Universidad Autónoma de Querétaro
Facultad de Ingeniería

Análisis de exámenes del lenguaje para la detección temprana de Alzheimer: Un enfoque de Aprendizaje Automático.

Tesis

Que como parte de los requisitos para obtener el grado de

Maestro en Ciencias en Inteligencia Artificial

Presenta:
Brandon Alejandro Llaca Sánchez

Dirigido por:
Dr. Saúl Tovar Arriaga

Santiago de Querétaro, Querétaro, México, 2025

Universidad Autónoma de Querétaro
Facultad de Ingeniería
Maestría en Ciencias en Inteligencia Artificial

Análisis de exámenes del lenguaje para la detección temprana de Alzheimer: Un enfoque de Aprendizaje Automático.

Tesis

Que como parte de los requisitos para obtener el grado de

Maestro en Ciencias en Inteligencia Artificial

Presenta
Brandon Alejandro Llaca Sánchez

Dirigido por:
Dr. Saúl Tovar Arriaga

Dr. Saúl Tovar Arriaga
Presidente

Med. Esp. Humberto Güendulain Arenas
Secretario

Dr. Sebastián Salazar Colores
Vocal

M. en C. Luis Roberto García Noguez
Suplente

Dr. Andras Takacs
Suplente

Santiago de Querétaro, Querétaro, México
Octubre 2025

A mi esposa, familia y amigos.
Gracias totales.

Agradecimientos

En primer lugar, me gustaría agradecer a mi esposa Daniela y a mi madre todo el apoyo recibido durante el desarrollo de este proyecto. Sin duda, ellas fueron parte fundamental para poder culminarlo de buena manera. Quiero agradecer también a mi asesor, el Dr. Saúl Tovar Arriaga por su guía precisa y apoyo durante el desarrollo de este trabajo. De igual forma, deseo expresar mi agradecimiento al M. en C. Luis Roberto García Noguez, por su invaluable acompañamiento en el camino y por las charlas de discusión en torno a los modelos del lenguaje y a la detección de la demencia, que sin duda enriquecieron esta investigación.

Asimismo, quiero expresar mi gratitud al Dr. Sebastián Salazar Colores, por facilitar generosamente el acceso a los recursos de cómputo de alto rendimiento utilizados para generar los *embeddings* del modelo Linq-Embed-Mistral. Agradezco también a la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI, anteriormente CONAHCYT) por la beca recibida durante el desarrollo de este proyecto. Es preciso mencionar también que este trabajo fue apoyado por las subvenciones AG03705 y AG05133 del *National Institute on Aging* (NIA), en concordancia con el reconocimiento requerido por el uso de la base de datos Pitt Corpus.

Por último, quiero darle las gracias a Jack y a Johnson, por siempre estar ahí cuando los necesité y permitirme sentir tranquilidad aun en momentos tormentuosos. Quizás también sería preciso cerrar esta sección agradeciéndome a mí, por haberme aventurado en este nuevo horizonte y permitirme disfrutar de nuevos saberes.

Resumen

La demencia es una enfermedad neurodegenerativa que afecta las funciones cognitivas de las personas, deteriorando progresivamente su calidad de vida. Entre sus distintas formas, el Alzheimer se destaca como la más común. Dada su naturaleza incurable, la detección temprana es crucial para facilitar cuidados y una atención médica oportuna. Ante este escenario, la aplicación de técnicas de aprendizaje profundo en el análisis de exámenes del lenguaje se presenta como una solución eficaz, especialmente ante la laboriosa y falible evaluación manual. En este trabajo se implementaron y compararon cinco enfoques automatizados de Procesamiento del Lenguaje Natural (PLN) para identificar indicios de Alzheimer a partir de transcripciones de audio de la prueba del robo de la galleta, en la base de datos Pitt Corpus. Se evaluaron cuatro enfoques basados en *embeddings* de modelos grandes del lenguaje (GloVe, BERT, Gemma-2B y Linq-Embed-Mistral), así como una representación clásica estadística Tf–Idf, cada uno integrado con un clasificador final de regresión logística. Para su comparación, se realizó una validación cruzada estratificada 5-fold, obteniéndose los resultados más destacados con los *embeddings* de BERT (84.73 % de exactitud), seguidos de cerca por el enfoque clásico Tf–Idf (83.73 % de exactitud), y el modelo de última generación Linq-Embed-Mistral (83.54 % de exactitud). Contrario a las expectativas iniciales, estos hallazgos sugieren que la elección y frecuencia de las palabras podrían ser tan o más determinantes que la información semántica o contextual en la detección del Alzheimer. Ahora bien, la falta de una base de datos accesible y en español de registros médicos de pacientes con esta condición crea la necesidad urgente de construir una, contribuyendo así a la investigación de esta neuropatía en México. En conjunto, este estudio aborda la importancia de mejorar la detección temprana del Alzheimer, particularmente en personas hispanohablantes, buscando utilizar inteligencia artificial para aumentar la eficiencia de los métodos actuales y avanzar hacia un software fácil de usar capaz de ofrecer un primer indicador de riesgo de la enfermedad, reduciendo así la necesidad inicial de una consulta médica presencial.

Abstract

Dementia is a neurodegenerative disease that affects individuals' cognitive functions, progressively deteriorating their quality of life. Among its various forms, Alzheimer's disease stands out as the most common. Given its incurable nature, early detection is crucial to facilitate care and timely medical attention. In this context, the application of deep learning techniques to the analysis of language assessments emerges as an effective solution, especially in contrast to the laborious and fallible manual evaluation. In this work, five automated Natural Language Processing (NLP) approaches were implemented and compared to identify Alzheimer's indicators based on audio transcriptions from the Cookie Theft picture description task, using the Pitt Corpus dementia database. Four approaches based on embeddings from large language models (GloVe, BERT, Gemma-2B, and Linq-Embed-Mistral) were evaluated, along with a classical statistical representation, Tf–Idf, each integrated with a final logistic regression classifier. For their comparison, a 5-fold stratified cross-validation was performed, with the most notable results obtained using BERT embeddings (84.73 % accuracy), followed closely by the classical Tf–Idf approach (83.73 % accuracy) and the state-of-the-art Linq-Embed-Mistral model (83.54 % accuracy). Contrary to initial expectations, these findings suggest that word choice and frequency might be as or even more decisive than semantic or contextual information in detecting Alzheimer's. Now then, the lack of an accessible Spanish-language database containing medical records of patients with this condition highlights the urgent need to build one, thereby contributing to the study of this neuropathy in Mexico. Taken together, this study addresses the importance of improving early detection of Alzheimer's, particularly among Spanish-speaking individuals, aiming to leverage artificial intelligence to increase the efficiency of current methods and move toward user-friendly software capable of offering an initial risk indicator for the disease, thus potentially mitigating the initial need for face-to-face medical consultations.

Índice general

Lista de figuras

Lista de tablas

Capítulo 1

Introducción

La demencia es una enfermedad neurodegenerativa que se caracteriza por una pérdida progresiva de las funciones cognitivas. Es actualmente la séptima causa de muerte a nivel mundial entre todas las enfermedades, y constituye una de las principales razones de discapacidad y dependencia entre las personas [1]. Según datos de la Organización Mundial de la Salud (OMS), se proyecta que más de dos mil millones de personas tendrán 60 años o más para el año 2050. Dado que este grupo de edad es el más afectado por la neuropatía, esto subraya la importancia de desarrollar técnicas para detectar y tratar la demencia de manera precisa y temprana.

En este contexto, entre los diferentes tipos de demencia, la Enfermedad del Alzheimer (EA) se destaca como la más común, representando aproximadamente del 50 % al 75 % de los casos [2]. Este tipo de demencia se caracteriza por una pérdida o alteración de la memoria, acompañada de un deterioro de las habilidades visoespaciales, razonamiento, capacidades ejecutivas y en general de las funciones cognitivas. Se cree que la EA puede empezar a desarrollarse en el cerebro desde 20 años antes de que los síntomas aparezcan [3], sin embargo, en la mayoría de los casos, estos suelen comenzar a manifestarse alrededor de los 60 años, momento en el que el deterioro se vuelve evidente y se acelera [2]. El curso progresivo de la EA conlleva, en sus últimas etapas, a un estado de dependencia total para la persona que la padece y, finalmente, a la muerte.

La conexión entre la demencia, específicamente la EA, y la Inteligencia Artificial (IA) se revela como un área crucial de investigación. La IA ha constituido un desarrollo reciente en la historia humana, surgiendo en los años cincuenta con la idea de utilizar computado-

ras para simular el pensamiento inteligente [4, 5]. Desde sus inicios, el objetivo fue crear máquinas capaces de imitar el comportamiento humano y, eventualmente, llegar a resolver problemas complejos que pudieran ser automatizados. A lo largo de las décadas, la IA ha experimentado un progreso significativo, especialmente en los últimos años, destacándose la rama del aprendizaje automático (*Machine Learning* (ML)), y dentro de ésta, la subrama del aprendizaje profundo (*Deep Learning* (DL)), que se caracteriza por utilizar redes neuronales artificiales para imitar el funcionamiento del cerebro humano [6]. Este enfoque ha demostrado ser efectivo en la reproducción de procedimientos complejos [7, 8].

Con el aumento de la capacidad de cómputo, la Inteligencia Artificial y el *Machine Learning* han sido aplicados para abordar una variedad cada vez más amplia de problemas, incluyendo aquellos de índole biológica, médica y de salud mental, como el plegamiento de proteínas, la detección de convulsiones y la evaluación de la salud mental en niños mediante el uso de redes neuronales profundas [9–11]. En este sentido, la IA y el ML han sido utilizados también para abordar el problema de la detección temprana de la demencia y el Alzheimer, como se verá a profundidad en el siguiente capítulo.

En el marco de lo expuesto, en este trabajo se buscó explorar y comparar el desempeño de diferentes técnicas para generar representaciones vectoriales (*embeddings*) a partir de transcripciones de audio, enfocadas a la detección temprana de la demencia. El objetivo fue verificar qué metodología, entre las comparadas, ofrece un mejor rendimiento para esta tarea en concreto y, de esta forma, contribuir a la búsqueda de alternativas eficientes para desarrollar una técnica capaz de detectar de manera temprana indicios de demencia, asegurando además su practicidad y escalabilidad (idealmente, integrable en un software o aplicación móvil). De este modo, se podría prescindir, en la medida de lo posible, de la necesidad de acudir a una consulta médica presencial para una primera valoración de la condición, lo cual resultaría especialmente útil en contextos rurales o en donde la atención médica especializada sigue siendo sumamente escasa.

Así pues, la necesidad previa planteó la pregunta de si los *embeddings* más complejos y semánticamente ricos —como los derivados de modelos basados en *Transformers* (e.g., BERT, *Bidirectional Encoder Representations from Transformers* [12]; Gemma, un modelo *decoder* introducido recientemente por Google DeepMind [13]; y el modelo de última generación Linq-Embed-Mistral, optimizado específicamente para tareas de *embedding* [14])— ofrecen efectivamente una ventaja significativa frente a enfoques estadísticos más sim-

ples (como Tf–Idf) o a representaciones no contextuales —como los de *Global Vectors for Word Representation* (GloVe) [15]— en este contexto de diagnóstico, considerando tanto el desempeño predictivo como la practicidad de la implementación.

1.1. Descripción del problema

El análisis manual de exámenes del lenguaje en pacientes con demencia es laborioso y presenta limitaciones y desafíos. Estos exámenes pueden ser extensos, además de complejos de aplicar en sitio en algunas ocasiones, lo que dificulta su ejecución e interpretación precisa y oportuna [16]. La falta de una metodología automatizada y eficiente complica una detección temprana del degeneramiento de las funciones cognitivas de los pacientes, lo cual puede retrasar una intervención médica adecuada y afectar negativamente la calidad de vida de los mismos.

Los enfoques tradicionales basados en métodos de análisis estadístico o reglas predefinidas por parte de los especialistas, tienen limitaciones en términos de su capacidad para capturar patrones sutiles y variaciones en el lenguaje, lo que puede conducir a una detección no óptima del proceso de degeneración de las funciones cognitivas.

En este contexto, la aplicación de técnicas de aprendizaje profundo ofrece una alternativa interesante para el análisis automatizado de exámenes del lenguaje en pacientes con neuropatías cognitivo-degenerativas. La falta de eficiencia en la identificación de marcadores indicativos de Alzheimer y características clave en el lenguaje de los pacientes constituye un desafío significativo, obstaculizando la aplicación de intervenciones tempranas y tratamientos efectivos. Esta situación subraya la importancia de explorar técnicas de IA como medio para abordar esta problemática.

A su vez, la falta de una base de datos accesible y en español de México de registros de datos clínicos, audio y texto de exámenes del lenguaje de pacientes con Alzheimer (en la medida del conocimiento actual de los autores), hace que la tarea de aplicar dichas técnicas adaptadas al contexto hispanohablante, y más concretamente al mexicano, resulte imposible.

Este proyecto plantea realizar el análisis necesario para construir un repositorio con

estas características, además de desarrollar un modelo de aprendizaje profundo basado en *embeddings* de oraciones que permita identificar con una buena exactitud indicadores de Alzheimer, agilizando la intervención clínica y mejorando la calidad de vida de quienes padecen esta enfermedad.

1.2. Justificación

Proyecciones estadísticas sugieren que alrededor de 152 millones de personas podrían padecer demencia para el año 2050 [17]. Dado que el Alzheimer es la forma más prevalente de esta neuropatía [2], y considerando su naturaleza incurable, la estrategia más eficaz radica en su detección en las etapas más tempranas posibles para facilitar el inicio de los cuidados necesarios y una atención médica más oportuna. En este contexto, la identificación temprana del Alzheimer resulta fundamental para proporcionar tratamientos efectivos y mejorar la calidad de vida de los individuos. Sin embargo, la evaluación manual por parte de los especialistas puede resultar laboriosa y en ocasiones inviable [16]. Desde esta perspectiva, la aplicación de técnicas de aprendizaje profundo para analizar los resultados de los exámenes del lenguaje realizados a estos pacientes, se presenta como una solución natural y útil.

Asimismo, la falta de una base de datos accesible y en español, de registros de conversaciones y exámenes del lenguaje y de las funciones cognitivas de pacientes con EA, hacen que la tarea de crear una con dichas características se convierta en algo imperativo en el momento actual. Surge una necesidad inmediata de construirla, y que de esta manera pueda ser utilizada como una referencia para futuros estudios en el tema en el idioma español, contribuyendo así al campo de la investigación de esta neuropatía en México.

Por otro lado, automatizar el análisis de exámenes del lenguaje mediante técnicas de *Machine Learning* permitiría extraer de manera eficiente y confiable información relevante, identificar patrones sutiles y marcadores indicativos de demencia en el lenguaje utilizado por los pacientes, así como enriquecer la interpretación de los resultados.

Así, esta investigación justifica su relevancia al abordar la necesidad crítica de mejorar la detección temprana del Alzheimer. La IA ofrece la posibilidad de agilizar la logística

y aumentar la eficiencia de los métodos actuales, liberando a los profesionales de tareas repetitivas y permitiéndoles centrarse directamente en la atención a los pacientes.

A largo plazo, se proyecta que el presente estudio sirva como antecedente para futuras investigaciones en donde se analicen también las imágenes de fondo de ojo de las personas a las cuales se les realicen este tipo de exámenes del lenguaje; esto con el objetivo de buscar una segunda serie de indicadores de EA o demencia en la vasculización de las arterias de la retina.

Se prevé que esta investigación evolucione hacia un único estudio *ad hoc*, llevado a cabo con un dispositivo compacto o una aplicación móvil, que integre exámenes del lenguaje y un análisis de imagen de fondo de ojo automatizados que puedan ayudar a dar indicadores rápidos de la presencia de EA.

1.3. Hipótesis

Un modelo de aprendizaje profundo para el análisis de exámenes del lenguaje de pacientes con Alzheimer, basado en *embeddings* de oraciones, permite discriminar correctamente a los pacientes sanos de los pacientes enfermos.*

*Para efectos del estudio, se entiende por "discriminar" a una clasificación correcta de los pacientes en al menos un 75 % de los casos.

1.4. Objetivos

1.4.1. Objetivo general

Construir un modelo de aprendizaje profundo basado en *embeddings* de oraciones que analice exámenes del lenguaje, para la detección temprana de Alzheimer.

1.4.2. Objetivos específicos

- Llevar a cabo el análisis para la creación de una base de datos de audio y de texto en el idioma español, de exámenes del lenguaje para el entrenamiento de modelos de

Inteligencia Artificial enfocados a la detección temprana de Alzheimer.

- Desarrollar un modelo para la detección de Alzheimer, mediante la selección y prueba de técnicas de IA encontradas en la literatura.
- Validar y evaluar la eficiencia del modelo para comprobar su utilidad.

Capítulo 2

Antecedentes

2.1. Estado del arte

Recientemente, se han propuesto y aplicado múltiples técnicas de ML para el diagnóstico automatizado de pacientes con Alzheimer y/o demencia, utilizando diferentes enfoques para desarrollar los modelos de acuerdo al tipo de datos recabados para hacer la detección, los cuales pueden abarcar desde imágenes cerebrales, datos clínicos, y/o audio o texto de estrevistas y exámenes del lenguaje [18].

En relación con los modelos de ML entrenados con imágenes de resonancia magnética (MRI) del cerebro para detectar estos padecimientos, en Salvatore et al. [19] en el 2015, emplearon una Máquina de Soporte Vectorial (SVM) para identificar marcadores en la progresión temprana del Alzheimer. Su enfoque incluyó la identificación y clasificación de regiones cerebrales críticas (como el hipocampo, la corteza entorrinal, los ganglios basales y el cerebelo) en pacientes con EA y deterioro cognitivo leve, logrando una exactitud en validación cruzada *k-fold* del 76 %.

Asimismo, en Bidani et al. [20] en el año 2019, se propuso un enfoque novedoso basado en una Red Neuronal Convolucional Profunda (DCNN) y *Transfer Learning* para detectar demencia en el conjunto de datos OASIS (una colección longitudinal de MRI's). Estas imágenes fueron preprocesadas utilizando la técnica de *Bag of Features* y métodos de clasificación para identificar la demencia en sus diferentes etapas. El modelo DCNN logró una exactitud significativa del 81.94 % reconociendo la presencia del padecimiento. Luego, en Basheer et al. [21] en 2021, propusieron un modelo de Red Neuronal Convolucional

(CNN) modificado con cambios en la estructura de la red en cápsulas, para clasificación de las imágenes del conjunto de datos OASIS en dos grupos, con demencia y sin demencia, logrando una exactitud del 92.39 %. En este estudio se avanzó en la identificación de características importantes en un modelo de CNN.

Referente a modelos de aprendizaje automático entrenados a partir de datos clínicos, se tiene por ejemplo a Alam et al. [22] en 2016, en donde se buscó mejorar la evaluación automática de la demencia mediante el uso de sensores fisiológicos y ambientales, utilizando datos de sensores portátiles (*Electrodermal Activity* (EDA), *Photoplethysmogram* (PPG), acelerómetro (ACC), entre otros) y un modelo de *Random Forest* (RF). Se correlacionaron deficiencias cognitivas con el deterioro de la salud funcional en 17 adultos mayores en una comunidad de retiro en Baltimore, validando los resultados con puntuaciones clínicas observadas. El estudio logró una exactitud en validación cruzada *k-fold* del 91.5 %.

En esta misma línea, Chiu et al. [23] en 2019 plantearon un modelo de selección de características basado en la ganancia de información, con el objetivo de desarrollar un cuestionario de preguntas clave destinado a auxiliar a neurólogos y neuropsicólogos en la detección de deterioro cognitivo leve (DCL) y demencia. El modelo (que resultó de 12 preguntas) demostró una alta exactitud (área bajo la curva ROC, AUC, entre 0.94 y 0.97) para discriminar entre cognición normal, deterioro cognitivo leve y demencia. Por su parte, en Hsiu et al. [24] en 2022 se utilizó un modelo de *K-Nearest Neighbors* (KNN) en índices espectrales de la forma de onda del pulso arterial, para identificar tempranamente el deterioro cognitivo. Se encontraron diferencias significativas entre pacientes con EA y controles, logrando una exactitud del 70.32 % mediante validación cruzada *3-fold*, y se estableció una correlación (con coeficiente de determinación $R^2 = 0.36$) entre la probabilidad de padecer la enfermedad según el modelo y la puntuación cognitiva en el *Mini-Mental-State-Examination* (MMSE) de los pacientes.

Por otro lado, en cuanto a modelos de aprendizaje automático entrenados a partir de audio o texto de estrevistas y exámenes del lenguaje, se puede mencionar a Sadeghian et al. [25] en el 2017, donde se propone un sistema de análisis automático del habla para diagnosticar el Alzheimer temprano. Utilizando características acústicas y lingüísticas extraídas de 72 grabaciones en donde los sujetos de prueba describían una imagen, se logró una discriminación efectiva entre pacientes con Alzheimer y controles del 91.7 %, obtenida a través de una red neuronal perceptrón multicapa, que fue usada como clasificador binario.

Adicionalmente, en 2022 Ilias y Askounis [26] abordaron la falta de interpretabilidad en modelos basados en transformadores para diagnosticar EA. Se emplearon diversos modelos basados en este tipo de redes en la base de datos *The ADReSS dataset*, consistente en 156 muestras de habla y transcripciones asociadas de sujetos angloparlantes, divididos en dos grupos: pacientes con EA (N = 78) y pacientes sin EA (N = 78), siendo el modelo BERT el modelo que tuvo la mayor exactitud con un 87.50 %. En este sentido, se presentan también métodos para identificar patrones lingüísticos en pacientes con EA y sin ella, revelando diferencias significativas mediante análisis detallados del lenguaje y técnicas de explicabilidad como LIME.

Existen también metodologías para la detección temprana de demencia empleando *embeddings* de oraciones generados por redes *Siamese BERT*. En Santander-Cruz et al. [27] también en 2022, se implementaron diferentes modelos de ML (SVM, KNN, RF y NN) entrenados a partir de 17 características demográficas, léxicas, sintácticas y semánticas, que fueron extraídas de 550 muestras de producción oral de controles ancianos y pacientes con EA de la base de datos *DementiaBank Pitt Corpus Database* [28]. La relevancia de estas características se evaluó mediante el puntaje de información mutua, evidenciando su asociación con el puntaje MMSE. Los resultados señalan un rendimiento superior (exactitud del 77 %) en comparación con enfoques basados en sintaxis y BERT al utilizar exclusivamente características lingüísticas (exactitud del 74 %). En particular, las combinaciones SBERT+SVM y SBERT+NN demostraron ser las que se desempeñaron mejor.

2.2. Marco teórico

2.2.1. Demencia

La demencia es una neuropatía que se caracteriza por la disminución de las habilidades cognitivas de una persona. Este deterioro afecta primordialmente la memoria, aunque también puede involucrar el lenguaje, atención, orientación, juicio y planificación. Su origen puede estar en patologías o lesiones cerebrales, así como en condiciones médicas que afectan otras partes del cuerpo [29]. Se estima que cada tres segundos se diagnostica un caso de demencia en el mundo, con 50 millones de casos en 2018 y una proyección de 152 millones para 2050 [17, 30].

Los síntomas de la demencia incluyen principalmente problemas de memoria a corto plazo y dificultades para realizar tareas diarias, aunque también pueden llegar a abarcar retirada de actividades sociales, ansiedad y depresión. El diagnóstico implica evaluar el historial médico, la declinación cognitiva y los problemas funcionales [16]. Si bien existen medicamentos que pueden ayudar con las habilidades cognitivas y el estado de ánimo, no detienen la progresión de la enfermedad.

El tratamiento varía según la situación individual, y se recomienda un enfoque centrado en el paciente y la familia. Además, es importante tomar medidas preventivas como controlar factores de riesgo, tales como presión arterial alta, realizar actividades mentales, físicas y sociales, mantener una dieta equilibrada, beber alcohol con moderación y garantizar un sueño de calidad. Los médicos también pueden brindar educación a la familia y cuidadores, así como ayudar en la planificación a largo plazo para el cuidado de la persona [29].

La Figura 1 ofrece una visión integral sobre la demencia, abordando su clasificación, síntomas, causas, tratamiento y prevención.

Alzheimer

La enfermedad del Alzheimer, responsable del 50-75 % de los casos de demencia, provoca daño celular en el cerebro y una disminución en la memoria y funciones cognitivas [2]. Su progresión abarca desde un deterioro cognitivo leve, hasta etapas graves que implican un estado de dependencia total para la persona, llegando finalmente a la muerte [3]. La EA inicia con síntomas tenues, y puede avanzar hasta afectar las habilidades físicas y la conciencia.

A pesar de que los factores de riesgo incluyen la edad, genética y entorno, la causa exacta de la EA sigue siendo desconocida, y no existe cura ni manera de detener su progresión [30].

Figura 1: Visión integral sobre la demencia. (Imagen propia).

2.2.2. Inteligencia Artificial

La Inteligencia Artificial (IA) es una disciplina que busca desarrollar sistemas capaces de realizar tareas que requieren inteligencia humana. A lo largo de su evolución, ha experimentado avances desde sus enfoques iniciales en lógica formal y razonamiento simbólico, hasta la llegada del aprendizaje automático y, más recientemente, del aprendizaje profundo [5].

A mediados del siglo XX, Alan Turing planteó la idea de una "máquina universal" capaz de imitar el comportamiento humano, y propuso el famoso "Test de Turing" para determinar si una máquina podía ser considerada inteligente. En décadas siguientes, surgieron enfoques fundamentados en conocimiento experto y representación simbólica, como los sistemas basados en reglas y los sistemas expertos [4].

La Figura 2 muestra un diagrama que ilustra diversas ramas y paradigmas de la IA en la actualidad, algunos de las cuales se indagarán a continuación.

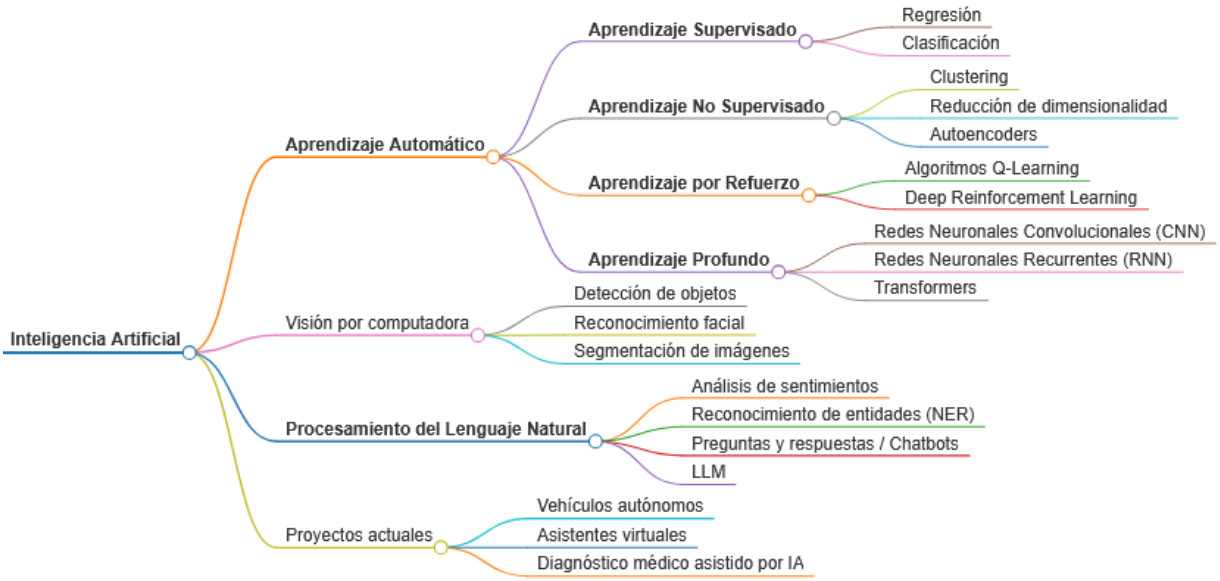


Figura 2: Algunas ramas y paradigmas de la IA en la actualidad. (Imagen propia).

2.2.3. Aprendizaje Automático

El Aprendizaje Automático (ML) es una rama de la Inteligencia Artificial que se caracteriza por permitir a los sistemas aprender sin necesidad de una programación explícita para este fin [31]. Este campo tuvo su florecimiento a finales del siglo XX, derivado del aumento en el poder de cómputo y de las investigaciones en nuevos modelos de IA, empezando una revolución en el área al posibilitar que las máquinas aprendieran patrones a través de algoritmos y modelos estadísticos, sin requerir programación escrita con este propósito.

Recientemente, el Aprendizaje Profundo (DL), una rama del aprendizaje automático basada en redes neuronales artificiales profundas, ha permitido explotar el potencial de la IA en diversas áreas, destacando el reconocimiento de voz, la visión por computadora y el procesamiento del lenguaje natural [6, 32].

Dentro del Aprendizaje Automático existen cuatro enfoques fundamentales: Supervisado, No Supervisado, por Refuerzo y Profundo, y cada uno de ellos engloba diversos algoritmos. A continuación se presenta una breve explicación de las características primor-

diales de cada enfoque, así como de los modelos principales dentro de cada categoría [33]:

1. **Aprendizaje Supervisado:** Este paradigma del ML se caracteriza por algoritmos que aprenden a predecir (regresión) o clasificar (clasificación) un valor de acuerdo a ejemplos que el usuario le proporciona, de ahí que se trate de aprendizaje "supervisado". El algoritmo aprende patrones que le ayudan a predecir o clasificar de manera correcta los datos, mediante métodos estadísticos y reglas que aprende de los ejemplos brindados; el proceso de aprendizaje está basado en un entrenamiento que el algoritmo realiza con los datos de entrada y sus respectivas etiquetas o salidas, dados por el usuario.

    - Árbol de Decisión (DT): Un árbol de decisión es un grafo compuesto de nodos, que representan sucesos o eventos sobre los cuales se está trabajando (prediciendo), y aristas, que representan las posibles decisiones u opciones que se tienen para cada uno de esos eventos. Existe también el Bosque Aleatorio (*Random Forest* (RF)), que corresponde a un modelo de varios árboles de decisión utilizados en conjunto para realizar predicciones.
    - Máquina de Soporte Vectorial (SVM): Una Máquina de Soporte Vectorial es un algoritmo de ML que funciona realizando una regresión o clasificación lineal (según corresponda al tipo de problema) en varias dimensiones. En problemas de clasificación, funciona estableciendo una división (en forma de hiperplano) entre las diferentes clases o grupos a clasificar, buscando que la distancia entre las diferentes clases y la división sea máxima, y como consecuencia, minimizando el error.
    - Redes Neuronales (NN): Una red neuronal es un algoritmo de aprendizaje automático que se caracteriza por imitar la forma en que funciona el cerebro humano. En este sentido, se trata de una malla interconectada de neuronas, cada una con un peso (un número) asociado, divididas en varias capas, generalmente siendo la primera llamada capa de entrada ("*input layer*"), la última llamada capa de salida ("*output layer*"), y las capas de en medio llamadas capas ocultas ("*hidden layers*").

      Los datos ingresan al modelo a través de la capa de entrada, son procesados por esta capa mediante operaciones matemáticas que involucran a los pesos, y los resultados ingresan después a las capas ocultas, en donde nuevamente son

procesados y posteriormente utilizados para alimentar a la capa de salida, que procesa los datos una última vez para realizar una predicción, ya sea de regresión o de clasificación. Estas predicciones se comparan con los datos reales, se calcula el error y éste se utiliza para realizar un ajuste de los pesos de las neuronas de manera retroactiva, para nuevamente realizar una predicción con los pesos ajustados. Este proceso se repite hasta que se alcanza un error establecido o bien después de repetirse un cierto número de veces. Cabe resaltar que este funcionamiento se refiere a las redes neuronales supervisadas (siendo estas las más comunes), pero también existen las redes neuronales no supervisadas y las redes neuronales que funcionan mediante aprendizaje por refuerzo.

- *K-Nearest Neighbors* (KNN): Este algoritmo de ML funciona asignando etiquetas a los datos según la mayoría de las etiquetas de sus K vecinos más cercanos, siendo K un número elegido por el usuario. Es un algoritmo simple y útil que puede ayudar a resolver problemas tanto de clasificación como de regresión, sin embargo, tiene la desventaja de que su implementación consume muchos recursos en cuanto la cantidad de datos y el valor de K incrementan.

2. **Aprendizaje No Supervisado:** A diferencia del paradigma anterior, este enfoque dentro del ML se distingue porque el usuario no sabe cuál es la respuesta "correcta" de antemano, por lo que no le proporciona ejemplos al algoritmo, sino más bien le brinda los datos en sí, y el algoritmo es el encargado de encontrar los patrones subyacentes en los mismos. Los algoritmos de aprendizaje no supervisado son útiles para encontrar características que relacionan a los datos, así como para reducir la dimensionalidad de los mismos.

   - *Principal Component Analysis* (PCA): En el algoritmo de PCA, se realiza una transformación matemática de los datos para reducir la dimensión de los mismos. Es comúnmente utilizado con este propósito, como un método para lidiar con la covarianza de variables en los datos, y tratar con ésta como una combinación lineal de otras variables linealmente independientes, llamadas "componentes principales".
   - *K-Means Clustering*: Este algoritmo consiste en agrupar los datos disponibles en *K clusters*, haciendo una partición de los mismos. El usuario decide dónde colocar inicialmente los K centros de los *clusters*, se realiza una clasificación de los datos calculando su distancia respecto a los K centros y agrupando cada

uno con el centro cuya distancia sea menor al dato. Una vez terminado este procedimiento, se vuelven a calcular K centros para los *K clusters*, calculando el gravicentro de los puntos dentro de cada *cluster*. Se vuelve a realizar el proceso de clasificación, y se repite el funcionamiento hasta alcanzar un error deseado. Dado el proceso iterativo a partir de los centros, la elección inicial de los K centros por parte del usuario influye en gran manera en la clasificación final.

3. **Aprendizaje por Refuerzo:** Se trata de una rama dentro del ML en la cual se estudia cómo un agente debería comportarse en un ambiente establecido para alcanzar un objetivo dado, dadas una recompensa, la cual se le dará si el agente se acerca a su objetivo, y un castigo, el cual se le dará si el agente se aleja del mismo. La meta del agente siempre es entonces maximizar la recompensa y minimizar el castigo, basado en sus acciones dentro del ambiente.

4. **Aprendizaje Profundo:** Constituye una rama del ML fundamentada en arquitecturas de redes neuronales con múltiples capas, lo que les permite aprender representaciones estructuradas de los datos. Debido a su capacidad de modelar relaciones no lineales y a su arquitectura jerárquica, este enfoque es comúnmente utilizado para tareas complejas de procesamiento secuencial.

    - Redes Neuronales Convolucionales (CNN): Las CNN son una arquitectura de red neuronal que se caracteriza por el uso de capas convolucionales, lo cual les permite extraer características locales de los datos. Son especialmente utilizadas en el procesamiento de imágenes. Su principal ventaja radica en la capacidad para detectar patrones espaciales, como bordes y texturas, de manera jerárquica. Por consiguiente, estas redes han constituido la base de avances importantes en visión por computadora, tales como la clasificación de imágenes y la detección y segmentación de objetos.
    - Redes Neuronales Recurrentes (RNN): Las RNN están diseñadas para trabajar con datos secuenciales. Integran conexiones recurrentes que les permiten conservar información sobre estados previos; de esta forma, pueden modelar dependencias temporales de los datos, y por esta razón son populares en tareas como procesamiento del lenguaje y análisis de series temporales. Cabe notar, sin embargo, que estas redes cuentan con limitaciones en el aprendizaje de dependencias de largo plazo, debido al problema del desvanecimiento y explosión del gradiente, lo cual dio lugar a variantes como *Long Short-Term Memory*

(LSTM) [34] y *Gated Recurrent Unit* (GRU) [35], que lograron mitigar parcialmente dichas dificultades.

- *Transformers*: Los *Transformers* son una arquitectura de red neuronal basada en mecanismos de atención, que les permite capturar dependencias de largo alcance de manera más efectiva que las RNN [36]. Además, durante el entrenamiento y en la etapa de *encoding*, pueden procesar secuencias en paralelo. Dentro de los *Transformers*, se distinguen tres principales tipos de arquitecturas [37]:
  1) Los modelos *encoder* (por ejemplo, BERT [12]), que reciben tokens como entrada y producen una representación vectorial contextual (una codificación) para cada token como salida, y se utilizan principalmente para tareas de comprensión del lenguaje, como clasificación de textos o reconocimiento de entidades.
  2) Los modelos *decoder* (por ejemplo, GPT [38]), que toman tokens como entrada y generan tokens como salida de manera autorregresiva token a token, y se emplean principalmente en tareas de generación de lenguaje, como chatbots.
  3) Los modelos *encoder-decoder* (por ejemplo, BART [39]), que reciben tokens como entrada y producen una serie de tokens como salida, y son utilizados principalmente para respuestas a preguntas con contexto, traducción automática y resumen de textos. En esta arquitectura, cada capa del *decoder* usa atención cruzada sobre las representaciones de la última capa oculta del *encoder* [39], para después generar texto token a token de manera autorregresiva [36].
  En la figura 3, se puede consultar una comparativa visual del funcionamiento de las 3 arquitecturas descritas. Los *Transformers* son la base de los grandes modelos del lenguaje (LLM), núcleo de múltiples aplicaciones de IA de última generación.

2.2.4. Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural (PLN, más comúnmente reconocido en la literatura como NLP, por sus siglas en inglés) es una rama de la inteligencia artificial que utiliza técnicas computacionales para estudiar el lenguaje humano. Desde sus inicios en la lingüística informática en la década de 1950, el PLN ha evolucionado de reglas gramaticales y sistemas simbólicos, hacia enfoques estadísticos y, más recientemente, hacia modelos

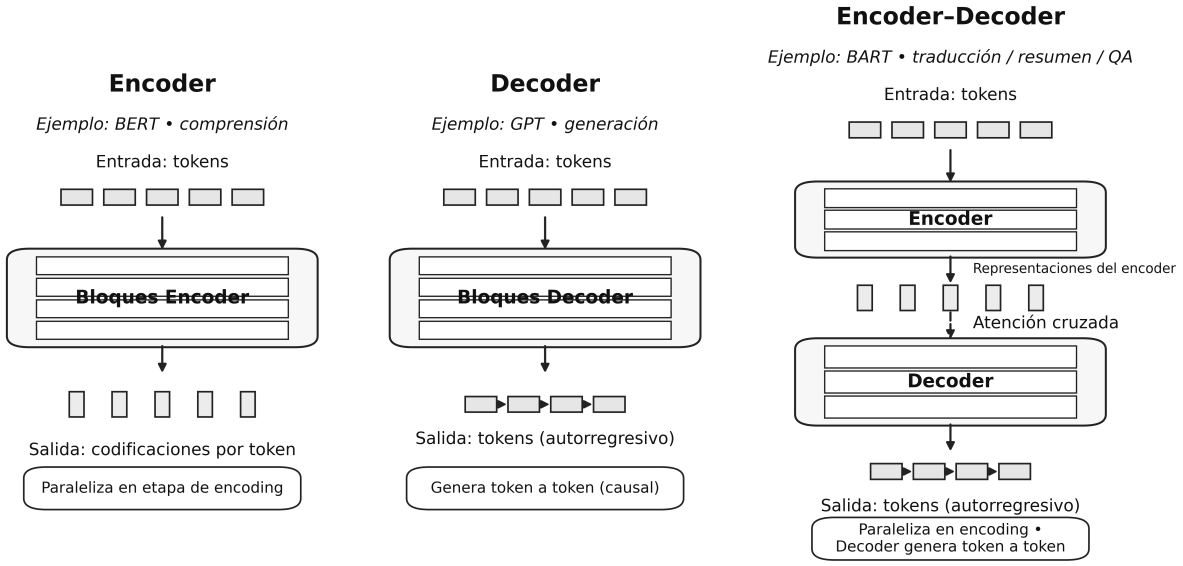


Figura 3: Comparativa visual de los 3 tipos de arquitectura *Transformer*. (Imagen propia).

de aprendizaje profundo [37]. Sus fundamentos incluyen el preprocesamiento de textos, la limpieza y normalización del conjunto de datos (del corpus), la tokenización por palabras y subpalabras, así como la representación numérica de texto mediante vectores. Estos pasos constituyen la base para que los algoritmos puedan analizar y extraer información relevante a partir del lenguaje natural.

El PLN se ha aplicado ampliamente en contextos prácticos que requieren la interpretación eficaz de grandes volúmenes de información textual. Ejemplos de ello incluyen la clasificación de textos [40], la identificación de discurso de odio [41], el análisis de sentimientos [42], la detección de *phishing* [43] y la detección de enfermedades mentales [44]. Asimismo, se han aplicado técnicas de PLN en múltiples idiomas y en una cantidad de dominios bastante variados, desde lingüística hasta ciencia, pasando por ciberseguridad y estudios de género [45].

- Aplicaciones clásicas: Entre las aplicaciones clásicas del PLN se pueden identificar la búsqueda semántica, la traducción automática, la síntesis de textos, la extracción de palabras clave, la generación de *embeddings* de texto, el análisis de sentimientos y el reconocimiento de entidades nombradas (NER) [46]. Profundizando en estas últimos dos, respecto al análisis de sentimientos, se trata de una técnica que busca identificar

la polaridad emocional de un texto (positivo, negativo o neutro), utilizando para ello primero técnicas de preprocesamiento y limpieza de datos, de codificación vectorial y clasificadores finales basados en técnicas como regresión logística o *Naive Bayes*. El segundo, NER, consiste en detectar y clasificar entidades como nombres de personas, organizaciones, lugares o fechas en grandes volúmenes de texto, teniendo gran importancia para ello las decisiones en el preprocesamiento del corpus, la selección del vocabulario y los sistemas de búsqueda inteligente.

- Aplicaciones contemporáneas: En la actualidad, las aplicaciones más avanzadas del PLN se concentran en los grandes modelos del lenguaje (LLM) y en los chatbots conversacionales. Estos modelos, fundamentados en arquitecturas *Transformer*, como BERT (*encoder*) o GPT (*decoder*), han demostrado un rendimiento excepcional en tareas de comprensión y generación de lenguaje [38, 40], posibilitándoles encontrar aplicaciones en campos interdisciplinarios como la educación [47], la salud [48] y el trabajo [49].

2.2.5. Aprendizaje Automático para detección de Alzheimer

El aprendizaje automático ha demostrado ser una herramienta útil en diversas aplicaciones médicas, incluida la detección de enfermedades relacionadas con el cerebro, como el autismo [11, 50, 51], la epilepsia [10], la depresión y la ansiedad [52, 53], y la demencia [25–27]. Para realizar esta última tarea, en particular, para diagnosticar la EA, se emplean tanto métodos supervisados como no supervisados, técnicas de visión por computadora, especialmente resonancia magnética (MRI) [30], así como datos clínicos y análisis de audio y texto de entrevistas y exámenes del lenguaje [18].

Capítulo 3

Metodología

Esta sección se divide en dos secciones principales. La primera corresponde a la descripción detallada del procedimiento seguido en el estudio comparativo de *embeddings* para la detección de demencia en la base de datos Pitt Corpus, mientras que en la segunda se describe la metodología seguida para llevar a cabo el análisis para la construcción de la base de datos en español.

3.1. Estudio comparativo entre diferentes métodos de *embeddings*.

3.1.1. Software

Este estudio se trabajó en el lenguaje de programación Python, versión 3.11.11, a través de la plataforma Google Colab. Asimismo, se utilizó un servidor dedicado de cómputo de alto rendimiento específicamente para la obtención de los *embeddings* generados con el modelo Linq-Embed-Mistral. El código utilizado para ejecutar los experimentos se encuentra disponible públicamente en GitHub (`https://github.com/Placanbero/dementia-embeddings`, a fecha de 3 de julio de 2025). Entre las librerías principales utilizadas destacan `torch`, `transformers`, `nltk` y `sklearn` para la descarga y manipulación de los *embeddings* de los modelos de lenguaje, además de la función `re` para la limpieza y preprocesamiento de los datos, y las funciones comúnmente utilizadas para el tratamiento matemático y gráfico en Python, tales como `numpy`, `pandas`, `matplotlib` y `seaborn`.

3.1.2. Metodología

La metodología empleada en este trabajo consistió en la siguiente serie de pasos, en los cuales se profundizará seguidamente. A continuación, se presenta un diagrama que resume detalladamente cada uno de ellos (Figura 4).

1. Preprocesamiento del conjunto de datos, la base de datos de demencia Pitt Corpus.
2. Generación de *embeddings* de modelos de lenguaje grande para las transcripciones utilizadas, y alimentación de un clasificador logístico con estas características, con el objetivo de diferenciar un resultado positivo de uno negativo en cuanto a la presencia de demencia en los datos.
3. Comparación del desempeño de estas técnicas entre sí y en contraste con la técnica Tf–Idf.
4. Análisis de los resultados obtenidos y deducción de conclusiones.

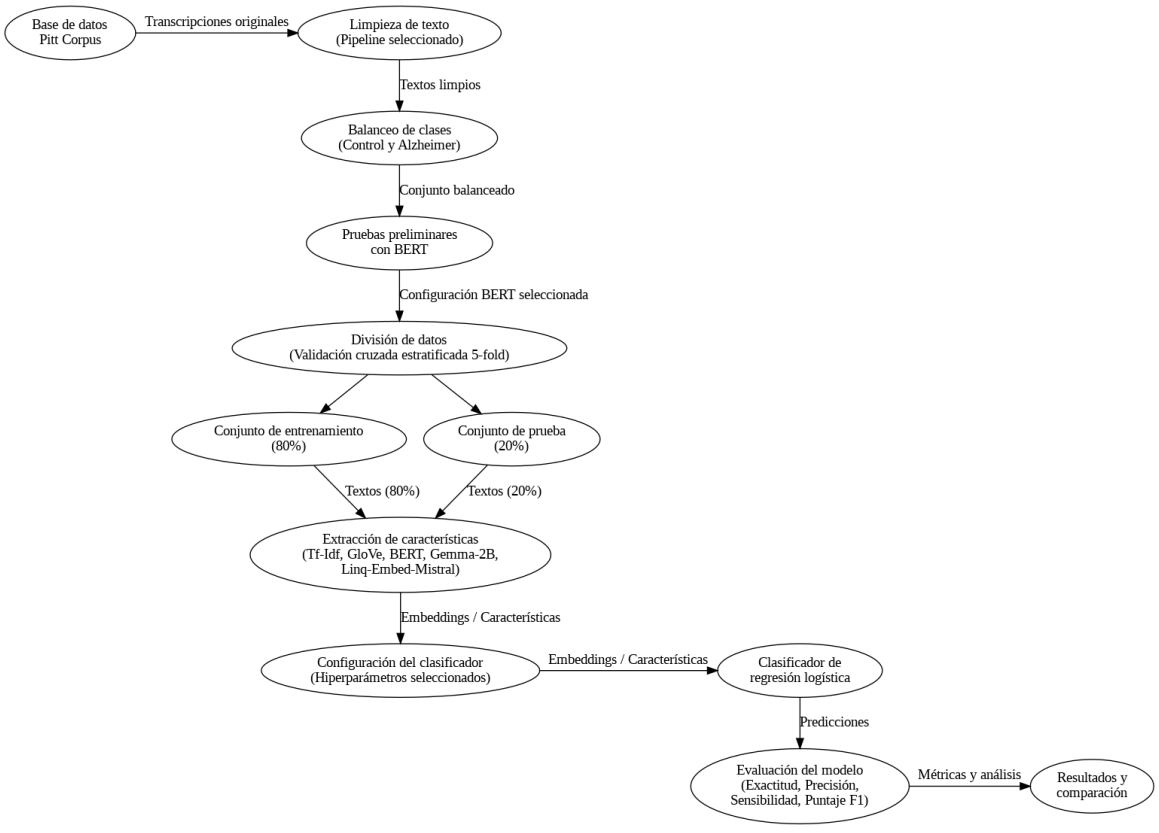


Figura 4: Diagrama detallado de la metodología utilizada en el estudio comparativo de *embeddings*.

3.1.3. Preparación del conjunto de datos.

Esta investigación se realizó con la base de datos de demencia Pitt Corpus, una base de datos en inglés de diferentes exámenes del lenguaje realizados a una serie de pacientes de control (sanos) y a pacientes diagnosticados con distintos tipos de demencia, construida por investigadores de la Universidad de Pittsburgh [54].

Este corpus es un compendio de producciones de lenguaje oral, transcripciones manuales de dichas producciones, datos demográficos y resultados de pruebas de funciones cognitivas (*Mini-Mental State Examination*), recopilados a lo largo de una serie de visitas clínicas sucesivas en las que participaron sujetos sanos de control (108 en total), pacientes con algún tipo de demencia diagnosticado (208 en conjunto) y pacientes con un diagnóstico no definido (un total de 85) [28].

Para efectos del presente estudio, se trabajó específicamente en las transcripciones de los audios de la prueba conocida como "El robo de la galleta" (*The Cookie Theft Picture*) [55] de la *Boston Diagnostic Aphasia Examination* [56], la cual se puede consultar en la Figura 5. En dicha prueba se le solicita al sujeto describir todo lo que puede observar en una imagen en blanco y negro que se le presenta, en la que se aprecia a dos niños robando galletas de un frasco, mientras su madre, de pie en el fregadero lavando trastes, parece no percatarse del agua que se desborda. Esta imagen contiene algunos elementos sobresalientes que una persona sana normalmente tiende a no pasar por alto.

La base de datos cuenta en total con 549 registros de visitas clínicas con sus respectivas transcripciones de la prueba del robo de la galleta, clasificados, según su diagnóstico, como sigue: 243 transcripciones de pacientes de control (*Control*), 234 de pacientes con diagnóstico de Alzheimer probable (*ProbableAD*), 42 de pacientes con diagnóstico de deterioro cognitivo leve (*MCI*), 21 de pacientes con diagnóstico de Alzheimer posible (*PossibleAD*), 5 de pacientes diagnosticados con demencia vascular (*Vascular*) y 3 de pacientes diagnosticados con algún problema de memoria (*Memory*), además de un registro de una persona con un diagnóstico no definido (*Other*).

Se realizó una discriminación de los datos, con la cual se buscó enfocar la detección en la presencia de Alzheimer, el tipo más común de demencia, además de contar con clases balanceadas. Debido a esto, se unificaron aquellos registros marcados como "*ProbableAD*"

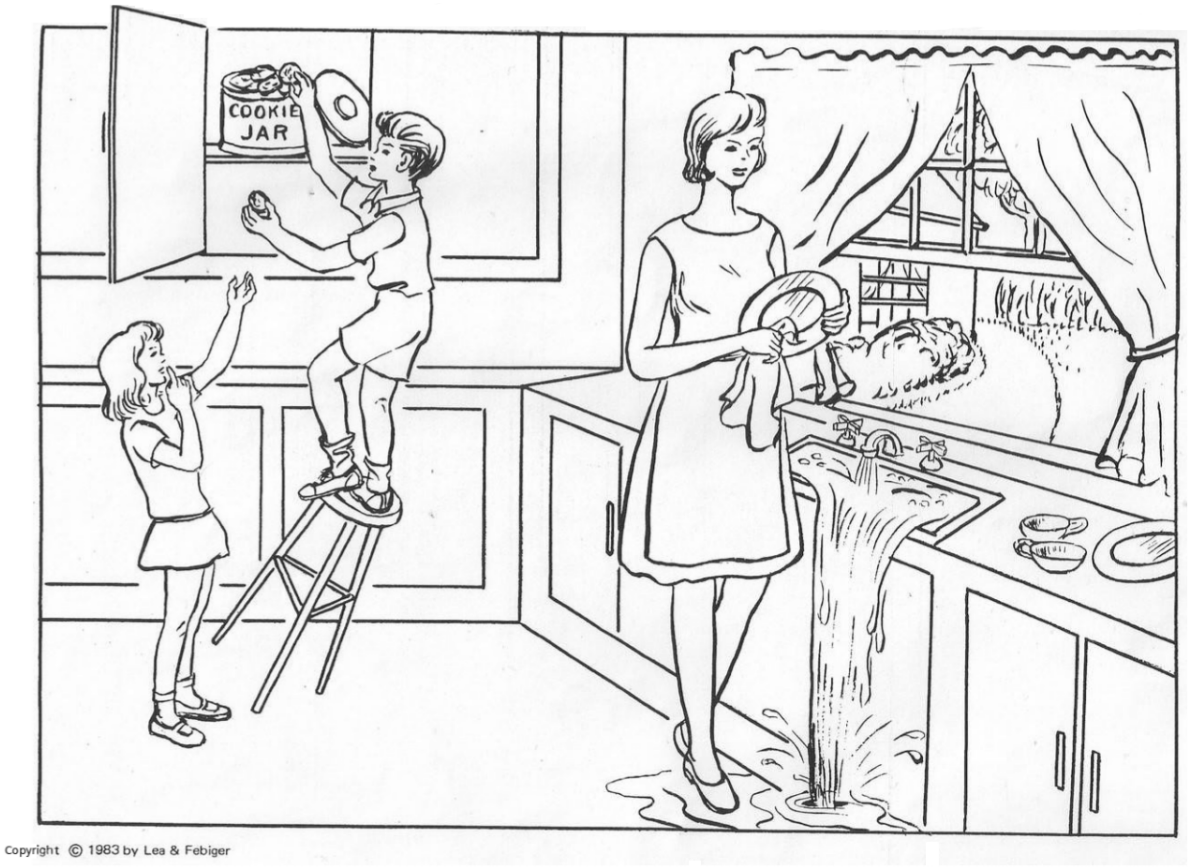


Figura 5: La imagen del robo de la galleta, de la *Boston Diagnostic Aphasia Examination* [55].

y "*PossibleAD*" en una sola categoría, y se utilizaron también los registros marcados como "Control" para aprender a diferenciar pacientes sanos de pacientes con alguna posible presencia de Alzheimer. Por tanto, el conteo final de transcripciones utilizadas resultó en dos clases balanceadas, una de 255 pacientes con Alzheimer (234 de *ProbableAD* y 21 de *PossibleAD*), y 243 de pacientes sanos (*Control*).

Este conjunto de transcripciones fue evaluado utilizando validación cruzada *5-fold*. En cada iteración, se utilizó el 80 % de los datos para entrenamiento y el 20 % restante para prueba. En promedio, cada conjunto de entrenamiento incluyó 204 transcripciones de pacientes con indicios de Alzheimer y 194 de pacientes de control, mientras que cada conjunto de prueba incluyó en promedio 51 registros de pacientes con indicios de Alzheimer y 49 de pacientes de control (Tabla 1). Es importante señalar que las divisiones de los conjuntos de entrenamiento y de prueba no fueron independientes del hablante. Esta decisión

fue intencional, ya que todos los métodos de *embeddings* fueron evaluados sobre las mismas particiones, asegurando una comparación justa. Además, múltiples transcripciones de un mismo participante fueron registradas en distintos momentos (incluso bajo diferentes diagnósticos), lo cual introduce una variabilidad útil en los datos en lugar de redundancia.

Partición del conjunto de datos	**Indicio de Alzheimer**	**Control**
Conjunto de entrenamiento (promedio por iteración)	204	194
Conjunto de prueba (promedio por iteración)	51	49

Tabla 1: Distribución promedio de etiquetas por iteración en la validación cruzada *5-fold*.

Adicionalmente, puesto que los archivos que contienen las transcripciones de las visitas clínicas se encuentran en el formato de transcripción CHAT (el cual incluye los metadatos de la misma, así como la transcripción del audio del investigador que realizó la entrevista, además de símbolos de anotaciones realizadas por los investigadores respecto al habla del paciente, tales como pausas, autocorrecciones, errores de pronunciación, etc.), un exhaustivo proceso de limpieza y preprocesamiento de las transcripciones fue llevado a cabo antes de utilizar estas transcripciones en los modelos. Tras este proceso de limpieza, se trabajó con las transcripciones completas y no únicamente con fragmentos de ellas.

En primer lugar, se eliminaron tanto los metadatos como la transcripción de lo dicho por el investigador, de tal forma que permaneciese únicamente la transcripción de lo dicho por el paciente. A partir de aquí, se exploraron distintos *pipelines* de preprocesamiento, con el objetivo de evaluar el impacto de diversos niveles de eliminación y conservación de los diferentes elementos lingüísticos presentes en las transcripciones. En concreto, se evaluaron variantes del proceso de limpieza que incluían la eliminación completa de anotaciones especiales (como errores gramaticales), de reformulaciones fonéticas, autocorrecciones y pausas, así como también un esquema en donde estos elementos eran reemplazados por descriptores lingüísticos explícitos en inglés (por ejemplo, reemplazando `[//]` por "*self-correction*" o `[+ gram]` por "*grammatical error*").

Después de múltiples pruebas, se optó por una función de limpieza que preserva en las transcripciones información potencialmente útil para el análisis del lenguaje natural relacionado con la demencia. En específico, se preservaron símbolos de pausas (`&-uh`), reformulaciones y autocorrecciones (`[/]` y `[//]`), autointerrupciones (`< ... >`) y errores gramaticales (`[+ gram]`). Esta elección estuvo fundamentada empíricamente en pruebas

que mostraron un mejor desempeño general en las métricas obtenidas en las predicciones con los distintos enfoques utilizados.

3.1.4. Generación de *embeddings* y clasificador logístico.

Para llevar a cabo la identificación de pacientes con indicios de Alzheimer, se generaron representaciones vectoriales de las transcripciones (*embeddings*) mediante distintos enfoques: un método clásico estadístico basado en la frecuencia de palabras (Tf–Idf), y métodos más recientes basados en modelos grandes del lenguaje; específicamente, se incluyeron tanto *embeddings* no contextuales (GloVe), como *embeddings* contextuales construidos a partir de *Transformers*. Dentro de este grupo, se utilizó una arquitectura *Transformer encoder* (BERT), una arquitectura *Transformer decoder* (Gemma-2B), así como una arquitectura *Transformer decoder* ajustada específicamente para tareas de generación de *embeddings* (Linq-Embed-Mistral). Estos *embeddings* se utilizaron después para alimentar un clasificador de regresión logística, buscando un balance entre el potencial representativo de los *embeddings* y la rapidez y eficacia de un modelo de clasificación logístico.

Cabe señalar que se evitó deliberadamente el ajuste fino (*fine-tuning*) de los modelos evaluados, por dos razones principales: en primer lugar, para garantizar una evaluación controlada y reproducible, sin la variabilidad asociada al ajuste de hiperparámetros específicos de cada modelo; y en segundo lugar, para valorar métodos ligeros que pudieran integrarse en entornos con recursos limitados, evitando así los costos computacionales típicamente asociados al ajuste de modelos o dominios.

Para ilustrar de manera más clara las diferencias conceptuales entre los distintos enfoques de *embeddings* analizados en este trabajo, en la Figura 6 se presenta una comparación esquemática del método Tf–Idf y de los métodos de *embeddings* basados en GloVe, BERT, Gemma-2B y Linq-Embed-Mistral. Esta representación visual resalta las distinciones entre los enfoques estadísticos basados en frecuencia, los *embeddings* no contextuales, y los *embeddings* contextuales derivados de distintos tipos de arquitecturas *Transformer*.

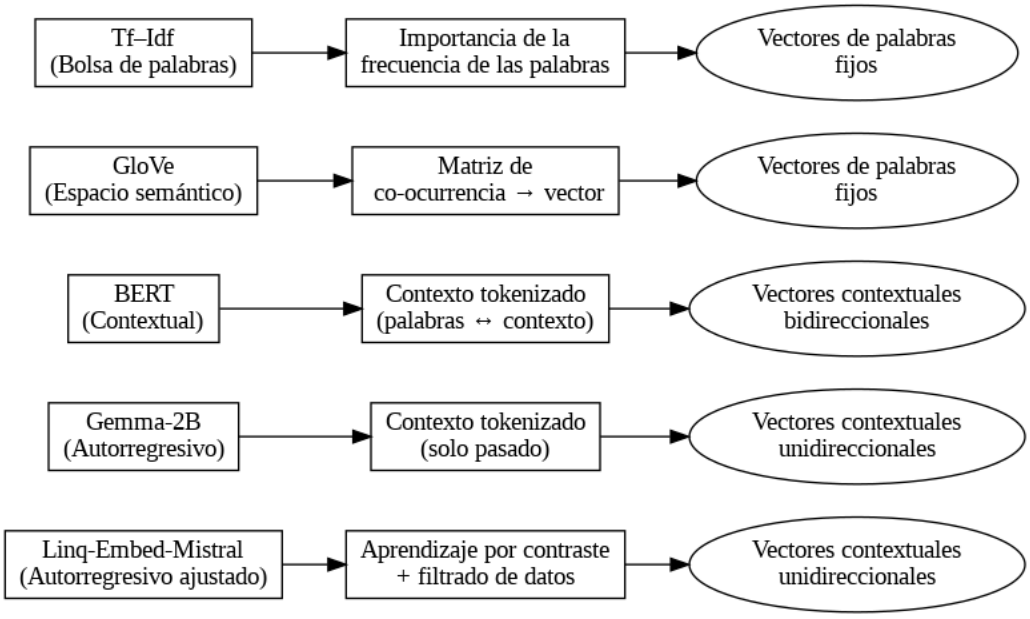


Figura 6: Comparación conceptual de los cinco enfoques de generación de *embeddings* de transcripciones evaluados en este estudio: Tf–Idf, GloVe, BERT, Gemma-2B y Linq-Embed-Mistral. (Imagen propia).

3.1.5. Enfoque Tf-Idf

En primer lugar, el enfoque Tf–Idf implicó la generación de una matriz numérica de frecuencia de palabras, la cual busca capturar la importancia relativa de palabras individuales en cada transcripción respecto al conjunto total de transcripciones analizadas. Esta técnica consiste en asignar un peso a cada término en función de su frecuencia en una transcripción determinada (frecuencia de término, *Tf*) y su frecuencia inversa en todos los documentos del corpus (frecuencia inversa de documento, *Idf*). Matemáticamente, el peso de un término t en un documento d dentro de un corpus D se calcula como:

$$\text{Tf–Idf}(t, d, D) = \text{Tf}(t, d) \cdot \log\left(\frac{N}{\text{Df}(t)}\right)$$

donde $\text{Tf}(t, d)$ representa la frecuencia del término t en el documento (transcripción, en este contexto) d, N es el número total de documentos en el corpus, y $\text{Df}(t)$ es el número de documentos (transcripciones) en los que aparece el término t. De esta forma, se busca penalizar aquellos términos que aparecen con alta frecuencia en todos los documentos (como conectores o artículos) por considerarse poco informativos, y se da mayor importancia a los términos que son característicos de documentos específicos.

Para facilitar una comprensión más clara de cómo el método Tf–Idf asigna pesos numéricos a las palabras en función de su frecuencia a lo largo de los documentos, en la Figura 7 se presenta una visualización esquemática ilustrativa del mecanismo de ponderación aplicado a un corpus de documentos sintético simplificado, pero contextualmente plausible. La figura muestra cómo Tf–Idf reduce la importancia relativa de los términos que aparecen frecuentemente en múltiples documentos, al mismo tiempo que resalta aquellos términos distintivos dentro de documentos individuales.

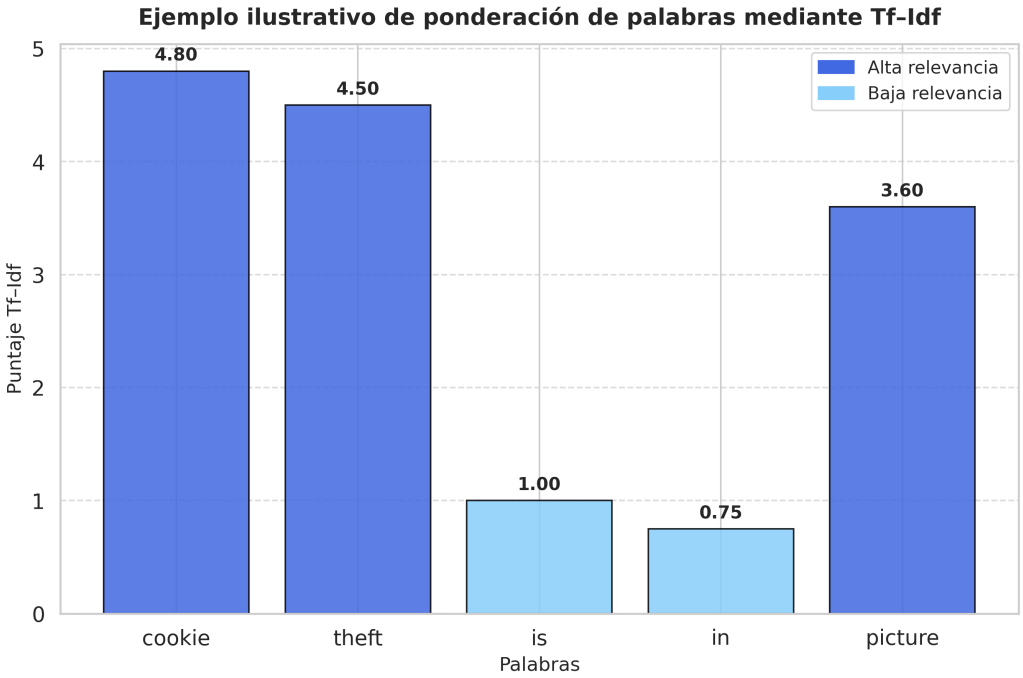


Figura 7: Visualización esquemática ilustrativa de la ponderación de palabras mediante Tf–Idf en un ejemplo simplificado de corpus de transcripciones. Los términos con mayor relevancia semántica (por ejemplo, “cookie”, “theft”, “picture”) reciben puntuaciones Tf–Idf más altas, en contraste con palabras frecuentes y menos informativas (por ejemplo, “is”, “in”). (Imagen propia).

Este enfoque ha sido ampliamente utilizado como base para tareas de recuperación de información y clasificación de textos, esto debido a su simplicidad y efectividad en contextos donde la presencia o ausencia de palabras clave puede ser determinante [57, 58]. En este sentido, está técnica es utilizada en este trabajo para explorar el impacto que tiene la utilización (o no utilización) de palabras específicas en el contexto de la prueba

del robo de la galleta, para determinar si una persona muestra (o no) signos de demencia, y comparar su desempeño con técnicas más avanzadas de NLP.

3.1.6. Enfoque *embeddings* de GloVe

Respecto a los *embeddings* no contextuales, en esta investigación se utilizaron vectores preentrenados de 300 dimensiones correspondientes al modelo GloVe (*Global Vectors for Word Representation*) [59], específicamente la versión entrenada con los corpus *Wikipedia 2014* y *Gigaword 5* (6000 millones de parámetros, 400000 palabras en el vocabulario, en minúsculas), con el objetivo de comparar la capacidad de representación y el desempeño de este modelo no contextual versus modelos contextuales para esta tarea específica.

El modelo GloVe —entrenado empleando datos de *Wikipedia 2014*, el archivo de textos de noticias *Gigaword 5* y el conjunto de páginas web *Common Crawl*— está fundamentado en una matriz de coocurrencias construida a partir de un gran corpus de texto, la cual captura cuántas veces dos palabras aparecen juntas dentro de un mismo contexto [15]. En este sentido, GloVe se entrena mediante la información contenida en esta matriz, optimizando una función de pérdida sobre una factorización implícita de la matriz de coocurrencias, que permite representar cada palabra como un vector en un espacio semántico continuo [15]. De esta forma, el modelo aprende representaciones vectoriales de palabras, capaces de preservar idealmente relaciones semánticas "proporcionales" entre las mismas (aunque sin ser capaz de diferenciar contextos diferentes para una misma palabra).

Por ejemplo, "*king - man + woman* $\approx$ *queen*", corresponde a un ejemplo típico del tipo de relaciones semánticas que surgen de forma natural dentro del espacio de *embeddings* generado por GloVe; esta relación se ilustra en una representación esquemática tridimensional en la Figura 8. Sin embargo, GloVe no es capaz de diferenciar palabras homógrafas, tales como *banco* (entidad financiera) y *banco* (lugar para sentarse), dado que siempre asigna la misma representación vectorial para una palabra dada sin importar su contexto. Para este proyecto, se realizó la conversión de las transcripciones a letra minúscula, tokenización de las mismas y posterior eliminación de las *stop-words*, previo a generar los *embeddings* de éstas con GloVe. Los *embeddings* finales de cada transcripción se obtuvieron promediando los vectores individuales de cada palabra contenida en la transcripción, y normalizando el resultado.

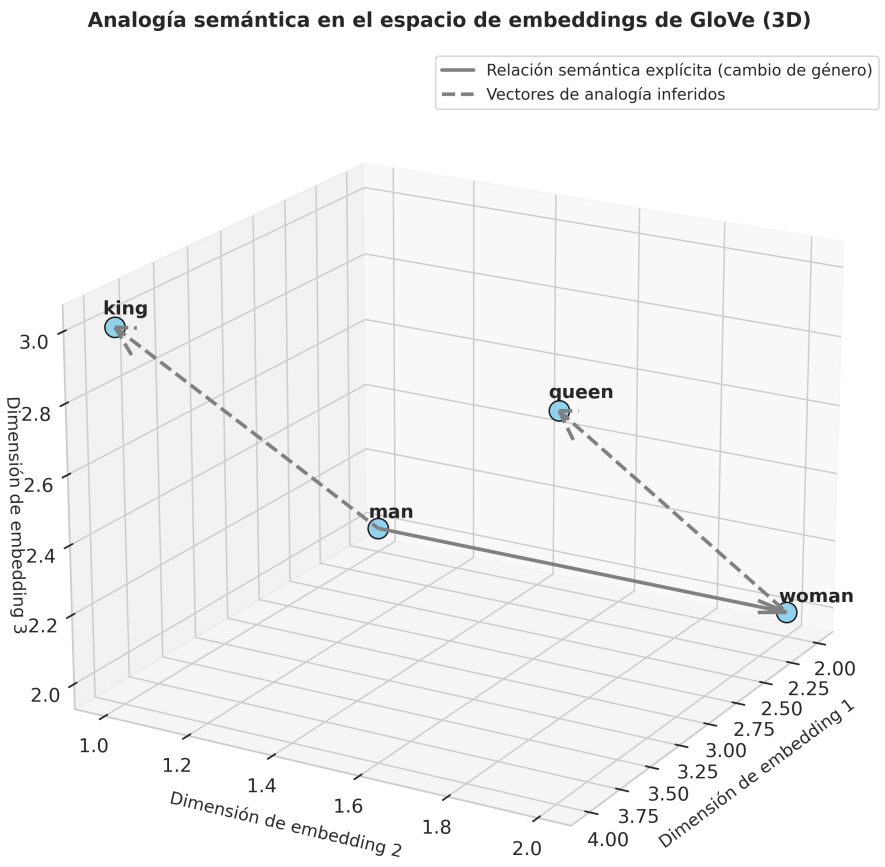


Figura 8: Representación esquemática tridimensional de analogías semánticas capturadas por los *embeddings* de GloVe, ejemplificando la aritmética vectorial semántica `king` − `man + woman` ≈ `queen`. (Imagen propia).

3.1.7. Enfoque *embeddings* de BERT

En el caso de los *embeddings* contextuales de BERT, se utilizaron implementaciones con los modelos *bert-base-uncased* y *bert-large-uncased* disponibles en la librería `transformers`. Cada transcripción fue tokenizada empleando el tokenizador correspondiente, limitando la longitud a la secuencia máxima permitida para gestionar adecuadamente los textos largos (512 tokens o subpalabras). Seguidamente se obtuvieron *embeddings* contextuales para cada token, a partir de los estados ocultos de la última capa del modelo. Finalmente, los *embeddings* para cada transcripción fueron generados mediante el promedio aritmético de los *embeddings* individuales de sus tokens, aunque también se realizaron pruebas utilizando el token especial `[CLS]` como representación vectorial de toda la transcripción, así como

una tercera configuración basada en la concatenación de los estados ocultos de las últimas cuatro capas para cada token, y luego tomando nuevamente el promedio aritmético de esos vectores para generar el *embedding* de la transcripción, como se discutirá más adelante.

El modelo BERT (*Bidirectional Encoder Representations from Transformers*) fue introducido en 2019 [12] como un modelo basado en la arquitectura *Transformer* tipo *encoder*, el cual se entrenó sobre los grandes corpus de *Wikipedia* en inglés y el dataset de libros *BookCorpus*, mediante dos tareas principales: la predicción de palabras enmascaradas dentro de una oración (*Masked Language Modeling*, este enfoque se basa en que, en lugar de intentar predecir la siguiente palabra en una secuencia dada, se entrene al modelo para predecir una palabra faltante dentro de la secuencia misma), y la predicción de coherencia entre pares de oraciones adyacentes (*Next Sentence Prediction*) [12]. Debido a este entrenamiento particular, a diferencia de modelos unidireccionales desarrollados previamente, BERT se caracteriza por realizar un entrenamiento bidireccional, lo que le permite incorporar información tanto del contexto anterior como del contexto posterior de cada palabra.

En cuanto a su arquitectura, BERT se compone exclusivamente de bloques *encoder* basados en *Transformer*; esto le permite construir representaciones jerárquicas del significado a lo largo de varias capas [12]. En este estudio, se utilizaron dos variantes del modelo: *bert-base-uncased*, con 12 capas y 110 millones de parámetros (que genera *embeddings* de 768 dimensiones), y *bert-large-uncased*, con 24 capas y 340 millones de parámetros (que genera *embeddings* de 1024 dimensiones) [12]. Ambas variantes generan, para cada token, una representación numérica (embedding) contextualizada, que depende directamente de las palabras que tiene tanto a su derecha como a su izquierda. Asimismo, BERT introduce un token especial `[CLS]` al inicio de cada *embedding*, cuyo vector asociado en la última capa del modelo está diseñado para capturar una representación integral del contenido de la secuencia; por esta razón, es comúnmente utilizado como la manera estándar para generar representaciones numéricas representativas de los enunciados en problemas de clasificación [60]. Siendo así, en este estudio se compararon tres enfoques para generar *sentence embeddings* con BERT: el uso exclusivo del *embedding* del token `[CLS]`, el promedio aritmético de todos los *embeddings* de los tokens de la transcripción, y la concatenación de los *embeddings* de las últimas cuatro capas ocultas (como fue propuesto en el artículo original de *BERT* [12]).

3.1.8. Enfoque *embeddings* de Gemma

Para los *embeddings* contextuales a partir del modelo Gemma de Google DeepMind, se empleó el modelo *gemma-2b*, disponible igualmente mediante la librería `transformers`. La generación de *embeddings* siguió un procedimiento análogo al de BERT, promediando las representaciones de la última capa oculta del modelo y adaptando la longitud máxima de tokenización a los límites dados para este modelo en específico.

El modelo Gemma es en realidad una familia *estado del arte* (2024) de modelos de lenguaje abiertos entrenados con arquitecturas del tipo *Transformer decoder* [13]. Estos modelos fueron diseñados para tareas generales de generación y comprensión del lenguaje natural; en concreto, el modelo *Gemma-2B* (modelo Gemma con 2 mil millones (2B) de parámetros), empleado en este estudio, fue entrenado de manera autorregresiva con el objetivo de predecir el siguiente token en una secuencia dada con una ventana de contexto de 8192 tokens y una dimensión de *embedding* de 2048 [13]. De esta forma, y en contraste con BERT, que emplea exclusivamente bloques de tipo *encoder* y está optimizado mediante una tarea de enmascaramiento de tokens (y por tanto, está entrenado para tomar información tanto de delante como de detrás del token para realizar una predicción), Gemma sigue un enfoque de modelado del lenguaje causal probabilístico a partir de la información previa, similar al entrenamiento de modelos como GPT.

Sin embargo, la elección de una arquitectura autorregresiva (y por tanto, unidireccional) implica que las representaciones obtenidas para cada token dependen únicamente de su contexto previo en la secuencia, lo cual podría afectar su capacidad para capturar relaciones semánticas bidireccionales (a diferencia de BERT), si bien este enfoque puede ofrecer ventajas en tareas donde la fluidez generativa y continuidad semántica hacia adelante es relevante (como por ejemplo, en la generación de texto). Esta es la razón por la cual se decidió comparar el desempeño de Gemma con el de BERT y GloVe (además del enfoque clásico Tf-Idf y del modelo de vanguardia Linq-Embed-Mistral) en esta investigación. Por una parte, para contrastar el desempeño de *embeddings* contextuales (BERT y Gemma) *versus* no contextuales (GloVe) en la tarea de la detección de indicios de Alzheimer a partir de transcripciones de audio en pruebas de descripción de imágenes, y por otra, para hacer un análisis comparativo entre dos tipos de arquitectura *Transformer* en esta tarea; una arquitectura *Transformer* tipo *encoder* con contexto bidireccional (BERT), en contraste con un modelo *estado del arte* tipo GPT (que han probado ser polifuncionales en tareas

de *NLP* [13]) como lo es Gemma, con una arquitectura *Transformer* tipo *decoder* con contexto autorregresivo unidireccional, con el objetivo final de evaluar por qué un enfoque resulta mejor (o peor) que el otro para esta tarea en concreto.

En esta línea, dado que uno de los propósitos centrales de esta investigación es analizar y comparar el rendimiento de un clasificador entrenado con *embeddings* obtenidos a partir de distintos modelos basados en *Transformers*, la Figura 9 presenta una visualización esquemática que contrasta las características arquitectónicas de BERT y Gemma-2B. Esta comparación visual se incluye para facilitar una mejor comprensión de cómo las diferencias estructurales influyen en el proceso de generación de *embeddings* y, en consecuencia, en el rendimiento predictivo del clasificador que se entrene con ellos.

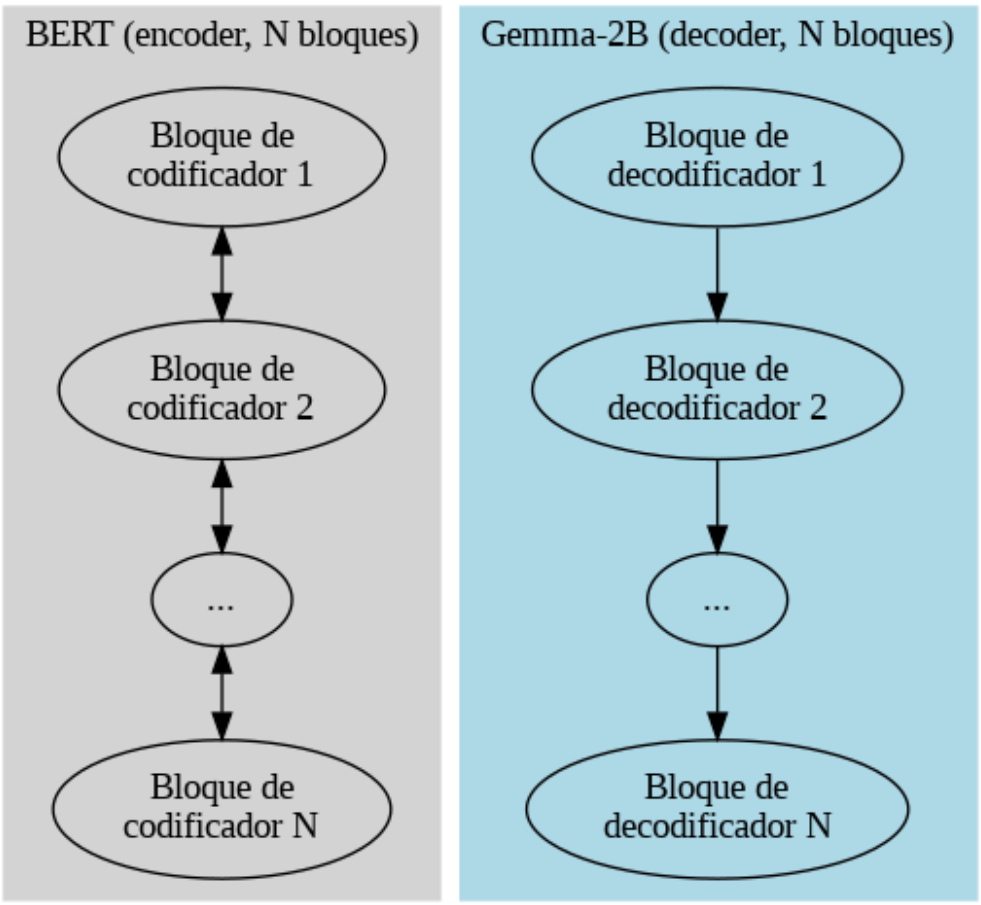


Figura 9: Contraste arquitectónico entre BERT y Gemma-2B. Nótese la bidireccionalidad de BERT y la diferencia de enfoques *encoder* contra *decoder*. (Imagen propia).

3.1.9. Enfoque *embeddings* de Linq-Embed-Mistral

Además de los modelos anteriores, en este estudio también se evaluaron *embeddings* derivados del modelo recientemente liberado y de acceso abierto *Linq-Embed-Mistral*, disponible mediante la librería `transformers` con el nombre `Linq-AI-Research/Linq-Embed-Mistral` [61]. Este modelo fue seleccionado debido a su desempeño *estado del arte* en el *Massive Text Embedding Benchmark* (MTEB) [62], donde se ubicó como el mejor modelo de acceso abierto y segundo en la clasificación general en una amplia gama de tareas de *embedding* (al 14 de mayo de 2025), como se muestra en la Tabla 2.

El *benchmark* MTEB ofrece una suite integral de evaluación para modelos de *sentence embeddings*, abarcando más de 100 tareas en nueve categorías, incluyendo recuperación, clasificación, agrupamiento, clasificación por pares y similitud semántica [63]. Se ha convertido en un estándar ampliamente utilizado para evaluar la capacidad de generalización de los modelos de *sentence embeddings* en un conjunto diverso de aplicaciones [62]. La alta posición de Linq-Embed-Mistral en este *benchmark* proporcionó una fuerte motivación para probar su aplicabilidad en tareas de clasificación de lenguaje clínico, como la detección de indicadores de demencia a partir de transcripciones de audio.

Técnicamente, Linq-Embed-Mistral se basa en una arquitectura *Transformer decoder* derivada de Mistral-7B [64], un modelo *decoder* diseñado para manejar de forma eficiente entradas de contexto largo con alta precisión semántica. El modelo fue optimizado para recuperación de texto mediante métodos avanzados de refinamiento de datos [61] y ajustado específicamente para generar *embeddings* de alta calidad a nivel de oración; emplea una metodología de entrenamiento que incluye aprendizaje por contraste con *hard negative mining*, así como un extenso proceso de filtrado de datos, lo que le permite generar *sentence embeddings* semánticamente ricos y a la vez robustos en diversas tareas de *embedding* [14]. Este enfoque de *fine-tuning* le permite superar a modelos previos de *embeddings* basados en *LLM* en múltiples aplicaciones prácticas.

Debido a los altos costos computacionales de este modelo para la generación de *embeddings*, estos no fueron generados en plataformas estándar como Google Colab. En su lugar, se empleó un servidor de cómputo de alto rendimiento equipado con cuatro GPUs NVIDIA RTX 3060 (24 GB cada una), localizado en el Centro de Investigaciones en Óptica (CIO) en León, México. Esta infraestructura computacional permitió procesar de forma eficiente

MTEB Rank	Modelo	Memoria (MB)	#Params.	Dim.	Max. Tokens	Prom.	Rec.	Clas.
1	gemini-embedding-exp-03-07 (consultado el 14 de mayo de 2025)	Desconocido	Desconocido	3072	8192	**68.37**	**67.71**	**71.82**
2	Linq-Embed-Mistral (consultado el 14 de mayo de 2025)	13563	7B	4096	32768	61.47	58.69	62.24
3	gte-Qwen2-7B-instruct (consultado el 14 de mayo de 2025)	29040	7B	3584	32768	62.51	60.08	61.55
4	multilingual-e5-large-instruct (consultado el 14 de mayo de 2025)	1068	560M	1024	514	63.22	57.12	64.94
5	SFR-Embedding-Mistral (consultado el 14 de mayo de 2025)	13563	7B	4096	32768	60.90	59.44	60.02

Tabla 2: Los 5 mejores modelos en el ranking de MTEB [63] al 14 de mayo de 2025. La tabla incluye uso de memoria, tamaño del modelo y métricas principales de evaluación del *benchmark*: desempeño promedio, recuperación y clasificación. Los modelos se ordenan utilizando el *Borda rank*.

y completa todas las transcripciones con el modelo Linq-Embed-Mistral. Los *embeddings* de las transcripciones se obtuvieron combinando las representaciones a nivel de token de la última capa del modelo y normalizando, tal como se recomienda en la documentación oficial [61], para luego ser utilizados como características de entrada para el clasificador de regresión logística.

3.1.10. Clasificador logístico

Una vez generadas las representaciones vectoriales correspondientes a cada enfoque (*embeddings* para los modelos del lenguaje y vectores de características para la técnica Tf-

Idf), estas fueron utilizadas para entrenar un modelo de clasificación basado en regresión logística. El clasificador logístico fue utilizado como una herramienta comparativa imparcial entre los diferentes *embeddings*, más que como un modelo predictivo optimizado. En esta dirección, el clasificador fue implementado con la librería `sklearn` utilizando los hiperparámetros recomendados por defecto, excepto por los que a continuación se detallan, los cuales fueron declarados explícitamente con el propósito de gestionar apropiadamente múltiples características (derivadas de los *embeddings* con alta dimensionalidad), el balance de clases (si bien en este estudio se buscó mitigar este aspecto en el pre-procesamiento de los datos) y buscando garantizar la convergencia del modelo durante el entrenamiento: `solver='lbfgs'`, `class_weight='balanced'`, `max_iter=1000`.

La regresión logística es un modelo lineal ampliamente utilizado para tareas de clasificación binaria [65]. En este estudio, se buscó distinguir entre dos clases posibles: pacientes con indicios de Alzheimer (*ProbableAD* o *PossibleAD*, codificada como 1) y pacientes sanos (*Control*, codificada como 0). Para ello, se empleó la función sigmoide, que estima la probabilidad $p(y = 1|\mathbf{x})$ de que una transcripción dada $\mathbf{x}$, representada por sus vectores de características o *embeddings*, pertenezca a la clase positiva (indicio de Alzheimer), como sigue:

$$p(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}}$$

donde $\mathbf{x}$ es el vector de características (*embeddings* o vector Tf–Idf), $\mathbf{w}$ es el vector de pesos o coeficientes aprendido por el modelo, y b es el término de sesgo del modelo logístico.

Con el objetivo de proporcionar una comprensión intuitiva de cómo la regresión logística realiza la clasificación binaria, la Figura 10 ilustra conceptualmente la frontera de decisión aprendida por el clasificador en un espacio bidimensional de *embeddings* simplificado. Aunque en la práctica los espacios de *embeddings* generados en este estudio son de alta dimensionalidad, esta figura resulta útil como una visualización de comportamiento al reproducir de forma efectiva cómo el clasificador de regresión logística separa los casos con indicio de Alzheimer de los casos *Control*.

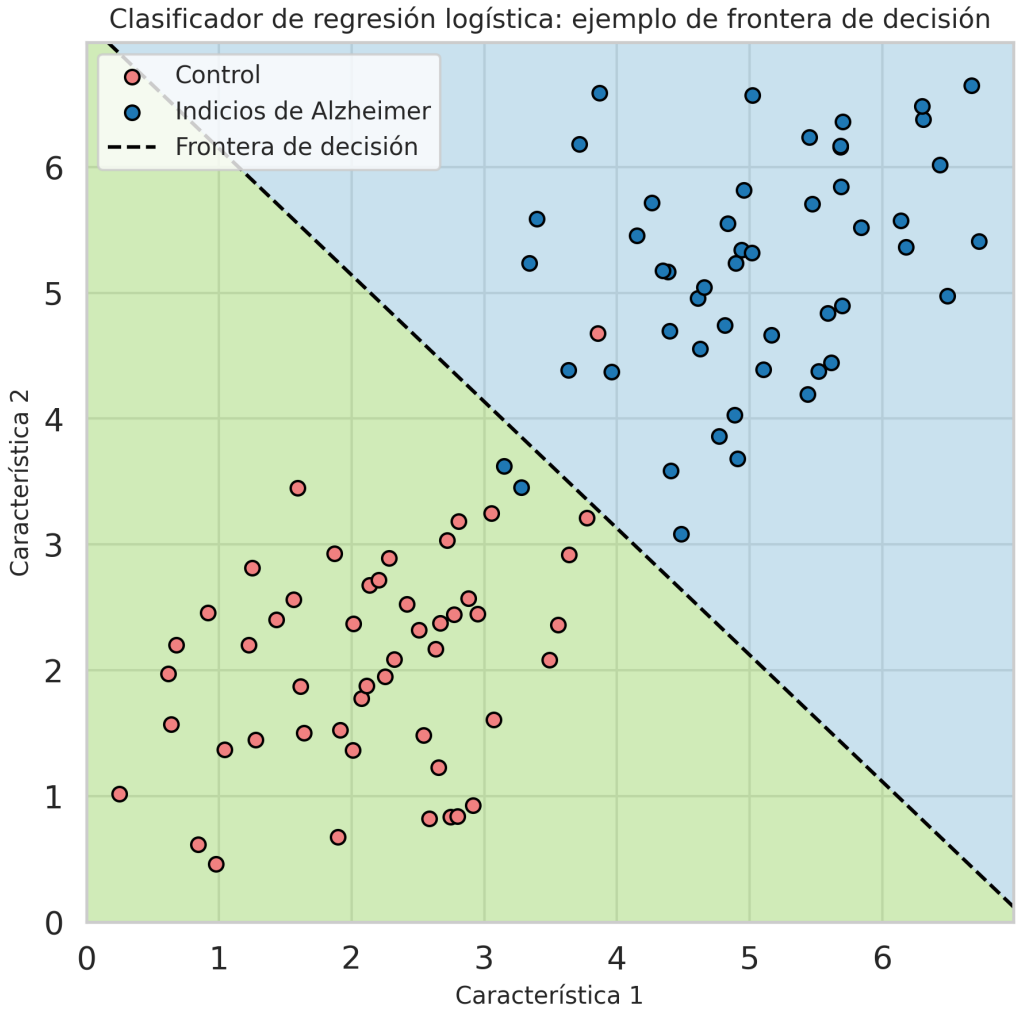


Figura 10: Ejemplo sintético de un clasificador de regresión logística aplicado a *embeddings* de transcripciones en 2D. La línea punteada representa la frontera de decisión entre las clases *indicios de Alzheimer* y *Control*. (Imagen propia).

3.1.11. Comparación del desempeño de los diferentes enfoques.

Para comparar de manera apropiada el desempeño de los cinco enfoques considerados para generar las representaciones vectoriales de las transcripciones, se llevó a cabo una validación cruzada estratificada tipo *5-fold*. En este contexto, "estratificada" significa que cada partición mantuvo una distribución balanceada entre casos con indicio de Alzheimer y casos de control, lo que redujo el sesgo de muestreo y, en consecuencia, el sesgo en los resultados. Se mantuvieron constantes las particiones generadas para cada *fold* en todas las comparaciones, con el objetivo de asegurar una evaluación justa y replicable entre los

distintos métodos.

Cada método se evaluó mediante un conjunto estándar de métricas de desempeño que incluyen exactitud (*accuracy*), precisión (*precision*), sensibilidad (*recall*) y puntaje F1 (*F1-score*). Dichas métricas se promediaron y se reportaron con su correspondiente desviación estándar, con el propósito de ofrecer un panorama confiable sobre el rendimiento general y la robustez de cada técnica.

3.2. Análisis para la construcción de la base de datos en español.

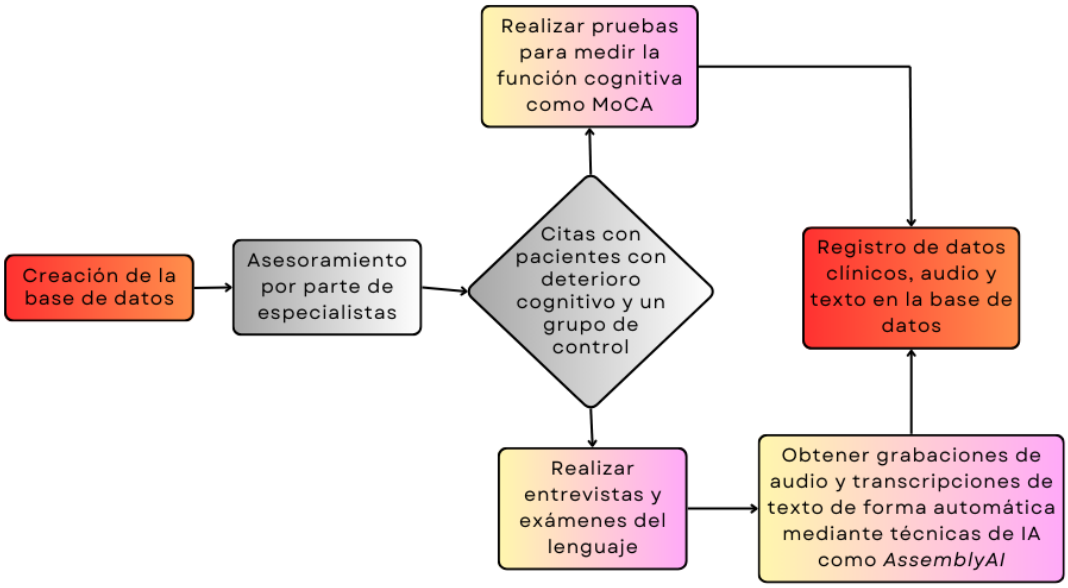


Figura 11: Diagrama de la metodología seguida para el análisis de la construcción de la base de datos en español.

Como parte de esta investigación, se inició con la construcción de una base de datos que servirá para aplicar técnicas de ML a la detección temprana de demencia en pacientes hispanohablantes. Los datos consisten en registros de audio y sus transcripciones en texto de exámenes del lenguaje, tanto de pacientes con Alzheimer diagnosticado como de aquellos con indicios de algún deterioro cognitivo, así como de un grupo de control.

En la Figura 11 se explora en detalle la metodología utilizada para llevar a cabo el análisis de la creación de esta base de datos en español. Como se puede apreciar, el primer paso consistió en el asesoramiento por parte de especialistas respecto a la mejor manera de realizar este tipo de procedimientos; consideraciones a tener en cuenta para el desarrollo exitoso del proceso. En específico, se contó con el apoyo y supervisión del Médico Especialista en Geriatría, el Dr. Humberto Güendulain Arenas para la creación de los exámenes del lenguaje, así como de la Psicóloga Lorena García Noguez, Maestra en Neurodesarrollo, para la aplicación de exámenes del lenguaje. Aunado a ello, el Maestro en Ciencias Luis Roberto García Noguez también contribuyó grandemente al desarrollo de este análisis, y a su posterior implementación práctica en la aplicación de exámenes del lenguaje.

Se solicitó asistencia por parte de los especialistas para poder concertar citas tanto con pacientes con Alzheimer diagnosticado, como con aquellos con indicios de algún deterioro cognitivo, y a su vez, se concertaron citas con sujetos que formarían parte de un grupo de control. En todas estas citas se realizaron exámenes del lenguaje, para los cuales se siguió el siguiente protocolo de aplicación, desarrollado en conjunto con el Maestro en Ciencias Luis Roberto García Noguez, con el apoyo del Dr. Humberto Güendulain Arenas. Un esbozo de las pruebas y preguntas de este protocolo se puede encontrar en la sección de Anexos.

3.2.1. Protocolo de aplicación de exámenes del lenguaje

1. **Preguntas iniciales:** Fueron utilizadas para evaluar variables sociodemográficas, de funcionalidad y hábitos de salud, con el objetivo de recabar información valiosa para la historia clínica de los sujetos (lo cual es un apoyo para identificar comorbilidades que pudieran estar causando demencias reversibles, un criterio de exclusión). Además, fueron utilizadas también para conseguir más pistas de audio para el entrenamiento de los modelos.

2. **Descripción de imágenes:** En específico, la prueba del robo de la galleta [55]

(Figura 5) y la escena del picnic [66] fueron empleadas para identificar marcadores de deterioro cognitivo.

3. **Prueba MoCA (*Montreal Cognitive Assessment*)** [67]: Se trata de una evaluación integral de las funciones cognitivas de una persona. En ella, se califican áreas como la memoria, la atención, el lenguaje, la abstracción y la orientación. Para la creación de esta base de datos, se aplicó la versión 8.1 de la prueba MoCA en español [68], con el propósito de evaluar el estado cognitivo de los sujetos en diversas áreas, y a su vez, buscando que esta prueba pudiera funcionar como una guía del correcto funcionamiento de futuros modelos que se construyesen en base a estos datos. Esta evaluación tiene una puntuación que va de 0 a 30 puntos, con puntajes posibles que van desde *desempeño cognitivo normal* (26 puntos o más) hasta *deterioro cognitivo severo* (de 0 a 9 puntos) [69]. En la versión de la prueba aplicada, es necesario además agregar un punto al puntaje total del sujeto si este cuenta con 12 años o menos de educación posterior al jardín de niños [68].

En el caso de que un deterioro cognitivo fuera sugerido por el puntaje MoCA, el siguiente paso consisitió en evaluar hábitos de sueño, estado de ánimo y capacidad para realizar actividades instrumentales de la vida diaria, a fin de identificar posibles demencias reversibles, mediante la aplicación de la Escala de Depresión Geriátrica [70], del Instituto Nacional de Geriatría, el Índice de Calidad de Sueño de Pittsburgh (PSQI) [71], del Departamento de Psiquiatría de la Universidad de Pittsburgh, y el Índice de LAWTON [72], del Instituto Nacional de Geriatría, respectivamente. Tanto las preguntas iniciales como la descripción de ambas imágenes fueron registradas en audio, mientras que para la prueba MoCA, esto sólo se realizó para la sección de "Lenguaje", debido a que es la única parte de la prueba en donde el sujeto habla sustancialmente.

Actualmente se cuenta con 4 registros de pacientes con deterioro cognitivo y 6 de pacientes sanos (a fecha de septiembre de 2025). Cabe resaltar que se tomó y aprobó la certificación oficial para la correcta aplicación de la prueba MoCA, disponible en la sección de Anexos.

3.2.2. Base de datos

Como se mencionó anteriormente, los datos consisten en registros de audio y sus transcripciones en texto de exámenes del lenguaje, así como de puntuaciones de la prueba de

evaluación de las funciones cognitivas MoCA. Esta base de datos servirá en estudios futuros en donde se apliquen técnicas de ML a la detección temprana de demencia en pacientes hispanohablantes. Por tanto, el idioma de la base de datos es el español, y los criterios de inclusión y exclusión que se están utilizando para su creación (los cuales fueron definidos en conjunto con el Médico Especialista Humberto Güendulain Arenas) son los siguientes:

1. **Pacientes de 60 años o más.**
2. **Pacientes que no cuenten con delirium o depresión diagnosticados (en general, pacientes sin comorbilidades o demencias reversibles).**
3. **Exclusión de casos avanzados.**

3.2.3. Desarrollos futuros

Se plantea el continuo desarrollo de la base de datos en estudios posteriores, buscando llegar a la meta inicial de aproximadamente 100 registros de individuos, con, idealmente, 50 de cada categoría. Se plantean también trabajos en donde se realice el preprocesamiento de los datos, mientras se continuan recabando registros de pacientes y de sujetos en el grupo de control. Este preprocesamiento, se anticipa, incluirá la obtención de las transcripciones del audio de los exámenes en forma de texto de manera automática, mediante la aplicación de técnicas de IA como el software *AssemblyAI* [73]. Este software posee la capacidad de diarizar el texto de la transcripción, y ya se ha utilizado para transcribir automáticamente algunos de los registros de audio de los exámenes del lenguaje aplicados hasta el momento, realizando un curado manual posterior de las transcripciones.

Se espera también que las técnicas desarrolladas en este trabajo para la detección de demencia en la base de datos Pitt Corpus (los distintos métodos de *embeddings*) sean aplicables en el futuro también en esta base de datos.

Capítulo 4

Resultados y discusión

4.1. Resultados

4.1.1. Métricas de evaluación

Antes de presentar el rendimiento obtenido por cada uno de los métodos de representación evaluados, esta sección comienza con una breve descripción de las métricas de evaluación empleadas. Dado que la tarea abordada en este estudio corresponde a un problema de clasificación binaria (Indicio de Alzheimer vs Control), se utilizaron las siguientes métricas estándar para evaluar el desempeño del modelo: exactitud, precisión, sensibilidad y puntaje F1.

Sea TP el número de verdaderos positivos (casos con "indicio de Alzheimer" correctamente clasificados), TN el número de verdaderos negativos (casos "control" correctamente clasificados), FP el número de falsos positivos (casos "control" clasificados incorrectamente como Alzheimer) y FN el número de falsos negativos (casos con "indicio de Alzheimer" clasificados incorrectamente como "control"). Entonces, las métricas de evaluación se definen como sigue:

$$\text{Exactitud } (\textit{Accuracy}) = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

$$\text{Precisión } (\textit{Precision}) = \frac{TP}{TP + FP} \tag{4.2}$$

$$\text{Sensibilidad } (\textit{Recall}) = \frac{TP}{TP + FN} \tag{4.3}$$

$$\text{Puntaje F1 } (\textit{F1-score}) = 2 \cdot \frac{\text{Precisión} \cdot \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}} \tag{4.4}$$

Estas métricas ofrecen perspectivas complementarias sobre el comportamiento del clasificador. Si bien la exactitud proporciona una indicación general del número de casos clasificados correctamente, puede ser engañosa en conjuntos de datos desbalanceados (aunque este no es el caso en el presente estudio). La precisión y la sensibilidad permiten una evaluación más matizada, particularmente importante cuando los costos de los falsos positivos y falsos negativos son asimétricos, como ocurre frecuentemente en tareas de diagnóstico médico [74]. Esto es especialmente relevante en el contexto de la presente investigación, ya que un falso negativo podría representar una demora significativa en el inicio del tratamiento para un paciente que eventualmente presentará síntomas claros de demencia. En esta misma línea, el puntaje F1 actúa como un indicador combinado —específicamente, la media armónica— de la precisión y la sensibilidad. Métricas como ROC-AUC y curvas de calibración, si bien valiosas en entornos clínicos sensibles al umbral de decisión, quedaron fuera del alcance de esta comparación a nivel de *embeddings* y se reservan para trabajo futuro.

4.1.2. Análisis preliminar de configuraciones de BERT

Como se estableció en la metodología del proyecto (Figura 4), previo al análisis principal se realizó un estudio preliminar con el objetivo de evaluar el desempeño de cuatro configuraciones posibles para la generación de *embeddings* contextuales bidireccionales mediante BERT. En específico, se compararon tres formas de obtener *sentence embeddings* de las transcripciones: (1) utilizando directamente el token especial `[CLS]`, que condensa información global de la oración y es considerado la manera estándar de obtener representaciones vectoriales de las oraciones en problemas de clasificación con BERT [60]; (2) promediando aritméticamente los *embeddings* individuales de cada token; y (3) concatenando los estados ocultos de las últimas cuatro capas del modelo para cada token (obteniendo vectores de 3072 dimensiones para BERT-base) y luego promediándolos para generar un único *sentence embedding* de 3072 dimensiones por transcripción.

Asimismo, se exploraron dos variantes del modelo BERT, que difieren en la dimen-

sionalidad de sus *embeddings*: BERT-base, que genera *embeddings* de 768 dimensiones, y BERT-large, que genera *embeddings* de 1024 dimensiones, presentando este último una arquitectura más compleja y con mayor capacidad potencial para capturar matices semánticos finos, tal como se discutió en la Subsección 3.1.7, pero a expensas de un mayor coste computacional. Este análisis preliminar resulta de gran importancia, pues las diferencias entre estas configuraciones podrían influir significativamente en el desempeño predictivo del modelo BERT.

Para garantizar una evaluación justa y consistente, todas las configuraciones fueron comparadas empleando exactamente la misma partición de entrenamiento-prueba (80–20), que fue obtenida de manera independiente a las particiones utilizadas posteriormente en la validación cruzada *5-fold*. Los resultados de este análisis preliminar (mostrados en la tabla 3) permitieron seleccionar objetivamente la configuración que mostró un mejor desempeño predictivo en la métrica de exactitud, siendo esta aquella que utilizó el modelo BERT-base con *sentence embeddings* generados mediante el promedio aritmético de sus *token embeddings*. Esta configuración fue utilizada posteriormente en el estudio principal en el cual se implementó validación cruzada.

Modelo de *Embedding*	**Exactitud (%)**	**Tiempo de *Embedding***
BERT-base (promedio)*	**84**	∼5 min
BERT-large (promedio)	79	∼20 min
BERT-base (`CLS`)	80	∼5 min
BERT-large (`CLS`)	70	∼20 min
BERT-base (concat. últimas 4)	81	∼5 min
BERT-large (concat. últimas 4)	81	∼20 min
Tf–Idf	**84**	**<1 s**
GloVe (300d)	80	∼1 s
Gemma-2B	80	∼2 h
Linq-Embed-Mistral	83	∼10 min †

Tabla 3: Exactitud para cada configuración de *embedding* en el análisis preliminar. El tiempo de *embedding* se refiere al tiempo total requerido para generar los *embeddings* de todas las transcripciones del conjunto de datos.

* Configuración de BERT seleccionada para el estudio principal. † Obtenido usando un servidor de alto rendimiento (4x GPU RTX 3060).

En consonancia con los hallazgos reportados en [12], donde este método mostró un buen desempeño al generar *embeddings*, la configuración que concatena las últimas cuatro capas ocultas produjo resultados competitivos. Sin embargo, en el contexto de este estudio, su rendimiento fue ligeramente inferior al del enfoque basado en el promedio de la última capa oculta del modelo BERT-base. Esto sugiere que, para esta tarea, la riqueza semántica a lo largo de capas puede no compensar la simplicidad y eficiencia de promediar los *token embeddings* finales. De forma similar, el enfoque con el token `[CLS]` arrojó resultados globalmente inferiores respecto de las otras configuraciones, indicando que, para esta tarea en concreto, la información contenida en el token `[CLS]` es menos informativa que el promedio de *embeddings* de tokens. La mejor configuración —BERT-base con *sentence embeddings* generados promediando los *embeddings* de sus tokens— se utilizó posteriormente en el estudio principal, el cual incorporó validación cruzada.

4.1.3. Análisis con validación cruzada *5-fold*

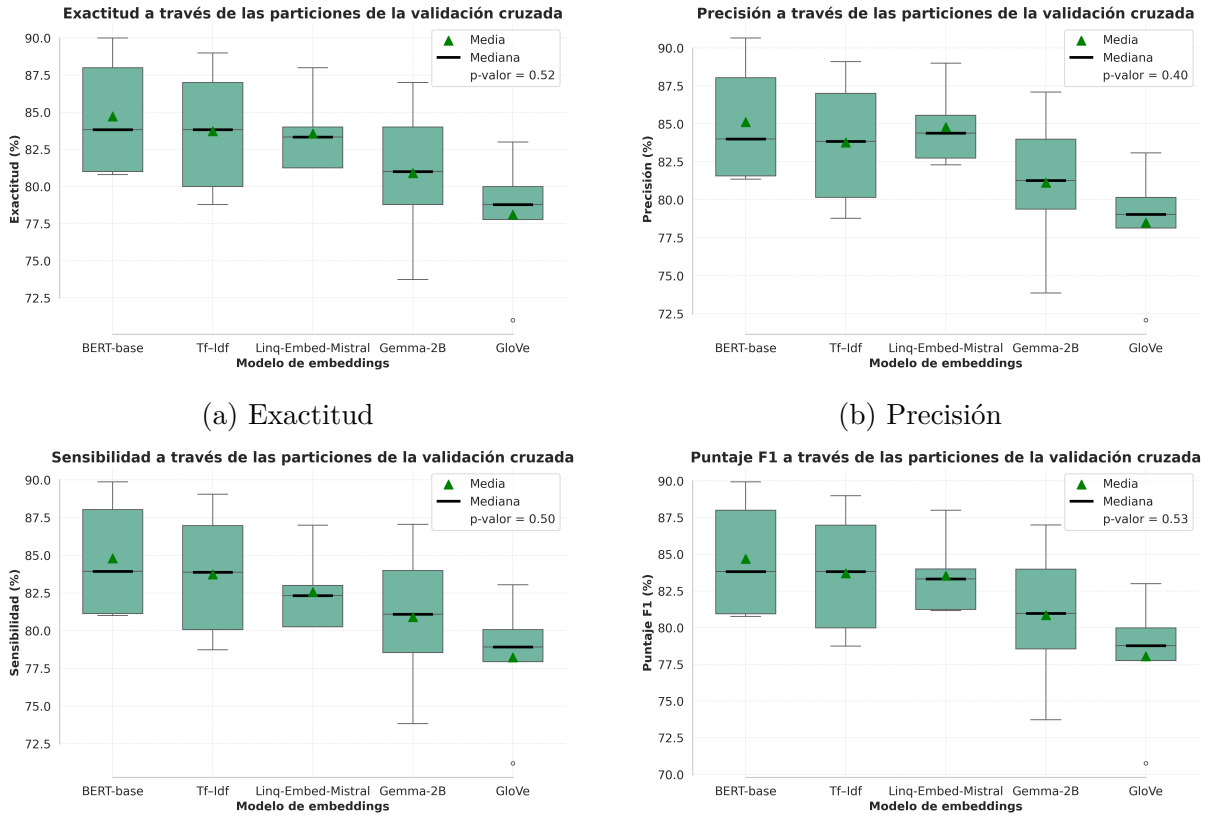
A continuación se presentan los resultados obtenidos mediante una validación cruzada estratificada *5-fold* (Tabla 4), mostrando los promedios y desviaciones estándar alcanzados para cada métrica de evaluación, para cada método de *embedding* analizado. Precisión, sensibilidad y el puntaje F1 se calcularon usando el promedio macro, aunque el balance de clases (véase Subsección 3.1.3) se preservó por estratificación.

Embedding	**Exactitud (%)**	**Precisión (%)**	**Sensibilidad (%)**	**Puntaje F1 (%)**
BERT-base (prom.)	**84.73 ± 3.70**	**85.13 ± 3.66**	**84.80 ± 3.60**	**84.70 ± 3.70**
Tf–Idf	83.73 ± 3.92	83.79 ± 3.92	83.74 ± 3.93	83.71 ± 3.93
Linq-Embed-Mistral	83.54 ± 2.67	84.71 ± 2.58	82.54 ± 2.67	83.52 ± 2.68
Gemma-2B	80.91 ± 4.53	81.13 ± 4.46	80.91 ± 4.54	80.86 ± 4.55
GloVe (300 d)	78.11 ± 3.97	78.51 ± 3.62	78.24 ± 3.91	78.06 ± 4.05

Tabla 4: Resumen de los valores promedio y desviaciones estándar de las métricas de evaluación obtenidas mediante validación cruzada *5 fold*.

De acuerdo con los resultados obtenidos, el método basado en *embeddings* contextuales bidireccionales generados por BERT (con la configuración de BERT seleccionada previamente) mostró el mejor rendimiento general, alcanzando una exactitud promedio de 84.73 % con una desviación estándar de 3.70 % en los 5 *folds*. De manera notable, y contrariamente a las expectativas iniciales, el método clásico Tf–Idf obtuvo resultados bas-

tante competitivos, con un 83.73 % de exactitud promedio y una desviación estándar de 3.92 %, apenas por debajo de BERT. De forma similar (y quizá por debajo de lo esperado) el modelo Linq-Embed-Mistral alcanzó una *accuracy* promedio de 83.54 % con desviación estándar de 2.67 %, equiparable al desempeño de BERT y Tf–Idf. La magnitud de las desviaciones estándar sugiere que la diferencia de rendimiento entre estos tres enfoques no fue realmente significativa, indicando que serían necesarias más pruebas para determinar si los *embeddings* de BERT y Linq-Embed-Mistral superan sustancialmente a Tf–Idf en este contexto específico. De hecho, una prueba t pareada entre BERT-base y Tf–Idf a lo largo de los 5 *folds* confirmó que la diferencia observada no fue estadísticamente significativa ($p > 0.05$) para ninguna de las métricas reportadas (e.g., exactitud: $t(4) = 0.70$, $p = 0.52$). Estos p-valores se reportan en la Figura 12.



(a) Exactitud

(b) Precisión

(c) Sensibilidad

(d) Puntaje F1

Figura 12: Diagramas de caja y bigotes que ilustran la distribución de las métricas de evaluación a lo largo de los 5 *folds* de la validación cruzada para cada método de *embedding*. El triángulo verde denota la media y la barra negra la mediana. El p-valor de una prueba t pareada que compara BERT-base contra Tf–Idf (por *fold*) se muestra como etiqueta en cada gráfico; $p > 0.05$ indica ausencia de diferencia estadísticamente significativa.

Los *embeddings* del modelo Gemma-2B, por su parte, alcanzaron una exactitud promedio del 80.91 %, desempeño que se sitúa entre el desempeño obtenido por BERT y del obtenido por los *embeddings* no contextuales de GloVe, con un 78.11 % de *accuracy* promedio.

Para brindar una visión más completa de la distribución de métricas a lo largo de los 5 *folds* de la validación cruzada, la Figura 12 presenta diagramas de caja y bigotes (*boxplots*) comparativos para cada métrica de evaluación (exactitud, precisión, sensibilidad y puntaje F1), agrupados por método de *embedding*.

4.2. Discusión de resultados

Estos resultados indican que, aunque los *embeddings* contextuales bidireccionales de LLM como BERT ofrecen una ligera ventaja predictiva, la diferencia con el método tradicional de Tf–Idf parece no ser significativa para esta tarea en concreto. Este hallazgo sugiere fuertemente que, en este contexto específico de la detección de Alzheimer mediante el análisis de transcripciones de la prueba del robo de la galleta, factores como la frecuencia y la elección específica de las palabras podrían ser tan o más determinantes para la detección de indicadores de demencia que la información semántica o contextual capturada por modelos más complejos de *embeddings*.

En cuanto a las posibles causas de este fenómeno, puede avanzarse la siguiente hipótesis. Como se señaló anteriormente, la prueba del robo de la galleta forma parte de la *Boston Diagnostic Aphasia Examination* [56], y fue incluida originalmente para evidenciar problemas de denominación de palabras, un rasgo típico de la afasia, pero también de la demencia [55]. Por lo tanto, es razonable esperar que esta prueba ponga de manifiesto problemas de recuperación léxica y la frecuencia relativa con que se emplean ciertas palabras en el discurso del paciente, exactamente la información que Tf–Idf cuantifica. Esto proporciona una explicación plausible del rendimiento inesperadamente alto de esta técnica clásica en el presente estudio: Tf–Idf se centra en la frecuencia y relevancia de las palabras, y se aplica aquí a una prueba diseñada explícitamente para exponer dificultades de acceso a las mismas, permitiéndole rivalizar con *embeddings* bidireccionales o semánticamente enriquecidos.

El análisis anterior destaca por su coherencia con los resultados hallados por Santander-Cruz et al. [27], en donde, al comparar la importancia de diferentes características sintácticas, léxicas y semánticas de las transcripciones presentes en la base de datos Pitt Corpus, hallaron, mediante la técnica de información mutua, que el conteo de palabras clave era el rasgo más informativo para la detección de la demencia entre las variables que consideraron, superando a otras características lingüísticas, así como a rasgos demográficos como la edad y el nivel de estudios.

En línea con lo anterior, el desempeño obtenido por el modelo Linq-Embed-Mistral, comparable tanto a BERT como a Tf–Idf, aporta perspectivas adicionales en este sentido. Aunque este modelo de última generación sobresale como uno de los mejores modelos de *sentence embeddings* en el *Massive Text Embedding Benchmark* (MTEB), sus métricas de evaluación en este estudio no superaron de forma significativa a las de Tf–Idf. Esto sugiere que, si bien modelos como Linq-Embed-Mistral exhiben una gran capacidad de generalización en una amplia gama de tareas de *embedding*, su riqueza semántica adicional no necesariamente se traduce en ventajas sustanciales en escenarios clínicos especializados. Desde un punto de vista práctico, este hallazgo enfatiza el valor de modelos más simples y de bajo costo computacional, como Tf–Idf, en aplicaciones de diagnóstico en el mundo real.

A su vez, el hecho de que el desempeño del modelo Gemma-2B resultara inferior al de BERT, sugiere que la bidireccionalidad de BERT le concedió una ventaja significativa, probablemente debido a la forma en que fue entrenado, pues una de las tareas para las que se entrenó fue precisamente la predicción de coherencia entre pares de oraciones adyacentes (*Next Sentence Prediction*), como se discutió en la Subsección 3.1.7. Este entrenamiento, se puede inferir, le permitió evaluar con mayor eficacia qué pacientes presentaron indicadores asociados a la demencia, tales como indicios de pérdida de memoria a corto plazo, de reducción de vocabulario o de alteraciones de atención, a menudo reflejados en la coherencia entre las oraciones que el paciente pronuncia.

Un punto de interés adicional es que tres de los modelos, a saber, BERT-large, Gemma-2B y Linq-Embed-Mistral, pese a su mayor dimensionalidad de *embedding* (1024, 2048 y 4096 características, respectivamente), no demostraron un rendimiento superior en esta tarea respecto al mostrado por BERT-base, que genera *embeddings* de 768 dimensiones. De hecho, los tres modelos de mayor dimensionalidad obtuvieron, en general, resultados inferiores a BERT-base en las métricas de evaluación (véanse las Tablas 3 y 4), en contra

de la creencia común de que un mayor número de dimensiones implica necesariamente mayor capacidad representacional y, por ende, mejor rendimiento. Como se observa, no siempre es el caso. A modo ilustrativo, considérese una esfera S en $\mathbb{R}^3$: ajustarla con un modelo de 6 dimensiones deja 3 coordenadas redundantes, introduciendo un pequeño error de aproximación, mientras que un modelo de 3 dimensiones evita dichas redundancias y ofrece en general un ajuste más exacto. Así, quizá una de las razones por las que BERT-base funciona mejor en esta tarea, es que proporciona una dimensionalidad de *embedding* más apropiada para representar estas transcripciones, en comparación con BERT-large, Gemma-2B y Linq-Embed-Mistral. En otras palabras, dado que todas las transcripciones provienen de la misma tarea de descripción de una imagen, emitidas por sujetos con diagnóstico de demencia o en general, de edad avanzada, es posible que no requieran de *embeddings* de mayor dimensionalidad. Futuros trabajos deberán poner a prueba esta hipótesis.

Asimismo, el hecho de que el enfoque que utilizó los *embeddings* generados por GloVe haya sido el de peor rendimiento relativo, parece indicar que en esta tarea en particular, el contexto en el que las palabras son producidas por el paciente tiene un peso específico elevado en cuanto a generar una adecuada representación vectorial de ellas, y por ende de toda la transcripción. Por ejemplo, se puede inferir que la palabra "robar", muy probablemente mencionada en alguna de sus variantes por el paciente al describir la imagen del robo de la galleta, posea en general una connotación bastante negativa, cuando en el contexto específico de la imagen, no lo es tanto, y esto es información que GloVe no alcanza a capturar. Esta incapacidad contextual, le proporciona, por tanto, una visión un tanto distorsionada (con ruido) de la transcripción en general, lo cual, se puede presumir, afecta la capacidad de predicción del clasificador logístico, que recibe sus *embeddings* como características para realizar inferencias.

En suma, los resultados obtenidos abren posibilidades a futuro para profundizar aún más en características lingüísticas específicas que contribuyan a diferenciar de manera efectiva pacientes sanos de aquellos con posibles indicios de Alzheimer. De igual modo, abren el panorama sobre técnicas útiles de análisis de exámenes del lenguaje, fáciles de calcular y con relativamente bajo costo computacional (características que mostró Tf-Idf en este estudio), que posean además una exactitud elevada, rivalizando con las técnicas del *estado del arte*. Esto convierte a Tf-Idf (y técnicas similares) en una posible candidata a ser utilizada en software o aplicaciones de teléfonos celulares destinadas a la detección

temprana de indicios de demencia.

Capítulo 5

Conclusiones

Los resultados de este trabajo constituyen un primer acercamiento hacia la tesis de que el enfoque Tf-Idf es, en general, comparable en cuanto a rendimiento (siendo en ocasiones superior), al rendimiento de los *embeddings* de modelos grandes del lenguaje, cuando son utilizados como características para alimentar un clasificador logístico con el objetivo de detectar Alzheimer en transcripciones de audio de pacientes angloparlantes. A partir de lo anterior, se puede inferir una conclusión muy interesante: los resultados sugieren que, al menos en ciertos contextos, la elección y la frecuencia de las palabras parecen ser igual (o más) relevantes que la información semántica o contextual para detectar demencia, en particular para la detección del Alzheimer. Este resultado es, en cierta medida, sorprendente, dado que es sabido que la enfermedad del Alzheimer degrada no sólo el vocabulario, sino también la estructura y la coherencia general del discurso que produce el sujeto.

Esto surge como un hallazgo inesperado y a la vez alentador, pues la técnica Tf-Idf es en general mucho menos costosa computacionalmente y más expedita en su aplicación que la carga que significa, por una parte, entrenar, y por otra, simplemente descargar y utilizar uno de estos grandes modelos del lenguaje, especialmente los más recientes *estado del arte*, como Linq-Embed-Mistral y Gemma. Esto podría conducir a una integración más sencilla de este enfoque en una futura aplicación móvil, dispositivo o sistema de software destinado a proporcionar un primer indicador de la enfermedad del Alzheimer, potencialmente sin requerir de una primera visita clínica presencial, algo particularmente valioso en contextos con acceso limitado a atención médica.

Adicionalmente, este estudio comprueba que las técnicas utilizadas en él para la de-

tección del Alzheimer son efectivas (alcanzando una exactitud de alrededor del 80 % al 85 %), en transcripciones de audio de pacientes angloparlantes de la prueba del robo de la galleta. Dicha prueba resalta por su sencillez de aplicación, por lo que estas técnicas emergen así como un procedimiento no invasivo, eficaz y rápido de aplicar (especialmente el enfoque Tf-Idf) para el escaneo expedito de un paciente en búsqueda de indicadores iniciales de demencia, pues en principio, solamente requieren de la transcripción del audio de la prueba del paciente.

Por otro lado, el análisis realizado para la construcción de una base de datos en español, que pueda ser utilizada para crear modelos de IA enfocados a la detección temprana de la demencia en el futuro, sumado al inicio de la construcción de esta base, resultó en la creación de un protocolo de aplicación de exámenes del lenguaje en el idioma español (el cual fue realizado con el apoyo de especialistas). Este protocolo es aplicable a personas de la tercera edad, con el objetivo de recabar información sobre sus funciones cognitivas y, de esa forma, que mediante estudios venideros pueda contribuir a la creación de una técnica de IA (posiblemente, alguna de las implementadas en este trabajo) para evaluar de manera oportuna a sujetos hispanohablantes con sospecha de desarrollar algún tipo de deterioro cognitivo.

5.1. Trabajos a futuro

Como futuras líneas de trabajo, se propone explorar otros enfoques de generación de *embeddings*, así como otras técnicas de ML para el clasificador final, tales como redes neuronales o una SVM. El *fine-tuning* de las últimas capas de modelos basados en *Transformers*, como BERT o Linq-Embed-Mistral, también surge como una vía prometedora. En la misma línea, técnicas de *transfer learning* orientadas a aprender representaciones más eficientes y específicas a partir de las transcripciones de esta tarea parecen valiosas para estudio futuro. Asimismo, sería pertinente ampliar el conjunto de datos para mejorar las fases tanto de entrenamiento como de evaluación, idealmente mediante colaboraciones con instituciones médicas que permitan incrementar significativamente el número de muestras disponibles (incluida la investigación en otros idiomas) y contar con la retroalimentación de especialistas clínicos, con miras a desarrollar sistemas de diagnóstico iniciales accesibles y fáciles de usar.

Adicionalmente, trabajos futuros podrían explorar avances recientes como la generación sintética de transcripciones para mitigar la escasez de muestras, especialmente para clases subrepresentadas de demencia [75, 76]. Además, aplicaciones futuras podrían integrar técnicas de explicabilidad agnósticas al modelo, como *SHapley Additive exPlanations* (SHAP) [77, 78] o *Local Interpretable Model-agnostic Explanations* (LIME) [78, 79], para aumentar la confianza clínica y la interpretabilidad de las predicciones basadas en NLP. De manera adicional, para reforzar la traslación a la práctica clínica, experimentos futuros podrían adoptar validación cruzada independiente del hablante, a fin de evitar contaminación entre pruebas de sujetos. Por otra parte, debe considerarse también la posible incorporación de otras modalidades de datos clínicos o exámenes de las funciones cognitivas (por ejemplo *Mini-Mental State Examination*, MMSE, o *Montreal Cognitive Assessment*, MoCA) además de las transcripciones lingüísticas, en pro de mejorar en mayor medida la exactitud diagnóstica y robustez general del sistema propuesto. De forma similar, enfoques multimodales que incluyan características acústicas o prosódicas, podrían emerger como una estrategia potencial para fortalecer aún más la robustez del modelo.

Bibliografía

[1] W. H. Organization, "Global status report on the public health response to dementia," World Health Organization, Geneva, Switzerland, Tech. Rep., 2021. [Online]. Available: https://iris.who.int/bitstream/handle/10665/344701/9789240033245-eng.pdf

[2] T. Ayodele, E. Rogaeva, J. T. Kurup, G. Beecham, and C. Reitz, "Early-onset alzheimer's disease: What is missing in research?" *Current Neurology and Neuroscience Reports*, vol. 21, no. 4, p. 4, 2021. [Online]. Available: https://doi.org/10.1007/s11910-021-01102-5

[3] A. Association, "2019 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 15, no. 3, pp. 321–387, March 2019.

[4] M. Haenlein and A. Kaplan, "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence," *California Management Review*, vol. 61, no. 4, pp. 5–14, 2019.

[5] V. Kaul, S. Enslin, and S. Gross, "History of artificial intelligence in medicine," *Gastrointestinal Endoscopy*, vol. 92, no. 4, pp. 807–812, 2020.

[6] K. Yamazaki, V.-K. Vo-Ho, D. Bulsara, and N. Le, "Spiking neural networks and their applications: A review," *Brain Sciences*, vol. 12, no. 7, p. 863, Jun. 2022.

[7] N. Ettehadi, X. Zhang, Y. Wang, D. Semanek, J. Guo, and J. Posner, "Automatic volumetric quality assessment of diffusion mr images via convolutional neural network classifiers," in *Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Virtual, 1–5 November 2021, 2021, pp. 2756–2760. [Online]. Available: https://doi.org/10.1109/EMBC46164.2021.9630834

[8] N. Liu, L. Wan, Y. Zhang, T. Zhou, H. Huo, and T. Fang, "Exploiting convolutional neural networks with deeply local description for remote sensing image classification," *IEEE Access*, vol. 6, pp. 11 215–11 228, 2018.

[9] K. M. Ruff and R. V. Pappu, "Alphafold and implications for intrinsically disordered proteins," *Journal of Molecular Biology*, vol. 433, no. 20, p. 167208, 2021, from Protein Sequence to Structure at Warp Speed: How Alphafold Impacts Biology. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283621004411

[10] M. Rashed-Al-Mahfuz, M. A. Moni, S. Uddin, S. A. Alyami, M. A. Summers, and V. Eapen, "A deep convolutional neural network method to detect seizures and characteristic frequencies using epileptic electroencephalogram (eeg) data," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1–12, 2021, art no. 2000112.

[11] M. Eni, I. Dinstein, M. Ilan, I. Menashe, G. Meiri, and Y. Zigel, "Estimating autism severity in young children from speech signals using a deep neural network," *IEEE Access*, vol. 8, pp. 139 489–139 500, 2020.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Minneapolis, MN, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[13] T. Mesnard, D. Keysers, Y. Jernite, X. Wang, S. Venkatesh, K. Clark, J. Wei, N. Stiennon, D. Dohan, Y. Kilcher, J. Uszkoreit, B. A. y Arcas, O. Vinyals, Q. V. Le, N. Shazeer, J. Dean, S. Petrov, D. Eck, M. Bosma, and C. Raffel, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295v4*, Apr. 2024, accessed: April 2025. [Online]. Available: https://arxiv.org/abs/2403.08295

[14] C. Choi, J. Kim, S. Lee, J. Kwon, S. Gu, Y. Kim, M. Cho, and J. yong Sohn, "Linq-embed-mistral technical report," 2024. [Online]. Available: https://arxiv.org/abs/2412.03223

[15] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[16] Z. Arvanitakis, R. Shah, and D. Bennett, "Diagnosis and management of dementia: Review," *JAMA*, vol. 322, no. 16, p. 1589, Oct. 2019.

[17] J. Vrijsen, T. F. Matulessij, T. Joxhorst, S. E. de Rooij, and N. Smidt, "Knowledge, health beliefs and attitudes towards dementia and dementia risk reduction among the dutch general population: a cross-sectional study," *BMC Public Health*, vol. 21, no. 857, 2021.

[18] A. Javeed, A. L. Dallora, J. S. Berglund, A. Ali, L. Ali, and P. Anderberg, "Machine learning for dementia prediction: A systematic review and future research directions," *Journal of Medical Systems*, vol. 47, no. 17, 2023.

[19] C. Salvatore, A. Cerasa, P. Battista, M. Gilardi, A. Quattrone, and I. Castiglioni, "Magnetic resonance imaging biomarkers for the early diagnosis of alzheimer's disease: A machine learning approach," *Frontiers in Neuroscience*, vol. 9, p. 307, 2015.

[20] A. Bidani, M. S. Gouider, and C. M. Travieso-González, "Dementia detection and classification from mri images using deep neural networks and transfer learning," in *Proceedings of the International Work-Conference on Artificial Neural Networks (IWANN)*. Munich, Germany, 17–19 September 2019: Springer, 2019, pp. 925–933.

[21] S. Basheer, S. Bhatia, and S. Sakri, "Computational modeling of dementia prediction using deep neural network: Analysis on oasis dataset," *IEEE Access*, vol. 9, pp. 42 449–42 462, 2021.

[22] M. Alam, N. Roy, S. Holmes, A. Gangopadhyay, and E. Galik, "Automated functional and behavioral health assessment of older adults with dementia," in *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 2016, pp. 140–149.

[23] P.-Y. Chiu, H. Tang, C.-Y. Wei, C. Zhang, G.-U. Hung, and W. Zhou, "Nmd-12: A new machine-learning derived screening instrument to detect mild cognitive impairment and dementia," *PLOS ONE*, vol. 14, no. 3, p. e0213430, 2019.

[24] H. Hsiu, S.-K. Lin, W.-L. Weng, C.-M. Hung, C.-K. Chang, C.-C. Lee, and C.-T. Chen, "Discrimination of the cognitive function of community subjects using the arterial pulse spectrum and machine-learning analysis," *Sensors*, vol. 22, no. 3, p. 806, 2022.

[25] R. Sadeghian, J. Schaffer, and S. Zahorian, "Speech processing approach for diagnosing dementia in an early stage," in *Proceedings of INTERSPEECH 2017*, Stockholm, Sweden, 20–24 August 2017, 2017, pp. 2705–2709. [Online]. Available: https://doi.org/10.21437/Interspeech.2017-1712

[26] L. Ilias and D. Askounis, "Explainable identification of dementia from transcripts using transformer networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4153–4164, 2022.

[27] Y. Santander-Cruz, S. Salazar-Colores, W. Paredes-García, H. Guendulain-Arenas, and S. Tovar-Arriaga, "Semantic feature extraction using sbert for dementia detection," *Brain Sciences*, vol. 12, no. 2, p. 270, 2022.

[28] B. MacWhinney, "Dementiabank: Pitt corpus," https://dementia.talkbank.org/access/English/Pitt.html, 2020, accessed: April 5, 2025.

[29] Z. Arvanitakis and D. Bennett, "What is dementia?" *JAMA*, vol. 322, no. 17, p. 1728, Nov. 2019.

[30] G. Mirzaei and H. Adeli, "Machine learning techniques for diagnosis of alzheimer disease, mild cognitive disorder, and other types of dementia," *Biomedical Signal Processing and Control*, vol. 72, p. 103293, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421008909

[31] Q. Bi, K. Goodman, J. Kaminsky, and J. Lessler, "What is machine learning? a primer for the epidemiologist," *American Journal of Epidemiology*, vol. 188, no. 12, pp. 2222–2239, December 2019. [Online]. Available: https://doi.org/10.1093/aje/kwz189

[32] B. Macukow, "Neural networks – state of art, brief history, basic models and architecture," in *Computer Information Systems and Industrial Management*, ser. Lecture Notes in Computer Science, K. Saeed and W. Homenda, Eds. Springer, Cham, 2016, vol. 9842, pp. 1–1.

[33] B. Mahesh, "Machine learning algorithms - a review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, 2020.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[35] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1724–1734. [Online]. Available: https://aclanthology.org/D14-1179

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[37] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd ed., 2025, online manuscript released August 24, 2025. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[38] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[39] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703/

[40] M. V. Koroteev, "Bert: A review of applications in natural language processing and understanding," 2021. [Online]. Available: https://arxiv.org/abs/2103.11943

[41] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, p. 126232, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223003557

[42] K. P. Gunasekaran, "Exploring sentiment analysis techniques in natural language processing: A comprehensive review," *arXiv preprint arXiv:2305.14842*, may 2023. [Online]. Available: https://arxiv.org/abs/2305.14842

[43] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A systematic literature review on phishing email detection using natural language processing techniques," *IEEE Access*, vol. 10, pp. 65 703–65 727, 2022.

[44] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," *npj Digital Medicine*, vol. 5, no. 1, p. 46, 2022. [Online]. Available: https://www.nature.com/articles/s41746-022-00589-7

[45] J. Sawicki, M. Ganzha, and M. Paprzycki, "The state of the art of natural language processing—a systematic automated review of nlp literature using nlp techniques," *Data Intelligence*, vol. 5, no. 3, pp. 707–749, 08 2023. [Online]. Available: https://doi.org/10.1162/dint_a_00213

[46] A. A. Abro, M. S. H. Talpur, and A. K. Jumani, "Natural language processing challenges and issues: A literature review," *Gazi University Journal of Science*, vol. 36, no. 4, p. 1522–1536, 2023.

[47] Z. Zhang, D. Zhang-Li, J. Yu, L. Gong, J. Zhou, Z. Hao, J. Jiang, J. Cao, H. Liu, Z. Liu, L. Hou, and J. Li, "Simulating classroom education with llm-empowered agents," 2024. [Online]. Available: https://arxiv.org/abs/2406.19226

[48] Y. Kim, X. Xu, D. McDuff, C. Breazeal, and H. W. Park, "Health-llm: Large language models for health prediction via wearable sensor data," 2024. [Online]. Available: https://arxiv.org/abs/2401.06866

[49] A. Cambon, B. Hecht, B. Edelman, D. Ngwe, S. Jaffe, A. Heger, M. Vorvoreanu, S. Peng, J. Hofman, A. Farach, M. Bermejo-Cano, E. Knudsen, J. Bono, H. Sanghavi, S. Spatharioti, D. Rothschild, D. G. Goldstein, E. Kalliamvakou, P. Cihon,

M. Demirer, M. Schwarz, and J. Teevan, "Early llm-based tools for enterprise information workers likely provide meaningful boosts to productivity," Microsoft, Microsoft Technical Report MSR-TR-2023-43, Dec. 2023. [Online]. Available: https://www.microsoft.com/en-us/research/publication/early-llm-based-tools-for-enterprise-information-workers-likely-provide-meaningful-boosts-to-productivity/

[50] P. Lanillos, D. Oliva, A. Philippsen, Y. Yamashita, Y. Nagai, and G. Cheng, "A review on neural network models of schizophrenia and autism spectrum disorder," *Neural Networks*, vol. 122, pp. 338–363, 2020.

[51] F. Ke, S. Choi, Y. Kang, K.-A. Cheon, and S. Lee, "Exploring the structural and strategic bases of autism spectrum disorders with deep learning," *IEEE Access*, vol. 8, pp. 153 341–153 352, 2020.

[52] M. Aidid and R. Musa, "Accuracy of supervised machine learning in predicting depression, anxiety and stress using web-based big data: Preserving the humanistic intellect," *Malaysian Journal of Medicine and Health Sciences*, vol. 18, pp. 87–92, 2022.

[53] M. Pandit, M. Azwaan, S. Wani, A. A. Ibrahim, R. A. Abdulghafor, and Y. Gulzar, "Examining factors for anxiety and depression prediction," *International Journal on Perceptive and Cognitive Computing*, vol. 9, no. 1, pp. 70–79, Jan. 2023.

[54] J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[55] L. Cummings, "Describing the cookie theft picture: Sources of breakdown in alzheimer's dementia," *Pragmatics and Society*, vol. 10, pp. 151–174, 03 2019.

[56] O. Spreen and A. H. Risser, "Assessment of aphasia," in *Acquired Aphasia*, 3rd ed., M. T. Sarno, Ed. San Diego, CA, USA: Academic Press, 1998, pp. 71–156. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780126193220500075

[57] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1988.

[58] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008.

[59] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," Available online: https://nlp.stanford.edu/projects/glove/ (accessed on 2 October 2024), 2014, pre-trained word vectors: Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors). [Online]. Available: https://nlp.stanford.edu/projects/glove/

[60] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how bert works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.54

[61] Linq AI Research, "Linq-embed-mistral (hugging face repository)," https://huggingface.co/Linq-AI-Research/Linq-Embed-Mistral, 2024, accessed: 2025-05-14.

[62] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive text embedding benchmark," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2014–2037. [Online]. Available: https://aclanthology.org/2023.eacl-main.148/

[63] N. Muennighoff, A. Webson, T. Liu, T. Schick, M. Ott, I. Gurevych, and P. Lewis, "Mteb: Massive text embedding benchmark," https://huggingface.co/spaces/mteb/leaderboard, 2023, accessed: 2025-05-14.

[64] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed, "Mistral 7b," *CoRR*, vol. abs/2310.06825, 2023. [Online]. Available: https://arxiv.org/abs/2310.06825

[65] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ: Wiley, 2013.

[66] K. D. Mueller, B. Hermann, J. Mecollari, and L. S. Turkstra, "Connected speech and language in mild cognitive impairment and alzheimer's disease: A review of picture description tasks," *Journal of Clinical and Experimental*

Neuropsychology, vol. 40, no. 9, pp. 917–939, 2018, pMID: 29669461. [Online]. Available: https://doi.org/10.1080/13803395.2018.1446513

[67] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[68] Z. S. Nasreddine, *Montreal Cognitive Assessment (MoCA), Versión 8.1 en español: Instrucciones para la administración y puntuación*, MoCA Cognition, 2018, moCA-8.1-Instructions — Spain/Spanish; versión del 19 feb. 2018 (basada en MoCA v8.1, 28 jun. 2017). Consultado: 14 feb. 2024. [Online]. Available: https://championsforhealth.org/wp-content/uploads/2018/12/MOCA-8.1-Spanish.pdf

[69] MoCA Cognition, "The MoCA test," https://mocacognition.com/faq/, 2024, consultado: 14 feb. 2024.

[70] *Escala de Depresión Geriátrica (GDS-15): ficha de aplicación*, Instituto Nacional de Geriatría (INGER), Ciudad de México, México, 2020, entidad de Certificación y Evaluación; material con licencia Creative Commons BY-NC-ND 4.0. [Online]. Available: https://gc.scalahed.com/recursos/files/r161r/w25740w/Guia_InstrumentosGeriatrica2020.pdf

[71] D. J. Buysse, C. F. Reynolds III, T. H. Monk, S. R. Berman, and D. J. Kupfer, "Índice de calidad de sueño de pittsburgh (psqi): cuestionario e instrucciones (versión española)," Folleto en línea alojado por la Universitat de Barcelona (Departamento de Psicobiología). [Online]. Available: https://www.ub.edu/psicobiologia/Pmemlleng/images/Index%20de%20Pittsburgh.pdf

[72] *Actividades instrumentales de la vida diaria (Índice de Lawton): ficha de aplicación*, Instituto Nacional de Geriatría (INGER), Ciudad de México, México, 2020, entidad de Certificación y Evaluación; Comité de Gestión por Competencias de la Geriatría; material con licencia Creative Commons BY-NC-ND 4.0. [Online]. Available: https://gc.scalahed.com/recursos/files/r161r/w25740w/Guia_InstrumentosGeriatrica2020.pdf

[73] AssemblyAI, "Assemblyai: Speech-to-text to powerful outcomes," https://www.assemblyai.com/, 2024, consultado: 14 feb. 2024.

[74] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[75] A. Hlédiková, D. Woszczyk, A. Akman, S. Demetriou, and B. Schuller, "Data augmentation for dementia detection in spoken language," *arXiv preprint arXiv:2206.12879*, 2022.

[76] L. García-Noguez, S. Salazar-Colores, S. Mondragón-Rodríguez, and S. Tovar-Arriaga, "A novel methodology for data augmentation in cognitive impairment subjects using semantic and pragmatic features through large language models," *Technologies*, vol. 13, p. 344, 08 2025.

[77] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.

[78] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of lime and shap in alzheimer's disease detection," *Brain Informatics*, vol. 11, no. 1, p. 10, 2024.

[79] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

Anexos

Anexo 1: Certificado de aprobación del curso oficial para aplicación de la prueba MoCA.

Figura 13: Certificado de aprobación del curso oficial para administrar y puntuar la prueba MoCA.

Anexo 2: Esbozo del protocolo de aplicación de exámenes del lenguaje.

La siguiente prueba consiste en 3 partes, listadas a continuación:

1. Preguntas iniciales.
2. Descripción de imagen del robo de la galleta e imagen de la escena del picnic.
3. Prueba MoCA.

    a) Si el paciente obtiene un puntaje en la prueba MoCA de 25 o menos, se procede a la aplicación de las siguientes pruebas, con el objetivo de tener un registro de posibles causas de deterioro cognitivo:

        1) Escala de Depresión Geriátrica, del Instituto Nacional de Geriatría [70].
        2) Índice de Calidad de Sueño de Pittsburgh (PSQI), del Departamento de Psiquiatría de la Universidad de Pittsburgh [71].
        3) Índice de LAWTON, del Instituto Nacional de Geriatría [72].

Tanto las preguntas iniciales como la descripción de ambas imágenes serán registradas en audio, mientras que para la prueba MoCA esto sólo será necesario para la sección de "Lenguaje".

Primera parte - Preguntas iniciales (en este punto empieza el diálogo con el paciente):

Entrevistador: Qué tal, buenos días. Vamos a comenzar con unas breves preguntas, con la intención de conocer un poco más sobre usted.

1. ¿Qué edad tiene?
2. ¿Cuál es su nivel de estudios?
3. ¿Podría por favor describir si actualmente tiene alguna dificultad para llevar a cabo sus actividades diarias (poner ejemplos para que la persona conozca a qué nos referimos)?

4. ¿Podría por favor describir si sus familiares cuentan o contaron con alguna enfermedad hereditaria? Por ejemplo, diabetes, demencia, hipertensión o alguna enfermedad relacionada con la tiroides. Revisar historia clínica.

5. ¿Podría por favor describir cuál es su estado de salud actual y si cuenta con alguna enfermedad diagnosticada? Por ejemplo, depresión, cirrosis, problemas en la tiroides, etc.

6. ¿Podría mencionar por favor cómo considera sus hábitos de salud actualmente, es decir, si considera que tiene una alimentación balanceada, si hace ejercicio regularmente, si quizás fuma o toma alcohol y de ser así, con qué frecuencia lo hace?

7. ¿Nos podría comentar cómo ha sido su estado de ánimo en las últimas semanas, sin considerar situaciones lamentables extraordinarias como lo podría ser la pérdida de un familiar? Es decir, ¿cómo se encuentra su estado anímico, se ha sentido triste últimamente?

8. Ahora, ¿nos podría mencionar cómo considera sus hábitos de sueño en las últimas semanas? Es decir, ¿ha descansado apropiadamente o ha tenido problemas para descansar o conciliar el sueño?

Muy bien, gracias, con respecto a las preguntas, por esta sección hemos terminado.

Segunda parte - Descripción de imágenes:

Entrevistador: Ahora vamos a pasar a la siguiente sección. En esta parte, le vamos a mostrar dos imágenes, y lo que buscamos es que usted pueda describirlas lo mejor que pueda, dando todos los detalles y características que pueda ver.

Esta es la primera imagen: (En este momento, se muestra la imagen de la prueba del robo de la galleta [55] (Figura 5)).

Muy bien, muchas gracias. Ahora, le voy a presentar la siguiente imagen, recordándole que lo que buscamos es que usted pueda describirla con el mayor detalle posible: (En este momento, se muestra la imagen de la escena del picnic [66]).

Excelente, muchas gracias.

Tercera parte - Aplicación de la prueba MoCA:

Entrevistador: Por último, vamos a pasar a la aplicación de una prueba para evaluar sus habilidades cognitivas. A esta prueba se le conoce como la prueba MoCA. Vamos a iniciar con la evaluación. Consiste en una serie de ejercicios de memoria, lenguaje, atención, etc., que vamos a puntuar y, al final, le brindaremos su calificación.

(En este punto, se procede a la aplicación de la prueba MoCA [68]).

Muy bien, hemos terminado con la prueba MoCA y con esto terminamos la sesión. Le agradecemos mucho su tiempo y su disposición a contribuir en este proyecto.