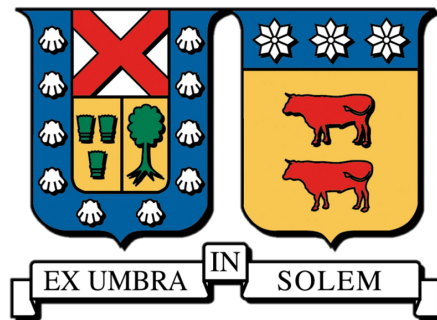


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE ELECTRÓNICA
VALPARAÍSO - CHILE



“Desarrollo de un modelo predictivo basado en procesamiento de lenguaje natural para la identificación de pacientes candidatos a derivación a Asesoramiento Genético Oncológico”

Tomás Felipe de Jesús Vega González

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL TELEMÁTICO

PROFESOR GUÍA:
PROFESOR CORREFERENTE:

Mohamed Abdelhamid
Mauricio Araya

ABRIL 2026



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título; ☐ Tesis de Postgrado;

Título del trabajo: Desarrollo de un modelo predictivo basado en procesamiento de lenguaje natural para la identificación de pacientes candidatos a derivación a Asesoramiento Genético Oncológico

Nombre del candidato(a): Tomás Felipe de Jesús Vega González

Carrera / Grado: Ingeniería Civil Telemática

Campus: Casa Central Valparaíso ; **Departamento:** Electrónica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Mohamed Abdelhamid, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL

☒ El trabajo **NO contiene información que amerite confidencialidad** y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (embargo) por:

☐ 6 meses; ☐ 12 meses; ☐ 2 años; ☐ 3 años; ☐ 5 años; ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 21-07-2026

; Firma:

M. Abdelhamid

Estudiante o Candidato(a):

Fecha: 21-07-2026

; Firma:

Tomás F. de J.

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Agradecimientos

A mis padres, por entregarme la vida, por su amor incondicional y dedicarme toda su sabiduría, dedicación para ser una mejor persona en la actualidad y en el futuro.

A mis queridos amigos de la vida que me acompañaron, enseñaron y guiaron en esto que se llama la vida. Gracias a cada uno por formar parte de ella, cada uno es una pieza para la formación de esta voluntad.

Por ultimo, a los compañeros que me encontré en la vida que no nos acompañan ahora, en un futuro nos volveremos a encontrar. Gracias

Resumen

Este trabajo presenta el desarrollo de un modelo predictivo basado en técnicas de procesamiento de lenguaje natural (NLP) con el objetivo de identificar pacientes que requieren derivación a Asesoramiento Genético Oncológico (AGO). La investigación fue realizada en colaboración con la Fundación Arturo López Pérez (FALP) y responde a la creciente necesidad de optimizar la detección de riesgos genéticos vinculados al cáncer de mama y de colon.

La metodología contempla varias etapas, incluyendo el preprocesamiento de datos clínicos, el análisis de datos estructurados y no estructurados, y el entrenamiento de modelos utilizando BioBERT, un modelo avanzado de NLP diseñado específicamente para el ámbito biomédico.

En conjunto, el modelo propuesto constituye un aporte significativo hacia una atención médica más personalizada y eficiente, al disminuir los errores en la derivación clínica y mejorar el uso de los recursos en diagnóstico y tratamiento.

Palabras Claves Procesamiento de Lenguaje Natural - Banderas Rojas - Asesoramiento Genético Ontológico - BioBERT - Datos Estructurados - Transformers

Abstract

This study presents the development of a predictive model based on natural language processing (NLP) techniques aimed at identifying patients who require referral to Oncogenetic Counseling (AGO). The research was conducted in collaboration with the Fundación Arturo López Pérez (FALP) and addresses the growing need to enhance the detection of genetic risks associated with breast and colorectal cancer.

The methodology comprises several stages, including clinical data preprocessing, analysis of both structured and unstructured data, and the training of models using BioBERT, an advanced NLP model specifically designed for biomedical applications.

Overall, the proposed model represents a significant contribution toward more personalized and efficient healthcare by reducing errors in clinical referral and optimizing the use of diagnostic and treatment resources.

Keywords Natural Language Processing - Red Flags - Oncogenetic Counseling - BioBERT - Unstructured Data - Transformers

Índice general

Índice de Figuras	8
Índice de Tablas	9
1 Introducción	10
1.1 Contexto	11
1.2 Definición del problema	12
1.3 Hipótesis	16
1.4 Usuarios	17
1.5 Objetivos	17
1.6 Estructura	17
2 Estado del Arte	19
2.1 Estado del arte y técnica	19
3 Diseño de la solución	28
3.1 Diagrama de Contexto	28
3.1.1 Diagrama de Contexto	28
3.2 Diagrama de Arquitectura	29
3.2.1 Diagrama de Arquitectura	29
3.3 Enumeración de Módulos	30
3.4 Matriz de Requisitos Funcionales y Módulos	30
3.5 Módulos	31
3.5.1 Módulo de Preprocesamiento de Datos Clínicos	31
3.5.2 Módulo de NLP de preentrenamiento	32
3.5.3 Módulo de Análisis	33
3.5.4 Módulo de Ejecución	34
3.5.5 Módulo de Resolución de AGO	35
3.6 Gestión de Riesgos	36
3.6.1 Supuestos	36
3.6.2 Dependencias	36
3.6.3 Restricciones	36
3.6.4 Riesgos	37
4 Desarrollo de la Solución	38
4.1 Preprocesamiento y Preparación de Datos	39

4.1.1	Estructuración de los datos	39
4.1.2	Oncotext	41
4.1.3	Estandarización de datos no estructurados	42
4.1.4	Modulo análisis de los datos estructurados	45
4.1.5	Módulo preparación de datos	46
4.1.6	Modulo de entrenamiento del modelo	47
5	Resultados y Discusión	50
5.1	Entrenamiento y Resultados del Modelo	50
5.1.1	1° Entrenamiento	50
5.1.2	2° Entrenamiento	53
5.1.3	3° Entrenamiento	54
5.1.4	4° Entrenamiento	54
5.2	Comparación del modelo entrenado v/s Sistemas de Inteligencia Artificial Conversacional	56
5.2.1	Modelo Desarrollado v/s ChatGPT GPT-5.3	56
5.2.2	Modelo Desarrollado v/s Gemini 3	60
6	Conclusión y Trabajo a Futuro	64
6.1	Conclusión	64
6.2	Trabajos a futuro	65
	Bibliografía	66

Índice de figuras

1.1	Defunciones cáncer de mama en Chile	13
1.2	Defunciones cáncer de colon en Chile	14
2.1	Proceso NLP aplicado en Radiología	20
2.2	Diagrama de arquitectura BERT	21
2.3	Diagrama de arquitectura XLNet	22
2.4	Diagrama de arquitectura CancerBERT	23
2.5	Modelo BI-LSTM-CRF	24
2.6	Modelo SpaCy	24
2.7	Diagrama Textblob	25
2.8	Diagrama medspaCy	25
2.9	Diagrama de extraccion de texto mediante OCR	26
3.1	Diagrama de contexto del sistema	28
3.2	Diagrama de arquitectura del sistema	29
3.3	Diagrama de Módulo de Preprocesamiento de Datos Clínicos	31
4.1	Vista de un término en Oncotext	41
4.2	Resultado de búsqueda	42
4.3	Flujo de estandarización	43
4.4	Flujo de estandarización	45
4.5	Dataframe para entrenamiento del modelo	47
4.6	Figura BioBERT	47
4.7	Entrenamiento del modelo con BioBert	48
5.1	Matriz de confusión del entrenamiento	51
5.2	Curva de aprendizaje 1° entrenamiento	52
5.3	Curva de aprendizaje 2° entrenamiento	53
5.4	Curva de aprendizaje 3° entrenamiento	54
5.5	Curva de aprendizaje 4° entrenamiento	55
5.6	Matriz de confusión del 4° entrenamiento	56
5.7	Prompt para contextualizar a ChatGPT	57
5.8	Prompt de prueba para ChatGPT	57
5.9	Respuesta de ChatGPT	58
5.10	Prompt con el archivo a analizar	58
5.11	Matriz de confusión ChatGPT	59
5.12	Respuesta del primer prompt de contextualización a Gemini	60

5.13 Respuesta del prompt de prueba a Gemini	61
5.14 Matriz de confusión Gemini	62

Índice de cuadros

2.1	Comparación de trabajos relacionados en NLP aplicado al dominio clínico.	27
3.1	Módulos de la arquitectura del sistema	30
3.2	Matriz de requisitos funcionales y módulos	30
3.3	Módulo de Preprocesamiento de Datos Clínicos	31
3.4	Módulo de NLP de preentrenamiento	32
3.5	Módulo de Análisis	33
3.6	Módulo de Ejecución	34
3.7	Módulo de Resolución de AGO	35
3.8	Riesgos y Medidas de Mitigación	37
4.1	Descripción de las hojas de cálculo de <code>Datos_anonimizados.xlsx</code>	39
4.2	Descripción de las hojas de cálculo de <code>Solicitud_ago_1.xlsx</code>	39
4.3	Descripción de las hojas de datos anonimizados del archivo <code>Solicitud_ago_2.xlsx</code>	40
4.4	Estructura de las hojas de cálculo de <code>Datos_anonimizados.xlsx</code>	40
4.5	Muestra de información del diccionario primario de Oncotext	44
4.6	Hiperparámetros usados y sus definiciones.	48
5.1	Comparación de métricas entre modelos	59
5.2	Comparación de métricas entre modelos	62

1. Introducción

Una de las enfermedades más comunes en el mundo en la actualidad es el cáncer de mama que afecta mayormente a mujeres y a hombre en menor cantidad, por otro lado una de las patologías con mayor tasa de mortalidad en el mundo es el cáncer de colon. Según la *Organización Mundial de Salud* (WHO), en el año 2020 hubo alrededor de 2 millones de mujeres con diagnostico de cáncer de mama, en donde 680000 perecen ante esta enfermedad[1]. Por otra parte, el cáncer colorrectal es el tercer tipo de cáncer más frecuente en el mundo, este representa aproximadamente 10 % de todos los casos de cáncer y es la segunda causa de muertes relacionada con este tipo de enfermedad del mundo según *Organización Mundial de Salud* (WHO)[2].

En Chile, el cáncer de mama se mantiene como la principal causa de muerte oncológica en personas de genero femenino. Según el Departamento de Estadísticas e Información de Salud (DEIS), en el año 2023 se registraron 2.263 fallecimientos por este padecimiento [3]. Por otro lado, datos del Observatorio Global del Cáncer (GLOBOCAN 2022) se estiman aproximadamente 5.640 nuevos casos anuales de cáncer de mama en Chile, considerándolo como el tipo de cáncer más habitual que padece población femenina nacional[4].

Por otra parte, el cáncer colorrectal se posiciona como uno de los tipo de cáncer de mayor impacto en Chile. Según reportes médicos ontológicos recientes, durante el año 2022 este tipo de cáncer se ubicó entre los más frecuentes tanto en hombres como en mujeres [5]. Además, la tasa de mortalidad ajustada en mujeres alcanzó 9,7 por cada 100.000 habitantes en 2019 [4]. En conjunto, tanto el cáncer de mama como el colorrectal destacan por su alta incidencia y mortalidad en Chile actualmente, lo que refuerza la necesidad de fortalecer estrategias de decisión o detección precoz y derivación oportuna.

1.1. Contexto

La *Fundación Arturo López Pérez* (FALP) [6] es una fundación sin fines de lucros que trabaja en Chile desde 1954. La FALP cuenta con un equipo médico altamente especializado en oncología el cual ofrece una amplia variedad de servicios y apoyo para sus pacientes. También esta institución cuenta con diversas misiones en donde se destacan los programas de educación, prevención y detección precoz del cáncer. Dentro de una de estas misiones se encuentra el programa de *Asesoramiento Genético Oncológico* (AGO), este permite determinar si existe algún factor hereditario del cáncer tanto para el paciente a evaluar como también para alguien de la rama hereditaria de este, a modo de prevención para cada uno.

En el día de hoy la FALP cuenta con distintas formas de acceder al programa AGO, una de las vías más importantes para acceder a este programa es la derivación realizada a través de consultas médicas realizada por los médicos oncológicos, estas consultas son denominadas evoluciones. Es importante el análisis de este medio de ingreso es de gran importancia, ya que un buen estudio de las evoluciones de los pacientes puede llevar a un correcto y oportuno derivado a AGO, lo que permite una detección temprana de un cáncer hereditario a un paciente y/o su rama familiar.

La *Fundación Arturo López Pérez*, es una institución oncológica de alta especialización en donde la gran parte de los pacientes con patologías relacionadas al cáncer son derivados a esta. Donde queda en registro todos los reportes médicos que han pasado por el paciente pertenecientes a la FALP o otra institución.

1.2. Definición del problema

En relación a lo descrito anteriormente, uno de los principales problemas para identificar si un paciente debe ser derivado a AGO, este debe esperar a que el médico oncólogo revise manualmente el reporte médico y discernir si el paciente es apto para ser derivado a *Asesoramiento Genético Oncológico* lo que esto presenta un problema, dado que afecta directamente al paciente dada a la espera que puede llevar la revisión adecuada de sus expedientes médico, como la detección de estas patologías que hacen que el paciente sea derivado a AGO. Una detección temprana de un cáncer hereditario permite tratar con tiempo al paciente o alguien de su rama familiar.

En este contexto, se propone implementar un sistema que tenga como entrada los reportes clínicos y datos del paciente. Obteniendo esta información, utilizando herramientas y técnicas de *Aprendizaje de Maquinas* para poder derivar a un paciente a AGO. Para poder derivar a un paciente a AGO existe las denominadas *Banderas Rojas* que en simples palabras son banderas de alerta que contienen los posibles síntomas que muestran si un paciente debe ser derivado, estas banderas van cambiando según el tipo de cáncer, en cuanto a este trabajo de trabaja con las Banderas Rojas de cáncer de mama y cáncer de colon.

Los dos tipos de cáncer recién mencionados son los que se presentan en mayor cantidad de padecimiento y alta mortalidad dentro de los pacientes que representa FALP. Esto se puede evidenciar en las figuras 1.1 y 1.2 estas fueron obtenidas en la pagina web del Departamento de Estadísticas e Información de Salud del Ministerio de Salud de Chile[3], donde podemos ver como ha ido evolucionando el numero de defunciones por estos 2 canceres a medida que ha pasado el tiempo. En este se puede apreciar que esta tendencia seguirá en alza y se espera que continué al pasar los años.

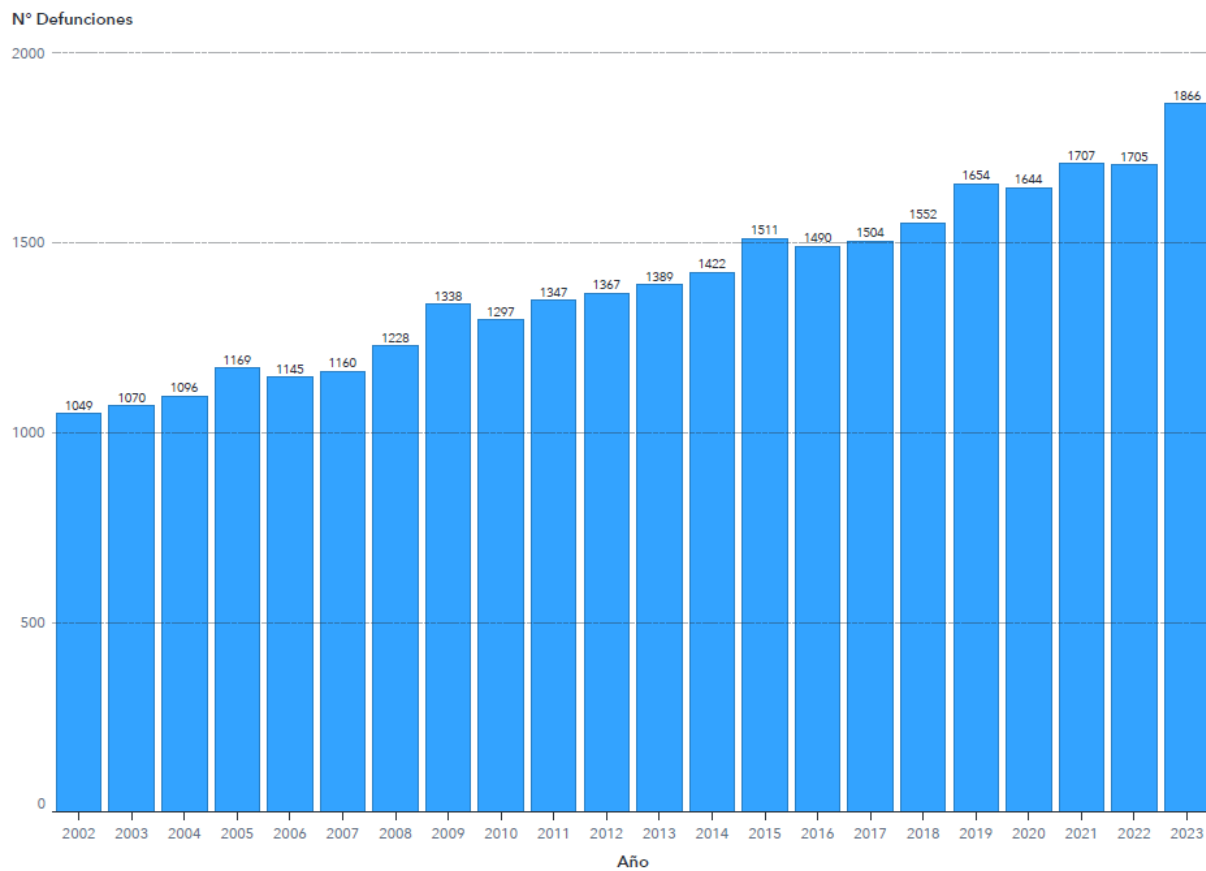


Figura 1.1: Defunciones cáncer de mama en Chile

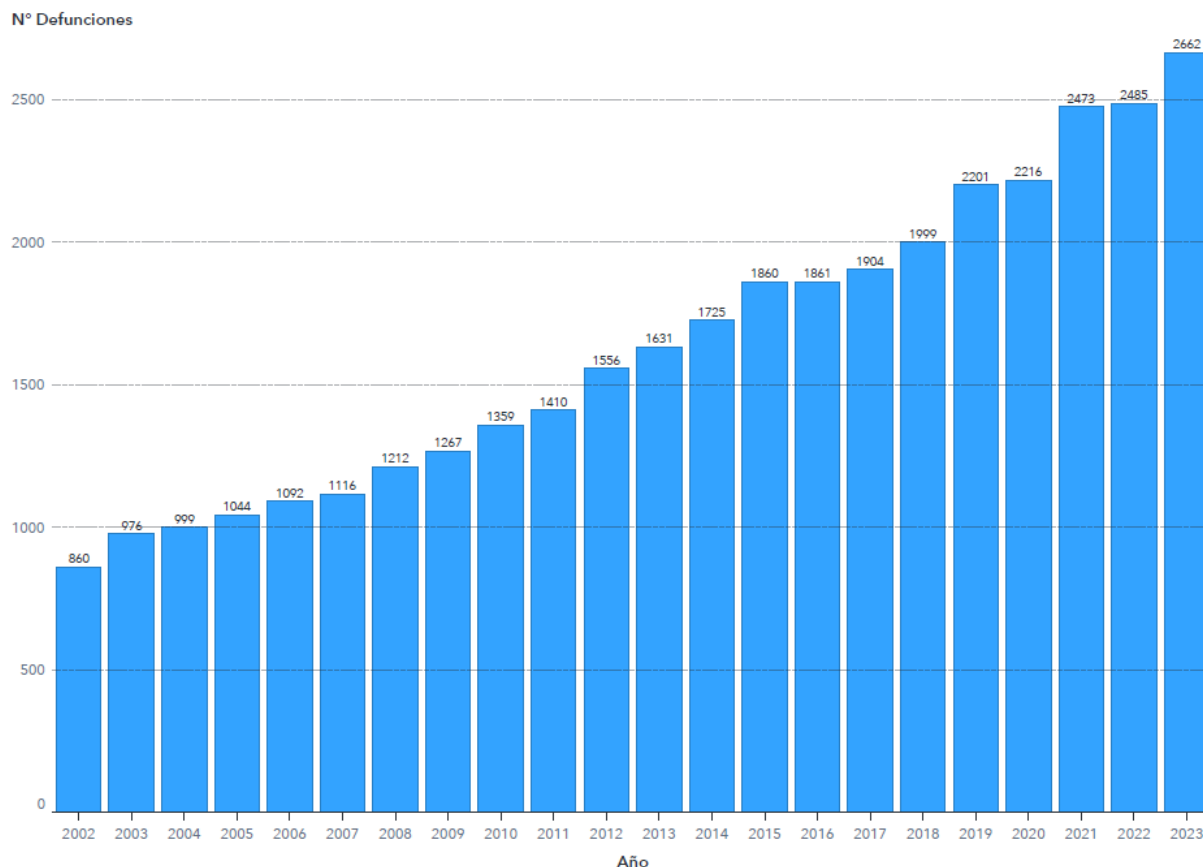


Figura 1.2: Defunciones cáncer de colon en Chile

Es por esto que la FALP, en su necesidad de modernizar y agilizar los procesos internos , necesita buscar una solución para que estas derivaciones se hagan de forma eficiente, en otras palabras, que los pacientes derivados efectivamente cumplan al menos una bandera roja, y que la detección de estas banderas rojas se haga de forma automática, ayudando así al equipo de asesoramiento oncológico a tomar decisiones con tiempo.

Para facilitar la solución al problema la FALP proporciono datos en base a documentos clínicos e informes de biopsias, estos datos vienen como datos estructurados y no estructurados. Estos son de suma importancia ya que nos proveerá el entrenamiento de nuestro modelo de predicción que permitirá definir si un paciente debe ser derivado a Asesoramiento Genético Oncológico.

En la presente memoria, se hará énfasis en la aplicación de un modelo de Procesamiento de Lenguaje Natural (NLP). Sin embargo, el desafío original fue abordado por tres memoristas, cuyos roles y funciones principales se detallan a continuación:

- **Axel Pérez:** Análisis de datos estructurados a partir de evoluciones clínicas y biopsias de pacientes pertenecientes a la Fundación Arturo López Pérez (FALP).
- **Eric Torres:** Implementación del modelo NLP y utilización de la herramienta Power BI para la visualización de resultados.

- **Tomás Vega:** Integración de la API de FALP para la identificación y procesamiento de términos médicos.

1.3. Hipótesis

- Es posible predecir si un paciente debe ser derivado a AGO mediante el entrenamiento de modelo de procesamiento de lenguaje natural para el caso de los exámenes médicos que representen a un dato no estructurado. Se espera enriquecer estos datos no estructurado realizando una limpieza y estandarización de estos, con el fin de realizar un entrenamiento de un modelo NLP con precisión aceptable. Esto permitirá con un sistema capaz de analizar exámenes médicos y biopsias para posteriormente asignar de manera correcta si un paciente debe ser derivado a AGO.
- Es posible definir antes si un paciente debe ser derivado a AGO sin necesidad de utilizar este modelo. Se espera examinar los datos estructurados dados por FALP y verificar si contiene una bandera roja para poder derivar a un paciente. Esto facilitaría el proceso de asignación, la toma de decisiones y optimizar los tiempos.

1.4. Usuarios

Se identificaron los siguientes perfil de usuarios:

1. Paciente: Persona natural que se atiende en FALP y que proporciona datos e informaciones referentes a su situación médica e historial clínico, lo cual es analizado por los profesionales de la institución.
2. Médico Oncológico: Es quien atiende al paciente, revisa su ficha clínica, levanta información, toma nota y determina si debe ser derivado a Asesoramiento Genético Oncológico (AGO).
3. Personal de equipo AGO: Revisa con quien se atendió el paciente y el resultado de dicha atención, coordina exámenes con el usuario y posteriormente verifica si posee cáncer hereditario.

1.5. Objetivos

El objetivo general del proyecto es crear un modelo de Machine Learning capaz de leer documentos clínicos y que mediante el entrenamiento de un modelo de predicción, alerte banderas rojas para que un paciente sea derivado a Asesoramiento Genético Oncológico

- Elaborar un plan de acción adecuado para abordar el desafío planteado, tras realizar un análisis previo exhaustivo del problema
- Búsqueda de metodologías de procesamiento de lenguaje natural para trabajar los datos no estructurados. Se enfatiza la limpieza y enriquecimiento de los datos, se evalúa obtener un modelo NLP para poder obtener palabras claves con el fin de detectar palabras importantes
- Analizar modelos de predicción, vinculando su uso con las técnicas de NLP. Se pretende realizar entrenamiento y prueba con los datos no estructurados dados por FALP a estos modelos encontrados.
- Desarrollar un sistema de predicción de banderas rojas, utilizando los datos disponibles por la FALP. Se escoge el modelo con la mejor precisión

1.6. Estructura

En el capítulo 2, Estado del Arte. Se hizo una investigación de lactucarios con métodos pertinentes en procesar datos clínicos no estructurados, desde técnicas de comparación de texto como *Pattern Matching* en Procesamiento Natural de Lenguaje como hasta métodos más complejos como *Word Embeddings*, en este caso con focal en BERT.

Capítulo 3, Diseño de la solución. En este capítulo se explica como se planifica los módulos para el desarrollo de la solución de la problemática estipulada, este capítulo se dividió en distintos steps.

Capítulo 4, Desarrollo de la solución. Se presenta el proceso de desarrollo del modelo creado, obteniendo métricas de evaluación. Denotan los steps descritos en el capítulo anterior y como se desarrollaron estos.

Capítulo 5, Resultados y Discusión. Por último, este capítulo se realiza las conclusiones obtenidas en el desarrollo e implementación del modelo propuesto, considerando el análisis realizado con los resultados obtenidos en el inciso anterior. Por otro lado, se discute las principales mejoras y limitaciones del desarrollo e implementación del modelo estipulado, proponiendo un trabajo a futuro para aumentar la robustez del modelo presentado.

2. Estado del Arte

2.1. Estado del arte y técnica

Una basta cantidad de datos clínicos de los pacientes vienen en formato de texto libre, es decir, en forma no estructurada. Dado lo mencionado esto presenta un gran desafío en el área de investigación por lo que han surgido una serie de metodologías o técnicas para extraer dicha información no estructuradas para realizar los análisis y predicciones posteriores.

Ref. [7]: los autores presentan una revisión integral sobre el uso de técnicas de Natural Language Processing (NLP) aplicadas al ámbito de la radiología. El estudio aborda el problema de la de los informes radiológicos dentro de los sistemas de registros médicos electrónicos (EMR), siendo estos datos no estructurados, lo que dificulta la extracción de información relevante de manera automática. Para ello, se presentan diversas técnicas de Natural Language Processing, incluyendo el reconocimiento de patrones, análisis de lenguaje y métodos basados en aprendizaje automático, cuyo fin es poder transformar datos no estructurados, es decir, reportes médicos en datos estructurados. Asimismo, los autores analizan múltiples aplicaciones, clasificación de reportes, identificación de patologías y optimización de procesos médicos. Los resultados evidencian que el uso de NLP permite automatizar proceso complejos con niveles de precisión aceptables a la revisión humana, posicionándose como una herramienta clave en el análisis masivo de datos clínicos y toma de decisiones.

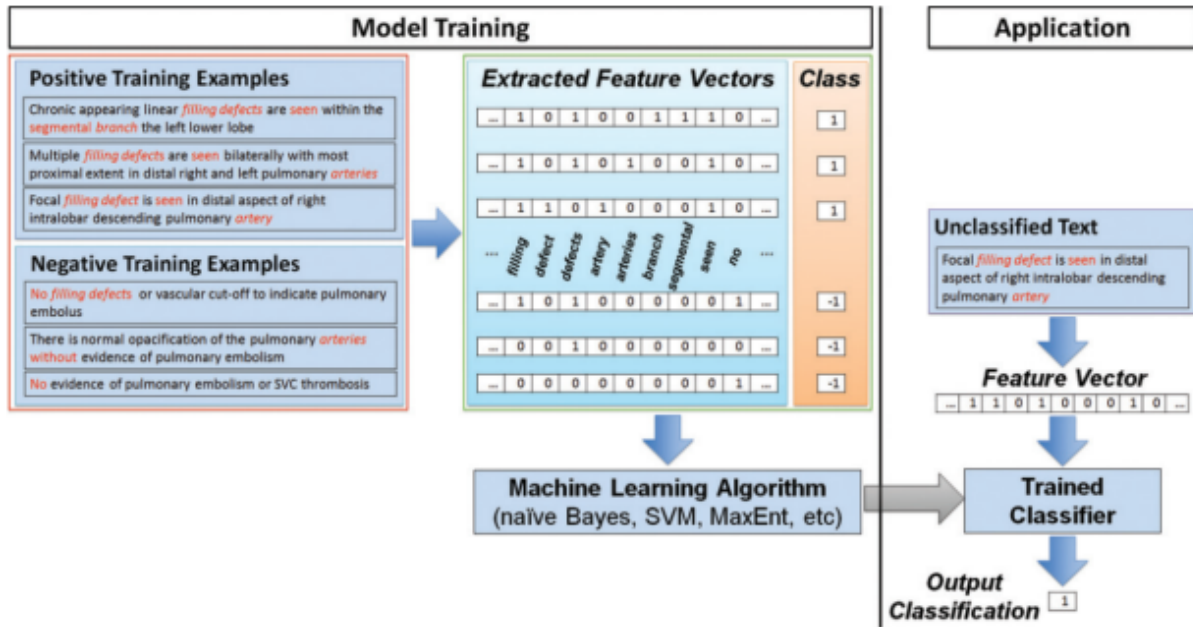
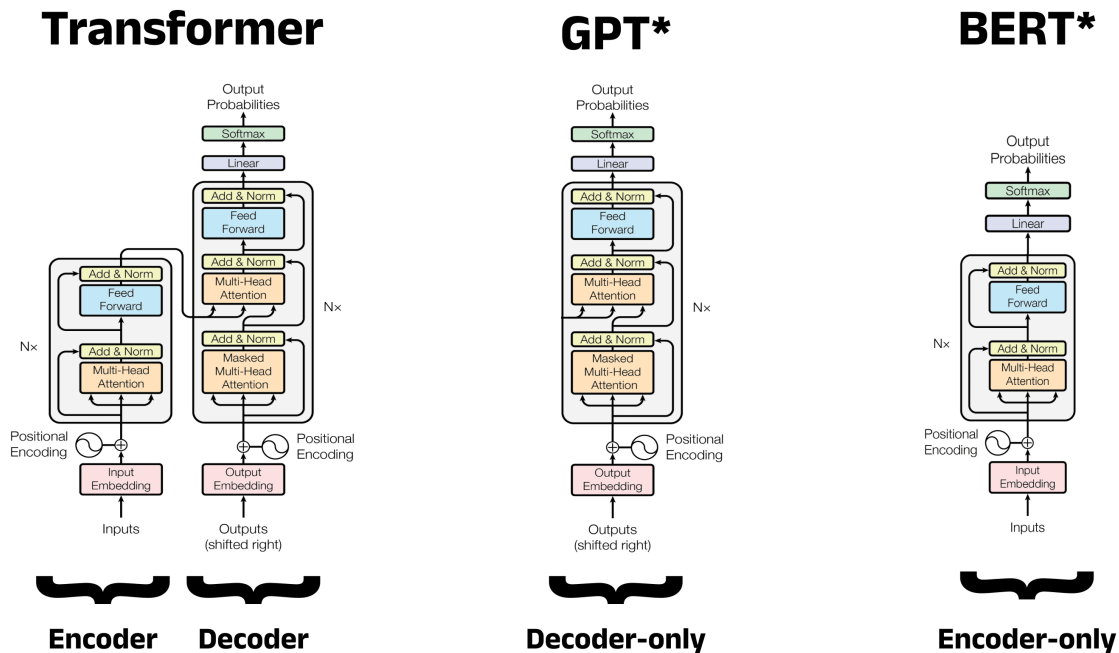


Figura 2.1: Proceso NLP aplicado en Radiología

Ref. [8]: los autores explican como funciona el modelo BERT (Bidirectional Encoder Representation from Transformers), en esta se describe en un modelo basado en redes neuronales profundas, donde este ha presentado un notorio avance en el Procesamiento de Lenguaje Natural. En este modelo se utiliza la arquitectura Transformers para entender el contexto de las palabras de forma bidireccional, es decir tomar el contexto anterior y el contexto posterior a la oración, en simples palabras de derecha a izquierda y de izquierda a derecha.

BERT emplea un mecanismo denominado *self-attention*, este mecanismo permite obtener relaciones entre palabras dentro de un texto, obteniendo mejoras en la clasificacion de textos y en el reconocimiento de entidades en estos. Por otro lado, el modelo BERT es pre-entrenado mediante técnicas tales como *Masked Language Modeling* y *Next Sentence Prediction*, lo que estas permiten aprender representaciones contextuales profundas del lenguaje antes de ajustarlo a tareas específicas.



*Illustrative example, exact model architecture may vary slightly

Figura 2.2: Diagrama de arquitectura BERT

Ref. [9]: los autores proponen un modelo denominado XLNet, este es un modelo pre-entrenado para el procesamiento natural de lenguaje basado con focal auto regresivo que busca superar las limitaciones que abordar las limitaciones que presenta el modelo BERT. Este trabajo identifica los principales que tiene el modelo BERT, esto es la dependencia de tokens enmascarados durante el entrenamiento de este y la dependencia entre palabras predichas.

Para abordar las limitaciones de BERT, el modelo XLNet introduce el concepto de *Permutation Language Modeling*, este permite modelar el contexto bidireccional sin necesidad de enmascarar la entrada. Esto se consigue maximizando la probabilidad de una secuencia observando todas las posibles permutaciones del orden de las palabras, permitiendo capturar dependencias tanto hacia adelante como hacia atrás en el texto.

Por otro lado, este modelo integra una arquitectura basada en Transformers-XL. incorporando metodologías de doble flujo **two-stream self-attention**, lo que permite separar la representación del contenido y la posición objetivo, mejorando la capacidad del modelo para predecir tokens de manera contextual.

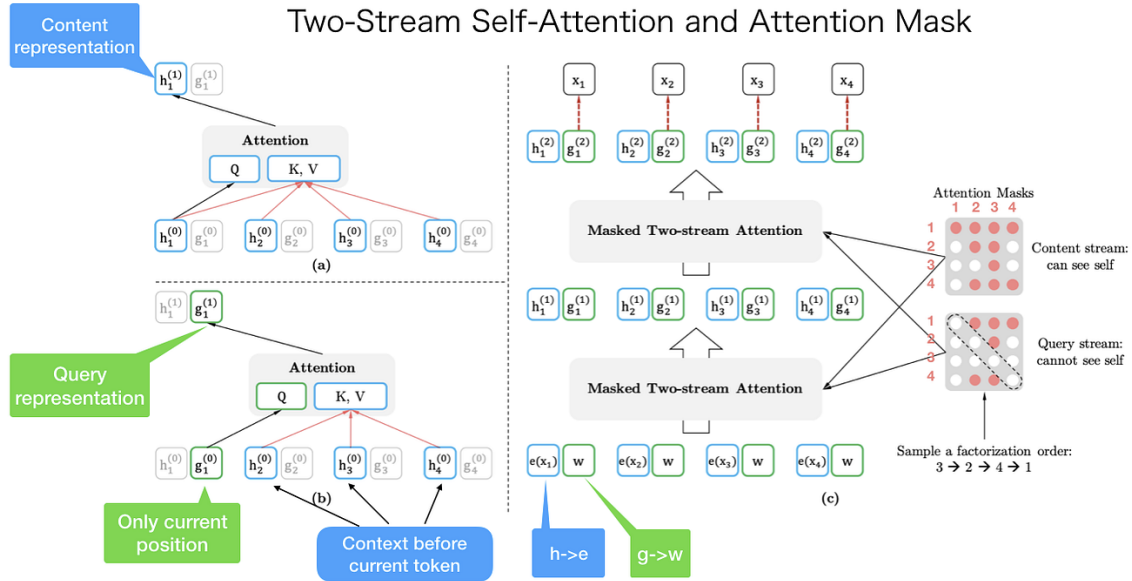


Figura 2.3: Diagrama de arquitectura XLNet

Ref[10]: los autores exponen CancerBERT, un modelo de lenguaje específico de índole oncológica diseñado para la extracción de fenotipos de cáncer de mama a partir de reportes médicos electrónicos (EHR, Electronic Health Record). El estudio aborda la problemática de procesar información clínica no estructurada (datos no estructurados), como reportes médicos y de patología, los cuales contienen información crítica para la toma de decisiones clínicas.

El modelo expuesto se basa en la arquitectura BERT y se entrena adicionalmente sobre un contexto clínico específico de cáncer, compuesto por millones de documentos provenientes de registros recintos clínicos. Este enfoque permite capturar de manera precisa la semántica del lenguaje clínico especializado. Además, el estudio introduce una mejora mediante la incorporación de un vocabulario específico de índole ontológica, con el objetivo de mitigar el problema de palabras fuera de vocabulario (out-of-vocabulary, OOV), lo cual impacta directamente en el rendimiento del modelo.

Para evaluar el modelo, los autores formulan una tarea de reconocimiento de entidades nombradas (NER), con focal en la identificación de ocho tipos de fenotipos de cáncer de mama, tales como tamaño tumoral, estadio del cáncer y tipo histológico. Los resultados denotan que CancerBERT supera a modelos vistos como BERT, alcanzando valores de F1 superiores (hasta 0.876), lo que demuestra la efectividad de los modelos preentrenados en dominios específicos para tareas clínicas.

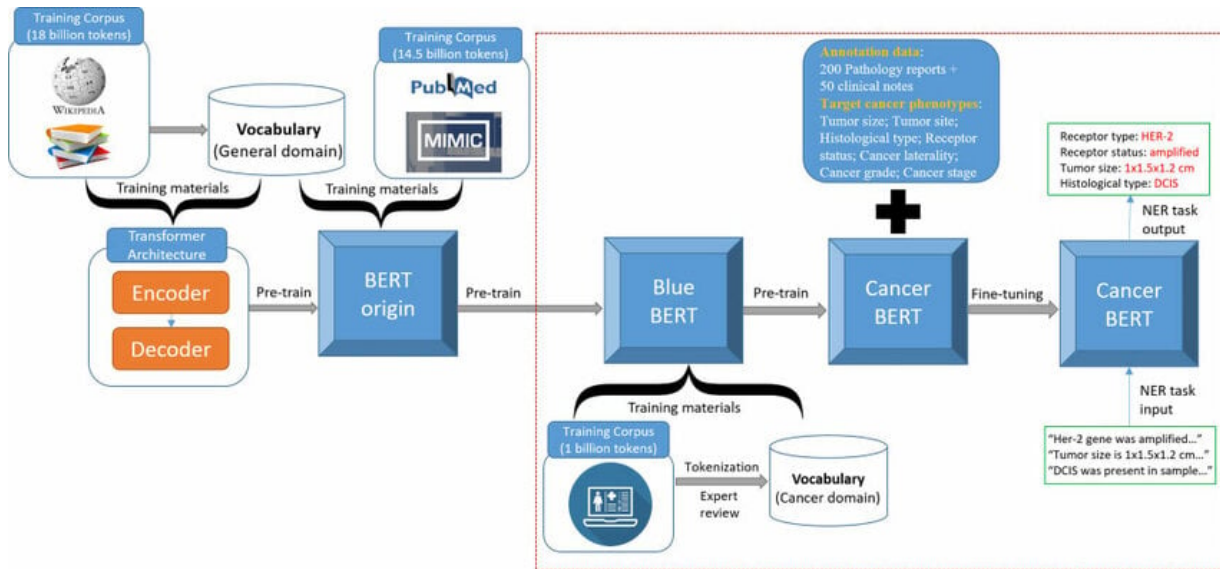


Figura 2.4: Diagrama de arquitectura CancerBERT

Ref[11]: los autores presentan un modelo llamado *HEterogeneous-target Robust MIX-ture regression*, el cual expone los enfoques tradicionales de regresión mediante la incorporación de técnicas de Finite Mixture Regression (FMR) para abordar casos con alta diversidad en los datos.

El principal objetivo del modelo es capturar relaciones complejas entre variables de entrada y múltiples variables objetivo, considerando que estas pueden pertenecer a diferentes distribuciones. Para esto, el enfoque plantea la existencia de múltiples sub-poblaciones dentro de los datos, donde cada una es modelada mediante una regresión específica.

Este artículo se descarta su utilidad, dado que el modelo BERT presenta una mejor potencial solución para abordar la problemática presentada.

Ref[12]: los autores proponen un modelo basado en redes neuronales profundas para tareas de *sequence tagging*, en este artículo proponen distintos modelos que combinan arquitecturas como *Long Short-Term Memory (LSTM)* con *Conditional Random Fields (CRF)*.

El estudio introduce y compara varias arquitecturas: LSTM, Bidirectional LSTM (BI-LSTM), LSTM-CRF y BI-LSTM-CRF. La principal innovación radica en la combinación de redes neuronales recurrentes con modelos probabilísticos, permitiendo aprovechar tanto la información contextual de las palabras como las dependencias entre etiquetas.

En particular, el modelo BI-LSTM-CRF destaca por integrar tres componentes clave:

- LSTM bidireccional: permite capturar información pasada y futura dentro de la secuencia de texto.
- Capas recurrentes: modelan dependencias de largo alcance entre palabras.
- Capa CRF: considera la dependencia entre etiquetas, optimizando la secuencia completa en lugar de predicciones individuales.

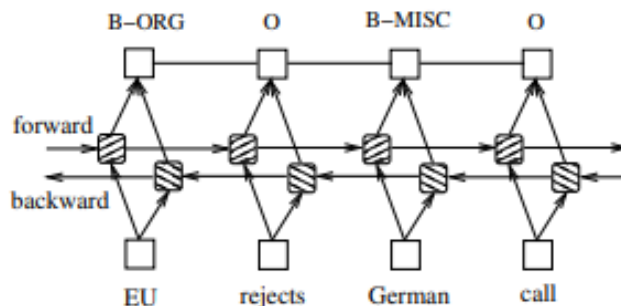


Figura 2.5: Modelo BI-LSTM-CRF

En la Ref. [13], los autores presentan spaCy, una biblioteca de procesamiento de lenguaje natural de Python la cual se encuentra en código abierto, esta es desarrollada para facilitar la implementación eficiente de sistemas de análisis de textos. Este framework está diseñado para ofrecer alto rendimiento en tareas fundamentales de NLP, tales como tokenización, análisis sintáctico (Dependency Parsing) y reconocimiento de entidades nombradas (NER).

Una de las principales características de spaCy es su enfoque en la eficiencia y escalabilidad, permitiendo procesar grandes volúmenes de texto con baja latencia. Además, la herramienta incorpora modelos pre-entrenados optimizados y permite su personalización mediante técnicas de aprendizaje supervisado, facilitando su adaptación a dominios específicos como en nuestro caso clínico.

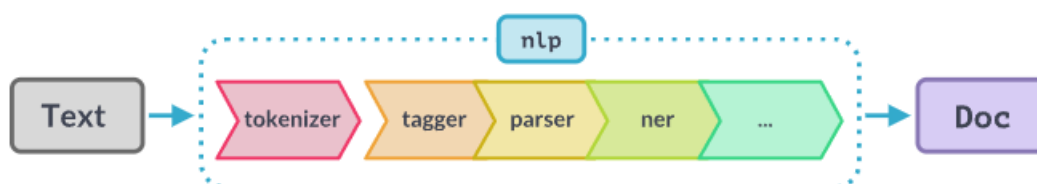


Figura 2.6: Modelo SpaCy

En la Ref. [14], los autores proponen TextBlob, una biblioteca de procesamiento de lenguaje natural en Python. Esta herramienta se basa sobre librerías ya existentes como NLTK y Pattern, proporcionando funcionalidades accesibles para tareas como tokenización, análisis de sentimiento, clasificación de texto, etiquetado gramatical (part-of-speech tagging) y traducción de texto.

A diferencia de frameworks más complejos, TextBlob se caracteriza por su fácil implementación, lo que la convierte en una opción adecuada para prototipado y desarrollo de aplicaciones simples de NLP. Sin embargo, su enfoque simplificado implica ciertas limitaciones en términos de rendimiento y precisión frente a modelos más avanzados basados en aprendizaje profundo.

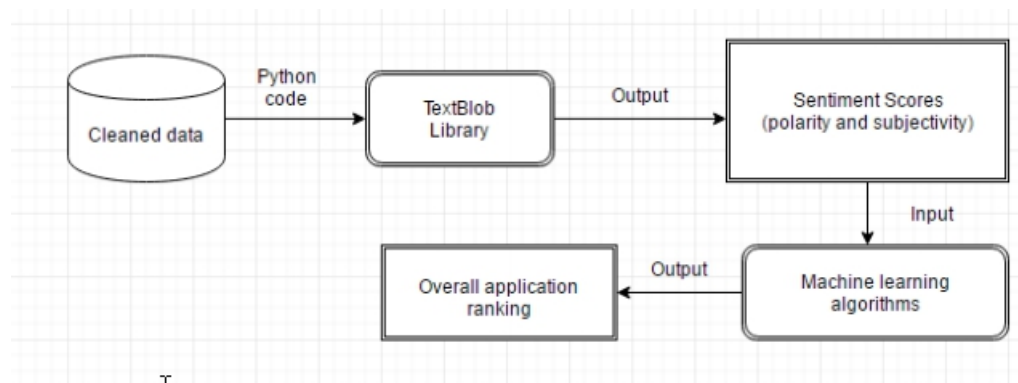


Figura 2.7: Diagrama Textblob

En la Ref. [15], los autores presentan medspaCy, una extensión de la biblioteca ya mencionada anteriormente spaCy, medspaCy está diseñada específicamente para el procesamiento de texto clínico. El objetivo principal de esta herramienta es facilitar la implementación de sistemas de NLP en contextos médicos, abordando las particularidades del lenguaje clínico, tales como abreviaciones, terminología especializada y estructuras lingüísticas complejas.

El framework medspaCy incorpora módulos orientados con objetivo médico, este incluye reconocimiento de entidades médicas, detección de negación, identificación de contexto clínico (por ejemplo, historial familiar o condiciones actuales) y extracción de información estructurada desde registros médicos electrónicos (EHR). Estos componentes permiten mejorar la interpretación de textos clínicos no estructurados.

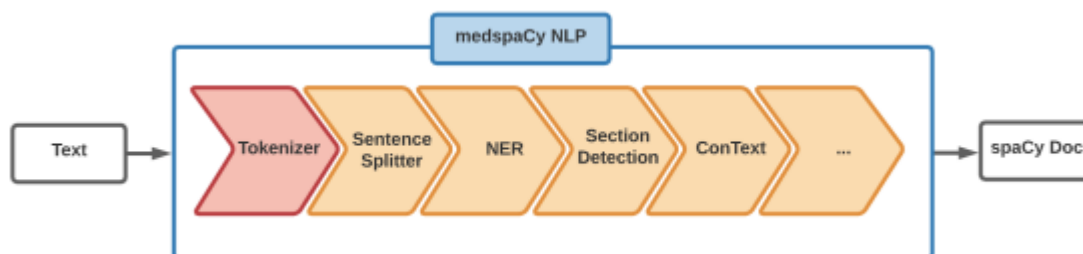


Figura 2.8: Diagrama medspaCy

En la Ref. [16], los autores exponen un enfoque para la extracción de texto desde imágenes de documentos médicos, combinando técnicas de reconocimiento óptico de caracteres (Optical Character Recognition, OCR) con métodos de procesamiento de lenguaje natural. El objetivo principal del estudio es transformar documentos médicos en formato imagen, tales como informes escaneados, en información estructurada que pueda ser procesada automáticamente.

El enfoque planteado consta de múltiples etapas. En primer lugar, se aplica un proceso de OCR para convertir el contenido visual en texto digital. Posteriormente, este texto es procesado mediante técnicas de NLP para identificar información pertinente, tales como entidades médicas, diagnósticos y términos clínicos. Este pipeline permite abordar uno de

los principales desafíos en el ámbito de la salud: la existencia de grandes volúmenes de datos clínicos en formatos no estructurados y no digitalizados.

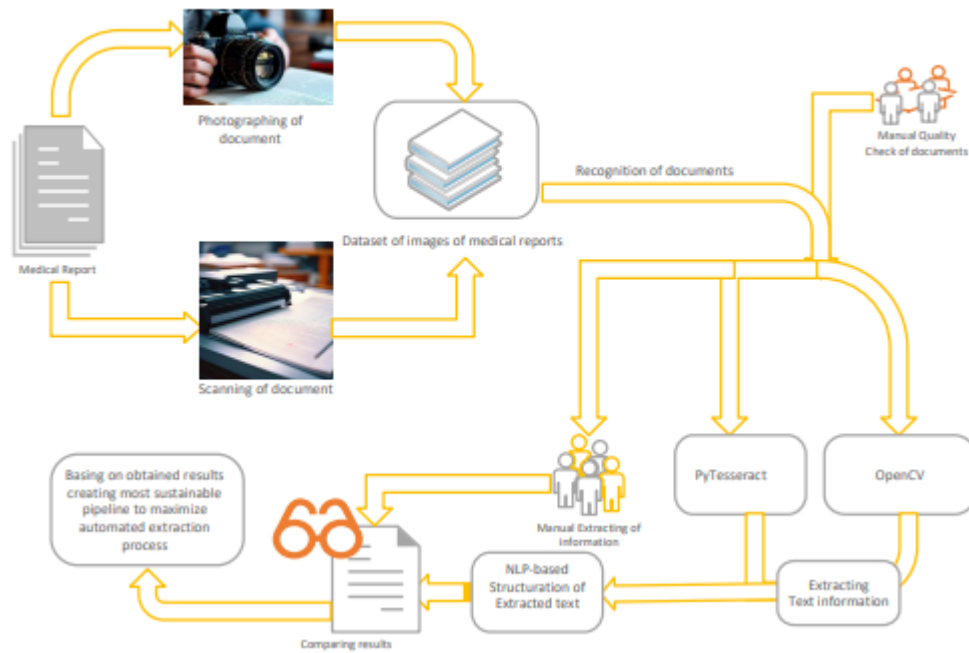


Figura 2.9: Diagrama de extracción de texto mediante OCR

Cuadro 2.1: Comparación de trabajos relacionados en NLP aplicado al dominio clínico.

Ref.	Año	Enfoque	Aplicación	Técnicas	Contribución principal
[1]	2016	NLP en radiología	Análisis clínico	Pattern Matching, NLP, ML	Estructuración de texto médico no estructurado.
[2]	2018	Modelo de lenguaje	NLP	Transformers	Comprensión contextual bidireccional del lenguaje.
[3]	2019	Modelo avanzado	NLP	Transformer-XL, Permutation LM	Mejora sobre BERT sin uso de enmascaramiento.
[4]	2020	NLP clínico	Oncología	BERT, NER	Extracción de fenotipos clínicos desde EHR.
[5]	2020	Machine Learning avanzado	Datos heterogéneos	FMR, EM, MTL	Modelado de heterogeneidad y subpoblaciones.
[6]	2015	Deep Learning NLP	Sequence Tagging	BI-LSTM, CRF	Modelo de alto desempeño para NER y etiquetado secuencial.
[7]	2023	Framework NLP	Aplicaciones industriales	Pipeline NLP, NER	Implementación eficiente de procesamiento de lenguaje.
[8]	2021	Framework NLP	Procesamiento básico	NLP clásico, NLTK	Simplificación de tareas básicas de procesamiento de texto.
[9]	2021	NLP clínico aplicado	EHR	spaCy, reglas, ML	Adaptación de NLP en el contexto médico.
[10]	2024	OCR + NLP	Reportes médicos	OCR, NLP	Extracción de texto desde imágenes médicas.
Trabajo desarrollado	2026	NLP	Oncología	NLP, clasificación	Derivación automática de pacientes a AGO, mejora en la toma decisiones oncológicas.

3. Diseño de la solución

3.1. Diagrama de Contexto

El diagrama de contexto nos permite visualizar las interacciones entre los diferentes componentes y actores con respecto al sistema a desarrollar. En la Figura 4.1 se puede observar el diagrama en como los usuarios interactúan con el sistema.

3.1.1. Diagrama de Contexto

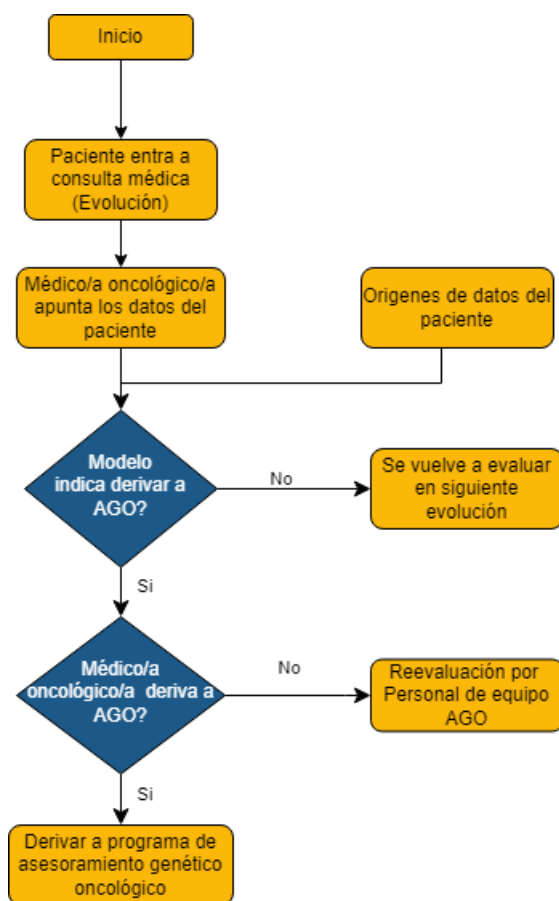


Figura 3.1: Diagrama de contexto del sistema

3.2. Diagrama de Arquitectura

La Figura 3.2 ilustra los componentes y módulos del sistema, junto con sus interacciones. El proceso comienza con la entrada de reportes clínicos del paciente, que incluyen datos estructurados y no estructurados. Luego, se realiza el filtrado y transformación de información para generar los dataframes.

Los datos procesados pasan por el módulo de limpieza, seguido del pre-entrenamiento y entrenamiento del modelo principal. Finalmente, el sistema predice si el paciente debe ser derivado a AGO, indicando su ID, bandera roja y tipo de cáncer.

3.2.1. Diagrama de Arquitectura

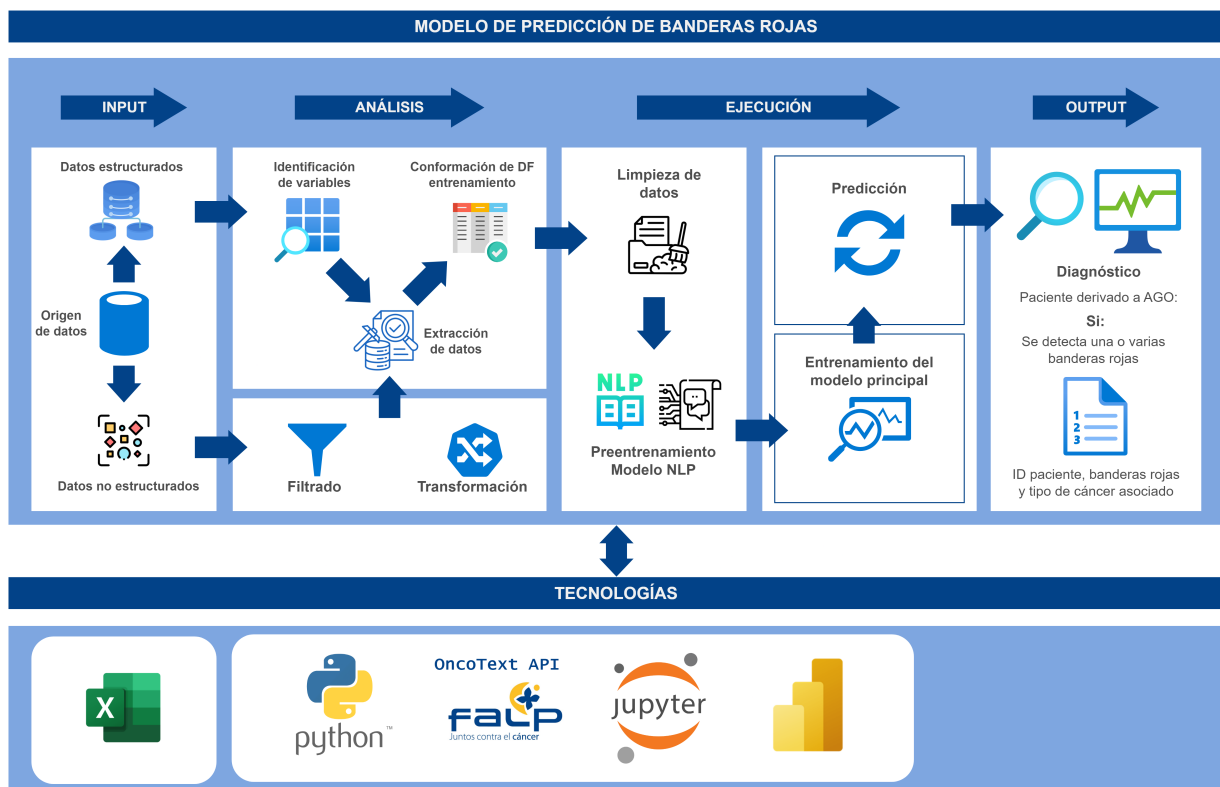


Figura 3.2: Diagrama de arquitectura del sistema

3.3. Enumeración de Módulos

Cuadro 3.1: Módulos de la arquitectura del sistema

Módulo	Propósito	Sección
Módulo de Preprocesamiento de Datos Clínicos	Se encarga del preprocesamiento de los datos del paciente, los cuales se presentan como el origen de datos ingresado en la entrada o input. Estos se clasifican como estructurados o no estructurados.	5.1.1
Módulo de NLP	Utiliza las técnicas de procesamiento del lenguaje natural (NLP) en los datos no estructurados para la identificación de los datos críticos en los textos de las evoluciones de los pacientes.	5.1.2
Módulo de Análisis	Permite identificar y extraer los datos obtenidos en el módulo de preprocesamiento de datos clínicos y en el módulo de NLP, para luego seleccionar el algoritmo adecuado para su tratado en el módulo de ejecución, encargado de la predicción.	5.1.3
Módulo de Ejecución	Su finalidad es filtrar, transformar y predecir con una alta probabilidad si el paciente posee banderas rojas que indiquen que este sea derivado a AGO, entregando la información pertinente en donde se justifique dicha recomendación. Para ello es utilizado un modelo previamente entrenado.	5.1.4
Módulo de Resolución de AGO	Se encarga de informar los resultados de los módulos anteriores, entregando las respectivas métricas calculadas en su ejecución. Informa por pantalla la precisión con respecto a la incógnita de si el paciente debe ser derivado a AGO.	5.1.5

3.4. Matriz de Requisitos Funcionales y Módulos

Cuadro 3.2: Matriz de requisitos funcionales y módulos

	M1	M2	M3	M4	M5
RF1	x	x			
RF2				x	x
RF3				x	x
RF4	x	x	x	x	x
RF5		x	x	x	x
RF6					x

3.5. Módulos

3.5.1. Módulo de Preprocesamiento de Datos Clínicos

Cuadro 3.3: Módulo de Preprocesamiento de Datos Clínicos

Propósito	Se encarga del preprocesamiento de los datos del paciente, los cuales se presentan como el origen de datos ingresados en la entrada o input. Estos se clasifican como estructurados o no estructurados.
Alcance	Es la entrada principal de datos, ya que la información recepcionada de este módulo será fundamental en el análisis. Este módulo se encarga de decidir con qué algoritmo trabajar de acuerdo al tipo de datos.
Marco Técnico del Módulo	Se realiza la lectura de el origen de datos y se clasifican de acuerdo a su estructura. Los datos son enviados a los siguientes módulos considerando el tipo de clasificación.
Estructura General	El módulo de Preprocesamiento de Datos Clínicos recibirá como entrada los distintos datos clínicos del paciente, identificando la estructura de ellos y luego aplicando las técnicas mencionadas, para clasificar la información de acuerdo al contexto entregado.

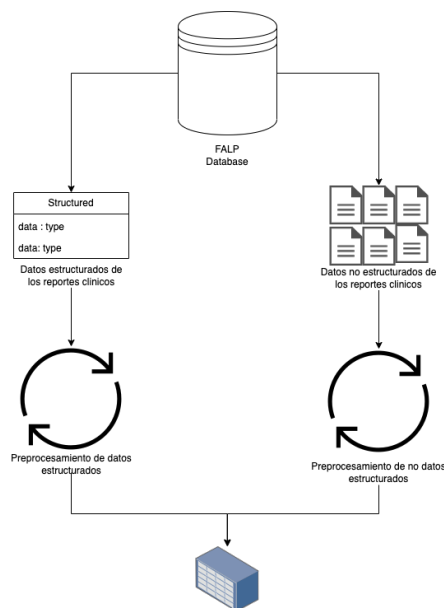


Figura 3.3: Diagrama de Módulo de Preprocesamiento de Datos Clínicos

3.5.2. Módulo de NLP de preentrenamiento

Cuadro 3.4: Módulo de NLP de preentrenamiento

Propósito	Utiliza las técnicas de procesamiento del lenguaje natural (NLP) en los datos no estructurados para la identificación de los datos críticos en los textos de las evoluciones de los pacientes.
Alcance	Es la entrada y procesamiento de los datos no estructurados, los cuales vienen del origen de datos. Su apropiado funcionamiento es fundamental para la identificación de las variables en los textos de las evoluciones de los pacientes, y así identificar posibles banderas rojas.
Marco Técnico del Módulo	Los datos no estructurados son procesados mediante metodologías NLP, tales como BERT, además se apoya en los términos y tags proporcionados por Oncotext, los cuales están alojados a modo de diccionario y servirán como “vocabulario” para el funcionamiento del Transformer.
Estructura General	El módulo de NLP recibe los datos no estructurados, los cuales son textos con las evoluciones médicas de los pacientes. Luego, utilizando las técnicas de NLP, permite identificar las variables críticas para el análisis, generando el modelo de preentrenamiento que será utilizado posteriormente por el módulo de ejecución para el modelo principal.

3.5.3. Módulo de Análisis

Cuadro 3.5: Módulo de Análisis

Propósito	Permite identificar, extraer y ordenar los datos obtenidos en el módulo de preprocesamiento de datos clínicos y los resultados preliminares del módulo de NLP de preentrenamiento.
Alcance	La recopilación de las variables identificadas, tanto en la data estructurada como en la no estructurada, permite la conformación de un dataframe con resultados preliminares para la derivación de pacientes a asesoramiento genético oncológico.
Marco Técnico del Módulo	Se extraen los datos de la parte estructurada y no estructurada, permitiendo la conformación de un dataframe para ser utilizado como datos de entrenamiento del modelo principal.
Estructura General	El módulo organiza y extrae las variables relevantes de los datos estructurados y no estructurados. Posteriormente, selecciona los resultados para conformar el modelo preentrenado.

3.5.4. Módulo de Ejecución

Cuadro 3.6: Módulo de Ejecución

Propósito	Su finalidad es mejorar los resultados de el modelo preentrenado, utilizando como base dicho modelo y además la información de los tipos de banderas rojas de cáncer de mama y colon.
Alcance	Este módulo se encarga de definir, con los datos obtenidos en los módulos anteriores, la existencia de banderas rojas, además de clasificar a cual tipo de cáncer corresponde. Luego su resultado se entrega al siguiente módulo.
Marco Técnico del Módulo	Se encarga de entrenar el modelo principal en base a los resultados de los módulos anteriores, entregando además las métricas necesarias para concluir sobre su rendimiento.
Estructura General	Recibe las variables identificadas en el módulo de análisis y haciendo uso del modelo previamente entrenado, se genera el resultado del modelo principal con sus respectivos rendimientos.

3.5.5. Módulo de Resolución de AGO

Cuadro 3.7: Módulo de Resolución de AGO

Propósito	Se encarga de informar los resultados de los módulos anteriores, entregando las respectivas métricas calculadas en su ejecución. Informa por pantalla la precisión con respecto a la incógnita de si el paciente debe ser derivado a AGO.
Alcance	Recibe la información concluyente de los módulos anteriores y, en base a ello, determina el resultado del modelo, recomendando al médico oncológico si el paciente debe ser derivado a AGO.
Marco Técnico del Módulo	Se reciben los resultados de los módulos anteriores, los cuales permiten en el análisis y contraste con el listado de banderas rojas proporcionadas por FALP, definir si un paciente debe ser derivado a AGO.
Estructura General	El módulo de Resolución de AGO recibirá como entrada la información concluyente de los módulos anteriores y determinará, mediante las técnicas seleccionadas, la precisión con que un paciente puede o no ser derivado a asesoramiento genético oncológico (AGO).

3.6. Gestión de Riesgos

3.6.1. Supuestos

1. Contaremos con los requisitos, tanto de hardware como de software para la realización del modelo.
2. Todas las partes del equipo de trabajo estarán en conocimiento sobre la problemática, propuestas y sus soluciones.
3. La información otorgada por la FALP será la suficiente para poder realizar un modelo funcional.
4. El modelo será usable por cualquier persona que tenga en sus manos la documentación, por lo que no requerirá una capacitación extra.
5. La información necesaria para ejecutar el modelo estará en el formato correcto.

3.6.2. Dependencias

1. Nuestro cliente accederá a enviar información extra de ser necesario.
2. Además de obtener información sobre la cual debe trabajar el modelo, debemos tener disponible la información que nos ayudará a modelar la solución.

3.6.3. Restricciones

1. Deberemos trabajar sobre Python y complementarlo con Jupyter, para realizar una buena documentación.
2. Primeramente debemos centrarnos en la creación del modelo, una vez hecho, si es posible, avanzar en la creación de una aplicación.

3.6.4. Riesgos

Cuadro 3.8: Riesgos y Medidas de Mitigación

Riesgo	Medida de Mitigación
No se encuentra una forma adecuada de analizar los datos no estructurados	Se coordina el uso de Oncotext para la identificación del lenguaje médico utilizado por los profesionales de la FALP. También se investigan papers sobre el uso de algoritmos de procesamiento de datos estructurados y no estructurados.
El modelo arroja resultados erróneos	Se investigan distintos modelos de clasificación y predicción, además de coordinar reuniones semanales para la definición de datos con los que trabajar y que puedan aportar al correcto funcionamiento del modelo.
La planificación de los hitos puede ser imprecisa y pueden haber retrasos	Coordinación de reuniones semanales para muestra de avances y uso de la metodología ágil de Scrum en el equipo de trabajo.
La integración del sistema	Se coordina la entrega de un modelo desarrollado en lenguaje de programación Python y ejecutado en notebooks de Jupyter Notebook. Se presentan los instructivos necesarios para la instalación de las bibliotecas empleadas.
Seguridad e integridad de los datos	Trabajo con queries de bases de datos a modo de no afectar los centros de recursos de la FALP. Se procede además con una limpieza de datos a modo de resguardar la identidad de los pacientes y se firma un contrato de confidencialidad.

4. Desarrollo de la Solución

Este capítulo describe en detalle el trabajo realizado, el cual se conforma en distintos módulos para el desarrollo del modelo de predicción, cuyo focal es identificar que pacientes requieren ser derivados a AGO. El enfoque principal se basa en el uso de datos clínicos brindados por FALP, estos son la base del diseño, implementación y evaluación del modelo.

4.1. Preprocesamiento y Preparación de Datos

Los datos clínicos brindados por FALP se presentan en archivos .XLSX (Hojas de Cálculo de Microsoft Excel), estas hojas contienen la información de los pacientes, sus evoluciones medicas y reporte de exámenes. Para utilizar los datos entregados en modelos de procesamiento de lenguaje natural (NLP), es necesario estructurarla y establecer una disposición uniforme de los datos. En simples términos, se debe recoger, organizar y preparar los datos que servirán como parámetro de entrada al modelo a entrenar.

4.1.1. Estructuración de los datos

Inicialmente, se detallará qué información contiene cada archivo entregado por FALP, los cuales corresponden a un total de tres archivos en formato .xlsx, estos son: *Datos_anonimizados*, *Solicitud_ago_1* y *Solicitud_ago_2*.

Cuadro 4.1: Descripción de las hojas de cálculo de *Datos_anonimizados.xlsx*

Nombre hoja	Contenido
Evoluciones 2022	Listado de todas las evoluciones realizadas en 2022. Cada registro contiene el ID del paciente, fecha de evolución, CIE10, evolución y diagnóstico.
Primera evolución	Listado de las primeras evoluciones realizadas a los pacientes. Cada registro contiene el ID del paciente, fecha de evolución, CIE10 y evolución.
Fish	Listado de los exámenes Fish realizados a los pacientes. Cada registro contiene el ID del paciente, fecha de realización y resultado.
Biopsias	Listado de biopsias realizadas a los pacientes. Cada registro contiene el ID del paciente, fecha de realización y diversos resultados.
Pacientes	Listado de pacientes. Cada registro contiene el ID del paciente, fecha de nacimiento, sexo, CIE10 y diagnóstico.

Cuadro 4.2: Descripción de las hojas de cálculo de *Solicitud_ago_1.xlsx*

Nombre hoja	Contenido
Evoluciones	Listado de evoluciones de pacientes que fueron derivados a AGO en los años 2020-2021. Cada registro contiene el ID del paciente, fecha de evolución, CIE10, evolución y diagnóstico.
Biopsias	Listado de biopsias de pacientes que fueron derivados a AGO en los años 2020-2021. Cada registro contiene el ID del paciente, fecha de realización y diferentes resultados.

Cuadro 4.3: Descripción de las hojas de datos anonimizados del archivo *Solicitud_ago_2.xlsx*

Nombre hoja	Contenido
solicitud_interconsulta	Listado de pacientes que fueron derivados a AGO en los años 2022-2023. Cada registro contiene el ID del paciente, fecha de solicitud, diagnóstico y comentario.
prestaciones	Listado de pacientes que fueron atendidos por AGO, luego de su derivación.

El archivo principal que ha sido utilizado en el desarrollo del modelo es *Datos_anonimizados.xlsx*, este contiene múltiples datos con respecto al paciente, los cuales ciertos datos no están asociados directamente a las banderas rojas, no obstante, se elimina los registro que contengan al menos un datos en blanco, como las hojas de calculo que no tengan relación directa e indirecta con las banderas rojas. EL contenido de este archivo en cuanto datos se puede definir de la siguiente forma:

Cuadro 4.4: Estructura de las hojas de cálculo de *Datos_anonimizados.xlsx*

Nombre hoja	Contenido
Pacientes	Sexo, fecha_nacimiento, paciente_id, CIE10
Evoluciones	paciente_id, fecha_evolucion, evolucion.
Biopsias	paciente_id, fecha_biopsia, informe, inf_macro, inf_micro, informe_add

Este archivo se divide en tres hojas de calculo: pacientes, evoluciones y biopsias, cada una de estas hojas contiene información importante para el modelo, que se define a continuación:

1. **Hoja de pacientes:** Contiene las siguientes columnas:

- a) **sexo:** Corresponde al sexo del paciente y tiene dos posibles valores; MA y FE, en donde MA corresponde a masculino y FE a femenino.
- b) **fecha_nacimiento:** Es la fecha de nacimiento del paciente, es importante que esté en formato AAAA-MM-DD.
- c) **paciente_id:** Corresponde al id del paciente.
- d) **CIE10:** Acrónimo de Clasificación Internacional de Enfermedades, 10.^a edición. Este código representa la clasificación de la enfermedad. En este caso, si el CIE10 de un paciente comienza con CIE18, corresponde a un paciente con cáncer de colon, y si comienza con CIE50, corresponde a cáncer de mama.

2. **Hoja de evoluciones:** Contiene todas las evoluciones posibles, describiéndose con las siguientes columnas:

- a) **paciente_id**: Corresponde al id del paciente.
- b) **fecha_evolución**: Es la fecha en la que el paciente tuvo la evolución.
- c) **evolución**: Corresponde a la descripción clínica que el médico escribe en formato texto libre, durante la consulta con el paciente.

3. **Hoja de biopsias**: Contiene las biopsias realizadas al paciente, contiene las siguientes columnas:

- a) **paciente_id**: Corresponde al id del paciente.
- b) **fecha_biopsia**: Fecha de realización de la biopsia.
- c) **informe**: Informe de la biopsia.
- d) **inf_macro**: Informe macro acerca de la biopsia.
- e) **inf_micro**: Informe micro acerca de la biopsia.
- f) **informe_add**: Informe adicional.

4.1.2. Oncotext

Oncotext es una herramienta desarrollada por el IMDS (Departamento de Informática Médica y Data Science) de la FALP para búsqueda y extracción de términos y mediciones desde un gran volumen de documentos de texto libre.

El uso de esta herramienta nos permite extraer información desde textos a través de tareas a procesar, por ejemplo, a partir de una evolución de un paciente podemos buscar términos asociados a antecedentes familiares.

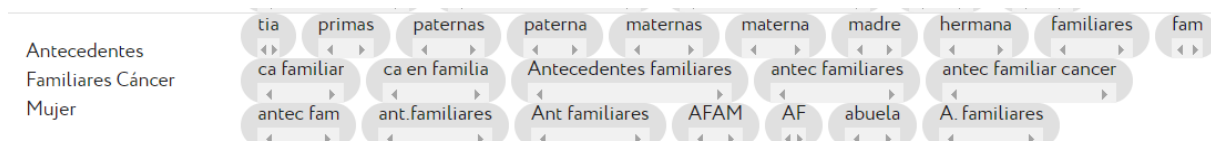


Figura 4.1: Vista de un término en Oncotext

En la figura anterior podemos observar que el término 'Antecedentes Familiares Cáncer Mujer' tiene varios sub-términos asociados a ella, el uso de oncotext nos permite buscar estos sub-términos en un texto a procesar de manera sencilla.

Accediendo a la API de Oncotext, se realizó la tarea de búsqueda de sub- términos que nos permitan obtener información relevante para la detección de las banderas rojas, por ejemplo, una salida para una búsqueda asociada al término anteriormente descrito sería la siguiente

```
Información encontrada:

Término: Antecedentes Familiares Cáncer Mujer

Ocurrencias:
A. familiares
hermana
familiares
```

Figura 4.2: Resultado de búsqueda

Esto quiere decir que para el texto de entrada, se han encontrado los sub-términos 'A. familiares', 'hermana' y 'familiares' asociados al Término 'Antecedentes Familiares Cáncer Mujer'

Si bien Oncotext es una herramienta funcional y operativa, su uso se encuentra limitado al conjunto de términos previamente definidos, sin la capacidad de realizar búsquedas más allá de los contenidos que posee en su base de datos.

4.1.3. Estandarización de datos no estructurados

En el contexto de análisis la calidad y uniformidad de los datos desempeñan un papel crítico en alcance en los modelos de procesamiento natural de lenguaje (NLP). En el contexto médico los diagnósticos clínicos, en particular los relacionados con el cáncer, presentan una variabilidad significativa en el formato o en el estilo de documentación. En el caso de los datos brindados por FALP, existe considerable inconsistencia su formato de biopsias y evoluciones, presentado símbolos no deseados, como abreviaciones de termino médicos. Por lo que es necesario para abordar esta variación, es necesario garantizar uniformidad en los datos en el entrenamiento del modelo. Para realizar esto se hicieron dos módulos para la estandarización de los datos.

El modulo de *Limpieza de ruido* consta del preprocesamiento de los textos dados por las evoluciones y biopsias. En este se identificara y eliminara los símbolos, caracteres especiales y elementos no deseados. La limpieza de ruido es esencial para que los datos sean consistentes para el entrenamiento del modelo NLP. Para esto en "Python" se hace uso de las siguientes librerías:

- Text-unidecode: Esta librería de Python es una herramienta que permite la transliteración de texto Unicode a texto ASCII. La transliteración es el proceso de convertir caracteres de un sistema de escritura en otro. En este caso, text-unidecode toma texto en Unicode que puede contener caracteres especiales, acentos y otros caracteres no ASCII y los convierte en su equivalente en texto ASCII, que generalmente no contiene caracteres especiales ni acentos. Esta representación es útil cuando se trabaja con texto con caracteres especiales, como es el caso de las evoluciones y

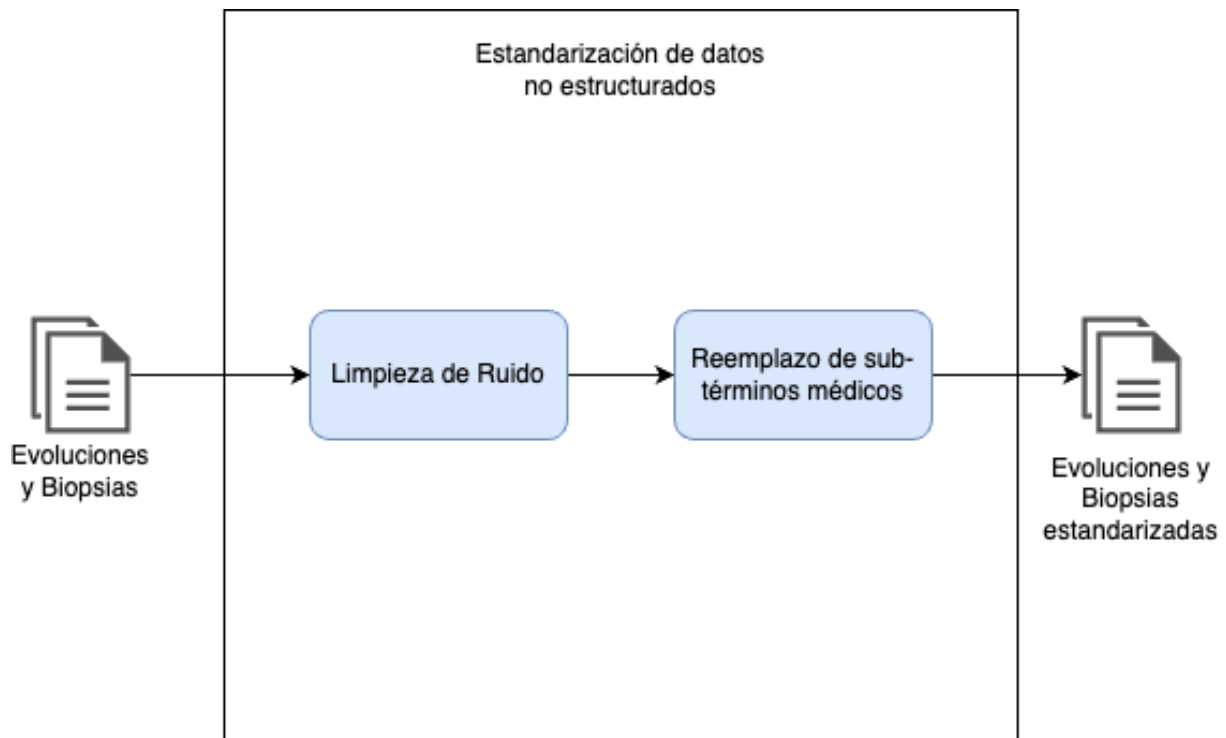


Figura 4.3: Flujo de estandarización

biopsias, es importante eliminar estos para dar uniformidad al formato de los textos para el entrenamiento.

- **Re:** Este es un modulo de expresiones regulares. Esta permite trabajar con patrones de texto, realizando búsqueda, extracciones y manipulaciones avanzadas de cadenas de texto. Estas son utilizadas para eliminar los símbolos no deseados, las cadenas de textos pasan por un filtro en donde solo se admiten letras, números y símbolos deseados (símbolos que son útiles para las fechas, porcentajes, etc), resultando un texto estandarizado.

El módulo Reemplazo de sub-términos médicos hace del uso de Oncotext. Como se mencionó anteriormente, Oncotext es una herramienta que permite obtener sub-términos médicos asociados a un término médico, por lo tanto, permitió reemplazar los sub-términos por su término asociado con el fin de estandarizar aún más los datos, dando más legibilidad a estos. Dado que las evoluciones y biopsias utilizan abreviaciones (sub-términos) estas son reemplazadas por su término como tal. Oncotext brinda una cantidad limitada de sub-términos útiles a reemplazar, pero que son significativos para la estandarización.

Dentro de las restricciones identificadas en el uso de Oncotext, se destaca particularmente el límite en la cantidad de sub-términos que pueden ser reemplazados, lo cual presenta un desafío significativo para el proyecto. Es importante mencionar que Oncotext se presenta como una API de procesamiento de lenguaje natural que se consume a través de solicitudes HTTP. Esto plantea una limitación crítica en el contexto de la aplicación, ya que se debe abordar una gran cantidad de textos médicos no estructurados. El problema

radica en que se debe procesar una gran cantidad de datos (del orden de los miles) en un intervalo de tiempo muy reducido. Esto significa que los datos se deben enviar prácticamente de manera simultánea, uno tras otro. Dado el volumen y la velocidad con la que se realizan las solicitudes a la API de Oncotext, este enfoque no resulta eficiente ni sostenible desde el punto de vista técnico, lo que ha motivado la búsqueda de una solución alternativa.

En respuesta a esta restricción técnica, se optó por solicitar el diccionario base de sub-términos ofrecido por Oncotext. Este diccionario se presenta en un archivo con formato csv y consta de tres columnas. La primera columna contiene el término original, la segunda columna corresponde a la forma escrita de las palabras que deben ser traducidas de acuerdo con el sub-término y, por último, la tercera columna proporciona la forma escrita limpia, que representa la traducción ideal del sub-término.

Es relevante señalar que, debido a la calidad variable de las traducciones en la segunda columna (Forma escritura), se decidió escoger en la tercera columna (Forma escritura limpia), que nos brinda una escritura pulcra del termino, sin errores ortográficos ni simbología no deseada. Este enfoque nos permitirá llevar a cabo un análisis exhaustivo de cada sub-término que sea pertinente en el contexto de nuestra aplicación.

Cuadro 4.5: Muestra de información del diccionario primario de Oncotext

#	Término	Forma escritura	Forma escritura limpia
1	APE	APE	ape
2	APE	PSA	psa
3	APE	PSA Total	psa total
4	APE	PSA Libre	psa libre
5	APE	Prostático Específico	prostático específico
6	Tabaquismo	Tabaquismo	tabaquismo
7	Tabaquismo	Tabaco	tabaco

Para optimizar la calidad de los datos textuales destinados al entrenamiento, se procedió a una selección detallada de los sub-términos que requerían ser reemplazados. La atención se dirigió específicamente a los términos indicadores de cáncer de colon y cáncer de mama. Esta estrategia de selección garantiza la relevancia y precisión de la terminología médica utilizada en el proceso de sustitución.

Complementariamente, se efectuó una revisión exhaustiva de las traducciones de algunos sub-términos. Se determinó la exclusión de aquellas traducciones que no se ajustaban al contexto o que podrían impactar negativamente la claridad del texto tras su aplicación. En la siguiente figura (Figura 5.4), muestra el resultante del modulo estandarización de datos para el entrenamiento del modelo NLP.

Evolución original v/s Evolución estandarizado	
evolucion	evolucion estandarizada
Ricardo,59\nParral\nANTECEDENTES:\nMédicos: HTA, DM2NIR, Cardiopatía coronaria(SCA + PTCA 2018)\nQx: Fx Tobillo derecho\nFármacos: LST,MTF, AMLODIPINO, ATV, AAS, CARVEDILOL, FLUOXETINA\nAlergias:-\nTabaco:-\nOH: Ocasional\nFamiliares: Madre Ca de Pulmón\n\nHISTORIA:\nPaciente consulta por rectorragia\nColono: Tu Colon sigmoides\nBiop: (15.4.22) ADC\nTC TAP: (4.5.22) Nódulo 4 mm LSD\nSe decide Qx RAB el 15.6.22\nBiopsia: ADC infiltrante hasta la serosa, PLV:+ PLN:+ TB: Moderado, Ganglios: 5/42 ,\nMSI:-\nSe discute en comité se decide QT adyuvante FOLFOX por 6 meses\nEx lab: (11.7.22) Hb:11,6 Gb: 8780 Plaq: 333.000 Crea:0.7 CEA:0,59 \n\nActualmente en BCG ECOG:1\n\nAl ex fisico:\nPeso:120\nTalla:181\nBien perfundido, hidratado\nAdenop:-\nMP:+ SRA, RR2TSS\nABD: BDI, RHA:+, herida operatoria con abundante secreción serosa y zona de pequeña dehiscencia\nEEII: edema:-, TVP:- \n\nPLAN:\nPaciente 59 años con Ca de Colon op PT3N2MO\nQT adyuvante FOLFOX por 12 ciclos\nAún no iniciar QT en contexto de herida op en curaciones\nDejo derivación para reservorio y pla n de tto\nControl presencial para evaluar herida en 2 semanas	ricardo,59 parral antecedentes: medicos: hta, dm2nir, cardiopatia coronaria sca + ptca 2018 qx: fx tobillo derecho farmacos: lst,mtf, amlodipino, atv, aas, carvedilol, fluoxetina alergias:- tabaco:- oh: ocasional familiares: madre cancer de pulmon historia: paciente consulta por disminucion hemoglobina colono: tu colon sigmoides biop: 15.4.22 adc tc tap: 4.5.22 nodule sospechosos ecografia 4 mm lsd se decide cirugia prostata rab el 15.6.22 biopsia: adc infiltrante hasta la serosa, plv:+ pln:+ tb: moderado, ganglios: 5/42 , msi:- se discute en comite se decide quimioterapia adyuvante folfox por 6 meses ex lab: 11.7.22 hb:11,6 gb: 8780 plaq: 333.000 crea:0.7 cea:0,59 actualmente en bcg ecog:1 al ex fisico: peso:120 talla:181 bien perfundido, hidratado adenop:- mp:+ sra, rr2tss abd: bdi, rha:+, herida operatoria con abundante secrecion serosa y zona de pequena dehiscencia eeii: edema:-, tvp:- plan: paciente 59 anos con cancer de colon op pt3n2mo quimioterapia adyuvante folfox por 12 ciclos aun no iniciar quimioterapia en contexto de herida op en curaciones dejo derivacion para reservorio y pla n de tto control presencial para evaluar herida en 2 semanas

Figura 4.4: Flujo de estandarización

4.1.4. Modulo análisis de los datos estructurados

La solución a la problemática se centra predominantemente en el análisis de datos no estructurados. Se propone un análisis inicial de variables específicas para comprender el funcionamiento del modelo. La información demográfica del paciente, como sexo y fecha de nacimiento, permite establecer un filtro preliminar para identificar casos que requieran derivación a AGO antes de la intervención del modelo,

Para esto, existen dos banderas rojas para el cáncer de mamá que se pueden corroborar analizando los datos estructurados. La primera bandera roja corresponde al diagnóstico de cáncer de mama en personas menores a 50 años y por último la otra bandera roja es la existencia del cáncer de mama en pacientes de sexo masculino.

Tras la realización de un análisis comparativo exhaustivo, se determinó que, de una población total de 5.767 pacientes, 1.462 individuos cumplen con al menos uno de los criterios definidos como "bandera roja". En consecuencia, estos pacientes satisfacen las condiciones necesarias para su derivación al servicio de AGO.

4.1.5. Módulo preparación de datos

Para poder realizar el entrenamiento del modelo, se debe realizar una preparación de los datos.

Para la preparación se procesaron los datos ya consolidados con Python, donde se creó un dataframe con todas las evoluciones y biopsias de los pacientes, con un total de 52.257 datos (46.132 evoluciones y 6.125 biopsias)

Posteriormente, con el id de cada paciente para cada registro, se realiza una asociación binaria con respecto a que pacientes deben ser derivados a AGO con los datos obtenidos en el Análisis de datos estructurados. Si una evolución y/o biopsia de un paciente cumple al menos con dos banderas rojas, o este haya sido derivado a AGO anteriormente, se fija con un valor 1 en el resultado, en otro caso se asigna con 0.

Luego de realizar esta asociación binaria, se obtiene el siguiente dataframe para el entrenamiento del modelo:

	texto	resultado
0	exámenes de rutina. prima paterna cancer ovari...	1
1	telemed viviana, 54 antecedentes: medicos: epo...	1
2	se revisa hemograma ok plan: a 2do ciclo qt	1
3	hombre de 46 años antecedentes: medicos: no qx...	1
4	telemedicina. eco falp: nodule sospechosos eco...	1
...	...	...
15372	mt+gc+reconstruccion retropectoral con implant...	0
8361	ca colon sigmoides operado hace 3 años y medio...	0
16587	ca mama triple negativo iv qmt paliativa carbo...	0
34096	mp+gc izq 27/10/22 mama con leve equimosis res...	0
19906	telemedicina cancer colon pt4an0m1c c12 folfox...	0

Figura 4.5: Dataframe para entrenamiento del modelo

4.1.6. Modulo de entrenamiento del modelo

BioBERT es una variante de BERT (Bidirectional Encoder Representations from Transformers) que ha sido preentrenada específicamente en datos biomédicos y de salud. En comparación con modelos de lenguaje generales, como BERT, GPT, etc., BioBERT a menudo supera en rendimiento en tareas biomédicas específicas debido a su enfoque en el dominio médico.

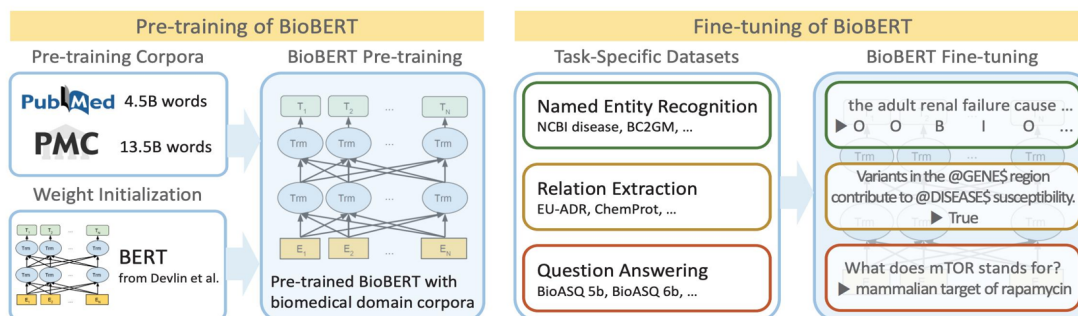


Figura 4.6: Figura BioBERT

Antes de proceder con el uso de los datos de entrenamiento (datos consolidados), se realizan una serie de pasos antes de pasar al entrenamiento del modelo, los cuales se detallan a continuación:

- Reducción de datos para balanceo de estos acorde a su clasificación de resultado (1 o 0) en proporción 50/50.
- Extracción de datos para testeo con proporción adecuada de 1s y 0s
- División de los datos en 80 % Train, 20 % Test

- Ajuste de hiperparámetros para el entrenamiento
- Elaboración de función de detención de entrenamiento

También se realizó un ajuste en los hiperparámetros para el entrenamiento del modelo. Estos se definen y se asignan los siguientes valores:

Cuadro 4.6: Hiperparámetros usados y sus definiciones.

Hiperparámetro	Valor usado	Nombre	Definición
batch_size_train	16	Tamaño de lote para entrenamiento	Número de muestras utilizadas para actualizar los parámetros en cada iteración durante el entrenamiento.
batch_size_test	32	Tamaño de lote para pruebas	Número de muestras utilizadas en cada iteración para evaluar el modelo en los datos de prueba.
initial_lr	1×10^{-5}	Tasa de aprendizaje inicial	Valor inicial de la tasa de aprendizaje que controla el ajuste de los pesos del modelo.
weight_decay	1×10^{-4}	Decaimiento del peso	Regularización aplicada a los pesos del modelo para evitar el sobreajuste, penalizando grandes valores en los pesos.
epoch	10	Épocas	Número de veces que todo el conjunto de datos es procesado durante el entrenamiento.

```

Epoch 1/10 - Train Loss: 0.6302 - Validation Loss: 0.5913
Epoch 2/10 - Train Loss: 0.5383 - Validation Loss: 0.5405
Epoch 3/10 - Train Loss: 0.4795 - Validation Loss: 0.5284
Epoch 4/10 - Train Loss: 0.4073 - Validation Loss: 0.5098
Epoch 5/10 - Train Loss: 0.3439 - Validation Loss: 0.5212
EarlyStopping counter: 1 out of 3
Epoch 6/10 - Train Loss: 0.2913 - Validation Loss: 0.5369
EarlyStopping counter: 2 out of 3
Epoch 7/10 - Train Loss: 0.2564 - Validation Loss: 0.5625
EarlyStopping counter: 3 out of 3
Early stopping

```

Figura 4.7: Entrenamiento del modelo con BioBert

Por otro lado, se crearon funciones para la optimización de los mencionados hiperparámetros. Esto ajusta la función de pérdida para calcular el `weight_decay`, como también el optimizador para ajustar los pesos del modelo durante el entrenamiento.

```
initial_lr = 1e-5
weight_decay = 1e-4

optimizer = AdamW(model.parameters(), lr = initial_lr,
                   weight_decay = weight_decay)

criterion = nn.CrossEntropyLoss()
```

Listing 4.1: Optimizador y Función de Pérdida

Por otro lado, se agrega la ultima función. Esta permite detener el entrenamiento en caso de que se este generando un sobre-ajuste. También, se incorpora otra función para el ajuste de la tasa de aprendizaje de este, iterando en los bucles de entrenamiento

```
early_stopping = EarlyStopping(patience = 3, verbose = True)

step_size = 3
gamma = 0.1
scheduler = StepLR(optimizer, step_size = step_size, gamma =
gamma)
```

Listing 4.2: Detención de entrenamiento y ajuste del Learning-State

5. Resultados y Discusión

En este capítulo se los resultados obtenidos en las iteraciones del modelo, en cada una se irán modificando los parámetros necesarios para el entrenamiento, observando su comportamiento con la ausencia de todos o algunos de ellos, se mostraran los resultados del Validation Loss y Training Loss. Por otro lado, también se mostraran el Accuracy, Precision, Recall y F1 Score, estos son fundamentales para la evaluación del modelo. Además, se realiza una comparación con chatbots potenciados con IA para poder evaluar el funcionamiento de nuestro modelo con respecto a estos, tales como Chat GPT y Gemini de Google.

5.1. Entrenamiento y Resultados del Modelo

5.1.1. 1° Entrenamiento

El primer entrenamiento realizado del modelo, trata de un simple test del modelo BioBERT con los datos ya estandarizados de la sección 4.1.5.

La matriz de confusión nos permite evaluar el rendimiento de un modelo de clasificación. El número de predicciones correctas e incorrectas se pueden apreciar en la siguiente imagen:

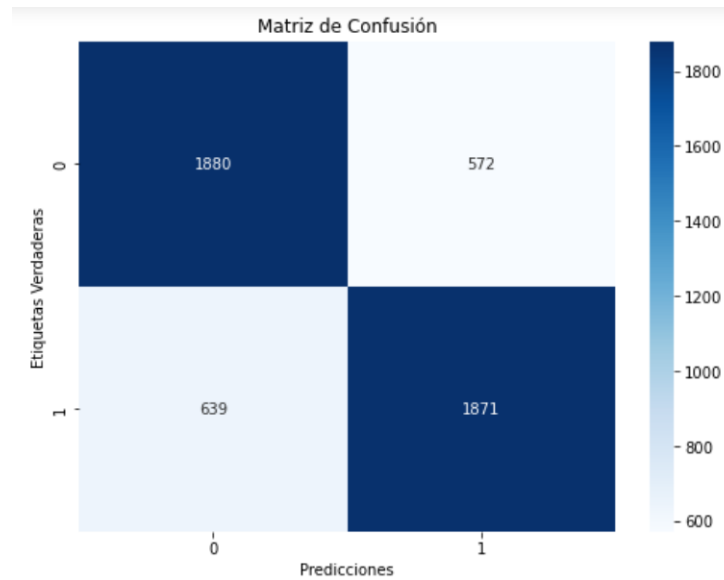


Figura 5.1: Matriz de confusión del entrenamiento

Además de la matriz de confusión, existen otras métricas relevantes para evaluar el rendimiento del modelo. Cada una de estas métricas ofrece diferentes perspectivas sobre la calidad de las predicciones realizadas:

- **Precisión (Precision):** Mide la proporción de verdaderos positivos entre todas las instancias predichas como positivas. Es decir, indica cuántas de las predicciones positivas fueron correctas. Se calcula como:

$$\text{Precisión} = \frac{VP}{VP + FP}$$

La precisión es útil cuando se desea minimizar los falsos positivos, es decir, cuando es más importante no etiquetar incorrectamente una instancia como positiva.

- **Exhaustividad (Recall):** Mide la proporción de verdaderos positivos entre todas las instancias que realmente son positivas. Indica qué tan bien el modelo es capaz de identificar todas las instancias positivas. Se calcula como:

$$\text{Exhaustividad} = \frac{VP}{VP + FN}$$

Esta métrica es relevante cuando es crucial capturar todas las instancias positivas, aunque se permita un mayor número de falsos positivos.

- **Puntuación F1 (F1 Score):** Es la media armónica entre precisión y exhaustividad. Se utiliza cuando se busca un equilibrio entre ambas métricas, ya que otorga el mismo peso a la precisión y a la exhaustividad. Se calcula como:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precisión} \cdot \text{Exhaustividad}}{\text{Precisión} + \text{Exhaustividad}}$$

Es particularmente útil cuando se tiene un conjunto de datos con clases desbalanceadas, donde uno de los dos tipos de errores (falsos positivos o falsos negativos) es más perjudicial que el otro.

- **Exactitud (Accuracy):** Mide la proporción de predicciones correctas (tanto positivas como negativas) sobre el total de instancias. Se calcula como:

$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN}$$

La exactitud es útil en conjuntos de datos equilibrados, pero puede no ser representativa en situaciones donde hay un gran desbalance entre clases.

Estas métricas complementan el análisis de la matriz de confusión, proporcionando una evaluación más completa sobre el rendimiento del modelo en diversas situaciones. Dependiendo del contexto, puede ser más importante priorizar alguna de estas métricas sobre las demás para obtener una interpretación más adecuada de los resultados del modelo.

Luego, los resultados del entrenamiento del modelo NLP, para la clasificación de evoluciones y biopsias de pacientes para la derivación a AGO, se resume en las siguientes métricas:

- Accuracy Score en el conjunto de prueba: 75.59 %
- Precision Score en el conjunto de prueba: 76.59 %
- Recall Score en el conjunto de prueba: 74.54 %
- F1 Score en el conjunto de prueba: 75.55 %

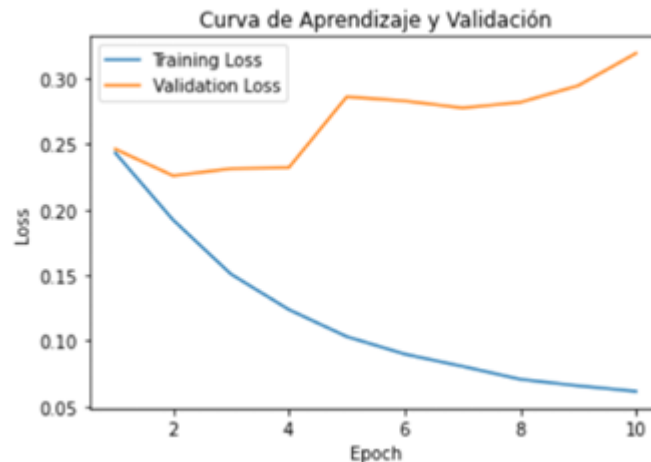


Figura 5.2: Curva de aprendizaje 1° entrenamiento

Dado que los posibles valores de salidas son de tipo booleano (0 o 1, true o false), un modelo de tipo aleatorio entregaría alrededor de un 50 % al calcular las métricas,

esto es una base para evaluar el modelo desarrollado. Este porcentaje es importante para establecer un desempeño mínimo deseado y sirve para determinar si el modelo tiene un rendimiento significativo o este esta funcionando aleatoriamente.

Por lo visto en este caso, los valores obtenidos son superiores al porcentaje aleatorio, mejoraron su valor alrededor de un 25 %, este aumento indica un avance con respecto al comportamiento base. Este aumento nos indica que el modelo tiene la capacidad sustancial de aprender patrones significativos a través de los datos y tomar decisiones con una precisión mucho mayor que con aleatoriedad simple.

Por otro lado, considerando que esta seria la primera versión del modelo desarrollado, la posibilidad de mejora es posible, ya que coloca al modelo como una potencial solución efectiva. Si bien, el resultado dado debe interpretarse como base para futura iteraciones y/o optimizaciones. Con ajustes en los datos, un mayor volumen de estos, mejoras en los valores de los hiperparámetros, se podrían alcanzara futuro niveles de desempeños aun mayores.

5.1.2. 2° Entrenamiento

El segundo entrenamiento, se realiza ciertos cambios a comparación del primero. En este entrenamiento se agrega a comparación al primer entrenamiento se hace uso del dataset del modulo de Preprocesamiento de Datos. Ademas se hace el uso del Modulo de Oncotext para obtener una estandarización de los textos.

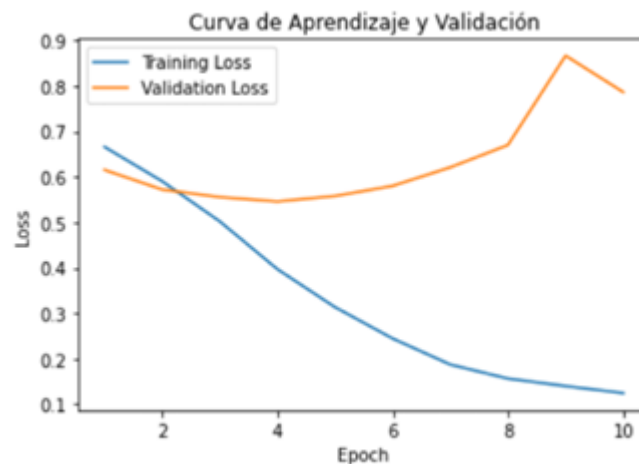


Figura 5.3: Curva de aprendizaje 2° entrenamiento

Por lo visto en la gráfica adjuntada recientemente, presenta una leve mejora en el entrenamiento, se puede observar en el Validation Loss, como en las siguientes métricas de evaluación:

- Accuracy Score en el conjunto de prueba: 79.01 %
- Precision Score en el conjunto de prueba: 79.71 %

- Recall Score en el conjunto de prueba: 78.16 %
- F1 Score en el conjunto de prueba: 78.93 %

5.1.3. 3° Entrenamiento

Para el tercer entrenamiento se agrega hiperparámetros para el entrenamiento, este permite ajustar como se realiza el entrenamiento. Estos valores de los hiperparámetros son los descritos en el Cuadro 4.6.

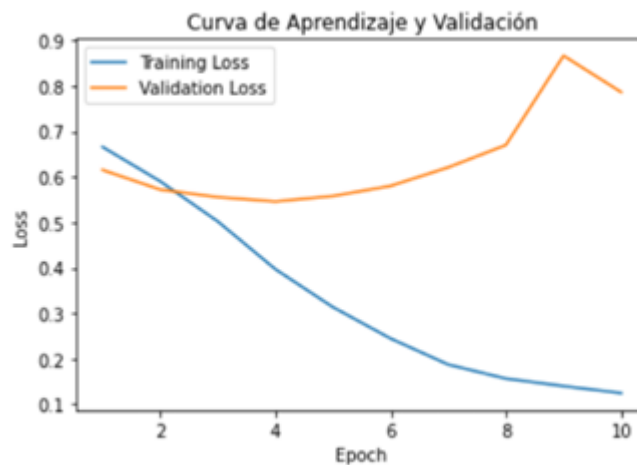


Figura 5.4: Curva de aprendizaje 3° entrenamiento

Con la aplicación de los hiperparametros en el entrenamiento, se puede observar una pequeña mejora en los *Epochs*. No obstante, al observar al decimo *Epochs* se puede apreciar todavía un sobreajuste. Por otro lado, los resultados de las métricas para evaluar el entrenamiento son:

- Accuracy Score en el conjunto de prueba: 76.91 %
- Precision Score en el conjunto de prueba: 75.32 %
- Recall Score en el conjunto de prueba: 80.23 %
- F1 Score en el conjunto de prueba: 77.70 %

5.1.4. 4° Entrenamiento

Para el cuarto y ultimo entrenamiento realizado, se agrega el Optimizador, Función de pérdida de la figura 3.8 y la función de detención de entrenamiento y ajuste de learning-rate, de la figura 3.9. Esto permite obtener un Validation Loss mas adecuado para la solución

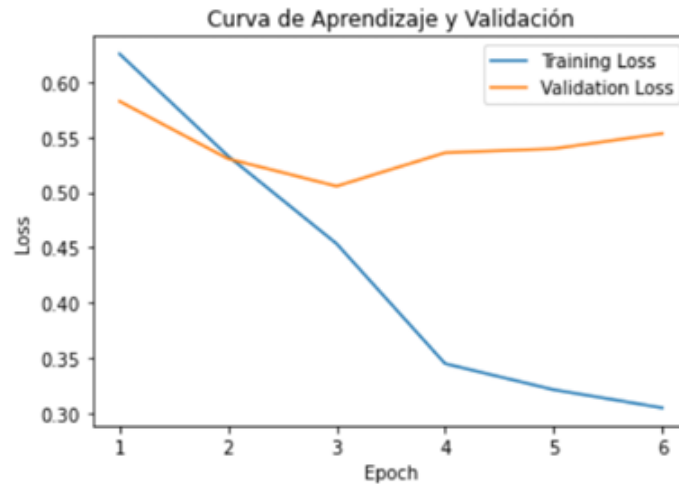


Figura 5.5: Curva de aprendizaje 4° entrenamiento

Se puede apreciar que el resultante en la gráfica adjunta, por lo visto mantiene un orden de magnitud adecuado de las métricas de rendimiento elegidas. Luego, los resultados para las métricas de evaluación son:

- Accuracy Score en el conjunto de prueba: 75.70 %
- Precision Score en el conjunto de prueba: 74.33 %
- Recall Score en el conjunto de prueba: 79.36 %
- F1 Score en el conjunto de prueba: 76.76 %

Posteriormente, este ultimo entrenamiento realizado es el escogido para la implementación del modelo de Banderas Rojas. Se adjunta la matriz de confusión que permite analizar le comportamiento del modelo con respecto al primer entrenamiento realizado.

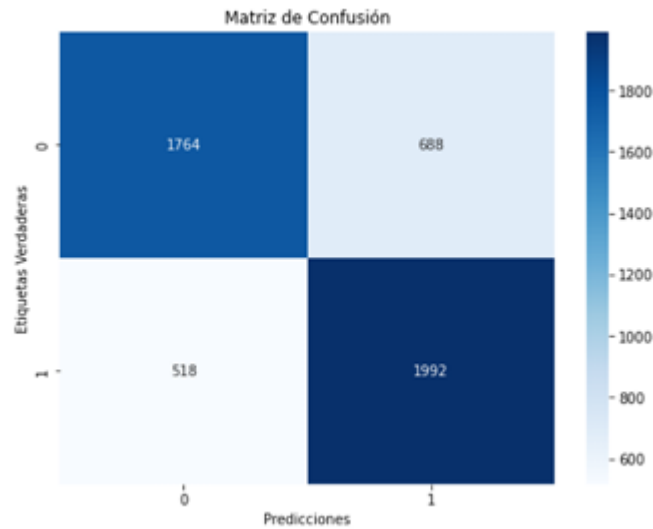


Figura 5.6: Matriz de confusión del 4° entrenamiento

De acuerdo a lo observado en la matriz de confusión, se pueden desprender los siguientes valor:

- Verdaderos Negativos (TN): 1.764
- Falsos Negativos (FN): 688
- Verdaderos Positivos (TP): 518
- Falsos Positivos (FP): 1.992

5.2. Comparación del modelo entrenado v/s Sistemas de Inteligencia Artificial Conversacional

Para poder realizar una comparativa y mejorar la magnitud de los resultados obtenido con el modelo realizado, se realizara una comparación con Sistemas de Inteligencia Artificial Conversacional para hacer un análisis con los mismos datos de test que utilizo el modelo entrenado con BioBert, pero con los textos en duro, es decir los reportes iniciales entregado por FALP, por lo tanto sin haber pasado por los módulos de preprocesamiento y preparación. Para esto se utilizara un el mismo dataset de testing de las evoluciones medicas del modelo desarrollado en donde compararemos el funcionamiento del modelo desarrollado v/s a la predicción realizada por los Sistemas de IA Conversacional.

5.2.1. Modelo Desarrollado v/s ChatGPT GPT-5.3

Para empezar, primero se le da contexto a ChatGPT sobre los reportes médicos y lo que se busca reconocer con estos:

Hola, necesito saber si un paciente debe ser derivado a Asesoramiento Genetico Oncologico en base a la evolucion clinica de este. Por lo tanto, te entregare textos que corresponden a la evolucion medica. Necesito que respondas con 1 si el paciente debe ser derivado a asesoramiento genetico oncologico, con sus banderas rojas y en caso contrario de que no deba ser derivado arroja 0. Esto es para hacer comparativa con un modelo de NLP entrenado

¡Perfecto! Entiendo lo que necesitas. 😊

Para confirmar, el flujo sería así:

1. Me compartes uno o varios textos con la **evolución clínica** del paciente.
2. Yo analizo la información y respondo con:
 - "1" → si el paciente **debe ser derivado a Asesoramiento Genético Oncológico**, indicando las **banderas rojas** detectadas.
 - "0" → si **no** corresponde derivarlo, sin banderas rojas.

Cuando quieras, envíame el primer texto de evolución clínica y lo evaluamos.

Figura 5.7: Prompt para contextualizar a ChatGPT

Luego, le entregamos un ejemplo de evolución medica para verificar si ChatGPT discernie correctamente si un paciente debe ser derivado a AGO, esta prueba tiene como valor real que el paciente debe ser derivado.

87 años. en muy buenas condiciones. comorbilidades cr. hipotiroidismo. fa en taco con xarelto y amiodarona. tabaco: no a. familiares: hija cancer de mama a los 51 años. 2 hijos fallecidos por cancer de colon 43 años y otro de pancreas 52, misma que tuvo cancer mama. refiere autopalpacion de zona nodular mama derecha desde hace 4 meses. mamo: nodulos sospechosos ecografia mama derecha periareolar birads 5. eco: nodulos sospechosos ecografia cuadrante mama derecha de 28 mm. birads 5. bp. core: histologia g3. re-, rp-, her2 3+. ki-67 70%. tac cuello tap: negativo. al ex: tumor en cuadrante periareolar mama derecha de unos 4 x 3 cm indurado, no adherido a planos profundos pero muy cercano a la piel. examen fisico axilar impresiona libre. plan: ago fish. discutir en comite. quimioterapia neo si fish + versus mt + blc

Figura 5.8: Prompt de prueba para ChatGPT

Procesé el archivo y generé una columna con la clasificación solicitada:

- **1** = derivar a Asesoramiento Genético Oncológico
- **0** = no derivar

Te dejé un Excel con dos columnas:

- **texto** (la evolución original)
- **resultado** (0 o 1)

Puedes descargarlo aquí:

Figura 5.9: Respuesta de ChatGPT

Una vez realizado el paso anterior, ChatGPT entendió lo que buscamos con su análisis, por lo tanto se le entrega un archivo que contiene los reportes médicos para que realice el mismo análisis que el paso anterior.

	Bp. CORE: CDI G3 Trip
1	ca de mama T2N0Mx, triple negativo. CDI G3 Triple negativo, Ki 67 60%
2	Paciente tratada por c Función OK y sin comp
3	57 años Asintomatica Eco mamaria enero 20

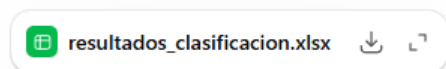
Procesé el archivo y generé una columna con la clasificación solicitada:

- **1** = derivar a Asesoramiento Genético Oncológico
- **0** = no derivar

Te dejé un Excel con dos columnas:

- **texto** (la evolución original)
- **resultado** (0 o 1)

Puedes descargarlo aquí:



Si quieres, también puedo:

- generar **solo la columna de 0/1** para pegarla directo en tu Excel,
- o ajustar los **criterios clínicos de derivación** para que el modelo quede más preciso para AGO (por ejemplo: **triple negativo <60, antecedentes familiares, ovario, colon precoz, múltiples primarios, etc.**).

Figura 5.10: Prompt con el archivo a analizar

Una vez entregado los resultados por ChatGPT obtenemos las siguientes métricas:

- Accuracy Score en el conjunto de prueba: 47.30 %
- Precision Score en el conjunto de prueba: 47.0 %
- Recall Score en el conjunto de prueba: 40.90 %
- F1 Score en el conjunto de prueba: 43.70 %

Posteriormente, al obtener los resultados dados por ChatGPT procedemos a crear su matriz de confusión:

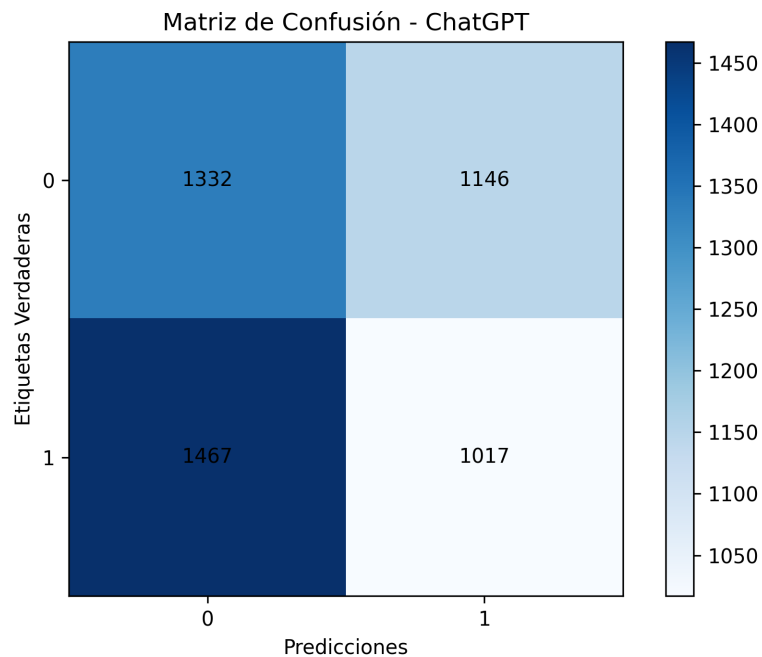


Figura 5.11: Matriz de confusión ChatGPT

De esta obtenemos:

- Verdaderos Negativos (TN): 1.332
- Falsos Negativos (FN): 1.146
- Verdaderos Positivos (TP): 1.467
- Falsos Positivos (FP): 1.017

Resumiendo ambos resultados, obtenemos la siguiente tabla a modo de comparación:

Cuadro 5.1: Comparación de métricas entre modelos

	Modelo Entrenado	Modelo ChatGPT
Accuracy Score	75.70 %	47.30 %
Precision Score	74.33 %	47.0 %
Recall Score	79.36 %	40.90 %
F1-Score	76.76 %	43.70 %

En base a los resultados obtenidos a partir de las métricas ChatGPT vs Modelo entrenado, es posible establecer diversas conclusiones respecto al desempeño de estos:

- **Accuracy Score:** El modelo entrenado con BioBERT presenta un desempeño significativamente superior al modelo basado en ChatGPT, evidenciando una mayor capacidad en clasificar correctamente los casos en términos generales. En particular, alcanza un 75,70 % de aciertos, lo que lo posiciona como una alternativa más confiable.
- **Precision Score:** Se observa una diferencia aproximada al 37 % entre ambos modelos, lo que indica que BioBERT logra identificar con mayor exactitud los casos positivos reales. Esto refleja una menor proporción de falsos positivos en comparación con la predicción realizada por ChatGPT.
- **Recall Score:** Esta métrica resulta especialmente relevante en el ámbito clínico, ya que mide la capacidad del modelo para detectar correctamente los casos que requieren derivación oncológica. En este sentido, el modelo basado en ChatGPT presenta un rendimiento considerablemente inferior, con un 41.40 %, lo que implica una detección deficiente de pacientes que sí requieren atención. Este comportamiento representa una limitación crítica, dado el impacto que podría tener en los pacientes que si deben ser derivados.
- **F1 Score:** Los resultados muestran que BioBERT mantiene un equilibrio aceptado entre precisión y sensibilidad, lo que se traduce en una correcta identificación de casos positivos. En contraste, el modelo ChatGPT se puede evidenciar un desempeño inferior en esta métrica.

5.2.2. Modelo Desarrollado v/s Gemini 3

Primero se le entrega contexto a Gemini sobre los reportes médicos, el mismo reporte que se le entrego a ChatGPT:

Entendido. Estoy listo para colaborar con la validación de tu modelo de NLP. Analizaré los textos de evolución médica que me proporciones para identificar criterios de sospecha de cáncer hereditario.

Para cada caso, responderé con:

- **1:** Si cumple con criterios para derivación a **Asesoramiento Genético Oncológico (AGO)**, detallando las **banderas rojas** detectadas (ej. edad temprana al diagnóstico, múltiples primarios, antecedentes familiares específicos, o histologías sugerentes).
- **0:** Si la evolución clínica no presenta indicadores actuales para derivación.

Quedo a la espera del primer texto de evolución médica para procesarlo.

Figura 5.12: Respuesta del primer prompt de contextualización a Gemini

Luego, le entregamos un ejemplo de evolución medica para verificar si Gemini es capaz de discernir correctamente si un paciente debe ser derivado a AGO, esta prueba tiene como valor real que el paciente debe ser derivado, se utiliza la misma evolución que se le entrego a ChatGPT.

1

Banderas Rojas Detectadas:

- **Historia Familiar de Cáncer de Inicio Temprano:** El paciente reporta un hijo fallecido por cáncer de colon a los **43 años**.
- **Agrupación Familiar de Cánceres Relacionados:** Presencia de un hijo con cáncer de páncreas (52 años) que también tuvo cáncer de mama, junto con una hija con cáncer de mama a los 51 años. Esta combinación (mama, páncreas, colon) es altamente sugerente de síndromes de predisposición hereditaria, como el **Síndrome de Cáncer de Mama y Ovario Hereditario (HBOC)** o el **Síndrome de Lynch**.
- **Diagnóstico Actual de Cáncer de Mama con Histología Específica:** Hallazgo de nódulo BIRADS 5 con inmunohistoquímica HER2 3+ y Ki-67 del 70%, que en el contexto de la carga familiar descrita, refuerza la necesidad de estudio genético a pesar de la edad del paciente (87 años).

Figura 5.13: Respuesta del prompt de prueba a Gemini

Una vez realizado el paso anterior, Gemini logro entender lo que buscamos con su análisis, por lo tanto se le entrega el archivo de testing en duro para que realice el mismo análisis que el paso anterior. Una vez entregado los resultados por Gemini obtenemos las siguientes métricas:

- Accuracy Score en el conjunto de prueba: 45.99 %
- Precision Score en el conjunto de prueba: 20.64 %
- Recall Score en el conjunto de prueba: 42.95 %
- F1 Score en el conjunto de prueba: 27.88 %

A continuación, al obtener los resultados dados por Gemini procedemos a crear su matriz de confusión:

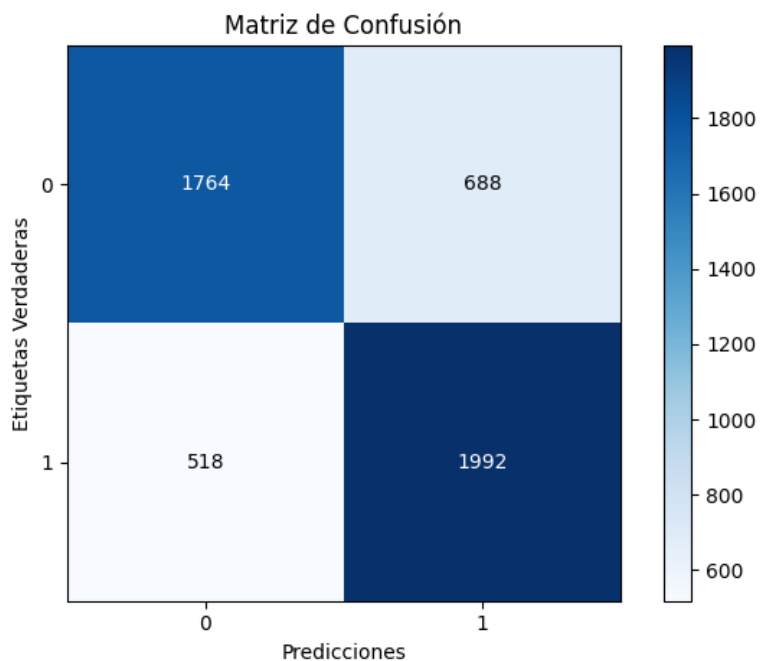


Figura 5.14: Matriz de confusión Gemini

De esta matriz de confusión obtenemos:

- Verdaderos Negativos (TN): 1.764
- Falsos Negativos (FN): 688
- Verdaderos Positivos (TP): 518
- Falsos Positivos (FP): 1.992

Resumiendo ambos resultados, obtenemos la siguiente tabla a modo de comparación:

Cuadro 5.2: Comparación de métricas entre modelos

	Modelo Entrenado	Modelo Gemini
Accuracy Score	75.70 %	45.99 %
Precision Score	74.33 %	20.64 %
Recall Score	79.36 %	42.95 %
F1-Score	76.76 %	27.88 %

- **Accuracy Score:** El modelo entrenado con BioBERT presenta un desempeño significativamente superior al modelo basado en Gemini, se puede apreciar una mayor capacidad en clasificar correctamente los casos en términos generales. En particular, alcanza un 75,70 % de aciertos, lo que lo posiciona como una opción fiable.
- **Precision Score:** Se observa una diferencia aproximada al 50 % entre ambos modelos, lo que indica que BioBERT logra identificar con mayor exactitud los casos positivos reales. Esto se concluye que es fiable para el contexto que estamos abordando, ya que garantiza que una gran parte de pacientes que si son candidatos sean derivados.
- **Recall Score:** Esta métrica resulta relevante en el ámbito clínico, ya que mide la capacidad del modelo para detectar correctamente los casos que requieren derivado a AGO. Por consiguiente, el modelo basado en Gemini presenta un rendimiento considerablemente deficiente, con un 42.95 %, lo que implica una detección inferior de pacientes que sí requieren ser derivados. Este comportamiento representa una limitación importante, dado el impacto que podría tener en los pacientes que si deben ser derivados.
- **F1 Score:** Los resultados muestran que BioBERT mantiene un equilibrio aceptado entre precisión y sensibilidad, lo que se traduce en una correcta identificación de casos positivos.

6. Conclusión y Trabajo a Futuro

6.1. Conclusión

Según lo mostrado y descrito anteriormente, los resultados obtenidos permiten concluir que el trabajo desarrollado, que abarca desde el diseño del módulo de preprocesamiento y estructuración de evoluciones medicas hasta la implementación del análisis de los datos estructurados contribuyó de manera significativa a la construcción de un modelo predictivo robusto. Este modelo no solo presenta un desempeño aceptable, sino que también constituye un aporte relevante para el apoyo en la toma de decisiones clínicas.

El desarrollo de la solución para identificar que pacientes figura una o varias banderas rojas de cáncer de colon o mama ha sido un proceso que ha abarcado múltiples aspectos, desde la estandarización y estructuración de los datos clínicos hasta la implementación de un modelo de procesamiento de lenguaje natural.

En el módulo de preprocesamiento y preparación de datos clínicos, se logró estructurar adecuadamente la información contenida en los archivos .xlsx proporcionados por la Fundación Arturo López Pérez. Una de las principales metas, la cual se logro realizar fue la creación de un archivo consolidado que integra datos relevantes de pacientes, estos son las evoluciones y biopsias.

Por otro lado, la estandarización de datos no estructurados se afronto con una limpieza de ruido y reemplazo de sub-términos médicos ayudados con el diccionario de Oncotext, como también utilizando técnicas de procesamiento que nos brinda Python. Se analizo e identifico la importancia variar los datos clínicos para poder garantizar una calidad de los datos utilizados en el entrenamiento del modelo.

La elección del modelo BioBert como modelo de pre entrenamiento se noto ser acertado, ya que este modelo pre entrenado de datos biomédicos muestra un buen rendimiento en tareas especifica con el área de la salud. EL entrenamiento de este modelo se realizo mediante un conjunto de datos balanceados con un ajustes en los hiperparámetros para obtener resultados esperados.

Los resultados obtenidos por el modelo de NLP ya entrenado, mostraron un desempeño deseable con métricas como Recall, F1 Score y Precisión en el conjunto de testing. La generación de la matriz de confusión nos detalla el funcionamiento del rendimiento del modelo, permitiendo la interpretación de los verdaderos positivo, verdaderos negativos, falsos positivos y falsos negativos.

En conclusión, el desarrollo de esta solución es una buena base para una herramienta en la detección de banderas rojas, por lo que es un indicador prometedor para las personas con riesgo de cáncer de mama y colon, proporcionando un enfoque eficiente ne el proceso

de reportes médicos, ya que este procesa datos estructurados como no estructurados. Con este enfoque se ira mejorando la eficacia de la derivación de pacientes a AGO, generando una detección temprana de un potencial cáncer, todo esto contribuyendo a una atención medica eficaz.

6.2. Trabajos a futuro

Durante el análisis y desarrollo del trabajo propuesto se implemento una solución, la cual utiliza técnicas de Procesamiento de Lenguaje Natural (NLP), para poder identificar los pacientes que deben ser derivado a Asesoramiento Genético Ontológico (AGO). Pero, como es común en las soluciones tecnológicas como en nuestro caso basadas en Inteligencia Artificial y Análisis de Datos, donde existen diversos incisos de mejora que se podrían abordar con futuras investigaciones o tecnología. Estas podrían mejorar la performance del modelo y como también este modelo propuesto puede facilitar la integración en otras problemáticas relacionadas con la salud o semejantes a estas. Dado el análisis y conclusión de la solución dada se propone las siguientes propuestas de mejora:

- **Ampliación del conjunto de datos:** Aunque el conjunto de datos inicial constaba de 52.257 muestras, el proceso de balanceo redujo significativamente esta cantidad a menos de la mitad. Por ello, una línea futura de trabajo consiste en la recolección de un mayor volumen de datos clínicos, lo que permitiría aumentar la capacidad del modelo para generalizar y abordar una mayor diversidad de casos de pacientes.
- **Optimización de hiperparámetros:** Si bien en la solución realizada se emplearon ciertos hiperparámetros definidos para el entrenamiento del modelo, resulta pertinente explorar alternativas adicionales o ajustes más finos de estos hiperparámetros. Esta investigación permitiría realizar una evaluación comparativa que podría conducir a mejoras en el rendimiento del modelo.
- **Evaluación de modelos NLP alternativos:** Aunque el modelo BioBERT fue seleccionado por ser una de las mejores opciones indagadas en el momento de realizar este trabajo, existen en la actualidad otros modelos de procesamiento de lenguaje natural (NLP) que podrían ofrecer un rendimiento superior. Investigar y comparar dichos modelos constituye una interesante línea de investigación futura. decisiones médicas.

Bibliografía

- [1] World Health Organization, “Colorectal cancer,” disponible en: <https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer> (Marzo, 2024).
- [2] World Health Organization, “Breast cancer,” disponible en: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer> (Marzo, 2024).
- [3] Departamento de Estadísticas e Información de Salud (DEIS), “Defunciones por cáncer en Chile,” disponible en: <https://deis.minsal.cl> (2023).
- [4] International Agency for Research on Cancer, “Cancer today – Chile fact sheet (globocan 2022),” disponible en: <https://gco.iarc.who.int/media/globocan/factsheets/populations/152-chile-fact-sheet.pdf> (Febrero, 2022).
- [5] Pan American Health Organization, “Perfil de país: Chile,” disponible en: <https://hia.paho.org/es/perfiles-de-pais/chile> (Junio, 2019).
- [6] Fundación Arturo López Pérez, “Fundación arturo lópez perez (falp),” disponible en: <https://www.falp.org> (2024).
- [7] T. Cai, A. A. Giannopoulos, S. Yu, T. Kelil, B. Ripley, K. K. Kumamaru, F. J. Rybicki, and D. Mitsouras, “Natural language processing technologies in radiology research and clinical applications,” *RadioGraphics*, vol. 39, no. 4, pp. 176–191, disponible en: <https://pubs.rsna.org/doi/10.1148/rg.2016150080> (Julio, 2019).
- [8] R. Horev, “Bert explained: State of the art language model for nlp,” disponible en: <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270> (Enero, 2023).
- [9] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining,” disponible en: <https://arxiv.org/abs/1906.08237> (Junio, 2019).
- [10] S. Zhou, L. Wang, N. Wang, H. Liu, and R. Zhang, “Cancerbert: a bert model for extracting breast cancer phenotypes from electronic health records,” disponible en: <https://arxiv.org/abs/2108.11303> (Agosto, 2021).
- [11] S. Raza and B. Schwartz, “Entity and relation extraction from clinical case reports of covid-19,” disponible en: <https://arxiv.org/abs/2010.05430> (Diciembre, 2020).

- [12] Z. Huang, W. Xu, and K. Yu, “Bidirectional lstm-crf models for sequence tagging,” disponible en: <https://arxiv.org/abs/1508.01991> (Agosto, 2015).
- [13] Explosion AI, “spacy: Industrial-strength natural language processing in python,” disponible en: <https://spacy.io> (Enero, 2023).
- [14] S. Loria, “Textblob: Simplified text processing,” disponible en: <https://textblob.readthedocs.io> (Enero, 2021).
- [15] H. Eyre, A. B. Chapman, K. S. Peterson, J. Shi, P. R. Alba, M. M. Jones, T. L. Box, S. L. DuVall, and O. V. Patterson, “Launching into clinical space with medspacy,” disponible en: <https://arxiv.org/abs/2106.07799> (Junio, 2021).
- [16] I. Malashin, I. Masich, V. Tynchenko, A. Gantimurov, V. Nelyub, and A. Borodulin, “Image text extraction and nlp from medical reports,” *Machine Learning and Knowledge Extraction*, vol. 6, no. 2, pp. 1361–1377, disponible en: <https://www.mdpi.com/2504-4990/6/2/64> (Abril, 2024).
- [17] T. Merino, J. Torres, M. L. Garmendia, and G. Acuña, “Carga del cáncer en chile y desafíos actuales,” *Revista Médica de Chile*, vol. 151, no. 12, pp. 1654–1662, disponible en: https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0034-98872023001201654 (Diciembre, 2023).
- [18] G. McGoldrick, Y. Cao, and S. Prince, “Understanding xlnet,” disponible en: <https://arxiv.org/abs/2002.03760> (Enero, 2020).