Centro Universitario UAEM Texcoco

Doctorado en Ciencias de la Computación

Generación de grafos de conocimiento a partir de textos médicos en español

Tesis Doctoral

M. en C. Gabriela Alejandra García Robledo

Directora de tesis
Dra. Alma Delia Cuevas Rasgado

Co-directora de tesis
Dra. Maricela Claudia Bravo Contreras

Tutor
Dr. Farid García Lamont

18 de noviembre de 2025

DEDICATORIA

Para mi más hermoso ángel, mi mamá; para mi superhéroe, mi papá; y para mi compañera de aventuras desde la infancia, mi hermana: gracias por ser mi refugio en cada paso, por sostenerme cuando flaqueo y por regalarme una vida llena de amor.

Y para el amor de mi vida, mi compañero incondicional, Josué: gracias por estar siempre detrás de escena, impulsándome con tu cariño en cada etapa de este camino. Eres una pieza esencial en este logro, y agradezco a la vida, cada día, por cruzarte en mi destino.

Agradecimientos

Agradezco profundamente a mis directoras de tesis, la Dra. Maricela y la Dra. Alma Delia, por su guía, apoyo y por acompañarme con tanta paciencia en este camino. Extiendo también mi gratitud al Dr. Alejandro, quien fungió como un tercer director; cada uno de sus consejos y su mirada objetiva fueron luz en los momentos más desafiantes de la investigación. A los tres los considero mis mentores y ahora también mis amigos. Ha sido un verdadero honor y un placer trabajar con ustedes.

A mi comité revisor, gracias por dedicar tiempo, atención y compromiso para evaluar mi trabajo.

A SECIHTI, gracias por los recursos que hicieron posible alcanzar esta meta. Espero, con todo el corazón, poder retribuir a mi país el apoyo brindado durante esta etapa de formación.

Resumen

En esta investigación se desarrolló un sistema para la representación de conocimiento médico a partir de textos en español, centrado en el reconocimiento de entidades y relaciones médicas. El trabajo abarca desde la identificación precisa de entidades hasta la construcción y validación de un grafo médico. En primer lugar, se implementó un reconocedor de entidades médicas que alcanzó una precisión del 97 %, identificando entidades de tipo Anatomía, Medicamento y Enfermedad. Este componente se entrenó mediante algoritmos de aprendizaje automático (como árboles de decisión y random forest) y fue validado tanto en datos internos como externos, demostrando su robustez. En segundo lugar, se diseñó un modelo para la detección de relaciones entre entidades médicas, utilizando modelos de lenguaje de gran escala ajustados con fine-tuning para tareas de clasificación binaria. El modelo alcanzó una precisión del 90,6 % sobre un corpus anotado manualmente por expertos del dominio. Para apoyar este proceso, se desarrolló una aplicación especializada para el etiquetado eficiente de relaciones, y se definieron siete patrones semánticos que permitieron la generación automática de tripletas informativas a partir de los textos. En tercer lugar, se propuso una metodología para la generación automática de grafos de conocimiento en el dominio médico. Esta fue aplicada a un corpus de 990 artículos científicos, y evaluada mediante cinco casos de uso diversos. La validación realizada por un experto en medicina confirmó la validez general del enfoque. Esta investigación contribuye significativamente al desarrollo de herramientas de Procesamiento de Lenguaje Natural aplicadas a la medicina en español, con potencial para aplicaciones en sistemas de pregunta y respuesta o descubrimiento de conocimiento.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En la actualidad, la información puede encontrarse en una amplia variedad de formatos, tales como archivos digitales, imágenes, grabaciones de audio o simples oraciones redactadas en lenguaje natural. Este último hace referencia a la forma en que los seres humanos se comunican de manera cotidiana, utilizando expresiones propias de su idioma. No obstante, gran parte de esta información digital se presenta de manera no estructurada o semiestructurada, lo que dificulta su procesamiento automático. En este contexto, surgen representaciones estructuradas como las ontologías y los grafos de conocimiento, que permiten modelar el significado de las expresiones lingüísticas y facilitar procesos de razonamiento automatizado.

Para lograr una interacción efectiva entre el lenguaje natural empleado por los humanos y el lenguaje formal binario entendido por las computadoras, es necesario contar con métodos especializados de extracción de información. Estos métodos permiten identificar y aislar características relevantes del contenido que se desea procesar. Su diseño depende en gran medida del tipo de datos: por ejemplo, en archivos de audio se aplican técnicas de análisis de señales; en imágenes, se recurre a la detección de patrones visuales como formas o colores; mientras que en textos planos se emplean enfoques de reconocimiento de entidades. El desarrollo de un sistema de extracción de información plantea múltiples desafíos, especialmente cuando se busca garantizar precisión, relevancia y robustez, independientemente del volumen o la naturaleza del contenido procesado.

La representación de la información es una tarea complementaria e igualmente crucial dentro del campo del Procesamiento de Lenguaje Natural (PLN). Ambas tareas, extracción y representación, requieren una secuencia de fases

interdependientes, tales como el preprocesamiento de los datos, la identificación de conceptos clave, la interpretación contextual del lenguaje, la desambiguación semántica, y el enriquecimiento progresivo de modelos conceptuales. En este sentido, los grafos de conocimiento se han consolidado como una de las herramientas más eficaces para la representación semántica, ya que permiten estructurar la información de manera que sea comprensible para las máquinas, facilitando tanto la búsqueda eficiente como el análisis profundo de relaciones entre entidades.

Los sistemas de extracción de información que incorporan grafos de conocimiento han demostrado ser especialmente útiles en el dominio de la salud. Aplicaciones como los sistemas de pregunta-respuesta centrados en enfermedades, medicamentos o tratamientos se benefician de este tipo de representaciones al reducir significativamente el tiempo computacional necesario para encontrar respuestas pertinentes y confiables.

En este contexto, el presente trabajo propone el desarrollo de un sistema para la generación automática de grafos de conocimiento a partir de textos semiestructurados en español, específicamente artículos científicos del ámbito médico. El sistema contempla dos componentes fundamentales, un reconocedor de entidades médicas capaz de identificar enfermedades, partes del cuerpo y medicamentos, y un identificador de relaciones que permite establecer vínculos semánticos entre dichas entidades. La integración de ambos componentes permite la construcción de un grafo de conocimiento que representa de forma estructurada la información médica contenida en los textos procesados, facilitando así su análisis y utilidad por parte de aplicaciones en el dominio de la salud.

1.1. Planteamiento del problema

La cantidad de investigaciones médicas ha experimentado un crecimiento acelerado en las últimas décadas, impulsado tanto por el surgimiento de nuevas enfermedades como por los constantes avances en diagnóstico, tratamiento y prevención. Esto se traduce en un aumento significativo en la cantidad de artículos científicos publicados diariamente. No obstante, una gran parte de esta información permanece representada exclusivamente en lenguaje natural, sin una estructura semántica formal como la que proporcionan las ontologías o los grafos de conocimiento. Esta ausencia de organización semántica dificulta la búsqueda y recuperación efectiva del conocimiento contenido

en dichas publicaciones. Las publicaciones científicas, aunque redactadas en un lenguaje técnico y estructurado, emplean el lenguaje natural como medio de expresión. Para facilitar el aprovechamiento automático de su contenido, es necesario extraer conceptos y relaciones relevantes que puedan organizarse y representarse mediante estructuras semánticas formales. Este proceso, sin embargo, plantea desafíos significativos en el ámbito del Procesamiento de Lenguaje Natural (PLN), especialmente en lenguas complejas como lo puede ser el español. Entre los principales retos se encuentran la diversidad de construcciones gramaticales, la ambigüedad léxica y semántica, la identificación precisa de entidades y relaciones, así como la redundancia o sinonimia entre conceptos.
En este contexto, se ha desarrollado un sistema orientado a la extracción automática de conocimiento a partir de artículos científicos en español, centrado en el dominio médico. El sistema permite identificar y estructurar la información relevante contenida en los textos, representándola en forma de grafos de conocimiento, lo que facilita su posterior análisis, integración y uso en aplicaciones especializadas.

1.2. Pregunta de investigación

¿Cómo puede desarrollarse un sistema de Procesamiento de Lenguaje Natural que permita la representación automática de grafos de conocimiento a partir de artículos científicos médicos en español?

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar un sistema para la generación automática de grafos de conocimiento a partir de artículos científicos médicos en español.

1.3.2. Objetivos específicos

- Identificar entidades médicas a partir de artículos científicos médicos en español para su clasificación en categorías de tipo enfermedad, anatomía y medicamento.

- Extraer relaciones semánticas entre entidades médicas identificadas para la vinculación entre conceptos relevantes en artículos científicos médicos.
- Implementar un método para la generación automática de grafos de conocimiento a partir de la información médica reconocida y recuperada de artículos científicos médicos en español.

1.4. Justificación

En el Informe sobre la Ciencia 2021 elaborado por la UNESCO [2], se identificó una tendencia global sostenida hacia el incremento en la producción científica, con un crecimiento del 21 % en el número de publicaciones respecto al año 2015. En el ámbito de la salud, las revistas científicas constituyen el principal medio de difusión del conocimiento y desempeñan un papel crucial tanto en la comunicación entre investigadores como en el desarrollo profesional continuo del personal sanitario [3]. Un ejemplo de esta aceleración en la generación de conocimiento se observó durante la pandemia por COVID-19, periodo en el que se registraban más de 1,000 nuevas publicaciones científicas semanales a nivel mundial [4].
A pesar de este crecimiento exponencial, la mayoría de la información médica disponible sigue careciendo de una representación semántica formal, lo cual limita significativamente su accesibilidad. Esta carencia impide, por ejemplo, una búsqueda precisa y rápida de información especializada que permita conectar conceptos relacionados dentro del dominio de la salud. Según datos del Banco Mundial [5], en México existen aproximadamente 4 médicos por cada 1,000 habitantes, y la lengua principal de comunicación es el español. En este contexto, contar con representaciones estructuradas del conocimiento médico, como los grafos de conocimiento, podría ser de gran utilidad para los profesionales del dominio médico, al facilitar la recuperación de información de manera eficiente. Estas estructuras permiten no solo organizar la información, sino también habilitar mecanismos de inferencia automática que expanden el significado y la interrelación entre conceptos médicos.
Los sistemas tradicionales de búsqueda de información implican altos costos computacionales, pues requieren múltiples etapas, que incluyen la recuperación de documentos relevantes y la posterior localización de los datos específicos dentro de los mismos. Por lo tanto, disponer de una representación formal

basada en conceptos médicos y sus relaciones permite una recuperación directa y semánticamente más precisa de la información. En este contexto, se propone el desarrollo de un sistema que genere automáticamente grafos de conocimiento a partir de artículos científicos médicos redactados en español. Esta representación estructurada tiene como propósito servir de apoyo a profesionales del sector salud, tal como, médicos, estudiantes de medicina y personal de enfermería facilitando su acceso a información relevante, precisa y organizada.

1.5. Hipótesis de investigación

La combinación de áreas del Procesamiento de Lenguaje Natural permitirán el desarrollo de un sistema para la representación automática de información en español mediante grafos de conocimiento en el dominio médico.

1.6. Estructura de la tesis

El resto del documento se estructura de la siguiente manera:

- El Capítulo 2 presenta los fundamentos teóricos que sustentan esta investigación, abarcando las áreas clave relacionadas con el Procesamiento de Lenguaje Natural, la extracción de información y la representación del conocimiento.
- El Capítulo 3 presenta un análisis detallado del estado del arte, en el cual se examinan los avances más significativos en tres áreas fundamentales para esta investigación: el reconocimiento de entidades nombradas, la identificación de relaciones semánticas y la representación del conocimiento.
- El Capítulo 4 describe la metodología propuesta, el diseño del sistema y los procesos implementados para la generación del grafo de conocimiento.
- El Capítulo 5 aborda la evaluación del sistema, incluyendo los experimentos realizados, los resultados obtenidos y su interpretación en función de los objetivos planteados.

- El Capítulo 6 presenta las conclusiones generales y propone diversos trabajos a futuro que podrían ampliar o perfeccionar las aportaciones realizadas en esta investigación.

Capítulo 2

Marco teórico

En esta sección, se presentan las bases teóricas que sustentan esta investigación. Los temas principales incluyen Procesamiento de Lenguaje Natural, Extracción de información y Representación de conocimiento.

2.1. Procesamiento de Lenguaje Natural (PLN)

El procesamiento de lenguaje natural es un área de la Inteligencia Artificial (IA) en la cual se estudian y desarrollan métodos y técnicas basadas en el lenguaje para una mejor comprensión y procesamiento desde una computadora.
[6] menciona que todo sistema de PLN se puede dividir en gramática, diccionario y un sistema de programación donde se unen todas las partes. El *léxico* es la división entre la gramática y diccionario.
Augusto Cortez en [7] menciona que una de las tareas fundamentales de la IA es la manipulación de lenguajes naturales usando herramientas de computación, de acuerdo con este, los lenguajes de programación juegan un papel importante porque forman el enlace necesario entre lingüística y su manipulación por una máquina. El PLN consiste en utilizar un lenguaje natural (lenguaje entendido por el ser humano) para comunicarse con la computadora. Utilizar el lenguaje natural (LN) en la comunicación entre una persona y una ventaja es una ventaja, ya que el locutor no tiene que esforzarse para aprender el medio de comunicación a diferencia de la interacción con lenguajes de comando o interfaces gráficas.
De acuerdo con [7] se describen los niveles de un sistema PLN en:

- Nivel fonológico. Describe la relación entre palabras y los sonidos que representan.
- Nivel morfológico. En este nivel se trata de la construcción de palabras a partir de morfemas (unidades de significado más pequeñas).
- Nivel sintáctico. Permite identificar la manera en que las palabras se unen para formar oraciones, así como el papel estructural que tiene cada palabra en la oración y los sintagmas que forman parte de otros sintagmas. Un sintagma se define como una palabra o grupo de palabras que conforman una unidad sintáctica y su función con respecto a otras palabras de la oración.
- Nivel semántico. Demuestra el significado de las palabras y como junto con otros pueden dar un sentido a la oración. También se refiere al significado independiente del contexto.
- Nivel pragmático. Este nivel describe el uso de una misma oración en distintas situaciones y cómo su uso puede afectar su significado.

2.1.1. Extracción de información

La extracción de información estructura el contenido relevante de un texto para estudiar un escenario específico, llamado dominio de extracción. El objetivo de este tipo de sistemas es identificar y unir la información relevante y a su vez ignorar la irrelevante. Algunos autores consideran que la extracción de información es una etapa posterior de la recuperación de información [8]. Alberto Téllez [9] explica que la principal diferencia entre la extracción y recuperación es que la primera brinda la información que interesa, mientras que la segunda puede proporcionar todos los textos en los que aparece la información solicitada. Además, define a una arquitectura general para construir sistemas de extracción de información como "una cascada de módulos que en cada paso agregan estructura al documento, y algunas veces, filtran información relevante por medio de aplicar reglas".
En [10] se describen los componentes típicos de un Sistema de Extracción de Información (SEI):

- Nivel de texto. Determina la importancia de los textos o partes a partir de estadísticas de ocurrencias utilizando patrones de palabras.

- Nivel de palabras. Selecciona las palabras de acuerdo con su función utilizando métodos estadísticos entrenados con textos previamente etiquetados
- Nivel de sentencias. Establece una relación entre las frases a partir de una estructura que muestra las relaciones.
- Nivel inter-sentencias. Identifica y unifica expresiones de referencia con estructuras anteriores.
- Nivel de plantillas. Reanuda la salida en una forma preestablecida.

Los SEI se pueden abordar con diversos enfoques basados en patrones, entre los que destacan los descritos en [10]:

- Patrones léxicos. Palabras utilizadas para la búsqueda de información. Su análisis es independiente al contexto.
- Patrones sintácticos. El etiquetado más utilizado es el POS (part-of-speech), se consideran las características morfológicas y sintácticas del lenguaje.
- Patrones semánticos. Se basan en metadatos semánticos añadidos a la web que ayudan a describir el contenido, significado y relación de los datos.
- Patrones de discurso. Sus palabras hacen referencia a característica de unidades de información dentro de un marco de escritura o estilo.

En [11] clasifican la estructura extraída de una fuente no estructurada en entidades, relaciones y adjetivos que describen entidades.

Identificación de entidades

Se refiere a distinguir los tokens dentro de un texto que contienen un significado relevante en un contexto. El reconocimiento de entidades es la representación de un objeto en el mundo real, por ejemplo, un concepto, una persona o una localidad. La identificación de entidades de un texto tiene como objetivo localizar y clasificar en distintos tipos dentro de un dominio. Actualmente, el término de entidades es ampliado para incluir nombres de

enfermedades, nombres de proteínas, títulos de artículos o nombres de revistas.
Para realizar una identificación de entidades se utiliza un estilo de anotado que trata de indicar con un estándar la parte de la entidad dentro del texto, las partes de una entidad pueden ser inicio, parte o final en el caso de los términos que se conforman de más de una palabra. Los más comúnmente utilizados son:

- IO. Se utiliza para indicar el token dentro de una entidad (Inside) con *I-entidad* y fuera de la misma (Outside) con *O-entidad*.
- BIO. Indica el token inicio de una entidad (Beginning) con *B-entidad*, parte de una entidad (Inside) con *I-entidad* y fuera de ella (Outside) con *O-entidad*.
- BILOU. Utilizado para anotar el token que forma el inicio de una entidad (Beginning) con *B-entidad*, parte de una entidad (Inside) con *I-entidad*, fuera de la misma (Outside) con *O-entidad* y en caso de que la entidad se forme de un solo token (Unique) se anota con *U-entidad*.

Identificación de relaciones semáticas

Las relaciones son aquellas que conectan semánticamente a dos entidades, mismas que ayudan a realizar un análisis más profundo de los textos. Algunos tipos de los más comunes son la sinonimia, hiperonimia, hiponimia o meronimia. Sin embargo, el tipo de relaciones semánticas dependen de las características léxico-semánticas y sintáctico-semánticas que se establecen entre los conceptos o del fin que se persiga con respecto al dominio tratado [12] como lo es la aplicación de relaciones de plantilla (*template relations*) donde se determinan de acuerdo con el tipo de entidades involucradas. El tipo de relaciones semánticas que se desea extraer depende del dominio, grado de estructura del texto y el estilo de escritura.
En [13] se plantean tres casos posibles al realizar una identificación de relaciones:

1. Se cuenta con las entidades identificadas previamente y con la manipulación de pares fijos de entidades se encuentra la relación que existe entre cada una.

2. Se tiene una relación de tipo $R1$ y una entidad identificada A, el objetivo es localizar las entidades X con las que se tiene una relación $R1$.

3. Se tiene un corpus no estructurado de gran tamaño sin entidades identificadas y una relación fija $R1$, se busca obtener todas las entidades que tienen una relación $R1$.

2.2. Representación del conocimiento

De acuerdo con [14] la representación del conocimiento es el uso de símbolos formales para representar una colección de proposiciones creídas por un agente. Por otro lado, en [15] mencionan que la representación del conocimiento actúa como una representación interna de la realidad dentro de un agente. En resumen, la representación del conocimiento es un campo de la inteligencia artificial que nos permite traducir la información del mundo real en una estructura formal procesable por una computadora.
La representación del conocimiento se puede dividir en dos tipos: el conocimiento que nos habla de la información disponible en el mundo en el que se vive y la representación que se encarga de como colocar lo que se sabe. La manera en la que se representa el conocimiento determina lo que se puede manejar.
Uno de los principales objetivos de la representación de conocimiento es realizar la inferencia (conclusión), por esta razón es necesario expresarlo en una forma computable. Para esto se requiere de un esquema de representación que es un instrumento para transformar el conocimiento de un dominio en un lenguaje formal construido por una sintaxis y semántica para que pueda procesarse de manera computacional.
La sintaxis describe las formas de construir o hacer una combinación de los símbolos de un lenguaje y la semántica determina el significado de dichos elementos y la relación entre ellos a partir de su referencia del mundo real.
Según Jordi Duran Cals en [14] un esquema de representación debe contar de ciertas propiedades que son relativas al uso de la representación:

- Representación apropiada. Debe de ser capaz de representar todo el conocimiento que sea necesario para el dominio en que se trabaja.

- Inferencia apropiada. Manipular las estructuras para que en todo momento se pueda derivar nuevas estructuras asociadas con conocimiento nuevo o inferido.
- Eficiencia inferencial. Se puede implementar el proceso de inferencias mediante el uso de heurísticas
- Eficiencia de adquisición. Se permite la incorporación de conocimiento nuevo

El proceso de representación de conocimiento comienza con el planteamiento y análisis del problema que se quiere resolver, el cual permite al ingeniero de conocimiento identificar el dominio. Requiere después adquirir el conocimiento lo cual es catalogada como una tarea compleja que necesita de características especiales como el de la percepción, comunicación asociación y razonamiento para obtenerlo. Después se debe codificar el conocimiento obtenido en un esquema desarrollado lo que permite decidir si es brindado a terceros u otras fuentes de información. El motor de inferencia es necesario en esta parte del proceso para poder obtener las conclusiones del problema a partir del conocimiento obtenido anteriormente. El último paso es el como se interpretan las conclusiones para tener una solución de un problema inicial.

2.2.1. Grafos de conocimiento

Existen distintas definiciones de un grafo de conocimiento, en [16] se define como un grafo de datos destinado a acumular y transmitir conocimiento del mundo real donde los nodos representan entidades y las aristas representan relaciones entre ellas. Su conocimiento puede obtenerse de fuentes externas o extraerse de otro grafo de conocimiento.
Los grafos de conocimiento se pueden formar a partir de numerosas fuentes y resultar altamente diversos en términos de estructura, se requieren de métodos efectivos para la extracción, enriquecimiento y evaluación de calidad para que un grafo de conocimiento crezca y mejore con el tiempo. En [17] describe que debido a que los grafos de conocimiento tienen la capacidad incorporar información de diferentes fuentes de datos (estructurados, no estructurados o semiestructurados) se entiende que está construido sobre distintas bases de datos que al unirse enriquecen su significado. Este incremento procede de la ontología donde se definen entidades y propiedades, además de representar el tipo y clase del conocimiento en el dominio. De esta forma un sistema

demuestra que un tipo de entidad puede estar relacionada con otra. El modelado de un grafo de un dominio tiene la posibilidad de conectarlo con otro modelo de dominio.
En [18] explica que la base de conocimiento de un grafo es construida y escrita por millones de hechos. También afirma que los grafos de conocimiento son redes semánticas muy grandes que integran varias y heterogéneas fuentes de información para representar el conocimiento sobre ciertos dominios del discurso. La adquisición de conocimiento describe el proceso de extraer información de diferentes fuentes, estructurarla y crear conocimiento útil. El conocimiento podría representarse de manera diferente entre sistemas, como documentos de texto, páginas web, bases de datos de relaciones y bases de datos. Los datos semánticos proporcionan una forma de superar esta discrepancia en la representación de datos y proporcionar los datos de forma estructurada.

Lenguaje de representación de conocimiento - RDF

RDF (Resource Description Framework) se define en [19] como un modelo conceptual que proporciona información que describe recursos que se encuentran en la Web y permite el intercambio de información con ayuda de aplicaciones para que los datos no pierdan su significado. RDF permite la reutilización de recursos.
RDF permite la mezcla de diferentes fuentes de datos, aunque los esquemas sean distintos. Su idea general es describir las cosas con ciertas propiedades que tienen valores. Se realizan siempre declaraciones de la forma sujeto-predicado-objeto que se conoce como tripletas.
Para hacer uso de RDF se necesita un sistema de identificadores únicos (Identificadores Uniformes de Recursos - URIs) y un lenguaje que sea procesable de forma automática para representar sus relaciones.

2.3. Aprendizaje automático

El aprendizaje es aumentar los conocimientos o habilidades para realizar una tarea o mantenerse en un entorno. En el caso de las máquinas se busca que tengan un aprendizaje artificial de una tarea para que puedan realizarlo de forma automática.
En [20] definen aprendizaje automático como un proceso de fases. Comien-

za cuando el sistema selecciona las características relevantes de un objeto o evento, las compara con otras ya conocidas con un proceso de cotejamiento y si las diferencias son significativas adapta el modelo con el objeto de acuerdo con el resultado. Para hacer que un sistema conozca características (entrenamiento) se pueden aplicar métodos matemáticos complejos, métodos de búsqueda en repositorios, entre otros.
Los paradigmas del aprendizaje automático de acuerdo con la selección y transformación de la información son:

- Aprendizaje deductivo. A partir de características de hechos ya conocidas se realizan inferencias para derivar nuevos hechos.
- Aprendizaje analítico. Formulan generalizaciones a partir del análisis de instancias. Requiere que se proporcione un amplio conocimiento del dominio, se centran en la mejora de la eficiencia y no en obtener nuevas descripciones.
- Aprendizaje analógico. Busca imitar algunas capacidades humanas. Resolver un problema a partir de las semejanzas con otros vistos anteriormente.
- Aprendizaje inductivo. Utilizado en problemas que a partir de un conjunto de ejemplos y contraejemplos busca inducir una descripción.

Los métodos de aprendizaje en IA se clasifican de la siguiente forma:

- Supervisados. Los datos de entrada etiquetados son necesarios para poder tener un aprendizaje de una tarea
- No supervisados. Desarrollan nuevos conocimientos por descubrimiento. Los datos de entrada no es necesario que estén etiquetados.

2.3.1. Algoritmos de aprendizaje automático

A continuación, se describen algunos de los algoritmos más utilizados para la tarea de clasificación.

Árboles de decisión

Un árbol de decisión es un algoritmo de aprendizaje supervisado utilizado para tareas de clasificación. Maneja una estructura de árbol jerárquico con parte conocidas como nodo raíz, ramas, nodos internos y nodos hoja [21].

Random Forest

Random Forest (Bosque aleatorio) es un algoritmo que combina la salida de múltiples árboles de decisión para alcanzar un solo resultado. Se utiliza para problemas de clasificación y regresión [21].

2.3.2. Aprendizaje profundo

El aprendizaje profundo (Deep learning) es una subárea del aprendizaje automático que se basa en el uso de redes neuronales artificiales con múltiples capas (también llamadas redes profundas) para modelar representaciones jerárquicas de los datos. Estas redes son capaces de aprender automáticamente características complejas a partir de grandes volúmenes de datos, sin necesidad de una ingeniería de características manual intensiva.
En [22] explican cómo el aprendizaje profundo ha transformado áreas como la visión por computadora, el reconocimiento de voz y el procesamiento del lenguaje natural, al permitir que los modelos extraigan automáticamente patrones complejos sin intervención humana directa.

Grandes modelos de lenguaje

Los Grandes Modelos de Lenguaje (LLM) son sistemas de inteligencia artificial basados en aprendizaje profundo, diseñados para generar texto de manera coherente y contextual. Estos modelos han sido preentrenados con grandes volúmenes de datos generalmente de dominio general, lo que les permite aprender las estructuras y patrones del lenguaje natural.
Gracias a su capacidad para producir texto fluido y semánticamente relevante, los LLM se aplican en una amplia variedad de tareas del procesamiento del lenguaje natural (PLN), tales como clasificación de texto, generación automática de contenido, traducción automática, entre otros.

Estos modelos se basan en la arquitectura Transformer, prrsentada en 2017 en el artículo "Attention Is All You Need" [1], y mostrada en la Figura 2.1.

La arquitectura Transformer constituye una innovación significativa en el diseño de redes neuronales, al basarse exclusivamente en mecanismos de atención, eliminando por completo la recurrencia y las convoluciones. Esta estructura permite procesar secuencias de entrada y salida capturando eficientemente relaciones de largo alcance entre sus elementos, independien-

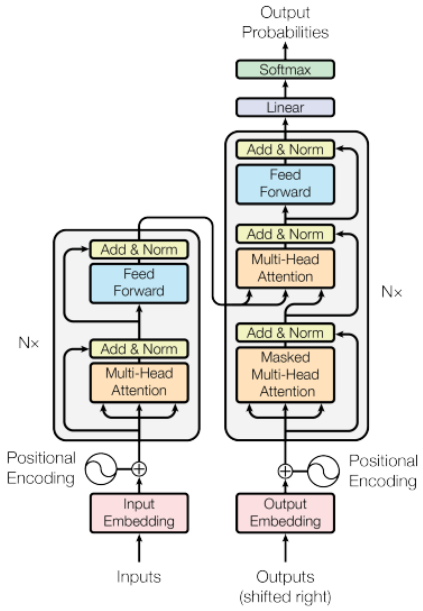

Figura 2.1: Arquitectura Transformer. Obtenida del artículo "Attention Is All You Need" [1].

temente de su posición relativa.
El modelo está compuesto por dos módulos principales: el codificador (encoder) y el decodificador (decoder). Ambos están formados por bloques repetidos que incluyen dos subcapas fundamentales: mecanismos de autoatención (self-attention) y redes neuronales feed-forward, lo que permite representar de forma jerárquica la información contextual en diferentes niveles de abstracción.

Capítulo 3

Estado del arte

En esta sección, se presenta el estado del arte de las áreas clave abordadas en esta investigación, tales como grafos de conocimiento, identificación de entidades y identificación de relaciones.

3.1. Grafos de conocimiento

A continuación, se describe una selección de trabajos relacionados cuyo objetivo principal es la representación de información mediante grafos de conocimiento.
La plataforma de extremo a extremo propuesta en [23] construye grafos de conocimiento a partir de textos no estructurados utilizando tecnologías como Apache Solr, Stanford CoreNLP, Apache Spark y Neo4j. Este sistema se divide en dos fases: la primera es la extracción, en la que se identifican entidades nombradas, relaciones entre ellas y enlaces a entidades en un grafo de conocimiento externo. Esta extracción puede realizarse en todos los documentos o en un subconjunto específico proporcionado por el usuario. La segunda fase, el enriquecimiento, consiste en ampliar el grafo con hechos provenientes de fuentes externas.
En [24], se propone la construcción de un subgrafo denominado PDD (pacientes-medicamentos-enfermedades), que se construye a partir de un conjunto de datos de medicamentos (DrugBank) y una ontología ICD-9. Este enfoque conecta datos clínicos del MIMIC-III sobre prescripciones y notas médicas de pacientes, permitiendo encontrar nuevas relaciones entre medicamentos y diagnósticos.

El trabajo presentado en [25] aborda la construcción de un grafo de conocimiento histórico, que representa la evolución de un término médico a lo largo del tiempo. Su objetivo principal es mostrar cómo cambian las ontologías, utilizando un algoritmo que puede ser ejecutado de manera ascendente y descendente para mapear las anotaciones semánticas y su impacto a lo largo del tiempo.
En [26], se introduce el enfoque OIE4KGC, que transforma documentos no estructurados en un grafo de conocimiento, donde los vértices representan conceptos y documentos, y las aristas las relaciones entre ellos. El proceso incluye la identificación de conceptos relevantes y su enlace utilizando el Sistema Unificado de Lenguaje Médico (UMLS) y los Identificadores de Conceptos Únicos (CUI), mejorando así la organización y recuperación de información.
El modelo propuesto en [27] integra conocimiento médico textual heterogéneo con datos de salud para soportar consultas semánticas y razonamiento. Utiliza un algoritmo de poda para filtrar inferencias irrelevantes, lo que permite representar relaciones semánticas complejas entre entidades médicas de diversas fuentes, como la web y libros médicos.
En [28], se construye un grafo de conocimiento a partir de texto plano mediante técnicas de Procesamiento de Lenguaje Natural, extracción de información y relaciones semánticas. Este enfoque utiliza tripletas RDF y un etiquetado de roles semánticos para representar la estructura de las relaciones y la organización coherente de las entidades.
El trabajo en [29] propone una estructura de cuadrupletas para representar el conocimiento médico, mejorando la clasificación de entidades mediante un algoritmo de traducción probabilística. Este enfoque permite construir grafos de conocimiento a gran escala a partir de registros médicos electrónicos.
El modelo presentado en [30] utiliza características faciales y la frecuencia cardiaca (FC) para predecir estados emocionales. A través de un aprendizaje profundo, se integran datos de videos faciales y señales fisiológicas en un grafo de conocimiento, permitiendo predecir emociones complejas.
En [31], se emplean grafos de conocimiento para la clasificación de textos breves. Este método enriquece la información del texto con un grafo de conocimiento externo y utiliza redes neuronales de grafos para capturar interacciones implícitas entre conceptos.
En [32], se propone un sistema de análisis de poesía clásica china basado en un grafo de conocimiento. Este enfoque permite analizar el tema y las emociones de los poemas mediante la construcción de un grafo ontológico de la

poesía clásica a partir de la combinación de modelos como BERT.
El método descrito en [33] resuelve la correferencia en textos biológicos mediante un grafo de conocimiento, proporcionando un flujo de trabajo que incluye la resolución de términos de referencia y la visualización de la información en un grafo.
Finalmente, [34] aborda la detección y resolución de datos anormales en grafos de genealogía, utilizando procedimientos específicos para tratar relaciones críticas entre entidades y mejorar la precisión del análisis.
A continuación, se presenta una comparativa de los trabajos analizados en esta sección en la Tabla 3.1.

Tabla 3.1: Estado del arte

	Tipo de ambiente	**Fuente de datos**	**Idioma**	**Enfoque**	**Evaluación**
Clancy, R. et al. [23]	General	TREC Washington Post Corpus	Inglés	Tripletas	580K documentos
Wang, M. et al. [24]	Medicina	MIMIC-III, ICD-9, DrugBank	Inglés	Tripletas RDF	Precisión de 94 %
Domingos S. et al. [25]	Medicina	SNOMED-CT, MESH, NCIt, ICD-9	Inglés	Algoritmo heurístico	26 % - 70 %
Muhammad I. et al. [26]	Medicina	ORRCA, ClauseIE	Inglés	Tripletas	Precisión - 47 % (ClauseIE) 78 % (ORRCA)
Longxiang S. et al. [27]	Medicina	Sistema de Información de Salud en Zhejiang, China	Chino	Algoritmo de poda	Precisión 92 %
Martinez J. et al. [28]	General	Páginas web de noticias TI	Inglés	Tripletas RDF	Precisión: sujeto 72 %, predicado 89 %, objeto 64 %, relación 82 % y tripletas 51 %
Linfeng L. et al. [29]	Medicina	Registros electrónicos médicos	Inglés	Cuadrupletas, redes neuronales y algoritmo de aprendizaje (PrTransH)	Precisión 97 %
Wenying Y. et al. [30]	General	Dataset DEAP	Inglés	Aprendizaje profundo, redes neuronales 3D profundas	Desv. est. (placer: 73 %, excitación: 53 % y dominio 72 %)
Xuhui J. et al. [31]	Noticias	Títulos de noticias de China de NLPCC2017	Chino	Redes neuronales	Precisión 80 %
Yuting W. et al. [32]	Poesía clásica	Website de clásicos chinos	Chino	Algoritmo A priori y combinación modelos (BERT)	Precisión: Rand 62 %, +BERT 71 %
Tai W. and Huan L. [33]	Académico	Libro de biología de 8º grado	Inglés	Reglas + semánticas	Precisión 90 %
Jianxuan S. et al. [34]	General	BD Huapu Big Knowledge Graph	Inglés	Algoritmo propio	Precisión: 84 %

3.2. Identificación de entidades

El reconocimiento de entidades nombradas ha sido ampliamente estudiado en diversos contextos, incluyendo el ámbito biomédico y los textos en español. A continuación, se presentan diferentes enfoques y modelos utilizados en este campo.

En [35] desarrolla un modelo para el reconocimiento de entidades nombradas en chino a partir del conjunto de datos del Diario del Pueblo. Este modelo utiliza BERT para extraer vectores dinámicos de palabras y mejorar el aprendizaje de características con BiLSTM. Posteriormente, se emplea CRF y un mecanismo de atención para aprender la probabilidad de transición entre etiquetas y mejorar la consistencia del etiquetado. Por su parte, NeuroNER [36] es una herramienta basada en redes neuronales artificiales para el reconocimiento de entidades nombradas. Acepta como entrada conjuntos de datos etiquetados para entrenamiento, validación y prueba, además de permitir la clasificación de nuevos textos sin etiquetar.

En el ámbito del turismo, [37] propone un modelo de reconocimiento de entidades turísticas utilizando BERT preentrenado y codificación BiLSTM para representar contextos. Se introduce un mecanismo de smoothing cross para mejorar la precisión en la asignación de etiquetas, aplicando el modelo en datasets con entidades como localización, persona y organización.

Dentro del campo de la salud en [38], se aborda el NER en registros médicos electrónicos en idioma chino utilizando un modelo BiLSTM-CRF complementado con un mecanismo de atención. El objetivo es reducir la inconsistencia del etiquetado, identificando entidades escritas de distintas formas pero que pertenecen al mismo tipo. Además, se implementa un algoritmo de autocorrección y un diccionario de fármacos con sus nombres comerciales y de producto para mejorar la detección de entidades.

CollaboNet [39] introduce un enfoque basado en la colaboración entre múltiples modelos especializados en diferentes tipos de entidades (sustancias químicas, enfermedades y genes). Cada modelo actúa como experto en su tipo de entidad y colabora con los otros para minimizar la ambigüedad y reducir los falsos positivos. Por otro lado, [40] desarrolla un sistema híbrido basado en reglas para la detección automática de pacientes, enfermedades, síntomas, elementos de laboratorio, medicamentos y tratamientos en registros médicos. Este sistema combina procesamiento de texto, análisis gramatical y descubrimiento de entidades con un módulo de razonamiento del conocimiento para normalizar entidades y relaciones.

En el dominio médico y específicamente en textos en español, se han desarrollado diversos modelos para el reconocimiento de entidades nombradas. Por un lado, el modelo presentado en [41] emplea información de definición de entidades mediante los métodos SQuad y SOne, y se basa en BERT para identificar distintos tipos de entidades, incluyendo términos normalizables y no normalizables. Su capacidad para clasificar términos médicos con variabilidad en su escritura lo hace especialmente útil en este campo. Por otro lado, PharmacoNER Tagger [42] adapta el reconocedor de entidades NeuroNER a textos clínicos en español, incorporando etiquetado POS y componentes de aprendizaje profundo para mejorar la identificación de entidades en casos clínicos.

Finalmente, en [43] se exploran técnicas de redes neuronales recurrentes y grandes modelos de lenguaje para el reconocimiento de entidades en español a partir de historias clínicas electrónicas en España. Se implementa un procedimiento para ampliar la cantidad de textos disponibles y mejorar el entrenamiento de los modelos. Asimismo, [44] presenta un marco de aprendizaje por transferencia basado en modelos de lenguaje preentrenados con documentos biomédicos multilingües, permitiendo la identificación de enfermedades y organismos en datos clínicos en español. Este modelo se complementa con una aplicación web para minería de textos clínicos.

Estas investigaciones reflejan la diversidad de enfoques y técnicas utilizadas en el reconocimiento de entidades nombradas, desde modelos basados en reglas hasta sofisticadas arquitecturas de aprendizaje profundo que aprovechan modelos de lenguaje preentrenados. A continuación, la Tabla 3.2 presenta una comparación de los distintos trabajos analizados en esta sección.

Tabla 3.2: Estado del arte identificación de entidades

Autores	Dominio	Idioma	Enfoque	Entidades identificadas	Evaluación
Bin, J. et al. [38]	Médico	Chino	BiLSTM-CRF Attention-BiLSTM-CRF Reglas postprocesamiento	Anatomía Descripción de los síntomas Síntoma independiente Fármaco Cirugía	P: 91.2 % R: 90.3 % F1: 90.8 %
Tang, X. et al. [35]	General	Chino	BERT-BiLSTM-AM-CRF	Persona Organización Ubicación	P: 95.7 % R: 96 % F1: 95.8 %
Yoon, W. et al. [39]	Médico	Inglés	BiLSTM-CRF	Enfermedad Sustancia química gen/proteína	P: 85.1 % R: 85.2 % F1: 85.1 %
Xiong, Y. et al. [41]	Médico	Español	BERT	Normalizables Proteínas No normalizables	P: 92.2 % R: 91.2 % F1: 91.3 %
Armengol-Estapé, J. et al. [42]	Médico	Español	RNA-CRF	Generales Normalizables Proteínas Sustancias químicas medicamentos	A: 99.5 % P: 92.3 % R: 89.7 % F1: 91 %
Dernoncourt, F. et al. [36]	General	Inglés	BRAT, RNA	Ubicación Organización Personas Otras	F1: 90.5 % y 97.7 %
Chen, L. et al. [40]	Médico	Inglés	Reglas basadas en patrones y LSTM	Enfermedades Síntomas Elementos de laboratorio Medicamentos Tratamientos	Basado en reglas F1: 93.8 % Híbrido F1: 84.9 %
Kai, G. et al. [37]	Turismo	Inglés	BERT y BiLSTM	Personas Ubicación	P: 81.9 % R: 81.4 % F1: 81.6 %
Moreno-Barea, F. et al. [43]	Médico	Español	RNN y XML-RoBERTa	Entidades de expedientes médicos: Centro de Salud Historial Médico Contacto de paciente Generales: Persona Ubicación	RNN P: 86.1 % R: 88.9 % F1: 86.9 % XML-RoBERTa P: 96.3 % R: 98.5 % F1: 97.3 %
Tamayo, A. et al. [44]	Médico	Español	RoBERTa y BERT multilingüe	Enfermedades Organismos	Organismos P: 85 % R: 90 % F1: 88 % Enfermedades P: 62 % R: 75 % F1: 68 %

3.3. Identificación de relaciones

La identificación de relaciones semánticas es una tarea del lenguaje natural cuyo objetivo es identificar y clasificar las relaciones existentes entre entidades en un texto. A continuación, se presentan diversos enfoques desarrollados en diferentes dominios y tipos de datos.

Un enfoque basado en etiquetas múltiples es propuesto en [45], el cual consta de tres capas: una de extracción de características utilizando una red neuronal BiLSTM sobre una oración con dos entidades objetivo, otra de agrupación de entidades mediante un módulo de atención y, finalmente, una capa de predicción de relaciones que incorpora una función de pérdida de margen deslizante con una etiqueta adicional de "sin relación".

En [46], se implementa un mecanismo de incrustación en tres modelos de identificación de relaciones existentes, mejorando su rendimiento y escalabilidad. Este enfoque se centra en extraer relaciones dentro de una misma oración, generando ternas que representan las relaciones entre entidades previamente identificadas.

Por otro lado, en [47], se analiza la correlación de coocurrencia de relaciones a nivel de documento, argumentando que las relaciones reflejan la distancia semántica. Se emplea una incrustación de relaciones y dos subtareas de predicción de ocurrencia para identificar correlaciones de relaciones, aplicadas en la extracción de hechos. Mientras que en [48], se desarrolla un clasificador de relaciones semánticas utilizando Máquinas de Soporte Vectorial y combinando características léxicas, de entidades nombradas y estructuras sintácticas, como árboles de análisis sintáctico. Para mejorar el preprocesamiento y el entrenamiento, se traduce el corpus al español y se evalúan diversas combinaciones de características.

Abordando dominios específicos en [49], se introduce una red neuronal iterativa diseñada para la identificación de relaciones anidadas con sensibilidad a menciones, generando de manera iterativa tuplas candidatas para clasificación y permitiendo la identificación de relaciones en capas arbitrarias. En el ámbito del análisis de noticia, [50] propone un método basado en una Máquina de Soporte Vectorial (SVM), en el cual las oraciones son transformadas en matrices numéricas mediante incrustaciones de palabras y posiciones, extrayendo características clave para la generación de vectores. Además, este enfoque incorpora modelos basados en Análisis de Componentes Principales (PCA) y Redes Neuronales Convolucionales (CNN) para mejorar la precisión en la clasificación.

Por otra parte, en el dominio geográfico, [51] compara el desempeño de distintas configuraciones de hiperparámetros en modelos de lenguaje previamente entrenados con datos en chino, abordando tanto la identificación de entidades como la identificación de relaciones espaciales. Asimismo, se evalúa un enfoque basado en reglas, destacando sus limitaciones en comparación con modelos más avanzados. En el ámbito legal, [52] aplica modelos de lenguaje para la identificación de relaciones en documentos sobre leyes laborales en España, con el propósito de construir de manera semiautomática un grafo de conocimiento que facilite la estructuración y análisis de la información jurídica.

En el dominio médico, la identificación de relaciones semánticas es crucial para comprender la interacción entre entidades como enfermedades, tratamientos y pruebas. En este contexto, [53] propone un sistema basado en aprendizaje profundo para la identificación de relaciones entre entidades médicas a partir de recetas de pacientes. Este sistema identifica relaciones entre enfermedades, pruebas y tratamientos médicos mediante una red neuronal convolucional compuesta por cuatro módulos: incrustación de palabras, extracción de características, convolución y clasificación con softmax. Por otro lado, el modelo convolucional jerárquico presentado en [54] aborda la segmentación de las oraciones según su estructura semántica para capturar relaciones de manera más precisa. Este modelo divide inicialmente la oración en cinco canales para extraer información estructural, permitiendo que cada palabra aprenda diferentes representaciones semánticas. Luego, emplea un modelo BERT para generar representaciones contextualizadas y realizar convoluciones tanto a nivel de token como de canal, codificando de este modo las dependencias semánticas a lo largo de toda la oración, mejorando así la identificación de relaciones en el ámbito médico.

Estos enfoques brindan una visión de la variedad de técnicas empleadas en la identificación de relaciones semánticas, que van desde métodos basados en reglas hasta modelos de aprendizaje profundo que utilizan representaciones contextuales y arquitecturas avanzadas para identificar relaciones complejas en diversos dominios. A continuación, en la Tabla 3.3, se presenta una comparativa de los trabajos analizados en esta sección.

Tabla 3.3: Estado del arte identificación de relaciones

Autores	Dominio	Idioma	Enfoque	Tipo de relaciones	Evaluación
Zhang, X. et al. [45]	General	Inglés	BiLSTM con módulo de atención	Generales	P: 89.9 % R: 93.7 % F1: 91.8 %
Xiang, L. et al. [46]	General	Inglés	BERT	Generales	P: 91.5 % R: 91.8 % F1: 91.7 %
Yixuan, C. et al. [49]	Finanzas	Inglés	BiLSTM	Causa-efecto	P: 81.63 % R: 80.99 % F1: 81.31 %
Ridong, H. et al. [47]	General	Inglés	BERT	Generales	F1: 73.98 %
Libo, Y. et al. [50]	Noticias	Inglés	SVM PCA CNN	Generales	P: 98.3 % R: 89.2 %
Wu, K. et al. [51]	Geografía	Chino	Modelos de lenguaje (PURE y CasREL)	Espaciales geográficas	P: 88.3 % R: 87.3 % F1: 87.8 %
Ruchi, P. et al. [53]	Médico	Inglés	Red neuronal convolucional	Relaciones entre enfermedades, pruebas y tratamientos	P: 74.5 % R: 70.9 % F1: 72.6 %
Ying, H. et al. [54]	Médico	Inglés	BERT y modelo convolucional jerárquico	Proteína - Proteína Fármaco - Fármaco Químico - Proteína	P: 86 % R: 88.3 % F1: 87.2 %
Serrano, C. et al. [48]	General	Español	SVM	Generales	P: 80.1 % R: 70.8 % F1: 75.2 %
Revenko, A. et al. [52]	Legal	Español	GRIT Text2Event	Legales (derecho, deberes, entre otras)	F1: 80 %

Capítulo 4

Metodología de solución

Para alcanzar los objetivos de esta investigación, se empleó una metodología de solución que se muestra en la Figura 4.1. Esta metodología está compuesta por tres fases principales: Procesamiento de información, Extracción de información y Representación de información, las cuales se describen detalladamente a continuación.

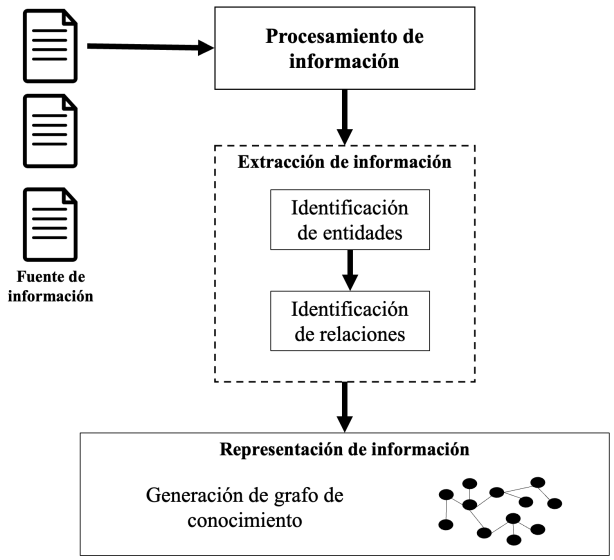

Figura 4.1: Diagrama de la metodología de solución para la generación automática de grafos de conocimiento en español en el dominio médico.

4.1. Fuente de información

La fuente de información utilizada es extraída de un desafío en la indexación semántica biomédica a gran escala y la respuesta a preguntas conocida como BioASQ [55].
Este dataset pertenece a la tarea MESINESP que tiene como objetivo principal promover el desarrollo de herramientas de indexación semántica relevantes en la práctica para contenido biomédico en un idioma diferente al inglés. Todos los documentos fueron etiquetados con descriptores DeCS [56] que es un vocabulario estructurado y controlado creado por BIREME para indexar publicaciones científicas en BvSalud [57].
Se decidió utilizar este conjunto de datos debido a que es uno de los pocos disponibles en español dentro del dominio médico. Además, su tamaño considerable permite realizar actividades de aprendizaje automático de manera más eficiente y facilita la segmentación del conjunto para diversas pruebas. Cabe destacar que este conjunto ha sido validado por expertos, lo que lo distingue de otros disponibles, lo que reforzó la decisión de emplearlo en esta investigación.
El dataset de esta tarea se encuentra disponible para su descarga y contiene 249,473 artículos científicos que alberga registros de las bases de datos LILACS [58], MEDLINE [59], IBECS [60], entre otras. El número de palabras total es de 45,322,119 y el tamaño del vocabulario es de 373,851.
Los artículos científicos se encuentran en formato JSON y en la Figura 4.2 se puede observar un ejemplo de su estructura.

{"id": "ibc-194909", "title": "La resiliencia importa: explicación de la asociación entre personalidad y funcionamiento psicológico durante la pandemia de COVID-19", "abstractText": "ANTECEDENTES/OBJETIVO: El objetivo fue dilucidar el mecanismo subyacente a través del cual las dimensiones básicas de la personalidad predicen indicadores del funcionamiento psicológico durante la pandemia de COVID-19, incluido el bienestar subjetivo y el estrés percibido. Como característica de la personalidad altamente contextualizada en circunstancias estresantes, se esperaba que la resiliencia tuviera un papel mediador en esta relación. MÉTODO: Una muestra de 2.722 adultos eslovenos (18-82 años), completó el Big Five Inventory, la Connor-Davidson Resilience Scale, la Perceived Stress Scale y el Mental Health Continuum. Se realizó un análisis de ruta con el procedimiento de estimación Bootstrap para evaluar el efecto mediador de la resiliencia en la relación entre la personalidad y el funcionamiento psicológico. RESULTADOS: La resiliencia medió total o parcialmente las relaciones entre los Cinco Grandes, y la extraversión con bienestar subjetivo y el estrés experimentado, al comienzo del estallido de COVID-19. El neuroticismo fue el predictor más fuerte de un funcionamiento psicológico menos adaptativo, tanto directamente como a través de la disminución de la capacidad de resiliencia. CONCLUSIONES: La resiliencia puede ser un factor de protección importante y requerido para una respuesta adaptativa de un individuo en situaciones estresantes como la pandemia y el confinamiento asociado", "journal": "Int. j. clin. health psychol. (Internet)", "year": 2021, "db": "IBECS", "decsCodes": ["D058873", "D006801", "D008875", "D010555", "D000369", "D010551", "D055500", "D008297", "D000075384", "D055815", "D005260", "D011024", "D000328", "D018352", "D011594", "D000293", "D000368"]},

Figura 4.2: Ejemplo de artículo científico de dataset BioASQ

Esta investigación hace uso del contenido de la etiqueta "abstractText", redactada en español, junto con la etiqueta "id", que facilita el control y seguimiento del resumen procesado en cada etapa.

4.2. Procesamiento de información

El procesamiento adecuado de los datos es fundamental para extraer información relevante y garantizar la precisión en los análisis. En el contexto de los artículos científicos, la capacidad de manejar grandes volúmenes de datos de manera eficiente es crucial para obtener resultados útiles y precisos. Un manejo inadecuado de los datos puede generar ineficiencias, tiempos de procesamiento prolongados y resultados erróneos. Por ello, el uso de técnicas de limpieza y depuración es esencial para optimizar el flujo de trabajo y asegurar la calidad del análisis posterior.

En esta fase, se incorporan los artículos científicos en lenguaje natural con el fin de llevar a cabo un proceso de depuración de la información, optimizando así su manipulación en las etapas subsiguientes.

Las actividades realizadas durante la limpieza de los resúmenes de los artículos científicos tienen como objetivo prepararlos para un análisis eficiente,

minimizando el consumo de recursos computacionales y mejorando los tiempos de procesamiento.

- **Tokenización**. Se realiza un proceso de segmentación del texto en tokens para cada artículo científico, considerando en esta investigación un *token* como la unidad mínima de una oración, es decir, una palabra. La tokenización permite reducir la complejidad del texto y facilita su procesamiento en etapas posteriores. La Figura 4.3 muestra un ejemplo de un fragmento de un artículo científico tokenizado a nivel de palabra.

[*El* , *objetivo* , *fue* , *dilucidar* , *el* , *mecanismo* , *subyacente* , *a* , *través* , *de* , *dimensiones* , *básicas*]

Figura 4.3: Fragmento de artículo científico tokenizado.

- **Eliminación de ruido**. En esta etapa se lleva a cabo la depuración de caracteres irrelevantes para el contexto de los datos, comúnmente denominados ruido. Para ello, se implementó un método que elimina signos de puntuación y caracteres especiales, tales como: punto, coma, punto y coma, guion bajo, barra diagonal, corchetes, paréntesis, dos puntos, así como comillas tanto españolas como inglesas ¿ ! " , / . ; : ∗ $ & [ ] ( ) «». La selección de estos símbolos se basó en un análisis previo de los artículos científicos utilizados, identificando aquellos caracteres que aparecían con mayor frecuencia en los contextos analizados y que no aportaban información significativa. Asimismo, se evitó eliminar símbolos que podrían ser relevantes en términos médicos específicos, con el fin de preservar la integridad semántica del contenido.

- **Expresiones regulares**. La aplicación de expresiones regulares permite identificar términos que combinan cifras con caracteres especiales, tales como punto, barra diagonal, guion, signo más y numeral (. − + #). También facilita la detección de combinaciones entre caracteres alfabéticos y símbolos como el signo más/menos, igual, mayor que, menor que, porcentaje y ampersand (± = > < % &). Para la definición de estos patrones, se llevó a cabo un análisis detallado del corpus de artículos científicos con el objetivo de evitar la eliminación de expresiones relevantes que pudieran representar entidades médicas. Dado que muchos

términos del ámbito biomédico incorporan números o símbolos específicos como parte de su sintaxis, se establecieron criterios que permitieran conservar aquellas estructuras necesarias para no afectar la integridad de la información ni dificultar el reconocimiento posterior de entidades médicas.

- **Normalización**. En esta etapa del procesamiento se transforma todo el contenido textual del artículo científico a letras minúsculas, con el propósito de evitar la duplicidad en la identificación de términos. Por ejemplo, palabras como "COVID" y "covid" podrían ser interpretadas como entidades distintas, a pesar de referirse a la misma enfermedad. Adicionalmente, se eliminan los espacios en blanco redundantes entre palabras, así como aquellos ubicados al inicio o al final del texto, contribuyendo así a una representación más limpia y uniforme de los datos.

En la Figura 4.4 se presenta el texto original de un resumen de artículo científico antes de ser sometido al procesamiento, mientras que la Figura 4.5 muestra el mismo texto tras la aplicación de las etapas de limpieza y normalización.

```
ANTECEDENTES/OBJETIVO: El objetivo fue dilucidar el mecanismo
subyacente a través del cual las dimensiones básicas de la
personalidad predicen indicadores del funcionamiento psicológico
durante la pandemia de COVID-19, incluido el bienestar subjetivo y el
estrés percibido. Como característica de la personalidad altamente
contextualizada en circunstancias estresantes, se esperaba que la
resiliencia tuviera un papel mediador en esta relación. MÉTODO: Una
muestra de 2.722 adultos eslovenos (18-82 años), completó el Big Five
Inventory, la Connor-Davidson Resilience Scale, la Perceived Stress
Scale y el Mental Health Continuum. Se realizó un análisis de ruta con
el procedimiento de estimación Bootstrap para evaluar el efecto
mediador de la resiliencia en la relación entre la personalidad y el
funcionamiento psicológico. RESULTADOS: La resiliencia medió total o
parcialmente las relaciones entre los Cinco Grandes, y la extraversión
con bienestar subjetivo y el estrés experimentado, al comienzo del
estallido de COVID-19. El neuroticismo fue el predictor más fuerte de
un funcionamiento psicológico menos adaptativo, tanto directamente
como a través de la disminución de la capacidad de resiliencia.
CONCLUSIONES: La resiliencia puede ser un factor de protección
importante y requerido para una respuesta adaptativa de un individuo
en situaciones estresantes como la pandemia y el confinamiento asociado
```

Figura 4.4: Resumen de artículo científico antes de ser procesado

antecedentes objetivo el objetivo fue dilucidar el mecanismo subyacente a través del cual las dimensiones básicas de la personalidad predicen indicadores del funcionamiento psicológico durante la pandemia de covid-19 incluido el bienestar subjetivo y el estrés percibido como característica de la personalidad altamente contextualizada en circunstancias estresantes se esperaba que la resiliencia tuviera un papel mediador en esta relación método una muestra de adultos eslovenos años completó el big five inventory la connor-davidson resilience scale la perceived stress scale y el mental health continuum se realizó un análisis de ruta con el procedimiento de estimación bootstrap para evaluar el efecto mediador de la resiliencia en la relación entre la personalidad y el funcionamiento psicológico resultados la resiliencia medió total o parcialmente las relaciones entre los cinco grandes y la extraversión con bienestar subjetivo y el estrés experimentado al comienzo del estallido de covid-19 el neuroticismo fue el predictor más fuerte de un funcionamiento psicológico menos adaptativo tanto directamente como a través de la disminución de la capacidad de resiliencia conclusiones la resiliencia puede ser un factor de protección importante y requerido para una respuesta adaptativa de un individuo en situaciones estresantes como la pandemia y el confinamiento asociado

Figura 4.5: Resumen de artículo científico después de ser procesado

4.3. Extracción de información

La fase de extracción de información está conformada por dos módulos fundamentales para la construcción del grafo de conocimiento: la identificación de entidades, que conforman los nodos, y la identificación de relaciones, que representan los enlaces entre dichas entidades. A continuación, se describe en detalle el funcionamiento de cada uno de estos módulos.

4.3.1. Identificación de entidades

El primer módulo de la fase de extracción de información corresponde a la identificación de entidades, cuyo objetivo es localizar los conceptos médicos presentes en el texto. Esta sección se encarga de realizar la detección de entidades dentro del artículo científico de entrada, el cual ha sido previamente procesado. Para ello, se emplea la metodología ilustrada en la Figura 4.6, la cual se describe a continuación.

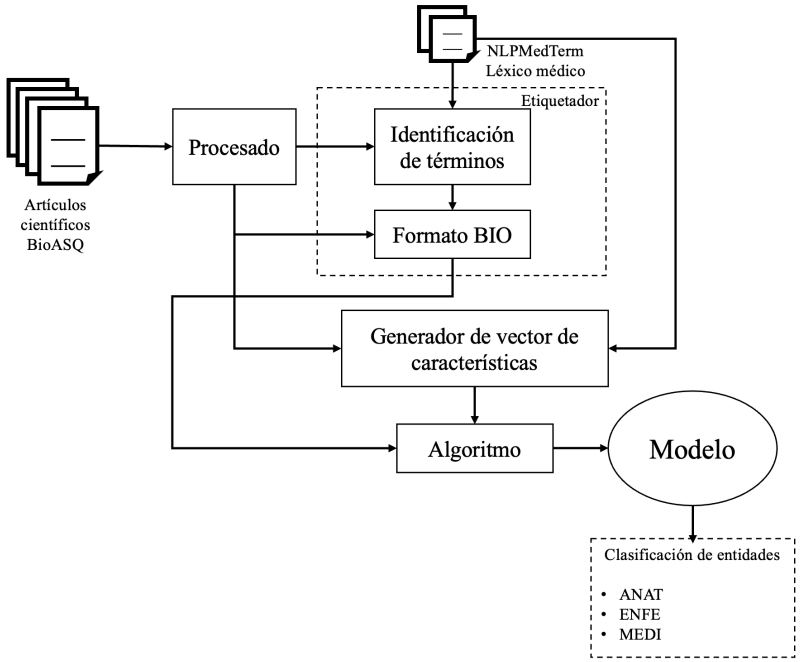

Figura 4.6: Método de identificación de entidades médicas.

Léxico médico NLPMedTerm

El diseño e implementación de este etiquetador se realizó utilizando el recurso NLPMedTerm [61]. El primer entregable del proyecto consiste en un léxico médico en español, desarrollado por expertos en la materia, y generado como parte del proyecto NLPMedTerm (Procesamiento del Lenguaje Natural para Terminología Médica) de la Universidad Autónoma de Madrid.
Este etiquetador está diseñado para identificar tres tipos principales de entidades: enfermedad, medicamento y anatomía. La descripción de estas categorías, basada en lo establecido por NLPMedTerm, se presenta en la Tabla 4.1.

Tabla 4.1: Entidades identificadas

Entidad	Descripción
ENFE	Patologías, enfermedades, síndromes o anormalidades lesiones, virus o bacterias
MEDI	Sustancias farmacológicas, antibióticos o sustancias clínicas
ANAT	Parte del cuerpo, órgano o sistema corporal

El primer entregable también contiene 127 grupos semánticos médicos. Para este desarrollo, se seleccionaron 11 de ellos por cumplir con los criterios necesarios para identificar entidades de tipo enfermedad, medicamento y anatomía. Los grupos semánticos utilizados se detallan en la Tabla 4.2.

Tabla 4.2: Grupos semánticos de entidades identificadas

Entidad	Grupo semántico	Ejemplo
Anatomía ANAT	Región corporal	Epigastrio
	Parte del cuerpo u órgano	Músculo
	Sistema corporal	Sistema urinario
Medicamento MEDI	Sustancia farmacológica	Opioide
	Antibiótico	Penicilina
	Fármaco clínico	Tacrólimus
Enfermedad ENFE	Enfermedad o síndrome	Diabetes
	Herida o intoxicación	Traumatismo
	Anormalidad anatómica	Arruga

Etiquetador de términos médicos

El etiquetador de términos genera un documento en formato XML que organiza los términos identificados en cada documento procesado, proporcionando la siguiente información:

- **Documento:** ID del documento, extraído del conjunto de artículos científicos de entrada.

- **Nombre:** Token del término identificado.
- **Inicio del término:** Posición inicial del término en el texto procesado.
- **Final del término:** Posición final del término en el texto procesado.
- **Clase:** Tipo de entidad (ANAT: Anatomía, ENFE: Enfermedad, MEDI: Medicamento) de acuerdo a su grupo semántico.
- **Multipalabra:** Valor booleano que indica si el término se compone de varios tokens.

La Figura 4.7 ilustra un fragmento del documento XML resultante, que contiene todos los términos identificados en los documentos de entrada.

Figura 4.7: XML con términos identificados

El documento XML generado se utiliza para realizar la anotación BIO de los textos, una técnica esencial para preparar datos de entrenamiento en tareas de aprendizaje automático. Esta anotación se lleva a cabo a nivel de token, proporcionando a los algoritmos información explícita sobre el inicio y el final de las entidades nombradas.
Por ejemplo, en el fragmento de texto "potente inhibidor de la proteasa principal del SARS-CoV-2", se identifican dos entidades relevantes:

- Un medicamento multipalabra, "inhibidor de la proteasa", clasificado como MEDI.

- Una enfermedad, “SARS-CoV-2”, clasificada como ENFE.

La anotación BIO (*Begin, Inside, Outside*) permite etiquetar de manera precisa cada token, indicando si marca el comienzo (B-*term*) o la continuación (I-*term*) de una entidad, o si no pertenece a ninguna entidad (O). Esta estrategia es útil para que los algoritmos de aprendizaje automático puedan reconocer entidades compuestas por varias palabras, tratándolas como un solo concepto en lugar de interpretar cada palabra de forma aislada.
Contar con un conjunto de datos etiquetado de esta forma resulta fundamental para el entrenamiento, ya que proporciona ejemplos claros y estructurados que ayudan al algoritmo durante el aprendizaje.
La anotación BIO correspondiente a este fragmento de texto se detalla en la Tabla 4.3.

Tabla 4.3: Ejemplo de anotado BIO

Token	Etiqueta
potente	O
inhibidor	B-MEDI
de	I-MEDI
la	I-MEDI
proteasa	I-MEDI
principal	O
del	O
sarscov-2	B-ENFE

En el ejemplo se observa que, al inicio de un término, se emplea el prefijo *“B-”* en la etiqueta para indicar el inicio de una entidad clasificada. Los tokens subsiguientes que forman parte de un término multipalabra se etiquetan con el prefijo *“I-”*, señalando que continúan la misma entidad. Por su parte, la etiqueta *“O”* se utiliza para indicar que un token no pertenece a ninguna clase o entidad reconocida.

Generador de vector de características

Los textos se segmentan a nivel de palabra, ya que esta granularidad permite capturar de manera precisa las unidades mínimas de significado que

forman las entidades de interés. Trabajar a nivel de palabra facilita la identificación de términos relevantes, especialmente en tareas donde las entidades pueden estar formadas por una o varias palabras consecutivas. Además, segmentar a nivel de token es una práctica estándar en el procesamiento de lenguaje natural, ya que permite asociar características específicas a cada unidad léxica.
Para generar automáticamente los valores numéricos que conforman el vector de características de cada token, se emplearon tanto el Entregable 1 de [61] como la herramienta de procesamiento de lenguaje natural SpaCy[1]. Cada vector está compuesto por ocho valores numéricos que representan características semánticas, morfológicas y sintácticas. Esta combinación permite incorporar información contextual relevante sobre las palabras, con el objetivo de mejorar la identificación de términos en el texto. Así, a cada token se le asigna un vector numérico que incluye las siguientes características:

1. **Término multipalabra con un grupo semántico médico.** La primera posición del vector indica si el token pertenece a un grupo semántico médico.

   - Se asigna un 1 si el token pertenece a un grupo semántico médico y es el inicio del término.
   - Se asigna un 2 si el token pertenece a un grupo semántico médico y forma parte de un término multipalabra sin ser el inicio.
   - Se asigna un 0 si el token no pertenece a ningún grupo semántico.

   Los grupos semánticos empleados se detallan en la Tabla 4.2.

2. **Tipo de grupo semántico médico.** Valor que indica el grupo semántico del token actual; de lo contrario, se asigna un cero. Los posibles valores de esta característica se presentan en la tabla 4.4.

3. **Tipo de dependencia.** Indica el tipo de relación de dependencia del token, basado en la tabla hash determinada por SpaCy.

4. **Tipo de dependencia del nodo padre sintáctico.** Representa el entero que identifica la relación de dependencia con el nodo padre sintáctico del token actual.

[1] https://spacy.io/

5. **Etiqueta POS.** Este entero indica la categoría gramatical general del token, según las etiquetas POS proporcionadas por SpaCy.

6. **Etiqueta POS del nodo padre sintáctico.** Representa la categoría gramatical del nodo padre sintáctico, conforme al conjunto de etiquetas POS de SpaCy.

7. **Nodos hijos a la izquierda.** Indica si existe un nodo hijo a la izquierda del token dentro de una ventana de dos tokens. Se asigna un 1 si existe, y un 0 en caso contrario.

8. **Nodos hijos a la derecha.** Indica si existe un nodo hijo a la derecha del token dentro de una ventana de dos tokens. Se asigna un 1 si existe, y un 0 en caso contrario.

Tabla 4.4: Asignación de valores numéricos asociados al grupo semántico

Valor de grupo semántico	Grupo semántico
1	Región corporal
2	Parte del cuerpo u órgano
3	Sistema corporal
4	Sustancia farmacológica
5	Antibiótico
6	Fármaco clínico
7	Enfermedad o síndrome
8	Herida o intoxicación
9	Anormalidad anatómica
10	Virus
11	Bacterias

Algoritmos de aprendizaje automático

La combinación de características para la generación de vectores de características se aplicó en algoritmos de clasificación basados en árboles de decisión y bosques aleatorios (Random Forest) para el reconocimiento de entidades médicas en textos en español. Se emplearon técnicas clásicas de clasificación de entidades para evaluar el rendimiento de dicha combinación de características.

Se generaron 1,182,663 vectores de características a partir de 13,067 resúmenes científicos extraídos del BioASQ Challenge. Los resúmenes fueron anotados utilizando los términos proporcionados por el recurso NLPMedTerm, validado por expertos en el ámbito médico, lo que permite una consistencia de las etiquetas aplicadas.

Uno de los principales desafíos durante el entrenamiento y validación de los algoritmos fue el desbalance de las clases en el conjunto de datos, un problema habitual pero importante en la tarea de reconocimiento de entidades. Aunque los resúmenes científicos contienen terminología especializada, la proporción de términos médicos respecto al total de palabras es considerablemente baja, esto se debe a la presencia de numerosas palabras de propósito general, como conectores, verbos y expresiones auxiliares necesarias para estructurar el contenido y transmitir la información de manera completa.

La Figura 4.8 muestra claramente este desbalance, la clase *"O"* representa palabras generales que no forman parte de las entidades médicas clasificadas en ANAT, ENFE o MEDI. Además, se evidencia un desequilibrio entre las etiquetas B-*term* e I-*term*, atribuible a que la mayoría de las entidades médicas en el conjunto de datos corresponden a términos de una sola palabra, lo que plantea un reto adicional para la clasificación. A pesar de este desequilibrio, el conjunto de datos fue dividido utilizando una validación cruzada, con un 80 % destinado al entrenamiento y un 20 % a la validación.

La elección de algoritmos basados en árboles de decisión y random forest se debe a su capacidad para manejar tanto variables categóricas como numéricas, facilitando la interpretación de los resultados. Además, random forest al construir múltiples árboles y promediar sus decisiones, se vuelven efectivos en escenarios con datos desbalanceados.

Los resultados obtenidos fueron prometedores, el algoritmo de árbol de decisión alcanzó una precisión del 97.84 %, mientras que el algoritmo de random forest logró superarlo ligeramente, alcanzando una precisión del 97.86 %.

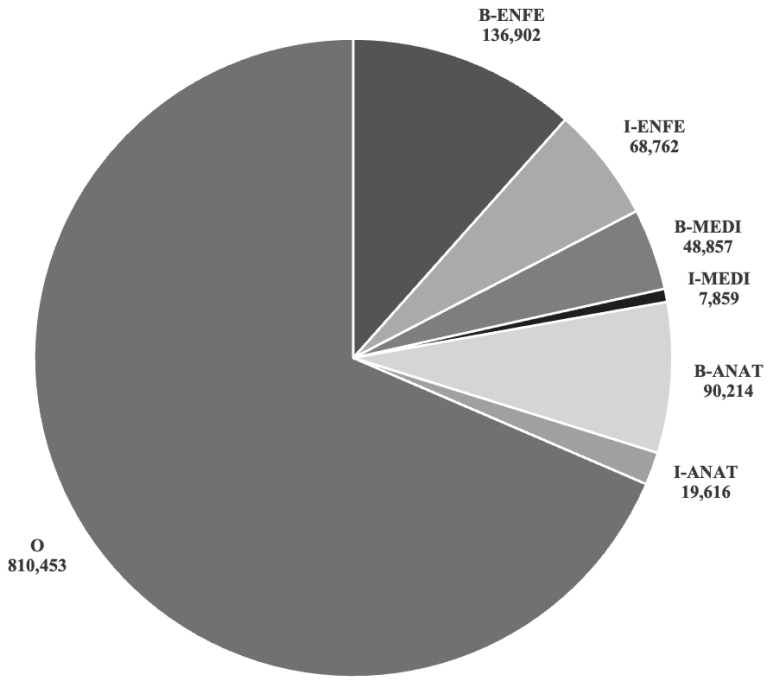

Figura 4.8: Distribución de clases en conjunto de datos BioASQ Challenge.

4.3.2. Identificación de relaciones

Las entidades identificadas en la sección anterior funcionan como nodos dentro de un grafo de conocimiento. La siguiente etapa de la extracción de información consiste en determinar las relaciones existentes entre dichas entidades. Para ello, se desarrolló la metodología representada en la Figura 4.9, cuyas etapas se describen a continuación.

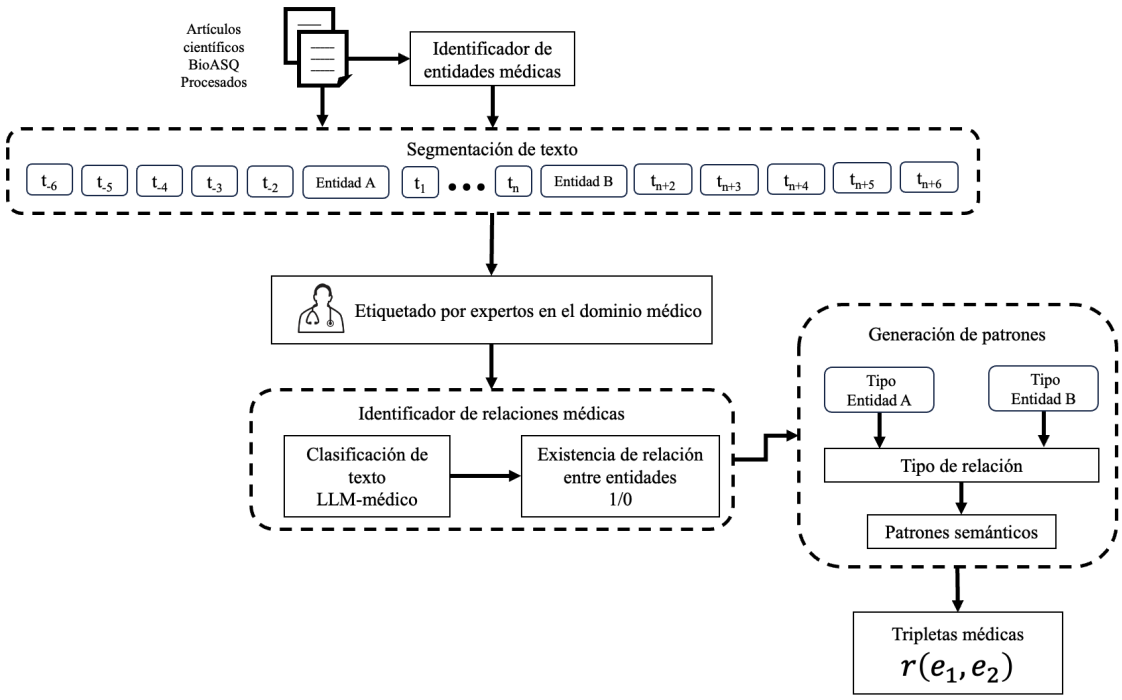

Figura 4.9: Método de identificación de relaciones médicas.

Segmentación de texto

Para esta fase se utilizaron los mismos textos procesados descritos en la Sección 4.2, junto con su versión etiquetada en formato BIO, generada por el identificador de entidades médicas desarrollado en la Sección 4.3. A partir de esta información, se implementaron representaciones sencillas y combinadas de entidades para el enmascaramiento de términos.

Dado que el formato BIO permite la identificación de entidades de múltiples palabras a través de la combinación de etiquetas B-*term* e I-*term*, se unifican dichas entidades en una única cadena, encerrándolas entre el símbolo *"@"* y la etiqueta B-*term*, por ejemplo: un termino conformado por *"B-MEDI, I-MEDI, I-MEDI"* su representación se convierte en *"@B-MEDI@"*. Esta transformación tiene como objetivo facilitar el enmascaramiento en fases posteriores y evitar conflictos en las tareas de clasificación de texto.

Se combinan todos los términos identificados en cada texto con el objetivo de calcular la distancia entre ellos. Esta estrategia se fundamenta en la Teoría

de la Distribución del Significado [62], la cual sostiene que las palabras que aparecen en contextos similares tienden a compartir significados. Bajo esta premisa, se asume que una menor distancia entre dos términos refleja una relación semántica más estrecha. Esta idea es consistente con enfoques como Word2Vec, donde las palabras ubicadas dentro de una ventana de contexto reducida tienden a tener significados relacionados [63], o como en los modelos basados en Transformer, donde la atención se enfoca inicialmente en tokens cercanos, capturando primero dependencias locales antes de atender relaciones más lejanas [1].
Se realiza un análisis del número de palabras entre pares de términos identificados, determinando que, en promedio, el contexto no se pierde hasta una distancia de 20 términos. Asimismo, se observó que no existían términos diferentes con una distancia menor a 10 sin relación aparente. Con base en estos hallazgos y en lo reportado en la literatura, se estableció una ventana contextual de 5 palabras a la izquierda y a la derecha del término.
Se generaron combinaciones entre todas las entidades presentes en un mismo texto, con el fin de identificar posibles relaciones médicas entre cada par. Esto permitió construir frases específicas para cada combinación de entidades detectadas, facilitando su análisis posterior.

Etiquetado de conjunto de textos

Se obtuvieron un total de 1,912 frases que contenían dos entidades médicas enmascaradas, con el propósito de determinar si existía una relación médica y contextual entre ellas. Estas frases fueron extraídas de un corpus de 600 textos previamente procesados con sus respectivas entidades identificadas. Solo se consideraron aquellas frases que cumplían con los criterios definidos en la sección anterior.
Dado que no existen conjuntos de datos previamente etiquetados que incluyan los tipos de entidades contemplados en esta investigación (ANAT, MEDI y ENFE) y que puedan ser utilizados para tareas de identificación de relaciones, se procedió a realizar un proceso de etiquetado manual por parte de tres expertos en el dominio médico. El resultado fue un conjunto de datos de entrenamiento validado por expertos, el cual está disponible en [64].
Cada frase fue etiquetada con un valor de 1 cuando las entidades involucradas presentaban una relación médica y contextual, y con un 0 en caso contrario. Para facilitar este proceso a los expertos, se desarrolló una aplicación denominada MedRel, diseñada para asistir en el etiquetado de relaciones médicas

en conjuntos de textos. Esta herramienta, mostrada en la Figura 4.10, está compuesta por cuatro secciones:

1. La frase etiquetada con la representación uniforme.
2. La oración original (sin etiquetar).
3. Las dos entidades involucradas.
4. Dos botones que permiten indicar si existe o no una relación médica y contextual entre las entidades.

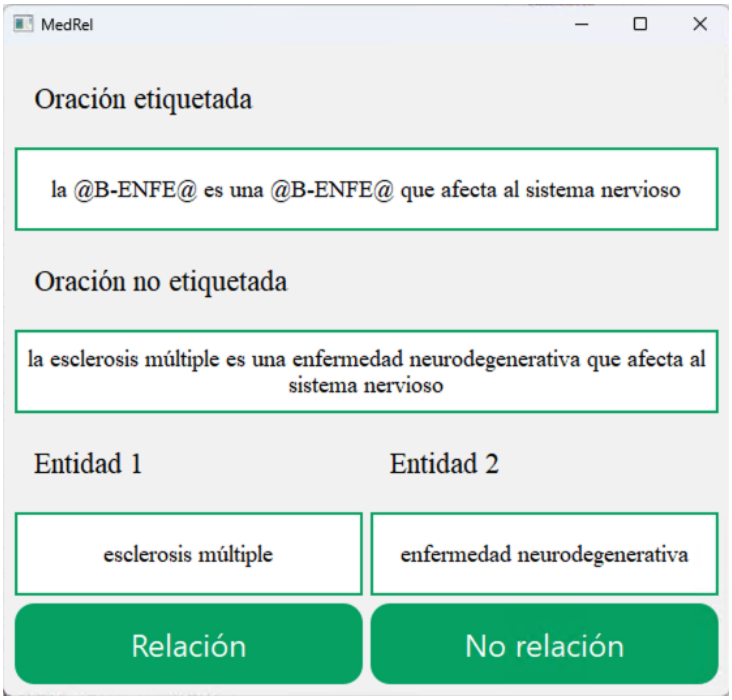

Figura 4.10: Aplicación MedRel para el etiquetado de conjunto de datos.

Como resultado del proceso de etiquetado, los expertos identificaron 1,106 frases que sí contenían una relación y 806 frases que no la contenían.

Identificador de relaciones médicas

Para clasificar la existencia o no de una relación médica entre dos entidades dentro de una oración, se realizó un ajuste fino (fine-tuning) de un modelo preentrenado en el dominio médico, denominado MedicoBERT, desarrollado en [65]. Este proceso de adaptación tiene como objetivo especializar el modelo en una tarea concreta de PLN, en este caso, la clasificación binaria de relaciones médicas y contextuales.
El ajuste fino consiste en proporcionar al modelo preentrenado un conjunto de datos más pequeño y etiquetado específicamente para la tarea objetivo. Durante este proceso, los pesos de las conexiones neuronales del modelo son ajustados para adaptarse mejor a las características y patrones del nuevo conjunto de datos.
Como punto de partida, se consideraron diversas configuraciones de entrenamiento propuestas en el estado del arte en tareas similares de clasificación de relaciones, lo que permitió definir un espacio de búsqueda inicial más informado. A partir de estas configuraciones, se empleó un enfoque exploratorio para identificar la combinación de hiperparámetros que ofreciera el mejor desempeño. Esta búsqueda se realizó mediante la exploración aleatoria del espacio de hiperparámetros, evaluando el rendimiento del modelo en cada configuración probada.
Se realizaron experimentos con la distribución de 80 % para el entrenamiento y 20 % para su evaluación que incluye el entrenamiendo de un LLM base, uno especializado y una calibración fina de hiperparámetros, cuyos resultados se presentan en la sección de Evaluación y análisis de resultados. A partir de estos, se identificó la siguiente combinación de hiperparámetros como la más adecuada para esta tarea obteniendo un 90 % de precision:

- Tasa de aprendizaje: $2{,}392911e^{-05}$
- Número de épocas: 20
- Tamaño del lote: 16
- Peso: 0.01
- Optimizador: Adam

La Tabla 4.5 muestra la distribución de los tipos de relaciones utilizadas en cada experimento entre las entidades ANAT (Anatomía), ENFE (Enfermedad) y MEDI (Medicamento), así como los valores asignados a las frases,

indicando la existencia (1) o ausencia (0) de una relación. Estas relaciones corresponden a aquellas validadas previamente por expertos médicos.

Tabla 4.5: Distribución de tipos de relaciones utilizados en el entrenamiento

Tipo de entidad origen	Tipo de entidad destino	Etiqueta 0	Etiqueta 1	Total
ANAT	ANAT	201	91	292
ANAT	ENFE	111	123	234
ANAT	MEDI	79	109	188
ENFE	ANAT	109	119	228
ENFE	ENFE	88	147	235
ENFE	MEDI	74	115	189
MEDI	ANAT	56	120	176
MEDI	ENFE	59	130	189
MEDI	MEDI	29	152	181

Generación de patrones semánticos

A partir de la identificación de una relación médica existente entre dos entidades, se utilizan únicamente aquellas que presentan una relación positiva para analizar el patrón subyacente entre ellas. Según el tipo de cada entidad, se determina el patrón semántico que mejor representa su relación, con el objetivo de generar una tripleta compuesta por dos entidades médicas y una relación entre ellas. Se definieron siete patrones posibles: seis específicos del dominio médico y uno generalizado. A continuación, se presentan en su notación formal.
Los seis patrones específicos son:

- $r_1 = \{(m, a) | m \in MEDI, a \in ANAT, MEDI\ trata\_condiciones\ ANAT\}$
- $r_2 = \{(m, e) | m \in MEDI, e \in ENFE, MEDI\ es\_tratamiento\ ENFE\}$
- $r_3 = \{(a, m) | a \in ANAT, m \in MEDI, ANAT\ se\_trata\_con\ MEDI\}$
- $r_4 = \{(a, e) | a \in ANAT, e \in ENFE, ANAT\ afectado\_por\ ENFE\}$
- $r_5 = \{(e, a) | e \in ENFE, a \in ANAT, ENFE\ tiene\_efecto\_en\ ANAT\}$
- $r_6 = \{(e, m) | e \in ENFE, m \in MEDI, ENFE\ usa\_tratamiento\ MEDI\}$

El patrón general corresponde a una relación del tipo *"es un"*, útil para los casos en que ambas entidades pertenecen al mismo tipo. Su notación es la siguiente:

- $r_7 = \{(e_1, e_2) | e_1 \in T, e_2 \in T, tipo(e_1) = t, tipo(e_2) = t, e_1\ es\_un\ t, e_2\ es\_un\ t\}$

Donde:
$T = ANAT \cup ENFE \cup MEDI$
$t \in \{ANAT, ENFE, MEDI\}$

La identificación de uno de estos patrones semánticos permite construir una tripleta informativa, considerando además el sentido contextual de las entidades para lograr una comprensión más precisa de la relación entre ellas. Por ejemplo, si el sistema de identificación de relaciones médicas determina que existe una relación entre las siguientes dos entidades:

```
Entidad A: COVID-19
Tipo de Entidad A: ENFE
Entidad B: pulmones
Tipo de Entidad B: ANAT
```

El patrón identificado sería:

$$r_5 = \{(e, a) | e \in ENFE, a \in ANAT, ENFE\ tiene\_efecto\_en\ ANAT\}$$

Por lo tanto, la tripleta generada es:

$$tiene\_efecto\_en(COVID - 19, pulmones)$$

Dado que se realiza un proceso de enriquecimiento con cada nuevo texto, al generar nuevas tripletas se implementa una verificación para evitar duplicados con respecto a las ya existentes. Además, si se identifica uno de los seis patrones específicos, se comprueba si las entidades involucradas ya forman parte de tripletas con el patrón generalizado. Este procedimiento busca asegurar un enriquecimiento significativo del conjunto de tripletas.

4.4. Representación de información

La representación de la información en el contexto de esta investigación se basa en la visualización de tripletas generadas por el módulo de identificación de relaciones, las cuales son representadas en un grafo de conocimiento. En dicho grafo, los nodos corresponden a las entidades médicas identificadas en el texto, mientras que las relaciones entre dichas entidades conforman las aristas que conectan los nodos.
Por ejemplo, una tripleta que involucra la entidad COVID-19 (de tipo *ENFE*) y la entidad pulmones (de tipo *ANAT*), unidas mediante la relación *tiene_efecto_en*, se representa como:

$$tiene\_efecto\_en(COVID-19, pulmones)$$

Esta tripleta puede visualizarse gráficamente como se muestra en la Figura 4.11.

Figura 4.11: Ejemplo de representación de una tripleta.

Los textos utilizados se encuentran en formato plano, sin ningún tipo de preprocesamiento previo. El procesamiento automático se llevó a cabo mediante una serie de etapas, detalladas en la Figura 4.1, que incluyen:

1. Preprocesamiento de texto
2. Identificación de entidades médicas
3. Identificación de relaciones semánticas

La etapa final corresponde a la representación de la información en el grafo de conocimiento, proceso que se detalla en la Figura 4.12.

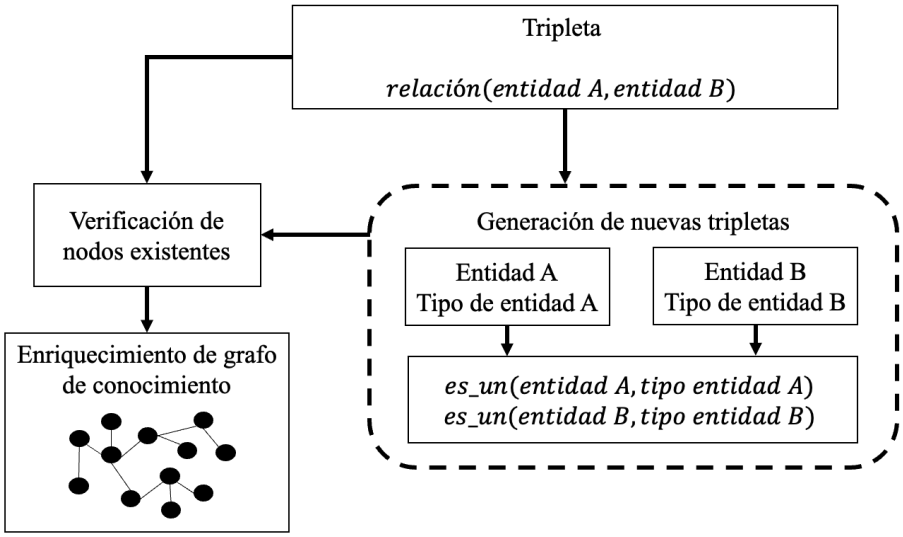

Figura 4.12: Método de representación de información.

Para ello, cada tripleta es sometida a un procedimiento dividido en dos componentes esenciales, definidos a continuación, que garantizan la coherencia estructural del grafo.

Generación de nueva tripletas

Se realiza una validación de las entidades involucradas en cada tripleta. Además de representar explícitamente la relación contenida en la tripleta, se generan nuevas relaciones del tipo es_un, que vinculan cada entidad con su nodo general correspondiente (Enfermedad, Anatomía o Medicamento), según el tipo de entidad identificado.
Este enriquecimiento se lleva a cabo bajo el supuesto de que si una entidad ha sido correctamente clasificada, entonces debe existir una relación que la vincule a su categoría general. Por ejemplo, a partir de la tripleta:

$$tiene\_efecto\_en(COVID-19, pulmones)$$

Se generan automáticamente las siguientes relaciones:

- $es\_un(COVID-19, Enfermedad)$

- $es\_un(pulmones, Anatoma)$

En caso de que la relación contenida en la tripleta analizada sea, desde un inicio, del tipo es_un, el procedimiento omite la etapa de generación de nuevas relaciones y pasa directamente a la verificación de nodos existentes.

Verificación de nodos existentes

Para evitar redundancia, se realiza una verificación previa de existencia de los nodos correspondientes a las entidades A y B de cada tripleta. Esto garantiza que no se creen múltiples instancias de una misma entidad y que el grafo permanezca interconectado en torno a las tres categorías principales: Medicamento, Enfermedad y Anatomía.
El grafo de conocimiento final fue construido a partir de 990 abstract de artículos científicos pertenecientes al conjunto de datos de BioASQ [55]. Es importante señalar que dichos artículos no fueron utilizados en ninguna de las fases de entrenamiento ni validación de los módulos anteriores.
El grafo resultante está compuesto por un total de 4355 nodos, clasificados por tipo tal como se resume en la Tabla 4.6.

Tabla 4.6: Clasificación de nodos representados en el grafo de conocimiento final.

Tipo de nodo	Cantidad
ENFE	2,217
MEDI	969
ANAT	1,169

En cuanto a las relaciones, se identificaron 12,294 relaciones, cuya distribución por categoría se presenta en la Tabla 4.7.

Tabla 4.7: Clasificación de relaciones representadas en el grafo de conocimiento final.

Tipo de entidad origen	Tipo de relación	Tipo de entidad destino	Cantidad
MEDI	ES_TRATAMIENTO	ENFE	1,485
MEDI	ES_UN	MEDI	968
ENFE	ES_UN	ENFE	2,216
ENFE	TIENE_EFECTO_EN	ANAT	1,795
ANAT	ES_UN	ANAT	1,168
ANAT	AFECTADO_POR	ENFE	1,597
ANAT	SE_TRATA_CON	MEDI	756
ENFE	USA_TRATAMIENTO	MEDI	1,669
MEDI	TRATA_CONDICIONES	ANAT	640

La visualización del grafo final se realizó utilizando la herramienta Neo4j[1], tal como se muestra en la Figura 4.13. En dicha representación, las entidades de tipo Anatomía (ANAT) se distinguen en color lila, las entidades correspondientes a Medicamento (MEDI) en azul y las de tipo Enfermedad (ENFE) en color naranja.

[1]https://neo4j.com/

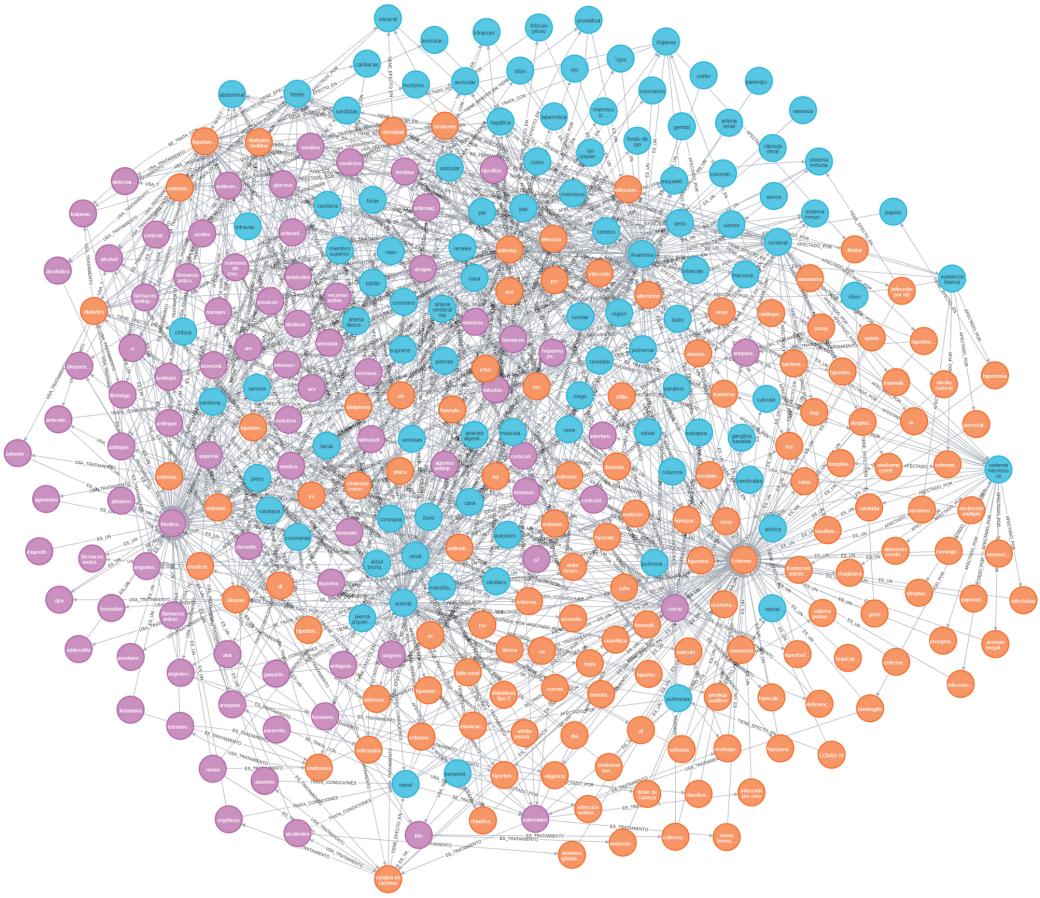

Figura 4.13: Grafo de conocimiento final generado con la metodología propuesta.

Capítulo 5

Evaluación y discusión de resultados

En este capítulo se presenta un análisis detallado de los resultados obtenidos en cada etapa de la metodología de esta investigación.

5.1. Identificación de entidades

Como parte de la evaluación de la etapa de identificación de entidades, se analizó el rendimiento de la combinación de características utilizando modelos de clasificación basados en los algoritmos de Árbol de Decisión y Random Forest. Para asegurar la robustez y capacidad de generalización de los resultados, se implementó una validación cruzada con cinco experimentos.
A pesar de la aleatoriedad en la partición de los datos, la métrica de precisión se mantuvo estable a lo largo de los diferentes experimentos, lo que evidencia la consistencia del enfoque desarrollado. Se eligió emplear exclusivamente la precisión como métrica de evaluación debido al marcado desbalance de clases presente en el conjunto de datos y mostrado anteriormente en la Figura 4.8. En este contexto, la precisión resulta una medida más adecuada para evaluar la capacidad del modelo de identificar correctamente los casos positivos, algo especialmente relevante tratándose de entidades médicas como se muestra en la ecuación 5.1.

$$Precisión = \frac{VP}{VP + FP} \tag{5.1}$$

La Tabla 5.1 presenta la distribución de precisión por clase obtenida por el modelo de Árbol de Decisión a lo largo de los cinco experimentos realizados con el conjunto de datos BioASQ. De manera similar, la Tabla 5.2 muestra los resultados del modelo de Random Forest bajo las mismas condiciones experimentales. En ambas tablas también se incluye la distribución de clases en cada partición del conjunto de datos durante la evaluación que corresponde al 20 %.

Tabla 5.1: Precisión obtenida en los experimentos utilizando modelo de árbol de decisión.

Clase	Precisión en Experimento				
	1	2	3	4	5
B-ENFE	100 %	100 %	100 %	100 %	100 %
I-ENFE	96.5 %	97 %	96.8 %	96.7 %	96.8 %
B-MEDI	100 %	100 %	100 %	100 %	100 %
I-MEDI	18.1 %	17.2 %	18.9 %	16.8 %	18.7 %
B-ANAT	100 %	100 %	100 %	100 %	100 %
I-ANAT	7.7 %	7.8 %	7.3 %	8 %	8 %
O	100 %	100 %	100 %	100 %	100 %
Total	**97.69 %**	**97.76 %**	**97.6 %**	**97.73 %**	**97.73 %**

Tabla 5.2: Precisión obtenida en los experimentos utilizando modelo de random forest.

Clase	Precisión en Experimento				
	1	2	3	4	5
B-ENFE	100 %	100 %	100 %	100 %	100 %
I-ENFE	96.8 %	97.3 %	97.2 %	97 %	97.4 %
B-MEDI	100 %	100 %	100 %	100 %	100 %
I-MEDI	19 %	17.5 %	19.7 %	17.6 %	19.5 %
B-ANAT	100 %	100 %	100 %	100 %	100 %
I-ANAT	7.2 %	7 %	7 %	7.1 %	7.4 %
O	100 %	100 %	100 %	100 %	100 %
Total	**97.71 %**	**97.77 %**	**97.69 %**	**97.76 %**	**97.76 %**

Para complementar estos resultados, las Figuras 5.1 y 5.2 muestran las matrices de confusión correspondientes a cada uno de los cinco experimen-

tos. Estas matrices reflejan una clara tendencia de las predicciones hacia la diagonal principal, lo cual indica que los modelos clasifican correctamente la mayoría de los casos. Sin embargo, la aparente ausencia de errores fuera de la diagonal puede sugerir un modelo preciso, posiblemente influenciado por el desbalance de clases en los datos de entrenamiento.
Este desbalance se puede observar en algunas categorías, como "I-MEDI", que presentan valores considerablemente más altos, en comparación con otras clases con menos representación. Este desequilibrio plantea un reto significativo para el modelo, que podría beneficiarse de una mayor representación de las clases minoritarias. Una posible solución consistiría en recolectar más datos pertenecientes a estas clases poco representadas, con el fin de mejorar el aprendizaje del modelo en dichas categorías. No obstante, este proceso requiere tiempo adicional y la colaboración de expertos en el dominio para garantizar la calidad y relevancia de los nuevos datos.

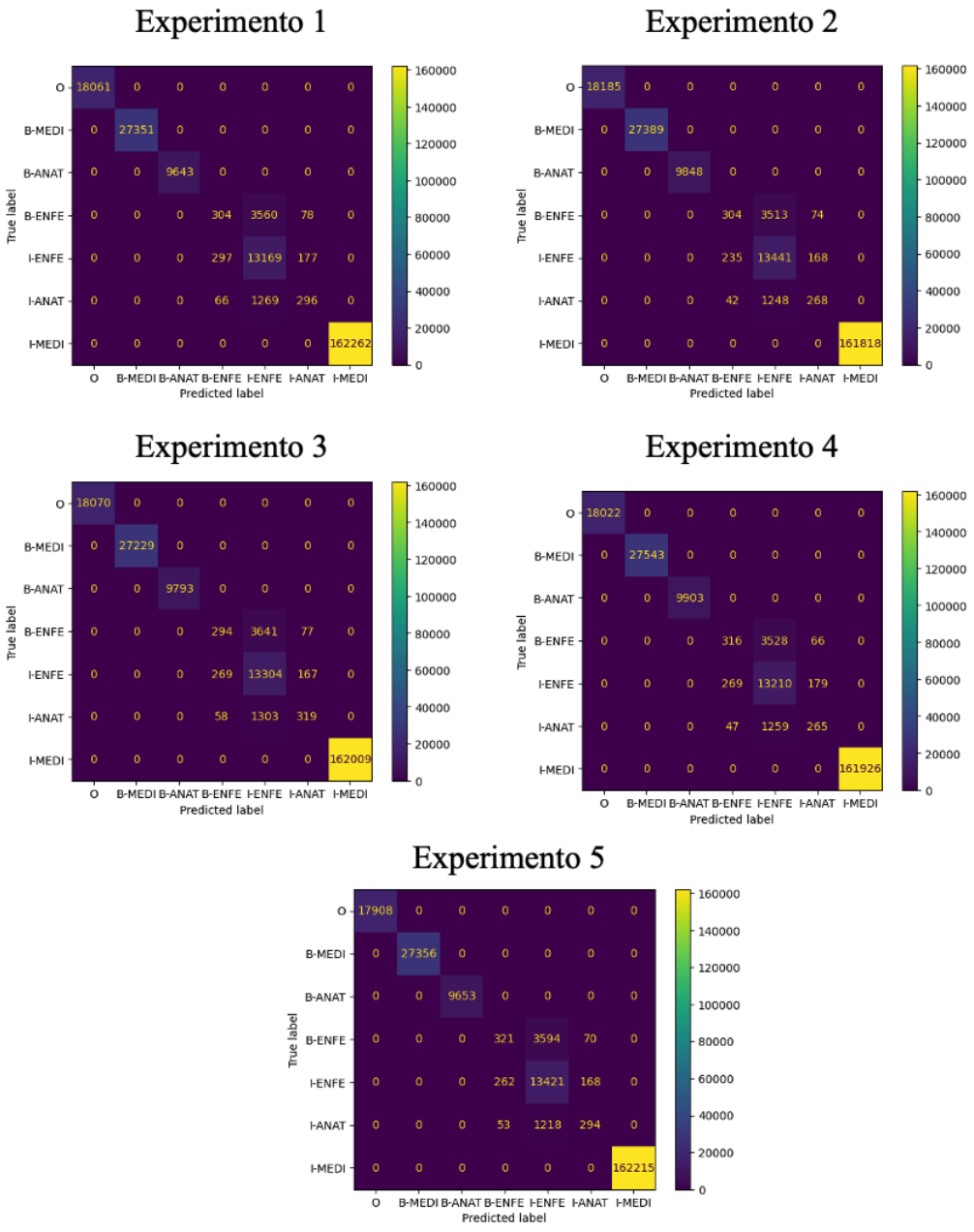

Figura 5.1: Matrices de confusión obtenidas en los experimentos del modelo de árbol de decisión.

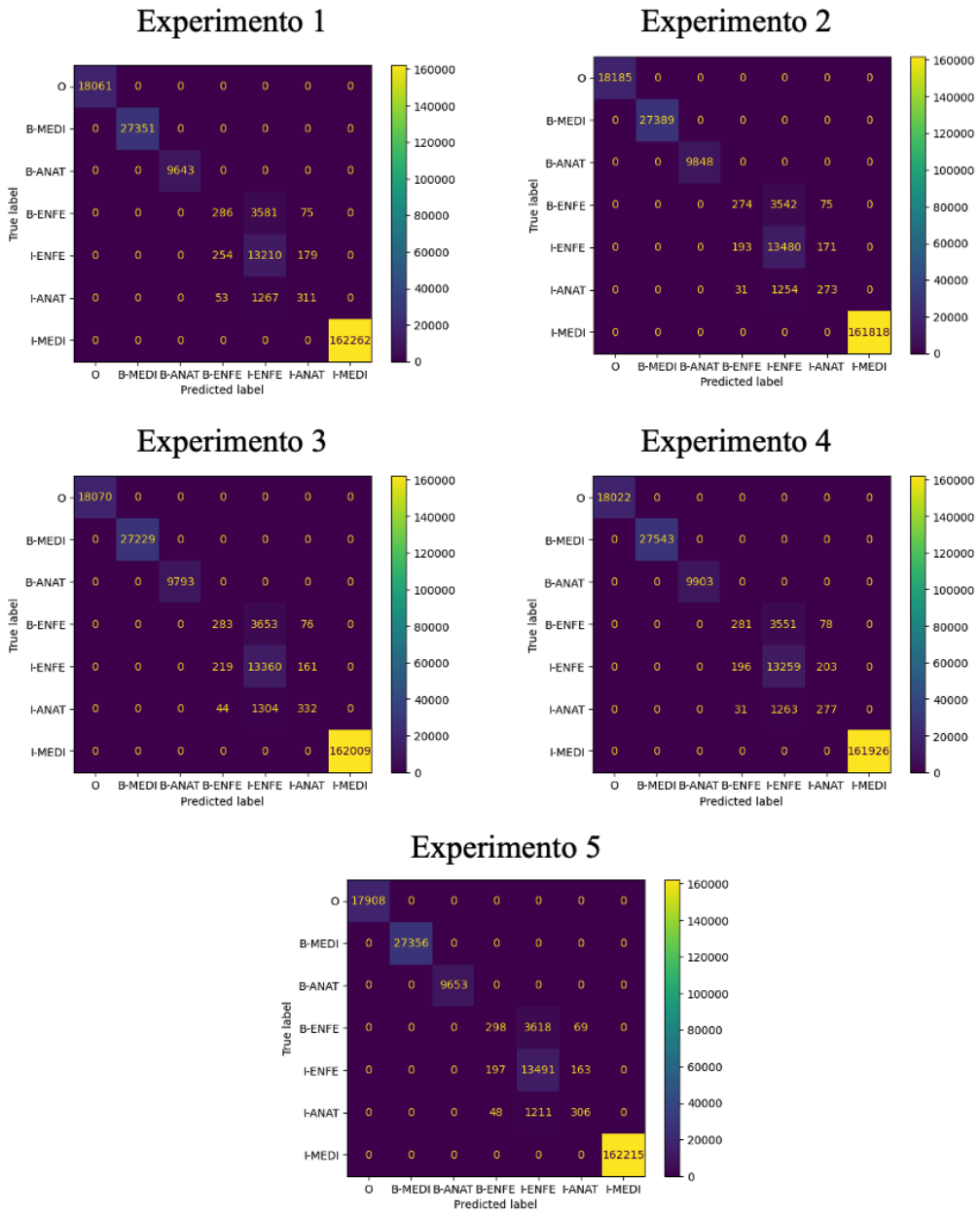

Figura 5.2: Matrices de confusión obtenidas en los experimentos del modelo de random forest.

5.1.1. Evaluación en conjunto de datos médico CoWeSe

Los modelos generados con ambos algoritmos fueron evaluados en un conjunto de textos médicos distinto al utilizado durante el entrenamiento.
La evaluación se realizó utilizando textos del corpus biomédico en español CoWeSe [66], que contiene 1.5 millones de documentos en texto plano y preprocesado. Este corpus fue creado a partir de un rastreo de aproximadamente 3,000 dominios españoles en 2020, llevado a cabo por la Text Mining Unit del Barcelona Supercomputing Center. La extracción de textos se limitó a etiquetas HTML de tipo párrafo y encabezado, abarcando una amplia variedad de fuentes, incluyendo comunidades médicas y científicas, revistas especializadas, centros de investigación, compañías farmacéuticas, portales informativos de salud, asociaciones de pacientes, blogs de profesionales sanitarios, hospitales y organizaciones de salud pública.
Para esta fase, se seleccionaron 5,000 textos del corpus CoWeSe, los cuales fueron anotados utilizando el etiquetador de términos médicos desarrollado en la fase de identificación de entidades, en formato BIO, mediante NLP-MedTerm, un recurso validado por expertos médicos, lo que garantiza la consistencia y fiabilidad de las anotaciones. En la Figura 5.3 se muestra la distribución de clases en el conjunto de datos, evidenciando un desbalance en las categorías. Esto confirma que, incluso tratándose de textos médicos, predominan las palabras generales sobre los términos médicos específicos.

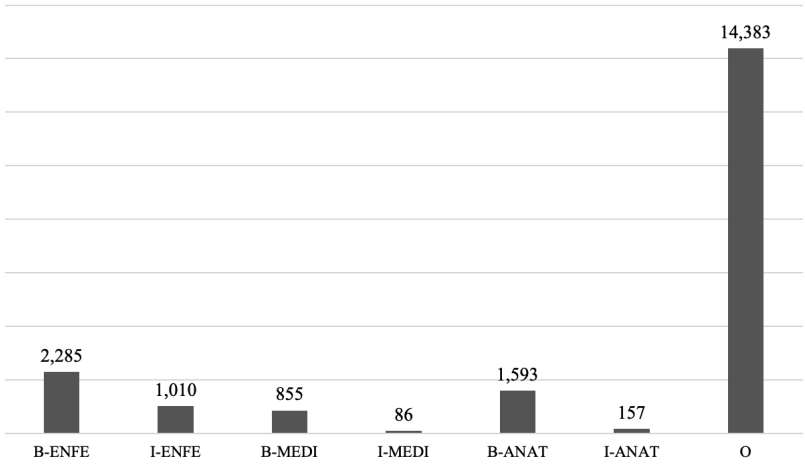

Figura 5.3: Distribución de clases en conjunto de datos CoWeSe.

En total, se generaron 20,369 vectores de características, alcanzando una precisión del 98.8 % tanto en el modelo de Árbol de Decisión como en el de Random Forest.
La Tabla 5.3 presenta la distribución de la precisión por clase para los modelos de Árbol de Decisión y Random Forest evaluado sobre el conjunto de datos CoWeSe, ambos en las mismas condiciones.
La diferencia de rendimiento entre ambos modelos fue mínima en ambas evaluaciones; no obstante, el modelo de Random Forest mostró un desempeño ligeramente superior, evidenciado por puntuaciones de precisión ligeramente más altas.

Tabla 5.3: Precisión obtenida con el conjunto de datos CoWeSe.

Clase	Precisión por clase en modelos	
	Árboles de decición	Random Forest
B-ENFE	100 %	100 %
I-ENFE	97.8 %	98.3 %
B-MEDI	100 %	100 %
I-MEDI	17.4 %	16.2 %
B-ANAT	100 %	100 %
I-ANAT	7 %	5.1 %
O	100 %	100 %
Total	**98.83 %**	**98.83 %**

5.1.2. Ejemplos utilizando el modelo árboles de decisión

A continuación, se presentan ejemplos de la identificación de términos médicos en distintos textos médicos, con el objetivo de mostrar el desempeño del modelo Árboles de Decisión. Las palabras resaltadas corresponden a las entidades reconocidas dentro de cada texto.

- Texto 1. *la* ***enfermedad del coronavirus 2019 covid-19*** *es una* ***enfermedad viral*** *que afecta a varios* ***órganos y sistemas*** *los estudios publicados recientemente sobre el potencial quimioprofilático de la quercetina contra el* ***sars-cov-2*** *metodología se realizó una búsqueda de la literatura en bases como pubmed medline scielo scorpus web of science cochrane library y clinical trials gov se incluyeron y evaluaron críticamente estudios que abordan la quercetina contra el* ***sars-cov-2****.*

La Figura 5.4 presenta la clasificación por token del texto 1 realizada por el modelo.

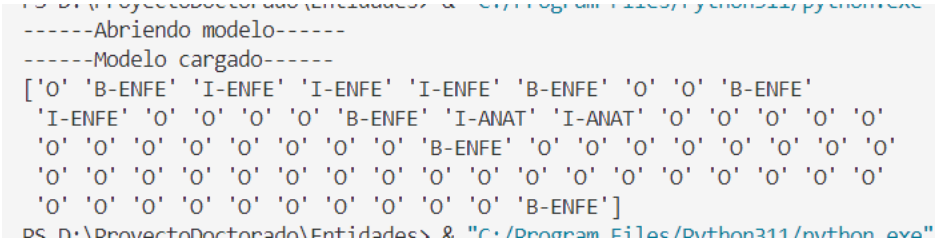

Figura 5.4: Ejemplo 1 del modelo de Árboles de decisión

- Texto 2. *los enfermos de* ***diabetes tipo 2*** *son mas propensos al contagio de* ***coronavirus*** *por el* ***virus*** *de sars-cov-2*

La Figura 5.5 presenta la clasificación por token del texto 2 realizada por el modelo.

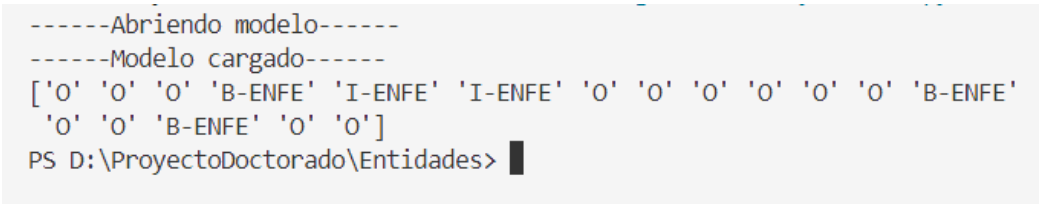

Figura 5.5: Ejemplo 2 del modelo de Árboles de decisión

- Texto 3. *con el tiempo la* ***diabetes de tipo 2*** *puede causar daños graves al organismo sobre todo a los nervios y los* ***vasos*** *sanguíneos*

La Figura 5.6 presenta la clasificación por token del texto 3 realizada por el modelo.

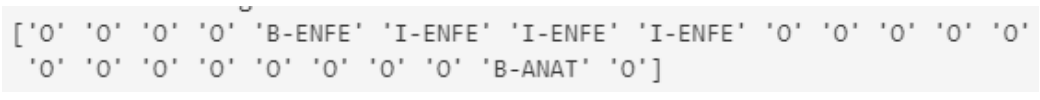

Figura 5.6: Ejemplo 3 del modelo de Árboles de decisión

- Texto 4. *las **enfermedades de los ojos** más frecuentes son las **conjuntivitis** el **ojo seco** los **errores de refracción miopía astigmatismo** e **hipermetropía** y las cataratas otras patologías también frecuentes son los **trastornos del nervio óptico** que incluyen el **glaucoma** y las **enfermedades** de la retina*

La Figura 5.7 presenta la clasificación por token del texto 4 realizada por el modelo.

```
['O' 'B-ENFE' 'I-ENFE' 'I-ENFE' 'I-ENFE' 'O' 'O' 'O' 'O' 'B-ENFE' 'O'
 'B-ENFE' 'I-ENFE' 'O' 'B-ENFE' 'I-ENFE' 'I-ENFE' 'B-ENFE' 'B-ENFE' 'O'
 'B-ENFE' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'B-ENFE' 'I-ENFE' 'I-ENFE'
 'I-ENFE' 'O' 'O' 'O' 'B-ENFE' 'O' 'O' 'B-ENFE' 'O' 'O' 'O']
```

Figura 5.7: Ejemplo 4 del modelo de Árboles de decisión

- Texto 5. *la **cabeza** está formada principalmente por el **cráneo** el hueso que protege a los principales **órganos** del **sistema nervioso central** el **cerebro** y el cerebelo*

La Figura 5.8 presenta la clasificación por token del texto 5 realizada por el modelo.

```
['O' 'B-ANAT' 'O' 'O' 'O' 'O' 'O' 'B-ANAT' 'O' 'O' 'O' 'O' 'O' 'O' 'O'
 'B-ANAT' 'O' 'B-ANAT' 'I-ENFE' 'I-ENFE' 'O' 'B-ANAT' 'O' 'O' 'O']
```

Figura 5.8: Ejemplo 5 del modelo de Árboles de decisión

- Texto 6. *metformina fortamet glumetza y otros es por lo general el primer **medicamento** recetado para la **diabetes tipo 2** funciona principalmente disminuyendo la producción de **glucosa** en el **hígado** y mejorando la sensibilidad del cuerpo a la **insulina** de modo que el organismo utilice la **insulina** de una manera más eficaz*

La Figura 5.9 presenta la clasificación por token del texto 6 realizada por el modelo.

```
['O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'B-MEDI' 'O' 'O' 'O' 'B-ENFE'
 'I-ENFE' 'I-ENFE' 'O' 'O' 'O' 'O' 'O' 'O' 'B-MEDI' 'O' 'O' 'B-ANAT' 'O'
 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'B-MEDI' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'B-MEDI'
 'O' 'O' 'O' 'O' 'O']
```

Figura 5.9: Ejemplo 6 del modelo de Árboles de decisión

Las pruebas con diversos textos utilizando el modelo de árboles decisión revelan una identificación mayoritaria de entidades médicas. No obstante, se identificaron pequeñas omisiones. Por ejemplo, en el texto 1, *quercetina* no se reconoció como medicamento (B-MEDI), y en el texto 5, *cerebelo* no se identificó como parte anatómica (B-ANAT). Esta inconsistencia se atribuye a la ausencia de estas entidades en el recurso Entregable 1 de NLPMedTerm, utilizado para la asignación de ciertas características en el vector generado. Adicionalmente, se observaron identificaciones incorrectas. En el texto 1, *órganos* se clasificó erróneamente como enfermedad (B-ENFE) en lugar de parte anatómica (B-ANAT). Asimismo, en algunos textos, como el 2, el término *sars-cov-2* no se identificó consistentemente como enfermedad (B-ENFE), posiblemente debido a su ubicación al final de la oración, lo que limitó el enriquecimiento de su vector mediante las características de nodos hijos a la derecha y dependencias sintácticas con otros tokens.

5.1.3. Ejemplos utilizando el modelo Random Forest

A continuación, se presentan ejemplos de la identificación de términos médicos en distintos textos médicos, con el objetivo de mostrar el desempeño del modelo Random Forest. Las palabras resaltadas corresponden a las entidades reconocidas dentro de cada texto.

- Texto 1. *la* ***enfermedad del coronavirus 2019 covid-19*** *es una* ***enfermedad viral*** *que afecta a varios* ***órganos y sistemas*** *los estudios publicados recientemente sobre el potencial quimioprofiláticо de la quercetina contra el* ***sars-cov-2*** *metodología se realizó una búsqueda de la literatura en bases como pubmed medline scielo scorpus web of science cochrane library y clinical trials gov se incluyeron y evaluaron críticamente estudios que abordan la quercetina contra el* ***sars-cov-2****.*

La Figura 5.10 presenta la clasificación por token del texto 1 realizada por el modelo.

```
------Abriendo modelo------
------Modelo cargado------
['O' 'B-ENFE' 'I-ENFE' 'I-ENFE' 'I-ENFE' 'B-ENFE' 'O' 'O' 'B-ENFE'
 'I-ENFE' 'O' 'O' 'O' 'O' 'B-ANAT' 'I-ANAT' 'I-ANAT' 'O' 'O' 'O' 'O' 'O'
 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'B-ENFE' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O'
 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O'
 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'B-ENFE']
```

Figura 5.10: Prueba 1 del modelo de Random Forest

- Texto 2. *los enfermos de* ***diabetes tipo 2*** *son mas propensos al contagio de* ***coronavirus*** *por el* ***virus*** *de sars-cov-2*

La Figura 5.11 presenta la clasificación por token del texto 2 realizada por el modelo.

```
------Abriendo modelo------
------Modelo cargado------
['O' 'O' 'O' 'B-ENFE' 'I-ENFE' 'I-ENFE' 'O' 'O' 'O' 'O' 'O' 'O' 'B-ENFE'
 'O' 'O' 'B-ENFE' 'O' 'O']
```

Figura 5.11: Prueba 2 del modelo de Random Forest

- Texto 3. *con el tiempo la* ***diabetes de tipo 2*** *puede causar daños graves al organismo sobre todo a los nervios y los* ***vasos*** *sanguíneos*

La Figura 5.12 presenta la clasificación por token del texto 3 realizada por el modelo.

```
------Abriendo modelo------
------Modelo cargado------
['O' 'O' 'O' 'O' 'B-ENFE' 'I-ENFE' 'I-ENFE' 'I-ENFE' 'O' 'O' 'O' 'O' 'O'
 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'B-ANAT' 'O']
```

Figura 5.12: Prueba 3 del modelo de Random Forest

- Texto 4. *las **enfermedades de los ojos** más frecuentes son las **conjuntivitis** el **ojo seco** los **errores de refracción miopía astigmatismo** e **hipermetropía** y las cataratas otras patologías también frecuentes son los **trastornos del nervio óptico** que incluyen el **glaucoma** y las **enfermedades** de la retina*

La Figura 5.13 presenta la clasificación por token del texto 4 realizada por el modelo.

```
------Abriendo modelo------
------Modelo cargado------
['O' 'B-ENFE' 'I-ENFE' 'I-ENFE' 'I-ENFE' 'O' 'O' 'O' 'O' 'B-ENFE' 'O'
 'B-ENFE' 'I-ENFE' 'O' 'B-ENFE' 'I-ENFE' 'I-ENFE' 'B-ENFE' 'B-ENFE' 'O'
 'B-ENFE' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'B-ENFE' 'I-ENFE' 'I-ENFE'
 'I-ENFE' 'O' 'O' 'O' 'B-ENFE' 'O' 'O' 'B-ENFE' 'O' 'O' 'O']
```

Figura 5.13: Prueba 4 del modelo de Random Forest

- Texto 5. *la **cabeza** está formada principalmente por el **cráneo** el hueso que protege a los principales **órganos** del **sistema nervioso central** el **cerebro** y el cerebelo*

La Figura 5.14 presenta la clasificación por token del texto 5 realizada por el modelo.

```
------Abriendo modelo------
------Modelo cargado------
['O' 'B-ANAT' 'O' 'O' 'O' 'O' 'O' 'B-ANAT' 'O' 'O' 'O' 'O' 'O' 'O' 'O'
 'B-ANAT' 'O' 'B-ANAT' 'I-ENFE' 'I-ENFE' 'O' 'B-ANAT' 'O' 'O' 'O']
```

Figura 5.14: Prueba 5 del modelo de Random Forest

- Texto 6. *metformina fortamet glumetza y otros es por lo general el primer **medicamento** recetado para la **diabetes tipo 2** funciona principalmente disminuyendo la producción de **glucosa** en el **hígado** y mejorando la sensibilidad del cuerpo a la **insulina** de modo que el organismo utilice la **insulina** de una manera más eficaz*

La Figura 5.15 presenta la clasificación por token del texto 6 realizada por el modelo.

```
------Abriendo modelo------
------Modelo cargado------
['O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'B-MEDI' 'O' 'O' 'O' 'B-ENFE'
 'I-ENFE' 'I-ENFE' 'O' 'O' 'O' 'O' 'O' 'O' 'B-MEDI' 'O' 'O' 'B-ANAT' 'O'
 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'B-MEDI' 'O' 'O' 'O' 'O' 'O' 'O' 'O' 'B-MEDI'
 'O' 'O' 'O' 'O' 'O']
```

Figura 5.15: Prueba 6 del modelo de Random Forest

Las pruebas se realizaron utilizando los mismos textos empleados en el modelo anterior. En general, se obtuvieron resultados similares, destacando una mejora en el Texto 1, donde el término *órganos* fue correctamente identificado como parte anatómica (B-ANAT). Por otro lado, al igual que con el modelo de árboles de decisión, se observó que algunos términos ubicados al final de los textos no fueron correctamente reconocidos.

5.2. Identificación de relaciones

Para desarrollar un modelo capaz de clasificar la existencia o ausencia de una relación entre dos términos médicos, se llevó a cabo el entrenamiento de un LLM previamente entrenado en el dominio biomédico. Este proceso se realizó utilizando una combinación de hiperparámetros seleccionados mediante una estrategia experimental sistemática, orientada a optimizar el desempeño del modelo en la tarea específica de identificación de relaciones semánticas en artículos científicos del dominio médico.
Los textos empleados en esta fase fueron generados a partir del enfoque propuesto, el cual incluyó las etapas de identificación de entidades médicas, enmascaramiento contextual y segmentación de oraciones. Cada uno de estos textos fue validado por expertos en el área de la salud, garantizando así la calidad y la pertinencia del corpus utilizado.
Cabe destacar que el conjunto de datos construido durante esta etapa ha sido puesto a disposición de la comunidad, con el objetivo de fomentar la reproducibilidad y promover futuras investigaciones en el campo del PLN en

el ámbito biomédico. Este corpus fue empleado como base para las experimentaciones subsiguientes con LLM´s.

5.2.1. Entrenamiento de modelo de lenguaje

Con el objetivo de evaluar el rendimiento del modelo en la tarea de clasificación binaria, se optó por utilizar cuatro métricas ampliamente reconocidas en el campo de la extracción de información (Accuracy, Precision, Recall y F1-score). Estas métricas fueron seleccionadas por su capacidad para proporcionar una evaluación robusta y complementaria del desempeño del modelo. En conjunto, permiten no solo medir la proporción de predicciones correctas, sino también evaluar la calidad de las predicciones positivas, la cobertura de las relaciones reales detectadas y el equilibrio entre ambas. Esta elección resulta particularmente pertinente en tareas sensibles como la detección de relaciones entre conceptos médicos, donde tanto los falsos positivos como los falsos negativos pueden tener implicaciones significativas.
Inicialmente, se utilizó un LLM base, específicamente BERT [67], diseñado para comprender el contexto completo de una palabra dentro de una oración, considerando tanto las palabras precedentes como las subsecuentes. Este modelo fue preentrenado con grandes corpus textuales, incluyendo Wikipedia en inglés y BooksCorpus, permitiendo así una comprensión profunda del lenguaje general. Para la etapa inicial de experimentación, se empleó la siguiente configuración de hiperparámetros:

- Tasa de aprendizaje: $5e^{-5}$
- Número de épocas: 20
- Tamaño del lote: 8
- Peso: 0.01
- Optimizador: Adam

Los resultados obtenidos con esta configuración se presentan en la Tabla 5.4, donde se observa que la precisión alcanzada fue del 81.2 %.

Tabla 5.4: Resultados de entrenamiento LLM BERT para la identificación de relaciones médicas

Época	Accuracy	Precisión	Recall	F1-score
1	0.624	0.75	0.4	0.55
2	0.728	0.77	0.75	0.76
3	0.744	0.756	0.818	0.786
4	0.765	0.815	0.763	0.788
5	0.697	0.795	0.636	0.707
6	0.765	0.753	0.877	0.81
7	0.788	0.79	0.859	0.823
8	0.78	0.79	0.84	0.814
9	**0.817**	**0.812**	**0.886**	**0.847**
10	0.798	0.832	0.813	0.822
11	0.812	0.839	0.831	0.835
12	0.814	0.846	0.827	0.836
13	0.783	0.851	0.754	0.8
14	0.791	0.843	0.781	0.811
15	0.804	0.808	0.863	0.835
16	0.809	0.823	0.85	0.836
17	0.809	0.823	0.85	0.836
18	0.809	0.818	0.859	0.838
19	0.812	0.813	0.872	0.842
20	0.806	0.811	0.863	0.837

Posteriormente, se experimentó con un modelo especializado en el dominio biomédico, denominado MedicoBERT [65]. De acuerdo con sus autores, este modelo fue entrenado mediante las tareas de enmascaramiento de palabras (Masked Language Modeling) y predicción de la siguiente oración (Next Sentence Prediction), utilizando más de tres millones de textos médicos en español. Dichos textos provienen de tres conjuntos de datos ampliamente reconocidos en el área: BioASQ, CoWeSe y CORD-19, lo que representa un corpus total de aproximadamente 1.1 mil millones de palabras.
Asimismo, se empleó el tokenizador propio de MedicoBERT, diseñado específicamente para el lenguaje médico, el cual cuenta con un vocabulario de más de 50,000 tokens especializados, lo que permite una representación más precisa y contextualizada de los términos del dominio. El objetivo de esta fase fue evaluar el impacto de un vocabulario específico del dominio en la capacidad del modelo para aprender relaciones médicas de manera más efectiva.

En esta etapa se llevó a cabo una calibración gruesa, entendida como un enfoque exploratorio destinado a identificar un rango de valores prometedores para cada hiperparámetro. Este proceso fue ejecutado de forma sistemática mediante una búsqueda manual, guiada por la revisión de literatura especializada en tareas similares. Se utilizaron como punto de partida los valores comúnmente reportados en estudios previos, adaptándolos al contexto específico de la tarea de clasificación de relaciones médicas. Los hiperparámetros evaluados incluyeron el número de épocas, tamaño del lote, tasa de aprendizaje y peso.
Estos parámetros fueron seleccionados por su influencia crítica en la dinámica de aprendizaje, afectando tanto la convergencia como la estabilidad del modelo. Los valores seleccionados en esta fase fueron los siguientes:

- Tasa de aprendizaje: $5e^{-5}$
- Número de épocas: 20
- Tamaño del lote: 8
- Peso: 0.01

Los resultados de esta experimentación se presentan en la Tabla 5.5, donde se aprecia una mejora en la precisión, alcanzando un 88.8 %, lo cual sugiere que la especialización en el vocabulario del dominio médico contribuye significativamente a una clasificación más precisa.

Tabla 5.5: Resultados de entrenamiento LLM MedicoBERT para la identificación de relaciones médicas

Época	Accuracy	Precisión	Recall	F1-score
1	0.738	0.78	0.766	0.773
2	0.783	0.799	0.838	0.818
3	0.809	0.804	0.887	0.844
4	0.817	0.862	0.816	0.838
5	0.822	0.824	0.883	0.852
6	0.788	0.873	0.744	0.803
7	0.832	0.819	0.914	0.864
8	0.835	0.847	0.874	0.86
9	0.84	0.868	0.856	0.862
10	0.851	0.864	0.883	0.873
11	0.848	0.857	0.887	0.872
12	**0.853**	**0.888**	**0.856**	**0.872**
13	0.832	0.829	0.896	0.862
14	0.843	0.849	0.887	0.868
15	0.861	0.882	0.878	0.88
16	0.835	0.838	0.887	0.862
17	0.856	0.868	0.887	0.878
18	0.859	0.872	0.887	0.88
19	0.861	0.882	0.878	0.88
20	0.864	0.883	0.883	0.883

Con base en estos resultados preliminares, se procedió a una calibración fina de los hiperparámetros, delimitando un espacio de búsqueda más acotado. Para esta etapa se empleó optimización bayesiana, una técnica avanzada que permite explorar eficientemente configuraciones de hiperparámetros al modelar probabilísticamente la función objetivo, en este caso, maximizar la precisión del modelo. La optimización bayesiana se basa en la construcción de un modelo probabilístico sustituto, generalmente un proceso gaussiano (Gaussian Process, GP), que estima el rendimiento del modelo en función de diferentes configuraciones, como se muestra en la ecuación 5.2.

$$f(x) \sim \mathcal{GP}(\mu(x), k(x, x')) \tag{5.2}$$

Donde:

$\mu(x)$: *valor esperado del desempeño*

$k(x, x')$: *kernel que representa la similitud entre configuraciones*

A partir de este modelo, se define una función de adquisición, como Expected Improvement (EI), la cual guía la selección de nuevos puntos en el espacio de búsqueda, se expresa en la ecuación 5.3:

$$\alpha_{\mathrm{EI}}(x) = E\left[\max\left(f(x) - f(x^{+}), 0\right)\right] \tag{5.3}$$

Esta estrategia permitió reducir de manera significativa el número de evaluaciones necesarias para encontrar configuraciones óptimas, lo que es especialmente relevante cuando cada evaluación representa un alto costo computacional. Los hiperparámetros encontrados mediante este enfoque fueron los siguientes:

- Tasa de aprendizaje: $2{,}39291e^{-5}$
- Número de épocas: 15
- Tamaño del lote: 16
- Peso: 0.01

Los resultados finales, presentados en la Tabla 5.6, muestran que la precisión se incrementó hasta un 90.6 %, lo cual confirma que una adecuada configuración de hiperparámetros, sumada a la especialización del modelo en el dominio médico, permite alcanzar niveles de rendimiento comparables con los reportados en la literatura especializada para tareas similares.

Tabla 5.6: Resultados de entrenamiento LLM MedicoBERT con calibración de hiperparámetros para la identificación de relaciones médicas

Época	**Accuracy**	**Precisión**	**Recall**	**F1-score**
1	0.626	0.854	0.474	0.61
2	0.788	0.889	0.75	0.813
3	0.822	0.806	0.936	0.866
4	0.843	0.838	0.923	0.879
5	0.83	0.854	0.872	0.863
6	0.843	0.863	0.885	0.874
7	0.838	0.845	0.902	0.872
8	0.848	0.867	0.889	0.878
9	0.838	0.875	0.86	0.867
10	0.832	0.894	0.826	0.859
11	0.835	0.881	0.847	0.863
12	0.843	0.872	0.872	0.872
13	0.848	0.864	0.894	0.879
14	0.835	0.882	0.817	0.859
15	**0.859**	**0.906**	**0.889**	**0.886**
16	0.856	0.881	0.885	0.883
17	0.859	0.895	0.872	0.884
18	0.851	0.874	0.885	0.88
19	0.859	0.892	0.877	0.884
20	0.856	0.881	0.885	0.883

En la Tabla 5.7 se presenta un resumen de los valores máximos alcanzados durante la época 15, evaluados conforme a las métricas previamente descritas.

Tabla 5.7: Valores máximos en métricas durante entrenamientos de grandes modelos de lenguaje.

Modelo	Métrica			
	A	P	R	F1
BERT	0.817	0.812	0.886	0.847
MedicoBERT	0.853	0.888	0.856	0.872
MedicoBERT + calibración	**0.859**	**0.906**	**0.889**	**0.886**

El modelo BERT general presenta un rendimiento aceptable, con una precisión de 0.812 y un F1-score de 0.847. Sin embargo, su desempeño en términos de accuracy (0.817) y recall (0.886) resulta inferior en comparación

con los modelos especializados en el dominio médico.
Al incorporar un modelo preentrenado específicamente en lenguaje médico en español, MedicoBERT, se observa una mejora significativa en todas las métricas evaluadas. Este modelo demuestra una mayor capacidad para capturar relaciones médicas relevantes, lo que se refleja en un incremento en la precisión (0.888) y el F1-score (0.872), junto con una mejora en la accuracy (0.853) respecto a BERT.
Por su parte, el modelo MedicoBERT calibrado mediante optimización bayesiana de hiperparámetros alcanza los valores más altos en las métricas clave, con una precisión de 0.906 y un recall de 0.889, evidenciando una notable capacidad para identificar correctamente las relaciones médicas sin aumentar los falsos positivos. Asimismo, obtiene el mayor F1-score (0.886), lo que indica un equilibrio óptimo entre precisión y sensibilidad. Es importante destacar que, a pesar de que la accuracy (0.859) de este modelo es ligeramente superior a la de MedicoBERT sin calibración (0.853) no representa un cambio significativo, ya que las métricas más relevantes en contextos de clasificación como precisión, recall y F1-score muestran mejoras significativas. Esto podría deberse a un ligero incremento en los errores sobre clases neutras o mayoritarias, sin comprometer el rendimiento general en las clases clínicas de interés.
En conjunto, los resultados muestran que la especialización del modelo en el dominio médico a través de MedicoBERT, junto con la calibración fina de hiperparámetros, tiene un impacto positivo en el rendimiento del modelo. A pesar de una leve reducción en la accuracy, el aumento en las métricas más representativas valida la eficacia de esta estrategia. Estos hallazgos confirman que el vocabulario, los datos válidados por expertos para la tarea de clasificación realizados y el contexto propios del lenguaje médico son elementos determinantes para mejorar el desempeño de los modelos de PLN en tareas especializadas.

5.2.2. Verificación del desempeño del modelo en la identificación de relaciones

Para la verificación del modelo entrenado encargado de la identificación de relaciones entre entidades médicas, se reservó un conjunto semilla extraído del corpus CoWeSe [66], compuesto por 100 frases médicas. Estas frases incluían pares de entidades correspondientes a las categorías utilizadas en esta investigación: Anatomía, Medicamento y Enfermedad. El objetivo fue

evaluar la capacidad del modelo para generalizar sobre nuevos ejemplos y clasificar correctamente la existencia o ausencia de relaciones entre dichas entidades.
Como resultado, el modelo alcanzó un 76 % de clasificaciones correctas, lo que indica un rendimiento general sólido. En la Figura 5.16, se presenta la matriz de confusión correspondiente, la cual permite observar con mayor detalle el tipo de aciertos y errores cometidos durante la evaluación.

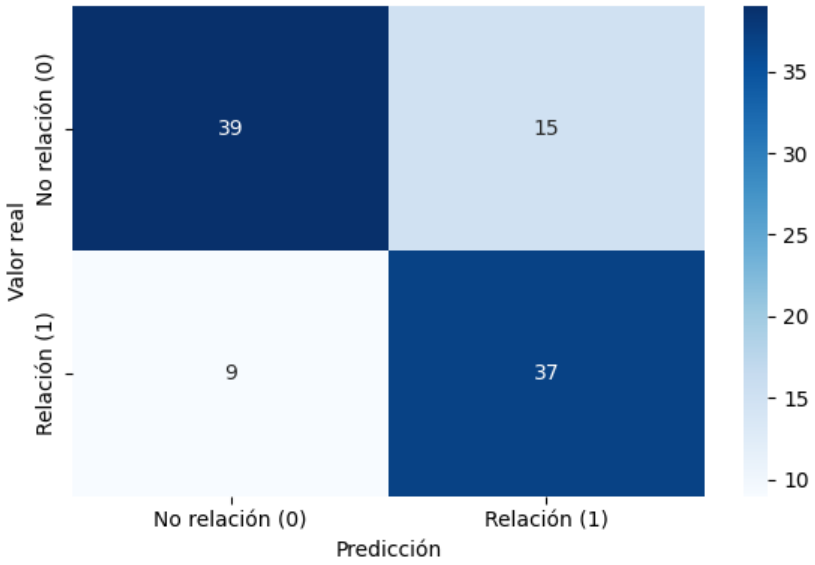

Figura 5.16: Matriz de confusión de la fase de verificación del modelo utilizando un subconjunto de datos del corpus CoWeSe.

A partir de esta matriz, se pueden calcular las cuatro métricas de evaluación previamente descritas: Accuracy (ecuación 5.4), Precisión (ecuación 5.5), Recall (ecuación 5.6) y F1-score (ecuación 5.7), cuyos valores obtenidos fueron los siguientes:

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN} = \frac{37 + 39}{37 + 39 + 15 + 9} = \frac{76}{100} = 0{,}76 \tag{5.4}$$

$$Precisión = \frac{VP}{VP + FP} = \frac{37}{37 + 15} = \frac{37}{52} = 0{,}71 \tag{5.5}$$

$$Recall = \frac{VP}{VP + FN} = \frac{37}{37 + 9} = \frac{37}{46} = 0{,}80 \tag{5.6}$$

$$F1 - score = 2 \times \frac{Precisión \times Recall}{Precisión + Recall} = 2 \times \frac{0{,}71 \times 0{,}80}{0{,}71 + 0{,}80} = 0{,}75 \tag{5.7}$$

Estos resultados muestran que el modelo tiene una alta capacidad para detectar relaciones verdaderas, logrando identificar correctamente el 80 % de las relaciones reales. Además, se observa un balance aceptable entre precisión y exhaustividad, reflejado en un valor de F1-score del 75 %, lo que indica un rendimiento equilibrado a pesar de que el modelo tiende a sobrepredecir relaciones con 15 falsos positivos.
Finalmente, los resultados obtenidos en esta fase de verificación validan la efectividad del enfoque incluyendo el uso de un modelo de lenguaje preentrenado y la incorporación de datos etiquetados y validados por expertos médicos. Esto respalda la viabilidad del uso de modelos de lenguaje en tareas como la identificación automática de relaciones semánticas en el dominio de la salud.

5.3. Generación de grafo de conocimiento

Con el propósito de evaluar la calidad del grafo de conocimiento generado, se diseñaron cinco casos de uso a partir de textos médicos provenientes de fuentes diversas. Esta etapa permitió analizar si el lenguaje original del texto, su redacción o el contexto de origen influían en el desempeño del sistema, tanto en su metodología como en la generación de tripletas.
Los textos utilizados incluyeron artículos científicos, sitios web especializados en salud y noticias médicas. Las relaciones extraídas por el sistema y representadas en el grafo de conocimiento fueron sometidas a un proceso de validación por parte de un experto en el dominio médico, quien identificó si las relaciones eran correctas, permitiendo así una evaluación de la calidad del grafo final.

Caso de uso 1. Resumen de artículo científico

- **Título.** Enfermedad pulmonar obstructiva crónica [68].

- **Texto.** La enfermedad pulmonar obstructiva crónica (EPOC), Se define como un estado patológico que se caracteriza por una limitación del flujo de aire que no es del todo reversible. El EPOC incluye el enfisema, un cuadro que se define en términos anatómicos, y que se caracteriza por destrucción y ensanchamiento de los alveolos pulmonares; La bronquitis crónica, un cuadro que se define en términos clínicos por tos crónica productiva, y finalmente la enfermedad de las vías respiratorias finas, en la que se estrechan los bronquiolos finos.

Las tripletas extraídas y representadas en el grafo de conocimiento del caso de uso 1 se detallan a continuación. Su visualización gráfica se presenta en la Figura 5.17, la cual ilustra de manera estructurada las relaciones entre las entidades identificadas.

1. $es\_un(vias\ respiratorias, anatomia)$
2. $es\_un(enfermedad\ pulmonar\ obstructiva\ cronica, enfermedad)$
3. $es\_un(epoc, enfermedad)$
4. $es\_un(bronquitis, enfermedad)$
5. $es\_un(bronquitis\ cronica, enfermedad)$

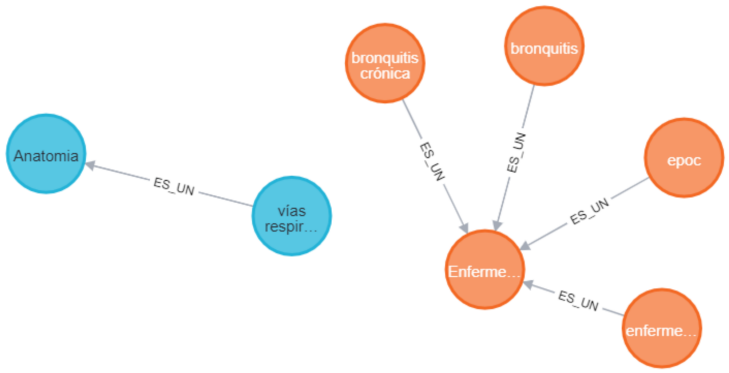

Figura 5.17: Representación de tripletas del caso de uso 1.

Caso de uso 2. Noticia médica

- **Título.** La pastilla contra la hipertensión en adultos también ayuda a los niños con piel de mariposa [69].
- **Texto.** La epidermólisis bullosa es una enfermedad genética rara que afecta la integridad de la piel y las mucosas, caracterizada por la formación de ampollas y heridas ante mínimos traumatismos. Esta condición compromete estructuras anatómicas como la epidermis y la dermis, debilitando la cohesión entre ellas. Recientemente, se ha explorado el uso del losartán, un medicamento antihipertensivo, como terapia potencial para mejorar la cicatrización en pacientes con epidermólisis bullosa. Estudios preliminares sugieren que el losartán podría reducir la fibrosis y promover la regeneración de la piel, ofreciendo una alternativa terapéutica prometedora. Aunque los resultados iniciales son alentadores, se requieren ensayos clínicos controlados para validar la eficacia y seguridad del losartán en este nuevo contexto terapéutico.

A continuación, se presentan las tripletas que conforman el grafo de conocimiento respecto al caso de uso 2. La Figura 5.18 ofrece una representación visual de dichas tripletas, permitiendo observar las relaciones establecidas entre las entidades reconocidas.

1. $es\_un(piel, anatomia)$
2. $trata\_condiciones(losartan, piel)$
3. $es\_un(losartan, medicamento)$

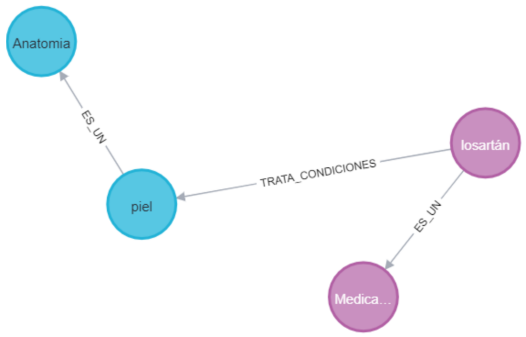

Figura 5.18: Representación de tripletas del caso de uso 2.

Caso de uso 3. Resumen de artículo científico

- **Título.** Anemia hemolítica autoinmune (AIHA) tras trasplante alogénico de células madre hematopoyéticas (TCMH): Un análisis retrospectivo y una propuesta de tratamiento por parte del Grupo Español De Trasplante de Médula Osea en Niños (GETMON) y el Grupo Español de Trasplante Hematopoyético (GETH) [70].

- **Texto.** La anemia hemolítica autoinmune (AHAI) es una complicación poco común pero grave que puede presentarse tras un trasplante alogénico de células madre hematopoyéticas (TCMH). Esta condición se caracteriza por la destrucción inmunomediada de los eritrocitos, lo que conduce a una anemia severa. El estudio retrospectivo analiza la incidencia, características clínicas y respuesta al tratamiento de la AHAI en pacientes pediátricos que han recibido un TCMH. Los hallazgos indican que la AHAI post-TCMH se asocia con una alta morbilidad y requiere un manejo terapéutico intensivo. Los tratamientos utilizados incluyen corticosteroides, inmunoglobulina intravenosa y, en casos refractarios, rituximab. El reconocimiento temprano y el tratamiento adecuado son esenciales para mejorar los resultados clínicos en esta población.

Las tripletas del caso de uso 3 y que conforman el grafo de conocimiento generado se muestran a continuación. Su representación visual, mostrada en la Figura 5.19, permite examinar la estructura relacional entre las entidades y facilita el análisis de la organización.

1. $es\_un(rituximab, medicamento)$
2. $es\_un(corticosteroides, medicamento)$
3. $es\_un(inmunoglobulina\ intravenosa, medicamento)$
4. $usa\_tratamiento(AHAI, corticosteroides)$
5. $usa\_tratamiento(AHAI, inmunoglobulina\ intravenosa)$
6. $es\_un(AHAI, enfermedad)$
7. $es\_un(anemia\ hemoltica\ autoinmune, enfermedad)$
8. $es\_un(anemia, enfermedad)$

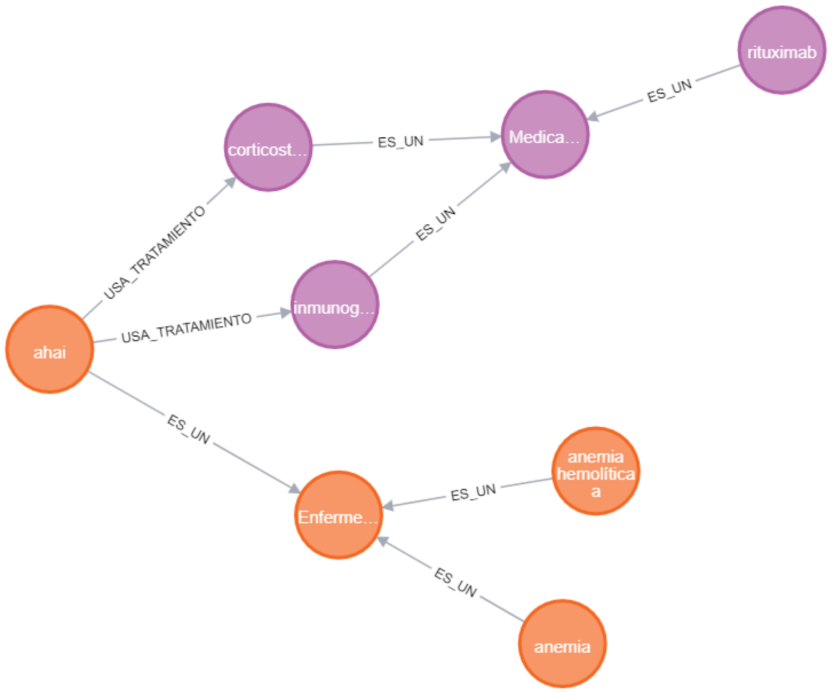

Figura 5.19: Representación de tripletas del caso de uso 3.

Caso de uso 4. Informes médicos

- **Título.** Informes actuales sobre diabetes [71].
- **Texto.** La diabetes y la depresión se presentan juntas con una frecuencia aproximadamente dos veces mayor de la que se predeciría por casualidad. La diabetes y la depresión comórbidas representan un desafío clínico importante, ya que los resultados de ambas afecciones se ven agravados por la otra. Si bien la carga psicológica de la diabetes puede contribuir a la depresión, esta explicación no explica completamente la relación entre estas dos afecciones. Ambas afecciones pueden estar impulsadas por mecanismos biológicos y conductuales subyacentes compartidos, como la activación del eje hipotálamo-hipofisario-adrenal, la inflamación, los trastornos del sueño, el estilo de vida inactivo, los malos hábitos alimenticios y los factores de riesgo ambientales y culturales. La depresión con frecuencia pasa desapercibida en personas con

diabetes a pesar de la disponibilidad de herramientas de detección efectivas. Tanto las intervenciones psicológicas como los antidepresivos son eficaces para tratar los síntomas depresivos en personas con diabetes, pero tienen efectos mixtos en el control glucémico. Se necesitan vías de atención claras que involucren a un equipo multidisciplinario para obtener resultados médicos y psiquiátricos óptimos para las personas con diabetes y depresión comórbidas.

A continuación, se presentan las tripletas del caso de uso 4 que conforman el grafo de conocimiento. La Figura 5.20 muestra su representación visual, lo que permite una comprensión más clara de las relaciones establecidas entre las entidades médicas y de la estructura general del conocimiento representado.

1. $es\_un(antidepresivos, medicamento)$
2. $usa\_tratamiento(diabetes, antidepresivos)$
3. $es\_tratamiento(antidepresivos, diabetes)$
4. $es\_un(diabetes, enfermedad)$
5. $es\_un(adrenal, anatomia)$
6. $es\_un(hipofisario, anatomia)$

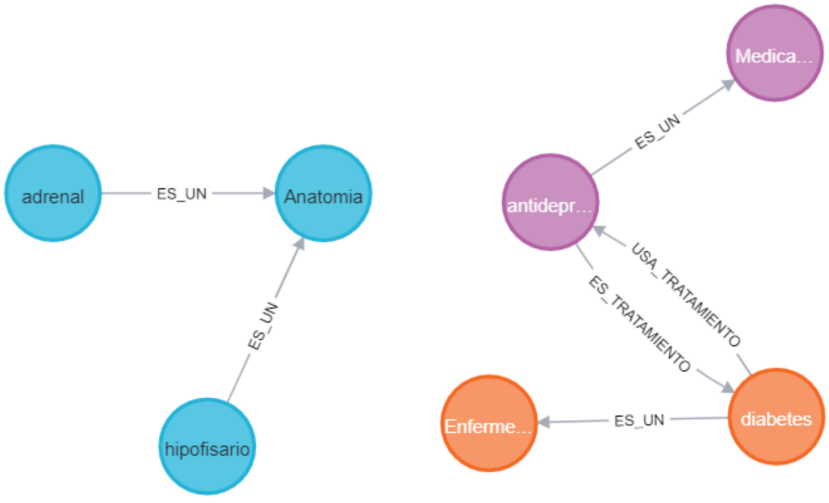

Figura 5.20: Representación de tripletas del caso de uso 4.

Caso de uso 5. Resumen de artículo científico

- **Título.** Lupus eritematoso sistémico [72].
- **Texto.** El lupus eritematoso sistémico (LES) es una enfermedad autoinmune compleja, caracterizada por afectar los anticuerpos, algunos de ellos claramente relacionados con manifestaciones típicas de la enfermedad. El LES puede aparecer a cualquier edad, pero afecta fundamentalmente a mujeres jóvenes en edad fértil. La patogénesis del lupus continúa sin conocerse, no obstante se han relacionado factores genéticos, ambientales, hormonales, así como también diversas alteraciones celulares y una pérdida en el equilibrio de las citoquinas. El cuadro clínico es muy heterogéneo, pudiendo afectar a casi cualquier órgano. Las principales manifestaciones clínicas son la afectación articular, la cutánea, la glomerulonefritis, la serositis (pleuritis y/o pericarditis), la afectación del sistema nervioso central y en ocasiones la trombosis. El compromiso renal marca claramente el pronóstico de los pacientes con LES. Estudios recientes muestran una mejoría en las tasas de sobrevida de los pacientes con LES, así como también una reducción en el número de brotes. Durante los últimos años se vienen desarrollando nuevas terapias para el LES que parecen tener unas tasas de respuestas esperanzadoras.

Las tripletas resultantes del caso de uso 5 se detallan a continuación. Su visualización, presentada en la Figura 5.21, ofrece una perspectiva estructurada de las relaciones identificadas entre las entidades, facilitando el análisis del grafo construido.

1. $es\_un(rion, anatomia)$
2. $es\_un(riones, anatomia)$
3. $es\_un(snc, anatomia)$
4. $es\_un(renal, anatomia)$
5. $es\_un(sistema\ nervioso\ central, anatomia)$
6. $tiene\_efecto\_en(pericarditis, renal)$
7. $afectado\_por(renal, lupus)$

8. $afectado\_por(renal, glomerulonefritis)$
9. $es\_un(lupus, enfermedad)$
10. $es\_un(pericarditis, enfermedad)$
11. $es\_un(glomerulonefritis, enfermedad)$
12. $tiene\_efecto\_en(glomerulonefritis, sistema\ nervioso\ central)$
13. $tiene\_efecto\_en(pericarditis, sistema\ nervioso\ central)$

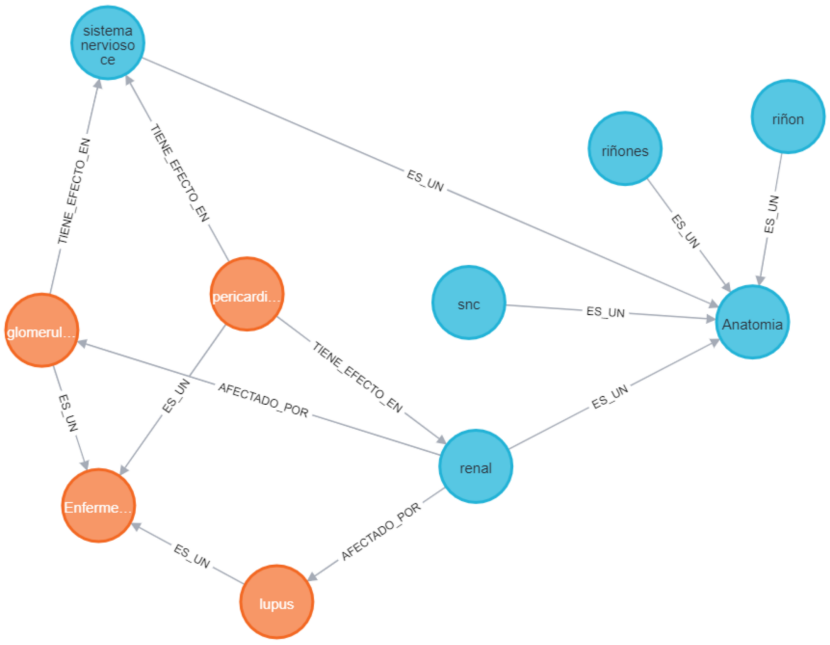

Figura 5.21: Representación de tripletas del caso de uso 5.

En los 5 casos de uso presentados el experto en el dominio médico indicó que todas las relaciones extraídas del texto y contenidas en el grafo de conocimiento eran correctas. No obstante, también señaló la ausencia de algunas tripletas esperadas. Esta limitación podría superarse mediante la incorporación de un más de datos, tanto para mejorar la identificación de entidades como para enriquecer la detección de relaciones semánticas entre ellas.

Capítulo 6

Conclusiones y trabajo a futuro

Como resultado de esta investigación, se desarrolló un sistema integral orientado a la extracción de conocimiento médico a partir de texto plano en español, centrado en el reconocimiento de entidades y relaciones médicas, y su posterior representación en grafos de conocimiento. En primer lugar, se logró implementar un reconocedor de entidades médicas capaz de identificar con una precisión de 97 % con tres tipos de entidades (Anatomía, Medicamento y Enfermedad). Esta evaluación se basó en una experimentación cruzada sobre un conjunto de datos desbalanceado, utilizando algoritmos de aprendizaje automático como árboles de decisión y random forest. Adicionalmente, se validó el modelo en un conjunto de datos externo al utilizado en el entrenamiento, demostrando su robustez. En segundo lugar, se diseñó un modelo para la identificación de relaciones entre entidades médicas, utilizando grandes modelos de lenguaje (LLM) ajustados mediante fine-tuning para la tarea específica de clasificación binaria (relación o no relación) con una precisión de 90.6 %. El modelo fue entrenado sobre un corpus anotado manualmente y validado por tres expertos en el dominio médico, lo cual garantiza la fiabilidad del conjunto de datos utilizado. Para facilitar este proceso, se desarrolló una aplicación que permitió a los expertos etiquetar de manera eficiente las relaciones entre entidades, contribuyendo así a la construcción de un conjunto de datos de alta calidad. Además, se propusieron y aplicaron siete patrones semánticos posibles entre las entidades involucradas, para la generación automática de tripletas a partir de los textos. Asimismo, se realizaron múltiples experimentos exploratorios para optimizar los hiperparámetros del LLM, y se llevó a cabo una validación adicional con un subconjunto semilla también revisado por especialistas, lo que refuerza la validez del enfoque.

En tercer lugar, se propuso una metodología general para la generación automática de grafos de conocimiento a partir de textos médicos en español. Esta metodología fue aplicada a un corpus compuesto por 990 artículos científicos, lo que permitió generar un grafo de conocimiento médico representativo y estructurado. Además, se evaluó el grafo mediante cinco casos de uso distintos, seleccionados por su variedad en cuanto a fuentes (artículos científicos, noticias, páginas médicas especializadas) y estilo de redacción. Un experto en el dominio médico verificó las relaciones extraídas en cada uno de los casos, confirmando la validez general del enfoque, aunque también se señalaron oportunidades de mejora del sistema.

En cuanto a las líneas futuras de trabajo, se identifican diversas áreas de mejora. En primer lugar, se plantea la necesidad de abordar las ambigüedades en el reconocimiento de entidades, especialmente en casos donde un término puede estar contenido en otro de distinta categoría, como ocurre con "pulmones" en "cáncer de pulmones". En segundo lugar, se recomienda enriquecer los conjuntos de datos utilizados, tanto para el reconocimiento de entidades como para la identificación de relaciones, a través de la colaboración directa con expertos del área médica. Esto permitiría incrementar la precisión del sistema y mejorar la cobertura semántica.

Otra línea de trabajo consiste en mejorar el reconocedor de entidades médicas en su capacidad de manejar sinónimos, mediante la incorporación de diccionarios médicos oficiales, lo cual evitaría la creación de nodos duplicados que representan una misma entidad con distintos nombres. Asimismo, se propone explorar nuevos enfoques basados en modelos de lenguaje para el desarrollo de NER médicos más robustos.

También se plantea la extensión del sistema hacia la incorporación de nuevos tipos de entidades y relaciones, con el objetivo de enriquecer aún más la estructura del grafo de conocimiento. Además, existe la necesidad de normalizar abreviaturas y siglas mediante un glosario especializado, lo que facilitaría la unificación de términos y el descubrimiento de nuevas relaciones implícitas entre entidades.

Finalmente, una propuesta prometedora consiste en utilizar el grafo de conocimiento resultante como entrada para el entrenamiento de Graph Neural Networks (GNN), con el fin de abordar tareas más complejas como la predicción de nuevas relaciones, la clasificación de entidades o la recomendación de tratamientos. Esta línea de investigación permitiría el desarrollo de aplicaciones avanzadas en áreas como el diagnóstico asistido por inteligencia artificial.

Bibliografía

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 6000–6010, Curran Associates Inc., 2017.

[2] *UNESCO - https://www.unesco.org/.*

[3] L. Navarrete and C. Pérez, "Revistas biomédicas: desarrollo y evolución," *Revista Médica Clínica Las Condes*, vol. 30, no. 3, pp. 219–225, 2019.

[4] D. Torres-Salinas, "Ritmo de crecimiento diario de la producción científica sobre covid-19. análisis en bases de datos y repositorios en acceso abierto," *El Profesional de la Información*, vol. 29, Apr. 2020.

[5] *Banco de datos mundial - https://datos.bancomundial.org/.*

[6] K. Jensen, "Natural language processing: The plnlp approach," *Kluwer Academic Publishers*, 1993.

[7] A. Cortez, H. Vega, and J. Pariona, "Procesamiento de lenguaje natural," *Revista Ingeniería de Sistemas e Informática*, vol. 6, no. 2, pp. 45–54, 2009.

[8] J. Cowie and W. Lehnert, "Information extraction," *Communications of the ACM*, vol. 39, pp. 80–91, Jan. 1996.

[9] A. Téllez, "Extracción de información con algoritmos de clasificación," Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica. Especialidad de Ciencias Computacionales, 2005.

[10] N. Oyarzún and P. Salas, "Construcción automática de diccionarios de patrones para sistemas de ei," *Sistemas de Extracción de Información*, pp. 1–10, 2014.

[11] S. Srawagi, "Information extraction," *NOW the essence of knowledge*, pp. 261–377, 2007.

[12] J. Llorens, S. Sinchez, and J. Morato, "Identificación automática de relaciones de jerarquia a partir de texto libre,"

[13] R. Jaime, "Aprendizaje activo para la extracción de relaciones en textos," tech. rep., 2019.

[14] R. Brachman and H. Levesque, *Knowledge representation and reasoning*. Morgan Kaufmann Publishers, 2004.

[15] G. Jakus, V. Milutinovic, S. Omerovic, and S. Tomazic, *Concepts, Ontologies, and Kowledge Representation*. Springer, 2013.

[16] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, and R. Navigli, *Knowledge graphs*. arXiv preprint arXiv:2003.02320, 2020.

[17] T. Saorín, *Grafos de conocimiento y bases de datos en grafo: conceptos fundamentales a partir de una obra maestra del Museo del Prado*. Anuario Think EPI, 2019.

[18] D. Fensel, U. Simsek, K. Angele, E. Huaman, E. Kärle, O. Panasiuk, I. Toma, J. Umbrich, and A. Wahler, *Knowledge Graphs*. Springer, 2020.

[19] *Sitio web Euskadi - https://www.euskadi.eus/contenidos/informacion/opendata_rdf_euskadi/es_info/adjuntos/RDF.pdf*.

[20] A. Moreno, E. Armengol, J. Béjar, L. Belanche, U. Cortés, R. Gavalda, J. M. Gimeno, B. López, M. Martín, and M. Sanchez, *Aprendizaje automático*. Edicions UPC, 1994.

[21] *Algoritmo de aprendizaje - https://www.ibm.com/mx-es/topics/*.

[22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[23] R. Clancy, I. F. Ilyas, and J. Lin, "Scalable knowledge graph construction from text collections," in *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, Association for Computational Linguistics, 2019.

[24] M. Wang, J. Zhang, J. Liu, W. Hu, S. Wang, X. Li, and W. Liu, "PDD graph: Bridging electronic medical records and biomedical knowledge graphs via entity linking," in *Lecture Notes in Computer Science*, pp. 219–227, Springer International Publishing, 2017.

[25] S. Domingos Cardoso, M. Da Silveira, and C. Pruski, "Construction and exploitation of an historical knowledge graph to deal with the evolution of ontologies," *Knowledge-Based Systems*, vol. 194, p. 105508, Apr. 2020.

[26] I. Muhammad, A. Kearney, C. Gamble, F. Coenen, and P. Williamson, "Open information extraction for knowledge graph construction," in *Communications in Computer and Information Science*, pp. 103–113, Springer International Publishing, 2020.

[27] L. Shi, S. Li, X. Yang, J. Qi, G. Pan, and B. Zhou, "Semantic health knowledge graph: Semantic integration of heterogeneous medical knowledge and services," *BioMed Research International*, vol. 2017, pp. 1–12, 2017.

[28] J. L. Martinez-Rodriguez, I. Lopez-Arevalo, and A. B. Rios-Alvarado, "OpenIE-based approach for knowledge graph construction from text," *Expert Systems with Applications*, vol. 113, pp. 339–355, Dec. 2018.

[29] L. Li, P. Wang, J. Yan, Y. Wang, S. Li, J. Jiang, Z. Sun, B. Tang, T.-H. Chang, S. Wang, and Y. Liu, "Real-world data medical knowledge graph: construction and applications," *Artificial Intelligence in Medicine*, vol. 103, p. 101817, Mar. 2020.

[30] W. Yu, S. Ding, Z. Yue, and S. Yang, "Emotion recognition from facial expressions and contactless heart rate using knowledge graph," in *2020 IEEE International Conference on Knowledge Graph (ICKG)*, IEEE, Aug. 2020.

[31] X. Jiang, Y. Shen, Y. Wang, X. Jin, and X. Cheng, "BaKGraSTeC: A background knowledge graph based method for short text classification,"

in *2020 IEEE International Conference on Knowledge Graph (ICKG)*, IEEE, Aug. 2020.

[32] Y. Wei, H. Wang, J. Zhao, Y. Liu, Y. Zhang, and B. Wu, "GeLaiGeLai: A visual platform for analysis of classical chinese poetry based on knowledge graph," in *2020 IEEE International Conference on Knowledge Graph (ICKG)*, IEEE, Aug. 2020.

[33] T. Wang and H. Li, "Coreference resolution improves educational knowledge graph construction," in *2020 IEEE International Conference on Knowledge Graph (ICKG)*, IEEE, Aug. 2020.

[34] J. Shao, G. Liu, and S. Ji, "An abnormal data analysis and processing method for genealogy graph databases," in *2020 IEEE International Conference on Knowledge Graph (ICKG)*, IEEE, Aug. 2020.

[35] X. Tang, Y. Huang, M. Xia, and C. Long, "A multi-task BERT-BiLSTM-AM-CRF strategy for chinese named entity recognition," *Neural Processing Letters*, vol. 55, pp. 1209–1229, July 2022.

[36] F. Dernoncourt, J. Y. Lee, and P. Szolovits, "Neuroner: an easy-to-use program for named-entity recognition based on neural networks," 2017.

[37] K. Gao, J. Zhou, Y. Chi, and Y. Wen, "Tourismner: A tourism named entity recognition method based on entity boundary joint prediction," *Intelligent Systems with Applications*, vol. 25, p. 200475, 2025.

[38] J. Bin, L. Rui, L. Shasha, Y. Jie, W. Qingbo, T. Yusong, and W. Jiaju, "A hybrid approach for named entity recognition in chinese electronic medical record," *BMC Medical Informatics and Decision Making*, vol. 19, Apr. 2019.

[39] W. Yoon, C. Ho So, J. Lee, and J. Kang, "CollaboNet: collaboration of deep neural networks for biomedical named entity recognition," *BMC Bioinformatics*, vol. 20, May 2019.

[40] L. Chen, Y. Gu, X. Ji, C. Lou, Z. Sun, H. Li, Y. Gao, and Y. Huang, "Clinical trial cohort selection based on multi-level rule-based natural language processing system," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1218–1226, 2019.

[41] X. Ying, C. Shuai, T. Buzhou, C. Qingcai, W. Xiaolong, Y. Jun, and Z. Yi, "Improving deep learning method for biomedical named entity recognition by using entity definition information," *BMC Bioinformatics*, vol. 22, Dec. 2021.

[42] J. Amengol-Estapé, F. Soares, M. Marimon, and M. Krallinger, "PharmacoNER tagger: a deep learning-based tool for automatically finding chemicals and drugs in spanish medical texts," *Genomics & Informatics*, vol. 17, p. e15, June 2019.

[43] F. J. Moreno-Barea, G. López-García, H. Mesa, N. Ribelles, E. Alba, J. M. Jerez, and F. J. Veredas, "Named entity recognition for deidentifying spanish electronic health records," *Computers in Biology and Medicine*, vol. 185, p. 109576, 2025.

[44] A. J. Tamayo Herrera, D. A. Burgos, and A. Gelbukh, "Clinical text mining in spanish enhanced by negationdetection and named entity recognition," *Computación y Sistemas*, vol. 27, no. 4, pp. 1169–1181, 2023.

[45] X. Zhang, P. Li, W. Jia, and H. Zhao, "Multi-labeled relation extraction with attentive capsule network," AAAI'19/IAAI'19/EAAI'19, AAAI Press, 2019.

[46] X. Li, Y. Li, J. Yang, H. Liu, and P. Hu, "A relation aware embedding mechanism for relation extraction," *Applied Intelligence*, pp. 1–10, 2022.

[47] H. Ridong, P. Tao, W. Benyou, L. Lu, T. Prayag, and W. Xiang, "Document-level relation extraction with relation correlations," *Neural Networks*, vol. 171, pp. 14–24, 2024.

[48] J. P. Torres, R. G. de Piñerez Reyes, and V. A. Bucheli, "Support vector machines for semantic relation extraction in spanish language," in *Advances in Computing* (J. E. Serrano C. and J. C. Martínez-Santos, eds.), (Cham), pp. 326–337, Springer International Publishing, 2018.

[49] Y. Cao, D. Chen, H. Li, and P. Luo, "Nested relation extraction with iterative neural network," CIKM '19, (New York, NY, USA), p. 1001–1010, Association for Computing Machinery, 2019.

[50] L. Yin, X. Meng, J. Li, and J. Sun, "Relation extraction for massive news texts.," *Computers, Materials & Continua*, vol. 60, no. 1, 2019.

[51] W. Kehan, Z. Xueying, D. Yulong, and Y. Peng, "Deep learning models for spatial relation extraction in text," *Geo-spatial Information Science*, vol. 26, no. 1, pp. 58–70, 2023.

[52] A. Revenko and P. Martín-Chozas, "Extraction and semantic representation of domain-specific relations in spanish labour law," *Procesamiento del Lenguaje Natural*, vol. 69, pp. 105–116, 2022.

[53] R. Patel, S. Tanwani, and C. Patidar, "Relation extraction between medical entities using deep learning approach," *Informatica*, vol. 45, no. 3, 2021.

[54] H. Ying, C. Yanping, H. Ruizhang, Q. Yongbin, and Z. Qinghua, "A hierarchical convolutional model for biomedical relation extraction," *Information Processing & Management*, vol. 61, no. 1, p. 103560, 2024.

[55] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artieres, A. Ngonga, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, and G. Paliouras, "An overview of the bioasq large-scale biomedical semantic indexing and question answering competition," *BMC Bioinformatics*, vol. 16, p. 138, 2015.

[56] *Descriptores en Ciencias de la Salud: DeCS [Internet]. ed. 2024. Sao Paulo (SP): BIREME / OPS / OMS. 2024 [actualizado 2024 Feb 08; citado 2023 Mayo 2]. Disponible en: https://decs.bvsalud.org/es/.*

[57] *Bioasq 2021 MESINESP - https://huggingface.co/datasets/bigbio/bioasq_2021_mesinesp.*

[58] *LILACS, Información Científica y Técnica en Salud de América Latina y el Caribe - https://lilacs.bvsalud.org/es/.*

[59] *MEDLINE, National Library of Medicine's - https://www.nlm.nih.gov/medline/medline$_h$ome.html.*

[60] J. Veiga de Cabo, "El indice bibliogrÃ¡fico espaÃde ciencias de la salud. CooperaciÃcon LatinoamÃ," *Revista EspaÃde Salud PÃ*, vol. 73, pp. 529 – 532, 09 1999.

[61] L. Campillos-Llanos, "First steps towards building a medical lexicon for spanish with linguistic and semantic information," in *Proc. of BioNLP 2019*, 2019.

[62] Z. S. H. and, "Distributional structure," *WORD*, vol. 10, no. 2-3, pp. 146–162, 1954.

[63] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *International Conference on Learning Representations*, 2013.

[64] G. A. García-Robledo, J. A. Reyes-Ortiz, M. Bravo, and A. D. Cuevas Rasgado, "MedRel_Spanish (versión 1)," 2025. Data set.

[65] J. Padilla Cuevas, J. A. Reyes-Ortiz, A. D. Cuevas-Rasgado, R. A. Mora-Gutiérrez, and M. Bravo, "Médicobert: A medical language model for spanish natural language processing tasks with a question-answering application using hyperparameter optimization," *Applied Sciences*, vol. 14, no. 16, 2024.

[66] M. Krallinger, J. Armengol-Estapé, O. De Gibert, C. P. Carrino, A. Gonzalez-Agirre, A. Gutiérrez-Fandiño, and M. Villegas, "Spanish biomedical crawled corpus," Feb. 2021.

[67] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[68] M. Barboza Hernández, "Enfermedad pulmonar obstructiva cronica," *Revista Medica Sinergia*, vol. 2, pp. 10–14, jun. 2017.

[69] El País, "La pastilla contra la hipertensión en adultos también ayuda a los niños con piel de mariposa," *El País*, May 2025. Accedido el 14 de mayo de 2025.

[70] M. González-Vicent, J. Sanz, J. L. Fuster, J. Cid, C. D. de Heredia, D. Morillo, J. M. Fernández, A. Pascual, I. Badell, D. Serrano, L. Fox, J. de la Serna, A. Benito, J. M. Couselo, B. Molina, M. Díaz, and M. Sanz, "Autoimmune hemolytic anemia (aiha) following allogeneic hematopoietic stem cell transplantation (hsct): A retrospective analysis and a proposal of treatment on behalf of the grupo español de trasplante

de medula osea en niños (getmon) and the grupo español de trasplante hematopoyetico (geth)," *Transfusion Medicine Reviews*, 2018. Advance online publication.

[71] R. I. G. Holt, M. de Groot, and S. H. Golden, "Diabetes and depression," *Current Diabetes Reports*, vol. 14, no. 6, p. 491, 2014. Published 18 April 2014.

[72] J. A. Gómez-Puerta and R. Cervera, "Lupus eritematoso sistémico," *Medicina & laboratorio*, vol. 14, no. 05-06, pp. 211–223, 2008.