



Centro Universitario UAEM Texcoco

DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN

ALGORITMO DE PREGUNTA-RESPUESTA A PARTIR DE TEXTOS CLÍNICOS

Tesis Doctoral

M. en C. Josué Padilla Cuevas

Tutor académico interno
Dra. Alma Delia Cuevas Rasgado

Tutor académico externo
Dr. José Alejandro Reyes Ortiz

16 de septiembre de 2024

Índice general

1. Introducción	1
1.1. Planteamiento del problema	2
1.2. Justificación	3
1.3. Hipótesis de investigación	4
1.4. Objetivos	4
1.4.1. Objetivo general	4
1.4.2. Objetivos específicos	5
1.5. Estructura de la tesis	5
2. Marco teórico	6
2.1. Procesamiento de lenguaje natural	6
2.2. Recuperación de información	8
2.3. Herramientas computacionales de búsqueda de respuestas	10
2.4. Aprendizaje automático	12
2.4.1. Tipos de aprendizaje automático	12
2.5. Aprendizaje profundo	13
2.5.1. Grandes Modelos de Lenguaje	16
3. Marco de referencia	18
3.1. Trabajos relacionados con los Sistemas de Pregunta-Respuesta Generales	18
3.2. Trabajos relacionados con la tarea de Recuperación de infor- mación	24
3.3. Trabajos relacionados con la extracción de respuestas	27
4. Metodología	31
4.1. Conjuntos de datos	32
4.2. Pre-procesado	32

4.3.	Análisis de la pregunta	33
4.4.	Recuperación de documentos científicos	35
4.4.1.	Artículos científicos médicos	36
4.4.2.	Pre-procesado de artículos científicos	37
4.4.3.	Recuperador de Documentos	39
4.4.4.	Evaluación del Sistema de Recuperación de artículos científicos	48
4.5.	Extracción de las respuestas candidatas	51
4.5.1.	Corpus para el entrenamiento del LLM médico	52
4.5.2.	Aprendizaje Adaptativo del gran modelo de lenguaje médico	54
4.5.3.	Ajuste fino del LLM médico para la tarea de pregunta- respuesta	60
4.5.4.	Calibración por optimización de hiperparámetros para el Ajuste Fino del Modelo Médico	61
5.	Resultados	65
5.1.	Resultados de la Recuperación de Documentos científicos	66
5.2.	Resultados de la extracción de las respuestas	71
5.2.1.	Resultados del entrenamiento del LLM médico	71
5.2.2.	Resultados del ajuste fino del LLM médico	73
6.	Análisis y discusión de resultados	79
7.	Conclusiones y trabajo futuro	83
	Referencias	85
A.	Comparativa de herramientas de PLN	95
B.	Palabras vacías del español	96

Índice de figuras

2.1. Sistema de Recuperación de Información descrito por Salton, G & Mc Gill.	9
2.2. Arquitectura básica de una Red Neuronal Recurrente.	15
2.3. Arquitectura Transformer. Obtenida del artículo “Attention Is All You Need” [1].	17
4.1. Metodología general para el algoritmo de pregunta-respuesta. .	31
4.2. Análisis sintáctico de una pregunta generado con la herramienta Freeling.	35
4.3. Recuperación de documentos científicos médicos.	36
4.4. Extracto de un artículo científico del corpus BioASQ MESINESP.	37
4.5. Artículo procesado del conjunto de datos BioASQ MESINESP. .	39
4.6. Nube de palabras de las top-words del tópico de COVID-19. .	40
4.7. Coeficientes de similitud del modelo vectorial	43
4.8. Coeficientes de similitud del modelo probabilístico	45
4.9. Coeficientes de similitud obtenidos con el modelo LSI	47
4.10. Preguntas de COVID-19 en español	49
4.11. Metodología para la extracción de la respuesta.	52
4.12. Ejemplos de una pregunta del conjunto de entrenamiento. . . .	53
4.13. Ejemplo de la salida del Tokenizador base.	55
4.14. Conversión de texto a valores numéricos realizada por el Tokenizador entrenado.	56
4.15. Arquitectura para la calibración de hiperparámetros.	62
5.1. Resultados de la comparativa de modelos de RI aplicados a textos científicos sobre COVID-19	71
5.2. Entrenamiento del modelo MédicoBERT.	72

5.3.	Comparativa de los resultados de extracción de respuestas basados en la métrica F1.	77
5.4.	Ejemplos de los tipos de preguntas que puede responder el modelo médicoBERT.	78

Índice de tablas

3.1. Estado del arte	22
3.2. Estado del arte de Recuperación de Información.	26
3.3. Estado del arte de modelos de lenguaje	30
4.1. Clasificación de preguntas realiza para el algoritmo de pregunta- respuesta.	34
4.2. Valores TF-IDF obtenidos a partir de la consulta y de los textos preprocesados y lematizados.	42
4.3. Matriz de frecuencia de Términos	47
4.4. Lista de corpus de texto utilizados para el preentrenamiento del modelo médico.	53
5.1. Resultados de la primera configuración experimental ordena- dos por ranking	67
5.2. Resultados experimentales de la segunda configuración expe- rimental.	67
5.3. Resultados de la tercera configuración experimental.	68
5.4. Tabla de resultados del experimento 4.	69
5.5. Resultados generales, se muestran los 3 mejores resultados pa- ra cada Top N.	70
5.6. Resultados de modelo de lenguaje basado en BioAsq MESINESP	73
5.7. Partición para el ajuste fino del conjunto de datos BioAsq10 .	74
5.8. Resultados del entrenamiento de la configuración inicial	75
5.9. Resultados del entrenamiento de la tarea de QA con la cali- bración gruesa de hiperparámetros	75
5.10. Resultados del entrenamiento de la tarea de QA con la bús- queda heurística de hiperparámetros	76

5.11. Resultados del entrenamiento de la tarea de QA con la cali- bración fina de hiperparámetros	76
A.1. Comparación de herramientas de PLN	95

Capítulo 1

Introducción

La investigación médica en español está experimentando un auge sin precedentes, por ejemplo, debido a la pandemia provocada por la enfermedad de COVID-19 causada por el virus SARS-CoV-2 que se detectó en diciembre de 2019 en la ciudad de Wuhan, China, se han infectado a nivel mundial más de 176 millones de personas ¹. Los esfuerzos por difundir y contener a dicha enfermedad han generado una gran cantidad de información de forma extraordinaria, tal como, documentos científicos.

Este gran volumen de información y conocimiento almacenado en los textos médicos proporciona a los expertos del dominio una herramienta fiable para sus diversas actividades de investigación donde requieren localizar información precisa de manera rápida y eficiente. Sin embargo, dada la estructura de los documentos y el volumen de datos, la tarea de buscar, recuperar y extraer información específica que requieren los expertos no es trivial. Además, realizar estas tareas de manera manual consume mucho tiempo, son tediosas y costosas.

Por lo anterior, en este trabajo de investigación se propone un algoritmo computacional de pregunta-respuesta para buscar, recuperar y extraer información del dominio médico a partir de textos científicos. El algoritmo está basado en modelos de recuperación de información y en grandes modelos de lenguaje neuronales. Además, las preguntas serán realizadas en lenguaje natural en el idioma español, aportando al problema de carencia de recursos para el procesamiento de textos en este idioma. Finalmente, las respuestas serán recuperadas de manera automática y a partir de un gran volumen de

¹<https://github.com/CSSEGISandData/COVID-19/>

textos científicos en formato no estructurado.

Este proyecto de investigación va dirigido tanto a los investigadores hispanohablantes de la salud que necesitan una asistencia eficaz para acceder de manera rápida, a la información médica. Por ejemplo, los investigadores en este campo siempre quieren estar al tanto de la evidencia clínica actualizada para el diagnóstico y el tratamiento de enfermedades bajo el esquema de Medicina Basada en la Evidencia. Además, el proyecto también está dirigido a investigadores en el área de las Ciencias de la Computación que requieran conocer técnicas de Procesamiento de Lenguaje Natural que incluyan la Recuperación y Extracción de Información.

En este capítulo se revisará la descripción del problema, la justificación, los objetivos, la hipótesis de investigación y por último la organización del documento de tesis.

1.1. Planteamiento del problema

La manipulación y tratamiento automatizado de información a partir de grandes conjuntos de datos digitales no estructurados plantea un reto computacional considerable. Un ejemplo es el ritmo acelerado de la investigación del COVID-19, que ha dado lugar a un volumen sustancial de literatura científica biomédica sobre la sintomatología, las pruebas de diagnóstico, las medidas de prevención, las vacunas y los tratamientos de esta nueva enfermedad. Esta vasta investigación se encuentra disponible públicamente para la comunidad médica en repositorios digitales. Según un recuento de la base de datos *Dimensions*², hasta la fecha se han publicado más de 2 millones de artículos científicos relacionados con el COVID-19. No obstante, localizar información relevante de manera manual para la comunidad de investigadores médicos dentro del volumen tan amplio de publicaciones, conlleva una gran cantidad de tiempo y esfuerzo. El problema descrito se puede abordar con un enfoque computacional, lo que requiere un análisis para proponer un algoritmo de búsqueda y recuperación de información

Así mismo, la incorporación del lenguaje natural al realizar consultas en estos algoritmos de búsqueda y recuperación de información presenta desafíos como la ambigüedad léxica y sintáctica, así como la presencia de anáforas directas e indirectas en los textos. Abordar estos fenómenos mediante métodos computacionales resulta crucial para mejorar la precisión en esta tarea.

²<https://reports.dimensions.ai/covid-19/>

De la misma forma, abordar la heterogeneidad de las fuentes de datos es otro desafío fundamental. La vasta cantidad de información sobre literatura médica se encuentra dispersa en diversas fuentes y formatos, incluyendo bases de datos, ontologías, archivos XML, JSON y artículos de investigación en texto plano. Estos datos residen en repositorios médicos abiertos como PubMed, IBECS, SCIELO y LILACS. La diversidad de formatos exige el desarrollo de procedimientos computacionales para normalizar la representación de los textos, lo cual es esencial para facilitar su procesamiento y análisis.

Finalmente, la optimización de los recursos computacionales para reducir los tiempos de respuesta y minimizar el impacto ambiental es otro desafío crucial en el desarrollo de este tipo de herramientas de búsqueda y extracción de información.

1.2. Justificación

A lo largo de su historia, la humanidad ha tenido que enfrentar numerosas enfermedades que han puesto en riesgo a la población. En este contexto, la enfermedad por coronavirus del 2019 acumula hasta la fecha un estimado global de más de 6.9 millones de muertes. Solo en México, se contabilizan 7,621,062 casos positivos y 334,167 defunciones³. La pandemia de provocada por esta enfermedad ha puesto de manifiesto las limitaciones existentes en el conocimiento científico sobre diversos aspectos de la enfermedad, incluyendo sus mecanismos de transmisión, fisiopatología, respuesta inmune, tratamientos farmacológicos y no farmacológicos, y estrategias de control efectivas.

Sin embargo, la imperiosa necesidad de comprender y combatir esta enfermedad ha impulsado un avance significativo en la investigación científica. Muchos de los conocimientos necesarios ya han sido desarrollados o se encuentran en pleno proceso de generación [2]. Todo este conocimiento que se realizó a un ritmo muy acelerado y en grandes volúmenes, tiene la necesidad de una correcta manipulación y una interpretación precisa, hecho por el cual se derivó el interés de realizar esta investigación. Asociado a lo anterior, surgió la necesidad de recuperar información relevante a cualquier enfermedad infecciosa por ejemplo el COVID-19 de manera automática que se encontraba en representaciones complejas.

³<https://datos.covid-19.conacyt.mx/>

El objetivo de este trabajo de investigación es reducir el tiempo y el esfuerzo que conlleva la búsqueda de información a los expertos en el área médica y científica. Para conseguirlo, se diseñó y desarrolló un algoritmo de pregunta-respuesta las cuales serán realizadas en lenguaje natural y localizadas a partir de documentos científicos en formato no estructurado.

Por lo anterior, en este trabajo de investigación se presenta un algoritmo computacional para responder preguntas en lenguaje natural sobre enfermedades, síntomas, tratamientos y otros temas médicos. El algoritmo se basa en la implementación de una metodología de recuperación de información para identificar documentos relevantes basados en una consulta específica y en el desarrollo de un gran modelo de lenguaje para extraer información clave de los artículos científicos. Este último componente representa una contribución novedosa en el área de las Ciencias de la Computación.

Por último, el trabajo está delimitado en el idioma español puesto que según [3] es una de las 5 lenguas más importantes a nivel mundial y actualmente los escasos sistemas de extracción y recuperación de información en el área médica están enfocados en el idioma inglés. Debido a lo anterior la investigación tiene un impacto mundial, puesto que es de gran ayuda para la comunidad investigadora hispanohablante de Centro, Sudamérica, el Caribe y Europa

1.3. Hipótesis de investigación

La implementación de técnicas de Procesamiento de Lenguaje Natural, Extracción de Información y un enfoque de aprendizaje profundo en un algoritmo de preguntas en lenguaje natural permitirá generar respuestas precisas y relevantes sobre temas médicos a partir de un corpus de textos científicos en español.

1.4. Objetivos

1.4.1. Objetivo general

Desarrollar un algoritmo de pregunta-respuesta utilizando un enfoque de aprendizaje profundo para extraer y recuperar información precisa y relevante sobre temas médicos a partir de un gran conjunto de textos científicos en español.

1.4.2. Objetivos específicos

1. Recolectar publicaciones científicas provenientes de repositorios de la literatura médica para la creación de un conjunto de datos no estructurados en español.
2. Adaptar herramientas de procesamiento de lenguaje natural para la limpieza y tratamiento de los textos provenientes de la literatura médica en español en formato no estructurado.
3. Implementar modelos de Recuperación de Información probabilísticos, semánticos y vectoriales para obtener documentos relevantes a partir de una pregunta en lenguaje natural en español.
4. Adaptar y ajustar un gran modelo de lenguaje neuronal para extraer las respuestas de los documentos relevantes a partir de una pregunta en español.

1.5. Estructura de la tesis

El resto del documento de tesis se organiza de la siguiente manera:

- En el capítulo 2 se exponen las bases teóricas que fundamentan este trabajo de investigación.
- El capítulo 3 presenta el estado del arte con los avances actuales más importantes en el área de conocimiento relacionado con esta tesis.
- El capítulo 4 muestra el desarrollo, la metodología implementada y los recursos utilizados para llevar a cabo este trabajo.
- El capítulo 5 presenta los resultados obtenidos y se realiza un análisis y discusión de estos.
- Por último, el capítulo 6 contiene las conclusiones generales y presenta una propuesta de trabajos a futuro.

Capítulo 2

Marco teórico

En este capítulo se presentan los fundamentos teóricos que sustentan la investigación, proporcionando un marco conceptual sólido para el desarrollo del algoritmo de pregunta-respuesta.

2.1. Procesamiento de lenguaje natural

El Procesamiento de Lenguaje Natural (PLN) es un conjunto de técnicas computacionales para el análisis y representación de lenguajes humanos [4]. El PLN consiste en la manipulación de un lenguaje natural para la comunicación con las máquinas.

Desarrollar un programa que comprenda el lenguaje natural es un problema difícil, porque la mayoría de estos lenguajes son grandes, complejos y pueden tener una cantidad infinita de oraciones. Muchas palabras tienen distintos significados como vino o banco y las mismas frases tienen significados diferentes en varios contextos.

La importancia del PLN se debe a que existe una enorme cantidad de datos en la red y que pueden utilizarse como un gran recurso de información importante. Las aplicaciones de PLN más conocidas para el manejo de estos datos son:

- Indización y búsqueda de grandes textos
- Recuperación de documentos relevantes (RI)
- Clasificación de textos

- Extracción de información (EI)
- Traducción automática
- Reconocimiento de voz
- Resumen automático de textos
- Respuesta a preguntas (QA)
- Adquisición de conocimientos
- Generación de textos/diálogos

El pipeline de un sistema de PLN se divide en los niveles de análisis:

Fonológico: Este nivel relaciona a las palabras con los sonidos que representan.

Morfológico: Este nivel de análisis lingüístico explora la estructura interna de las palabras, descomponiéndolas en sus unidades mínimas de significado, los morfemas, y revelando así los procesos de formación de palabras.

Semántico: Este nivel de análisis se centra en la semántica, es decir, en el estudio del significado de las palabras, las frases y las oraciones en un contexto determinado. Asimismo, explora las relaciones de significado entre las palabras y cómo estas contribuyen a la construcción del sentido global de un texto. El objetivo de esta fase es determinar el significado de una frase, busca las condiciones en las que la oración es verdadera.

Pragmático: Este nivel de análisis se centra en la pragmática, explorando cómo el significado de las oraciones se ve influenciado por el contexto de enunciación, las intenciones del hablante y las inferencias que realiza el oyente. En este sentido, se estudia cómo las oraciones se relacionan entre sí en un discurso y cómo el significado de una oración puede depender del contexto previo.

Análisis sintáctico. Cumple con dos funciones: **determinar** la estructura de las oraciones, en esta parte se debe identificar el sujeto, el objeto de cada verbo y determinar las palabras modificadoras de las frases. La segunda tarea es regularizar la estructura sintáctica de una oración.

Otra de las etapas del PLN es el análisis del discurso y las gramáticas, los cuales se describen a continuación.

Análisis del discurso. Hasta ahora se ha descrito la estructura y el significado de las frases de manera individual, pero una parte fundamental del PLN es entender el significado de todo el texto. Porque, la información que transmite un texto es claramente más que la suma del significado de sus oraciones individuales. Por ejemplo, si un texto cuenta una historia, describe un procedimiento o presenta un argumento, se deben comprender las conexiones entre las frases que lo componen para entender plenamente la historia. El análisis del discurso es el estudio de estas conexiones entre las oraciones. Dado que estas conexiones están implícitas en el documento, identificarlas puede ser una tarea difícil.

Gramática. La gramática de una lengua puede considerarse como una teoría de la estructura de esa lengua. Como cualquier teoría matemática, la teoría de la gramática se basa en un conjunto de número finito de oraciones observadas, pero lo proyecta a un conjunto infinito de oraciones gramaticales. Esto es posible gracias a la creación de leyes generales (reglas gramaticales), enmarcadas en términos de construcciones hipotéticas como los fonemas, palabras, y frases, de la lengua hablada que se analiza. Una gramática bien formulada debe ser capaz de determinar sin ambigüedad el conjunto de todas las oraciones gramaticales en esa lengua [5].

2.2. Recuperación de información

La recuperación de información (RI) se define como a la necesidad de un usuario por obtener una serie de documentos relevantes ante una pregunta en realizada en lenguaje natural. Actualmente la RI tiene una gran importancia debido al gran volumen de información digital disponible y a la que se puede acceder de forma sencilla. En este sentido, Pérez Carballo define a la recuperación de información como “una tarea de tarea documentos relevantes desde un gran archivo en respuesta a una pregunta formulada y ordenarlos de acuerdo con su relevancia” [6]. Para Ricardo Baeza [7] la RI consiste en la representación, el almacenamiento, la organización y el acceso a los ítems de una colección de información. A partir de este concepto se han creado a lo largos de los años Sistemas de Recuperación de Información SRI, para solventar este tipo de requerimientos de datos, para Salton [8] cualquier SRI es descrito como una serie de de documentos o ítems, un conjunto de peticiones y algún mecanismo de similitud que determine los documentos satisfacen las necesidades de información expresadas por el usuario a partir de una pre-

gunta o consulta, en la (Figura 4.3) se puede observar el esquema del SRI descrito por Salton.

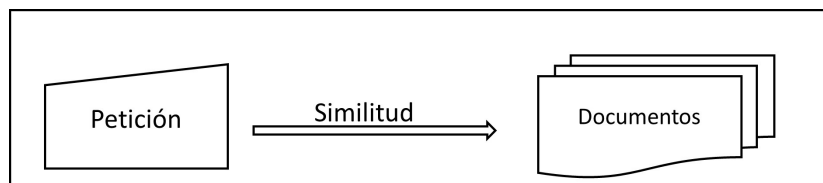


Figura 2.1: Sistema de Recuperación de Información descrito por Salton, G & Mc Gill.

Para el diseño y elaboración de un SRI es necesario aplicar algún modelo de recuperación, actualmente existen diversas variantes algunos de los más utilizados son los siguientes [9]:

- **Modelo booleano**, está fundamentado en la teoría de conjuntos, representa un enfoque clásico para la recuperación de información, en el que las consultas se expresan como expresiones booleanas y los documentos se representan como conjuntos de términos.
- **Modelo vectorial**, se encuentra fundamentado en el álgebra lineal, representa los documentos y las consultas como vectores en un espacio multidimensional. Esta representación permite calcular la similitud entre documentos y consultas mediante operaciones vectoriales, como el producto escalar, la distancia de Jaccard o del coseno.
- **Modelos lógicos**, están basados en la lógica formal, proporcionan un marco teórico riguroso para la representación del conocimiento y el razonamiento. Estos modelos utilizan lenguajes formales y sistemas de inferencia para capturar la estructura lógica de las afirmaciones y realizar deducciones válidas.
- **Modelos probabilísticos**, se fundamentan en el teorema de Bayes, proporcionan un marco probabilístico para la representación del conocimiento incierto y la inferencia. Estos modelos permiten actualizar las creencias a medida que se adquiere nueva evidencia en la recuperación de información.

- **Modelos de RI basados en la inteligencia artificial**, se basan en técnicas como redes neuronales artificiales, algoritmos genéticos y bases de conocimiento, para mejorar la precisión y relevancia de los resultados de búsqueda.

2.3. Herramientas computacionales de búsqueda de respuestas

Un enfoque de Recuperación de Información (RI) tiene como objetivo devolver los documentos relevantes a partir de una consulta realizada en lenguaje natural por un usuario, estos documentos pueden pertenecer a una colección o una biblioteca digital y por lo general son localizados por algún buscador de similitud.

Una herramienta de pregunta-respuesta o *Question-Answering* (QA) se puede clasificar como un tipo de recuperación de información donde a partir de una consulta expresada en lenguaje natural debe devolver ya no una serie de documentos relevante, sino la propia respuesta en forma de hechos [10]. La herramienta de QA tiene como objetivo obtener respuestas concretas a preguntas precisas realizadas por los usuarios, se pueden considerar como una evolución los buscadores WEB tradicionales. Para que una herramienta de QA sea precisa y eficaz debe adaptarse a dominios concretos mediante el uso de recursos semánticos y sintácticos adecuados. Los enfoques o herramientas de Pregunta-Respuesta pueden buscar información en diferentes formatos de datos, por ejemplo:

- Datos estructurados como bases de datos, ontologías y grafos de conocimiento.
- Datos semi-estructurados, donde pueden ser XML, JSON, estructuras de texto en BD.
- No estructurados o texto libre.
- Datos heterogéneos que se considera como una combinación de todos los tipos de datos anteriores.

Estos formatos de datos pueden estar localizados en la Web, en un conjunto de documentos digitales o en un texto plano. Las herramientas de QA se

enfocan en dominios libres o en dominios específicos, estos últimos se utilizan para realizar consultas de alta precisión, en donde también se toma en cuenta el perfil de los usuarios.

Las respuestas de este tipo de sistemas se pueden presentar en los siguientes formatos:

- Texto
- Imágenes
- Voz
- Video

Las características más importantes que deben tener este tipo de herramientas de extracción y recuperación de información fueron debatidas dentro del *Workshop Question Answering: Strategy and Resources* [11] y se describen a continuación:

Precisión. Una de las características trascendentes es la precisión, porque una herramienta de QA que arroja una respuesta incorrecta se considera peor que una que no responde. Es decir, la herramienta computacional de pregunta respuesta debe ser capaz de detectar si la respuesta no está disponible dentro del corpus, así como descubrir las contradicciones en las fuentes y la información conflictiva, por lo tanto, debe imitar la inferencia del sentido común.

Respuesta en tiempo razonable. La provisión de respuestas en tiempo real constituye un requisito fundamental en este tipo de enfoques. Para garantizar la pertinencia y actualidad de las respuestas, es imperativo mantener actualizados los datos con los hechos y eventos más recientes del dominio específico en cuestión.

Usabilidad. Una herramienta de pregunta-respuesta debe ser capaz de buscar y encontrar respuestas independientemente del formato de origen de la información, es decir, debe poder trabajar con fuentes heterogéneas. Además, debe ofrecer al usuario la posibilidad de proporcionar contexto para la pregunta, facilitando así una recuperación de información más precisa y relevante.

Complejidad. Las respuestas proporcionadas a los usuarios deben ser completas. Para esto se necesita realizar una fusión de la respuesta porque en ocasiones esta se encuentra distribuida a lo largo de un texto o incluso en

varios documentos de una colección. Algunas veces se deberá que encontrar analogías con otras preguntas en distintas bases de conocimiento y adaptar las respuestas con ayuda del contexto y el perfil del usuario. A esto se le conoce como la incorporación de capacidades de razonamiento y uso bases de conocimiento de gran rendimiento.

Relevancia. Los seres humanos son los evaluadores de la utilidad de una herramienta de pregunta-respuesta. Por lo que las respuestas deben ser precisas y relevantes acorde a las necesidades de una pregunta con el contexto particular. En algunas ocasiones será necesario utilizar de una secuencia de preguntas para ayudar a clarificar la información necesaria para devolver una respuesta.

2.4. Aprendizaje automático

El aprendizaje es una de las actividades más importantes que realizan los seres vivos, lo cual les permite adaptarse al entorno donde viven. En particular, el ser humano aprende de la naturaleza y de entornos especiales a través de diferentes técnicas que varían en complejidad, y que requieren transformaciones de ideas y estructuras de información en la mente.

Según Mitchell [12] el aprendizaje automático (ML) “*Una máquina aprende con respecto a una tarea T , una métrica de rendimiento P y un tipo de experiencia E , si mejora de forma fiable su rendimiento P en la tarea T , tras la experiencia E* ”. Son principios de entrenamiento que permiten a las máquinas aprender por si mismas o que mejoran automáticamente su rendimiento gracias a la experiencia obtenida. Sus principales características son:

- Imitar el comportamiento humano.
- Extraer patrones y relaciones de los datos.

2.4.1. Tipos de aprendizaje automático

El campo del Aprendizaje Automático (ML) se dedica al desarrollo de algoritmos y modelos computacionales que permiten a las máquinas aprender a partir de datos. Los principales paradigmas de aprendizaje automático son el supervisado, el no supervisado y el semi-supervisado.

- **Aprendizaje supervisado.** Exige la elaboración de un oráculo, de forma que pronostique las clases (categorías) a las que pertenecerían los ejemplos dados y, a continuación, clasifique los datos en esas categorías. En general, un método de aprendizaje se denomina supervisado si el proceso requiere alguna intervención del usuario. Algunos enfoques requieren que se proporcionen ejemplos de entrenamiento. Esto se hace como un pre-proceso para etiquetar adecuadamente los ejemplos que aparecen en el conjunto de entrenamiento y con esta información poder clasificar de manera adecuada a los nuevos datos a los que se somete el algoritmo. Entre los modelos de aprendizaje supervisado más utilizados se encuentran las Redes Neuronales Artificiales (RNA), las Máquinas de Soporte Vectorial (SVM) y los algoritmos de k-vecinos más cercanos. Cada uno de estos modelos presenta características y fortalezas particulares, adaptándose a diferentes tipos de problemas y conjuntos de datos.
- **Aprendizaje no supervisado.** Los métodos no supervisados tienen que descubrir las clases de conceptos a las que pertenecen los ejemplos dados, es decir, asignar los ejemplos a las clases de conceptos. A diferencia del aprendizaje supervisado, no se conocen las etiquetas de las clases (no hay clases predefinidas). El objetivo principal es identificar patrones inherentes en los datos para segmentarlos en grupos homogéneos basados en sus atributos. Estas técnicas, conocidas como clustering, permiten descubrir estructuras latentes en los datos y facilitar su análisis e interpretación. Un ejemplo destacado de algoritmo de *clustering* es *k-means*.
- **Aprendizaje semi-supervisado.** Este tipo de aprendizaje constituye un enfoque híbrido que combina las ventajas del aprendizaje supervisado y no supervisado. Al aprovechar tanto datos etiquetados como no etiquetados, este paradigma permite construir modelos de clasificación más precisos y robustos, especialmente en escenarios donde la adquisición de datos etiquetados resulta costosa o limitada.

2.5. Aprendizaje profundo

Según [4] a partir de 2005, el aprendizaje profundo (un enfoque basado en las RNA), empezó a cobrar protagonismo, y ahora se ha convertido en

una tarea que impulsa la investigación en la Inteligencia Artificial. El aprendizaje profundo se ocupa de la simulación de redes neuronales artificiales que “aprenden” gradualmente haciendo uso de múltiples capas. La técnica elemental se basa en los principios generales de las matemáticas que les permiten aprender a partir de ejemplos para reconocer personas u objetos en una fotografía, y traducir el lenguaje hablado de uno a otro. Las tecnologías basadas en el aprendizaje profundo han transformado la investigación y han producido resultados muy precisos en el reconocimiento del habla, la visión por computadora, el PLN y la robótica.

En el ámbito del aprendizaje profundo, destacan las Redes Neuronales Recurrentes (RNN) y las Redes Neuronales Convolucionales (CNN) como arquitecturas de gran relevancia. A continuación se detallarán las características y aplicaciones de cada una de ellas.

El diseño de las CNN se inspira en la estructura multicapa de la corteza visual (segmento del cerebro que recibe la información por medio de los ojos). Demasiadas capas de neuronas virtuales en una CNN es lo que hace que sea “profunda”, por lo tanto, sea más capaz de aprender sobre el mundo que la rodea.

Las RNN se caracterizan por no tener una estructura de capas definida, permiten realizar conexiones arbitrarias entre sus neuronas creando ciclos, esto hace que la red tenga una especie de memoria, las RNN hacen uso de la información de manera secuencial, con esto, La salida dependerá de los datos actuales, de los datos previamente procesados y también de los cálculos realizados (Figura 2.2). En cuanto más larga sea la secuencia temporal que se analiza puede aparecer el problema del desvanecimiento del gradiente, para solucionar esto se han incorporado unidades de memoria a largo plazo (LSTM, por sus siglas en inglés).

Las Redes Neuronales Recurrentes de Memoria a Largo Plazo son un tipo específico de RNN diseñadas para abordar el problema del desvanecimiento del gradiente. Su arquitectura incorpora mecanismos de puertas (puerta de entrada, puerta de olvido y puerta de salida) que regulan el flujo de información a través de la red, permitiendo aprender dependencias a largo plazo en secuencias de datos [13].

- **Puerta de entrada.** Determina qué porción de la nueva información será almacenada en la celda de memoria de la LSTM. Actúa como un filtro que regula el flujo de datos entrantes, permitiendo que solo los aspectos más relevantes sean incorporados al estado interno de la red.

- **Puerta de olvido.** Regula el grado en que se conserva o descarta la información almacenada en la celda de memoria de la LSTM. Actúa como un mecanismo de control que permite a la red olvidar información irrelevante o obsoleta, concentrándose en los aspectos más significativos para la tarea en cuestión.
- **Puerta de salida.** Determina qué parte de la información almacenada en la celda de memoria se utilizará para calcular la salida de la red en un instante de tiempo dado. Actúa como un filtro que selecciona los aspectos más relevantes de la información interna para generar la predicción final.

Las Redes Neuronales Recurrentes de Memoria a Largo Plazo (LSTM) han demostrado ser altamente efectivas en una amplia gama de tareas, incluyendo el Procesamiento del Lenguaje Natural, el reconocimiento de escritura a mano, el reconocimiento de voz, el reconocimiento de gestos y el análisis de secuencias de imágenes y video.

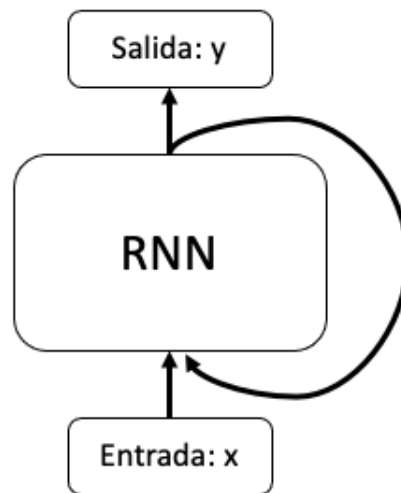


Figura 2.2: Arquitectura básica de una Red Neuronal Recurrente.

2.5.1. Grandes Modelos de Lenguaje

Los Grandes Modelos de Lenguaje (LLM), como BERT, Llama o GPT, son sistemas de inteligencia artificial basados en aprendizaje profundo capaces de generar texto humano de alta calidad. Estos modelos han sido pre-entrenados en grandes cantidades de datos de texto por lo regular de dominio general, lo que les permite aprender las complejidades y patrones del lenguaje. Gracias a su capacidad para generar texto coherente y contextualmente relevante, los LLM se están utilizando en una amplia variedad de aplicaciones, desde chatbots, asistentes virtuales, clasificadores de texto, entre otras. Además, estos modelos se basan en la arquitectura Transformer que fue presentada para resolver tareas de PLN en el 2017 en el artículo “Attention Is All You Need” [1] y se puede observar en la Figura 2.3. En este trabajo se presenta una novedosa arquitectura de red neuronal que se basa exclusivamente en mecanismos de atención. A diferencia de las redes recurrentes y convolucionales tradicionales, esta propuesta elimina por completo la recurrencia y las convoluciones, ofreciendo una alternativa innovadora para el procesamiento de secuencias. El Transformer introduce una arquitectura de red neuronal revolucionaria, prescindiendo de las operaciones recurrentes y convolucionales tradicionales. Su diseño se basa en la exclusiva utilización de mecanismos de autoatención, lo que le permite capturar relaciones de larga distancia entre los elementos de las secuencias de entrada y salida. Sus principales componentes son el codificador y decodificador.

- Codificador (izquierda): El codificador del Transformer procesa la secuencia de entrada y construye una representación vectorial de alto nivel, capturando las relaciones semánticas y sintácticas entre los elementos de la secuencia. De esta manera, el modelo adquiere una profunda comprensión del significado y contexto de la entrada.
- Decodificador (derecha): El decodificador del Transformer utiliza la representación contextual generada por el codificador como insumo para generar secuencialmente la secuencia de salida deseada. Este proceso involucra un mecanismo de atención que permite al decodificador enfocarse en las partes relevantes de la representación del codificador en cada paso de la generación, asegurando así la coherencia y relevancia de la salida.

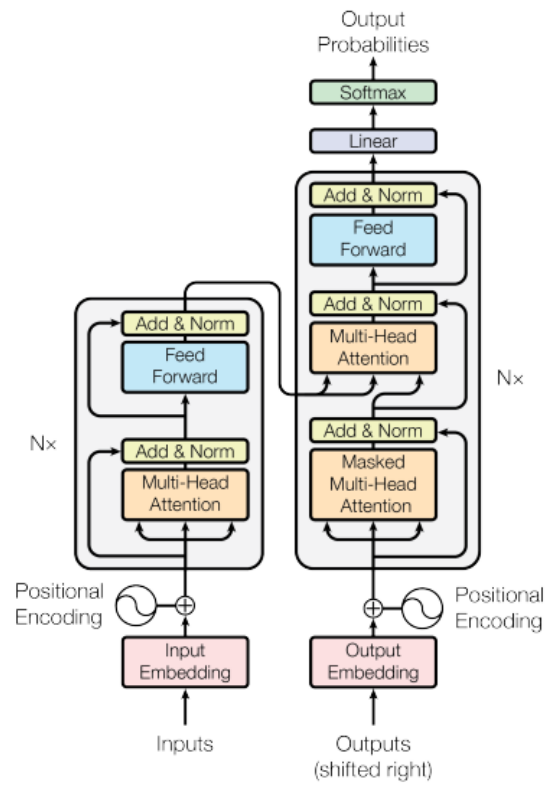


Figura 2.3: Arquitectura Transformer. Obtenida del artículo “Attention Is All You Need” [1].

Capítulo 3

Marco de referencia

En este capítulo se presenta una revisión exhaustiva de los trabajos relacionados con el desarrollo de herramientas computacionales para responder a preguntas formuladas en lenguaje natural. La revisión se organiza en tres categorías: Sistemas de Pregunta-Respuesta Generales, Recuperación de Información y Extracción de Respuestas.

Al examinar estas tres categorías, la revisión pretende proporcionar una comprensión completa del estado actual de la técnica en la respuesta a preguntas en lenguaje natural. Los conocimientos adquiridos a partir de esta revisión sirvieron de base para el desarrollo del algoritmo de respuestas a preguntas propuesto en este trabajo de investigación.

3.1. Trabajos relacionados con los Sistemas de Pregunta-Respuesta Generales

En la siguiente sección se describen los trabajos relacionados con el enfoque de algoritmos de pregunta-respuesta (QA).

En el año 2020 se desarrolló un sistema llamado *COVIDASK* [14] que combina técnicas de minería de textos biomédicos, técnicas de QA en tiempo real y enfoques de recuperación de información (IR) para proporcionar respuestas en lenguaje natural. La evaluación de *COVIDASK* se lleva a cabo utilizando un conjunto de datos creado manualmente por expertos en el dominio llamado “*COVID-19 Questions*” que se basa en información de varias fuentes, incluyendo el Centro para el Control y Prevención de Enfermedades (CDC) y la Organización Mundial de la salud (OMS).

En [15] los autores proponen un sistema semántico de respuesta exactas e ideales a preguntas biomédicas realizadas en lenguaje natural. El sistema consta de los módulos de clasificación de preguntas, recuperación de documentos, recuperación de pasajes y extracción de respuestas.

El sistema *SemBioNLQA*, se basa en patrones léxico-sintácticos elaborados a mano y un algoritmo de aprendizaje automático para la clasificación de preguntas, también utiliza el motor de búsqueda de PubMed y similitud de conceptos del Sistema Unificado de Lenguaje Médico para la recuperación de documentos, el modelo de clasificación BM25, los sinónimos de BioPortal, el análisis de sentimientos y la métrica de frecuencia para la extracción de respuestas. La evaluación manual de este sistema consta de 30 preguntas de las cuales 10 son de tipo hecho, 11 de lista y 9 de sí o no. *SemBioNLQA* obtuvo 18 respuestas correctas, es decir, una precisión del 60 %.

En [16] se presenta un sistema para encontrar respuestas a preguntas en textos biomédicos *Bio-AnswerFinder*. Utiliza una distancia ponderada de palabras o frases aprendidas de los resúmenes de PubMed y clasifica las respuestas después de categorizar el tipo de la pregunta.

Este enfoque recupera documentos relevantes de manera iterativa a través de consultas, utilizando palabras clave mejoradas de un motor de búsqueda tradicional. Para mejorar el rendimiento de la recuperación de documentos, introduce una red neuronal supervisada de memoria a corto plazo, para seleccionar las palabras clave de la pregunta y facilitar la búsqueda iterativa de las respuestas. El sistema fue evaluado por tres expertos, otorgándole una precisión del 71 %, 56 % y 66 % respectivamente.

AskHERMES [17] es un sistema de pregunta respuesta en línea para preguntas clínicas complejas, por ejemplo, ¿Cuál es la tasa de mortalidad de los Síndromes Coronarios Agudos en 1994? El sistema realiza un análisis semántico de las preguntas con la finalidad de producir resúmenes extractivos centrados en la pregunta como respuesta. El sistema recupera las respuestas de resúmenes de MEDLINE, artículos de texto completo de PubMed, documentos de eMedicine, además, guías clínicas y artículos de Wikipedia.

AskHERMES recibe una pregunta en lenguaje natural, después realiza un análisis de la pregunta para extraer automáticamente las necesidades de información, posteriormente produce una lista de términos de consulta utilizando el recurso de conocimiento UMLS, finalmente, devuelve una serie de documentos que coinciden con la búsqueda e identifica los pasajes relevantes.

Utiliza 4654 preguntas clínicas como datos de entrenamiento y prueba y responde a las preguntas en un tiempo promedio de 20 segundos. *AskHER-*

MES fue evaluado con 60 preguntas realizadas por 3 expertos y en una escala del 1 al 5 obtuvo los siguientes resultados: facilidad de uso 4, calidad de la respuesta 2.5, rendimiento general 3.

En [18] los autores proponen un prototipo de sistema pregunta respuesta para el dominio académico, las preguntas como la base de conocimiento están en español. El sistema ha sido completamente desarrollado desde cero y combina herramientas superficiales de procesamiento del lenguaje natural con técnicas de redundancia de datos. Los autores concluyen que debido a la simplicidad del enfoque que utilizan, el sistema arrojó un 43.5 % de respuestas correctas.

CAiRE-COVID [19] Es un sistema de respuestas a preguntas relacionadas con el COVID-19, las preguntas son de tipo resumen y realiza las consultas en múltiples documentos médicos. Este proyecto ganó una de las 10 tareas en el desafío de conjuntos de datos de investigación abiertos *Kaggle COVID-19*. El sistema combina la extracción de información con técnicas de garantía de calidad, además, suministra la respuesta uniendo múltiples pasajes de distintos documentos para proporcionarle al usuario la información más relevante relacionada con su consulta.

En [20] se presenta una plataforma integrada de múltiples fuentes para responder a preguntas clínicas. *MiPACQ* realiza un procesamiento de preguntas basado en la semántica y la extracción de información. Utilizando técnicas PLN y Recuperación de Información (RI). La plataforma acepta preguntas clínicas de texto libre y presenta al usuario respuestas breves de fuentes como los recursos enciclopédicos médicos generales y los datos del paciente que residen en la historia clínica electrónica, además, de implementar un sistema basado en aprendizaje automático obteniendo una mejora notable en comparación de sus versiones anteriores.

En 2015 los autores [21] proponen un *framework* novedoso de análisis semántico para responder preguntas utilizando una base de conocimientos. El análisis semántico se reduce a la generación de consulta de grafos, formulado como un problema de búsqueda por etapas.

El método que utilizan es aprovechar la base de conocimientos desde una etapa temprana para reducir el espacio de búsqueda y con esto simplificar el problema de coincidencia semántica. También utilizan un modelo de red neuronal convolucional para hacer coincidir a las preguntas con las secuencias de los predicados. Este sistema obtuvo en sus primeras pruebas una precisión del 52.8 %.

En el trabajo [22] elaborado en el 2017 proponen una traducción de una

pregunta en lenguaje natural (NLQ) para hacer consultas en un grafo de conocimiento, con el reto de entender con claridad la pregunta y pasarla a una consulta estructurada, este sistema obtuvo una precisión del 84 % utilizando el conjunto de pregunta respuesta *QALD*. Un trabajo similar se puede consultar en el artículo [23] donde presentan el sistema *Sematch* para la búsqueda de entidades en un grafo de conocimiento que combina procesamiento de consultas en lenguaje natural, vinculación de entidades, consultas basadas en enlaces y similitudes semánticas. El sistema ha sido validado en un conjunto de datos y se ha desarrollado un prototipo que traduce consultas de lenguaje natural a *SPARQL* y tiene una precisión del 76 %.

En el desarrollo del sistema [24] se traducen preguntas de voz a texto en lenguaje natural y se localizan las respuestas en una ontología del dominio académico, utilizando patrones estructurales de *SQWRL*, se realiza una clasificación de las preguntas de tipo persona, evento y espacio físico con una precisión cercana del 97 %.

Finalmente, los autores del trabajo realizado en [25] presentan *AUEB*, que participó en la competencia *BioASQ* que se enfoca en sistemas de pregunta y respuesta en el dominio biomédico. *AUEB* utiliza como enfoque dos métodos de aprendizaje profundo, una red neuronal convolucional básica y posit-drmm obteniendo resultados alrededor del 75 % de precisión.

A continuación, se muestra en la Tabla 3.1 una comparativa con los diversos trabajos analizados para la tarea de pregunta-respuesta.

Tabla 3.1: Estado del arte

Sistema QA/ referencia	Idioma	Formato de la Pregunta	Tipo de preguntas	Tipos de datos de entrada	Fuentes de datos	Dominio	Enfoque	Evaluación
CodAsk [14]	Inglés	Lenguaje natural	Oraciones interrogativas y basadas en palabras clave	Semiestructurado y No estructurado	Cord19	Covid19	Minería de textos	Análisis cuantitativo, cualitativo
SemBioNLQA [15]	Inglés	Lenguaje natural	Sí/no, hechos, lista y preguntas de resumen	No estructurado y Semiestructurado	PubMed BioASQ	Biomédico	Aprendizaje automático	Evaluación manual, precisión, exhaustividad, medida F
Bio-AnswerFinder [16]	Inglés	Lenguaje natural	Sí/no, hechos, lista y resumen	No estructurado y Semiestructurado	PubMed BioASQ	Biomédico	Red neuronal	Evaluación a 936 preguntas utilizando precisión
AskHERMES [17]	Inglés	Lenguaje natural	Resumen	No estructurados	MEDLINE PubMed eMedicine Wikipedia	Biomédico	Máquinas de Soporte Vectorial	Manual por expertos: facilidad, tiempo, rendimiento y calidad de la respuesta.
Vicedo et al [18]	Español	Plantilla de pregunta	Preguntas que requieren respuestas cortas (número, fecha y otra)	No estructurados	Documentos locales y Documentos web	Académico	PLN Técnicas de redundancia de datos	Rango recíproco medio (MRR)
CAiRE-COVID [19]	Inglés	Lenguaje natural Palabras clave	Resumen	No estructurado y Semiestructurado	Cord19 PubMed	Covid19	Red Neuronal Artificial	Rango recíproco medio (MRR)

Sistema QA/ referencia	Idioma	Formato de la Pregunta	Tipo de preguntas	Tipos de datos de entrada	Fuentes de datos	Dominio	Enfoque	Evaluación
MiPACQ [20]	Inglés	Lenguaje natural	Resumen	No estructurado y semiestructurado	Medpedia EMR Clinques	Médico	Red neuronal Artificial	Precisión, Rango recíproco medio (MRR) y análisis cualitativo
STAGG [21]	Inglés	Lenguaje natural	Hechos	Estructurado y semiestructurado	Freebase WEBQUESTIONS	Dominio abierto	Redes Neuronales Artificiales	Cantidad de preguntas 2032. Métricas: precisión, exhaustividad, medida F
Zheng et al [22]	Inglés	Lenguaje natural	Oraciones interrogativas y basadas en palabras clave	Estructurado y semiestructurado	QALD	Dominio abierto	SPARQL	Precisión y Exhaustividad
SEMATCH [23]	Inglés	Lenguaje natural	Hechos	Estructurado y semiestructurado	DBpedia INEX XER	Entretenimiento (personas, música, películas, lugares y videojuegos)	SPARQL y Similitud Semántica	Precisión y exhaustividad
García-Robledo et al [24]	Español	Plantilla de pregunta	Interrogativas (tiempo, espacio y persona)	Estructurado	Ontología	Académico	SQWRL	Cantidad de preguntas, precisión
AUEB [25]	Inglés	Lenguaje natural	Hechos	Estructurado y semiestructurado	PubMed BioASQ	Biomédico	Red Neuronal Convolutcional	Precisión y MAP
Este trabajo	Español	Lenguaje natural	Hechos	No estructurado y semiestructurado	BioASQ CORD-19 CoWeSe Preguntas-respuestas COVID	Médico	Aprendizaje profundo y Enfoques probabilísticos, semánticos y estadísticos	Precisión, F1 y Exhaustividad

Después de la revisión de esta categoría se puede observar que los trabajos relacionados abarcan una amplia gama de enfoques para la respuesta a preguntas, incluyendo sistemas basados en reglas, métodos estadísticos y técnicas de aprendizaje automático. El objetivo de la revisión se centra en comprender los principios fundamentales y los desafíos que implica el desarrollo de algoritmos o sistemas de respuesta a preguntas eficaces, además se observa que existen dos grandes componentes para crear un algoritmo de pregunta-respuesta que incluyen a la Recuperación de Información y la extracción de respuestas.

3.2. Trabajos relacionados con la tarea de Recuperación de información

En esta sección se presenta una revisión de la literatura para la tarea de Recuperación de Información.

En [26], se realiza Recuperación de documentos del dominio biomédico y se plantean dos estrategias, la primera un etiquetado de documentos que combina varios modelos tradicionales como BM25 y de espacio vectorial. Además, utilizan un segundo enfoque basado en métodos de aprendizaje automático, debido a la abundancia de terminologías específicas en el dominio biomédico. En [27] se propone un Sistema de Recuperación de Información entre idiomas (CLIR), devuelve una respuesta escrita en un idioma diferente al de la consulta, en este caso entre árabe e inglés. Utiliza el Corán como conjunto de datos, además realiza una técnica de expansión de la consulta para mejorar el impacto en el rendimiento de los métodos de IR (booleano, espacio vectorial y BM25).

Los trabajos [28], [29] y [30] utilizan el conjunto de datos de investigación abierta del Instituto Allen, CORD-19. El primero presenta un Sistema de Recuperación de Información a consultas formuladas por expertos biosanitarios sobre la enfermedad COVID-19 y el virus causal SARS-CoV-2. El enfoque utilizado para la IR está basado en modelos de lenguaje, retroalimentación y redes bayesianas. En la segunda investigación se presenta el motor de búsqueda de información COVIDEX, que proporciona una infraestructura robusta basada en palabras clave en idioma inglés y técnicas de aprendizaje automático. El sistema ha sido evaluado en el desafío TREC-COVID y obtuvo las puntuaciones más altas. Finalmente, CO-Search, es un sistema de búsqueda

semántico de múltiples etapas, está diseñado para manejar consultas sobre el coronavirus, contiene un recuperador de palabras clave que toma una consulta de entrada y devuelve una lista ordenada de los documentos relevantes. Utiliza un enfoque de aprendizaje profundo junto con los modelos BM25 y TF-IDF. Para evaluar la efectividad del sistema de búsqueda se utilizó el conjunto de datos de preguntas de la competencia TREC.

En [31] se muestra un estudio de la influencia de la longitud de los textos en tareas de RI por medio de un modelo probabilístico de tópicos (MPT). La finalidad es reducir las dimensiones de los vectores en los modelos de RI para que las operaciones sean más sencillas sin comprometer el rendimiento de los sistemas. El trabajo [32] muestra un nuevo enfoque probabilístico para la RI basado en métodos de traducción automática estadística, la idea principal es convertir un documento a una consulta ideal. La evaluación de este trabajo utilizó la colección de documentos de la competencia TREC.

El sistema de recuperación de información médica [33] para la tarea de QA, incorpora asociaciones semánticas y aplica un enfoque heurístico para la retroalimentación de las consultas para mejorar la recuperación en bases de conocimiento estructuradas. Se utilizó la base de datos digital de PubMed. En [34] se comparó el desempeño del modelo BM25 contra el booleano y el de espacio vectorial, los resultados arrojaron que los documentos mejor clasificados como relevantes se obtuvieron con el algoritmo de ranking basado en probabilidades OkapiBM25.

CO-SE [35] es un motor de búsqueda de respuestas sobre un corpus de publicaciones científicas. Contiene un sistema algorítmico que devuelve documentos relevantes a consultas sobre COVID-19, está basado en TF-IDF y un modelo de Transformers, su precisión es de 71.45 %.

En [36] se realiza una Recuperación de Información multimedia (imagen, audio, video y texto). La recuperación de documentos se realiza con los modelos de espacio vectorial, LSI y un enfoque de redes neuronales. Finalmente, en [37] se realiza una comparativa para evaluar el desempeño de los algoritmos de espacio vectorial, probabilístico y LSI. Los resultados indican que la técnica de LSI tiene una baja precisión y exhaustividad, en comparación de los modelos probabilístico y vectorial.

Se presenta a continuación una Tabla 3.2 comparativa que resume las características más relevantes de los trabajos relacionados con la tarea de Recuperación de Información.

Tabla 3.2: Estado del arte de Recuperación de Información.

Referencia	Idioma	Fuente de datos	Enfoque	Evaluación
Bo Xu 2018 [26]	Inglés	Documentos Biomédicos TREC	Aprendizaje Automático, modelo BM25	NA
CLIR 2021 [27]	Árabe	Documentos del Corán	Modelo booleano, vectorial y BM25	Precisión 72 %
Otegi 2022 [28]	Inglés	Documentos médicos CORD-19	LLM y redes bayesianas	Exhaustividad 85 %
Covidex 2020 [29]	Inglés	Documentos médicos CORD-19	Pyserini, BM25	Precisión 82 %
CO-Search 2021 [30]	Inglés	Documentos médicos CORD-19	BM25 y TF-IDF	Precisión 84 %
C. Badenes 2021 [31]	Inglés y español	Textos legislativos JRCAcquis	Modelos probabilísticos de Tópicos	Precisión 84 %
A. Berger 1999 [32]	Inglés	Noticias TREC	TF-IDF, Modelo binomial	Precisión 62 %
Wang 2017 [33]	Inglés	Bases de conocimiento médicas	Algoritmos Heurísticos	Precisión 41 %
G. A. Tinega 2018 [34]	Inglés	Artículos científicos	Modelo booleano, vectorial, BM25 y TF-IDF	Exhaustividad 78 %
CO-SE 2022 [35]	Inglés	Artículos científicos	Redes neuronales y TF-IDF	Precisión 71.45 %
Mahmood 2019 [36]	Inglés	Imágenes, audio y video Flickr	Modelos de espacio vectorial, LSI y redes neuronales	Exhaustividad 61.5 %
V. Rodríguez 2019 [37]	Inglés	Documentos de turismo	Modelos de espacio vectorial LSI y BM25	Exhaustividad 34 %

Después de la revisión de esta categoría en el estado del arte, se puede observar que los artículos referentes a la recuperación de información exploran técnicas para localizar de manera eficiente documentos relevantes que puedan contener las respuestas a las consultas de los usuarios. El objetivo de la revisión es comprender cómo los métodos y enfoques de recuperación de información pueden aprovecharse para apoyar los sistemas de respuesta a preguntas.

3.3. Trabajos relacionados con la extracción de respuestas

El análisis de los trabajos relacionados con los enfoques de respuesta a preguntas reveló una tendencia predominante hacia el empleo de técnicas de aprendizaje profundo, particularmente grandes modelos de lenguaje neuronales, para la extracción de respuestas. En consecuencia, esta sección se dedica a explorar estos modelos en profundidad, examinando los datos de entrenamiento utilizados, las tareas de procesamiento del lenguaje natural que pueden abordar y su eficiencia en general. La revisión abarca diversas técnicas de extracción de información y sus aplicaciones en el ámbito de los sistemas de respuesta a preguntas.

En [38] se describe el modelo *Transformer* BioBert el cual es un modelo de representación del lenguaje pre-entrenado con BERT, y adaptado a corpus biomédicos en inglés. Según los autores este nuevo modelo supera ampliamente a BERT y a modelos de última generación en tareas de minería de textos biomédicos como: reconocimiento de entidades nombradas biomédicas (mejora de la puntuación F1 en un 0,62 %), extracción de relaciones biomédicas (mejora de la puntuación F1 en un 2,80 %) y respuesta a preguntas biomédicas (mejora de la MRR en un 12,24 %). BioBERT fue el primer modelo basado en BERT de dominio específico preentrenado en corpus biomédicos durante 23 días. Los corpus utilizados para este modelo fueron obtenidos de resúmenes de artículos científicos del repositorio médico PubMed y artículos completos del repositorio PMC. El trabajo [39] desarrolla y evalúa una representación continua de las notas clínicas mediante transformadores bidireccionales (ClinicalBERT). La finalidad es predecir el reingreso hospitalario de los pacientes en 30 días. La base es el modelo BERT, que está preentrenado con un corpora estándar, de Wikipedia y BookCorpus. Posteriormente se ajusta

el modelo utilizando 2,083,180 de notas clínicas que contienen resúmenes de alta y el historial de los primeros días en la unidad de cuidados intensivos. De la misma forma en el dominio médico los autores del trabajo [40] proponen adaptar al modelo Roberta con textos de la literatura biomédica y clínica para el español, con la intención de potenciar las aplicaciones de PLN del español nativo en biomedicina. Los autores de RobertaClinical confirman que el preentrenamiento específico del dominio es fundamental para obtener mejores resultados en tareas específicas como NER y reportan que obtuvieron un F-score del 88.21 %. Asimismo, en el trabajo [41] se presenta a med-BERT que también reconoce entidades médicas con un F1 del 93 % y fue entrenado con textos clínicos en inglés y chino. En los trabajos [42] y [43] se presentan modelos para facilitar la investigación en el desarrollo de representaciones lingüísticas en el área de la biomedicina y son sometidos a una evaluación de cinco tareas para la comprensión del lenguaje biomédico (BLUE y BLURB). Por otro lado, SciBERT [44] es un modelo adaptado a las publicaciones científicas del repositorio Semantic Scholar, el modelo está entrenado con 18 % de artículos que corresponden a ciencias de la computación y el 82 % a biomedicina, obteniendo un modelo híbrido que fue evaluado con tareas para la extracción de información. En el artículo [45] se presenta a MarIA que consta de un compendio de experimentos con varios modelos de lenguaje en español como RoBERTa-base, RoBERTa-large [46], GPT2 y GPT2-large [47], los cuales fueron pre-entrenados utilizando un corpus con 135 mil millones de palabras extraídas del Archivo Web del Español construido por la Biblioteca Nacional de España entre los años 2009 y 2019, se evalúa el rendimiento de los modelos utilizando nueve diferentes corpus de datos para diversas tareas de PLN en el dominio general, así como un conjunto de pregunta-respuesta (SQAC) obtenido de artículos de Wikipedia en español, enciclopedias y de Wikinoticias. Los mejores resultados en los experimentos arrojan un 88.51 % de F1 en RoBERTa-base para la tarea de NER y un 98.56 % de F1 para el modelo RoBERTa-large en la tarea de POS. MarIA obtuvo una perplexidad en el conjunto de evaluación de 10.08 para el modelo basado en GPT2 y del 6.68 para el modelo GPT2-Large. En [48] presenta el modelo BETO que se compone de 12 capas de auto atención, 12 cabezas de atención cada una, una capa oculta de tamaño 1024 y fue entrenado con un total de 110 millones de parámetros. Este modelo se pre-entrenó con los textos de la parte en español de Wikipedia y con un fragmento del proyecto OPUS19. En BETO los autores reportan que su modelo supera a la variante de BERT multilingüe en las tareas de etiquetado POS, NER e Inferencia de Lenguaje Natural (XNLI)

obteniendo un F-score de 88.43 %, exactitud del 98.97 % y del 82.01 % para cada tarea. Por su parte en los trabajos BERTIN [49] y ELECTRICIDAD que son grandes modelos de lenguaje entrenados con textos de dominio general en español, por ejemplo, BERTIN se entrenó desde cero con la parte en español del conjunto mC4 y ELECTRICIDAD fue entrenado con el corpus OSCAR 20. BERTIN obtuvo un F-score para la tarea de NER del 87.79 % y del 96.44 % para el etiquetado POS. Por otro lado, el modelo de lenguaje ELECTRICIDAD obtuvo un 80.5 % de F-score para la tarea de NER, un 98.16 % para POS y una exactitud del 78.78 % para XNLI. En los trabajos de [50], [51] y [52] se presentan modelos de lenguaje basados en BERT y RoBERTa entrenados en idioma francés, holandés e italiano respectivamente y que se describe a continuación, FlauBERT, fue adaptado al francés utilizando un corpus amplio y heterogéneo, además fue aplicado a diversas tareas de PLN como: Clasificación de Textos (CLS), paráfrasis (PAWS-X), inferencia del lenguaje natural, análisis sintáctico y desambiguación del sentido de las palabras (WSD), y lo evalúan con el *benchmark* FLUE. Por su parte RobBERT, fue ajustado para aumentar el rendimiento de las tareas de PLN específicamente en neerlandés, además los autores también evalúan la importancia de los tokenizadores para el entrenamiento de los LLM, en el proceso de entrenamiento utilizan el optimizador Adam y una tasa de aprendizaje preestablecida en 0.000001. Por último, el modelo AlBERTo se centra en el lenguaje utilizado en las redes sociales en italiano, concretamente en Twitter. Los autores avalúan a AlBERTo en la competencia EVALITA 2016 en la tarea de SENTIPOLC (*SENTiment POLarity Classification*) situándose entre los mejores resultados en la detección de polaridad, subjetividad e ironía en tuits italianos.

A continuación, se muestra en la Tabla 3.3 una comparativa con las características más relevantes de los grandes modelos de lenguaje. Este trabajo de investigación se enfoca en comparar el idioma de los textos con los que fue entrenado, la fuente de los datos, el dominio o área de aplicación del modelo, las tareas que puede resolver y su evaluación.

Después de la revisión de los trabajos relacionados y dada la capacidad observada de los grandes modelos de lenguaje, en este trabajo de investigación se propone adaptar LLM basado en la arquitectura Transformers al dominio médico para realizar la extracción de las respuestas.

Tabla 3.3: Estado del arte de modelos de lenguaje

Modelo/Referencia	Idioma	Conjunto de datos	Dominio	Tarea	Evaluación para cada Tarea
BioBERT [38]	Inglés	Resúmenes PubMed, artículos PMC	Médico	-REL -QA	-F1 86.51 % -F1 60.0 %
ClinicalBERT [™] [39]	Inglés	MIMIC-III	Clínico/Médico	-Predicción de reingreso hospitalario	-AUROC 71.4 %
RobertaClinical [40]	Español	Medical crawler Clinical cases mespen_Medline PubMed	Clínico/Médico	-NER	-F1 88.21 %
Med-BERT [41]	Inglés y chino	Registros Clínicos	Clínico	-NER	-F1 93 %
BlueBERT [42]	Inglés	PubMed MIMIC-III	Biomédico	-Similitud de oraciones -REL -NER	-Pearson 84.8 % -F1 74.4 % -F1 86.6 %
PubMedBERT [43]	Inglés	PubMed	Biomédico	-NER - QA	-F1 85.62 % -F1 55.84 %
SciBERT [44]	Inglés	Semantic Scholar	Computación, Biomédico	-NER - CLS -REL	-F1 67.57 % -F1 70.98 % -F1 79.97 %
ELECTRICIDAD [45]	Español	OSCAR 20	General	-POS -NER	-F1 98.16 % -F1 80.5 %
ROBERTA [46]	Español	Wikipedia, Wikinews	General	-NER -POS	-F1 88.51 % -F1 98.56 %
BETO [48]	Español	Wikipedia OPUS19	General	-POS	-Exactitud 82.01 %
BERTIN [49]	Español	mC4	General	-NER -POS	-F1 87.79 % -F1 96.44 %
FlauBERT [50]	Francés	Wikipedia, Libros, Common Craw	General	-PAWS-X -WSD	-Exactitud 89.34 % -Exactitud 50.48 %
RobBERT [51]	Neerlandés	OSCAR, Common Crawl, SoNaR-500	General	-Análisis de sentimientos -NER	-F1 94.378 % -F1 89.08 %
ALBERTO [52]	Italiano	TWITA	Redes sociales	-Detección de ironía -Análisis de sentimientos -Clasificación de subjetividad	-F1 60.90 % -F1 72.23 % -F1 79.06 %

Capítulo 4

Metodología

Este capítulo detalla el desarrollo del algoritmo de pregunta-respuesta en textos médicos en español. El algoritmo se basa en la metodología que se observa en la Figura 4.1, la cual comprende cuatro fases principales: pre-procesamiento de textos, análisis de preguntas, recuperación de documentos relevantes y extracción de respuestas. Estas fases fueron diseñadas tras una exhaustiva revisión de trabajos relacionados, con el objetivo de identificar y contribuir a las técnicas y enfoques más relevantes para abordar el problema planteado. Las fases principales de la metodología se describen en las siguientes secciones.

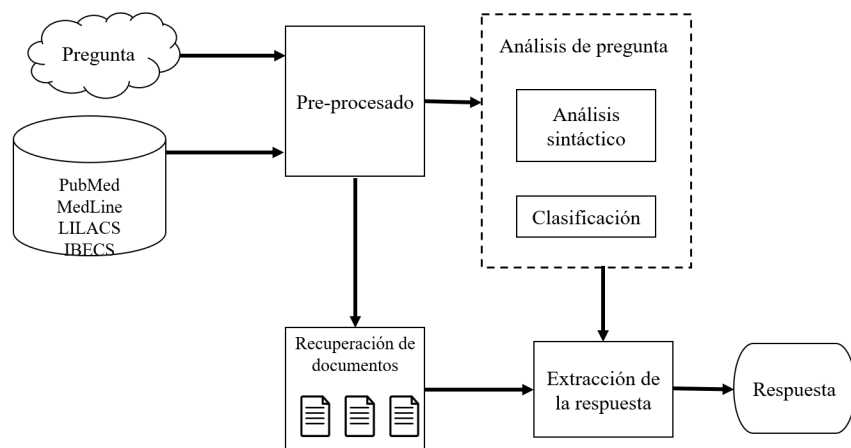


Figura 4.1: Metodología general para el algoritmo de pregunta-respuesta.

4.1. Conjuntos de datos

El desarrollo de este trabajo de investigación se basa en un conjunto de datos de textos médicos en formato no estructurado o semi estructurado, compuesto por más de tres millones de documentos recopilados de diversos repositorios médicos como PubMed, MedLine, LILACs e IBECS. Estos textos fueron utilizados en las diferentes fases de la construcción del algoritmo de respuesta a preguntas. Para cada fase del algoritmo, se describirá en detalle el conjunto de datos utilizado, incluyendo la cantidad de textos, el número de palabras, el vocabulario y el proceso aplicado.

4.2. Pre-procesado

La fase de pre-procesamiento tiene como objetivo limpiar y preparar los textos provenientes de los repositorios médicos para optimizar la búsqueda de documentos relevantes en respuesta a una consulta en lenguaje natural. Esta fase también se encarga de depurar la pregunta original, extrayendo los términos más importantes comúnmente denominados "foco de la pregunta". En esencia, esta fase prepara la información para las etapas posteriores del algoritmo de respuesta a preguntas.

Las tareas específicas realizadas en la fase de pre-procesamiento incluyen:

- **Tokenización.** Consiste en dividir los textos en unidades básicas o tokens, como palabras o multipalabras, esto ayudará para simplificar el procesamiento de los textos.
- **Limpieza.** Consiste en eliminar cualquier fragmento, palabra o carácter que no sea relevante para el contexto de los datos y que comúnmente se les conoce como ruido, incluye eliminar caracteres especiales o puntuación no necesaria.
- **Eliminación de palabras vacías.** Consiste en remover palabras utilizadas comúnmente en el lenguaje y que tiene un bajo contenido semántico.
- **Estandarización.** Consiste en llevar a los textos en una representación común como convertir todas las palabras a minúsculas para unificar formatos de texto inconsistentes.

- **Lematización.** Consiste en la reducción de las palabras a su forma canónica o lema, con el objetivo de representar a todas sus flexiones. En el caso de los sustantivos su forma singular y masculino y en los verbos su forma en infinitivo.
- **Derivación.** Consiste en identificar y agrupar palabras a partir de su raíz.

Es importante mencionar que cada una de las fases posteriores del algoritmo de respuesta a preguntas requiere un pre-procesamiento específico, adaptado a las necesidades y características de la tarea en cuestión. En cada fase, se describirán las tareas de pre-procesamiento particulares utilizadas.

4.3. Análisis de la pregunta

Esta fase se centra en el análisis de las preguntas para identificar sus componentes esenciales y prepararlas para las etapas posteriores del proceso. Se determinó que los elementos clave de una pregunta formulada en lenguaje natural son la partícula interrogativa y el foco de la pregunta.

Las partículas interrogativas son elementos importantes para clasificar las preguntas y determinar la respuesta adecuada. Cada partícula interrogativa, como “quién”, “dónde”, “cuándo”, “por qué”, “cuál”, “cómo” y “cuánto”, orienta la búsqueda de información hacia un tipo específico de respuesta, como “persona”, “lugar”, “tiempo”, “razón”, “causa”, “alternativa”, “forma” o “cantidad”. Esta clasificación se basa en la taxonomía de preguntas propuesta por [53]. La clasificación de preguntas que se realizó en este trabajo de investigación se muestra en la Tabla 4.1.

Tabla 4.1: Clasificación de preguntas realiza para el algoritmo de pregunta-respuesta.

Partícula Interrogativa	Tipo de Respuesta	Ejemplo de pregunta
¿Qué?	Definición, cantidad, localización, hechos (acciones, eventos o acontecimientos), temporalidad (fecha), título, sustantivo (sustantivos).	¿Qué es el síndrome inflamatorio multisistémico del COVID19 en niños?
¿Quién?	Persona, agrupamiento de personas por características afines.	¿Quién recibió el Premio Nobel de Desarrollo de CRISPR?
¿Cómo?	Proceso (desarrollo), manera.	¿Cómo identificar los casos de COVID-19 clínicamente?
¿Dónde?	Localización.	¿Dónde se encuentra la base J en el genoma de Leishmania tarentolae?
¿Cuándo?	Temporalidad (fecha).	¿Cuándo se usó ceritinib en lugar de crizotinib?
¿Cuál?	Lista de hechos (acciones, eventos o acontecimientos), sustantivo (sustantivos), localización, organización, cantidad.	¿Cuál es el mecanismo de acción de la decitabina?
¿Por qué?	Razón (causa o motivo).	¿Por qué daña a los pulmones la enfermedad del COVID-19?
¿Cuánto?	Cantidad.	¿Cuántos pares de bases hay en el genoma mitocondrial humano o ADNmt?

EL foco o núcleo de la pregunta es el concepto más importante para realizar la búsqueda de documentos candidatos donde pueda estar la respuesta. En este trabajo, se utiliza como foco de la pregunta a los sintagmas nominales. Para identificar estos sintagmas, se realizó un análisis sintáctico superficial de las preguntas empleando la herramienta Freeling ¹ y en la Figura 4.2 se ilustra un ejemplo de este análisis sintáctico superficial para una pregunta en español.

La elección de Freeling para el análisis sintáctico se sustenta en una comparación detallada de cinco herramientas de PLN y los resultados se presentan en el Anexo 1.

Tras identificar el núcleo de la pregunta, se eliminan los demás componentes mediante un procesamiento de texto que incluye la tokenización y eliminación de ruido (todo lo que no contenga a los sintagmas). A continuación, se presentan ejemplos de preguntas del dominio médico junto con sus respectivos núcleos, los cuales servirán como entradas para la búsqueda de documentos relevantes.

Pregunta: *¿Cómo diagnosticar casos graves de COVID-19?*

Foco de la pregunta: *casos graves de covid-19*

Pregunta: *¿Cuándo se usó ceritinib en lugar de crizotinib?*

Foco de la pregunta: *ceritinib crizotinib*

¹<https://nlp.lsi.upc.edu/freeling/demo/demo.php>

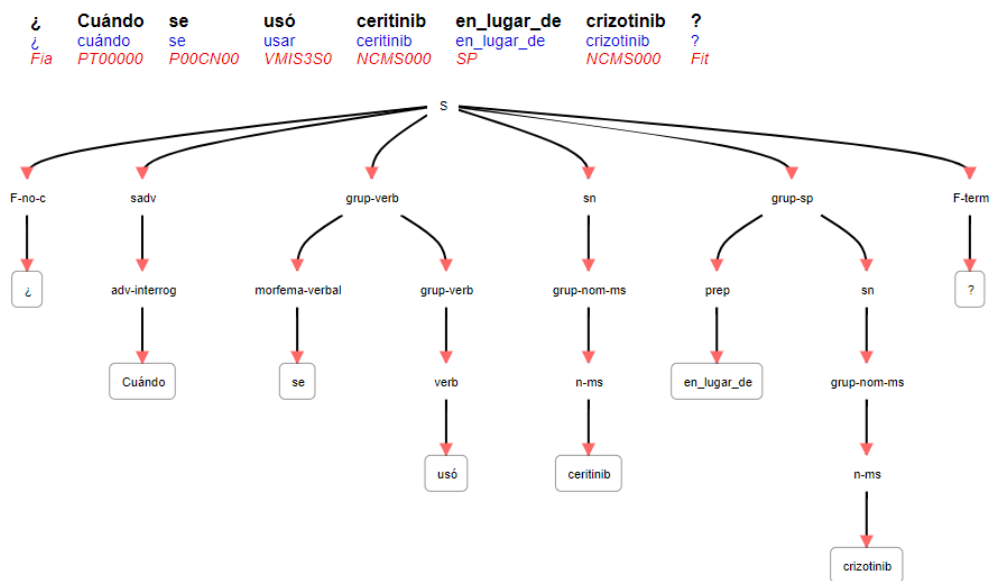


Figura 4.2: Análisis sintáctico de una pregunta generado con la herramienta Freeling.

4.4. Recuperación de documentos científicos

La fase de Recuperación de Documentos es una de las etapas más importantes del algoritmo de pregunta-respuesta, ya que en esta es donde se identifican los documentos médicos pertinentes a partir de una pregunta formulada en lenguaje natural. Para conseguir el objetivo, se diseñó e implementó una Sistema de Recuperación de Información (SRI) basado en una metodología propuesta que opera sobre un conjunto de documentos médicos y una serie de peticiones de información. El SRI emplea un mecanismo de similitud de textos para determinar qué documentos satisfacen las necesidades de información expresadas en las peticiones. En la Figura 4.3 se muestran los componentes del SRI que incluyen un preprocesado particular de artículos científicos, un recuperar de documentos que adapta modelos de RI probabilísticos, vectoriales y estadísticos, además se realiza una evaluación del SRI con un conjunto de 100 preguntas para determinar la precisión del sistema propuesto.

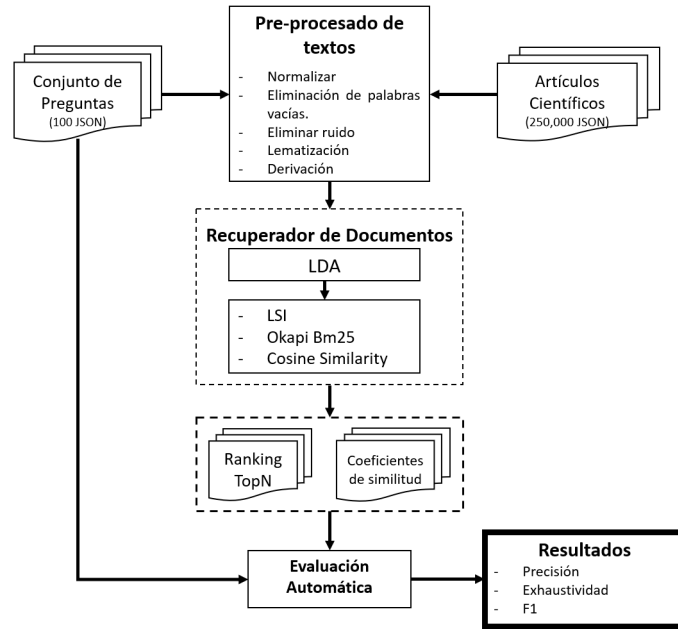


Figura 4.3: Recuperación de documentos científicos médicos.

4.4.1. Artículos científicos médicos

Para responder a las preguntas del ámbito médico, este proyecto de investigación utilizó artículos científicos extraídos del conjunto de datos de la tarea MESINESP de la competencia BioASQ [54]. La Tarea BioASQ MESINESP, patrocinada por la Secretaría de Estado para el Avance Digital (SEAD) y el Plan de Impulso de las Tecnologías del Lenguaje (Plan TL), ofrece un valioso conjunto de datos para la investigación en el área del PLN aplicado al dominio Médico.

El conjunto de datos de BioASQ tiene más de 249 mil resúmenes de artículos científicos en español, extraídos de diversos repositorios médicos como PubMed, MedRxiv y OMS. Estos resúmenes, en formato JSON, están etiquetados con los siguientes campos: id, title, year, abstractText, db y decsCodes. La Figura 4.4 ilustra un ejemplo de este conjunto de datos.

```
{
  "articles": [
    {
      "id": "ibc-ET6-1764",
      "title": "La Quercetina como un potencial  
nutracéutico contra la enfermedad por coronavirus 2019 (COVID-19)",
      "abstractText": "INTRODUCCIÓN: La enfermedad del coronavirus 2019  
(COVID-19) es una enfermedad viral que afecta a varios órganos y sistemas.  
Los tratamientos preventivos o profilácticos son especialmente útiles en  
enfermedades infecciosas emergentes como COVID-19 porque reducen la  
necesidad de hospitalización y el gasto en salud pública. Aunque el efecto  
preventivo del SARS-CoV-2 de varios agentes terapéuticos (e.g.,  
hidroxicloroquina/cloroquina, remdesivir, lopinavir y ritonavir) se ha  
evaluado ampliamente, ninguno de ellos ha demostrado una gran eficacia  
clínica. MÉTODO: Por lo tanto, aquí nuestro objetivo es abordar y discutir  
los estudios publicados recientemente sobre el potencial quimioprofiláctico  
de la quercetina contra el SARS-CoV-2. METODOLOGÍA: Se realizó una  
búsqueda de la literatura en bases como PubMed/MEDLINE, Scielo, Scopus,  
Web of Science, Cochrane Library y Clinical Trials.gov. Se incluyeron y  
evaluaron críticamente estudios que abordan la quercetina contra el SARS-  
CoV-2 u otros tipos de coronavirus. RESULTADOS: Algunos estudios han  
demostrado que la quercetina, un flavonoide aprobado por la FDA que se  
utiliza como agente antioxidante y antiinflamatorio, inhibe la entrada del  
coronavirus (SARS-CoV) en la célula huésped. Además, un estudio in silico  
mostró que la quercetina es un potente inhibidor de la proteasa principal  
del SARS-CoV-2 (Mpro), lo que sugiere que este flavonoide también es activo  
contra COVID-19. CONCLUSIONES: Debido a que la quercetina podría prevenir  
y disminuir la duración de las infecciones por SARS-CoV-2, es plausible  
suponer que el uso profiláctico de este flavonoide produce varios  
beneficios clínicos. Pero, estas pruebas preliminares deben ser  
confirmadas mediante ensayos in vitro y, posteriormente, en un ensayo  
clínico aleatorizado",
      "journal": "Ars pharm",
      "year": 2021,
      "db": "IBECs",
      "decsCodes": ["D006801", "D018352", "D015203", "D019587",  
"D065129", "D000998", "D011480", "D011794"]
    }
  ]
}
```

Figura 4.4: Extracto de un artículo científico del corpus BioASQ MESINESP.

4.4.2. Pre-procesado de artículos científicos

En esta subsección se presenta el pre-procesado de textos para la Recuperación de Información. Para optimizar la recuperación de artículos científicos relevantes, es necesario reducir la cantidad de datos que deben procesarse y como consecuencia disminuir el espacio de búsqueda. Esto se logra mediante el pre-procesamiento de textos, una etapa que minimiza el tiempo de procesamiento y optimiza los recursos computacionales.

Para el pre-procesado de artículos científicos se utilizaron diversas tareas de PLN, como la tokenización, eliminación de palabras vacías, normalización de los textos y eliminación de caracteres especiales y signos de puntuación. Además, se aplicaron técnicas de lematización y derivación para eliminar información irrelevante y redundante de los textos.

- **Eliminación de palabras vacías.** Se eliminaron preposiciones, artículos, pronombres, conjunciones y otras palabras que no aportan un

significado importante al contexto científico. Por ejemplo: “es”, “yo”, “tu”, “de”, “en”, “para”, “por”, “y”, “o”, “esto”, “esos”, “todo”. La lista completa de las palabras vacías en español utilizadas en este trabajo se muestran en el Anexo 2.

- **Normalización.** Para la normalización de los artículos científicos se transformaron todas las palabras al mismo nivel de carácter, es decir, para este trabajo todas las palabras se convirtieron a minúsculas, además se eliminaron las palabras con longitud menor o igual a 2.
- **Eliminación de caracteres especiales y signos de puntuación.** Para preparar los textos para el proceso de la Recuperación de Documentos, se eliminaron los siguientes caracteres y signos de puntuación: “?” “.” “,” “!” “;” “/” “,” “:” “(” “)” “%” “<” “>” “=” “ [” “] ” “+” “#” “&” “|” “\$” “{” “}” “@”
- **Lematización.** Se obtuvieron los lemas de todas las palabras contenidas en los artículos científicos, unificando las diferentes flexiones o conjugaciones. Por ejemplo, para las palabras: “vacunó”, “vacunará”, “vacunaste”, “vacunado”, el lema obtenido es “vacunar”. Para esta tarea se realizó un análisis morfológico para identificar y eliminar de manera correcta los sufijos y prefijos de las palabras. Para ello se empleó el diccionario en español del modelo “es_core_news_sm” incluido en la biblioteca **spaCy**.
- **Derivación.** Para aplicar la técnica de derivación a los artículos científicos se obtuvieron las raíces de las palabras contenidas en los textos. Por ejemplo, la raíz de “vacunó”, “vacunará”, “vacunaste”, “vacunado” es “vacun”. Para esta tarea se utilizó el algoritmo **Porter2**.

En la Figura 4.5 se muestra la salida de un artículo científico médico en español después de aplicarle el procesado descrito.

introducción enfermedad coronavirus covid enfermedad viral afectar varios órganos sistema
tratamiento preventivo profiláctico especialmente útil enfermedad infeccioso emergente covid-19
reducir necesidad hospitalización gasto salud público aunque efecto preventivo sars-cov-2 varios
agente terapéutico eg hidroxiclороquina cloroquin remdesivir lopinavir ritonavir evaluar ampliamente
ninguno demostrar gran eficacia clínico aquí objetivo abordar discutir estudio publicado recientemente
potencial quimioprofiláctico quercetina metodología realizar búsqueda literaturair base pubmed medline
scielo scopus of science incluir evaluar críticamente estudio abordar quercetina sars-cov-2 tipo coronavirus
resultado alguno estudio demostrar quercetina flavonoide aprobado fda utilizar agente antioxidante
antiinflamatorio inhibir entrada coronavirus sars-cov-2 célula huésped además estudio silico mostrar quercetina
potente inhibidor proteasa principal sarscov mpro sugerir flavonoide activo covid conclusión debido quercetina
poder prevenir disminuir duración infección sarscov plausible suponer uso profiláctico flavonoide producir
varios beneficio clínico prueba preliminar deber confirmar mediante ensayo vitro posteriormente ensayo
clínico aleatorizado

Figura 4.5: Artículo procesado del conjunto de datos **BioASQ MESINESP**.

4.4.3. Recuperador de Documentos

En la Recuperación de documentos se realiza una similitud textual entre la consulta y los artículos científicos para recuperar documentos relevantes. Para ello se utiliza el núcleo de la consulta y algoritmos o modelos computacionales. En este trabajo de investigación se realiza una combinación de un algoritmo de descubrimiento de tópicos, modelos probabilísticos, estadísticos y vectoriales para mejorar la recuperación de documentos científicos en español. A continuación, se presenta la descripción y la implementación los enfoques utilizados.

Algoritmo de descubrimiento de tópicos

En este trabajo se adaptó un algoritmo de descubrimiento de tópicos para agrupar a los documentos relacionados con un tema médico en específico, por ejemplo, la enfermedad de COVID-19. El objetivo es mejorar la recuperación de documentos relevantes con la reducción del espacio de búsqueda, por medio de un filtrado de información. Se utilizó el algoritmo de descubrimiento de tópicos LDA, ya que es uno de los algoritmos más populares para esta tarea en la literatura y se basa en una simple suposición de intercambiabilidad de palabras y temas en un documento. Es una técnica de reducción de la dimensionalidad con semántica probabilística [55]. La adaptación de este algoritmo se realizó utilizando los enfoques propuestos en el trabajo STTM [56].

En la Figura 4.6 se muestra la nube de palabras que contiene a las top-words después del filtrado de documentos relacionados con la enfermedad del coronavirus, En este ejemplo, el número de documentos se redujo en un 60 % y las 10 primeras palabras del tópico COVID-19 son: “paciente”, “SARS-CoV2”,

“pandemia”, “vacuna”, “covid”, “infección”, “tratamiento”, “virus” y “salud”.



Figura 4.6: Nube de palabras de las top-words del t3pico de COVID-19.

Modelo vectorial

El modelo vectorial, presentado por Gerard Salton [8], es un método fundamental en el ámbito de la RI. Este modelo se basa en la representación de documentos y consultas como vectores en un espacio multidimensional, donde cada dimensión representa un término del vocabulario. La similitud entre la pregunta y los documentos se calcula utilizando medidas de distancia o similitud entre sus vectores respectivos. En este proyecto de investigación, se implementó la similitud del coseno debido a su eficiencia computacional, además permite identificar documentos que comparten temas o conceptos similares.

En este modelo de RI se considera un espacio vectorial que consta de un número N de documentos D_i cada uno identificado por uno o más términos o índices T_j como se muestra en la Ecuación 4.1, los términos pueden ponderarse según su importancia entre 0 y 1 [8]. En consecuencia, cada artículo científico se transforma como un vector numérico que expresa una relación con cada ocurrencia de determinadas palabras o términos dentro del texto.

$$D_i = (d_{i1}, d_{i2}, \dots, d_{il}) \quad (4.1)$$

Para encontrar el valor de la ponderación de cada palabra dentro del vector numérico del documento se utiliza la frecuencia de las palabras de cada documento por la frecuencia inversa del término en la colección (TF-IDF).

Esta ponderación permite diferenciar entre términos comunes y especializados, asignando mayor peso a aquellos que son más relevantes para el tema del documento médico. Para realizar la ponderación mediante la técnica TF-IDF se utiliza la Ecuación 4.2.

$$tfidf_{i,j} = tf_{i,j} \times \log(Ndf_i) \quad (4.2)$$

TF-IDF combina dos medidas:

1. Frecuencia de Término (TF): Representa la frecuencia con la que aparece un término específico dentro de un documento. Un término con mayor frecuencia en un documento indica que es más relevante para el tema del mismo.
2. Frecuencia Inversa del Término (IDF): Representa la frecuencia con la que un término aparece en el conjunto de documentos. Un término con menor IDF indica que es más especializado y distintivo, lo que le otorga mayor peso en la ponderación.

En la Tabla 4.2 se muestra un fragmento de los valores obtenidos al aplicar la medida TF-IDF a los artículos científicos preprocesados y lematizados del conjunto de datos de BioAsq y a la consulta "¿Qué quimioprofiláctico se puede utilizar contra el SARS-CoV-2?". La primera entrada de la matriz de pesos le corresponde al vector la consulta y las demás filas corresponden a los vectores de los artículos científicos.

Tabla 4.2: Valores TF-IDF obtenidos a partir de la consulta y de los textos preprocesados y lematizados.

	clínico	cloroquina	cobertura	coronavirus	covid	cuidado	diagnóstico
D1	0	0	0.65232	0	0.15721	0	0
D2	0.29677	0.06148	0	0.20275	0.16454	0	0
D3	0.51139	0.05398	0	0	0.14176	0	0
D4	0	0	0.45987	0	0.05590	0	0
D5	0	0	0	0	0.04412	0	0
D6	0	0.27772	0.37018	0	0.06170	0.15491	0.14016
D7	0	0	0.04762	0	0.09355	0.23490	0
D8	0.16547	0	0	0	0.27523	0	0
D9	0.15679	0	0	0.16068	0.17387	0	0
D10	0.07683	0.26657	0.23652	0	0.08520	0	0
D11	0	0	0	0	0.08109	0	0
D12	0	0	0	0	0	0	0
D13	0	0	0	0	0.05651	0.14188	0
D14	0	0.12634	0.36589	0	0	0	0
D15	0	0.36454	0	0	0.09953	0	0

Después de obtener los vectores numéricos que representan tanto el conjunto de documentos como la consulta del usuario, se procede a calcular la similitud de la correlación entre los vectores, y puede ser cuantificada por el coseno del ángulo entre ellos. La finalidad es identificar aquellos documentos relevantes para la consulta. La similitud del coseno es una métrica intuitiva de semejanza entre dos vectores que se basa en el cálculo del coseno del ángulo formado entre ellos. Esta medida está normalizada por la longitud de los vectores, lo que la hace independiente de su magnitud y permite comparar vectores de diferentes tamaños. Para este proyecto se utilizó la Ecuación 4.3 para implementar la similitud del coseno.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (4.3)$$

En la Figura 4.7 se pueden observar los coeficientes de similitud obtenidos después de aplicar la similitud del coseno del vector de la consulta contra los vectores de los artículos científicos.

[0.51798, 0.37671, 0.34602, 0.34303, 0.27589,
0.26553, 0.24317, 0.24315, 0.23133, 0.22982,
0.22890, 0.21926, 0.17992, 0.17507, 0.16103,
0.15403, 0.15041, 0.13591, 0.13525, 0.13329]

Figura 4.7: Coeficientes de similitud del modelo vectorial

Los coeficientes obtenidos de la similitud del coseno del modelo vectorial se pueden interpretar de la siguiente manera:

- Valor 1: Indica que los vectores son idénticos, es decir, la consulta y el artículo científico comparten exactamente los mismos términos y en las mismas proporciones.
- Valor 0: Indica que los vectores son ortogonales, es decir, no comparten ningún término en común. En este caso, no existe relación alguna entre la consulta y el artículo científico.
- Valores entre 0 y 1: Indican grados variables de similitud entre la consulta y el artículo científico. Cuanto mayor sea el valor, mayor será la semejanza entre ellos.

Modelo probabilístico

Los modelos probabilísticos en el ámbito de la Recuperación de Documentos se basan en el principio de ordenación por probabilidad [57] donde los documentos se clasifican en un conjunto de relevantes y un conjunto de documentos no relevantes con base en una consulta inicial. Además, se considera lo siguiente:

- Existe una respuesta ideal a la consulta planteada, esta respuesta se encuentra dentro del conjunto de documentos relevantes.
- Existe una consulta ideal, la cual proporciona un conjunto de documentos relevantes.
- En principio se desconoce cuál es la consulta ideal, pero se sabe que es una combinación de 0 y 1. Se asignan valores iniciales de 0.5 denominados de "Máxima incertidumbre" que el documento recuperado puede estar en el conjunto de relevantes o en el de no relevantes.

En consecuencia, los modelos probabilísticos se fundamentan en el teorema de Bayes, donde se tiene la probabilidad de que un término o palabra se encuentre en el conjunto de documentos relevantes entre la probabilidad que se encuentre en el conjunto de documentos que no sean relevantes (Ecuación 4.4).

$$sim(d_j, q) = \frac{P(d_j \text{ sea relevante})}{P(d_j \text{ no sea relevante})} \quad (4.4)$$

En este proyecto, tras una exhaustiva revisión de la literatura, se optó por implementar el algoritmo probabilístico BM25 (Best Match 25). Este algoritmo se fundamenta en los principios esenciales de los modelos probabilísticos clásicos, pero se diferencia en que clasifica el conjunto de documentos en función de la presencia de términos de consulta en cada uno de ellos, independientemente de su interrelación o proximidad relativa.

El algoritmo Okapi BM25 [58] incorpora información relativa a la frecuencia de términos tanto en los documentos como en la consulta desde una perspectiva probabilística, utilizando la distribución 2-Poisson. Esta distribución permite modelar la probabilidad de que un término aparezca en un documento de manera más precisa que otros modelos, tomando en cuenta la longitud del documento y la frecuencia general del término en la colección. Para implementar el cálculo de la puntuación BM25 se utilizó la Ecuación 4.5. Donde dada una pregunta Q que contiene las palabras clave $q_1, \dots, q_n$, la puntuación BM25 de un documento D :

$$score(D, Q) = \sum_{i=1}^n IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} \quad (4.5)$$

Donde $f(q_i, D)$ es la frecuencia del término q_i en el documento D , $|D|$ es el número de palabras que contiene el documento D , $avgdl$ es el promedio de palabras que existen en la colección de documentos, k_1 y b son parámetros que por defecto tienen los valores de 1.2 y 0.75 respectivamente. Por último, $IDF(q_i)$ es la frecuencia inversa del término de la consulta q_i y se calcula con la Ecuación 4.6.

$$IDF(q_i) = \ln \left(\frac{N - n(q_i) + 0,5}{n(q_i) + 0,5} + 1 \right) \quad (4.6)$$

Donde N es el número de documentos del corpus y $n(q_i)$ es el número de artículos médicos que contienen el término de pregunta q_i

En este trabajo de investigación se ajustaron los parámetros k_1 y b del modelo probabilístico. Los parámetros k_1 y b desempeñan un papel crucial en el cálculo de la similitud entre un documento y una consulta. Para mejorar la efectividad de la recuperación de documentos médicos, se optó por calibrar estos parámetros con valores reducidos de 1.0 y 0.5, respectivamente. Esta estrategia tiene como objetivo principal reducir el impacto excesivo de la frecuencia de términos en la puntuación de similitud, tanto en el documento como en la consulta. El impacto de la reducción de los parámetros k_1 y b es el siguiente:

Reducción del Sesgo hacia Documentos con Alta Frecuencia de Términos: Al disminuir el valor de k_1 , se mitiga el impacto desmedido de la frecuencia de términos en la puntuación de similitud. Esto permite que los documentos con menor cantidad de términos relevantes, pero que aún son pertinentes para la consulta, logren tener una mejor ranking o posición en los resultados de la búsqueda.

Comparación Justa entre Documentos de Diferentes Longitudes: El ajuste del parámetro b a un valor más bajo reduce la influencia de la longitud del documento en la puntuación de similitud. De esta manera, se logra una comparación más justa entre documentos de diferentes tamaños, evitando que los textos extensos dominen los resultados de búsqueda en detrimento de documentos más concisos que puedan ser más relevantes para la consulta.

Por último, se aplica el modelo probabilístico a los documentos científicos procesados para obtener los coeficientes de similitud entre los artículos y la consulta de entrada. Además, un ranking es construido de una manera dinámica con el objetivo de obtener la relevancia de cada artículo respecto a la consulta. En la Figura 4.8 se aprecia un fragmento la puntuación de similitud obtenidos con el modelo BM25.

```
[0.84636, 0.82860, 0.80777, 0.80308, 0.77788,
0.77645, 0.77427, 0.75500, 0.75495, 0.75148,
0.73185, 0.72687, 0.72449, 0.71406, 0.70642,
0.69996, 0.69060, 0.68356, 0.67987, 0.67805]
```

Figura 4.8: Coeficientes de similitud del modelo probabilístico

La puntuación de similitud BM25 mostrada calcula la relevancia de los documentos con respecto a la pregunta en lenguaje natural y se encuentra

ordenada, de manera que los artículos con mayor puntuación o más cercanos a 1 se consideraron más relevantes para la consulta.

Modelo semántico-estadístico

Para la etapa de recuperación de información (RI) en este trabajo de investigación se adaptó un modelo fundamentado en la semántica y la estadística de ocurrencia de términos en artículos científicos. Con el objetivo de superar las limitaciones de los métodos tradicionales en la coincidencia exacta de términos, se implementó el modelo de Indexación Semántica Latente (LSI). LSI aprovecha la estructura latente subyacente en el uso de las palabras, revelando relaciones semánticas profundas [59]. Este modelo utiliza la SVD para descomponer una matriz de términos-documentos en tres matrices: U , V y D . La matriz U representa los documentos, contiene los valores singulares y V representa los términos [60]. Al proyectar los documentos y las consultas en el espacio vectorial definido por las primeras k columnas de V , se obtiene una representación semántica que permite identificar documentos relevantes de manera más precisa. En la Ecuación 4.7 se muestra la SVD, donde M es la matriz de frecuencia de palabras K es la matriz ortogonal, S la matriz diagonal y D la matriz transpuesta.

$$M = KSD^t \quad (4.7)$$

Al aplicar el modelo LSI que se basa en la indexación de términos, ubicándolos en un contexto semántico se evitan fenómenos como la sinonimia o polisemia. Además, permite la recuperación de información sobre el significado de un documento y así superar los problemas que conllevan realizar una búsqueda de información basada en la coincidencia lexicográfica. Una parte fundamental de la búsqueda estadística del modelo LSI es la creación de la Matriz de asociación de término-documento (véase la Tabla 4.3) de los artículos médicos y la consulta realizada en lenguaje natural. Además, se aplicó el método de reducción de dimensionalidad o descomposición de vectores singulares (SVD) para crear la matriz ortogonal que representa el vector de palabras, la matriz transpuesta que representa el vector de documentos y la diagonal que representa el valor singular de la coincidencia palabra-documento.

Tabla 4.3: Matriz de frecuencia de Términos .

Raíces	D1	D2	D3	D4	D5	D6	D7	D8	D9
cloroquin	0	0	1	0	0	1	0	0	0
virus	1	0	1	1	1	0	1	1	0
vacun	0	1	0	0	1	1	1	1	1
pacient	0	1	0	0	0	1	0	0	0
enfermed	1	1	0	1	1	1	1	1	1
chin	0	0	1	0	0	0	0	0	1
cov	1	0	1	1	0	1	1	1	0
contagi	0	1	0	0	0	0	0	1	1
sars-cov	1	0	1	0	1	1	1	0	0
wuh	0	0	1	0	0	0	0	0	1
modern	0	1	0	0	1	1	1	0	0
muert	1	1	1	0	1	0	0	0	1
tratamient	1	1	1	1	1	1	1	0	0
fiebr	1	0	1	0	1	0	1	0	0
cas	0	0	1	0	0	0	0	0	0

Finalmente se obtuvieron los coeficientes de similitud del modelo LSI aplicados a la matriz de dimensión reducida que se muestran en la Figura 4.9 para recuperar a los artículos médicos relevantes.

```
[0.89885, 0.85362, 0.81604, 0.71477, 0.53148,
 0.51873, 0.49105, 0.42096, 0.29095, 0.35269,
 0.31985, 0.26645, 0.20226, 0.18050, 0.16392,
 0.11632, 0.11552, 0.10947, 0.01560, 0.00236]
```

Figura 4.9: Coeficientes de similitud obtenidos con el modelo LSI

Cada coeficiente representa la similitud de cada documento médico del corpus con respecto a la pregunta realizada en lenguaje natural y se interpreta que los documentos con los valores más altos satisfacen de mejor forma la necesidad de información del investigador en el dominio médico.

4.4.4. Evaluación del Sistema de Recuperación de artículos científicos

Para determinar la eficacia de las diferentes técnicas de PLN en la recuperación de información médica, se diseñó un experimento utilizando un corpus de 249,00 artículos científicos en español y un conjunto de 100 preguntas sobre COVID-19. Se evaluaron los diversos modelos de RI presentados que combinan enfoques probabilísticos, estadísticos y semánticos. Las métricas de precisión, exhaustividad y medida F se emplearon para medir el desempeño de cada modelo combinado con las técnicas de PLN y seleccionar el más adecuado para la tarea de Recuperación de Documentos médicos en español.

Conjuntos de datos para la evaluación del SRI

En la evaluación del Recuperador de Documentos se utilizó un conjunto de datos en el que se realizó la búsqueda de los artículos relevantes. Este conjunto de datos es BioAsq, que consta de 249,474 textos científicos, con un número de palabras de 45,322,119 y un vocabulario de 373,851 términos.

Se diseñó un conjunto de 100 preguntas en español sobre COVID-19, basado en el corpus BioAsq, para evaluar el rendimiento del recuperador de documentos. Este conjunto, en formato JSON y con las etiquetas “id_Pregunta”, “id_Documento” y “texto de Pregunta”, garantiza que las respuestas se encuentren dentro del corpus. La creación manual de las preguntas asegura la relevancia y calidad de la evaluación. Un extracto de este conjunto de datos puede verse en la Figura [4.10](#).

```

{"PreguntasCovid": [
  {"id_Pregunta": 1,
   "TextoPregunta": "¿Qué quimioprofiláctico se puede utilizar contra el SARS-CoV-2? ",
   "id_Documento": "ibc-ET6-1764"},

  {"id_Pregunta": 2,
   "TextoPregunta": "¿Qué atención requieren los pacientes con FRA?",
   "id_Documento": "ibc-194912"},

  {"id_Pregunta": 3,
   "TextoPregunta": "¿Qué manifestaciones oculares se pueden presentar con el COVID-19?",
   "id_Documento": "ibc-192560"},

  {"id_Pregunta": 4,
   "TextoPregunta": "¿Existen estudios de glucocorticoides y COVID-19?",
   "id_Documento": "ibc-196134"}
]}

```

Figura 4.10: Preguntas de COVID-19 en **español**

Como se puede observar en la etiqueta `id_Documento` se encuentra el identificador del artículo médico donde se encuentra la respuesta a la pregunta realizada en lenguaje natural. Dicha información sirve para realizar una evaluación del recuperador de información de manera automática.

Métricas de evaluación para el SRI

Para evaluar si los documentos recuperados por cada enfoque (modelos de RI combinado con técnicas de PLN y el algoritmo de descubrimiento de tópicos) son respuestas exactas a las consultas, se realizó una clasificación según su relevancia [61]. Esta clasificación o ranking se basa en las métricas de precisión, exhaustividad y medida-F, que permiten determinar si los documentos recuperados son relevantes a la consulta realizada.

- La precisión se define como la fracción de documentos recuperados que es relevante y proporciona una medida para evitar los documentos irrelevantes o comúnmente llamados ruido, para lo cual se utiliza la Ecuación 4.8.

$$Precisión = \frac{\#Documentos\ relevantes\ recuperados}{\#Documentos\ recuperados} \quad (4.8)$$

- La exhaustividad se considera como la fracción de documentos relevantes que se recuperaron, y es una medida para la recuperación de documentos, o comúnmente denominada silencio (véase la Ecuación

4.9). En esta métrica, se ha añadido un término para determinar si el documento relevante está en la lista de los documentos mejor clasificados.

$$Exhaustividad = \frac{\#Documentos\ relevantes\ recuperados}{\#Documentos\ relevantes} \quad (4.9)$$

- F1 es una medida armónica donde combina los resultados de la precisión y la exhaustividad. Se puede observar en la Ecuación 4.10

$$F1 = \frac{2 \times Precisión \times Exhaustividad}{Precisión \times Exhaustividad} \quad (4.10)$$

Después de una revisión detallada y extensa de la literatura, se eligieron las métricas descritas para la evaluación porque son las más utilizadas en el área del PLN.

Configuración de experimentos para la evaluación del SRI

Con el objetivo de evaluar el desempeño del recuperador de documentos médicos relevantes, se diseñaron múltiples configuraciones experimentales. Estas configuraciones exploraron la combinación de modelos vectoriales, probabilísticos y semántico-estadísticos, junto con técnicas de PLN y un algoritmo de descubrimiento de tópicos.

1. Se combinaron los modelos probabilístico, semántico y vectorial con técnicas de preprocesado-derivación y preprocesado-lematización. Se devuelve el top 3 y 5 de los documentos mejor ranqueados. En esta configuración, no se empleó el algoritmo LDA.
2. El modelo probabilístico se combinó con preprocesado simple, preprocesado-lematización y preprocesado-derivación. Además, se devuelve el top 3, 5, 10, 15 y 20 de los documentos candidatos a relevantes, en esta configuración se utiliza el algoritmo de descubrimiento de tópicos LDA.
3. El modelo semántico-estadístico se combinó con preprocesado de textos simple, preprocesado-lematización, preprocesado-derivación. Se recupera el top 3, 5, 10, 15 y 20 de los documentos candidatos, en esta configuración se utiliza el algoritmo LDA.

4. El modelo de vectorial se combinó con preprocesado simple, preprocesado-lematización, preprocesado-derivación. Se recupera el top 3, 5, 10, 15 y 20, en esta configuración se utiliza el algoritmo LDA.

Empleando las configuraciones experimentales especificadas, se procedió a evaluar la efectividad de la fase de recuperación de documentos médicos relevantes. Los resultados obtenidos de esta evaluación se detallan en el capítulo 5.

4.5. Extracción de las respuestas candidatas

Una vez concluida la fase de recuperación de documentos médicos relevantes, se procede a la etapa de extracción de respuestas precisas mediante un modelo de lenguaje de gran escala basado en la arquitectura Transformer. Esta elección metodológica, fundamentada en una exhaustiva revisión bibliográfica, se justifica por la capacidad de los modelos Transformer de capturar dependencias de largo alcance y generar representaciones semánticas enriquecidas, características esenciales para la tarea de respuesta a preguntas en el dominio médico.

Con el objetivo de abordar el desafío de la respuesta a preguntas en el ámbito médico, se ha diseñado una metodología para la extracción de respuestas basado en grandes modelos de lenguaje. La metodología propuesta, detallada en la Figura 4.11, aprovecha las capacidades de estos modelos para generar representaciones vectoriales de alta calidad de los documentos médicos y aplicar técnicas de atención para identificar los segmentos de texto más relevantes en relación con la consulta realizada en español.

Las fases más importantes de la metodología que se ha implementado combina técnicas de transferencia de aprendizaje y ajuste fino un gran modelo de lenguaje (LLM). El modelo base se entrenó inicialmente en un corpus de texto general y posteriormente se ajustó a un corpus especializado en el dominio médico, compuesto por más de 3 millones de documentos. El objetivo de este proceso fue enriquecer el vocabulario del modelo y mejorar su capacidad para comprender el lenguaje médico. Finalmente, se llevó a cabo un ajuste fino específico para la tarea de extracción de respuestas, optimizando los hiperparámetros del modelo mediante el uso de técnicas de calibración.

En este trabajo de investigación se seleccionó el modelo BERT-Base-Cased preentrenado [62] para la tarea de extracción de respuestas. Esta ar-

quitectura, preentrenada en un corpus masivo de texto, proporciona una representación contextual rica de las palabras, lo que facilita la identificación de las respuestas relevantes. Además, siguiendo las recomendaciones de [51], se optó por un modelo preentrenado para minimizar los costos computacionales y ambientales asociados con el entrenamiento desde cero.

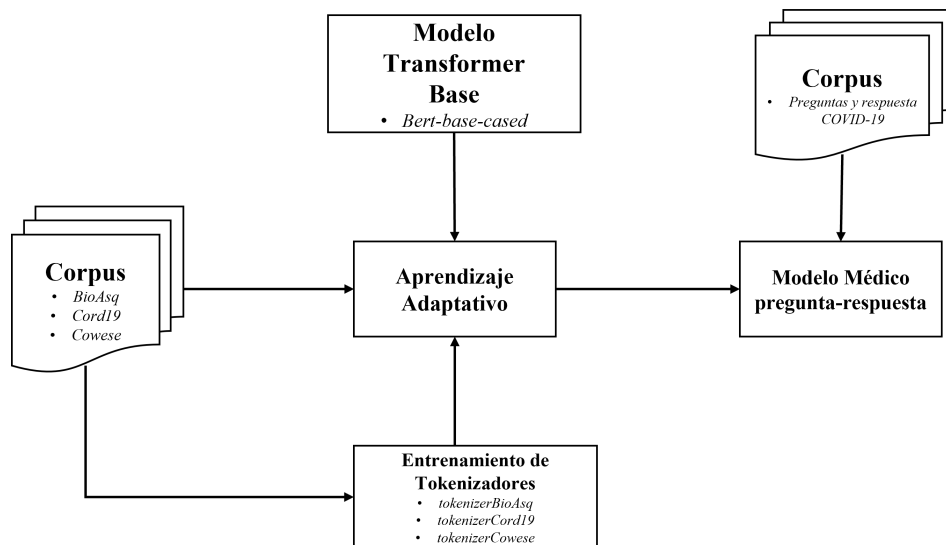


Figura 4.11: Metodología para la extracción de la **respuesta**.

4.5.1. Corpus para el entrenamiento del LLM médico

Aún cuando el español es uno de los 5 idiomas más hablados en el mundo, entrenar grandes modelos de lenguaje en este idioma no es una tarea sencilla, puesto que encontrar grandes conjuntos de datos o recursos para entrenar o evaluar modelos en español se vuelve una tarea complicada. En esta sección se presentan los conjuntos de datos utilizados para entrenar y evaluar el modelo de lenguaje médico, los cuales constan de tres conjuntos de datos para realizar el aprendizaje adaptativo y uno más para llevar a cabo el ajuste fino del modelo en la tarea de pregunta-respuesta (QA).

Los conjuntos utilizados para adaptar el modelo al vocabulario médico en español son: BioAsq, Cord-19 y Cowese. El primer conjunto de datos utilizado es de la competencia internacional centrada en la biomedicina, extracción y la recuperación de información biomédica del año 2021. El segundo conjunto de datos utilizado es CORD-19 [21] el cual es un repositorio con más de

un millón de artículos científicos de texto completo que incluye contenido sobre COVID-19 y la investigación histórica relacionada con los coronavirus, los documentos están estructurados en formato CSV y están divididos en: *cord_uid*, *source_x*, *title*, *doi*, *abstract*, *authors*, *journal*, *url*, entre otros.

Para adaptar al modelo médico se utilizaron 814,402 resúmenes del conjunto de datos CORD-19 traducidos al español. El último conjunto de datos para el preentrenamiento del modelo médico es Cowese [63] y según sus autores es el mayor corpus biomédico en español hasta la fecha, tiene casi dos millones de textos preprocesados y 750 millones de tokens o palabras, además fue recolectado en 3000 sitios en español. En la Tabla 2 4.4 se muestra la lista de los corpus y el vocabulario de cada conjunto utilizado para realizar el preentrenamiento del modelo médico.

Tabla 4.4: Lista de corpus de texto utilizados para el preentrenamiento del modelo médico.

Corpus	Número de Textos	Vocabulario
BioAsq	249,473	48,217
Cord-19	814,402	104,186
Cowese	1,973,048	227,667

Para el ajuste fino y entrenamiento del modelo médico en la tarea de responder preguntas médicas, se utilizó el conjunto de datos BioAsq10 del año 2022, específicamente de la tarea 7b. Este conjunto de datos, compuesto por 39,680 preguntas biomédicas de tipo factoides, se divide en las siguientes secciones: context, question, id, title y answers. La Figura 4.12 muestra un fragmento de ejemplo de las preguntas contenidas en este conjunto de datos.

```
{ "id": 35,532 }
{ "question":
¿Se ha evaluado tocilizumab frente a Covid-19? }
{ "answer_start": [ 478 ], "text": [ "Los resultados clínicos preliminares han
indicado que el antagonismo del receptor IL-6 (IL-6R), incluyendo el anticuerpo
monoclonal humanizado tocilizumab aprobado por la FDA, puede mejorar los resultados
de los pacientes con COVID-19 grave o crítico manteniendo un buen perfil de seguridad." ] }
```

Figura 4.12: Ejemplos de una pregunta del conjunto de entrenamiento.

Estos conjuntos de datos constituyeron el corpus de entrenamiento para crear el modelo de lenguaje médico propuesto para la extracción de las respuestas MédicoBERT.

4.5.2. Aprendizaje Adaptativo del gran modelo de lenguaje médico

El aprendizaje adaptativo de un LLM se refiere a la capacidad de estos modelos para mejorar su rendimiento y precisión en función de los datos a los que se expone o que se entrena. Esto implica que los LLM poseen la habilidad de aprender y mejorar de manera continua, adaptándose a las necesidades, características y al dominio específico de las tareas que se les encomienda resolver. Los LLM pueden adaptarse a cambios en su entrenamiento y a nuevos datos, esta característica los convierte en una herramienta robusta y confiable para resolver diversas tareas de PLN.

En este trabajo de investigación se adapta un modelo de lenguaje de gran tamaño (LLM) al dominio médico. Para esta tarea, se emplea el modelo LLM de Representación de Codificador Bidireccional de Transformadores, también conocido como BERT, desarrollado por Google. BERT, basado en redes neuronales, está diseñado específicamente para resolver tareas de Procesamiento del Lenguaje Natural.

BERT aprovecha el componente codificador de la arquitectura Transformer [1] y se preentrena utilizando Wikipedia y un corpus de libros digitales. Este modelo es un codificador Transformer bidireccional multicapa, y su arquitectura base posee 12 capas de codificación, una longitud de secuencia de 512, 768 estados dimensionales y 12 cabezas de atención, sumando un total de 110 millones de parámetros [62]. Para garantizar la máxima confiabilidad en las respuestas, se seleccionó BERT, un modelo de comprensión del lenguaje. De esta manera, las respuestas se obtienen directamente de los documentos relevantes, evitando posibles sesgos o imprecisiones propias de la generación de texto.

Con el objetivo de adaptar el LLM a la terminología médica, se llevó a cabo un proceso de entrenamiento tanto de los tokenizadores como del modelo base, el cual se describe en detalle en las siguientes secciones.

Entrenamiento de los tokenizadores

Un tokenizador es un componente fundamental del LLM que permite transformar el texto en una representación numérica para poder procesarlos. El modelo base BERT que se adaptó, utiliza un tokenizador basado en subpalabras (*wordpiece*) [64], el objetivo principal de este enfoque es evitar que las palabras de uso frecuente se fragmenten, preservando su significado completo. Por otro lado, las palabras poco comunes o inusuales se descomponen en subpalabras más pequeñas, permitiendo al modelo reconocerlas y evitar clasificarlas como desconocidas como se muestra en la Figura 4.13.

```
texto = "Qué quimioprofiláctico se puede utilizar contra el coronavirus"

Tokens = ['Qué', 'quimio', '##prof', '##il', '##áctico', 'se', 'puede', 'utilizar', 'contra',
          'el', 'corona', '##virus']

Ids= [8177, 18754, 22084, 5347, 17448, 5304, 5554, 7703, 5941, 5292, 22621, 14612]
```

Figura 4.13: Ejemplo de la salida del Tokenizador base.

Adaptar BERT al dominio médico requiere reemplazar el tokenizador pre-determinado. Esto se debe a que un tokenizador entrenado con un conjunto de datos de un dominio general y en otro idioma no es adecuado para realizar tareas de extracción de información en la literatura médica. El vocabulario específico del ámbito médico en español difiere significativamente del lenguaje general, lo que limita el rendimiento de un tokenizador genérico. La finalidad de entrenar nuevos tokenizadores es mejorar la precisión para comprender la estructura y composición de las abreviaturas, así como los términos médicos en español, por ejemplo, nombres de medicamentos o enfermedades.

En este trabajo de investigación se entrenaron tres tokenizadores, uno para cada conjunto de datos utilizado (BioAsq, Cord-19 y CoWeSe), utilizando una GPU NVIDIA 3090 y el marco de aprendizaje profundo Pytorch. El objetivo de este entrenamiento fue dotar al modelo de la capacidad de comprender el lenguaje médico de manera precisa y efectiva al contar con el vocabulario adecuado para el dominio médico. Cada uno de los tokenizadores entrenados posee un tamaño de vocabulario de 50,000 tokens.

La Figura 4.14 ilustra la salida de uno de los tokenizadores entrenados, mostrando cómo divide una oración en subpalabras. El propósito de esta segmentación es obtener valores que sean concisos, pero a la vez altamente

significativos para el modelo en el contexto médico. Además, la Figura también presenta los valores numéricos de la conversión de los textos científicos, esta conversión sirve para que el modelo puede procesar y analizar de manera eficiente la información.

```
Text: "una respuesta inmunitaria en pacientes con enfermedad
      del coronavirus causa una tormenta de citocinas inflamatorias"

Tokens: ['[CLS]',
         'una', 'respuesta', 'inmunitaria', 'en', 'pacientes', 'con', 'enfermedad',
         'del', 'coronavirus', 'causa', 'una', 'tormenta', 'de', 'citocinas', 'inflamatorias',
         '[SEP]']

ids: [2,
      1882, 931, 12120, 319, 432, 330, 539, 349, 11127, 1711, 1882, 18659, 305, 4296, 4868,
      3]
```

Figura 4.14: Conversión de texto a valores numéricos realizada por el Tokenizador entrenado.

El entrenamiento del tokenizador con un corpus médico especializado permitió al modelo construido médicoBERT generar subpalabras y tokens más representativos para el dominio, lo que se tradujo en una mejora significativa en la representación vectorial de términos médicos y, por ende, en su desempeño en tareas específicas de PLN de este dominio, por ejemplo, las respuestas a preguntas.

Entrenamiento del LLM médico

En este trabajo de investigación se realizó el entrenamiento del modelo médico donde se utilizaron los tokenizadores descritos anteriormente y los conjuntos de datos de literatura médica. Además, a los conjuntos se les hizo un procesamiento moderado que consiste en la eliminación de los signos de puntuación como: *¿*, *¡*, *í*, *ı*, */*, *;*, *:*, ***, *Ç*, *\$*, y *&*. El motivo de eliminar estos signos se debe a que no aportan información relevante para el entrenamiento del modelo y podrían generar ruido durante el proceso. Por último, para preparar el texto para el entrenamiento, se realizó una normalización convirtiéndolo a minúsculas. Esta etapa es crucial, porque garantiza que todas las palabras se representen de la misma manera, evitando que variaciones en el caso (mayúsculas/minúsculas) afecten el proceso de aprendizaje.

Para entrenar el modelo médico encargado de extraer las respuestas a las preguntas, se utilizó la técnica de Modelado de Lenguaje Enmascarado

(MLM), que consiste en entrenar el LLM para predecir las palabras faltantes en una secuencia de texto [65], para la generación del modelo se reemplazaron de manera aleatoria el 15 % de las palabras del corpus de literatura médica, utilizando el token especial <mask>. La finalidad es aprender las relaciones entre las palabras y su contexto. Es importante mencionar que esta técnica no requiere que el conjunto de datos esté etiquetado. A continuación, se muestra un fragmento del enmascaramiento de un texto proveniente del dataset CORD-19, el objetivo es predecir que palabra reemplaza al token especial y le da un mejor sentido a la oración con base en el contexto médico.

La actividad in vitro frente al SARS-CoV-2 de la <mask> es algo mayor que la que posee la cloroquina y también se ha descrito su uso para el tratamiento de la infección por COVID-19 como mono-droga y asociado al uso del antibiótico azitromicina

$$P(m) = \{\text{hidroxicloroquina}|\text{urticaria}|\text{enfermedad}\}$$

Con el objetivo de enriquecer el entrenamiento del modelo médico, también se utilizó la técnica de Predicción de la Siguiente Oración (NSP). Esta técnica consiste en dividir cada conjunto de datos en oraciones y crear dos conjuntos distintos los cuales se describen a continuación:

- Conjunto de pares relacionados: Este conjunto contiene el 50 % de las oraciones del conjunto de datos original, donde la siguiente oración de cada par es la continuación lógica de la primera.
- Conjunto de pares aleatorios: Este conjunto contiene el 50 % restante de las oraciones del conjunto de datos original, donde las oraciones se combinan aleatoriamente sin ninguna relación lógica entre ellas.

Asimismo, se muestra un ejemplo del entrenamiento del modelo médico con NSP, donde se presenta una “oración A” y una “oración B”. El objetivo es que el modelo médico clasifique o prediga (P) si “B” es la siguiente oración en la secuencia “A”, evaluando la coherencia y relación semántica entre ambas.

“Oración A”: *De los 88 pacientes incluidos en el estudio, 24 pacientes con COVID-19 fallecieron: 10/23 (43,4 %) eran pacientes con diagnóstico de demencia y 14/65 (21,5 %) pacientes del grupo control*

“Oración B”: *La letalidad entre los pacientes con demencia degenerativa primaria por COVID-19 es significativamente mayor en comparación con otros pacientes con edades medias y comorbilidades similares.*

(P): Es _siguiente

Esta tarea es fundamental porque obliga al modelo a realizar un análisis semántico profundo de los textos médicos, permitiéndole comprender las relaciones entre las oraciones y mejorar su capacidad para generar texto coherente y con sentido.

Hiperparámetros y métricas de evaluación del LLM médico

Para evaluar al modelo médico se utilizaron las métricas de perplejidad y la pérdida de validación, según [66] es una de las métricas más importantes de PLN para evaluar los modelos lingüísticos y es abreviada como (PP) y se muestra en la Ecuación 4.11, se define como la probabilidad inversa del conjunto de pruebas, normalizada por el número de palabras.

$$PP(W) = \frac{1}{P(w_1, w_2, \dots, w_N)^{\frac{1}{N}}} = \sqrt[N]{\frac{1}{P(w_1, w_2, \dots, w_N)}} \quad (4.11)$$

-Cuanto más grande es $PP(W)$ en el modelo, es menor el valor de la perplejidad.

-Los mejores modelos de lenguaje tienen valores más pequeños de perplejidad.

Por otro lado, la pérdida de validación, definida en la Ecuación 4.12, se calcula como la suma de los errores individuales de cada muestra perteneciente al conjunto de validación, y esta métrica se evalúa al finalizar cada época de entrenamiento.

$$Loss = \frac{-1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} * \log(p_{ij}) \quad (4.12)$$

Con la finalidad de determinar la mejor configuración de hiperparámetros para el entrenamiento del modelo médico, se llevó a cabo un proceso de calibración inicial o también llamada calibración gruesa. Esta estrategia consistió en dos etapas sucesivas, las cuales se describirán a continuación:

- Obtener la configuración inicial de Hiperparámetros: En esta etapa, se utilizaron valores de hiperparámetros reportados en los trabajos relacionados como punto de partida. Posteriormente, se realizó una búsqueda manual exhaustiva para ajustar cada hiperparámetro y mejorar el rendimiento del entrenamiento.
- Búsqueda Manual de Hiperparámetros: la búsqueda manual de hiperparámetros implicó probar diferentes valores para cada hiperparámetro y evaluar el rendimiento del modelo en cada caso. Esta técnica permitió identificar la configuración de hiperparámetros que mejoraba el rendimiento del modelo.

La configuración de hiperparámetros encontrada después de la calibración gruesa se presenta a continuación:

```
training_args = {
    épocas = 45,
    tasa_aprendizaje =  $2e - 5$ ,
    peso = 0,01,
    lote = 16,
    acumulación_gradiente = train_batch_size * 4,
    fp16 = True,
    optimizador = "adafactor",
}
```

Donde cada época corresponde a una iteración completa del modelo sobre todo el conjunto de datos de entrenamiento (más de 3 millones de textos médicos). El peso de una conexión neuronal indica la importancia que se asigna a una determinada característica del texto al realizar una predicción. La tasa de aprendizaje actúa como un factor de escala que determina la magnitud de los cambios en los pesos durante el entrenamiento del LLM. El optimizador, por su parte, es el algoritmo que selecciona la dirección y el tamaño de estos ajustes, con el objetivo de minimizar la función de pérdida de entrenamiento.

Los resultados de la evaluación del entrenamiento del modelo médico se muestran en la Sección 5.

4.5.3. Ajuste fino del LLM médico para la tarea de pregunta-respuesta

A continuación se presenta el proceso utilizado en esta investigación para ajustar el modelo de lenguaje médico a la tarea de respuesta a preguntas médicas en español (QA). Este proceso de adaptación o también conocido como ajuste fino tiene como objetivo resolver una tarea específica de PLN. Para ello, se proporciona al modelo pre-entrenado un conjunto de datos más pequeño y etiquetado específicamente para la tarea de QA. Al ajustar los pesos de las conexiones neuronales del modelo, se logra que este se adapte mejor a las características y patrones del conjunto de datos de preguntas-respuestas.

La tarea de QA es una tarea fundamental del procesamiento del lenguaje natural que busca obtener respuestas precisas a partir de grandes cantidades de texto. Para adaptar el gran modelo de lenguaje médico generado a esta tarea, se implementó un proceso de ajuste fino en varias etapas, con el objetivo de optimizar su capacidad para comprender consultas médicas en español y proporcionar respuestas extractivas relevantes y coherentes.

- **Arquitectura base del modelo de lenguaje médico** Para adaptar el LLM BERT a las particularidades del lenguaje médico, se realizaron modificaciones en la arquitectura original las cuales ya fueron descritas en las secciones anteriores, de las cuales se obtuvo el modelo Médico-BERT. Además, se utilizó una longitud de secuencia de 512 tokens, lo cual es suficiente para procesar la mayoría de las consultas médicas y textos médicos cortos. Para aquellos casos en los que se requiera procesar textos más largos, se implementó una técnica de ventana deslizante, siguiendo las recomendaciones presentadas en el trabajo propuesto en [67].
- **Adaptación a preguntas generales de tipo factoides en español.** Con el objetivo de adaptar el modelo a la tarea de responder preguntas factuales en español, se realizó un ajuste fino de la última capa utilizando un subconjunto de 64,000 preguntas del dataset SQuAD [68]. Este proceso permitió al modelo familiarizarse con la estructura y el vocabulario característicos de las preguntas en español, mejorando significativamente su capacidad para extraer respuestas precisas de los textos médicos.

- **Entrenamiento del modelo MédicoBERT con datos de preguntas médicas.** Con el objetivo de mejorar la precisión del modelo médico (MédicoBERT) en la tarea de preguntas-respuestas, se llevó a cabo un ajuste fino utilizando el conjunto de datos BioAsq10 de 2022². Este corpus, compuesto por 39,680 instancias con preguntas de tipo factoides o de hechos sobre temas médicos, esto permitió refinar la capacidad del modelo para identificar y extraer información relevante de textos especializados. Al seleccionar un conjunto con datos o textos independientes de la fase del preentrenamiento, se garantizó la diversidad y se minimizó el riesgo de introducir sesgos inherentes a las fases previas de entrenamiento.
- **Calibración fina de hiperparámetros.** Con el objetivo de maximizar el desempeño del modelo en la tarea de QA médica, se realizó una exhaustiva búsqueda de los hiperparámetros óptimos. Para evaluar la calidad de las respuestas generadas por el modelo, se utilizó la métrica F1, un promedio armónico entre la precisión y el recall, indicadores estándar en la evaluación de extracción de información. La importancia de esta etapa en el contexto de la investigación justifica una descripción más detallada del proceso, la cual se presenta en la siguiente sección.

4.5.4. Calibración por optimización de hiperparámetros para el Ajuste Fino del Modelo Médico

Para realizar el entrenamiento o ajuste fino del modelo médico se buscó la configuración adecuada de hiperparámetros. Para esta tarea se empleó una técnica de calibración por optimización. El objetivo de esta calibración es maximizar el valor de la métrica F1, la cual permite evaluar la precisión y exhaustividad del modelo en la tarea de respuesta a preguntas médicas. Para ello, se llevó un proceso el cual se describe en la Figura 4.15.

²<https://huggingface.co/datasets/avacaondata/bioasq22-es>

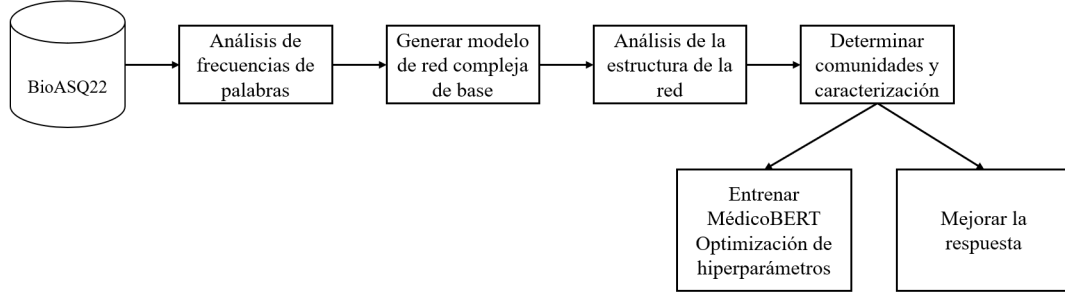


Figura 4.15: **Arquitectura para la calibración de hiperparámetros.**

A continuación, se presenta la configuración experimental empleada en el ajuste fino del modelo, donde se buscó optimizar el valor de la métrica F1. Se detallan los hiperparámetros ajustados y los rangos de valores explorados en la búsqueda.

maxF1

$S.A$

$\theta : \{L_r, w_0, b_z, e\}$

$L_r : \text{Tasa aprendizaje} \quad L_r = [0, 1], R^+$

$w_0 : \text{Peso} \quad w_0 = [0, 1], R^+$

$b_z : \text{Lotes} \quad b_z = \{2, 4, 8, 16, 32, 64\}$

$e : \text{Epocas} \quad e = [3, 7], R^+$

Mediante un análisis preliminar se identificaron los hiperparámetros que, de acuerdo a la literatura y a experimentos previos, tendrían un mayor impacto en la métrica F1 de la tarea de respuesta a preguntas médicas. Estos hiperparámetros fueron seleccionados para su optimización.

Para obtener la mejor configuración mediante la calibración por optimización de los hiperparámetros se realizaron las siguientes etapas.

1. **Hiperparámetros obtenidos del estado del arte.** En esta primera etapa, se realizó un entrenamiento inicial del modelo con una configuración de hiperparámetros estándar, ampliamente utilizada en la literatura para el ajuste fino de modelos en dominios similares. El objetivo de este entrenamiento fue obtener una solución inicial y establecer un punto de referencia para futuras comparaciones. La configuración encontra-

da es: $Epocas = 5$, $Lotes = 8$, $Peso = 0,001$, $Tasa_de_Aprendizaje = 1e - 5$ y los resultados de este entrenamiento se detallan en la sección 5.

2. **Calibración gruesa de Hiperparámetros.** Se empleó un método exploratorio para identificar la configuración que mejoraba la solución preliminar y reducía el espacio de búsqueda de la configuración óptima. Esta búsqueda se logró explorando aleatoriamente el espacio de hiperparámetros y evaluando el rendimiento del modelo para cada configuración. Se realizaron un total de 56 experimentos, variando los hiperparámetros en cada iteración de la siguiente manera: $pocas \pm 0,25$, $Lotes * /2$, $peso \pm 0,005$, $Tasa_de_Aprendizaje \pm 0,5e - 5$. Los resultados del entrenamiento para la tarea de respuesta a preguntas con esta configuración encontrada se presentan en la sección 5.
3. **Calibración Fina por optimización de Hiperparámetros.** Se empleó un enfoque metaheurístico obtenido de [69] para identificar la mejor configuración de hiperparámetros para el modelo médico entrenado (MédicoBERT) con el objetivo de optimizar su rendimiento en la tarea de QA en español en el dominio médico. Los enfoques utilizados se describen a continuación.

Búsqueda heurística (algoritmo de ascenso en la montaña) . Para optimizar la configuración de hiperparámetros del modelo MédicoBERT para la tarea de QA en español, se utilizó una técnica heurística inspirada en el ascenso en la montaña. La estrategia de búsqueda se basó en un algoritmo iterativo de búsqueda local. La búsqueda se inició con una configuración aleatoria de los hiperparámetros. En cada iteración, se generaba una nueva configuración aleatoria dentro de una vecindad de la configuración actual. Después, se evaluaba el rendimiento del modelo con ambas configuraciones. Si el rendimiento con la nueva configuración era mejor que con la actual, se actualizaba la configuración. Este proceso se repitió durante 67 iteraciones, y los resultados del entrenamiento con la configuración de hiperparámetros encontrada se muestran la sección 5.

Regresión no lineal (algoritmo de búsqueda armónica). El procedimiento consistió en generar un modelo de regresión no lineal de tipo logNormal, con la métrica F1 designada como variable de respuesta y los valores de los hiperparámetros como variables predictoras. En este modelo

de regresión se seleccionó en función del objetivo de minimizar el error generado. Posteriormente, se formuló un problema de optimización en el que la función objetivo era el modelo de regresión y las restricciones eran los intervalos de cada hiperparámetro. Debido a la naturaleza no lineal del modelo, se empleó la metaheurística de búsqueda armónica propuesta en el trabajo [70] para identificar la configuración óptima de los hiperparámetros (épocas, lotes, peso y tasa de aprendizaje). Los resultados obtenidos con la configuración de hiperparámetros se presentan en la Sección 5.

Capítulo 5

Resultados

Este capítulo presenta una evaluación exhaustiva del algoritmo de pregunta-respuesta diseñado específicamente para el dominio médico en español. Se empleó un corpus textual de gran tamaño, compuesto por más de 3 millones de documentos provenientes de diferentes repositorios médicos. El proceso de respuesta a preguntas se dividió en las siguientes etapas:

- **Recopilación y preprocesamiento:** Se recopilaron documentos médicos de diversas fuentes (PubMed, MedLine, BiosAsq, Cowese, entre otros.) y se sometieron a un proceso de limpieza y normalización, incluyendo tokenización, eliminación de palabras vacías, lematización y derivación.
- **Análisis de las preguntas:** Las consultas realizadas en español fueron analizadas mediante técnicas de PLN, para encontrar las palabras clave o partícula interrogativa, con el objetivo de comprender la intención de clasificar las preguntas y encontrar la mejor respuesta.
- **Recuperación de documentos:** Los documentos y las consultas fueron representados en un espacio vectorial mediante técnicas de caracterización como: TF-IDF y Word embeddings con el objetivo de calcular la similitud textual. Además, se emplearon tres modelos de recuperación de información basados en ranking para identificar los documentos más relevantes a cada consulta, considerando tanto la similitud como factores contextuales. Los modelos implementados son: vectorial, semántico y probabilístico.
- **Extracción de respuestas:** Se aplicaron técnicas de extracción de información basados en grandes modelos de lenguaje para localizar las

partes de los documentos recuperados que contenían la respuesta más precisa a la pregunta generada en lenguaje natural.

Los resultados obtenidos en esta investigación se presentan en dos subsecciones principales. En primer lugar, se analizan los resultados de la fase de recuperación de documentos, donde se evaluó la eficacia de los modelos vectoriales, probabilísticos y semánticos empleados para identificar los documentos más relevantes a cada consulta. Por otro lado, se presentan los resultados de la fase de extracción de respuestas, donde se evalúa la precisión del modelo de lenguaje creado para localizar la información específica que responde a la pregunta médica realizada en español.

5.1. Resultados de la Recuperación de Documentos científicos

En esta sección se presentan los resultados obtenidos para la recuperación de documentos relevantes a partir de una pregunta médica realizada en lenguaje natural. Las configuraciones experimentales se pueden observar de detalle en [4.4.4](#).

La Tabla [5.1](#) muestra los resultados de la implementación de la primera configuración experimental que consta de los tres modelos de RI combinados con un procesamiento utilizando técnicas de PLN y sin el algoritmo LDA. La precisión para la similitud del coseno fue de 22 % y 23 %, mientras que para el modelo probabilístico se obtuvo un valor de 60 % y 63 %. Sin embargo, los resultados para un experimento de LSI no son concluyentes debido a que el proceso de lematización excedió el tiempo de ejecución, interrumpiéndose después de 4 horas. Es importante resaltar como se describió anteriormente que en esta configuración no se utiliza el algoritmo de descubrimiento de tópicos LDA.

Tabla 5.1: Resultados de la primera configuración experimental ordenados por ranking .

Modelo	Configuración experimental	Ranking	Precisión
Probabilístico	preprocesado-lematización	5	0.63
Probabilístico	preprocesado-derivación	3	0.60
LSI	preprocesado-derivación	3	0.50
Similitud del coseno	preprocesado-derivación	3	0.22
Similitud del coseno	preprocesado-lematización	5	0.23
LSI	preprocesado-lematización	5	NA

La Tabla 5.2 muestra los resultados de aplicar la configuración 2 que combina el modelo probabilístico Okapi BM25 con técnicas de PLN, además, a partir de esta configuración experimental se incluye el algoritmo de descubrimiento de tópicos LDA para reducir el espacio de búsqueda de documentos relevantes. Los resultados presentados están ordenados por el valor de la medida F en orden descendente. Se resalta que el experimento mejor posicionado para este enfoque probabilístico fue aplicar la configuración de preprocesado-derivación a los documentos y a la consulta, además de recuperar los 20 mejores documentos. Con ello se obtuvo una precisión del 87 %, una exhaustividad del 69 % y una medida F del 77 %.

Tabla 5.2: Resultados experimentales de la segunda configuración experimental.

Configuración experimental	Top N	Precisión	Exhaustividad	Medida-F
preprocesado-derivación	20	0.87	0.6957	0.7731
preprocesado	20	0.83	0.7007	0.7599
preprocesado-derivación	15	0.82	0.6952	0.7525
preprocesado	15	0.83	0.6773	0.7459
preprocesado-derivación	10	0.82	0.6722	0.7388
preprocesado-lematización	20	0.83	0.6438	0.7252
preprocesado-lematización	15	0.81	0.6337	0.7111
preprocesado	10	0.76	0.6616	0.7074
preprocesado-lematización	10	0.78	0.6194	0.6905
preprocesado-derivación	5	0.75	0.6324	0.6862
preprocesado	5	0.71	0.6290	0.6670
preprocesado-lematización	5	0.72	0.5897	0.6484
preprocesado-derivación	3	0.59	0.5264	0.5564
preprocesado	3	0.57	0.5329	0.5508
preprocesado-lematización	3	0.58	0.5039	0.5393

La Tabla 5.3 muestra los resultados de aplicar la configuración experimental 3 ordenados por la medida-F, está configuración combina el modelo LSI con técnicas de PLN y el algoritmo LDA. El mejor resultado para esta configuración estadística-semántica se obtuvo al aplicar el preprocesado-derivación y recuperar los 20 documentos más relevantes. Esta combinación obtuvo una precisión del 89 %, exhaustividad del 82 % y medida-F del 85 %.

Tabla 5.3: Resultados de la tercera configuración experimental.

Configuración experimental	Top N	Precisión	Exhaustividad	Medida-F
preprocesado-derivación	20	0.89	0.827	0.857
preprocesado	20	0.87	0.814	0.841
preprocesado-derivación	15	0.86	0.807	0.832
preprocesado	15	0.85	0.803	0.826
preprocesado-derivación	10	0.85	0.801	0.825
preprocesado-lematización	20	0.85	0.792	0.820
preprocesado	10	0.84	0.794	0.816
preprocesado-lematización	15	0.83	0.780	0.804
preprocesado-derivación	5	0.82	0.780	0.799
preprocesado-lematización	10	0.81	0.766	0.787
preprocesado-lematización	5	0.79	0.751	0.770
preprocesado	5	0.78	0.751	0.765
preprocesado-derivación	3	0.75	0.726	0.737
preprocesado-lematización	3	0.73	0.703	0.716
preprocesado	3	0.72	0.702	0.711

La Tabla 5.4 contiene los resultados al aplicar la configuración experimental 4 que combina el último modelo de RI que utiliza la similitud del coseno, técnicas de PLN y el algoritmo de descubrimiento de tópicos LDA. Se muestra que la mejor combinación es esta configuración experimental es el preprocesado-derivación y recuperar los 20 documentos top o también llamados relevantes. Para este enfoque de similitud se obtuvo una precisión del 88 %, una exhaustividad del 78 % y una medida-F del 83 %.

Tabla 5.4: Tabla de resultados del experimento 4.

Configuración experimental	Top N	Precisión	Exhaustividad	Medida-F
preprocesado-derivación	20	0.88	0.786	0.830
preprocesado-derivación	15	0.86	0.784	0.820
preprocesado-lematización	20	0.85	0.744	0.793
preprocesado-derivación	10	0.82	0.766	0.792
preprocesado-lematización	15	0.81	0.741	0.774
preprocesado	20	0.83	0.700	0.760
preprocesado	10	0.77	0.727	0.748
preprocesado	15	0.8	0.698	0.746
preprocesado-derivación	5	0.74	0.713	0.726
preprocesado	10	0.75	0.683	0.715
preprocesado-lematización	5	0.71	0.686	0.698
preprocesado-derivación	3	0.65	0.638	0.644
preprocesado	5	0.63	0.606	0.618
preprocesado-lematización	3	0.62	0.611	0.615
preprocesado	3	0.56	0.548	0.554

Con el propósito de visualizar de manera clara los resultados obtenidos, en la Tabla 5.5 se presentan los tres mejores resultados para cada valor de Top-N, ordenados según la medida F. Es importante destacar que la combinación del modelo probabilístico Okapi BM25 con técnicas de PLN, derivación de textos y descubrimiento de tópicos, resultó ser la más efectiva, alcanzando una medida F máxima del 85 %, lo cual evidencia la relevancia de estas técnicas para la tarea de recuperación de información en el contexto médico en español.

Tabla 5.5: Resultados generales, se muestran los 3 mejores resultados para cada Top N.

Modelo	Método	Top N	Precisión	Exhaustividad	Medida-F
Probabilístico	preprocesado-derivación	20	0.89	0.827	0.857
Probabilístico	preprocesado	20	0.87	0.814	0.841
Semántico	preprocesado-derivación	20	0.88	0.786	0.830
Probabilístico	preprocesado-derivación	15	0.86	0.807	0.832
Probabilístico	preprocesado	15	0.85	0.803	0.825
Semántico	preprocesado-derivación	15	0.86	0.783	0.819
Probabilístico	preprocesado-derivación	10	0.85	0.801	0.825
Probabilístico	preprocesado	10	0.84	0.794	0.816
Semántico	preprocesado-derivación	10	0.82	0.765	0.791
Probabilístico	preprocesado-derivación	5	0.82	0.780	0.799
Probabilístico	preprocesado-lematización	5	0.79	0.751	0.770
Probabilístico	preprocesado	5	0.78	0.751	0.765
Probabilístico	preprocesado-derivación	3	0.75	0.726	0.737
Probabilístico	preprocesado-lematización	3	0.73	0.702	0.716
Probabilístico	preprocesado	3	0.72	0.702	0.712

Finalmente, la Figura 5.1 muestra los resultados globales ordenados por la medida-F de la aplicación de las configuraciones experimentales que incluyen modelos de RI, con enfoques probabilísticos, vectoriales y estadísticos, además de la combinación con técnicas de PLN y el algoritmo LDA. Se puede observar que las puntuaciones más bajas de cada enfoque se obtuvieron al recuperar solamente el Top-3 de documentos relevantes. Con este ranking el modelo vectorial obtuvo una medida-F del 55 %, el LSI del 64 % y el probabilístico del 73 %. En contraste, con los valores obtenidos de la medida-F al recuperar el Top-20 de los documentos relevantes, los cuales fueron del 77 % para el modelo vectorial, el semántico del 83 % y el probabilístico del 85 %.

Los resultados indican que la combinación óptima se logró al utilizar el modelo probabilístico OkapiBM25, junto con el preprocesado-derivación y el algoritmo LDA. Esta configuración se seleccionó para la fase de recuperación de información y sus resultados sirven como entrada para la etapa de extracción de respuestas que será evaluada a continuación.

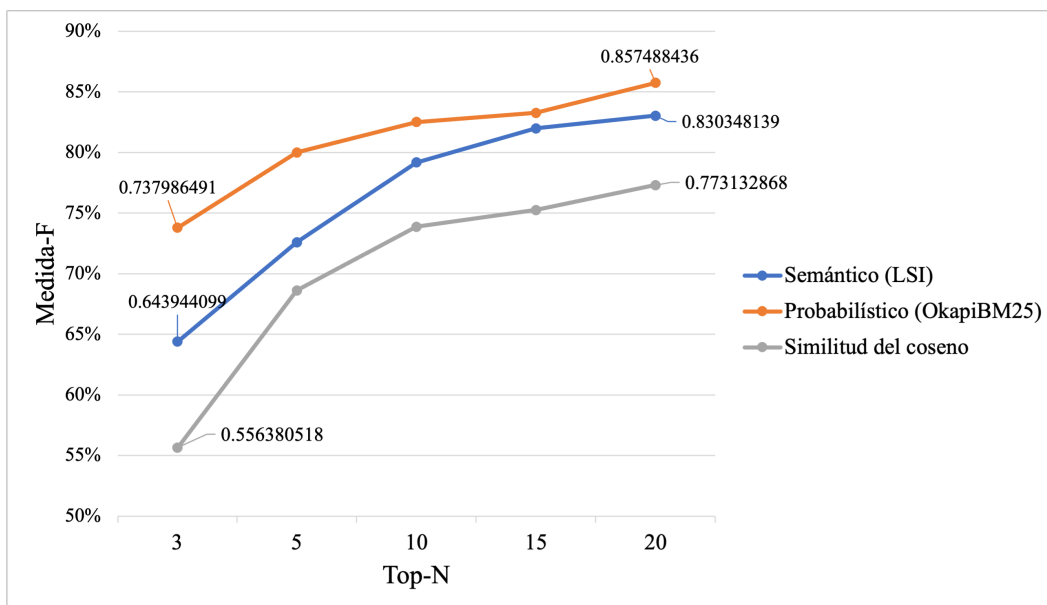


Figura 5.1: Resultados de la comparativa de modelos de RI aplicados a textos científicos sobre COVID-19

5.2. Resultados de la extracción de las respuestas

Una vez identificados los documentos relevantes, se procede a la etapa de extracción de respuestas concisas. Para ello, se ha adaptado el modelo de lenguaje preentrenado BERT, finamente ajustado a un corpus de documentos médicos, para que sea capaz de identificar el fragmento textual que mejor responde a la pregunta del usuario.

Los resultados de esta fase se dividen en dos etapas: la adaptación de un gran modelo de lenguaje preentrenado al dominio médico el cual se denomina médicoBERT y el ajuste fino del modelo resultante para responder preguntas extractivas en español.

5.2.1. Resultados del entrenamiento del LLM médico

La presente sección detalla los resultados obtenidos tras el proceso de entrenamiento del modelo de lenguaje, el cual fue adaptado al dominio médico mediante la incorporación de un corpus textual de 3,036,923 documentos en

español. El modelo, basado en la arquitectura BERT, fue entrenado durante un periodo de aproximadamente 15 días en una GPU NVIDIA 3090 utilizando la librería PyTorch versión 32.4. A fin de evaluar el desempeño del modelo, se emplearon las métricas y los hiperparámetros especificados en secciones anteriores.

Como se observa en la Figura 5.2, el modelo converge sobre los datos de entrenamiento y validación en un número mayor de 40 épocas, lo que indica que alcanza un punto de rendimiento máximo y no se observa una mejora significativa en las siguientes épocas. Este comportamiento sugiere que la configuración de hiperparámetros presentada es la mejor para el entrenamiento del modelo de lenguaje médico en español.

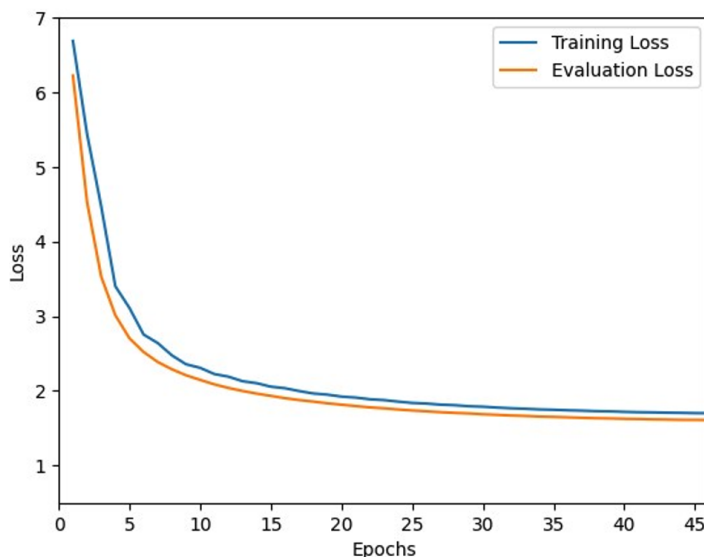


Figura 5.2: Entrenamiento del modelo MédicoBERT.

Las Tabla 5.6 presentan los valores de perplejidad más bajos alcanzados en el entrenamiento del modelo de lenguaje médico en español. Estos resultados se obtuvieron a partir de una serie de experimentos en los que se variaron sistemáticamente el tamaño del conjunto de datos y los hiperparámetros del modelo (número de épocas, tasa de aprendizaje y tamaño de lote). Entre los hiperparámetros utilizados, se observó que la época tuvo el mayor impacto en el rendimiento del modelo al realizar el entrenamiento.

En consecuencia se puede observar que a través de la calibración gruesa

Tabla 5.6: Resultados del entrenamiento del modelo MédicoBERT.

Conjunto de textos	Cantidad de textos	Número de palabras	Tamaño del vocabulario	Épocas	Tiempo de entrenamiento		Pérdida de evaluación	Pérdida de entrenamiento	Perplejidad
					Tokenizador	Modelo			
BioAsq	249,473	45,322,119	373,851	4	00:03:15	15:33:57	4.1030	4.4023	65.41
BioAsq + CoWeSe	1,973,048	903,577,457	752,412	30	00:20:50	100:11:08	2.1091	2.2535	8.25
BioAsq + CoWeSe + CORD-19	3,036,923	1,155,535,281	1,215,013	50	00:30:15	133:20:10	1.9782	2.0411	7.25
BioAsq + CoWeSe + CORD-19	3,036,923	1,155,535,281	1,215,013	40	00:30:15	120:03:00	1.6325	1.6412	4.28

del espacio de hiperparámetros se identificó la configuración adecuada para la adaptación del modelo de lenguaje al dominio médico. Los resultados obtenidos indican que la configuración seleccionada, caracterizada por un bajo valor de perplejidad al procesar textos de mayor longitud y después de 40 épocas de entrenamiento, permitió al modelo adquirir una comprensión profunda del contexto médico. Esta capacidad de comprensión es esencial para la etapa, enfocada en la extracción precisa de respuestas a preguntas médicas, para lo cual se realizó un ajuste fino del modelo médico denominado médicoBERT.

5.2.2. Resultados del ajuste fino del LLM médico

En última instancia, para abordar el objetivo central de esta investigación, se llevó a cabo un ajuste fino del modelo de lenguaje de gran escala al dominio médico. Este proceso tuvo como finalidad habilitar al modelo para la tarea específica de extracción de respuestas. En esta etapa, se emplearon las diversas configuraciones de hiperparámetros optimizadas y detalladas en secciones anteriores.

La etapa de evaluación de este trabajo de investigación se basa en la definición de Magnini [71] “*Dada una pregunta Q y un candidato a respuesta A , decidir si A es la respuesta correcta para Q* ”. En este sentido, la evaluación del algoritmo propuesto se llevó a cabo por medio de colecciones de datos de prueba las cuales se encuentran etiquetados y verificados por expertos en el dominio médico y contienen pares de preguntas y respuestas médicas.

Para evaluar la efectividad del modelo en la tarea final de extracción de respuestas, se calcularon las métricas de precisión, exhaustividad y medida F1, consideradas estándares en el área de PLN. Los experimentos se llevaron a cabo sobre el conjunto de datos BioASQ10, el cual fue dividido en conjuntos de entrenamiento, validación y prueba siguiendo una estrategia de partición estratificada que se muestra la Tabla 5.7. La configuración experimental

consistió en una GPU NVIDIA 3090, el framework PyTorch, la biblioteca Transformers de Hugging Face, el tokenizador subpalabra descrito anteriormente entrenado sobre el corpus médico y las diferentes configuraciones de hiperparámetros obtenidas en la etapa de optimización.

Tabla 5.7: Partición para el ajuste fino del conjunto de datos BioAsq10

Partición	Cantidad de ejemplos	Porcentaje
Entrenamiento	23,808	60 %
Pruebas	7,936	20 %
Validación	7,936	40 %

Para evaluar la capacidad de generalización del modelo médico en la tarea de extracción de respuestas médicas en español, se realizó un procedimiento de validación cruzada de cinco repeticiones por cada configuración de hiperparámetros. En cada repetición, el conjunto de datos BioASQ10 se dividió aleatoriamente en conjuntos de entrenamiento, pruebas y validación, manteniendo una proporción constante entre las particiones. Se calculó la medida F1 para cada repetición y se reportó el promedio de las cinco medidas, lo que permitió estimar la capacidad del modelo para realizar predicciones correctas sobre datos no vistos durante el entrenamiento.

Los resultados de la extracción de las respuestas médicas en español utilizando diferentes estrategias de calibración de hiperparámetros se presentan en las siguientes Tablas.

La Tabla 5.8 presenta los resultados de la extracción de respuestas a partir de textos médicos en español, utilizando dos modelos de lenguaje: el modelo médico (médicoBERT) ajustado finamente y el modelo preentrenado BERT-base. En esta primera fase experimental, se exploró el impacto del ajuste fino en el desempeño de un modelo genérico para una tarea específica. Ambos modelos fueron evaluados utilizando las mismas configuraciones de hiperparámetros reportadas en la literatura para tareas similares, tal como se detalla en la etapa 1 de configuración por optimización de hiperparámetros. Los resultados obtenidos permiten evaluar que tan bueno fue el ajuste fino para mejorar la precisión de la extracción de las respuestas en el dominio médico.

Tabla 5.8: Resultados del entrenamiento de la configuración inicial

Modelo	Épocas	Lotes	Tasa	Peso	Promedio F1
modeloMédico	5	8	1e-5	0.001	48.05
BERT-base	5	8	1e-5	0.001	21.33

La Tabla 5.9 muestra los resultados destacados obtenidos en la fase de búsqueda por calibración gruesa de hiperparámetros, la cual consistió en una búsqueda exploratoria en un espacio de hiperparámetros predefinido y este proceso se describe con mayor detalla en la etapa 2 de optimización de hiperparámetros. Se evaluaron un total de 56 combinaciones diferentes de hiperparámetros, variando la tasa de aprendizaje, el tamaño de lote, el peso y las épocas. De este conjunto de experimentos, se seleccionaron los 10 que obtuvieron el valor más alto en la métrica del promedio de F1.

Tabla 5.9: Resultados del entrenamiento de la tarea de QA con la calibración gruesa de hiperparámetros

Épocas	Lotes	Tasa	Peso	Promedio F1
3.25	8	2.5e-5	0.015	57.5
3.5	4	3.0e-5	0.02	56.29
3.75	4	3.5e-5	0.025	56.01
4	4	4.00e-5	0.03	55.61
4.5	4	5.00e-5	0.04	54.89
3.25	16	2.00e-5	0.01	54.74
4.25	4	4.5e-5	0.035	54.46
3.25	32	2.00e-5	0.01	54.57
3.25	64	2.00e-5	0.01	53.61
3.75	4	5.0e-5	0.08	51.47

En la Tabla 5.10 se muestran los resultados del ajuste fino del modelo de lenguaje médico, empleando las configuraciones de hiperparámetros obtenidas a través de una búsqueda heurística basada en un algoritmo de ascenso por la montaña. Esta etapa, descrita en 3, permitió identificar la combinación de hiperparámetros que maximiza el rendimiento del modelo en la tarea de

QA en español respecto a la métrica promedio de F1. Los resultados presentados en la Tabla corresponden a la mejor configuración encontrada durante el proceso de investigación de esta tesis cuyo objetivo es extraer respuestas a preguntas médicas en español.

Tabla 5.10: Resultados del entrenamiento de la tarea de QA con la búsqueda heurística de hiperparámetros

Épocas	Lotes	Tasa	Peso	Promedio F1
3.5	16	2e-5	0.001	62.35

La Tabla 5.11 presenta los resultados obtenidos en la etapa 3 de la optimización de hiperparámetros, donde se empleó un algoritmo de búsqueda armónica inspirado en la composición musical para encontrar la mejor configuración de que combinen las épocas, lotes, tasa de aprendizaje y peso que ajusten el modelo de mejor forma a la tarea de extracción de respuestas médicas en español. Los resultados alcanzados en esta etapa muestran una similitud estadística con aquellos obtenidos mediante el algoritmo de ascenso por la montaña, lo cual sugiere que ambas técnicas son capaces de identificar configuraciones de hiperparámetros de alta calidad.

Tabla 5.11: Resultados del entrenamiento de la tarea de QA con la calibración fina de hiperparámetros

Épocas	Lotes	Tasa	Peso	Promedio F1
3.25	8	2e-5	0.015	61.72

Finalmente, la Gráfica 5.3 presenta un comparativo del desempeño de diferentes configuraciones de hiperparámetros para la tarea de extracción de respuestas a preguntas médicas en español. Se evaluaron más de 100 configuraciones divididas en 5 categorías, obtenidas mediante diversas estrategias de calibración, incluyendo búsqueda exploratoria, búsqueda armónica y algoritmos de optimización basados en métodos heurísticos. La medida F1, que combina precisión y recall, se utilizó como métrica principal para evaluar el desempeño del modelo.

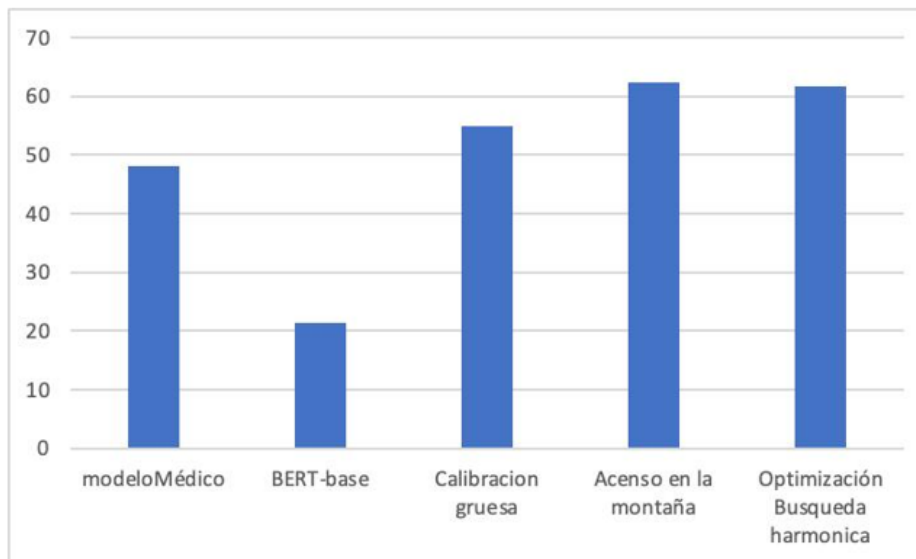


Figura 5.3: Comparativa de los resultados de extracción de respuestas basados en la métrica F1.

Los resultados obtenidos permiten identificar las configuraciones de hiperparámetros que maximizan la precisión y el recall del gran modelo de lenguaje médico para la tarea de extracción de respuestas. La configuración de hiperparámetros obtenida mediante el algoritmo de ascenso por la montaña demostró ser la más efectiva para la tarea de QA. Estos resultados evidencian la sensibilidad del modelo a la configuración de hiperparámetros y la necesidad de optimizar de manera conjunta parámetros como épocas, peso, tasa de aprendizaje y tamaño de lote para alcanzar el máximo rendimiento y que pequeñas variaciones en sus valores pueden tener un impacto significativo en el desempeño del modelo para tarea de extracción de respuestas a partir de textos médicos en español.

En la Figura 5.4, se presentan dos categorías de preguntas que el modelo de pregunta-respuesta puede responder: preguntas cerradas (sí/no) y preguntas abiertas que solicitan información de tipo factoides (hechos). Esta clasificación permite ilustrar la diversidad de interrogantes que el modelo puede procesar y responder de manera efectiva.

Pregunta: <i>¿Cuál es la indicación para Favipiravir?</i> Contexto: <i>En la era de la COVID-19, ha habido una carrera por encontrar fármacos eficaces contra el SARS-CoV-2. Hay informes sobre los usos de Remdesivir y Favipiravir como antivirales existentes contra virus...</i> Respuesta: Covid-19, influenza
Pregunta: <i>¿Se ha evaluado tocilizumab frente a Covid-19?</i> Contexto: <i>Los resultados clínicos preliminares han indicado que el antagonismo del receptor de la IL-6 (IL-6R), incluso con el anticuerpo monoclonal humanizado tocilizumab aprobado por la FDA, puede mejorar los resultados de los pacientes con COVID-19 grave o crítica, manteniendo al mismo tiempo un buen perfil de seguridad...</i> Respuesta: Sí

Figura 5.4: Ejemplos de los tipos de preguntas que puede responder el modelo médicoBERT.

Capítulo 6

Análisis y discusión de resultados

En este capítulo se presenta un análisis de los resultados del trabajo de investigación desarrollado y expuesto, cuyo objetivo principal fue evaluar la hipótesis de que la implementación de técnicas de PLN como la Extracción de Información y un enfoque de aprendizaje profundo en un algoritmo de preguntas en lenguaje natural permitiría generar respuestas precisas y relevantes sobre temas médicos a partir de un corpus de textos científicos en español. Para lograr este objetivo, se diseñó y desarrolló una herramienta computacional que abarca diversas etapas o pasos que incluyen el procesado de los textos, la recuperación de información, el análisis de la pregunta y la extracción de las respuestas.

Entre las diversas etapas o fases descritas, aquellas dedicadas a la recuperación y extracción de información tienen una especial relevancia debido a las contribuciones que aportan al campo de las ciencias de la computación. A continuación, se analizarán en detalle los resultados obtenidos en cada una de estas etapas, las cuales, en conjunto, permiten abordar el problema de pregunta-respuesta en textos científicos médicos en español.

La evaluación de la metodología propuesta de Recuperación de Documentos médicos relevantes se llevó a cabo mediante un estudio que involucró un corpus de gran escala y múltiples combinaciones experimentales. Se exploraron tres modelos de RI que son probabilístico, vectorial y semántico, además se utilizaron técnicas de PLN para el procesamiento de los datos y se incluyó el algoritmo de descubrimiento de tópicos LDA, con el objetivo de identificar reducir el espacio de búsqueda. Los resultados obtenidos indican que la combinación de LDA, procesamiento de texto y stemming con el modelo probabilístico Okapi BM25 resultó en el mejor desempeño, alcan-

zando una precisión del 89 %, un recall del 82 % y una medida F1 del 85 %. Estos resultados superan a los trabajos previos reportados en el estado del arte como [17] y en [34], que combinaron el mismo algoritmo probabilístico Okapi BM25 con patrones léxico-sintácticos o índices invertidos. A diferencia de estos estudios, nuestra propuesta incorpora el algoritmo LDA para filtrar palabras clave y reducir el espacio de búsqueda, lo que ha permitido alcanzar mayor precisión, recall y medida F1. Asimismo, la metodología propuesta supera los enfoques heurísticos de [33] y se compara favorablemente con las redes bayesianas y los modelos lingüísticos presentados en [28]. Incluso en comparación con investigaciones en inglés como la de [72], que utiliza clasificadores de texto basados en aprendizaje profundo, nuestra propuesta ofrece resultados competitivos con una menor demanda computacional.

La metodología propuesta en la etapa de Recuperación de Documentos Relevantes de este trabajo de investigación es altamente reproducible y puede ser aplicada a una amplia variedad de dominios. Si bien se ha evaluado en un corpus de documentos médicos en español, su adaptabilidad a otros campos del conocimiento es evidente. El proceso de recuperación de documentos, tal como se describe, puede ser fácilmente replicado para otros dominios, dada la creciente digitalización de la información y la necesidad de acceder a información relevante de manera eficiente. En consecuencia, la evaluación de la etapa de recuperación de documentos médicos relevantes ha permitido identificar sus fortalezas y debilidades. Si bien los resultados obtenidos son prometedores, es fundamental reconocer las limitaciones inherentes a cualquier sistema de recuperación de información, como el tiempo de procesamiento y la necesidad de recursos computacionales cuando crece de manera considerable el conjunto de búsqueda de documentos. Estas limitaciones deben ser tomadas en cuenta al poner a prueba las fases o etapas proporcionadas en esta tesis en otras aplicaciones prácticas.

Por otra parte, como resultado de la etapa de la extracción de las respuestas se construyó un gran modelo de lenguaje médico el cual tiene por nombre médicoBERT. Este modelo basado en la arquitectura Transformers y aprendizaje profundo es capaz de resolver diferentes tareas de PLN incluyendo QA. Para crear este modelo se realizó un aprendizaje adaptativo y un ajuste fino para habilitarlo en la tarea de contestar a preguntas médicas en español, en ambos casos se refinó el proceso de entrenamiento por medio de la búsqueda de la mejor configuración hiperparámetros para mejorar la eficiencia del modelo.

En el caso del aprendizaje adaptativo del modelo médico, los resultados

demostraron que el entrenamiento de tokenizadores de un dominio específico garantiza una transformación precisa del texto a valores numéricos. Además, se observó un mejor reconocimiento del vocabulario y la terminología especializada en el ámbito médico, lo que permite una extracción de información más efectiva y, en consecuencia, un análisis más profundo y significativo de la literatura médica. El entrenamiento del modelo base con literatura médica ofrece un gran potencial para mejorar el rendimiento y la aplicabilidad de los LLM en áreas específicas. Esto se evidenció al obtener un valor de perplejidad muy bajo de 4.28, lo que indica una adaptación exitosa del modelo al dominio médico en español. Asimismo, se observó que el rendimiento de los modelos de lenguaje en un dominio específico mejora al aumentar la cantidad de textos y palabras disponibles en esa área.

En comparación con estudios previos, el modelo médicoBERT generado en este trabajo de investigación para la etapa de extracción de respuestas se destaca por ser uno de los primeros LLM adaptado al dominio médico en español y específicamente diseñado para responder preguntas médicas. A diferencia de los modelos de lenguaje presentados en [38], [39], [41], [42], [43] y [44] que se enfocan en el dominio médico o clínico, pero en inglés, nuestro modelo brinda la ventaja de estar en español, lo que lo hace más accesible y útil para profesionales de la salud hispanohablantes. Asimismo, se diferencia de modelos en español como los presentados en [44], [?], [47] y [48] que están adaptados a dominios no especializados o generales, ya que este modelo se centra específicamente en el ámbito médico. Cabe destacar que si bien el trabajo [73] presenta un modelo adaptado al dominio médico en español, este fue diseñado para resolver la tarea de reconocimiento de entidades con nombre (NER), mientras que médicoBERT aborda una de las tareas más complejas de PLN como es la extracción de respuestas donde obtuvo resultados prometedores.

Así mismo, los resultados de las diferentes estrategias de calibración de hiperparámetros demostraron que la calibración por optimización mejora sustancialmente los resultados obtenidos en el rendimiento y la eficacia de los grandes modelos de lenguaje para la tarea de pregunta-respuesta. Además, se evidenció que la calibración fina de un modelo de lenguaje médico en español ya sea mediante búsqueda heurística o armónica, produce resultados estadísticamente similares.

Es importante destacar que, para la última fase del algoritmo de pregunta-respuesta, se empleó un gran modelo de lenguaje, médicoBERT, entrenado específicamente con textos médicos en español. Sin embargo, a pesar de su

desempeño mostrado en los resultados, al igual que otros modelos de lenguaje de gran escala, médicoBERT esta sujeto a sesgos inherentes a los datos de entrenamiento y presentar dificultades en la generalización a casos no contemplados durante el entrenamiento. Estos aspectos deben ser cuidadosamente considerados al interpretar los resultados o al utilizarlo en otras aplicaciones en la medicina. Finalmente, Los hallazgos de esta investigación respaldan la idea de que las técnicas de PLN y aprendizaje profundo pueden ser utilizadas para desarrollar enfoques de extracción y recuperación de información médica eficientes y precisos. Pero se debe tomar en cuenta que el campo de los grandes modelos de lenguaje aún se encuentra en una etapa de desarrollo, lo que implica ciertas limitaciones en términos de generalización, robustez, sesgo y recursos computacionales. A pesar de estos desafíos, los resultados obtenidos en esta tesis constituyen un paso hacia el desarrollo de herramientas de apoyo a la investigación en ciencias de la computación y la práctica en el dominio médico.

Capítulo 7

Conclusiones y trabajo futuro

En este trabajo de investigación se desarrolló una serie de pasos ordenados y finitos para responder preguntas en español dentro del dominio médico. Se empleó un enfoque de aprendizaje profundo para la extracción de información, complementado por modelos estadísticos, probabilísticos y vectoriales, con el fin de recuperar información precisa y relevante a partir de un gran corpus de textos en español.

Para cumplir con los objetivos de esta investigación, se recopilaron publicaciones y textos científicos de diversos repositorios especializados en literatura médica, creando un conjunto de datos compuesto por más de 3 millones de textos no estructurados en español. Asimismo, se adaptaron herramientas de Procesamiento del Lenguaje Natural para la limpieza y tratamiento de estos textos, optimizando su formato para el análisis y procesamiento posterior.

Asimismo, en esta investigación se propone una metodología para la recuperación de información o documentos médicos relevantes, basada en la combinación de tres enfoques clásicos: similitud de coseno, búsqueda probabilística y búsqueda semántica, integrados con técnicas de Procesamiento del Lenguaje Natural. La metodología fue evaluada mediante una serie de experimentos utilizando un conjunto de datos compuesto por preguntas médicas. En el proceso, se analizó el desempeño de diversas configuraciones, empleando el algoritmo Asignación Latente de Dirichlet para la reducción de la dimensionalidad del corpus y mejorar la relevancia de los resultados.

Los resultados obtenidos demuestran que la combinación del modelo probabilístico, a través del algoritmo BM25, con LDA y técnicas de PLN, como el stemming y el procesamiento de texto, alcanzó los mejores resultados, logrando un F-score del 85.7%. En esta investigación no solo se implemen-

tan distintas configuraciones experimentales, sino que también introduce un conjunto de datos de preguntas en español creado y evaluado por humanos. Estas contribuciones fortalecen el campo de la Recuperación de Información, al proponer una metodología para recuperar información relevante a partir de grandes corpus de texto en español.

En la fase de extracción de respuestas a partir de preguntas en español dentro de documentos relevantes, se adaptó y ajustó un modelo de lenguaje neuronal de gran tamaño, basado en BERT-base. Para ello, se crearon tres tokenizadores específicos, diseñados para el vocabulario de la literatura médica en español. Posteriormente, el modelo se entrenó utilizando las tareas de MLM y NSP. Los resultados obtenidos indican una exitosa adaptación del modelo al lenguaje médico, con una perplejidad de 4.25 como métrica clave. Las contribuciones de esta fase son tres. Primero, se desarrolló un modelo de lenguaje entrenado con más de 3 millones de textos y 1.1 mil millones de palabras procedentes de la literatura médica en español, capaz de adaptarse a diversas tareas de Procesamiento del Lenguaje Natural mediante técnicas de ajuste fino. Segundo, se implementaron mejoras en la eficiencia del ajuste fino de un gran modelo de lenguaje a través de la optimización de hiperparámetros. Tercero, se creó una herramienta computacional para responder preguntas médicas en español, la cual mostró resultados prometedores al alcanzar un F-score de 62.35 %.

Tras realizar esta investigación, se concluye que la combinación de modelos clásicos de RI con herramientas de vanguardia, como los LLM, permite una recuperación y extracción de información eficaz en grandes corpus en español. Esto es especialmente relevante, dado que actualmente existen pocas herramientas diseñadas específicamente para este propósito en dicho idioma. Además, a medida que las técnicas de aprendizaje adaptativo continúen avanzando, es previsible que los LLM se vuelvan aún más sofisticados, permitiendo una interacción más natural e inteligente en una amplia variedad de tareas y dominios.

Como trabajo futuro, se propone mejorar los resultados en la evaluación de respuestas a preguntas mediante la mejora de la fase de recuperación de documentos relevantes, para ello se utilizarán algoritmos de aprendizaje profundo. Asimismo, se sugiere enriquecer el proceso de recuperación incorporando características lingüísticas y semánticas basadas en recursos como WordNet y MedLexSP. En cuanto a la fase de extracción de respuestas, se plantea optimizar el entrenamiento del modelo mediante una técnica innovadora centrada en la configuración de hiperparámetros, empleando el algo-

ritmo de recocido simulado para mejorar su rendimiento. Además, se busca contestar a preguntas de mayor complejidad, como las de tipo resumen, ampliando así la capacidad del algoritmo para abordar tareas más exigentes.

Asimismo, se evaluará la metodología de Recuperación de Información en un dominio diferente al médico para analizar su adaptabilidad. Por otro lado, se buscará ampliar la aplicación del modelo de lenguaje a gran escala generado, no solo para la tarea de extracción de respuestas a preguntas, sino también para otras tareas de procesamiento del lenguaje natural en el ámbito médico. En particular, se trabajará en la identificación de entidades médicas, como estructuras anatómicas, enfermedades y fármacos. Además, el modelo MédicoBERT se adaptará para la extracción de relaciones semánticas dentro del dominio médico, lo que incluirá la comprensión de relaciones causales y asociaciones entre entidades. Esto abarca, por ejemplo, la identificación de la indicación de un fármaco específico para una enfermedad concreta, o la determinación de la relación de una enfermedad con una o más partes anatómicas. Finalmente, se explorará la posibilidad de utilizar el modelo para la clasificación de textos médicos.

Bibliografía

- [1] V. Ashish, S. Noam, P. Niki, U. Jakob, J. Llion, G. A. N, K. L. ukasz, and P. Illia, “Attention is all you need,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [2] C. Cabezas, “Pandemia de la COVID-19: Tormentas y retos,” *Revista Peruana de Medicina Experimental y Salud Pública*, vol. 37, pp. 603–4, Dec. 2020.
- [3] R. G. Gordon, *Ethnologue: Languages of the world*. SIL International, 2005.
- [4] K. Chowdhary, *Fundamentals of Artificial Intelligence*. Springer India, 2020.
- [5] N. Chomsky, “Three models for the description of language,” *IEEE Transactions on Information Theory*, vol. 2, pp. 113–124, Sept. 1956.
- [6] J. Perez-Carballo and T. Strzalkowski, “Natural language information retrieval: progress report,” *Information Processing & Management*, vol. 36, pp. 155–178, Jan. 2000.
- [7] B.-Y. Ricardo and R.-N. Berthier, *Modern information retrieval*. New York: ACM press, 1999.
- [8] G. Salton, A. Wong, and C. S. Yang, “A vector space model for information retrieval,” *Communications of ACM*, vol. 18, pp. 613–620, 1975.
- [9] S. Dominich, “A unified mathematical definition of classical information retrieval,” *Journal of the American Society for Information Science*, vol. 51, no. 7, pp. 614–624, 2000.

- [10] P. Martínez-Barco, J.-L. Vicedo, E. S. Boró, and D. Tomás, “Sistemas de pregunta-respuesta,” october 2007.
- [11] M. T. Maybury, “Toward a question answering roadmap,” in *New Directions in Question Answering*, 2003.
- [12] T. M. Mitchell, “Does machine learning really work?,” *AI Magazine*, vol. 18, p. 11, Sep. 1997.
- [13] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, pp. 1735–1780, Nov. 1997.
- [14] J. Lee, S. S. Yi, M. Jeong, M. Sung, W. Yoon, Y. Choi, M. Ko, and J. Kang, “Answering questions on COVID-19 in real-time,” in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Association for Computational Linguistics, 2020.
- [15] M. Sarrouiti and S. O. E. Alaoui, “SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions,” *Artificial Intelligence in Medicine*, vol. 102, p. 101767, Jan. 2020.
- [16] I. B. Ozyurt, A. Bandrowski, and J. S. Grethe, “Bio-AnswerFinder: a system to find answers to questions from biomedical texts,” *Database*, vol. 2020, Jan. 2020.
- [17] Y. Cao, F. Liu, P. Simpson, L. Antieau, A. Bennett, J. J. Cimino, J. Ely, and H. Yu, “AskHERMES: An online question answering system for complex clinical questions,” *Journal of Biomedical Informatics*, vol. 44, pp. 277–288, Apr. 2011.
- [18] J. L. Vicedo, R. Izquierdo, F. Llopis, and R. Muñoz, “Question answering in spanish,” in *Comparative Evaluation of Multilingual Information Access Systems*, pp. 541–548, Springer Berlin Heidelberg, 2004.
- [19] D. Su, Y. Xu, T. Yu, F. B. Siddique, E. Barezi, and P. Fung, “CAiRE-COVID: A question answering and query-focused multi-document summarization system for COVID-19 scholarly information management,” in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Association for Computational Linguistics, 2020.

- [20] B. Cairns, R. Nielsen, J. Masanz, J. Martin, M. Palmer, W. Ward, and G. Savova, “The mipacq clinical question answering system,” *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2011, pp. 171–80, 01 2011.
- [21] W. tau Yih, M.-W. Chang, X. He, and J. Gao, “Semantic parsing via staged query graph generation: Question answering with knowledge base,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, 2015.
- [22] W. Zheng, H. Cheng, L. Zou, J. X. Yu, and K. Zhao, “Natural language question/answering,” in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ACM, Nov. 2017.
- [23] G. Zhu and C. A. Iglesias, “Sematch: Semantic entity search from knowledge graph,” in *SumPre-HSWI@ESWC*, 2015.
- [24] G. A. García-Robledo, B. A. González-Beltrán, J. A. Reyes-Ortiz, and M. Bravo, “Extracción de respuestas a partir de ontologías utilizando patrones estructurales en sqwrl,” 2020.
- [25] D. Pappas, R. McDonald, G.-I. Brokos, and I. Androutsopoulos, “AUEB at BioASQ 7: Document and snippet retrieval,” in *Machine Learning and Knowledge Discovery in Databases*, pp. 607–623, Springer International Publishing, 2020.
- [26] B. Xu, H. Lin, Y. Lin, Y. Ma, L. Yang, J. Wang, and Z. Yang, “Improve biomedical information retrieval using modified learning to rank methods,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, pp. 1797–1809, Nov. 2018.
- [27] A. A. Taan, S. U. R. Khan, A. Raza, A. M. Hanif, and H. Anwar, “Comparative analysis of information retrieval models on quran dataset in cross-language information retrieval systems,” *IEEE Access*, vol. 9, pp. 169056–169067, 2021.
- [28] A. Otegi, I. S. Vicente, X. Saralegi, A. Peñas, B. Lozano, and E. Agirre, “Information retrieval and question answering: A case study on COVID-

- 19 scientific literature,” *Knowledge-Based Systems*, vol. 240, p. 108072, Mar. 2022.
- [29] E. Zhang, N. Gupta, R. Tang, X. Han, R. Pradeep, K. Lu, Y. Zhang, R. Nogueira, K. Cho, H. Fang, and J. Lin, “Covidex: Neural ranking models and keyword search infrastructure for the COVID-19 open research dataset,” in *Proceedings of the First Workshop on Scholarly Document Processing*, Association for Computational Linguistics, 2020.
 - [30] A. Esteva, A. Kale, R. Paulus, K. Hashimoto, W. Yin, D. Radev, and R. Socher, “COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization,” *npj Digital Medicine*, vol. 4, Apr. 2021.
 - [31] C. Badenes-Olmedo, B. Lozano-Álvarez, and O. Corcho, “Impact of text length for information retrieval tasks based on probabilistic topics,” *Procesamiento del Lenguaje Natural*, p. 27–36, 2021.
 - [32] A. Berger and J. Lafferty, “Information retrieval as statistical translation,” in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, Aug. 1999.
 - [33] H. Wang, Q. Zhang, and J. Yuan, “Semantically enhanced medical information retrieval system: A tensor factorization based approach,” *IEEE Access*, vol. 5, pp. 7584–7593, 2017.
 - [34] G. A. Tinega, P. W. Mwangi, and D. R. Rimiru, “Text mining in digital libraries using OKAPI BM25 model,” *International Journal of Computer Applications Technology and Research*, vol. 7, pp. 398–406, Oct. 2018.
 - [35] S. Raza, “A COVID-19 search engine (CO-SE) with transformer-based architecture,” *Healthcare Analytics*, vol. 2, p. 100068, Nov. 2022.
 - [36] M. Mahmood, W. J. AL-kubaisy, and B. Al-Khateeb, “Using artificial neural network for multimedia information retrieval,” *Journal of Southwest Jiaotong University*, vol. 54, June 2019.
 - [37] D. V. Rodriguez and D. L. Carver, “Comparison of information retrieval techniques for traceability link recovery,” in *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT)*, IEEE, Mar. 2019.

- [38] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, pp. 1234–1240, Sept. 2019.
- [39] K. Huang, J. Altsaas, and R. Ranganath, “Clinicalbert: Modeling clinical notes and predicting hospital readmission,” 2019.
- [40] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, and M. Villegas, “Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario,” *ArXiv*, vol. abs/2109.03570, 2021.
- [41] N. Liu, Q. Hu, H. Xu, X. Xu, and M. Chen, “Med-bert: A pretraining framework for medical records named entity recognition,” *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5600–5608, 2022.
- [42] Y. Peng, S. Yan, and Z. Lu, “Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets,” in *Proceedings of the 18th BioNLP Workshop and Shared Task* (D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, eds.), (Florence, Italy), pp. 58–65, Association for Computational Linguistics, Aug. 2019.
- [43] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, “Domain-specific language model pretraining for biomedical natural language processing,” *ACM Trans. Comput. Healthcare*, vol. 3, oct 2021.
- [44] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A pretrained language model for scientific text,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 3615–3620, Association for Computational Linguistics, Nov. 2019.
- [45] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano Oller, C. Rodrí-

- guez Penagos, A. Gonzalez-Agirre, and M. Villegas Montserrat, “Maria: Spanish language models,” 2022-03.
- [46] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized BERT pretraining approach,” *CoRR*, vol. abs/1907.11692, 2019.
 - [47] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019.
 - [48] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, “Spanish pre-trained bert model and evaluation data,” *arXiv preprint arXiv:2308.02976*, 2023.
 - [49] J. d. l. Rosa, E. G. Ponferrada, P. Villegas, P. González de Prado Salas, M. Romero, and M. Grandury, “Bertin: Efficient pre-training of a spanish language model using perplexity sampling,” 2022-03.
 - [50] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab, “FlauBERT: Unsupervised language model pre-training for French,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference* (N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Marseille, France), pp. 2479–2490, European Language Resources Association, May 2020.
 - [51] P. Delobelle, T. Winters, and B. Berendt, “Robbert: a dutch roberta-based language model,” *arXiv preprint arXiv:2001.06286*, 2020.
 - [52] M. Polignano, P. Basile, M. Degemmis, G. Semeraro, and V. Basile, “Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets,” in *Italian Conference on Computational Linguistics*, 2019.
 - [53] D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Girju, R. Goordrum, and V. Rus, “The structure and performance of an open-domain question answering system,” in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, (Hong Kong), pp. 563–570, Association for Computational Linguistics, Oct. 2000.

- [54] L. Gasco, A. Nentidis, A. Krithara, D. Estrada-Zavala, R. T. Murasaki, E. Primo-Peña, C. Bojo Canales, G. Paliouras, M. Krallinger, *et al.*, “Overview of bioasq 2021-mesinesp track. evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials,” CEUR Workshop Proceedings, 2021.
- [55] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, p. 993–1022, mar 2003.
- [56] Q. Jipeng, Q. Zhenyu, L. Yun, Y. Yunhao, and W. Xindong, “Short text topic modeling techniques, applications, and performance: A survey,” 2019.
- [57] S. M. Ross, *Introduction to probability models*. CA: Academic Press, 2014.
- [58] A. Trotman, A. Puurula, and B. Burgess, “Improvements to BM25 and language models examined,” in *Proceedings of the 2014 Australasian Document Computing Symposium*, ACM, Nov. 2014.
- [59] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American Society for Information Science*, vol. 41, pp. 391–407, Sept. 1990.
- [60] G. E. Forsythe, M. A. Malcom, and C. B. Moler, *Computer Methods for Mathematical Computations*. Englewood Cliffs: Prentice Hall, 1997.
- [61] C. Cyril and K. Michael, *Factors Determining the Performance of Indexing Systems*. ASLIB, 1966.
- [62] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2018.
- [63] C. P. Carrino, J. Armengol-Estapé, O. de Gibert Bonet, A. Gutiérrez-Fandiño, A. Gonzalez-Agirre, M. Krallinger, and M. Villegas, “Spanish Biomedical Crawled Corpus: A Large, Diverse Dataset for Spanish Biomedical Language Models,” *arXiv e-prints*, p. arXiv:2109.07765, Sept. 2021.

- [64] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. R. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. S. Corrado, M. Hughes, and J. Dean, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *ArXiv*, vol. abs/1609.08144, 2016.
- [65] A.-A. J., *Large Language Models: Concepts, Techniques and Applications*. CRC Press, 2024.
- [66] J. Daniel and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson India, 2000.
- [67] S. Wu and M. Dredze, “Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 833–844, Association for Computational Linguistics, Nov. 2019.
- [68] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ questions for machine comprehension of text,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (J. Su, K. Duh, and X. Carreras, eds.), (Austin, Texas), pp. 2383–2392, Association for Computational Linguistics, Nov. 2016.
- [69] M. Feurer and F. Hutter, *Hyperparameter Optimization*, pp. 3–33. Cham: Springer International Publishing, 2019.
- [70] Z. W. Geem, J. H. Kim, and G. Loganathan, “A new heuristic optimization algorithm: Harmony search,” *SIMULATION*, vol. 76, no. 2, pp. 60–68, 2001.
- [71] A. Vallin, B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peñas, M. de Rijke, B. Sacaleanu, D. Santos, and R. Sutcliffe, “Overview of the CLEF 2005 multilingual question answering track,” in *Accessing Multilingual Information Repositories*, pp. 307–331, Springer Berlin Heidelberg, 2006.

- [72] M. Khadhraoui, H. Bellaaj, M. B. Ammar, H. Hamam, and M. Jmaiel, “Survey of bert-base models for scientific text classification: Covid-19 case study,” *Applied Sciences*, vol. 12, no. 6, 2022.
- [73] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Villegas, “Maria: Spanish language models,” *Procesamiento del Lenguaje Natural*, p. 39–60, 2022.

Anexo A

Comparativa de herramientas de PLN

En la Tabla [A.1](#) se muestra una comparativa de diversas herramientas para el Procesamiento de Lenguaje Natural PLN, donde se revisaron las tareas que tienen soporte para el idioma español.

Tabla A.1: Comparativa de herramientas de PLN

Soporte de español	CoreNLP	OpenNLP	FreeLing	NLTK	IXA pipes
Segmentación de frases	X		X		X
Tokenización	X		X	X	X
Etiquetado morfológico	X	X	X		X
Lematización			X		X
Desambiguación léxica			X		X
Reconocimiento de Entidades Nombradas	X	X	X		X
Sintaxis superficial			X		
Sintaxis de constituyentes	X		X		X
Sintaxis de dependencias			X		X
Correferencia			X		X
Similitud de textos				X	

Anexo B

Palabras vacías del español

“ser”, “estar”, “tener”, “haber”, “él”, “del”, “qué”, “serías”, “tendréis”, “estad”, “ellas”, “se”, “estando”, “hayáis”, “vosotros”, “hube”, “este”, “tuyo”, “mi”, “del”, “tendríamos”, “estaremos”, “poco”, “estuvieron”, “durante”, “era”, “eras”, “habrá”, “estaríamos”, “cual”, “hubieron”, “les”, “tuviésemos”, “vuestro”, “estas”, “tengas”, “otro”, “sentido”, “estuvieseis”, “hubieseis”, “yo”, “tuvisteis”, “a”, “teníais”, “tenemos”, “hubiéramos”, “tuviera”, “nosotros”, “antes”, “pero”, “habías”, “donde”, “seríais”, “habrían”, “estadas”, “habríais”, “eso”, “tuvimos”, “fueras”, “estaría”, “hubieras”, “en”, “ellos”, “estuvisteis”, “hay”, “esa”, “habiendo”, “con”, “todos”, “sí”, “seréis”, “tuvieran”, “sería”, “estuviera”, “erais”, “hubimos”, “sin”, “estos”, “habéis”, “esos”, “habían”, “tuviese”, “muy”, “estáis”, “seáis”, “tengo”, “está”, “habíais”, “una”, “estuvo”, “habidos”, “estarían”, “hubierais”, “quien”, “vuestra”, “hubo”, “porque”, “no”, “sintiendo”, “tuvieseis”, “fuésemos”, “tuvieramos”, “habrán”, “seremos”, “sois”, “esta”, “sentidos”, “le”, “unos”, “estamos”, “ya”, “tienes”, “contra”, “algo”, “tened”, “estar”, “algunos”, “ella”, “fuera”, “has”, “tuvieses”, “hubisteis”, “tu”, “tiene”, “estoy”, “estuviéramos”, “de”, “hasta”, “estuviste”, “hubiste”, “fueron”, “sobre”, “fuimos”, “tenidos”, “he”, “por”, “somos”, “estabas”, “habíamos”, “estés”, “e”, “soy”, “nuestros”, “teniendo”, “tus”, “sentidas”, “la”, “tuve”, “hayas”, “me”, “estará”, “están”, “desde”, “hubiesen”, “nosotras”, “ante”, “su”, “mi”, “estado”, “sus”, “habidas”, “un”, “estaban”, “mis”, “o”, “habréis”, “tienen”, “también”, “quienes”, “haya”, “mucho”, “estéis”, “suyo”, “el”, “estarías”, “otros”, “estás”, “habida”, “serás”, “tanto”, “fueseis”, “estarás”, “estaréis”, “estuviese”, “mías”, “ese”, “estabais”, “tenidas”, “nos”, “mía”, “hubieses”, “fuesen”, “nada”, “tenga”, “estábamos”, “tenida”, “habría”, “tengan”, “él”, “suyos”, “tuvierais”, “teníamos”, “qué”, “tenéis”, “tuvieras”, “estuvieran”, “estuviesen”, “tenido”, “tengamos”, “eran”, “estare”, “vosotras”, “habrías”, “hubieran”,

“tuvo”, “lo”, “nuestras”, “eres”, “son”, “uno”, “esté”, “estén”, “éramos”, “más”,
 “vuestros”, “habido”, “estuvieses”, “para”, “sentida”, “estuviésemos”, “cuando”,
 “tuviesen”, “tendría”, “hayan”, “tendrían”, “al”, “estada”, “serían”, “estemos”,
 “fuerais”, “estuve”, “estarán”, “seré”, “fuisteis”, “nuestra”, “ni”, “habré”, “suyas”,
 “estados”, “tuyos”, “ti”, “nuestro”, “tendré”, “tendrán”, “otras”, “tenían”, “esta-
 ba”, “entre”, “algunas”, “os”, “como”, “tengáis”, “te”, “estuvieras”, “seríamos”,
 “tuviste”, “vuestras”, “y”, “tenías”, “estuvierais”, “míos”, “fuese”, “habríamos”,
 “sea”, “todo”, “suya”, “los”, “habremos”, “fueran”, “hemos”, “hubiésemos”, “es”,
 “había”, “hayamos”, “que”, “serán”, “tenía”, “otra”, “hubiese”, “tendríais”, “hu-
 biera”, “tuvieron”, “fuiste”, “tú”, “seas”, “tendremos”, “siente”, “tuyas”, “fueses”,
 “sentid”, “las”, “sean”, “habrás”, “muchos”, “tuya”, “esas”, “seamos”, “han”, “es-
 to”, “fui”, “tendrá”, “será”, “estuvimos”, “tendrás”, “tendrías”, “fue”, “fuéramos”,
 “estaríais”, “mío”, “ha’

Acrónimos

XML	Extensible Markup Language (Lenguaje de Marcado Extensible)
JSON	Java Script Object Notation (Notación de Objeto de JavaScript)
IBECS	Índice Bibliográfico Español en Ciencias de la Salud
SCIELO	Scientific Electronic Library Online (Biblioteca Científica Electrónica en Línea)
LILACS	Literatura Latinoamericana y del Caribe en Ciencias de la Salud
RI	Recuperación de Información
QA	Question Answering (Pregunta-respuesta)
BD	Base de Datos
PLN	Procesamiento de Lenguaje Natural
EI	Extracción de Información
ML	Machine Learning (Aprendizaje Automático)
RNN	Red Neuronal Recurrente
CNN	Convolutional Neural Network (Red Neuronal Convolutiva)
LSTM	Long Short-Term Memory (Memoria a Largo Plazo)
OMS	Organización Mundial de la Salud
MRR	Mean Reciprocal Rank (Rango Recíproco Medio)
MAP	Mean Average Precision (Precisión Media Media)
UMLS	Unified Medical Language System (Sistema Unificado de Lenguaje Médico)
SQWRL	Semantic Query Enhanced Web Rule Language
POS	Part Of Speech