Automatización de codificación y resumen de informes de exploraciones radiológicas de próstata

Trabajo Fin de Máster

Mariia Chizhikova

Trabajo de investigación para el

Máster en Tecnologías del Lenguaje

Universidad Nacional de Educación a Distancia

Dirigido por los profesores

Prof. Dra. Lourdes Araujo Serna
Prof. Dr. Juan Martínez Romo

Junio 2023

Agradecimientos

Quisiera expresar mi más sincero agradecimiento a todas las personas e instituciones que contribuyeron de manera significativa a la realización de este Trabajo Fin de Máster.

En primer lugar, deseo agradecer a mis tutores, la Dra. Lourdes Araujo Serna y el Dr. Juan Martínez Romo, por su orientación experta y sus valiosas sugerencias a lo largo de todo el proceso de investigación. Su dedicación y conocimiento fueron fundamentales para este trabajo.

También quiero agradecer a todos los investigadores del grupo Sistemas Inteligentes de Acceso a la Información (SINAI) de la Universidad de Jaén, en especial a Maite Martín Valdivia y Alfonso Ureña López, quienes me apoyaron desde el inicio de mi camino en Procesamiento de Lenguaje Natural y me brindaron una base sólida para comenzar con mis estudios en este máster lo cual es un mérito destacable dado que me gradué en Filología Hispánica. Su conocimiento, confianza y paciencia infinitos han sido y siguen siendo indispensables para mi crecimiento académico y personal.

Mi reconocimiento se extiende a HTMédica, principalmente a Pilar López Úbeda y Teodoro Martín Noguerol, por proporcionar el corpus de informes radiológicos en el que se fundamenta mi trabajo. Agradezco sinceramente su tiempo y conocimientos compartidos.

Asimismo, deseo agradecer a mis compañeros de trabajo por sus discusiones, aportes y sugerencias que enriquecieron mi trabajo. La cafeína de los desayunos de todos los días no haría el mismo efecto si no fuera por ellos.

Finalmente, quiero expresar mi infinita gratitud a toda mi familia que me brindó apoyo incondicional y motivación durante todo el proceso. No encuentro palabras que fueran suficientes para darles crédito a mis padres, mi hermana y mi pareja. Todo lo que he logrado no habría sido posible sin ellos.

Resumen

La radiología constituye uno de los pilares fundamentales de la medicina hoy en día al apoyar tanto el diagnóstico, como el tratamiento. El informe radiológico, a su vez, es un componente esencial del estudio en el que se fundamenta la interpretación de los hallazgos por parte del médico remitente. La introducción de registros electrónicos de salud y la digitalización de la información acumulada por el sistema sanitario han creado una necesidad de estructuración de la información contenida en formato de texto libre en los informes clínicos para habilitar su aprovechamiento tanto en la práctica clínica como en la investigación médica.

Con el fin de contribuir tanto a la calidad de la comunicación entre los radiólogos y los médicos remitentes, como a la estructuración de la información contenida en el texto de los informes radiológicos, en el presente trabajo se abordan las tareas la automatización de la codificación clínica y la generación de conclusiones de informes de exploraciones radiológicas de próstata escritos en español.

La tarea de codificación se centra en la clasificación PI-RADS v.2.1 que implementa una escala de 5 puntos basada en la probabilidad de que una combinación de hallazgos esté en correlación con la presencia de un cáncer clínicamente significativo. Con el fin de asignar estos códigos se realizaron 3 experimentos, de los cuales el mejor resultado (0,9372 de macro F1) ha mostrado un sistema basado en el ajuste fino de un modelo pre-entrenado sobre una combinación de textos médicos y clínicos.

Para la tarea de generación automática de conclusiones de informes radiológicos se realizaron experimentos con un modelo de arquitectura codificador-decodificador, resultando su ajuste fino la opción preferente que proporcionó resultados prometedores (0,7545 de ROUGE-L).

Abstract

Nowadays, radiology is one of the fundamental pillars of medicine by supporting both diagnosis and treatment. The radiological report, in turn, is an essential component of the study on which the interpretation of the findings by the referring physician is based. The introduction of electronic health records and the digitization of information accumulated by the health system have created a need for structuring the information contained in free text format in clinical reports to enable their use in clinical practice and medical research.

In order to contribute both to the quality of communication between radiologists and referring physicians and to the structuring of the information contained in the text of the radiological reports, the present paper addresses the tasks of automation of clinical coding and the generation of conclusions of reports of radiological examinations of prostate written in Spanish.

The coding task focuses on the PI-RADS v.2.1 classification that implements a 5-point scale based on the probability that a combination of findings is correlated with the presence of a clinically significant cancer. In order to assign these codes 3 experiments were performed, of which the best result (0.9372 macro F1) has been reached by a system based on fine-tuning of a pre-trained model on a combination of medical and clinical corpora.

For the task of automatic generation of radiological report conclusions, experiments were carried out with an encoder-decoder architecture model, resulting its fine-tuning to be the preferred option that yields promising results (0.7545 ROUGE-L).

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

1.1. Motivación

La introducción de los registros electrónicos de salud (EHR por sus siglas de inglés *Electronic Health Records*) y la digitalización de la información acumulada y generada por el sistema sanitario en general han creado nuevas oportunidades para impulsar la investigación y desarrollo de tecnologías biomédicas que permitan llevar el conocimiento y evidencias almacenadas en forma de texto libre a la práctica. La estructuración de los informes clínicos contribuye a agilizar varios aspectos tanto de la práctica clínica, tales como la formación y la preparación de personal médico, dado el aumento de importancia del papel que juega el creciente volumen de evidencias estructuradas (Olsen, Aisner, y McGinnis, 2007).

Las situación de emergencia sanitaria global provocada por la pandemia de la enfermedad por el virus SARS-CoV-2 destacó aún más la necesidad de disponer de herramientas eficientes de búsqueda y recuperación de la información así como estrategias de explotación para una gran diversidad de tipos de contenido médico (Miranda-Escalada et al., 2020).

La radiología constituye uno de los pilares fundamentales de la medicina hoy en día al apoyar tanto el diagnóstico, como el tratamiento. El informe radiológico, a su vez, es un componente esencial del estudio en el que se fundamenta la interpretación de los hallazgos por parte del médico remitente. El objetivo de un informe radiológico, por tanto, consiste en comunicar los resultados de una exploración al médico remitente y/o paciente (Kahn Jr et al., 2009). La información contenida en informes radiológicos puede ser explotada más allá de diagnóstico y tratamiento de casos concretos, llegando

a ser fuente de datos para la investigación médica o educación (Kahn Jr et al., 2009).

Todo lo expuesto anteriormente sugiere que, además de la evidente trayectoria del desarrollo tecnológico de la radiología encaminada hacia una mejora de las técnicas de imagen utilizadas, se tiene que focalizar la importancia de la calidad de transmisión de los resultados de estas exploraciones a médicos remitentes y su aprovechamiento más eficiente para investigación tanto clínica, como médica.

En los últimos años se estaban desarrollando muchas propuestas de diferente naturaleza para sobrepasar el obstáculo que supone la forma narrativa y la ausencia de una estructura estandarizada de los informes radiológicos. Muchas de estas soluciones involucran la indexación de los documentos o sus partes con códigos de tales ontologías como la Clasificación Internacional de Enfermedades (CIE) (Miranda-Escalada et al., 2020), la Nomenclatura Sistematizada de Medicina (Snomed) (Miranda-Escalada et al., 2022) o clasificaciones más centradas en dominios concretos de la radiología como serían los códigos BI-RADS, PI-RADS y NI-RADS, entre otros (An, Unsdorfer, y Weinreb, 2019). Otras propuestas se centran en mejorar la calidad de contenido legible por humanos abordando la tarea de generación automática de las conclusiones centrándose en aumentar la relevancia de las ideas presentadas en el contenido producido por tales sistemas (Zhang et al., 2018).

El ámbito de la recuperación de información clínica se ha aprovechado de los avances recientes del PLN, en especial del advenimiento de los modelos pre-entrenados transformer (Vaswani et al., 2017) que se han aplicado con éxito tanto a la codificación clínica, como a la generación de resúmenes, como se expondrá en la Sección 2.2. Sin embargo, se nota un importante desbalanceo en cuanto a los idiomas para los cuales se desarrollan los sistemas, dado que la inmensa mayoría de los trabajos se centra en textos escritos en inglés (Névéol et al., 2018). Además, los muchos trabajos que se han llevado a cabo no llegan a abarcar la totalidad de tareas pendientes de resolución.

Según la Asociación Española Contra el Cáncer (AECC), el cáncer de próstata es el tumor más frecuente entre los varones y constituye la segunda causa de mortalidad por cáncer entre hombres, cuya incidencia aumenta con la edad[1]. Según el Observatorio del Cáncer de la AECC, en el año 2022

[1]Fuente: https://www.contraelcancer.es/es/todo-sobre-cancer/tipos-cancer/

35.877 personas fueron diagnosticados de este tipo de cáncer con una clara evolución ascendente con respecto a los años anteriores según se puede ver en la Figura 1.1.

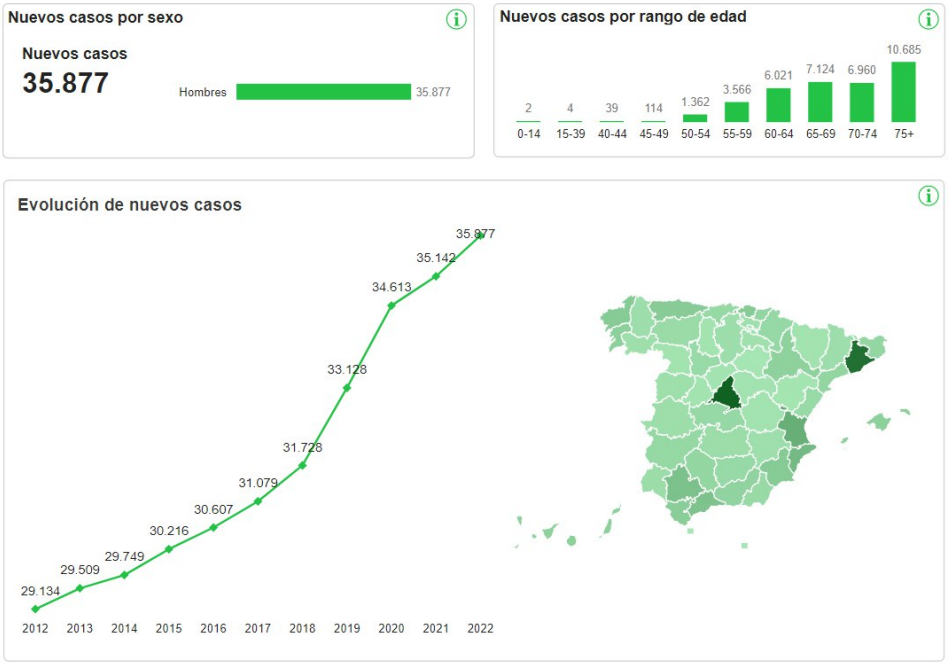


Figura 1.1: Evolución de la incidencia del cáncer de próstata en España. Fuente: https://observatorio.contraelcancer.es/explora/dimensiones-del-cancer. Fecha de consulta: 24/04/2023

La resonancia magnética multiparamétrica (RMNmp) es considerada como la herramienta más importante para la optimización de la biopsia y constituye la exploración previa a la biopsia recomendada para pacientes con alto riesgo de padecer cáncer de próstata clínicamente significativo, según la guía elaborada en conjunto por la Asociación Europea de Urología (EAU), la Asociación Europea de Medicina Nuclear(EANM), la Sociedad Europea de Radioterapia y Oncología (ESTRO), Sociedad Europea de la Radiología Urogenital (ESUR), y la Sociedad Internacional de Oncología Geriátrica (SIOG) (Mottet et al., 2021). Además, esta exploración radiológica tiene utilidad para la estadificación de la enfermedad, la monitorización terapéutica y el seguimiento de pacientes en vigilancia activa.

Con el fin de proporcionar una guía de interpretación de los resultados

cancer-prostata. Fecha de consulta: 24/04/2023

de las exploraciones radiológicas de próstata, la ESUR desarrolló en 2012 la nomenclatura PI-RADS (Barentsz et al., 2012) que en 2015 y 2019 fue actualizada para adaptarla a los desarrollos más recientes en el ámbito de la radiología (Turkbey et al., 2019). La evaluación de PI-RADS v.2.1 utiliza una escala de 5 puntos basada en la probabilidad de que una combinación de hallazgos de RMNmp se correlacione con la presencia de un cáncer clínicamente significativo para cada lesión en la glándula prostática [2].

La interpretación de informes y la asignación manual de códigos según una nomenclatura estandarizada como es PI-RADS es un proceso costoso en cuanto al tiempo y esfuerzo que se puede convertir en una responsabilidad adicional del radiólogo en el caso de que la institución no disponga de codificadores clínicos humanos contratados. Asimismo, la codificación clínica manual puede conllevar errores atribuidos al factor humano.

1.2. Propuesta y objetivos

Con el fin de abordar los problemas expuestos en el apartado anterior, el presente trabajo se centra en un dominio específico de la radiología oncológica de próstata y plantea soluciones tanto para la automatización de la asignación de códigos estandarizados PI-RADS a los informes radiológicos escritos en español, como para la generación automática de conclusiones. Con el fin de mejorar la eficiencia y precisión en estos procesos, se plantea una serie de objetivos específicos que guiarán el desarrollo de las soluciones propuestas. A continuación, se describen detalladamente los objetivos perseguidos en este estudio:

- Estudiar el estado de arte de los algoritmos del Procesamiento de Lenguaje Natural (PLN) aptos para la resolución de las dos tareas: la clasificación de textos según la nomenclatura PI-RADS (Turkbey et al., 2019) y la generación de resúmenes.
- Analizar en detalle el conjunto de informes clínicos con el que estamos trabajando.
- Evaluar la capacidad de los modelos de arquitectura transformer de

[2] Fuente: American College of Radiology: PI-RADS v.2.1: https://www.acr.org/-/media/ACR/Files/RADS/Pi-RADS/PIRADS-V2-1.pdf?la=en. Fecha de consulta: 24/04/2023

ajustarse a la tarea de codificación automática de informes radiológicos según la clasificación PI-RADS v.2.1.

- Estudiar la viabilidad de realización de una optimización de hiperparámetros de un modelo.
- Evaluar la capacidad de un modelo generativo de arquitectura codificador-decodificador de realizar la tarea de codificación clínica según PI-RADS v.2.1.
- Elegir, ajustar y evaluar un modelo generativo de arquitectura encoder-decoder a la tarea de generación de las conclusiones de informes radiológicos.
- Realizar un análisis de errores para conocer las limitaciones de los enfoques propuestos y esbozar trayectorias de su mejora en el futuro.

1.3. Estructura del documento

El presente trabajo se estructura de la siguiente manera.

Capítulo 1. Introducción. Este capítulo presenta los principales fundamentos que han impulsado la elaboración de este trabajo, junto con el análisis de la problemática y el estado actual de la disciplina. Por último, se exponen las diversas aportaciones del trabajo realizado.

Capítulo 2. Estado del arte. Este capítulo proporciona una descripción más detallada de la disciplina en cuestión, abarcando su origen e historia hasta la actualidad. Se exponen las técnicas actuales más comúnmente empleadas para abordar las tareas del tema tratado, así como sus limitaciones.

Capítulo 3. Sistemas propuestos. Este capítulo se enfoca en una descripción detallada de los métodos propuestos, empezando por el análisis del conjunto de datos en el que se fundamenta nuestra aproximación y continuando con una caracterización del sistema implementado para cumplir con los objetivos de este estudio.

Capítulo 4. Evaluación. Este capítulo describe la metodología utilizada para evaluar la propuesta realizada, a la vez que presenta los resultados obtenidos al evaluar el método propuesto en diferentes tareas y sobre colecciones de evaluación de distintos dominios.

Capítulo 5. Discusión. En este capítulo se explica la metodología empleada para evaluar la propuesta realizada, además de presentar los resultados obtenidos al evaluar el método propuesto en el conjunto de evaluación que, igual que los datos de entrenamientos se compone de informes clínicos reales.

Capítulo 6. Conclusiones y trabajo futuro. Este capítulo resume las diversas conclusiones obtenidas a partir del trabajo realizado y plantea posibles líneas de investigación futura.

Capítulo 2

Estado del arte

Este capítulo describe en mayor detalle la tarea que nos ocupa, presentando su origen y su historia hasta el presente. Se muestran las técnicas actuales más utilizadas para la generación automática de resúmenes y codificación de textos tanto de dominio general, como de dominios más específicos dentro de los cuales se desarrolla este trabajo: la biomedicina o la radiología.

2.1. Descripción del problema

El PLN se aplica en el ámbito de la biomedicina con resultados interesantes a tareas de un rango amplio, empezando por tareas intermedias, como el reconocimiento y normalización de entidades nombradas clínicamente relevantes (López-Úbeda et al., 2021) y llegando a implementación de sistemas de apoyo a la toma de decisiones clínicas, como sería el caso de un sistema que combina métodos basados en reglas y aprendizaje automático para detectar y clasificar alergias que constituyen contraindicaciones de la anestesia y relevantes para los cuidados intensivos (Berge et al., 2023).

Las tecnologías de codificación y clasificación de textos clínicos abarcan una variedad de enfoques computacionales que transforman el texto libre de los registros clínicos en datos estructurados asignándoles códigos de terminologías estándar (Stanfill et al., 2010). La acción de resumir, a su vez, se define como "reducir a términos breves y precisos, o considerar tan solo y repetir abreviadamente lo esencial de un asunto o materia"[1], lo cual sugiere que algunas de las características de un buen resumen serían su brevedad

[1]REAL ACADEMIA ESPAÑOLA: Diccionario de la lengua española, 23.ª ed., [versión 23.6 en línea]. https://dle.rae.es Fecha de la consulta: 19/04/2023

y capacidad de condensar ideas importantes. De este modo, un resumen podría definirse como un texto producido a partir de uno o más textos que transmite la información relevante contenida en los documentos originales de una forma más concisa.

Se podría afirmar que un código PI-RADS asignado a un informe radiológico constituye la síntesis muy breve de toda la información relevante relativa al cáncer de próstata. Sin embargo, una RMNmp de próstata suele ser una exploración que visualiza toda la región de la pelvis: desde la bifurcación aórtica hasta el suelo pélvico (Franiel et al., 2017), lo cual aumenta el riesgo de identificación de hallazgos inesperados. Se denomina "hallazgos inesperados" al conjunto de signos radiológicos identificados en un examen de determinada modalidad de imagen que cumplan con dos criterios: no están aparentemente relacionados con los resultados esperados a priori del examen radiológico e implican una emergencia clínica o una situación de urgencia de la que se tiene que informar al médico prescriptor u otro especialista así como al paciente con el fin de preservar su vida y/o prevenir sucesos peligrosos (López-Úbeda et al., 2020). En un informe radiológico, el resumen o conclusión (también llamado "sección de observaciones" o "sección de impresión" - *impression section* en inglés) es la parte que contiene las observaciones, inferencias y conclusiones clave, incluidas las recomendaciones (Kahn Jr et al., 2009). Debido a que una RMNmp de próstata constituye también una exploración de la pelvis, su conclusión podría contener datos relevantes no relativos a la oncología y hallazgos inesperados, lo que hace que la asignación de un código PI-RADS no pueda ser considerado un resumen completo.

En nuestro estudio abordaremos la asignación de códigos PI-RADS a los informes radiológicos como un problema de aprendizaje profundo supervisado, ajustando a ella un modelo de arquitectura Transformer (Vaswani et al., 2017) pre-entrenado sobre una combinación de textos biomédicos y clínicos (Carrino et al., 2021b). La clasificación PI-RADS v.2.1 define una escala de 5 puntos basada en la probabilidad de que la combinación de los hallazgos de la RMNmp esté en correlación con la presencia de un cáncer clínicamente relevante para cada lesión en la glándula prostática [2]:

- PI-RADS 1 - riesgo muy bajo

[2]Fuente: American College of Radiology: PI-RADS v.2.1: https://www.acr.org/-/media/ACR/Files/RADS/Pi-RADS/PIRADS-V2-1.pdf?la=en

- PI-RADS 2 - riesgo bajo
- PI-RADS 3 - riesgo intermedio (la presencia de un cancer clínicamente relevante es equívoca)
- PI-RADS 4 - riesgo alto
- PI-RADS 5 - riesgo muy alto

Dada la posibilidad de presencia de varias lesiones PI-RADS en un paciente, la codificación clínica se configura, en nuestro caso, como un problema de clasificación de textos multietiqueta, es decir, a cada informe se le pueden asignar varias etiquetas del conjunto.

En cuanto a la generación de resúmenes, para esta tarea se diseñará un sistema "sequence-to-sequence" (seq2seq) también basado en un modelo de arquitectura Transformer, con la diferencia de que en este caso se trabajará con un modelo compuesto por dos partes: codificador y decodificador, mientras que el modelo para la clasificación consta solo de un codificador. La generación de resúmenes también se plantea como un problema de aprendizaje supervisado. Con este fin se ajustará un modelo seq2seq pre-entrenado sobre un corpus plurilingüe de dominio general para tomar como entrada el cuerpo de un informe radiológico y devolver su conclusión. La elección de este modelo se debe al hecho de que no nos consta la existencia de un modelo entrenado con textos clínicos de dominio biomédico o clínico de arquitectura codificador-decodificador.

Asimismo, se realizarán experimentos de transformación del problema de clasificación a formato seq2seq, lo cual habilitará el uso del modelo codificador-decodificador para este fin. Nos adentraremos en los detalles de las arquitecturas en la Sección 3.

2.2. Trabajos previos

Como ya hemos mencionado en la Sección 1.1, un informe radiológico constituye la vía principal de comunicación entre los médicos de diferentes especialidades y/o entre los médicos y los pacientes. Es por ello que la radiología se considere un campo en el que la aplicación de las tecnologías de lenguaje sea especialmente beneficioso (Mozayan et al., 2021).

Recientemente, se desarrollaron numerosos métodos para aprovechar la información contenida en los informes radiológicos de texto libre que varían

entre sistemas de extracción de información que buscan detectar términos específicos clínicamente relevantes dentro del contexto de los informes (López-Úbeda et al., 2021) e incluso sistemas de apoyo al diagnóstico que integren varias metodologías como el reconocimiento de entidades y clasificación de textos (López-Úbeda et al., 2020b).

Esta sección describe con mayor detalle el origen y la historia tanto de la codificación clínica de informes radiológicos como la generación automática de resúmenes de estos textos. Se presentan los métodos contemporáneos de mayor aplicación para solventar las tareas de mayor relevancia en el ámbito abordado, así como los desafíos que conllevan y las investigaciones en curso en dichos ámbitos.

2.2.1. Codificación clínica de informes radiológicos

Con el término "codificación clínica" nos referimos al proceso de mapeo de conceptos relevantes de EHR de un paciente a códigos de una terminología o nomenclatura. Esta tarea puede realizarse tanto a nivel de texto completo, como a nivel de entidades nombradas relevantes, lo cual requiere su reconocimiento como paso anterior. Actualmente, la codificación se realiza manualmente, lo que implica la contratación por parte de clínicas y hospitales de trabajadores especializados. La principal desventaja de esto es que la codificación manual de textos es un proceso lento, lo que hace que la anotación de un texto clínico puede demorarse semanas desde su escritura haciendo imposible el aprovechamiento inmediato de información (Catling, Spithourakis, y Riedel, 2018). Es por ello que el desarrollo de sistemas capaces de realizar una codificación precisa automáticamente ha llamado la atención de la comunidad del PLN.

A día de hoy existen muchos esquemas estandarizados de codificación clínica distintos que pueden ser generales, como es el caso de la Clasificación Internacional de Enfermedades, 10ª revisión (CIE-10) y de la Nomenclatura Sistematizada de Medicina - Términos Clínicos (SNOMED-CT), o más específicos, como es el caso de la clasificación PI-RADS en la que nos centramos en el presente estudio. A diferencia de las ontologías más generales, los códigos PI-RADS no solo señalan la presencia de un tipo de cáncer, sino que se desarrollan categorías de evaluación que resumen los niveles de sospecha o riesgo y se pueden utilizar para seleccionar a los pacientes para biopsias y otras exploraciones (Turkbey et al., 2019).

Los orígenes de las nomenclaturas de enfermedades se remontan a los años 70 del siglo XIX cuando fue preparada la Nomenclatura Americana de Enfermedades que se publicó, debido a una discontinuidad en el trabajo, solo en 1933 bajo el título "Standard Classified Nomenclature of Disease" (Logie, 1933). Esta nomenclatura se planteó como base para la realización del registro estadístico de enfermedades con vistas a descubrir las verdades estadísticas relativas a su historia y naturaleza (Organization, 1957). En 1955 fue publicada la Clasificación Estadística Internacional de Enfermedades que ofrece categorías a las que puedan ser mapeadas todas las condiciones nosológicas, lo cual fue logrado utilizando el método de agrupación selectiva (Organization, 1957).

La preocupación por la automatización de la asignación de códigos clínicos tampoco es muy reciente: los trabajos tempranos se empezaron a publicar en los años 60-70 del siglo pasado. En 1968 se propuso uno de los primeros sistemas que automatizaban la clasificación clínica de informes según la ya citada Clasificación Estadística Internacional de Enfermedades que emplea un enfoque basado en reglas y *matching* con diccionarios de términos relevantes (Scott, 1968).

Recientemente ha habido un incremento tanto en el número de las nomenclaturas utilizadas para indexar los textos clínicos, como en la variedad de metodologías propuestas para la automatización de estos procesos. Entre las ontologías más estudiadas figuran los ya mencionados anteriormente CIE-10 (y sus versiones anteriores) y SNOMED-CT, así como Unified Medical Language System (UMLS) y Medical Subject Headings (MeSH) (Stanfill et al., 2010).

Es muy importante resaltar que la mayoría de los estudios realizados en el ámbito de la codificación clínica se centran en textos escritos en inglés: para el año 2017 se habían publicado en PubMed [3] solo 39 artículos relativos al PLN biomédico aplicado a textos en español. Sin embargo, la habilidad de acceder a información clínica estructurada en idiomas distintos al inglés abre acceso a datos que abarcan cohortes de pacientes tratados en países que no son angloparlantes (Névéol et al., 2018).

Con el fin de impulsar la investigación en este ámbito, en los últimos

[3]Pubmed es un motor de búsqueda de libre acceso que permite consultar la base de datos MEDLINE de citaciones y resúmenes de artículos de investigación biomédica. Es una fuente de información especializada en ciencias de la salud, con terminología biomédica y en constante actualización.

años se han desarrollado y liberado tanto conjuntos de datos anotados por expertos, como módulos que se pueden aprovechar para construir sistemas de PLN de diferente complejidad: herramientas para segmentación de secuencias (Costumero et al., 2014), detección de negación (Cotik et al., 2016), modelos de aprendizaje profundo pre-entrenados (Carrino et al., 2022) y ajustados para una tarea específica (Chizhikova et al., 2022), entre otros.

En el dominio específico de radiología se han desarrollado sistemas de distinta índole para textos en español. Para la tarea de codificación clínica se han propuesto enfoques basados en algoritmos de distinta naturaleza. En 2015 se propuso un sistema que combina el reconocimiento de términos clínicamente relevantes con la detección de negación y aplicación de reglas para clasificar, con un grado de sensibilidad bastante alto (0.96 de recall) los informes según la clasificación RadLex, una ontología en inglés desarrollada específicamente para indexación y recuperación de información (RI) radiológica (Cotik, Filippo, y Castano, 2015). Otros estudios recientes exploran la aplicación de técnicas basadas en el ajuste fino de modelos pre-entrenados para la clasificación de informes radiológicos según ontologías estandarizadas como el CIE-10 (Chizhikova et al., 2023) o códigos propios internos a una institución (López-Ubeda et al., 2020a).

Pese al reciente auge de las tecnologías de lenguaje aplicadas a textos médicos y clínicos en español, no nos consta la existencia de recursos, estudios y sistemas publicados que tengan como objetivo automatizar la codificación de informes radiológicos con códigos PI-RADS v.2.1. Los enfoques propuestos para la clasificación PI-RADS involucran tratamiento de imágenes mediante implementación de sistemas basados en redes neuronales convolucionales y transformers para asistir a los codificadores humanos en la asignación de clase PI-RADS a las patologías localizadas (Bijl, Blaumer, y Matuschek, 2022). Por otra parte, se propusieron también metodologías basadas en reglas (Zhang et al., 2022) o arboles de decisión que toman como entrada los bigramas relevantes previamente seleccionados por un sistema basado en reglas también (Ma et al., 2017).

2.2.2. Generación automática de resúmenes

En la Sección 1.1 hemos desarrollado la definición del concepto de "resumen" fundamentándonos en la premisa de que el objetivo de un resumen consiste en presentar las ideas principales de un documento en menos espa-

cio. De este modo, se puede afirmar que el mayor reto que plantea la acción de sintetizar consiste en resaltar segmentos informativos a expensas del resto (Radev, Hovy, y McKeown, 2002). Se pueden identificar muchos tipos distintos de resúmenes que se distinguen teniendo en cuenta diversos aspectos: según su autor, su objetivo y forma, entre otros. También se distingue entre resúmenes extractivos y abstractivos: lo primero se refiere a selección de oraciones y datos relevantes dentro del documento original, mientras que la elaboración de un resumen abstractivo requiere parafrasear estos conceptos significativos (Borko y Bernier, 1975). Dado el hecho de que un resumen abstractivo involucra la extracción de información relevante y su parafraseo, la tarea de generación de resúmenes estuvo inicialmente más ligada a la recuperación de información, pero se desligó de esta rama dado que los enfoques más recientes derivan del procesamiento de lenguaje natural y la generación de lenguaje en particular (Radev, Hovy, y McKeown, 2002).

Durante las etapas tempranas de investigación en el ámbito de la generación automática de resúmenes de textos se propusieron sistemas relativamente simples que resaltaban pasajes importantes dentro del texto original computando unas métricas específicas por términos, oraciones o párrafos. Así, se propuso una metodología basada en frecuencias de aparición de términos y su posición relativa dentro de la oración como medida de relevancia de esta oración dentro del texto (Luhn, 1958).

Más tarde, con el florecimiento del PLN estadístico se propusieron enfoques basados en el aprendizaje automático, cuyos pioneros implementaban clasificadores Bayesianos para combinar las características de un corpus de artículos científicos y sus resúmenes abstractivos (Kupiec, Pedersen, y Chen, 1995).

A causa del éxito de la revolución que desencadenó en el ámbito de la búsqueda Web la implementación del algoritmo PageRank basado en la teoría de grafos (Page et al., 1998), la comunidad de PLN se interesó por aprovechar esta teoría trabajando con grafos léxicos y semánticos. El algoritmo TextRank (Mihalcea y Tarau, 2004) es la implementación probablemente más exitosa de técnicas basadas en grafos para la generación de resúmenes que se presentan, en este caso, como una lista de términos y sintagmas clave (los llamados *keywords*). La principal ventaja de este método radica en que es un algoritmo no supervisado que no necesita datos anotados para su entrenamiento y es aplicable a otros idiomas (salvo los que utilicen el

sistema de escritura logográfico debido a la dificultad añadida en cuanto a la identificación de términos y sintagmas).

En cuanto a los enfoques abstractivos, se puede decir que en su origen se encuentran las metodologías que se pueden denominar "compresivos", entendiendo el término compresión como el proceso de 'exprimir' la información irrelevante. Uno de los ejemplos sería el sistema *Ultra-summarization* (Witbrock y Mittal, 1999) que parte de reconocimiento de unidades textuales relevantes con un modelo estadístico entrenado sobre un corpus de artículos periodísticos y sus resúmenes, e inicializa un modelo probabilístico de generación de lenguaje basado en predicción de secuencias de bigramas.

Tradicionalmente la evaluación de la calidad de los resúmenes generados implicaba esfuerzo humano y juicio en términos de unos criterios bastante abstractos y subjetivos: la coherencia, la consistencia, la legibilidad y la corrección. Este procedimiento, igual que codificación clínica, es un proceso que requiere mucho tiempo, lo que hace que las campañas de evaluación sean muy difíciles de organizar de forma frecuente. Es por ello que se desarrollaron métricas estadísticas basadas en co-occurencia de ciertas unidades textuales (palabras o caracteres). La métrica más utilizada actualmente para evaluar la calidad de los resúmenes generados automáticamente es ROUGE (Recall-Oriented Understudy for Gisting Evaluation) y más precisamente su variante ROUGE-L que mide la subsecuencia común más larga (Lin, 2004). Dedicaremos más espacio a las métricas y el procedimiento de su computación en la Sección 4.2.

Los métodos de resumen abstractivo de textos más recientes se basan en algoritmos de aprendizaje profundo. Uno de los primeros sistemas propuestos que implemente una red neuronal para esta tarea se centra en la tarea de generación de titulares periodísticos que se resuelve de una forma eficiente mediante un modelo de lenguaje neuronal inspirado en los avances de la traducción automática contemporánea (Rush, Chopra, y Weston, 2015). Este sistema implementa una adaptación de una red *feed-forward* combinada con un codificador contextual basado en un mecanismo de atención entrenado para minimizar la probabilidad logarítmica negativa de un conjunto de resúmenes aplicando el algoritmo de descenso de gradiente estocástico.

La introducción de la arquitectura Transformer (Vaswani et al., 2017) diseñada inicialmente para la tarea de la traducción automática, pero que se expandió a una gran variedad de tareas de PLN, supuso una transición

hacia enfoques que aprovechen modelos pre-entrenados y ajustados a tareas específicas mediante el aprendizaje de transferencia o *transfer learning* (Kenton y Toutanova, 2019). El éxito de los modelos pre-entrenados llevó a la exploración de los límites de esta metodología. Nuestra propuesta se basa en un modelo pre-entrenado que surge de un estudio de las posibilidades de aplicar un único modelo, objetivo, procedimiento de entrenamiento y estrategias de decodificación a la resolución de distintos problemas.

El modelo T5 ("Text-to-Text Transfer Transformer") sigue bastante fielmente la implementación original de la arquitectura Transformer (Vaswani et al., 2017) y fue entrenado sobre una variedad de tareas transformadas en formato "text-to-text", lo cual implica que el modelo reciba como entrada un texto con un prefijo que especifique la tarea a realizar.

Hasta ahora veníamos citando sistemas de generación de resúmenes desarrollados para textos de dominio general. En el dominio radiológico, las aplicaciones tempranas constituían enfoques extractivos de reconocimiento de términos relevantes (Hripcsak, Kuperman, y Friedman, 1998). La primera aproximación a la generación de conclusiones de informes radiológicos coherentes y legibles por el ser humano implementa un modelo neuronal "seq2seq" basado en redes neuronales LSTM (Long Short-Term Memory) bidireccionales (Graves y Graves, 2012) que incorpora también un codificador para la información de antecedentes no contenida en el informe mismo. Los autores reportan que además de buenos resultados mostrados por las métricas estadísticas ROUGE, han sometido las conclusiones generadas a una evaluación llevada a cabo por radiólogos que las calificaron favorablemente (Zhang et al., 2018). Asimismo, los autores de este trabajo probaron que este mismo sistema extendido por un modelo generador de cursor (*pointer generation model*) mejora la exactitud fáctica de los resultados. Siguiendo esta línea, posteriormente se propuso una extensión al generador de cursores que incorpora conocimiento específico al dominio mediante la codificación de términos relevantes con códigos de ontologías como RadLex o UMLS y los codifica en un vector contextual por separado que es usado luego en el proceso de generación (MacAvaney et al., 2019).

Otros enfoques para mejorar la corrección de los resúmenes de informes radiológicos generados se basan en el aprendizaje por refuerzo (RL por sus siglas en inglés *Reinforcement Learning*). Se propuso realizar un ajuste del sistema generativo mediante RL para optimizar la legibilidad y y la precisión

clínica mediante la metodología conocida como entrenamiento de secuencia autocrítica (SCST - *Self-Critical Sequence Training*) que utiliza el algoritmo REINFORCE y minimiza la recompensa negativa esperada como una función de los parámetros de la red (Liu et al., 2019a). Es notable que se utilizan dos tipos de recompensas, una que se fundamenta en un reconocedor de entidades clínicamente relevantes y otra que evalúa la calidad del texto generado.

Más recientemente se desarrolló un marco general dentro del cual la calidad del resumen generado se evalúa comparándolo con el resumen de referencia mediante una métrica específica que fue probada como una recompensa apta para optimizar la exactitud fáctica de las generaciones de un modelo mediante RL (Zhang et al., 2020).

Los enfoques más actuales de generación de sección de impresión están orientados hacia modelos multimodales que combinen el procesamiento de imágenes con técnicas de PLN con el fin de mejorar la calidad del resultado incorporando características visuales (Delbrouck, Zhang, y Rubin, 2021).

El ámbito de generación de resúmenes de informes radiológicos, pese a su consolidación muy reciente, cuenta con un trabajo bastante prolífico que continúa con organización de tareas compartidas en talleres de congresos como el Congreso de la Asociación de Lingüística Computacional (ACL). Por citar un ejemplo, podemos mencionar la tarea MEDIQA 2021 en el workshop BioNLP 2021 que atrajo la atención de la comunidad al desarrollo de sistemas para generación de resúmenes fácticamente correctos de informes radiológicos escritos en inglés (Abacha et al., 2021).

No obstante, el resumen automático de informes escritos en español todavía no cuenta con tanta atención por parte de la comunidad científica. Nos consta solo la existencia de una propuesta para generación de resúmenes abstractivos de artículos biomédicos utilizando distintos módulos que contienen reglas de una dimensión individual o una combinación de dimensiones de la descripción lingüística: las dimensiones textual, léxica, discursiva y sintáctica-comunicativa (Da Cunha, Wanner, y Cabré, 2007). En el dominio clínico no se han puesto disponibles propuestas de metodologías para resolver la tarea, lo cual influirá notablemente la selección de métodos implementados durante la realización del presente trabajo.

Capítulo 3

Sistemas propuestos

En este capítulo se describen en profundidad los sistemas propuestos para las dos tareas estudiadas. El primer lugar, se realizará un análisis descriptivo-exploratorio de los datos con los que estamos trabajando y se detalla el pre-procesamiento aplicado. En segundo lugar, se describen los sistemas diseñados, sus arquitecturas y los experimentos llevados a cabo y su configuración.

3.1. Conjunto de datos

En este apartado se ofrece una descripción detallada del conjunto de datos en el que se fundamenta el presente estudio: una colección de 5.000 informes clínicos no estructurados de exploraciones radiológicas de próstata. Nuestro estudio es retrospectivo y abarca informes generados entre abril de 2019 y junio de 2022 en HTMédica, una clínica radiológica privada con cedes en Andalucía, Asturias y Castilla la Mancha.

El conjunto de datos con el que estamos trabajando es bastante variado en cuanto a las exploraciones documentadas: como se puede ver en la Figura 3.1, la mayoría de los informes de las exploraciones por resonancia magnética son exámenes prostáticos multiparamétricos (49,8 % del total) o de vísceras pélvicas (38,8 % del total). Entre otros tipos de exámenes figuran RMs de zonas cervical y lumbar, región anal, así como las RMs de pecho y todo el cuerpo.

Los informes analizados se estructuran típicamente en cinco secciones distintas: información clínica, comparación con estudios anteriores, detalles técnicos, hallazgos y conclusiones. Para salvaguardar la privacidad del pa-

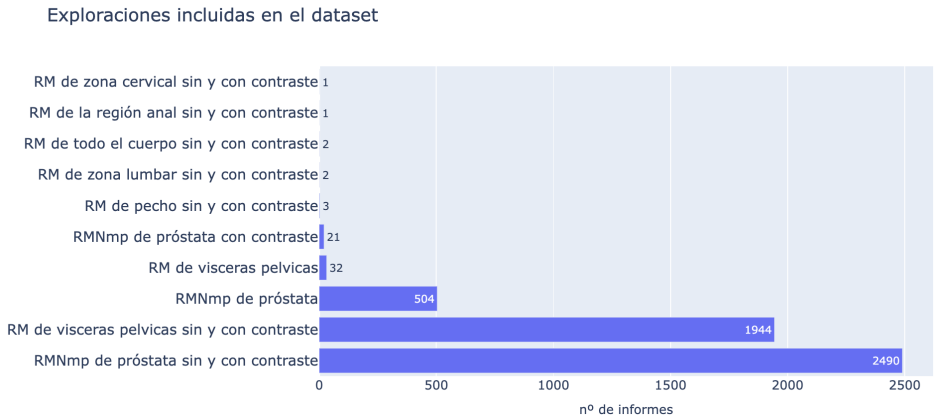


Figura 3.1: Pruebas radiológicas documentadas en el conjunto de datos

ciente y del especialista, se llevó a cabo un proceso de anonimización de los informes, donde se eliminaron los datos sensibles.

En cuanto el proceso de etiquetado, un radiólogo con doce años de experiencia fue el encargado de etiquetar cada uno de los 5.000 informes de radiología, clasificándolos según las categorías PI-RADS v2.1. Es importante destacar que cada informe radiológico puede contener más de una lesión, lo que plantea un desafío de etiquetado multietiqueta, ya que un informe puede estar asociado con más de un PI-RADS. El número mínimo de lesiones o PI-RADS encontrado en los informes fue de uno, mientras que el máximo registrado fue de tres lesiones en un solo informe. La Figura 3.2 muestra la comparativa entre el número de informes etiquetados con una, dos y tres categorías.

Con el fin de poder entrenar y evaluar los sistemas de aprendizaje profundo dividimos el dataset de modo que 3.490 informes (70 % del total) entraron en el conjunto de entrenamiento y 1,510 documentos formaron el conjunto de evaluación. Para crear una división que no sesgue de ninguna manera los subconjuntos resultantes hemos realizado un split estratificado. La Figura 3.3 ilustra la distribución de frecuencias de aparición de las distintas etiquetas en los conjuntos de entrenamiento y evaluación.

Para poder realizar la optimización de hiperparámetros del modelo y seguir llevando a cabo una evaluación final fiable y objetiva consideramos dividir en dos partes el conjunto de evaluación, también de manera estratificada. De este modo conseguimos el tercer subconjunto - el de validación o

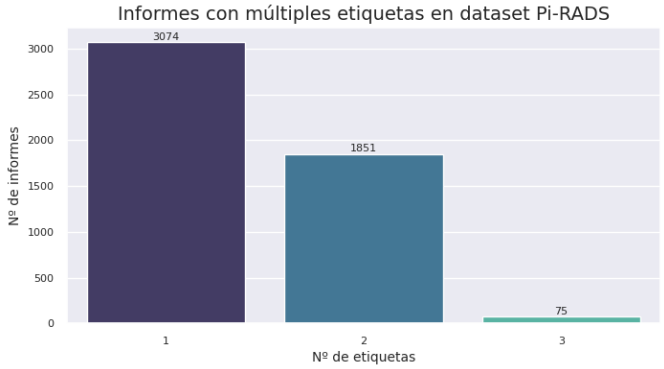


Figura 3.2: Estadísticas de número de etiquetas asignadas a los informes del dataset

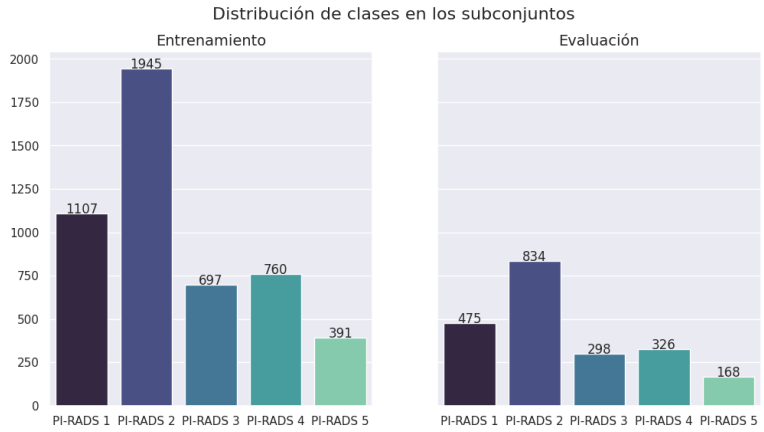


Figura 3.3: Distribución de etiquetas en los subconjuntos train y test iniciales

desarrollo. De este modo conseguimos 3 subconjuntos de 3.490, 753 y 757 informes en los datasets de entrenamiento, validación y evaluación respectivamente. La Figura 3.4 ilustra las frecuencias de aparición de las etiquetas en los tres conjunto de datos.

En cuanto a la longitud de los textos en el dataset, el informe más largo contiene 517 tokens[1], el más corto suma unos 76 tokens siendo la media 137,7 con la desviación estándar igual a 38,3. El ejemplo 3.1.1 muestra un informe extraído del dataset.

Ejemplo 3.1.1. *INFORMACION CLINICA Biopsia de prostata : pin 2, 2017. Psa en ascenso. HALLAZGOS: Volumen prostático: 45 cc. Hipertro-*

[1]Estas medidas fueron obtenidas realizando la división de textos en tokens usando espacios en blanco y signos de puntuación como separadores.

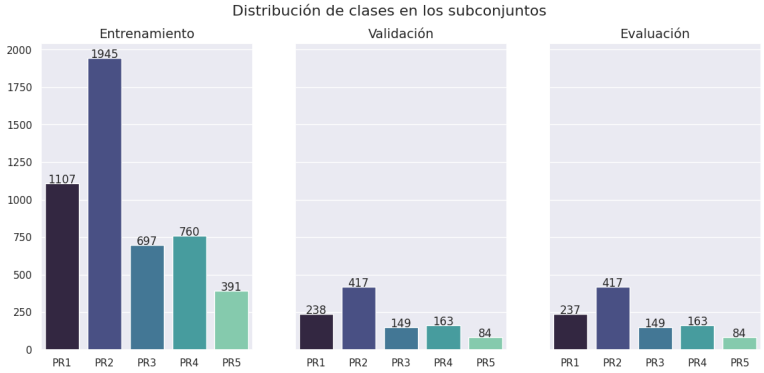


Figura 3.4: Distribución de etiquetas en los subconjuntos

fia de la zona de transición que comprime y adelgaza la zona periférica. No se observan áreas claramente sospechosas de carcinoma clínicamente significativo en secuencia T2, difusión ni contrastada. Múltiples imágenes nodulares en zona transicional compatibles en primer término con nódulos hiperplásicos (PIRADS 1 y 2). Cápsula y grasa periprostática sin alteraciones evidentes. Vesículas seminales y ampollas deferenciales presentes y simétricas. Sin alteraciones parietales ni del contenido. No se observan adenomegalias. Estructuras óseas sin lesiones sospechosas de malignidad. CONCLUSION: HPB. No se observan áreas claramente sospechosas de carcinoma clínicamente significativo demostrables por esta técnica.

3.1.1. Pre-procesamiento

El conjunto de datos fue proporcionado por la clínica radiológica en formato Excel, que es la herramienta que se utilizó para el etiquetado. Con el fin de convertir los datos en un formato coherente con las tareas a realizar tuvimos que pre-procesarlo.

Codificación automática Desarrollamos y aplicamos dos estrategias distintas de preparación de los datos para la codificación clínica dado que abordamos esta tarea tanto con el enfoque de clasificación automática, como con la metodología seq2seq.

En primer lugar, conviene notar que los radiólogos muchas veces incluyen menciones de los códigos PI-RADS en el cuerpo del texto. Con el fin de crear un algoritmo que no esté sesgado con las menciones textuales expresas de las etiquetas, como primer paso del pre-procesado se implementó un algoritmo basado en reglas para eliminar los códigos y sus acrónimos de los textos.

Anotación original	One-hot encoding	Etiqueta seq2seq
pr1, pr2	[1.0, 1.0, 0.0, 0.0, 0.0]	Pi-RADS-1 Pi-RADS-2

Tabla 3.1: Ejemplo de la anotación pre-procesada

Después de aplicar este procedimiento, la longitud máxima del texto en el corpus se redujo a 511 tokens, la mínima pasó a ser a 72 y la media se mantuvo en 137,7, mientras que la desviación estándar bajó a 36,9.

En segundo lugar, las etiquetas fueron codificadas mediante la técnica de *one-hot encoding* que es una técnica utilizada para transformar variables categóricas en un formato numérico que pueda ser procesado por algoritmos de aprendizaje automático. Mediante esta codificación, cada etiqueta se representa como un vector en el que las posiciones que tienen el valor 1 indican las clases correspondientes, mientras que todas las demás posiciones tienen el valor 0.

Para poder abordar la tarea de clasificación con un modelo seq2seq aplicamos un pre-procesamiento diferente en el que las etiquetas se representaban con sus denominaciones textuales: "Pi-RADS-1", por ejemplo. Para los textos que tenían varias etiquetas asignadas se generaba un *string* con sus códigos separados con espacios en blanco: "Pi-RADS-1 Pi-RADS-2", por ejemplo. Asimismo, debido a las peculiaridades de funcionamiento del sistema que explicaremos con más detalle en la Sección 3.2.3, consideramos añadir al principio de cada texto un prefijo específico a la tarea de clasificación. De este modo, a cada entrada del corpus le fue añadido al principio un prefijo "Clasifica: ".

La Tabla 3.1 muestra un ejemplo de anotación de una entrada del corpus pre-procesada de las dos maneras anteriormente descritas.

Generación automática de resúmenes Para conseguir entrenar un modelo de generación de resúmenes, tuvimos que separar las conclusiones de los informes. Como hemos mencionado en la Sección 3.1, los informes proporcionados generalmente tienen divisiones internas en secciones que se marcan con el título de sección escrito en mayúsculas. Por lo tanto, para conseguir un conjunto de datos válido para entrenar y evaluar un sistema de generación automática de resúmenes, hizo falta implementar un sistema basado en reglas que reconozca la sección de conclusión marcada típicamente con títulos como "CONCLUSION", "RESÚMEN" y sus variantes. Conviene

	max. tokens	min. tokens	promedio (STD)
Informes	477	57	111,8 (31,9)
Conclusiones	135	2	24,8 (14,9)

Tabla 3.2: Estadísticas de longitud de informes y sus resúmenes

aclarar que la preparación de los textos para esta tarea no implicó eliminación de menciones expresas de códigos PI-RADS. La Tabla 3.2 muestra la comparativa de las medidas de longitud de los informes y sus conclusiones.

3.2. Automatización de la codificación clínica

En esta sección se detallan los sistemas implementados para cumplir con el objetivo de la automatización de la codificación clínica según la ontología PI-RADS v.2.1. Se describen las arquitecturas implementadas, así como las configuraciones experimentales de cada una de las aproximaciones.

3.2.1. Primera aproximación

Como hemos mencionado en la Sección 2.2.1, recientemente, las metodologías basadas en el ajuste fino de los modelos pre-entrenados de arquitectura transformer se aplican con resultados muy prometedores a variadas tareas dentro del dominio del PLN médico y clínico.

En el marco de la primera aproximación que realizamos en este trabajo, formulamos el problema de la asignación de códigos PI-RADS v.2.1. como una tarea de clasificación automática de textos y en concreto, su variante multietiqueta.

Siguiendo la metodología propuesta en (Kenton y Toutanova, 2019), realizamos un ajuste fino (ing. *fine-tuning*) de un modelo pre-entrenado para la clasificación de textos. El modelo que hemos seleccionado para el ajuste sigue la arquitectura RoBERTa, una versión mejorada de BERT (Liu et al., 2019b). Las diferencias principales entre BERT y RoBERTa consisten en la introducción del tokenizador basado en la metodología Byte-Pair Encoding (Sennrich, Haddow, y Birch, 2015), eliminación de los embeddings de la codificación de tipo de token (eng. *token type id*) a favor de la introducción de caracteres especiales que marquen la división entre segmentos textuales distintos y eliminación del objetivo de la predicción de la oración siguiente

(eng. *next sentence prediction*) del proceso del pre-entrenamiento.

Dada la naturaleza de nuestros textos, seleccionamos un modelo pre-entrenado sobre una combinación de textos médicos y clínicos (RoBERTa-clinical) obtenidos de fuentes tales como MedicalCrawler - la colección de textos en español extraídos de sitios web médicos (Carrino et al., 2021a) o PubMed - un repositorio de artículos científicos de acceso abierto[2]. Estos autores señalan que las diferencias estructurales y léxicas que distinguen los textos médicos de los clínicos, hacen que se sea necesario entrenar modelos de lenguaje de dominio clínico para obtener un rendimiento óptimo. Sin embargo, la ausencia de un corpus clínico de gran escala no permite realizar un pre-entrenamiento, por lo que se tuvo que recurrir a una combinación de textos de ambos dominios (Carrino et al., 2021b).

Hemos realizado el ajuste fino del modelo RoBERTa-clinical a la tarea de clasificación multietiqueta mediante la adición de una cabeza de clasificación. Este módulo se compone de las siguientes capas:

1. Dropout - ajusta a 0 los valores de un porcentaje fijado (0.1 en nuestro caso) de las salidas. Esta técnica tiene como objetivo regularizar los datos de cara a su entrada en la siguiente capa para evitar el sobreajuste al modelo (Srivastava et al., 2014).

2. Capa lineal - aplica una transformación lineal a los datos de entrada: $y = xA^T + b$.

3. Aplicación de la función de tangente hiperbólica elemento por elemento: $Tanh(x) = tanh(x) = \frac{exp(x)-exp(-x)}{exp(x)+exp(-x)}$

4. Segunda capa dropout.

5. Capa lineal de salida que devuelve un vector de 5 dimensiones (una por cada código PI-RADS).

La Figura 3.5 muestra un esquema de la arquitectura del clasificador.

Para levar a cabo el entrenamiento del clasificador nos basamos en la biblioteca Huggigface `transformers` (Wolf et al., 2019) para Python. El algoritmo de optimización que fue seleccionado para realizar el ajuste es el AdamW, una modificación del algoritmo de descenso de gradiente estocástico con tasa de aprendizaje (eng. *learning rate*) adaptativa en función de cada parámetro optimizado que desvincula la regularización $L2$ del parámetro

[2] https://pubmed.ncbi.nlm.nih.gov/

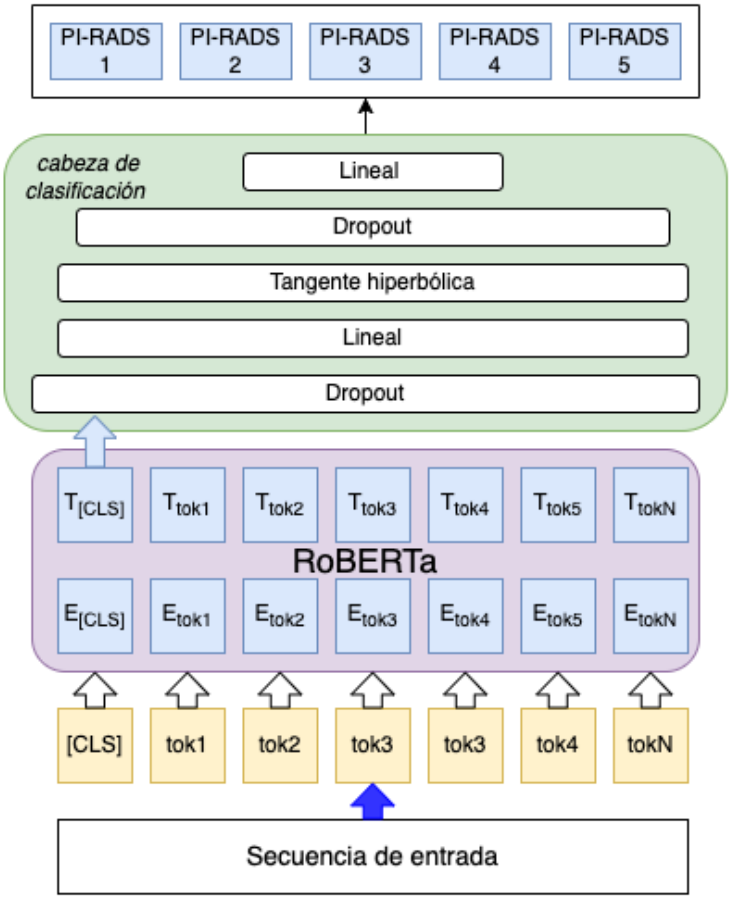


Figura 3.5: Arquitectura del clasificador basado en modelo de lenguaje pre-entrenado RoBERTa-clinical

`weight decay` (Loshchilov y Hutter, 2017). La función de pérdida definida para el entrenamiento es la comúnmente utilizada para problemas de clasificación multietiqueta: entropía cruzada binaria combinada con una capa de sigmoide, también conocida como `BCEWithLogitsLoss`. Se computa según la siguiente fórmula:

$$l(x, y) = \frac{\sum_{n=1}^{N} l_n}{N}$$

$$l = -w[y \cdot \log\sigma(x) + (1 - y) \cdot \log\sigma(1 - \sigma(x))]$$

donde n = batch size, x=predicciones, y=etiquetas y w=peso.

Realizamos el ajuste fino del modelo con todos los hiperparámetros por defecto de la librería `transformers`, menos el número de las épocas de en-

Parámetro	**Valor**
Learning rate	0,00005
Batch size de entrenamiento	8
Batch size de evaluación	16
Warmup steps	0
Weight decay	0
AdamW ϵ	0.00000001
AdamW β_1	0.9
AdamW β_2	0.999
Tipo de scheduler	lineal
Paciencia	5
Épocas de ajuste	10

Tabla 3.3: Hiperparámetros del clasificador de la 1ª aproximación

trenamiento que fue elegido mediante la implementación de una estrategia de interrupción temprana (ing. *early stopping*) que para el entrenamiento y recupera la mejor versión del modelo cuando se alcanza un determinado número de épocas sin mejorarse la métrica de referencia, que en nuestro caso es la variante macro de la medida F1 [3]. Para hacer posible la implementación de esta estrategia, realizamos la evaluación del modelo después de cada época. Los hiperparámetros, incluida la *paciencia* del algoritmo de la interrupción temprana (el número de épocas sin mejora de la métrica después del cual se para el entrenamiento) y el total de las épocas de ajuste completadas se resumen en la Tabla 3.3.

El ajuste fue realizado haciendo uso de una tarjeta gráfica NVIDIA A-100 y tardó alrededor de 14 minutos en ejecutarse.

3.2.2. Segunda aproximación

Los autores del artículo que introduce la arquitectura RoBERTa enfatizan en que el modelo es muy sensible a los cambios de tales hiperparámetros como el learning rate y el épsilon de optimizador (Liu et al., 2019b). La búsqueda de los hiperparámetros óptimos se ha señalado como uno de los pasos más difíciles de los proyectos de aprendizaje automático y muchos métodos y herramientas se han propuesto en los últimos años para aliviar esta carga (Akiba et al., 2019). Algunas de las dificultades asociadas a dicho

[3] Las métricas de evaluación se describen detalladamente en la Sección 4.2.1

proceso son la construcción de un espacio de búsqueda óptimo para balancear la duración del experimento y el rendimiento de los modelos conseguidos, así como la disponibilidad de recursos computacionales limitada. El principal objetivo de nuestra segunda aproximación a la codificación clínica es evaluar la viabilidad de la realización de una optimización de hiperparámentros de un clasificador de arquitectura igual a la que hemos descrito en la sección anterior.

Para cumplir con este objetivo se realizará una búsqueda de hiperparámetros basada en el *framework* Optuna[4], que integra estrategias para mitigar las dificultades mencionadas anteriormente. En primer lugar, proporciona una implementación del muestreador fundamentado en el enfoque de estimador Parzen estructurado en árbol (TPE por sus siglas en inglés: Tree-structured Parzen Estimator Approach) (Bergstra et al., 2011). Durante cada una de las pruebas, para cada uno de los parámetros a optimizar, TPE ajusta un modelo de mezcla gaussiana (GMM) $l(x)$ al conjunto de valores de parámetro asociados con los mejores valores objetivos, y otro GMM $g(x)$ a los valores de parámetro restantes. Se elige el valor del parámetro x que maximiza la relación $\frac{l(x)}{g(x)}$ (Bergstra et al., 2011; Akiba et al., 2019).

En el segundo lugar, Optuna proporciona estrategias de interrupción temprana de pruebas no prometedoras lo cual se puede considerar un elemento esencial para bajar el coste de la optimización. Dicho mecanismo, de forma muy parecida a como funciona la implementación de interrupción temprana descrita en la Sección 3.2.1, monitoriza los valores objetivo intermedios y termina las pruebas que no cumplen con los criterios establecidos por el algoritmo Hyperband (Li et al., 2018).

En esta segunda aproximación a la tarea de clasificación de textos según las categorías PI-RADS v.2.1 se realizará una optimización de hiperparámetros fundamentada en el *framework* Optuna y, específicamente, en las estrategias descritas en los párrafos anteriores. Se harán 10 pruebas con el fin de seleccionar valores óptimos de los siguientes hiperparámetros: learning rate, batch size de entrenamiento, weight decay, épsilon del optimizador y el número de los pasos de calentamiento (*warmup steps*). La métrica de referencia durante el proceso de optimización es la variante macro de la medida F1. El número de las épocas para cada una de las pruebas será determinado por la estrategia de la interrupción temprana con el mismo valor de

[4] https://optuna.org/

Parámetro	Espacio de búsqueda	Valor seleccionado
Learning rate	[1e-5, 2e-5, 1e-4, 1e-6]	2e-5
Weight decay	Flotante entre 0 y 1e-1	0,0243
Batch size	[16, 32]	16
Adamϵ	[1e-8, 1e-7, 1e-9]	1e-9
Warmup steps	Número entero entre 0 y 1000	433

Tabla 3.4: El espacio de búsqueda definido para la segunda aproximación y los parámetros seleccionados después de la optimización

paciencia que en la primera aproximación - 5 épocas.

De las 10 pruebas realizadas, 4 fueron interrumpidas por Optuna por no prometedoras. Todo el procedimiento de la optimización tardó 1 hora y 43 minutos en ejecutarse en la misma tarjeta gráfica NVIDIA A-100 de 40 GB.

La Tabla 3.4 resume el espacio de búsqueda definido para la optimización, así como los parámetros seleccionados.

3.2.3. Tercera aproximación

La tercera aproximación estudiada tiene como objetivo investigar la aplicabilidad de un modelo de lenguaje generativo a la tarea de la codificación clínica según la clasificación PI-RADS v.2.1.

El modelo Text-To-Text Transfer Transformer (T5) es, como lo hemos adelantado en la Sección 2.2.2, un modelo de arquitectura codificador - decodificador que sigue bastante fielmente la propuesta de los autores del modelo transformer original (Vaswani et al., 2017), esquematizada en la Figura 3.6.

Lo que distingue el modelo T5 del transformer original es la eliminación del sesgo de Layer Norm, colocación de la normalización de capa fuera de la ruta residual y el uso del esquema de codificación de posición relativa en vez de señalización sinusoide de la posición (Raffel et al., 2020).

El modelo original T5 fue pre-entrenado en una mezcla de tareas no supervisadas y supervisadas para lo que cada tarea se convirtió en un formato 'seq2seq'. El entrenamiento supervisado se llevó a cabo en tareas proporcionadas por los conjuntos de referencia GLUE (Wang et al., 2018) y SuperGLUE (Wang et al., 2019). Para que el modelo tenga constancia de la tarea a realizar, a cada secuencia de entrada se le añadía un prefijo específico. La Figura 3.7 ilustra el funcionamiento del modelo T5.

Un hecho muy relevante para nuestro estudio es que T5 mostró resultados

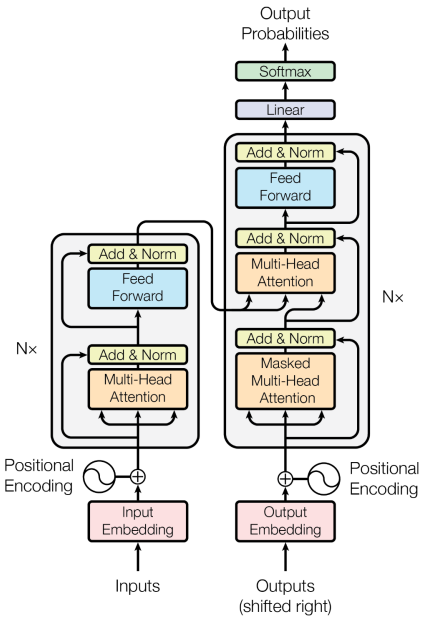


Figura 3.6: Arquitectura original del modelo transformer (Vaswani et al., 2017)

muy prometedores en una amplia gama de tareas, incluido el análisis de sentimientos formulado como una tarea de clasificación binaria 'seq2seq' (Raffel et al., 2020). La versión T5-base de 220 millones de parámetros diseñada para que el codificador y el decodificador sean de un tamaño comparable con el modelo $\text{BERT}_{\text{base}}$, alcanzó un 95.2 de accuracy [5]. Sin embargo, dado que este modelo es monolingüe y fue entrenado con datos en inglés, no es apto para ser utilizado en el presente trabajo.

El éxito de los modelos T5 en tareas de distinta índole alineado con los avances recientes en el ámbito del ajuste de los grandes modelos preentrenados mediante instrucciones formuladas en lenguaje natural desembocó en el desarrollo del procedimiento de ajuste denominado *Flan* (del inglés *finetuning language models*) (Chung et al., 2022). Este procedimiento engloba un total de 1.836 tareas y se demostró que es beneficioso para modelos T5 de todos los tamaños, puesto que mejora su rendimiento en una amplia gama de tareas multilingües. Dado que el modelo T5-base original es monolingüe, para poder hacer comparaciones sobre conjuntos de datos en idiomas distintos al inglés, fue entrenado con el objetivo estándar de mode-

[5] Fracción de predicciones correctas

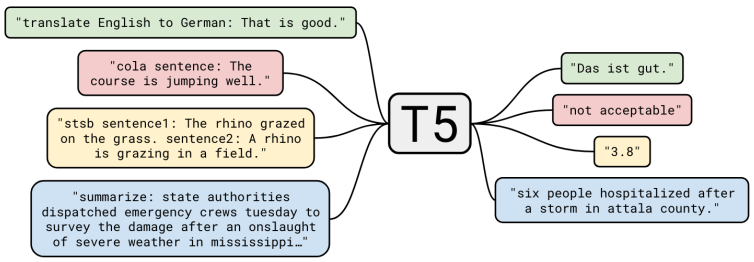


Figura 3.7: El marco 'seq2seq' de la propuesta original del modelo T5 (Raffel et al., 2020)

lado del lenguaje sobre una sub-colección plurilingüe de 100 mil millones de tokens muestreada del Colossal Clean Crawled Corpus (C4) (Raffel et al., 2020).

Seleccionamos para nuestra tercera aproximación la versión Flan-T5-base debido a que su rendimiento en marcos de evaluación como MMLU (de inglés *massive multitask language understanding*) (Hendrycks et al., 2020) mejora 11,5 puntos con respecto a T5-base (Chung et al., 2022).

Al igual que en las primeras dos aproximaciones, nos basamos en las herramientas proporcionadas por la librería `transformers` de HuggingFace (Wolf et al., 2019) para realizar el ajuste fino del modelo elegido. Una vez realizado el pre-procesamiento del corpus según la metodología descrita en la Sección 3.1.1, iniciamos el entrenamiento en una tarjeta gráfica NVIDIA A-100 40 GB con hiperparámetros por defecto que son idénticos a los que se muestran en la Tabla 3.3, menos el valor del número de épocas. Igual que en las aproximaciones anteriores, hacemos uso de la estrategia de interrupción temprana que para el entrenamiento en función de los valores de la macro F1 obtenidos en la evaluación realizada después de cada época. El entrenamiento fue detenido pasadas 8 épocas y duró aproximadamente 17 minutos. En cuanto a la función de pérdida computada para el ajuste de los parámetros del modelo, se calculó la Entropía Cruzada entre las salidas no normalizadas del modelo y los tokens de referencia:

$$l(x, y) = \sum_{n=1}^{N} \frac{1}{\sum_{n=1}^{N} w_{y_n} \cdot 1\{y_n \neq \text{ignore\_index}\}} l_n$$

$$l_n = -w_n \log \frac{\exp(x_n, y_n)}{\sum_{i=1}^{C} \exp(x_{n,c})} \cdot 1\{y_n \neq \text{ignore\_index}\}$$

donde x es la salida del modelo, y es la referencia, w es el peso, C la longitud de la salida, N es el tamaño del batch e `ignore_index` es el parámetro ajustado para ignorar los tokens especiales de separación.

3.3. Automatización de la generación de resúmenes de informes radiológicos

Esta sección ofrece una descripción detallada de los sistemas propuestos para abordar la tarea del resumen automático de informes radiológicos.

3.3.1. Primera aproximación

Nuestra propuesta para la generación automática de resúmenes abstractivos de informes clínicos de exploraciones radiológicas de próstata sigue la línea del enfoque directo empleado, por ejemplo, para resumir automáticamente artículos periodísticos con resultados bastante buenos (Zolotareva, Tashu, y Horváth, 2020). El enfoque directo difiere del indirecto por usar el texto a resumir como dato de entrada, sin aplicarle ningún procedimiento previo (Givchi, Ramezani, y Baraani-Dastjerdi, 2022).

El pre-entrenamiento multitarea con el posterior ajuste en el *framework Flan* habilitan los modelos Flan-T5 a resolver tareas sin haber sido entrenados para ello específicamente a partir de solo unos pocos ejemplos ya resueltos (lo que se denomina en la literatura en inglés *few-show learning*) o incluso sin ningún ejemplo previo (*zero-shot*) (Chung et al., 2022).

En nuestra primera aproximación a la tarea de la generación de resúmenes de informes radiológicos pondremos a prueba el rendimiento del modelo Flan-T5-base sin ningún tipo de ajuste previo. A pesar de que no podamos considerar este experimento totalmente *zero-shot*, debido a la presencia en el conjunto de entrenamiento y de ajuste de ejemplos de tareas de generación de resúmenes (Raffel et al., 2020), no nos consta la presencia entre estos datos de ningún tipo de conjunto de textos clínicos en español.

Con el fin de alcanzar esta meta, realizamos inferencia directamente sobre el conjunto de evaluación. Para asegurarnos de que el modelo "reconozca" el prefijo específico de la tarea, llevaremos a cabo dos pruebas, una adjuntando a los informes el prefijo "summarize: "señalado en la documentación[6] como

[6] https://huggingface.co/google/flan-t5-base/blob/main/config.json

el específico para la tarea de la generación de resúmenes y otra adjuntando la versión traducida de este mismo prefijo: 'resume: '.

Las dos pruebas fueron ejecutadas en la tarjeta gráfica NVIDIA A-100 40 GB y tardaron alrededor de 6 minutos en completarse.

3.3.2. Segunda aproximación

El propósito de la segunda aproximación a la tarea de la generación de resúmenes se propone llevar a cabo un ajuste fino del modelo Flan-T5-base a esta tarea. Para cumplir con ello se seguirá el mismo procedimiento que se describe en la Sección 3.2.3 a excepción de que la métrica usada para evaluar el modelo sobre el conjunto de validación después de cada época de entrenamiento es ROUGE-L. En cuanto al prefijo específico a la tarea, usamos la versión traducida `resume:` para no introducir términos de otros idiomas.

El ajuste del modelo se realizó en el mismo clúster de computación que todos los descritos anteriormente y el tiempo aproximado de ejecución fue de 23 minutos.

Capítulo 4

Evaluación

Este capítulo describe la metodología utilizada para evaluar el sistema propuesto, a la vez que presenta los resultados obtenidos en la evaluación final sobre el conjunto de test.

4.1. Metodología de evaluación

Los sistemas propuestas se evalúan computando métricas de referencia para cada una de las tareas que describiremos en los siguientes apartados. La evaluación se llevará a cabo sobre subconjuntos del corpus que no fueron vistos por los sistemas durante el entrenamiento y no influenciaron su ajuste directa ni indirectamente. Según se ha descrito en la Sección 3.1, el dataset de evaluación final contiene 757 informes clínicos y fue obtenido realizando una división estratificada con el fin de mantener la misma distribución de etiquetas PI-RADS que en el corpus entero.

4.2. Métricas de evaluación

En este apartado se ofrece una descripción de las métricas utilizadas para evaluar el rendimiento del sistema propuesto para cada una de las tareas: la codificación automática y la generación de resúmenes.

4.2.1. Codificación automática

Dado que el enfoque propuesto para abordar la automatización de la codificación clínica aplica la metodología de clasificación multietiqueta de textos,

las métricas elegidas para medir el rendimiento de nuestro sistema son las que se utilizan comúnmente para evaluar este tipo de sistemas de aprendizaje automático: precisión, cobertura (recall) y la medida F1. Estas métricas se computan de la siguiente manera:

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

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

$$\text{Medida F1} = 2 \cdot \frac{P \cdot C}{P + C}$$

donde VP = verdaderos positivos, VN = verdaderos negativos, FP - falsos positivos y FN = falsos negativos.

Las tres métricas tienen variantes micro y macro: la variante micro se calcula promediando las métricas individuales para cada clase o etiqueta en el conjunto de datos, mientras que la variante macro no toma en consideración el número de entradas por clase o etiqueta, ya que se calcula como la media aritmética de las puntuaciones de las clases individuales. Dado que nuestro dataset no está balanceado y queremos guiarnos por el rendimiento del sistema en todas las clases, pondremos especial énfasis en los valores de la variante macro de la medida F1.

4.2.2. Generación de conclusiones

Para la evaluación del sistema de generación de conclusiones nos basaremos, en los valores de las métricas ROUGE (Recall-Oriented Understudy for Gisting Evaluation) que son, como ya lo hemos adelantado en la Sección 2.2.2 muy frecuentemente utilizadas para ofrecer una medida de la calidad de los resúmenes generados por sistemas automáticos.

Las métricas ROUGE-1 y ROUGE-2 miden las coincidencias de unigramas y bigramas, respectivamente, entre el resumen del sistema y el resumen de referencia (Lin, 2004). Sea g un n-grama y R y S - los resúmenes de referencia y del sistema, respectivamente. El valor de la métrica se calcularía según la siguiente fórmula:

$$\text{ROUGE-}n(S) := \frac{\sum_{g \in S} |\{g | g \in S\} \cap \{g | g \in R\}|}{\sum_{g \in R} |\{g | g \in R\}|}$$

Cuando hay más de un par se resúmenes evaluados, se computa la métrica individual y se saca la media de todos los valores obtenidos.

La variante ROUGE-L se basa en estadísticas de la secuencia más larga común (LCS por sus siglas en inglés *Longest Common Subsequence*) que tiene en cuenta la similitud estructural de nivel de secuencias e identifica el n-grama común más largo. En particular, ROUGE-L es la media armónica ponderada (o f-medida) que combina la precisión de la LCS (el porcentaje de la secuencia de hipótesis cubierta por la LCS) y el recall (cobertura) de la LCS (el porcentaje de la secuencia de referencia cubierta por la LCS) (Lin, 2004).

Sean X el resumen de referencia de longitud m e Y el candidato para la evaluación cuya longitud equivale a n. $LCS(X, Y)$ sería, entonces, la secuencia común más larga. La presición y el recall se computan de la siguiente manera:

$$P_{LCS} = \frac{LCS(X, Y)}{m}$$

$$R_{LCS} = \frac{LCS(X, Y)}{n}$$

Sea $\beta = \frac{P_{LCS}}{R_{LCS}}$. La medida ROUGE-L se computa según la siguiente fórmula:

$$\text{ROUGE-L} = 2 \cdot \frac{(1 + \beta^6) \cdot R_{LCS} \cdot P_{LCS}}{R_{LCS} + \beta^2 \cdot P_{LCS}}$$

4.3. Resultados

En esta sección se presentan los resultados de los experimentos planteados en el Capítulo 3 y resumidos en la Tabla 4.1. Para cada una de las aproximaciones a las dos tareas se expondrán los valores de las métricas de rendimiento, que se compararán con el fin de identificar el enfoque óptimo. Posteriormente, se realizará un análisis de errores cometidos por el mejor sistema para identificar sus limitaciones y áreas de perfeccionamiento.

4.3.1. Automatización de la codificación clínica

La Tabla 4.2 resume el rendimiento de los tres sistemas planteados en Secciones 3.2.1, 3.2.2 y 3.2.3.

Tarea	Sistema	Resumen
Codificación clínica	RoBERTa-default	Modelo RoBERTa ajustado con parámetros por defecto
	RoBERTa-opt	Modelo RoBERTa resultante de la optimización de hiperparámetros
	T5-cls	Modelo Flan-T5 ajustado para realizar la codificación clínica en formato seq2seq
Generación de conclusiones	T5-zero-shot-eng	Generación de conclusiones con el modelo Flan-t5 sin ajuste previo y con el prefijo summarize:
	T5-zero-shot-es	Generación de conclusiones con el modelo Flan-t5 sin ajuste previo y con el prefijo resume:
	T5-sum	Modelo Flan-T5 ajustado a la tarea de generación de conclusiones

Tabla 4.1: Resumen de los sistemas a evaluar

Se puede observar que en la tarea de codificación clínica, el modelo RoBERTa pre-entrenado sobre un conjunto de textos de dominios biomédico y clínico y ajustado con parámetros por defecto muestra los mejores resultados según todas las métricas de evaluación. Este hecho sugiere que durante la optimización de hiperparámetros no se haya podido encontrar una combinación mejor que la de empleada para ajustar el modelo RoBERTa-default, por lo que el rendimiento del modelo resultante de la mejor prueba de optimización es inferior (0.9372 contra 0.9486 de medida macro F1).

Los resultados desglosados por clase que se pueden ver en la Tabla 4.3 prueban que el sistema de la primera aproximación es superior que el resultado de la optimización asignando etiquetas de cualquier categoría PI-RADS. Cabe mencionar que el sistema RoBERTa-opt en términos de precisión da peores resultados en todas las etiquetas menos la del riesgo más alto, pero en términos de recall supera ligeramente al sistema con parámetros por defecto en las clases 1 y 2, mientras que los valores para las demás etiquetas son en general más bajos.

Sistema	Micro-avg			Macro-avg		
	Precisión	**Recall**	**Medida F1**	**Precisión**	**Recall**	**Medida F1**
RoBERTa default	0,9733	0,9381	0,9554	0,9658	0,9328	**0,9486**
RoBERTa-opt	0,9516	0,9371	0,9443	0,9478	0,9272	0,9372
T5-cls	0,9541	0,9314	0,9427	0,9407	0,9164	0,9276

Tabla 4.2: Métricas de rendimiento de cada uno de los clasificadores propuestos.

Label	RoBERTa-default			RoBERTa-opt			support
	Precisión	**Recall**	**Medida F1**	**Precisión**	**Recall**	**Medida F1**	
PI-RADS 1	**0,9716**	0,8650	**0,9152**	0,9200	**0,8734**	0,8961	237
PI-RADS 2	**0,9927**	0,9832	**0,9880**	0,9764	**0,9904**	0,9833	417
PI-RADS 3	**0,9379**	**0,9128**	**0,9252**	0,9116	0,8993	0,9054	149
PI-RADS 4	**0,9623**	**0,9387**	**0,9503**	0,9554	0,9202	0,9375	163
PI-RADS 5	0,9643	**0,9643**	**0,9643**	**0,9756**	0,9524	0,9639	84

Tabla 4.3: Comparativa desglosada por clases entre el modelo RoBERTa-default y RoBERTa-opt

Por otro lado, el modelo generativo que emplea un enfoque de clasificación totalmente diferente muestra resultados de una calidad muy cercana al método considerado estado-de-arte en la tarea de clasificación de textos, aunque sea un poco inferior con respecto al mejor sistema (0,021 de diferencia en términos de medida macro F1).

En la Tabla 4.4 que muestra los resultados desglosados por clase se puede ver que rendimiento del modelo T5-cls falla con respecto al RoBERTa-default sobre todo en las clases de riesgo más alto: se nota una bajada de unos 0,0398 puntos de precisión en detección de la etiqueta PI-RADS 3, de 0,0675 puntos de recall a la hora de detectar la etiqueta PI-RADS 4 y de 0,0595 puntos de medida macro F1 en la detección de la clase de riesgo más alto, PI-RADS 5. Dicho en otras palabras, el modelo T5-cls da más falsos positivos para la clase PI-RADS 3, más falsos negativos para la clase PI-RADS 4 y comete más errores de los dos tipos para PI-RADS 5.

4.3.2. Automatización de la generación de conclusiones

La Tabla 4.5 muestra los resultados de evaluación en el conjunto de test de dos sistemas presentados en las Secciones 3.3.1 y 3.3.2.

Vemos que el ajuste fino mejora radicalmente el rendimiento del sistema, que suma 0,4787; 0,5125 y 0,5152 puntos en valores de las métricas

Label	RoBERTa-default			T5-cls			support
	Precisión	Recall	Medida F1	Precisión	Recall	Medida F1	
PI-RADS 1	**0,9716**	0,8650	**0,9152**	0,9450	**0,8692**	0,9055	237
PI-RADS 2	**0,9927**	0,9832	**0,9880**	0,9833	**0,9904**	0,9869	417
PI-RADS 3	**0,9379**	0,9128	**0,9252**	0,8981	**0,9463**	0,9216	149
PI-RADS 4	0,9623	**0,9387**	**0,9503**	**0,9726**	0,8712	0,9191	163
PI-RADS 5	**0,9643**	**0,9643**	**0,9643**	0,9048	0,9048	0,9048	84

Tabla 4.4: Comparativa desglosada por clases entre el modelo RoBERTa-default y T5-cls

Sistema	ROUGE-1	ROUGE-2	ROUGE-L
T5-zero-shot-en	0,2871	0,1822	0,2393
T5-zero-shot-es	0,2866	0,1820	0,2392
T5-sum	0,7658	0,6947	**0,7545**

Tabla 4.5: Resultados de evaluación final para los generadores de conclusiones

ROUGE-1, ROUGE-2 y ROUGE-L respectivamente si comparamos la la inferencia con prefijo `summarize:` con los resultados del modelo ajustado. Este hecho prueba que el modelo Flan-T5 no tiene la capacidad suficiente de generalización sobre textos de dominio clínico escritos en español que, como hemos mencionado en la Sección 2.2.2, rinde bien en tareas de generación de resúmenes de textos de dominio más general, como pueden ser los textos periodísticos.

Por otra parte, es notable que el cambio del prefijo específico a la tarea no influye en los resultados de la inferencia, que siguen siendo muy bajos.

4.4. Análisis de errores

En el dominio de medicina, donde los errores pueden tener consecuencias significativas, resulta crítico comprender las deficiencias del sistema. En esta sección, con el fin de evaluar las limitaciones del mejor sistema propuesto e identificar áreas de mejora, realizaremos un análisis de los errores cometidos durante la evaluación final.

4.4.1. Automatización de la codificación clínica

El sistema RoBERTa-default clasificó incorrectamente un total de 75 informes del conjunto de evaluación (9,9 % del total), de los cuales solamente 13 predicciones fueron totalmente incorrectas. Es decir, solo en 13 casos (1,7 % del total) el sistema no fue capaz de acertar ninguna de las etiquetas PI-RADS asignadas.

Los informes etiquetados con un número mayor de etiquetas suelen ser más complejos semánticamente al describir varias lesiones. Según se puede observar en la Figura 4.1, la mayoría de los errores (52) se cometieron en textos etiquetados originariamente con dos o tres clases.

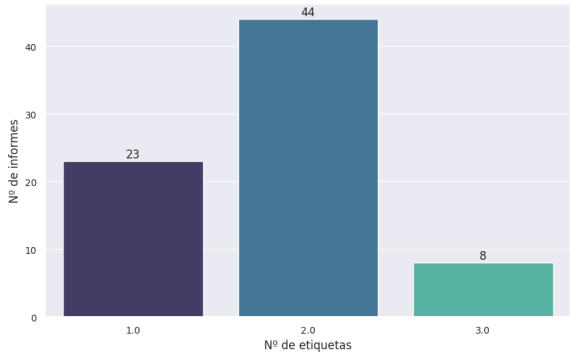
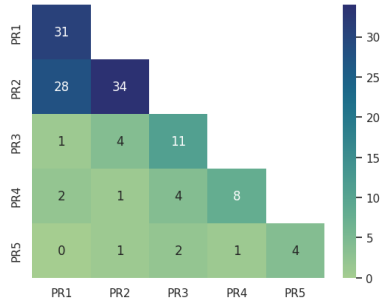


Figura 4.1: Número de informes con varias etiquetas entre los que fueron clasificados erróneamente

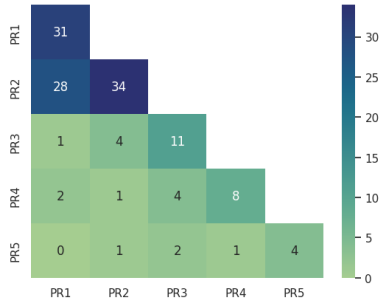
Entre las posibles razones de estos tipos de errores en etiquetado automático de conjuntos de datos multietiqueta se podría mencionar la falta de ejemplos adecuados en el conjunto de datos de entrenamiento: si el conjunto de datos de entrenamiento no contiene suficientes ejemplos que reflejen la diversidad y complejidad de los textos con múltiples etiquetas, el sistema puede tener dificultades para aprender patrones para su clasificación precisa. Sin embargo, las matrices de co-ocurrencia de etiquetas en los textos clasificados erróneamente mostradas en las Figuras 4.2a y 4.2b no difieren sustancialmente de la matriz de co-ocurrencia de códigos PI-RADS en el conjunto de entrenamiento que se puede ver en la Figura 4.3a.

Es por ello que podemos apelar a la complejidad semántica inherente de los textos procesados más que a una deficiencia del conjunto de datos. Conviene notar que la longitud media de los informes incorrectamente clasificados es inferior a la media de la totalidad del dataset (120,8 tokens

con desviación estándar de 55,8 contra 137,7 tokens con 38,3 de desviación estándar de 38,3). Los textos más cortos ofrecen la información compleja en una forma más sintetizada y la ausencia de mayor detalle puede dificultar la clasificación.



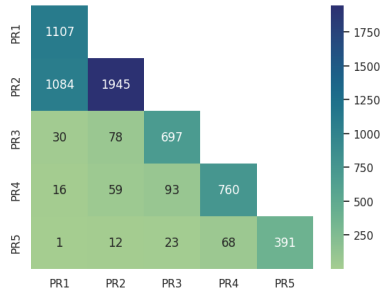
(a) Co-ocurrencia de etiquetas originales en informes pertenecientes a dos clases



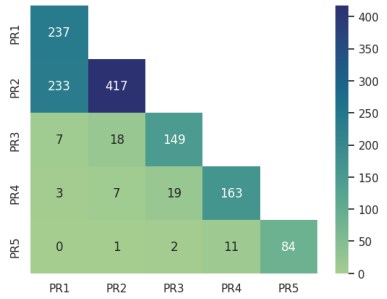
(b) Co-ocurrencia de etiquetas originales en informes pertenecientes a tres clases

Figura 4.2: Matrices de co-currencia de etiquetas originales en los informes erróneamente clasificados

Como se pudo ver en las Tablas 4.3 y 4.4, la única métrica cuyo valor baja del 0,9 es la cobertura (recall) para la clase PI-RADS 1, lo cual indica una tasa más elevada de los falsos negativos con respecto a los verdaderos positivos. Esto se puede deber a que la etiqueta de menor riesgo sea la que más veces aparece conjuntamente con otras, especialmente con la PI-RADS 2, según se puede ver en la matriz de co-ocurrencia de la Figura 4.3.



(a) Co-ocurrencia de etiquetas en el conjunto de entrenamiento



(b) Co-ocurrencia de etiquetas en el conjunto de evaluación

Figura 4.3: Matrices de co-currencia de etiquetas originales en los conjuntos de entrenamiento y evaluación

En total se dieron 38 falsos negativos en la etiqueta PI-RADS 1. No-

toriamente, en 32 casos de estos 38 el modelo fue capaz de detectar una etiqueta de riesgo más bajo del conjunto de las originariamente asignadas. La Tabla 4.6 ilustra este tipo de error que se puede considerar clínicamente irrelevante para el caso de un sistema de alerta, dado que PI-RADS v.2.1 es una clasificación que desarrolla categorías de monitorización y asesoramiento según la escala de riesgo de que el paciente padezca cáncer de próstata[1]. De este modo, a pesar de que un informe radiológico registre dos lesiones con nivel de riesgo cancerígeno diferente, la probabilidad total de que el paciente tenga cáncer corresponde a la de la etiqueta de mayor riesgo.

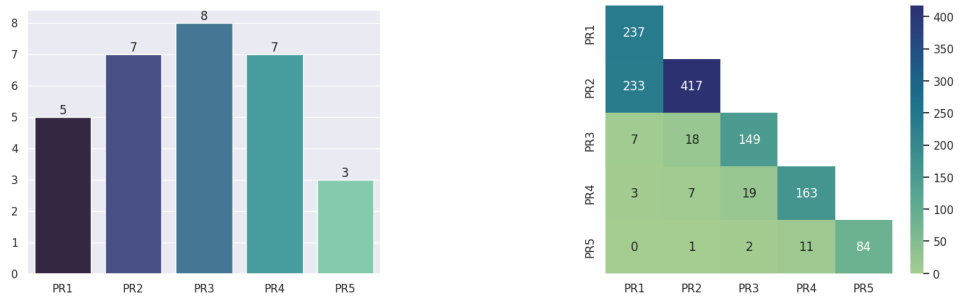
Anotación original	**Predicción**
pr1, pr2	pr2

Tabla 4.6: Ejemplo de un error irrelevante clínicamente

Entre todos los errores, 15 casos corresponden a errores clínicamente significativos de subestimación del grupo de riesgo del paciente. Entre ellos, a seis informes se les asingnó originariamente una etiqueta PI-RADS, tres fueron clasificados con dos códigos y otros seis con tres códigos. Entre las etiquetas confundidas figuran tanto las de alto riesgo, como las de bajo, como se puede ver en la Figura 4.4a. Es notable que en el caso de subestimación del nivel de riesgo, sí que podamos observar, en la Figura 4.4b que los informes presentan co-ocurrencias de etiquetas inusuales con respecto a los del conjunto de entrenamiento: hay varios informes etiquetados con códigos PI-RADS 3 y 5, 2 y 4, 1 y 4 que, al describir lesiones con niveles de riesgo tan dispares pueden resultar ambiguos para el sistema.

En cuanto a la sobreestimación, es decir, la predicción de una etiqueta correspondiente a un nivel de riesgo más alto del verdadero, se dieron 7 errores de este tipo. A cuatro de estos informes originariamente se les fue asignada una etiqueta y los tres restantes contaban con dos códigos PI-RADS anotados. Cinco de las siete veces el clasificador asignaba una etiqueta de riesgo alto (PI-RADS 4) y muy alto (PI-RADS 5). Los informes cuya clasificación automática resultó en una sobre-estimación de riesgo son más largos de la media del dataset, alcanzando en el caso del más extenso 219 tokens y 104 en el caso del más corto, con una media de 151 tokens por

[1]Fuente: American College of Radiology: PI-RADS v.2.1: https://www.acr.org/-/media/ACR/Files/RADS/Pi-RADS/PIRADS-V2-1.pdf?la=en



(a) Distribución de etiquetas (b) Co-ocurrencia de etiquetas

Figura 4.4: Estadísticas sobre etiquetas originales en el conjunto de informes en cuyo etiquetado se cometieron errores de subestimación de nivel de riesgo

informe, contra 137,7 tokens que es la media de la totalidad del dataset.

4.4.2. Automatización de la generación de conclusiones

A diferencia de la codificación clínica, la generación de conclusiones es una tarea más difícil de evaluar y plantea un desafío importante en cuanto al análisis de errores dada la complejidad del mismo término 'error' en este contexto. La inexistencia de una sola salida correcta y la dependencia de tales factores pragmáticos como la intención con la que se genera, la experiencia a la cualificación del lector (Lloret, Plaza, y Aker, 2018) así como criterios como los que mencionamos en la Sección 2.1: brevedad y transmisión de información relevante relativa a las observaciones, inferencias y conclusiones clave.

La evaluación realizada mediante el cálculo del valor de la métrica ROUGE está basada en la premisa de que la conclusión de referencia cumple con todos los criterios, por lo que cuanto más parecida a esta se genere la conclusión, mejor calidad tendrá.

En esta sección realizaremos un análisis de los resultados del mejor sistema de generación de resúmenes de todos los planteados, T5-sum que dio buenos resultados en términos de las tres métricas según vimos en la Sección 4.3. Estas métricas se computan, como hemos mencionado anteriormente, haciendo media de los resultados por cada par de conclusión original y la generada. La Figura 4.5 muestra el histograma de los valores de la métrica ROUGE-L para cada una de las conclusiones generadas. En esta gráfica podemos ver que una parte importante de las conclusiones generadas ha ano-

tado un valor de cerca de 0,9 de ROUGE-L y que la mayoría (en concreto, 526 casos) supera 0,7.

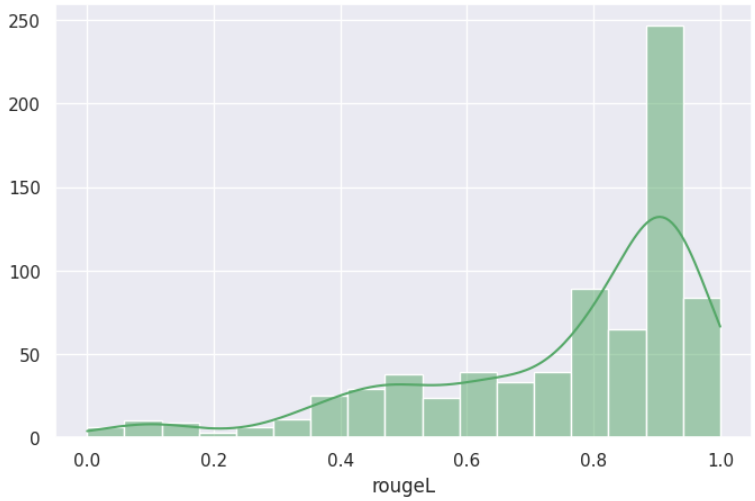


Figura 4.5: Histograma de los valores de la métrica ROUGE-L para cada una de las conclusiones generadas

En cuanto a la longitud de las conclusiones generadas, en la Figura 4.6 se puede ver una comparativa entre las conclusiones de referencia y las generadas por el modelo T5-sum. Podemos ver que la longitud de las salidas del generador de conclusiones no presenta disparidades con respecto a los originales, a excepción de una bajada de número de conclusiones largas (más de 40 tokens). Consecuentemente, se nota una bajada de la longitud media de las conclusiones: el sistema devuelve un promedio de 20,3 tokens, mientras que la extensión media de las conclusiones escritas por expertos radiólogos es de 24,7.

Si consideramos que una conclusión deficiente es la que obtiene menos de 0,5 de ROUGE-L, el mejor sistema propuesto ha generado 122 conclusiones de poca calidad. La longitud media de los 122 informes en cuestión es mayor que el promedio de longitud de todo el conjunto: 137,8 contra 111,8. Este hecho nos puede llevar a la conclusión de que una de las limitaciones de nuestro sistema es el resumen de textos largos: las conclusiones evaluadas con un bajo valor de ROUGE-L son, en general, más cortas que sus referentes: la longitud media de las conclusiones en este subconjunto es de 37 tokens, mientras entre las generadas el promedio es de 21,6 tokens.

Según hemos mencionado en la Sección 3.1, muchos informes del con-

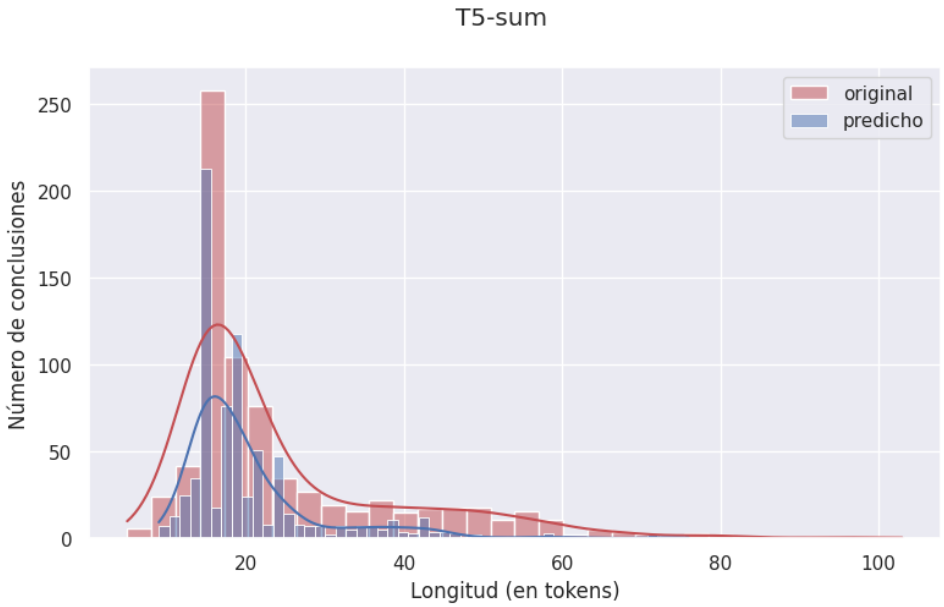


Figura 4.6: Histograma de longitud (número de tokens) de conclusiones de referencia y las generadas por el sistema

junto de datos contienen menciones explícitas de códigos PI-RADS v.2.1 asignados a los informes. Para la tarea de generación de conclusiones no hemos eliminado estas menciones, porque su presencia no interfería en la imparcialidad del entrenamiento, como ocurría en el caso de la codificación clínica. Podemos aprovechar esta característica para arrojar más luz sobre la calidad de las salidas de nuestro mejor sistema. Según se puede ver en la gráfica de la Figura 4.7 que muestra una comparativa del número de las conclusiones con menciones explícitas de un código PI-RADS v.2.1, en los textos generados por el modelo las menciones de códigos PI-RADS aparecen con más frecuencia.

La generación de resúmenes abstractivos puede llevar a errores relacionados con la exactitud fáctica del texto: a pesar de obtener una puntuación ROUGE alta, una conclusión generada puede contener una inconsistencia importante, lo cual es un problema recurrente de la aplicación de modelos generativos de lenguaje pre-entrenados (Cao et al., 2020). Evaluar automáticamente la exactitud fáctica de un texto generado es una tarea pendiente de resolución. En nuestro caso, vamos a comprobar que todos los códigos PI-RADS generados por el modelo T5-sum dentro de las conclusiones son correctos. Para ello, estos se extraerán mediante *matching* y se compararán

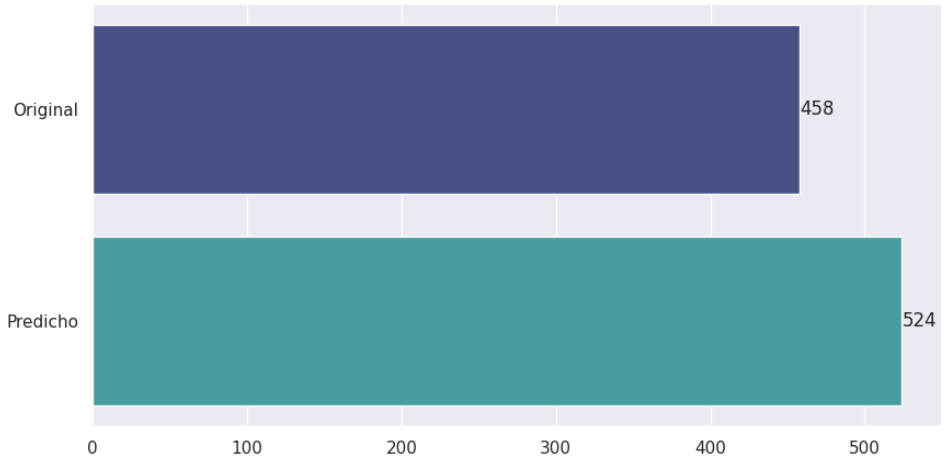


Figura 4.7: Número de conclusiones con menciones explícitas de códigos PI-RADS en el conjunto de referencia y el generado por el modelo T5-sum

mediante cálculo de la métrica de precisión explicada en la Sección 4.2 con las etiquetas asignadas por anotadores expertos. Elegimos esta métrica para asegurarnos de que el modelo no genera códigos aleatoriamente y no da lugar a ningún falso positivo.

Esta evaluación reveló que todos los códigos detectados en las conclusiones generadas automáticamente se corresponden con las anotaciones de los expertos, puesto que para todas las etiquetas el valor de precisión es igual a 1,0.

Además, conviene mencionar que en 92 casos el modelo generó una conclusión que mencionaba explícitamente el código PI-RADS correcto cuando la conclusión referente no lo contenía. Esto señala que el ajuste fino del sistema ha permitido producir un modelo que es capaz de otorgar importancia debida a las menciones de códigos PI-RADS y añadirlas en las síntesis producidas.

Capítulo 5

Discusión

Este capítulo analiza y discute en profundidad los resultados obtenidos en la evaluación presentada en el capítulo anterior. Más específicamente, se investigan las razones por las que aproximaciones como la optimización de hiperparámetros y la generación de conclusiones *zero-shot* no han llevado a buenos resultados. Además, se discuten las limitaciones de la evaluación de la generación de resúmenes con las métricas utilizadas en el presente estudio. Por último se discute la capacidad de cómputo requerida para llevar a cabo toda la experimentación presentada en el este trabajo.

5.1. Limitaciones de la optimización de hiperparámetros realizada

La evaluación del sistema de codificación clínica obtenido después de 10 pruebas de optimización de hiperparámetros basada en el *framework* Optuna (Akiba et al., 2019) ha demostrado que dicha optimización no ha permitido alcanzar el objetivo de encontrar una combinación de learning rate, batch size, weight decay, épsilon del optimizador Adam y el número de pasos de calentamiento mejor que la especificada por defecto.

La razón de esto probablemente se halla en el pequeño número de pruebas realizadas y en la forma en la que fue definido el espacio de búsqueda.

En primer lugar, la realización de 10 pruebas constituye aproximadamente la décima parte de todas las que se realizarían si se aplicara el muestreador Grid Search - una técnica muy usada para la selección de parámetros óptimos que consiste en probar todas las combinaciones posibles dentro de un

espacio de búsqueda (Sun et al., 2021). Si sustituyéramos el muestreo dentro de un rango de números por una lista de dos valores para parámetros del número de los pasos de calentamiento y el weight decay, y dejáramos el resto tal y como se define en la Tabla 3.4, tendríamos que realizar 96 pruebas hasta agotar todas las combinaciones posibles. Por otro lado, las particularidades del algoritmo de muestreo seleccionado explicadas en la Sección 3.2.2 hacen que una optimización de este tipo se beneficie de la subida del número de pruebas realizadas, dado que el muestreador ajusta una GMM al conjunto de valores de un parámetro asociados con los mejores valores objetivos obtenidos previamente. Conviene notar que con duración de de 14 minutos por prueba en la tarjeta gráfica que se utilizó para estos experimentos, una optimización de 96 pruebas tardaría aproximadamente 22,4 horas en ejecutarse, lo cual conllevaría un consumo energético mucho mayor y constituye una de las razones por las que este experimento no se llevó a acabo en marco del presente trabajo.

En segundo lugar, la reducción del espacio de búsqueda podría mejorar y agilizar la optimización de hiperparámetros, pero acarrearía la necesidad de una selección más meticulosa de valores a probar.

Es notable que después de las 10 pruebas han quedado sin utilizarse valores del espacio de búsqueda para el learning rate ($1e-6$) y épsilon del optimizador ($1e-8$), como se puede inferir de la Tabla 5.1 que resume todas las combinaciones probadas durante la optimización. También son destacables la recurrencia mayor al tamaño de lote igual a 16 (6 pruebas contra 4 con el tamaño igual a 32), el hecho de que la mayoría de los valores seleccionados para el weight decay sean menores de 0,1 y que los del número de pasos de calentamiento sean, en su mayoría, mayores de 500.

La técnica de muestreo seleccionada no nos ha llevado a una mejora de los resultados, además, impidió pruebas de combinaciones de parámetros más diversas, lo cual ha sido contraproducente, dado que de las 10 pruebas solo 6 no fueron interrumpidas por ser consideradas no prometedoras y la configuración óptima resultó ser que que se seleccionó durante la tercera prueba.

learning rate	batch size	weight decay	Adamϵ	warm-up steps
2e-5	32	0,2465	1e-7	615
1e-4	16	0,0016	1e-9	384
2e-5	16	0,0243	1e-9	433
1e-5	16	0,0034	1e-9	256
1e-4	32	0,0328	1e-9	765
1e-5	16	0,0413	1e-9	574
1e-4	16	0,0679	1e-7	930
2e-5	32	0,054	1e-7	924
1e-5	32	0,0679	1e-9	30
1e-4	16	0,099	1e-9	533

Tabla 5.1: Hiperparámetros seleccionados por el muestreador de Optuna para cada una de las pruebas de optimización realizadas

5.2. Escasa calidad de generación *zero-shot* de conclusiones

En la Sección 4.3.2 hemos señalado que la primera aproximación, es decir, la prueba de las capacidades del modelo Flan-T5-base de producir resúmenes de informes radiológicos sin haber sido ajustado específicamente a la realización de esta tarea, no dio buenos resultados. La Tabla 5.2 muestra un ejemplo de las conclusiones producidas por cada uno de los sistemas. Ya en este ejemplo se ve claramente que el sistema sin ajustar produce conclusiones muy extensas, con erratas ('mltiples', 'nodules') e introducción de términos en inglés ('CLINICAL INFORMATION'). Además, este intento de resumen, coincide casi por completo con una parte del informe truncada.

El modelo Flan-T5-base, ajustado mediante un procedimiento multitarea, es un modelo plurilingüe y el idioma español es el segundo más representado dentro del conjunto de las tareas del marco *Flan* (Chung et al., 2022), de modo que el idioma español no constituye un punto débil del modelo, al igual que no lo es la tarea misma de generación de resúmenes, porque forma parte importante de los conjuntos de datos utilizados para su creación.

El hecho de que Flan-T5-base no muestre buenos resultados en la tarea planteada en este trabajo deriva de las particularidades de la variante lingüística utilizada en los informes, así como de los requerimientos específicos que ponemos a la hora de comparar la salida del modelo con la conclusión escrita por un radiólogo.

Original	**T5-zero-shot**	**T5-sum**
HPB. No se observan áreas claramente sospechosas de carcinoma clínicamente significativo demostrables por esta técnica.	CLINICAL INFORMACION elevación de PSA HALLAZGOS: Volumen prostático: 47 cc. Hipertrofia de la zona de transición que comprime y adelgaza la zona periférica. No se observa áreas claramente sospechosas de carcinoma clnicamente significada en secuenciaT2, difusión ni contrasted. Mltiples imágenes nodules en zona transicional compatibles en primer término con nódulos hiperplásicos (PIRADS 1 y 2) Cápsula y grasa periprostática sin alteraciones evi	HPB. No se observan áreas claramente sospechosas de carcinoma clnicamente significativo demostrables por esta técnica (PIRADS 1 y 2).

Tabla 5.2: Comparación de salidas del modelo sin ajustar (T5-zero-shot) y el modelo ajustado (T5-sum)

Uno de los rasgos más evidentes que distingue las salidas del modelo no ajustado de las conclusiones de referencia es la longitud. La gráfica de la Figura 5.1 permite comparar las conclusiones generadas con las originales en función de este rasgo. Las salidas del modelo en este caso son mucho más largas que las conclusiones generadas, de lo cual podemos concluir que las tareas de generación de resúmenes incluidas en el procedimiento de ajuste del modelo en el marco *Flan* no comparten requerimientos con la tarea que planteamos nosotros en el presente trabajo.

Por último, conviene mencionar que el éxito del ajuste fino de este mismo modelo a la tarea de la generación de conclusiones muestra que el sistema tiene un gran potencial de adaptación al dominio y un vocabulario lo suficientemente amplio para llegar a producir resúmenes semejantes a los que fueron escritos por expertos humanos.

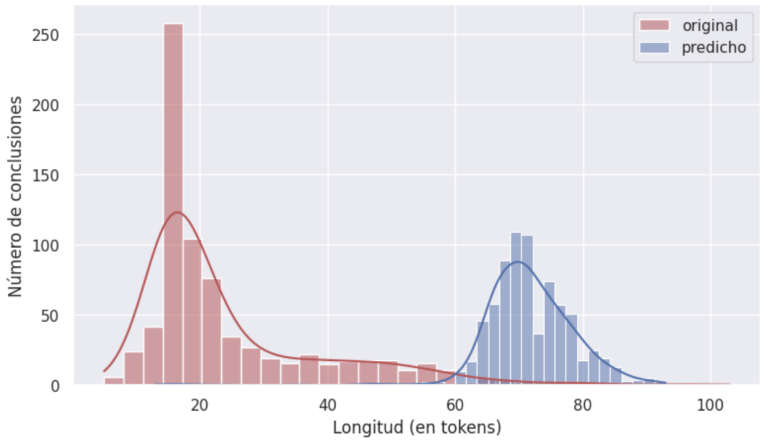


Figura 5.1: Histograma de longitud de las salidas del sistema T5-zero-shot comparada con el histograma de longitud de conclusiones originales

5.3. Limitaciones de la evaluación de resúmenes generados

Como hemos adelantado en la Sección 4.4.2, la generación automática de resúmenes abstractivos, al involucrar la generación de lenguaje plantea retos en cuanto a su evaluación. En esta sección discutiremos los inconvenientes del enfoque adoptado.

Las métricas ROUGE estiman la cobertura de conceptos apropiados en un resumen generado automáticamente en función de la cobertura (recall) de coincidencias de unigramas y bigramas para las variantes ROUGE-1 y ROUGE-2 y la medida F1 estimada en función de la secuencia común más larga para la variante ROUGE-L.

A pesar de ser una de las formas de evaluación de sistemas de generación de resúmenes más utilizadas, ROUGE es criticado por usar únicamente técnicas de *matching*, teniendo en cuenta solamente la forma de los términos y no su significado, de modo que la sustitución de una palabra por un sinónimo bajaría la puntuación. Es por ello que el ajuste fino de un modelo de generación de conclusiones con una métrica de referencia como ROUGE puede restringir la variedad de vocabulario utilizada en la generación de resúmenes y llevar a producción de salidas muy poco diversificadas.

El hecho de realizar *matching* y contabilizar las coincidencias entre n-gramas también subyace a que ROUGE se critique por la imposibilidad

de alcanzar un valor de 1.0, a no ser que se usen exactamente los mismos términos (Schluter, 2017).

Por otra parte, este enfoque se critica por el hecho de que un resumen que unos anotadores humanos consideren de escasa calidad pueda alcanzar un valor alto de ROUGE, debido al hecho de que esta metodología no evalúa la legibilidad del texto producido ni su exactitud fáctica (Lloret, Plaza, y Aker, 2018). Esto último hace necesario que un sistema de generación de conclusiones como el planteado en el presente trabajo pase por una evaluación exhaustiva por parte de personal cualificado capaz de detectar errores conceptuales y medir la calidad del uso de la terminología específica del dominio de la radiología oncológica.

5.4. Recursos computacionales requeridos para los experimentos

Para llevar a cabo experimentos con grandes modelos de lenguaje pre-entrenados se requiere una infraestructura de recursos computacionales sólida y adecuada. Los modelos utilizados en el presente estudio, RoBERTa-clinical y Flan-T5-base, ocupan aproximadamente 4,5 y 8,25 gigabytes de memoria, respectivamente. Esta memoria la ocupan, en su mayoría, los pesos del modelo, los del optimizador y los gradientes.

El entrenamiento eficiente de un modelo de aprendizaje profundo requiere un procesador gráfico que además de tener capacidad para cargar el modelo, necesita tener memoria suficiente para realizar las tres principales operaciones que tienen lugar durante entrenamiento de un modelo de arquitectura transformer: contracciones de tensores, normalizaciones estadísticas y operaciones elemento por elemento (ing. *element-wise operations*) (Ivanov et al., 2021).

En nuestro caso, disponíamos de una tarjeta gráfica potente, NVIDIA A100 40GB, que se puede considerar un hardware especializado que está lejos de las máquinas a escala de consumidor, como podría ser NVIDIA GeForce RTX 4080 de 16 GB.

Durante su ajuste (y durante cada una de las pruebas de la optimización de hiperparámetros), el modelo RoBERTa-clinical ocupaba con la totalidad de sus procesos aproximadamente el 28 % de la memoria de la GPU, mientras que el modelo Flan-T5-base, al constar de dos bloques Transformer,

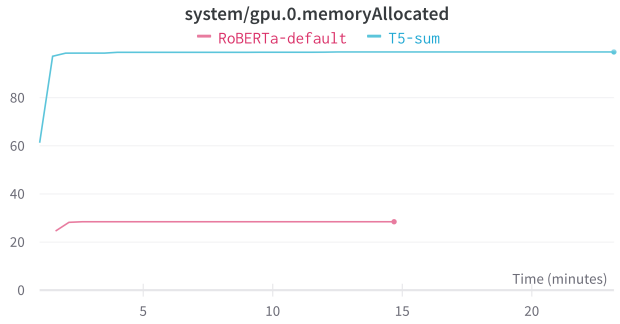


Figura 5.2: Memoria del procesador gráfico ocupada durante el proceso de ajuste de los sistemas T5-sum y RoBERTa-default.

codificador y decodificador, ocupaba casi toda la memoria disponible, como se puede ver en la gráfica de la Figura 5.2.

El hecho de que el modelo T5 para su ajuste necesite un procesador más potente se puede considerar un argumento más a favor del uso, para la tarea de codificación clínica, de un modelo más ligero, eficaz y eficiente como resultó ser RoBERTa-clinical.

Capítulo 6

Conclusiones y trabajo futuro

En este capítulo de conclusiones, se presentan los resultados y hallazgos obtenidos a lo largo de este estudio sobre la automatización de la codificación clínica según PI-RADS v.2.1 y generación de conclusiones de informes radiológicos. Estas conclusiones pretenden aportar una visión crítica y reflexiva sobre el trabajo realizado, así como proponer algunas líneas de trabajo futuro para mejorar el estado actual de la cuestión.

6.1. Conclusiones

En este trabajo se ha abordado la tarea de automatización de la codificación y resumen de informes de exploraciones radiológicas de próstata. Con el objetivo de contribuir a la mejora de la calidad de transmisión de información radiológica y su aprovechamiento tanto con metodologías computacionales como por un médico humano, se han propuesto soluciones fundamentadas en los avances recientes del estado del arte investigados y descritos en la Sección 2.2.

La indagación realizada sobre el trabajo previo en los campos de investigación sobre la codificación clínica y la generación de resúmenes ha permitido llegar a la conclusión de que el PLN aplicado al dominio de la radiología es un campo que se conformó recientemente y que está llamando cada vez más atención hoy en día. Sin embargo, la mayoría de los avances descritos no se enfocaban en un idioma distinto al inglés, lo cual subraya

la relevancia de la presente contribución que se centra en procesar informes radiológicos escritos en español.

El conjunto de datos que se ha manejado en el presente estudio es una colección de informes de exploraciones radiológicas de próstata realizadas entre abril de 2019 y junio de 2022 en HTMédica, una clínica radiológica privada. Dicho conjunto es bastante variado en cuando a las exploraciones documentadas, puesto que describe no solo RMs multiparamétricos de próstata, sino también exploraciones de vísceras pélvicas, zonas cervical y lumbar, región anal, pecho y todo el cuerpo.

Para abordar los problemas planteados se han propuesto sistemas de diferentes arquitecturas. Para la codificación clínica de informes con códigos PI-RADS v.2.1 se compararon dos metodologías totalmente diferentes. Por una parte, se ha ajustado un modelo transformer pre-entrenado sobre una combinación de textos médicos y clínicos cuya arquitectura contiene solo un bloque codificador (experimento RoBERTa-default). Además, se ha realizado un experimento con el objetivo de evaluar la viabilidad de realización de una optimización de hiperparámentros de este sistema para maximizar su rendimiento (experimento RoBERTa-opt). Por otra parte, se ha implementado un clasificador basado en la metodología 'seq2seq' para lo cual fue utilizado un modelo de lenguaje generativo (experimento T5-sum).

De las tres aproximaciones planteadas para la realización de clasificación de informes en función del riesgo cancerígeno de los hallazgos descritos según los códigos PI-RADS v.2.1, el ajuste de un modelo codificador pre-entrenado a la tarea de clasificación de texto multietiqueta ha dado mejores resultados: el sistema RoBERTa-default llegó a anotar un 0,9554 de la medida macro F1 sobre el conjunto de evaluación. Es remarcable el hecho de que de 757 informes que conformaban en conjunto de datos de la evaluación final, el sistema clasificó de forma errónea solo 75 textos. El análisis de estos errores realizado en la Sección 4.4 ha permitido concluir que no todos estos errores son clínicamente relevantes para un sistema de alerta de riesgo, dado que en 53 casos de clasificación errónea de los 75 la etiqueta del riesgo más alto fue predicha correctamente, lo cual no obstaculiza una monitorización correcta del paciente, pues refleja fielmente el nivel de riesgo de padecer cáncer que le corresponde.

En cuanto a la optimización de hiperparámetros planteada en la Sección 3.2.2, a pesar de que el modelo RoBERTa-opt haya conseguido unos

resultados comparables con el sistema de referencia ajustado con parámetros por defecto, no se ha conseguido encontrar una combinación de hiperparámentros óptima, por lo que el sistema RoBERTa-opt da resultados ligeramente inferiores en la tarea de la asignación de códigos PI-RADS: 0,9372 contra 0,9484. Esto puede deberse a que se han realizado menos pruebas de las requeridas y también a la configuración del espacio de búsqueda.

La aproximación 'seq2seq' a la tarea de clasificación de textos según PI-RADS v.2.1 resultó ser comparable en términos de calidad de sus salidas con el empleo de un modelo que conste solo de un codificador. En este enfoque se ajustó un modelo que tomaba como entrada el texto del informe y producía la salida en formato de texto que luego se evaluaba con la misma metodología de evaluación que las dos otras aproximaciones. A pesar de que este enfoque no ha conseguido superar la calidad de las predicciones hechas con el modelo RoBERTa-default: 0,9276 puntos de macro F1 contra los 0,9486 anotados por RoBERTa-default, el experimento ha permitido demostrar la aplicabilidad de una arquitectura generativa a tareas para las que mayoritariamente se usan sistemas que constan de un bloque codificador.

La tarea de generación de conclusiones de informes radiológicos fue abordada mediante la aplicación del mismo modelo generativo que se ajustó para la tarea de clasificación - Flan-T5-base. Se realizaron dos experimentos: uno orientado a la evaluación de las capacidades del modelo Flan-T5 de resumir textos de manera *zero-shot*, es decir, sin ningún ajuste previo y la segunda tenía como objetivo evaluar la capacidad que tiene el mismo sistema de adaptarse al dominio clínico y el formato de las conclusiones contenidas en el conjunto de datos.

El modelo Flan-T5-base no ha mostrado buenos resultados (0,2393 de ROUGE-L) en la aproximación *zero-shot* a pesar de que sus conjuntos de entrenamiento y ajuste multitarea con la metodología *Flan* contenían un número considerable de tareas de generación de resúmenes y textos en español, que es el segundo idioma más representado en el conjunto de sus datos de ajuste. Es por ello que se podría llegar a la conclusión de que las dificultades residen en la variedad lingüística utilizada en los informes con su sintaxis y terminología muy específicos del campo clínico.

A pesar de no tener capacidad de generar conclusiones sin ajuste previo, el modelo Flan-T5-base ha mostrado un buen potencial de adaptación al dominio, marcando valores altos de todas las métricas de evaluación (0,7545

de ROUGE-L) después de haber sido sometido a un ajuste fino. En cuanto a las limitaciones de este sistema, conviene mencionar que el análisis de errores realizado ha permitido determinar que la generación de conclusiones de textos largos es un reto que conlleva una bajada de calidad de las conclusiones predichas. Además, se han discutido las limitaciones que impone el hecho de usar la métrica ROUGE como el valor de referencia en el ajuste fino y como la métrica de evaluación final. Estas limitaciones, abordadas en la Sección 5.3 se desprenden, en su mayoría, del hecho de que ROUGE se basa en el *matching* de n-gramas y no tiene en cuenta la similitud semántica entre las conclusiones generadas y las de referencia.

6.2. Trabajo futuro

Considerando la tarea abordada en relación con el estado del arte y los resultados obtenidos, es posible identificar varias áreas de investigación futura para ampliar las soluciones propuestas en este documento.

En primer lugar, se precisa una investigación más profunda en el área de la optimización de hiperparámetros, ya que una experimentación más exhaustiva y con un mayor número de pruebas y configuraciones del espacio de búsqueda permitirá no solo mejorar el rendimiento del sistema, sino que también posibilitará la definición de pautas para una realización de una búsqueda de hiperparámetros óptimos más eficaz y eficiente.

El enfoque actual se centra en el análisis de informes radiológicos en formato de texto libre, pero existe un potencial sin explotar en la integración de datos multimodales. La combinación de imágenes médicas con datos de texto podría enriquecer la generación de conclusiones y facilitar una mejor interpretación y extracción de información relevante, puesto que la imágen médica es la fuente primaria de la información ofrecida por un radiólogo en formato de una narrativa. Explorar cómo incorporar y fusionar datos de diferentes modalidades podría ser un camino prometedor para las futuras investigaciones.

La interoperabilidad entre diferentes sistemas de información médica sigue siendo un desafío importante en el ámbito de la codificación clínica y la generación de informes radiológicos. Trabajar en estándares y protocolos comunes para compartir datos entre sistemas, como la mejora de estándares de codificación unificados y la implementación de interfaces estandarizadas,

puede facilitar la integración y el intercambio de información entre diferentes instituciones y sistemas de salud. Investigar sobre la mejor manera de integrar sistemas como los que planteamos en el presente trabajo en un sistema de información de una clínica radiológica es un paso importante que se tiene que dar en el desarrollo futuro del proyecto. También sería beneficioso evaluar los enfoques propuestos sobre los datos de otras exploraciones o sobre datos provenientes de otras instituciones médicas.

Muy relacionada con el punto anterior está la necesidad de definición de unos criterios objetivos de evaluación de una conclusión para hacer que los resultados de evaluación de estos sistemas sean más interpretables y se ajusten más al objetivo de crear un modelo que no solo genere unas secuencias que en su forma se parezcan a la referencia, sino que se tenga en cuenta la exactitud fáctica de las conclusiones generadas y su similitud semántica con los referentes. Además, conviene señalar que una evaluación experta del sistema de generación de resúmenes propuesto es otra tarea pendiente de realización.

Por último, se podría plantear un sistema unificado que sea capaz de realizar correctamente las dos tareas: codificación clínica y generación de conclusiones. Vista la aplicabilidad del modelo Flan-T5-base a la tarea de clasificación en formato 'seq2seq', se podría plantear un ajuste multitarea con el fin de que el modelo adquiera las dos capacidades.

Bibliografía

Bibliografía

[Abacha et al.2021] Abacha, Asma Ben, Yassine M'rabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, y Dina Demner-Fushman. 2021. Overview of the mediqa 2021 shared task on summarization in the medical domain. En *Proceedings of the 20th Workshop on Biomedical Language Processing*, páginas 74–85.

[Akiba et al.2019] Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, y Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. En *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[An, Unsdorfer, y Weinreb2019] An, Julie Y, Kyle ML Unsdorfer, y Jeffrey C Weinreb. 2019. Bi-rads, c-rads, cad-rads, li-rads, lung-rads, ni-rads, o-rads, pi-rads, ti-rads: Reporting and data systems. *Radiographics*, 39(5):1435–1436.

[Barentsz et al.2012] Barentsz, Jelle O, Jonathan Richenberg, Richard Clements, Peter Choyke, Sadhna Verma, Geert Villeirs, Olivier Rouviere, Vibeke Logager, y Jurgen J Fütterer. 2012. Esur prostate mr guidelines 2012. *European radiology*, 22:746–757.

[Berge et al.2023] Berge, GT, OC Granmo, TO Tveit, BE Munkvold, AL Ruthjersen, y J Sharma. 2023. Machine learning-driven clinical decision support system for concept-based searching: a field trial in a norwegian hospital. *BMC Medical Informatics and Decision Making*, 23(1):1–15.

[Bergstra et al.2011] Bergstra, James, Rémi Bardenet, Yoshua Bengio, y Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24.

[Bijl, Blaumer, y Matuschek2022] Bijl, Demster, Nils Blaumer, y David Matuschek. 2022. Pairads: Interaction of humans and technology rethought. *Trust, Professional Vision, and Diagnostic Work. 19*, página 51.

[Borko y Bernier1975] Borko, Harold y Charles L Bernier. 1975. Abstracting concepts and methods.

[Cao et al.2020] Cao, Meng, Yue Dong, Jiapeng Wu, y Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. En *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Noviembre.

[Carrino et al.2021a] Carrino, Casimiro Pio, Jordi Armengol-Estapé, Ona de Gibert Bonet, Asier Gutiérrez-Fandiño, Aitor Gonzalez-Agirre, Martin Krallinger, y Marta Villegas. 2021a. Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models. *arXiv preprint arXiv:2109.07765*.

[Carrino et al.2021b] Carrino, Casimiro Pio, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, y Marta Villegas. 2021b. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario.

[Carrino et al.2022] Carrino, Casimiro Pio, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, y Marta Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. En *Proceedings of the 21st Workshop on Biomedical Language Processing*, páginas 193–199, Dublin, Ireland, Mayo. Association for Computational Linguistics.

[Catling, Spithourakis, y Riedel2018] Catling, Finneas, Georgios P Spithourakis, y Sebastian Riedel. 2018. Towards automated clinical coding. *International journal of medical informatics*, 120:50–61.

[Chizhikova et al.2022] Chizhikova, Mariia, Jaime Collado-Montañez, Pilar López-Úbeda, Manuel C Díaz-Galiano, L Alfonso Ureña-López, y M Teresa Martín-Valdivia. 2022. Sinai at clef 2022: Leveraging biomedical transformers to detect and normalize disease mentions.

[Chizhikova et al.2023] Chizhikova, Mariia, Pilar López-Úbeda, Jaime Collado-Montañez, Teodoro Martín-Noguerol, Manuel C Díaz-Galiano, Antonio Luna, L Alfonso Ureña-López, y M Teresa Martín-Valdivia. 2023. Cares: A corpus for classification of spanish radiological reports. *Computers in Biology and Medicine*, 154:106581.

[Chung et al.2022] Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, y others. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

[Costumero et al.2014] Costumero, Roberto, Ángel García-Pedrero, Consuelo Gonzalo-Martín, Ernestina Menasalvas, y Socorro Millan. 2014. Text analysis and information extraction from spanish written documents. En *Brain Informatics and Health: International Conference, BIH 2014, Warsaw, Poland, August 11-14, 2014, Proceedings*, páginas 188–197. Springer.

[Cotik, Filippo, y Castano2015] Cotik, Viviana, Darío Filippo, y José Castano. 2015. An approach for automatic classification of radiology reports in spanish. En *MEDINFO 2015: eHealth-enabled Health*. IOS Press, páginas 634–638.

[Cotik et al.2016] Cotik, Viviana, Vanesa Stricker, Jorge Vivaldi, y Horacio Rodríguez Hontoria. 2016. Syntactic methods for negation detection in radiology reports in spanish. En *Proceedings of the 15th Workshop on Biomedical Natural Language Processing, BioNLP 2016: Berlin, Germany, August 12, 2016*, páginas 156–165. Association for Computational Linguistics.

[Da Cunha, Wanner, y Cabré2007] Da Cunha, Iria, Leo Wanner, y Teresa Cabré. 2007. Summarization of specialized discourse: The case of medical articles in spanish. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 13(2):249–286.

[Delbrouck, Zhang, y Rubin2021] Delbrouck, Jean-Benoit, Cassie Zhang, y Daniel Rubin. 2021. Qiai at mediqa 2021: Multimodal radiology report summarization. En *Proceedings of the 20th Workshop on Biomedical Language Processing*, páginas 285–290.

[Franiel et al.2017] Franiel, Tobias, Michael Quentin, Ullrich Gerd Mueller-Lisse, Lars Schimmoeller, Patrick Asbach, Stefan Rödel, Winfried Willinek, Katja Hueper, Dirk Beyersdorff, y Matthias Röthke. 2017. Mri of the prostate: Recommendations on patient preparation and scanning protocol. *RoFo : Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin*, 189(1):21—28, January.

[Givchi, Ramezani, y Baraani-Dastjerdi2022] Givchi, Azadeh, Reza Ramezani, y Ahmad Baraani-Dastjerdi. 2022. Graph-based abstractive biomedical text summarization. *Journal of Biomedical Informatics*, 132:104099.

[Graves y Graves2012] Graves, Alex y Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, páginas 37–45.

[Hendrycks et al.2020] Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, y Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

[Hripcsak, Kuperman, y Friedman1998] Hripcsak, George, Gilad J Kuperman, y Carol Friedman. 1998. Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods of information in medicine*, 37(01):01–07.

[Ivanov et al.2021] Ivanov, Andrei, Nikoli Dryden, Tal Ben-Nun, Shigang Li, y Torsten Hoefler. 2021. Data movement is all you need: A case study on optimizing transformers. *Proceedings of Machine Learning and Systems*, 3:711–732.

[Kahn Jr et al.2009] Kahn Jr, Charles E, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, y Daniel L Rubin. 2009. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856.

[Kenton y Toutanova2019] Kenton, Jacob Devlin Ming-Wei Chang y Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. En *Proceedings of naacL-HLT*, volumen 1, página 2.

[Kupiec, Pedersen, y Chen1995] Kupiec, J, J Pedersen, y F Chen. 1995. A trainable document summarizer. dans les actes de 18th annual international acm sigir conference on research and development in information retrieval, 68–73.

[Li et al.2018] Li, Lisha, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, y Ameet Talwalkar. 2018. Hyperband: A novel bandit-based approach to hyperparameter optimization.

[Lin2004] Lin, Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. En *Text summarization branches out*, páginas 74–81.

[Liu et al.2019a] Liu, Guanxiong, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, y Marzyeh Ghassemi. 2019a. Clinically accurate chest x-ray report generation. En *Machine Learning for Healthcare Conference*, páginas 249–269. PMLR.

[Liu et al.2019b] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, y Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[Lloret, Plaza, y Aker2018] Lloret, Elena, Laura Plaza, y Ahmet Aker. 2018. The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, 52:101–148.

[Logie1933] Logie, H.B, editor. 1933. *Standard Classified Nomenclature of Disease*. Commonwealth Fund.

[López-Ubeda et al.2020a] López-Ubeda, Pilar, Manuel Carlos Díaz-Galiano, L Alfonso Ureña López, M Teresa Martín-Valdivia, Teodoro Martín-Noguerol, y Antonio Luna. 2020a. Transfer learning applied to text classification in spanish radiological reports. En *Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultilingualBIO 2020)*, páginas 29–32.

[López-Úbeda et al.2020b] López-Úbeda, Pilar, Manuel Carlos Díaz-Galiano, Teodoro Martín-Noguerol, Antonio Luna, L Alfonso Ureña-López, y M Teresa Martín-Valdivia. 2020b. Covid-19 detection in radiological text reports integrating entity recognition. *Computers in Biology and Medicine*, 127:104066.

[López-Úbeda et al.2021] López-Úbeda, Pilar, Manuel Carlos Díaz-Galiano, L Alfonso Ureña-López, y M Teresa Martín-Valdivia. 2021. Combining word embeddings to extract chemical and drug entities in biomedical literature. *BMC bioinformatics*, 22:1–18.

[Loshchilov y Hutter2017] Loshchilov, Ilya y Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

[Luhn1958] Luhn, Hans Peter. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

[López-Úbeda et al.2020] López-Úbeda, Pilar, Manuel Carlos Díaz-Galiano, Teodoro Martín-Noguerol, Alfonso Ureña-López, María-Teresa Martín-Valdivia, y Antonio Luna. 2020. Detection of unexpected findings in radiology reports: A comparative study of machine learning approaches. *Expert Systems with Applications*, 160:113647.

[Ma et al.2017] Ma, Shuai, Yi Liu, Ge Gao, Rui Wang, Yahui Shi, Zuofeng Li, Juan Wei, y Xiaoying Wang. 2017. Using ngram-based features to explore the correlation of prostate mr findings and pi-rads classification. En *Proc. Intl. Soc. Mag. Reson. Med*, volumen 25, página 2077.

[MacAvaney et al.2019] MacAvaney, Sean, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, y Ross W Filice. 2019. Ontology-aware clinical abstractive summarization. En *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, páginas 1013–1016.

[Mihalcea y Tarau2004] Mihalcea, Rada y Paul Tarau. 2004. Textrank: Bringing order into text. En *Proceedings of the 2004 conference on empirical methods in natural language processing*, páginas 404–411.

[Miranda-Escalada et al.2022] Miranda-Escalada, Antonio, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios

Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, y Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. En *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

[Miranda-Escalada et al.2020] Miranda-Escalada, Antonio, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, y Martin Krallinger. 2020. Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020. *CLEF (Working Notes)*, 2020.

[Mottet et al.2021] Mottet, Nicolas, Roderick CN van den Bergh, Erik Briers, Thomas Van den Broeck, Marcus G Cumberbatch, Maria De Santis, Stefano Fanti, Nicola Fossati, Giorgio Gandaglia, Silke Gillessen, y others. 2021. Eau-eanm-estro-esur-siog guidelines on prostate cancer—2020 update. part 1: screening, diagnosis, and local treatment with curative intent. *European urology*, 79(2):243–262.

[Mozayan et al.2021] Mozayan, Ali, Alexander R Fabbri, Michelle Maneevese, Irena Tocino, y Sophie Chheang. 2021. Practical guide to natural language processing for radiology. *Radiographics*, 41(5):1446–1453.

[Névéol et al.2018] Névéol, Aurélie, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, y Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):1–13.

[Olsen, Aisner, y McGinnis2007] Olsen, LeighAnne, Dara Aisner, y J Michael McGinnis. 2007. The learning healthcare system: workshop summary.

[Organization1957] Organization, World Health. 1957. Manual of the international statistical classification of diseases, injuries, and causes of death : based on the recommendations of the seventh revision conference, 1955, and adopted by the ninth world health assembly under the who nomenclature regulations.

[Page et al.1998] Page, Lawrence, Sergey Brin, Rajeev Motwani, y Terry Wi-

nograd. 1998. The pagerank citation ranking: Bring order to the web. Informe técnico, technical report, Stanford University.

[Radev, Hovy, y McKeown2002] Radev, Dragomir, Eduard Hovy, y Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational linguistics*, 28(4):399–408.

[Raffel et al.2020] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, y Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

[Rush, Chopra, y Weston2015] Rush, Alexander M., Sumit Chopra, y Jason Weston. 2015. A neural attention model for abstractive sentence summarization. En *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, páginas 379–389, Lisbon, Portugal, Septiembre. Association for Computational Linguistics.

[Schluter2017] Schluter, Natalie. 2017. The limits of automatic summarisation according to rouge. En *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 41–45. Association for Computational Linguistics.

[Scott1968] Scott, Anne M. 1968. Automatic coding of a diagnosis. En Gordon McLachlan y RFA Shegog, editores, *Computers in the Service of Medicine*, volumen 2. Oxford University Press London, página 89.

[Sennrich, Haddow, y Birch2015] Sennrich, Rico, Barry Haddow, y Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

[Srivastava et al.2014] Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, y Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

[Stanfill et al.2010] Stanfill, Mary H, Margaret Williams, Susan H Fenton, Robert A Jenders, y William R Hersh. 2010. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6):646–651.

[Sun et al.2021] Sun, Yuting, Shifei Ding, Zichen Zhang, y Weikuan Jia. 2021. An improved grid search algorithm to optimize svr for prediction. *Soft Computing*, 25:5633–5644.

[Turkbey et al.2019] Turkbey, Baris, Andrew B Rosenkrantz, Masoom A Haider, Anwar R Padhani, Geert Villeirs, Katarzyna J Macura, Clare M Tempany, Peter L Choyke, Francois Cornud, Daniel J Margolis, y others. 2019. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European urology*, 76(3):340–351.

[Vaswani et al.2017] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, y Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

[Wang et al.2019] Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, y Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

[Wang et al.2018] Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, y Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

[Witbrock y Mittal1999] Witbrock, Michael J y Vibhu O Mittal. 1999. Ultra-summarization (poster abstract) a statistical approach to generating highly condensed non-extractive summaries. En *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 315–316.

[Wolf et al.2019] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, y others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

[Zhang et al.2022] Zhang, Dylan, Ben Neely, Joseph Y Lo, Bhavik N Patel, Terry Hyslop, y Rajan T Gupta. 2022. Utility of a rule-based algorithm in the assessment of standardized reporting in pi-rads. *Academic Radiology*.

[Zhang et al.2018] Zhang, Yuhao, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, y Curtis P Langlotz. 2018. Learning to summarize radiology findings. *arXiv preprint arXiv:1809.04698*.

[Zhang et al.2020] Zhang, Yuhao, Derek Merck, Emily Tsai, Christopher D. Manning, y Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 5108–5120, Online, Julio. Association for Computational Linguistics.

[Zolotareva, Tashu, y Horváth2020] Zolotareva, Ekaterina, Tsegaye Misikir Tashu, y Tomás Horváth. 2020. Abstractive text summarization using transfer learning. En *ITAT*, páginas 75–80.