Universidad del Valle de Guatemala
Data Science
Departamento de Ciencias de la Computación
Ciclo II, 2024

Proyecto 2

Resultados Finales

Guillermo Alfonso Furlán Estrada	Carné 20713
Roberto Francisco Ríos Morales	Carné 20979
Diego Andrés Alonzo Medinilla	Carné 20172
Diego Alejandro Perdomo Sagastume	Carné 20204

Guatemala, 18 de noviembre de 2024

Análisis exploratorio de datos

El proyecto consiste de tres diferentes bases de datos relacionales que contienen información acerca de diferentes artículos y sus contenidos. Siendo que el objetivo principal era crear un modelo capaz de captar la relación del extracto del artículo con ciertas palabras clave, fue necesario entender los contenidos de cada una de estas bases de datos de una manera conjunta. Los tres archivos que contenían la data del proyecto consistían en tres aspectos del proyecto: relaciones, extractos y entidades.

relations_df

	id	abstract_id	type	entity_1_id	entity_2_id	novel
0	0	1353340	Association	410	D007966	No
1	1	1353340	Positive_Correlation	rs74315458	D007966	Novel
2	2	1671881	Positive_Correlation	D010661	rs62514952	Novel
3	3	1671881	Positive_Correlation	D010661	rs62514953	Novel
4	4	1671881	Association	5053	D010661	No
...	...	...	...	...	...	...
4275	4275	30836660	Negative_Correlation	C009172	1278	Novel
4276	4276	30836660	Negative_Correlation	C009172	1277	Novel
4277	4277	30836660	Negative_Correlation	C009172	12825	Novel
4278	4278	30836660	Negative_Correlation	C009172	1281	Novel
4279	4279	30836660	Negative_Correlation	C009172	D007674	Novel

Tabla 1: Relaciones entre entidades y su tipo correspondiente

abstracts_df

	abstract_id	title	abstract
0	1353340	Late-onset metachromatic leukodystrophy: molec...	We report on a new allele at the arylsulfatase...
1	1671881	Two distinct mutations at a single BamHI site ...	Classical phenylketonuria is an autosomal rece...
2	1848636	Debrisoquine phenotype and the pharmacokinetic...	The metabolism of the cardioselective beta-blo...
3	2422478	Midline B3 serotonin nerves in rat medulla are...	Previous experiments in this laboratory have s...
4	2491010	Molecular and phenotypic analysis of patients ...	Eighty unrelated individuals with Duchenne mus...
...	...	...	...
395	28851297	An Ag-globin G->A gene polymorphism associated...	BACKGROUND: Increase of the expression of g-gl...
396	28883039	Disease-associated mutations in human BICD2 hy...	Bicaudal D2 (BICD2) joins dynein with dynactin...
397	29049388	An inducible mouse model of podocin-mutation-r...	Mutations in the NPHS2 gene, encoding podocin,...
398	29183288	Arginase 1 deletion in myeloid cells affects t...	BACKGROUND: (Over-)expression of arginase may ...
399	30836660	Salidroside Ameliorates Renal Interstitial Fib...	Salidroside (Sal) is an active ingredient that...

Tabla 2: Extractos de los diferentes artículos con su correspondiente ID y título

entities_df

	id	abstract_id	offset_start	offset_finish	type	mention	entity_ids
0	0	1353340	11	39	DiseaseOrPhenotypicFeature	metachromatic leukodystrophy	D007966
1	1	1353340	111	126	GeneOrGeneProduct	arylsulfatase A	410
2	2	1353340	128	132	GeneOrGeneProduct	ARSA	410
3	3	1353340	159	187	DiseaseOrPhenotypicFeature	metachromatic leukodystrophy	D007966
4	4	1353340	189	192	DiseaseOrPhenotypicFeature	MLD	D007966
...	...	...	...	...	...	...	...
3631	13631	30836660	2237	2241	GeneOrGeneProduct	TLR4	21898,7099
3632	13632	30836660	2242	2251	GeneOrGeneProduct	NF-kappaB	18033,4790
3633	13633	30836660	2256	2260	GeneOrGeneProduct	MAPK	26413,5594
3634	13634	30836660	2344	2347	ChemicalEntity	Sal	C009172
3635	13635	30836660	2398	2412	DiseaseOrPhenotypicFeature	renal fibrosis	D007674

Launchpad ⊗ 0 ⚠ 31 0

Tabla 3: Presencia de entidades, tipo de entidad y su mención

Para entender mejor la información que se tenía disponible para crear un modelo capaz de analizar y detectar la relación de ciertas entidades presentes en la información de un artículo y el contenido de este, fue necesario entender las estadísticas de cada uno de los grupos de información. Siendo que el extracto de artículo es lo que más información brinda, se consideró el punto de interés principal para entender la relación entre entidades. Con esto, se encuentra una distribución normal en el largo de los diferentes extractos.

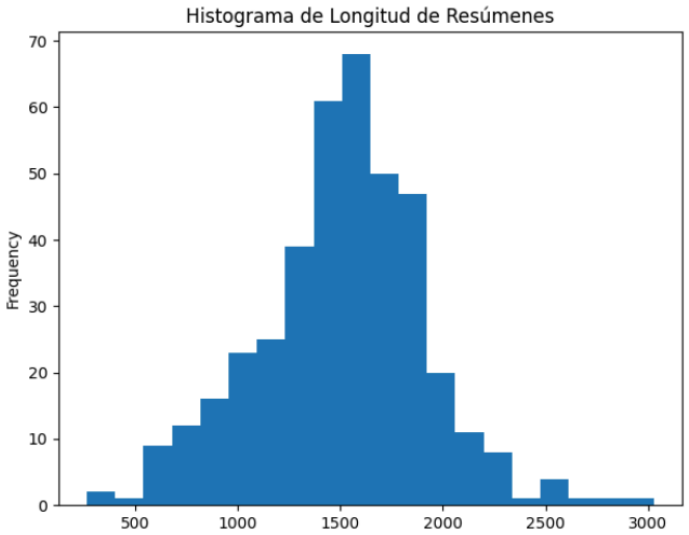

Figura 1: Distribución de longitud de resúmenes de artículos

Por otro lado, en el caso de las entidades presentes en la mayoría de abstractos, se encontró que su tipo tiene una tendencia clara: información genética, de enfermedad y fenotípica. Esto sirve para denotar la preferencia que tienen los diferentes artículos en cuanto a la información que trata y el ambiente en el que puede que este modelo esté mejor orientado.

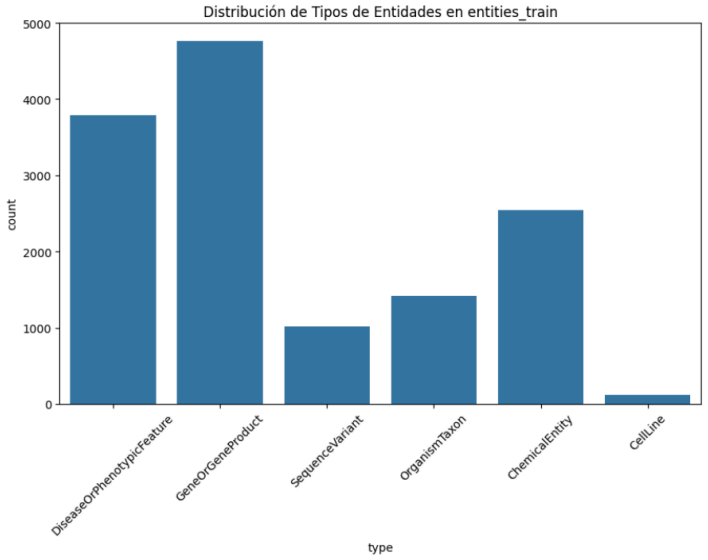

Figura 2: Distribución de tipos de entidades en los artículos

Finalmente, al ver las diferentes relaciones que existen entre las entidades, la más resaltable fue la de asociación. Seguida por correlación positiva, esto nos hace entender que la mayoría de relaciones

entre entidades nos permiten identificar más fácilmente la presencia de entidades relacionadas en el caso de una investigación concentrada en una entidad específica.

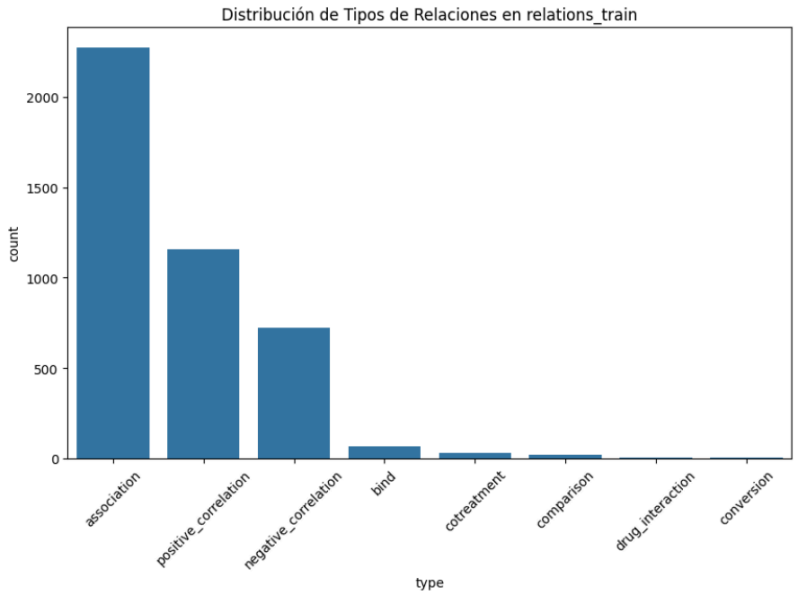

Figura 3: Distribución de tipos de relación entre entidades

Con el fin de poder visualizar mejor la presencia de las entidades en los diferentes extractos de los artículos, se decidió unificar la base de datos en un solo cuadro de información que proporcionaba la información más esencial para poder determinar la presencia de relaciones entre entidades de manera correcta.

	abstract_id	abstract	type_x	entity_1_mention	entity_2_mention	entity_1_type	entity_2_type	processed_abstract
0	1353340	We report on a new allele at the arylsulfatase...	Association	arylsulfatase A	metachromatic leukodystrophy	GeneOrGeneProduct	DiseaseOrPhenotypicFeature	We report on a new allele at the [ENTITY_1]ary...
2	1353340	We report on a new allele at the arylsulfatase...	Association	arylsulfatase A	MLD	GeneOrGeneProduct	DiseaseOrPhenotypicFeature	We report on a new allele at the [ENTITY_1]ary...
6	1353340	We report on a new allele at the arylsulfatase...	Association	ARSA	metachromatic leukodystrophy	GeneOrGeneProduct	DiseaseOrPhenotypicFeature	We report on a new allele at the arylsulfatase...
8	1353340	We report on a new allele at the arylsulfatase...	Association	ARSA	MLD	GeneOrGeneProduct	DiseaseOrPhenotypicFeature	We report on a new allele at the arylsulfatase...
12	1353340	We report on a new allele at the arylsulfatase...	Association	arylsulfatase	metachromatic leukodystrophy	GeneOrGeneProduct	DiseaseOrPhenotypicFeature	We report on a new allele at the [ENTITY_1]ary...

Figura 4: Base de datos unificada para el entrenamiento y evaluación del modelo

Modelado

El proceso de modelado se centró en la selección y entrenamiento de un modelo capaz de abordar con precisión y eficiencia las tareas de clasificación de relaciones semánticas entre entidades textuales. Después de un análisis comparativo, se optó por utilizar el modelo BERT (Bidirectional Encoder Representations from Transformers) como la base principal del sistema debido a sus características avanzadas y ventajas en tareas de procesamiento de lenguaje natural (PLN).

Se evaluaron múltiples enfoques, entre los cuales destacaron las Redes Neuronales Recurrentes (RNN) y los modelos basados en transformadores, como BERT.

Ventajas de BERT sobre RNN:

Mayor precisión: BERT demostró un mejor rendimiento en tareas de clasificación y extracción de relaciones gracias a su capacidad para comprender contextos bidireccionales. Esto significa que analiza las palabras en relación con las que aparecen tanto antes como después en la secuencia.

Mejor manejo de texto largo: Las RNN, aunque eficaces para tareas secuenciales, tienden a sufrir de problemas de gradientes desvanecientes y pérdida de contexto en secuencias extensas, lo que no ocurre con BERT gracias a su arquitectura de atención.

Generalización superior: El preentrenamiento de BERT en grandes corpus como Wikipedia permite que el modelo capture patrones lingüísticos robustos y generales, facilitando su adaptación a tareas específicas con fine-tuning.

Desempeño de RNN: Aunque las RNN fueron inicialmente probadas debido a su menor demanda computacional, su desempeño fue consistentemente inferior en métricas como la precisión y la capacidad para identificar relaciones complejas. Estas limitaciones resultaron en su descarte para la solución final.

Ventajas de BERT

Preentrenamiento avanzado: Al estar preentrenado en un corpus masivo, BERT requiere menos datos para ajustarse a tareas específicas, lo que lo hace ideal en dominios con datasets limitados.

Bidireccionalidad: Su comprensión contextual en ambas direcciones asegura un análisis más profundo y preciso de las relaciones entre términos.

Adaptabilidad: Su arquitectura permite la incorporación de tareas personalizadas mediante la adición de capas de clasificación, sin necesidad de modificar el modelo base.

Consideraciones Computacionales
El uso de BERT implica mayores requerimientos de hardware debido a su complejidad arquitectónica:

Demanda de recursos: Comparado con las RNN, BERT necesita un procesamiento computacional significativamente más alto, especialmente durante el ajuste fino (fine-tuning).

Solución implementada: Para mitigar este desafío, se utilizó una GPU (Unidad de Procesamiento Gráfico) durante el entrenamiento. Esto permitió reducir los tiempos de procesamiento y mejorar la eficiencia, haciendo viable el ajuste del modelo en un marco de tiempo razonable.

Evaluación

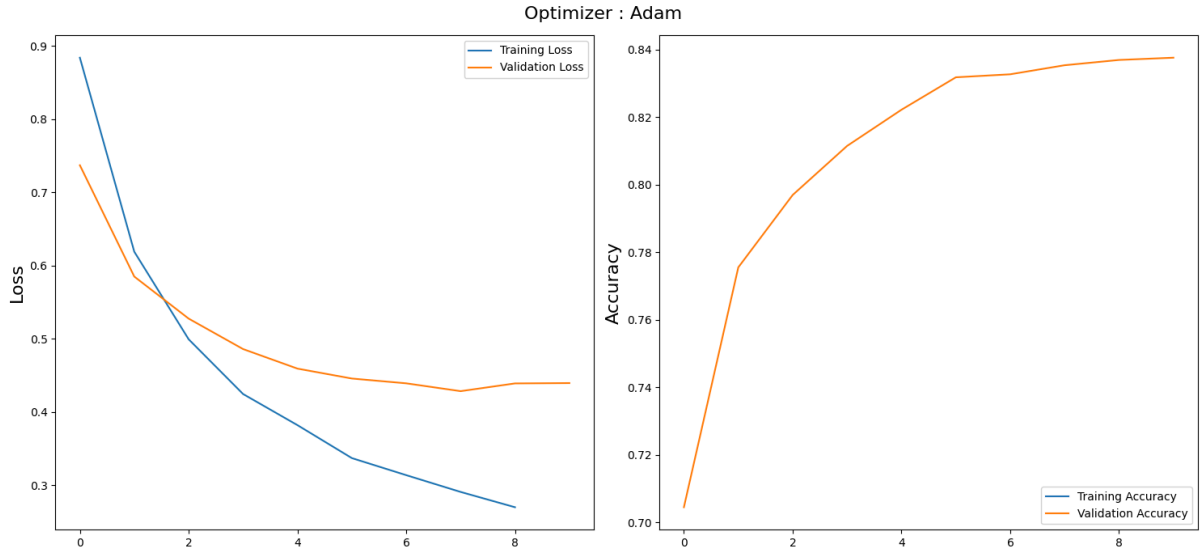

Figura 4: pérdida y exactitud a través de las épocas en el modelo de relaciones.

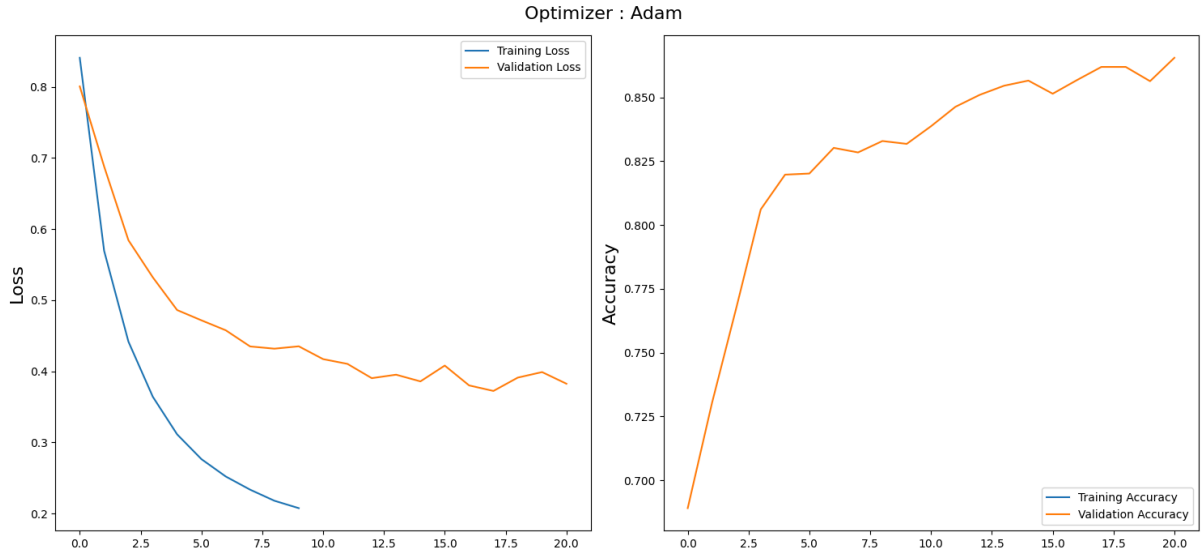

Figura 5: pérdida y exactitud a través de las épocas en el modelo de relaciones.

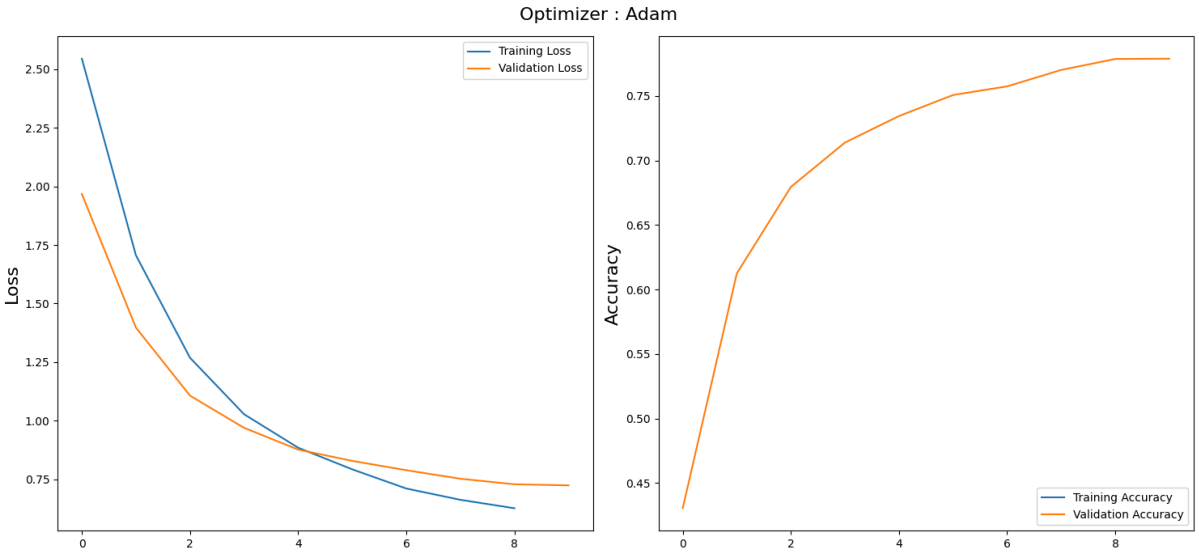

Figura 6: pérdida y exactitud a través de las épocas en el modelo de relaciones.

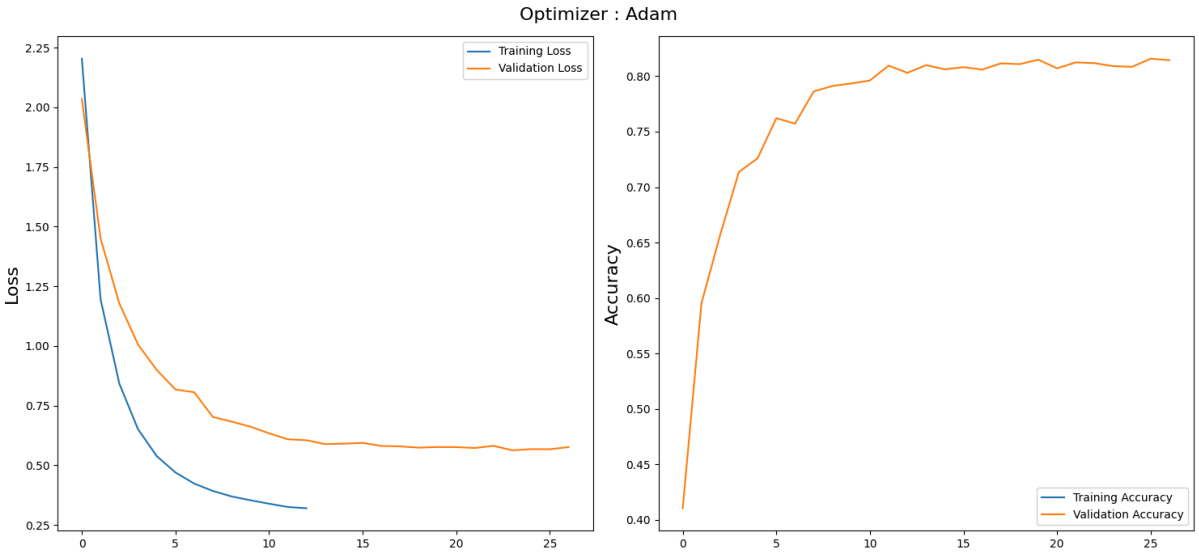

Figura 7: pérdida y exactitud a través de las épocas en el modelo de relaciones.

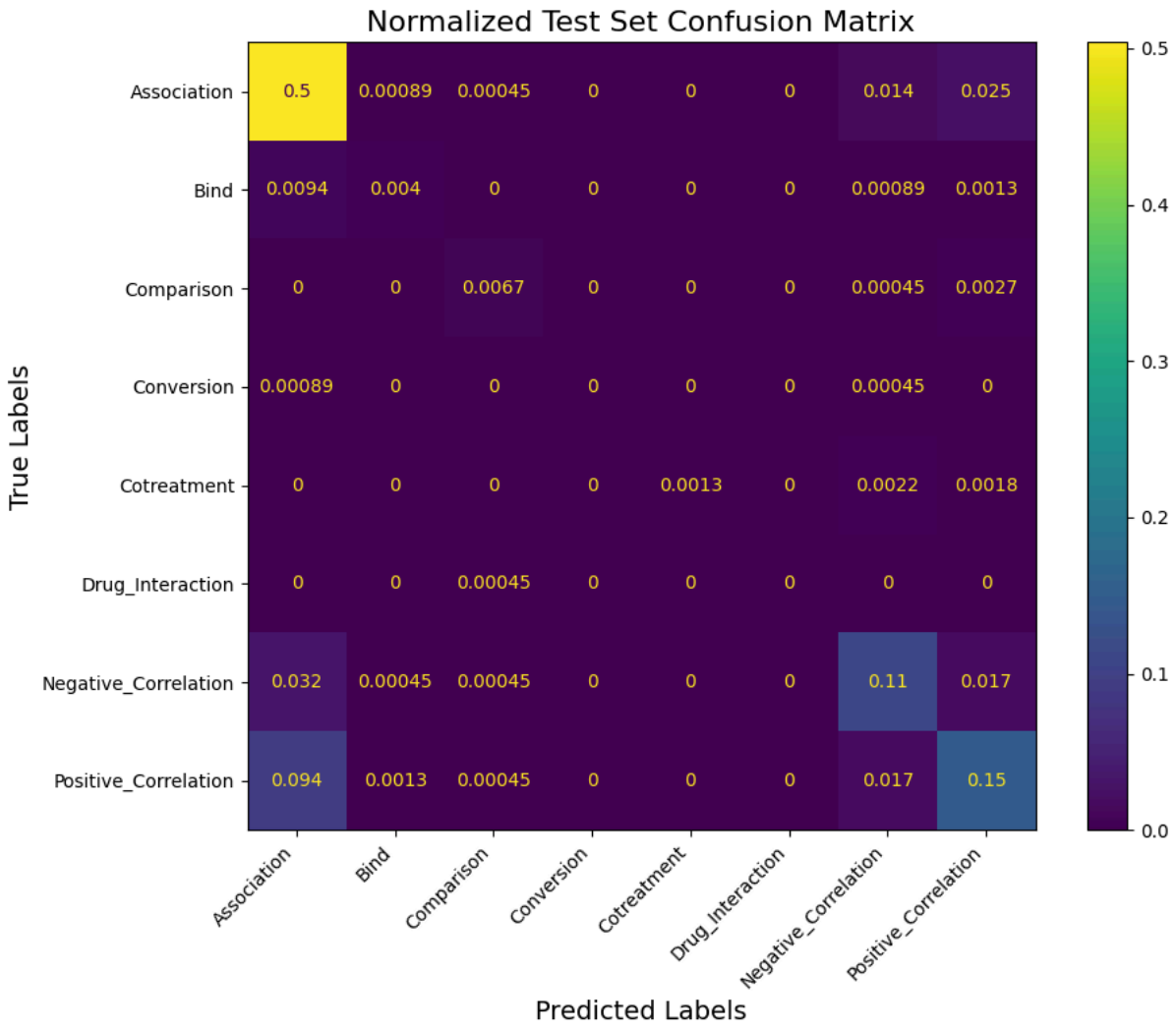

Figura 8: matriz de correlación entre las predicciones hechas por el modelo vs valores reales.

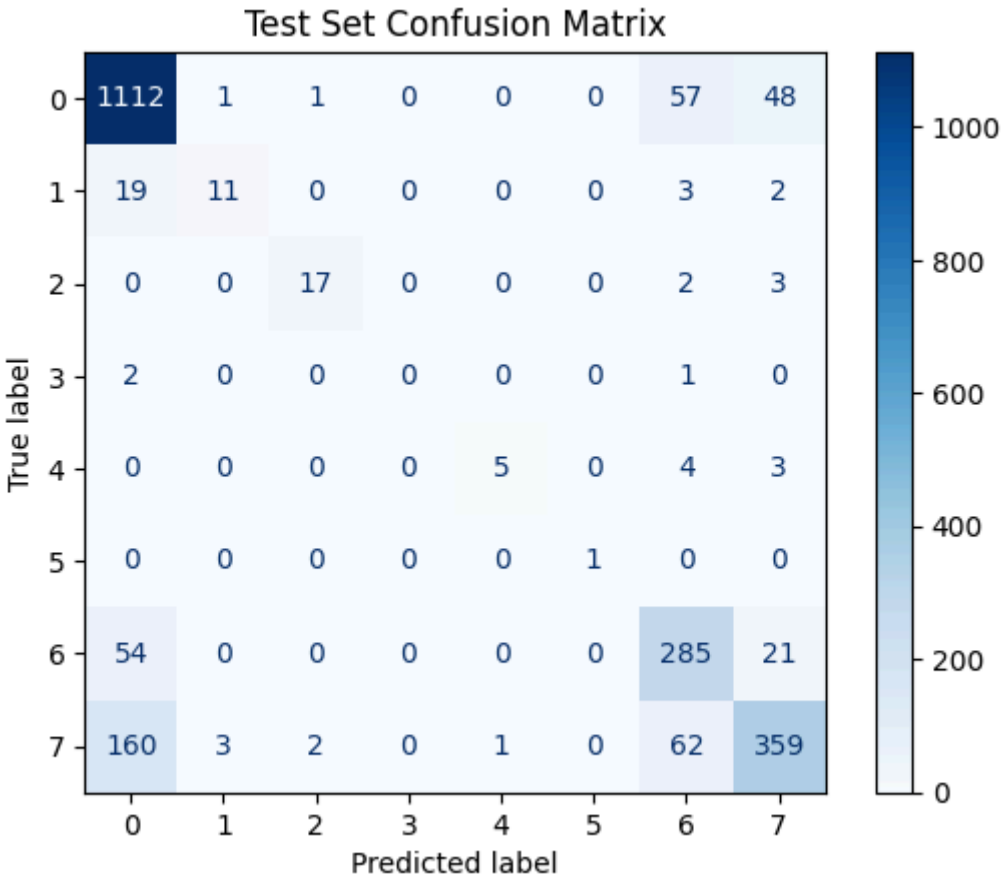

Figura 9: matriz de correlación entre las predicciones hechas por el modelo vs valores reales.

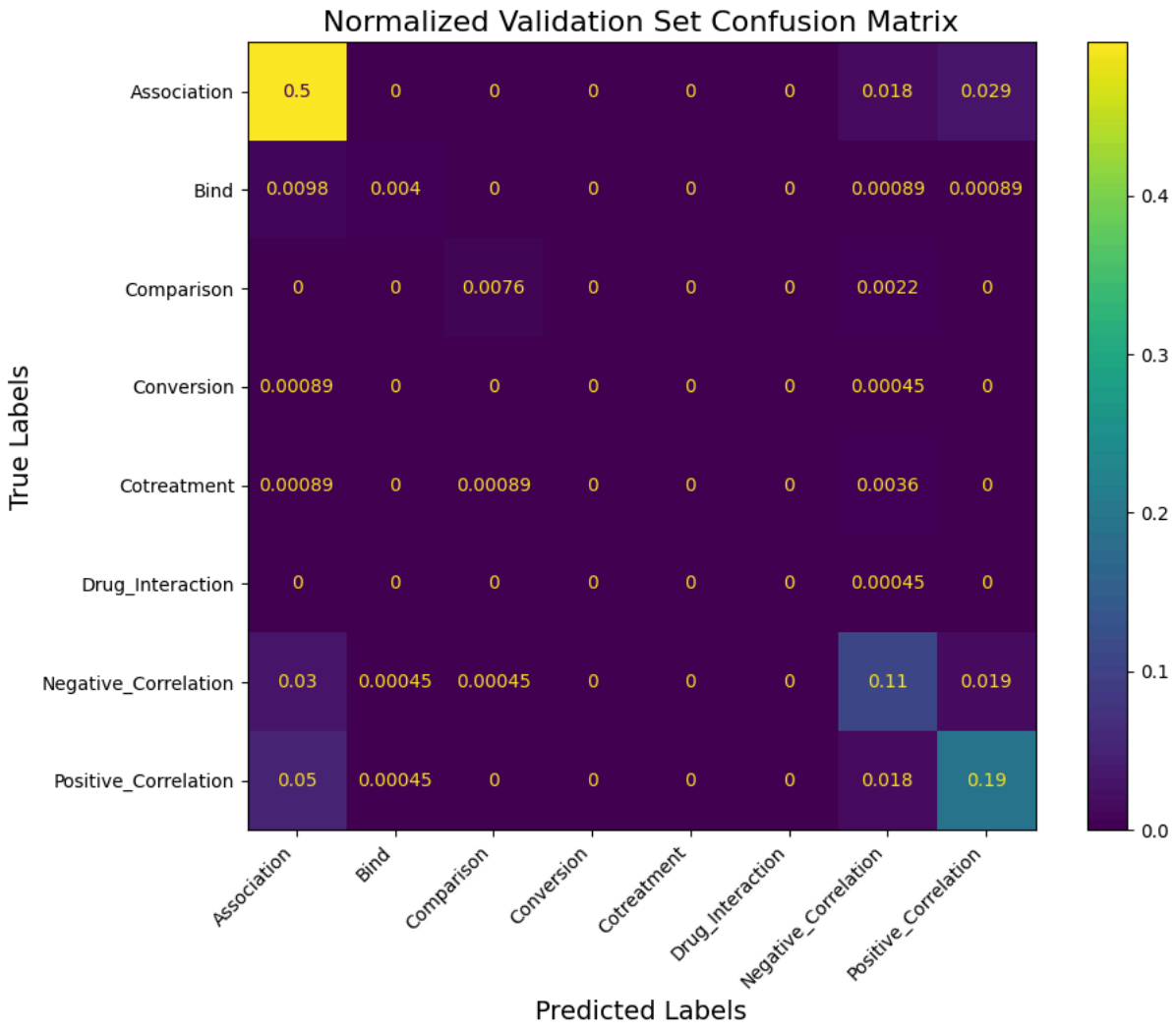

Figura 10: matriz de correlación entre las predicciones hechas por el modelo vs valores reales.

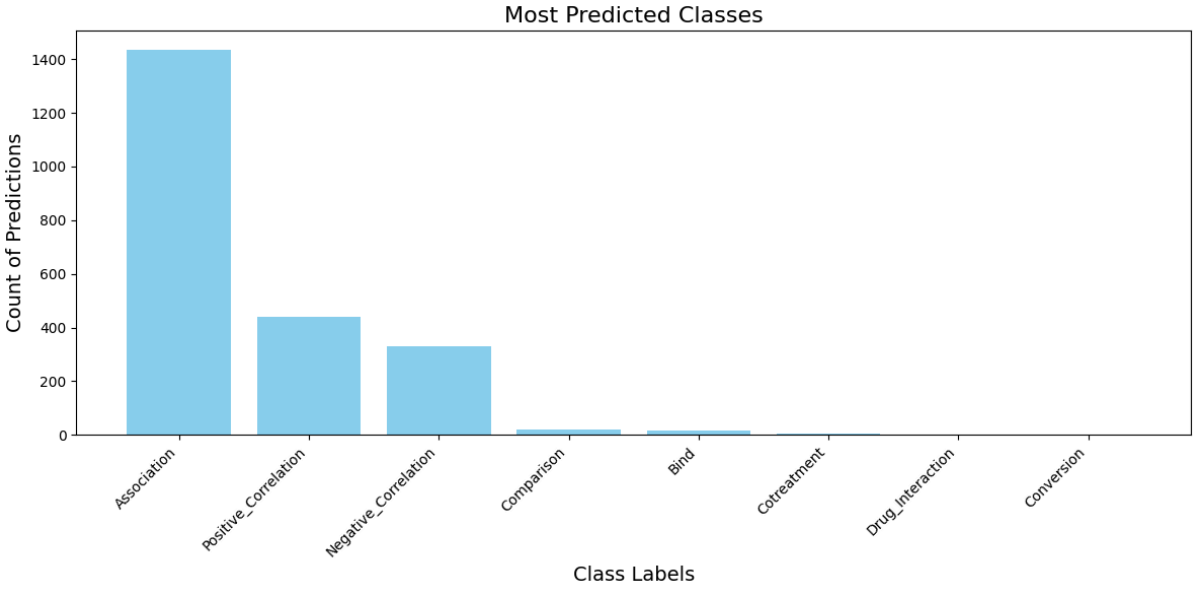

Figura 11: top de clases predichas por el modelo.

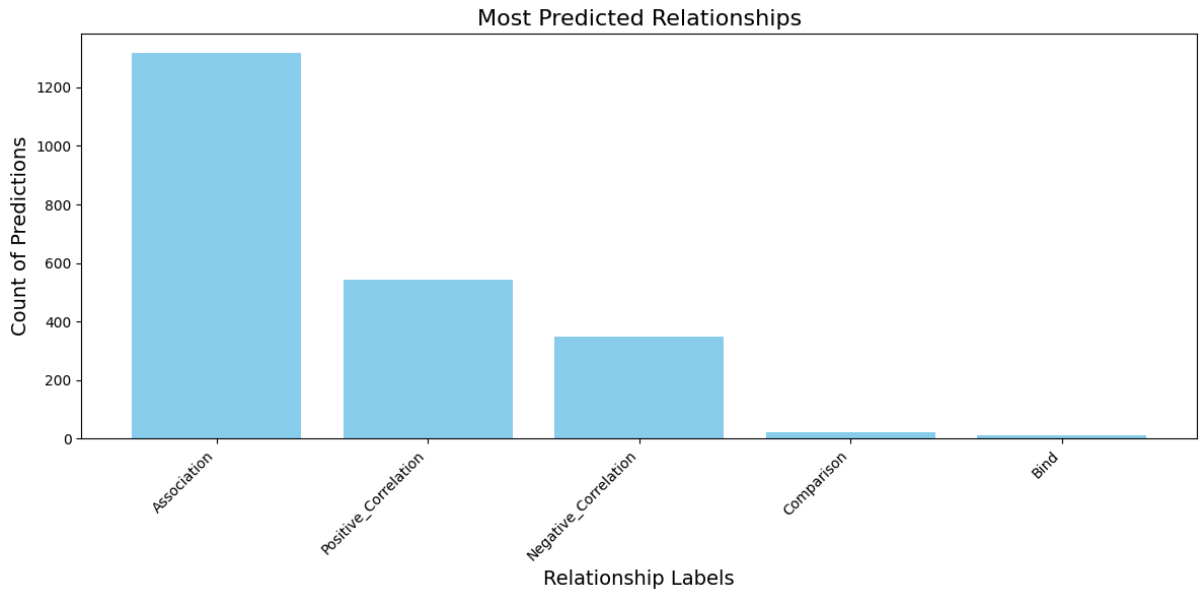

Figura 12: top de clases predichas por el modelo.

Desarrollo

Para la carga de datos y preparación para entrenamiento de modelos se usaron los resultados del EDA y se limpiaron para poder alimentar los modelos. En esta sección se describen los modelos BERT utilizados en la aplicación y su integración para cumplir con los objetivos específicos de la solución desarrollada. Se identificaron tres configuraciones principales, descritas a continuación:

1. **Modelo Base: bert-base-uncased**
Este modelo se seleccionó como punto de partida debido a su capacidad para realizar tareas generales de clasificación de texto.

Características:
bert-base-uncased es un modelo preentrenado en grandes corpus de texto en inglés, utilizando únicamente letras minúsculas para normalizar los datos. Es ampliamente utilizado como referencia en tareas de Procesamiento de Lenguaje Natural (PLN) debido a su robustez y generalidad.

Integración en la Aplicación:
En la solución, este modelo fue cargado con la clase AutoModelForSequenceClassification, ajustando el número de etiquetas (num_labels) para adaptarlo a las clases específicas del problema. Un tokenizer asociado (AutoTokenizer) se utilizó para preprocesar los datos textuales, asegurando consistencia en la tokenización de los textos de entrada.

Propósito:
Actuar como base para la clasificación inicial y permitir comparaciones con el modelo especializado.

2. **Modelo Personalizado: relationship_model**
Este modelo se ajustó específicamente para abordar la tarea de extracción de relaciones semánticas entre entidades, un componente central de la aplicación.

Características:
relationship_model es un modelo BERT fine-tuned entrenado en un conjunto de datos especializado, diseñado para identificar relaciones entre entidades textuales, como entity_1_type y entity_2_type, y clasificar dichas relaciones en categorías predefinidas.

Integración en la Aplicación:
Se cargó un tokenizer personalizado mediante la clase AutoTokenizer, optimizado para los textos del dominio específico de la aplicación.
El modelo es responsable de analizar descripciones procesadas (processed_abstract) y generar predicciones relacionadas con las relaciones semánticas.

Propósito:
Potenciar el módulo de extracción de relaciones, maximizando la precisión mediante un modelo ajustado al dominio.

3. Tokenizer Adicional
Se creó un tercer tokenizer que se utiliza en una función de preprocesamiento adicional (preprocess_function). Aunque no se menciona explícitamente el modelo asociado, parece ser una referencia intermedia para garantizar consistencia en los datos procesados.

Integración en la Aplicación:
Este tokenizer se emplea en tareas complementarias de preparación de datos, aplicando truncamiento y padding automático para textos largos.

Propósito:
Proveer soporte al flujo de procesamiento y garantizar que los datos sean compatibles con los modelos utilizados.

Comparación y Diferencias
Los modelos utilizados presentan características complementarias:

Generalidad vs. Especificidad:

bert-base-uncased proporciona una solución generalista que puede ser ajustada para diferentes tareas. relationship_model está optimizado para un propósito concreto dentro del dominio de la aplicación.

Entrenamiento Previo:

bert-base-uncased está preentrenado en corpus generales de texto. relationship_model fue ajustado utilizando datos especializados, mejorando su desempeño en tareas específicas.

Rol en la Aplicación:

bert-base-uncased se utiliza como referencia para análisis comparativos y pruebas iniciales.
relationship_model desempeña un rol crítico en la solución final al resolver el problema principal.
Con esta configuración, la aplicación garantiza un equilibrio entre robustez y especificidad, utilizando modelos que aprovechan tanto datos generales como información especializada para maximizar la precisión y efectividad en las tareas de clasificación y extracción de relaciones semánticas.

La implementación final fue una la aplicación web interactiva que permite a los usuarios seleccionar un modelo BERT pre entrenado, ingresar un texto científico y obtener predicciones sobre las relaciones entre entidades en el texto. La aplicación también muestra métricas de rendimiento y visualizaciones para cada modelo seleccionado. Se optó por hacer una interfaz con Streamlit que permitiera la interacción con el usuario para seleccionar un modelo y hacer predicciones con el seleccionado. Para evitar problemas con las dependencias se optó por usar un contenedor de docker. Los modelos se cargan en la aplicación y son llamados desde la interfaz cuando el usuario lo solicita. Las principales herramientas utilizadas fueron:

Streamlit: Para crear la interfaz web.
Pandas: Para manejar datos.
Torch: Para trabajar con modelos de aprendizaje profundo.
Transformers: Para cargar modelos y tokenizadores preentrenados de Hugging Face.
Time: Para medir el tiempo de predicción.

Discusión

Los principales desafíos encontrados fueron encontrar un balance entre la precisión de los modelos y el tiempo de entrenamiento ya que al principio resultaron muy tardados. Otro desafío en la fase de selección de datos fue asegurar la calidad y diversidad del corpus de entrenamiento para mejorar la precisión del modelo, manejar eficientemente los recursos computacionales para reducir el tiempo de respuesta, y garantizar la escalabilidad de la aplicación para soportar múltiples usuarios.

Para abordar este problema se decidió variar la cantidad de epochs para mejorar los tiempos. A su vez los problemas de calidad y diversidad del corpus de entrenamiento, se decidió recolectar un conjunto de datos más amplio y variado, asegurando que las palabras y frases relevantes estuvieran bien representadas. Se ajustaron los hiperparámetros de los modelos BERT, como el tamaño del vector, la ventana de contexto y el número de épocas, para mejorar el rendimiento. Se optó por utilizar modelos BERT preentrenados avanzados y se experimentó con diferentes configuraciones para encontrar la mejor combinación. Además, se implementaron técnicas de validación cruzada para asegurar que el modelo generalizara bien a datos no vistos. Estas decisiones ayudaron a mejorar la precisión de las predicciones y la eficiencia del modelo.

Desarrollar una aplicación de predicción de relaciones entre entidades utilizando Streamlit y modelos BERT implica varias lecciones aprendidas. Primero, es crucial el preprocesamiento de datos, asegurándose de que el texto esté limpio y tokenizado correctamente. La selección de modelos es vital, probando diferentes configuraciones para encontrar la mejor opción. Evaluar el modelo con métricas adecuadas y visualizaciones como matrices de confusión ayuda a entender su rendimiento. La optimización de hiperparámetros mediante experimentación y validación cruzada mejora la generalización del modelo. Streamlit facilita la creación de aplicaciones interactivas, y es importante

proporcionar una interfaz de usuario clara y útil. El manejo de errores y la robustez de la aplicación son esenciales para una buena experiencia del usuario. Además, optimizar el rendimiento y considerar la escalabilidad asegura que la aplicación pueda manejar múltiples usuarios. Finalmente, la seguridad y privacidad de los datos son fundamentales, implementando mecanismos de autenticación y cumpliendo con las regulaciones de privacidad. Estas lecciones ayudan a desarrollar aplicaciones más robustas y eficientes.

Para mejorar la aplicación, se consideran varias estrategias. Primero, aumentar la calidad de las predicciones de los modelos mediante la recolección de un corpus más grande y diverso, así como la implementación de técnicas de aumento de datos. Ajustar los hiperparámetros y probar diferentes arquitecturas de modelos también podría mejorar el rendimiento de los Berts. Además, incorporar técnicas de aprendizaje por transferencia utilizando modelos preentrenados más avanzados puede proporcionar mejores resultados. Mejorar la interfaz de usuario de Streamlit para hacerla más intuitiva y proporcionar visualizaciones adicionales de los resultados de las predicciones puede enriquecer la experiencia del usuario. Finalmente, optimizar el tiempo de respuesta y asegurar la escalabilidad de la aplicación garantizará que pueda manejar un mayor número de usuarios sin comprometer el rendimiento. Estas mejoras no solo aumentarán la precisión de las predicciones, sino que también harán que la aplicación sea más robusta y fácil de usar.

Como extensiones futuras podrían implementarse paralelismo con librerías que permitan un control profundo de vectores para acelerar el entrenamiento y automatizar el proceso de selección de parámetros con más librerías, también implementar más modelos a la aplicación para incrementar las opciones dadas al usuario. Se podría considerar varias mejoras. Primero, integrar capacidades de procesamiento de lenguaje natural más avanzadas, como modelos de lenguaje de última generación (por ejemplo, GPT-4 o T5), para mejorar la precisión y comprensión del texto. También podrías añadir soporte para múltiples idiomas, permitiendo a los usuarios analizar textos en diferentes lenguas. Implementar una base de datos para almacenar y gestionar las predicciones y entradas de los usuarios podría facilitar el análisis de tendencias y la mejora continua del modelo. Además, podrías desarrollar una API que permita a otras aplicaciones y servicios acceder a las capacidades de predicción de la aplicación. Finalmente, incorporar análisis de sentimientos y detección de entidades nombradas podría proporcionar un contexto adicional y enriquecer las predicciones, haciendo la aplicación aún más útil y versátil. Estas extensiones no solo mejorarían la funcionalidad y precisión de la aplicación, sino que también ampliarían su alcance y utilidad para los usuarios.

Referencias

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

Daniel. (2024, 20 mayo). *Memoria a largo plazo a corto plazo (LSTM): ¿Qué es?* Formación En Ciencia de Datos | DataScientest.com. https://datascientest.com/es/memoria-a-largo-plazo-a-corto-plazo-lstm

¿Qué es una RNN?: Explicación sobre redes neuronales recurrentes: AWS. (s. f.). Amazon Web Services, Inc. https://aws.amazon.com/es/what-is/recurrent-neural-network/#:~:text=Una%20red%20neuronal%20recurrente%20(RNN,salida%20de%20datos%20secuencial%20espec%C3%ADfica.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1706.03762