



Universidad
Rey Juan Carlos

Tesis Doctoral

Sistema de asistencia de investigación
médica basado en conocimiento.

Autor:

Adrián Alonso Barriuso

Directores:

Alberto Fernández Isabel, Isaac Martín de Diego

Programa de Doctorado en Tecnologías de la Información
y las Comunicaciones

Escuela Internacional de Doctorado

Diciembre de 2023

A mi familia.

Agradecimientos

Mucho ha acontecido en los 4 años que ha durado esta tesis. Frase manida, frase con la que probablemente empiece la sección de agradecimientos de la mayoría de tesis doctorales, pero esta, no puede comenzar con otra. A lo largo de este viaje no sólo he aprendido, dudado, fracasado, a veces desesperado y otras tantas triunfado. A lo largo de este viaje me he casado, aprendido a vivir en una pandemia global, afrontado multitud de desafíos profesionales, y lo más importante, he sido padre. Quizá, esta última, sea la mayor de las motivaciones para crecer y ser la mejor versión de uno mismo, porque, qué puede haber más importante que ser el mejor ejemplo posible para un hijo. Mientras escribo estas líneas, mi hijo tiene 9 meses recién cumplidos y, si haces las cuentas, ha habido 39 meses (más o menos) con los que no he contado con semejante motivación. No habría llegado tan lejos, en serio, de verdad, sin el apoyo de las personas que voy a mencionar a continuación. Mis palabras, me temo, quedarán muy lejos de hacerles justicia, pero dedicarles reconocimiento es lo mínimo que puedo hacer.

En primer lugar, como no podría ser de otra forma, gracias a mis directores, Alberto e Isaac, por darme la oportunidad de realizar esta tesis y por brindarme las herramientas necesarias para llevarla a cabo. Gracias también a Juan Fernando, por hacer accesibles tantos conceptos médicos.

Mención especial merece Oleg, gracias por ser mi mayor mecenas, por tu confianza y por la libertad creativa que siempre me has dado. Espero poder agradecértelo con mi trabajo durante muchos años más.

Gracias también, por supuesto, al resto de mis compañeros en Dezzai, vuestro apoyo, sabiduría y espíritu de equipo componen una gran familia. En particular, quisiera agradecer a Alfonso, cuyo excelente trabajo está reflejado en parte de esta tesis. Y cómo no, gracias a mi amigo Nacho, de quién tengo la suerte de aprender en el ámbito profesional y en el personal.

Gracias a mi hermano, Alberto. Tu ética profesional y tu eterna voluntad de hacer

las cosas bien siempre han sido para mi la mayor de las inspiraciones. Eres el espejo en el que quiero verme reflejado.

Gracias Andrea por los sacrificios que has hecho para darme la oportunidad de hacer este trabajo. Tu constante apoyo y enorme paciencia siempre han bastado para hacerme salir adelante en los momentos más duros.

Y para acabar, como tampoco podría ser de otra manera, gracias a mis padres. De vosotros he aprendido el mantra que trato de repetirme cada día: *querer es poder*. A vosotros os debo todo lo que tengo y todo lo que soy.

Gracias a todos de corazón.

Resumen

El progreso de la medicina ha sido enorme a lo largo de los últimos tiempos, transitando desde la transmisión de conocimiento por métodos tradicionales como la observación y la experiencia, hasta el advenimiento de la medicina basada en la evidencia, apoyada en estudios clínicos rigurosos. La creciente cantidad de información médica planteó la necesidad de sistemas de gestión de conocimiento efectivos. Con la digitalización, la información médica se ha hecho más accesible, aunque el desafío actual es filtrar y acceder a datos relevantes rápidamente.

Esta tesis, fruto de la colaboración entre la empresa Dezzai, la Universidad Rey Juan Carlos (URJC) y el Instituto de Investigación Biomédica de Salamanca (IBSAL), busca enfrentar este desafío. La sinergia entre el ámbito académico y el empresarial se manifiesta en un proyecto que integra investigación y aplicación práctica, con el fin de ayudar a mantener a los profesionales médicos actualizados en sus respectivos campos en el menor tiempo y con la menor interacción posibles. La tesis introduce un asistente de investigación médica basado en conocimiento compuesto por tres componentes que trabajan de manera orquestada: Un estimador de relevancia de artículos médicos, un recomendador de artículos médicos y un asistente conversacional que permite a los usuarios interactuar con los artículos recomendados mediante el uso del lenguaje natural.

El estimador de relevancia de artículos médicos utiliza técnicas de aprendizaje automático, lexicones especializados y fuentes de conocimiento externas para filtrar artículos médicos relevantes, mientras que el recomendador de artículos médicos se centra en extraer diagnósticos de registros de salud electrónicos y encontrar artículos científicos relacionados, llegando a un compromiso entre relevancia y similitud semántica. El asistente conversacional mejora la interacción con los artículos recomendados, proporcionando respuestas basadas en la evidencia más reciente y relevante a la vez que ahorra la necesidad de interacción del usuario con sistemas complejos de recuperación de información. En conjunto, estos componentes ofrecen una solución integral para navegar por el extenso volumen de información médica, apoyando

tanto la investigación académica como la práctica clínica.

Para llevar a cabo esta propuesta, se ha seguido una metodología de trabajo basada en la aplicación de técnicas de ciencia de datos, donde la fase de experimentación y evaluación de cada componente ha contado con la implicación de expertos en el dominio de la medicina de IBSAL, asegurando la máxima rigurosidad en los resultados obtenidos.

Abstract

The progress of medicine has been tremendous in recent times, transitioning from the transmission of knowledge through traditional methods such as observation and experience to the advent of evidence-based medicine, underpinned by rigorous clinical studies. The burgeoning volume of medical information has necessitated effective knowledge management systems. With digitalization, medical information has become more accessible; however, the current challenge lies in swiftly filtering and accessing relevant data.

This dissertation, a collaborative effort between the company Dezzai, Universidad Rey Juan Carlos (URJC), and Instituto de Investigación Biomédica de Salamanca (IB-SAL), aims to tackle this challenge. The synergy between academia and the business sector is evidenced in a project that merges research and practical application, aimed at keeping medical professionals up-to-date in their respective fields with minimal time and interaction. The thesis introduces a knowledge-based medical research assistant composed of three orchestrated components: a medical article relevance estimator, a medical article recommender, and a conversational assistant that allows users to interact with recommended articles using natural language.

The medical article relevance estimator employs machine learning techniques, specialized lexicons and external knowledge sources to filter relevant medical articles, while the medical article recommender focuses on extracting diagnoses from electronic health records and finding related scientific articles, striking a balance between relevance and semantic similarity. The conversational assistant enhances interaction with recommended articles, providing answers based on the most recent and pertinent evidence, simultaneously saving the user's need to engage with complex information retrieval systems. Collectively, these components offer a comprehensive solution for navigating the vast volume of medical information, supporting both academic research and clinical practice.

To realize this proposal, a work methodology based on the application of data science techniques has been followed, where the experimentation and evaluation

phase of each component involved the participation of medical domain experts from IBSAL, ensuring the utmost rigor in the results obtained.

Índice general

Agradecimientos	I
Resumen	III
Abstract	V
Índice general	IX
Índice de figurasXIII
Índice de tablasXVI
Listado de acrónimosXVII
1. Introducción	1
1.1. Contexto	1
1.2. Motivación.	3
1.3. Objetivos.	4
1.4. Estructura	5
2. Estado del arte	9
2.1. Búsqueda y recuperación de información	9
2.1.1. Relevancia de documentos	11
2.1.2. Conclusiones.	14
2.2. Técnicas de generación de diccionarios	14
2.2.1. Lexicones de análisis de sentimientos	16
2.2.2. Conclusiones.	17

2.3. Inteligencia artificial en medicina	18
2.3.1. Procesamiento automático de diagnósticos.	20
2.3.2. Conclusiones.	23
2.4. Procesamiento del lenguaje natural	23
2.4.1. Asistentes conversacionales	26
2.4.2. Conclusiones.	28
3. Propuesta general.	31
3.1. Arquitectura.	31
3.2. Metodología de trabajo	34
3.3. Distribución temporal	36
4. Estimador de relevancia de artículos	39
4.1. Introducción	39
4.2. Propuesta de marco de trabajo	41
4.2.1. Flujo del marco de trabajo.	43
4.2.2. Cálculo de relevancia.	45
4.2.3. Concepto de diseño del lexicón	47
4.2.4. Desarrollo del modelo de aprendizaje máquina	50
4.3. Experimentos	51
4.3.1. Parámetros neutrales.	53
4.3.2. Ajuste fino de parámetros	56
4.3.3. Validación a lo largo del tiempo	58
4.4. Conclusiones	61
5. Recomendador de artículos	63
5.1. Introducción	63
5.2. Propuesta de marco de trabajo	65
5.2.1. Módulo pre-procesador de artículos	66
5.2.2. Módulo extractor de diagnósticos	67
5.2.3. Módulo calculador de similitud semántica	70
5.2.4. Módulo calculador de relevancia	71

5.2.5. Módulo recomendador de artículos	72
5.3. Experimentos	73
5.3.1. Evaluación del módulo extractor de diagnósticos	73
5.3.2. Evaluación del módulo calculador de similitud semántica	75
5.3.3. Evaluación del sistema en un entorno real	78
5.4. Conclusiones	81
6. Asistente generativo contextualizado	83
6.1. Introducción	83
6.2. Propuesta	85
6.2.1. Flujo de funcionamiento del asistente	86
6.2.2. Módulo de interacción con el LLM	86
6.3. Experimentos	88
6.3.1. Selección de parámetros.	89
6.3.2. Batería de consultas pre-definidas	90
6.3.3. Evaluación con GPT-4	94
6.3.4. Batería de consultas a posteriori.	99
6.4. Conclusiones	99
7. Conclusiones	103
7.1. Contribuciones	103
7.2. Trabajo futuro.	106
Referencias	109

Índice de figuras

2.1. Arquitectura de SBERT, una modificación de BERT que utiliza redes siamesas para obtener representaciones semánticas vectoriales de frases (Reimers y Gurevych, 2019).	10
2.2. Ejemplo de búsqueda en PubMed. PubMed es un servicio gratuito de la Biblioteca Nacional de Medicina de los Estados Unidos que ofrece acceso a investigaciones biomédicas. Cuenta con opciones avanzadas de búsqueda y la capacidad de crear alertas. A la izquierda, los usuarios pueden filtrar los resultados por año, disponibilidad de texto, atributo del artículo y fecha de publicación. Los resultados centrales presentan el título del estudio, autores, fuente, fecha, un breve resumen y un identificador único. También hay opciones para citar, compartir o guardar cada artículo. La navegación adicional permite cambiar el orden de los resultados y ajustar cómo se muestran.	13
2.3. Relaciones semánticas en WordNet (Miller, 1995).	15
2.4. Representación gráfica adoptada por SentiWordNet para representar las propiedades relacionadas con la opinión de un sentido del término en consideración. (Sebastiani y Esuli, 2006).	17
2.5. Ejemplo de red bayesiana ficticia para la predicción de cáncer o tuberculosis (Arora <i>et al.</i> , 2019).	18

2.6. Marco de trabajo basado en la visión artificial para la predicción de nódulos pulmonares (Saba, 2020). El uso de técnicas de aprendizaje automático ayuda en el diagnóstico temprano y la evaluación de nódulos pulmonares al procesar imágenes de tomografía computarizada construidas mediante métodos de inteligencia artificial. Estos sistemas se denominan sistemas de apoyo a la decisión que investigan las imágenes a través de pre-procesamiento, segmentación, extracción de características y un proceso de clasificación.	21
2.7. Arquitectura <i>Transformer</i> original (Vaswani <i>et al.</i> , 2017).	26
2.8. Línea temporal de los <i>Large Language Model</i> (LLM) con mas de diez mil millones de parámetros surgidos desde 2019 (Zhao <i>et al.</i> , 2023). . .	27
2.9. Gráfico de la evolución de la investigación surgida desde la liberación de <i>Large Language Model Meta AI</i> (LLaMA) (Zhao <i>et al.</i> , 2023).	28
3.1. Arquitectura simplificada del asistente de investigación médica basado en conocimiento.	32
3.2. Prototipo de la interfaz de usuario del asistente. En la sección "Relevant Documents" se recogen los N (3 visibles en este ejemplo) artículos recomendados para un paciente cuyo diagnóstico es "COPD exacerbation". Se representa el título, parte del <i>abstract</i> y la URL de la fuente original de cada artículo. El usuario ha introducido una consulta en la sección "Query to the assistant" y el asistente genera la respuesta enlazando con la fuente que la fundamenta en la sección "Output".	33
3.3. Mapa conceptual de la ciencia de datos (Cao, 2017).	34
3.4. Distribución temporal de las principales tareas de la tesis.	37
4.1. Representación general de la arquitectura de <i>Webelance</i> (módulos, componentes y artefactos).	42
4.2. Proceso seguido por <i>Webelance</i> para calcular la relevancia de un artículo científico.	44
4.3. Vista general de la arquitectura del generador de lexicones.	48
4.4. Representación de las cuatro distribuciones bi-modales de la relevancia de artículos para cada año.	57
5.1. Vista general de la arquitectura del marco de trabajo <i>Medical Expert Linguist for Evaluating Nosology and Diagnosis Information</i> (MELENDI).	66
5.2. Arquitectura interna del módulo <i>Pre-procesador de artículos</i>	67

5.3. Proceso del módulo <i>Extractor de diagnósticos</i>	68
5.4. Arquitectura interna del módulo <i>Extractor de diagnósticos</i>	69
5.5. Arquitectura interna del módulo <i>Calculador de similitud semántica</i> . . .	70
5.6. Arquitectura del módulo <i>Calculador de relevancia</i> (para más detalle ver el Capítulo 4).	72
5.7. Arquitectura interna del módulo <i>Recomendador de artículos</i>	73
5.8. Representación de artículos distribuidos según su relevancia y similitud coseno con respecto al diagnóstico <i>COPD exacerbation</i>	79
5.9. Representación de artículos distribuidos según su relevancia y similitud coseno con respecto al diagnóstico <i>Community acquired pneumonia, Alcohol dependence, Acute on chronic renal failure</i>	80
5.10. Representación de artículos distribuidos según su relevancia y similitud coseno con respecto al diagnóstico <i>Amyotrophic Lateral Sclerosis</i> . .	80
6.1. Representación general de la arquitectura final del sistema (módulos, componentes y artefactos).	85
6.2. Flujo de funcionamiento del sistema completo.	87

Índice de tablas

4.1. Capas de la red neuronal convolucional.	51
4.2. Valores de saturación de los parámetros de las Ecuaciones 4.3 y 4.5. . .	53
4.3. Comparación de la opinión de los expertos con la relevancia predicha por <i>Webelance</i> para artículos publicados durante 2015.	54
4.4. Comparación de la opinión de los expertos con la relevancia predicha por <i>Webelance</i> para artículos publicados durante 2016.	55
4.5. Comparación de la opinión de los expertos con la relevancia predicha por <i>Webelance</i> para artículos publicados durante 2017.	55
4.6. Comparación de la opinión de los expertos con la relevancia predicha por <i>Webelance</i> para artículos publicados durante 2018.	55
4.7. Comparación de la opinión de expertos con la relevancia predicha por el marco <i>Webelance</i> para artículos publicados durante 2015.	58
4.8. Comparación de la opinión de expertos con la relevancia predicha por el marco <i>Webelance</i> para artículos publicados durante 2016.	59
4.9. Comparación de la opinión de expertos con la relevancia predicha por el marco <i>Webelance</i> para artículos publicados durante 2017.	59
4.10. Comparación de la opinión de expertos con la relevancia predicha por el marco <i>Webelance</i> para artículos publicados durante 2018.	59
4.11. Relevancia a lo largo del tiempo de artículos sobre el tratamiento de la alopecia androgénica.	60
5.1. Comparación de entidades y enfermedades detectadas en 500 <i>Electronic Health Records</i> (EHR) entre MedCAT, SciSpacy, BioBERT y el método de coincidencia exacta.	75

5.2. Diagnósticos aislados provenientes de diez EHR usados para evaluar el sistema al completo.	77
5.3. Comparativa entre la opinión de los expertos y los resultados proporcionados por el producto escalar y la similitud coseno.	77
5.4. Comparación entre las evaluaciones de los expertos y las recomendaciones realizadas por MELENDI. Cada fila muestra los tres artículos seleccionados por los expertos para cada diagnóstico junto con la posición del artículo en el ranking propuesto por el sistema (valor de similitud coseno).	81
6.1. Parámetros relevantes de <i>Generative Pre-trained Transformer</i> (GPT)-3.5-turbo y valores seleccionados.	89
6.2. Diagnósticos utilizados para generar las recomendaciones para evaluar el asistente.	91
6.3. Batería de consultas generalistas para la evaluación del asistente. . . .	91
6.4. Resultados de evaluación de las respuestas generalistas realizada por los expertos.	92
6.5. Consultas de control diseñadas después de conocer los abstracts recomendados para cada diagnóstico.	101

Listado de acrónimos

API *Application Programming Interface.*

BART *Bidirectional and Auto-Regressive Transformers.*

BERT *Bidirectional Encoder Representations from Transformers.*

CNN *Convolutional Neural Network.*

DOI *Digital Object Identifier.*

EHR *Electronic Health Records.*

GAMER *Generative Assistant for Medical Expert Requests.*

GPT *Generative Pre-trained Transformer.*

HIS *Hospital Information System.*

IA *Inteligencia Artificial.*

IBSAL *Instituto de Investigación Biomédica de Salamanca.*

IR *Information Retrieval.*

LLaMA *Large Language Model Meta AI.*

LLM *Large Language Model.*

LSTM *Long-Short Term Memory.*

MELENDI *Medical Expert Linguist for Evaluating Nosology and Diagnosis Information.*

MIMIC-III *Medical Information Mart for Intensive Care.*

ML *Machine Learning.*

NER *Named Entity Recognition.*

PaLM *Pathways Language Model.*

PLN *Procesamiento del lenguaje natural.*

PMC *Pubmed Central.*

RNN *Recurrent Neural Networks.*

RoBERTa *Robustly Optimized BERT Pretraining Approach.*

SABERMED *Swarm Agent-Based Environment For Reputation in MEDicine.*

SBERT *Sentence Bidirectional Encoder Representations from Transformers.*

T5 *Text to Text Transfer Transformer.*

TF-IDF *Term Frequency-Inverse Document Frequency.*

TF-IGM *Term Frequency Inverse Gravity Moment.*

UMLS *Unified Medical Language System.*

URJC *Universidad Rey Juan Carlos.*

Capítulo 1

Introducción

En este capítulo se realiza una introducción al trabajo realizado en esta tesis. Se comienza con una descripción del contexto en el que se desarrolla, dentro del ámbito de un doctorado industrial. Después, se describen las motivaciones que han dado lugar a la propuesta presentada y a la colaboración entre la universidad y el entorno empresarial. A continuación, se describen sus objetivos y, finalmente, se hace una descripción de su estructura.

1.1. Contexto

La historia de la medicina está llena de avances que han transformado la forma en que se practica la atención médica, muchos de los cuales han surgido gracias a la investigación rigurosa y metódica. A lo largo de los años, la metodología de investigación médica ha evolucionado para adaptarse a las necesidades cambiantes de la sociedad y para aprovechar los avances tecnológicos.

Durante siglos, el conocimiento médico se basó en la tradición oral y la observación directa. Los médicos confiaban en su experiencia personal y en la sabiduría transmitida de generación en generación. Con el tiempo, surgió la necesidad de un enfoque más sistemático y riguroso. En la era moderna, con la aparición de la medicina basada en la evidencia, los estudios clínicos y experimentales se convirtieron en el estándar de oro para determinar la eficacia y la seguridad de las intervenciones médicas.

A medida que el volumen de información médica aumentaba, surgía un nuevo desafío: gestionar, categorizar y acceder a este inmenso corpus de conocimiento. La digitalización de registros médicos y publicaciones permitió el almacenamiento de información en bases de datos electrónicas, pero acceder a la información relevante

y actualizada se convirtió en un reto.

La proliferación de documentos y publicaciones en distintos campos del conocimiento, especialmente en la ciencia médica, plantea desafíos significativos para la comunidad académica y los profesionales sanitarios. Esta tesis nace con el objetivo de proporcionar una solución a la creciente necesidad de acceder a información precisa, relevante y actualizada, permitiendo a los médicos mantenerse actualizados en sus respectivos campos en el menor tiempo posible. Este trabajo es un esfuerzo colaborativo entre la Universidad Rey Juan Carlos (URJC)¹ y la empresa Dezzai², patrocinado por una beca de doctorado industrial proveniente de las ayudas para Doctorados Industriales en la Comunidad de Madrid —Orden 1921/2019 con número de referencia IND2019/TIC-17169. Estas ayudas tienen como objetivo principal la construcción de puentes entre el entorno académico y las empresas, puentes que permitan construir sinergias entre ambas entidades. En el ámbito empresarial, la alta competitividad, especialmente en el sector tecnológico, y las apremiantes fechas de productificación motivadas por dicha competencia, dan lugar a la implementación de soluciones bien conocidas en el menor tiempo posible, dificultando el avance tecnológico. Dicho avance, por tanto, queda relegado casi exclusivamente al ámbito académico, cuyas novedosas contribuciones raramente son implementadas en el entorno empresarial. Así, iniciativas como las ayudas para Doctorados industriales, cobran una importancia capital en el desarrollo industrial tecnológico en España. Gracias a esta colaboración, la empresa Dezzai puede obtener una ventaja competitiva en la medida en que las soluciones propuestas en el presente trabajo se encuentran ya implementadas en varios de sus productos. Por otra parte, la novedad propuesta tiene valor en sí misma como contribución científica en sí, además de el impacto que puede tener a la hora de acelerar las investigaciones médicas de sus potenciales usuarios, contribuyendo positivamente en el trato a los pacientes y, por ende, en toda la sociedad.

El proyecto propone un asistente integral de investigación médica que comprende tres componentes clave: un sistema para estimar la relevancia de artículos médicos, un recomendador de artículos basado en dicha relevancia y similitud semántica, y finalmente, un asistente de conversación desarrollado sobre la arquitectura de un *Large Language Model* (LLM), que utiliza los artículos recomendados como contexto para responder consultas del usuario en lenguaje natural.

El primer componente, denominado *Webelance*, se centra en abordar el problema del filtrado de información en la comunidad científica. Frente al crecimiento exponencial de publicaciones, la necesidad de sistemas que ayuden a medir la relevancia de los textos en dominios específicos se hace cada vez más evidente. *Webelance* utili-

¹<https://dezzai.com/es/>

²<https://www.urjc.es/>

za un léxico basado en relevancia, modelos de aprendizaje automático y métricas de reputación para calcular la relevancia de artículos científicos en el ámbito médico. Los experimentos realizados demuestran la eficacia del sistema en diversos escenarios y ajustes de parámetros.

El segundo pilar de esta tesis, el sistema *Medical Expert Linguist for Evaluating Nosology and Diagnosis Information* (MELENDI), se enfoca de nuevo en el ámbito clínico y ofrece a los profesionales de la salud un medio para mantenerse actualizados en sus respectivos campos. Este recomendador extrae diagnósticos de Registros Electrónicos de Salud *Electronic Health Records* (EHR), busca artículos relevantes relacionados con esos diagnósticos y los recomienda de forma pro-activa al usuario. Su arquitectura modular aborda tareas específicas como el cálculo de similitudes entre artículos y la generación de un ranking de resultados. MELENDI ha sido evaluado en un escenario del mundo real, mostrando su viabilidad y utilidad en la práctica clínica.

Finalmente, el tercer componente, bautizado *Generative Assistant for Medical Expert Requests* (GAMER), utiliza un LLM para permitir que los médicos interactúen con los artículos recomendados por MELENDI a través de lenguaje natural. Al inyectar estos artículos como contexto en el LLM, los profesionales de la salud pueden realizar preguntas específicas y recibir respuestas fundamentadas en las publicaciones más recientes y relevantes en su ámbito.

Esta tesis ofrece un enfoque holístico para abordar los desafíos de acceder e interactuar con la información científica, tanto para la comunidad académica como para los profesionales de la salud. Los tres componentes trabajan de manera sinérgica para ofrecer una solución integral que facilita la tarea de mantenerse actualizado en el vertiginoso mundo de la ciencia médica.

1.2. Motivación

Al desarrollarse íntegramente en un entorno empresarial, la tesis tiene un enfoque eminentemente práctico, ajustándose a las necesidades de negocio de Dezzai. Dezzai es una *startup* tecnológica fundada en 2015 en Las Rozas, Madrid. Está especializada en inteligencia artificial y más concretamente en las áreas de Procesamiento del lenguaje natural (PLN) y visión artificial. Nació como Medlab Media Group, con la motivación de utilizar Inteligencia Artificial (IA) para proporcionar soluciones innovadoras a problemas relacionados principalmente con el dominio de la medicina, contando siempre con un asistente médico dentro de sus objetivos estratégicos. Más adelante, evolucionó en Dezzai y actualmente cuenta con clientes no sólo del dominio de la medicina sino también en el dominio legal, financiero y en el ámbito de la prevención de riesgos laborales. Además de consultoría tecnológica, Dezzai cuenta

con varios productos de software basados en IA, creados por un equipo de científicos de datos, ingenieros y lingüistas computacionales. En el año 2017, fruto del acercamiento entre Dezzai y miembros del laboratorio de ciencia de datos (DSLAB)³ de la Universidad Rey Juan Carlos y el Instituto de Investigación Biomédica de Salamanca (IBSAL), nace el proyecto *Retos Swarm Agent-Based Environment For Reputation in MEDicine* (SABERMED), con el objetivo de detectar páginas web potencialmente fraudulentas sobre contenidos médicos. A lo largo de esta colaboración, se establecieron sinergias entre las tres entidades más allá del desarrollo de SABERMED, dando lugar al presente trabajo. Cabe destacar el crucial rol de los expertos de IBSAL en calidad de expertos en el dominio médico de cara a la evaluación de los experimentos y las propuestas presentadas en la tesis.

Esta investigación surge de la necesidad identificada por Dezzai en las conversaciones con sus clientes. Dada la saturación de los centros médicos (R. Johnson *et al.*, 2018) y la cada vez mayor publicación de investigaciones científicas (Olivas-Martínez *et al.*, 2021), a los médicos les resulta cada vez más complicado mantenerse actualizados sobre los últimos avances científicos en sus especialidades. A esto hay que sumar el conocimiento que requieren los indexadores y buscadores científicos disponibles o sus limitaciones técnicas, como algoritmos desfasados orientados a encontrar coincidencias de palabras clave. Por estos motivos, la creación de un sistema de asistencia a la investigación médica completo, que sea capaz de ofrecer información relevante y actualizada sobre las características de los pacientes, con una mínima interacción y conocimiento técnico necesarios por parte del usuario, resulta de gran importancia.

1.3. Objetivos

El objetivo principal de la tesis es la creación de un asistente de investigación médica basado en conocimiento que permita a los médicos acelerar el proceso de actualización en sus respectivos campos. El sistema deberá ser capaz de proporcionar artículos científicos relevantes y actualizados relacionados con las enfermedades diagnosticadas a los pacientes del usuario. Estos artículos deberán proceder de fuentes de confianza y ser consolidados en una base de datos de conocimiento. Finalmente, el usuario deberá poder realizar consultas en lenguaje natural sobre el contenido de los artículos recomendados. Estas consultas le permitirán extraer conocimiento sobre los últimos avances relacionados con las características específicas de sus pacientes, sin ningún conocimiento técnico necesario.

Para lograr este objetivo, se han propuesto los siguientes sub-objetivos:

- **Estudio y documentación sobre el estado del arte.** Se realiza una revisión con-

³<https://www.datasciencelab.es/>

tinua de los últimos avances relacionados con los temas fundacionales de la tesis. La inteligencia artificial es una disciplina en auge y en los últimos tiempos se ha producido una explosión de nuevas aproximaciones, especialmente con el nacimiento de nuevas arquitecturas de aprendizaje profundo. Una de las arquitecturas más relevantes que han surgido en los últimos años son los *transformers*, que han dado lugar a enormes avances en el campo del procesamiento del lenguaje natural. El rápido avance en este campo ha obligado a replantearse algunas decisiones de planificación a lo largo de la tesis.

- **Creación de una métrica de relevancia de artículos científicos.** La base del asistente es contar con artículos médicos relevantes y actualizados, por lo que la creación de una métrica de relevancia lo más objetiva posible es un pilar fundamental. Se debe diseñar un marco de trabajo que permita estimar la relevancia de artículos científicos utilizando técnicas de aprendizaje máquina y métricas de reputación. Para ello se debe seleccionar y consolidar un corpus de documentos que permita entrenar y evaluar el sistema. El marco de trabajo se debe ajustar y evaluar sobre el dominio científico-médico.
- **Diseño de un sistema de recomendación de artículos científicos.** Una vez desarrollada la métrica de relevancia, esta se debe combinar con un sistema de recuperación de información semántico para obtener artículos relevantes y relacionados con las enfermedades de los pacientes. Para lograrlo, se debe desarrollar un marco de trabajo capaz de extraer los diagnósticos procedentes de EHR y realizar recomendaciones combinando diagnóstico, similitud semántica y relevancia. Se deberá obtener y consolidar un corpus de EHR para llegar a este fin.
- **Diseño de un asistente basado en lenguaje natural.** Como paso final, el sistema de recomendación se debe integrar con un LLM con el fin realizar consultas en lenguaje natural sobre las recomendaciones de artículos realizadas. Las respuestas del asistente deberán estar basadas exclusivamente en el contexto proporcionado por las recomendaciones y los artículos en los que se basan las respuestas deberán ser citados en cada caso.

1.4. Estructura

La estructura del documento se divide en siete capítulos, cuatro de los cuáles (Capítulos 3, 4, 5 y 6) son nucleares y representan los componentes principales de la propuesta de la tesis. En este capítulo se ha hecho una introducción al contexto en el que se desarrolla la tesis, así como sus motivaciones y objetivos. En el Capítulo 2 se hace una revisión de la literatura científica relacionada. Finalmente, el Capítulo 7 finaliza

la tesis con un análisis de los aportaciones conseguidas y posibles mejoras futuras. Después del Capítulo 7 hay una sección de referencias donde se puede encontrar la bibliografía utilizada.

- **Capítulo 2. Estado del arte.** En este capítulo se realiza una revisión bibliográfica sobre todos los ámbitos que componen los fundamentos teóricos de la investigación. Se comienza con una revisión de los enfoques de recuperación de información y del concepto de relevancia. Después, se abordan las técnicas de generación de diccionarios en general y los lexicones de sentimiento en específico. A continuación, se repasa el papel y los avances de la inteligencia artificial en medicina, profundizando en la procesamiento automático de diagnósticos. Finalmente, se revisan las principales técnicas de procesamiento del lenguaje natural y la evolución de los modelos conversacionales.
- **Capítulo 3. Propuesta general.** Este capítulo introduce la arquitectura de la propuesta general de la tesis con sus tres componentes principales, después se hace una descripción de la metodología empleada y se presenta una distribución temporal de las principales tareas en las que se ha segmentado la tesis.
- **Capítulo 4. Estimador de relevancia de artículos.** En este capítulo se presenta el marco de trabajo *Webelance*. *Webelance* es una propuesta de marco de trabajo para la estimación de relevancia de artículos científicos. Está basado en un léxico de relevancia, modelos de aprendizaje máquina y métricas de reputación. El marco de trabajo se ha ajustado al dominio científico-médico. El capítulo cuenta con una sección de propuesta, donde se describen los aspectos técnicos del marco de trabajo y las decisiones de diseño. Una sección de experimentos, donde se valida la propuesta en el dominio de la medicina utilizando más de 50.000 artículos de investigación biomédica. Y finalmente, una sección de conclusiones, donde se analizan los resultados de este componente.
- **Capítulo 5. Recomendador de artículos médicos.** Este capítulo describe el segundo componente principal de la tesis, MELENDI. Consiste en un marco de trabajo que permite hacer recomendaciones de artículos médicos en base a diagnósticos extraídos de EHRs. Este componente se basa en la hibridación de modelos de similitud semántica con las métricas de relevancia estimadas por *Webelance*. El componente cuenta además con un módulo de detección automática de diagnóstico aplicado a informes de alta provenientes de *Medical Information Mart for Intensive Care* (MIMIC-III). Gracias a este módulo, el sistema recomienda automáticamente artículos relevantes basados en los diagnósticos, sin la necesidad de la intervención activa del usuario. El capítulo, de forma análoga al anterior, cuenta con una sección de propuesta, experimentos y

conclusiones. Para realizar los experimentos, se han utilizado más de 1.000.000 de artículos médicos.

- **Capítulo 6. Asistente generativo contextualizado.** Aquí se presenta GAMER, el tercer y último componente principal del sistema. Consiste en un asistente basado en lenguaje natural que utiliza las recomendaciones proporcionadas por MELENDI para inyectar contexto a un LLM y restringir sus respuestas a dicho contexto. De esta forma, el usuario puede realizar consultas y recibir respuestas basadas en artículos relevantes y de confianza, evitando depender del conocimiento previo del LLM y posibles alucinaciones. El capítulo cuenta con la misma estructura que los dos anteriores. Para validar la propuesta se han realizado tres baterías de experimentos donde un equipo de expertos en el dominio médico ha analizado una serie de respuestas proporcionadas por el asistente.
- **Capítulo 7. Conclusiones.** En este capítulo se hace una reflexión sobre las aportaciones conseguidas en la tesis, limitaciones y posibles trabajos futuros. También se proporcionan las publicaciones derivadas de la investigación.

Capítulo 2

Estado del arte

En este capítulo se introducen los fundamentos teóricos empleados para la construcción del presente trabajo. En primer lugar se abordan los enfoques de *Information Retrieval* (IR) adentrándose en modelos semánticos que proporcionan similitud de documentos. Estos modelos se revisan considerando propósitos generales y, más tarde, específicamente en el ámbito sanitario. Después, se profundiza en una parte fundamental de los sistemas de IR, la relevancia de documentos. Se ofrece una visión general del concepto de relevancia, tanto aquello relacionado con su definición formal como revisando su evolución en el campo de la investigación durante los últimos años. En segundo lugar, se introducen algunas técnicas para la generación automática de diccionarios. Por otra parte, dado que el análisis de relevancia de artículos presentado en este trabajo está inspirado en el clásico enfoque del Análisis de Sentimientos (de Diego *et al.*, 2018), se hace también una revisión de la literatura relacionada con el tema. En tercer lugar, se aborda una revisión exhaustiva sobre la IA en medicina. Allí se consideran los enfoques y temas más típicos en el área de atención médica, después, se revisan y discuten enfoques más específicos centrados en la tarea de procesamiento automático de diagnósticos. Finalmente, se hace una revisión de las tareas más importantes del PLN y se recoge la evolución de las aproximaciones a los asistentes conversacionales de los últimos años, desde los clásicos *chatbots* hasta los modelos generativos más avanzados de la actualidad.

2.1. Búsqueda y recuperación de información

IR es uno de los procesos más importantes en informática (Chowdhury, 2010). Consiste en recuperar la información de interés a partir de un conjunto de documentos (es decir, un corpus). Estos documentos suelen contener datos textuales, que son

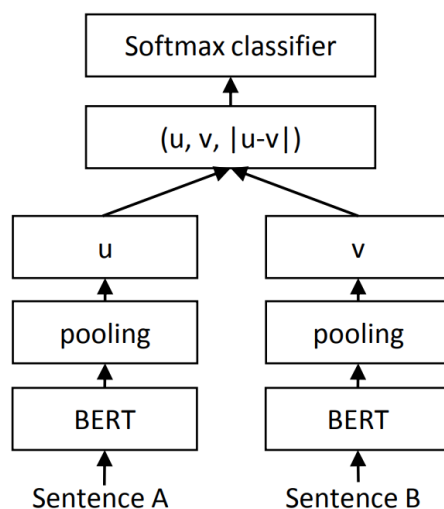


Figura 2.1: *Arquitectura de SBERT, una modificación de BERT que utiliza redes siamesas para obtener representaciones semánticas vectoriales de frases (Reimers y Gurevych, 2019).*

información no estructurada o semi-estructurada, pero también pueden ser información estructurada como vídeos e imágenes.

En el caso del contenido textual, las tareas de IR pueden organizarse en enfoques de procesamiento de texto y enfoques de no procesamiento de texto (Mala y Lobiyal, 2016). En el primer caso, la extracción de palabras clave (palabras relevantes) del texto es la operación más común. Este conjunto de palabras se estandariza transformando las palabras a su raíz léxica (es decir, reducción al lema o a la raíz). Esta tarea permite al computador encontrar palabras similares en los distintos textos y establecer similitudes de acuerdo con las coincidencias. Trabajos típicos son aquellos que utilizan algoritmos como *Term Frequency-Inverse Document Frequency* (TF-IDF), similitud coseno, semántica latente y combinaciones de ellos (Passalis y Tefas, 2018). En el segundo caso, se considera toda la estructura del contenido textual. Así, los algoritmos basados en semántica distribucional son los más típicos en este contexto. Estos algoritmos se centran en vectorizar palabras, lo que permite organizar el contenido textual en varias dimensiones según la co-ocurrencia de palabras. Enfoques típicos son aquellos que implementan redes neuronales recurrentes y modelos basados en *Transformers*. Ejemplos de ellos son las redes neuronales del tipo *Long-Short Term Memory* (LSTM) y las arquitecturas basadas en *Bidirectional Encoder Representations from Transformers* (BERT) (Jiang *et al.*, s.f.). Un ejemplo de las últimas sería *Sentence Bidirectional Encoder Representations from Transformers* (SBERT), que utiliza la similitud del coseno para calcular la similitud semántica entre dos cadenas de texto (ver Figura 2.1.)

Los sistemas de recomendación son una de las arquitecturas de software más típi-

cas que hacen uso de IR. Estos sistemas suelen proporcionar recomendaciones según un conjunto de palabras clave proporcionadas por los usuarios o elaboran un perfil siguiendo sus preferencias previamente expresadas.

En lo que respecta al ámbito sanitario, las técnicas de IR se incluyen en sistemas de recomendación relacionados con documentos que extraen similitudes entre síntomas, pacientes y enfermedades (Stark *et al.*, 2019). Por lo tanto, es un asunto clave para el experto en salud encontrar posibles soluciones ya documentadas para un problema específico. Así, el desarrollo de estos sistemas enfocados en el apoyo a la toma de decisiones ha sido una mejora importante (De Croon *et al.*, 2021).

Los enfoques típicos que abordan este hecho presentan dos perspectivas diferentes: pacientes y médicos. En el primer caso, los pacientes utilizan un sistema de recomendación para seleccionar el centro médico o médico más interesante. Ejemplos de estos trabajos son (Martinez *et al.*, 2014) y (Waqar *et al.*, 2019). En el segundo caso, los expertos en atención sanitaria utilizan recomendaciones proporcionadas por estos sistemas para encontrar nuevos medicamentos, tratamientos médicos y procedimientos. Ejemplos de estos enfoques son: (Katzman *et al.*, 2018) y (Q. Zhang *et al.*, 2015).

2.1.1. Relevancia de documentos

La relevancia en documentos representa un campo extendido en los sistemas de IR (MacDonald y Drew, 2018). Estos enfoques se centran en organizar los documentos (generalmente textos) según características específicas extraídas de ellos (Penha *et al.*, 2019).

La relevancia se define como el valor con el cual algo está relacionado o que muestra utilidad con respecto a lo que está ocurriendo o se está discutiendo (H. D. White, 2017). Generalmente, la relevancia de un artículo científico se calcula en función del factor de impacto de la revista relacionada y el número de citas relacionadas (ver, por ejemplo, (Bloching y Heinzl, 2013)).

La relevancia puede fluctuar con el tiempo produciendo tendencias. Estas tendencias ilustran la manera en que algo se vuelve irrelevante (es decir, olvidado) o relevante (es decir, en tendencia). En el caso de la comunidad científica, esto denota que un estudio o un descubrimiento podría volverse obsoleto o altamente importante para desarrollar nueva investigación (Abramo *et al.*, 2019). Este problema puede ser extrapolado a los nombres de técnicas específicas (es decir, una palabra o un conjunto de palabras que describen cierta técnica) indicando su relevancia en el dominio. Sin embargo, a la luz de la inmensa demanda que puede implicar el ejercicio de filtrado manual, automatizar dichas operaciones resulta fundamental para abordar el proceso.

El conocido algoritmo TF-IDF (Ramos *et al.*, 2003) es una de las técnicas más importantes para automatizar el cálculo de relevancia de las palabras que aparecen en un corpus de textos (Jabri *et al.*, 2018) (ver Ecuaciones 2.1, 2.2 y 2.3). Demuestra habilidad para filtrar las palabras más utilizadas (es decir, aquellas palabras que tienen más ocurrencias) centrándose solo en las más significativas. Vale la pena mencionar que, aunque se han desarrollado algoritmos novedosos basados en TF-IDF (por ejemplo, *Term Frequency Inverse Gravity Moment* (TF-IGM)) que muestran capacidades para proporcionar mejores resultados en áreas específicas, todos presentan un decremento en la calidad general cuando se aplican a dominios de propósito general (K. Chen *et al.*, 2016). Además, la incrustación de palabras (Zamani y Croft, 2017), los algoritmos evolutivos (Cecchini *et al.*, 2018) y la retroalimentación interactiva (Donoso-Guzmán y Parra, 2018) son técnicas comunes en el campo. En contraste, otras técnicas concentran sus esfuerzos en medir lo opuesto (es decir, la anti-relevancia) con el propósito de encontrar problemas en sistemas de recomendación (Sánchez y Bellogín, 2018).

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D) \quad (2.1)$$

Donde t es el término, d el documento y D el conjunto de documentos o corpus.

A su vez, la función TF se define como:

$$TF(t, d) = \frac{f(t, d)}{\max\{f(t, d) : t \in d\}} \quad (2.2)$$

Donde se obtiene la frecuencia normalizada de cada término en cada documento, y finalmente la función IDF:

$$IDF(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (2.3)$$

Profundizando en el proceso de filtrado de documentos por relevancia (Benedetti *et al.*, 2019), se observa que la mayoría de las técnicas empleadas proceden intensificando sus esfuerzos en etiquetas o palabras clave (Lu y Tanne, 2017). En el caso específico de los documentos médicos, existen varios enfoques (Lagopoulos *et al.*, 2018) cuya mayoría comúnmente utiliza PubMed (Goeckenjan *et al.*, 2011) (ver Figura 2.2) para realizar búsquedas basadas en la relevancia del documento utilizando sus palabras clave (Fiorini *et al.*, 2018). Además, algunas alternativas consideran la relevancia de las oraciones como un mecanismo para producir el ranking de documentos (Q. Chen *et al.*, 2018).

The screenshot displays the PubMed search results page for the query "ibuprofen for headache". The interface includes the NIH National Library of Medicine logo, a search bar with the query, and navigation options like "Save", "Email", and "Send to". The results are sorted by "Best match" and displayed on page 1 of 44. On the left, there are filters for "MY NCBI FILTERS", "RESULTS BY YEAR" (a bar chart from 1973 to 2023), "TEXT AVAILABILITY" (Abstract, Free full text, Full text), "ARTICLE ATTRIBUTE" (Associated data), "ARTICLE TYPE" (Books and Documents, Clinical Trial, Meta-Analysis, Randomized Controlled Trial, Review, Systematic Review), and "PUBLICATION DATE" (1 year, 5 years, 10 years, Custom Range).

437 results

Page 1 of 44

1 [Paracetamol versus Ibuprofen for the Acute Treatment of Migraine Headache in Children: A Blinded Randomized Controlled Trial.](#)
 Cite Pavithra V, Mishra D, Behera S, Juneja M.
 Indian J Pediatr. 2020 Oct;87(10):781-786. doi: 10.1007/s12098-020-03315-x. Epub 2020 May 26.
 Share PMID: 32451827 Clinical Trial.
 OBJECTIVE: To compare the efficacy of oral paracetamol and oral **ibuprofen** for the management of acute **headache** in children with migraine without aura. METHODS: This randomized-controlled trial was done at the Pediatric department of a public hospital in India between ...

2 [Ibuprofen for acute treatment of episodic tension-type headache in adults.](#)
 Cite Derry S, Wiffen PJ, Moore RA, Bendtsen L.
 Cochrane Database Syst Rev. 2015 Jul 31;2015(7):CD011474. doi: 10.1002/14651858.CD011474.pub2.
 Share PMID: 26230487 Free PMC article. Review.
 BACKGROUND: Tension-type **headache** (TTH) affects about one person in five worldwide. It is divided into infrequent episodic TTH (fewer than one **headache** per month), frequent episodic TTH (1 to 14 **headaches** per month), and chronic TTH (15 **headaches** a mon ...

3 [Intravenous ibuprofen for acute treatment of migraine: A double-blind, randomized, placebo-controlled pilot study.](#)
 Cite Yuan H, Curran JG, Keith SW, Hopkins MM, Silberstein SD.
 Headache. 2021 Oct;61(9):1432-1440. doi: 10.1111/head.14214. Epub 2021 Oct 3.
 Share PMID: 34601736 Clinical Trial.
 We performed a single-center, double-blind, randomized, placebo-controlled pilot study to evaluate the efficacy and safety of IV **ibuprofen** as an acute treatment of migraine (NCT01230411). METHODS: Individuals with episodic migraine were screened at the Jefferson **Headach** ...

4 [Caffeine in the management of patients with headache.](#)
 Cite Lipton RB, Diener HC, Robbins MS, Garas SY, Patel K.
 J Headache Pain. 2017 Oct 24;18(1):107. doi: 10.1186/s10194-017-0806-2.
 Share PMID: 29067618 Free PMC article. Review.
 Caffeinated **headache** medications, either alone or in combination with other treatments, are widely used by patients with **headache**. ...In this review, we consider the role of caffeine in the over-the-counter treatment of **headache**. The MEDLINE and Cochrane data ...

Figura 2.2: Ejemplo de búsqueda en PubMed. PubMed es un servicio gratuito de la Biblioteca Nacional de Medicina de los Estados Unidos que ofrece acceso a investigaciones biomédicas. Cuenta con opciones avanzadas de búsqueda y la capacidad de crear alertas. A la izquierda, los usuarios pueden filtrar los resultados por año, disponibilidad de texto, atributo del artículo y fecha de publicación. Los resultados centrales presentan el título del estudio, autores, fuente, fecha, un breve resumen y un identificador único. También hay opciones para citar, compartir o guardar cada artículo. La navegación adicional permite cambiar el orden de los resultados y ajustar cómo se muestran.

2.1.2. Conclusiones

En el caso de sistema de recomendación propuesto en la tesis, se trata de un sistema centrado en el apoyo a la toma de decisiones que proporciona los documentos científicos más relevantes según el procesamiento automático de los informes de alta. Así, es capaz de proporcionar los documentos más similares a enfermedades o síndromes específicos, facilitando el trabajo de investigación de los médicos. Para ello, una vez revisada la literatura científica, se ha optado por un sistema de similitud semántica basado en una arquitectura de tipo *transformer*, ya que proporciona mejor robustez ante la recuperación de sinónimos y se adapta muy bien al tamaño de los *abstracts* de artículos científicos, que por su parte, han sido elegidos sobre los artículos completos por contener la información más relevante de los documentos. Por otra parte, el algoritmo TF-IDF ha sido seleccionado como algoritmo de relevancia de propósito general para proceder con el ranking de las palabras recopiladas de un corpus de un dominio específico. El campo de la medicina ha sido elegido para dirigir los experimentos, lo que también ha resultado en el uso de documentos indexados en PubMed.

2.2. Técnicas de generación de diccionarios

La generación de diccionarios que incluyen información sobre varias palabras representa un problema ampliamente abordado (Y. Chen *et al.*, 2014), lo que ha resultado en un desarrollo incesante de diccionarios enfocados en diversos propósitos y campos de aplicación, como el cuidado de la salud (Were *et al.*, 2007), la economía (Perelet *et al.*, 2014) y la industria de las tecnologías de la información (Q. Deng *et al.*, 2017).

Dentro de las posibles clasificaciones para los diccionarios, una de las más comúnmente reconocidas es representada tanto por las categorías estándar como por las de tesaurus. Los diccionarios que pertenecen al primer grupo contienen definiciones (Lowe y Sayle, 2015), o en general, información relevante sobre palabras (Wang *et al.*, 2018). Por el contrario, los diccionarios que pertenecen a la segunda categoría organizan las palabras según la similitud del significado (Alfano *et al.*, 2018).

Para producir los diccionarios mencionados, algunas técnicas existentes aplican algoritmos que toman WordNet (Fellbaum, 2010) como entrada para construir exitosamente el artefacto resultante requerido (Bracewell, 2008). Además, WordNet proporciona métricas de similitud entre palabras, sinónimos y antónimos (ver Figura 2.3), lo que demuestra su alta adecuación a la instancia específica presentada por el tesaurus. Existen otros enfoques que involucran técnicas notoriamente conocidas en el dominio de las tecnologías de la información (por ejemplo, Programación Lineal

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponymy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry
<i>Note:</i> N = Nouns Aj = Adjectives V = Verbs Av = Adverbs		

Figura 2.3: Relaciones semánticas en WordNet (Miller, 1995).

Entera Mixta (Garrido-Jurado *et al.*, 2016)), que muestran evidencia clara del trabajo incesante que la comunidad científica invierte en este problema.

Junto con lo anterior, desarrollar sistemas capaces de automatizar la generación de diccionarios a partir de un corpus de texto se confirma como una tarea esencial (Cheng y Huang, 2017) y un siguiente paso inevitable tanto para reducir la complejidad general del proceso como para obtener resultados que cumplan con las restricciones de tiempo a menudo encontradas en escenarios del mundo real.

Con respecto a los fundamentos de estos sistemas, la selección de palabras clave (Kutsumi *et al.*, 2010) y la organización en temas (Macdonald *et al.*, 2018) se muestran como los enfoques predominantes. La primera estrategia logra regularmente mejores resultados mediante el pre-procesamiento de las palabras, siendo la derivación (Patil *et al.*, 2017) y la lematización (Rundell y Kilgarriff, 2011) las dos técnicas de soporte más determinantes para alcanzar dicho rendimiento. Mientras que la derivación opera extrayendo la raíz de las palabras (lo que puede inducir algunos problemas de ambigüedad), la lematización se centra en la selección de lemas (que se seleccionan por convención para representar el lexema de palabras específicas), permitiendo así tener una forma para cada una de las palabras. Este problema es muy relevante en dominios como la medicina y similares, donde se consideran conceptos específicos o formas latinas originales (H. Liu *et al.*, 2012). La segunda estrategia utiliza palabras

clave para organizar el diccionario en temas. A pesar de que este método muestra relevancia para los usuarios que buscan encontrar palabras específicas de diferentes dominios, se encuentran dificultades para recopilar términos infrecuentes de interés en los textos (Z. Chen *et al.*, 2016).

2.2.1. Lexicones de análisis de sentimientos

El análisis de sentimientos y la minería de opiniones es un área específica de esfuerzo en la disciplina de la minería de texto. Se centra en la evaluación de las emociones que pueden surgir de los textos. Entre los diversos métodos de clasificación para las emociones, aquellos basados en la polaridad (negativo, neutro y positivo) y aquellos basados en el estado de ánimo (Scherer *et al.*, 1984) (ira, miedo, repulsión, tristeza, sorpresa y alegría) ilustran los escenarios más representativos. Cabe señalar que el primero suele ser preferido al segundo cuando se pretende discernir si las emociones corresponden a un estado de ánimo específico.

Este dominio específico presenta tres perspectivas principales: basadas en diccionarios (Taboada *et al.*, 2011), enfoques basados en *Machine Learning* (ML) (Chikersal *et al.*, 2015) y aproximaciones combinadas (Fernández-Isabel *et al.*, 2023). La primera utiliza lexicones o léxicos (Khoo y Johnkhan, 2018) (es decir, diccionarios con dos campos para cada entrada: clave y valor), mientras que la segunda emplea modelos estadísticos para predecir valores de sentimiento.

Los léxicos suelen construirse utilizando un gran corpus en el que se consideran varios textos. Hay varias tácticas para lograr esta tarea. Algunas de ellas utilizan algoritmos genéticos (Mowlaei *et al.*, 2018) o heurísticas específicas (Akram y Tahir, 2018) para obtener el valor correspondiente de las palabras. Otros proponen una organización basada en temas (D. Deng *et al.*, 2019, 2018).

Una de las debilidades conocidas de los lexicones está relacionada con el dominio del corpus. Para propósitos generales, existen varios enfoques (por ejemplo, SentiWordNet (Baccianella *et al.*, 2010) (ver Figura 2.4) y SenticNet (Cambria *et al.*, 2014)) que ofrecen un rendimiento satisfactorio. Sin embargo, en el caso de los dominios específicos (Goeuriot *et al.*, 2012; H. Han *et al.*, 2018) (por ejemplo, salud o deportes) o el idioma (S. Zhang *et al.*, 2018) (por ejemplo, inglés o chino), se pueden generar resultados ruidosos (Bandhakavi *et al.*, 2017). Este hecho facilita el desarrollo de múltiples léxicos entrenados con diferentes corpus, lo que puede producir una amplia variedad de diccionarios con valores subjetivos (Freitas *et al.*, 2004).

Además, los léxicos suelen presentar una limitación importante ya que están limitados por el conjunto de palabras consideradas. En consecuencia, esto produce tanto una pérdida de información como resultados ruidosos al evaluar un texto que contiene palabras que no están en el léxico.

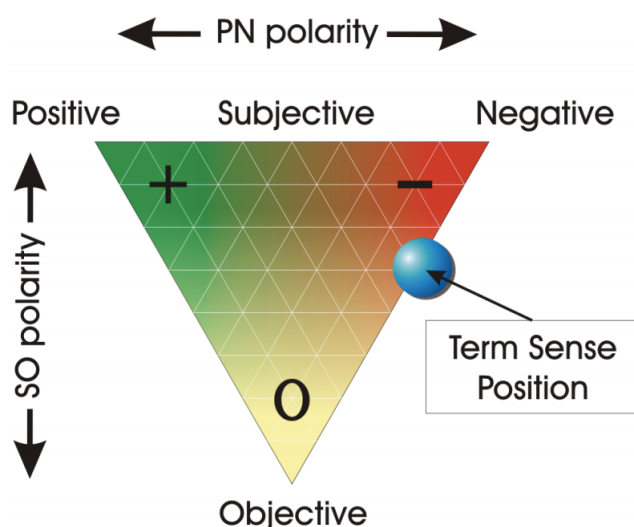


Figura 2.4: Representación gráfica adoptada por SentiWordNet para representar las propiedades relacionadas con la opinión de un sentido del término en consideración. (Sebastiani y Esuli, 2006).

Las redes neuronales ofrecen asistencia para resolver los fallos del léxico previamente descritos (Lei *et al.*, 2018; Fu *et al.*, 2018). Las redes neuronales también consideran años diferentes para proporcionar predicciones ajustadas. Alternativamente, hay otros métodos que se centran en técnicas de incrustación de palabras para mitigar los problemas del léxico (Rezaeinia *et al.*, 2019; Park *et al.*, 2018). Estas técnicas de incrustación de palabras también podrían emplearse para entrenar modelos de ML, lo que lleva a la producción de un marco completo para el análisis de sentimientos (Moussa *et al.*, 2018).

2.2.2. Conclusiones

En cuanto al marco de trabajo para la estimar la relevancia presentado en esta tesis, se produce un diccionario estándar siguiendo la selección de palabras clave y la lematización como estrategia de diseño principal y técnica de soporte, respectivamente. El diccionario resultante organiza sus palabras basándose en su relevancia.

La estimación de relevancia combina el uso de dos tipos de artefactos: un diccionario basado en relevancia previamente descrito y modelos de ML basados en redes neuronales (Cambria, 2016). El léxico utilizado en el marco de trabajo presentado basa sus fundamentos en las técnicas de análisis de sentimientos descritas anteriormente. Sin embargo, innova al almacenar la relevancia de la palabra (en lugar del valor del sentimiento) como el valor asociado a cada entrada de clave particular para un año específico (lo que potencialmente lleva a varias entradas para la misma palabra mostrando diferentes valores de relevancia dependiendo del año en consi-

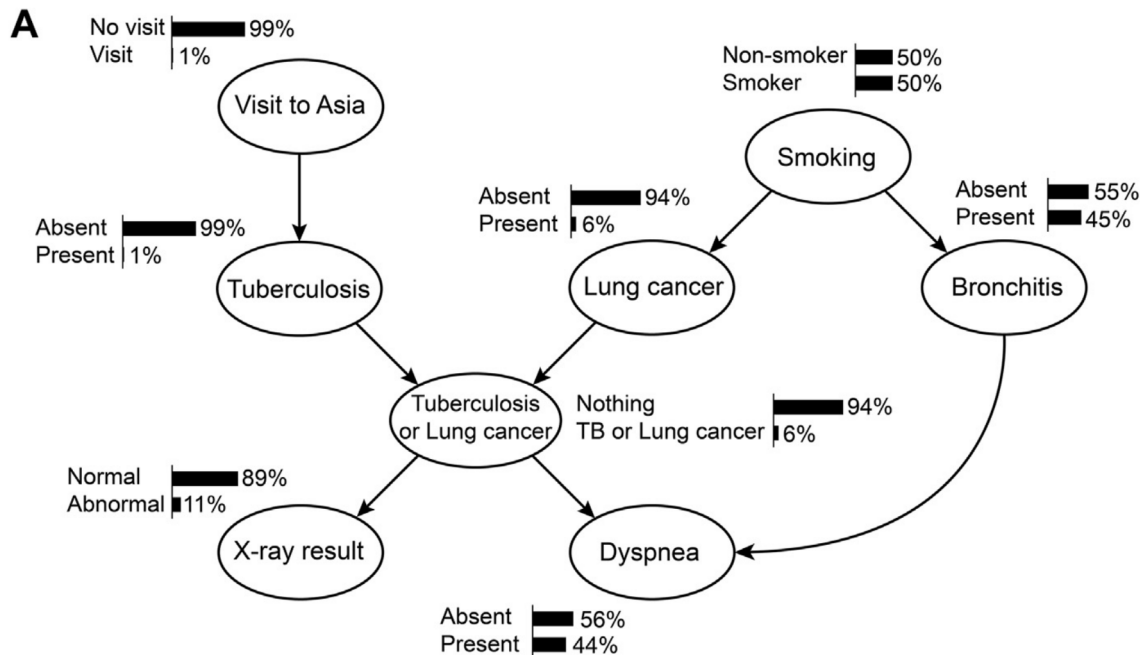


Figura 2.5: Ejemplo de red bayesiana ficticia para la predicción de cáncer o tuberculosis (Arora et al., 2019).

deración). La relevancia en sí se calcula utilizando métricas de relevancia (es decir, TF-IDF, ver Sección 2.1.1).

2.3. Inteligencia artificial en medicina

La IA es un campo de la informática centrado en emular los procesos mentales de individuos inteligentes para resolver problemas específicos mediante el uso de máquinas (Hunt, 2014).

Este campo puede organizarse en cuatro perspectivas principales: sistemas de razonamiento basados en casos, sistemas expertos, redes bayesianas y sistemas basados en comportamientos.

Profundizando en las perspectivas, los sistemas de razonamiento basados en casos resuelven problemas utilizando la experiencia adquirida a través del estudio de casos previos (Kolodner, 2014). Los sistemas expertos pueden recopilar conocimiento específico de seres humanos con experiencia en áreas concretas y luego aplicarlo para resolver problemas relacionados (Liebowitz, 2019). En el caso de las redes bayesianas, construyen gráficos que utilizan probabilidades para transitar entre estados (aristas). Estas probabilidades se utilizan para inferir situaciones futuras posibles de acuerdo con un estado actual (Marcot y Penman, 2019). Finalmente, los sistemas basados en comportamientos consisten en estructuras que establecen relaciones e

interacciones entre los elementos para resolver problemas complejos (Riedl, 2019). Aquí, los sistemas multiagente son el enfoque más típico. Estos sistemas inteligentes cooperan, compiten e interactúan entre sí y con el entorno para resolver problemas sencillos que pueden unirse para elaborar una solución más compleja.

En lo que respecta al ámbito médico, los sistemas inteligentes suelen considerar tanto la perspectiva del paciente (Ploug y Holm, 2020) como la del médico (Scheepers-Hoeks *et al.*, 2013). El desarrollo de sistemas complejos e inteligentes centrados en el ámbito de la salud es un tema ampliamente extendido.

Los sistemas de razonamiento basados en casos en el ámbito de la salud se han utilizado ampliamente para detectar posibles enfermedades e inferir la evolución de los pacientes (Bentaiba-Lagrid *et al.*, 2020). Este hecho tiene sus fundamentos en el progreso común de las enfermedades en la mayoría de los pacientes. Así, considerando los pasos iniciales, los siguientes pueden inferirse con cierto grado de certeza (Duan y Jiao, 2021).

Los sistemas expertos son enfoques bien conocidos centrados principalmente en la predicción y diagnóstico de enfermedades utilizando la experiencia de los médicos (Singla *et al.*, 2014). Así, acumulan conocimientos de expertos para inferir posibles soluciones según un conjunto de síntomas. Estos síntomas suelen incluirse manualmente en el sistema como entrada por parte de los usuarios.

Las redes bayesianas (ver Figura 2.5) son enfoques orientados principalmente a los pacientes. Permiten descubrir patrones en los tratamientos y también en el diagnóstico de enfermedades (Akila y Balaganesh, 2021). De esta manera, facilitan la generación de ontologías y relaciones entre medicamentos, síntomas y también efectos secundarios.

Los sistemas basados en comportamientos están presentes en múltiples perspectivas en el ámbito de la salud (Isern y Moreno, 2016). Pueden ser útiles para los pacientes, donde pueden analizar múltiples variables de una persona a través de sensores, y para los expertos en atención sanitaria, ya que pueden simular reacciones complejas a medicamentos. Además, son relevantes en la simulación de infraestructuras sanitarias y también en experimentos biomédicos.

Finalmente, es relevante destacar sistemas que pueden ofrecer apoyo a los profesionales de la salud durante el proceso de procedimientos complejos. Aunque generalmente no incluyen algoritmos de IA complejos, presentan elementos robóticos para simplificar las tareas durante operaciones y tratamientos invasivos (Sun *et al.*, 2020).

2.3.1. Procesamiento automático de diagnósticos

Existen sistemas específicos adaptados para detectar y realizar diagnósticos efectivos utilizando datos provenientes de pacientes anteriores. Estos sistemas son generalmente sistemas expertos, que se entrenan siguiendo un conjunto de reglas o etiquetas que se utilizan para elaborar una clasificación. Esta clasificación consta de una bandera positiva o una bandera negativa, donde la positiva confirma la detección de una posible enfermedad, y la negativa descarta la posibilidad con cierto grado de certeza. Todos estos sistemas también se llaman en medicina: sistemas de atención sanitaria inteligentes (Mansour *et al.*, 2021). Cabe destacar que estos sistemas tienen como punto común la explicabilidad de las decisiones tomadas (Khodabandehloo *et al.*, 2021). Esto es fundamental para los expertos en atención sanitaria, ya que deben explicar a los pacientes las posibles enfermedades detectadas. Además, ayuda durante el proceso de confirmación, ya que un ser humano debe estar de acuerdo con la detección realizada por el sistema.

Profundizando en los sistemas basados en reglas, siguen diferentes indicaciones (perspectiva de abajo-arriba) estableciendo el posible positivo según las reglas cumplidas. Estos sistemas actúan como profesionales de la salud, siguiendo los diferentes síntomas y puntos de interés para producir el diagnóstico final. Por lo tanto, son muy similares a los procedimientos de toma de decisiones humanos. Ejemplos típicos de estos sistemas son aquellos basados en lógica difusa (Mousavi *et al.*, 2021), árboles de decisión y grafos de criterio (Alves *et al.*, 2021).

Los sistemas basados en reglas también se han utilizado en la literatura para establecer relaciones entre un diagnóstico, medicamentos y posibles reacciones adversas. Así, el contenido textual se analiza para producir modelos generalmente basados en grafos que pueden reflejar y simplificar estas interacciones y problemas (Tan *et al.*, 2022). Este procedimiento facilita la selección del tratamiento más adecuado para los profesionales de la salud.

Por otro lado, los sistemas de razonamiento basados en casos son una parte específica de los sistemas basados en reglas. Estos sistemas utilizan casos estudiados previamente para inferir conocimiento y hacer suposiciones sobre un conjunto de síntomas (Duan y Jiao, 2021). También utilizan algunas reglas para discriminar entre los casos, filtrando para seleccionar los más adecuados.

En el caso de los sistemas basados en etiquetas, utilizan los datos para detectar patrones ocultos que generalmente no son detectados por los profesionales de la salud (perspectiva de arriba-abajo). Estos sistemas se basan en técnicas y modelos de ML. Luego, estos modelos se entrenan utilizando la perspectiva de aprendizaje supervisado para abordar el problema. Así, adaptan sus parámetros según los datos de entrada y la etiqueta deseada, obteniendo una configuración específica que detecta

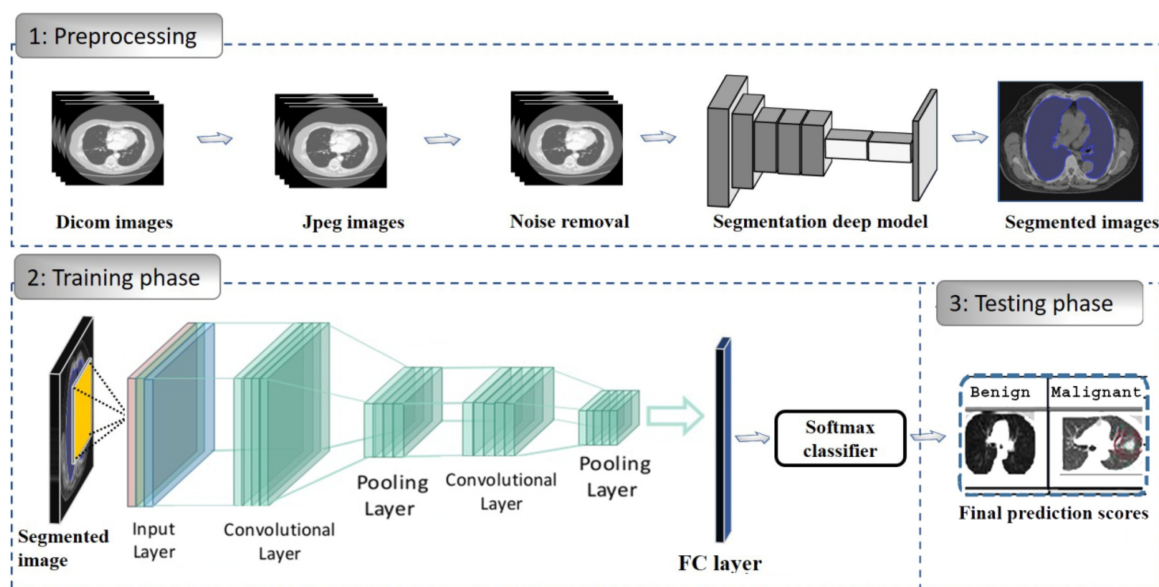


Figura 2.6: Marco de trabajo basado en la visión artificial para la predicción de nódulos pulmonares (Saba, 2020). El uso de técnicas de aprendizaje automático ayuda en el diagnóstico temprano y la evaluación de nódulos pulmonares al procesar imágenes de tomografía computarizada construidas mediante métodos de inteligencia artificial. Estos sistemas se denominan sistemas de apoyo a la decisión que investigan las imágenes a través de pre-procesamiento, segmentación, extracción de características y un proceso de clasificación.

la enfermedad predefinida en los datos. Ejemplos típicos de estos sistemas utilizan modelos de ML bien conocidos como redes neuronales convolucionales (Saha *et al.*, 2021), bosques aleatorios y máquinas de vector soporte (Rauber *et al.*, 2021).

Mención especial en esta categoría para los sistemas de visión artificial. Son sistemas que utilizan modelos de ML específicamente adaptados para detectar patrones visuales (ver Figura 2.6). Por lo tanto, son útiles para evaluar imágenes clínicas de pacientes, donde se reflejan posibles enfermedades, problemas y síntomas (I. Castiglioni *et al.*, 2021), (Solovyh *et al.*, 2021). Pueden centrarse en regiones detalladas y analizar los píxeles de forma equiparable o incluso mejor que un profesional humano, lo que permite utilizar estos sistemas como sistemas de apoyo a la toma de decisiones. Por ejemplo, la *Convolutional Neural Network* (CNN) propuesta por (Esteva *et al.*, 2017) se utilizó para clasificar lesiones cutáneas, entrenada con un conjunto de datos de 129.450 imágenes clínicas de 2.032 enfermedades diferentes. Su rendimiento se probó frente a 21 dermatólogos certificados en dos clasificaciones críticas: carcinomas de queratinocitos frente a queratosis seborreicas benignas y melanomas malignos frente a nevos benignos. La CNN demostró un rendimiento comparable a los expertos en ambas tareas, lo que indica que la inteligencia artificial puede clasificar el cáncer de piel con una competencia similar a la de los dermatólogos.

Extracto de un resumen de alta anonimizado proveniente de MIMIC-III

Admission Date: [**2167-4-28**] Discharge Date: [**2167-5-3**]

Date of Birth: [**2090-2-1**] Sex: F

Service: GEN [**Doctor First Name 147**]

ADMITTING DIAGNOSIS: 1. Pancreatic mass.

DISCHARGE DIAGNOSES: 1. Pancreatic mass.

PROCEDURES DURING ADMISSION: 1. Exploratory laparotomy, lysis of adhesions and enucleation of a neuro-endocrine pancreatic mass.

HISTORY OF PRESENT ILLNESS: The patient is a 77 year old female who presents with a history of a benign pancreatic mass, which causes her significant abdominal pain. The patient presents electively to have this resected.

PAST MEDICAL HISTORY: 1. Diabetes mellitus. 2. Increased cholesterol. 3. Idiopathic hypertrophic subaortic stenosis with an echocardiogram [**3-/2167**], revealing an ejection fraction of greater than 55 %.

PAST SURGICAL HISTORY: 1. Pancreatic resection in [**2155**].

ALLERGIES: No known drug allergies.

MEDICATIONS: 1. Aspirin. 2. Atenolol 125 mg p.o. twice a day. 3. Hydrochlorothiazide 25 mg p.o. q. day. 4. Zestril 40 mg p.o. q. day. 5. Prilosec 20 mg p.o. q. day. 6. Verapamil 240 mg p.o. twice a day.

PHYSICAL EXAMINATION: On examination the patient is awake and alert in no apparent distress. Her heart is regular rate and rhythm, S1, S2. Her lungs are clear to auscultation bilaterally. Abdomen soft. She has a well healed midline scar.

HOSPITAL COURSE: The patient was admitted to the hospital on [**2167-4-28**], and taken to the Operating Room for enucleation of the pancreatic mass and lysis of adhesions. She would also be discharged on:

DISCHARGE MEDICATIONS: 1. Protonix 40 mg p.o. q. day.

DISCHARGE INSTRUCTIONS: 1. She was told to follow-up with Dr. [**Last Name (STitle) 468**] in the office and to call for a follow-up appointment.

[**First Name8 (NamePattern2) 251**] [**Name8 (MD) **], M.D. [**MD Number(1) 4984**]

Dictated By:[**Last Name (NamePattern1) 4985**]

2.3.2. Conclusiones

En cuanto a la solución propuesta en este trabajo, se trata de un sistema de apoyo a la toma de decisiones basado en el conocimiento experto proporcionado a través de informes de alta de varios pacientes. El sistema procesa este contenido textual de forma semántica y luego utiliza ese conocimiento para recomendar los artículos científicos más adecuados a los profesionales de la salud. Este hecho simplifica su trabajo y facilita encontrar nuevos hallazgos relacionados con algunos medicamentos y tratamientos para enfermedades de interés. Además, la incorporación de un LLM, permite a sus usuarios interactuar con las recomendaciones mediante el uso del lenguaje natural. El sistema de recomendación propuesto en esta tesis utiliza los diagnósticos detectados en informes de alta provenientes de MIMIC-III para fundamentar las recomendaciones. Esta detección se basa en la combinación de reglas y patrones lingüísticos con un modelo de ML previamente entrenado para aislar dichos diagnósticos de los EHR. La similitud semántica entre los diagnósticos y los artículos médicos permite inferir conocimiento. Por lo tanto, podría considerarse un híbrido entre un sistema basado en reglas y un sistema basado en casos, donde los patrones para encontrar información son previamente conocidos y pueden utilizarse para lograr clasificaciones relevantes de enfermedades (es decir, tareas relacionadas con la nosología) y recomendaciones relacionadas.

2.4. Procesamiento del lenguaje natural

El PLN es un área específica de la inteligencia artificial que busca facilitar la interacción entre las computadoras y el lenguaje humano. Esta disciplina se centra en enseñar a las máquinas cómo interpretar, comprender y generar el lenguaje en una forma que sea valiosa y significativa para los seres humanos. El PLN tiene sus raíces en los años 50, con la traducción automática siendo uno de los primeros problemas abordados (Nadkarni *et al.*, 2011). Las primeras soluciones dependían de reglas gramaticales y diccionarios, pero la complejidad inherente del lenguaje hizo que estos métodos fueran insuficientes. Con la llegada de la era de la información y la disponibilidad de grandes cantidades de datos textuales, los métodos basados en estadísticas y más tarde en aprendizaje profundo han dominado la disciplina, permitiendo avances significativos en tareas como traducción automática, análisis de sentimientos y reconocimiento de voz (Otter *et al.*, 2020). A continuación se recoge una lista de las principales tareas del PLN:

- **Tokenización:** Divide el texto en palabras, frases u otros subcomponentes significativos llamados tokens. Es uno de los primeros pasos en muchos flujos de PLN.

- **Lematización:** Convierte una palabra a su forma base o raíz.
- **Stemming:** Reduce una palabra a su base truncando los sufijos.
- **Part-of-Speech Tagging (POS Tagging):** Asigna etiquetas gramaticales a palabras, identificando si son sustantivos, verbos, adjetivos, etc.
- **Reconocimiento de Entidades Nombradas (NER):** Identifica y clasifica entidades en el texto, como nombres de personas, organizaciones, lugares, fechas, etc.
- **Análisis sintáctico (Parsing):** Determina la estructura gramatical de una frase, construyendo un árbol que representa las relaciones de dependencia entre las palabras.
- **Semántica textual:** Se enfoca en comprender el significado del texto, incluyendo la identificación de sinónimos, la resolución de anáforas y la desambiguación del sentido de las palabras.
- **Traducción Automática:** Convierte texto de un idioma a otro.
- **Generación de Texto:** Crea texto a partir de datos.
- **Análisis de Sentimientos:** Detecta la polaridad (positiva, negativa, neutral) de un texto.
- **Reconocimiento de Voz:** Convierte el habla en texto escrito.
- **Síntesis de Voz (Text-to-Speech):** Convierte texto escrito en habla.
- **Resumen Automático:** Crea un resumen conciso de un texto más largo.
- **Respuesta a Preguntas:** Busca respuestas a preguntas en grandes conjuntos de datos o en la web.
- **Clasificación de Texto:** Asigna etiquetas o categorías a un texto según su contenido.
- **Extracción de Relaciones:** Identifica y clasifica relaciones entre entidades nombradas en el texto.
- **Desambiguación de Entidades:** Determina a qué entidad se refiere una palabra o frase.
- **Detección de Idioma:** Identifica qué idioma se está utilizando en un texto.
- **Corrección Ortográfica:** Identifica y corrige errores de escritura en el texto.

- **Diálogos y Sistemas Conversacionales:** Permite a las máquinas mantener conversaciones fluidas con los usuarios.

A pesar de los avances, el PLN aún enfrenta desafíos. La ambigüedad del lenguaje humano, la variabilidad en el uso del lenguaje entre diferentes grupos de personas y la vastedad de contextos y dominios son áreas en las que aún hay margen de mejora. La aparición de la arquitectura *Transformer* (ver Figura 2.7) y de modelos posteriores basados en ella, como BERT (Devlin *et al.*, 2018), *Robustly Optimized BERT Pretraining Approach* (RoBERTa) (Y. Liu *et al.*, 2019) y *Generative Pre-trained Transformer* (GPT) (Radford *et al.*, 2018), entre otros, ha sido revolucionaria en el campo. Estos modelos han establecido nuevos estándares de referencia en una amplia variedad de tareas, superando en muchos casos las técnicas tradicionales y anteriores basadas en otros enfoques, como las *Recurrent Neural Networks* (RNN), LSTM o CNN. La arquitectura *Transformer* introdujo el mecanismo de atención, lo que permite a los modelos centrarse en diferentes partes de un texto dependiendo de su relevancia para una palabra o frase en particular. Esto ha mejorado drásticamente la capacidad de los modelos para manejar dependencias a largo plazo en el texto. Por otra parte, BERT popularizó la idea de utilizar un modelo entrenado de forma auto supervisada para luego ajustar ese modelo pre-entrenado para diversas tareas específicas de PLN, este concepto es habitualmente conocido como *transfer learning* (Weiss *et al.*, 2016). Esto ha permitido a los investigadores aprovechar modelos altamente eficientes sin necesidad de grandes conjuntos de datos o recursos computacionales extensivos, dando lugar a un enorme avance en la comunidad científica del PLN en los últimos años. A partir de la arquitectura original *Transformer* (ver Figura 2.7) se han establecido tres variantes enfocadas a tareas diferentes: Los *encoders*, *decoders* y *encoders-decoders*, dando lugar a numerosos modelos cada una de ellas. Los modelos basados en *encoders*, como BERT, RoBERTa o *XLNet* (Yang *et al.*, 2019) se centran en procesar la entrada y generar una representación fija. Estos modelos son particularmente buenos para tareas que no necesitan generar texto, sino entenderlo o extraer información de él, como la clasificación de texto, dar respuesta a preguntas, extracción de entidades nombradas o desambiguación de significados de palabras. Por otra parte, los modelos basados en *decoders*, como GPT o *Large Language Model Meta AI* (LLaMA) (Touvron, Lavril, *et al.*, 2023) se centran en generar texto. Están diseñados para producir secuencias de salida y son útiles en tareas donde la generación de lenguaje es clave, como creación de contenido, traducción automática o generación de código a partir de instrucciones en lenguaje natural. Por último, Los modelos que combinan *encoders* y *decoders*, como la arquitectura original *Transformer*, *Text to Text Transfer Transformer* (T5) (Raffel *et al.*, 2020) o *Bidirectional and Auto-Regressive Transformers* (BART) (Lewis *et al.*, 2019) aprovechan las fortalezas de ambos para tareas que involucran tanto la comprensión como la generación de texto. Algunas de estas tareas son la generación de resúmenes y la traducción automática.

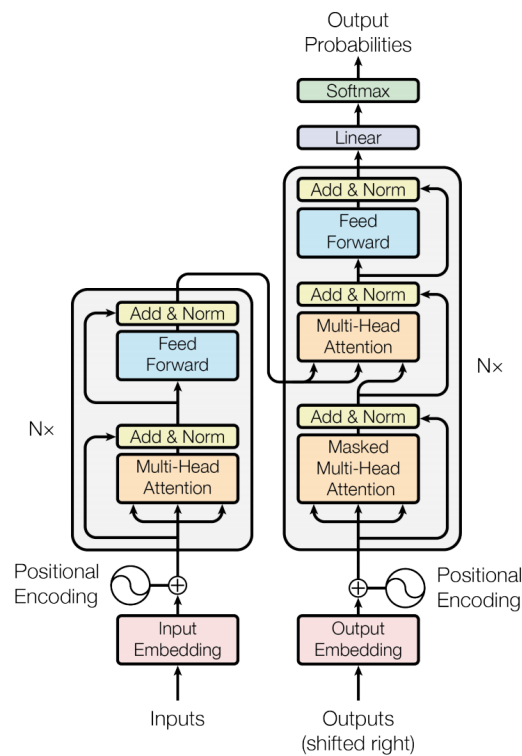


Figura 2.7: *Arquitectura Transformer original (Vaswani et al., 2017).*

Tratamiento especial merece la disciplina de la minería de texto. El PLN y la minería de texto están estrechamente relacionados y a menudo se solapan, pero tienen diferentes objetivos y áreas de enfoque. Mientras que el PLN se centra en entender y generar lenguaje, la minería de texto se centra en extraer información y descubrir patrones en grandes volúmenes de texto. Aunque ambas disciplinas pueden usarse juntas, por ejemplo, en un sistema de análisis de sentimientos, el PLN generalmente se utiliza para tareas de comprensión y generación de lenguaje, mientras que la minería de texto se utiliza para analizar grandes conjuntos de datos textuales en busca de patrones o información útil. Algunas de las tareas propias de la minería de textos más relevantes son el *clustering* o agrupación de documentos (Afzali y Kumar, 2019) y la extracción de temas, típicamente conocida como *topic modeling* (Kherwa y Bansal, 2019).

2.4.1. Asistentes conversacionales

Un asistente conversacional es un programa de software diseñado para interactuar con humanos a través del lenguaje natural, generalmente con el objetivo de ayudar en tareas específicas o proporcionar información. Estos asistentes pueden estar habilitados para voz o texto y suelen incorporar capacidades de aprendizaje automático y procesamiento del lenguaje natural para entender y responder a consultas de

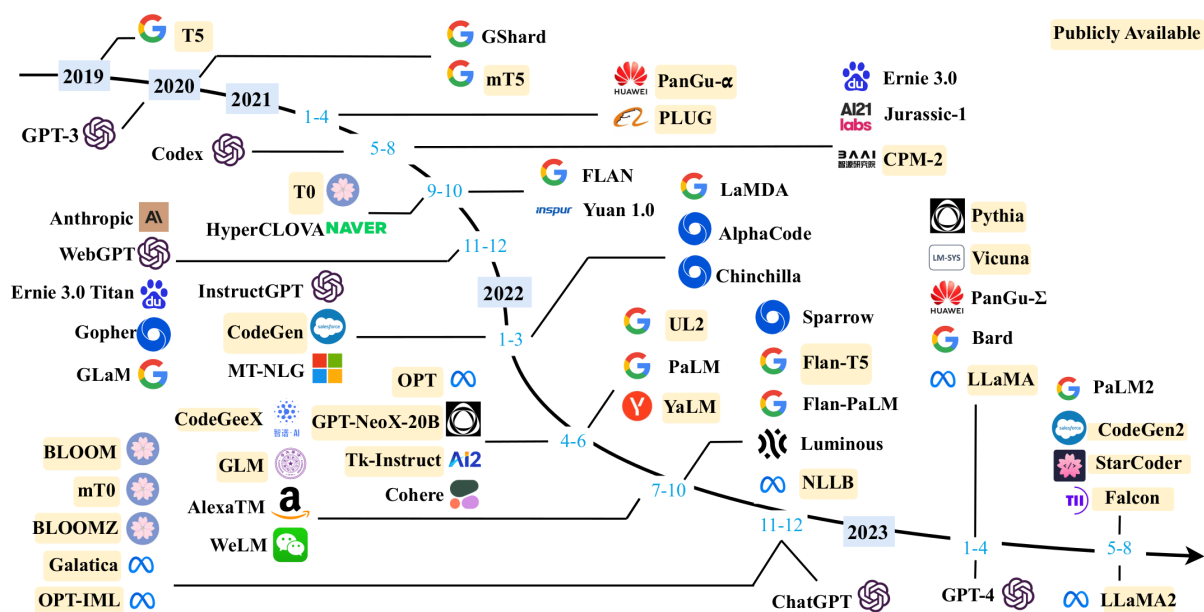


Figura 2.8: Línea temporal de los LLM con más de diez mil millones de parámetros surgidos desde 2019 (Zhao et al., 2023).

manera más efectiva. Un *chatbot* es un subconjunto de asistentes conversacionales que interactúa principalmente a través de interfaces de texto. Los primeros *chatbots* utilizaban enfoques basados en reglas y patrones de coincidencia para interactuar con los usuarios (Weizenbaum, 1966). Estos *chatbots* carecían de la habilidad para comprender y generar lenguaje natural de una manera que pudiera considerarse IA pero eran efectivos para tareas específicas y contextos limitados.

Con la evolución de las técnicas de aprendizaje automático y el PLN, los *chatbots* basados en algoritmos de aprendizaje automático empezaron a ser implementados. Estos modelos podían entrenarse en grandes conjuntos de datos y ofrecían respuestas más precisas en comparación con los sistemas basados en reglas, aunque carecían de la potencia necesaria para un correcto entendimiento del contexto (Suta et al., 2020).

Con el lanzamiento de GPT-1 (Radford et al., 2018), el campo vio un aumento significativo en la calidad de los asistentes conversacionales. Los modelos evolucionaron con GPT-2 (Radford et al., 2019) y GPT-3 (Brown et al., 2020), demostrando capacidades para tareas como traducción de idiomas, generación de código y más. No obstante, estos modelos generativos requerían de complejos *prompts* con ejemplos resueltos sobre las tareas que debían resolver, alejándolos por tanto de una interacción natural como asistente. Este problema fue solucionado con chatGPT y GPT-4 ha llevado esta evolución aún más lejos, mejorando la coherencia, la comprensión del contexto y la capacidad para realizar tareas más complejas. No obstante, la empresa que desarrolló estos modelos, denominados LLM, OpenAI, decidió privatizarlos,

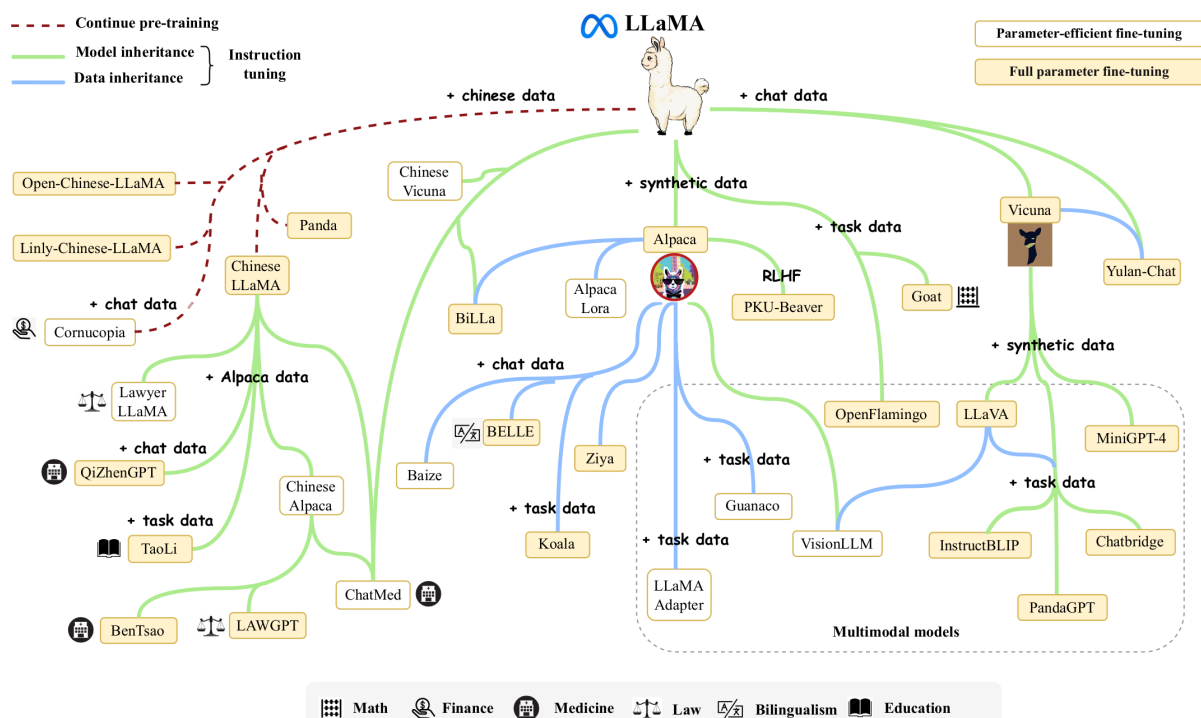


Figura 2.9: Gráfico de la evolución de la investigación surgida desde la liberación de LLaMA (Zhao et al., 2023).

limitando el avance académico. Por su parte, Google también lanzó modelos homónimos como *Pathways Language Model* (PaLM) (Chowdhery et al., 2022) y PaLM-2, aunque Google también decidió privatizar este último, publicando un reporte técnico que omitía gran parte de los detalles técnicos como ya hiciera OpenAI con GPT-4. A la carrera de la publicación de LLM (ver Figura 2.8) se unió Facebook con LLaMA y posteriormente con LLaMA-2 (Touvron, Martin, et al., 2023) liberando el modelo a la comunidad científica en su versión 1 y para uso industrial, aunque con ciertas limitaciones, en su versión 2. Esto dio lugar a una explosión de modelos publicados por diferentes instituciones basados en diferentes implementaciones y ajustes de los modelos de Facebook como Alpaca (Taori et al., 2023) y Vicuna (Chiang et al., 2023) (ver Figura 2.9).

2.4.2. Conclusiones

Diversas técnicas de PLN han sido utilizadas a lo largo de todos los componentes de la tesis. En primer lugar, técnicas como la lematización y el *Part-of-Speech Tagging* han sido seleccionadas para el cálculo de lemas y obtención de sustantivos para la construcción del diccionario de relevancia. Por otra parte, se ha utilizado una técnica de resumen automático extractivo para comprimir la información más relevante en los artículos utilizados con el mismo fin, así como *tokenización* de frases. Se ha utilizado también un modelo *Named Entity Recognition* (NER) especializado

en entidades médicas para apoyar el aislamiento de las entidades que componen los diagnósticos de los EHR. En último lugar, se han seleccionado GPT-3.5-turbo y GPT-4 para los experimentos en la implementación de un LLM para el asistente conversacional con inyección de contexto. Esta decisión está motivada por la buena habilidad demostrada por GPT-3.5-turbo en tareas relacionadas con el dominio médico, como la creación de resúmenes o el uso de lenguaje técnico (Casella *et al.*, 2023) así como por su velocidad y relativo bajo coste.

Capítulo 3

Propuesta general

En este capítulo se introduce la propuesta general de la tesis, un asistente de investigación médica basado en conocimiento. Primero, en la sección 3.1, se presenta su arquitectura global con sus tres componentes principales: *Webelance*, componente encargado de estimar la relevancia de artículos biomédicos, MELENDI, un sistema de recomendación de artículos médicos y finalmente GAMER, un asistente generativo con inyección de contexto. Después, en la sección 3.2 se detalla la metodología de trabajo empleada a lo largo de la tesis y una planificación temporal del desarrollo de la misma.

3.1. Arquitectura

El objetivo del asistente es proporcionar un marco completo de soporte a la investigación médica, que permita a los médicos mantenerse actualizados en sus respectivos campos y en materias relacionadas con las características de sus pacientes con la menor interacción, necesidad de conocimiento técnico y tiempo posibles. Para lograr este objetivo, el asistente se diseña sobre tres componentes fundamentales: En primer lugar, el asistente debe ser capaz de proporcionar artículos relevantes. En segundo lugar, el sistema debe ser capaz de recomendar artículos relevantes a sus usuarios médicos en base a las características de sus pacientes, sin necesidad de interacción. Y por último, el sistema debe permitir la recuperación de información proveniente de las recomendaciones sin la necesidad de conocimiento técnico, es decir, con lenguaje natural. La arquitectura de la propuesta general se diseña con un componente fundamental responsable de cada una de estas tareas. Así, el componente del asistente generativo con inyección de contexto se comunica con el componente de recomendación de artículos, que, a su vez, basa su métrica de relevancia de ar-

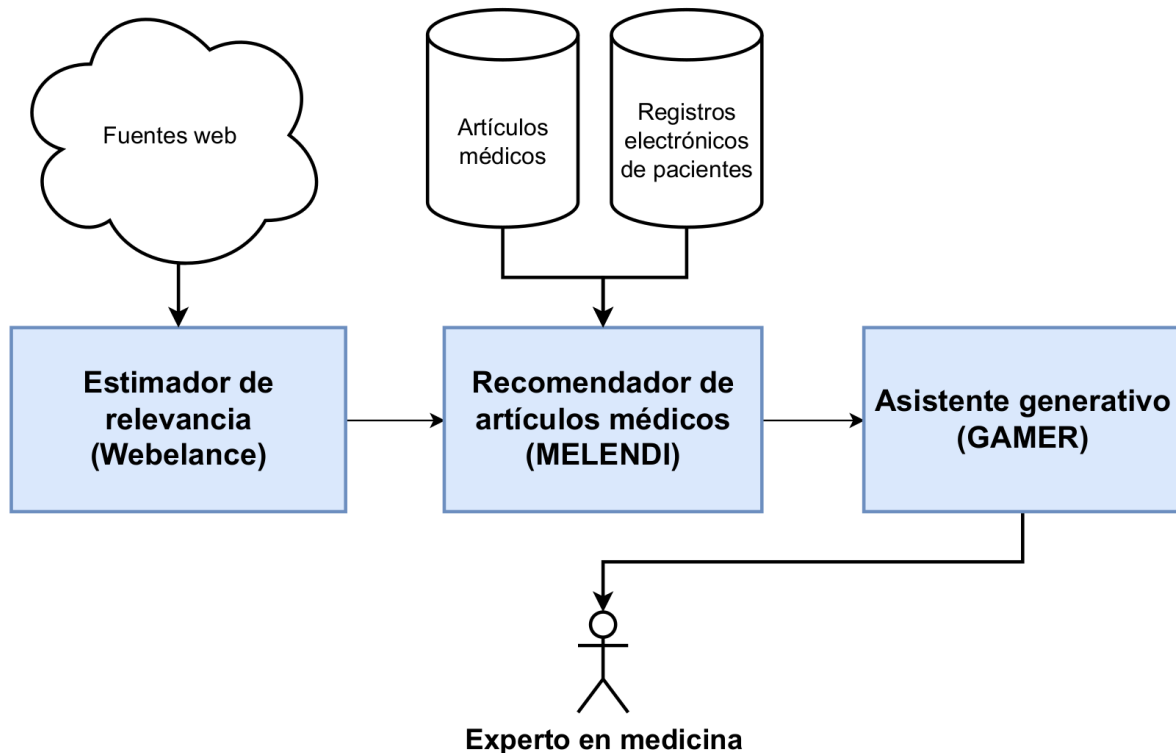


Figura 3.1: Arquitectura simplificada del asistente de investigación médica basado en conocimiento.

tículos en el componente de estimación de relevancia. En la Figura 3.1 se puede ver la arquitectura básica del sistema completo. *Webelance* consulta una fuente de conocimiento externa para apoyar el cálculo de relevancia. MELENDI consume tres fuentes de información, en primer lugar la base de datos con los artículos médicos, en segundo lugar la base de datos de historiales electrónicos de los pacientes del usuario y, en último lugar, consulta *Webelance* para el cálculo de relevancia como métrica necesaria para poder recomendar los artículos. El componente GAMER, por su parte, utiliza los artículos recomendados calculados por MELENDI en base a un diagnóstico concreto y una consulta del usuario para construir la respuesta. En la Figura 3.2 se puede ver un ejemplo de interacción del usuario, que realiza una consulta en base a los artículos recomendados sobre un paciente con el diagnóstico *COPD exacerbation*.

Debido a que cada uno de los componentes depende del anterior, se han construido en orden de dependencia, es decir, en primer lugar *Webelance*, después MELENDI y por último GAMER. Cada componente ha sido evaluado de forma independiente con experimentos diseñados específicamente. En capítulos sucesivos se profundiza en detalles de arquitectura, experimentos y evaluación de cada uno de los componentes.

Dezzai-Assistant

Diagnosis

COPD exacerbation

Query to the assistant

What are the benefits of whole body vibration training in COPD patients? Submit

Assistant Examples ▼

Publication year filter

1989 2021

1989 2021

Assistant Creativity

0.00 2.00

Relevant Documents

- **(Doc 1)** Benefits of whole body vibration training in patients hospitalised for COPD exacerbations - a randomized clinical trial. (2014) Patients with stable COPD show improvements in exercise capacity and muscular function after the application of whole body vibration. We aimed to evaluate whether this modality added to conventional physiotherapy in exacerbated hospitalised COPD pati... (<https://www.ncbi.nlm.nih.gov/pubmed/24725369>)
- **(Doc 2)** Trends in In-Hospital Outcomes Among Adults Hospitalized With Exacerbation of Chronic Obstructive Pulmonary Disease. (2016) Although substantial advances have been made in the treatment of chronic obstructive pulmonary disease (COPD), little is known regarding the impact of these advancements on inpatient outcomes over time. We sought to examine temporal trends in in-hosp... (<https://www.ncbi.nlm.nih.gov/pubmed/26263035>)
- **(Doc 3)** Relationship between bone resorption, oxidative stress and inflammation in severe COPD exacerbation. (2014) The natural course

Output

According to [1](#), whole body vibration training can lead to improvements in exercise capacity and quality of life in patients with stable COPD. The study aimed to evaluate whether this modality added to conventional physiotherapy in exacerbated hospitalised COPD patients would be safe and would improve exercise capacity and quality of life. The study found that whole body vibration during hospitalised exacerbations did not cause procedure-related adverse events and induced clinically significant benefits regarding exercise capacity and health-related quality of life that were associated with increased serum levels of irisin, a marker of muscle activity.

Figura 3.2: Prototipo de la interfaz de usuario del asistente. En la sección "Relevant Documents" se recogen los N (3 visibles en este ejemplo) artículos recomendados para un paciente cuyo diagnóstico es "COPD exacerbation". Se representa el título, parte del abstract y la URL de la fuente original de cada artículo. El usuario ha introducido una consulta en la sección "Query to the assistant" y el asistente genera la respuesta enlazando con la fuente que la fundamenta en la sección "Output".

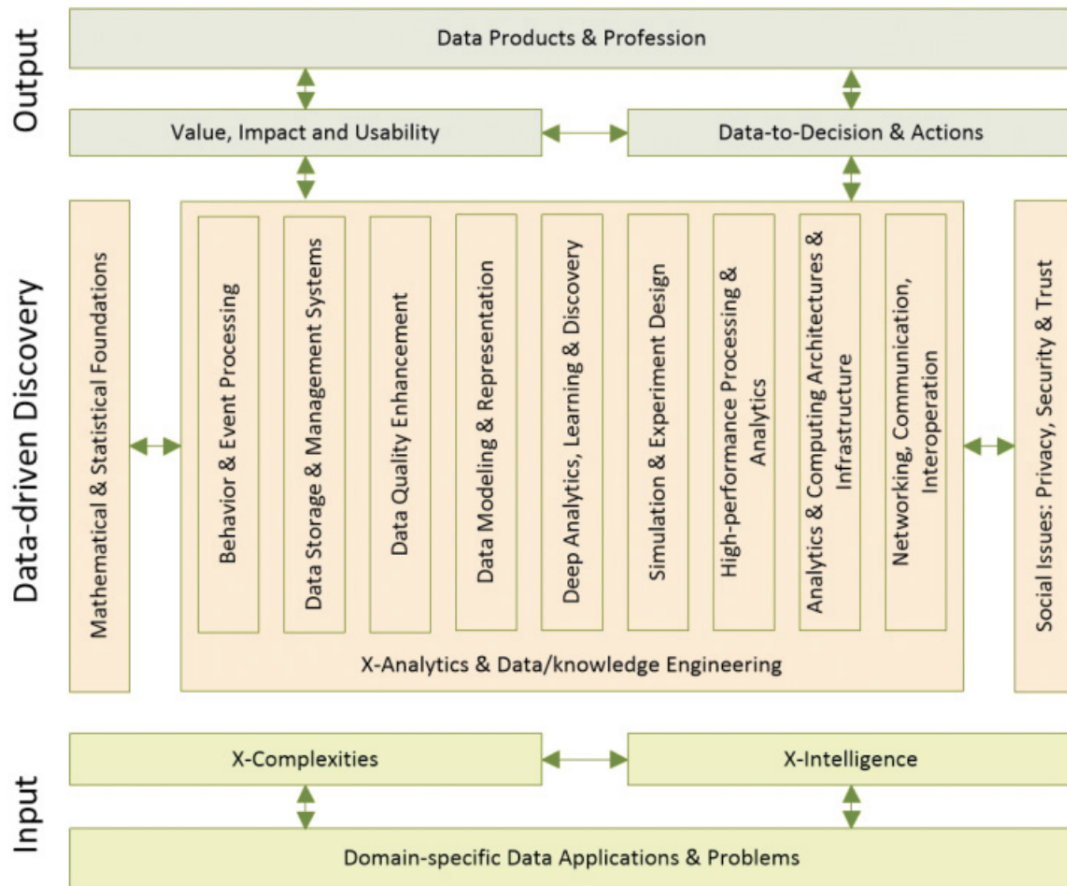


Figura 3.3: Mapa conceptual de la ciencia de datos (Cao, 2017).

3.2. Metodología de trabajo

Esta tesis se ha desarrollado bajo la metodología de la ciencia de datos, en la Figura 3.3 se puede ver un mapa conceptual que resume dicha metodología. Típicamente, la ciencia de datos se resume en tres fases principales: Obtención y limpieza de datos, modelización y evaluación, dando lugar a un bucle entre la evaluación y la modelización, ya que los resultados de las evaluaciones a menudo se utilizan para volver a ajustar los modelos y que estos puedan mejorar.

Profundizando en la metodología, se identifican los siguientes pasos:

- **Definición del Problema:** Antes de sumergirse en los datos, es esencial definir claramente el problema que se quiere resolver. Esta definición debe ser específica, medible y relevante para el objetivo de la investigación.
- **Obtención de datos:** Se recolectan datos de diversas fuentes, que pueden incluir bases de datos, sensores, encuestas, registros, entre otros. Es vital asegurar la calidad y pertinencia de los datos recopilados. En esta tesis se trabaja con da-

tos provenientes del dominio biomédico, tanto artículos científicos como historiales clínicos electrónicos de pacientes. Cabe señalar que estos últimos únicamente pueden ser obtenidos previa consecución de la certificación del curso *CITI Data or Specimens only*⁴, dada la sensibilidad de los mismos.

- **Tratamiento de datos:** Esta etapa involucra la limpieza y transformación de los datos para prepararlos para el análisis. Las tareas comunes incluyen la identificación y manejo de valores faltantes, detección de atípicos, y la normalización y estandarización de datos.
- **Exploración de datos:** Antes del análisis formal, se realiza un análisis exploratorio para entender las características y estructuras subyacentes en los datos. Esto implica visualizaciones, estadísticas descriptivas y pruebas de hipótesis.
- **Selección de características:** No todas las características o variables en un conjunto de datos son relevantes para el problema en cuestión. Esta etapa se centra en identificar y seleccionar las características más significativas para el análisis posterior.
- **Modelado:** Se construyen modelos predictivos o descriptivos utilizando técnicas de aprendizaje automático o estadística. Esta fase también incluye la selección de algoritmos, la configuración de parámetros y la validación cruzada.
- **Evaluación:** Una vez que se ha construido el modelo, se evalúa su rendimiento en datos no vistos anteriormente. Se utilizan métricas específicas para determinar la precisión, la sensibilidad y otras métricas relevantes.
- **Implementación:** Si el modelo es satisfactorio, se puede desplegar en un entorno de producción. Esto puede involucrar la integración con sistemas existentes y la creación de interfaces de usuario.
- **Monitorización y mantenimiento:** Los modelos de ciencia de datos no son estáticos. Es esencial monitorizar su rendimiento en el tiempo y reajustar o recalibrar según sea necesario.
- **Presentación de resultados:** Los hallazgos y resultados se presentan a las partes interesadas de manera comprensible, utilizando visualizaciones y narrativas claras.

En la siguiente sección se muestra una distribución temporal de las tareas que componen esta tesis, cada una de las cuales se ajusta a la metodología de ciencia de datos previamente descrita.

⁴<https://physionet.org/content/mimiciii/1.4/>

3.3. Distribución temporal

En la Figura 3.4 se puede ver un diagrama de Gantt del desarrollo de las principales tareas de la tesis. La consecución de cada uno de los tres componentes principales está señalada con un cuadrado en negro. Cabe destacar el estudio constante de la literatura científica a lo largo de la tesis, una tarea necesaria dado el enorme avance de la inteligencia artificial en los últimos años, especialmente en el campo del PLN. Cada uno de los componentes principales, *Webelance*, MELENDI y GAMER, se ha desarrollado bajo la metodología de ciencia de datos, comenzando por la adquisición y tratamiento de datos para construir cada componente, la construcción de modelos, la experimentación y la evaluación. Es ampliamente establecido que la obtención y el tratamiento de datos ocupa entre el cincuenta y el ochenta por ciento de un proyecto de ciencia de datos, no obstante, este no ha sido el caso por diferentes razones. Como se puede observar en el diagrama de Gantt, las tareas 2, 8 y 9 se corresponden con la obtención y limpieza de datos. La tarea 2 consiste en la obtención de un corpus de artículos médicos para construir *Webelance*, el componente estimador de relevancia. Al tratarse de artículos científicos, estos datos están limpios en origen, por lo que no han requerido de tiempo de limpieza. La tarea 8, por su parte, consiste en un corpus de artículos diferente y mucho mayor para adaptarse a las necesidades del componente recomendador de artículos. El mencionado mayor volumen de este corpus obedece a la necesidad de contar con la mayor cantidad de conocimiento manejable de forma ágil en términos computacionales para realizar los experimentos del recomendador. Finalmente, la tarea 9 lleva bastante más tiempo de forma comparativa a las tareas 2 y 8 por la obligatoriedad de realización de un curso de formación específico que permite el acceso a los datos. En este caso se trata de una base de datos de acceso público que contiene información detallada relacionada con la atención médica de pacientes. Una vez obtenida la base de datos, los datos de interés para esta tesis se encuentran limpios y anonimizados en texto plano dentro de un fichero CSV, por lo que aparte de la realización del curso y aislamiento de los mismos, su pre-procesado y limpieza no han requerido más tiempo.

Por su parte, las tareas 3, 4, 5, 10, 11, 12, 13 y 16 son las tareas de relativas a la modelización y creación de algoritmos para cada uno de los 3 componentes. Estas a menudo se solapan en el tiempo unas con otras debido a la iteración de mejora mientras se desarrollan varias en paralelo. Finalmente, las tareas de experimentación y evaluación se corresponden con los ID 6, 15 y 19. Comparativamente, estas tareas son las que han llevado más tiempo, ya que se han ajustado parámetros en función de los resultados de distintas evaluaciones de los sistemas realizados por los expertos en el dominio. La tarea 18 puede ser considerada como una tarea de obtención de datos, ya que consiste de un corpus de consultas construidos conjuntamente con los expertos en el dominio con el fin de evaluar la propuesta realizada en el componente



Figura 3.4: Distribución temporal de las principales tareas de la tesis.

GAMER.

La tarea 17 se corresponde con la propia escritura de la tesis, que comienza una vez definida la propuesta de GAMER y que se desarrolla mayoritariamente en paralelo a los experimentos y evaluación de la misma.

En el próximo capítulo se aborda en detalle *Webelance*, el estimador de relevancia y primer componente principal de la tesis. Este capítulo aborda las tareas comprendidas entre la 2 y la 6. Después, se profundiza en el segundo componente, MELENDI, encargado de recomendar artículos científicos, cuyas tareas comprenden entre la 8 y la 14. Finalmente, en el capítulo posterior se describe GAMER, que consiste en un asistente generativo con inyección de contexto y cierra la arquitectura principal. Las tareas relativas a este componente son la 16, 18 y 19.

Capítulo 4

Estimador de relevancia de artículos

En este capítulo se introduce el primer componente de la tesis, el marco de trabajo *Webalance*. El objetivo de *Webalance* es la estimación automática de relevancia de artículos. Es el componente básico del asistente de investigación basado en conocimiento presentado en esta tesis, ya que los artículos proporcionados deben resultar relevantes en su dominio de aplicación. El marco de trabajo se basa en un lexicón de relevancias, modelos de aprendizaje máquina y métricas de reputación para estimar la relevancia. Aquí se describe la metodología utilizada para construirlo así como los experimentos diseñados para validar la propuesta dentro del dominio de la medicina.

4.1. Introducción

Es ampliamente conocido que existe una gran cantidad de documentos en todo el mundo. El contenido de estos documentos resulta ser muy heterogéneo, teniendo información que puede organizarse en múltiples temas. Además, esta información suele ser estática (por ejemplo, vídeos o textos) y no puede ser actualizada o mejorada fácilmente. Estos hechos podrían crear dificultades para aquellos usuarios que intentan hacer una selección correcta de documentos para obtener conocimiento relacionado con los mismos.

En el caso de la comunidad científica, esta situación se ha agravado con el aumento del número de revistas y conferencias (Ware y Mabe, 2015) que se ha producido en los últimos años. Esto da lugar a escenarios donde se pueden encontrar textos obsoletos y estudios de baja calidad junto con temas altamente conocidos (Shojania *et al.*, 2007; Pattanittum *et al.*, 2012). Por lo tanto, los nuevos investigadores que buscan mejorar sus habilidades y base científica en un dominio específico suelen encon-

trar dificultades. En estos casos, la figura de un mentor que proporcione directrices y realice una primera discriminación para filtrar textos es fundamental (Williamson *et al.*, 1989). Además, este proceso de filtrado puede ser útil para expertos científicos en un tema específico o una comunidad científica determinada, ya que el tiempo y esfuerzo necesarios para realizar una investigación pueden verse reducidos considerablemente.

La relevancia de artículos científicos consiste en el hecho de resultar valiosos y útiles para los científicos en su trabajo. Sin embargo, la discriminación de textos en función de su relevancia siempre ha sido una tarea difícil de lograr debido a los diversos factores que influyen en el proceso. Algunos de estos factores son: la importancia inicial de los temas considerados, su evolución a través del tiempo (podrían quedar obsoletos), la reputación de los autores, los dominios afectados y la importancia del documento para la comunidad. Obsérvese que algunos de estos factores pueden estar sesgados por personas dependiendo de sus antecedentes, opiniones y habilidades relacionadas con los dominios y temas considerados (Kumar, 2016).

Por esta razón, desarrollar un sistema que apoye y asista durante dicho proceso de filtrado resulta fundamental. El sistema debe ser capaz de medir objetivamente la relevancia de un texto en un dominio específico. Además, el sistema también debe realizar una amplia investigación estudiando varios corpus de documentos, procesando la información recopilada y organizando el conocimiento (simulando los antecedentes y habilidades de un mentor). Finalmente, el sistema debería incluir una medida para clasificar y discriminar los textos.

En este capítulo se introduce el marco de trabajo *Webelance*. Hace uso de dos tipos principales de artefactos: un léxico o lexicón basado en relevancia y modelos de ML para medir la relevancia de los trabajos científicos. El léxico se construye procesando una gran cantidad de artículos científicos pertenecientes a un dominio específico. Este proceso obtiene los conceptos y también mide su relevancia. Las métricas de relevancia se basan en la ocurrencia de los conceptos y la reputación del artículo. Esta reputación se basa en medidas objetivas previas utilizadas por la comunidad científica (Fernández-Isabel *et al.*, 2018). De este modo, la comunidad científica actúa como experta para generar el conocimiento utilizado para entrenar el sistema. Los modelos de ML complementan el marco de trabajo haciendo predicciones de la relevancia de los conceptos no considerados por el léxico. Por lo tanto, *Webelance* sigue un flujo de trabajo bien conocido en el dominio de la minería de textos (Cambria, 2016).

Los experimentos realizados en este trabajo están orientados a validar la propuesta. Se configuran diferentes valores para los parámetros internos del sistema con el fin de probar el rendimiento del marco de trabajo. Primero, se realiza un experimento con valores neutrales de los parámetros. Después, un segundo experimento en el que los valores de los parámetros son ajustados por expertos evalúa la mejora lograda

cuando se considera el conocimiento del dominio. Finalmente, un tercer experimento valida los resultados proporcionados por el sistema a lo largo del tiempo.

El sistema se evalúa mediante una batería de pruebas de documentos previamente etiquetados como relevantes o no relevantes por los expertos. Se ha seleccionado el dominio de la medicina para realizar los experimentos. Esta decisión está motivada por tres fundamentos. Primero, es uno de los dominios más importantes para la humanidad (More, 2016). Segundo, está constantemente siendo actualizado y mejorado con los avances realizados por los investigadores (lo que implica que las tendencias se modifican en un corto período de tiempo, generando varios manuscritos obsoletos (A. Castiglioni, 2019)). Finalmente, es uno de los campos más explorados en la comunidad científica (Richa, 2018), lo que facilita el proceso de recopilación de documentos requerido para crear un amplio corpus organizado por años.

El resto del capítulo se organiza de la siguiente manera. La sección 4.2 presenta el marco de trabajo desarrollado detallando sus módulos y componentes. La sección 4.3 propone varios experimentos en el dominio de la medicina para ilustrar la viabilidad del sistema. La sección 4.4 concluye y proporciona algunas directrices futuras.

4.2. Propuesta de marco de trabajo

El propósito principal del marco de trabajo *Webelance* consiste en estimar la relevancia de textos científicos en un dominio de aplicación prefijado. *Webelance* es un sistema experto que obtiene información de expertos de la comunidad científica para ser construido. De esta manera, se utilizan conceptos como el número de citas y la reputación de los autores basada en su importancia relativa en el campo de aplicación. Nótese que el sistema necesita otros expertos independientes en el dominio de aplicación seleccionado para ser evaluado y demostrar su viabilidad.

El sistema ha sido dotado con la capacidad de evaluar manuscritos publicados y también fragmentos de los mismos. Estos objetivos se logran utilizando dos tipos de artefactos: un léxico (Mowlaei *et al.*, 2018) que almacena la relevancia de los términos del dominio objetivo y modelos de ML basados en redes neuronales para hacer predicciones. Ambos artefactos tienen en consideración los años en los que los artículos fueron publicados. Este hecho permite a *Webelance* incorporar el factor tiempo en todos sus cálculos dando lugar una funcionalidad adicional que fortalece la contribución de esta propuesta, ya que tanto la capacidad del sistema para modular la curva de olvido (Averell y Heathcote, 2011) como la detección de tendencias en la relevancia se vuelven viables.

En lo que respecta a la arquitectura del marco de trabajo, consiste en los artefactos mencionados anteriormente (es decir, el léxico y las redes neuronales), el módulo

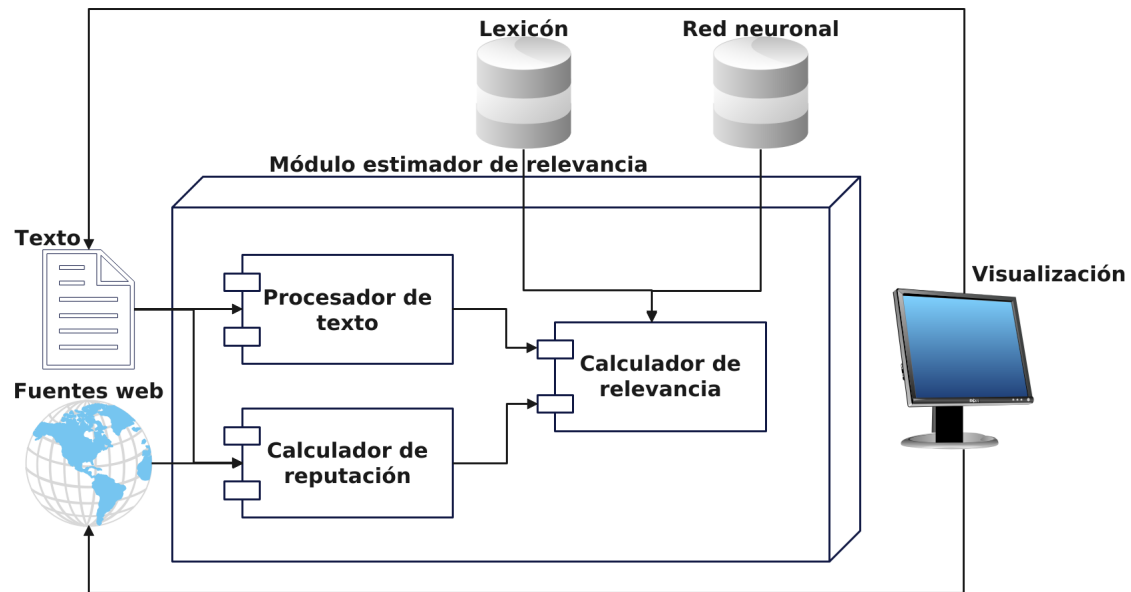


Figura 4.1: Representación general de la arquitectura de Webelance (módulos, componentes y artefactos).

Estimador de relevancia que actúa como núcleo del sistema y el módulo de *Visualización* para realizar representaciones gráficas de la información obtenida (ver Figura 4.1).

El módulo *Estimador de relevancia* es alimentado por el texto científico de interés que se va a analizar y comprende tres componentes: *Procesador de texto*, *Calculador de reputación* y *Calculador de relevancia*.

El componente *Procesador de texto* aplica técnicas de minería de texto (Manning *et al.*, 2014) para obtener las palabras relevantes del mismo. Estas palabras son los sustantivos en forma de uni-gramas. Cabe destacar que los n-gramas y los sintagmas nominales han sido descartados para facilitar el proceso de clasificación de relevancia. Esta decisión se basa en la suposición de que un *token* (es decir, uni-grama) es relevante a pesar de ser parte de una estructura nominal más compleja. Debido al hecho de que los conceptos científicos están definidos por sustantivos (Tolle y Chen, 2000), se ha tomado la decisión de diseño de descartar el resto de los elementos gramaticales.

El módulo *Calculador de reputación* solo toma acción cuando el texto científico presenta un *Digital Object Identifier* (DOI) (Paskin, 2010). En tal circunstancia, se utilizan *fuentes de información web* (en este caso, la *Application Programming Interface* (API) de Semantic Scholar (Allen Institute for Artificial Intelligence and Semantic Scholar, 2018)) para estimar la reputación del artículo de entrada.

Finalmente, el módulo *Calculador de relevancia* aborda el cálculo de la relevan-

cia considerando los resultados de los componentes anteriores. Esta relevancia se estima de acuerdo a tres procedimientos diferentes, no mutuamente excluyentes. La elegibilidad de estos procedimientos está determinada por cada caso particular bajo análisis, siendo sus resultados potencialmente combinables. El primer enfoque considera únicamente los valores de relevancia de las palabras almacenadas en el léxico para calcular la relevancia general. La segunda estrategia introduce redes neuronales para complementar los resultados derivados del léxico para aquellas oraciones cuya relevancia permanece desconocida y aún requiere predicción (es decir, aquellas oraciones que presentan palabras que no coinciden con las palabras almacenadas en el léxico). Por último, el tercer método amplía el cálculo y tiene en cuenta el hecho de que la información del DOI está disponible para evaluar la relevancia final combinando la relevancia calculada en el segundo enfoque con la reputación del texto científico.

4.2.1. Flujo del marco de trabajo

Esta sección ilustra el flujo de trabajo seguido por *Webelance* y las acciones que tienen lugar en cada componente específico.

El proceso se inicia seleccionando el texto científico de entrada cuya relevancia se va a evaluar y concluye proporcionando un valor de relevancia normalizado entre 0 y 1 (ver Figura 4.2). La interfaz gráfica del marco ofrece opciones para indicar si el texto de entrada corresponde a un formato en bruto o a un archivo de documento (por ejemplo, archivo PDF o similar). En el segundo caso, se dispara el proceso de análisis para extraer el contenido antes de proceder con la limpieza y división en oraciones (pasos *Parsear a texto crudo* y *Limpiar y tokenizar frases*).

A continuación, se extraen y lematizan los sustantivos relevantes, manteniéndolos organizados en listas que representan las oraciones originales donde aparecieron. Las propias oraciones originales también se procesan para transformarlas en una lista de lemas. Todas estas tareas se abordan durante el paso *Lematizar frases*. Cabe señalar que el componente *Procesador de texto* es responsable de todos estos pasos (análisis, limpieza y lematización).

Una vez que los sustantivos y sus oraciones correspondientes están organizados, se consulta el léxico de relevancia con el objetivo de obtener la relevancia asociada a los sustantivos de entrada que coincidan con su contenido. Cuando se encuentra el sustantivo de entrada en cuestión en el léxico, su valor de relevancia es acumulado. De lo contrario, la red neuronal procesa la oración donde aparece el sustantivo objetivo con el fin de proporcionar una estimación de relevancia. Todos los pasos involucrados son llevados a cabo por el componente *Calculador de relevancia*.

Posteriormente, si el texto procesado fue originalmente extraído de un archivo de

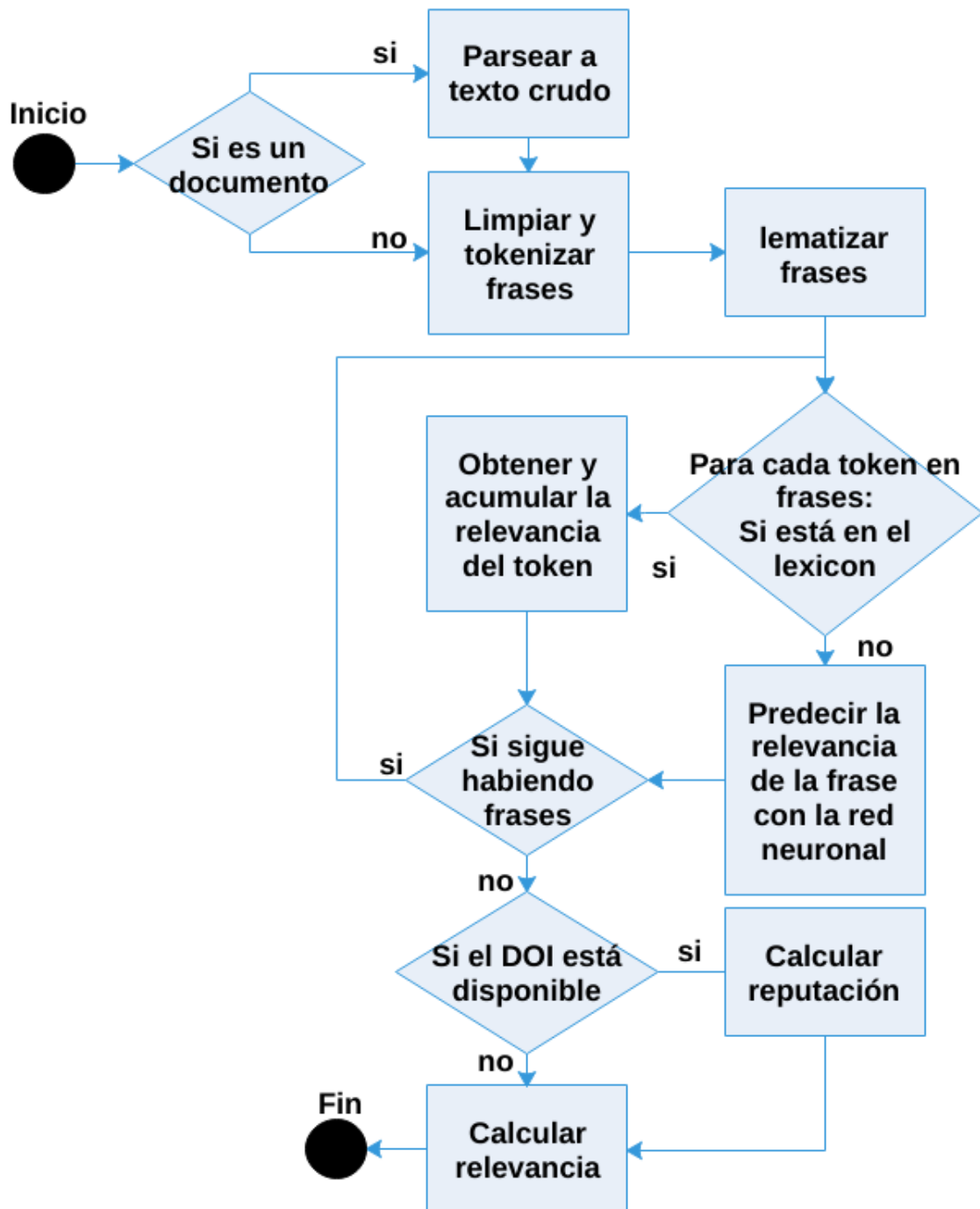


Figura 4.2: Proceso seguido por Webelance para calcular la relevancia de un artículo científico.

documento (fue necesaria la extracción del texto en bruto), el sistema intenta obtener su DOI para calcular su reputación basándose en la API de Semantic Scholar (Allen Institute for Artificial Intelligence and Semantic Scholar, 2018). Esta operación se lleva a cabo durante el paso de *Calcular reputación*, siendo el componente *Calculador de reputación* también responsable de ello.

Finalmente, tiene lugar el paso *Calcular relevancia* para evaluar la relevancia final, que puede considerar la reputación del artículo dependiendo de la presencia de la información del DOI. Como antes, el componente *Calculador de relevancia* está a cargo de este último cálculo.

4.2.2. Cálculo de relevancia

Examinando el cálculo de relevancia con más detalle, en primer lugar, la relevancia rel_lex_p de un texto científico p publicado en el año y se define como el promedio de la relevancia individual de aquellas palabras provenientes de p que coinciden con sus correspondientes entradas del léxico:

$$rel_lex_p = \frac{1}{T} \sum_{t=1}^T rel_lex_t^{(y)}, \quad (4.1)$$

donde T detalla el número total de palabras de entrada encontradas en el lexicón para el año específico y y $rel_lex_t^{(y)}$ denota la relevancia almacenada en el léxico para cada una de ellas.

Obsérvese que las entradas del léxico están organizadas en conjuntos etiquetados según los años a los que pertenecen. Por lo tanto, el conjunto correspondiente se consulta dependiendo del año y que se tome en consideración para así recuperar el valor de relevancia de las palabras. La siguiente sub-sección describe la manera en la que se calcula $rel_lex_t^{(y)}$ y también establece los procedimientos mediante los cuales se construyen los conjuntos anteriormente mencionados.

A continuación, la relevancia del contenido $rel_content_p$ de un artículo p publicado en el año y se calcula de la siguiente manera:

$$rel_content_p = \theta \cdot rel_lex_p + (1 - \theta) \cdot \frac{1}{S} \sum_{s=1}^S rel_neural_s^{(y)}, \quad (4.2)$$

donde rel_lex_p especifica la relevancia promedio proporcionada por el léxico para todo el texto, s denota la oración cuya relevancia es desconocida, S describe el número de oraciones cuya relevancia es desconocida, $rel_neural_s^{(y)}$ representa el valor de relevancia proporcionado por la red neuronal para cada oración s de acuerdo con el año y y $\theta \in [0, 1]$ modula la importancia relativa de cada predicción individual en el

resultado final.

La siguiente técnica corresponde a la reputación del artículo rep_p , el cálculo se realiza de acuerdo con el algoritmo introducido en (Fernández-Isabel *et al.*, 2018), que establece que la reputación de un texto científico p se define de esta manera:

$$rep_p = \alpha \cdot rep_authors_p + (1 - \alpha) \cdot citations_p, \quad (4.3)$$

donde $rep_authors_p$ representa la reputación de los autores, $citations_p$ detalla el número de veces que el texto p ha sido citado y $\alpha \in [0, 1]$ ajusta el peso e impacto de los términos mencionados anteriormente en la ecuación.

El término $rep_authors_p$ se formaliza promediando las reputaciones de cada uno de los autores rep_i del texto científico considerado:

$$rep_authors_p = \sum_{i=1}^n rep_i/n, \quad (4.4)$$

donde la reputación de cada uno de los autores rep_i también se define de la siguiente manera:

$$rep_i = \omega_1 \cdot inf_citation_count + \omega_2 \cdot citation_velocity + \omega_3 \cdot seniority + \omega_4 \cdot papers, \quad (4.5)$$

siendo $\sum_{i=1}^4 \omega_i = 1$ y donde $inf_citation_count$ corresponde al número de citas influyentes en trabajos publicados (Valenzuela *et al.*, 2015), $citation_velocity$ proporciona una medida de la popularidad y durabilidad de la publicación durante los tres últimos años, $seniority$ representa el número de años transcurridos entre la primera y la última publicación científica del autor, y $papers$ muestra la cifra correspondiente al número de publicaciones del autor. Vale la pena destacar que esta propuesta introduce una variación del algoritmo original presentado en (Fernández-Isabel *et al.*, 2018) como resultado de adoptar el parámetro $citation_velocity$ como sustituto del número total de citas recibidas en trabajos publicados al proceder con el cálculo de rep_i .

Todos los parámetros que intervienen en el cálculo de rep_i junto con $citations_p$ se recopilan utilizando la API de Semantic Scholar (Allen Institute for Artificial Intelligence and Semantic Scholar, 2018).

Finalmente, la relevancia de un texto científico p publicado en el año y se define como una combinación convexa:

$$relevance_p = \gamma \cdot rel_content_p + (1 - \gamma) \cdot rep_p, \quad (4.6)$$

donde el parámetro $\gamma \in [0, 1]$ regula la importancia relativa de ambos términos en la ecuación.

En resumen, el léxico siempre participa en la evaluación de la relevancia debido a que las coincidencias ocurren de manera constante con sus palabras almacenadas (es decir, el léxico está alineado con el dominio de aplicación de los textos científicos de entrada). En cuanto a la red neuronal, desempeña un papel de apoyo para proporcionar asistencia en el cálculo de las predicciones de relevancia cuando el léxico lo requiere. Finalmente, cuando la información del DOI está disponible, la reputación se utiliza para calcular la relevancia.

Para aquellos textos de entrada que presentan DOI, la propuesta manifiesta una contribución significativa y principal novedad al incluir la reputación en el cálculo general de la relevancia, siendo la red neuronal un método de apoyo contingente cuando el léxico no es capaz de proporcionar los valores de relevancia para palabras de entrada específicas.

Las siguientes secciones proporcionan una descripción detallada de los procesos seguidos para crear los componentes centrales del artefacto (el léxico y la red neuronal) junto con orientación sobre el flujo de trabajo de *Webelance*.

4.2.3. Concepto de diseño del lexicón

Esta sección muestra el marco de trabajo específico concebido para crear el léxico requerido. Como se presentó en las secciones anteriores, el artefacto resultante corresponde a un diccionario organizado donde, análogamente a los lexicones de sentimiento (Baccianella *et al.*, 2010), sus entradas toman la forma de pares que almacenan información clave (es decir, palabra) y valor (es decir, relevancia). Las entradas se clasifican en conjuntos que se ajustan a los años del calendario, siendo por lo tanto factible tener la misma palabra almacenada en diferentes conjuntos con diversos valores de relevancia asociados. Estos valores se definen dentro de una escala predefinida, entre 0 y 1, e indican la polaridad de la palabra.

Obsérvese que deben cumplirse dos condiciones previas antes de proceder con el proceso de creación del lexicón. En primer lugar, debe seleccionarse un dominio de aplicación concreto. En segundo lugar, los textos científicos pertenecientes al dominio elegido deben clasificarse según el año de publicación y deben tener suficientes muestras para poder crear conjuntos relevantes y evitar posibles defectos. Estos requisitos previos no deben ser percibidos como una limitación del alcance de la solución, sino como una alineación y cualificación para el dominio de aplicación específico al que se expone el marco de trabajo de trabajo *Webelance*.

Profundizando en la arquitectura (ver Figura 4.3), comprende tres módulos principales: un módulo de *Extracción, Transformación y Carga (ETL)* (Vassiliadis, 2009),

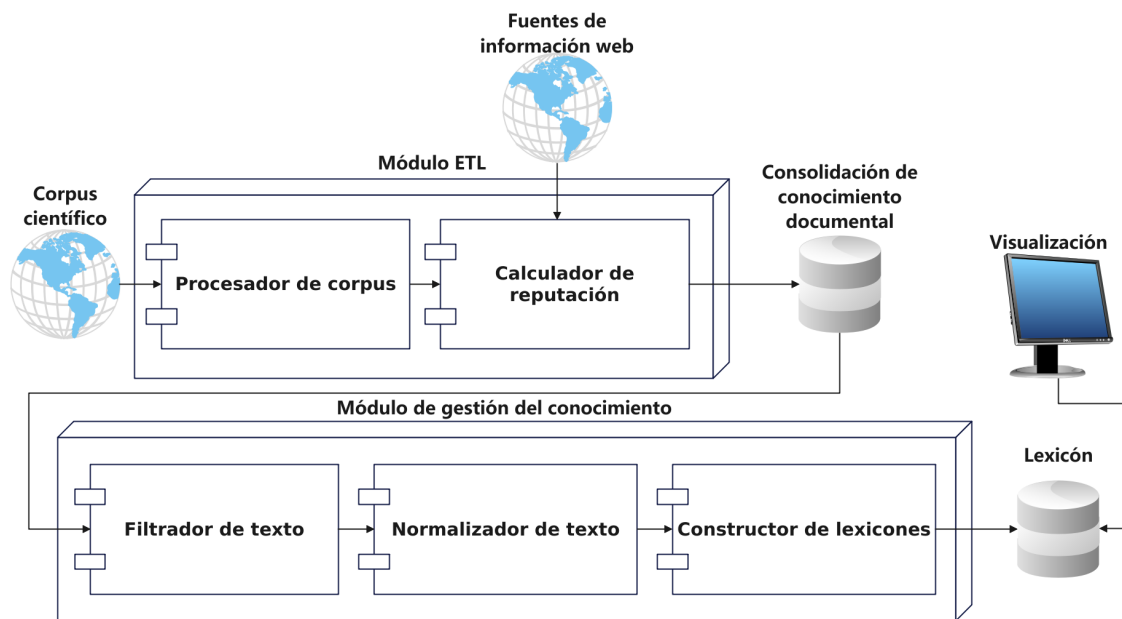


Figura 4.3: Vista general de la arquitectura del generador de lexicones.

un módulo de *Gestión del Conocimiento* y un módulo de *Visualización*. El marco de trabajo se completa con una base de datos *ElasticSearch* (Gormley y Tong, 2015) (es decir, una base de datos basada en documentos (J. Han *et al.*, 2011)) llamada *Consolidación de Conocimiento Documental*.

El módulo *ETL* es responsable de obtener y procesar los textos científicos de entrada para recopilar el conocimiento relevante y su almacenamiento en la base de datos de *Consolidación del Conocimiento Documental*. Dos componentes internos participan en estas tareas: un *Procesador de Corpus* y un *Calculador de Reputación*. El primer elemento recupera y analiza un corpus científico seleccionado utilizando técnicas de *web scraping* (Mitchell, 2018). Después de eso, se aplica un algoritmo *TextRank* (Mihalcea y Tarau, 2004) para reducir el tamaño del contenido. Este paso permite filtrar el posible texto no relevante. Con respecto al segundo elemento, se encarga de calcular la reputación de cada texto científico de entrada.

El objetivo del módulo de *Gestión del Conocimiento* es generar el léxico de relevancia a partir del conocimiento almacenado en la base de datos mencionada anteriormente. Comprende tres componentes: un *Filtrador de texto*, un *Normalizador de texto* y un *Constructor de lexicones*.

El *Filtrador de texto* es responsable de recopilar los textos almacenados en la base de datos de *Consolidación del Conocimiento Documental* con el fin de aplicar técnicas de minería de texto. Estas técnicas *tokenizan* y filtran todas las palabras para recopilar sustantivos. Obsérvese que la etapa de filtrado implica la eliminación de palabras vacías (Chandramouli, 2018), siendo este proceso totalmente dependiente

del dominio de aplicación real (por ejemplo, en el dominio médico (Gupta, 2015), palabras como *píldora*, *paciente* y *dolor* se consideran como palabras vacías, mientras que, en el dominio académico (Leefmann *et al.*, 2016), *hipótesis*, *base de datos* o *conclusiones* tienen esa consideración). Así, se obtienen los lemas de estos sustantivos. Este enfoque reduce la lista de palabras y también aumenta el número de coincidencias en el léxico final.

El *Normalizador de texto* produce una matriz término-documento (Anandarajan *et al.*, 2019) para calcular la medida TF-IDF (Ramos *et al.*, 2003). Los pesos resultantes se combinan con las reputaciones de los artículos con el fin de obtener la relevancia de cada término para un año específico. Así, la relevancia $rel_lex_t^{(y)}$ de un término t para los $N^{(y)}$ artículos publicados en el año y y pertenecientes al corpus es:

$$rel_lex_t^{(y)} = \log \left(\frac{1}{N^{(y)}} \cdot \sum_{p=1}^{N^{(y)}} (\beta \cdot tfidf_{t,p} + (1 - \beta) \cdot rep_p) \right), \quad (4.7)$$

donde rep_p es la reputación del artículo p , $tfidf_{t,p}$ es el valor TF-IDF para el término t en el artículo p y $\beta \in [0, 1]$ representa un parámetro de compensación para modular la importancia relativa de los términos TF-IDF y reputación. Obsérvese que $tfidf_{t,p} \in (0, 1)$, ya que su valor se normaliza utilizando los valores máximo y mínimo del término en el conjunto de artículos donde aparece. Con respecto a la función logaritmo, se aplica para normalizar la distribución.

En el marco de trabajo, la medida TF-IDF se utiliza en cada uno de los documentos procesados. De este modo, cada documento individual ha sido considerado como un corpus específico, y cada una de sus frases ha sido considerada como un documento determinado. De tal manera, se mitiga la alta ocurrencia de un concepto específico en el documento. Esto ajusta la relevancia de un concepto relacionado con una baja ocurrencia en todo el corpus de documentos. Como consecuencia, sólo aquellos conceptos que aparezcan en varios documentos serán relevantes.

Finalmente, el *Constructor de lexicones* completa el desarrollo del léxico de relevancia organizando las palabras alfabéticamente. Este componente es capaz de introducir ajustes adicionales relacionados con el dominio de aplicación proporcionando pesos específicos para enfatizar las palabras que aparecen en un diccionario (por ejemplo, en el dominio de la medicina (Merriam-Webster, 1995)).

Como una funcionalidad adicional, este elemento también otorga la posibilidad de considerar el factor de evolución al almacenar las entradas correspondientes en el léxico de relevancia. Más específicamente, el cálculo de la relevancia para un término particular t y año y podría considerar cálculos de años anteriores para el mismo término t aplicando la siguiente ecuación recursiva:

$$rel_lex_t^{(y)} = \rho \cdot rel_lex_t^{(y)} + (1 - \rho) \cdot rel_lex_t^{(y-1)}, \quad (4.8)$$

donde $rel_lex_t^{(y)}$ muestra la relevancia proporcionada por el marco de trabajo para el término t y el año y , $rel_lex_t^{(y-1)}$ ilustra la relevancia del término para el año anterior en consideración y el parámetro de compensación ρ controla la influencia de ambos factores.

El marco de trabajo propuesto presenta dos alternativas con respecto al cálculo de la relevancia asociada con cada término t : años aislados y años combinados. El primer enfoque obtiene la relevancia de un término t considerando únicamente los artículos correspondientes al año bajo análisis (ver Ecuación 4.7). El segundo método, en cambio, toma como base inicial los resultados derivados del primer procedimiento y actualiza la relevancia asociada a cada término t aplicando una estrategia recursiva (ver Ecuación 4.8). Este último enfoque permite modular la curva de olvido (Averell y Heathcote, 2011) y las posibles tendencias del dominio de aplicación.

4.2.4. Desarrollo del modelo de aprendizaje máquina

Los modelos de CNN (Poria *et al.*, 2015) se incluyen para respaldar el marco de trabajo y proporcionar cálculos de relevancia para aquellas oraciones que no presenten coincidencias con términos almacenados en el léxico. Análogamente al concepto de diseño del léxico, el marco de trabajo soporta la creación y uso de diferentes redes neuronales que contienen información relacionada con años específicos. Para este propósito, se entrenan redes neuronales separadas (una CNN por año). Obsérvese que los modelos CNN han sido seleccionados debido a que suelen ser mejores y más rápidos que otros tipos de redes neuronales para reconocer patrones clave (Lopez y Kalita, 2017).

La configuración de las capas (ver Tabla 4.1) es una variación mejorada de un enfoque bien conocido utilizado previamente en el dominio del PLN para proporcionar valores de relevancia (Bhavsar *et al.*, 2017).

Con respecto a la arquitectura de la CNN, la capa de *embeddings* utiliza una entrada de modelo pre-entrenado de Glove (Pennington *et al.*, 2014) con un vocabulario de 400.000 palabras tomadas de Wikipedia (O'Sullivan, 2016). La capa de salida contiene una función de activación *softmax* que proporciona valores entre 0 y 1. En las capas ocultas, también se han incluido convoluciones, funciones de activación *relu* y *dropouts*.

Se propone la siguiente metodología para entrenar la red neuronal. Primero, el usuario selecciona un corpus de documentos que no ha sido utilizado durante el proceso de construcción del léxico de relevancia. Después de este paso, cada texto

Capas
1. Embedding input_dim 400.000 output_dim 50
2. Dropout rate 0,4
3. Conv1D 250 filters of 3 with stride 1
4. Pool1D (max) with stride 1
5. Dense units 250
6. Dropout rate 0,4
7. Relu
8. Dense units 1
9. Softmax

Tabla 4.1: *Capas de la red neuronal convolucional.*

científico se divide en oraciones. Estas oraciones se procesan (para crear *tokens*) y se lematizan (se eliminan las palabras vacías encontradas). Este procedimiento origina listas de *tokens* que pueden ser evaluados por el léxico de relevancia. Una vez completada esta tarea, las oraciones con una relevancia promedio superior a un umbral de relevancia ϵ se etiquetan como 1 (es decir, relevantes). Esto facilita el proceso de selección de las oraciones con un valor de relevancia específico. El resto de las oraciones que muestran valores por debajo del umbral o que no tienen palabras relevantes se etiquetan como 0 (es decir, irrelevantes). Finalmente, las dos clases (relevantes e irrelevantes) deben equilibrarse seleccionando el mismo número de instancias. El modelo final debería ser capaz de predecir correctamente cuáles son los textos más relevantes.

4.3. Experimentos

Se ha seleccionado el dominio de la medicina con el objetivo de ejemplificar y ratificar la validez de la propuesta.

Dado un conjunto de documentos previamente evaluados por expertos en el dominio (es decir, investigadores médicos), los experimentos se centran tanto en evaluar la precisión del marco de trabajo *Webelance* en la estimación de la relevancia del documento como en verificar la sensibilidad general del sistema a variaciones en los valores de los parámetros internos.

Cuatro grupos diferentes de 10 documentos (para el primer y segundo experimentos) y 5 documentos específicos relacionados con la alopecia androgénica (para el tercer experimento) pertenecientes a los años 2015, 2016, 2017 y 2018 respectivamente han sido evaluados por 9 expertos procedentes de IBSAL durante la fase de preparación de los experimentos. En los dos primeros experimentos, por cada conjunto, 5 artículos han sido identificados como *relevantes* y 5 como *no relevantes* por

consenso. En el tercer experimento, los expertos no etiquetan los documentos. Todos los documentos presentan DOI disponible (por lo tanto, la relevancia calculada por el marco de trabajo *Webelance* sigue la Ecuación 4.6).

Los experimentos propuestos requieren una preparación previa del léxico y de los artefactos de la red neuronal antes de proceder. A pesar de que ambos elementos difieren en los valores de parámetros seleccionados, se comparten algunas líneas base y estrategias comunes. Este hecho otorga la capacidad de comparar los resultados de los experimentos y, en última instancia, deducir las conclusiones adecuadas con respecto al rendimiento del marco de trabajo *Webelance*.

Para el artefacto léxico, se han descargado y procesado cuatro conjuntos de 15.000 artículos pertenecientes a los años comprendidos entre 2015 y 2018, ambos inclusive. Se ha aplicado el algoritmo TextRank usando un factor de compresión del 20 % (ver Sección. 4.2.3). La relevancia de cada término ha sido calculada de manera recursiva (ver Ecuación 4.8), lo que ha llevado a considerar la relevancia almacenada de años anteriores (revelando tendencias y las fluctuaciones de relevancia a lo largo de los años). Nótese que los cálculos de relevancia correspondientes a 2015 no disponen del factor de memoria de años anteriores, ya que es el primer año considerado. Finalmente, cabe señalar que todos los artículos presentan DOI.

Con respecto a las redes neuronales, se han etiquetado previamente cuatro conjuntos de 320.000 oraciones extraídas de artículos pertenecientes a los años entre 2015 y 2018 como relevantes y no relevantes por el léxico de relevancia antes de proceder a crear las cuatro redes neuronales (es importante señalar que los artículos de los que se extraen estas oraciones no se han utilizado para crear el léxico). Para cada conjunto, el 80 % de las oraciones se utiliza para entrenar cada red neuronal y el 20 % restante se emplea para test. Se toma una decisión de diseño para establecer que la condición previa para que una oración se considere relevante es que debe contener un mínimo de 4 sustantivos y debe tener un valor mínimo de 0,7 para su relevancia asociada. La justificación detrás de este valor límite se basa en el hecho de que parece ser el punto de partida para la mayoría de los términos más relevantes en la población del diccionario médico. De esta manera, los conceptos considerados por el léxico se han utilizado para etiquetar las oraciones con el fin de entrenar las redes neuronales.

Además, ninguno de los 45 artículos evaluados por los expertos médicos se ha utilizado durante la fase de preparación del léxico y las redes neuronales, asegurando un análisis independiente y una exposición aislada de todo el marco de trabajo a esas entradas.

En todos los cálculos resultantes, los valores de relevancia van de 0 (baja relevancia) a 1 (alta relevancia). Un valor de relevancia menor a 0,5 indica que el artículo bajo

Parámetro	Valor de saturación
Citations	10
Influential citation count	208
Citation velocity	105
Seniority	34
Papers	196

Tabla 4.2: Valores de saturación de los parámetros de las Ecuaciones 4.3 y 4.5.

estudio no es relevante. En cuanto a la reputación, también va de 0 (baja reputación) a 1 (alta reputación).

4.3.1. Parámetros neutrales

Este experimento evalúa los resultados derivados del marco de trabajo *Webelance* cuando se seleccionan valores neutrales para los parámetros (es decir, se da igual peso a todos los términos de la ecuación).

Con respecto al léxico, el parámetro β se establece en 0,5 para equilibrar de manera uniforme la contribución de los términos que componen la ecuación (ver Ecuación 4.7). En cuanto a los parámetros necesarios para calcular el término rep_p (ver Ecuación 4.3), el parámetro α también se establece en 0,5 para asignar el mismo peso a la reputación de los autores y al número de citas (evitando una penalización excesiva a los artículos más recientes). En lo que respecta al término de reputación de los autores (ver Ecuación 4.5), los parámetros ω_n se establecen todos al mismo valor, siendo por lo tanto $\omega_1 = \omega_2 = \omega_3 = \omega_4 = 0,25$. Finalmente, en relación con el parámetro ρ necesario para proceder con los cálculos recursivos (ver Ecuación 4.8), se establece en 0,5 para dar la misma importancia a los cálculos de relevancia del año actual y de años anteriores.

Cabe señalar que una vez que se fija el parámetro α , algunos valores de saturación deben ser establecidos para ciertos parámetros (ver Ecuación 4.3 y Ecuación 4.5). Con este propósito, el percentil 90 de cada parámetro se calcula a partir de una muestra independiente de 5.000 artículos pertenecientes al dominio de la medicina (ver Tabla 4.2).

Una vez completado el proceso de preparación del léxico y las redes neuronales, el paso final a seguir antes de testear todo el marco de trabajo consiste en fijar el parámetro θ (ver Ecuación 4.2) y el parámetro γ (ver Ecuación 4.6). Debido a la naturaleza del experimento, θ se establece en 0,5 para dar igual importancia al léxico y a las predicciones de la red neuronal, y γ se establece de forma análoga en 0,5 para fijar igual peso en los cálculos de relevancia provenientes de los artefactos (léxico y red neuronal) y de la reputación del artículo.

ID	DOI	Relevante	Relevancia predicha
1	10.2147/JMDH.S93254	no	0,224
2	10.2147/CIA.S90296	no	0,249
3	10.3109/03009734.2015.1027430	no	0,283
4	10.5713/ajas.14.0403	no	0,292
5	10.4103/2230-973X.167696	no	0,292
6	10.2337/dc15-1174	sí	0,570
7	10.1007/s40629-015-0067-z	sí	0,562
8	10.1111/ene.12724	sí	0,579
9	10.1038/nm.3911	sí	0,567
10	10.1084/jem.20141555	sí	0,524

Tabla 4.3: Comparación de la opinión de los expertos con la relevancia predicha por *Webelance* para artículos publicados durante 2015.

Con todos los parámetros establecidos, el marco *Webelance* se testea con los artículos previamente evaluados por los expertos en el campo médico para verificar la correcta alineación del sistema con sus evaluaciones. Las Tablas 4.3 a 4.6 ilustran los resultados obtenidos para los cuatro conjuntos de 10 documentos pertenecientes a los años 2015-2018. En cada tabla, la primera columna asigna un valor de ID a cada artículo por facilidad de referencia, la segunda columna denota el DOI del artículo en particular, la tercera columna indica si los expertos consideran que el artículo es relevante y, finalmente, la cuarta columna muestra las cifras resultantes obtenidas del marco de trabajo *Webelance* después de su procesamiento. Como se indicó anteriormente, valores de relevancia por encima de 0,5 indican que el artículo en cuestión es relevante, mientras que valores por debajo de 0,5 indican lo contrario.

Los resultados de *Webelance* para 2015 concuerdan con la opinión de los expertos en el 100 % de los casos (ver Tabla 4.3). En cuanto a 2016, el sistema proporciona una evaluación diferente para el artículo referido por ID=6, lo que da lugar a una precisión del 90 % (ver Tabla 4.4). Se encuentra la misma precisión del 90 % para 2017, esta vez con una discrepancia en el artículo referido por ID=9 (ver Tabla 4.5). Por último, los resultados para 2018 presentan una precisión del 70 %, encontrando fallos en la evaluación en los artículos con ID=6, ID=7 e ID=10 (ver Tabla 4.6).

En resumen, una configuración de parámetros neutral demuestra un rendimiento de evaluación aceptable, alcanzando un total del 87,5 % de exactitud (*accuracy*) general durante la evaluación de los 40 artículos científicos revisados. Se logró también una precisión del 100 %, es decir, todos los artículos predichos como relevantes eran en realidad artículos relevantes. Finalmente, se alcanzó un índice de exhaustividad del 75 % (porcentaje de artículos relevantes correctamente predichos como tales).

ID	DOI	Relevante	Relevancia predicha
1	10.1016/j.atg.2016.03.005	no	0,261
2	10.1136/bmjopen-2015-010054	no	0,239
3	10.1136/rmdopen-2015-000222	no	0,269
4	10.1136/bmjopen-2015-009929	no	0,305
5	10.1136/bmjopen-2016-012286	no	0,291
6	10.1136/bmjopen-2016-011749	sí	0,343
7	10.4062/biomolther.2015.172	sí	0,569
8	10.1136/bmjopen-2015-010159	sí	0,530
9	10.1038/bjc.2016.17	sí	0,562
10	10.1038/ni.3462	sí	0,558

Tabla 4.4: Comparación de la opinión de los expertos con la relevancia predicha por Webelance para artículos publicados durante 2016.

ID	DOI	Relevante	Relevancia predicha
1	10.1136/bmjopen-2016-015110	no	0,193
2	10.1136/bmjopen-2017-016689	no	0,241
3	10.1080/10810730.2017.1414901	no	0,242
4	10.1136/bmjopen-2017-018856	no	0,287
5	10.1136/bmjopen-2016-014216	no	0,385
6	10.1038/s41559-017-0338-9	sí	0,501
7	10.1136/bmjdr-2016-000317	sí	0,532
8	10.1016/j.ymthe.2017.06.002	sí	0,588
9	10.12659/MSM.898297	sí	0,497
10	10.2337/dc17-1185	sí	0,511

Tabla 4.5: Comparación de la opinión de los expertos con la relevancia predicha por Webelance para artículos publicados durante 2017.

ID	DOI	Relevante	Relevancia predicha
1	10.1016/j.parepi.2018.e00073	no	0,229
2	10.1136/bmjopen-2017-020854	no	0,237
3	10.1136/bmjopen-2017-018513	no	0,244
4	10.1136/bmjopen-2018-025073	no	0,320
5	10.1016/j.celrep.2018.04.036	no	0,354
6	10.1038/s41559-018-0575-6	sí	0,449
7	10.4062/biomolther.2018.103	sí	0,366
8	10.1016/j.apsb.2018.01.008	sí	0,506
9	10.1038/s41588-018-0154-4	sí	0,521
10	10.1016/j.peptides.2017.12.015	sí	0,442

Tabla 4.6: Comparación de la opinión de los expertos con la relevancia predicha por Webelance para artículos publicados durante 2018.

4.3.2. Ajuste fino de parámetros

El segundo experimento se centra en la robustez del marco de trabajo. El experimento anterior ha proporcionado resultados de rendimiento de referencia basados en una selección neutral de parámetros. Entonces, la pregunta a responder aquí es si estos resultados de referencia se mejorarían mediante una selección adecuada de parámetros (realizada por expertos).

Con respecto a la creación del artefacto léxico, el parámetro β (ver Ecuación 4.7) y el parámetro α (ver Ecuación 4.3) se establecen en 0,5 al igual que en el experimento anterior. Sin embargo, guiados por las opiniones de los expertos, los parámetros ω_n se fijan de la siguiente manera: $\omega_1 = 0,2$, $\omega_2 = 0,1$, $\omega_3 = 0,3$, $\omega_4 = 0,4$. Es decir, se asignan pesos mayores a la antigüedad del investigador y al número de artículos publicados. Con respecto al parámetro ρ , ahora adopta un valor de 0,8 para dar menos peso a los cálculos de relevancia de los años anteriores para cada término específico. En cuanto a los parámetros en las Ecuaciones 4.3 y 4.5, los valores de saturación se establecen como se muestra en la Tabla 4.2.

Al igual que en el experimento anterior, una vez que se fijan todos los parámetros léxicos, el artefacto léxico de relevancia se puede crear utilizando los cuatro conjuntos de 15.000 artículos pertenecientes a los años comprendidos entre 2015 y 2018.

Dado que el constructor de léxico puede introducir ajustes adicionales, se utiliza un diccionario médico abierto (Merriam-Webster, 1995) como herramienta de apoyo externa durante el proceso de construcción del léxico para fortalecer la relevancia asociada con los términos estrictamente médicos almacenados en el léxico. Por lo tanto, aquellos términos léxicos también presentes en este diccionario externo ven su relevancia multiplicada por 0,6. Para aquellos términos léxicos no encontrados en él, su relevancia se multiplica por 0,4. Esta pequeña transformación contribuye a obtener una distribución bimodal para la relevancia de los términos almacenados en el léxico para cada año bajo estudio (ver Figura 4.4).

En relación con los artefactos de redes neuronales, se reproducen los mismos fundamentos y procedimientos aplicados en el experimento anterior durante este.

Finalmente, se aborda la asignación de valores para los parámetros θ y γ . Se fija θ a 0,8 para dar más peso a los resultados del léxico y reforzar la idea del papel de apoyo de la red neuronal. El parámetro γ se establece en 0,6 para proporcionar la mayor parte del peso a la reputación resultante.

Una vez establecidos todos los parámetros requeridos, el marco *Webelance* se prueba nuevamente con el mismo conjunto de 40 artículos previamente evaluados por expertos en el campo médico. Las Tablas 4.7 a 4.10 ilustran los resultados obtenidos para los cuatro conjuntos de 10 documentos pertenecientes a los años de 2015 a

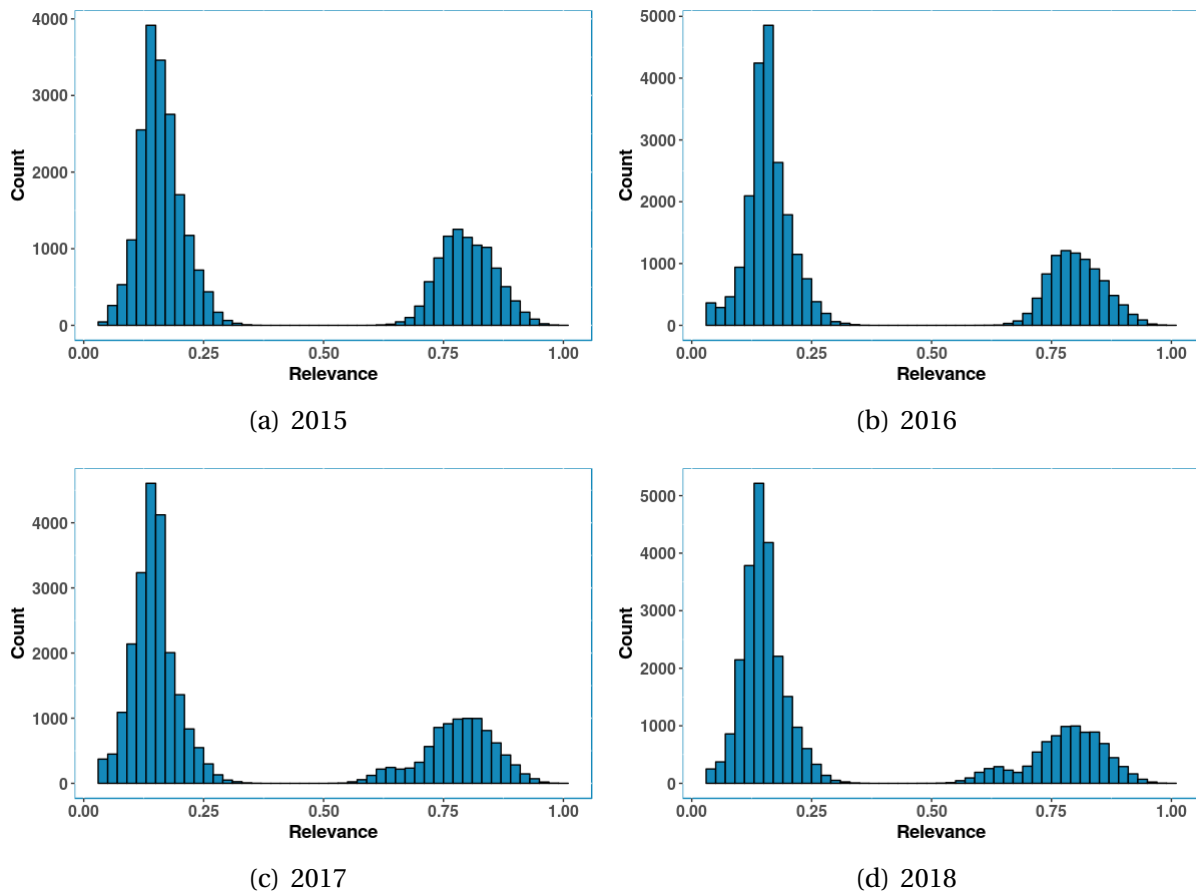


Figura 4.4: Representación de las cuatro distribuciones bi-modales de la relevancia de artículos para cada año.

2018.

Los resultados de 2015 muestran que el sistema *Webelance* está completamente alineado con los expertos. También es pertinente señalar que todas las cifras están lejos del valor neutral de 0,5, lo que demuestra un rendimiento positivo alejándose de resultados dudosos. Con respecto a 2016, el sistema concuerda con los expertos excepto para el artículo identificado con ID=6 (cuyo valor de relevancia predicho es menor a 0,5 por un pequeño margen). Las investigaciones adicionales sobre este caso particular revelan que el artículo corresponde a una revisión que contiene varios términos médicos diferentes que producen ruido. Sin embargo, el resto de los resultados están claramente separados de la neutralidad (es decir, el valor 0,5), dando una exactitud (*accuracy*) general del 90 % para 2016. Se obtiene el mismo rendimiento para el 2017, teniendo un 90 % de exactitud y una única relevancia predicha que no coincide con la opinión de los expertos (artículo ID=5). El artículo corresponde a una revisión con los mismos problemas que el perteneciente a 2016. Finalmente, los resultados de 2018 coinciden en un 100 % con las opiniones de los expertos. Es importante señalar que algunos de los valores predichos para este año están cerca del

ID	DOI	Relevante	Relevancia predicha
1	10.2147/JMDH.S93254	no	0,319
2	10.2147/CIA.S90296	no	0,355
3	10.3109/03009734.2015.1027430	no	0,394
4	10.5713/ajas.14.0403	no	0,401
5	10.4103/2230-973X.167696	no	0,403
6	10.2337/dc15-1174	sí	0,733
7	10.1007/s40629-015-0067-z	sí	0,736
8	10.1111/ene.12724	sí	0,738
9	10.1038/nm.3911	sí	0,747
10	10.1084/jem.20141555	sí	0,800

Tabla 4.7: Comparación de la opinión de expertos con la relevancia predicha por el marco *Webelance* para artículos publicados durante 2015.

valor 0,5, lo que indica que algunos artículos son propensos a generar dudas durante sus evaluaciones.

En conclusión, esta segunda configuración de parámetros resulta en un muy buen rendimiento en la evaluación, logrando una exactitud total del 95 % durante la evaluación de los 40 artículos científicos bajo revisión. En este caso, la precisión y la exhaustividad alcanzan el 95 %.

Tomando en consideración los resultados obtenidos en ambos experimentos, la mejora lograda por la segunda configuración con respecto a la configuración neutral es del 7,5 % (en exactitud general). Esto era de esperar ya que la segunda configuración fue ajustada por expertos en el dominio. Sin embargo, esta mejora implica que los resultados de referencia (87,5 % de exactitud general) son lo suficientemente altos como para considerar el marco de trabajo robusto. Así, los resultados proporcionan aplicabilidad directa y evidencia de apoyo para deducir que el sistema no presenta alta sensibilidad a los valores seleccionados para sus parámetros internos. En términos generales, tener sistemas con baja dependencia en la configuración de parámetros internos representa una característica altamente valorada a la que todos los marcos funcionales aspiran. Por lo tanto, los resultados derivados de los experimentos realizados refuerzan y demuestran verdaderamente los sobresalientes atributos internos que el marco de trabajo ha presentado. Este último hecho, considerado junto con la novedad de la propuesta, prueba que el marco *Webelance* está en una muy buena posición para abordar futuras líneas de mejora.

4.3.3. Validación a lo largo del tiempo

En este experimento, se estudian los cambios en la relevancia de los artículos con el paso del tiempo. Para ilustrar el experimento, se ha seleccionado un campo especí-

ID	DOI	Relevante	Relevancia predicha
1	10.1016/j.atg.2016.03.005	no	0,210
2	10.1136/bmjopen-2015-010054	no	0,334
3	10.1136/rmdopen-2015-000222	no	0,368
4	10.1136/bmjopen-2015-009929	no	0,392
5	10.1136/bmjopen-2016-012286	no	0,397
6	10.1136/bmjopen-2016-011749	sí	0,466
7	10.4062/biomolther.2015.172	sí	0,699
8	10.1136/bmjopen-2015-010159	sí	0,704
9	10.1038/bjc.2016.17	sí	0,711
10	10.1038/ni.3462	sí	0,718

Tabla 4.8: Comparación de la opinión de expertos con la relevancia predicha por el marco Webelance para artículos publicados durante 2016.

ID	DOI	Relevante	Relevancia predicha
1	10.1136/bmjopen-2016-015110	no	0,283
2	10.1136/bmjopen-2017-016689	no	0,343
3	10.1080/10810730.2017.1414901	no	0,344
4	10.1136/bmjopen-2017-018856	no	0,386
5	10.1136/bmjopen-2016-014216	no	0,509
6	10.1038/s41559-017-0338-9	sí	0,689
7	10.1136/bmjdr-2016-000317	sí	0,695
8	10.1016/j.ymthe.2017.06.002	sí	0,714
9	10.12659/MSM.898297	sí	0,769
10	10.2337/dc17-1185	sí	0,778

Tabla 4.9: Comparación de la opinión de expertos con la relevancia predicha por el marco Webelance para artículos publicados durante 2017.

ID	DOI	Relevante	Relevancia predicha
1	10.1136/bmjopen-2017-018513	no	0,283
2	10.1016/j.parepi.2018.e00073	no	0,326
3	10.1136/bmjopen-2017-020854	no	0,331
4	10.1136/bmjopen-2018-025073	no	0,438
5	10.1016/j.celrep.2018.04.036	no	0,449
6	10.1038/s41559-018-0575-6	sí	0,540
7	10.4062/biomolther.2018.103	sí	0,616
8	10.1016/j.apsb.2018.01.008	sí	0,676
9	10.1038/s41588-018-0154-4	sí	0,687
10	10.1016/j.peptides.2017.12.015	sí	0,704

Tabla 4.10: Comparación de la opinión de expertos con la relevancia predicha por el marco Webelance para artículos publicados durante 2018.

ID	DOI	2015	2016	2017	2018
1	10.4103/0973-6247.162713	0,646	0,642	0,633	0,639
2	10.4103/0974-7753.160098	0,593	0,591	0,584	0,589
3	10.4103/0970-0358.155271	0,678	0,662	0,650	0,678
4	10.1016/j.jdc.2015.07.004	0,635	0,630	0,615	0,617
5	10.5966/sctm.2015-0107	0,680	0,676	0,667	0,672

Tabla 4.11: Relevancia a lo largo del tiempo de artículos sobre el tratamiento de la alopecia androgénica.

fico relacionado con el dominio de la medicina: el tratamiento de la alopecia androgénica. La Tabla 4.11 muestra la relevancia de 5 artículos que tratan sobre este campo en la base de datos. En este caso, los parámetros del sistema son los propuestos por los expertos en el dominio de la medicina (ver Sección 4.3.2).

Los resultados del experimento muestran que la relevancia cambia con el tiempo siguiendo la misma tendencia para todos los artículos. Considerando todos los artículos bajo análisis, la relevancia máxima se alcanza en 2015, mientras que la mínima se logra en 2017. Durante los primeros tres años (2015, 2016 y 2017), el artículo más relevante es el identificado con ID=5. En cuanto a 2018, el artículo identificado por ID=3 muestra la mayor relevancia. Obsérvese que el artículo con ID=5 recibió un total de 3, 5, 15 y 9 citas respectivamente durante el período considerado. Para el mismo período, el artículo con ID=3 recibió 1, 12, 34 y 35 citas. Por lo tanto, queda claro que la relevancia del artículo con ID=3 fue afectada positivamente por el aumento en el número de citas. Por el contrario, la relevancia del artículo con ID=5 fue afectada negativamente por la disminución en el número de citas. A pesar de lo anterior, el número de citas no proporciona suficiente contexto para entender la relevancia. Como ejemplo, el número de citas del artículo con ID=3 es mayor que el número de citas del artículo con ID=5 en 2016 y 2017. Sin embargo, la relevancia calculada por el marco de trabajo *Webelance* muestra valores más bajos para el artículo con ID=3 para ambos años.

Considerando los valores de relevancia obtenidos en 2015, 2016 y 2017, el sistema recomendaría a un investigador novel seleccionar el artículo identificado por ID=5 ya que presenta la relevancia más alta. Sin embargo, el sistema recomendaría el artículo identificado por ID=3 cuando el año en consideración es 2018 (ya que muestra la mayor relevancia). Obsérvese que este cambio sería difícil de detectar por un mentor humano.

Los resultados del experimento permiten concluir que el marco *Webelance* es capaz de capturar las fluctuaciones y la evolución de la relevancia de los artículos a lo largo del tiempo. Es importante destacar que, gracias a la funcionalidad de actualización incluida en la Ecuación 4.8, el sistema proporciona una respuesta suave a

posibles cambios abruptos en la relevancia de los artículos.

4.4. Conclusiones

En este capítulo se ha introducido el marco de trabajo *Webelance* para ofrecer una solución a un problema existente en la comunidad científica: estimar la relevancia de artículos científicos. La relevancia (es decir, el grado en que algo está relacionado o es útil para lo que está ocurriendo o de lo que se está hablando) es un concepto subjetivo y, por lo tanto, es casi imposible de fijar en un momento concreto. Sin embargo, hay muchas métricas objetivas utilizadas para estimarlo. Estas estimaciones son objetivas, pero no son únicas. Por lo tanto, un marco de trabajo para combinarlas y proporcionar un único valor de relevancia resulta crucial.

El sistema propuesto presenta, como principal novedad, la capacidad de medir la relevancia de un texto científico perteneciente a un área de investigación específica. Obsérvese que *Webelance* puede ser entrenado para cualquier dominio específico de interés. Lo cual es señal de flexibilidad y usabilidad de cara a potenciales usuarios.

El sistema es capaz de estimar valores de relevancia según un conjunto de características recopiladas y palabras clave del dominio. Para llevar a cabo estas tarea, implementa un diccionario y modelos de ML para hacer predicciones cuando el diccionario no puede estimar la relevancia apropiada por sí solo. Ambos artefactos se complementan con un sistema que mide la reputación de los artículos según la reputación de los autores y el número de citas.

Se han desarrollado varios experimentos para probar la viabilidad de la propuesta. En este trabajo, se han centrado en el dominio de la medicina debido a su importancia para los seres humanos y la extensa investigación disponible en el área. Los parámetros del sistema fueron ajustados para construir adecuadamente tanto el diccionario como los artefactos de la red neuronal. La red neuronal ha demostrado su importancia en aquellos casos en los que los documentos presentaron palabras inusuales del dominio. Por otro lado, el cálculo de la reputación del artículo evaluado también ha demostrado su valor al reforzar los resultados proporcionados por el sistema. El sistema ha demostrado un buen rendimiento evaluando artículos previamente etiquetados por expertos en el dominio. Además, la evolución en el tiempo de la relevancia calculada en artículos de un campo específico de la medicina ha mostrado cómo el sistema puede detectar tendencias. Estas tendencias pueden ser utilizadas para seleccionar los mejores artículos para un año específico y hacer recomendaciones a investigadores noveles. Todos estos hechos dan lugar a la conclusión de que el prototipo es robusto y cumple satisfactoriamente sus propósitos.

El marco de trabajo *Webelance* ha sido concebido como una herramienta de pro-

pósito general. En este capítulo, se ha demostrado su utilidad en el dominio médico. Sin embargo, su aplicación en otros campos requeriría el ajuste de los parámetros según el conocimiento proporcionado por los expertos correspondientes. Además, se necesitaría la selección de otros conjuntos de artículos científicos para entrenar los léxicos y los modelos ML del sistema.

El estudio presentado incluye una limitación relacionada con el número de artículos disponibles previamente evaluados por expertos en el campo. Estos artículos se utilizan para evaluar el rendimiento del sistema y deberían incluirse experimentos más avanzados. Por lo tanto, los trabajos futuros deberían abordar este problema y proporcionar al sistema una base de datos de artículos evaluados. Este enfoque resultaría tanto en una posible detección de otras posibles debilidades del sistema y, lo que es más importante, en una probable perspectiva para encontrar una configuración óptima de parámetros. Finalmente, la evolución de la relevancia con el tiempo también se ha limitado a sólo cuatro años. Por lo tanto, son deseables futuros experimentos en los que se consideren períodos de tiempo más largos para mostrar la evolución temporal de la relevancia y la aparición de tendencias.

En el próximo capítulo, el marco de trabajo *Webelance* se integra en un marco de trabajo completo capaz de brindar un apoyo más automatizado a la comunidad científica. En el dominio específico de la medicina, la arquitectura propuesta se incluye en un sistema de recomendación específico para investigadores y profesionales de la medicina. Con respecto a las posibles actualizaciones y las futuras acciones que se podrían seguir, se debería considerar un conjunto más grande de textos científicos previamente evaluados por expertos en el dominio con el objetivo de realizar pruebas continuas del sistema y obtener valores óptimos para los parámetros de compensación del sistema. Además, la evaluación de métricas similares a TF-IDF capaces de generar perspectivas alternativas podría suponer otro ámbito de exploración. Una evaluación de los cambios de relevancia en el tiempo con el fin de validar la detección de posibles tendencias sobre un tema específico es otro punto de investigación futura. Finalmente, también podría ser interesante evaluar la posibilidad de producir diccionarios con n-gramas en lugar de solo uni-gramas. Vale la pena señalar que las pruebas anteriores relacionadas con este punto mostraron un mejor rendimiento en el caso de los uni-gramas, pero se considerarán experimentos al respecto en futuros trabajos.

Capítulo 5

Recomendador de artículos

En este capítulo se introduce MELENDI, el segundo componente principal del asistente de investigación médica basado en conocimiento que compone esta tesis. En el capítulo anterior se presentó *Webalance* como punto de partida para la estimación de relevancia de artículos, MELENDI combina la relevancia estimada por *Webalance* con un modelo de similitud semántica basado en una arquitectura de tipo *transformer* para proporcionar recomendaciones basadas en las características de los pacientes del usuario. Para extraer estas características, MELENDI cuenta con un módulo de extracción automática de diagnósticos provenientes de historiales clínicos electrónicos que combina patrones lingüísticos con un modelo pre-entrenado de aprendizaje máquina para realizar dicha tarea. En adelante, se introduce la propuesta y los experimentos realizados para validarla.

5.1. Introducción

Los profesionales de la salud tienen la obligación de tratar a sus pacientes, pero también de mantenerse actualizados en sus respectivas especialidades para asegurar los tratamientos más recientes basados en evidencia científica.

La inmensa y creciente cantidad de literatura científica, principalmente proveniente de revistas, congresos, ensayos clínicos y directrices que aparecen cada día, hacen que esta tarea sea cada vez más difícil (R. Johnson *et al.*, 2018). Este hecho se ve agravado en el caso de los últimos lanzamientos de medicamentos, avances médicos y técnicas, que requieren mucho tiempo para discriminar los enfoques más relevantes relacionados con sus respectivos campos.

La falta de actualizaciones de conocimiento basadas en buenas evidencias para los profesionales de atención médica tiene un impacto potencial en la calidad de los

tratamientos y en la posibilidad de salvar vidas. Esta situación se ha exacerbado con el brote de COVID-19, que ha dado lugar una explosión de literatura científica sobre la enfermedad y el virus (Gianola *et al.*, 2020), con una extrema congestión de los servicios de atención médica (Olivas-Martínez *et al.*, 2021).

La gravedad del virus y la necesidad de enfrentar una pandemia global de la manera más rápida posible ha llevado a empeorar la calidad de las publicaciones de investigación (Raynaud *et al.*, 2021). Por lo tanto, el desarrollo de un sistema que brinde apoyo a los profesionales de la salud para mantenerse al día en sus campos en el menor tiempo posible, ofreciendo publicaciones relevantes basadas en las características de sus pacientes, se ha convertido en un tema clave. Obsérvese que estas características son típicamente la información nosológica (es decir, clasificación de enfermedades) y el diagnóstico realizado, ya que estos elementos suelen incluir los detalles más relevantes de posibles enfermedades y síndromes.

Este capítulo presenta el marco de trabajo MELENDI. Consiste en un novedoso sistema que funciona automáticamente extrayendo diagnósticos de EHR para después buscar artículos relevantes relacionados con dichos diagnósticos y finalmente recomendarlos al usuario de forma pro-activa.

MELENDI podría integrarse potencialmente en un *Hospital Information System* (HIS) o ser consumido directamente por el médico a través de una aplicación móvil o web. Este hecho promueve un fácil acceso para los profesionales de la salud, permitiendo consultar artículos relevantes relacionados con las enfermedades y síndromes de sus pacientes, dando lugar a un valioso ahorro de tiempo.

En cuanto a la arquitectura del sistema, presenta cinco módulos que abordan cinco tareas específicas: procesamiento de artículos, detección de diagnósticos a partir de resúmenes de alta, cálculo de similitudes entre artículos y diagnósticos, cálculo de la relevancia de artículos y generación de un ranking de resultados.

El sistema ha sido adaptado al formato de los EHR procedentes de MIMIC-III (A. E. Johnson *et al.*, 2016). De esta manera, es capaz de detectar y extraer automáticamente los diagnósticos principales al combinar expresiones regulares con un modelo de detección de conceptos médicos (Kraljevic *et al.*, 2021).

Se han llevado a cabo varios experimentos para evaluar la viabilidad del sistema. Un grupo de expertos en el dominio de la salud del IBSAL ha participado en esta tarea. Primero, las pruebas realizadas evalúan los módulos más relevantes del sistema (es decir, aquellos que calculan similitudes entre artículos y detectan los principales diagnósticos en el contenido textual de los resúmenes de alta) y las diferentes decisiones de diseño. Luego, se considera el rendimiento completo del sistema y la calidad de los resultados.

El resto del capítulo está estructurado de la siguiente manera. La Sección 5.2 pre-

senta los detalles del sistema y sus componentes, la Sección 5.3 se centra en los experimentos realizados para validar el sistema y, finalmente, la Sección 5.4 concluye y propone posibles directrices futuras.

5.2. Propuesta de marco de trabajo

El marco de trabajo MELENDI es un sistema completo desarrollado con el objetivo de ahorrar tiempo a los profesionales de la salud. Su principal funcionalidad consiste en hacer recomendaciones de artículos científicos de acuerdo con los diagnósticos detectados en resúmenes de alta específicos.

Para lograr esta tarea, el sistema utiliza la información más relevante recopilada de varios resúmenes de alta de pacientes. Esta información conforma el conocimiento principal del sistema. Para obtener este conocimiento, se utilizan técnicas de procesamiento de lenguaje natural que permiten aislar los diagnósticos del resto de información contenida en los EHR. Estos diagnósticos forman parte del contenido textual, que presenta enfermedades, entidades nosológicas, síndromes y también cualquier condición patológica o de salud de los pacientes. *Unified Medical Language System (UMLS)* (Bodenreider, 2004), proporciona apoyo para lograr la detección y extracción de estos conceptos médicos. Luego, el conocimiento obtenido es utilizado por el sistema para buscar artículos científicos relacionados con estos conceptos de entrada según su similitud semántica. Finalmente, estos artículos se ordenan según un conjunto de características específicas extraídas de los metadatos que miden su relevancia en el dominio de la salud.

En cuanto a la arquitectura del sistema, está organizada en cinco módulos principales: módulo *Pre-procesador de artículos*, módulo *Extractor de diagnósticos*, módulo *Calculador de similitud semántica*, módulo *Calculador de relevancia* y módulo *Recomendador de artículos*. Estos módulos son complementados por dos bases de datos que consolidan la información recopilada de artículos científicos y un módulo de *Visualización* para interactuar con los usuarios (ver Figura 5.1).

El primer módulo se encarga de analizar los resúmenes de publicaciones científicas relacionadas con el ámbito de la salud. El segundo módulo aborda la detección de diagnósticos y las tareas de recopilación de información a partir de los resúmenes de alta proporcionados. El tercer módulo calcula la similitud entre los conceptos de entrada (es decir, los conceptos detectados en los diagnósticos y los resúmenes de los artículos científicos). El cuarto módulo calcula la relevancia de los artículos más similares obtenidos, y el quinto módulo clasifica estos artículos según los valores de similitud y relevancia.

Las bases de datos están formadas por un conjunto de *Artículos científicos* y su

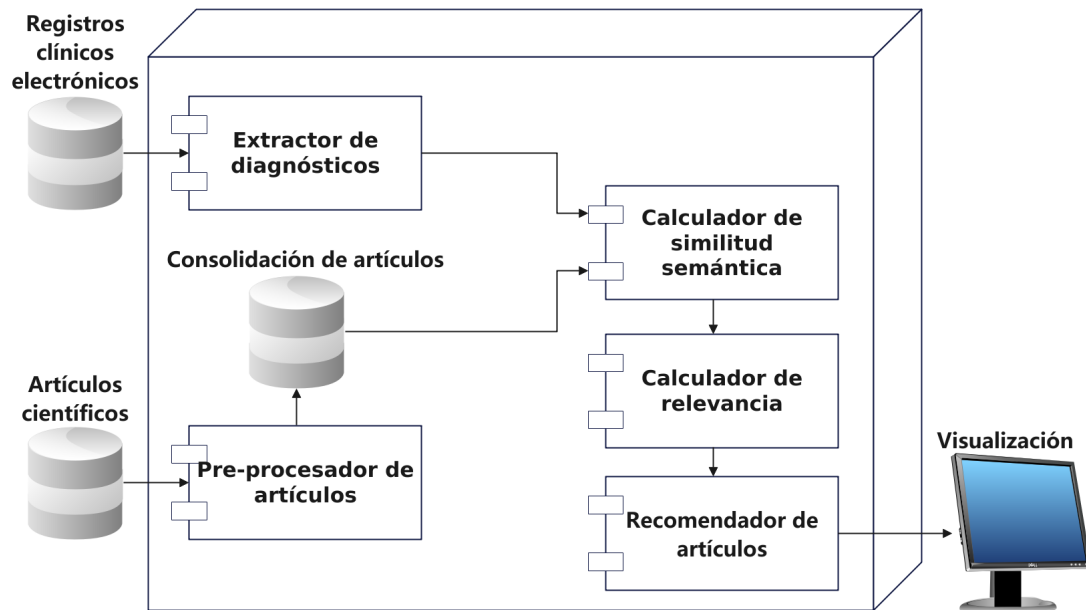


Figura 5.1: Vista general de la arquitectura del marco de trabajo MELENDI.

correspondiente *Consolidación de información* una vez pre-procesados además de un conjunto de historiales clínicos electrónicos. La primera comprende una colección de resúmenes (*abstracts*) a analizar mediante procedimientos semánticos para generar los *embeddings* o incrustaciones de sus oraciones correspondientes. Estos *embeddings* se almacenan en la segunda base de datos para ser utilizados por el sistema en respuesta a las solicitudes realizadas por los usuarios.

5.2.1. Módulo pre-procesador de artículos

La función principal de esta parte del sistema es obtener los vectores semánticos (*embeddings*) de los resúmenes de los artículos. El proceso se realiza en segundo plano una vez al mes. Esto permite actualizar la información sobre el estado del dominio de la salud considerando las últimas novedades científicas.

La fuente de información utilizada para seleccionar y filtrar los artículos científicos es Pubmed (J. White, 2020), ya que es un indexador de artículos científicos de acceso abierto. Por otro lado, el modelo *msmarco-distilbert-base-v4* (Reimers y Gurevych, 2019) ha sido el modelo semántico utilizado para calcular las incrustaciones.

El sistema también obtiene los metadatos de cada artículo. Estos metadatos están formados por el DOI, el año de publicación y los autores. Esta información se utiliza posteriormente por el módulo *Cálculador de Relevancia* para estimar la relevancia de los artículos. Es importante señalar que la recopilación de estos metadatos también se realiza durante el proceso en segundo plano, mientras que la estimación de la reputación se lleva a cabo en tiempo real en respuesta a una solicitud realizada por

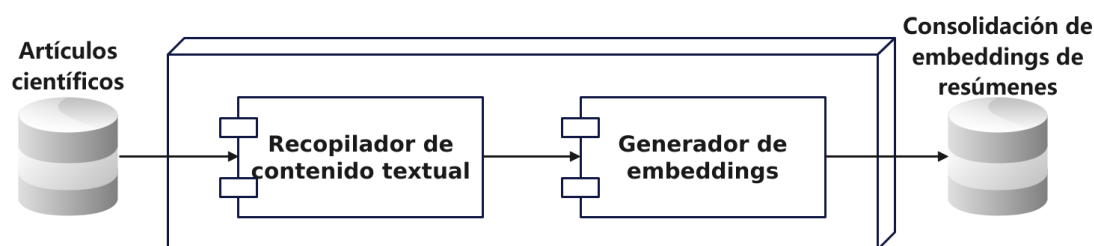


Figura 5.2: *Arquitectura interna del módulo Pre-procesador de artículos.*

el usuario.

En cuanto a la arquitectura del módulo (ver Figura 5.2), presenta dos componentes para abordar sus dos principales tareas: el *Recopilador de contenido textual* y el *Generador de embeddings*. El primer componente extrae los textos y los metadatos, mientras que el segundo genera las incrustaciones asociadas y las almacena en la base de datos de consolidación de *embeddings* de resúmenes.

5.2.2. Módulo extractor de diagnósticos

Este módulo ha sido diseñado específicamente para abordar las particularidades de los EHR, dado que este tipo de informe médico es conocido por su naturaleza no estructurada y compleja, lo que dificulta la extracción de información valiosa. Por lo tanto, el sistema ha sido adaptado para tratar la sección de resúmenes de alta de MIMIC-III en particular. Este último es un corpus médico bien conocido y un estándar a utilizar en varias tareas relacionadas con la atención sanitaria.

El módulo utiliza un enfoque basado en reglas que combina tanto criterios lingüísticos (en forma de búsqueda de ciertas palabras clave) como el uso de patrones textuales específicos de los EHR (en forma de expresiones regulares). Esta estrategia permite la clasificación automática de los datos internos del EHR en diferentes secciones (por ejemplo, Diagnóstico Principal, Fechas y Lugares, o Medicamentos). Este proceso es comúnmente conocido como un problema de identificación de secciones Pomares-Quimbaya *et al.* (2019).

Este enfoque se lleva a cabo a través de tres pasos secuenciales (ver Figura 5.3 para más detalles): *pre-procesamiento de texto*, *extracción de fragmentos relevantes (sección de diagnóstico principal)* y *NER de entidades médicas*.

En el paso de *preprocesamiento de texto*, se aplica una técnica de pre-procesamiento y normalización donde el texto se limpia para eliminar caracteres especiales como \t.

Hasta el paso *segmentación con expresiones regulares*, el contenido textual se divide en párrafos basados en líneas vacías, y luego se dividen aún más los párrafos en

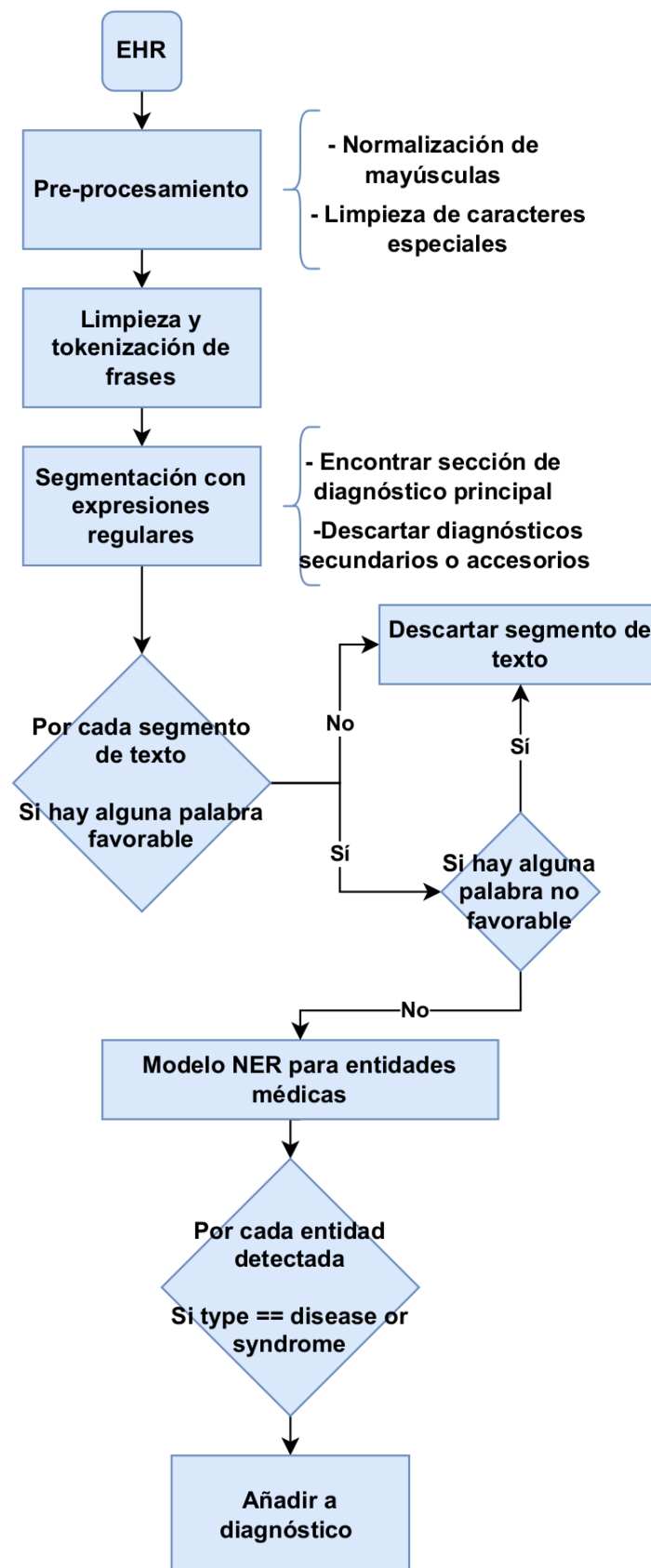


Figura 5.3: *Proceso del módulo* Extractor de diagnósticos.

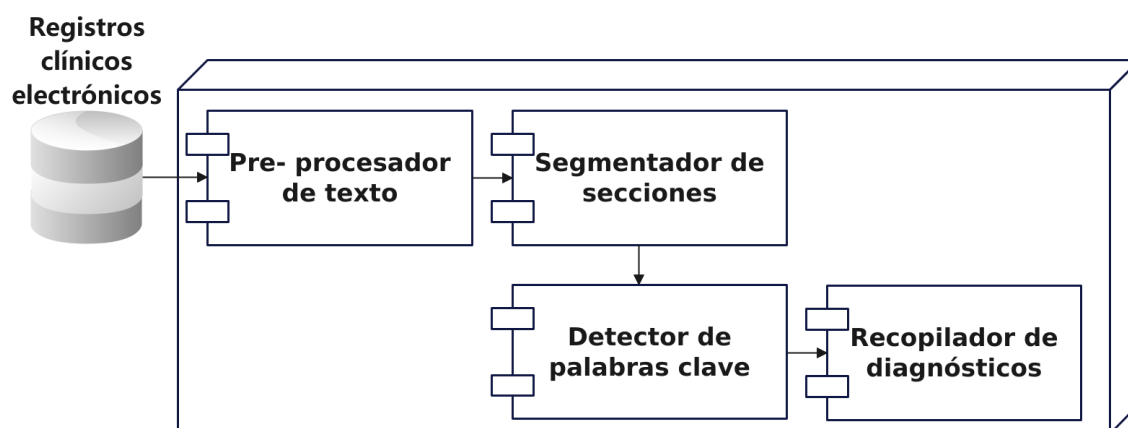


Figura 5.4: *Arquitectura interna del módulo* Extractor de diagnósticos.

secciones en función de la presencia de encabezados de sección. Esta tarea se lleva a cabo utilizando las siguientes dos expresiones regulares:

(RegEx 1) `r"\n\s*\n+"`

(RegEx 2) `r"(?i)(?:^|\n`

`(history\s+of\s+present\s+illness|physical\s+examination|
diagnosis|treatment|medications)
(?=\s+|\n|$)"`

Obsérvese que el objetivo de este paso es aislar cada región de texto y reducir el coste computacional de la herramienta NER que en este punto aún no ha sido utilizada. En el paso de *NER para entidades médicas*, para cada fragmento de texto, un clasificador basado en palabras clave filtra el diagnóstico principal del resto de los datos. Este filtro de palabras clave utiliza palabras favorables y no favorables. Las primeras incluyen aquellas que son propensas a aparecer en procedimientos de diagnóstico (por ejemplo, *diagnosis, diagnosed, symptoms* o *patient presents*). La segunda comprende aquellas que no suelen incluirse en procedimientos de diagnóstico (por ejemplo, *procedures, dates, medications* o *history of*). Cabe señalar que la lista de palabras clave ha sido creada como parte de un proceso heurístico y podría ser refinada en futuras versiones.

Este paso se completa mediante la detección de entidades nombradas del dominio de la medicina a través de la herramienta MedCAT Kraljevic *et al.* (2021). La principal ventaja de esta herramienta es su sistema NER+L, que mapea automáticamente cada una de las entidades encontradas con sus correspondientes metadatos de UMLS. Este proceso de mapeo es el componente clave de este sistema, ya que actúa como un filtro y permite compilar solo las entidades cuyo *semantic type* de UMLS es "T047" (disease or syndrome), que componen el diagnóstico. Obsérvese

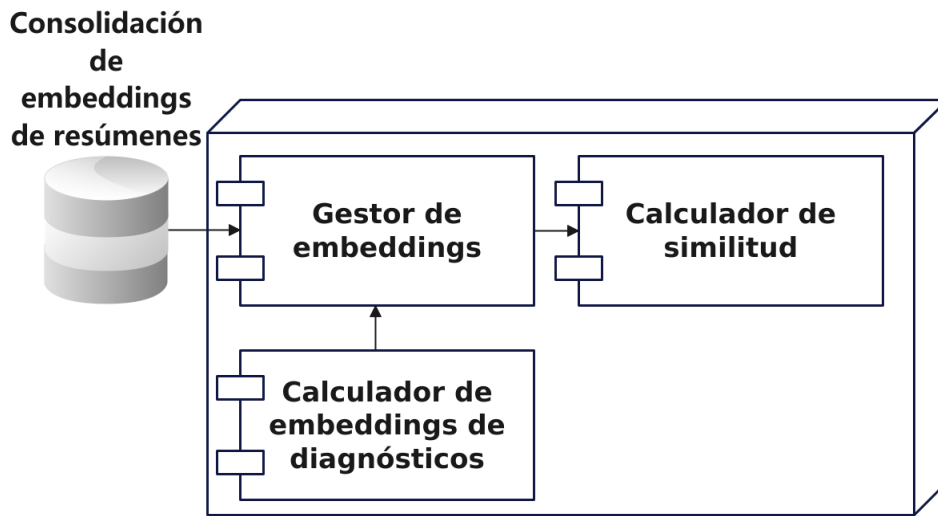


Figura 5.5: Arquitectura interna del módulo Calculador de similitud semántica.

que esta operación minimiza los posibles errores cometidos por la herramienta NER. Finalmente, el módulo produce una lista de las entidades deseadas encontradas en la sección de diagnóstico principal, cuyos *embeddings* se calculan para servir como entrada para el módulo *Calculador de similitud semántica*.

En cuanto a la arquitectura del módulo, consta de cuatro componentes: el *Pre-procesador de texto*, el *Segmentador de secciones*, el *Detector de palabras clave* y el *Recopilador de diagnósticos*. El paso de *pre-procesamiento* se logra mediante el componente *Pre-procesador de texto*, mientras que el paso de *segmentación con expresiones regulares* es abordado por el *Segmentador de secciones*. El paso de *NER para entidades médicas* hace uso de los otros dos componentes: el *Detector de palabras clave* y el *Recolector de diagnósticos*.

5.2.3. Módulo calculador de similitud semántica

Este módulo implementa un motor de búsqueda semántica basado en la similitud del coseno (Li y Han, 2013), que se aplica a los vectores provenientes de un modelo de búsqueda semántica simétrica (ver Ecuación 5.1).

$$\cos(\mathbf{t}, \mathbf{e}) = \frac{\mathbf{t} \cdot \mathbf{e}}{\|\mathbf{t}\| \|\mathbf{e}\|} = \frac{\sum_{i=1}^n t_i e_i}{\sqrt{\sum_{i=1}^n (t_i)^2} \sqrt{\sum_{i=1}^n (e_i)^2}} \quad (5.1)$$

Donde \mathbf{t} y \mathbf{e} son vectores que representan los *embeddings* de dos documentos.

Este modelo permite encontrar contenido textual similar a una consulta (en este caso, el diagnóstico detectado), identificando de manera efectiva sinónimos y palabras de significado cercano debido a la representación contextual proporcionada por una arquitectura de tipo a BERT ajustada para el cálculo de similitud semántica.

En este sentido, cabe señalar que aunque un modelo orientado a búsqueda asimétrica basado en el producto escalar podría parecer inicialmente más apropiado para este caso de uso debido a que los *abstracts* suelen ser más extensos que los diagnósticos, las evaluaciones empíricas y las ventajas de utilizar una métrica acotada llevaron a seleccionar un modelo de búsqueda semántica simétrica basado en la similitud del coseno.

En lo que respecta a la arquitectura del módulo, comprende tres componentes (ver Figura 5.5): el *Calculador de embeddings de diagnósticos*, el *Gestor de embeddings* y el *Calculador de similitud*.

El componente *Calculador de embeddings de diagnósticos* genera el vector para el diagnóstico detectado. Así, el texto se convierte en información numérica.

El componente *Gestor de embeddings* une los vectores pre-calculados de los *abstracts* con el vector correspondiente al diagnóstico.

El componente de *Cálculo de similitud* calcula la similitud entre los vectores de cada *abstract* del corpus y el perteneciente al diagnóstico. Finalmente, devuelve los n *abstracts* con mayor similitud semántica, concluyendo las tareas del módulo.

5.2.4. Módulo calculador de relevancia

Este módulo implementa la funcionalidad de una versión actualizada del sistema previamente desarrollado llamado *Webelance* (ver Capítulo 4). En este caso se ha sustituido un componente de la Ecuación 4.5, debido a cambios en el API de Semantic Scholar. Concretamente el campo *citation_velocity* ha sido sustituido por *h-index*, dando lugar a la Ecuación 5.2.

$$rep_i = \omega_1 \cdot inf_citation_count + \omega_2 \cdot h - index + \omega_3 \cdot seniority + \omega_4 \cdot papers, \quad (5.2)$$

donde $\sum_{i=1}^4 \omega_i = 1$. El parámetro *inf_citation_count* corresponde al número de citas altamente influyentes del autor, mientras que el parámetro *h-index* es una métrica que busca cuantificar la productividad y el impacto de un investigador mediante la evaluación del número de trabajos publicados y el número de citas que estos han recibido en la literatura científica. El parámetro *seniority* representa los años transcurridos entre la primera y la última publicación científica del autor, y el parámetro *papers* considera el número de publicaciones del mismo.

El componente *Calculador de relevancia* aborda el cálculo de la relevancia del artículo teniendo en cuenta los resultados de los componentes anteriores. De este modo, la reputación del artículo y la relevancia estimadas se combinan aquí para pro-

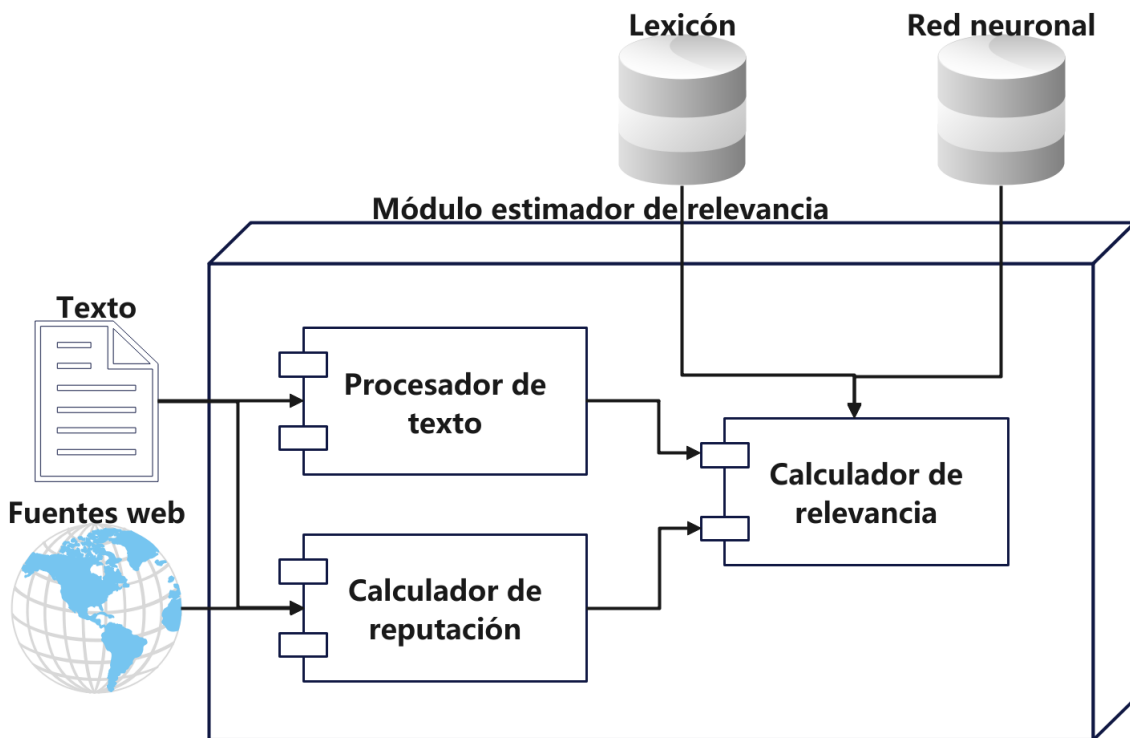


Figura 5.6: Arquitectura del módulo Calculador de relevancia (para más detalle ver el Capítulo 4).

ducir el resultado deseado.

5.2.5. Módulo recomendador de artículos

Este módulo produce el resultado final del sistema. Clasifica el conjunto de manuscritos filtrados según su similitud estimada con un diagnóstico previamente seleccionado y su relevancia asociada.

En lo que respecta a la arquitectura, presenta dos componentes: el *Colector de información* y el *Generador de ranking* (ver Figura 5.7). El primero recopila los valores de relevancia y similitud previamente calculados, mientras que el segundo genera la recomendación final.

Profundizando en el componente *Colector de información*, este organiza los n artículos más similares según la similitud estimada con el diagnóstico seleccionado utilizando los módulos correspondientes del sistema. Para cada uno de estos manuscritos, se calcula su relevancia. Nótese que el rango de valores es el mismo para la medida de similitud y la relevancia, siendo en ambos casos 0 el peor escenario y 1 el escenario óptimo.

El componente *Generador de Ranking* utiliza la distancia euclídea entre cada par: similitud-relevancia para alcanzar un compromiso entre el valor de similitud y el va-

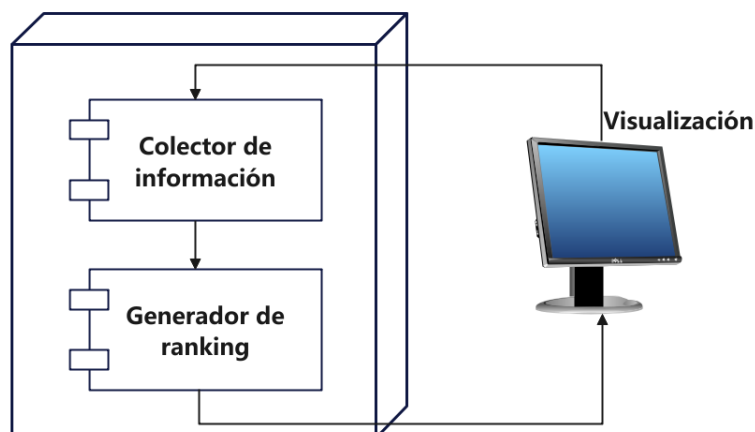


Figura 5.7: *Arquitectura interna del módulo Recomendador de artículos.*

lor de relevancia. Así, la distancia al punto óptimo [*similitud* = 1, *relevancia* = 1] es el valor utilizado para clasificar los artículos. Cabe destacar que solo se utilizan i artículos (siendo $i \leq n$) para hacer la recomendación final. Este valor i se aborda mediante un parámetro configurable del sistema.

5.3. Experimentos

Esta sección presenta los diferentes experimentos realizados para evaluar la viabilidad de la propuesta. En primer lugar, se ha considerado el módulo *Extractor de diagnósticos* con el objetivo de seleccionar la tecnología más adecuada para abordar las tareas correspondientes. En segundo lugar, se ha evaluado el módulo *Calculador de similitud semántica*, utilizando dos diferentes medidas estado del arte para evaluar la similitud entre documentos, y seleccionando el enfoque más adecuado según sus rendimientos obtenidos. Finalmente, se ha abordado una evaluación completa del sistema, ilustrando sus capacidades en un escenario del mundo real.

Cabe destacar que los tres experimentos han involucrado a un conjunto de 5 expertos en el ámbito médico de IBSAL. Han llegado a los resultados de las evaluaciones mediante consenso y votación. Han evaluado la elicitación de diagnósticos, la relevancia de los artículos devueltos por el modelo de similitud semántica de acuerdo a esos diagnósticos, así como el impacto de combinar la similitud semántica con la relevancia intrínseca de cada artículo devuelto.

5.3.1. Evaluación del módulo extractor de diagnósticos

El módulo *Extractor de diagnósticos* es uno de los módulos relevantes de MELEN-DI. Se encarga de encontrar los fragmentos de diagnóstico en los textos y posterior-

mente, extraer las entidades nombradas relevantes relacionadas con el ámbito médico. Este último es susceptible de lograrse mediante diferentes herramientas, lo que lleva a considerar algunos experimentos para seleccionar la más apropiada.

En la actualidad, han aparecido varias herramientas de NER especialmente enfocadas en el ámbito médico. Algunas de las más relevantes son: BioBERT (Lee *et al.*, 2020), MEdCAT (Kraljevic *et al.*, 2021) y Scispacy (Neumann *et al.*, 2019). Sin embargo, el sistema necesita solo aquellas que están vinculadas a UMLS para confirmar que las entidades nombradas seleccionadas forman parte del dominio médico y puedan ser debidamente filtradas de acuerdo a los tipos semánticos pertinentes. Este punto es principalmente relevante porque MELENDI incluye estos términos para producir representaciones vectoriales (*embeddings*) que se utilizarán para encontrar artículos científicos relacionados con el ámbito sanitario. En este caso, MedCAT y Scispacy son los que cumplen con ese requisito.

Por lo tanto, se ha diseñado un experimento completo para comparar el rendimiento de estas dos herramientas respecto a otras similares. Esto permite situar estos enfoques en el estado del arte según su capacidad para detectar entidades nombradas relacionadas con el ámbito médico (es decir, principalmente enfermedades, medicamentos y tratamientos).

En un primer paso, expertos en atención sanitaria han analizado 500 muestras aleatorias de resúmenes de alta del conjunto de datos MIMIC-III (A. E. Johnson *et al.*, 2016). Estos expertos tenían como tarea principal detectar el diagnóstico en los textos. Posteriormente, en un segundo paso, el sistema utilizó las diferentes herramientas de NER para evaluar las mismas muestras y medir el rendimiento en comparación con seres humanos.

Los resultados indican que MedCAT es la mejor herramienta para llevar a cabo la tarea de NER. Scispacy, utilizando el modelo *en_core_scibert*, es significativamente peor en la detección de entidades nombradas relacionadas con el ámbito médico. Los demás presentan resultados similares en particular cuando se refieren a enfermedades o síndromes, lo cual es un buen indicador del rendimiento de MedCAT (ver Tabla 5.1). Como se puede observar, Scispacy, un BioBERT ajustado para el reconocimiento de enfermedades y un método de coincidencia exacta (buscar una coincidencia exacta entre las descripciones de conceptos de UMLS y los textos del EHR) detectan significativamente menos entidades, enfermedades o síndromes.

Durante el experimento, varios factores necesitaron reconsideración para seleccionar el modelo a evaluar. En el caso de BERN2 (Sung *et al.*, 2022), resultó ser una opción desafiante debido a sus altos requisitos de recursos. Sin embargo, demandaba más de 70 GB de espacio en disco, 63,5 GB de RAM y 5,05 GB de GPU, lo que dificultaba su despliegue en un entorno práctico. Dado que la mayoría de los posibles

Sistema NER	Entidades detectadas	Enfermedades detectadas
MedCAT	322.114	15.664
Scispacy	197.435	12.807
BioBERT	10.910	10.910
Coincidencia exacta	65.475	7.493

Tabla 5.1: Comparación de entidades y enfermedades detectadas en 500 EHR entre MedCAT, SciSpacy, BioBERT y el método de coincidencia exacta.

sistemas médicos tendrían problemas para asignar los recursos necesarios, la perspectiva de integrar este modelo en MELENDI fue descartada. Además, en el caso del modelo Med-Flair (EIDin *et al.*, 2021), resultó inaccesible e intrazable, lo que también resultó en su exclusión de la evaluación.

Por otro lado, los resultados muestran que el sistema fue capaz de aislar completamente el diagnóstico principal en el 75 % de los casos usando MedCAT en comparación con seres humanos. Los casos problemáticos se revisaron manualmente, y los expertos concluyeron que no había una sección de diagnóstico claramente identificable, sino que el diagnóstico estaba disperso en el contenido textual. Además, el sistema usando MedCAT fue capaz de devolver información relacionada con el diagnóstico que no estaba presente en el fragmento de texto del diagnóstico principal en el 97 % de los casos. Este hecho ilustra un rendimiento bastante positivo del sistema.

5.3.2. Evaluación del módulo calculador de similitud semántica

El módulo *Calculador de similitud semántica* es otro módulo relevante de MELENDI. Se encarga de medir la similitud semántica entre documentos y las entidades nombradas presentes en el diagnóstico detectado. Siguiendo enfoques estado del arte, hay dos medidas de similitud más típicas: el producto escalar y la similitud coseno.

El producto escalar se usa generalmente en la búsqueda semántica asimétrica. Actualmente, los enfoques más interesantes en este campo, tanto teóricos como potenciales, residen en modelos que se han ajustado utilizando esta medida de similitud. Por otro lado, la similitud coseno se orienta generalmente a la búsqueda semántica simétrica. Este hecho, a priori, no encaja tan bien con el caso propuesto en el sistema, ya que los *abstracts* de los manuscritos científicos suelen ser más extensos que el conjunto de conceptos detectados en los fragmentos de diagnóstico.

Se ha propuesto un experimento para confirmar y validar esta suposición. Para lograrlo, se han seleccionado dos modelos pre-entrenados para representar cada una de las dos posibilidades. En el caso del producto escalar, *msmarco-distilbert-base-tas-b* (Hofstätter *et al.*, 2021) ha sido el candidato. Para el caso de la similitud coseno, se ha elegido *msmarco-distilbert-base-v4* (Reimers y Gurevych, 2019).

Cabe señalar que una de las características del producto escalar es que el resultado no está acotado, lo que dificulta su interpretación, combinación y comparación con otras métricas. Por lo tanto, la evaluación del experimento ha sido diseñada para comparar el rendimiento de ambos modelos con 10 diagnósticos aleatorios. Este proceso consiste en evaluar un conjunto de 1.000 manuscritos científicos obtenidos aplicando el producto escalar y la similitud coseno respectivamente. Luego, se seleccionan los 3 manuscritos científicos más similares a los diagnósticos y, además, los 3 menos similares (es decir, específicamente las posiciones 998, 999 y 1.000). Este filtro obtiene un total de 120 manuscritos (es decir, 60 para cada medida) que son evaluados por los 5 expertos médicos previamente seleccionados.

Así, para cada uno de los 10 diagnósticos, el experto recibió 12 manuscritos, 6 relacionados con la similitud coseno (m_i^c , $i = 1 \dots 6$) y 6 relacionados con el producto escalar (m_i^d , $i = 1 \dots 6$). Cabe señalar que algunos de los manuscritos pueden estar incluidos en ambos conjuntos resultantes (es decir, un manuscrito puede ser seleccionado utilizando el producto escalar y también la similitud coseno). Así, m_i^c y m_i^d ($i = 1 \dots 3$) corresponden a los manuscritos científicos más similares a los diagnósticos utilizando las medidas de similitud coseno y producto escalar, respectivamente. Además, m_i^c y m_i^d ($i = 4 \dots 6$) corresponden a los manuscritos científicos menos similares a los diagnósticos utilizando las medidas de similitud coseno y producto escalar, respectivamente.

A continuación, se pidió a los expertos que seleccionaran los 3 manuscritos que mejor representan el diagnóstico correspondiente. Finalmente, se comparan las opiniones de los expertos y el resultado del sistema utilizando ambas medidas. Esto permite seleccionar la medida que mejor se ajusta según la validación humana.

El resultado del experimento se puede ver en las Tablas 5.2 y 5.3. La primera tabla ilustra el id asignado de cada diagnóstico y las entidades nombradas detectadas en los fragmentos de diagnóstico. La segunda proporciona información detallada comparando las elecciones de las medidas y la opinión global de expertos humanos. Allí, cada diagnóstico también utiliza un id, teniendo los 6 manuscritos recomendados utilizando el producto escalar y los otros 6 recomendados utilizando la similitud coseno. Los manuscritos están ordenados de izquierda a derecha en orden decreciente según sus respectivas puntuaciones. Así, los manuscritos correspondientes a las primeras 3 celdas tienen las puntuaciones de similitud más altas, respectivamente, mientras que las últimas 3 corresponden a las posiciones 997, 998 y 999, respectivamente. Este arreglo se utiliza en ambos casos: producto escalar y similitud coseno.

Los manuscritos con una alta similitud que fueron seleccionados por los expertos aparecen en color verde. Por otra parte, los manuscritos con una baja similitud que fueron seleccionados por los expertos aparecen en color rojo. El resto de ellos aparece sin color, indicando que no han sido elegidos por parte de los expertos. El experto

EHR	Diagnóstico
1	coronary artery disease, Diabetes Mellitus Tyep 2, hypertension.
2	COPD exacerbation.
3	Acute pancreatitis, Hypertriglicidemia, Metastatic Breast Cancer.
4	Pneumonia, asthma, diabetes.
5	Gastric perforation from marginal, ulcer.
6	Acute on Chronic Diastolic Congestive Heart Failure, Atrial fibrillation with rapid ventricular response.
7	Community acquired pneumonia, Alcohol dependence, Acute on chronic renal failure.
8	Amyotrophic Lateral Sclerosis.
9	Chronic Type A Aortic Dissection, s/p Repair ascending dissection reanastamosis of grafts , Coronary Artery Disease, Atrial Fibrillation, s/p Coronary Artery Bypass Grafting and Maze Procedure on, Hypertension, Dyslipidemia, h/o of transient ischemic attack.
10	Hyponatremia secondary to Fanconis syndrome versus SIADH.

Tabla 5.2: Diagnósticos aislados provenientes de diez EHR usados para evaluar el sistema al completo.

Diagnóstico	Artículos usando DOT						Artículos usando COS						
	m_1^d	m_2^d	m_3^d	m_4^d	m_5^d	m_6^d	m_1^c	m_2^c	m_3^c	m_4^c	m_5^c	m_6^c	
1	OK	OK											KO
2	OK		OK				OK	OK	OK				
3			OK	KO			OK						
4	OK							OK	OK				
5	OK		OK				OK	OK					
6	OK		OK				OK		OK				
7		OK					OK	OK					
8	OK	OK		KO			OK						
9	OK	OK		KO			OK						
10		OK					OK						KO

Tabla 5.3: Comparativa entre la opinión de los expertos y los resultados proporcionados por el producto escalar y la similitud coseno.

seleccionó como relevante el manuscrito con la puntuación más alta 7 veces en el caso del producto escalar y 8 veces en el caso de la similitud coseno. Se puede observar que el enfoque de la similitud coseno tiene menos manuscritos de baja similitud seleccionados por expertos. Esto conduce a incluir la medida de similitud coseno en el módulo del sistema. Esta decisión, soportada por el presente experimento, es reseñable, ya que dicha medida proporciona una mejor interpretabilidad y es más fácil combinarla con la medida de relevancia, dado que ambas tienen sus límites entre 0 y 1.

Finalmente, es importante señalar la alta correlación entre las selecciones realizadas por los expertos y los manuscritos proporcionados por ambas medidas. Por lo tanto, se puede concluir que el módulo funciona con alta calidad y los resultados son fiables.

5.3.3. Evaluación del sistema en un entorno real

Este experimento muestra la viabilidad de la propuesta en un escenario del mundo real. Para lograrlo, se considera una configuración completa del sistema MELENDI. Se han utilizado representaciones vectoriales de 1.000.000 de resúmenes provenientes de manuscritos científicos recopilados aleatoriamente de PubMed como base de datos. Además, los parámetros relacionados con la relevancia (ver Sección 5.2.4) se han seleccionado. En este caso, los parámetros ω_n toman los siguientes valores: $\omega_1 = 0,1$, $\omega_2 = 0,2$, $\omega_3 = 0,3$, $\omega_4 = 0,4$. Estos pesos están relacionados con la importancia otorgada a los diferentes elementos que el módulo *Calculador de relevancia* utiliza para producir su resultado. El resto de los parámetros relacionados con el sistema *Webalance* mantienen su configuración estándar.

Una vez que el sistema está listo para funcionar, los 10 EHR y sus evaluaciones realizadas por los expertos utilizados en el experimento anterior han sido seleccionados para ser analizados aquí (ver Tabla 5.2).

Posteriormente, en este experimento, se evalúan conjuntamente la relevancia y la similitud. Para este propósito, se calcularon los 100 manuscritos con la mayor similitud coseno para cada uno de los 10 diagnósticos y su relevancia respectiva. Se obtienen los 10 manuscritos más cercanos al punto ideal ([relevancia, similitud coseno]=[1, 1]) para cada diagnóstico.

A continuación, se establece una comparación entre la relevancia de los 10 manuscritos seleccionados y la opinión del grupo de expertos.

En la Figura 5.8, se muestran los 100 manuscritos principales para el diagnóstico *COPD exacerbation* con id=2. Los 10 primeros clasificados están marcados dentro de las dos curvas verdes. Además, los artículos seleccionados como los más relevantes por los expertos en el experimento anterior (3 como máximo) están representados en rojo. En este caso particular, los 3 artículos se encuentran dentro de los 10 recomendados por el sistema.

Por otro lado, en la Figura 5.9, correspondiente al diagnóstico *Community acquired pneumonia, Alcohol dependence, Acute on chronic renal failure* con id=7, 2 (de 3) artículos están dentro de los 10 primeros. El otro artículo queda fuera de los 10 primeros, pero está dentro de los 100 primeros debido a su alta puntuación de similitud coseno. Obsérvese que es el artículo con la similitud más alta. Sin embargo, su puntuación de relevancia relativamente baja lo penaliza lo suficiente como para

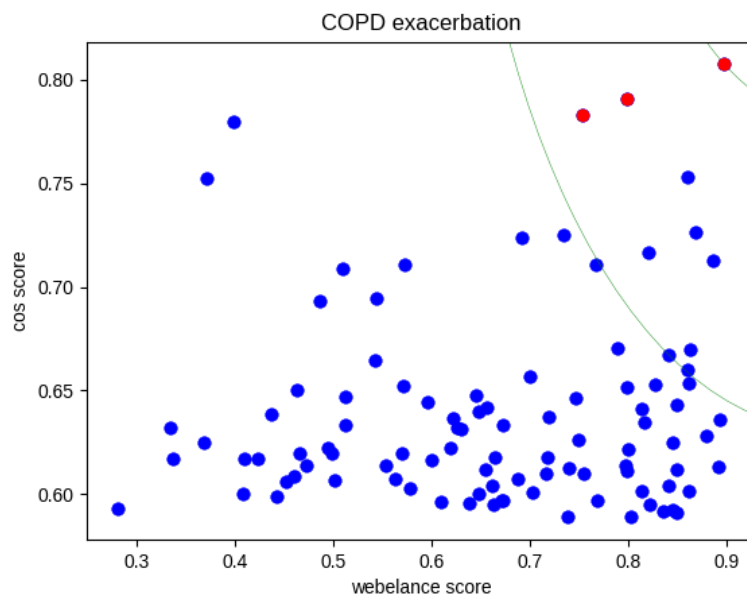


Figura 5.8: Representación de artículos distribuidos según su relevancia y similitud coseno con respecto al diagnóstico COPD exacerbation.

desplazarlo fuera de los 10 primeros.

Finalmente, los resultados para el diagnóstico *Amyotrophic Lateral Sclerosis* con $id=8$ se representan en la Figura 5.10. En este caso, solo 1 de los 3 artículos seleccionados por los expertos se encuentra dentro de los 10 primeros recomendados. Obsérvese que este aparece en la primera posición. Sin embargo, el segundo artículo tiene una puntuación de similitud coseno y relevancia relativamente alta, ubicándolo dentro de los primeros 100. El tercero está fuera de los primeros 100 manuscritos con una baja puntuación de similitud coseno de 0,414.

Los resultados de los 10 diagnósticos se han recopilado en la Tabla 5.4. Ahí, se puede ver la posición relativa de cada uno de los 3 artículos seleccionados por los expertos en comparación con la recomendación del sistema.

Después de concluir el experimento, se pueden extraer las siguientes conclusiones:

- En 9 de los 10 diagnósticos, el sistema ha recomendado al menos 1 de los 3 artículos seleccionados por los expertos, siempre dentro de los primeros 3 más cercanos al punto ideal.
- En 6 de los 10 diagnósticos, el artículo con la mejor puntuación del sistema fue seleccionado por los expertos humanos.
- En 5 de los 10 diagnósticos, hay al menos 2 artículos seleccionados por los ex-

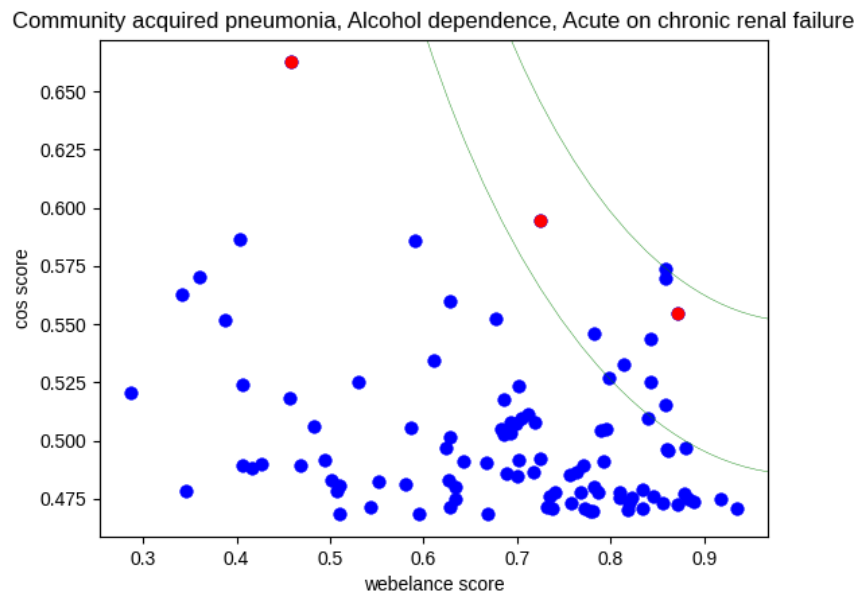


Figura 5.9: Representación de artículos distribuidos según su relevancia y similitud coseno con respecto al diagnóstico Community acquired pneumonia, Alcohol dependence, Acute on chronic renal failure.

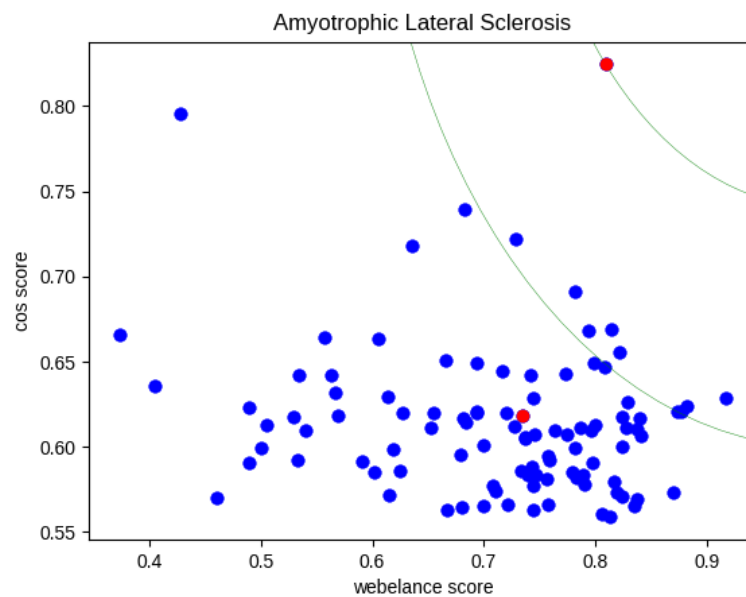


Figura 5.10: Representación de artículos distribuidos según su relevancia y similitud coseno con respecto al diagnóstico Amyotrophic Lateral Sclerosis.

Diagnóstico	Artículo 1	Artículo 2	Artículo 3
1	1	4	>100
2	1	2	6
3	>50	>100	>100
4	1	>10	>10
5	1	>10	>10
6	1	2	>100
7	3	4	>10
8	1	>10	>100
9	2	3	>100
10	2	>10	>100

Tabla 5.4: Comparación entre las evaluaciones de los expertos y las recomendaciones realizadas por MELENDI. Cada fila muestra los tres artículos seleccionados por los expertos para cada diagnóstico junto con la posición del artículo en el ranking propuesto por el sistema (valor de similitud coseno).

peros humanos dentro de los primeros 5 seleccionados por el sistema.

- En 6 de 10, los expertos humanos han seleccionado al menos un artículo fuera de los primeros 100 recomendados por el sistema.

En conclusión, el sistema puede proporcionar conocimiento de manera similar a los expertos en el ámbito médico. Por otro lado, se cumple la suposición de la importancia de la relevancia, ya que el sistema incluye algunos manuscritos en la recomendación de acuerdo con esta característica. Este aspecto es especialmente interesante, ya que la relevancia se calcula automáticamente, y es una tarea difícil de lograr para los seres humanos, ya que normalmente desconocen la calidad de los autores o la importancia de la revista donde se publicó el manuscrito. Así, el sistema permite a los usuarios actualizar automáticamente el conocimiento sobre un diagnóstico (y también sus tratamientos) sin tener que preocuparse por verificar los índices de calidad de las publicaciones donde se aborda ni tener que realizar búsquedas manuales en herramientas de recuperación de la información.

5.4. Conclusiones

En este capítulo ha introducido MELENDI, un sistema de recomendación de artículos científicos para profesionales de la salud que les permite mantenerse actualizados sobre los avances relacionados con los diagnósticos de sus pacientes.

La principal novedad del sistema consiste en la hibridación de un modelo de similitud semántica con un estimador automático de la relevancia de artículos científicos para producir la clasificación final de artículos científicos. De esta manera, la

recomendación logra un equilibrio entre la relación del artículo recomendado con su diagnóstico correspondiente y la relevancia del propio artículo. Este equilibrio garantiza que los artículos recomendados estén estrictamente relacionados con los pacientes y sean relevantes en su temática y origen.

Se han llevado a cabo experimentos con dos modelos de similitud semántica, uno ajustado para el producto escalar y otro para la distancia coseno. Estos experimentos, realizados por un grupo de especialistas médicos de IBSAL, han mostrado diferencias insignificantes entre ambos modelos, seleccionando el modelo ajustado para similitud coseno por su simplicidad e interpretabilidad de los resultados. Por otro lado, se calcularon los 100 artículos con la mejor similitud coseno respecto a 10 diagnósticos extraídos de 10 EHR de MIMIC-III para evaluar la efectividad de esta recomendación. Los resultados obtenidos indican que en 9 de los 10 diagnósticos, el sistema ha recomendado al menos uno de los 3 artículos más relevantes seleccionados por los expertos de IBSAL, todos ellos dentro de los 3 más cercanos al punto ideal de similitud y relevancia. Esto demuestra el buen rendimiento del sistema de recomendación y respalda su posible implementación en un entorno real.

A pesar del buen rendimiento encontrado, este trabajo ha tenido diferentes limitaciones. En primer lugar, el sistema de detección de diagnósticos se ha adaptado a la tipología y casuística de los informes de alta de MIMIC-III, lo que limita su implementación con otros tipos de EHR. Por esta razón, se contemplará el desarrollo de un sistema que automatice completamente la detección de diagnósticos independientemente de la tipología del EHR para trabajos futuros, sujeto a la disponibilidad de datos etiquetados para este propósito. Además, el modelo de similitud semántica es de dominio general, por lo que ajustar el modelo para el ámbito médico tendría un impacto positivo en su rendimiento. Este proceso de ajuste se planteará en trabajos futuros, ya que requiere que el sistema esté en producción y recoja comentarios de los usuarios. Finalmente, el equilibrio entre relevancia y similitud también podría ajustarse según la importancia relativa que el usuario dé a cada característica, resultando en un sistema adaptado a las particularidades del usuario. Por otra parte, el sistema está limitado a la búsqueda sobre *abstracts* y no artículos completos, en trabajos futuros se explorarán diferentes estrategias para abordar este problema, ya que los artículos completos no se pueden codificar en un único vector semántico por limitaciones de memoria. Además, esta planificado escalar el sistema para trabajar con todo Pubmed a nivel de *abstracts* y con *Pubmed Central* (PMC) al completo en el caso de artículos completos. En el próximo capítulo, el marco de trabajo MELENDI es implementado, de nuevo, en un sistema completo que incorpora un LLM como última etapa en después de la recomendación, permitiendo al usuario interactuar con las recomendaciones en lenguaje natural.

Capítulo 6

Asistente generativo contextualizado

En el capítulo anterior se presentó MELENDI con el objetivo de recomendar artículos relevantes basados en las características de los pacientes de forma automática. En este capítulo, se incorpora un componente adicional posterior a la recomendación proporcionada por MELENDI que permite interactuar y realizar consultas concretas sobre las recomendaciones mediante el uso del lenguaje natural. Este componente cierra el círculo y da lugar al asistente de investigación médica basado en conocimiento que da título a la tesis. Más adelante, se introduce la metodología utilizada para construirlo, los experimentos realizados y finalmente unas conclusiones y posibles directrices futuras.

6.1. Introducción

Tradicionalmente, los asistentes conversacionales o *chatbots* se basan en modelos de lenguaje entrenados con grandes volúmenes de datos, pero suelen carecer de la capacidad para considerar un contexto específico o actualizado que pueda ser crucial para ciertas aplicaciones. En el contexto médico, esta limitación puede ser especialmente problemática dado que las consultas a menudo requieren un conocimiento muy específico y actualizado. Los *chatbots* médicos tradicionales basados en conocimiento están ajustados por especialistas en el dominio y cuentan con bases de datos confiables, pero requieren de una actualización constante de sus modelos de detección de intenciones (Chow *et al.*, 2023; Thirunavukarasu *et al.*, 2023) y están limitados a flujos conversacionales pre-definidos. Con el nacimiento de los LLM (Zhao *et al.*, 2023) y en especial con chatGPT (J. Deng y Lin, 2022) debido a su popularidad, las posibilidades de interacción en lenguaje natural con sistemas basados en conocimiento se han disparado, pero también plantean diversos problemas técnicos y éticos (Javaid

et al., 2023). Algunas de las principales limitaciones técnicas de chatGPT que pueden tener un mayor impacto en asistentes conversacionales médicos son su fecha de entrenamiento y su tendencia a generar alucinaciones (Alkaissi y McFarlane, 2023). Aunque se ha utilizado una gran cantidad de contenidos científico-médicos de acceso libre, el entrenamiento de chatGPT llega únicamente hasta septiembre de 2021 (en el momento de escritura de la tesis), por lo que en ningún caso puede proveer a los usuarios de información actualizada sobre investigaciones posteriores a dicha fecha. Añadido a este problema, las mencionadas alucinaciones impiden que las respuestas proporcionadas por chatGPT y por los LLM en general, sean plenamente confiables, ya que son respuestas generadas que no son trazables ni se ajustan a la veracidad de datos reales. Los orígenes de las alucinaciones pueden ser diversos, como limitaciones en el conjunto de datos y en el entrenamiento (Ji *et al.*, 2023). Por tanto, un sistema que inyecte contexto a chatGPT para proveerle de información actualizada y ayudar a evitar alucinaciones resulta necesario para aprovechar las capacidades de chatGPT en un entorno profesional, donde cada respuesta necesita trazabilidad y confianza.

En este capítulo se aborda el tercer y último componente de la tesis, donde se propone una solución integral al problema previamente descrito. En él confluyen los dos componentes anteriores (ver Capítulos 4 y 5) para dar lugar a un asistente de investigación médica basado en conocimiento relevante y actualizado, además de centrado en las características específicas de los pacientes de sus potenciales usuarios.

La propuesta no solo mejora la precisión de las respuestas generadas, sino que también permite que el asistente cite las fuentes en las que basa sus respuestas. De esta manera, el usuario puede tener un nivel de confianza más alto en la información proporcionada y puede acceder a las fuentes originales para un examen más detallado. Por otra parte, la capacidad de interactuar con el sistema usando preguntas formuladas en lenguaje natural elimina la necesidad de que el usuario tenga conocimientos especializados en búsqueda de información, lo que hace que el sistema sea accesible y fácil de usar, permitiendo ahorrar tiempo a los facultativos.

Este capítulo profundiza en el diseño, la implementación y la evaluación del asistente generativo contextualizado o asistente generativo con inyección de contexto, denominado GAMER, delineando cómo este sistema representa un avance significativo en el campo de la asistencia a la investigación y la toma de decisiones en la medicina. Se han realizado una serie de experimentos en los que se evalúa la viabilidad del sistema, implicando en el proceso nuevamente a expertos en el dominio procedentes de IBSAL. El capítulo se estructura en tres puntos, de forma análoga a los dos componentes anteriores: En primer lugar, la Sección 6.2 hace una descripción de la arquitectura de la solución propuesta, sirviéndose de *Webelance* y MELENDI para la obtención de documentos relevantes y su recomendación en base a los diagnósticos

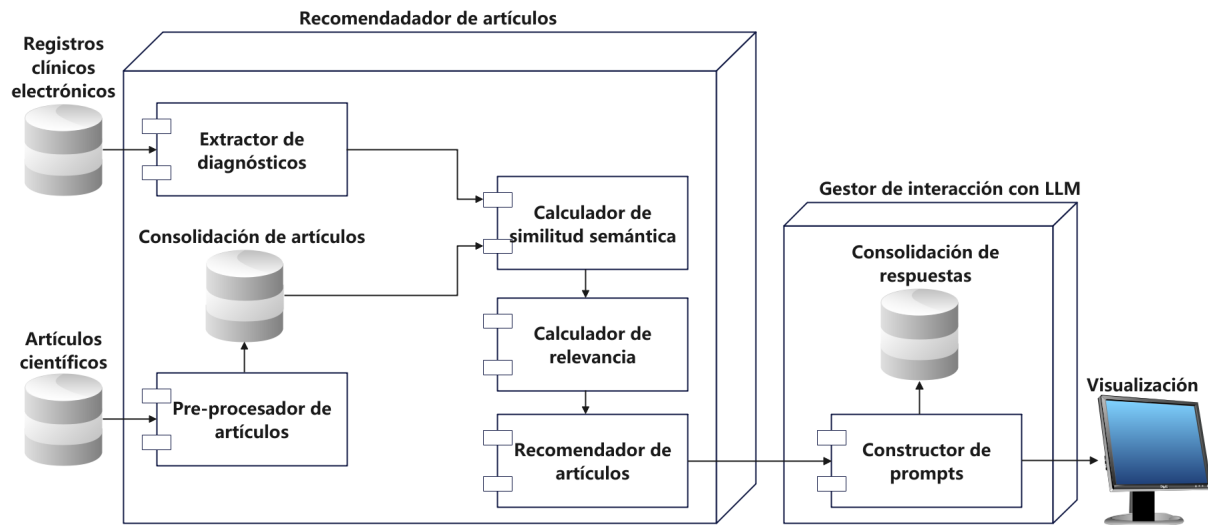


Figura 6.1: Representación general de la arquitectura final del sistema (módulos, componentes y artefactos).

extraídos. En la Sección 6.3 se presentan los experimentos realizados con la configuración final del sistema y, finalmente, la Sección 6.4 se hace una reflexión sobre los resultados obtenidos en este componente y posibles mejoras futuras.

6.2. Propuesta

La funcionalidad principal de la solución propuesta consiste en proporcionar respuestas a preguntas en lenguaje natural relacionadas con los artículos recomendados por MELENDI, de forma que el médico no solo disponga de artículos relevantes y actualizados relacionados con las características de sus pacientes sin la necesidad de hacer ninguna búsqueda, si no que además pueda realizar filtros o hacer preguntas concretas a los mismos, permitiéndole centrarse en aquellos que le resulten de mayor interés. En la Figura 6.1 se puede ver la arquitectura principal del sistema, que tiene 4 piezas fundamentales: MELENDI, que es la pieza para proporcionar artículos relevantes, un módulo de interacción con un LLM, que interconecta MELENDI con el LLM, un componente de consolidación de respuestas y un módulo de visualización.

En este punto, MELENDI mantiene la misma propuesta descrita en el Capítulo 5, con su correspondiente actualización de *Webelance* para adaptarse a los documentos más nuevos y a la evolución del API de *Semantic Scholar*. Por otra parte, el módulo de interacción con el LLM gestiona la configuración de los parámetros del mismo, así como la inyección de contexto, el establecimiento del rol del sistema y el envío de respuestas al módulo de visualización.

6.2.1. Flujo de funcionamiento del asistente

En la Figura 6.2, se puede ver una representación del flujo completo del sistema, en el que interactúan de forma orquestada sus 3 componentes principales, *Webe-lance* para estimar la relevancia de artículos, MELENDI para recomendar artículos relevantes en base a EHR y finalmente el asistente basado en un LLM. El sistema procesa automáticamente los EHR disponibles del usuario extrayendo sus diagnósticos y creando recomendaciones personalizadas de artículos previamente procesados y consolidados. Una vez que el usuario ve las recomendaciones de artículos en base al paciente que este mismo seleccione, interactúa con ellos a través del LLM. Esta interacción se fundamenta en la inyección de los artículos recomendados del paciente en cuestión más la consulta que el usuario quiere realizar sobre ellos. Una vez que el usuario envía su consulta sobre los artículos recomendados, esta es enviada al LLM junto con los textos de los *abstracts* de los artículos. Después, se obtiene la respuesta y el módulo de visualización se la presenta al usuario.

6.2.2. Módulo de interacción con el LLM

Este módulo es el más importante del asistente, ya que se encarga de la construcción de *prompts* (J. White *et al.*, 2023) y de gestionar la comunicación entre el recomendador, el LLM y el usuario. Se genera un *prompt* con el rol del sistema, el cuál le dice al asistente cómo debe actuar ante las peticiones del usuario, posteriormente se le añade la consulta del usuario. La consulta del usuario se genera concatenando en una cadena de caracteres todos los *abstracts* recomendados con su correspondiente título. Adicionalmente, se añade una numeración a cada *abstract* para que el LLM lo pueda referenciar con facilidad en sus respuestas. Cada una de las interacciones entre el usuario, el recomendador y el LLM es consolidada en el módulo de consolidación de respuestas, que permite recuperar consultas anteriores. Este flujo de ejecución se repite cada vez que el usuario selecciona una recomendación en particular con la que interactuar. Una parte importante de este módulo son los parámetros del LLM que entran en juego, por lo que cabe señalar los más relevantes y que son comunes en la mayoría de los LLM disponibles.

- **temperature:** Este parámetro controla el nivel de aleatoriedad de las respuestas generadas. Valores más altos (cerca de 1) resultarán en respuestas más creativas y menos predecibles. Valores más bajos (cerca de 0) harán que el modelo se centre más en la respuesta más probable.
- **max_tokens:** Este parámetro limita el número máximo de tokens en la salida. Reducir este número podría resultar en respuestas truncadas, mientras que un límite muy alto podría generar respuestas excesivamente verbosas.

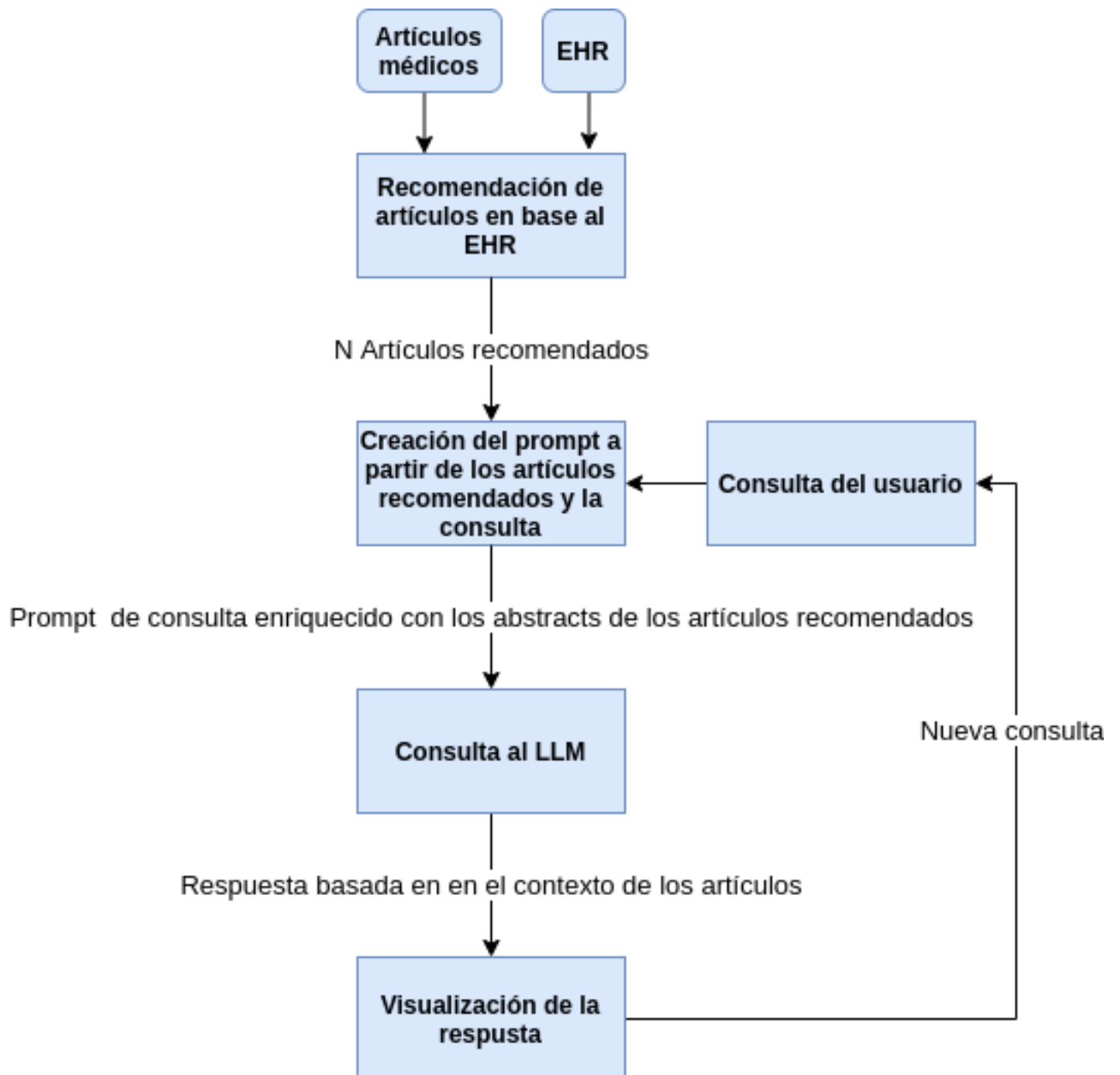


Figura 6.2: Flujo de funcionamiento del sistema completo.

- **top_p:** También conocido como "nucleus sampling", este parámetro controla la diversidad de la muestra. Un valor de top_p de 1 considerará todos los tokens posibles para la generación.
- **frequency_penalty:** Este valor se usa para penalizar palabras que aparecen con frecuencia. Un valor negativo favorecerá palabras que son menos comunes.
- **presence_penalty:** Este valor controla cuán propenso es el modelo a incluir nuevos conceptos en la conversación.

6.3. Experimentos

En esta sección se presentan los experimentos realizados para validar la propuesta del asistente. Los experimentos se han enfocado en validar que el LLM basa sus respuestas exclusivamente en el contexto proporcionado por los *abstracts* y que sus respuestas son técnicamente correctas. Para ello, se han desarrollado 3 experimentos con varias series de consultas relacionadas con las recomendaciones proporcionadas por MELENDI sobre *cinco* diagnósticos seleccionados de entre todos los detectados en el corpus de MIMIC-III. Una serie de expertos en el dominio de la medicina de IB-SAL han sido de nuevo los responsables de realizar las validaciones. Los expertos han diseñado una batería de *diez* preguntas sin conocer los resultados de los artículos recomendados. Este experimento está motivado por la naturaleza pro-activa del asistente. Es decir, el asistente proporciona recomendaciones al usuario sin la necesidad de interacción del mismo y, dado que las recomendaciones están frecuentemente basadas en diagnósticos compuestos de varias patologías y diagnósticos asociados, se han diseñado consultas que proporcionen información sobre posible relación o interacciones de las distintas patologías o diagnósticos asociados encontrados en los artículos recomendados. De esta manera, gracias a la información proporcionada por el asistente, el usuario podrá discriminar en primera instancia aquellos artículos más relacionados con las características de sus pacientes, es decir, aquellos que abordan más características en común con el paciente. En segundo lugar, se ha desarrollado un experimento para abordar el impacto de un LLM más potente evaluando las respuestas dadas por el mismo en aquellas respuestas del experimento anterior calificadas como inadecuadas. Por último, se ha diseñado otra batería de quince consultas extra a posteriori de conocer los *abstracts* recomendados para los cinco diagnósticos. Estas consultas son mucho más concretas y están orientadas a validar que las respuestas citan correctamente los *abstracts* específicos en los que están fundamentadas.

6.3.1. Selección de parámetros

La selección de parámetros y la configuración de MELENDI permanece igual en los tres experimentos. Los parámetros de MELENDI siguen la configuración estándar propuesta en la Sección 5.3.3. En cuanto a los parámetros del LLM, en primer lugar se ha seleccionado el modelo **GPT-3.5-turbo** teniendo en cuenta su popularidad actual y su buen compromiso entre rendimiento y coste. Los parámetros internos seleccionados del modelo se pueden ver en la Tabla 6.1 donde sin duda el más importante es la temperatura. La temperatura controla el nivel de aleatoriedad en las respuestas generadas. Un valor más alto, como 1, hará que las respuestas sean más diversas y menos predecibles, mientras que un valor más bajo como 0,1 hará que las respuestas sean más deterministas y enfocadas en la opción más probable. Una temperatura elevada suele dar lugar a una menor calidad cuando se busca precisión en las respuestas (Peng *et al.*, 2023). El sistema debe contestar ciñéndose exclusivamente al contexto proporcionado y ser lo más determinista posible para asegurar trazabilidad, por lo que la temperatura se fija a 0. Por otra parte, aunque `top_p` también tiene influencia sobre lo determinista que será la respuesta, se establece en su valor por defecto, 1. Esta decisión se toma siguiendo la recomendación de la documentación proporcionada por OpenAI⁵ de alterar la temperatura o `top_p`, pero no ambas a la vez. El número máximo de *tokens* se deja sin límite, aunque el máximo tamaño de contexto viene fijado por la versión del modelo, 4.096 en esta versión de GPT-3.5-turbo. Finalmente, `frequency_penalty` y `presence_penalty` se establecen en su valor por defecto, 0.

Parámetro	Descripción	Valor seleccionado
<code>temperature</code>	Controla la aleatoriedad de las respuestas	0
<code>max_tokens</code>	Máximo número de tokens en la salida	Sin límite
<code>top_p</code>	Controla la diversidad de la muestra	1.0
<code>frequency_penalty</code>	Penalización por frecuencia de palabra	0
<code>presence_penalty</code>	Penalización por presencia de palabra	0

Tabla 6.1: *Parámetros relevantes de GPT-3.5-turbo y valores seleccionados.*

Por otra parte, se establece el rol del sistema para el y la metodología para la creación del *prompt* de usuario con el contexto insertado más la consulta. El *prompt* del sistema se establece de la siguiente forma:

Rol del sistema

You are an assistant expert in biomedical research that uses several user-provided abstracts from scientific articles to answer questions related to them.

Este rol se puede dividir en 3 partes fundamentales que le indican a GPT como debe actuar ante las consultas del usuario:

⁵<https://platform.openai.com/docs/api-reference/chat/create>

- **"You are an assistant expert in biomedical research"**: Establece que el modelo debe actuar como si tuviera un alto nivel de conocimiento y competencia en el campo de la investigación biomédica.
- **"that uses several user-provided abstracts from scientific articles"**: Este fragmento especifica que el modelo deberá basar sus respuestas en los resúmenes (*abstracts*) de artículos científicos que el usuario proporcionará.
- **"to answer questions related to them"**: Finalmente, esto establece que el modelo deberá usar la información en los *abstracts* para responder preguntas que estén relacionadas con esos mismos *abstracts*.

Una vez definido el rol del sistema permanece fijo en cada interacción con el asistente. Por otra parte, el *prompt* del usuario es dinámico y se construye de la siguiente manera:

Rol del usuario

Given this diagnosis «texto del diagnóstico» and the the following abstracts from medical articles related to the diagnosis: «texto de los abstracts», answer the following question and cite the corresponding abstract or abstracts with its number with the format [Abstract N] where N represents the number of each provided abstract. The question is «texto de la consulta»

Donde «**texto del diagnóstico**» es una cadena de caracteres que contiene el diagnóstico concreto extraído del EHR del paciente en cuestión y «**texto de los abstracts**» contiene el *abstract* de los 10 artículos más relevantes recomendados por MELENDI para dicho diagnóstico. Los *abstracts* se concatenan incluyendo el título del artículo y un número del 1 al 10, ordenados de mayor a menor según su ranking en la recomendación, siendo el 1 el mejor posicionado y el 10 el último.

6.3.2. Batería de consultas pre-definidas

Los expertos médicos han diseñado una batería de 10 preguntas generalistas e independientes de los diagnósticos y los artículos recomendados. Estas preguntas están orientadas al descubrimiento de relaciones entre las diferentes patologías que pueden componer un diagnóstico y que el asistente no inventa las respuestas si estas no se encuentran en los *abstracts* recomendados. Para este experimento se han seleccionado 5 de los diagnósticos utilizados para los experimentos de MELENDI (ver Sección 5.3.2). Estos diagnósticos se pueden ver en la Tabla 6.2.

Por otro lado, las diez consultas se recogen en la tabla 6.3. Se ha obtenido un total de 10 artículos recomendados para cada uno de los 5 diagnósticos y posteriormente se han realizado cada una de las 10 consultas a cada conjunto de los conjuntos de

ID	Diagnóstico
1	Hyponatremia secondary to Fanconis syndrome versus SIADH.
2	Acute pancreatitis, Hypertriglicidemia, Metastatic Breast Cancer.
3	Community acquired pneumonia, Alcohol dependence, Acute on chronic renal failure.
4	Amyotrophic Lateral Sclerosis.
5	Gastric perforation from marginal, ulcer.

Tabla 6.2: *Diagnósticos utilizados para generar las recomendaciones para evaluar el asistente.*

ID	Consulta
1	Is there statistical relevance in the association of the diagnoses found?
2	What level of scientific evidence do the abstracts have?
3	What is the highest level of scientific evidence among the abstracts provided?
4	Does any of the associated diagnoses improve the main disease?
5	Does any of the associated diagnoses worsen the main disease?
6	Does any of the associated diagnoses increase the mortality of the main disease?
7	Does any of the associated diagnoses decrease the mortality of the main disease?
8	Is there any treatment of the associated diagnoses that improves the main disease?
9	Is there any treatment of the associated diagnoses that worsens the main disease?
10	Does any of the secondary diagnoses reduce the response to any of the treatments for the main disease?

Tabla 6.3: *Batería de consultas generalistas para la evaluación del asistente.*

artículos recomendados, haciendo un total de 50 respuestas sometidas a evaluación por parte de los expertos en el dominio. De las 50 respuestas, 43 han sido evaluadas como correctas, dando lugar a una precisión del **86 %**.

Este valor se puede derivar de los resultados reflejados en la tabla 6.4. Donde se han recogido las evaluaciones de las respuestas realizadas por los expertos. Esta evaluación se ha hecho cotejando las respuestas del asistente con el contenido de los *abstracts* recomendados para cada diagnóstico. A continuación cabe analizar una serie ejemplos evaluados como correctos:

Diagnóstico 1, Consulta 3

There is no clear answer to this question as the abstracts cover different topics and do not necessarily present scientific evidence in the same way. However, [Abstract 1] presents a large cohort study using Cox proportional hazards and competing risk models, which is a high level of scientific evidence.

Esta respuesta hace un razonamiento sobre una cuestión generalista alegando que las evidencias presentadas por los artículos cubren diferentes temas y que no

	C1	C2	C3	C4	C5	C6	C7	C8	C9	10
D1	x	✓	✓	✓	✓	✓	✓	✓	x	✓
D2	✓	✓	✓	x	x	✓	✓	✓	✓	✓
D3	✓	✓	x	✓	✓	✓	✓	✓	x	✓
D4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
D5	✓	✓	✓	✓	x	✓	✓	✓	✓	✓

Tabla 6.4: Resultados de evaluación de las respuestas generalistas realizada por los expertos.

serían fácilmente comparables. Una vez aclarado este hecho, señala que el *abstract 1* presenta métodos asociados a una alta evidencia científica.

Diagnóstico 1, Consulta 5

[Abstract 1] suggests that hyponatremia and hypernatremia are associated with increased mortality in patients with chronic kidney disease. Hyponatremia is specifically associated with increased risk for various cause-specific mortality categories, including cardiovascular, malignancy related, and non-cardiovascular/non-malignancy related deaths. [Abstract 3] also notes that hyponatremia as an initial presentation is an uncommon feature of malignancy-associated SIADH, and in the lung cancer population, hyponatremia has been identified as a negative prognostic factor in hospitalized patients and those with advanced-stage disease. Therefore, it is possible that the presence of hyponatremia in the diagnoses of Fanconi syndrome or SIADH could worsen the main disease. [Abstract 1, Abstract 3]

En esta interacción se responde a la información requerida citando en cada caso el *abstract* en el que se basa la respuesta, demostrando un buen ejemplo del comportamiento esperado por el asistente. Cabe destacar la conclusión final, que se fundamenta en los dos *abstracts* previamente citados.

Diagnóstico 2, Consulta 7

There is no information in the provided abstracts to suggest that any of the associated diagnoses (Acute pancreatitis, Hypertriglyceridemia, or Metastatic Breast Cancer) decrease the mortality of the main disease.

Esta respuesta es un buen ejemplo de cómo el sistema debe comportarse cuando la respuesta a la consulta no se encuentra en el contexto de *abstracts* proporcionados a GPT-3.5-turbo. Una de las principales motivaciones de la inyección de contexto es evitar las alucinaciones, por lo que la respuesta es adecuada.

Diagnóstico 2, Consulta 8

For Acute Pancreatitis, [Abstract 1] suggests that oral administration of apigenin in rats following experimentally induced acute pancreatitis seems to be protective on the lung tissue. However, caution is required for human use as more detailed studies are needed. For Hypertriglyceridemia-induced Pancreatitis, [Abstract 2] reports a case of a patient with clinical acute pancreatitis secondary to hypertriglyceridemia who was successfully managed with the addition of intravenous insulin. For Metastatic Breast Cancer, [Abstract 8] reports the results of a Phase I clinical trial to investigate the safety, immunostimulatory effects, and antineoplastic activity of a multi-target vaccine composed of four distinct peptides derived from cancer-testis (CT) antigens and vascular endothelial growth factor receptors (VEGFRs) in patients with advanced pancreatic cancer. The multi-peptide vaccine was well-tolerated, and clinical benefits were observed in four of the nine patients.

En esta interacción el asistente desgana cada uno de los tratamientos de los diagnósticos asociados que en los *abstracts* se señala que han mejorado la enfermedad principal, citando las correspondientes fuentes.

Diagnóstico 3, Consulta 8

There is evidence that early antimicrobial therapy and appropriate medications to cover typical pathogens are extremely important in treating community-acquired pneumonia in older individuals [Abstract 3]. Pneumococcal and influenza vaccination comprise one of the most important preventive approaches for community-acquired pneumonia in the elderly [Abstract 9]. However, there is no information in the provided abstracts about treatments for alcohol dependence or acute on chronic renal failure that improve the main disease of community-acquired pneumonia.

Esta respuesta recoge los tratamientos asociados a la mejora del diagnóstico principal presentes en los *abstracts*, pero al contrario que en el caso anterior, señala que no hay información sobre los tratamientos de los diagnósticos asociados que mejoran la enfermedad principal.

Diagnóstico 4, Consulta 2

The abstracts provided do not specify the level of scientific evidence they represent. However, they are all from peer-reviewed medical articles, which suggests that they are based on scientific research and have undergone a rigorous review process.

Aquí, el asistente realiza un razonamiento adecuado sobre el nivel de evidencia estadística, teniendo en cuenta que es una pregunta generalista y potencialmente subjetiva.

Diagnóstico 5, Consulta 2

The abstracts provided do not specify the level of scientific evidence they represent. However, they are all from peer-reviewed medical articles, which suggests that they are based on scientific research and have undergone a rigorous review process.

En esta interacción se repite la consulta anterior y, aunque la respuesta es correcta y análoga, muestra un razonamiento diferente, evidenciando la influencia que el contexto tiene en las respuestas de GPT-3.5-turbo aún cuando la temperatura es baja o, como en este caso, cero.

En general, las respuestas recibidas son más que correctas y fundamentadas en el contexto proporcionado, lo que sugiere que la metodología propuesta es robusta y eficaz. Nótese que el diagnóstico 4, *Amyotrophic Lateral Sclerosis*, contiene una única enfermedad, por lo que las consultas relacionadas con varias enfermedades siempre se contestan alegando insuficiencia de información, lo que puede haber influido en el hecho de ser el único diagnóstico cuyas consultas siempre han sido evaluadas como correctas. Por otra parte, el diagnóstico 5, únicamente está compuesto por dos enfermedades, lo cual puede haber influido asimismo en tener una única respuesta evaluada como incorrecta.

6.3.3. Evaluación con GPT-4

A pesar del buen rendimiento mostrado por el asistente en el apartado anterior, 7 de las 50 respuestas han sido evaluadas como incorrectas. Por tanto, en esta subsección se realiza una comparativa utilizando **GPT-4** en lugar de GPT-3.5-turbo para evaluar dichas respuestas.

GPT-3.5-turbo: Diagnóstico 1, Consulta 1

There is no abstract that directly addresses the statistical relevance of the association between the diagnoses of hyponatremia secondary to Fanconi syndrome versus SIADH. However, [Abstract 1] provides information on the association between hyponatremia and mortality in patients with chronic kidney disease, while Abstract 3 discusses the association between SIADH and malignancy, specifically lung cancer.

GPT-4: Diagnóstico 1, Consulta 1

The abstracts provided do not contain direct statistical information comparing the association between Hyponatremia, Fanconi's syndrome, and SIADH. However, they do provide individual insights into each condition. Hyponatremia is associated with increased mortality in patients with chronic kidney disease [Abstract 1]. SIADH, a common cause of hyponatremia, is often associated with malignancies and can be a negative prognostic factor in lung cancer [Abstract 3]. Fanconi's syndrome is a genetic disorder that affects the bone marrow and is associated with mutations in the FANC genes [Abstract 4]. However, the abstracts do not provide statistical relevance or direct comparisons between these diagnoses.

En este caso, GPT-3.5-turbo confundió el síndrome de Fanconi con la enfermedad o anemia de Fanconi que son patologías completamente distintas. Por otra parte, GPT-4 contestó en base al contexto proporcionado y no confundió ambas enfermedades.

GPT-3.5-turbo: Diagnóstico 1, Consulta 9

[Abstract 2] discusses the traditional management strategies for hyponatremia, such as fluid restriction, salt tablets, loop diuretics, and normal saline, and states that there are scant data to support their use. The article also mentions that data from a large hyponatremia registry reveal the limited efficacy of these therapies. However, the article goes on to discuss the potential benefits of vasopressin receptor antagonists in treating hyponatremia.

GPT-4: Diagnóstico 1, Consulta 9

Based on the provided abstracts, there is no explicit mention of a treatment for the associated diagnoses (Fanconi's syndrome, SIADH, or hyponatremia) that worsens the main disease. However, it is important to note that the management of these conditions can be complex and may require careful monitoring to avoid potential complications. For instance, the use of vasopressin receptor antagonists in the treatment of hyponatremia has shown promise, but there are reports of complications resulting from their misuse [Abstract 2]. Similarly, traditional treatments for SIADH such as fluid restriction and demeclocycline have been reported to fail, cause side effects, or increase the duration of hospital stay [Abstract 5]. Therefore, while not directly worsening the main disease, these treatments can potentially lead to other complications if not properly managed.

En este caso, GPT-3.5-turbo no responde a la consulta y hace una deriva sobre los posibles beneficios de los tratamientos asociados. En su lugar, GPT-4 responde direc-

tamente y de forma correcta a la cuestión y aporta información relevante obtenida de los artículos.

GPT3.5-turbo: Diagnóstico 2, Consulta 4

[Abstract 1] suggests that apigenin, a substance with antioxidant and anti-inflammatory effects, may be protective on lung tissue following experimental induced acute pancreatitis. However, caution is required for human use as more detailed studies are needed. There is no evidence in the other abstracts that the other associated diagnoses (Hypertriglyceridemia and Metastatic Breast Cancer) improve the main disease (Acute Pancreatitis).

GPT4: Diagnóstico 2, Consulta 4

None of the abstracts provided suggest that any of the associated diagnoses (Acute pancreatitis, Hypertriglyceridemia, Metastatic Breast Cancer) improve the main disease. Each of these conditions is a serious health concern on its own and can complicate the management and prognosis of the others. For instance, severe hypertriglyceridemia is a risk factor for acute pancreatitis [Abstract 2], and breast cancer can metastasize to the pancreas [Abstract 3, Abstract 5]. Furthermore, acute pancreatitis is a common disease with significant mortality [Abstract 1]. Therefore, these associated diagnoses do not improve the main disease, but rather complicate it.

Este caso la versión del asistente con GPT-3.5-turbo responde parcialmente mal a la consulta, ya que, aunque es cierto que responde correctamente al hecho de que hay evidencia en los *abstracts* de que los diagnósticos asociados mejoren la enfermedad principal, responde a la cuestión 8 cuando no debería hacerlo. Esto refleja la tendencia de GPT-3.5-turbo de proporcionar información extra aunque no se pida de forma explícita. Por otra parte la respuesta de GPT-4 se centra exclusivamente en la consulta y ha sido evaluada como correcta.

GPT3.5-turbo: Diagnóstico 2, Consulta 5

There is no clear evidence in the provided abstracts that any of the associated diagnoses worsen the main disease of acute pancreatitis, hypertriglyceridemia, or metastatic breast cancer. However, [Abstract 1] suggests that apigenin administration may be protective on the lung tissue following experimental induced acute pancreatitis, which could potentially ameliorate acute lung injuries.

GPT4: Diagnóstico 2, Consulta 5

Yes, the associated diagnoses can worsen the main disease. Severe hypertriglyceridemia is one of the many yet rare risk factors associated with acute pancreatitis [Abstract 2]. Additionally, metastatic breast cancer can spread to the pancreas, causing further complications [Abstract 3, Abstract 5].

GPT-3.5-turbo responde mal a esta cuestión cuando hay una clara evidencia en los *abstracts* proporcionados. GPT-4 responde adecuadamente, citando las correspondientes fuentes.

GPT3.5-turbo: Diagnóstico 3, Consulta 3

There is no clear answer to this question as the abstracts cover different topics and do not necessarily present scientific evidence in the same way. However, Abstract 2 provides a comprehensive search of the literature on community-acquired pneumonia in children and presents the latest evidence on its etiology, diagnosis, severity, and antibiotic management. Therefore, it could be considered a high-level scientific evidence abstract in its specific area of focus.

GPT4: Diagnóstico 3, Consulta 3

The highest level of scientific evidence among the abstracts provided appears to be from Abstract 4, which is a prospective observational study. This type of study design is generally considered to provide a higher level of evidence than retrospective studies or reviews, as it allows for the collection of data over time and can help establish cause-and-effect relationships. [Abstract 4]

El razonamiento proporcionado por GPT-4 en esta consulta es más acertado, citando el estudio observacional prospectivo de artículo 4.

GPT3.5-turbo: Diagnóstico 3, Consulta 9

There is no clear answer to this question based on the provided abstracts. However, Abstract 1 mentions that appropriate therapy is important in deciding on the site of care for elderly patients with community-acquired pneumonia and multiple comorbid illnesses. Abstract 3 also emphasizes the importance of early antimicrobial therapy and appropriate medications to cover typical pathogens in treating community-acquired pneumonia in older individuals. Additionally, Abstract 5 suggests that alcohol use disorder is associated with greater pneumonia severity in hospitalized HIV-infected patients.

GPT4: Diagnóstico 3, Consulta 9

Based on the provided abstracts, there is no explicit mention of a treatment for the associated diagnoses (Alcohol dependence, Acute on chronic renal failure) that worsens the main disease (Community acquired pneumonia). However, it is important to note that alcohol dependence can increase the severity of pneumonia [Abstract 5] and acute alcohol intoxication can suppress the phagocytic function of neutrophils, potentially worsening the body's response to pneumonia [Abstract 8]. Chronic kidney disease can also increase the risk of cardiovascular disease [Abstract 6], which could potentially complicate the management of pneumonia, but this is not directly stated in the provided abstracts.

En este caso los dos modelos responden parcialmente mal. Aunque ambos mencionan que los *abstracts* no disponen de información para responder a la consulta, después centran su argumento en lo que mejora o lo que no mejora la enfermedad principal, en lugar de en lo que la empeora.

GPT3.5-turbo: Diagnóstico 5, Consulta 5

According to [Abstract 1], patients with perforated gastric cancer had a poorer overall survival rate than those who had T3 tumors without perforation. Recurrence of peritoneum occurred more frequently in patients with perforation. However, none of the other abstracts provide information on whether any associated diagnoses worsen the main disease.

GPT4: Diagnóstico 5, Consulta 5

Yes, the presence of gastric cancer in patients with gastric perforation can worsen the overall survival rate. Patients with gastric cancer perforation had a poorer overall survival rate than those who had T3 tumors without perforation. In addition, in patients with perforation, recurrence of peritoneum occurred more frequently [Abstract 1]. Furthermore, low serum albumin levels can predict the need for gastric resection in patients with perforated peptic ulcer, which can complicate the disease [Abstract 3].

En esta interacción GPT-3.5-turbo confunde el impacto de la perforación gástrica en los pacientes con cáncer frente a los que tienen cáncer gástrico, pero no han tenido dicha perforación. Por su parte, GPT-4 comete el mismo error, ya que la respuesta deseada sería "los pacientes que presentan cáncer gástrico y sufren perforación tienen peor supervivencia".

En general se ha observado un razonamiento muy superior en las respuestas de GPT-4, respondiendo bien a 5 de las 7 preguntas que GPT-3.5-turbo había respon-

dido mal. Por tanto, GPT-4 se posiciona como un LLM más fiable en esta tarea, algo presumible al ser un modelo mucho más potente y lento.

6.3.4. Batería de consultas a posteriori

Adicionalmente, los expertos de IBSAL han diseñado una serie de 15 consultas concretas, 3 para cada diagnóstico, después de conocer el contenido de los *abstracts* recomendados para cada diagnóstico. Las consultas se pueden ver en la Tabla 6.5,

donde se recogen cada una de 3 preguntas diseñadas para cada uno de los 5 diagnósticos. La tasa de acierto conseguida en este experimento ha sido del 100 %, por lo que, a falta de un número mayor de experimentos, el sistema demuestra una gran robustez y precisión al responder preguntas basadas en contexto inyectado. Nótese que sólo se ha realizado este experimento con GPT-3.5-turbo, ya que ha demostrado la máxima fiabilidad. Para este caso de uso o similares, la viabilidad de GPT-3.5-turbo posibilita su implementación en un mayor número de escenarios, ya que, aunque GPT-4 ha demostrado ser más robusto en el experimento anterior, también es mucho más lento y caro en el momento de escritura de esta tesis.

6.4. Conclusiones

En este capítulo se ha presentado el componente final de la tesis, un asistente de investigación médica que utiliza un LLM y un sistema de recomendación de artículos médicos (MELENDI) para dar respuestas basadas en conocimiento a consultas predefinidas o en tiempo real por parte de los facultativos. Estas consultas se realizan en lenguaje natural y se utilizan documentos de Pubmed para fundamentar las respuestas de GPT-3.5-turbo y GPT-4 mediante la inyección de contexto. La principal novedad introducida es la propuesta de un marco de trabajo completo que combina métricas de relevancia, reputación, similitud semántica, una base de datos de artículos médicos de referencia en el ámbito de la investigación biomédica y un LLM estado del arte para asistir a los investigadores biomédicos. Además, la naturaleza pro-activa del asistente, permite realizar consultas relacionadas con las características de los pacientes del usuario gracias a la extracción automática de sus diagnósticos provenientes de sus EHR. Todas estas características permiten un enorme ahorro de tiempo a los médicos a la hora de mantenerse actualizados en sus respectivos campos.

Para validar la propuesta se ha realizado una batería de experimentos en los que se ha involucrado a una serie de expertos en el dominio médico provenientes de IBSAL. Los experimentos se han diseñado para evaluar la calidad de las respuestas del asistente y si las mismas están única y exclusivamente basadas en el contexto inyectado proveniente de los artículos recomendados. En primer lugar, se han definido

10 consultas generalistas relacionadas sobre todo con relaciones entre las diferentes enfermedades que componen los diagnósticos y sobre el nivel de evidencia científica de los artículos recomendados. Posteriormente, se han seleccionado 5 diagnósticos sobre los que se ha obtenido una recomendación de 10 artículos relacionados para cada uno. Una vez obtenido el contexto proporcionado por cada conjunto de artículos, se ha realizado cada una de las 10 consultas sobre cada conjunto utilizando GPT-3.5-turbo y un *prompt* diseñado específicamente, dando un total de 50 consultas. De las 50 consultas, 43 han sido evaluadas como correctas, llevando a una precisión del 86 %. Después, se han repetido las consultas evaluadas como incorrectas utilizando GPT-4, respondiendo bien a 5 de las 7 preguntas erróneas, lo que demuestra una mejora clara con respecto a GPT-3.5-turbo. Por último, el sistema se ha evaluado con una batería adicional de 15 consultas diseñadas después de conocer los *abstracts* recomendados para cada diagnóstico, con el fin de evaluar el sistema con consultas concretas sobre cada diagnóstico, conociendo dónde se encuentra el resultado de cada respuesta dentro de los *abstracts*. En este experimento se ha logrado un 100 % de precisión con GPT-3.5-turbo, demostrando una gran fiabilidad del LLM, a pesar de los errores cometidos en preguntas generalistas.

Aunque se han obtenido unos magníficos resultados, el asistente como pieza aislada del resto de la tesis tiene algunos posibles puntos de mejora para trabajos futuros. En primer lugar, las interacciones con el asistente son independientes unas de otras. En otras palabras, el asistente olvida las conversaciones y las respuestas anteriores y cada interacción funciona como una nueva conversación. Se podrían implementar estrategias para mantener parte de la conversación dentro del contexto del asistente, pero se ha decidido dejar fuera del alcance del presente trabajo debido a la gran cantidad de contexto que supone la inyección del texto de los artículos recomendados, limitando la cantidad de conversación potencial almacenada. Esta limitación puede solventarse en parte haciendo uso también de modelos con mayor capacidad de contexto, aunque estos modelos suelen ser más caros. Por otra parte, existe una limitación en términos de experimentos realizados variando los parámetros internos del LLM. Aunque una temperatura baja es a priori la mejor elección para este caso de uso, esta interactúa con el parámetro *top_p*, por lo que más experimentos con variaciones de estos parámetros, así como de *frequency_penalty* y *presence_penalty* resultan interesantes. En último lugar, un estudio sobre la influencia de la variedad y número de enfermedades que componen el diagnóstico sobre la calidad de las respuestas podría ser muy revelador, ya que los experimentos realizados parecen indicar la existencia de dicha influencia.

En el siguiente capítulo, el último de la tesis, se hace una descripción global sobre los logros, limitaciones y trabajos futuros del sistema propuesto en la misma.

ID de diagnóstico	ID	Consulta
1	1	Does hyponatremia increase the risk of cardiovascular mortality?
	2	Are vasopressin receptor antagonists useful in the treatment of hyponatremia?
	3	Is it true that small cell lung cancer has no relation to SIADH?
2	4	Does apigenin improve the prognosis of lung disease associated with acute pancreatitis?
	5	Is apigenin beneficial in the treatment of acute pancreatitis?
	6	Do serum CA125 levels predict worse outcomes after surgery in metastatic pancreatic cancer?
3	7	What is the main cause of community-acquired pneumonia?
	8	Does alcohol abuse disease improve the life expectancy of patients with pneumonia?
	9	Does alcohol enhance neutrophil-mediated cellular phagocytosis?
4	10	Is amyotrophic lateral sclerosis a reversible disease?
	11	Is it true that patients with amyotrophic lateral sclerosis experience no pain of any kind?
	12	Is it true that the TDP43 protein is never aggregated in patients with amyotrophic lateral sclerosis?
5	13	Is it true that gastric cancer patients who suffer from intestinal perforation have a worse prognosis?
	14	Is gastric perforation more common than duodenal perforation?
	15	Is serum albumin a predictor of the need for gastric surgery?

Tabla 6.5: Consultas de control diseñadas después de conocer los abstracts recomendados para cada diagnóstico.

Capítulo 7

Conclusiones

Aquí se presenta el capítulo final de la tesis. Se comienza repasando las principales contribuciones conseguidas a lo largo de la misma, tanto a nivel académico como industrial. Se continúa con un análisis de las limitaciones y posibles trabajos futuros. Finalmente, se aportan las publicaciones realizadas derivadas de la investigación.

7.1. Contribuciones

Este trabajo se ha realizado como un esfuerzo colaborativo conjunto entre la URJC, IBSAL y la empresa Dezzai, por tanto, los objetivos planteados al inicio de la tesis se han ajustado a las necesidades estratégicas y de negocio de Dezzai. Esto, sin perder de vista el afán por la innovación que caracteriza a las tres entidades. Este afán ha dado lugar a una serie de contribuciones científicas directas e indirectas del desarrollo de la tesis, que no sólo han servido como contribución propiamente científica, si no que sirven para acreditar a Dezzai como una empresa con una fuerte apuesta por la innovación tecnológica. Apuesta especialmente firme en el campo de la inteligencia artificial.

En la tesis se ha presentado una propuesta para un asistente de investigación médica basado en conocimiento. El objetivo del asistente es acelerar el proceso de actualización de los médicos en sus respectivos campos y especialidades. Dada la enorme cantidad de literatura científica existente, sumado al tiempo que los médicos deben pasar con sus pacientes, les deja poco margen para mantenerse actualizados sobre nuevos tratamientos, medicamentos o estudios clínicos. Además, la discriminación de artículos científicos relevantes resulta cada vez más difícil ante la creciente cantidad de publicaciones. De esta manera, el asistente pretende proporcionar una solución integral a toda esta problemática, ofreciendo un sistema capaz de hacer

recomendaciones de artículos científicos relevantes sobre las características de los pacientes. Estas recomendaciones se realizan sin la necesidad de intervención del usuario, ya que el sistema extrae las características de los pacientes (las enfermedades que componen su diagnóstico) de forma automática a través de los EHR. Una vez realizadas las recomendaciones sobre los diversos pacientes, el sistema permite realizar consultas en lenguaje natural sobre las mismas, de forma que el usuario pueda discriminar aquellos artículos de su interés sin la necesidad de conocer complejos sistemas de filtrado ni de leer todas las recomendaciones realizadas. Para lograr este objetivo, el asistente se ha dividido en tres componentes principales: El primer componente (ver Capítulo 4), bautizado como *Webelance*, es un marco de trabajo que permite estimar la relevancia de artículos científicos. El segundo (ver Capítulo 5), denominado MELENDI, combina la relevancia estimada por *Webelance* con similitud semántica para realizar recomendaciones de artículos. Finalmente, el componente denominado GAMER (ver Capítulo 6), incorpora un LLM controlado con inyección de contexto para realizar consultas en lenguaje natural sobre las recomendaciones.

Aunque la principal contribución del sistema completo es el potencial ahorro de tiempo proporcionado a los médicos o investigadores, cabe destacar algunas de las aportaciones más importantes logradas por cada componente de forma independiente:

- *Estimador de relevancia de artículos*
 - El sistema no se limita a una única métrica objetiva para estimar la relevancia, lo cuál puede dar lugar a limitaciones o sesgos. En su lugar, combina diferentes métricas y fuentes de información, lo que produce una estimación holística y robusta de la relevancia.
 - El sistema puede ser entrenado para cualquier dominio específico de interés, demostrando flexibilidad y usabilidad.
 - Al implementar un diccionario y modelos de aprendizaje máquina para las predicciones, el sistema asegura que puede manejar una amplia gama de textos, incluso aquellos que presentan palabras inusuales del dominio. El diccionario provee una base sólida, mientras que los modelos de ML pueden capturar las particularidades y palabras menos frecuentes del dominio.
 - Al considerar la reputación de los artículos basada en la reputación de los autores y el número de citas, el sistema introduce un elemento de conocimiento externo en la estimación de relevancia. Esto reconoce la importancia de la influencia y el reconocimiento comunitario en la valoración de trabajos científicos.

- El sistema ha demostrado un buen rendimiento evaluando artículos previamente etiquetados por expertos. Esta capacidad de evaluación y retroalimentación continua permite al sistema mejorar y adaptarse con el tiempo.
 - Más allá de evaluar la relevancia en un momento dado, el sistema puede detectar tendencias a lo largo del tiempo. Esta habilidad es valiosa para identificar cambios en el campo y orientar a investigadores sobre áreas emergentes o de interés creciente.
- *Recomendador de artículos*
- MELENDI combina un modelo de similitud semántica con un estimador automático de la relevancia de artículos científicos. Esta integración permite al sistema hacer recomendaciones que son relevantes tanto en términos académicos como contextualmente pertinentes para el diagnóstico en cuestión. Esto asegura que los profesionales de la salud reciban información que sea directamente aplicable a su contexto clínico.
 - Aunque en este trabajo el sistema está adaptado a los informes de alta de MIMIC-III, la naturaleza modular de la arquitectura propuesta permite desacoplar esta implementación y adaptarla a otros orígenes de datos, así como a consultas manuales por parte del usuario.
 - La solución propuesta prioriza un equilibrio entre relevancia y similitud semántica, no obstante este compromiso se puede alterar, priorizando una sobre otra en función de las necesidades y el dominio de aplicación.
 - El sistema recomienda artículos basados en los diagnósticos de los pacientes de forma automática, sin la necesidad de intervención del usuario. Este hecho evita la necesidad de conocimiento técnico específico propio de los motores de búsqueda y los sistemas de recuperación de información.
- *Asistente generativo contextualizado*
- Se combina un LLM estado del arte con un sistema de recomendación de artículos médicos basado en conocimiento, ofreciendo un enfoque completo para responder a consultas sobre los artículos recomendados.
 - Las respuestas proporcionadas por el asistente se basan en artículos biomédicos indexados en Pubmed, asegurando que la información proporcionada es relevante y actualizada.
 - Los usuarios pueden hacer consultas en lenguaje natural, lo que facilita su interacción con el asistente y elimina la necesidad de aprender comandos específicos.

- La naturaleza modular del sistema permite ajustar los parámetros internos del LLM o sustituirlo por otro.

La métrica de relevancia de artículos propuesta en *Webelance* se encuentra implementada ya en una solución integral de un motor de búsqueda en un hospital. Este motor de búsqueda es un producto llamado *Dezzai Search* que está evolucionando para la integración de MELENDI y GAMER, encontrándose en el momento de escritura de esta tesis en fase de implementación y escalado. Una serie de clientes procedentes de diversos laboratorios de investigación médica han manifestado su interés en la solución integral aquí propuesta, evidenciando que este trabajo responde a una necesidad real de la industria.

Cabe señalar que el componente basado en el asistente generativo no estaba en el plan inicial de investigación. No obstante, la súbita irrupción de los modelos generativos ha generado un enorme interés en la industria, obligando a Dezzai a replantear su estrategia comercial y, por tanto, adaptar el plan de investigación a la misma.

7.2. Trabajo futuro

A pesar los buenos resultados mostrados por cada uno de los componentes y el sistema completo, este trabajo ha contado con una serie de limitaciones. Esta sección pretende abordar estas limitaciones y otras posibles mejoras a través de trabajos futuros potencialmente implementables.

En cuanto a la estimación de la relevancia, el estudio tiene limitaciones en cuanto al número de artículos evaluados por expertos usados para testear el sistema, y sólo contempla la relevancia de estos a lo largo de cuatro años. Para trabajos futuros, es esencial expandir la base de datos con más artículos evaluados y considerar períodos de tiempo más extensos. Además, esto permitiría un mejor ajuste de los parámetros internos. Por otra parte, la métrica de relevancia se ha adaptado únicamente al dominio médico, sería interesante estudiar cómo se comporta en otros dominios, dado que la configuración neutral de parámetros ha dado unos resultados razonablemente buenos en el dominio médico. Además, sería interesante evaluar métricas alternativas al TF-IDF y explorar la eficacia de n-gramas en comparación con los uni-gramas, aunque los uni-gramas hayan mostrado mejores resultados previamente.

Por otra parte, MELENDI se ha adaptado y evaluado teniendo en cuenta la casuística observada en los informes de alta de MIMIC-III, por tanto, su aplicabilidad con otro tipo de EHR no es directa y requiere adaptación. La creación de un sistema generalista que detecte diagnósticos en narrativas clínicas de forma automática sería de gran interés, aunque probablemente requiera de un enorme volumen de datos etiquetados. Otra posible mejora vendría del uso o entrenamiento de un modelo

de similitud semántica adaptado al dominio médico, ya que el utilizado es de carácter generalista. Una vez que el sistema completo esté en producción, podrá recoger la retroalimentación de los usuarios para este fin así como para ajustar el equilibrio entre relevancia y similitud semántica. En cuanto a la escalabilidad del sistema de recomendación, está planeada la inclusión de la totalidad de *abstracts* indexados en *Pubmed* y los artículos completos de *Pubmed Central*, así como una interfaz para que el usuario pueda subir sus propios documentos.

Finalmente, las interacciones con el asistente generativo son independientes unas de otras. En otras palabras, el asistente olvida las conversaciones y las respuestas anteriores, de modo que cada interacción funciona como una nueva conversación. Se podrían implementar estrategias para mantener parte de la conversación dentro del contexto del asistente, pero se ha decidido dejar fuera del alcance del presente trabajo debido a la gran cantidad de contexto que supone la inyección del texto de los artículos recomendados, limitando la cantidad de conversación potencial almacenada. Esta limitación puede solventarse, en parte, haciendo uso también de modelos con mayor capacidad de contexto, aunque estos modelos suelen ser más caros. Por otra parte, existe una limitación en términos de experimentos realizados variando los parámetros internos del LLM. Aunque una temperatura baja es a priori la mejor elección para este caso de uso, esta interactúa con el parámetro *top_p*, por lo que más experimentos con variaciones de estos parámetros, así como de *frequency_penalty* y *presence_penalty* resultan interesantes. En último lugar, un estudio sobre la influencia de la variedad y número de enfermedades que componen el diagnóstico sobre la calidad de las respuestas podría ser muy revelador, ya que los experimentos realizados parecen indicar la existencia de dicha influencia.

Publicaciones derivadas de la investigación:

- Fernández-Isabel, A., Barriuso, A. A., Cabezas, J., de Diego, I. M., Pinheiro, J. V. (2020). Knowledge-based framework for estimating the relevance of scientific articles. *Expert Systems with Applications*, 161, 113692.
- Recommendation System of Scientific Articles from Discharge Summaries. Barriuso, A. A., Fernández-Isabel, A., de Diego, I. M., Galaz, A. A., Pinheiro, J. V. (2023). *Artificial Intelligence In Medicine* (En Revisión).

Referencias

- Abramo, G., D'Angelo, C. A., y Felici, G. (2019). Predicting publication long-term impact through a combination of early citations and journal impact factor. *Journal of Informetrics*, 13(1), 32–49.
- Afzali, M., y Kumar, S. (2019). Text document clustering: issues and challenges. En *2019 international conference on machine learning, big data, cloud and parallel computing (comitcon)* (pp. 263–268).
- Akila, D., y Balaganesh, D. (2021). Semantic web-based critical healthcare system using bayesian networks. *Materials today: proceedings*.
- Akram, J., y Tahir, A. (2018). Lexicon and heuristics based approach for identification of emotion in text. En *2018 international conference on frontiers of information technology (fit)* (pp. 293–297).
- Alfano, M., Lenzitti, B., Lo Bosco, G., y Taibi, D. (2018). Development and practical use of a medical vocabulary-thesaurus-dictionary for patient empowerment. En *Proceedings of the 19th international conference on computer systems and technologies* (pp. 88–93).
- Alkaissi, H., y McFarlane, S. I. (2023). Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).
- Allen Institute for Artificial Intelligence and Semantic Scholar. (2018). *Semantic Scholar API*. <https://api.semanticscholar.org/>. ([En línea: último acceso 20-11-2023])
- Alves, M. A., Castro, G. Z., Oliveira, B. A. S., Ferreira, L. A., Ramírez, J. A., Silva, R., y Guimarães, F. G. (2021). Explaining machine learning based diagnosis of covid-19 from routine blood tests with decision trees and criteria graphs. *Computers in Biology and Medicine*, 132, 104335.
- Anandarajan, M., Hill, C., y Nolan, T. (2019). Term-document representation. En *Practical text analytics* (pp. 61–73). Springer.

- Arora, P., Boyne, D., Slater, J. J., Gupta, A., Brenner, D. R., y Druzdzal, M. J. (2019). Bayesian networks for risk prediction using real-world data: a tool for precision medicine. *Value in Health*, 22(4), 439–445.
- Averell, L., y Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, 55(1), 25–35.
- Baccianella, S., Esuli, A., y Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. En *Lrec* (Vol. 10, pp. 2200–2204).
- Bandhakavi, A., Wiratunga, N., Massie, S., y Padmanabhan, D. (2017). Lexicon generation for emotion detection from text. *IEEE intelligent systems*, 32(1), 102–108.
- Benedetti, F., Beneventano, D., Bergamaschi, S., y Simonini, G. (2019). Computing inter-document similarity with context semantic analysis. *Information Systems*, 80, 136–147.
- Bentaiba-Lagrid, M. B., Bouzar-Benlabiod, L., Rubin, S. H., Bouabana-Tebibel, T., y Hanini, M. R. (2020). A case-based reasoning system for supervised classification problems in the medical field. *Expert Systems with Applications*, 150, 113335.
- Bhavsar, K., Kumar, N., y Dangeti, P. (2017). *Natural language processing with python cookbook: Over 60 recipes to implement text analytics solutions using deep learning principles*. Packt Publishing Ltd.
- Bloching, P. A., y Heinzl, H. (2013). Assessing the scientific relevance of a single publication over time. *South African Journal of Science*, 109(9), 1–2.
- Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1), D267–D270.
- Bracewell, D. B. (2008). Semi-automatic creation of an emotion dictionary using wordnet and its evaluation. En *2008 ieee conference on cybernetics and intelligent systems* (pp. 1385–1389).
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2), 102–107.
- Cambria, E., Poria, S., Gelbukh, A., y Kwok, K. (2014). Sentic api: a common-sense based api for concept-level sentiment analysis.
- Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, 50(3), 1–42.

- Cascella, M., Montomoli, J., Bellini, V., y Bignami, E. (2023). Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems*, 47(1), 33.
- Castiglioni, A. (2019). *A history of medicine*. Routledge.
- Castiglioni, I., Ippolito, D., Interlenghi, M., Monti, C. B., Salvatore, C., Schiaffino, S., ... Sardanelli, F. (2021). Machine learning applied on chest x-ray can aid in the diagnosis of covid-19: a first experience from lombardy, italy. *European radiology experimental*, 5(1), 1–10.
- Cecchini, R. L., Lorenzetti, C. M., Maguitman, A. G., y Ponzoni, I. (2018). Topic relevance and diversity in information retrieval from large datasets: A multi-objective evolutionary algorithm approach. *Applied Soft Computing*, 69, 749–770.
- Chandramouli, A. (2018, agosto 9). *Domain-specific stopword removal from unstructured computer text using a neural network*. Google Patents. (US Patent App. 15/426,958)
- Chen, K., Zhang, Z., Long, J., y Zhang, H. (2016). Turning from tf-idf to tf-igm for term weighting in text classification. *Expert Systems with Applications*, 66, 245–260.
- Chen, Q., Kim, S., Wilbur, W. J., y Lu, Z. (2018). Sentence similarity measures revisited: Ranking sentences in pubmed documents. En *Proceedings of the 2018 acm international conference on bioinformatics, computational biology, and health informatics* (pp. 531–532).
- Chen, Y., Beynon, J. A., Perlov, B., Ghatare, S. P., Bolivar, A., Parikh, N., ... Song, G. (2014, marzo 18). *Methods and apparatus for generating a data dictionary*. Google Patents. (US Patent 8,676,829)
- Chen, Z., Cafarella, M., y Jagadish, H. (2016). Long-tail vocabulary dictionary extraction from the web. En *Proceedings of the ninth acm international conference on web search and data mining* (pp. 625–634).
- Cheng, Y., y Huang, Y. (2017). Research and development of domain dictionary construction system. En *Proceedings of the international conference on web intelligence* (pp. 1162–1165).
- Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., ... others (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. Ver <https://vicuna.lmsys.org> (En línea: último acceso 20-11-2023).
- Chikersal, P., Poria, S., y Cambria, E. (2015). Sentu: sentiment analysis of tweets by combining a rule-based classifier with supervised learning. En *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)* (pp. 647–651).
- Chow, J. C., Sanders, L., y Li, K. (2023). Impact of chatgpt on medical chatbots as a

- disruptive technology. *Frontiers in Artificial Intelligence*, 6, 1166014.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... others (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Chowdhury, G. G. (2010). *Introduction to modern information retrieval*. Facet publishing.
- De Croon, R., Van Houdt, L., Htun, N. N., Štiglic, G., Abeele, V. V., Verbert, K., y cols. (2021). Health recommender systems: systematic review. *Journal of Medical Internet Research*, 23(6), e18035.
- de Diego, I. M., Fernández-Isabel, A., Ortega, F., y Moguerza, J. M. (2018). A visual framework for dynamic emotional web analysis. *Knowledge-Based Systems*, 145, 264–273.
- Deng, D., Jing, L., Yu, J., y Ng, M. K. (2018). Topic-adaptive sentiment lexicon construction. En *2018 first asian conference on affective computing and intelligent interaction (acii asia)* (pp. 1–6).
- Deng, D., Jing, L., Yu, J., Sun, S., y Ng, M. K. (2019). Sentiment lexicon construction with hierarchical supervision topic model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4), 704–718.
- Deng, J., y Lin, Y. (2022). The benefits and challenges of chatgpt: An overview. *Frontiers in Computing and Intelligent Systems*, 2(2), 81–83.
- Deng, Q., Hine, M., Ji, S., y Sur, S. (2017). Building an environmental sustainability dictionary for the it industry. En *Proceedings of the 50th hawaii international conference on system sciences*.
- Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Donoso-Guzmán, I., y Parra, D. (2018). An interactive relevance feedback interface for evidence-based health care. En *23rd international conference on intelligent user interfaces* (pp. 103–114).
- Duan, J., y Jiao, F. (2021). Novel case-based reasoning system for public health emergencies. *Risk Management and Healthcare Policy*, 14, 541.
- ElDin, H. G., AbdulRazek, M., Abdelshafi, M., y Sahlol, A. T. (2021). Med-flair: medical named entity recognition for diseases and medications based on flair embedding. *Procedia Computer Science*, 189, 67–75.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., y Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *natu-*

- re, 542(7639), 115–118.
- Fellbaum, C. (2010). Wordnet. En *Theory and applications of ontology: computer applications* (pp. 231–243). Springer.
- Fernández-Isabel, A., Cabezas, J., Moctezuma, D., y de Diego, I. M. (2023). Improving sentiment classification performance through coaching architectures. *Cognitive Computation*, 15(3), 1065–1081.
- Fernández-Isabel, A., Prieto, J. C., Ortega, F., de Diego, I. M., Moguerza, J. M., Mena, J., ... Napalkova, L. (2018). A unified knowledge compiler to provide support the scientific community. *Knowledge-Based Systems*, 161, 157–171.
- Fiorini, N., Canese, K., Starchenko, G., Kireev, E., Kim, W., Miller, V., ... others (2018). Best match: new relevance search for pubmed. *PLoS biology*, 16(8), e2005343.
- Freitas, C. O., Bortolozzi, F., y Sabourin, R. (2004). Study of perceptual similarity between different lexicons. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(07), 1321–1338.
- Fu, X., Yang, J., Li, J., Fang, M., y Wang, H. (2018). Lexicon-enhanced lstm with attention for general sentiment analysis. *IEEE Access*, 6, 71884–71891.
- Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F. J., y Medina-Carnicer, R. (2016). Generation of fiducial marker dictionaries using mixed integer linear programming. *Pattern Recognition*, 51, 481–491.
- Gianola, S., Jesus, T. S., Bargerri, S., y Castellini, G. (2020). Characteristics of academic publications, preprints, and registered clinical trials on the covid-19 pandemic. *PloS one*, 15(10), e0240123.
- Goeckenjan, G., Sitter, H., Thomas, M., Branscheid, D., Flentje, M., Griesinger, F., ... others (2011). Pubmed results. *Pneumologie*, 65(8), e51–e75.
- Goeriot, L., Na, J.-C., Min Kyaing, W. Y., Khoo, C., Chang, Y.-K., Theng, Y.-L., y Kim, J.-J. (2012). Sentiment lexicons for health-related opinion mining. En *Proceedings of the 2nd acm sighth international health informatics symposium* (pp. 219–226).
- Gormley, C., y Tong, Z. (2015). *Elasticsearch: The definitive guide: A distributed real-time search and analytics engine*. O'Reilly Media, Inc.
- Gupta, S. (2015). *Distantly supervised information extraction using bootstrapped patterns* (Tesis Doctoral no publicada). Stanford University.
- Han, H., Zhang, J., Yang, J., Shen, Y., y Zhang, Y. (2018). Generate domain-specific sentiment lexicon for review sentiment analysis. *Multimedia Tools and Applications*, 77(16), 21265–21280.
- Han, J., Haihong, E., Le, G., y Du, J. (2011). Survey on nosql database. En *2011 6th*

- international conference on pervasive computing and applications* (pp. 363–366).
- Hofstätter, S., Lin, S.-C., Yang, J.-H., Lin, J., y Hanbury, A. (2021). Efficiently teaching an effective dense retriever with balanced topic aware sampling. *arXiv preprint arXiv:2104.06967*.
- Hunt, E. B. (2014). *Artificial intelligence*. Academic Press.
- Isern, D., y Moreno, A. (2016). A systematic literature review of agents applied in healthcare. *Journal of medical systems*, 40(2), 1–14.
- Jabri, S., Dahbi, A., Gadi, T., y Bassir, A. (2018). Ranking of text documents using tf-idf weighting and association rules mining. En *2018 4th international conference on optimization and applications (icoa)* (pp. 1–6).
- Javaid, M., Haleem, A., y Singh, R. P. (2023). Chatgpt for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(1), 100105.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
- Jiang, Z., El-Jaroudi, A., Hartmann, W., Karakos, D., y Zhao, L. (s.f.). Cross-lingual information retrieval with bert. En *Lrec 2020 language resources and evaluation conference 11–16 may 2020* (p. 26).
- Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., ... Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1), 1–9.
- Johnson, R., Watkinson, A., y Mabe, M. (2018). The STM report. *An overview of scientific and scholarly publishing. 5th edition October*, 94.
- Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., y Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1), 1–12.
- Kherwa, P., y Bansal, P. (2019). Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, 7(24).
- Khodabandehloo, E., Riboni, D., y Alimohammadi, A. (2021). HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline. *Future Generation Computer Systems*, 116, 168–189.
- Khoo, C. S., y Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4), 491–511.
- Kolodner, J. (2014). *Case-based reasoning*. Morgan Kaufmann.

- Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., ... others (2021). Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117, 102083.
- Kumar, K. R. (2016). Mathematical modeling of consciousness: Subjectivity of mind. En *2016 international conference on circuit, power and computing technologies (iccpct)* (pp. 1–5).
- Kutsumi, H., Naito, E., Ozawa, J., y Yoshida, H. (2010, noviembre 23). *Dictionary creation device and dictionary creation method*. Google Patents. (US Patent 7,840,565)
- Lagopoulos, A., Anagnostou, A., Minas, A., y Tsoumakas, G. (2018). Learning-to-rank and relevance feedback for literature appraisal in empirical medicine. En *International conference of the cross-language evaluation forum for european languages* (pp. 52–63).
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., y Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240.
- Leefmann, J., Levallois, C., y Hildt, E. (2016). Neuroethics 1995–2012. a bibliometric analysis of the guiding themes of an emerging research field. *Frontiers in human neuroscience*, 10, 336.
- Lei, Z., Yang, Y., y Yang, M. (2018). Sentiment lexicon enhanced attention-based lstm for sentiment classification. En *Thirty-second aaaa conference on artificial intelligence*.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Li, B., y Han, L. (2013). Distance weighted cosine similarity measure for text classification. En *Intelligent data engineering and automated learning—ideal 2013: 14th international conference, ideal 2013, hefei, china, october 20-23, 2013. proceedings 14* (pp. 611–618).
- Liebowitz, J. (2019). *The handbook of applied expert systems*. cRc Press.
- Liu, H., Christiansen, T., Baumgartner, W. A., y Verspoor, K. (2012). Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of biomedical semantics*, 3(1), 3.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Lopez, M. M., y Kalita, J. (2017). Deep learning applied to nlp. *arXiv preprint ar-*

Xiv:1703.03091.

- Lowe, D. M., y Sayle, R. A. (2015). Leadmine: a grammar and dictionary driven approach to entity recognition. *Journal of cheminformatics*, 7(1), S5.
- Lu, Y., y Tanne, M. (2017, julio 25). *Systems for and methods of finding relevant documents by analyzing tags*. Google Patents. (US Patent 9,715,542)
- Macdonald, E., Chu, A., Cooper, M., Croxford, R., Duchon, R., Fung, K., ... others (2018). Development of a concept dictionary to standardize definitions and classifications while working with a common repository of linked administrative data. *International Journal of Population Data Science*, 3(4).
- MacDonald, K., y Drew, W. (2018). An improvement of information retrieval systems. *International Journal of Software Systems Research and Methodology*, 2(1).
- Mala, V., y Lobiyal, D. (2016). Semantic and keyword based web techniques in information retrieval. En *2016 international conference on computing, communication and automation (iccca)* (pp. 23–26).
- Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., y McClosky, D. (2014). The stanford corenlp natural language processing toolkit. En *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55–60).
- Mansour, R. F., El Amraoui, A., Nouaouri, I., Díaz, V. G., Gupta, D., y Kumar, S. (2021). Artificial intelligence and internet of things enabled disease diagnosis model for smart healthcare systems. *IEEE Access*, 9, 45137–45146.
- Marcot, B. G., y Penman, T. D. (2019). Advances in bayesian network modelling: Integration of modelling technologies. *Environmental modelling & software*, 111, 386–393.
- Martinez, M. L., Vazquez, D. A., Maya, D., Olvera, X., Guzman, G., Torres, M., ... Moreno, M. (2014). Geospatial recommender system for the location of health services. En *2014 14th international conference on computational science and its applications* (pp. 200–203).
- Merriam-Webster, I. (1995). *Merriam-webster's medical dictionary*. Merriam-Webster.
- Mihalcea, R., y Tarau, P. (2004). Textrank: Bringing order into text. En *Proceedings of the 2004 conference on empirical methods in natural language processing*.
- Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11), 39–41.
- Mitchell, R. (2018). *Web scraping with python: Collecting more data from the modern web*. O'Reilly Media, Inc.

- More, B. (2016, 11). Overview of medicine- its importance and impact. *DJ International Journal Medical Research*, 1, 1-8.
- Mousavi, S. M., Abdullah, S., Niaki, S. T. A., y Banihashemi, S. (2021). An intelligent hybrid classification algorithm integrating fuzzy rule-based extraction and harmony search optimization: Medical diagnosis applications. *Knowledge-Based Systems*, 220, 106943.
- Moussa, M. E., Mohamed, E. H., y Haggag, M. H. (2018). A generic lexicon-based framework for sentiment analysis. *International Journal of Computers and Applications*, 1–11.
- Mowlaei, M. E., Abadeh, M. S., y Keshavarz, H. (2018). Lexicon generation using genetic algorithm for aspect-based sentiment analysis. En *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)* (pp. 000133–000138).
- Nadkarni, P. M., Ohno-Machado, L., y Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5), 544–551.
- Neumann, M., King, D., Beltagy, I., y Ammar, W. (2019). Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.
- Olivas-Martínez, A., Cárdenas-Fragoso, J. L., Jiménez, J. V., Lozano-Cruz, O. A., Ortiz-Brizuela, E., Tovar-Méndez, V. H., ... others (2021). In-hospital mortality from severe covid-19 in a tertiary care center in Mexico City; causes of death, risk factors and the impact of hospital saturation. *Plos one*, 16(2), e0245772.
- O’Sullivan, D. (2016). *Wikipedia: a new community of practice?* Routledge.
- Otter, D. W., Medina, J. R., y Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2), 604–624.
- Park, D., Kim, S., Lee, J., Choo, J., Diakopoulos, N., y Elmqvist, N. (2018). Concept-vector: text visual analytics via interactive lexicon building using word embedding. *IEEE transactions on visualization and computer graphics*, 24(1), 361–370.
- Paskin, N. (2010). Digital object identifier (doi®) system. *Encyclopedia of library and information sciences*, 3, 1586–1592.
- Passalis, N., y Tefas, A. (2018). Learning bag-of-embedded-words representations for textual information retrieval. *Pattern Recognition*, 81, 254–267.
- Patil, H. B., Mhaske, N. T., y Patil, A. S. (2017). Design and development of a dictionary based stemmer for marathi language. En *International conference on next generation computing technologies* (pp. 769–777).
- Pattanittum, P., Laopaiboon, M., Moher, D., Lumbiganon, P., y Ngamjarus, C. (2012).

- A comparison of statistical methods for identifying out-of-date systematic reviews. *PLoS One*, 7(11), e48894.
- Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., ... Tao, D. (2023). Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*.
- Penha, G., Campos, R., Canuto, S., Gonçalves, M. A., y Santos, R. L. (2019). Document performance prediction for automatic text classification. En *European conference on information retrieval* (pp. 132–139).
- Pennington, J., Socher, R., y Manning, C. (2014). Glove: Global vectors for word representation. En *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).
- Perellet, R., Mason, P., Markandya, A., y Taylor, T. (2014). *Dictionary of environmental economics*. Routledge.
- Ploug, T., y Holm, S. (2020). The four dimensions of contestable ai diagnostics-a patient-centric approach to explainable ai. *Artificial Intelligence in Medicine*, 107, 101901.
- Pomares-Quimbaya, A., Kreuzthaler, M., y Schulz, S. (2019). Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. *BMC medical research methodology*, 19, 1–20.
- Poria, S., Cambria, E., y Gelbukh, A. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. En *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2539–2544).
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., y cols. (2018). Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., y cols. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1), 5485–5551.
- Ramos, J., y cols. (2003). Using tf-idf to determine word relevance in document queries. En *Proceedings of the first instructional conference on machine learning* (Vol. 242, pp. 133–142).
- Rauber, T. W., da Silva Loca, A. L., de Assis Boldt, F., Rodrigues, A. L., y Varejão, F. M. (2021). An experimental methodology to evaluate machine learning methods for fault diagnosis based on vibration signals. *Expert Systems with Applications*, 167,

114022.

- Raynaud, M., Zhang, H., Louis, K., Goutaudier, V., Wang, J., Dubourg, Q., ... others (2021). Covid-19-related medical research: a meta-research and critical appraisal. *BMC Medical Research Methodology*, 21(1), 1–11.
- Reimers, N., y Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Rezaeinia, S. M., Rahmani, R., Ghodsi, A., y Veisi, H. (2019). Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications*, 117, 139–147.
- Richta, R. (2018). *Civilization at the crossroads: Social and human implications of the scientific and technological revolution (international arts and sciences press): Social and human implications of the scientific and technological revolution*. Routledge.
- Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1), 33–36.
- Rundell, M., y Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end. *A Taste for Corpora. In Honour of Sylviane Granger*, 257–282.
- Saba, T. (2020). Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. *Journal of Infection and Public Health*, 13(9), 1274–1289.
- Saha, P., Sadi, M. S., y Islam, M. M. (2021). Emcnet: Automated covid-19 diagnosis from x-ray images using convolutional neural network and ensemble of machine learning classifiers. *Informatics in medicine unlocked*, 22, 100505.
- Sánchez, P., y Bellogín, A. (2018). Measuring anti-relevance: a study on when recommendation algorithms produce bad suggestions. En *Proceedings of the 12th acm conference on recommender systems* (pp. 367–371).
- Scheepers-Hoeks, A.-M. J., Grouls, R. J., Neef, C., Ackerman, E. W., y Korsten, E. H. (2013). Physicians' responses to clinical decision support on an intensive care unit—comparison of four different alerting methods. *Artificial intelligence in medicine*, 59(1), 33–38.
- Scherer, K. R., y cols. (1984). On the nature and function of emotion: A component process approach. *Approaches to emotion*, 2293, 317.
- Sebastiani, F., y Esuli, A. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. En *Proceedings of the 5th international conference on language resources and evaluation* (pp. 417–422).
- Shojania, K. G., Sampson, M., Ansari, M. T., Ji, J., Doucette, S., y Moher, D. (2007). How quickly do systematic reviews go out of date? a survival analysis. *Annals of internal*

- medicine*, 147(4), 224–233.
- Singla, J., Grover, D., y Bhandari, A. (2014). Medical expert systems for diagnosis of various diseases. *International Journal of Computer Applications*, 93(7).
- Solovyh, E., Obrubov, A., Arranz, I., Pérez, F., y Tejedor, M. (2021). Artificial intelligence dentomo: Opportunities and prospects for interpretation of cone beam ct in dentistry. *Bulletin of Experimental Biology and Medicine*, 170, 686–688.
- Stark, B., Knahl, C., Aydin, M., y Elish, K. (2019). A literature review on medicine recommender systems. *International journal of advanced computer science and applications*, 10(8).
- Sun, Y., Wang, L., Jiang, Z., Li, B., Hu, Y., y Tian, W. (2020). State recognition of decompressive laminectomy with multiple information in robot-assisted surgery. *Artificial Intelligence in Medicine*, 102, 101763.
- Sung, M., Jeong, M., Choi, Y., Kim, D., Lee, J., y Kang, J. (2022). Bern2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, 38(20), 4837–4839.
- Suta, P., Lan, X., Wu, B., Mongkolnam, P., y Chan, J. H. (2020). An overview of machine learning in chatbots. *International Journal of Mechanical Engineering and Robotics Research*, 9(4), 502–510.
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., y Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2), 267–307.
- Tan, H. X., Teo, C. H. D., Ang, P. S., Loke, W. P. C., Tham, M. Y., Tan, S. H., ... others (2022). Combining machine learning with a rule-based algorithm to detect and identify related entities of documented adverse drug reactions on hospital discharge summaries. *Drug safety*, 45(8), 853–862.
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... Hashimoto, T. B. (2023). Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6), 7.
- Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., y Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, 1–11.
- Tolle, K. M., y Chen, H. (2000). Comparing noun phrasing techniques for use with medical digital library tools. *Journal of the American society for information science*, 51(4), 352–370.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... others (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Valenzuela, M., Ha, V., y Etzioni, O. (2015). Identifying meaningful citations. En *Workshops at the twenty-ninth aaii conference on artificial intelligence*.
- Vassiliadis, P. (2009). A survey of extract–transform–load technology. *International Journal of Data Warehousing and Mining (IJDWM)*, 5(3), 1–27.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, M., Rao, Q., y Wang, H. (2018). Construction and application of chinese generation lexicon for chinese irregular collocation between verbs and nouns. En *Workshop on chinese lexical semantics* (pp. 667–678).
- Waqar, M., Majeed, N., Dawood, H., Daud, A., y Aljohani, N. R. (2019). An adaptive doctor-recommender system. *Behaviour & Information Technology*, 38(9), 959–973.
- Ware, M., y Mabe, M. (2015). The stm report: An overview of scientific and scholarly journal publishing.
- Weiss, K., Khoshgoftaar, T. M., y Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1), 1–40.
- Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45.
- Were, M. C., Mamlin, B. W., Tierney, W. M., Wolfe, B., y Biondich, P. G. (2007). Concept dictionary creation and maintenance under resource constraints: lessons from the ampath medical record system. En *Amia annual symposium proceedings* (Vol. 2007, p. 791).
- White, H. D. (2017). Relevance in theory. En *Encyclopedia of library and information sciences* (pp. 3926–3939). CRC Press.
- White, J. (2020). Pubmed 2.0. *Medical reference services quarterly*, 39(4), 382–387.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.
- Williamson, J. W., German, P. S., Weiss, R., Skinner, E. A., y Bowes, F. (1989). Health science information management and continuing education of physicians: a survey of us primary care practitioners and their opinion leaders. *Annals of Internal*

Medicine, 110(2), 151–160.

- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., y Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.
- Zamani, H., y Croft, W. B. (2017). Relevance-based word embedding. En *Proceedings of the 40th international acm sigir conference on research and development in information retrieval* (pp. 505–514).
- Zhang, Q., Zhang, G., Lu, J., y Wu, D. (2015). A framework of hybrid recommender system for personalized clinical prescription. En *2015 10th international conference on intelligent systems and knowledge engineering (iske)* (pp. 189–195).
- Zhang, S., Wei, Z., Wang, Y., y Liao, T. (2018). Sentiment analysis of chinese microblog text based on extended sentiment dictionary. *Future Generation Computer Systems*, 81, 395–403.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... others (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.