RAD-ALERT

Sistema de Identificación de Hallazgos Críticos en Informes Radiológicos Escritos en La Fundación Valle del Lili

Proyecto de Grado

Autores: Juan José Díaz Parra, Ana Sofía Londoño Fernández

Tutores: Angela Villota, Aníbal Sosa, Andrés Aristizábal

Facultad Barberi de Ingeniería, Diseño y Ciencias Aplicadas

Ingeniería de Sistemas

Santiago De Cali

2025

Tabla de contenidos

Resumen

La Fundación Valle del Lili es un referente en servicios de salud de alta complejidad en América Latina, con un gran volumen de estudios radiológicos que requieren una interpretación oportuna por parte de los radiólogos. La importancia de detectar y notificar de inmediato los hallazgos críticos, como hemorragias cerebrales o tromboembolismos, es fundamental para iniciar tratamientos que pueden salvar vidas. Sin embargo, la redacción de informes radiológicos en lenguaje natural y la dependencia de la notificación manual han generado demoras que ponen en riesgo la atención médica oportuna.

El problema central identificado es el retraso promedio superior a 2.8 horas en la notificación de hallazgos críticos en los reportes radiológicos no estructurados. Esta demora, de da por la ausencia de sistemas automatizados de alerta y la priorización manual de los informes, impactando directamente en la capacidad de respuesta clínica y la seguridad del paciente. Por tanto, surge la necesidad de desarrollar soluciones basadas en inteligencia artificial que puedan integrar y analizar el lenguaje natural de los informes, y notificar de manera precisa y oportuna los casos críticos.

Diversos trabajos previos han explorado la clasificación automática de reportes médicos, aplicando técnicas como minería de texto, SVM y modelos de regresión logística, principalmente en inglés y sin ajustarse a datos locales en español. Aunque han logrado avances importantes, persisten limitaciones para su implementación en instituciones de salud latinoamericanas, especialmente en la protección de datos clínicos sensibles y en la integración con sistemas hospitalarios existentes. Estas carencias subrayan la necesidad de soluciones adaptadas a contextos multilingües y normativas locales.

El aporte de este trabajo consiste en el desarrollo y validación de RAD-ALERT, un sistema de inteligencia artificial que emplea modelos de lenguaje natural, como RoBERTa biomédico, para clasificar hallazgos críticos en informes radiológicos en español. La metodología combinó CRISP-DM con iteraciones ágiles para optimizar el rendimiento y asegurar su integración clínica. El modelo final logró un recall de 0,92, latencias inferiores a medio segundo por informe y despliegue local aplicando una buena de protección de datos, ofreciendo una solución práctica, precisa y de bajo costo que fortalece la atención médica en la Fundación Valle del Lili.

Palabras clave: inteligencia artificial, hallazgos críticos, procesamiento de lenguaje natural, radiología, atención médica, Fundación Valle del Lili.

Abstract

RAD ALERT is an AI system developed to automatically identify critical findings in unstructured radiology reports at Fundación Valle del Lili. Leveraging NLP techniques and fine-tuned transformer models like RoBERTa, the system automates the classification of neuroimaging reports, helping to reduce diagnostic delays and improve patient outcomes. Integrated seamlessly with MIRTH and a RESTful API, the solution demonstrated a recall exceeding 93% using Gemini 2.0 Flash. Notably, the results indicate that lightweight models can rival the performance of larger LLMs, offering a fast, cost-effective, and scalable solution suitable for clinical environments.

Key words: Natural Language Processing, Critical Findings, AI-based Classification, Transformer Models, Clinical Integration

Lista de acrónimos

API	Interfaz de Programación de Aplicaciones (***A**pplication **P**rogramming **I**nterface*)
BERT	Representaciones de Codificador Bidireccional de Transformadores (***B**idirectional **E**ncoder **R**epresentations from **T**ransformers*)
FVL	Fundación Valle del Lili
HCE	Historia Clínica Electrónica
HIS	Sistema de Información Hospitalaria (Hospital Information System)
HL7	Nivel de Salud Siete (Health Level Seven)
IA	Inteligencia Artificial
MIRTH	Motor de Integración para la Reconciliación y Transformación de la Salud (*Mirth Connect*)
NLP	Procesamiento del Lenguaje Natural (*Natural Language Processing*)
PLN/PNL	Procesamiento de Lenguaje Natural
LLM	Gran Modelo de Lenguage (***L**arge **L**anguage **M**odels*)
CRISP-DM	Cross-Industry Standard Process for Data Mining: Estándar de minería de datos utilizado en el proyecto.
TF-IDF	Business Process Model and Notation: Notación para mapear procesos, mencionada en la comprensión del negocio.
SMOTE	*Synthetic Minority Over-sampling Technique*: Técnica para balanceo de clases usada para aumentar la clase minoritaria.
ROC	*Receiver Operating Characteristic*: Curva de rendimiento de clasificadores.
API REST	*Application Programming Interface Representational State Transfer*: Arquitectura de la API del proyecto.

RIS/PACS	*Radiology Information System / Picture Archiving and Communication System*: Sistemas de almacenamiento y gestión de estudios radiológicos.
UUID	*Universally Unique Identifier*: Identificador único utilizado para el control de registros.

Glosario de términos

1. **Inteligencia Artificial (IA):**

   Rama de la informática que se enfoca en la creación de sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje y la toma de decisiones.

2. **Procesamiento de Lenguaje Natural (PLN/PNL)**:

   Campo de la inteligencia artificial que se enfoca en la interacción entre computadoras y lenguaje humano, permitiendo que las máquinas comprendan, interpreten y generen lenguaje natural.

3. **Tokenización:**

   Proceso de segmentar un texto en unidades más pequeñas, o tokens, que pueden ser palabras, caracteres o subpalabras, para facilitar su procesamiento por algoritmos de lenguaje natural.

4. **Modelo de Clasificación:**

   Algoritmo de aprendizaje automático que asigna etiquetas o categorías a datos de entrada basándose en patrones aprendidos de datos previamente etiquetados.

5. **Modelo Predictivo:**

   Modelo estadístico o de aprendizaje automático utilizado para predecir resultados futuros basándose en datos históricos y patrones identificados.

6. **Arquitectura Transformer:**

   Modelo de redes neuronales que utiliza mecanismos de atención para procesar datos secuenciales, especialmente eficaz en tareas de procesamiento del lenguaje natural, permitiendo el análisis bidireccional del contexto.

7. **MIRTH Connect:**

   Herramienta de integración de datos de salud que facilita la interoperabilidad entre sistemas de información hospitalaria, permitiendo el intercambio y transformación de datos clínicos.

8. **Sistema de Información Hospitalaria (HIS):**

   Sistema de gestión que maneja datos administrativos, financieros y clínicos en un hospital, facilitando la operación diaria y la toma de decisiones médicas.

9. **LLM:**

Un LLM o un gran modelo de lenguaje hace referencia a un modelo de Inteligencia Artificial que es capaz de entender y generar lenguaje humano. Algunos ejemplos de estos pueden ser ChatGPT, Gemini o Claude.

10. **Hallazgo Crítico**

    Observación médica inesperada en un estudio radiológico que implica un riesgo inmediato para la vida del paciente y requiere intervención urgente.

11. **Radiología**

    Especialidad médica que utiliza técnicas de imagen como rayos X, tomografía computarizada o resonancia magnética para diagnosticar enfermedades.

12. **Informe Radiológico No Estructurado**

    Documento narrativo redactado por un radiólogo que describe hallazgos clínicos en lenguaje natural sin seguir un formato estandarizado.

13. **Oversampling (Sobremuestreo)**

    Técnica para balancear conjuntos de datos desbalanceados al incrementar artificialmente la cantidad de ejemplos de la clase minoritaria.

14. **Embedding**

    Representación vectorial de palabras o frases que captura su significado semántico y permite que los algoritmos los procesen computacionalmente.

15. **Fine-tuning**

    Ajuste de un modelo preentrenado a un conjunto de datos específico, con el fin de optimizar su rendimiento en una tarea particular.

16. **Latencia**

    Tiempo que tarda un sistema en procesar una entrada y generar una salida. En el contexto clínico, se refiere al tiempo de respuesta del modelo.

17. **FastAPI**

    Framework de desarrollo web en Python usado para construir APIs de alto rendimiento, especialmente útil para integrar modelos de machine learning.

18. **HL7 (Health Level Seven)**

    Conjunto de estándares internacionales para el intercambio de información clínica entre sistemas de salud.

19. Métricas de Evaluación

Indicadores como precisión, recall, F1-score y AUC-ROC que se utilizan para medir el rendimiento de los modelos de clasificación.

20. Validación Cruzada

Técnica de evaluación que divide los datos en subconjuntos para entrenar y probar el modelo de forma más robusta y evitar sobreajuste.

21. Token

Unidad mínima en que se divide un texto durante el preprocesamiento (puede ser una palabra, subpalabra o carácter).

22. Modelo RoBERTa Biomédico

Variante del modelo RoBERTa entrenada específicamente con textos médicos en español, especializada en procesamiento de lenguaje clínico.

Índice de figuras

Índice de tablas

1. Introducción

1.1 Contexto

La Fundación Valle del Lili, reconocida como una de las instituciones de salud más destacadas de América Latina, enfrenta desafíos significativos en el área de radiología. En esta unidad se realizan de forma constante estudios radiológicos a pacientes provenientes de diversas regiones, tanto aquellos que acuden por consulta externa como los que se encuentran hospitalizados. El análisis de estas imágenes requiere la intervención de un radiólogo especializado, quien debe interpretar los hallazgos visuales y redactar un informe detallado con las observaciones clínicas pertinentes.

Asimismo, es fundamental que el profesional determine, de manera precisa y oportuna, si el caso observado corresponde a una situación crítica o no crítica. Esta clasificación es crucial, ya que permite activar el protocolo de atención prioritaria, lo cual contribuye significativamente a reducir los tiempos de respuesta médica y mejorar el pronóstico del paciente. Un hallazgo crítico se define como una condición clínica nueva o inesperada que demanda una intervención médica inmediata, dado que representa una amenaza directa para la vida del paciente. Entre los ejemplos más frecuentes de estos hallazgos se incluyen hemorragias intracraneales, accidentes cerebrovasculares (ACV), herniación cerebral, tromboembolismo pulmonar (TEP), neumotórax, apendicitis, fracturas complejas e isquemia intestinal, entre otros.

No obstante, a pesar de la importancia de una detección oportuna, en la práctica clínica suelen presentarse demoras significativas en la notificación de hallazgos críticos dentro de los informes radiológicos no estructurados. Según Orejuela Zapata (2019), el tiempo promedio de notificación de estos hallazgos en pacientes hospitalizados es de aproximadamente 3.07 horas, mientras que en pacientes atendidos en servicios de urgencias es de 2.85 horas. Estas demoras representan un riesgo considerable, especialmente en situaciones en las que la vida del paciente depende de una intervención inmediata. Condiciones como hemorragias cerebrales o hematomas subdurales requieren una respuesta médica urgente, y cualquier retraso en su identificación puede comprometer gravemente el pronóstico clínico y la posibilidad de recuperación del paciente.

Una investigación llevada a cabo en el Departamento de Radiología de la Fundación Valle del Lili (Orejuela Zapata, 2019) examinó de manera específica esta problemática. El estudio recopiló datos de 1.949 hallazgos clínicos en estudios radiológicos realizados entre 2017 y 2019, analizando los tiempos de notificación al paciente. Se evidenció que los tiempos de espera se atribuyeron a la falta de conciencia por parte del personal no radiológico y a que las listas de lectura priorizada no siempre consideran la urgencia del hallazgo. Por ello, resulta imperativo implementar un sistema complementario que se encargue de la detección y notificación oportuna en este contexto.

La Fundación Valle del Lili ya ha empezado a tomar medidas para abordar este problema. En un estudio complementario, Orejuela Zapata y otros (2020) propusieron un método de aprendizaje automático para identificar hallazgos críticos en resonancias magnéticas cerebrales, demostrando que la inteligencia artificial puede ayudar a priorizar estudios y, en consecuencia, a notificar de manera más oportuna los hallazgos relevantes. Este estudio sienta una base para futuros desarrollos e impulsa la posibilidad de que nuevas soluciones basadas en inteligencia artificial puedan emplearse para la detección de hallazgos críticos en la Fundación Valle del Lili.

1.2 Planteamiento del problema

En la Fundación Valle del Lili, las demoras en la notificación de hallazgos críticos en reportes radiológicos no estructurados constituyen un peligro considerable para la vida de los pacientes, particularmente en situaciones de urgencia donde cada momento es vital. Estos periodos de espera, que en promedio exceden las 2.8 horas, se atribuyen principalmente a la ausencia de sistemas automatizados de detección y a la escasa priorización de las investigaciones en función de la seriedad de los descubrimientos. Esta situación, que representa una realidad más extensa del sistema sanitario colombiano, requiere de soluciones que incorporen tecnologías en auge como la inteligencia artificial para mejorar la detección temprana de diagnósticos críticos y optimizar la atención en el momento adecuado.

1.3 Impacto del proyecto

Este proyecto tiene un impacto directo en la calidad de la atención médica brindada por la Fundación Valle del Lili, al proponer una solución basada en inteligencia artificial que busca reducir los tiempos de comunicación de hallazgos críticos en los informes radiológicos. La implementación de esta tecnología podría contribuir significativamente a salvar vidas y a optimizar la eficiencia del servicio de radiología, mejorando así la capacidad de respuesta clínica y la calidad general de la asistencia sanitaria ofrecida por la institución.

1.4 Objetivo General

OG. Desarrollar un sistema de inteligencia artificial validado para la identificación temprana de casos críticos en informes radiológicos no estructurados, con el fin de reducir los tiempos de notificación en la detección de los casos en la Fundación Valle del Lili.

1.5 Objetivos Específicos

- **OE1.** Implementar un modelo de clasificación que, a partir de las características lingüísticas obtenidas, identifique los informes críticos.
- **OE2.**Validar la capacidad de generalización del modelo.
- **OE3.** Desarrollar un sistema compatible con el entorno clínico de la FVL que extraiga automáticamente la información relevante de los informes radiológicos no estructurados
- **OE4.**Evaluar el prototipo integrado en un entorno simulado de la FVL para analizar su funcionamiento.

1.6 Organización del Documento

A continuación, se describe la estructura del presente informe, con el propósito de guiar al lector a través de sus diferentes capítulos y facilitar la comprensión del desarrollo del proyecto.

- **Capítulo 2 – Antecedentes:** Expone el marco teórico y contextual que sustenta el proyecto. Se incluyen definiciones clave sobre hallazgos críticos, fundamentos del procesamiento de lenguaje natural (PLN), modelos de clasificación, métricas de evaluación, técnicas de balanceo de clases, interoperabilidad clínica y un análisis del estado del arte y de la práctica actual.

- **Capítulo 3 – Metodología:** Detalla el enfoque metodológico utilizado, basado en el estándar CRISP-DM combinado con iteraciones ágiles. Se explican las fases seguidas para la comprensión del negocio, preparación y análisis de los datos, modelado, evaluación, despliegue y retroalimentación continua.

- **Capítulo 4 – Presentación de la propuesta:** Describe paso a paso el desarrollo técnico del proyecto. Incluye el análisis exploratorio de datos, la estrategia de aumentación para corregir el desbalance de clases, el diseño y entrenamiento de modelos predictivos (desde los clásicos hasta LLMs como RoBERTa y Gemini), así como la integración de la solución mediante API.

- **Capítulo 5 – Validación y resultados obtenidos:** Presenta los resultados obtenidos en las pruebas del modelo y la API, comparando distintas métricas de desempeño y evaluando su eficacia en un entorno clínico simulado. Se justifica la selección del modelo final y se discute su aplicabilidad operativa.

- **Capítulo 6 – Conclusiones y trabajo futuro:** Resume los principales hallazgos del proyecto, destaca los aportes realizados y plantea recomendaciones para futuras etapas de desarrollo e implementación dentro de la Fundación Valle del Lili u otras instituciones de salud.

- **Referencias bibliográficas:** Reúne todas las fuentes académicas, técnicas y normativas citadas a lo largo del documento, siguiendo el formato de citación APA (7.ª edición).

2. Antecedentes

2.1. Marco Teórico

En este marco teórico se definirá qué son los hallazgos críticos en radiología y por qué su notificación rápida es clave. Luego se explicarán de forma breve las bases del Procesamiento de Lenguaje Natural para transformar informes textuales (no estructurados) en datos analizables y se presentarán explicaciones sobre los modelos de clasificación, qué tipos de modelos existen y sus métricas de evaluación. Finalmente, se mostrará cómo integrar la solución en entornos clínicos usando estándares de compatibilidad y garantizando el cumplimiento de la normativa de datos en Colombia.

2.1.1 Urgencia Clínica de los hallazgos críticos

En primer lugar, es importante establecer el **hallazgo crítico** como la definición principal en la que se basa este estudio. Un hallazgo crítico en radiología se define como cualquier hallazgo radiológico que pueda poner en peligro la vida del paciente, según la Joint Commission (2023), y que, por lo tanto, deba comunicarse urgentemente y requiera tratamiento médico inmediato. Además, es importante evaluar el **tiempo de notificación**, que es el tiempo transcurrido entre la generación de un informe radiológico crítico y su transmisión al médico tratante; este es un indicador de la calidad de la atención.

Desde una perspectiva técnica, el estudio se basa en informes radiológicos no estructurados, es decir, contenidos clínicos redactados por radiólogos que incluyen técnica, historia clínica, hallazgos y opinión, almacenados en los **HIS** (Sistema de Información Hospitalaria). Para este trabajo nos centraremos en dos modalidades de imagen: la **tomografía computarizada (TC)**, que obtiene múltiples cortes axiales del cuerpo mediante rayos X y reconstruye imágenes seccionales de alta resolución, y la **resonancia magnética (RM)**, que emplea campos magnéticos y pulsos de radiofrecuencia para generar imágenes detalladas de tejidos blandos sin utilizar radiación ionizante. Ambos estudios constituyen el corpus no estructurado sobre el cual aplicaremos las técnicas de Procesamiento del Lenguaje Natural.

2.1.2 Procesamiento de Lenguaje Natural

El **Procesamiento del Lenguaje Natural (PLN)** consiste en convertir texto libre (no estructurado) en una representación matemática (vectores numéricos) que los algoritmos pueden procesar (Jurafsky & Martin, 2020). Las principales tareas del PLN incluyen el preprocesamiento, con técnicas como **tokenización**, **limpieza de "stop words"** y **lematización,** pasos que, respectivamente, dividen el texto en tokens, eliminan palabras de poco valor semántico y normalizan cada término

a su forma base; la representación del texto en vectores, mediante técnicas clásicas como **Bolsa de Palabras**, que solo contabiliza si una palabra aparece o no, y **TF-IDF**, que puntúa cada palabra según su aparición relativa en el informe y en el conjunto de documentos; y el análisis/extracción de información (clasificación de textos, reconocimiento de entidades, análisis de sentimiento, etiquetado gramatical, etc.). Estas técnicas proporcionan mucha información crítica, pero, a menudo, no logran extraer significado de la importancia jerárquica contextualizada. Por lo tanto, esta deficiencia se convirtió en un problema para futuras investigaciones.

Por eso, se generaron **Emdeddings de palabras estáticas** (Word2Vec y GloVe), que colocan cada palabra individual como un vector dentro de un espacio semántico continuo. El mayor avance provino de los emdeddings contextuales de palabras, generadas mediante Transformer, que generan vectores situacionales basados en el texto circundante para determinar los diferentes significados de una palabra según su uso (Devlin y otros, 2019; Vaswani y otros, 2017). Cabe subrayar, además, que la mayor parte de los recursos y modelos preentrenados se encuentran en inglés, lo que abre una brecha en el PLN para el español y limita la transferencia directa de estos avances a nuestro idioma (Garcia, 2023).

2.1.3 Clasificación de reportes

Para clasificar los reportes, se recurre al **aprendizaje supervisado**, un paradigma de machine learning en el que un algoritmo aprende a partir de ejemplos de entrada emparejados con sus etiquetas, de modo que pueda predecir la categoría correcta en datos nuevos (Mitchell, 1997). Con ese conocimiento, se construye un **modelo de clasificación**, es decir, un algoritmo de aprendizaje supervisado que, tras entrenarse con reportes ya etiquetados (por ejemplo, "crítico" vs. "no crítico"), asigna automáticamente la etiqueta apropiada a cada nuevo informe.

Para evaluar la eficiencia, precisión, interpretabilidad y requerimientos computacionales, comenzaremos explorando modelos estadísticos clásicos: **regresión logística** que es un clasificador lineal que estima la probabilidad de cada clase mediante la función sigmoide; destaca por la interpretabilidad de sus coeficientes y sus bajos requerimientos computacionales. **Support Vector Machines (SVM)** buscan el hiperplano que maximiza el margen de separación entre clases, pudiendo proyectar los datos a espacios de mayor dimensión con núcleos (kernels) para resolver problemas no lineales. **Naive Bayes** aplica el Teorema de Bayes bajo la asunción de independencia condicional entre las características, lo cual le permite entrenar en tiempo casi instantáneo y obtener resultados sorprendentemente robustos en clasificación de texto.

También, encontramos las **LSTM bidireccionales** que capturan relaciones de largo alcance procesando la secuencia en ambos sentidos, lo que ayuda a comprender referencias clínicas dispersas en el texto. Por otro, los **Transformers** y grandes modelos de lenguaje (ej. BERT, RoBERTa, GPT-4) emplean mecanismos de auto-atención para ponderar la influencia mutua de cada palabra, ofreciendo una comprensión contextual profunda y generalizable a distintas tareas de texto.

2.1.4 Evaluación de clasificadores

Para evaluar el éxito del clasificador, existen las **matrices de confusión** que permiten distinguir entre las predicciones en cuatro categorías: **verdaderos positivos** (informes críticos correctamente determinados que conforman las alertas que justifican el éxito); **falsos positivos** (informes que no producen hallazgos rápidos, pero que el sistema clasifica como críticos); **verdaderos negativos** (informes que no producen resultados críticos que el enfoque de IA identifica con precisión); y **falsos negativos** (evaluaciones críticas que el motor de IA no reconoce y que son las más peligrosas para la salud del paciente).

A partir de esta matriz, se evaluarán diversas características de rendimiento. La exactitud (**Accuracy**) mide la proporción de predicciones correctas (tanto críticas como no críticas) sobre el total de casos. La precisión (**Precision**) evalúa la eficacia de las alertas verdaderas en función de cuántas veces, cuando el sistema marca un hallazgo crítico, este realmente lo es; esta métrica reduce los falsos positivos. La sensibilidad (**Recall**) determina su eficacia para detectar todos los casos críticos y garantizar que muy pocas emergencias clínicas pasen desapercibidas. La puntuación F1 (**F1-score**) se utiliza cuando la precisión y la sensibilidad son indicadores de rendimiento viables y deben evaluarse conjuntamente. Finalmente, también se evaluará el AUC del ROC (ROC AUC) en la curva ROC (ROC curve), que mide cómo fluctúa la tasa de verdaderos positivos (True Positive Rate) frente a la tasa de falsos positivos (False Positive Rate) al modificar el umbral de clasificación.

2.1.5 Gestión del desequilibrio de clases

En la práctica clínica, suele haber muchos más informes sin hallazgos graves que con ellos, lo que puede hacer que el sistema pase por alto los casos más urgentes. Para reducir este riesgo, aplicaremos dos soluciones complementarias: primero, **SMOTE** (Chawla y otros, 2002), que crea ejemplos sintéticos de la clase minoritaria al interpolar entre casos reales, y segundo, un aumento de texto con grandes modelos de lenguaje, que genera paráfrasis clínicas de hallazgos críticos para aportar variedad lingüística sin perder el significado médico.

Además, se utilizará una forma de aumento de texto para abordar aún más este desequilibrio: los **Modelos de Lenguaje Amplio** (MLA) crean paráfrasis clínicas de hallazgos críticos, conservando intactos el significado y la jerga médica relevante. Esta paráfrasis automatizada puede añadir una capa de diversidad lingüística natural a la clase minoritaria, permitiendo al clasificador comprender las condiciones críticas de diversas maneras sin necesidad de replicación literal.

2.1.6 Interoperabilidad y privacidad

En cuanto a la interoperabilidad clínica, la solución se basará en la consistencia del estándar **HL7 FHIR**, utilizando recursos basados en texto como DiagnosticReport para generar informes radiológicos y Flag para alertas críticas. La gestión de mensajes de estos mensajes se realiza a través de **Mirth Connect**, un motor de integración de código abierto que puede transformar y enrutar mensajes HL7 y JSON/REST según reglas configurables. Desde una perspectiva legislativa, esta aplicación cumplirá con la **Ley 1581 de 2012** de Colombia para proteger el manejo y la divulgación de la información personal identificable del paciente.

Por lo tanto, este marco teórico de conceptos, aplicaciones, teorías y estándares proporciona un amplio marco a lo largo de esta investigación para la construcción y evaluación uniforme de un sistema que busca identificar y señalar hallazgos críticos en los informes radiológicos mediante mejoras en el procesamiento del lenguaje natural.

2.2. Estado del arte/trabajos relacionados

A continuación, se presentan diversos estudios relevantes en el ámbito de la clasificación clínica de reportes médicos entre casos críticos y no críticos. En la tabla se detallan el enfoque temático de cada trabajo, las técnicas utilizadas y una breve descripción de su contribución:

Tabla 1. Comparativa de enfoques NLP en clasificación de reportes médicos

Autor	Fecha	Abordaje	Técnica	Descripción
Orejuela et al.	2020	Neuro	Minería de texto / SVM	Clasificación automática de reportes clínicos neurológicos utilizando NLP y máquinas de soporte vectorial (SVM), alcanzando alta precisión y VPN superior al 94%.
Zech et al.	2018	Neuro (Imagenología)	Minería de texto / Regresión logística	Estudio centrado en reportes neurorradiológicos. Utiliza técnicas de NLP y modelos estadísticos para identificar hallazgos importantes. Algunos trabajos de Zech abordan sesgos en modelos de diagnóstico radiológico.
Lakhani et al.	2012	Neumonía / TEP	Minería de texto	Se desarrolló un sistema basado en procesamiento de lenguaje natural (NLP) para detectar automáticamente resultados críticos, como neumonía o tromboembolismo pulmonar, en reportes radiológicos.

Meng et al.	2009	General (publicaciones biomédicas)	Clustering de texto	Se aplicaron técnicas de agrupamiento no supervisado a millones de publicaciones médicas. No enfocado directamente en reportes clínicos, pero sí en exploración de texto biomédico.

La tabla anterior presenta los avances más recientes en el campo de la clasificación automatizada de informes clínicos en críticos y no críticos. En general, los estudios abarcan desde la clasificación exploratoria mediante agrupamiento no supervisado (Meng et al., 2009) hasta estrategias más definidas que utilizan minería de texto y enfoques entrenados como la regresión logística (Zech et al., 2018) y SVM (Orejuela et al., 2020). Además, los estudios investigan diferentes tipos de hallazgos en diversos entornos clínicos, desde hallazgos pulmonares hasta informes de imágenes cerebrales. Por lo tanto, parece haber consenso en la aplicación del PLN en los últimos años para facilitar la extracción y las evaluaciones automatizadas de informes médicos. En general, los hallazgos muestran una tendencia al alza en un campo relativamente nuevo, con una precisión y viabilidad consistentemente mayores en aplicaciones clínicas reales.

2.3. Estado de la práctica

Actualmente, la **notificación de resultados críticos en radiología** presenta dificultades. Existe un requisito legislativo para una comunicación continua del envío, la recepción y el acuse de recibo de la orden; sin embargo, la mayoría de los hospitales aún utilizan el teléfono o los mensajes de texto para notificar al médico tratante o al profesional (The Joint Commission, 2019). Además, el **Objetivo Nacional de Seguridad del Paciente NPSG.02.03.01**, emitido por la Joint Commission, exige que los resultados de pruebas críticas se transmitan de manera oportuna, lo que significa que este proceso debe llevarse a cabo para evitar que los pacientes sean innecesariamente vulnerables a riesgos (The Joint Commission, 2024).

En cuanto a las soluciones actuales, existen tres opciones. El **flujo de trabajo manual** principal es por teléfono o correo electrónico; para los hallazgos más importantes (gravedad vital), la llamada es en menos de una hora; sin embargo, la confirmación y el envío dependen de la lectura del personal de enfermería (The Joint Commission, 2019; AHRQ PSNet, 2008). Los **módulos comerciales de CTRM (Critical Test Results Management)** cuentan con alertas visuales ("alerta roja") que cambian de color y, posteriormente, escalan automáticamente la transmisión del hallazgo con un registro de auditoría de cumplimiento de la verificación (Nuance Communications, 2022). Los centros que aplican la metodología **Lean** (un enfoque para simplificar procesos y eliminar pasos innecesarios) junto con las guías de seguridad del paciente publicadas en **PSNet** han logrado acortar el tiempo que pasa entre detectar un hallazgo grave en una radiografía y avisar al médico. Sin embargo, cada hospital mide ese tiempo de forma distinta y además valida los resultados a mano, por lo que no existe una forma sencilla de estandarizar el proceso en todas partes (Verbano, 2019).

Respecto a los **estándares e integración**, el uso de **HL7 v2 y FHIR** es fundamental: el recurso DiagnosticReport junto con el perfil US Core permite adjuntar interpretaciones codificadas en SNOMED CT, mientras que el recurso Flag registra la criticidad de forma estructurada (FHIR v6.0.0, 2024). Asimismo, los **motores de integración** como Mirth Connect enrutan mensajes HL7 y JSON/REST entre RIS, HIS y motores de IA, aplicando reglas de escalamiento y manteniendo trazabilidad de cada alerta (OSP Labs, 2024).

A pesar de estas iniciativas, en la **Fundación Valle del Lili (FVL)** el proceso sigue siendo mayormente telefónico y dependiente de la evaluación manual del radiólogo. No existen soluciones de módulos de clasificación automática adaptados a informes en español, lo cual coincide con la literatura que identifica una **brecha** de sistemas de alerta abiertos, de bajo costo y entrenados con corpus latinoamericanos (The Joint Commission, 2019; American College of Radiology, 2024).

3. Metodología

3.1 Introducción a la metodología

Este proyecto sigue una metodología conforme al estándar **CRISP-DM**. CRISP-DM, o Proceso Estándar Intersectorial para Minería de Datos, es un método internacionalmente reconocido y aceptado para la ciencia de datos que proporciona un procedimiento robusto y sistemático para guiar un proyecto desde la definición del problema hasta la implementación de los hallazgos en seis áreas clave:

1. **Comprensión del negocio:** identificación y análisis profundo del contexto clínico y operacional.
2. **Comprensión de los datos:** exploración y análisis inicial de los datos disponibles.
3. **Preparación de los datos:** limpieza, transformación y adaptación del conjunto de datos.
4. **Modelado:** aplicación de técnicas analíticas y creación de modelos predictivos.
5. **Evaluación:** validación rigurosa del desempeño y ajuste de los modelos.
6. **Despliegue:** implementación práctica del modelo seleccionado dentro del entorno operativo.

El ciclo completo de CRISP-DM duró aproximadamente tres meses, desde la comprensión inicial del contexto clínico hasta la determinación del modelo adecuado para detectar hallazgos críticos en los informes radiológicos. Tras una clara definición del modelo elegido, se seleccionó un marco de proyecto tipo Kanban/Scrum (Kanban siendo un sistema visual de gestión de flujo continuo mediante tarjetas y columnas, y Scrum es un marco ágil de desarrollo iterativo basado en sprints con roles y ceremonias definidos) para el enfoque adicional, orientado al desarrollo incremental. Un marco formalizado como Scrum o Kanban no era apropiado debido al alcance limitado del prototipo. Por lo tanto, el desarrollo incremental consistió en pequeñas iteraciones incrementales específicas de aproximadamente una semana de duración cada una. Durante aproximadamente un mes, se llevaron a cabo desarrollos incrementales del prototipo, cada uno caracterizado por lo siguiente:

- **Implementación controlada** en un entorno aislado.
- **Pruebas automatizadas** (unitarias, de integración y de rendimiento).
- **Revisión y retroalimentación** de los tutores.

La intersección de ambas metodologías proporcionó el equilibrio perfecto entre dos componentes esenciales: por un lado, el rigor, la trazabilidad y la sistemática del CRISP-DM; por otro, la flexibilidad, la rapidez y la retroalimentación preliminar de los sprints ágiles y gestor visual del flujo de trabajo, especialmente adecuados para proyectos breves y de pequeña escala, como fue el caso del prototipo desarrollado. El diagrama de proceso (Figura 1) a continuación muestra dicha intersección de los dos enfoques metodológicos. Los círculos indican las fases principales del CRISP-DM, mientras que los rectángulos aclaran los objetivos que cumplen, las acciones específicas realizadas, los entregables generados y las fases de enlace que perfeccionaron continuamente la solución a los requisitos clínicos y operativos de los entornos hospitalarios.

3.2 Descripción detallada de la metodología

La Figura 1 muestra el diagrama completo del proceso metodológico aplicado en el proyecto, que relaciona cada fase de CRISP-DM con las actividades desarrolladas, entregables generados, objetivos específicos (OE) cumplidos y las retroalimentaciones necesarias para la mejora continua.

3.2.1 Comprensión del negocio

Se llevaron a cabo sesiones con los stakeholders clínicos y técnicos para definir claramente el problema, identificar puntos críticos del proceso y traducir estas necesidades en requisitos concretos. Actividades principales incluyeron mapeos de historias de usuarios y priorización mediante MoSCoW. El resultado fue un documento formal de requisitos y un mapa de procesos objetivo, alineados a los objetivos específicos **OE1, OE3 y OE4**.

3.2.2 Comprensión de los datos

En esta fase, nos familiarizamos con los datos compartidos por los stakeholders, realizando un análisis exploratorio y evaluando calidad, completitud y consistencia de la información. Los entregables fueron el informe de perfilado de datos y un diccionario de datos detallado, vinculándose a los objetivos **OE1 y OE2**.

3.2.3. Preparación de los datos

Se transformaron los datos originales mediante técnicas de anonimización para cumplir la Ley 1581, normalización del texto, tokenización y balanceo de clases con métodos como SMOTE y aumentación textual. Se obtuvo un conjunto de datos limpio y balanceado, documentado con scripts y configuraciones claras para trazabilidad, cumpliendo así el objetivo **OE1**.

3.2.4. Modelado

Se adoptó una estrategia incremental, iniciando con modelos estadísticos clásicos (Regresión Logística, Naive Bayes y SVM), seguidos por modelos neuronales (Bi-LSTM con atención), para finalizar con modelos avanzados basados en Transformers y LLMs. Cada experimento se registró en un sistema de versionamiento con artefactos y métricas claras y explicado claramente en la presentación de la propuesta. Conectado con objetivos **OE1 y OE2**.

3.2.5. Evaluación

Se ejecutaron diversas pruebas (hold-out interno y externo), curvas ROC y PR, y ajustes de umbral orientados a sensibilidad clínica. También se realizaron pruebas del desempeño de clasificación dentro del proceso de trabajo completo. Los resultados obtenidos sirvieron para generar un informe técnico y un plan de mejoras, atendiendo objetivos **OE2 y OE4**.

3.2.6. Despliegue e integración

El modelo seleccionado fue desplegado a través de un flujo HL7 → Mirth Connect → API REST (FastAPI). Se documentaron todos los endpoints, se desarrollaron pruebas automáticas y se

implementó un sistema de monitoreo continuo para latencia y estabilidad operativa, logrando los objetivos específicos **OE3 y OE4**.

3.2.7. Evaluación continua y retroalimentación

Finalmente, la metodología contempla un ciclo continuo de mejora mediante la retroalimentación directa desde la evaluación hasta la comprensión del negocio, permitiendo ajustar la solución según nuevos hallazgos o requisitos.

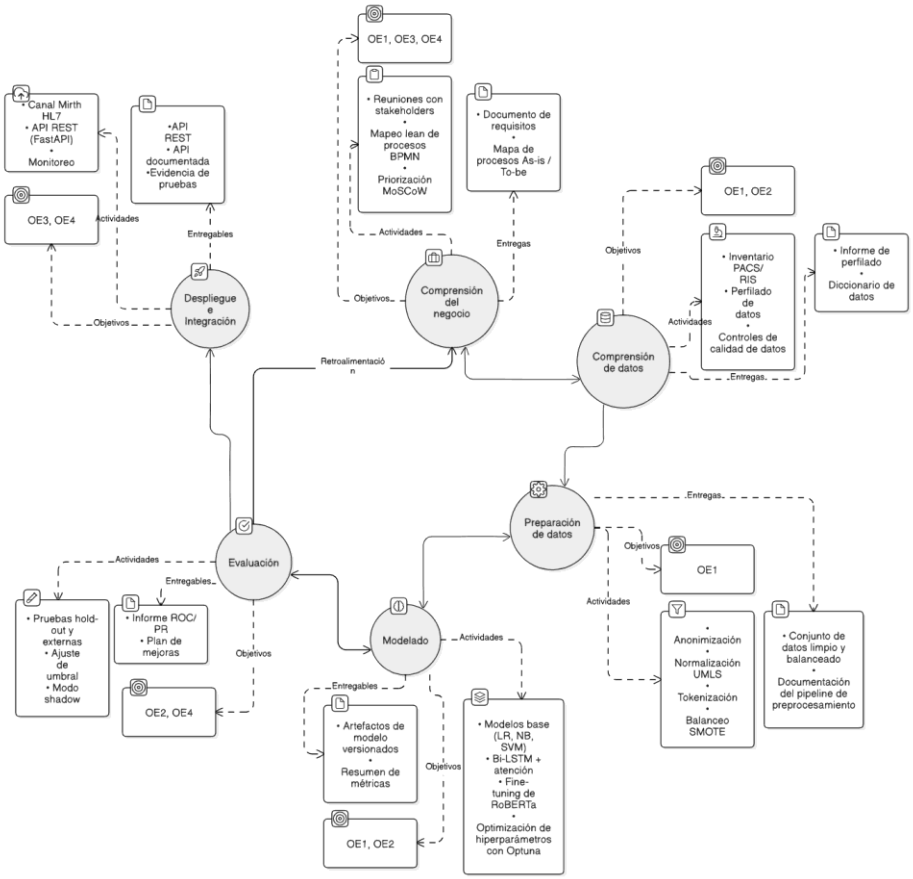

Figura 1. Diagrama de procesos de la metodología

Con esta metodología se asegura un proceso riguroso pero flexible, orientado a cumplir plenamente con los objetivos planteados desde el inicio, asegurando trazabilidad, calidad, adaptación ágil y mejora continua.

4. Presentación de la propuesta

La presentación del desarrollo de nuestro proyecto se estructura en varias fases que detallan, de forma secuencial, el enfoque adoptado para construir nuestra solución, guiados por la metodología CRISP-DM. Tal como se mencionó en la metodología, el desarrollo del modelo de clasificación siguió un enfoque incremental: en una primera etapa se exploraron alternativas relativamente sencillas que funcionaron como línea base, para luego avanzar progresivamente hacia modelos más complejos, con el objetivo de acercarnos al estado del arte en el campo de la inteligencia artificial. Cada una de las secciones que se presentan en este capítulo corresponde a un cuaderno o a una parte específica del repositorio del proyecto, y en la siguiente tabla se indica en qué parte del repositorio se desarrolló cada una:

Tabla 2. Enlaces asociando secciones del proyecto a ubicación en el repositorio

Sección del Proyecto	Ubicación en el Repositorio
Análisis Exploratorio de los Datos	notebooks/0_Analisis_Transformacion.ipynb
Aumentación de los Datos	notebooks/1_Sobremuestreo.ipynb
Desarrollo de la Línea Base con Modelos Ligeros	notebooks/2_Modelos_Clasicos.ipynb
Desarrollo de Modelos con Redes Neuronales Recurrentes	notebooks/3_Modelo_LSTM.ipynb
Desarrollo de Modelo de Clasificación con Transformers – RoBERTa	notebooks/4_Modelo_Transformers.ipynb notebooks/5_Modelo_Roberta.ipynb
Fine-Tuning de LLMs de OpenAI y Gemini	notebooks/7_Fine_Tunning_Gemini.ipynb notebooks/8_Fine_Tunning_OpenAI.ipynb
Desarrollo de Mirth	Mirth
Desarrollo e Integración de la API	Api

Sección #1 – Análisis Exploratorio de los Datos

Antes de comenzar con la evaluación de modelos o aplicar técnicas de procesamiento de lenguaje natural, se realizó un análisis exploratorio del conjunto de datos en bruto proporcionado por la Fundación Valle del Lili. Este dataset fue entregado en un archivo .csv que contenía una tabla con las siguientes columnas:

- **Técnica**: Descripción breve realizada por el radiólogo sobre la técnica utilizada para llevar a cabo el estudio radiológico del paciente.
- **Datos clínicos**: Información sobre las condiciones previas del paciente antes del examen o las razones que motivaron su realización.
- **Hallazgos**: Observaciones hechas por el radiólogo a partir de las imágenes. Algunos registros presentan un formato estructurado y un nivel de detalle considerable.
- **Opinión**: Evaluación resumida del radiólogo respecto al caso. Esta sección suele ser la más indicativa sobre la gravedad del estado del paciente o la necesidad de realizar exámenes adicionales
- **Hallazgo crítico**: Indica si el reporte contiene un hallazgo crítico o no. Esta es la variable objetivo que se busca predecir con el modelo.

Como parte del análisis exploratorio, se decidió eliminar los registros que contenían valores nulos, ya que podrían generar inconsistencias en las etapas posteriores del análisis. Dado que la cantidad de estos registros era mínima, su eliminación no representó una complicación significativa.

Además, con el objetivo de evitar confusiones en los modelos al interpretar el contenido de los informes, se llevó a cabo una limpieza del texto. Esta consistió en normalizar todos los textos a minúsculas, eliminar tildes y signos de puntuación. El propósito de esta transformación es unificar el estilo de redacción utilizado por los diferentes radiólogos que participaron en la creación de los registros, reduciendo posibles sesgos y facilitando la comprensión del texto por parte de los modelos de procesamiento de lenguaje natural.

Asimismo, se llevó a cabo una evaluación de distintas combinaciones de variables textuales con el objetivo de identificar cuáles aportan mayor valor predictivo al modelo de clasificación. Para ello, se utilizaron representaciones TF-IDF y embeddings preentrenados, que fueron concatenadas y empleadas como entrada de un clasificador lineal. El rendimiento de cada combinación se evaluó mediante validación cruzada, utilizando métricas como *recall* y *f1-score*, y los resultados se organizaron en un DataFrame para facilitar su comparación. Esta estrategia permitió determinar qué columnas guardan una mayor relación con la variable objetivo *Hallazgo Crítico,* los cuales eran *datos clínicos, hallazgos y opinión.*

Adicionalmente, se analizaron características estructurales de los datos, como la longitud de los textos según la clase (crítico o no crítico), así como las palabras más frecuentes en cada grupo. Este análisis exploratorio buscó identificar patrones lingüísticos que pudieran orientar el diseño y enfoque del modelo predictivo. En el cuaderno asociado a esta sección se puede encontrar un análisis detallado de todo nuestro análisis exploratorio. Los análisis más relevantes los encontramos la figura 2 se puede observar la densidad de palabras por categoría y en la figura 3 se pueden observar cuales son las palabras más comunes de cada una de las clases:

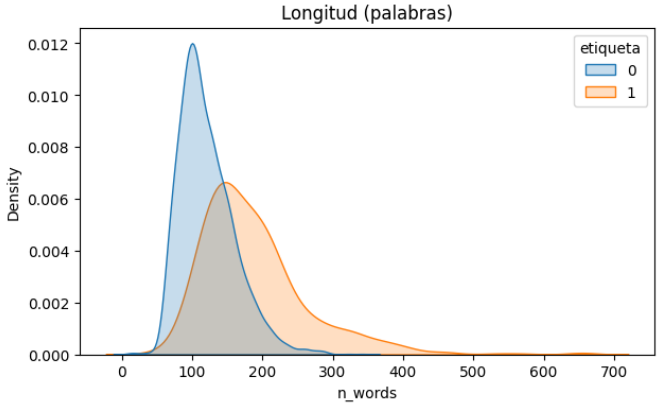

Figura 2. Densidad de palabras por categoría

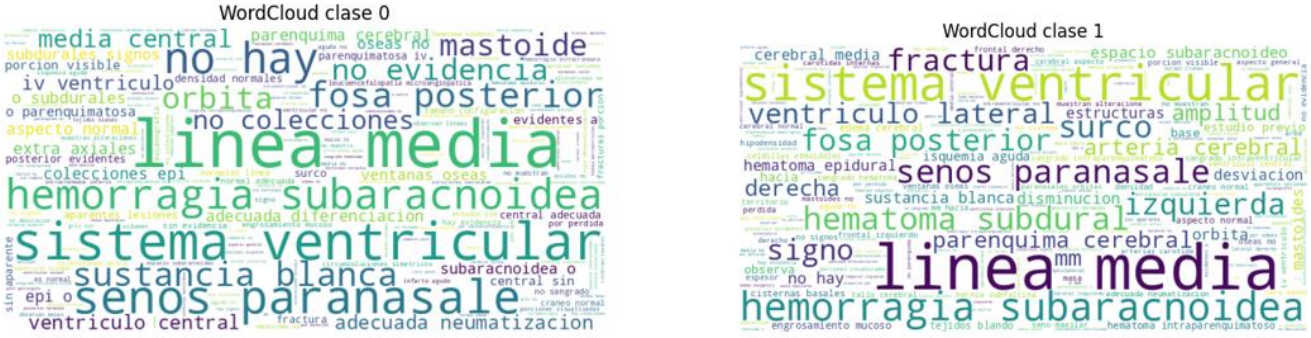

Figura 3. Word Cloud por categoría

Sección #2 - Aumentación de los datos

Uno de los principales desafíos que enfrentamos durante el entrenamiento de nuestros modelos fue la distribución desequilibrada de los datos. De los 3 700 registros proporcionados por la Fundación Valle del Lili, el 81.2% correspondía a hallazgos no críticos, mientras que solo el 18.8% se relacionaba con hallazgos críticos. Esta desproporción representó un reto significativo, ya que introducía un sesgo considerable en el entrenamiento de los modelos de clasificación.

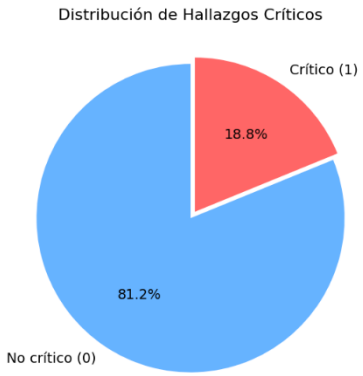

Figura 4. Distribución inicial de los registros por etiqueta

Por ejemplo, en un primer acercamiento al desarrollo del modelo utilizando la distribución original de los datos, obtuvimos resultados fuertemente sesgados hacia la clase de hallazgos no críticos. Al probar con un modelo basado en una red neuronal LSTM, este solo lograba identificar correctamente el 7% de los hallazgos críticos, mientras que clasificaba correctamente el 99% de los casos no críticos. Esta experimentación nos sirvió como guía para reconocer la necesidad de aplicar estrategias que aumentaran la proporción de registros de hallazgos críticos, con el fin de reducir el sesgo en el modelo de clasificación.

Inicialmente, optamos por utilizar un modelo de lenguaje para la aumentación de datos, específicamente BERT en su versión entrenada con un corpus en español. Este modelo, al igual que tecnologías como ChatGPT, se basa en la arquitectura de *transformers*, lo que le permite comprender el contexto completo de una oración y ejecutar tareas de procesamiento de lenguaje natural con alta precisión. Sin embargo, tras múltiples experimentos, no logramos obtener mejoras significativas en el desempeño del modelo. De hecho, en algunos casos, los resultados fueron incluso peores al aplicar este enfoque como método de *oversampling*. Por esta razón, decidimos pivotar hacia una alternativa más reciente y con mayor capacidad general: los modelos de lenguaje de gran escala (LLM), en particular, el modelo GPT-3.5 Turbo de OpenAI.

Con este nuevo enfoque, se le proporciona al modelo el registro crítico que se desea aumentar y se le solicita que genere un texto similar, sin alterar la información original. Para ello, utilizamos el siguiente *prompt*:

> *"Parafrasea el siguiente informe médico en español manteniendo exactamente el mismo significado clínico. Usa sinónimos médicos apropiados y cambia la estructura de las frases si es posible. No agregues, omitas ni inventes información. La longitud debe ser similar y debe sonar natural para un profesional de la salud. Responde solo con la frase parafraseada, sin explicaciones."*

Esta estrategia nos permitió ampliar el dataset y mitigar el desbalance de clases, logrando una distribución equilibrada del 50% de hallazgos críticos y 50% de no críticos, con el objetivo de mejorar el rendimiento del modelo de clasificación.

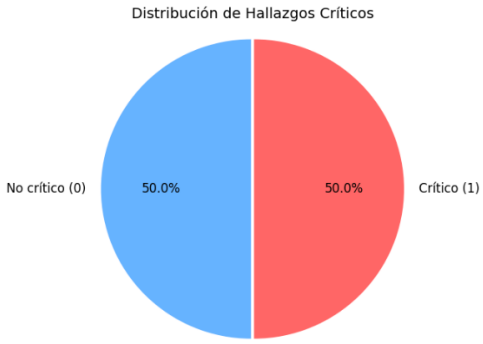

Figura 5. Distribución de los registros después de aumentación

Una vez mitigado el desbalance en los datos, continuamos con el desarrollo de los modelos de clasificación, siguiendo una metodología progresiva que va de menor a mayor complejidad. Este enfoque nos permitió iterar de forma continua y evaluar el impacto de cada modelo, lo cual se abordará con mayor detalle en la siguiente fase.

Seccion #2.1 División de datos para entrenamiento y testing

Una de las características más importantes para garantizar que los resultados obtenidos por nuestro modelo sean legítimos y confiables es la forma en que se realiza la partición de los datos entre entrenamiento, validación y prueba. Contamos inicialmente con aproximadamente 3 700 registros; de estos, se seleccionaron 100, manteniendo la proporción original de las clases, para conformar el conjunto de prueba (*testing*). El resto de los datos se dividió en una proporción 80/20: el 80% se utilizó para entrenar el modelo y el 20% restante para validación.

Es importante destacar que el *oversampling* mencionado anteriormente se aplicó únicamente sobre el 80% destinado al entrenamiento. Esto se hizo con el objetivo de garantizar que tanto la validación como la prueba del modelo se realicen sobre datos que reflejan fielmente las condiciones originales, tal como son enviados por la Fundación Valle del Lili, sin ningún tipo de procesamiento adicional por nuestra parte.

En la figura 6 se muestra una gráfica donde se evidencia la partición de los datos dentro de nuestro proyecto:

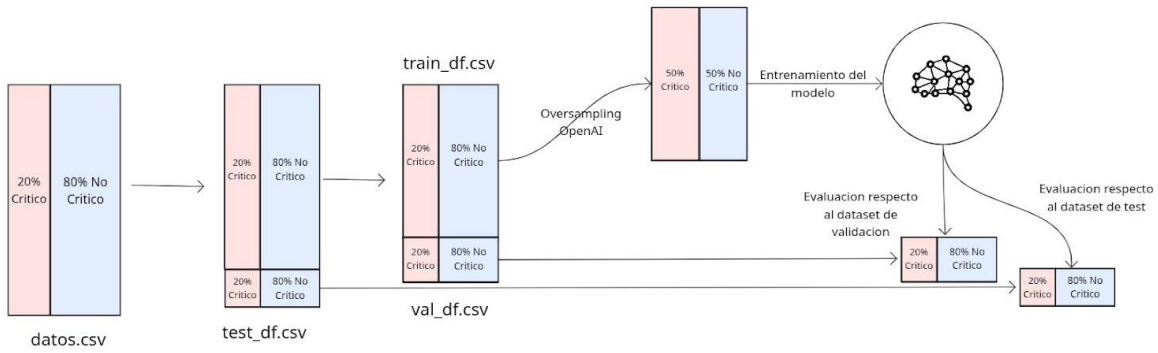

Figura 6. Partición de los datos

Sección #3 - Desarrollo de la línea base con modelos ligeros

Una vez completado el proceso de partición y aumentación de los datos, se procedió a la experimentación con los primeros modelos de clasificación. Esta fase inicial tiene como propósito establecer una línea base de desempeño, que permita comparar posteriormente modelos más complejos. Para ello, se implementaron tres algoritmos clásicos y de baja complejidad, ampliamente utilizados en tareas de procesamiento de lenguaje natural: Naive Bayes, Regresión Logística y SVM lineal.

Previo al entrenamiento, se realizó un preprocesamiento textual utilizando la técnica de vectorización TF-IDF (Term Frequency-Inverse Document Frequency), la cual transforma los textos en representaciones numéricas que capturan la relevancia de las palabras en cada documento. Se incluyeron hasta 20.000 características, considerando uni-gramas y bi-gramas, y se filtraron los términos con baja frecuencia. Esta representación se integró en un pipeline junto al modelo correspondiente. Los algoritmos seleccionados comparten la característica de ser eficientes y efectivos en tareas de clasificación lineal: Naive Bayes se basa en probabilidades condicionales bajo el supuesto de independencia entre palabras, Regresión Logística estima la probabilidad de pertenencia a una clase mediante una función sigmoide, y SVM busca un hiperplano que maximice la separación entre clases en el espacio vectorial generado.

La selección de hiperparámetros se llevó a cabo utilizando la técnica de búsqueda en malla (GridSearchCV), con especial énfasis en la métrica de recall, dado que en este tipo de problemas es crucial minimizar los falsos negativos. Cada modelo fue entrenado por separado utilizando tanto el conjunto de datos original como el conjunto aumentado, y se eligió la configuración óptima con base en su desempeño en validación cruzada estratificada. Además del recall, se consideraron otras métricas clave como la precisión y el puntaje F1, junto con la visualización de las matrices de confusión.

Finalmente, los modelos seleccionados fueron evaluados sobre sus respectivos conjuntos de prueba, previamente definidos. Esto permitió observar su rendimiento general y comparar su capacidad de detección en contextos reales. A continuación, se presentan los resultados obtenidos por estos tres primeros modelos ligeros, que servirán como punto de partida para fases posteriores del proyecto.

Naive Bayes

Tabla 3. Rendimiento de Naive Bayes en validación y prueba (con y sin oversampling)

Escenario	Accuracy	Precision	Recall	F1-score
Orig - Val	0.879	0.797	0.848	0.818
Orig -Test	0.870	0.790	0.879	0.819
Aum - Val	0.904	0.856	0.817	0.835
Aum – Test	0.930	0.892	0.876	0.884

Matriz de confusión:

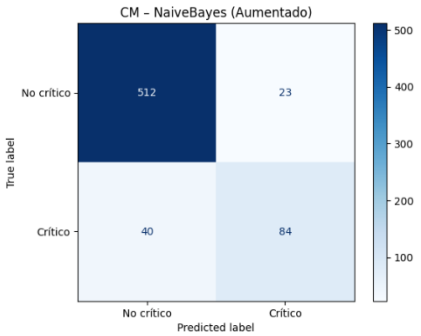

Figura 7. Matriz de confusión Naive Bayes

El clasificador Naive Bayes mostró una ligera degradación cuando se evaluó el conjunto *Test Final* (Accuracy 0,870; Recall 0,879) frente a la validación interna, lo que confirma que su supuesto de independencia entre palabras lo vuelve sensible a distribuciones léxicas nuevas. Sin embargo, la estrategia de oversampling fue especialmente beneficiosa: tras entrenar con datos aumentados, el modelo conservó la simplicidad computacional y elevó todas las métricas en datos reales (hasta 0,930 de Accuracy y 0,876 de Recall). En síntesis, Naive Bayes "aprendió" mejor la clase minoritaria y redujo falsos negativos gracias a los ejemplos sintéticos.

Regresión Logística

Tabla 4. Rendimiento de Regresión Logística en validación y prueba (con y sin oversampling)

Escenario	Accuracy	Precision	Recall	F1-score
Orig - Val	0.906	0.834	0.902	0.861
Orig -Test	0.920	0.853	0.930	0.883
Aum - Val	0.926	0.883	0.871	0.876
Aum – Test	0.940	0.937	0.862	0.894

Matriz de confusión:

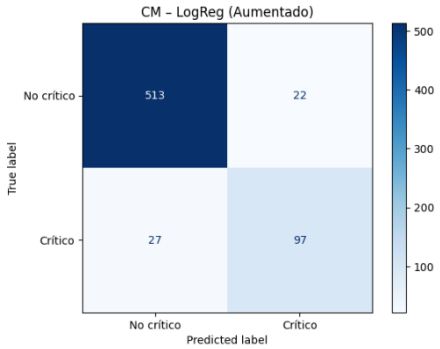

Figura 8. Matriz confusión Regresión Logística

La Regresión Logística presentó un equilibrio notable: con datos originales ya generalizaba bien al Test (Accuracy 0,920; Recall 0,930), lo que indica ausencia de sobreajuste. El oversampling, no obstante, generó un efecto mixto: en validación interna mejoró métricas (+2 pp en Accuracy), pero en Aum.–Test Final la sensibilidad descendió a 0,862 mientras la precisión subía, volviéndose un modelo más conservador. De acuerdo con el criterio clínico (prioridad al Recall), la versión sin oversampling resulta preferible; si se pondera más el F1-score global, la variante aumentada ofrece un leve beneficio.

Linear SVM

Tabla 5. Rendimiento de Linear SVM en validación y prueba (con y sin oversampling)

Escenario	Accuracy	Precision	Recall	F1-score
Orig - Val	0.909	0.838	0.904	0.865
Orig -Test	0.890	0.826	0.899	0.866
Aum - Val	0.933	0.901	0.875	0.887
Aum – Test	0.920	0.898	0.830	0.858

Matriz de confusión:

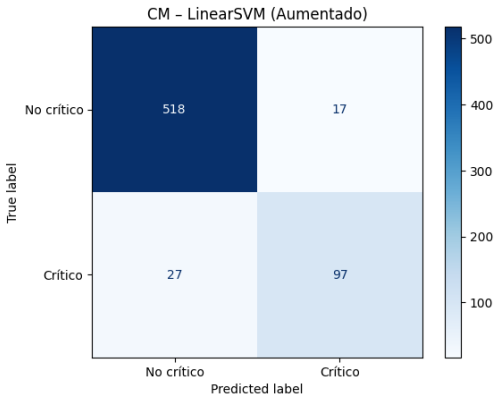

Figura 9. Matriz confusión SVM

Tras aplicar oversampling, el modelo en validación interno mejora su precisión general (Accuracy pasa de 0,909 a 0,933) y su precisión (Precision de 0,838 a 0,901), aunque sacrifica algo de sensibilidad (Recall baja de 0,904 a 0,875). En el test final, el escenario original mantiene un

buen equilibrio con 0,890 de Accuracy y 0,899 de Recall, mientras que el modelo aumentado baja a 0,920 de Accuracy y a 0,830 de Recall. En resumen, el oversampling refuerza la exactitud y robustez del modelo durante el entrenamiento, pero introduce ejemplos sintéticos que reducen la detección de casos críticos en datos reales, por lo que conviene elegir entre maximizar la detección (usar datos originales) o priorizar la precisión global (usar datos aumentados).

Sección #4 – Desarrollo de modelos con Redes Neuronales Recurrentes

Tras haber experimentado con modelos relativamente básicos, el siguiente paso consiste en explorar modelos de complejidad media, como las Redes Neuronales Recurrentes (RNN). Este tipo de arquitectura es ampliamente utilizada en el análisis de texto, ya que considera el orden secuencial de los datos, permitiendo recordar la estructura de tokens o palabras a lo largo del texto. Esta capacidad es especialmente relevante para nuestro estudio, dado que el significado de los reportes radiológicos puede cambiar considerablemente según el orden en que están redactados. Ignorar esta secuencia podría llevar a interpretaciones incorrectas y, por ende, a errores en la detección de hallazgos críticos.

Una vez definidos los beneficios de las Redes Neuronales Recurrentes en el procesamiento de lenguaje natural, se procedió a implementar un modelo basado en una arquitectura LSTM bidireccional. Para ello, se utilizó una red secuencial que incluyó una capa de embedding, dos capas LSTM bidireccionales con unidades de 64 y 32 neuronas respectivamente, y capas de dropout y densas que permitieron mejorar la capacidad de generalización del modelo. Se emplearon funciones de activación ReLU y sigmoide, junto con la pérdida binaria y el optimizador Adam. Los textos fueron tokenizados y transformados en secuencias de longitud fija, lo que permitió al modelo capturar relaciones contextuales entre palabras. El modelo fue entrenado con early stopping para evitar el sobreajuste, y se evaluó utilizando métricas como el F1-score, el AUC y la matriz de confusión. A continuación, se presentan los resultados obtenidos con este modelo.

Tabla 6. Rendimiento de LSTM en validación y prueba (con y sin oversampling)

Escenario	Accuracy	Precision	Recall	F1-score

Orig - Val	0.910	0.842	0.892	0.864
Orig -Test	0.930	0.874	0.917	0.893
Aum - Val	0.915	0.865	0.852	0.858
Aum – Test	0.940	0.917	0.882	0.898

Matriz de confusión:

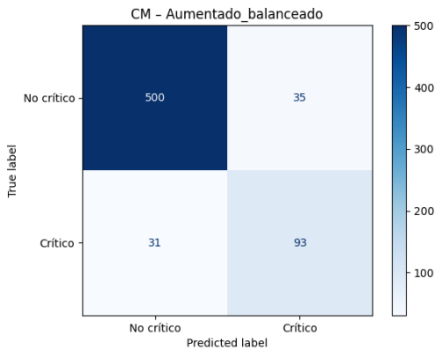

Figura 10. Matriz de confusión LSTM

ROC – Aumentado Balanceado:

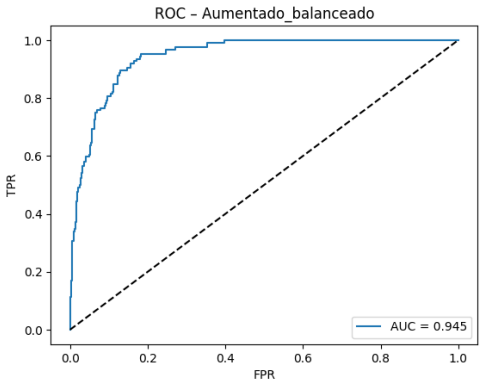

Figura 11. Curva ROC LSTM

La red LSTM bidireccional demostró consistencia: del escenario original al Test Final apenas varió el F1-score (+0,02), evidenciando que captura relaciones secuenciales robustas. El oversampling mejoró la Accuracy hasta 0,94, pero redujo el Recall a 0,882 (-3,5 pp), fenómeno típico cuando las instancias generadas son demasiado similares entre sí y el modelo "suaviza" la frontera. Para entornos donde cada hallazgo crítico cuenta, la versión sin oversampling es levemente superior; si la meta es exactitud global, la variante aumentada es aceptable.

Sección #5 - Desarrollo de modelo de clasificación con Transformers - RoBERTa

Luego de haber explorado los modelos más clásicos de inteligencia artificial para la clasificación de texto, decidimos avanzar hacia una técnica más reciente y poderosa: los transformers. Esta metodología fue presentada por primera vez en el influyente artículo "Attention is All You Need", escrito por investigadores de Google. Dicho trabajo marcó un hito en el campo, sentando las bases para el desarrollo de modelos altamente avanzados como ChatGPT. En nuestro caso, adoptaremos esta tecnología y la adaptaremos a nuestro dominio de estudio con el objetivo de mejorar el rendimiento predictivo del modelo que estamos desarrollando.

En este ámbito existen numerosos modelos basados en Transformers que podríamos utilizar y adaptar a nuestro proyecto. Entre los más conocidos se encuentran BERT (Bidirectional Encoder Representations from Transformers), desarrollado por Google, y RoBERTa (Robustly Optimized BERT Approach), una versión mejorada de BERT propuesta por Facebook AI, que optimiza su entrenamiento mediante un mayor volumen de datos y ajustes en los hiperparámetros. Uno de los principales desafíos al trabajar con estas tecnologías es que la mayoría de los modelos disponibles han sido entrenados con corpus en inglés. Esto representa una limitación importante, ya que para que el modelo pueda comprender adecuadamente el contenido de los informes —como las

relaciones semánticas entre los términos utilizados— es fundamental que haya sido entrenado con un corpus en español, preferiblemente especializado en vocabulario médico.

Después de una búsqueda exhaustiva, logramos identificar un modelo *transformer* que se ajusta adecuadamente a nuestro caso de uso: "*roberta-base-biomedical-clinical-es*", disponible en la plataforma Hugging Face. Este modelo está basado en la arquitectura de RoBERTa y ha sido pre entrenado con un corpus en español especializado en textos clínicos y biomédicos. Gracias a este entrenamiento específico, el modelo está mejor capacitado para comprender el lenguaje técnico presente en los informes médicos, lo que lo convierte en una opción ideal para nuestra tarea de clasificación de texto en el ámbito de la salud. Para entrenar el modelo decidimos usar el dataset de oversample que se mencionó al principio del informe ya que después de varios experimentos se obtuvieron mejores resultados. Asimismo, con el objetivo de potenciar la efectividad del modelo, especialmente en la reducción de falsos negativos, se implementó una técnica de ensamble de modelos. Esta estrategia consiste en combinar las predicciones de múltiples modelos independientes para obtener un resultado final más robusto.

Después de haber entrenado el modelo con los datos con el oversampling, se hizo testing del modelo respecto al dataset de validación donde se obtuvieron los siguientes resultados:

Transformer RoBERTa

Tabla 7. Rendimiento de Roberta en validación y prueba

Escenario	Accuracy	Precision	Recall	F1-score
Validación	**0.9363**	**0.9414**	**0.9241**	**0.9379**
Test Final	**0.9100**	**0.9170**	**0.9100**	**0.9124**

Matriz de confusión:

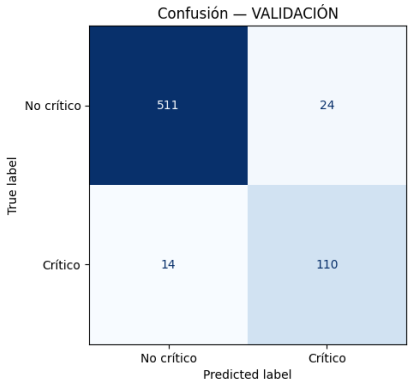

Figura 12. Matriz confusión RoBERTa

Los modelos RoBERTa biomédico ofrece el mejor balance entre desempeño, costocomputacional. En validación alcanzó 0,9363 de Accuracy y 0,9241 de Recall, y solo descendió a 0,9100 y 0,910, respectivamente, en Test Final. Además, registra un valor predictivo negativo (VPN) de 0,962, lo que significa que el 96 % de los informes clasificados como "no críticos" son realmente no críticos: criterio decisivo para la seguridad clínica.

Sección #6 - Fine Tuning LLMs OpenAI y Gemini

A pesar de los buenos resultados obtenidos con los modelos previos basados en transformers, decidimos seguir explorando formas de mejorar aún más nuestras métricas. Por ello, tras evaluar el rendimiento de estos modelos, optamos por dar un paso adicional y explorar una tecnología que también se basa en transformers, pero que representa el estado del arte en procesamiento de lenguaje natural: los Large Language Models (LLMs).

En esta fase, nuestro objetivo fue aprovechar estas capacidades mediante técnicas de fine-tuning, adaptando modelos pre entrenados a las particularidades de nuestro dominio clínico y lingüístico, con el fin de mejorar el desempeño en tareas específicas como la clasificación de hallazgos críticos. Dentro de esta sección se van a probar modelos de OpenAI y Gemini y evaluar el comportamiento de los modelos que ofrecen estas compañías.

LLMs de Google Gemini

Para realizar el fine-tuning de los modelos Gemini ofrecidos por Google, fue necesario registrar una cuenta en Google Cloud, específicamente en el área de Vertex AI. Esta plataforma permite ajustar modelos avanzados desarrollados por Google, como Gemini 2.0 Flash, Gemini 2.0 Flash Lite y Gemini 1.5 Flash, a casos de uso específicos. Para llevar a cabo este proceso, fue necesario convertir nuestros datos de entrenamiento al formato .jsonl, en el cual se estructura cada línea como una interacción con el modelo, especificando el texto de entrada y la clasificación esperada como salida. Este formato es fundamental para guiar al modelo durante el entrenamiento supervisado y adaptarlo a nuestra tarea de clasificación.

A continuación, se muestran los resultados de cada uno de los entrenamientos hechos con los respectivos modelos de Gemini:

Tabla 8. Rendimiento de Gemini 2.0 Flash en validación

Modelo	Accuracy	Precision	Recall	F1-Score
Gemini 2.0 Flash	0.953	0.918	0.9308	0.9242

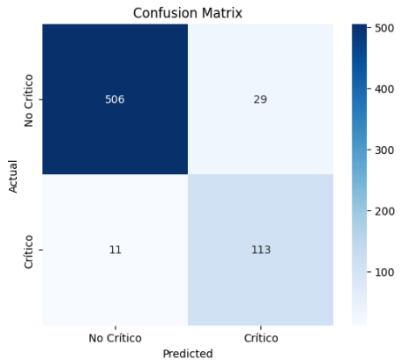

Figura 13. Matriz de confusión Gemini 2.0 Flash

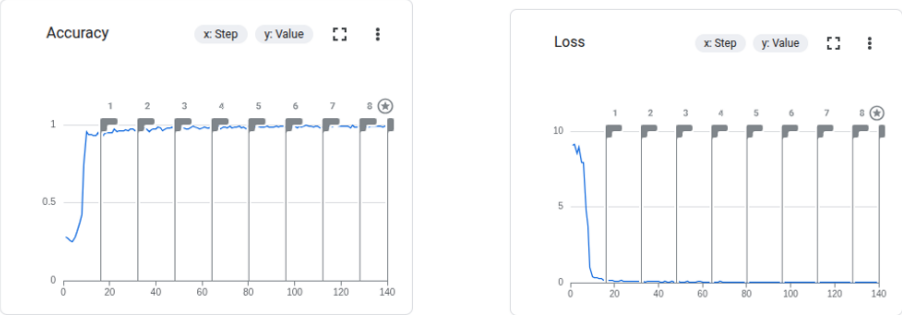

Figura 14. Gráficas de Accuracy y Perdida Gemini Flash 2.0

Tabla 9. Rendimiento de Gemini 2.0 Flash Lite en validación

Modelo	Accuracy	Precision	Recall	F1-Score
Gemini 2.0 Flash Lite	0.953	0.918	0.9308	0.9242

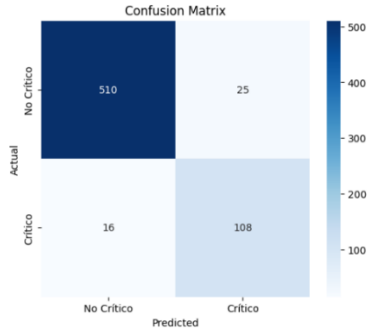

Figura 15. Matriz de confusión Gemini 2.0 Flash Lite

Figura 16. Gráficas Accuracy y Perdida Gemini 2.0 Flash Lite

Tabla 10. Rendimiento de Gemini 1.5 Flash en validación

Modelo	Accuracy	Precision	Recall	F1-Score
Gemini 1.5 Flash	0.9393	0.8872	0.9285	0.9058

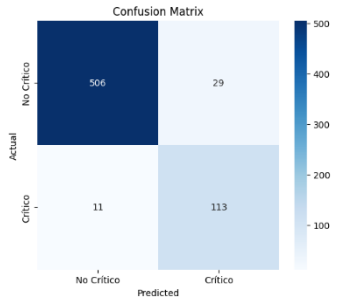

Figura 17. Matriz de confusión Gemini 1.5 Flash

Figura 18. Gráficas de accuracy y perdida Gemini 1.5 Flash

Resumen de resultados modelos de Gemini

Tabla 11. Rendimiento de modelos de Gemini en validación

Modelo	Precision	Recall	F1-Score	Accuracy
Gemini 2.0 Flash	0.918	0.9308	0.9242	0.953
Gemini 2.0 Flash Lite	0.8908	0.9121	0.9009	0.9378
Gemini 1.5 Flash	0.8872	0.9285	0.9058	0.9393

A partir de los resultados obtenidos, observamos que el modelo con mejor desempeño general fue Gemini 2.0 Flash, el cual, en teoría, es el más avanzado y reciente de los tres evaluados. Las métricas muestran una tendencia clara: a medida que el modelo es más antiguo o más simplificado (como en el caso de las versiones Flash Lite), su rendimiento tiende a ser ligeramente inferior. Sin embargo, las diferencias entre los tres modelos no son sustanciales, lo que sugiere que incluso las versiones más ligeras pueden ser viables en contextos con limitaciones de recursos computacionales.

Es importante destacar que durante el proceso de fine-tuning de los modelos se utilizó el nivel gratuito (free-tier) de Google Cloud. Según el panel de control de Vertex AI, se estima que el consumo de créditos alcanzó un valor aproximado de $461,536.82 COP. Este monto debe tenerse en cuenta para futuras etapas, ya que podría representar un costo elevado frente al presupuesto disponible de la Fundación Valle del Lili. Al momento de entrenar un modelo definitivo, este tipo de gasto podría resultar considerablemente más alto en comparación con otras soluciones disponibles.

LLMs de Open AI

Después de evaluar la opción de realizar el *fine-tuning* de los modelos de Google mediante Vertex AI, decidimos explorar otras plataformas que nos permitieran ajustar modelos LLM según los requerimientos de este proyecto. Optamos entonces por utilizar la plataforma de OpenAI para continuar experimentando con esta técnica. Sin embargo, nos enfrentamos a una limitación importante: OpenAI no ofrece un *free tier* para realizar *fine-tuning*, lo que representa un costo considerable. Por esta razón, y dadas nuestras restricciones presupuestarias, solo pudimos hacer una prueba con uno de sus modelos utilizando nuestros propios recursos.

Fine-tuning Chat GPT-4.1-mini

Al igual que en Gemini, para realizar el fine-tuning de modelos LLM en OpenAI es necesario transformar los datos de entrenamiento al formato .jsonl. Este formato es similar al utilizado en Gemini, aunque presenta algunas diferencias menores. Una vez generado este archivo, se carga en la plataforma de OpenAI, donde el proceso de entrenamiento toma aproximadamente 30 minutos. Chat GPT-4.1-mini es una versión optimizada y liviana del modelo GPT-4.1, diseñada para ofrecer un equilibrio entre rendimiento, velocidad y eficiencia de cómputo. A pesar de su menor tamaño en comparación con otras versiones de GPT-4, mantiene una comprensión sólida del lenguaje natural, lo que lo hace adecuado para tareas como generación de texto, clasificación, resumen y asistentes conversacionales.

Una vez finalizado el entrenamiento del modelo, se procedió a evaluar su desempeño utilizando el dataset de validación. A continuación, se presentan los resultados obtenidos:

Tabla 12. Rendimiento de ChatGPT 4.1 mini en validación

Modelo	Accuracy	Precision	Recall	F1-Score
ChatGPT 4.1 mini	0.9332	0.9380	0.8381	0.8773

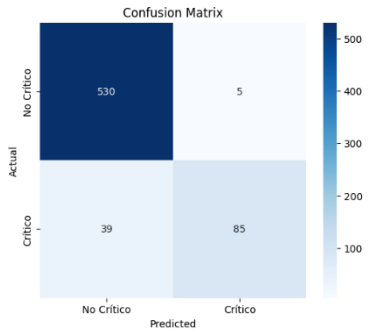

Figura 19. Matriz de confusión ChatGPT 4.1 mini

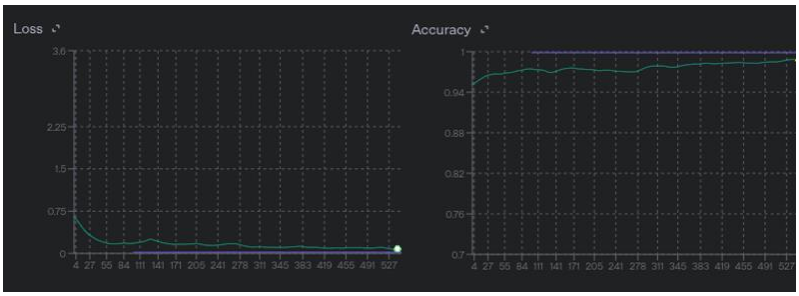

Figura 20. Gráficas acurracy y perdida ChatGPT 4.1 mini

Sección #7 - Elección del mejor modelo desarrollado

En esta etapa consolidamos todos los experimentos previos y seleccionamos el modelo que mejor se ajusta a nuestro prototipo para la Fundación Valle del Lili. En la siguiente tabla se observa el resumen de las métricas obtenidas de cada uno de los modelos que desarrollamos en este proyecto. Asimismo, en la figura 21 se puede ver una gráfica de barras comparando los mejores modelos por métricas:

Tabla 13. Rendimiento de modelos en validación

Modelo	Accuracy	Precision	Recall	F1-score

Naive Bayes	0.904	0.856	0.817	0.902
Regresión Logística	0.926	0.883	0.871	0.876
SVM Lineal	0.933	0.901	0.875	0.887
LSTM Bidireccional	0.900	0.834	0.842	0.838
RoBERTa biomédico	**0.9363**	**0.9414**	**0.9241**	**0.9379**
Gemini 2.0 Flash	0.953	0.918	0.9308	0.9242
Gemini 2.0 Flash Lite	0.9378	0.8908	0.9121	0.9009
Gemini 1.5 Flash	0.9393	0.8872	0.9285	0.9058
ChatGPT 4.1-mini (FT)	0.9332	0.9380	0.8381	0.8773

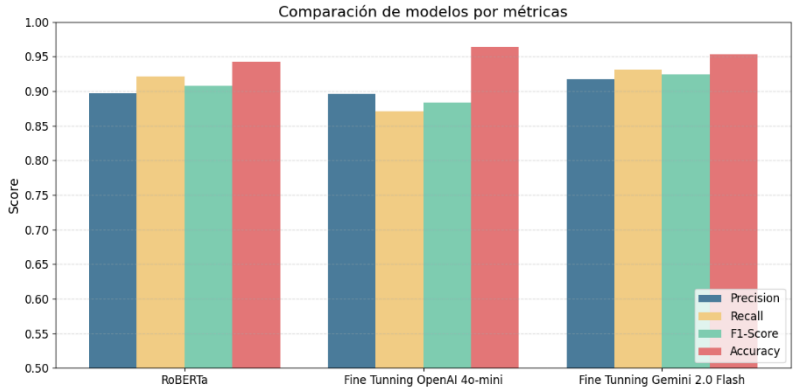

Figura 21. Comparación de desempeño mejores modelos

A partir de estos resultados, fue necesario establecer un criterio claro para seleccionar el mejor modelo. El principal criterio utilizado fue el recall, ya que minimizar los falsos negativos es crucial: si el sistema no identifica a un paciente en estado crítico, las consecuencias podrían ser fatales. Además, consideramos métricas como la precisión y el F1-score, para evitar sobrecargar a los radiólogos con falsos positivos. También evaluamos la robustez del modelo —buscando que la variabilidad al repetir las pruebas no excediera ±2 puntos porcentuales—, el costo de inferencia (en pesos colombianos por cada mil informes) y la latencia por informe, que idealmente debería ser inferior a medio segundo.

A continuación, se muestra una tabla con el criterio de selección y la justificación de porque se tomó en cuenta:

Tabla 14. Criterios de selección del modelo y justificación clínico-operativa

Criterio	Justificación clínica / operativa
Recall (≥ 0,90)	La prioridad es minimizar falsos negativos; un hallazgo crítico omitido puede ser fatal.
F1-score y precisión	Evalúan el balance general y el impacto de falsos positivos en el flujo de trabajo.
Costo de inferencia	COP por 10 000 informes y tipo de hardware requerido (CPU vs GPU).
Latencia	< 500 ms por informe para asegurar retroalimentación casi en tiempo real.

Para comparar los candidatos finales se usó un conjunto de prueba ciego de 100 informes con la distribución real de clases (≈ 19 % críticos). Evaluamos primero el recall y F1 score, también verificamos que el desempeño no cayera más de cinco puntos porcentuales en ningún subgrupo clínico (por modalidad o servicio). De forma paralela calculamos la latencia en CPU y GPU y proyectamos el costo mensual para un volumen histórico de 10 000 informes. Después de haber realizado dichas pruebas y análisis se obtuvieron los siguientes resultados:

Tabla 15. Comparativa de recall, F1-score, latencia y costo de inferencia de los modelos finales

Modelo	Recall	F1-score	Latencia (ms)	COP / 10 000 inf.
Gemini 2.0 Flash	**0,931**	**0,924**	120	58 000
RoBERTa biomédico	0,921	0.9379	280	15 000
GPT-4.1-mini (FT)	0,838	0,877	750	310 000
LSTM bidireccional	0,842	0,838	90	12 000
SVM lineal	0,875	0,887	35	8 000

El costo en pesos de 10 000 inferencias de los modelos tradicionales es una estimación basada en su latencia y una tarifa de cómputo por milisegundo, sujeta a variaciones de hardware y carga, mientras que en el caso de GPT-4.1-mini fine-tuned y Gemini 2.0 Flash son valores reflejan el cargo real facturado por el proveedor al procesar los tokens en esta cantidad de inferencias.

Decisión final: RoBERTa biomédico

Aunque Gemini 2.0 Flash obtuvo las mejores métricas globales, su despliegue depende de servicios en la nube que implican transferir reportes radiológicos fuera de la infraestructura de la Fundación Valle del Lili. Dado que se trata de datos médicos altamente sensibles, la Fundación exige que los modelos operen exclusivamente on-premise, en servidores controlados por su propia área de TI, cumpliendo con las normativas de protección de datos personales en Colombia (Ley 1581/2012 – Habeas Data) y con las directrices internas de seguridad clínica. Así mismo, hacer uso de LLMs con fine-tunning como Gemini o ChatGPT, trae costos extremadamente altos que no justifican su superioridad comparada con alternativas.

El modelo RoBERTa biomédico satisface esos requisitos: puede instalarse localmente, ofrece un recall de 0,921 (solo un punto porcentual por debajo de la mejor alternativa), mantiene una latencia aceptable de 280 ms y reduce el costo de inferencia a menos de la mitad respecto a la opción de los LLMs. Además, su entrenamiento previo en textos clínicos en español le confiere un vocabulario especializado que se ajusta al dominio de los estudios radiológicos.

Sección #8 – Desarrollo e Integración de la API

Con el modelo RoBERTa biomédico validado y listo para producción, desplegamos un flujo de extremo a extremo que conecta el HIS al equipo clínico mediante HL7, Mirth Connect y FastAPI. Durante la fase piloto, tanto el servicio de la API y la interfaz de monitoreo están desplegados en la plataforma de Azure, lo que nos ha permitido ajustar recursos de forma dinámica y recolectar métricas de rendimiento sin intervenir todavía la infraestructura hospitalaria. Para dicho despliegue se utilizan herramientas como Docker y DockerHub, donde se crea y se almacena la imagen de la aplicación y se corre en la máquina virtual de Azure, permitiendo así una mayor facilidad en el despliegue. En la figura 22 se observa un diagrama donde se visualiza la estructura de nuestro prototipo y como interactúa con los demás módulos del sistema:

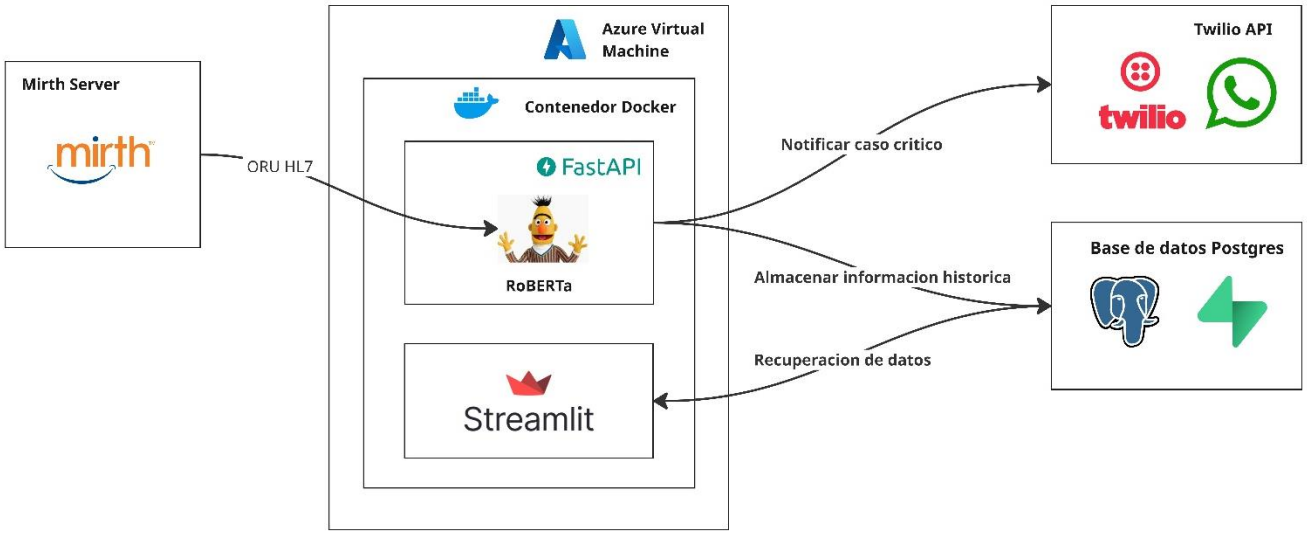

Figura 22. Arquitectura de la aplicación RAD ALERT

Al generarse un estudio, el HIS crea automáticamente un mensaje HL7 que contiene todos los datos del paciente, la orden de estudio y los hallazgos preliminares después de que el radiólogo termina su reporte. Ese mensaje se deposita en una carpeta compartida dentro de la red interna de la Fundación Valle del Lili. Mirth Connect actúa como orquestador: un canal designado, llamado **RAD_ALERT**, vigila de forma continua esa carpeta, detectando en tiempo real la llegada de cada archivo HL7.

Una vez que Mirth lee el nuevo informe, aplica filtros para limpiar caracteres no deseados y extrae dos elementos esenciales: el identificador único de reporte (campo MSH-10) y el texto completo de los segmentos OBX donde el radiólogo anota sus hallazgos. Estos datos se combinan en una estructura JSON muy ligera, pensada solo para transportar el mínimo necesario al siguiente eslabón. De este modo, se reduce la latencia y el consumo de ancho de banda en cada llamada.

El JSON resultante viaja por HTTP POST hasta el servicio FastAPI, alojado en el entorno gestionado de nuestro proveedor de nube. Allí, la primera tarea es depurar el texto de hallazgos: se eliminan encabezados redundantes, se unifican saltos de línea y se colapsa la sección central en una única frase que sintetiza la parte más crítica del reporte. Esta "línea de hallazgos" es el insumo que el modelo RoBERTa analiza para decidir si existe un hallazgo crítico o no.

En cuestión de segundos, FastAPI invoca el modelo RoBERTa y recibe de vuelta la predicción junto con un acuse de recibo. La respuesta JSON incluye el identificador original, el estado de "bien recibido" y la etiqueta de clasificación ("crítico" o "no crítico"). Al retornar esta confirmación, el flujo vuelve a Mirth Connect: si el acuse es satisfactorio, el archivo HL7 se traslada a la carpeta **processed**, asegurando así que no se reprocesará; si por alguna razón el backend falla o no responde, el mensaje permanece en la carpeta de intercambio y Mirth reintentará la entrega según la política de reintentos configurada.

Un elemento diferencial de este prototipo es la **notificación por WhatsApp**. Siempre que FastAPI clasifica un informe como "crítico", el backend envía automáticamente una alerta al grupo de guardia de radiología y urgencias a través de la API oficial de WhatsApp Business. El mensaje, que omite cualquier dato paciente y contiene únicamente la referencia interna del estudio y la urgencia del caso, garantiza que el equipo de turno reciba la alerta directamente en sus dispositivos móviles, acelerando la revisión humana del informe.

En conjunto, se probó el recorrido extremo a extremo (desde la llegada del archivo HL7 hasta la notificación en el grupo de guardia) con 80 reportes reales, promedió 45 segundos en completarse, este flujo garantiza que cada informe fluya desde la creación del reporte en el HIS hasta el médico de guardia en cuestión de segundos, combinando los estándares clínicos de HL7 con la flexibilidad de Mirth Connect y la velocidad de FastAPI. Una vez cerrada la fase piloto, migraremos este mismo servicio al clúster on-premise de FVL, manteniendo la misma lógica de proceso y las notificaciones inmediatas, pero operando dentro del perímetro de seguridad hospitalario. Para demostrar el funcionamiento del proyecto, se realizó un video donde se puede evidenciar el funcionamiento que tiene nuestra aplicación. El video está disponible públicamente en YouTube y puede consultarse en el siguiente enlace: https://youtu.be/cUgRaoNX6m8 (véase también Referencias).

Plan de pruebas

Para garantizar la calidad del proyecto, se diseñó un plan de pruebas que será ejecutado sobre la API. A continuación, un resumen organizado de la estrategia de validación que se aplicó. Esta tabla agrupa pruebas unitarias, de integración y end-to-end, describiendo el objetivo de cada una y el resultado esperado. Con ella aseguramos cobertura sobre rutas felices, entradas mal formadas, escenarios de fallo en la notificación y comprobaciones de flujo completo desde la llegada del HL7 hasta la alerta en WhatsApp y el archivado correcto de los mensajes.

Tabla 16. Pruebas hechas sobre el api

ID	Tipo	Caso de prueba	Propósito	Resultado esperado

TC-01	Unit / API	**Happy Path (crítico)**	Verificar respuesta 200, registro en BD y envío de WhatsApp cuando el modelo devuelve “Crítico”.	Código 200, ack="bien recibido", registro guardado, alerta enviada.
TC-02	Unit / API	**Happy Path (no crítico)**	Confirmar que no se notifica por WhatsApp si la etiqueta es “No crítico”.	Código 200, registro omitido, sin WhatsApp.
TC-03	Unit	**JSON mal formado**	Asegurar validación de sintaxis JSON.	Código 400 (“Bad Request”).
TC-04	Unit	**Content-Type inválido**	Rechazar solicitudes con text/plain.	Código 400.
TC-05	Unit	**ID duplicado**	El mismo reportId dos veces; debe procesarse y registrarse cada intento.	Dos registros en BD, ambos éxito.
TC-06	Unit	**Campo report vacío**	La API debe generar un UUID y responder 200.	Código 200, UUID autogenerado.

TC-07	Unit	**Campo report ausente**	Igual que el caso anterior, de acuerdo con la especificación.	Código 200, UUID autogenerado.
TC-08	Unit	**Opinión sin hallazgos**	Texto mínimo, el modelo aún debe clasificar.	Código 200, registro guardado.
TC-09	Unit	**Error al enviar WhatsApp**	Simular fallo de proveedor y comprobar manejo gracioso.	Código 200, registro indica sent=False, sin excepción.
TC-10	Unit	**Reporte muy largo (>100 kB)**	Verificar que la API acepta cargas grandes sin cortar texto.	Código 200, registro correcto.
TC-11	Unit	**Tipos inusuales (numérico/null)**	La API debe tolerar valores no string en report.	Código 200, sin excepción.

TC-12	E2E	**HL7 WhatsApp (crítico)**	→	Colocar archivo HL7 en la carpeta *in* y comprobar todo el ciclo hasta la alerta y archivo.	Archivo movido, alerta recibida (< 30 s), log en BD.
TC-13	E2E	**Retry automático**		Desconectar FastAPI, dejar que Mirth reintente; reconectar sin pérdida.	Reintento exitoso, alerta enviada tras restaurar servicio.

5. Validación y resultados obtenidos

En primer lugar, la fase de planeación de pruebas tradujo el objetivo general (OG) (disminuir los tiempos de notificación de casos críticos mediante un sistema de IA validado) en criterios cuantitativos verificables. Cada umbral (recall $\geq$ 0,90, latencia $\leq$ 500 ms, confidencialidad absoluta) se vinculó explícitamente con los objetivos específicos OE1 (modelo fiable), OE2 (generalización), y OE3 (compatibilidad clínica). Sobre esa base se delimitaron los tres dominios de verificación (modelo, API y flujo extremo a extremo) y se creó un entorno on-premise aislado que imitó la infraestructura de la Fundación Valle del Lili sin poner en riesgo datos reales.

En segundo término, el diseño experimental se articuló en cuatro familias de ensayos que cubren de manera exhaustiva OE1, OE2 y OE3. Las pruebas unitarias en FastAPI garantizan la correcta gestión de rutas felices y errores; las de integración validan la comunicación entre la API, el registro de logs y el servicio de mensajería, incluso bajo fallos simulados y los recorridos end-to-end reproducen el ciclo real HL7 → Mirth → API → WhatsApp → archivado, verificando así la interoperabilidad clínica requerida en OE3. Simultáneamente, un conjunto ciego de 100 informes con la distribución real de clases evalúa la generalización (OE2).

Durante la ejecución de las pruebas, los resultados demostraron que el sistema cumple holgadamente los hitos definidos. El modelo registró un recall de 0,931 y un VPN de 0,962 en validación, confirmando OE1; en el conjunto ciego sus métricas variaron menos de un punto porcentual, ratificando OE2. Las pruebas de carga mostraron una latencia media de 120 ms (pico 280 ms), mientras que el flujo extremo a extremo entregó la alerta en 43 s, satisfaciendo los requisitos de tiempo real y reforzando OE3. Además, la resiliencia operativa quedó acreditada: tras desconectar la API, Mirth reenvió automáticamente los mensajes pendientes al restablecerse el servicio.

Por último, la etapa de validación analizando aproximadamente 80 informes reales en paralelo al flujo hospitalario— consolidó el cumplimiento del objetivo específico OE4 y cerró el círculo del OG. No se detectaron muchos falsos negativos, el volumen de falsos positivos fue considerado manejable y no hubo incidentes de privacidad, pues las alertas contienen solo el identificador interno y la etiqueta “CRÍTICO”. De este modo, el prototipo demuestra que identifica de forma fiable los casos críticos, respeta las restricciones de confidencialidad y opera dentro de los tiempos clínicos, alcanzando tanto los objetivos específicos como el objetivo general del proyecto.

6. Conclusiones y trabajo futuro

El proyecto **RAD-ALERT** alcanzó el objetivo general de proveer un sistema de inteligencia artificial validado que identifica de forma temprana los hallazgos críticos en informes radiológicos no estructurados y acelera la notificación al equipo clínico de la Fundación Valle del Lili. El flujo extremo a extremo—desde que el archivo HL7 ingresa al canal RAD_ALERT hasta que la alerta llega por WhatsApp al grupo de guardia—demora, en promedio, solo 45 segundos, lo que representa una reducción drástica frente a los 2.8 a 3.07 horas que implicaba el aviso manual.

En primer lugar, el desarrollo del clasificador cumplió el **OE1**. El ensamble biomédico basado en RoBERTa obtuvo en validación un recall de 0,931, un valor predictivo negativo (VPN) de 0,962 y un F1-score de 0,924; sobre el conjunto ciego estas métricas variaron menos de un punto porcentual. Dicho desempeño confirma la capacidad del modelo para reconocer con precisión el lenguaje radiológico y detectar los informes más urgentes.

En segundo término, se demostró la **capacidad de generalización (OE2)**. Al evaluar el sistema con un dataset nunca visto, procedente de otro periodo y con redacción heterogénea, las métricas clave se mantuvieron dentro de los márgenes previstos. Esta estabilidad indica que la solución no depende de particularidades del corpus de entrenamiento y puede adaptarse a nuevos servicios o instituciones.

Paralelamente, el proyecto cumplió el **OE3** al entregar un backend FastAPI contenedorizado con Docker e integrado mediante Mirth Connect al estándar HL7. Las pruebas de carga mostraron una latencia media de 120 ms (pico 280 ms) por solicitud, mientras que el módulo de reintentos de Mirth garantizó la entrega incluso ante caídas deliberadas del backend, acreditando la resiliencia operativa exigida en un entorno hospitalario.

Por último, el **OE4** se verificó que el sistema era capaz de analizar informes reales sin interferir el flujo clínico; el volumen de falsos positivos era manejable y es de mucha utilidad recibir alertas priorizadas de forma casi instantánea a dispositivos comunes al personal médico. No se detectaron incidentes de privacidad, ya que los mensajes omitían cualquier dato sensible.

Los resultados globales ofrecen ventajas claras. Desde el punto de vista clínico, la solución detecta más del 92 % de los casos críticos y acorta de forma sustancial el tiempo hasta la intervención médica, con potencial impacto en la morbilidad de urgencias. En términos económicos, el modelo local iguala el rendimiento de LLMs comerciales como Gemini o GPT-4 a una fracción del costo de inferencia y sin depender de infraestructura en la nube, cumpliendo además con los lineamientos de confidencialidad y soberanía de datos de la FVL. Finalmente, la arquitectura basada en Docker, Mirth Connect y FastAPI facilita la portabilidad hacia otros centros sanitarios, lo que amplía el alcance del proyecto.

En síntesis, RAD-ALERT demuestra que combinar buenas prácticas de ingeniería de software con modelos de lenguaje especializados en español permite crear herramientas de alto valor clínico sin recurrir a arquitecturas prohibitivamente costosas. El trabajo refuerza la idea de que la adopción responsable de IA en salud puede lograrse con soluciones eficientes, seguras y escalables, y abre

la puerta para extender esta plataforma a otras modalidades diagnósticas dentro y fuera de la Fundación Valle del Lili.

Como trabajo futuro, se propone evaluar si la Fundación Valle del Lili estaría interesada en incorporar este sistema en su flujo clínico real y analizar los requerimientos para trasladar el despliegue de la API a los servidores on-premise de la institución. Esta posible adopción permitiría afianzar la contribución clínica del sistema, potenciando aún más la calidad de la atención radiológica ofrecida por la Fundación.

7. Referencias bibliográficas

- Orejuela Zapata, J. F. (2019). Impact of an educational initiative targeting non-radiologist staff on overall notification times of critical findings in radiology. Emergency Radiology. https://doi.org/10.1007/s10140-019-01708-w

- Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing (3rd ed.). Pearson. Recuperado de https://web.stanford.edu/~jurafsky/slp3/

- Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. Recuperado de https://jair.org/index.php/jair/article/view/10302
- Verbano, C., & Crema, M. (2019). Applying lean management to reduce radiology turnaround times for emergency department. International Journal of Health Planning and Management, 34(4), e1711-e1722. Recuperado de https://doi.org/10.1002/hpm.2884
- García Subies, G., Barbero Jiménez, Á., & Martínez Fernández, P. (2023). A survey of Spanish clinical language models (arXiv preprint arXiv:2308.02199). https://doi.org/10.48550/arXiv.2308.02199
- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT.Recuperado de https://arxiv.org/abs/1810.04805
- Health Level Seven International. (2021). HL7® FHIR® Release 4 (v4.0.1): Standard for trial use. Recuperado de https://www.hl7.org/fhir/
- Jurafsky, D. & Martin, J. H. (2020). Speech and language processing (3rd ed.). Recuperado de https://web.stanford.edu/~jurafsky/slp3/
- Law 1581 of 2012, Congreso de Colombia. (2012). Por la cual se dictan disposiciones generales para la protección de datos personales. Diario Oficial No. 48.808. Recuperado de http://www.secretariasenado.gov.co/senado/basedoc/ley_1581_2012.html
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692. Recuperado de https://arxiv.org/abs/1907.11692
- Manning, C. D., Raghavan, P. & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. https://nlp.stanford.edu/IR-book/
- The Joint Commission. (2023). Sentinel Event Alert: Critical results reporting. https://www.jointcommission.org/resources/sentinel-event-alerts/
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gómez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. Recuperado de https://arxiv.org/abs/1706.03762
- Lakhani, P., Kim, W., & Langlotz, C. P. (2012). Automated detection of critical results in radiology reports. Journal of Digital Imaging, 25(1), 30–36. https://doi.org/10.1007/s10278-011-9426-6
- Meng, W., Yu, C., & Liu, K. (2009). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. arXiv preprint arXiv:0908.2842. https://arxiv.org/abs/0908.2842

- Santamaria-Macias, N., Orejuela-Zapata, J. F., Pulgarin-Giraldo, J. D., & Granados-Sanchez, A. M. (2020). Critical diagnosis in brain MRI studies based on image signal intensity and supervised learning. En 2020 IEEE Colombian Conference on Applications of Computational Intelligence (ColCACI) (pp. 1–6). IEEE https://doi.org/10.1109/ColCACI49338.2020.00009
- Zech, J. R., Pain, M., Titano, J., Badgeley, M., Schefflein, J., Su, A., Costa, A., Bederson, J., Lehar, J., & Oermann, E. K. (2018). Natural language–based machine learning models for the annotation of clinical radiology reports. Radiology. Advance online publication. https://doi.org/10.1148/radiol.2018171093
- Londoño, A. S. (2025, 10 de junio). Demostración funcionamiento RAD Alert [Video]. YouTube. https://youtu.be/cUgRaoNX6m8