
Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

# Aplicación de embeddings de BERT para detección automática de Alzheimer

Tesis de Licenciatura en Ciencias de la Computación

Sofía Milena Goldberg

Director: Pablo Brusco
Codirectora: Lara Gauder
Buenos Aires, 2025

# APLICACIÓN DE EMBEDDINGS DE BERT PARA DETECCIÓN AUTOMÁTICA DE ALZHEIMER

La detección temprana del Alzheimer representa un desafío clave en el ámbito médico, ya que un diagnóstico preciso en las primeras etapas de la enfermedad puede facilitar intervenciones más efectivas y mejorar la calidad de vida de los pacientes. En este contexto, el análisis del habla y el lenguaje ha surgido como una herramienta prometedora para identificar patrones lingüísticos asociados con el deterioro cognitivo. En este estudio, investigamos la efectividad de los embeddings generados con BERT para la clasificación de transcripciones de habla, con el propósito de distinguir entre individuos con Alzheimer y controles sanos, en inglés y en español. Además de replicar un trabajo previo, ampliamos el análisis comparando el desempeño de distintas representaciones de los textos, agregando métricas de evaluación y observando el impacto de utilizar modelos entrenados con texto capitalizado (cased) y modelos entrenados únicamente con texto en minúsculas (uncased).

Tanto en inglés como en español, nuestros resultados superaron a los reportados en el trabajo replicado. En inglés, el mejor F1-score obtenido fue de 0,76 con Random Forest, superando el 0,69 reportado en el trabajo original con XGBoost. En español, SVM alcanzó un F1-score de 0,70, mejorando significativamente el 0,53 reportado.

Nuestros experimentos revelaron que el F1-score no es una métrica adecuada para evaluar el desempeño de los clasificadores, especialmente en conjuntos de datos desbalanceados. Por ello, analizamos métricas adicionales como precision, recall, accuracy, specificity, ROC AUC y PR AUC, que permitieron una evaluación más detallada del rendimiento de los clasificadores.

Los hallazgos obtenidos evidencian que las representaciones contextuales derivadas de la última capa de BERT superan a los embeddings extraídos de la primera capa, gracias a su capacidad de capturar información semántica más rica y dependiente del contexto. En el Pitt Corpus (inglés), los embeddings de la última capa lograron un ROC AUC de hasta 0,89, mientras que los embeddings de la primera capa alcanzaron un máximo de 0,84. En Chile AD (español), aunque el rendimiento general fue inferior, los embeddings de la última capa de BERT obtuvieron un ROC AUC de 0,72, superando ampliamente a los embeddings provenientes de la primera capa, cuyo mejor desempeño fue 0,54.

Asimismo, observamos que en español, los modelos cased mejoran el rendimiento de los clasificadores, mientras que en inglés, los modelos uncased resultan más eficaces. En el Pitt Corpus, SVM con embeddings de la última capa logró un ROC AUC de 0,90 con el modelo uncased, mientras que con el modelo cased obtuvo 0,87. En contraste, en Chile AD, el uso de la versión uncased redujo significativamente el desempeño, con una caída en ROC AUC de 0,72 a 0,41 en Random Forest con los embeddings de la primera capa.

Nuestros resultados indican que, en general, el desempeño en inglés fue superior al obtenido en español, lo que podría atribuirse a la menor cantidad de datos disponibles, el desbalance entre las clases en el conjunto en español o a las diferencias entre los modelos de BERT empleados en cada idioma.

**Palabras claves:** Alzheimer, Embeddings, BERT, Clasificación, Análisis Cruzado entre Lenguajes

# APPLICATION OF BERT EMBEDDINGS FOR ALZHEIMER'S AUTOMATIC DETECTION

Early detection of Alzheimer's disease is a key challenge in the medical field, as an accurate diagnosis in the early stages of the disease can enable more effective interventions and improve patients' quality of life. In this context, speech and language analysis has emerged as a promising tool for identifying linguistic patterns associated with cognitive decline. In this study, we investigate the effectiveness of BERT-generated embeddings for classifying speech transcriptions to distinguish between individuals with Alzheimer's and healthy controls in both English and Spanish. In addition to replicating a previous study, we extend the analysis by comparing the performance of different text representations, incorporating additional evaluation metrics, and examining the impact of using models trained on capitalized text (cased) versus models trained exclusively on lowercase text (uncased).

In both English and Spanish, our results outperformed those reported in the replicated study. In English, the best F1-score obtained was 0,76 with Random Forest, surpassing the 0,69 reported in the original study with XGBoost. In Spanish, SVM achieved an F1-score of 0,70, significantly improving upon the 0,53 reported.

Our experiments revealed that the F1-score is not an adequate metric to fully evaluate classifier performance, especially in imbalanced datasets. Therefore, we analyzed additional metrics such as precision, recall, accuracy, specificity, ROC AUC, and PR AUC, which provided a more detailed assessment of classifier performance.

The findings show that contextual representations derived from BERT's last layer outperform embeddings extracted from the first layer, thanks to their ability to capture richer, context-dependent semantic information. In the Pitt Corpus (English), last-layer embeddings achieved a ROC AUC of 0,89, whereas first-layer embeddings reached a maximum of 0,84. In Chile AD (Spanish), although overall performance was lower, last-layer BERT embeddings obtained a ROC AUC of 0,72, significantly surpassing the 0,54 achieved with first-layer embeddings.

Furthermore, we observed that in Spanish, cased models improve classifier performance, whereas in English, uncased models are more effective. In the Pitt Corpus, SVM with last-layer embeddings achieved a ROC AUC of 0,90 with the uncased model, while the cased model yielded 0,87. In contrast, in Chile AD, using the uncased version significantly reduced performance, with a drop in ROC AUC from 0,72 to 0,41 in Random Forest with first-layer embeddings.

Our results indicate that, overall, performance in English was superior to that in Spanish. This could be attributed to the smaller amount of available data, the class imbalance in the Spanish dataset, or differences between the BERT models used for each language.

**Keywords:** Alzheimer, Embeddings, BERT, Classification, Cross-Lingual Analysis

# AGRADECIMIENTOS

A mis directores, Pablo y Lara, por su guía y apoyo durante todo este proceso. Gracias Pablo por enseñarme todo lo que sé de aprendizaje automático y por transmitirme tu pasión por este área. Gracias Lara por elegirme como tu primer tesista y confiar en mí. Está de más decir que sin ustedes este trabajo no hubiera sido posible.

A mi familia, por su apoyo incondicional. A mi papá, que así lloviera, tronara o jugara Boca siempre estuvo ahí para buscarme de la facultad o esperarme con la cena lista. A mi mamá, y ahora colega, que desde el día uno estuvo ahí, tanto para ayudarme como para escuchar todas mis anécdotas y contarme las suyas, si no fuera por vos nunca hubiera estudiado esta carrera hermosa. A mi hermano, por estar siempre ahí para ser molestado o molestarme cuando más lo necesitaba (aunque no lo supiera).

A mis amigos, o mejor dicho los de Berto, que me acompañaron en cada paso de esta carrera. Son lo más lindo que me llevo de estos años en la facultad.

A Tomás, por mantenerme cuerda y ser mi compañero de estudio, de tps y de todo. Gracias por estar siempre ahí para escucharme, ayudarme y apoyarme. No podría haberlo hecho sin vos.

*A mi abuelo Jorge, por ser mi fan número uno y siempre creer en mí.*

# Índice general

# 1. INTRODUCCIÓN

La enfermedad de Alzheimer (AD por sus siglas en inglés) es la forma más común de demencia (Omura y col. 2022) y uno de los trastornos neurodegenerativos con mayor prevalencia a nivel mundial, siendo la causa del 60-70 % de los casos de demencia (Simon y col. 2009). Este trastorno se caracteriza por una progresiva disminución de las habilidades cognitivas y funcionales, asociada con cambios en el tejido cerebral y la reducción de neurotransmisores como la acetilcolina (Knopman y col. 2021), lo que afecta negativamente la memoria, el lenguaje, la comprensión y el comportamiento. Según estimaciones, en 2020 aproximadamente 50 millones de personas en el mundo vivían con AD (Breijyeh y Karaman 2020), y este número podría duplicarse cada 20 años debido al envejecimiento de la población (Mayeux y Stern 2012).

El diagnóstico de AD suele ser complejo, ya que requiere la evaluación exhaustiva de un médico especialista basada en el historial clínico del paciente, observaciones de familiares y diversas pruebas neuropsicológicas, como el Mini-Mental State Examination (MMSE) (Arevalo-Rodriguez y col. 2015). Sin embargo, estas pruebas pueden ser poco confiables y requieren tiempo, recursos y personal especializado (Lee y col. 2022). Por lo tanto, es crucial desarrollar herramientas diagnósticas accesibles y eficientes que permitan la detección temprana de AD, ya que la intervención oportuna puede retrasar su progresión y mitigar los efectos negativos en los pacientes.

Entre los síntomas más destacados del AD en etapas iniciales se encuentran la pérdida de memoria y los problemas de comunicación, como alteraciones en la fluidez del habla, prosodia y narrativa (Knopman y col. 2021). Estos síntomas han motivado el estudio del habla y el lenguaje como herramientas potenciales para el diagnóstico temprano de la enfermedad. En los últimos años, se han intensificado los esfuerzos para desarrollar sistemas automáticos de detección de AD utilizando tecnologías de análisis del habla y el lenguaje, así como técnicas avanzadas de aprendizaje automático (Qi y col. 2023). Estos sistemas tienen el potencial de ofrecer una alternativa no invasiva, escalable y más económica frente a los métodos convencionales, contribuyendo significativamente a mejorar el diagnóstico y monitoreo de esta enfermedad.

A pesar de su potencial, la mayoría de los estudios en este área se han realizado utilizando únicamente datos en inglés (Yang y col. 2022; Wang y col. 2022; Warnita y col. 2018; Balagopalan y col. 2020), limitando su aplicabilidad en contextos multilingües y en países donde el inglés no es la lengua principal. En este sentido, es fundamental investigar y desarrollar sistemas de detección de AD en otros idiomas, que permitan ampliar el alcance y la generalización de estos modelos a nivel global. Algunos estudios recientes han explorado la detección de AD en otros idiomas, como el español europeo y latinoamericano (Pérez-Toro y col. 2023; Pérez-Toro y col. 2022; Melistas y col. 2023), el griego (Melistas y col. 2023), el alemán (Pérez-Toro y col. 2023), el húngaro (Gosztolya y col. 2019) y el mandarín (Chien y col. 2019), entre otros, mostrando resultados prometedores y abriendo nuevas oportunidades de investigación en este campo.

En particular, el español, como uno de los idiomas más hablados en el mundo (Julian 2020), representa un campo de interés significativo para el desarrollo de métodos de diagnóstico en idiomas diferentes al inglés. El trabajo de Pérez-Toro y col. (2022) aborda esta necesidad al combinar datos en inglés y español para la detección automática de AD.

Los autores investigan la posibilidad de discriminar AD, usando embeddings acústicos (Wav2Vec (Baevski y col. 2020)) y lingüísticos (BERT Devlin y col. (2019), Roberta (Liu y col. 2019)). En el trabajo se presentan dos experimentos. El primero consiste en la clasificación de Alzheimer vs. controles sanos, usando los embeddings acústicos y lingüísticos para entrenar distintos algoritmos de clasificación y comparar sus desempeños en ambos idiomas. En el segundo experimento se propone un método de transferencia de aprendizaje, en el cual utilizan datos en inglés para mejorar la discriminación de AD en datos en español. Los autores reportan que la información lingüística es más importante en inglés, donde alcanzaron un F1-score de 0,76, mientras que en el español es más relevante la información acústica, con un F1-score de 0,80. Al realizar la transferencia de aprendizaje, obtuvieron un F1-score de 0,85.

En este trabajo nos proponemos replicar y extender el experimento de Pérez-Toro y col. (2022) -al que de ahora en más llamaremos *trabajo original*- donde obtienen embeddings lingüísticos usando BERT y evalúan el desempeño de distintos clasificadores en la tarea de discriminar Alzheimer de controles sanos. Para esto, emplearemos las mismas bases de datos utilizadas en el trabajo original: Pitt Corpus (Becker y col. 1994) y Chile AD (Sanz y col. 2022). Estas consisten en transcripciones de audios de participantes describiendo una imagen, etiquetadas según el diagnóstico final del participante.

En el código provisto en el trabajo original pudimos observar que para la obtención de embeddings se usa la salida de la capa inicial de embeddings de BERT. Debido a eso, extenderemos el análisis comparando los resultados de la clasificación obtenidos en la replicación con los resultados de usar embeddings contextuales, provenientes de la última capa de BERT. Utilizaremos un clasificador que predice siempre la clase mayoritaria, lo que nos permitirá contar con un punto de referencia sencillo para comparar los resultados y evaluar la mejora obtenida en la tarea de clasificación con ambas representaciones textuales. Para evaluar la significancia de los resultados obtenidos con los distintos conjuntos de embeddings, realizaremos *tests de permutación*[1], comparando los resultados obtenidos con los distintos conjuntos de embeddings contra una distribución nula generada mediante permutaciones aleatorias de las etiquetas. Esto nos permitirá evaluar si las diferencias observadas en las métricas de clasificación con el punto de referencia son estadísticamente significativas o si podrían haber ocurrido por azar.

Además, profundizaremos en el análisis de evaluación del rendimiento de los clasificadores. Mientras que el trabajo original se centra exclusivamente en la métrica F1-score, nosotros incorporaremos otras métricas como precision, recall, accuracy y el área bajo las curvas ROC y Precision-Recall. En particular, nos interesa analizar si el F1-score es una métrica adecuada para este problema o si otras métricas ofrecen una mejor representación del desempeño de los clasificadores, reflejando con mayor precisión lo que realmente ocurre en los datos.

Por último, exploraremos el impacto de utilizar las variantes cased y uncased de los modelos de BERT para la tarea de detección de Alzheimer. Evaluaremos ambos enfoques en inglés y español, tanto con las representaciones de la primera capa de BERT como con las representaciones extraídas de la última capa, analizando cómo la elección de los modelos influye en los resultados de la clasificación.

[1] Este procedimiento se detalla en la Sección 3.2.

## 1.1. Trabajo previo

Los estudios sobre la detección automática del Alzheimer a partir del habla han aumentado en los últimos años y es un área de estudio de gran interés debido a la preocupación global causada por esta enfermedad.

En Warnita y col. (2018), proponen la utilización de una red neuronal convolucional con mecanismo de compuerta (GCNN) para la detección automática de AD. Las GCNN pueden ser entrenadas con pocos datos, en el orden de cientos de muestras, como es común en contextos clínicos donde la recolección de datos etiquetados resulta costosa o limitada. Esta arquitectura es especialmente adecuada para escenarios con datos limitados debido a su menor cantidad de parámetros en comparación con modelos recurrentes, y al uso de mecanismos de compuerta que permiten filtrar información irrelevante y mejorar la generalización. Usando embeddings acústicos extraídos del Pitt Corpus, el método captura información temporal relevante para la clasificación entre pacientes con AD y controles sanos. Los experimentos mostraron que la GCNN alcanzó una *accuracy* promedio del $73,6\,\%$. Es importante destacar que el conjunto de datos presenta un desbalance de clases, con un total de 488 audios (255 AD, 233 controles sanos) provenientes de 267 sujetos (169 AD, 98 controles sanos). Al no depender de información lingüística, este enfoque tiene la ventaja de ser aplicable a idiomas con pocos recursos.

En Balagopalan y col. (2020) los autores comparan dos enfoques principales para la detección del AD utilizando el conjunto de datos del ADReSS Challenge (Luz y col. 2020): uno basado en embeddings lingüísticos y acústicos diseñados manualmente, y otro utilizando BERT *fine-tuned* específicamente para esta tarea. El enfoque basado en embeddings manuales incluye 509 atributos extraídos de transcripciones y audios, divididos en léxico-sintácticos, embeddings acústicos y semánticos. El modelo de BERT *fine-tuned* da mejores resultados en la tarea de detección de AD que el basado en embeddings diseñados a mano.

El estudio de Wang y col. (2022) examina cómo la combinación de diferentes atributos y modelos puede fortalecer la robustez de los codificadores de texto preentrenados, como BERT y Roberta, mediante *fine tuning*, especialmente en escenarios con datos limitados. Los embeddings generados se evalúan utilizando varios clasificadores, y la decisión final se hace a través de votación mayoritaria. Los experimentos, realizados con el conjunto de datos del ADReSS Challenge, demostraron mejoras consistentes en el rendimiento gracias a estas combinaciones. En particular, se alcanzaron *accuracy* de $91{,}67\,\%$ y $93{,}75\,\%$ al trabajar con transcripciones manuales y automáticas, respectivamente. Este último resultado representa el mejor desempeño reportado hasta ahora para transcripciones generadas automáticamente por sistemas de reconocimiento de voz.

Muchos de los avances en este campo han sido presentados en el contexto del *Interspeech*, la conferencia líder a nivel mundial en ciencia y tecnología del procesamiento del lenguaje hablado. En el marco de los *challenges* presentados en esta conferencia en los años 2020, 2021 y 2023, respectivamente, se presentaron dos de los datasets más utilizados para la detección automática de AD: ADReSS y ADReSSo (Luz y col. 2021). El challenge ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) de 2020 propuso la tarea de clasificación binaria entre pacientes con AD y controles sanos, a partir de grabaciones del test del Desenlace del Picnic del DementiaBank Pitt Corpus, cuidadosamente balanceadas por edad y género. El desafío ADReSSo (AD Recognition through Spontaneous Speech only), introducido en 2021, planteó un escenario aún más realista y exigente: detección de AD y regresión de la puntuación MMSE, utilizando exclusivamente

señales de audio (sin transcripciones), con datos también derivados del Pitt Corpus pero con un enfoque más centrado en condiciones naturales y sin alineamiento perfecto entre texto y audio.

Uno de los trabajos presentados en el Interspeech de 2022 es Yang y col. (2022). Los autores proponen un enfoque innovador para abordar el problema de la detección de AD en etapas tempranas utilizando datos de habla limitados: combinan aprendizaje autosupervisado, adversarial y aumentado, capaces de capturar patrones característicos del AD en un espacio latente, aprovechando una gran cantidad de datos de habla normal fácilmente obtenibles y un conjunto de datos aumentados de pacientes con AD. El método alcanzó un F1-score de $83,15\,\%$, destacándose por su capacidad para manejar la escasez de datos y mejorar la precisión en la detección de patrones asociados al AD.

En el Interspeech de 2023 se presentaron varios trabajos relacionados a la detección automática de AD, entre ellos Melistas y col. (2023), donde se explora un enfoque multilingüe para la detección de la enfermedad basado en atributos extraídos del habla e interacciones conversacionales. Los autores evalúan en escenarios de *zero-shot* (sin datos del idioma objetivo durante el entrenamiento) y *few-shot* (con datos limitados del idioma objetivo), para poner a prueba el desempeño del modelo en distintos idiomas. Los experimentos realizados utilizaron datos en inglés, griego y español, destacándose por incorporar atributos independientes del lenguaje, como señales paralingüísticas, dinámicas de interacción entre hablantes y metadatos demográficos (edad, género, años de educación, etc.). Entrenando en inglés y evaluadas en griego obtuvieron un *accuracy* de $78{,}3\,\%$, superando en un $4{,}4\,\%$ al *baseline* propuesto por los organizadores del *challenge*. En español, los resultados mostraron que el enfoque mantiene un rendimiento competitivo incluso con pocos datos de entrenamiento.

En Pérez-Toro y col. (2023), también presentado en el Interspeech de 2023, los autores proponen un análisis multilingüe para la detección automática del AD empleando atributos acústicos (duración, ritmo, *pleasure-arousal-dominance* (PAD) y embeddings de Wav2Vec) y lingüísticos (basados en gramática y embeddings de BERT). El estudio considera datos en inglés, español y alemán, y explora dos enfoques de clasificación: entrenamiento y prueba en uno o más idiomas, y entrenamiento en un idioma y prueba en otro. En el primer enfoque, los mejores resultados de clasificación fueron obtenidos realizando la detección automática de AD en cada idioma (sin mezclar). En cambio, en el caso del segundo enfoque, los resultados muestran que la información dada por los atributos no es transferible entre idiomas.

En Qi y col. (2023) se presenta un resumen detallado de los trabajos realizados en los últimos años. Describen las tecnologías utilizadas para detectar AD a partir del habla, incluyendo bases de datos y herramientas usadas para la extracción de *features* y para la clasificación. Además, incluyen estrategias de optimización para sistemas de detección de AD y abordan desafíos relacionados con el tamaño de los datasets, la explicabilidad de los modelos, entre otros. En particular, destacan que la escasez de datos especializados dificulta el entrenamiento de modelos robustos, dado que recolectar grabaciones transcritas de pacientes con AD requiere la participación de profesionales clínicos y es costoso en tiempo y recursos. Por otro lado, subrayan que muchos modelos basados en redes neuronales profundas funcionan como "cajas negras", lo que complica la interpretación de sus decisiones y limita su adopción en entornos clínicos donde la trazabilidad y la comprensión del diagnóstico son fundamentales.

## 1.2. Estructura de la tesis

Este trabajo se compone de cuatro capítulos, incluido el presente, y cuatro apéndices.

En el Capítulo 2, describimos las bases de datos utilizadas en la experimentación, explicamos el funcionamiento de BERT —el modelo empleado para la obtención de embeddings a partir de los textos—, detallamos el proceso de replicación del estudio de Pérez-Toro y col. (2022) y presentamos el método utilizado para extraer embeddings contextuales de los textos.

En el Capítulo 3, mostramos los resultados de la replicación del trabajo original y ampliamos el análisis comparando estos resultados con aquellos obtenidos mediante embeddings contextuales. Además, incorporamos métricas de evaluación adicionales para determinar cuál es la más adecuada para este problema. Por último, analizamos el impacto de utilizar modelos de BERT entrenados con texto capitalizado (cased) frente a modelos entrenados únicamente con texto en minúsculas (uncased).

En el Capítulo 4, sintetizamos los hallazgos obtenidos a lo largo de la tesis, destacamos las principales contribuciones del estudio y planteamos posibles líneas de investigación futura.

En el Apéndice A, presentamos las tablas con los hiperparámetros óptimos encontrados para cada algoritmo de clasificación, con cada tipo de embeddings, junto con mapas de calor que ilustran los valores de accuracy obtenidos para cada configuración evaluada.

En el Apéndice B, mostramos *box plots* con los resultados de 10 ejecuciones de tests de permutación, realizados para evaluar la significancia estadística de los resultados obtenidos con los distintos tipos de embeddings, tanto en inglés como en español.

En el Apéndice C, ampliamos el análisis de los modelos BERT entrenados sin capitalización (uncased), presentando los valores obtenidos para todas las métricas consideradas en el estudio.

En el Apéndice D, exploramos en mayor profundidad las diferencias entre las distintas representaciones textuales, visualizando cómo se distribuyen conjuntos de palabras en ambos idiomas dentro de un espacio bidimensional.

# 2. MATERIALES Y MÉTODOS

## 2.1. Bases de Datos

En esta sección describiremos las dos bases de datos utilizadas para realizar los experimentos. En ambos casos, los datos consisten en transcripciones de grabaciones de audio de entrevistas a participantes, en las que se les pidió que describieran una imagen. Las transcripciones están etiquetadas como Alzheimer (AD) o controles sanos (CTR), según el diagnóstico final del paciente.

### Pitt Corpus

El Pitt Corpus del DementiaBank (Becker y col. 1994) contiene audios y transcripciones de hablantes nativos de inglés (Estados Unidos), recolectados por la Universidad de Pittsburgh. Esta es una de las bases de datos en inglés más utilizadas para la tarea de detección de Alzheimer a partir del habla.

Los participantes fueron entrevistados como máximo una vez al año durante un período de cinco años, y en cada sesión se les solicitó que describieran todo lo que veían en la imagen conocida como *"cookie theft picture"* (Figura 2.1). Las respuestas de los participantes fueron grabadas y transcritas textualmente. En total, la base de datos contiene 309 grabaciones/transcripciones de 194 participantes con AD y 243 grabaciones/transcripciones de 99 controles sanos.

Los audios de los participantes con AD tienen una duración promedio de $76 \pm 35$ segundos, mientras que los audios de los participantes etiquetados como CTR tienen una duración promedio de $63 \pm 24$ segundos, en ambos casos incluyendo las intervenciones del entrevistador.

### Chile AD

Esta base de datos fue recolectada por la Universidad de Chile y el Hospital del Salvador (Chile) (Sanz y col. 2022). Contiene audios y transcripciones de 39 hablantes nativos de la variante chilena del español, a quienes se les pidió que describieran la imagen *"cookie theft picture"*.

Los participantes fueron evaluados y diagnosticados por neurólogos, siguiendo los criterios clínicos del *National Institute of Neurological and Communicative Disorders and Stroke - Alzheimer's Disease and Related Disorders Association*. De los 39 participantes, 21 fueron etiquetados con Alzheimer y 18 como controles sanos. Se seleccionaron participantes con características similares en cuanto a sexo, edad y nivel educativo. En la Tabla 2.4 se detallan las características demográficas de los participantes de Chile AD.

Los audios fueron segmentados y transcritos manualmente, excluyendo las intervenciones del entrevistador. Eliminando el habla del entrevistador, la duración promedio de los audios es de $84 \pm 29$ segundos para los participantes con AD y de $73 \pm 15$ segundos para los participantes etiquetados como CTR. Las transcripciones se encuentran en minúsculas y sin signos de puntuación.

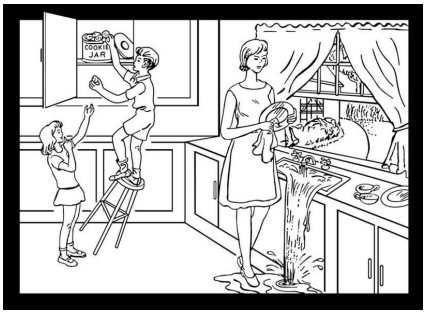



Fig. 2.1: *"cookie theft picture"* (Goodglass y col. 1983)

## 2.2. BERT

En esta tesis buscamos replicar el experimento de Pérez-Toro y col. (2022) en el que los autores emplearon BERT (Devlin y col. 2019) (*Bidirectional Encoder Representations from Transformers*), un modelo que fue entrenado para aprender a representar texto como una secuencia de vectores, usando aprendizaje autosupervisado. Sin embargo, un aspecto metodológico particular de dicho trabajo es que el código provisto para la obtención de embeddings utiliza las salidas de la capa inicial de embeddings de BERT, en lugar de las salidas de la última capa, que capturan información contextual más profunda.

El propósito de esta sección es analizar las diferencias fundamentales entre estas dos representaciones textuales. Esto resulta especialmente relevante ya que se realizará una comparación directa entre los resultados obtenidos con las representaciones de la capa de embeddings y los obtenidos con las representaciones de la última capa de BERT.

BERT fue entrenado en varios corpus de texto no etiquetado, como English Wikipedia y BookCorpus, utilizando dos tareas de aprendizaje supervisado:

- **Masked Language Model (MLM)**: Se enmascaran aleatoriamente algunas palabras en la entrada y se entrena al modelo para predecir las palabras enmascaradas.
- **Next Sentence Prediction (NSP)**: El modelo recibe dos oraciones y aprende a predecir si la segunda oración sigue directamente a la primera en el texto original. Es una tarea de clasificación binaria.

Estas tareas permiten que BERT funcione tanto a nivel de palabra como a nivel de oración, lo que lo hace adecuado para una amplia variedad de tareas de procesamiento de lenguaje natural.

Además, BERT cuenta con variantes cased y uncased, que difieren en cómo manejan la distinción entre mayúsculas y minúsculas. Los modelos cased preservan esta diferencia, lo que puede ser útil en tareas donde la capitalización aporta significado, como el reconocimiento de entidades nombradas o la desambiguación de términos (por ejemplo, en español, "Tierra" no es lo mismo que "tierra"). En contraste, los modelos uncased convierten todo el texto a minúsculas antes de ser procesado, lo que reduce el tamaño del vocabulario y elimina posibles variaciones causadas por la capitalización. La diferencia entre los modelos cased y uncased de BERT no solo radica en la normalización del texto, sino también en su entrenamiento. BERT cased se entrena con texto que mantiene las mayúsculas originales, lo que le permite capturar mejor la información que depende de la capitalización. En cambio, BERT uncased es entrenado con texto completamente en minúsculas, lo que lo

hace más robusto frente a variaciones en la escritura, pero menos preciso en tareas donde la capitalización es clave.

Para entender cómo se pueden obtener embeddings a partir de BERT, es fundamental comprender su arquitectura. En este trabajo utilizaremos la versión BASE de BERT. En la Figura 2.2 se ilustra el proceso de generación de embeddings con BERT-BASE, donde las secuencias de palabras se transforman en representaciones densas a través de múltiples etapas:

1. **Tokenizador**
   Divide el texto en subpalabras utilizando WordPiece (Wu y col. 2016), un algoritmo de tokenización que divide las palabras en unidades más pequeñas (por ejemplo, `overflowing` → `[over, ##flow, ##ing]`). Luego, mapea los tokens a identificadores únicos: números enteros que corresponden al índice del token en el vocabulario de WordPiece (por ejemplo, el token `##happy` corresponde al índice 5423).

   Además, el tokenizador reemplaza cualquier token que no esté en su vocabulario por `[UNK]` ("desconocido") y añade tokens especiales: `[CLS]` al principio de la secuencia para representar la entrada completa y `[SEP]` para marcar el final o separar partes de la secuencia.

2. **Capa de Embeddings**
   La capa de embeddings en BERT se encarga de transformar los tokens de entrada en representaciones vectoriales que puedan ser procesadas por las capas posteriores. Este proceso consta de tres componentes principales: embeddings de tokens, embeddings de posición y embeddings de segmento, que se combinan para generar una representación inicial rica en información contextual.

   Para obtener los **embeddings de token**, cada ID generado por el tokenizador se mapea a un vector único. Estos vectores provienen de un vocabulario preentrenado que contiene representaciones para cada token. Por ejemplo, el token `##happy` corresponde al ID 5423, y, a su vez, el ID 5423 corresponde al vector $[0{,}7, 0{,}4, -0{,}2, ..., -0{,}6]$ de 768 dimensiones, en el caso de BERT-base.

   Los modelos como BERT no procesan secuencias de manera estrictamente secuencial, por lo que necesitan información sobre el orden de las palabras. Cada posición en la secuencia de entrada tiene un vector asociado que indica su lugar relativo en la oración, llamado **embedding de posición**.

   BERT es preentrenado para manejar pares de oraciones (por ejemplo, para tareas como clasificación de relaciones entre frases). Para distinguir los diferentes segmentos, el modelo incluye **embeddings de segmento**, tales que, si la entrada contiene un segmento A y un segmento B, todos los tokens del segmento A reciben un vector correspondiente a ese segmento, y los del segmento B reciben otro, correspondiente al segmento B. En el caso de que haya un solo segmento, todos los tokens tendrán el mismo embedding de segmento.

   Una vez generados los embeddings de tokens, posición y segmento, estos se suman elemento a elemento para formar un único vector de dimensión fija para cada token. Esta combinación encapsula información sobre el significado del token, su posición en la oración y a qué segmento pertenece.

La salida de esta capa es una matriz de dimensiones `(n_tokens, d_model)`, donde `n_tokens` es el número de tokens en la secuencia (incluidos `[CLS]` y `[SEP]`) y `d_model` es la dimensión fija del embedding (768 en BERT base).

3. **Encoder**
   Luego de la capa de embeddings, BERT tiene una pila de encoders. En el modelo BERT-base, esta pila consta de 12 capas de codificador, cada una de las cuales es una unidad independiente pero interconectada que procesa las representaciones del texto para capturar relaciones contextuales y semánticas más profundas entre las palabras.

   Cada capa de encoder tiene primero un mecanismo de *self-attention*, donde cada palabra representada por su embedding interactúa con todas las demás palabras en la secuencia de entrada. Esto permite que cada token ajuste su representación según su relación con otros tokens en el texto. Por ejemplo, en una oración ambigua, como "el banco está cerca del centro", donde la palabra "banco" puede referirse tanto a una institución financiera como a un asiento, el mecanismo de *self-attention* ayuda a una palabra a entender su contexto considerando las palabras previas y posteriores. Esto significa que, al procesar la palabra "banco", el mecanismo de *self-attention* asignará mayor peso a palabras como "cerca" y "del centro", ya que estas aportan pistas contextuales que pueden ayudar a desambiguar el significado. Por ejemplo, si en el contexto previo o posterior hay referencias relacionadas con dinero, cuentas o transacciones, el modelo ajustará el embedding de "banco" para que represente una institución financiera. En cambio, si el contexto incluye palabras como "sentarse" o "descansar", el embedding reflejará que se refiere a un asiento.

   De esta forma, el mecanismo de *self-attention* no solo captura relaciones locales entre palabras adyacentes, sino también dependencias globales en toda la oración, lo que le permite al modelo construir representaciones precisas y contextualmente relevantes para cada token.

   En BERT-base, hay 12 cabezas de atención que trabajan en paralelo, cada una enfocándose en diferentes aspectos de las relaciones entre palabras.

   Una vez que los embeddings son ajustados por el mecanismo de atención, la salida se combina con la entrada original de la capa (es decir, la salida de la capa de embeddings) mediante una conexión residual, una técnica introducida en He y col. (2015) que se utiliza para evitar el problema conocido como *vanishing gradients*. Este paso asegura que la información original no se pierda y que los embeddings mantengan características clave de la entrada inicial. Después de esta combinación, los valores resultantes se normalizan mediante una capa de normalización. Esto estabiliza las representaciones, facilitando el entrenamiento del modelo y evitando que los valores se desestabilicen durante las múltiples iteraciones.

   Luego, los embeddings pasan por una red neuronal *feed-forward* que refina aún más sus representaciones. Esta red consta de dos capas: una expansión lineal, que aumenta temporalmente la dimensión del embedding para capturar características más complejas, seguida de una reducción que lo devuelve a su dimensión original. Este paso añade capacidades no lineales a las transformaciones, enriqueciendo la información capturada por los embeddings. Al igual que en el paso anterior, se aplica una conexión residual y normalización para mantener la estabilidad y coherencia de las

representaciones.

La salida final de una capa de codificador sirve como entrada para la siguiente capa. Este proceso se repite a lo largo de las 12 capas del codificador en BERT base. En cada capa, los embeddings se refinan progresivamente, capturando relaciones más profundas y complejas entre los tokens del texto. Al final de este proceso, cada token tiene una representación contextualizada que incorpora tanto las relaciones locales (con palabras cercanas) como las globales (con palabras más lejanas en la secuencia).

Al igual que en la capa de embeddings, la salida es una matriz de dimensiones `(n_tokens, d_model)`.

A lo largo de esta tesis será necesario tener presente la arquitectura de BERT y cómo transforma los tokens a lo largo de sus capas. La diferencia entre la salida de la capa de embeddings y la salida final del codificador radica en el nivel de información que encapsulan: mientras que los embeddings iniciales representan una combinación estática de información léxica, posicional y segmental, la salida final del codificador proporciona representaciones profundamente contextualizadas que incorporan relaciones complejas entre las palabras en un contexto bidireccional.

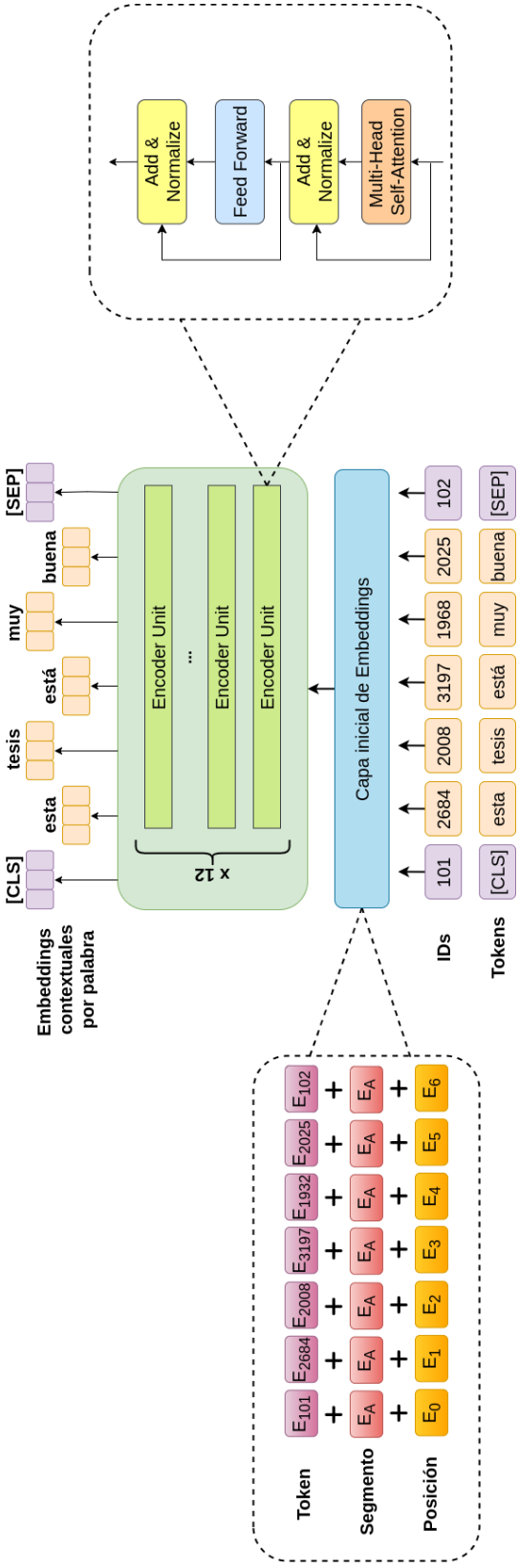



Fig. 2.2: Arquitectura de BERT

## 2.3. Implementación

En esta sección describiremos el proceso de replicación del trabajo original. Durante el desarrollo, realizamos ciertos ajustes y tomamos decisiones metodológicas en base a la información disponible, ya que el trabajo original no describe completamente su implementación.

### Detalles implementativos reportados

En Pérez-Toro y col. (2022), utilizan transcripciones provenientes de dos bases de datos, una en español y otra en inglés, etiquetadas según el diagnóstico final del participante (AD vs. CTR):

- **Pitt Corpus** (inglés): Emplean un subconjunto de 186 hablantes (93 con Alzheimer y 93 controles). Sin embargo, desconocemos el criterio de selección de los participantes y el número de visita utilizado.
- **Chile AD** (español): Usan todos los datos disponibles (21 participantes con Alzheimer y 18 controles).

En ambos casos, las intervenciones de los entrevistadores fueron eliminadas. No obstante, no tenemos detalles adicionales sobre el preprocesamiento del texto.

Para obtener las representaciones lingüísticas utilizan BERT. El código utilizado está disponible públicamente (Perez-Toro 2020). En adelante, nos referimos a los embeddings obtenidos con esta metodología como BERT-FIRST, ya que las representaciones provienen de la primera capa del codificador. Para obtener las representaciones de los datos del Pitt Corpus usan la versión de BERT preentrenada exclusivamente en inglés y, para Chile AD, emplean la versión multilingüe, preentrenada en 104 idiomas, incluyendo el español. Los embeddings se obtienen a partir de la salida de la capa inicial de embeddings de BERT. Para generar una única representación por participante, promedian los embeddings de las palabras dichas por cada participante, resultando en un vector de 768 dimensiones por participante.

Los algoritmos de aprendizaje automático que reportan haber entrenado para la tarea de clasificación son Support Vector Machine (SVM) (Cortes y Vapnik 1995) con kernel radial, XGBoost (XGB) (Chen y Guestrin 2016) y Multi-Layer Perceptron (MLP) (Rumelhart y col. 1986) con una capa de entrada de 768 neuronas y, suponemos, una única neurona de salida, con la función de pérdida sigmoide. Dado que la capa de entrada de 768 neuronas es equivalente a la dimensión de los embeddings de BERT, que van a ser la entrada del modelo en este caso, podemos decir que MLP es equivalente a Logistic Regression (LR) (Cox 1958).

Para la búsqueda de hiperparámetros utilizaron *grid search*. En la Tabla 2.1 se detallan las configuraciones exploradas que reportaron en el trabajo original. No especifican los valores óptimos obtenidos para cada clasificador.

La evaluación de los clasificadores la llevaron a cabo mediante *4-fold cross-validation*, con *folds* definidos con muestras al azar y estratificados según la cantidad de participantes etiquetados con Alzheimer y controles sanos. Para incrementar la robustez de los resultados, el experimento lo repitieron 10 veces. Reportan el promedio y la desviación estándar de los resultados utilizando F1-score como métrica de evaluación, los cuales se presentan en la Tabla 2.2.

| Clasificador | Hiperparámetros |
|---|---|
| **SVM** | $C \in \{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$<br>$\gamma \in \{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$ |
| **XGB** | Número de estimadores $\in \{50, 70, \ldots, 300\}$<br>Profundidad máxima $\in \{1, 3, 5, 7\}$ |

Tab. 2.1: Grillas de hiperparámetros reportadas en el trabajo original.

| Idioma | Modelo de BERT | SVM | XGBoost | LR |
|---|---|---|---|---|
| Inglés | BERT-ENG | 0.68 (0.01) | 0.69 (0.02) | 0.68 (0.01) |
| Español | BERT-MUL | 0.53 (0.08) | 0.46 (0.14) | 0.51 (0.10) |

Tab. 2.2: Resultados reportados en el trabajo original, promedio (desviación estándar) de F1-score para los embeddings de BERT. BERT-ENG corresponde al modelo de BERT entrenado exclusivamente en inglés, mientras que BERT-MUL hace referencia al modelo de BERT entrenado en más de 100 idiomas, incluyendo al español.

### Decisiones tomadas en este trabajo

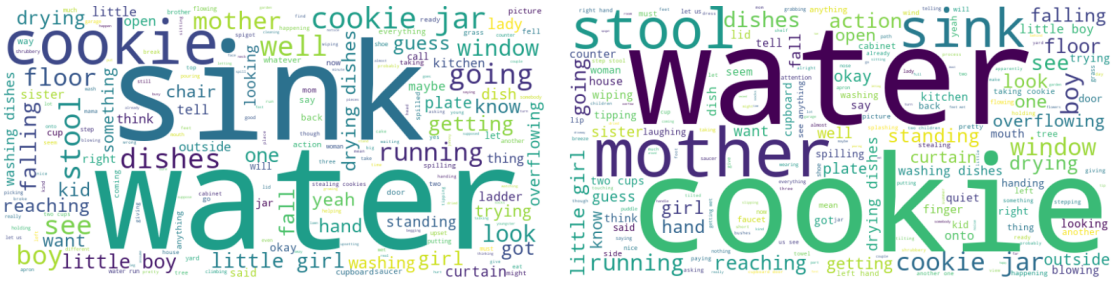
En el caso del Pitt Corpus, utilizamos únicamente los datos correspondientes a la primera visita de cada participante. El corpus incluye 99 individuos etiquetados como controles sanos y 194 diagnosticados con Alzheimer. Dado que en el trabajo original las clases están balanceadas, seleccionamos aleatoriamente un subconjunto de 99 participantes con Alzheimer. Esto resultó en un total de 198 muestras balanceadas (99 CTR y 99 AD).

Las transcripciones originales de Pitt están en un formato estándar, diseñado para producir transcripciones computarizadas de interacciones conversacionales cara a cara. Realizamos un proceso de limpieza de los textos, el cual incluyó: (a) eliminación de intervenciones del entrevistador, (b) eliminación de signos de puntuación, (c) eliminación de palabras no pertenecientes al idioma inglés, (d) eliminación de muletillas (por ejemplo: mmh, ooh), (e) eliminación de anotaciones fonéticas, (f) eliminación de marcadores de eventos no verbales (por ejemplo, `<cough>`, `<laugh>`), (g) conversión del texto a minúsculas y (h) expansión de contracciones comunes (por ejemplo, *"isn't"* a *"is not"*).

Como resultado, obtuvimos una base de datos final compuesta por 198 textos en inglés, en minúsculas y sin signos de puntuación, con la misma cantidad de sujetos en ambas clases. En la Tabla 2.3 se presentan las características demográficas de los participantes del Pitt Corpus. En la Figura 2.3 se muestran las nubes de palabras más frecuentes en las transcripciones de los participantes con Alzheimer y controles sanos.

| Condición | Género | Cantidad | Edad ($\bar{\mu} \pm$ std) | Educación ($\bar{\mu} \pm$ std) |
|---|---|---|---|---|
| AD | Mujeres | 64 | $73{,}0 \pm 8{,}6$ | $11{,}1 \pm 2{,}2$ |
| | Hombres | 35 | $69{,}3 \pm 8{,}2$ | $12{,}8 \pm 2{,}9$ |
| CTR | Mujeres | 58 | $63{,}2 \pm 8{,}0$ | $14{,}0 \pm 2{,}5$ |
| | Hombres | 41 | $64{,}2 \pm 8{,}9$ | $13{,}7 \pm 2{,}4$ |

Tab. 2.3: Características demográficas de los participantes del Pitt Corpus. La edad y la educación se expresan en años.




(a) (b)

Fig. 2.3: Nube de palabras más frecuentes en las transcripciones de los participantes con Alzheimer (a) y controles sanos (b) del Pitt Corpus.

Las transcripciones de Chile AD son textos en minúsculas, sin signos de puntuación ni caracteres especiales, y sin intervenciones del entrevistador. La base de datos Chile AD cuenta con una única transcripción por participante y las clases están levemente desbalanceadas: 54 % de los participantes están etiquetados como AD y 46 % como CTR. Como en el trabajo original utilizaron todos los datos disponibles, y teniendo en cuenta que son muy pocos, decidimos no balancear las clases. En la Tabla 2.4 se presentan las características demográficas de los participantes de Chile AD. En la Figura 2.4 se muestran las nubes de palabras más frecuentes en las transcripciones de los participantes con Alzheimer y controles sanos de Chile AD.

| Condición | Género | Cantidad | Edad ($\bar{\mu} \pm$ std) | Educación ($\bar{\mu} \pm$ std) |
|---|---|---|---|---|
| AD | Mujeres | 13 | $79{,}1 \pm 6{,}8$ | $11{,}1 \pm 3{,}9$ |
| | Hombres | 8 | $77{,}0 \pm 7{,}1$ | $11{,}4 \pm 3{,}7$ |
| CTR | Mujeres | 14 | $74{,}5 \pm 1{,}7$ | $7{,}5 \pm 3{,}6$ |
| | Hombres | 4 | $76{,}4 \pm 4{,}6$ | $14{,}4 \pm 3{,}1$ |

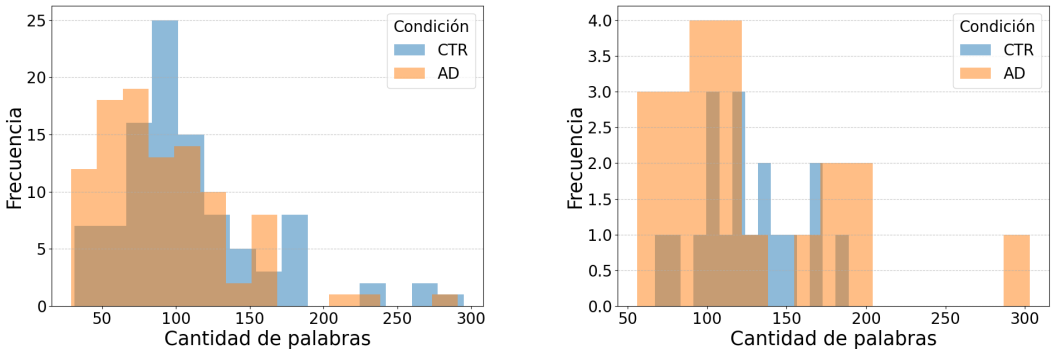
Tab. 2.4: Características demográficas de los participantes de Chile AD. La edad y la educación se expresan en años.


(a) (b)

Fig. 2.4: Nube de palabras más frecuentes en las transcripciones de los participantes con Alzheimer (a) y controles sanos (b) de Chile AD.

En ambas bases de datos, cada transcripción corresponde a un participante distinto, es decir, no hay dos instancias distintas que provengan de un mismo hablante.

El Pitt Corpus cuenta con un total de 1000 palabras únicas entre todas sus transcripciones, y la longitud promedio de las transcripciones es de $100{,}23 \pm 48{,}34$ palabras. Por su parte, en Chile AD se identificaron un total de 673 palabras únicas en las transcripciones, con una longitud promedio de $125{,}05 \pm 47{,}69$ palabras. En la Figura 2.5 se muestran histogramas de la cantidad de palabras por transcripción en cada base de datos.




(a) (b)

Fig. 2.5: Histogramas de la cantidad de palabras por transcripción en el Pitt Corpus (a) y Chile AD (b).

Es importante destacar que no se utilizó un *Hold-Out set*, es decir, una porción del conjunto de datos reservada exclusivamente para la evaluación final del modelo. Esta decisión se tomó por dos razones principales: en primer lugar, el trabajo original que replicamos tampoco empleó un conjunto de test separado; en segundo lugar, los datasets disponibles contienen una cantidad muy limitada de muestras, lo que hace que separar una parte para testeo implique perder datos valiosos para el entrenamiento y la validación.

En contextos con pocos datos, usar un *Hold-Out set* puede llevar a estimaciones inestables del rendimiento de los clasificadores, ya que la muestra de test es pequeña y poco representativa. Por eso, se opta por estrategias como la validación cruzada repetida, que permite aprovechar al máximo el conjunto disponible y obtener métricas más robustas. Sin embargo, la ausencia de un *Hold-Out set* también representa una desventaja importante: no se puede estimar de forma completamente imparcial el rendimiento de los algoritmos de clasificación sobre datos no vistos, lo que podría llevar a una sobrestimación de los resultados.

En la Figura 2.6 se puede ver el pipeline que utilizamos para replicar el trabajo original.

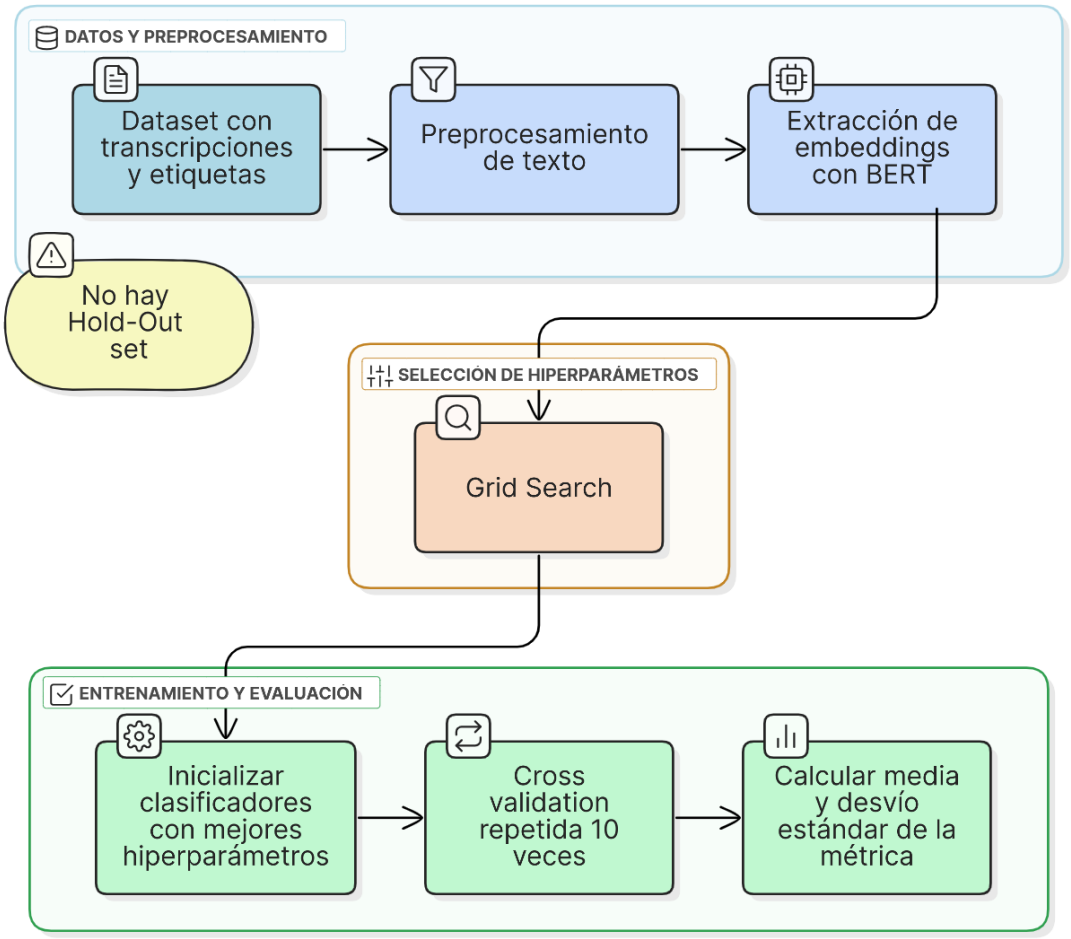



Fig. 2.6: Pipeline utilizado para replicar el primer experimento del trabajo original. El proceso comienza con un dataset de transcripciones etiquetadas, que se somete a preprocesamiento textual y extracción de embeddings utilizando BERT. No se utilizó conjunto Hold-Out. La selección de hiperparámetros se realiza mediante *Grid Search*, y los clasificadores se evalúan con validación cruzada repetida 10 veces. Finalmente, se calcula la media y el desvío estándar de la métrica de evaluación.

Para obtener los embeddings lingüísticos a partir de las transcripciones, utilizamos el código proporcionado en el trabajo original (BERT-FIRST). Usamos los modelos `bert-base-cased` para las transcripciones en inglés y `bert-base-multilingual-cased` (Devlin y col. 2018) para las transcripciones en español.

Para la clasificación empleamos los mismos algoritmos de aprendizaje automático que reportaron haber usado en el trabajo original: SVM con kernel radial, XGBoost y reemplazamos MLP por Logistic Regression dada su equivalencia funcional en este caso. Además, agregamos Random Forest (RF) (Breiman 2001) como clasificador adicional, para explorar su desempeño en este problema.

Las configuraciones de hiperparámetros exploradas para SVM y XGBoost fueron las mismas que en el trabajo original (2.1). Para Logistic Regression y Random Forest, utilizamos los valores predeterminados de la librería `scikit-learn` (Pedregosa y col. 2011),

con la única excepción de que en Random Forest fijamos el número de estimadores en 200.

Para la búsqueda de los mejores hiperparámetros utilizamos *grid search*, donde entrenamos y evaluamos los clasificadores con todas las combinaciones posibles de hiperparámetros de la Tabla 2.1, usando todos los datos.

Para la evaluación utilizamos *k-fold cross-validation* con $k = 4$. Este proceso consiste en dividir el conjunto de datos en 4 subconjuntos o *folds*. En cada iteración, se usa uno de estos folds como conjunto de validación, mientras que los restantes se emplean para el entrenamiento del clasificador. Este procedimiento se repite 4 veces, asegurando que cada subconjunto sea utilizado una vez para la validación.

Con el fin de que los conjuntos de entrenamiento y validación estén balanceados en cuanto a la proporción de participantes con y sin AD, utilizamos *folds* estratificados (preservan el porcentaje de muestras para cada clase).

Nuestro método de evaluación consiste en almacenar las predicciones generadas en cada *fold*. Una vez finalizadas todas las iteraciones de la validación cruzada, se obtiene una predicción por cada participante en el conjunto de datos. Utilizamos estas predicciones, junto con las etiquetas reales, para calcular la métrica deseada. En el caso del *grid search*, para cada clasificador seleccionamos una de las configuraciones de hiperparámetros que maximizó la métrica accuracy. En el Anexo A, presentamos mapas de calor para los valores de accuracy obtenidos para cada combinación de hiperparámetros (Figura A.1), junto con los resultados de la búsqueda de hiperparámetros (Tabla A.1).

Tras finalizar el *grid search*, entrenamos los clasificadores inicializados con los mejores hiperparámetros encontrados para cada uno. Para evaluarlos, realizamos una nueva validación cruzada de 4 *folds*. Siguiendo la misma metodología que la reportada en el trabajo original, realizamos 10 repeticiones de la validación cruzada para cada clasificador, utilizando diferentes semillas en la partición de los datos. Como métrica de evaluación, utilizamos F1-score, ya que es la métrica reportada en el trabajo original. Para facilitar la comparación de los resultados con los reportados, calculamos el promedio de los F1-scores obtenidos en las 10 repeticiones para cada clasificador, así como su desviación estándar.

## 2.4. Embeddings contextuales

Los embeddings BERT-FIRST son la salida de la capa inicial de embeddings de BERT. Para más adelante realizar una comparación entre representaciones, obtuvimos embeddings provenientes de la última capa de BERT. De ahora en más, llamaremos BERT-LAST a los embeddings obtenidos con esta metodología.

Estos embeddings contienen representaciones contextualizadas de cada *token*, integrando información tanto semántica como sintáctica del texto. Para reducir esta representación a un único vector por transcripción, calculamos el promedio de los embeddings de todos los *tokens* de la secuencia, lo que resulta en un vector de 768 dimensiones por texto.

Al igual que en el proceso de replicación, realizamos *grid search* para obtener los mejores hiperparámetros, y entrenamos los clasificadores inicializados con las mejores configuraciones encontradas. Para evaluar el desempeño de los clasificadores, corrimos *4-fold cross validation* 10 veces para cada clasificador, con distintas separaciones de los datos.

Las configuraciones exploradas en el *grid search* están en la Tabla 2.1 y en el Anexo A se encuentran los mapas de calor con los valores de accuracy (Figura A.2) obtenidos para cada combinación de hiperparámetros, y los mejores hiperparámetros obtenidos para cada clasificador (Tabla A.2).

A partir de ahora, denominaremos BERT-FIRST-ENG-CASED y BERT-LAST-ENG-CASED a los embeddings BERT-FIRST y BERT-LAST, respectivamente, obtenidos a partir del modelo `bert-base-cased` de BERT. De la misma forma, llamaremos BERT-FIRST-ENG-UNCASED y BERT-LAST-ENG-UNCASED a los embeddings obtenidos a partir del modelo `bert-base-uncased`. A los embeddings obtenidos a partir del modelo `bert-base-multilingual-cased` los denominaremos BERT-FIRST-MUL-CASED y BERT-LAST-MUL-CASED para BERT-FIRST y BERT-LAST, respectivamente. De manera análoga, llamaremos BERT-FIRST-MUL-UNCASED y BERT-LAST-MUL-UNCASED a los embeddings obtenidos a partir del modelo `bert-base-multilingual-uncased`.

# 3. EXPERIMENTACIÓN

## 3.1. Resultados de la replicación

La Tabla 3.1 presenta los F1-scores promedio y las desviaciones obtenidas al replicar el trabajo original, junto con los resultados reportados en este.

| **Clasificador** | **Resultados replicados** | | **Resultados reportados** | |
|---|---|---|---|---|
| | **Pitt Corpus** | **Chile AD** | **Pitt Corpus** | **Chile AD** |
| SVM | 0.75 (0.02) | **0.70 (0.00)** | 0.68 (0.01) | **0.53 (0.08)** |
| XGBoost | 0.72 (0.03) | 0.60 (0.07) | **0.69 (0.02)** | 0.46 (0.14) |
| Logistic Regression | 0.75 (0.02) | 0.57 (0.04) | 0.68 (0.01) | 0.51 (0.10) |
| Random Forest | **0.76 (0.02)** | 0.51 (0.09) | - | - |

Tab. 3.1: Comparación de resultados replicados con los reportados en el trabajo original. Promedio y desvío estándar del F1-score. Los resultados replicados corresponden a los obtenidos con los embeddings BERT-FIRST-ENG-CASED para el Pitt Corpus y BERT-FIRST-MUL-CASED para Chile AD. Se resaltan los valores de F1-Score más altos obtenidos en cada idioma.

Los resultados replicados muestran un F1-score superior a los reportados en el trabajo original, tanto en inglés como en español. Esto puede deberse a múltiples factores, entre ellos nuestra interpretación de los aspectos técnicos del trabajo original, y diferencias de implementación, descritas en la Sección 2.3. Además, el trabajo original fue publicado en 2022, mientras que los experimentos de esta tesis se llevaron a cabo en 2024. Durante estos dos años, es posible que las versiones de las librerías utilizadas hayan cambiado, lo que podría haber afectado los resultados. Otro factor que podría influir son las semillas empleadas, especialmente en clasificadores como SVM, que no son completamente determinísticos.

Un aspecto clave que probablemente influye en las diferencias observadas, particularmente para el Pitt Corpus, es que en el trabajo original no se especifican los IDs de los participantes ni el número de visita utilizados. Debido a esta falta de información, seleccionamos aleatoriamente a los participantes con AD para equilibrar las clases y utilizamos únicamente la primera visita de cada participante. Por lo tanto, es altamente probable que el conjunto de entrenamiento que empleamos no sea el mismo que el utilizado en el trabajo original. Además, aplicamos nuestro propio proceso de procesamiento a las transcripciones en inglés antes de obtener los embeddings, y desconocemos el procesamiento aplicado por los autores. Estos factores combinados podrían explicar las diferencias observadas en los resultados.

## 3.2. Extensión 1: Representaciones contextuales

En la Tabla 3.2 se detallan la media y desviación estándar de los F1-scores obtenidos para cada clasificador usando los embeddings BERT-LAST y los BERT-FIRST.

Incluimos el F1-score obtenido utilizando un clasificador que predice siempre la clase mayoritaria. El uso de este clasificador nos sirve como *baseline* para evaluar el rendimiento de los clasificadores entrenados con los diferentes embeddings. Este baseline proporciona un punto de referencia mínimo, ya que representa el desempeño que se obtendría sin considerar ninguna información de los datos más allá de la distribución de clases. Comparar nuestros resultados con los de este clasificador nos permite cuantificar cuánto contribuyen los embeddings a mejorar la capacidad predictiva. Además, al utilizar este enfoque, se busca que cualquier mejora en las métricas de clasificación sea atribuible al aprendizaje de características útiles a partir de los datos.

| **Clasificador** | **Inglés** | **Español** | **Inglés** | **Español** |
|---|---|---|---|---|
| **Embeddings** | BERT-LAST-ENG-CASED | BERT-LAST-MUL-CASED | BERT-FIRST-ENG-CASED | BERT-FIRST-MUL-CASED |
| Baseline | 0.00 (0.00) | 0.70 (0.00) | 0.00 (0.00) | 0.70 (0.00) |
| SVM | 0.78 (0.01) | 0.66 (0.07) | 0.75 (0.02) | **0.70 (0.00)** |
| XGBoost | 0.73 (0.03) | 0.64 (0.09) | 0.72 (0.03) | 0.60 (0.07) |
| Logistic Regression | **0.79 (0.01)** | 0.68 (0.07) | 0.75 (0.02) | 0.57 (0.04) |
| Random Forest | 0.77 (0.02) | **0.71 (0.07)** | **0.76 (0.02)** | 0.51 (0.09) |

Tab. 3.2: Promedio (desviación estándar) del F1-score. El baseline representa un clasificador que siempre retorna la clase mayoritaria. Se resaltan los valores de F1-Score más altos obtenidos en cada idioma, con cada conjunto de embeddings.

Los resultados obtenidos para el Pitt Corpus muestran desempeños similares entre ambas representaciones, con valores de F1-score consistentemente por encima del baseline. En todos los clasificadores, los embeddings BERT-LAST-ENG-CASED lograron un desempeño ligeramente superior al de los embeddings BERT-FIRST-ENG-CASED. Específicamente, para los embeddings BERT-LAST-ENG-CASED, el mejor desempeño fue alcanzado por Logistic Regression, con un F1-score de 0,79. Por otro lado, en las representaciones BERT-FIRST-ENG-CASED, Random Forest se destacó como el clasificador con mejor rendimiento, obteniendo un F1-score de 0,76.

Con ambas representaciones, XGB fue el clasificador con el F1-score más bajo en el Pitt Corpus. Esto podría deberse a su sensibilidad a la elección de hiperparámetros y a la naturaleza de los datos. A diferencia de SVM o LR, XGB tiende a beneficiarse de una cantidad mayor de datos y de características bien diferenciadas, lo que puede no ser el caso en este conjunto específico. Por otro lado, a diferencia de Random Forest, que construye múltiples árboles de manera independiente y promedia sus predicciones, XGB usa un enfoque secuencial de *boosting*, donde cada árbol intenta corregir los errores del anterior. Esto hace que XGB sea más sensible al ruido y a la redundancia en los datos, lo que podría estar afectando su desempeño en este caso. Dado que los embeddings de BERT ya contienen información rica y densa, XGBoost podría estar sobreajustando a ciertos patrones específicos en lugar de generalizar bien. En cambio, Random Forest, al ser un método basado en *bagging*, puede ser más robusto frente a posibles correlaciones o redundancias en los embeddings, lo que explicaría su mejor desempeño.

Con los embeddings BERT-LAST-MUL-CASED el clasificador que obtuvo el F1-score más alto en Chile AD fue Random Forest, con 0,71. Si bien este valor técnicamente supera al baseline (0,70), la diferencia es mínima. Para los embeddings BERT-FIRST-MUL-CASED,

el clasificador SVM logró el F1-score más alto con un valor de 0,70. Sin embargo, la varianza nula hace pensar que siempre predice lo mismo, obteniendo siempre el mismo F1. Al revisar las predicciones, notamos que efectivamente SVM aprendió a predecir siempre la clase mayoritaria. Esto puede deberse a la falta de información discriminativa en los embeddings BERT-FIRST-MUL-CASED para este conjunto de datos. Los demás clasificadores presentan F1-scores menores, con Random Forest mostrando el desempeño más bajo de 0,51.

Aunque los embeddings BERT-LAST-MUL-CASED tienden a obtener mejores resultados para Chile AD, su desempeño es limitado y está por debajo del baseline. Por otro lado, los embeddings de BERT-FIRST-MUL-CASED no parecen aportar suficiente información discriminativa, y su mejor resultado con SVM tampoco proporciona evidencia confiable de que el clasificador sea efectivo en esta configuración.

Para evaluar si los desempeños obtenidos con los distintos embeddings son estadísticamente significativos, y en particular si existe alguna diferencia entre los resultados de los embeddings BERT-LAST y los resultados de los embeddings de BERT-FIRST, realizamos *tests de permutación* (Ojala y Garriga 2010). Estos tests permiten determinar si los valores de F1-score observados son superiores a los que se esperarían por azar, proporcionando una perspectiva adicional sobre la robustez de los resultados obtenidos.

Un test de permutación es una técnica no paramétrica utilizada para evaluar la significancia estadística de un resultado. Consiste en permutar aleatoriamente las etiquetas de los datos múltiples veces para generar una distribución nula de los valores de una métrica elegida, que se podrían obtener por azar. Comparando los resultados reales con esta distribución, es posible determinar si los valores observados son significativamente mejores que los esperados bajo un escenario de azar.

En este trabajo, los tests de permutación se llevaron a cabo siguiendo el mismo esquema de validación cruzada utilizado en la replicación del trabajo original. Para cada una de las 10 repeticiones del cross-validation, se generaron 100 permutaciones aleatorias de las etiquetas, respetando las divisiones de datos usadas para obtener los resultados de la Tabla 3.2. Esto permitió obtener una distribución nula de los valores de F1-score para cada clasificador y conjunto de datos evaluado, con ambas representaciones.

Comparando los valores reales de F1-score con estas distribuciones nulas, calculamos los p-valores, que indican la probabilidad de que el desempeño observado sea atribuible al azar. En este análisis, consideramos como significativo cualquier resultado con un p-valor menor o igual a 0,05. Estos valores proporcionan una medida estadística de la robustez de los resultados obtenidos y sirven como una validación adicional para interpretar la significancia de los mismos.

Decidimos realizar el análisis solamente para Random Forest y SVM, ya que son los que mejores resultados obtuvieron en Chile AD, y en el Pitt Corpus los resultados son similares para todos los clasificadores. Los resultados obtenidos para otros clasificadores están detallados en el Anexo B.

Los resultados de los tests de permutación con SVM y Random Forest para el Pitt Corpus se pueden observar en la Figura 3.1. Para todos los clasificadores, tanto para los embeddings BERT-LAST-ENG-CASED como para los embeddings BERT-FIRST-ENG-CASED, los tests obtuvieron siempre p-valores significativos, con valores menores a 0,01. Esto indica que, independientemente del clasificador utilizado, ambos tipos de embeddings fueron efectivos para capturar las diferencias entre las clases en este corpus. En particular, los embeddings BERT-FIRST-ENG-CASED lograron resultados igualmente sólidos que los em-

beddings BERT-LAST-ENG-CASED, lo que sugiere que las representaciones semánticas obtenidas por estos métodos son lo suficientemente robustas como para diferenciar las clases de manera confiable.

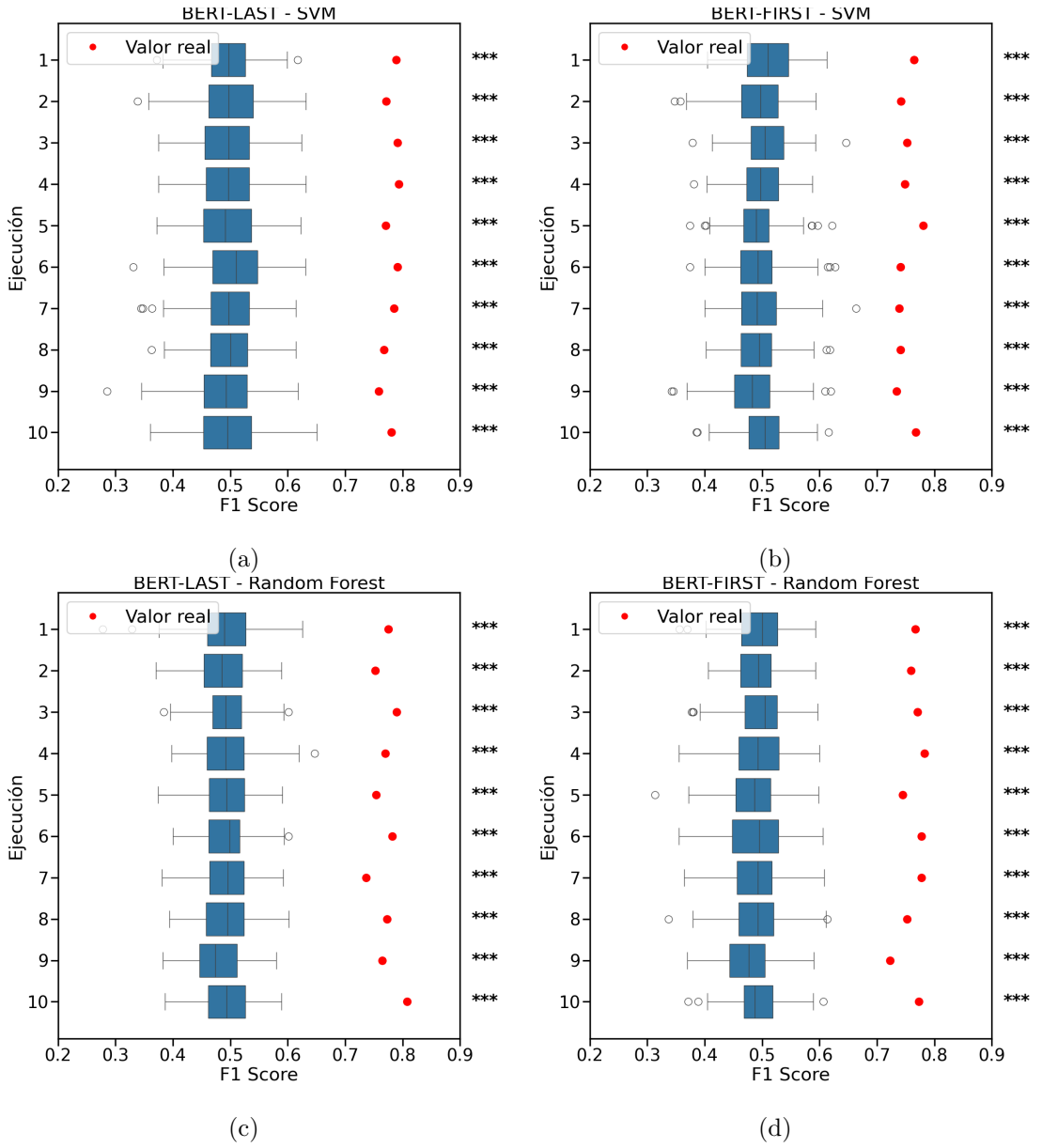



Fig. 3.1: Resultados de 10 ejecuciones de tests de permutación con los datos del Pitt Corpus, utilizando diferentes semillas para la separación de los datos. El punto rojo indica la puntuación obtenida con las etiquetas reales. En las subfiguras (a) y (b), se muestran los resultados para SVM con embeddings BERT-LAST-ENG-CASED y embeddings BERT-FIRST-ENG-CASED, respectivamente. En las subfiguras (c) y (d), se presentan los resultados para Random Forest con embeddings BERT-LAST-ENG-CASED y embeddings BERT-FIRST-ENG-CASED, respectivamente. Los p-valores asociados indican la significancia estadística de la puntuación observada en comparación con la distribución generada por permutaciones aleatorias, y se usan asteriscos (*) para denotar significancia estadística, utilizando más si el p-valor ($p$) es más chico. **ns**: $p > 0{,}05$; *: $p \leq 0{,}05$; **: $p \leq 0{,}01$; ***: $p \leq 0{,}001$.

En la Figura 3.2 se presentan los resultados de los tests de permutación realizados con Random Forest y SVM para los datos de Chile AD. En el caso de los embeddings BERT-

LAST-MUL-CASED, los valores de F1 obtenidos con SVM en las permutaciones presentan una distribución variada, con el F1 obtenido con las etiquetas reales (puntos rojos) en la parte superior del rango en varias ejecuciones. El p-valor más bajo obtenido es 0,01, lo que indica que en al menos una ejecución, el resultado real es significativamente mejor que los valores esperados por azar ($p \leq 0{,}05$). Sin embargo, otros p-valores son más altos, lo que sugiere cierta variabilidad en la significancia de los resultados dependiendo de la ejecución.

Por otro lado, los embeddings BERT-FIRST-MUL-CASED muestran una distribución muy diferente: el valor de F1 en todas las ejecuciones es consistentemente 0,7, tanto para los datos reales como para las permutaciones. Esta falta de variabilidad indica que SVM está produciendo un resultado fijo, sin importar la estructura de los datos. El hecho de que todos los p-valores sean 1 sugiere que el rendimiento de SVM con estos embeddings es indistinguible de una asignación de la etiqueta más probable, sin importar la instancia.

Para Random Forest, en el caso de los embeddings BERT-LAST-MUL-CASED, se observan varias ejecuciones con p-valores significativos, incluso obteniendo un p-valor menor a 0,01 en algunos casos. En cambio, con los embeddings BERT-FIRST-MUL-CASED, Random Forest nunca obtiene un p-valor significativo, con el resultado real cayendo casi siempre en la parte inferior del rango. En general, Random Forest parece captar mejor los patrones en los datos en comparación con SVM.

Este análisis sugiere que los embeddings BERT-LAST-MUL-CASED de BERT capturan mejor la información relevante para la tarea de clasificación en la base de datos Chile AD en comparación con los embeddings BERT-FIRST-MUL-CASED, que no muestran ninguna diferencia con respecto a los valores generados aleatoriamente.

Estos resultados reflejan una diferencia marcada entre los desempeños observados en inglés y español. Aunque ambas bases de datos son chicas, la base de datos en inglés es mucho más grande que la de español. En consecuencia, para el inglés, los clasificadores tienen más datos para aprender patrones generales y evitar el sobreajuste. En contraste, la base en español es más pequeña, lo que hace que los clasificadores sean más sensibles a la variabilidad en la partición de los datos, reduciendo la estabilidad de los resultados y aumentando la probabilidad de que el desempeño observado se deba al azar. Por otro lado, en inglés, se emplea un modelo de bert, entrenado exclusivamente en corpus en inglés, mientras que en español se utiliza un modelo multilingüe entrenado en textos de más de 100 idiomas, lo que podría afectar su capacidad para capturar matices específicos del español.

En resumen, al comparar los embeddings BERT-FIRST, extraídos de la capa inicial de BERT con los BERT-LAST, obtenidos de la capa final, observamos diferencias importantes en el rendimiento de los clasificadores, particularmente en la base de datos en español. En el conjunto de datos en inglés, ambas estrategias de extracción de embeddings ofrecieron mejoras respecto al baseline. Sin embargo, los resultados en español revelaron un panorama distinto. En este caso, los embeddings BERT-FIRST-MUL-CASED mostraron un desempeño comparable al azar, mientras que las representaciones BERT-LAST-MUL-CASED mejoraron el rendimiento en algunos clasificadores, aunque no de manera consistente.

La inclusión del clasificador que predice siempre la clase mayoritaria como baseline nos permitió tener una referencia sólida para entender el contexto de estos resultados, especialmente en español: SVM con los embeddings de BERT-FIRST-MUL-CASED obtiene los mismos resultados que el baseline, ya que predice siempre la clase mayoritaria. Dado que en Chile AD las clases están desbalanceadas, obtiene un F1-score de 0,7 en todas las ejecuciones, el cual es más alto que casi todos los valores obtenidos con los embeddings

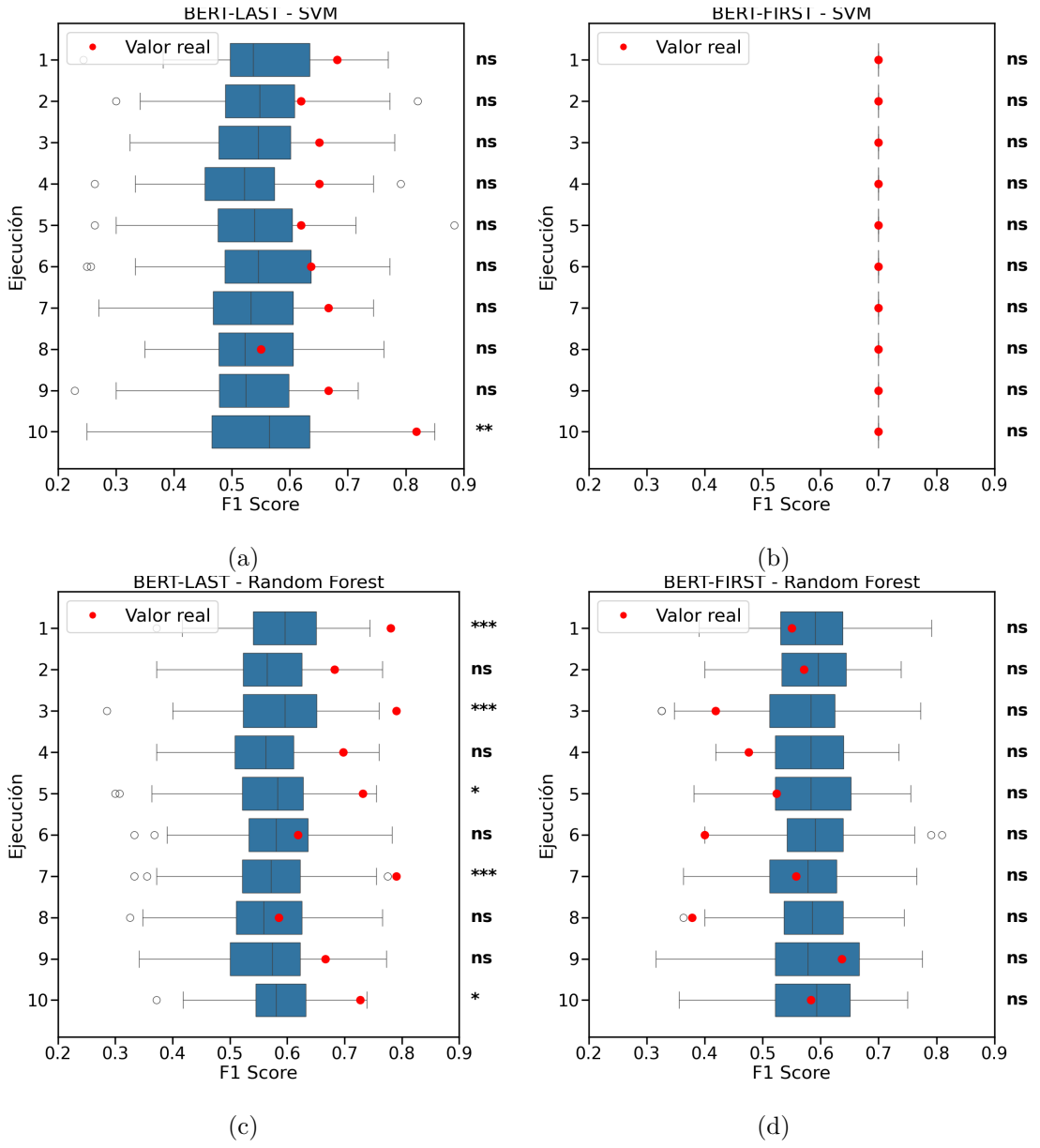



Fig. 3.2: Resultados de 10 ejecuciones de tests de permutación con los datos de Chile AD, utilizando diferentes semillas para la separación de los datos. El punto rojo indica la puntuación obtenida con las etiquetas reales. En las subfiguras (a) y (b), se muestran los resultados para SVM con embeddings BERT-LAST-MUL-CASED y embeddings BERT-FIRST-MUL-CASED, respectivamente. En las subfiguras (c) y (d), se presentan los resultados para Random Forest con embeddings BERT-LAST-MUL-CASED y embeddings BERT-FIRST-MUL-CASED, respectivamente. Los p-valores asociados indican la significancia estadística de la puntuación observada en comparación con la distribución generada por permutaciones aleatorias, y se usan asteriscos (*) para denotar significancia estadística, utilizando más si el p-valor ($p$) es más chico. **ns**: $p > 0{,}05$; *: $p \leq 0{,}05$; **: $p \leq 0{,}01$; ***: $p \leq 0{,}001$.

BERT-LAST, a excepción de Random Forest. Sin embargo, al analizar los tests de permutación, observamos que los resultados con los embeddings BERT-LAST-MUL-CASED son

en ocasiones significativos, mientras que los de BERT-FIRST-MUL-CASED siempre resultan similares al azar. Dado este panorama, surge la necesidad de analizar si el F1-score es suficiente para capturar la complejidad de los datos o si, por el contrario, podría estar ocultando información relevante.

## 3.3. Extensión 2: Otras métricas de evaluación

La elección de métricas de evaluación en problemas de clasificación binaria es un aspecto fundamental que influye directamente en la interpretación de los resultados y en la selección del mejor modelo. En tareas de detección de enfermedades como el Alzheimer (AD), el impacto de los errores no es simétrico: un falso negativo (no detectar un paciente con AD) puede tener consecuencias clínicas más graves que un falso positivo (diagnosticar erróneamente un control sano como paciente con AD).

Si bien el F1-score es una métrica ampliamente utilizada en este tipo de problemas, presenta limitaciones importantes en este contexto. Como discute Ferrer (2025), el F1-score combina precision y recall en una sola métrica, pero asigna el mismo peso a ambos errores sin considerar su impacto diferencial en la tarea. Además, la interpretación de los valores de F1-score depende de la distribución de clases en los datos. Un mismo valor de F1-score puede significar desempeños muy diferentes dependiendo de los *priors* del conjunto de datos, lo que dificulta la comparación entre diferentes datasets y puede llevar a interpretaciones erróneas de los resultados.

Este problema se observa claramente en nuestros experimentos con el conjunto de datos Chile AD (Tabla 3.2), donde el baseline que predice siempre la clase mayoritaria obtiene un F1-score de 0,7. Esto demuestra que F1-score puede dar valores altos incluso cuando el clasificador no aprende nada útil y simplemente reproduce el sesgo del conjunto de datos (en este caso, que el dataset está desbalanceado). De hecho, el mismo valor de 0,7 es obtenido por SVM, que también predice siempre la clase mayoritaria, lo que genera la ilusión de un buen desempeño cuando, en realidad, el clasificador no está diferenciando entre clases. Esto refuerza la idea de que F1-score por sí solo no es suficiente para evaluar correctamente el desempeño de un clasificador y que es necesario complementar el análisis con otras métricas que reflejen mejor la capacidad discriminativa del clasificador. Como señala Ferrer (2025), F1-score tiende a favorecer modelos que detectan más muestras como pertenecientes a la clase de interés, independientemente de si están correctamente o incorrectamente clasificadas. En este caso, dado que AD es la clase de interés, F1-score puede inflar la aparente calidad del clasificador cuando este predice mayoritariamente dicha clase, incluso si lo hace de manera incorrecta.

Para abordar estas limitaciones, en vez de utilizar solo F1-score, incorporamos otras métricas como precision, recall, accuracy, specificity (también conocida como true negative rate), el área bajo la curva ROC (ROC AUC) y el área bajo la curva Precision-Recall (PR AUC). Estas métricas permiten evaluar de manera más completa la capacidad de los modelos para distinguir entre clases y optimizar la toma de decisiones en entornos médicos.

La Tabla 3.3 presenta la media y la desviación estándar de cada métrica obtenida en el Pitt Corpus, comparando los embeddings BERT-LAST-ENG-CASED con los BERT-FIRST-ENG-CASED. Al igual que con el F1-score, las demás métricas superan al baseline en casi todos los casos para ambas representaciones. En el caso de Pitt, todas las métricas siguen los mismos patrones observados con F1, confirmando la tendencia general. Los resultados muestran que los embeddings BERT-LAST ofrecen un mejor desempeño en comparación con los embeddings BERT-FIRST-ENG-CASED en la mayoría de los clasificadores, aunque la diferencia es poca. Además, los clasificadores entrenados con embeddings BERT-LAST-ENG-CASED tienden a lograr mayor precision y recall, lo que indica una mejor capacidad para identificar correctamente a los pacientes con AD y reducir falsos positivos.

| Embeddings | Clasificador | F1-Score | Precision | Recall | Accuracy | Specificity | ROC AUC | PR AUC |
|---|---|---|---|---|---|---|---|---|
| | **Baseline** | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) | 0.50 (0.00) | 1.00 (0.00) | 0.50 (0.00) | 0.50 (0.00) |
| BERT-FIRST-ENG-CASED | **SVM** | 0.75 (0.02) | 0.79 (0.03) | 0.72 (0.03) | 0.76 (0.02) | 0.80 (0.04) | **0.84 (0.01)** | 0.86 (0.01) |
| | **XGBoost** | 0.72 (0.03) | 0.75 (0.03) | 0.70 (0.03) | 0.73 (0.03) | 0.76 (0.04) | 0.82 (0.02) | 0.84 (0.02) |
| | **Logistic Regression** | 0.75 (0.02) | **0.81 (0.03)** | 0.70 (0.02) | **0.77 (0.02)** | **0.83 (0.02)** | **0.84 (0.01)** | **0.88 (0.01)** |
| | **Random Forest** | **0.76 (0.02)** | 0.79 (0.02) | **0.74 (0.02)** | **0.77 (0.02)** | 0.80 (0.02) | **0.84 (0.01)** | 0.86 (0.01) |
| BERT-LAST-ENG-CASED | **SVM** | 0.78 (0.01) | **0.85 (0.01)** | 0.72 (0.02) | **0.80 (0.01)** | **0.87 (0.01)** | 0.87 (0.01) | 0.89 (0.01) |
| | **XGBoost** | 0.73 (0.03) | 0.74 (0.03) | 0.73 (0.03) | 0.74 (0.03) | 0.74 (0.04) | 0.84 (0.01) | 0.85 (0.01) |
| | **Logistic Regression** | **0.79 (0.01)** | 0.83 (0.01) | **0.76 (0.01)** | **0.80 (0.01)** | 0.85 (0.01) | **0.89 (0.01)** | **0.90 (0.01)** |
| | **Random Forest** | 0.77 (0.02) | 0.78 (0.01) | **0.76 (0.03)** | 0.77 (0.02) | 0.79 (0.02) | 0.86 (0.01) | 0.87 (0.01) |

Tab. 3.3: Promedio y desvío estándar obtenidos para el Pitt Corpus. Se resaltan los valores más altos obtenidos para cada métrica, con cada conjunto de embeddings.

La Tabla 3.4 muestra el desempeño de los clasificadores en Chile AD. Al igual que en el Pitt Corpus, los embeddings BERT-LAST-MUL-CASED superan a los de BERT-FIRST-MUL-CASED en la mayoría de las métricas, aunque en este caso encontramos diferencias muy marcadas entre lo reflejado por F1 y lo reflejado por las demás métricas. Con los embeddings BERT-LAST-MUL-CASED, los clasificadores logran un mejor equilibrio entre precisión y recall, lo que indica que no solo detectan mejor a los pacientes con AD, sino que también reducen la cantidad de falsos positivos. Este efecto es particularmente notable en el caso de Random Forest, que alcanza los valores más altos de precision y recall, sugiriendo que este clasificador es capaz de identificar correctamente ambas clases con mayor consistencia.

Los embeddings BERT-LAST-MUL-CASED producen valores superiores en ROC AUC y Precision-Recall AUC, lo que indica una mejor capacidad para distinguir entre clases y manejar el compromiso entre precision y recall. En cambio, los embeddings de BERT-FIRST generan clasificadores con valores de specificity (True Negative Rate) más bajos, sugiriendo una menor capacidad para identificar correctamente a los controles sanos.

En particular, SVM con BERT-FIRST-MUL-CASED obtiene un specificity nulo y un recall máximo, sugiriendo nuevamente que el clasificador predice siempre la clase mayoritaria (AD en este caso). Al observar las predicciones confirmamos que SVM con los embeddings BERT-FIRST-MUL-CASED predice siempre AD. Una posible razón es que, dado el pequeño desbalance, SVM puede estar sesgando su frontera de decisión hacia la clase mayoritaria para minimizar el error global. Esto es especialmente probable si las características extraídas no logran separar de forma clara las dos clases, lo que lleva al clasificador a optar por la solución "trivial" de predecir siempre la clase con mayor cantidad de instancias.

En problemas de detección de enfermedades como el Alzheimer, es crucial diferenciar entre falsos positivos (clasificar erróneamente a un control como paciente con AD) y falsos negativos (no detectar a un paciente con AD). F1-score combina precision y recall en una sola métrica, pero da igual peso a ambos, lo que puede ser problemático si el costo de los errores es asimétrico. Adicionalmente, el F1-score no es simétrico respecto a la asignación de clases. Esto significa que si intercambiamos las etiquetas de las clases, el valor del F1-score podría cambiar significativamente. Al probar considerar CTR como la clase positiva y AD como la clase negativa, SVM con los embeddings BERT-FIRST-MUL-CASED obtiene un F1-score muy cercano a cero. Esto subraya que el F1-score es sensible a la asignación de las clases y no puede considerarse una métrica completamente simétrica cuando las etiquetas se cambian.

Además, en este caso se usa SVM con kernel radial, cuyo desempeño depende fuertemente de la correcta elección de sus hiperparámetros, en particular el parámetro de regularización (C) y el parámetro gamma. Si estos parámetros no se ajustan adecuadamente,

el clasificador puede generar fronteras de decisión poco representativas, que favorecen la predicción de la clase mayoritaria. Esto es especialmente problemático en conjuntos de datos pequeños, como en este caso, donde la variabilidad y la representatividad de las muestras son limitadas. Si bien realizamos *grid search* donde se buscó obtener el C y el gamma que maximizaran el accuracy, hubo varias combinaciones que alcanzaron el valor más alto, lo cual se puede ver en la Figura A.1c. Tal vez, si se hubiera elegido otra de estas combinaciones, o incluso una con un accuracy ligeramente inferior, el clasificador podría haber mostrado un desempeño más equilibrado, ya que se podría haber logrado un mejor compromiso entre recall y specificity. Estos resultados muestran el problema de optimizar utilizando una métrica como accuracy, en vez de, por ejemplo, ROC AUC, como sugiere Ferrer (2025), quien argumenta que las métricas basadas en curvas de decisión ofrecen una mejor evaluación del desempeño del clasificador en diferentes umbrales y condiciones de distribución de clases.

Dentro de los clasificadores entrenados con embeddings BERT-FIRST-MUL-CASED, los mejores resultados en Chile AD los obtienen XGBoost y Logistic Regression, mientras que, con los embeddings BERT-LAST-MUL-CASED, Random Forest supera al resto en todas las métricas.

| Embeddings | Clasificador | F1-Score | Precision | Recall | Accuracy | Specificity | ROC AUC | PR AUC |
|---|---|---|---|---|---|---|---|---|
| | **Baseline** | 0.70 (0.00) | 0.54 (0.00) | 1.00 (0.00) | 0.54 (0.00) | 0.00 (0.00) | 0.50 (0.00) | 0.54 (0.00) |
| BERT-FIRST-MUL-CASED | **SVM** | **0.70 (0.00)** | 0.54 (0.00) | **1.00 (0.00)** | 0.54 (0.00) | 0.00 (0.00) | 0.42 (0.02) | 0.52 (0.02) |
| | **XGBoost** | 0.60 (0.07) | **0.58 (0.06)** | 0.63 (0.08) | **0.55 (0.07)** | **0.46 (0.10)** | **0.54 (0.10)** | 0.59 (0.09) |
| | **Logistic Regression** | 0.57 (0.04) | 0.53 (0.03) | 0.61 (0.07) | 0.50 (0.04) | 0.37 (0.06) | **0.54 (0.05)** | **0.60 (0.03)** |
| | **Random Forest** | 0.51 (0.09) | 0.51 (0.07) | 0.51 (0.11) | 0.47 (0.07) | 0.43 (0.09) | 0.51 (0.06) | 0.57 (0.05) |
| BERT-LAST-MUL-CASED | **SVM** | 0.66 (0.07) | 0.64 (0.05) | 0.67 (0.09) | 0.62 (0.07) | 0.57 (0.06) | 0.61 (0.07) | 0.67 (0.07) |
| | **XGBoost** | 0.64 (0.09) | 0.62 (0.07) | 0.68 (0.13) | 0.60 (0.09) | 0.52 (0.08) | 0.61 (0.09) | 0.66 (0.08) |
| | **Logistic Regression** | 0.68 (0.07) | 0.66 (0.06) | 0.70 (0.09) | 0.65 (0.07) | 0.58 (0.07) | 0.68 (0.05) | 0.71 (0.05) |
| | **Random Forest** | **0.71 (0.07)** | **0.71 (0.07)** | **0.71 (0.08)** | **0.68 (0.07)** | **0.66 (0.08)** | **0.72 (0.05)** | **0.74 (0.05)** |

Tab. 3.4: Promedio y desvío estándar obtenidos para Chile AD. Se resaltan los valores más altos obtenidos para cada métrica, con cada conjunto de embeddings.

En ambos idiomas, los embeddings BERT-LAST muestran un desempeño superior, aunque en español la diferencia entre estos y los embeddings BERT-FIRST es aún más marcada que en inglés.

Volvimos a realizar los tests de permutación, pero esta vez utilizando accuracy como métrica de evaluación, para comparar los resultados con los obtenidos usando F1-score. En el Pitt Corpus todos los tests volvieron a ser significativos con ambas representaciones y para todos los clasificadores. En Chile AD, los resultados se mantuvieron respecto a F1-score: mientras que algunos tests con los embeddings BERT-LAST-MUL-CASED fueron significativos, los de BERT-FIRST-MUL-CASED siguieron mostrando un desempeño equivalente al azar. Específicamente, con embeddings BERT-LAST-MUL-CASED, SVM, XGB y Logistic Regression lograron la misma cantidad de tests significativos, mientras que Random Forest obtuvo 6/10, una leve mejora respecto a los 5/10 obtenidos con F1-score. Con BERT-FIRST-MUL-CASED, los resultados no cambiaron.

Los resultados presentados muestran que, si bien F1-score puede ser útil en ciertos contextos, su interpretación es limitada cuando los costos de los errores son desiguales o cuando la distribución de clases influye en su valor. Como se argumenta en Ferrer (2025), es crucial utilizar métricas que capturen de manera más completa el comportamiento del clasificador, en particular cuando se busca minimizar falsos negativos en enfermedades como el Alzheimer. El uso de ROC AUC y PR AUC permite evaluar de manera más robusta la capacidad discriminativa de los clasificadores, mientras que incorporar specificity y

accuracy complementa el análisis al reflejar la habilidad del clasificador para manejar ambas clases.

## 3.4. Extensión 3: Comparación de modelos base

En esta sección, analizaremos el impacto de utilizar modelos de BERT en sus variantes *cased* y *uncased* para la tarea de clasificación de Alzheimer a partir de texto. Como explicamos en la Sección 2.2, la distinción entre modelos cased y uncased radica en cómo estos modelos fueron entrenados para distinguir o no la información de la capitalización de las palabras en el texto, y por lo tanto, las diferencias en las representaciones que generan. Estas diferencias se dan incluso para las mismas palabras cuando éstas se encuentran en minúscula, ya que los modelos han sido entrenados con distintos conjuntos de datos. En este trabajo, no estamos interesados en observar el efecto de la capitalización, sino sólo medir el impacto en los resultados al usar las representaciones generadas por los modelos cased y uncased para la tarea de detección automática de Alzheimer.

Hasta ahora, empleamos el modelo `bert-base-cased` para los textos en inglés y el modelo `bert-base-multilingual-cased` para los textos en español, pero las transcripciones de ambas bases de datos fueron preprocesadas y convertidas a minúsculas previo a obtener los embeddings. Resulta relevante evaluar si el uso de modelos cased aporta ventajas en este contexto o si un modelo uncased podría desempeñarse de manera similar, o mejor.

Para este experimento, repetimos el mismo proceso de búsqueda de hiperparámetros, entrenamiento y evaluación que en las secciones anteriores, usando las mismas separaciones de datos durante las validaciones cruzadas, pero esta vez entrenando con los embeddings BERT-FIRST-ENG-CASED y BERT-LAST-ENG-CASED para el Pitt Corpus y los embeddings BERT-FIRST-MUL-CASED y BERT-LAST-MUL-CASED para Chile AD. Los resultados de la búsqueda de hiperparámetros con ambas representaciones, junto con mapas de calor mostrando el accuracy obtenido para cada configuración de hiperparámetros, se detallan en el Anexo A.

A partir de ahora utilizaremos solamente ROC AUC como métrica de evaluación, dado que, como explicamos anteriormente, el F1-score no es una métrica adecuada en este caso, y accuracy tampoco lo es, ya que el dataset Chile AD está desbalanceado. Tanto ROC AUC como PR AUC podrían ser usadas en este caso. En el Anexo C se encuentran tablas con los resultados de todas las métricas, obtenidos para el Pitt Corpus y Chile AD, usando las versiones uncased de los modelos de BERT. En este, extendemos el análisis de por qué el F1-score no es una métrica adecuada para Chile AD.

En la Tabla 3.5 se presentan los ROC AUC obtenidos para el Pitt Corpus. Tanto con en BERT-LAST como en BERT-FIRST observamos una leve mejora al utilizar el modelo uncased, aunque con los embeddings BERT-ENG-LAST-UNCASED la diferencia es más marcada, mejorando el desempeño con todos los clasificadores. Los embeddings BERT-FIRST-ENG-UNCASED solo muestran una leve mejora respecto a los embeddings BERT-FIRST-ENG-CASED con el clasificador Random Forest. Aunque ambas representaciones generan resultados muy por encima del baseline, los embeddings BERT-LAST-ENG-UNCASED, al igual que en el caso cased, exhiben un rendimiento superior en comparación con los embeddings BERT-FIRST-ENG-UNCASED.

| Embeddings | Clasificador | ROC AUC | |
|---|---|---|---|
| | | Uncased | Cased |
| | **Baseline** | 0.50 (0.00) | 0.50 (0.00) |
| BERT-FIRST | **SVM** | 0.84 (0.01) | **0.84 (0.01)** |
| | **XGBoost** | 0.82 (0.01) | 0.82 (0.02) |
| | **Logistic Regression** | 0.84 (0.01) | **0.84 (0.01)** |
| | **Random Forest** | **0.86 (0.01)** | **0.84 (0.01)** |
| BERT-LAST | **SVM** | **0.90 (0.01)** | 0.87 (0.01) |
| | **XGBoost** | 0.87 (0.01) | 0.84 (0.01) |
| | **Logistic Regression** | **0.90 (0.01)** | **0.89 (0.01)** |
| | **Random Forest** | 0.88 (0.01) | 0.86 (0.01) |

Tab. 3.5: Promedio y desvío estándar de ROC AUC, obtenidos para el Pitt Corpus, comparando los resultados de usar modelos cased y uncased de BERT. Se resaltan los ROC-AUC más altos obtenidos para cada conjunto de embeddings.

La Tabla **??** presenta los valores de ROC AUC obtenidos para Chile AD. A diferencia de lo observado en el Pitt Corpus, en este caso la versión uncased de BERT no supera el baseline con ninguno de los clasificadores, ni siquiera con los embeddings BERT-LAST-MUL-UNCASED. Sin embargo, estos embeddings siguen mostrando un desempeño superior al de los embeddings BERT-FIRST-MUL-UNCASED. Tanto en los resultados con BERT-LAST-MUL-UNCASED como con BERT-FIRST-MUL-UNCASED, se aprecia una disminución notable en el rendimiento de los clasificadores en comparación con los resultados obtenidos con los embeddings de la versión cased.

En resumen, la comparación entre las variantes cased y uncased de BERT muestra que el desempeño de los modelos depende del idioma. Para el Pitt Corpus (inglés), tanto los embeddings de BERT-FIRST como los embeddings BERT-LAST se benefician de la versión uncased. En cambio, para los textos en español, especialmente con embeddings BERT-LAST, el uso de la versión uncased ocasiona una degradación considerable del desempeño.

Con el objetivo de analizar con mayor detalle las diferencias entre los embeddings generados con los modelos cased y uncased de BERT, generamos visualizaciones de palabras tanto en inglés como en español. Utilizamos cuatro conjuntos de palabras: las 20 palabras más comunes del Pitt Corpus y las 20 más comunes de Chile AD, un conjunto de palabras arbitrarias en inglés y el mismo conjunto en español. En el Anexo D se encuentran las visualizaciones de las palabras, junto con el análisis de las palabras arbitrarias.

La Figura 3.3 corresponde a las proyecciones de las representaciones BERT-FIRST-ENG-CASED y BERT-FIRST-ENG-UNCASED de las 20 palabras más comunes del Pitt Corpus, y las proyecciones de las representaciones BERT-FIRST-MUL-CASED y BERT-FIRST-MUL-UNCASED de las 20 palabras más comunes de Chile AD. En el caso del Pitt Corpus observamos que, tanto con BERT-FIRST-ENG-CASED como con BERT-FIRST-ENG-UNCASED, se agrupan palabras semánticamente similares como “girl” y “boy”, “cookie” y “cookies”, o mismos tipos de palabras como los verbos “falling” y “running” o los sustantivos (y objetos inanimados) “water” y “floor”. No se aprecian clústeres tan marcados como los que obtuvimos con el conjunto arbitrario de palabras en inglés, lo cual es esperable dado que estos términos provienen de conversaciones reales y, por ende, reflejan una mayor variabilidad lingüística. Sin embargo, aunque de forma levemente distinta, los modelos parecen

agrupar de forma coherente las palabras del Pitt Corpus.

Por el contrario, en el caso de las palabras de Chile AD , observamos clústeres mucho más marcados, aunque no muy coherentes. Si bien, en el caso de los embeddings BERT-FIRST-MUL-CASED, términos como "platos", "lavaplatos" y "plato" se encuentran cercanos, en el mismo clúster vemos palabras como "niño" y "veo", que no son semánticamente similares. Con los embeddings BERT-FIRST-MUL-UNCASED es incluso peor, ya que palabras como "piso" y "arriba" o "parece" y "plato" están totalmente superpuestos, como si fueran equivalentes.

La Figura 3.3 corresponde a las proyecciones de las representaciones BERT-LAST-ENG-CASED y BERT-LAST-ENG-UNCASED de las 20 palabras más comunes del Pitt Corpus, y las proyecciones de las representaciones BERT-LAST-MUL-CASED y BERT-LAST-MUL-UNCASED de las 20 palabras más comunes de de Chile AD. En el caso del Pitt Corpus, si bien no se formaron clústeres muy distinguibles, podemos ver que palabras como "boy" y "girl" o "cookie" y "cookies" se encuentran cercanas, tanto con los embeddings BERT-LAST-ENG-CASED como con los embeddings BERT-LAST-ENG-UNCASED, al igual que sucedía con los embeddings BERT-FIRST. Sin embargo, mientras que con los últimos se agrupaban palabras similares en términos del tipo de palabra (verbo, sustantivo, etc.), en este caso notamos que las agrupaciones se dan por uso común de las palabras y significado dentro de contextos. Por ejemplo, los términos "falling" y "water", que comúnmente se utilizan juntos, están cercanos tanto en la versión uncased como en la cased. Lo mismo sucede con "going" y "well" con BERT-LAST-ENG-CASED, o "sink", "dishes" y "water" con BERT-LAST-ENG-UNCASED. Es difícil determinar si hay una superioridad del modelo cased por sobre el uncased o viceversa en este caso, ya que estas agrupaciones son subjetivas.

Al observar los resultados para las palabras de Chile AD con los embeddings BERT-LAST, notamos que las palabras están dispersas, sin agrupaciones muy marcadas. En los embeddings BERT-LAST-MUL-CASED se agrupan términos como "lavando" y "secando", ambos gerundios y que se usan en contextos similares, pero también cercanos a estos está "niño", que no se relaciona. Lo mismo sucede con BERT-LAST-MUL-UNCASED, donde se agrupan palabras como "niños", "niño" y "señora" junto con "agua".

En resumen, para los términos extraídos de los corpus Pitt y Chile AD, observamos que los embeddings BERT-LAST tienden a reflejar asociaciones basadas en el uso contextual de las palabras más que en categorías gramaticales, aunque en español las representaciones siguen siendo más dispersas. Estos resultados sugieren que la capacidad de BERT para modelar similitudes semánticas varía según el idioma y el tipo de embedding utilizado, con una mayor robustez en inglés y un desempeño más inconsistente en español, especialmente en la versión uncased.

Los resultados que obtuvimos en este experimento reflejan en su mayoría los mismos patrones observados en las métricas de desempeño de los clasificadores. En inglés, la versión uncased de BERT parece organizar las palabras de manera más coherente, lo que se traduce en un mejor desempeño de los clasificadores. En el caso del español, los embeddings BERT-FIRST no logran capturar la información semántica de las palabras de forma adecuada, independientemente de si se usa un modelo cased o uncased, lo que explicaría el bajo desempeño de los clasificadores en Chile AD. Si bien los embeddings BERT-LAST-MUL-UNCASED organizan de forma coherente las palabras aisladas, evidenciada en la claridad de los clústeres en las visualizaciones, esta separación no se refleja en un desempeño óptimo de los clasificadores. Esto podría deberse a la elección de hiperparámetros o a la baja cantidad de datos, lo que impide a los clasificadores aprender patrones significativos.

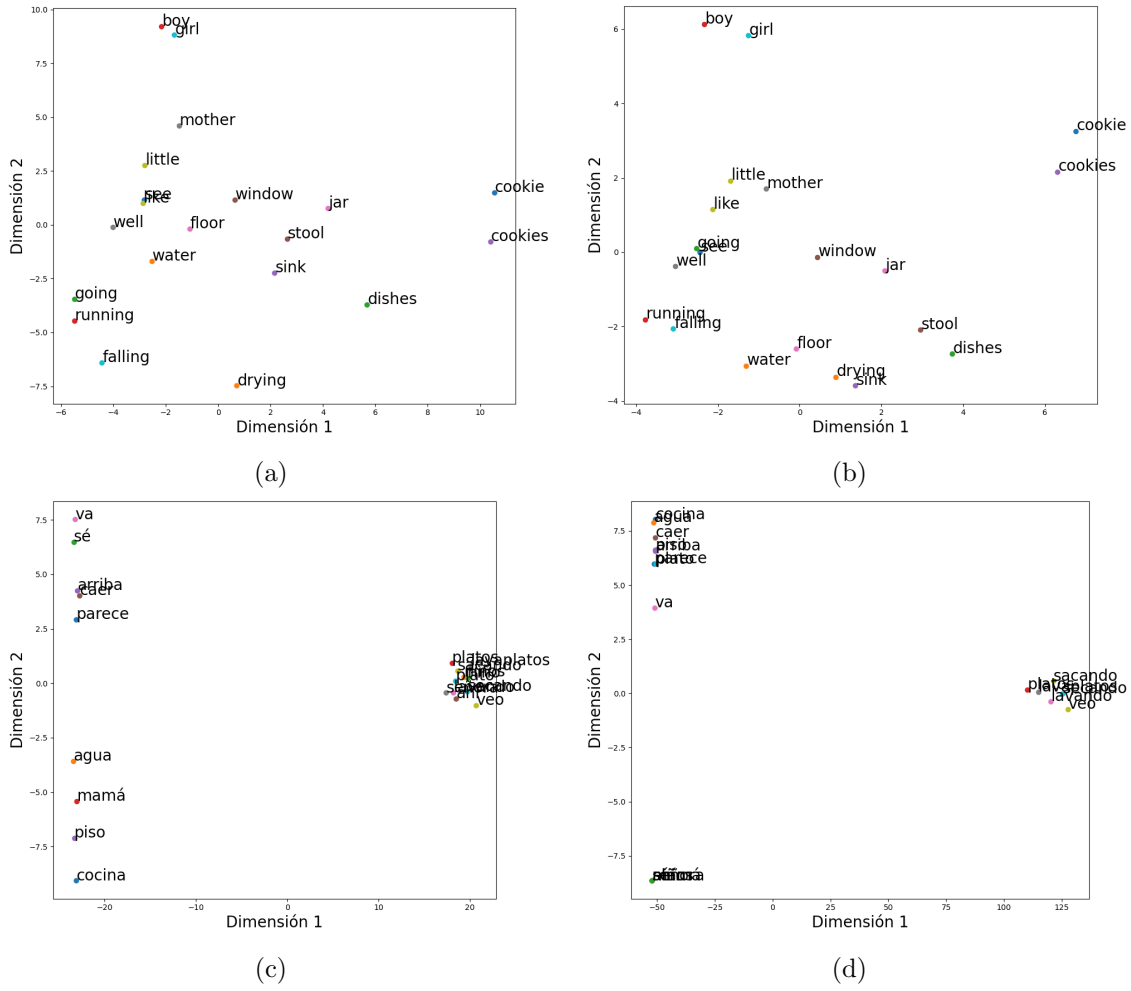



Fig. 3.3: Visualización de embeddings BERT-FIRST de las 20 palabras más comunes de los datasets en un espacio bidimensional, obtenido mediante reducción de dimensionalidad con PCA. En (a) y (b) se muestran los resultados para el Pitt Corpus, con los embeddings BERT-FIRST-ENG-CASED y BERT-FIRST-ENG-UNCASED, respectivamente. En (c) y (d), se presentan los resultados para Chile AD, con los embeddings BERT-FIRST-MUL-CASED y BERT-FIRST-MUL-UNCASED, respectivamente.

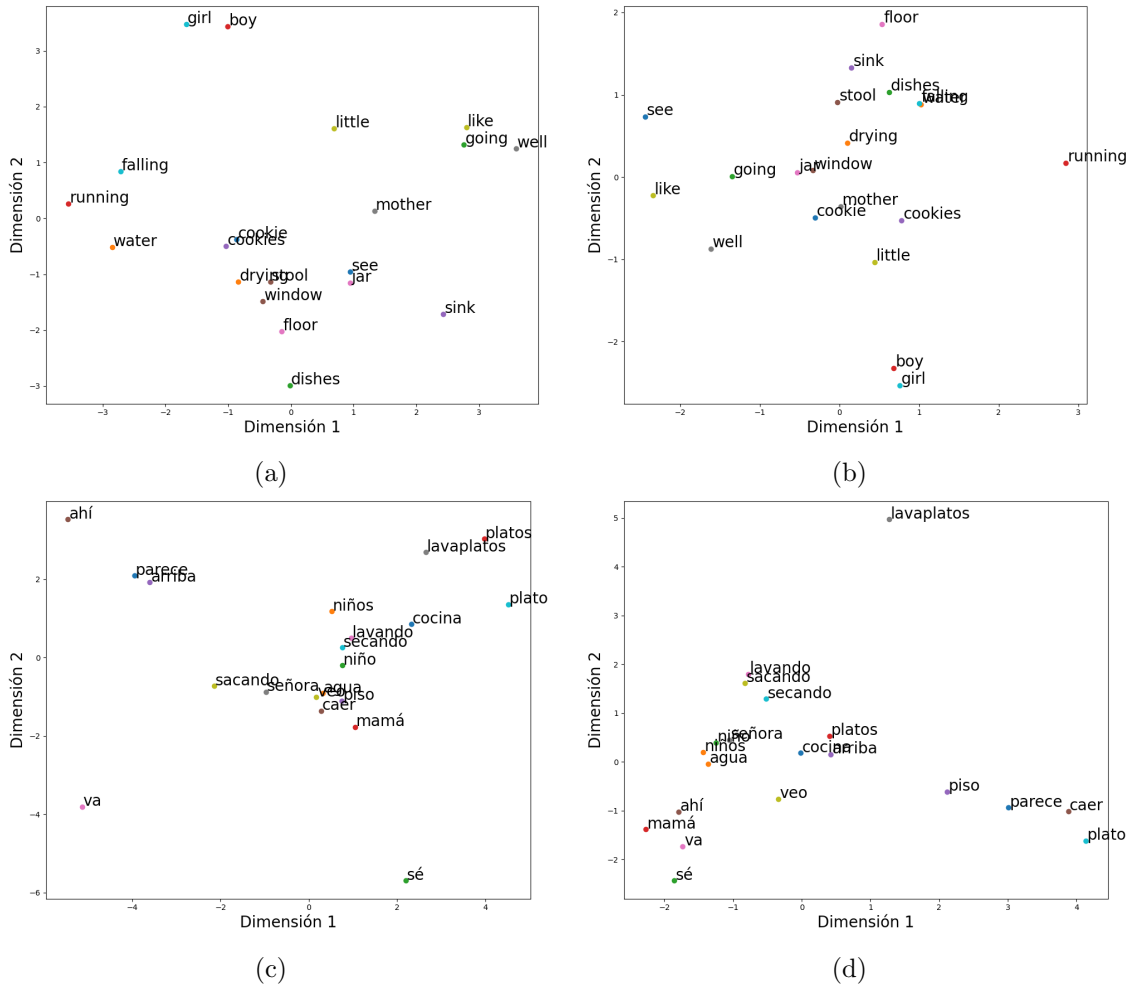



Fig. 3.4: Visualización de embeddings BERT-LAST de las 20 palabras más comunes de los datasets en un espacio bidimensional, obtenido mediante reducción de dimensionalidad con PCA. En (a) y (b) se muestran los resultados para el Pitt Corpus con los embeddings BERT-LAST-ENG-CASED y BERT-LAST-ENG-UNCASED, respectivamente. En (c) y (d), se presentan los resultados para Chile AD, mostrando los embeddings BERT-LAST-MUL-CASED y BERT-LAST-MUL-UNCASED, respectivamente.

# 4. CONCLUSIONES

En este trabajo, hemos replicado y extendido el estudio de Pérez-Toro y col. (2022) sobre detección automática de la enfermedad de Alzheimer a partir de transcripciones de habla utilizando embeddings generados con BERT. A lo largo del proceso, evaluamos el desempeño de diferentes representaciones del texto, extendimos el análisis con otras métricas de evaluación, y analizamos la influencia de los modelos cased y uncased en la clasificación de los datos en inglés y español.

Al replicar la metodología del trabajo original, obtuvimos resultados superiores a los reportados por los autores, tanto en la base de datos en inglés como en la base de datos en español. En el Pitt Corpus, el mejor F1-score reportado fue 0,69 con XGBoost, mientras que en nuestra replicación alcanzamos un F1-score de 0,72 con el mismo clasificador. Además, el clasificador con el mayor desempeño tras la replicación fue Random Forest, con un F1-score de 0.76, el cual no había sido evaluado en el trabajo original. Pueden haber muchas razones por las cuales no obtuvimos los mismos resultados, entre ellas posibles diferencias en la implementación debido a nuestra interpretación del trabajo original, semillas empleadas en clasificadores no determinísticos y separación de datos, versiones de las librerías utilizadas, y diferencias en los datos. Además, en el trabajo de Pérez-Toro y col. (2022) reportan resultados obtenidos con embeddings acústicos, los cuales decidimos no utilizar para enfocar el estudio únicamente en embeddings lingüísticos.

Extendimos el análisis comparando el desempeño de dos representaciones de embeddings extraídas de BERT para la replicación: una proveniente de la capa inicial de embeddings (BERT-FIRST) y otra de la capa final (BERT-LAST). Utilizando el F1-score como métrica de evaluación, encontramos que los embeddings BERT-LAST superaron sistemáticamente a los BERT-FIRST en casi todos los clasificadores y en ambas bases de datos.

En el Pitt Corpus, el mejor desempeño se obtuvo con Logistic Regression utilizando embeddings BERT-LAST-ENG-CASED, alcanzando un F1-score de 0,79. En contraste, el mejor resultado con los embeddings BERT-FIRST-ENG-CASED fue un F1-score de 0,76, obtenido con Random Forest. Todos los clasificadores entrenados con ambas representaciones del Pitt Corpus superaron al baseline (un clasificador que siempre predice la clase mayoritaria). XGBoost fue el clasificador con el desempeño más bajo para los dos conjuntos de embeddings, con un F1-score de 0,73 con los embeddings BERT-LAST-ENG-CASED y un F1-score de 0,76 con los embeddings BERT-FIRST-ENG-CASED, posiblemente debido a su sensibilidad a los hiperparámetros y a la redundancia en los datos. A diferencia de Random Forest, que es más robusto a correlaciones al usar *bagging*, XGBoost emplea *boosting*, lo que podría haberlo hecho más susceptible al sobreajuste en este conjunto específico.

En Chile AD, la mayoría de los clasificadores obtuvo resultados similares al baseline con ambas representaciones. Sin embargo, los embeddings BERT-LAST-MUL-CASED lograron F1-scores ligeramente superiores en general. Aunque la diferencia entre el mejor F1-score con BERT-LAST-MUL-CASED (0,71 con Random Forest) y el mejor con BERT-FIRST-MUL-CASED (0,70 con SVM) es mínima, observamos que SVM siempre predice la clase mayoritaria. Dado que Chile AD es un dataset desbalanceado, este comportamiento explica su F1-score relativamente alto.

Además, realizamos tests de permutación para evaluar la significancia de los resultados. En el Pitt Corpus, ambas representaciones produjeron resultados estadísticamente

significativos. Sin embargo, en Chile AD, los embeddings BERT-FIRST-MUL-CASED nunca lograron superar el azar, mientras que los embeddings BERT-LAST-MUL-CASED sí alcanzaron resultados significativos en algunas pruebas.

Dado que el F1-score no parecía ser la métrica más adecuada para este problema, como se evidenció en los resultados de SVM con los embeddings BERT-FIRST-MUL-CASED, decidimos incluir métricas adicionales para un análisis más completo. En el Pitt Corpus, ambas representaciones lograron resultados altos y muy superiores al baseline en todos los clasificadores. Sin embargo, los embeddings BERT-LAST-ENG-CASED siguieron mostrando un mejor desempeño: por ejemplo, con esta representación, Logistic Regression alcanzó un ROC AUC de 0,89 y un Precision-Recall AUC de 0,90, mientras que con los embeddings BERT-FIRST-ENG-CASED obtuvo 0,84 y 0,88, respectivamente.

Contrario a lo que sucede en el Pitt Corpus, donde el F1-score es consistente con otras métricas y no parece ocultar información relevante, en Chile AD la situación es diferente. Un caso claro es el de SVM con los embeddings BERT-FIRST-MUL-CASED, que obtiene un F1-score de 0,70. Sin embargo, este valor es engañoso, ya que SVM predice siempre la clase mayoritaria, resultando en un Recall máximo y un True Negative Rate nulo. Esto indica que el F1-score está inflado debido al alto Recall, causado por el leve desbalance de clases en un dataset de tamaño reducido, lo que limita su utilidad como métrica única para evaluar el desempeño en este conjunto.

Los resultados obtenidos con todas las métricas en Chile AD confirman que los embeddings BERT-LAST-MUL-CASED superan a los BERT-FIRST-MUL-CASED. Con los primeros, Random Forest se destaca como el mejor clasificador en todas las métricas, alcanzando un ROC AUC de 0,72. En contraste, con los embeddings BERT-FIRST-MUL-CASED, XGBoost mostró el mejor desempeño, pero con un ROC AUC un 33 % menor.

En este trabajo, también exploramos el impacto de utilizar modelos cased y uncased de BERT para la detección automática de Alzheimer a partir de texto. Evaluamos ambas variantes en inglés y español, considerando representaciones BERT-FIRST y BERT-LAST. Los resultados mostraron que la elección entre cased y uncased influye en el desempeño del clasificador, pero de manera diferente según el idioma. En el Pitt Corpus (inglés), los embeddings BERT-LAST-ENG-UNCASED mejoraron respecto a los embeddings BERT-LAST-ENG-CASED, donde por ejemplo, SVM obtuvo un ROC AUC de $0,90$, mientras que en la versión cased alcanzaba un valor de 0,87. Con los embeddings BERT-FIRST, los resultados entre las versiones cased y uncased fueron muy similares, con una leve mejora en el caso de Random Forest, que obtuvo un ROC AUC de 0,86 con BERT-LAST-ENG-UNCASED, cuando los embeddings BERT-LAST-ENG-CASED presentaban un valor de 0,84.

Por el contrario, en español (Chile AD), el uso de la versión uncased afectó negativamente el rendimiento, especialmente con los embeddings BERT-LAST. En particular, el ROC AUC de Random Forest se redujo drásticamente de 0,72 con los embeddings BERT-LAST-MUL-CASED a 0,41 con los embeddings BERT-LAST-MUL-UNCASED, y esta caída se observó de manera consistente con todos los clasificadores. Los embeddings BERT-FIRST también se vieron afectados, como lo evidencia la disminución del ROC-AUC en Logistic Regression, que pasó de 0,54 con BERT-LAST-MUL-CASED a 0,34 con BERT-LAST-MUL-UNCASED.

En general, nuestros resultados indican que las representaciones contextuales obtenidas de la última capa de BERT ofrecen una ventaja sobre los embeddings extraídos de la primera capa, especialmente en la base de datos en español, y usando el modelo cased. Esta diferencia se debe a que las representaciones contextuales integran información semántica más rica y dependiente del contexto, permitiendo una mejor discriminación entre pacientes

con Alzheimer y controles sanos.

A pesar de las mejoras observadas con los embeddings contextuales, los resultados obtenidos en español fueron generalmente inferiores a los obtenidos en inglés. Esto podría deberse a la menor cantidad de datos disponibles, al desbalance de clases o a las diferencias en los modelos de BERT utilizados. En particular, el modelo multilingüe utilizado para los textos en español podría no capturar con la misma precisión los matices lingüísticos específicos del idioma en comparación con un modelo entrenado exclusivamente en inglés.

Como trabajo futuro, proponemos varias líneas de investigación para mejorar y ampliar este estudio. En primer lugar, dado que el rendimiento en Chile AD fue notoriamente inferior al obtenido en inglés, una posible mejora sería utilizar un modelo de BERT entrenado exclusivamente en español, en lugar de un modelo multilingüe. Esto podría generar embeddings más representativos del idioma y mejorar la clasificación. También sería interesante probar arquitecturas distintas a BERT para generar los embeddings o modelos híbridos que combinen embeddings semánticos con características estructurales del lenguaje. Por modelos híbridos nos referimos a aquellos que integran distintas fuentes de información lingüística: por un lado, los embeddings semánticos capturan el significado contextual de las palabras y frases; por otro lado, las características estructurales del lenguaje reflejan aspectos como la complejidad sintáctica, el uso de pausas, repeticiones, longitud de las oraciones o errores gramaticales. La combinación de ambos tipos de información puede enriquecer la representación del discurso y resultar especialmente útil en tareas como la detección de deterioro cognitivo, donde no solo importa lo que se dice, sino también cómo se dice. Un enfoque híbrido podría, por ejemplo, concatenar los vectores generados por modelos como BERT con vectores de características lingüísticas extraídas manual o automáticamente, alimentando esa representación conjunta a un modelo de clasificación.

Por otro lado, en el Pitt Corpus, si bien este trabajo utilizó únicamente la primera visita de cada participante, se podrían incluir visitas posteriores (segunda, tercera, última, o todas ellas), asegurando que los hablantes no se repitan entre los conjuntos de entrenamiento y prueba. Esto permitiría evaluar la estabilidad de los modelos a lo largo del tiempo y analizar cómo evoluciona el lenguaje en pacientes con Alzheimer. Además, se podría incluir en el análisis a los participantes con diagnóstico de Alzheimer que fueron excluidos del Pitt Corpus en este trabajo, aunque esto generaría un desbalance en las clases que requeriría técnicas específicas para su manejo. Asimismo, tanto en Pitt Corpus como en Chile AD, además de la tarea de descripción de Cookie Theft, existen otras tareas lingüísticas que podrían incorporarse al análisis, obteniendo así más datos de entrenamiento.

Otro aspecto a profundizar es el análisis de los embeddings generados. Aunque en este trabajo se exploró la separación de palabras individuales en el espacio de embeddings, sería valioso examinar las representaciones a nivel de texto completo para comprender mejor cómo los modelos capturan la estructura lingüística de los pacientes con Alzheimer.

También sería relevante reconsiderar la optimización de los hiperparámetros en clasificadores como SVM, particularmente en Chile AD, donde hubo casos en los cuales este predecía siempre la clase mayoritaria. En lugar de seleccionar los hiperparámetros que maximizan accuracy, se podría priorizar otra métrica, como ROC-AUC, o probar configuraciones alternativas de regularización y kernel.

Finalmente, una de las limitaciones más importantes de este estudio es el tamaño reducido de los conjuntos de datos utilizados. Ampliar la cantidad de datos, ya sea mediante la recolección de nuevas muestras o la integración de bases de datos adicionales, permitiría entrenar modelos más robustos y evaluar su capacidad de generalización de manera más

precisa.

En conclusión, este estudio confirma la relevancia del uso de modelos de lenguaje profundo para la detección de Alzheimer y destaca la importancia de elegir adecuadamente la representación del texto, las métricas de evaluación y los modelos lingüísticos utilizados en cada idioma.

# Apéndice

# A. RESULTADOS DE LA BÚSQUEDA DE HIPERPARÁMETROS

A continuación presentamos los resultados de las distintas búsquedas de hiperparámetros realizadas a lo largo de este trabajo. Para cada conjunto de embeddings se muestra una tabla con los mejores hiperparámetros para SVM y XGBoost, junto con mapas de calor mostrando el accuracy obtenido para cada combinación de hiperparámetros.

| **Clasificador** | **Idioma** | **Hiperparámetros** |
|---|---|---|
| SVM | Inglés | $C = 10000$, $\gamma = 0,001$ |
| | Español | $C = 0,0001$, $\gamma = 0,0001$ |
| XGBoost | Inglés | Número de estimadores = 50, Profundidad máxima = 5 |
| | Español | Número de estimadores = 50, Profundidad máxima = 1 |

Tab. A.1: Hiperparámetros óptimos para SVM y XGBoost, utilizando BERT-FIRST-ENG-CASED para el Pitt Corpus (inglés) y BERT-FIRST-MUL-CASED para Chile AD (español).

| **Clasificador** | **Idioma** | **Hiperparámetros** |
|---|---|---|
| SVM | Inglés | $C = 1$, $\gamma = 0,1$ |
| | Español | $C = 1000$, $\gamma = 0,001$ |
| XGBoost | Inglés | Número de estimadores = 50, Profundidad máxima = 7 |
| | Español | Número de estimadores = 50, Profundidad máxima = 3 |

Tab. A.2: Hiperparámetros óptimos para SVM y XGBoost, utilizando BERT-LAST-ENG-CASED para el Pitt Corpus (inglés) y BERT-LAST-MUL-CASED para Chile AD (español).

| **Clasificador** | **Idioma** | **Hiperparámetros** |
|---|---|---|
| SVM | Inglés | $C = 1000$, $\gamma = 0,01$ |
| | Español | $C = 1$, $\gamma = 1$ |
| XGBoost | Inglés | Número de estimadores = 70, Profundidad máxima = 1 |
| | Español | Número de estimadores = 50, Profundidad máxima = 3 |

Tab. A.3: Hiperparámetros óptimos para SVM y XGBoost, utilizando BERT-FIRST-ENG-UNCASED para el Pitt Corpus (inglés) y BERT-FIRST-MUL-UNCASED para Chile AD (español).

| **Clasificador** | **Idioma** | **Hiperparámetros** |
|---|---|---|
| SVM | Inglés | $C = 100$, $\gamma = 0,001$ |
| | Español | $C = 0,0001$, $\gamma = 0,0001$ |
| XGBoost | Inglés | Número de estimadores = 130, Profundidad máxima = 1 |
| | Español | Número de estimadores = 50, Profundidad máxima = 3 |

Tab. A.4: Hiperparámetros óptimos para SVM y XGBoost, utilizando BERT-LAST-ENG-UNCASED para el Pitt Corpus (inglés) y BERT-LAST-MUL-UNCASED para Chile AD (español).

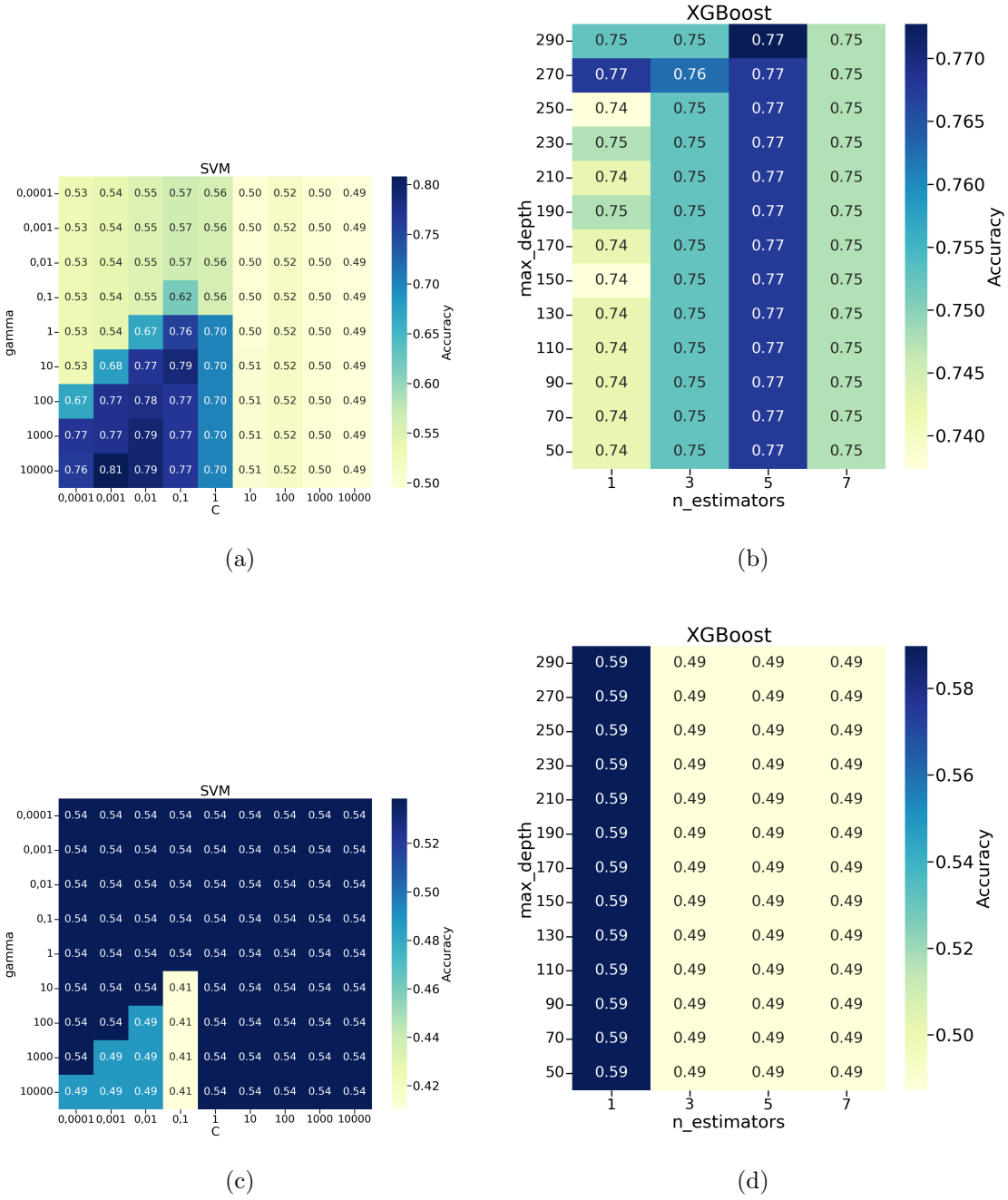
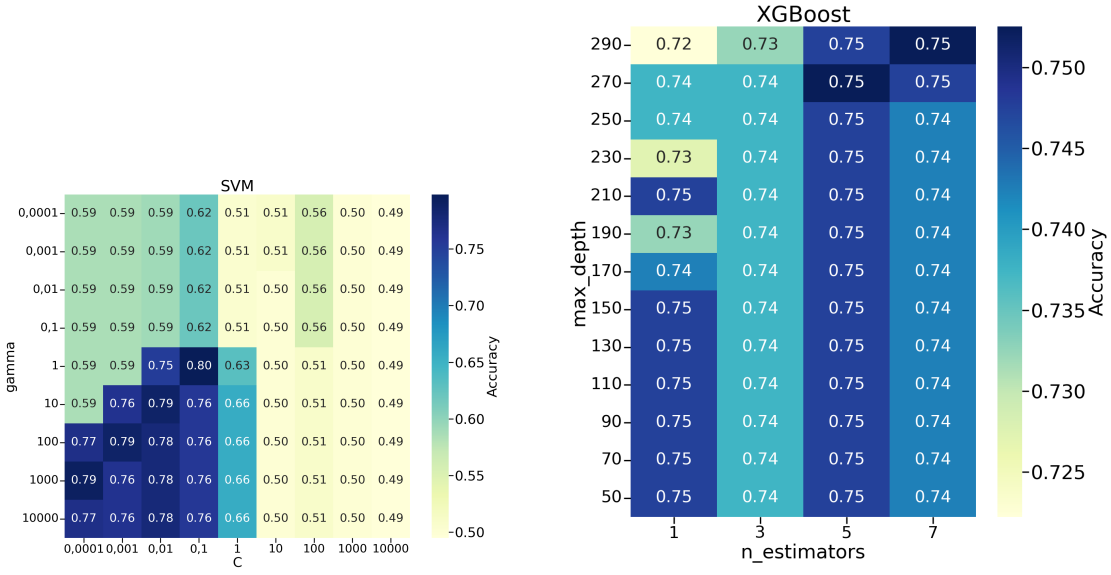



Fig. A.1: Accuracy para cada combinación de hiperparámetros utilizando embeddings BERT-FIRST-ENG-CASED para el Pitt Corpus (inglés) y BERT-FIRST-MUL-CASED para Chile AD (español). En (a) y (b) se muestran los resultados para el Pitt Corpus con SVM y XGBoost, respectivamente. En (b) y (c), se presentan los resultados para Chile AD con SVM y XGBoost, respectivamente. Cada celda indica la accuracy alcanzada para una configuración específica de hiperparámetros.




(a) (b)

(c) (d)

Fig. A.2: Accuracy para cada combinación de hiperparámetros utilizando embeddings BERT-LAST-ENG-CASED para el Pitt Corpus (inglés) y BERT-LAST-MUL-CASED para Chile AD (español). En (a) y (b) se muestran los resultados para Pitt Corpus con SVM y XGBoost, respectivamente. En (c) y (c), se presentan los resultados para Chile AD con SVM y XGBoost, respectivamente. Cada celda indica la accuracy alcanzada para una configuración específica de hiperparámetros.

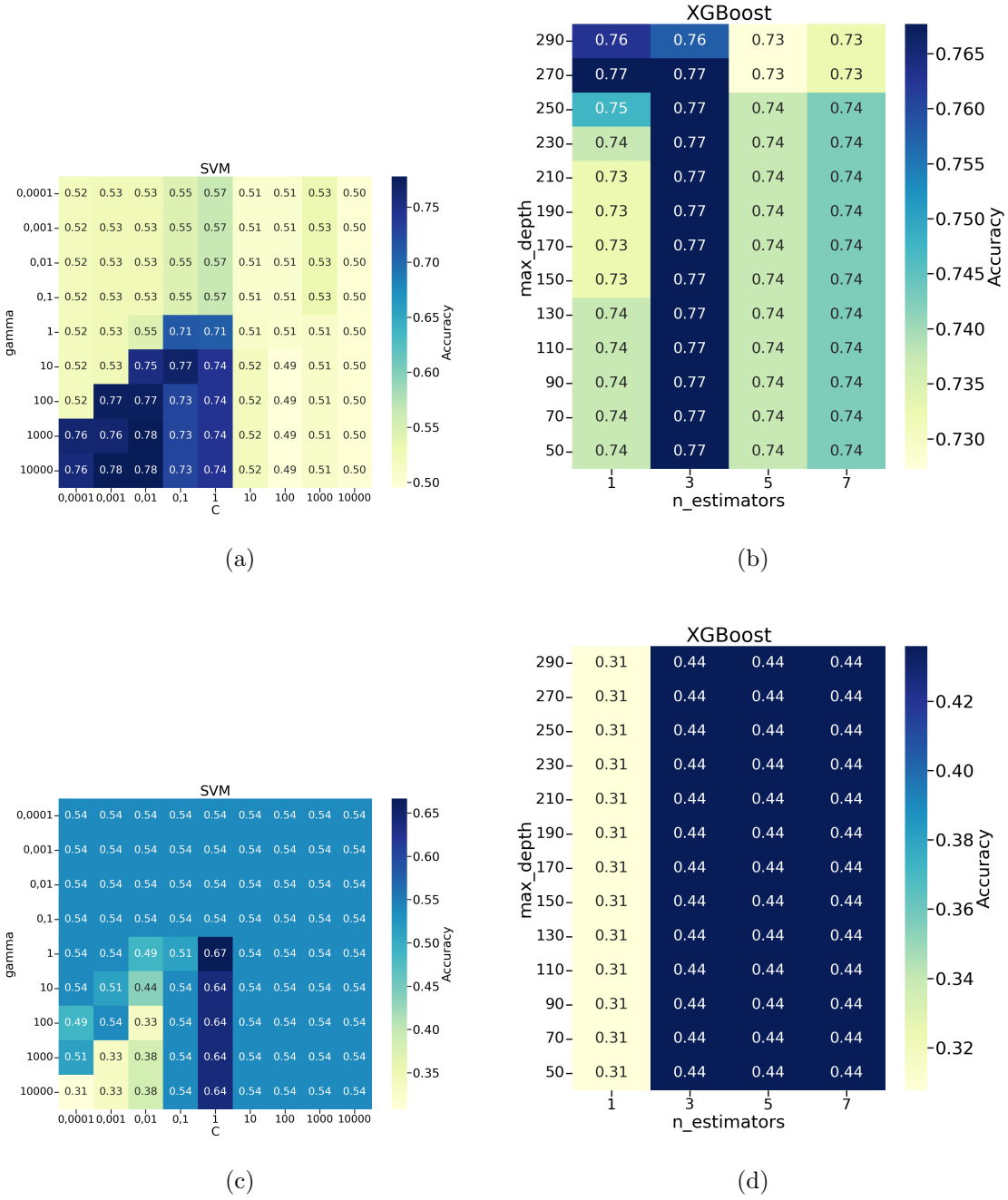
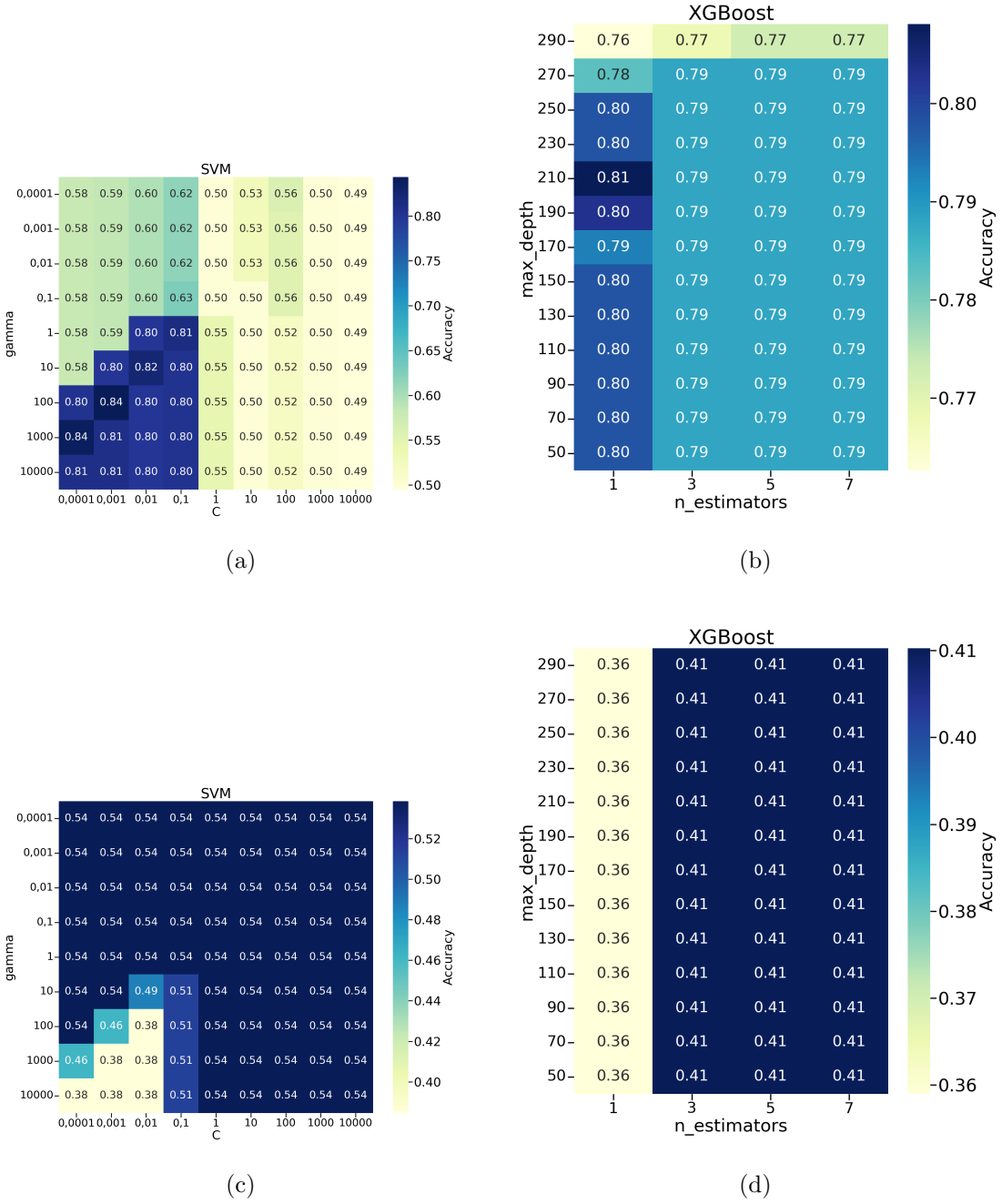



Fig. A.3: Accuracy para cada combinación de hiperparámetros utilizando embeddings BERT-FIRST-ENG-UNCASED para el Pitt Corpus (inglés) y BERT-FIRST-MUL-UNCASED para Chile AD (español). En (a) y (b) se muestran los resultados para Pitt Corpus con SVM y XGBoost, respectivamente. En (c) y (c), se presentan los resultados para Chile AD con SVM y XGBoost, respectivamente. Cada celda indica la accuracy alcanzada para una configuración específica de hiperparámetros.




(a) (b)

(c) (d)

Fig. A.4: Accuracy para cada combinación de hiperparámetros utilizando embeddings BERT-LAST-ENG-UNCASED para el Pitt Corpus (inglés) y BERT-LAST-MUL-UNCASED para Chile AD (español). En (a) y (b) se muestran los resultados para Pitt Corpus con SVM y XGBoost, respectivamente. En (c) y (c), se presentan los resultados para Chile AD con SVM y XGBoost, respectivamente. Cada celda indica la accuracy alcanzada para una configuración específica de hiperparámetros.

# B. RESULTADOS DE LOS TESTS DE PERMUTACIÓN

A continuación se muestran los resultados de los tests de permutación realizados para cada conjunto de embeddings BERT-FIRST-ENG-CASED y BERT-LAST-ENG-CASED en el Pitt Corpus, y BERT-FIRST-MUL-CASED y BERT-LAST-MUL-CASED en Chile AD. Se presentan los resultados de 10 ejecuciones de los tests de permutación, utilizando diferentes semillas para la separación de los datos. En cada subfigura se muestra la puntuación obtenida con las etiquetas reales, y se marcan en negrita aquellos p-valores menores o iguales a 0,05.

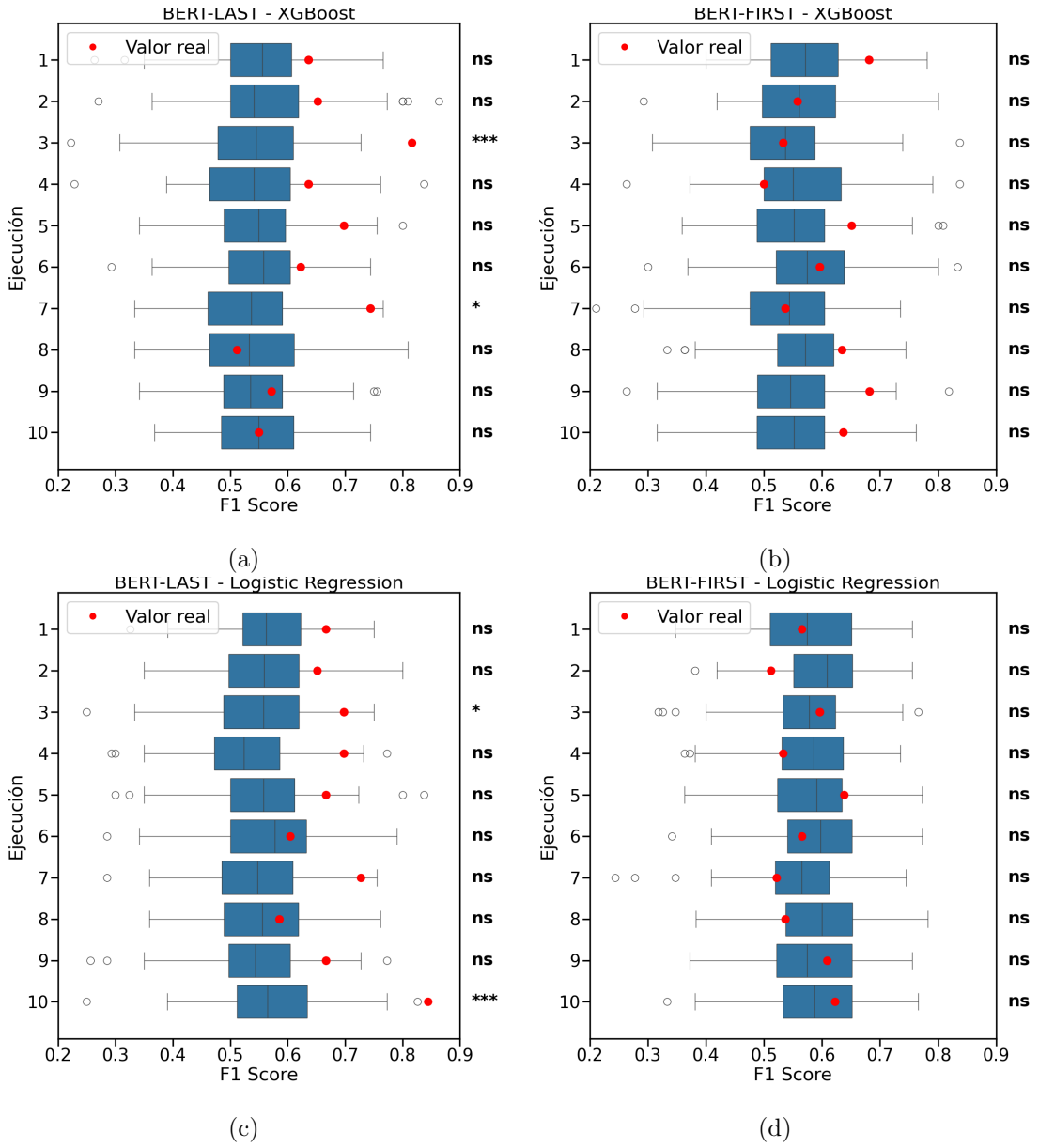



Fig. B.1: Resultados de 10 ejecuciones de tests de permutación con los datos de Chile AD, utilizando diferentes semillas para la separación de los datos. El punto rojo indica la puntuación obtenida con las etiquetas reales. En las subfiguras (a) y (b), se muestran los resultados para XGBoost con embeddings BERT-LAST-MUL-CASED y embeddings BERT-FIRST-MUL-CASED, respectivamente. En las subfiguras (c) y (d), se presentan los resultados para Logistic Regression con embeddings BERT-LAST-MUL-CASED y embeddings BERT-FIRST-MUL-CASED, respectivamente. Los p-valores asociados indican la significancia estadística de la puntuación observada en comparación con la distribución generada por permutaciones aleatorias, y se usan asteriscos (*) para denotar significancia estadística, utilizando más si el p-valor ($p$) es más chico. **ns**: $p > 0{,}05$; *: $p \leq 0{,}05$; **: $p \leq 0{,}01$; ***: $p \leq 0{,}001$.

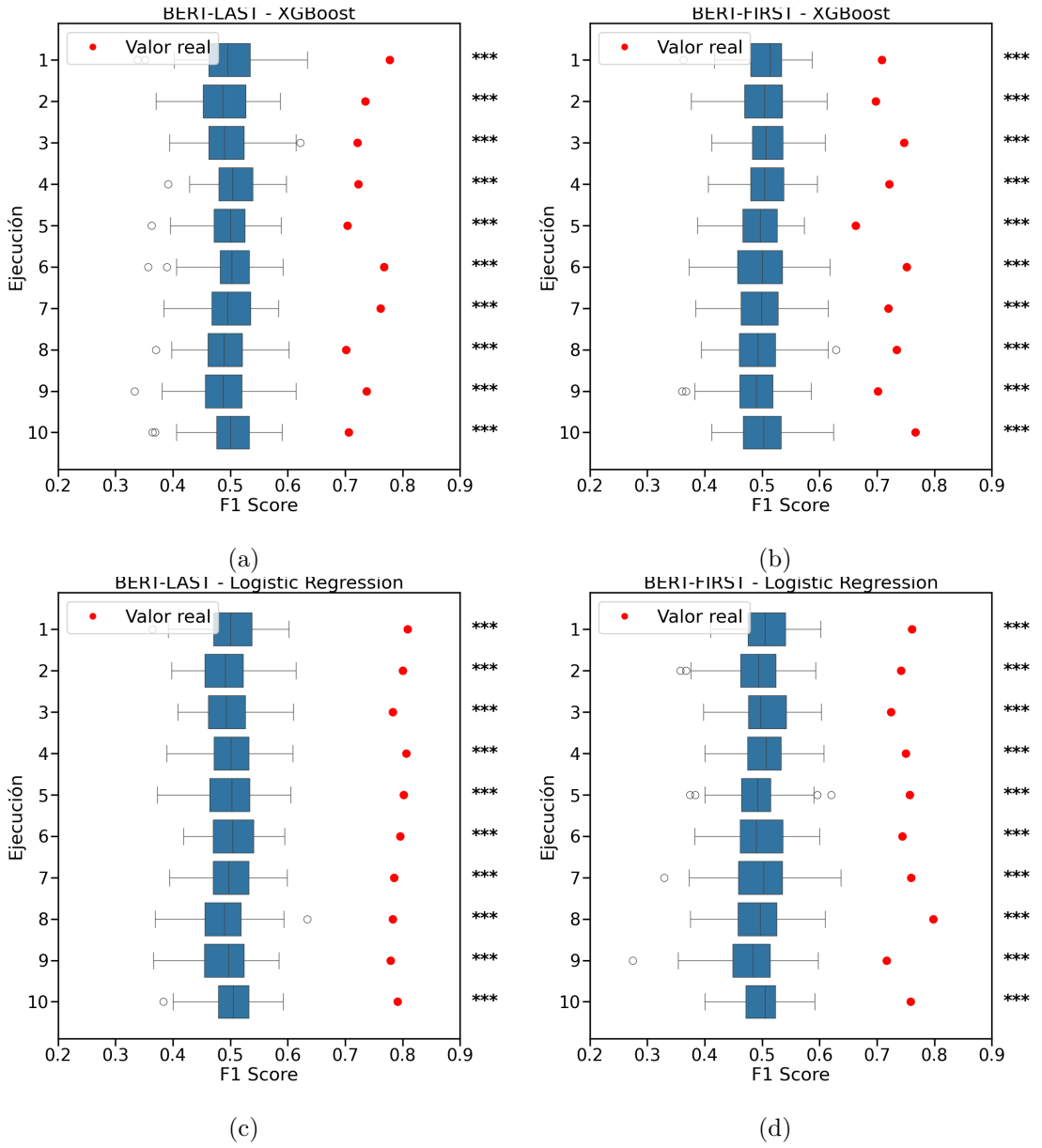



Fig. B.2: Resultados de 10 ejecuciones de tests de permutación con los datos del Pitt Corpus, utilizando diferentes semillas para la separación de los datos. El punto rojo indica la puntuación obtenida con las etiquetas reales. En las subfiguras (a) y (b), se muestran los resultados para XGBoost con embeddings BERT-LAST-ENG-CASED y embeddings BERT-FIRST-ENG-CASED, respectivamente. En las subfiguras (c) y (d), se presentan los resultados para Logistic Regression con embeddings BERT-LAST-ENG-CASED y embeddings BERT-FIRST-ENG-CASED, respectivamente. Los p-valores asociados indican la significancia estadística de la puntuación observada en comparación con la distribución generada por permutaciones aleatorias, y se usan asteriscos (*) para denotar significancia estadística, utilizando más si el p-valor ($p$) es más chico. **ns**: $p > 0{,}05$; *: $p \leq 0{,}05$; **: $p \leq 0{,}01$; ***: $p \leq 0{,}001$.

# C. RESULTADOS CON BERT UNCASED

A continuación, se presentan los resultados de la clasificación con todas las métricas de evaluación, utilizando los embeddings BERT-FIRST-ENG-UNCASED y BERT-LAST-ENG-UNCASED para el Pitt Corpus, y BERT-FIRST-MUL-UNCASED y BERT-LAST-MUL-UNCASED para Chile AD.

En la Tabla C.1 se encuentran los resultados obtenidos para el Pitt Corpus. Al comparar estos resultados con los de la Tabla 3.3, se observa que los resultados obtenidos con los embeddings BERT-FIRST-ENG-CASED y BERT-FIRST-ENG-UNCASED son muy similares. Si bien se nota una mejora con ROC-AUC, con las demás métricas pareciera ser indistinto usar la versión cased o uncased.

En contraste, la comparación entre los resultados obtenidos con los embeddings BERT-LAST-ENG-UNCASED y los embeddings BERT-LAST-ENG-CASED en el Pitt Corpus revela diferencias más marcadas. Aunque ambas versiones generan resultados competitivos, los embeddings BERT-LAST-ENG-UNCASED consistentemente muestran un rendimiento superior al de los embeddings BERT-LAST-ENG-CASED. Por ejemplo, el clasificador SVM alcanza un F1-score de 0,81 con BERT-LAST-ENG-UNCASED, en comparación con 0,78 con BERT-LAST-ENG-CASED, y mejoras similares se evidencian en métricas como ROC AUC y PR AUC. Esta diferencia sugiere nuevamente que, al trabajar con embeddings BERT-LAST, la versión uncased de BERT podría capturar de forma más robusta la información semántica relevante.

| Embeddings | Clasificador | F1-Score | Precision | Recall | Accuracy | Specificity | ROC AUC | PR AUC |
|---|---|---|---|---|---|---|---|---|
| | **Baseline** | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) | 0.50 (0.00) | 1.00 (0.00) | 0.50 (0.00) | 0.50 (0.00) |
| BERT-FIRST-ENG-UNCASED | **SVM** | **0.76 (0.02)** | 0.79 (0.03) | 0.73 (0.02) | **0.77 (0.02)** | 0.80 (0.04) | 0.84 (0.01) | 0.87 (0.01) |
| | **XGBoost** | 0.74 (0.03) | 0.77 (0.02) | 0.72 (0.03) | 0.75 (0.02) | 0.78 (0.03) | 0.82 (0.01) | 0.83 (0.02) |
| | **Logistic Regression** | 0.75 (0.02) | **0.80 (0.03)** | 0.71 (0.01) | **0.77 (0.02)** | **0.83 (0.03)** | 0.84 (0.01) | **0.88 (0.01)** |
| | **Random Forest** | **0.76 (0.02)** | 0.79 (0.02) | **0.74 (0.03)** | **0.77 (0.02)** | 0.80 (0.02) | **0.86 (0.01)** | 0.87 (0.01) |
| BERT-LAST-ENG-UNCASED | **SVM** | **0.81 (0.02)** | 0.83 (0.02) | 0.78 (0.02) | 0.81 (0.02) | 0.84 (0.03) | **0.90 (0.01)** | **0.91 (0.01)** |
| | **XGBoost** | 0.80 (0.02) | 0.81 (0.03) | 0.79 (0.03) | 0.80 (0.02) | 0.81 (0.03) | 0.87 (0.01) | 0.87 (0.02) |
| | **Logistic Regression** | **0.81 (0.01)** | **0.84 (0.02)** | 0.79 (0.02) | **0.82 (0.01)** | **0.85 (0.02)** | **0.90 (0.01)** | **0.91 (0.01)** |
| | **Random Forest** | **0.81 (0.01)** | 0.81 (0.01) | **0.81 (0.02)** | 0.81 (0.01) | 0.81 (0.02) | 0.88 (0.01) | 0.88 (0.01) |

Tab. C.1: Promedio y desvío estándar obtenidos para el Pitt Corpus, usando la versión uncased de BERT.

La Tabla C.2 muestra los resultados obtenidos para Chile AD. En los resultados de ambas representaciones se observa una notable disminución en el desempeño de la mayoría de los clasificadores, respecto a los resultados de la versión cased (3.4).

Usando los embeddings BERT-LAST-MUL-UNCASED, casi todas las métricas disminuyen notablemente, para todos los clasificadores, en comparación con los resultados obtenidos con la versión cased. En particular, SVM muestra un comportamiento similar al observado con BERT-FIRST-MUL-CASED, ofreciendo resultados equivalentes a los de un clasificador que predice sistemáticamente la clase mayoritaria. Al observar las predicciones, corroboramos que SVM predice siempre AD en este caso. Este fenómeno, al igual que en el de los embeddings BERT-FIRST-MUL-CASED, podría atribuirse tanto al tipo de kernel empleado como a la elección de hiperparámetros. Sin embargo, considerando el bajo desempeño obtenido con los demás clasificadores, pareciera que los embeddings BERT-LAST-MUL-UNCASED no logran separar claramente las dos clases, lo que provoca que SVM incline su frontera de

decisión hacia la clase mayoritaria.

| Embeddings | Clasificador | F1-Score | Precision | Recall | Accuracy | Specificity | ROC AUC | PR AUC |
|---|---|---|---|---|---|---|---|---|
| | **Baseline** | 0.70 (0.00) | 0.54 (0.00) | 1.00 (0.00) | 0.54 (0.00) | 0.00 (0.00) | 0.50 (0.00) | 0.54 (0.00) |
| BERT-FIRST-MUL-UNCASED | **SVM** | **0.74 (0.03)** | **0.60 (0.03)** | **0.95 (0.04)** | **0.64 (0.04)** | 0.27 (0.09) | 0.33 (0.05) | 0.45 (0.03) |
| | **XGBoost** | 0.48 (0.09) | 0.48 (0.07) | 0.49 (0.11) | 0.44 (0.07) | **0.38 (0.10)** | **0.42 (0.07)** | **0.52 (0.05)** |
| | **Logistic Regression** | 0.47 (0.08) | 0.44 (0.07) | 0.51 (0.10) | 0.39 (0.08) | 0.26 (0.09) | 0.34 (0.06) | 0.48 (0.03) |
| | **Random Forest** | 0.45 (0.07) | 0.44 (0.05) | 0.47 (0.10) | 0.40 (0.05) | 0.32 (0.06) | 0.34 (0.05) | 0.47 (0.04) |
| BERT-LAST-MUL-UNCASED | **SVM** | **0.70 (0.00)** | **0.54 (0.00)** | **1.00 (0.00)** | **0.54 (0.00)** | 0.00 (0.00) | **0.50 (0.03)** | **0.57 (0.03)** |
| | **XGBoost** | 0.46 (0.08) | 0.45 (0.07) | 0.48 (0.11) | 0.41 (0.08) | **0.33 (0.12)** | 0.42 (0.10) | 0.52 (0.06) |
| | **Logistic Regression** | 0.51 (0.07) | 0.47 (0.06) | 0.56 (0.09) | 0.43 (0.08) | 0.27 (0.08) | 0.38 (0.07) | 0.48 (0.03) |
| | **Random Forest** | 0.52 (0.07) | 0.49 (0.06) | 0.56 (0.08) | 0.45 (0.07) | **0.33 (0.08)** | 0.41 (0.08) | 0.50 (0.05) |

Tab. C.2: Promedio desvío estándar obtenidos para Chile AD, usando la versión uncased de BERT.

En el caso de los embeddings BERT-FIRST-MUL-UNCASED, el clasificador SVM experimenta una ligera mejora respecto a la versión cased, ya que ahora no parece estar prediciendo siempre AD, lo que se ve reflejado en la specificity (que pasó de ser nulo a 0,27) y en el recall (antes era máximo y ahora es de 0,95). Aunque parece obtener buenos resultados, si observamos los valores de ROC AUC y Precision-Recall AUC vemos que están muy por debajo del baseline, evidenciando nuevamente que F1-score no es una métrica adecuada en este caso, ya que está fuertemente influenciado por los valores de precision y recall. Generamos matrices de confusión (véase Figura C.1), donde se evidenció que SVM casi siempre predice la clase AD, reflejándose en un recall de 0,95, que su vez resulta en un F1-score de 0,76.

Los demás clasificadores (XGBoost, LR y RF) entrenados con los embeddings BERT-FIRST-MUL-UNCASED muestran un rendimiento considerablemente inferior a los de los embeddings BERT-FIRST-MUL-CASED, con una disminución general en casi todas las métricas, y casi todos los valores por debajo del baseline.

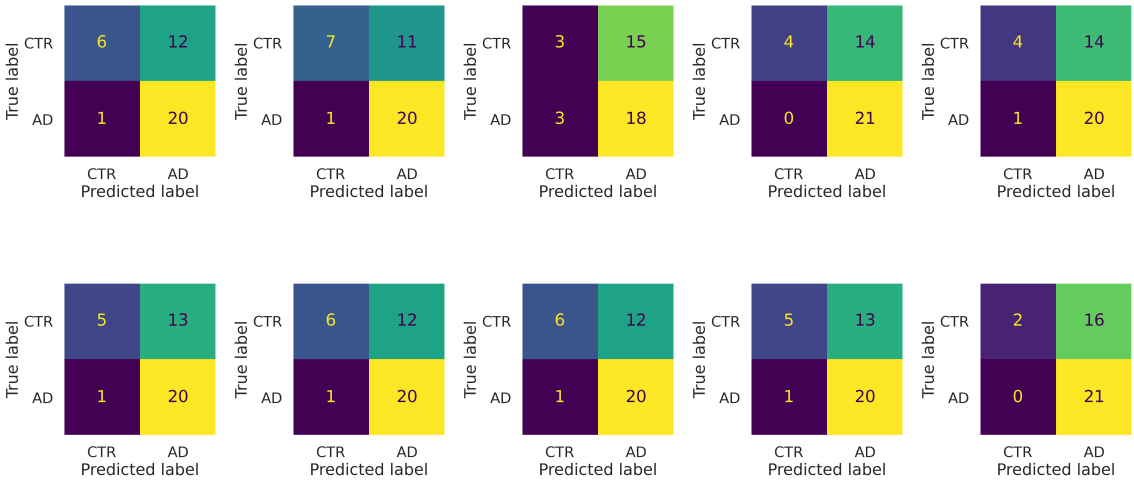



Fig. C.1: Matrices de confusión para cada iteración para el clasificador SVM entrenado con embeddings BERT-FIRST-MUL-UNCASED.

Si bien a primera vista SVM con los embeddings BERT-FIRST-MUL-UNCASED parece mejorar respecto a los resultados obtenidos con BERT-FIRST-MUL-CASED, este incremento se debe principalmente a la forma en que el clasificador maneja la clase mayoritaria, inflando así las métricas de F1-score y accuracy. En realidad, no existe una mejora genuina

en la capacidad de discriminación del clasificador, lo que pone de manifiesto la necesidad de complementar estas métricas con otras que reflejen mejor el equilibrio entre la identificación de casos positivos y negativos, especialmente en escenarios con pocos datos o clases desbalanceadas como este.

# D. ANÁLISIS DE EMBEDDINGS DE PALABRAS

A continuación presentamos visualizaciones de embeddings BERT-FIRST y BERT-LAST de distintos conjuntos de palabras, generados usando las versiones cased y uncased de los modelos de BERT. El objetivo de este experimento es observar cómo se agrupan las mismas palabras con las distintas representaciones, lo cual nos permite evaluar la capacidad de cada variante para capturar las relaciones semánticas intrínsecas entre términos.

Para representar las relaciones entre las palabras en un espacio bidimensional, aplicamos Análisis de Componentes Principales (PCA) (Jolliffe 2011) a los embeddings. PCA es una técnica de reducción de dimensionalidad que proyecta los datos en un espacio de menor dimensión, preservando la mayor cantidad posible de información. De esta manera, los gráficos resultantes nos permiten visualizar cómo cada modelo organiza las palabras en función de su similitud semántica.

Para llevar a cabo este análisis, usamos diversos conjuntos de palabras. Uno de ellos incluye palabras arbitrarias agrupadas en categorías definidas, como términos relacionados con la realeza ("rey", "reina", "príncipe", "princesa") y frutas ("banana", "manzana"), tanto en español como en inglés. Además, incorporamos dos conjuntos adicionales: uno con las 20 palabras más comunes en el corpus de Chile AD y otro con las 20 más frecuentes en el Pitt Corpus. Esto nos permitió comparar cómo se comportan las representaciones de BERT en diferentes contextos semánticos.

En la Figura D.1, mostramos las proyecciones de los embeddings BERT-FIRST para los conjuntos de palabras arbitrarias, en inglés y español. En las subfiguras (a) (BERT-FIRST-ENG-CASED) y (b) (BERT-FIRST-ENG-UNCASED), correspondientes al inglés, tanto BERT-FIRST-ENG-CASED como BERT-FIRST-ENG-UNCASED organizan las palabras de forma coherente con su significado, agrupando términos relacionados, como "king" y "queen" o "banana" y "apple". En cambio, en las subfiguras (c) (BERT-FIRST-MUL-CASED) y (d) (BERT-FIRST-MUL-UNCASED), correspondientes al español, observamos que en ninguno de los casos las las palabras se organizan de manera clara, y se agrupan términos semánticamente distintos en regiones cercanas, como "roca" y "príncipe" en el caso cased, o "roca" y "banana" en el caso uncased. Este resultado sugiere que, en el caso del español, los embeddings BERT-FIRST no logran capturar adecuadamente la información semántica de las palabras, independientemente de si el modelo es cased o uncased.

En la Figura D.2, presentamos las proyecciones de los embeddings BERT-LAST para los mismos conjuntos de palabras arbitrarias. Al igual que en el caso de los embeddings BERT-FIRST, las subfiguras (a) (BERT-LAST-ENG-CASED) y (b) (BERT-LAST-ENG-UNCASED) corresponden al inglés, mientras que (c) (BERT-LAST-MUL-CASED) y (d) (BERT-LAST-MUL-UNCASED) exhiben los resultados para el español. Para el inglés, tanto con los embeddings BERT-LAST-ENG-UNCASED como con los embeddings BERT-LAST-ENG-CASED, observamos que las palabras semánticamente afines (por ejemplo, "king", "queen", "prince"y "princess") se agrupan en clústeres definidos, pero de forma distinta. Por ejemplo, con BERT-LAST-ENG-UNCASED los términos de la realeza están ubicados más cerca entre sí, pero con BERT-LAST-ENG-CASED "dog" y "cat" se ven más cercanos que en la versión uncased. En cambio, en el caso de las palabras en español, si bien también se aprecia una agrupación por categorías semánticas (términos de la realeza, frutas o animales) con ambos modelos, la versión uncased muestra clústeres más definidos.

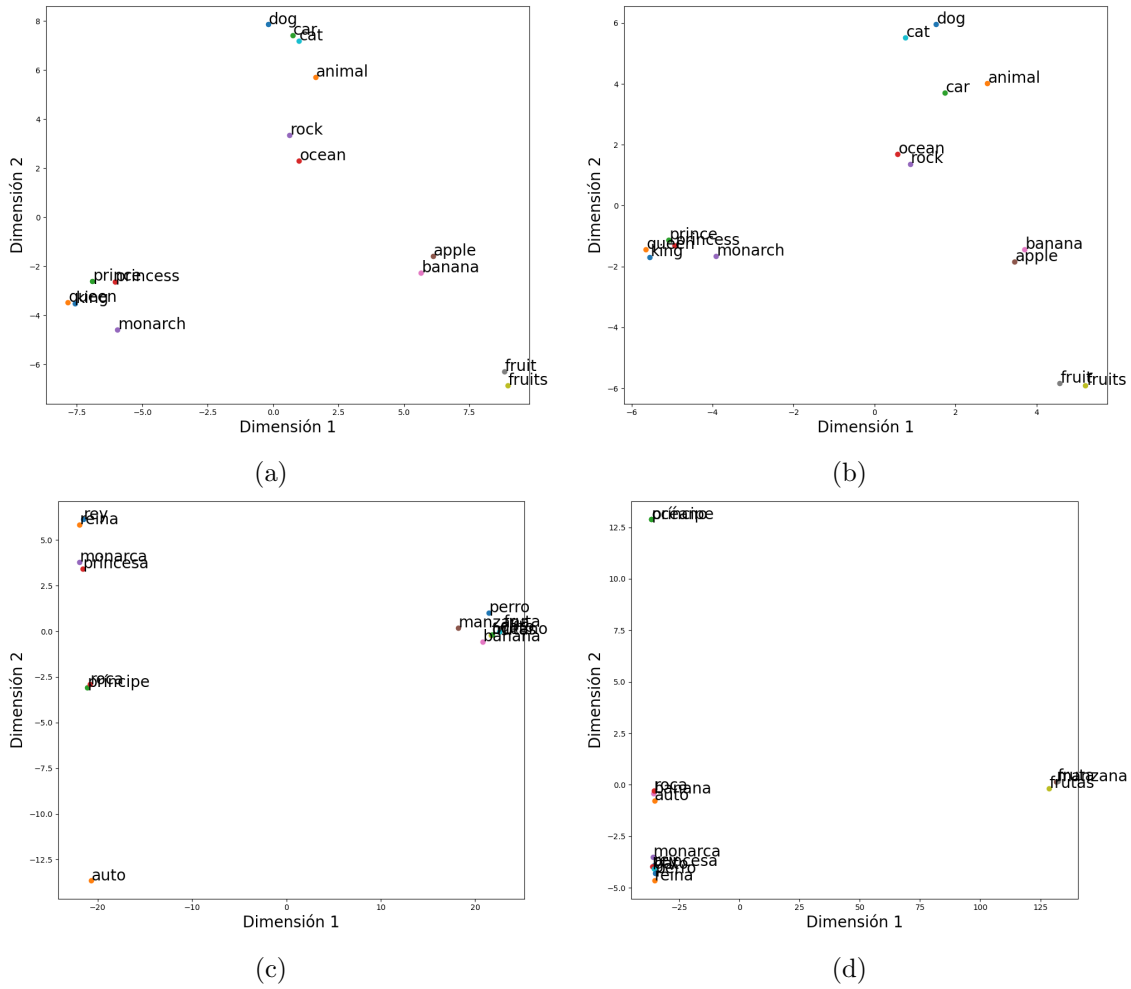



Fig. D.1: Visualización de embeddings BERT-FIRST de palabras en un espacio bidimensional obtenido mediante reducción de dimensionalidad con PCA. En (a) y (b) se muestran las proyecciones de los embeddings BERT-FIRST-ENG-CASED y BERT-FIRST-ENG-UNCASED de las palabras en inglés. En (c) y (d), se muestran las proyecciones de los embeddings BERT-FIRST-MUL-CASED y BERT-FIRST-MUL-UNCASED de las palabras en español.

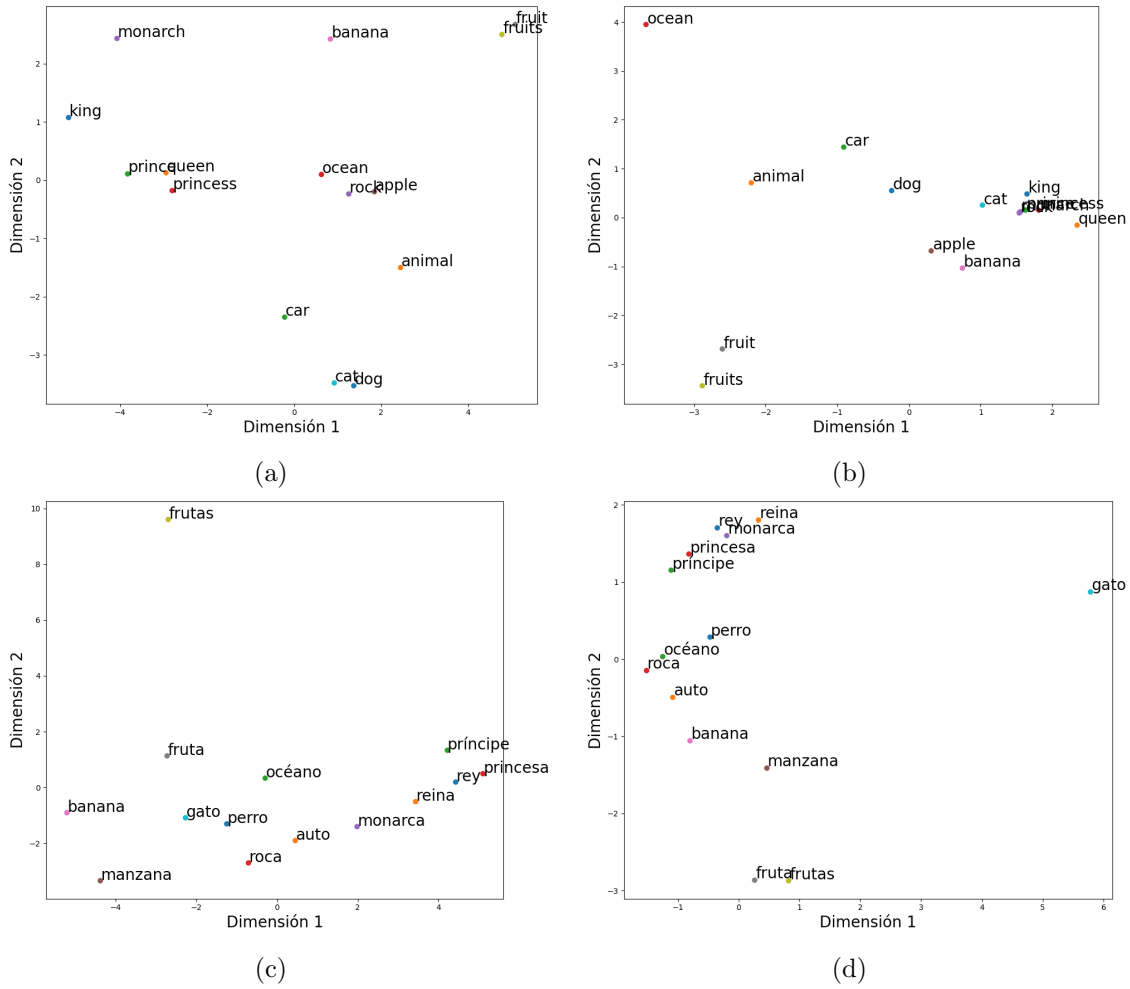



Fig. D.2: Visualización de embeddings BERT-LAST de palabras en un espacio bidimensional obtenido mediante reducción de dimensionalidad con PCA. En (a) y (b) se muestran las proyecciones de los embeddings BERT-LAST-ENG-CASED y BERT-LAST-ENG-UNCASED de las palabras en inglés. En (c) y (d), se muestran las proyecciones de los embeddings BERT-LAST-MUL-CASED y BERT-LAST-MUL-UNCASED de las palabras en inglés.

En resumen, el análisis comparativo de las proyecciones de los embeddings BERT-FIRST y BERT-LAST revela diferencias significativas en la capacidad de los modelos para capturar relaciones semánticas en inglés y español. Mientras que en inglés, tanto la versión cased como la uncased logran organizar palabras semánticamente afines en clústeres bien definidos, en español la agrupación es menos coherente, especialmente con los embeddings BERT-FIRST.

# BIBLIOGRAFíA


Arevalo-Rodriguez, Ingrid y col. (2015). «Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI)». En: *Cochrane database of systematic reviews* 3.

Baevski, Alexei y col. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. arXiv: 2006.11477 [cs.CL]. URL: https://arxiv.org/abs/2006.11477.

Balagopalan, Aparna y col. (2020). «To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection». En: *arXiv preprint arXiv:2008.01551*.

Becker, James T. y col. (jun. de 1994). «The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis». En: *Archives of Neurology* 51.6, págs. 585-594.

Breijyeh, Zeinab y Rafik Karaman (2020). «Comprehensive review on Alzheimer's disease: causes and treatment». En: *Molecules* 25.24, pág. 5789.

Breiman, Leo (2001). «Random Forests». En: *Machine Learning* 45.1, págs. 5-32. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

Chen, Tianqi y Carlos Guestrin (2016). «XGBoost: A Scalable Tree Boosting System». En: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

Chien, Yi-Wei y col. (2019). «An automatic assessment system for Alzheimer's disease based on speech using feature sequence generator and recurrent neural network». En: *Scientific Reports* 9.1, pág. 19597.

Cortes, Corinna y Vladimir Vapnik (1995). «Support-vector networks». En: *Machine Learning* 20.3, págs. 273-297.

Cox, D. R. (1958). «The Regression Analysis of Binary Sequences». En: *Journal of the Royal Statistical Society. Series B (Methodological)* 20.2, págs. 215-242. ISSN: 00359246. URL: http://www.jstor.org/stable/2983890 (visitado 09-02-2025).

Devlin, J. y col. (jun. de 2019). «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». En: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, págs. 4171-4186. URL: https://www.aclweb.org/anthology/N19-1423.

Devlin, Jacob y col. (2018). «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». En: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

Ferrer, Luciana (2025). «No Need for Ad-hoc Substitutes: The Expected Cost is a Principled All-purpose Classification Metric». En: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: https://openreview.net/forum?id=5PPbvCExZs.

Goodglass, H. y col. (1983). *Cookie Theft picture*. Philadelphia, PA: Lea & Febiger.

Gosztolya, Gábor y col. (2019). «Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features». En: *Computer Speech & Language* 53, págs. 181-197.

He, Kaiming y col. (2015). «Deep Residual Learning for Image Recognition». En: *CoRR* abs/1512.03385. arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

Jolliffe, Ian (2011). «Principal Component Analysis». En: *International Encyclopedia of Statistical Science*. Ed. por Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, págs. 1094-1096. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_455. URL: https://doi.org/10.1007/978-3-642-04898-2_455.

Julian, George (2020). «What are the most spoken languages in the world». En: *Retrieved May* 31.2020, pág. 38.

Knopman, David S y col. (2021). «Alzheimer disease». En: *Nature reviews Disease primers* 7.1, pág. 33.

Lee, Ya-Chen, Shu-Chun Lee y En-Chi Chiu (2022). «Practice effect and test-retest reliability of the Mini-Mental State Examination-2 in people with dementia». En: *BMC geriatrics* 22.1, pág. 67.

Liu, Yinhan y col. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. cite arxiv:1907.11692. URL: http://arxiv.org/abs/1907.11692.

Luz, Saturnino y col. (2020). «Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge». En: *Interspeech 2020*, págs. 2172-2176. DOI: 10.21437/Interspeech.2020-2571.

— (2021). *Detecting cognitive decline using speech only: The ADReSSo Challenge*. arXiv: 2104.09356 [eess.AS]. URL: https://arxiv.org/abs/2104.09356.

Mayeux, Richard y Yaakov Stern (2012). «Epidemiology of Alzheimer disease». En: *Cold Spring Harbor Perspectives in Medicine* 2.8, a006239. DOI: 10.1101/cshperspect.a006239.

Melistas, Thomas y col. (2023). «Cross-Lingual Features for Alzheimer's Dementia Detection from Speech». En: *Proceedings of Interspeech 2023*. ISCA, págs. 1-5. URL: https://www.isca-archive.org/interspeech_2023/melistas23_interspeech.pdf.

Ojala, Markus y Gemma C Garriga (2010). «Permutation tests for studying classifier performance». En: *Journal of Machine Learning Research* 11.Jun, págs. 1833-1863.

Omura, John D. y col. (2022). «Modifiable Risk Factors for Alzheimer Disease and Related Dementias Among Adults Aged ≥ 45 Years — United States, 2019». En: *MMWR. Morbidity and Mortality Weekly Report* 71, págs. 680-685. DOI: 10.15585/mmwr.mm7120a2. URL: http://dx.doi.org/10.15585/mmwr.mm7120a2.

Pedregosa, F. y col. (2011). «Scikit-learn: Machine Learning in Python». En: *Journal of Machine Learning Research* 12, págs. 2825-2830.

Perez-Toro, P. A. (jul. de 2020). *PauPerezT/WEBERT: Word Embeddings using BERT*. Ver. V0.0.1. DOI: 10.5281/zenodo.3964244. URL: https://doi.org/10.5281/zenodo.3964244.

Pérez-Toro, PA y col. (2023). «Automatic Assessment of Alzheimer's across Three Languages Using Speech and Language Features». En: *Proceedings of Interspeech 2023*. ISCA, págs. 1748-1752.

Pérez-Toro, Paula Andrea y col. (2022). «Alzheimer's Detection from English to Spanish Using Acoustic and Linguistic Embeddings.» En: *Interspeech*, págs. 2483-2487.

Qi, Xiaoke y col. (2023). «Noninvasive automatic detection of Alzheimer's disease from spontaneous speech: a review». En: *Frontiers in Aging Neuroscience* 15, pág. 1224723.

Rumelhart, David E., Geoffrey E. Hinton y Ronald J. Williams (1986). «Learning representations by back-propagating errors». En: *Nature* 323, págs. 533-536. URL: https://api.semanticscholar.org/CorpusID:205001834.

Sanz, Camila y col. (2022). «Automated text-level semantic markers of Alzheimer's disease». En: *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 14.1, e12276.

Simon, Roger P, Michael Jeffrey Aminoff, David A Greenberg y col. (2009). *Clinical neurology*. Vol. 20. Lange Medical Books/McGraw-Hill.

Wang, Yi y col. (2022). «Exploring linguistic feature and model combination for speech recognition based automatic ad detection». En: *arXiv preprint arXiv:2206.13758*.

Warnita, Tifani, Nakamasa Inoue y Koichi Shinoda (2018). «Detecting Alzheimer's disease using gated convolutional neural network from audio data». En: *arXiv preprint arXiv:1803.11344*.

Wu, Yonghui y col. (2016). «Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation». En: *CoRR* abs/1609.08144. arXiv: 1609.08144. URL: http://arxiv.org/abs/1609.08144.

Yang, Longfei y col. (2022). «Augmented Adversarial Self-Supervised Learning for Early-Stage Alzheimer's Speech Detection.» En: *INTERSPEECH*, págs. 541-545.