UNIVERSIDAD AUTÓNOMA DE QUERÉTARO
FACULTAD DE INGENIERÍA

PROTOCOLO DE TESIS

UNIVERSIDAD AUTÓNOMA DE QUERÉTARO
FACULTAD DE INGENIERÍA

PROTOCOLO DE TESIS

Ingeniería Física

Evaluación de un Modelo Basado en Transformers para el Pronóstico Temporal de Concentraciones de Material Particulado Coarse (PMCO)

Luis Eduardo Mauricio Álvarez
258957

Dr. Marco Antonio Aceves Fernández

12/06/2023

1. RESUMEN

Los preocupantes efectos adversos de la contaminación atmosférica en la salud y el medio ambiente han impulsado la monitorización y control efectivo del material particulado que se encuentra suspendido en el ambiente. Esta tesis explora la aplicación de un modelo avanzado basado en Transformers para predecir las concentraciones de Material Particulado Coarse (PMCO).

Aunque los modelos de Transformers han probado su valía en el procesamiento de lenguaje natural, en los últimos años han mostrado resultados prometedores en la predicción de series temporales. Esta investigación se centra en su eficacia para predecir las concentraciones de PMCO. El objetivo principal es evaluar el potencial de estos modelos como herramientas precisas y eficientes de predicción, contribuyendo así al desarrollo de estrategias para controlar la contaminación atmosférica y proteger la salud y el medio ambiente.

2. ANTECEDENTES Y/O FUNDAMENTACIÓN TEÓRICA

2.1. Contaminación del Aire y Material Particulado

2.1.1. Definición y Clasificación de la Contaminación del Aire

La contaminación del aire es un problema mundial que afecta a la población generando enfermedades y altos costos en el sector de la salud como en el medio ambiente (European Environment Agency [EEA], 2023). Los contaminantes atmosféricos pueden provenir de diversas fuentes, como la industria, el transporte, la agricultura y la quema de combustibles fósiles. Estos contaminantes pueden afectar la calidad del aire en diferentes áreas, tanto en espacios interiores como exteriores (World Health Organization [WHO], 2023).

Clasificación de los contaminantes del aire:

a) **Las fuentes principales**: Esta categoría incluye las grandes instalaciones industriales, como las centrales eléctricas, refinerías de petróleo, plantas químicas y de fertilizantes, y otras plantas industriales (Manisalidis et al., 2020).
b) **Las fuentes de área interior**: Esta categoría incluye actividades que ocurren dentro de edificios, como las tintorerías, las tiendas de impresión y las estaciones de servicio (Manisalidis et al., 2020).
c) **Las fuentes móviles**: Esta categoría incluye los vehículos, como automóviles, camiones, trenes y aviones. Estos vehículos emiten gases de escape y partículas finas que pueden tener efectos perjudiciales en la salud humana y el medio ambiente (Manisalidis et al., 2020).
d) **Las fuentes naturales**: Esta categoría incluye desastres naturales, como incendios forestales, erosión volcánica, tormentas de polvo y quema agrícola (Manisalidis et al., 2020).

La contaminación del aire se determina por la presencia de contaminantes en el aire en grandes cantidades por periodos largos. Los contaminantes del aire incluyen partículas dispersadas, hidrocarburos, CO, CO2, NO, NO2, SO3, entre otros (Manisalidis et al., 2020).

2.1.2. Definición y Características de Material Particulado

El Material Particulado (PM) es una combinación de sólidos y líquidos en forma de gotas. Algunas partículas se emiten directamente, mientras que otras se forman cuando los contaminantes emitidos por diversas fuentes reaccionan en la atmósfera (Agencia de Protección Ambiental de Estados Unidos, 2022). Las partículas contaminantes vienen en diferentes tamaños, siendo aquellas más pequeñas que 10 micrómetros capaces de ingresar a nuestros pulmones y causar serios problemas de salud (European Environment Agency, 2020).

2.1.3. Clasificación de material particulado según tamaño

El material particulado se clasifica en función de su tamaño: PM10, PM2.5 y PMCO.

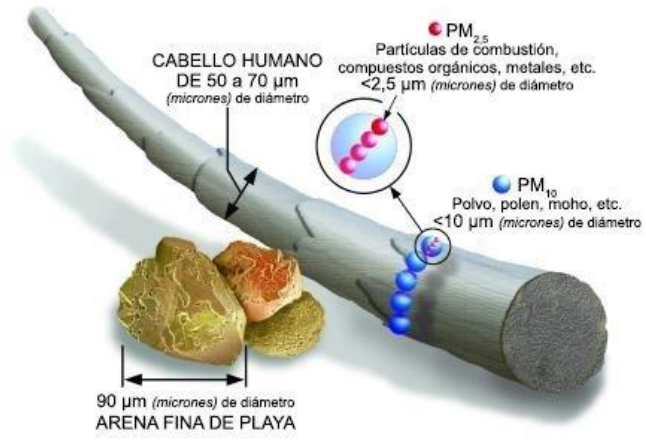

Figura 1. Comparación de tamaño de las partículas de PM
(Agencia de Protección Ambiental de Estados Unidos, 2022).

- PM10: se refiere a partículas menores o iguales a 10 micrómetros de diámetro (Agencia de Protección Ambiental de Estados Unidos, 2022).
- PM2.5: se refiere a partículas menores o iguales a 2.5 micrómetros de diámetro (Agencia de Protección Ambiental de Estados Unidos, 2022).
    - ¿Qué son 2,5 micrómetros? Para ponerlo en perspectiva, considere un solo cabello de su cabeza. En promedio, el cabello humano mide aproximadamente 70 micrómetros de diámetro, lo que lo hace 30 veces más grande que la partícula fina más grande (Agencia de Protección Ambiental de Estados Unidos, 2022).

- PMCO: se refiere a partículas con un diámetro entre 2.5 y 10 micrómetros (Preiss, P., & Roos, J., 2013).

En resumen las escalas quedan en referencia a la siguiente imagen:

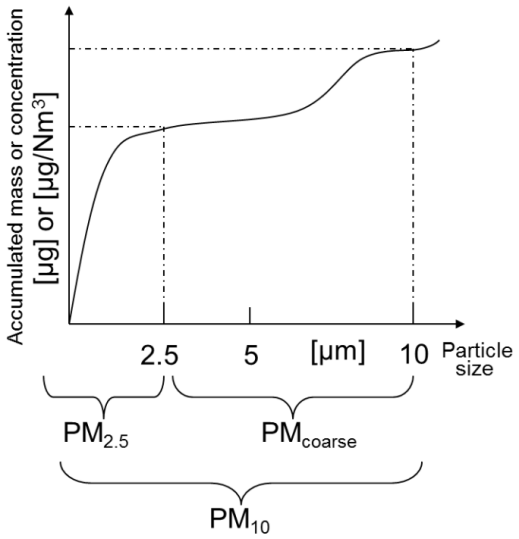

Figura 2. Comparativa entre tamaños de diversas categorías de material particulado (Preiss, P., & Roos, J., 2013).

2.1.4. Importancia de monitorear y controlar PM

El monitoreo y control de las partículas en suspensión en el aire, conocidas como PM10 y PM2.5, es de suma importancia debido a los significativos impactos que tienen en la salud humana (Organización Mundial de la Salud, 2022). Las partículas PM10 son las más grandes y provienen de fuentes como el polvo de carreteras, construcciones, minas, entre otros (Organización Mundial de la Salud, 2022). Aunque estas partículas pueden causar irritación en los ojos, la nariz y la garganta, son las partículas más pequeñas, PM2.5, las que representan un riesgo mayor (Centers for Disease Control and Prevention, 2023).

Los efectos en la salud relacionados con la exposición a la contaminación por partículas incluyen:

- Irritación de los ojos, los pulmones y la garganta.
- Dificultad para respirar.
- Exacerbación de los síntomas de asma.
- Aumento del riesgo de cáncer de pulmón.
- Problemas relacionados con el nacimiento, como bajo peso al nacer.

Dada la gravedad de estos efectos en la salud, es imperativo que se implementen medidas de monitoreo y control de las partículas PM10 y PM2.5. Esto implica vigilar de cerca los niveles de estas partículas en el aire y tomar acciones para reducir su presencia cuando sea necesario (Organización Mundial de la Salud, 2022).

2.2. Estaciones de Monitoreo PM de la CDMX

El Sistema de Monitoreo Atmosférico (SIMAT) mide permanentemente los principales contaminantes en más de 40 sitios de monitoreo en la Ciudad de México y la zona conurbada del Estado de México. El monitoreo tiene como objetivos evaluar el cumplimiento de las Normas Oficiales Mexicanas, cuantificar la exposición de la población, informar y prevenir sobre los riesgos, activar o desactivar alertas de emergencia y generar datos confiables para la evaluación y seguimiento de las estrategias de gestión de la calidad del aire (SEDEMA, 2023).

En términos operativos, el Sistema de Monitoreo Atmosférico en su conjunto está conformado por cuatro subsistemas (RAMA, REDMA, REDMET y REDDA) (SEDEMA, 2023), de los cuales describiremos a RAMA:

- La **Red Automática de Monitoreo Atmosférico** (RAMA) utiliza equipos continuos para la medición de dióxido de azufre, monóxido de carbono, dióxido de nitrógeno, ozono, PM10, PM2.5 y PMCO. Está integrada por 34 estaciones de monitoreo y cuenta con un laboratorio para el mantenimiento y calibración de los equipos de monitoreo.

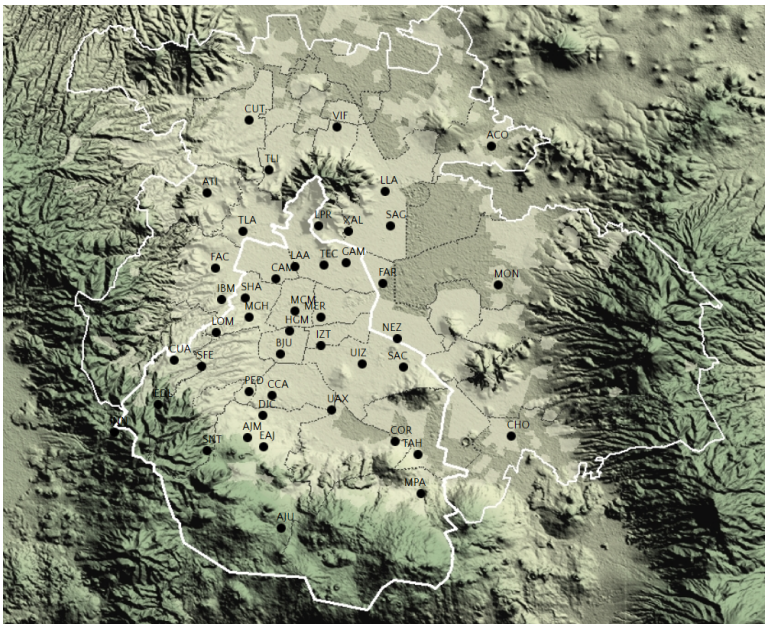

Figura 3. Ubicación de las estaciones de monitoreo(SEDEMA, 2023)

2.3. Redes Neuronales en el Pronóstico de Series Temporales

2.3.1. Concepto y arquitectura de redes neuronales artificiales

Las redes neuronales son una arquitectura que imita el funcionamiento de las redes neuronales biológicas del cerebro humano. Aprenden identificando patrones y se usan en tareas como predicción, clasificación y regresión. Sin embargo, tienen algunas desventajas, como la necesidad de convertir los datos a formato numérico y la necesidad de muchos datos para obtener buenas predicciones, especialmente en el caso de redes neuronales profundas (Aceves-Fernández, 2021).

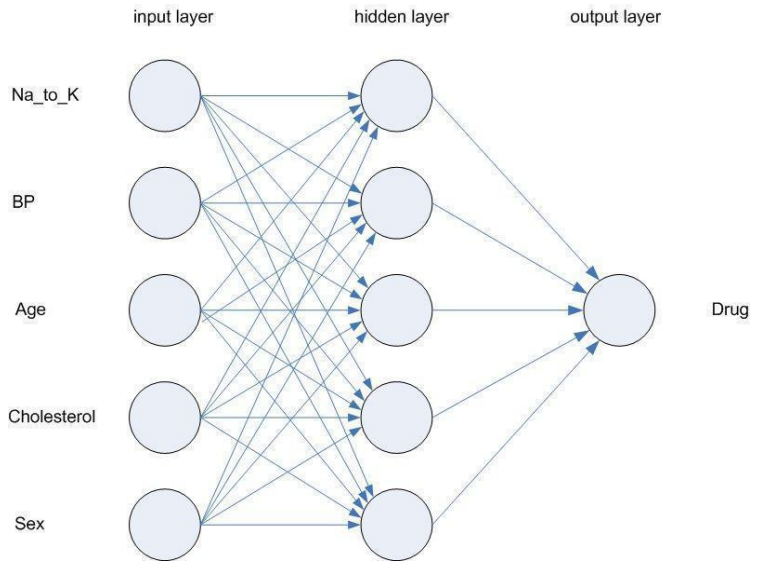

Figura 4. Esquema de Red Neuronal (IBM Documentation, 2021).

Para construir y entrenar una red neuronal, se necesita una capa de entrada con N neuronas para los datos de entrada de N dimensiones y una capa de salida con M neuronas para los M tipos de datos de entrenamiento. Las capas ocultas, que determinan la topología de la red, pueden variar en número. Una red sencilla tiene dos capas ocultas, mientras que una red profunda tiene muchas más. Para clasificar, es necesario preparar y etiquetar los datos de entrenamiento. Luego, la red neuronal se entrena para reducir el error entre su salida y la salida predicha hasta que se alcanza un determinado umbral (Aceves-Fernández, 2021) .

2.3.2. Perceptrón

Un perceptrón es un modelo matemático y computacional simple basado en una neurona biológica. Fue desarrollado por Frank Rosenblatt en 1957 (IBM Documentation, 2021) y es considerado como una de las primeras formas de redes neuronales artificiales. El perceptrón funciona como un clasificador lineal binario, es

decir, puede separar conjuntos de datos linealmente separables en dos clases distintas (Rokach & Maimon, 2007). Matemáticamente, la salida del perceptrón (y) se calcula como:

$$y = f(\sum(w_i * x_i) + b) \quad (1)$$

donde w_i son los pesos, x_i son las entradas, b es el sesgo y f es la función de activación, típicamente la función escalón unitario.

El perceptrón es la estructura básica de la red neuronal, como se muestra en la figura.

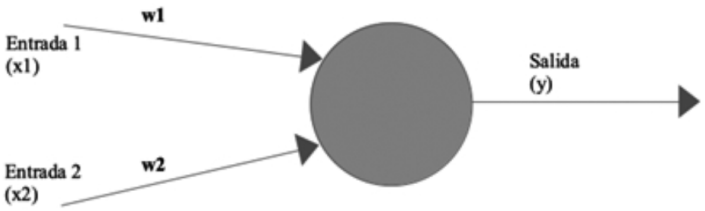

Figura 5. Topología de una red neuronal simple (Aceves-Fernández, 2021)

Aunque el perceptrón es una red neuronal simple, las redes más complejas se basan en versiones con más capas y neuronas por capa del perceptrón, tales como redes convolutivas, recurrentes, generativas, de celda de memoria, de creencia profunda, transformadores y codificadores (Aceves-Fernández, 2021).

2.3.3. Redes Neuronales Recurrentes (RNN)

Una red neuronal recurrente (RNN) es una clase de red neuronal artificial diseñada para manejar datos secuenciales o series temporales (Rokach & Maimon, 2007). Estos algoritmos de aprendizaje profundo se aplican frecuentemente en situaciones ordenadas o temporales, como la traducción automática, el procesamiento del lenguaje natural (NLP), el reconocimiento de voz y la generación de subtítulos para imágenes; y se encuentran integrados en aplicaciones conocidas como Siri, búsqueda por voz y Google Translate. Al igual que las redes neuronales de propagación hacia adelante y las redes neuronales convolucionales (CNN), las RNN emplean datos de entrenamiento para aprender. Sin embargo, se diferencian por su capacidad de "memoria", ya que utilizan información de entradas previas en los datos de entrada actuales y en los resultados obtenidos (Rokach & Maimon, 2007). La salida de una RNN se calcula como:

$$h_t = f(W_{hh} * h_{t-1} + W_{xh} * x_t + b_h) \quad (2)$$

donde h_t es el estado oculto en el tiempo t, W_{hh} y W_{xh} son matrices de pesos, x_t es la entrada en el tiempo t, y b_h es el sesgo.

A diferencia de las redes neuronales profundas convencionales, que asumen que los datos de entrada y los resultados son independientes entre sí, los resultados de las RNN dependen de los elementos previos en la secuencia. Aunque los eventos futuros también serían útiles para determinar los resultados de una secuencia específica, las RNN unidireccionales no pueden considerar estos eventos en sus predicciones (Goodfellow et al., 2016).

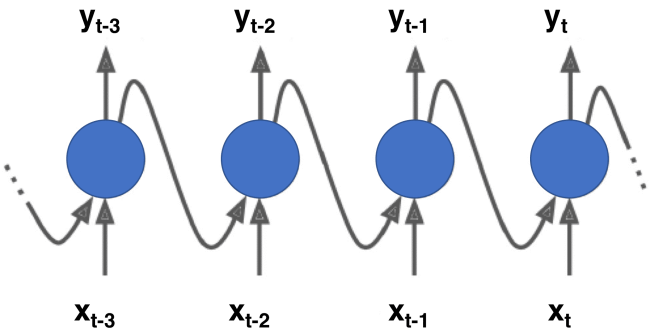

Figura 6. Esquema de Red Neuronal Recurrente (RNN) (Goodfellow et al., 2016).

2.3.4. Long Short-Term Memory (LSTM)

Las redes LSTM (Long Short-Term Memory) son un tipo especial de red neuronal recurrente (RNN) diseñadas para abordar el problema del desvanecimiento del gradiente en las RNNs tradicionales. Este problema ocurre cuando las redes tratan de aprender dependencias temporales a largo plazo en secuencias de datos (de la Fuente et al., 2022). Las LSTM introducen unidades de memoria llamadas celdas, con puertas de entrada, olvido y salida para controlar el flujo de información. Las ecuaciones que definen una LSTM son:

$$i\_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \quad (3)$$

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \quad (4)$$

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c * [h_{t-1}, x_t] + b_c) \quad (6)$$

$$h_t = o_t \odot tanh(c_t) \quad (7)$$

donde i_t, f_t y o_t son las puertas de entrada, olvido y salida respectivamente, y c_t y h_t son el estado de la celda y la salida.

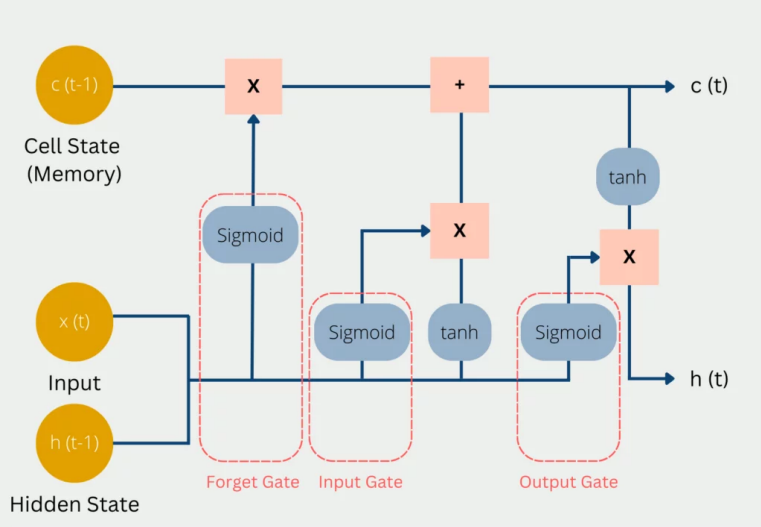

Figura 7. Estructura de Red Neuronal Long-Short Term (LSTM) (de la Fuente et al., 2022).

Las LSTM se componen de unidades de memoria llamadas celdas, que permiten almacenar información a lo largo del tiempo y regular el flujo de información mediante estructuras llamadas puertas. Estas puertas deciden cuándo agregar, modificar o eliminar información en la celda de memoria, lo que facilita el aprendizaje de dependencias a largo plazo en los datos secuenciales. Las LSTM son ampliamente utilizadas en tareas de procesamiento del lenguaje natural, traducción automática, reconocimiento de voz y predicción de series temporales, entre otras aplicaciones (Lindemann et al., 2021).

2.3.5. Secuencia a Secuencia (Seq2Seq)

El modelo Seq2Seq es un enfoque de aprendizaje automático utilizado en tareas como traducción automática y resumen de texto. Funciona analizando partes de las entradas y salidas sin considerar información futura, lo que lo hace eficiente en procesar lenguaje natural al entender el contexto (Sánchez Gozalo, J., 2020).

Consiste en un codificador que transforma la secuencia de entrada en un vector de contexto (c), y un descodificador que genera la secuencia de salida a partir de este contexto. El codificador y el descodificador suelen ser RNN o LSTM (Sánchez Gozalo, J., 2020).

$$c = f_{enc}(x_1, x_2, \dots, x_n) \qquad (8)$$

$$y_t = f_{dec}(y_{t-1}, c) \quad (9)$$

donde x_i son los elementos de la secuencia de entrada, y y_t son los elementos de la secuencia de salida.

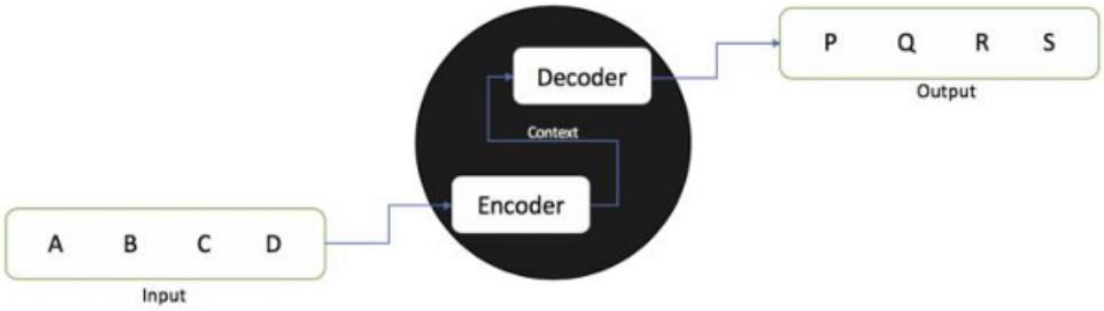

Figura 8. Esquema general del modelo Seq2Seq (Sánchez Gozalo, J., 2020).

En la figura que precede, se ilustra que un modelo Seq2Seq es un modelo que recibe una secuencia de elementos, como palabras, letras o series temporales, y produce una secuencia distinta de elementos.

2.4. Modelo Transformer

2.4.1. Orígenes y Evolución

Los Transformers fueron introducidos por Vaswani et al. en 2017 en su trabajo "Attention is All You Need"(Vaswani et al., 2017). Los Transformers se basan en el mecanismo de atención de "auto-atención" o "auto-atención de múltiples cabezas", que permite a la red neuronal enfocarse en diferentes partes de la entrada simultáneamente. Esta arquitectura ha demostrado ser muy eficaz para tareas de procesamiento del lenguaje natural (NLP), como la traducción automática, el resumen de texto, y la generación de texto (Vaswani et al., 2017) .

En el campo de la predicción de series de tiempo, los Transformers también han demostrado su potencial. Con su capacidad para capturar dependencias a largo plazo en las secuencias de datos y manejar eficientemente secuencias largas, podrían ofrecer ventajas significativas para la predicción de series de tiempo. Según un estudio reciente titulado "Are Transformers Effective for Time Series Forecasting?"(Zeng, Chen, Zhang, & Xu, 2022), parece que los Transformers pueden ser efectivos para estas tareas. Sin embargo, es importante destacar que la eficacia de los Transformers puede variar en función de la naturaleza específica de la serie de tiempo y de los detalles de implementación del modelo (Zeng, Chen, Zhang, & Xu, 2022).

2.4.2. Arquitectura

El modelo Transformer establece una arquitectura innovadora, cuya configuración se ilustra en la figura que sigue:

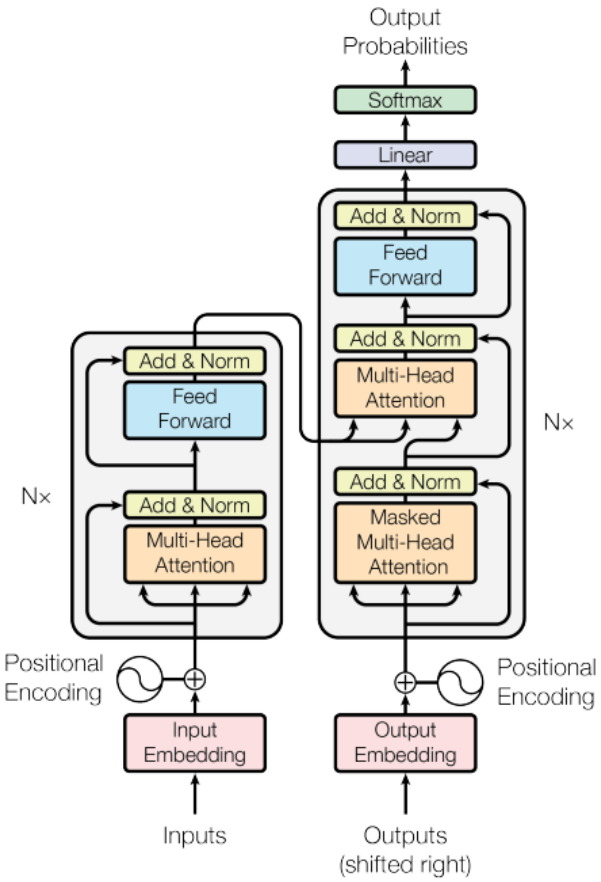

Figura 8. Transformers - Arquitectura del modelo (Vaswani et al., 2017).

El esquema Transformer se fundamenta en dos elementos clave: un módulo de codificación y otro de decodificación (Sánchez Gozalo, J., 2020). La representación gráfica que se observa en la parte superior ilustra el módulo de codificación hacia la izquierda y el módulo de decodificación hacia la derecha.

Los módulos de codificación tienen una estructura homogénea, aunque es posible que difieran en términos de pesos asignados. Además, cada módulo de codificación se subdivide en dos secciones, como se puede observar en la gráfica siguiente:

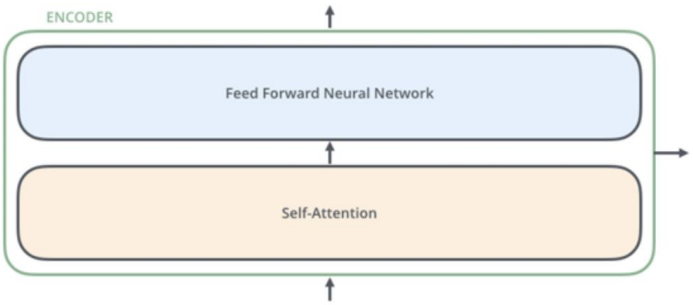

Figura 9. Codificador del modelo Transformer (Sanchez-Gonzalo, J.,2020).

En su aplicación, el módulo de codificación procesa la secuencia completa de una sola vez, en lugar de procesar cada palabra de forma incremental. Esto confiere a estos modelos una mayor potencia, ya que pueden retener información global en un solo instante (Sánchez Gozalo, J., 2020).

La salida generada por el módulo de decodificación se retroalimenta como entrada, procesando la secuencia completa en un solo instante, de manera similar a como se hace con el módulo de codificación. Durante el inicio del entrenamiento, cuando no hay entradas disponibles para el decodificador, se emplea un token especial de inicio (Sánchez Gozalo, J., 2020).

La gráfica subsiguiente ilustra la operativa global del esquema Transformer para dos entradas dadas:

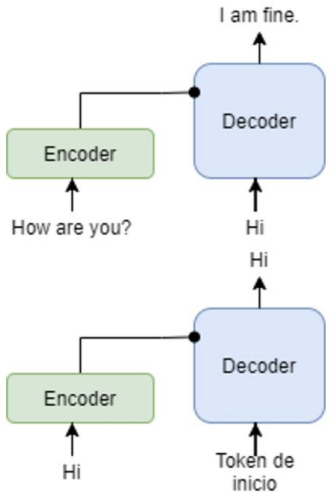

Figura 10. Funcionamiento de las entradas del modelo de Transformer (Sanchez-Gonzalo, J.,2020).

Para entender el rol de cada unidad dentro de la arquitectura Transformer, introducimos las siguientes explicaciones (Sánchez Gozalo, J., 2020) (Vaswani et al., 2017):

- **Codificación Posicional**: Dado que cada palabra en una oración es procesada en paralelo por la serie de módulos de codificación/decodificación en el Transformer, el modelo no tiene inherente un sentido de orden o posición para cada palabra. Por lo tanto, se incorporó este proceso para asignar una posición a cada token.
- **Sumar y Normalizar**: Tanto las entradas como las salidas de la capa de atención de múltiples cabezales (Multi-Head Attention), así como de la red de propagación directa por posición, pasan por dos capas de tipo "sumar y normalizar", compuestas por una capa que amalgama los resultados de la capa anterior y, posteriormente, una capa de normalización. Esto permite que el esquema Transformer procese la secuencia completa en un solo instante, abordando la limitación del modelo Seq2Seq y evitando demoras en el procesamiento realizado por el modelo.
- **Lineal**: Es una red neuronal básica (sin capa oculta) y completamente conectada que tiene como finalidad transformar el vector resultante de la serie de módulos de decodificación en un vector de mayor dimensión, llamados vector logits.
- **Propagación hacia Adelante (FF)**: Es la red neuronal convencional que consta de varias capas (entrada, salida y al menos una capa oculta) donde la señal de entrada se propaga hacia adelante y puede ser entrenada.
- **Softmax**: Es un algoritmo que calcula la probabilidad de que cada palabra esté en la posición asociada.

El componente de codificación se alimenta a partir de la capa de incrustación (embedding), mientras que la función de atención de múltiples cabezales (Multi-Head Attention) se encarga de discernir las variadas interrelaciones existentes entre las palabras que conforman esa entrada. Acto seguido, las capas denominadas "Sumar y Normalizar" se encargan de consolidar cada token en relación a su estado inicial y normalizan sus valores antes de proceder a la capa de Propagación hacia Adelante (FF). Una vez combinados nuevamente en la última capa de "Sumar y Normalizar", el modelo se entrena a través de la diferenciación de la composición de funciones de la siguiente unidad. En cuanto al componente de decodificación, su cometido principal es la realización de la predicción final (Vaswani et al., 2017).

Al observar los elementos internos del componente de codificación o decodificación, se podría pensar en ellos como si fueran dos cajas enigmáticas: la de atención y la de propagación hacia adelante. La función de atención se centra en capturar la esencia del significado, mientras que el componente de propagación hacia adelante se ocupa de dar forma a la solución que el modelo específico generará (Vaswani et al., 2017).

Con base en lo anterior, el componente de decodificación integra dos capas de atención, y su tarea consiste en manejar dos entradas de manera simultánea en el componente de propagación hacia adelante. La intención del componente de decodificación radica en la utilización de resultados anteriores para el propósito de su entrenamiento (Vaswani et al., 2017).

3. DESCRIPCIÓN DEL PROBLEMA

La contaminación del aire por partículas PMCO (Material Particulado Coarse) es un fenómeno creciente que amenaza la salud humana y el equilibrio del medio ambiente. Según el Centers for Disease Control and Prevention (2023), la exposición a partículas PMCO puede aumentar el riesgo de enfermedades cardiovasculares, enfermedades respiratorias, y tener efectos negativos en el embarazo. Además, la Agencia de Protección Ambiental de Estados Unidos (2022) señala que la contaminación del aire compromete la calidad del agua y la biodiversidad. Para contrarrestar estos efectos, es imperativo monitorear y controlar las concentraciones de PMCO. Sin embargo, las herramientas de pronóstico actuales, basadas en modelos de series temporales, pueden presentar limitaciones en precisión y eficiencia. Los modelos basados en Transformers han mostrado un alto rendimiento en tareas de procesamiento de lenguaje natural, pero su aplicación en el pronóstico de concentraciones de PMCO es un campo aún poco explorado. Esto plantea la necesidad de investigar si los modelos basados en Transformers pueden ser aplicados de manera efectiva para mejorar la precisión y eficiencia en la predicción de las concentraciones de PMCO.

4. JUSTIFICACIÓN

Esta investigación se justifica en la urgencia de desarrollar herramientas de pronóstico más precisas y eficientes para el monitoreo de concentraciones de PMCO, dada la amenaza que representan para la salud humana y el medio ambiente. Al explorar la aplicación de modelos basados en Transformers en el pronóstico de concentraciones de PMCO, esta investigación busca contribuir con una metodología novedosa que podría superar las limitaciones de los enfoques tradicionales. El desarrollo y validación de un modelo basado en Transformers para el pronóstico temporal de concentraciones de PMCO podrían resultar en una herramienta valiosa para los responsables de la toma de decisiones en políticas de control de calidad del aire. Esto tendría un impacto directo en la protección de la salud pública y la conservación del medio ambiente, áreas que ya han sido identificadas como críticamente afectadas por las partículas PMCO (Centers for Disease Control and Prevention, 2023; Agencia de Protección Ambiental de Estados Unidos, 2022). Además, esta investigación enriquecerá el campo académico al expandir el conocimiento sobre las aplicaciones de los modelos basados en Transformers fuera del ámbito del procesamiento de lenguaje natural.

5. HIPÓTESIS

Al utilizar datos históricos de concentraciones de Material Particulado Coarse (PMCO) y características temporales como variables de entrada, los modelos basados en Transformers, que han demostrado ser efectivos en capturar dependencias a largo plazo en tareas de procesamiento de lenguaje natural, serán capaces de generar predicciones precisas de concentraciones futuras de PMCO como variable de salida. Esto se basa en el supuesto de que existen patrones temporales en las concentraciones de PMCO que pueden ser capturados por estos modelos.

6. OBJETIVOS

Objetivo General:

- Evaluar la eficacia de un modelo basado en Transformers en el pronóstico temporal de las concentraciones de Material Particulado Coarse (PMCO) con el fin de determinar su potencial como herramienta de predicción de alta precisión y eficiencia.

Objetivos Específicos:

- Recopilar y preparar un conjunto de datos históricos de concentraciones de Material Particulado Coarse (PMCO) y características temporales relevantes que servirán como entrada para el modelo.
- Realizar una revisión bibliográfica exhaustiva sobre la contaminación del aire por partículas PMCO, sus efectos sobre la salud y el medio ambiente, y los métodos existentes para su monitoreo y pronóstico.
- Evaluar el rendimiento del modelo basado en Transformers utilizando métricas de evaluación apropiadas.
- Identificar y proponer posibles mejoras y extensiones al modelo, así como sugerir áreas de investigación futura relacionadas con la aplicación de modelos basados en Transformers en pronósticos ambientales.

7. METODOLOGÍA

7.1. Materiales

1. **Computadora con acceso a Internet**: Esto es esencial para acceder a Google Colab, que es una plataforma de Jupyter Notebook en la nube que no requiere configuración y te brinda acceso a recursos de cómputo gratuitos, incluidas GPUs.
2. **Google Colab**: Usarás Google Colab para escribir y ejecutar el código. Asegúrate de tener una cuenta de Google para poder acceder a este servicio.
3. **Librerías de Python**:
    a. **Pandas**: Para la manipulación de datos y análisis. Esto será útil en la etapa de recopilación de datos y preprocesamiento.
    b. **Numpy**: Para realizar operaciones matemáticas y manipular matrices.
    c. **PyTorch**: Es la librería que usarás para construir y entrenar el modelo de Transformers como se menciona en la metodología.
    d. **Hugging Face Transformers**: Es una librería construida sobre PyTorch y es fundamental para trabajar con la arquitectura de Transformers.
    e. **Scikit-learn**: Es útil para la normalización de datos y también para evaluar el modelo usando métricas como RMSE.
    f. **FancyImputation**: Para datos faltantes con el método MICE.
    g. **Matplotlib y Seaborn**: Para visualización de datos y análisis de resultados.
    h. **Datos**: Necesitarás acceder a la base de datos de la Red Automática de Monitoreo Atmosférico (RAMA) de la Ciudad de México para extraer los datos de calidad del aire. Asegúrate de tener los permisos necesarios para acceder a estos datos.

7.2. Metodología

La metodología propuesta para este proyecto es la siguiente:

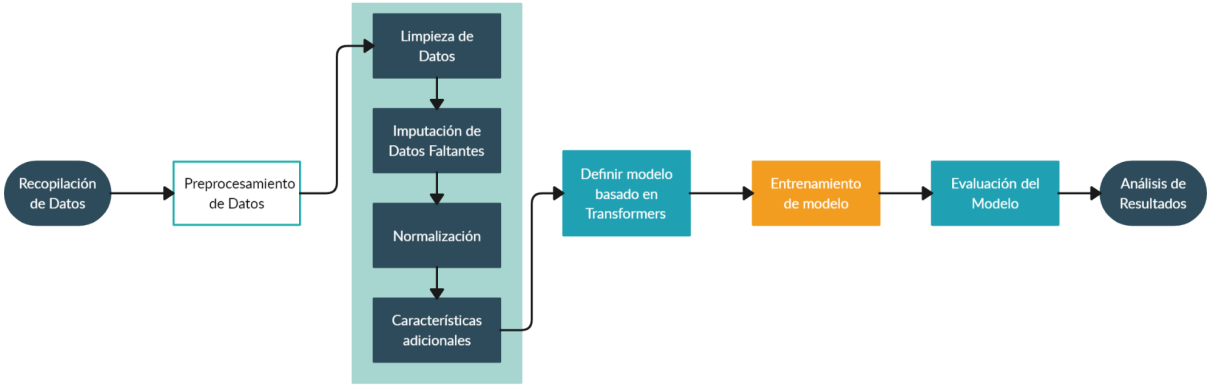

Figura 11. Diagrama de bloques de la metodología propuesta para este proyecto de investigación.

Recopilación de Datos: Los datos de calidad del aire se extraerán de la base de datos de la Red Automática de Monitoreo Atmosférico (RAMA) de la Ciudad de México (SEDEMA, 2023). Específicamente, se centrará en las concentraciones de partículas PMCO registradas de 2012 a 2022 en las cuatro estaciones más optimas para trabajar; entre algunas de sus cualidades seran las que menos datos faltantes contenga.

Preprocesamiento de Datos: Se llevará a cabo una serie de pasos para preparar los datos para el entrenamiento del modelo (Wilson, 2021):

- **Limpieza de Datos**: Se realizará un proceso de limpieza de datos para identificar y manejar cualquier error o anomalía en el conjunto de datos. En este caso la base de la Red Automática de Monitoreo Atmosférico (RAMA) de la Ciudad de México cuenta con datos faltantes representados con los numeros -99 que ocuparán ser tratados.
- **Imputación de Datos Faltantes**: Se emplea el método MICE (Multiple Imputation by Chained Equations) para imputar cualquier dato faltante en las series de tiempo.
- **Normalización**: Los datos se normalizaron para tener una media de cero y una desviación estándar de uno. Esta es una práctica común para facilitar el entrenamiento de los modelos de aprendizaje automático.
- **Características adicionales**: Se generarán características adicionales que podrían ser relevantes para el pronóstico, como tendencias estacionales y promedios móviles.

Desarrollo y Entrenamiento del Modelo: Se utilizará la arquitectura de los Transformers, y con ayuda de las librerías que ofrece Hugging Face (Hugging Face, 2023) y PyTorch (PyTorch, 2023) para el pronóstico de series de tiempo. Para ello:

- Se definirá la arquitectura del modelo Transformer, que incluye capas de auto-atención, capas de punto de atención y capas de feed-forward. El modelo se construirá utilizando la biblioteca PyTorch, que proporciona soporte para la arquitectura de Transformers (Hugging Face, 2023).
- Se entrenará el modelo utilizando un algoritmo de optimización Adam. Durante este proceso, se ajustarán los hiperparámetros del modelo para maximizar su rendimiento (Hugging Face, 2023).

Evaluación del Modelo: La eficacia del modelo se evaluará utilizando técnicas de validación cruzada. Los datos se dividirán en conjuntos de entrenamiento, validación y prueba. La métrica de rendimiento será el Error Cuadrático Medio (RMSE), que es comúnmente utilizado en tareas de pronóstico. Además, se comparará el rendimiento del modelo de Transformers con el de los métodos convencionales de pronóstico de la calidad del aire (Wackerly, et al., 2017).

Análisis de Resultados: Se realizará una evaluación exhaustiva del modelo desarrollado, analizando su capacidad para predecir densidades en diversos lapsos temporales y la utilidad de estas predicciones para la toma de decisiones. Se discutirán las limitaciones del estudio y se propondrán posibles mejoras y trabajos futuros. Este análisis proporcionará una base sólida para la validación de la hipótesis del estudio.

8. CRONOGRAMA

Mes	Actividad
Julio	**Investigación Teórica**: Realizar una revisión de la literatura relevante y estudiar los conceptos básicos de la arquitectura Transformer y la predicción de la calidad del aire
	Recopilación de Datos: Adquirir los datos necesarios de la Red Automática de Monitoreo Atmosférico (RAMA) de la Ciudad de México.
Agosto	**Preprocesamiento de Datos**: Limpieza de datos, manejo de valores nulos, imputación de datos faltantes y normalización de datos.
	Diseño del Modelo de Algoritmo de Atención Transformer: Comenzar a diseñar y desarrollar el modelo, basándose en la investigación teórica realizada.
Septiembre	**Diseño y Optimización del Modelo de Algoritmo de Atención Transformer**: Continuar con el desarrollo del modelo y comenzar a optimizarlo.
	I**ngeniería de Características**: Crear nuevas características que puedan mejorar el rendimiento del modelo.
Octubre	**Entrenamiento del Modelo**: Entrenar el modelo utilizando los datos preprocesados y las características desarrolladas.
	Evaluación del Modelo: Utilizar técnicas de validación cruzada y el Error Cuadrático Medio (MSE) como métrica de rendimiento para evaluar el modelo.
Noviembre	**Ajuste del Modelo**: Basado en los resultados de la evaluación, realizar ajustes en el modelo para mejorar su rendimiento.
	Análisis de Resultados: Analizar los resultados obtenidos y compararlos con los resultados de modelos existentes.
Diciembre	**Escritura de la Tesis**: Recopilar todos los hallazgos y escribir la tesis.
	Revisión y Presentación: Revisar la tesis y prepararse para la presentación.

9. ALCANCE DEL PROYECTO

Desarrollo de un Modelo Predictivo: El proyecto se enfocará en el desarrollo y evaluación de un modelo basado en Transformers para predecir concentraciones de Material Particulado Coarse (PMCO). No se abordarán otros contaminantes atmosféricos.

10. RESULTADOS ESPERADOS

Modelo Optimizado de Transformers: Se espera desarrollar un modelo basado en Transformers altamente optimizado que demuestre un rendimiento competitivo en términos de precisión y eficiencia.

Mejor Comprensión de los Datos: A través del análisis de cómo el modelo procesa tendencias y patrones, se espera obtener una comprensión más profunda de los datos y cómo pueden influir en las concentraciones de PMCO.

Marco de Evaluación Robusta: Se espera desarrollar un marco de evaluación que permita una comparación objetiva y rigurosa entre diferentes modelos predictivos.

11. REFERENCIAS BIBLIOGRÁFICAS

Agencia de Protección Ambiental de Estados Unidos. (2022). Conceptos básicos sobre el material particulado (PM, por sus siglas en inglés). Recuperado de https://espanol.epa.gov/espanol/conceptos-basicos-sobre-el-material-particulado-pm-por-sus-siglas-en-ingles

Centers for Disease Control and Prevention. (2023). Particle pollution. Recuperado de https://www.cdc.gov/air/particulate_matter.html

de la Fuente, A. R., Sanchez-Ramirez, E., Calderón-Alvarado, M. P., Segovia-Hernandez, J. G., & Hernández-Vargas, E. A. (2022). Development of deep learning architectures for forecasting distillation columns dynamic behavior of biobutanol purification. En Computer Aided Chemical Engineering (pp. 49–54). Elsevier.

European Environment Agency. (2020). What is particulate matter and what are its effects on human health? Recuperado de https://www.eea.europa.eu/help/faq/what-is-particulate-matter-and

Aceves-Fernández, M. (2021). Inteligencia artificial para programadores con prisa. UNIVERSO DE LETRAS.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. Londres, Inglaterra: MIT Press.

Hugging Face. (2023). Preprocessing data. Recuperado de https://huggingface.co/transformers/preprocessing.html

IBM Documentation. (2021). ¿Qué son las redes neuronales recurrentes?. Recuperado de https://www.ibm.com/es-es/topics/recurrent-neural-networks

IBM Documentation. (2021). El modelo de redes neuronales. Recuperado de: https://www.ibm.com/docs/es/spss-modeler/saas?topic=networks-neural-model

Larios, I. (s/f). Series de Tiempo. Recuperado de http://www.estadistica.mat.uson.mx/Material/seriesdetiempo.pdf

Lindemann, B., Müller, T., Vietz, H., Jazdi, N., & Weyrich, M. (2021). A survey on long short-term memory networks for time series prediction. Procedia CIRP, 99, 650–655. doi:10.1016/j.procir.2021.03.088

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: A review. Frontiers in Public Health, 8, 14. doi:10.3389/fpubh.2020.00014

Organización Mundial de la Salud. (2022). Ambient (outdoor) air pollution. Recuperado de https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

Organización Mundial de la Salud (WHO). (2023). Air pollution. Recuperado de https://www.who.int/health-topics/air-pollution#tab=tab_1

Preiss, P., & Roos, J. (2013). Global characterization factors for damage to human health due to particulate matter – based on the TM5-FASST model. Universität Stuttgart. Recuperado de https://lc-impact.eu/doc/deliverables/fine_particular_matter.pdf

Rokach, L., & Maimon, O. Z. (2007). Data mining with decision trees: Theory and applications: Theory and applications. Singapur, Singapur: World Scientific Publishing.

Sánchez Gozalo, J. (2020). Análisis del estado del arte de la generación de texto con redes neuronales mediante modelos de Transformer. Trabajo de Grado. Escuela de Ingeniería Informática de Segovia, Universidad de Valladolid. Recuperado de https://uvadoc.uva.es/bitstream/handle/10324/43030/TFG-B.%201562.pdf?sequence=1&isAllowed=y

SEDEMA. (2023). Bases de datos - Red Automática de Monitoreo Atmosférico (RAMA) [Base de datos]. Calidad del aire. http://aire.cdmx.gob.mx

SEDEMA. (2023). Dirección de Monitoreo Atmosférico. Recuperado de http://www.aire.cdmx.gob.mx/default.php?opc='ZaBhnml='

Transformer — PyTorch 2.0 documentation. (2023). Recuperado de: https://pytorch.org/docs/stable/generated/torch.nn.Transformer.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wackerly, D., Mendenhall, W., & Scheaffer, R. L. (2007). Mathematical statistics with applications (7th ed.). Florence, KY: Brooks/Cole.

Zeng, A., Chen, M., Zhang, L., & Xu, Q. Are transformers effective for time series forecasting? arXiv 2022. arXiv preprint arXiv:2205.13504.