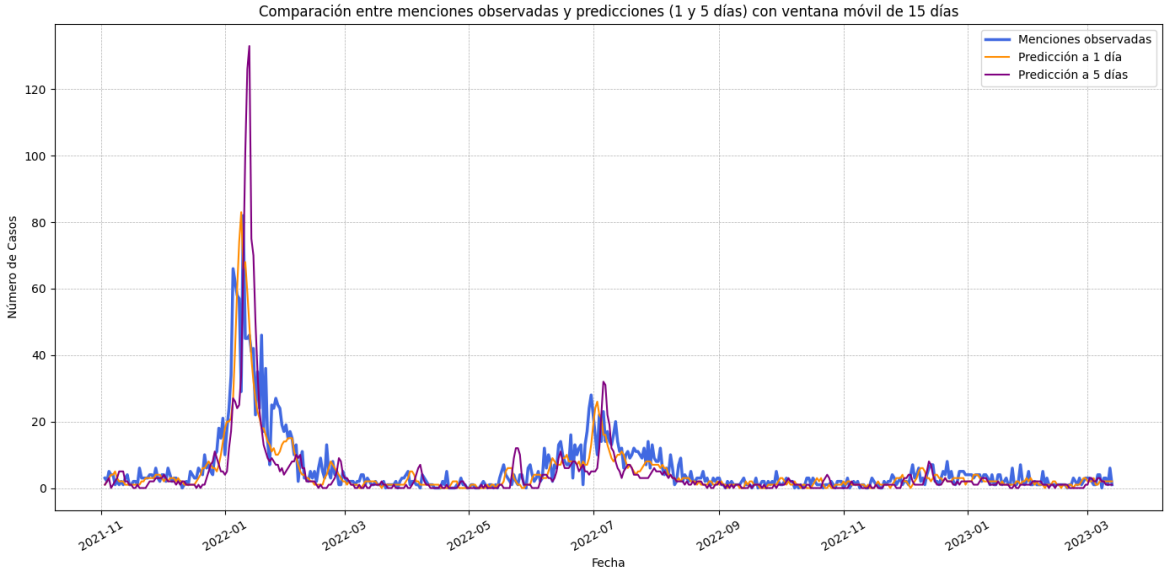

Figura 5.10 Comparación superpuesta de las predicciones a 1 y 5 días, utilizando una ventana móvil de 15 días.

En la Figura 5.11, se presenta la comparación entre los valores observados y predichos del modelo con una ventana móvil de 30 días y horizonte de predicción de 1 día.

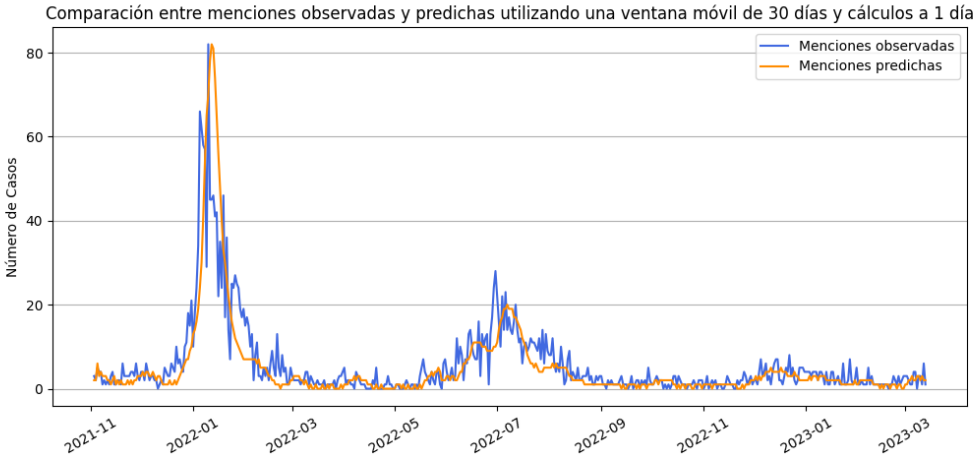

Figura 5.11 Comparación entre los valores observados y predichos de afirmaciones de contagio utilizando el modelo de Gompertz. Se muestra el desempeño del modelo con una ventana móvil de 30 días y un horizonte de predicción de 1 día.

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 = $\pm$ 5.98 y MAE = 2.96.

En la Figura 5.12, se muestran los resultados del modelo entrenado con una ventana móvil de 30 días y horizonte de predicción a 5 días.

Figura 5.12 Comparación entre los valores observados y predichos de afirmaciones de contagio utilizando el modelo de Gompertz. Se muestra el desempeño del modelo con una ventana móvil de 30 días y un 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 = $\pm$ 10.08 y MAE = 3.89.

Finalmente, en la Figura 5.13, se ilustra la sobreposición de las predicciones a 1 y 5 días, ambas con una ventana móvil de 30 días, para comparar su respectivo desempeño frente a los datos observados.

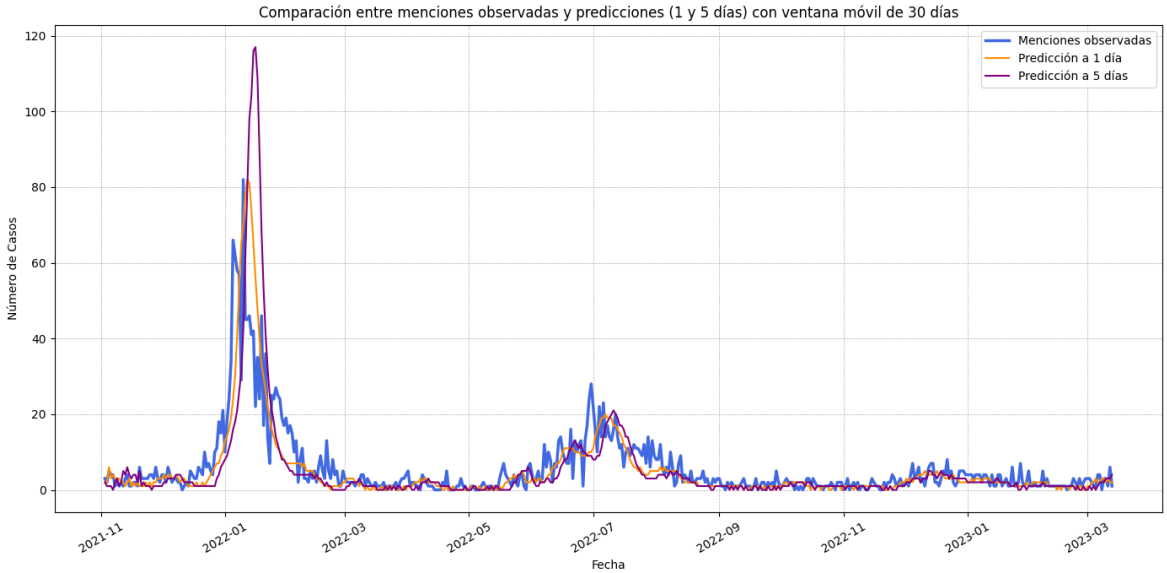

Figura 5.13 Comparación superpuesta de las predicciones a 1 y 5 días, utilizando una ventana móvil de 30 días.

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 términos generales, reproduce con buena precisión 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 identificó una tendencia a la sobreestimación, atribuida probablemente a la intensificación repentina de la atención mediática y a la consecuente amplificación en la conversación digital, que incrementó de forma desproporcionada las menciones de contagio.

También se observó que el modelo ofrece mayor precisión al realizar predicciones con un día de anticipación en comparación con horizontes de cinco días. Asimismo, el modelo entrenado con una ventana móvil de 15 días mostró el mejor desempeño tanto para el horizonte de un día como para el de cinco, logrando un equilibrio óptimo entre estabilidad y sensibilidad frente a cambios recientes. Estos hallazgos permiten orientar futuros ajustes del sistema hacia combinaciones más adecuadas de ventana y horizonte de pronóstico.

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). Esta combinación ofrece sensibilidad semántica, capacidad de adaptación y claridad en la evolución temporal, permitiendo realizar predicciones a corto plazo ajustadas dinámicamente. De esta manera, se configura como una herramienta potencialmente efectiva para la alerta temprana en contextos epidemiológicos reales, al facilitar intervenciones oportunas basadas en los cambios detectados en redes sociales.

5.5.1 Resultados de la predicción

El modelo de Gompertz mostró un buen ajuste para anticipar la evolución diaria de las afirmaciones de contagio.

Los valores de error promedio obtenidos se resumen en la Tabla 5.9:

Tabla 5.9 Desempeño del modelo Gompertz con diferentes ventanas móviles y horizontes de predicción.

Ventana móvil	Horizonte	RMSE	MAE
7 días	1 día	5.34	2.83
7 días	5 días	11.47	5.18
15 días	1 día	5.14	2.61
15 días	5 días	8.57	3.73
30 días	1 día	5.98	2.96
30 días	5 días	10.08	3.89

En general, el modelo alcanzó mayor precisión en las predicciones a un día, con errores más bajos y una representación fiel de la tendencia observada.
El mejor desempeño se obtuvo con la ventana de 15 días, que logró un equilibrio adecuado entre sensibilidad y estabilidad.

Las predicciones capturaron correctamente los picos de contagio y las variaciones generales de la serie temporal, aunque se detectó ligera sobreestimación en periodos de alta atención mediática.

Estos resultados confirman que la combinación de señales sociales (ConBiBER) y el modelo Gompertz constituye una herramienta efectiva de alerta temprana, capaz de anticipar cambios relevantes en la dinámica epidémica con pocos días de anticipación.

5.6 Identificación de zonas de contagio

La identificación de zonas geográficas con alta concentración de afirmaciones de contagio fue un componente clave de este monitoreo epidemiológico digital. Para lograrlo, se desarrolló un índice de riesgo geoespacial destinado a generar mapas dinámicos diarios que visualizan la distribución del riesgo por regiones específicas.

El cálculo diario del índice se basó en dos componentes principales: (i) los casos activos, identificados a partir de publicaciones que el modelo ConBiBER clasificó como afirmaciones positivas, y (ii) un número máximo esperado de casos *(M)*, que se estimó mediante el siguiente procedimiento heurístico:

1. Se contabilizó el número de usuarios únicos *(N)* que publicaron en la red social X durante un periodo de siete días en cada estado, sin importar la temática de sus publicaciones.
2. Se definió el máximo esperado de contagios *(M)* como el 10% del total de usuarios únicos (N):
$$M = 0.10 \cdot N$$
3. Se calcularon diariamente los casos activos *(C_t)* como la suma de los nuevos casos detectados en el día actual y los cinco días previos, lo cual concuerda con el periodo de contagio estimado de un individuo:
$$C_t = \sum_{i=0}^{5} C_{t-1}$$
donde c_{t-i} corresponde al número de casos nuevos detectados en el día $t-i$.
4. Se normalizaron los casos activos con respecto al máximo esperado para obtener el valor de riesgo (r_t):
$$r_t = \frac{C_t}{M}$$

5. Finalmente, se asignó un nivel de riesgo y un color asociado según los siguientes umbrales:

$$\begin{cases} \text{Verde:} & r_t \leq 0.10 \\ \text{Amarillo:} & 0.10 < r_t \leq 0.25 \\ \text{Naranja:} & 0.25 < r_t \leq 0.55 \\ \text{Rojo:} & r_t > 0.55 \end{cases}$$

Para validar cualitativamente la eficacia del índice propuesto, se realizaron comparaciones visuales con el semáforo epidemiológico oficial emitido por la Secretaría de Salud en tres periodos específicos. La primera comparación correspondió al intervalo del 1 al 14 de noviembre de 2021 (Figura 5.14). El panel A muestra el semáforo epidemiológico oficial, mientras que los paneles B, C y D ilustran los mapas generados por el índice propuesto para los días 1, 7 y 14 de noviembre, respectivamente. El panel E ilustra la sobreposición promedio de los mapas B, C y D en el periodo del 1 al 14 de noviembre de 2021.

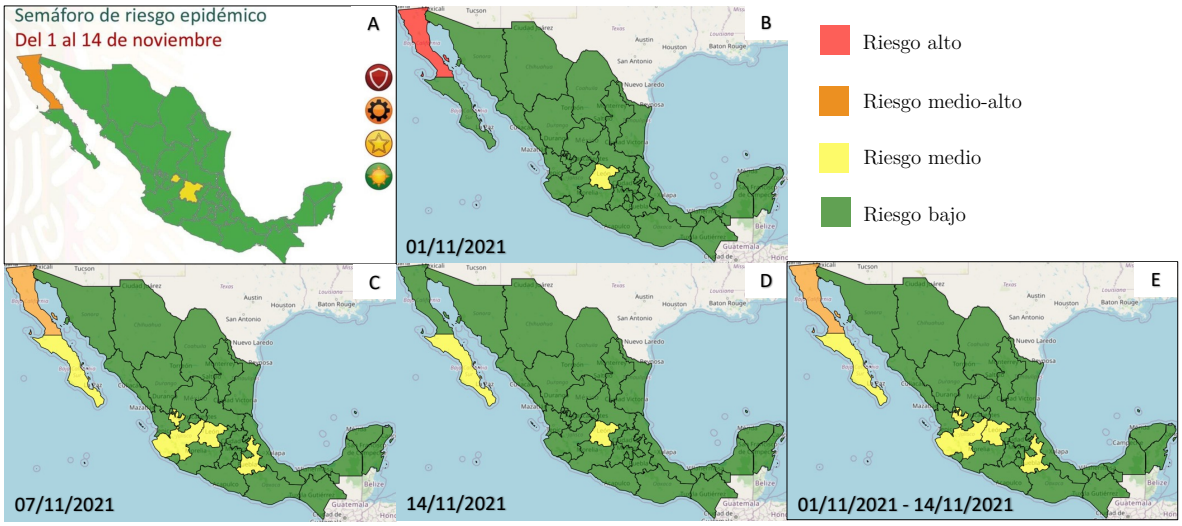

Figura 5.14 Comparación visual entre el semáforo epidemiológico oficial (panel A) y el índice de riesgo estimado a partir de afirmaciones clasificadas (paneles B–E). Fechas representadas: 01/11/2021 (B), 07/11/2021 (C), 14/11/2021 (D), 01/11/2021 - 14/11/2021 (E).

De manera análoga, se llevó a cabo una segunda comparación correspondiente al periodo del 29 de noviembre al 12 de diciembre de 2021 (Figura 5.15). Nuevamente, el panel A corresponde al semáforo oficial, mientras que los paneles B, C y D muestran los mapas del índice propuesto para los días 29 de noviembre, 6 de diciembre y 12 de diciembre, respectivamente. El panel E ilustra la sobreposición promedio de los mapas B, C y D en el periodo del 29 de noviembre al 12 de diciembre de 2021.

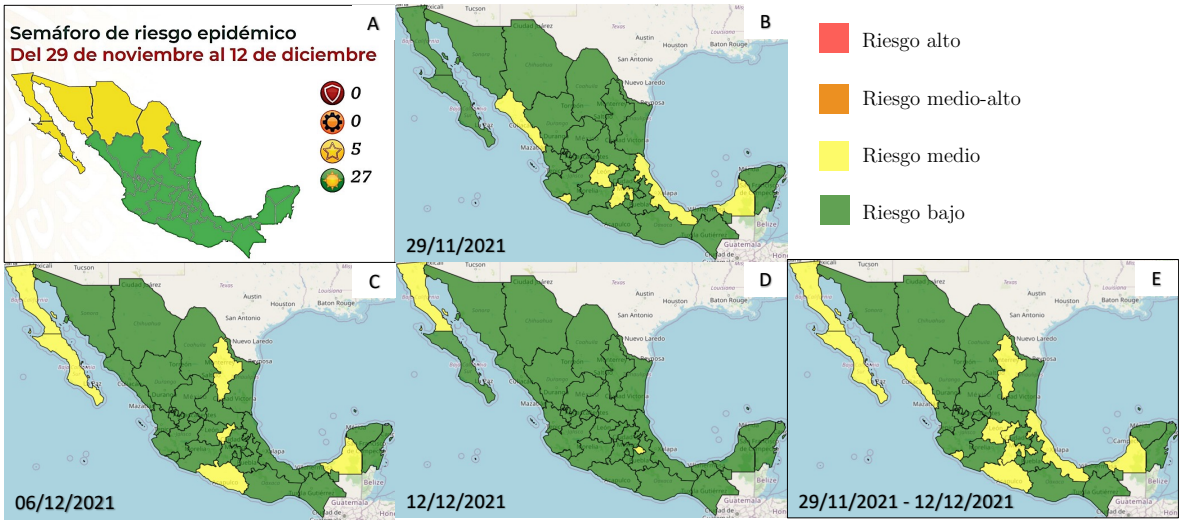

Figura 5.15 Comparación visual entre el semáforo epidemiológico oficial (panel A) y el índice de riesgo estimado a partir de afirmaciones clasificadas (paneles B–E). Fechas representadas: 29/11/2021 (B), 06/12/2021 (C), 12/12/2021 (D), 29/11/2021 - 12/12/2021 (E).

Una tercera comparación se realizó en un periodo de mayor incidencia epidemiológica, comprendido entre el 10 y el 23 de enero de 2022 (Figura 5.16). El panel A ilustra el semáforo oficial, y los paneles B, C y D presentan los mapas del índice propuesto para los días 10, 16 y 23 de enero, respectivamente. El panel E ilustra la sobreposición promedio del índice durante dicho periodo.

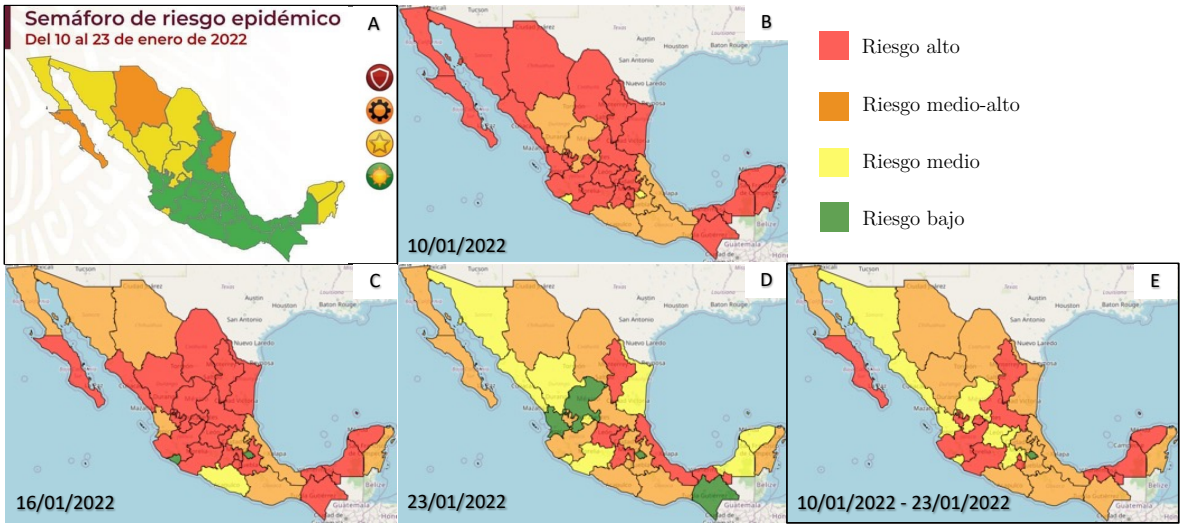

Figura 5.16 Comparación visual entre el semáforo epidemiológico oficial (panel A) y el índice de riesgo estimado a partir de afirmaciones clasificadas (paneles B–D). Fechas representadas: 10/01/2022 (B), 16/01/2022 (C), 23/01/2022 (D), 10/01/2022 - 23/01/2022 (E).

Aunque el índice desarrollado se fundamenta en afirmaciones personales derivadas de publicaciones en la red social X y no en indicadores clínicos tradicionales,

las comparaciones visuales realizadas en los tres periodos analizados mostraron tanto coincidencias como divergencias con el semáforo epidemiológico oficial. En los escenarios de incidencia moderada se observó una alta correspondencia entre ambos esquemas, mientras que durante el periodo de mayor actividad (enero de 2022) el incremento atípico de menciones en redes sociales generó una mayor divergencia, evidenciando la sensibilidad del índice ante variaciones abruptas en la conversación digital. Esta evaluación global respalda la utilidad del índice propuesto como una herramienta complementaria y eficaz para la vigilancia geoespacial.

Una ventaja sustancial del índice propuesto frente al semáforo oficial radica en su frecuencia de actualización diaria, en contraste con la periodicidad quincenal del esquema implementado por la Secretaría de Salud. Dicho esquema se sustentaba en un índice de riesgo compuesto por diez indicadores ponderados, diseñados para reflejar tanto la transmisión como la gravedad de la epidemia. Los indicadores empleados fueron: (1) la tasa de reproducción efectiva, (2) la tasa de incidencia de casos activos estimados, (3) la tasa de mortalidad, (4) la tasa de casos hospitalizados, (5) el porcentaje de ocupación de camas generales, (6) el porcentaje de ocupación de camas con ventilador, (7) el porcentaje de positividad de pruebas diagnósticas, (8) la tendencia de casos hospitalizados, (9) la tendencia de casos con síndrome COVID-19, y (10) la tendencia de la tasa de mortalidad.

Cada indicador recibía un puntaje parcial de 0 a 4, según su nivel de riesgo (de bajo a crítico), y se multiplicaba por un factor de ponderación que daba mayor relevancia a los indicadores asociados con hospitalización y mortalidad, y menor a los vinculados con transmisión comunitaria. La suma ponderada de los diez indicadores generaba un puntaje total entre 0 y 40, el cual clasificaba el nivel de riesgo en cuatro categorías: verde (0–9), amarillo (10–19), naranja (20–29) y rojo (30–40). Este método reflejaba la situación epidémica con base en datos clínicos e institucionales, actualizados de manera quincenal [66].

En contraste, el índice propuesto en este estudio captura la percepción ciudadana y la dinámica del contagio a partir de datos generados en tiempo real, lo que le otorga una mayor sensibilidad para detectar variaciones abruptas en la propagación o percepción del riesgo. Su actualización diaria permite anticipar cambios epidemiológicos con horas o incluso días de ventaja respecto a los mecanismos institucionales más lentos y dependientes de procesos administrativos.

Adicionalmente, al basarse exclusivamente en publicaciones clasificadas como afirmaciones positivas de contagio (a diferencia de enfoques que consideran todas las menciones sobre COVID-19), el índice propuesto ofrece una representación precisa y específica del riesgo activo de contagio. Finalmente, la metodología propuesta amplía los alcances de enfoques previos centrados solo en el análisis textual para identificar síntomas o diagnósticos autorreportados (por ejemplo, [67] y [68]), al integrar

componentes semánticos, temporales y geoespaciales. De este modo, se presenta un marco metodológico integral, dinámico y adaptable a distintas regiones y escenarios epidemiológicos, que además puede servir como base replicable para futuras emergencias sanitarias.

5.6.1 Resultados del índice geoespacial

Los resultados del índice propuesto respecto al semáforo epidemiológico oficial se obtuvieron a partir de tres comparaciones realizadas en periodos representativos del comportamiento epidémico nacional: 1–14 de noviembre de 2021, 29 de noviembre al 12 de diciembre de 2021 y 10 al 23 de enero de 2022.

En cada intervalo se analizaron los niveles de riesgo asignados por ambos esquemas, codificados de forma ordinal (verde = 1, amarillo = 2, naranja = 3, rojo = 4), y se calcularon tres métricas principales: el porcentaje de coincidencia (IoU), el error porcentual regional y la correlación de Pearson (r).

Los valores obtenidos se resumen en la Tabla 5.10, la cual presenta la comparación general entre el índice propuesto y el semáforo epidemiológico oficial en los tres periodos evaluados.

Tabla 5.10 Resultados comparativos entre el índice propuesto y el semáforo epidemiológico oficial.

Periodo	IoU / % de coincidencia	Error porcentual regional	Correlación Pearson
01 nov – 14 nov 2021	87.5 %	12.5 %	r = 0.89
29 nov – 12 dic 2021	59.4 %	40.6 %	r = 0.022
10 ene – 23 ene 2022	21.9 %	78.1 %	r = −0.0279

Durante el primer periodo (1–14 de noviembre de 2021), el índice geoespacial mostró una alta coincidencia (87.5 %) y una correlación significativa (r = 0.89) con el semáforo oficial, reproduciendo de manera consistente los patrones espaciales de riesgo reportados por la Secretaría de Salud.

En el segundo periodo (29 de noviembre al 12 de diciembre de 2021), la coincidencia disminuyó a 59.4 % (r = 0.022), evidenciando un aumento en la sensibilidad del modelo frente a fluctuaciones locales en las publicaciones de redes sociales, incluso cuando los indicadores oficiales permanecían estables.

Finalmente, durante el periodo de alta incidencia (10–23 de enero de 2022), la correspondencia descendió notablemente (IoU = 21.9 %, r ≈ 0), lo que indica que el índice detectó una expansión más amplia del riesgo, reflejada en el aumento de menciones afirmativas de contagio.

Esta divergencia no representa un error del modelo, sino una reacción anticipada a la dinámica social de la pandemia: el índice respondió al incremento en la percepción pública de contagios antes de que se consolidaran los reportes clínicos oficiales.

En conjunto, los resultados evidencian que el índice propuesto mantiene alta concordancia en periodos de estabilidad epidemiológica, pero mayor sensibilidad y anticipación durante etapas de rápida propagación, lo que respalda su utilidad como herramienta complementaria de vigilancia y alerta temprana.

5.7 Comparación con el estado del arte

Con el fin de contextualizar y validar los resultados obtenidos en la presente tesis, en esta sección se realiza una comparación cuantitativa y cualitativa del rendimiento de los modelos desarrollados (ConBiBER y Gompertz) frente al trabajo relacionado.

5.7.1 Rendimiento del modelo de clasificación (ConBiBER)

La clasificación precisa de afirmaciones de contagio es fundamental en este nuevo enfoque de monitoreo. El rendimiento del modelo ConBiBER se comparó con otros clasificadores reportados en la literatura para tareas análogas de minería de texto en salud pública. La comparativa se visualiza en la Tabla 5.11.

Tabla 5.11 Comparativa de rendimiento en clasificación de textos.

Modelo	Tarea (Fuente)	Precisión	Sensibilidad (Recall)	F1-Score
ConBiBER	**Afirmaciones COVID-19**	**73.9 %**	**91.0 %**	**81.6 %**
SMO (en [40])	Ideación suicida	89.5 %	-	89.3 %
Random Tree (en [41])	Rumores de Zika (mejor caso)	74.6 %	86.9 %	68.8 %
Sistema (en [36])	Alerta temprana COVID-19	-	78.0 %	-

Los resultados de ConBiBER demuestran un rendimiento robusto y altamente competitivo.

- Alta Sensibilidad: La métrica más destacada es la Sensibilidad (Recall) del 91.0 %. Este valor es crucial para un sistema de vigilancia epidemiológica, ya que indica que el modelo es capaz de identificar correctamente al 91% de las afirmaciones positivas reales. Este resultado supera la sensibilidad del 78.0 % reportada por el sistema de alerta temprana en [36] y el 86.9 % de recall del modelo en [41]. En este contexto, es preferible un modelo con mayor

sensibilidad, aun a costa de una precisión ligeramente menor, para minimizar los falsos negativos.

- Equilibrio (F1-Score): El F1-Score de 81.6 % muestra un balance sólido entre precisión y sensibilidad. Este valor es comparable al 89.3 % del modelo SMO para un dominio diferente (ideación suicida) y es significativamente superior al F1-Score máximo de 68.8 % obtenido en la detección de rumores de Zika.
- Precisión: La Precisión (73.9 %) es comparable al 74.6 % del mejor caso reportado en [41], indicando un desempeño alineado con el estado del arte en tareas semánticamente complejas sobre redes sociales.

5.7.2 Rendimiento del modelo predictivo

Para la predicción temporal, se comparó el Error Medio Absoluto (MAE) del modelo Gompertz (utilizando la ventana óptima de 15 días) con los enfoques de series temporales mencionados en el estado del arte, como los utilizados en el sistema DEFENDER [50].

Tabla 5.12 Comparativa de rendimiento en predicción.

Modelo	Tarea	Métrica (MAE)
Nuestro modelo (Ventana 15 días, Horiz. 1 día)	**Predicción de menciones**	**2.61**
Nuestro modelo (Ventana 15 días, Horiz. 5 días)	**Predicción de menciones**	**3.73**
DEFENDER (ARIMA + X) [50]	Nowcasting de casos	8.20
ARIMA Tradicional [50]	Nowcasting de casos	13.05

El desempeño del modelo Gompertz resulta notablemente positivo. Aunque los modelos en [50] realizan *nowcasting* de casos y este trabajo predice menciones, las métricas de error son análogas.

El MAE de 2.61 (a 1 día) y 3.73 (a 5 días) obtenido por la función de Gompertz es significativamente más bajo que el MAE de 8.20 reportado por el modelo ARIMA enriquecido con datos de X y muy inferior al MAE de 13.05 del ARIMA tradicional.

Esto sugiere que, para la tarea específica de modelar la evolución de las afirmaciones de contagio, un modelo de crecimiento como Gompertz (que es más simple e interpretable) puede ser más eficaz y preciso que modelos de series temporales más complejos.

5.7.3 Validación de la correlación

La correlación de Pearson de 0.83 observada entre las menciones clasificadas por ConBiBER y los casos oficiales de la Secretaría de Salud, valida cuantitativamente la idoneidad de X como fuente de datos.

Este hallazgo refuerza lo reportado en el estado del arte, donde estudios como [34] (enfermedad de Lyme) encontraron una "fuerte correlación" y [36] (COVID-19) y [103] (Zika) determinaron que los datos de redes sociales podían anticipar los reportes oficiales. El valor de 0.83 aporta una métrica robusta a estas conclusiones, confirmando el potencial de la metodología para la anticipación de tendencias epidemiológicas.

6

Conclusiones y trabajos futuros

En este capítulo se presentan las conclusiones derivadas de la investigación, las cuales sintetizan los principales hallazgos obtenidos a lo largo del trabajo. Asimismo, se destacan las contribuciones más relevantes en términos metodológicos y aplicados, así como las limitaciones identificadas durante el desarrollo del estudio. Finalmente, se plantean posibles líneas de trabajo futuro orientadas a la mejora y ampliación del enfoque propuesto, con el fin de fortalecer su aplicabilidad en contextos epidemiológicos diversos.

6.1 Conclusiones

6.1.1 Cumplimiento de los objetivos

La presente investigación tuvo como objetivo general desarrollar un modelo computacional para el monitoreo y predicción de zonas de contagio. Este objetivo se ha satisfecho mediante la integración exitosa de una metodología que abarcó desde la recolección de datos en la red social X hasta la generación de visualizaciones de riesgo. La combinación de técnicas de procesamiento de lenguaje natural, aprendizaje profundo (BERT) y modelización (función de Gompertz) ha resultado en una herramienta funcional que complementa los sistemas de vigilancia epidemiológica, tal como se buscaba.

En cuanto a los objetivos específicos, estos fueron alcanzados de la siguiente manera:

- El primer objetivo, enfocado en diseñar un mecanismo de extracción y preprocesamiento de datos de X, se cumplió mediante el desarrollo de un pipeline computacional que permitió recolectar, filtrar y normalizar grandes volúmenes de texto, preparando la información relevante para el análisis semántico.
- El segundo objetivo, que buscaba clasificar las afirmaciones de contagio, se logró a través de la implementación y ajuste fino de un modelo BERT. Esta "modelización mediante BERT" fue crucial para lograr una detección semánticamente precisa, permitiendo al sistema discernir entre ruido y afirmaciones indicativas de contagio.
- El tercer objetivo, orientado a desarrollar un modelo predictivo temporal y espacial, se alcanzó mediante la aplicación de la función de Gompertz sobre los datos clasificados. Esta aproximación permitió generar una "caracterización predictiva" y, mediante el uso de "ventanas móviles", se obtuvo el dinamismo necesario para una "estimación ajustada y continua" de la evolución de los contagios.
- El cuarto objetivo, que buscaba diseñar un algoritmo heurístico de localización dinámica, se cumplió con el desarrollo del "índice de riesgo geoespacial". Este índice fue el componente algorítmico que permitió generar los "mapas dinámicos diarios" y, con ello, determinar el estado de riesgo epidemiológico en la República mexicana.

6.1.2 Discusión

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 pone de manifiesto el potencial de las redes sociales para capturar tendencias epidemiológicas en tiempo real.

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 estudio de caso, ya sea en el seguimiento de diferentes enfermedades o en la aplicación en otros países.

6.2 Trabajo futuro

Los resultados obtenidos en esta tesis abren diversas líneas de investigación que permitirán fortalecer y ampliar el enfoque propuesto. A continuación, se describen las principales direcciones sugeridas para trabajos futuros:

- Mejora del etiquetado automático de afirmaciones
  Se plantea optimizar la fase de clasificación de afirmaciones relacionadas con contagio mediante el uso de modelos de lenguaje preentrenados de última generación. Aunque el clasificador ConBiBER (BERT + CNN) demostró ser eficaz, la rápida evolución en esta área abre la posibilidad de incorporar arquitecturas más avanzadas, como RoBERTa, DeBERTa, LLaMA o modelos generativos como GPT. Estas arquitecturas pueden ofrecer una comprensión semántica más precisa y una mayor capacidad para detectar expresiones ambiguas, informales o no literales típicas del lenguaje empleado en redes sociales.
- Extensión a otras enfermedades infecciosas
  La metodología desarrollada puede ser adaptada a otras enfermedades transmisibles de relevancia epidemiológica, como influenza, dengue o viruela del mono. Para ello, será necesario ajustar los esquemas de recolección, clasificación y predicción en función de las características particulares de cada enfermedad y su representación lingüística en redes sociales.
- Combinación de modelos predictivos
  Se contempla integrar el modelo Gompertz con otros enfoques de predicción temporal, tales como el modelo autorregresivo integrado de media móvil

(ARIMA) o redes neuronales recurrentes (RNN). Esta combinación busca aprovechar la capacidad del modelo Gompertz para capturar la fase inicial de crecimiento y complementar su desempeño en fases posteriores mediante técnicas más adaptativas, lo que permitiría realizar proyecciones más precisas en horizontes superiores a cinco días.

- Construcción de un índice de riesgo más robusto
  Se propone enriquecer el índice de riesgo desarrollado mediante la incorporación de variables adicionales como factores socioeconómicos, patrones de movilidad y características demográficas. Estos elementos permitirán construir un índice más contextualizado, que refleje de manera más precisa las dinámicas locales de contagio y contribuya a una mejor interpretación del riesgo epidemiológico en distintas regiones.

6.3 Aportaciones científicas

Durante el desarrollo de la tesis doctoral se trabajaron áreas como el procesamiento de lenguaje natural, aprendizaje profundo, análisis temporal y geoespacial, redes sociales basadas en ubicación, y vigilancia epidemiológica. Se adquirió conocimiento y experiencia significativa en técnicas como el uso de modelos de lenguaje preentrenados (BERT), redes neuronales convolucionales (CNN), modelado matemático con la función de Gompertz y generación de mapas de riesgo. Estas contribuciones se concretaron en distintas publicaciones científicas y en el desarrollo de un sistema de monitoreo, las cuales se mencionan a continuación.

1. **P. Wences-Olguin**, A. Martínez-Rebollar, and H. Estrada-Esquivel, “Monitoreo y Predicción de Enfermedades Infecciosas a través del Análisis de Redes Sociales,” Revista Mexicana de Ingeniería Biomédica, vol. 46, no. 3, Nov. 2025.
2. **P. Wences-Olguin**, A. Martínez-Rebollar, and H. Estrada-Esquivel, “Redes sociales y salud pública: Exploración de Twitter como indicador de tendencias de COVID-19 en México,” Tecnología y Ciencia Aplicada, vol. 7, no. 1, p.23 –, Jan.–Jun. 2024.
3. Estrada-Esquivel, A. Martínez-Rebollar, **P. Wences-Olguin**, Y. Hernandez-Perez, and J. Ortiz-Hernandez, “A smart information system for passengers of urban transport based on IoT,” Electronics, vol. 11, no. 5, p. 834, 2022. [Online]. Available: https://doi.org/10.3390/electronics11050834
4. A. Martínez-Rebollar, **P. Wences-Olguin**, G. Palacios-Gonzalez, H. Estrada-Esquivel, and Y. Hernandez-Perez, “Generation of Twitter information

databases: A case study for the mobility of people infected with COVID-19," Research in Computing Science, vol. 151, no. 1, pp. 43–50, Jan. 2022.

Referencias

[1] SAS, "Solutions for real-time disease surveillance and data-driven decision making". Consultado: el 4 de noviembre de 2025. [En línea]. Disponible en: https://www.sas.com/content/dam/SAS/documents/briefs/solution-brief/en/situational-awareness-health-crisis-response-analysis-111295.pdf?utm_source=economist&utm_medium=referral&utm_campaign=ana-gen-gbc-c19-global

[2] Sasha Walek, "New COVID local risk index identifies neighborhoods at highest risk of pandemic's impact". Consultado: el 4 de noviembre de 2025. [En línea]. Disponible en: https://nyulangone.org/news/new-covid-local-risk-index-identifies-neighborhoods-highest-risk-pandemics-impact

[3] S. Masri, J. Jia, C. Li, G. Zhou, M. C. Lee, G. Yan, y J. Wu, "Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic", *BMC Public Health*, vol. 19, núm. 1, jun. 2019, doi: 10.1186/s12889-019-7103-8.

[4] A. Hortaçsu, J. Liu, y T. Schwieg, "Estimating the fraction of unreported infections in epidemics with a known epicenter: An application to COVID-19", *J. Econom.*, núm. January, 2020, doi: 10.1016/j.jeconom.2020.07.047.

[5] J. C. Salvador, "Covid-19: ¿Nos podemos creer las estadísticas oficiales del coronavirus?" Consultado: el 4 de noviembre de 2025. [En línea]. Disponible en: https://www.redaccionmedica.com/opinion/covid-19-nos-podemos-creer-las-estadisticas-oficiales-del-coronavirus--9422

[6] L. Simonsen, J. R. Gog, D. Olson, y C. Viboud, "Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems", *J. Infect. Dis.*, vol. 214, núm. suppl 4, pp. S380–S385, dic. 2016, doi: 10.1093/infdis/jiw376.

[7] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, y D. Newth, "Understanding Human Mobility from Twitter", *PLOS ONE*, vol. 10, núm. 7, pp. e0131469–e0131469, jul. 2015, doi: 10.1371/journal.pone.0131469.

[8] J. Devlin, M.-W. Chang, K. Lee, y K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", en *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, y T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[9] A. Rogers, O. Kovaleva, y A. Rumshisky, "A Primer in BERTology: What We Know About How BERT Works", doi: 10.1162/tacl.

[10] Mayank Mishra, “Convolutional neural networks, explained”, Towards Data Science. Consultado: el 4 de noviembre de 2025. [En línea]. Disponible en: https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939

[11] MATLAB, “Introducing deep learning with MATLAB”, Deep learning book. Consultado: el 4 de noviembre de 2025. [En línea]. Disponible en: https://la.mathworks.com/campaigns/offers/next/deep-learning-ebook.html

[12] Purwono, A. Ma’arif, W. Rahmaniar, H. I. K. Fathurrahman, A. Z. K. Frisky, y Q. M. U. Haq, “Understanding of Convolutional Neural Network (CNN): A Review”, *Int. J. Robot. Control Syst.*, vol. 2, núm. 4, pp. 739–748, 2022, doi: 10.31763/ijrcs.v2i4.888.

[13] R. Yamashita, M. Nishio, R. K. G. Do, y K. Togashi, “Convolutional neural networks: an overview and application in radiology”, *Insights Imaging*, vol. 9, núm. 4, pp. 611–629, ago. 2018, doi: 10.1007/s13244-018-0639-9.

[14] N. Alsadi, S. A. Gadsden, y J. Yawney, “Intelligent estimation: A review of theory, applications, and recent advances”, *Digit. Signal Process. Rev. J.*, vol. 135, abr. 2023, doi: 10.1016/j.dsp.2023.103966.

[15] D. N. Dwivedi y A. Gupta, “Artificial intelligence-driven power demand estimation and short-, medium-, and long-term forecasting”, en *Artificial Intelligence for Renewable Energy Systems*, Elsevier, 2022, pp. 231–242. doi: 10.1016/B978-0-323-90396-7.00013-4.

[16] P. Kumari y D. Toshniwal, “Deep learning models for solar irradiance forecasting: A comprehensive review”, *J. Clean. Prod.*, vol. 318, oct. 2021, doi: 10.1016/j.jclepro.2021.128566.

[17] K. M. C. Tjørve y 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, núm. 6, jun. 2017, doi: 10.1371/journal.pone.0178691.

[18] B. Gompertz, “On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies”, vol. 115, pp. 513–583, 1825, doi: https://www.jstor.org/stable/107756.

[19] J. S. Kim, H. Jin, H. Kavak, O. C. Rouly, A. Crooks, D. Pfoser, C. Wenk, y A. Zufle, “Location-Based Social Network Data Generation Based on Patterns of Life”, presentado en Proceedings - IEEE International Conference on Mobile Data Management, Institute of Electrical and Electronics Engineers Inc., jun. 2020, pp. 158–167. doi: 10.1109/MDM48529.2020.00038.

[20] Y. Wang, H. Sun, Y. Zhao, W. Zhou, y S. Zhu, "A Heterogeneous Graph Embedding Framework for Location-Based Social Network Analysis in Smart Cities", *IEEE Trans. Ind. Inform.*, vol. 16, núm. 4, pp. 2747–2755, abr. 2020, doi: 10.1109/TII.2019.2953973.

[21] A. Cerdan Schwitzguébel y O. Romero Bartomeus, "Location-Based Social Network Data for Exploring Spatial and Functional Urban Tourists and Residents Consumption Patterns", 2018.

[22] S. Costanzo y A. Flores, "COVID-19 Contagion Risk Estimation Model for Indoor Environments", *Sensors*, vol. 22, núm. 19, oct. 2022, doi: 10.3390/s22197668.

[23] J. F. Sortino Barrionuevo, H. Castro Noblejas, y M. J. Perles Roselló, "Mapping the Risk of COVID-19 Contagion at Urban Scale", *Land*, vol. 11, núm. 9, sep. 2022, doi: 10.3390/land11091480.

[24] C. A. Mackenzie, "Summarizing risk using risk measures and risk indices", *Risk Anal.*, vol. 34, núm. 12, pp. 2143–2162, dic. 2014, doi: 10.1111/risa.12220.

[25] Organizacion Mundial de la Salud, "Coronavirus". Consultado: el 4 de noviembre de 2025. [En línea]. Disponible en: https://www.who.int/es/health-topics/coronavirus#tab=tab_1

[26] F. J. Díaz-Castrillón y A. I. Toro-Montoya, "SARS-CoV-2/COVID-19: el virus, la enfermedad y la pandemia", *Med. Lab.*, vol. 24, núm. 3, pp. 183–205, may 2020, doi: 10.36384/01232576.268.

[27] D. Murthy, "Sociology of Twitter/X: Trends, Challenges, and Future Research Directions", *Annu. Rev. Sociol.*, vol. 25, pp. 51–51, 2024, doi: 10.1146/annurev-soc-031021.

[28] J. Ruiz-Soler, "Twitter research for social scientists: A brief introduction to the benefits, limitations and tools for analysing Twitter data", *Dígitos Rev. Comun. Digit.*, vol. 1, núm. 3, pp. 17–32, may 2017, doi: 10.7203/rd.v1i3.87.

[29] H. Zhang, S. Li, Y. Chen, J. Dai, y Y. Yi, "A Novel Encoder-Decoder Model for Multivariate Time Series Forecasting", *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/5596676.

[30] N. M. Norwawi, "Sliding window time series forecasting with multilayer perceptron and multiregression of COVID-19 outbreak in Malaysia", en *Data Science for COVID-19 Volume 1: Computational Perspectives*, Elsevier, 2021, pp. 547–564. doi: 10.1016/B978-0-12-824536-1.00025-3.

[31] Geo Joy, "How to Transform Time Series to a Supervised Learning Problem", Deep Learning for Time Series Forecasting. Consultado: el 4 de noviembre de

2025. [En línea]. Disponible en: https://github.com/Geo-Joy/Deep-Learning-for-Time-Series-Forecasting/blob/master/C4%20-%20How%20to%20Transform%20Time%20Series%20to%20a%20Supervised%20Learning%20Problem.md

[32] David Andrés Sánchez, "Métodos de pronóstico de Series Temporales", Series temporales. Consultado: el 4 de noviembre de 2025. [En línea]. Disponible en: https://mlpills.dev/series-temporales/pronostico-de-series-temporales/

[33] 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, núm. 30, pp. 25–47, dic. 2022, doi: 10.21138/GF.789.

[34] J. S. P. Tulloch, R. Vivancos, R. M. Christley, A. D. Radford, y 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, dic. 2019, doi: 10.1016/j.yjbinx.2019.100060.

[35] A. Fernández-Rodríguez, I. Casas, E. Culebras, E. Morilla, M. C. Cohen, y J. Alberola, "COVID-19 and post-mortem microbiological studies", *Rev. Espanola Med. Leg.*, vol. 46, núm. 3, pp. 127–138, jul. 2020, doi: 10.1016/j.reml.2020.05.007.

[36] N. E. Kogan, L. Clemente, P. Liautaud, J. Kaashoek, N. B. Link, A. T. Nguyen, F. S. Lu, P. Huybers, B. Resch, C. Havas, A. Petutschnig, J. Davis, M. Chinazzi, B. Mustafa, W. P. Hanage, A. Vespignani, y M. Santillana, "An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time", *Sci. Adv.*, vol. 7, núm. 10, p. eabd6989, mar. 2021, doi: 10.1126/sciadv.abd6989.

[37] D. I. Fallatah y H. A. Adekola, "Digital epidemiology: harnessing big data for early detection and monitoring of viral outbreaks", *Infect. Prev. Pract.*, vol. 6, núm. 3, p. 100382, sep. 2024, doi: 10.1016/j.infpip.2024.100382.

[38] A. Gupta y R. Katarya, "Social media based surveillance systems for healthcare using machine learning: A systematic review", *J. Biomed. Inform.*, vol. 108, ago. 2020, doi: 10.1016/j.jbi.2020.103500.

[39] A. E. Aiello, A. Renson, y P. N. Zivich, "Social Media-and Internet-Based Disease Surveillance for Public Health", *Annu Rev Public Health*, vol. 41, pp. 1–1, 2025, doi: 10.1146/annurev-publhealth.

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

[41] A. Ghenai y Y. Mejova, "Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter", el 12 de julio de 2017, *arXiv*: arXiv:1707.03778. doi: 10.48550/arXiv.1707.03778.

[42] L. C. Erlandsson, "Use of social networks as a source of information on health and digital health literacy in the Spanish general population", *Rev Esp Salud Pública*, vol. 98, p. 10, may 2025.

[43] M. U. Hoque, K. Lee, J. L. Beyer, S. R. Curran, K. S. Gonser, N. S. N. Lam, V. V. Mihunov, y K. Wang, "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: 10.1109/ACCESS.2022.3189670.

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

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

[46] R. P. D. Redondo, C. Garcia-Rubio, A. F. Vilas, C. Campo, y A. Rodriguez-Carrion, "A hybrid analysis of LBSN data to early detect anomalies in crowd dynamics", *Future Gener. Comput. Syst.*, vol. 109, pp. 83–94, 2020, doi: 10.1016/j.future.2020.03.038.

[47] O. Alonso, V. Kandylas, S. E. Tremblay, y S. Whiting, "Answering recreational web searches with relevant things to do results", *Inf. Process. Manag.*, vol. 57, núm. 2, pp. 102184–102184, 2020, doi: 10.1016/j.ipm.2019.102184.

[48] X. Xiong, S. Qiao, Y. Li, N. Han, G. Yuan, y Y. Zhang, "A point-of-interest suggestion algorithm in Multi-source geo-social networks", *Eng. Appl. Artif. Intell.*, vol. 88, núm. September 2019, pp. 103374–103374, 2020, doi: 10.1016/j.engappai.2019.103374.

[49] S. Chae, S. Kwon, y D. Lee, "Predicting infectious disease using deep learning and big data", *Int. J. Environ. Res. Public. Health*, vol. 15, núm. 8, ago. 2018, doi: 10.3390/ijerph15081596.

[50] N. Thapen, D. Simmie, C. Hankin, y J. Gillard, "DEFENDER: Detecting and forecasting epidemics using novel data-analytics for enhanced response", *PLoS ONE*, vol. 11, núm. 5, may 2016, doi: 10.1371/journal.pone.0155417.

[51] W. Yang y X. Chang, "Time series analysis and prediction of the trends of COVID-19 epidemic in Singapore based on machine learning", *Comput. Methods Programs Biomed. Update*, vol. 7, ene. 2025, doi: 10.1016/j.cmpbup.2025.100190.

[52] A. Tarafdar, J. Mahato, R. K. Upadhyay, y P. Bhattacharya, "Exploring the synergy of media awareness and quarantine classes in SiSAQEIH model for pandemic control: A Deep LSTM-RNN predictions", *Phys. Nonlinear Phenom.*, vol. 474, pp. 134563–134563, abr. 2025, doi: 10.1016/j.physd.2025.134563.

[53] A. Chhabra, S. K. Singh, A. Sharma, S. Kumar, B. B. Gupta, V. Arya, y K. T. Chui, "Sustainable and intelligent time-series models for epidemic disease forecasting and analysis", *Sustain. Technol. Entrep.*, vol. 3, núm. 2, may 2024, doi: 10.1016/j.stae.2023.100064.

[54] S. Y. Ilu y R. Prasad, "Improved autoregressive integrated moving average model for COVID-19 prediction by using statistical significance and clustering techniques", *Heliyon*, vol. 9, núm. 2, feb. 2023, doi: 10.1016/j.heliyon.2023.e13483.

[55] W. Angulo, J. M. Ramírez, D. De Cecchis, J. Primera, H. Pacheco, y 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, núm. 1, dic. 2021, doi: 10.1038/s41598-021-95785-y.

[56] R. Ramalingam, A. J. Gnanaprakasam, y S. Boulaaras, "Stability and control analysis of COVID-19 spread in India using SEIR model", *Sci. Rep.*, vol. 15, núm. 1, dic. 2025, doi: 10.1038/s41598-025-93994-3.

[57] "Use Cases, Tutorials, & Documentation". Consultado: el 16 de junio de 2025. [En línea]. Disponible en: https://developer.x.com/en

[58] A. Martínez-Rebollar, P. Wences-Olguin, G. Palacios-Gonzalez, H. Estrada-Esquivel, y Y. Hernandez-Perez, "Generation of Twitter Information Databases: A Case Study for the Mobility of People Infected with COVID-19", *Res. Comput. Sci.*, vol. 151, núm. 151(1) (2022), pp. 43–50, ene. 2022.

[59] "Tweepy". Consultado: el 16 de junio de 2025. [En línea]. Disponible en: https://www.tweepy.org/

[60] "re — Regular expression operations", Python documentation. Consultado: el 4 de noviembre de 2025. [En línea]. Disponible en: https://docs.python.org/3/library/re.html

[61] "NLTK :: Natural Language Toolkit". Consultado: el 4 de noviembre de 2025. [En línea]. Disponible en: https://www.nltk.org/

[62] J.-D. Morel, J.-M. Morel, y L. Alvarez, "Time warping between main epidemic time series in epidemiological surveillance", *PLOS Comput. Biol.*, vol. 19, núm. 12, p. e1011757, dic. 2023, doi: 10.1371/journal.pcbi.1011757.

[63] H. Nishiura, G. Chowell, H. Heesterbeek, y J. Wallinga, "The ideal reporting interval for an epidemic to objectively interpret the epidemiological time course", *J. R. Soc. Interface*, vol. 7, núm. 43, pp. 297–307, feb. 2010, doi: 10.1098/rsif.2009.0153.

[64] Q. Xu, Y. R. Gel, L. L. Ramirez Ramirez, K. Nezafati, Q. Zhang, y K.-L. Tsui, "Forecasting influenza in Hong Kong with Google search queries and statistical model fusion", *PLOS ONE*, vol. 12, núm. 5, p. e0176690, may 2017, doi: 10.1371/journal.pone.0176690.

[65] M. J. Paul, M. Dredze, y D. Broniatowski, "Twitter Improves Influenza Forecasting", *PLoS Curr.*, 2014, doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.

[66] Secretaría de salud México, "Lineamiento para la estimación de riesgos del semaforo por regiones COVID-19", México, 2021. [En línea]. Disponible en: https://coronavirus.gob.mx/wp-content/uploads/2021/08/2021.8.18-Metodo_semaforo_COVID.pdf

[67] A. Z. Klein, S. Kunatharaju, K. O'Connor, y 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: 10.2196/46484.

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

Apéndice A

Lista de términos empleados en el filtrado semántico

cobi, cobicho, cobid, coronavirus, coronavid, coronabirus, coronaviru, coronavis, coronovis, covid, covid-19, covid_19, covid19, coronavirus, covi, covicho, covir, covis, covit, covit19, covitt, covyd, covyd19, sars-cov2, cov2, bicho.

Apéndice B

Glosario de términos empleados en esta tesis.

BERT (Bidirectional Encoder Representations from Transformers)

Es un modelo de representación del lenguaje desarrollado por Google AI. Su característica principal es que preentrena representaciones "condicionando conjuntamente en el contexto tanto a la izquierda como a la derecha en todas las capas". En esta tesis, se utilizó como base para el clasificador ConBiBER.

Bounding Box (Caja Delimitadora)

Un rectángulo geográfico definido por dos coordenadas: el vértice suroeste (longitud mínima, latitud mínima) y el vértice noreste (longitud máxima, latitud máxima). En esta investigación, se utilizaron para delimitar las áreas geográficas de las cuales se recolectaron publicaciones en la red social X.

ConBiBER (Convolucional Binary Classifier with BERT)

Es el modelo de clasificación supervisado desarrollado en esta tesis. Fue diseñado específicamente para "identificar afirmaciones de contagio en textos coloquiales" de la red social X. Su arquitectura combina un modelo BERT multilingüe con una Red Neuronal Convolucional (CNN).

Función Gompertz (Modelo Gompertz)

Un modelo matemático que describe el crecimiento que "crece rápidamente y se desacelera con el tiempo". En esta tesis, se empleó para predecir el número de afirmaciones positivas de contagio.

Índice de Riesgo (Geoespacial)

Una herramienta desarrollada en esta tesis para cuantificar y resumir el riesgo de contagio en zonas geográficas específicas.

LBSN (Location-Based Social Networking / Redes Sociales Basadas en la Ubicación)

Una evolución de las redes sociales que "integrando la dimensión espacial en las interacciones sociales". Permiten que el contenido generado por el usuario sea asociado con "ubicaciones geográficas específicas".

Ventanas Móviles (Sliding Windows)

Un método de análisis de series temporales que divide los datos en "segmentos de tamaño fijo que se desplazan secuencialmente a lo largo de la serie". Los valores anteriores dentro de la ventana se usan para estimar valores futuros. En esta tesis, se aplicaron ventanas de 7, 15 y 30 días junto con el modelo Gompertz para generar predicciones dinámicas.

Zonas de Contagio

En el contexto específico de esta tesis, se definen como "un área geográfica en la cual se concentra un número significativo de publicaciones en la red social X que contienen afirmaciones personales de infección por COVID-19". Estas zonas son identificadas por el clasificador ConBiBER y georreferenciadas para el cálculo del índice de riesgo

Procesamiento de Lenguaje Natural (PLN)

Un campo de la inteligencia artificial enfocado en la interacción entre las computadoras y el lenguaje humano.

Redes Neuronales Convolucionales (CNN)

Un tipo de red neuronal profunda especializada en "procesar datos con una topología de cuadrícula" mediante el uso de filtros para "extraer características esenciales". Aunque son comúnmente usadas para imágenes, en esta tesis se utilizaron como un componente clave del clasificador ConBiBER.

Matriz de Confusión

Una tabla utilizada para evaluar el desempeño del clasificador ConBiBER. Visualiza el número de Verdaderos Positivos (afirmaciones de contagio predichas correctamente), Verdaderos Negativos (comentarios generales predichos correctamente), Falsos Positivos (comentarios generales clasificados erróneamente como contagios) y Falsos Negativos (afirmaciones de contagio que el modelo no detectó).

Exactitud (Accuracy)

Métrica que mide el porcentaje total de predicciones correctas (tanto afirmaciones de contagio como comentarios generales) que realizó el clasificador ConBiBER.

Precisión (Precision)

Mide la calidad de las predicciones positivas del clasificador. De todas las publicaciones que el modelo ConBiBER identificó como "afirmaciones de contagio", esta métrica indica qué porcentaje realmente lo eran.

Sensibilidad (Recall)

Mide la capacidad del clasificador para encontrar todos los casos positivos. De todas las publicaciones que realmente eran "afirmaciones de contagio".

Especificidad (Specificity)

Mide la capacidad del clasificador para identificar correctamente los casos negativos.

F1-Score

La media armónica entre la Precisión y la Sensibilidad.

AUC-ROC (Área bajo la Curva ROC)

Métrica que mide la "capacidad discriminativa" global del clasificador ConBiBER. Evalúa qué tan bien puede el modelo "distinguir entre afirmaciones personales de contagio y comentarios generales". Un valor cercano a 1.0 indica un mejor "equilibrio óptimo entre sensibilidad y especificidad".

RMSE (Raíz del Error Cuadrático Medio)

Métrica utilizada para "validar" el modelo predictivo de Gompertz. Mide la magnitud promedio de los errores de predicción (la diferencia entre las "menciones observadas" y las "menciones predichas"), dando una penalización mayor a los errores más grandes. Un RMSE más bajo indica un mejor ajuste del modelo.

MAE (Error Absoluto Medio)

Métrica utilizada junto con el RMSE para "cuantificar el grado de ajuste" del modelo Gompertz. Mide el promedio de las diferencias absolutas entre las predicciones y los valores reales, lo que la hace más fácil de interpretar en la escala de los datos.

Correlación de Pearson (r)

Una medida estadística utilizada en la tesis para dos propósitos clave:

1) Para validar la "sincronicidad temporal" entre las afirmaciones de contagio clasificadas por ConBiBER y los "casos oficiales reportados por la Secretaría de Salud".
2) Para comparar los niveles de riesgo del "índice geoespacial" con el "semáforo epidemiológico oficial" en diferentes periodos.