UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO

DETECCIÓN AUTOMÁTICA DE METÁSTASIS A DISTANCIA DESCRITA EN REPORTES DE IMAGENOLOGÍA MEDIANTE EL USO DE PROCESAMIENTO DE LENGUAJE NATURAL

Ricardo Ignacio Ahumada Oliva

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN INFORMÁTICA MÉDICA.

Directora de Tesis: Prof. Jocelyn Dunstan E. PhD.

Codirector de Tesis: Prof. Pablo Báez B. PhD.

2022

Agradecimientos

A Javiera por el apoyo incondicional en todos los años del magíster. A Ivancito que tuvo paciencia cada vez que le decía que tenía que trabajar. A toda la familia que se preocuparon y preguntaron siempre.

A Pablo Báez por su guía, compañerismo y preocupación. A Jocelyn Dunstan por haberme dado la oportunidad de entrar en el mundo del NLP. Al grupo de PLN clínico del CMM, principalmente Matías Rojas y Fabián Villena.

A la Fundación Arturo López Pérez por los datos proporcionados y por los recursos para la anotación. A Inti, Sergio y Marcela, por acompañar el proceso y entregarnos su crítica y visión pensando en los y las pacientes. A Jocelyn Garay y Gisselle Caamaño por su gran trabajo de anotación que es la base de esta tesis.

A mis compañeras Natalia, Macarena y Camila, por el apañe en los ramos y en la tesis. A mis colegas del SSMS que me apoyaron y aguantaron en todo momento.

Finalmente, al DAAD y al Heidelberg Center para América Latina (Universidad de Heidelberg) por la beca que me proporcionaron y que me permitió estudiar este postgrado con tranquilidad.

ÍNDICE

RESUMEN

Antecedentes: La mortalidad por cáncer se produce principalmente por la progresión del tumor a la etapa de metástasis a distancia. Uno de los criterios para determinar progresión a esta etapa es un examen de imagenología o medicina nuclear. Realizar un tratamiento curativo depende, muchas veces, de la ausencia de metástasis a distancia. Por lo que contar con esta información permite la gestión y priorización de pacientes en lista de espera por intervenciones o tratamientos.

Problema: La metástasis a distancia no se encuentra estandarizada en el registro clínico electrónico del Instituto Oncológico de la Fundación Arturo López Pérez, sino que está en formato de texto libre. Esto dificulta el acceso para la gestión clínica y, detectar manualmente estos hallazgos, consume horas de personal. Por otro lado, la pesquisa de metástasis a distancia se realiza con el análisis de cada reporte de imagenología o medicina nuclear en texto libre, desde donde tampoco se puede extraer la condición de metástasis a distancia de manera automática.

Solución: En esta tesis se propone el desarrollo de un modelo de procesamiento del lenguaje natural capaz de detectar la metástasis a distancia en reportes de imagenología y medicina nuclear y clasificarlos según la presencia o ausencia de esta.

Método: A partir de un corpus anotado con menciones de metástasis afirmativas, negadas o inciertas, se entrenó un modelo de reconocimiento de entidades nombradas basado en una red neuronal recurrente capaz de extraer automáticamente los hallazgos de metástasis a distancia y a partir de ellos, clasificar a nivel de documento cada reporte. Se comparó el rendimiento, medido en precisión, exhaustividad y *F1-score*, de este modelo con un algoritmo basado en reglas, que se utilizó como línea de base.

Resultados: Es posible detectar metástasis a distancia de cáncer de próstata, cáncer de mama y cáncer colorrectal en reportes de imagenología y medicina nuclear, utilizando métodos de procesamiento de lenguaje natural. Se logró detectar entidades de metástasis a distancia al interior del texto clínico con una media balanceada de rendimiento de 0,856 medido en *F1-score*. Además, se clasificó los documentos utilizando aprendizaje profundo con rendimientos máximos, medidos en *F1-score*, de 0,90 para documentos sin metástasis a distancia (M0) y 0,87 para documentos que sí presentaban metástasis a distancia (M1).

ABSTRACT

Background: Cancer mortality is mainly caused by progression of the tumor to the distant metastatic stage. One of the criteria to determine progression to this stage is an imaging or nuclear medicine examination. Curative treatment often depends on the absence of distant metastases. Having this information allows the management and prioritization of patients on the waiting list for interventions or treatments.

Problem: Distant metastasis is not standardized in the electronic health record of the Oncology Institute of the Arturo López Pérez Foundation, but is in free text format. This makes it difficult to access for clinical management, and manually detecting these findings consumes staff hours. On the other hand, the screening for distant metastases is performed with the analysis of each radiology or nuclear medicine report in free text, from where the distant metastasis condition cannot be extracted automatically either.

Solution: In this thesis we propose the development of a natural language processing model capable of detecting distant metastasis in radiology and nuclear medicine reports and classifying them according to the presence or absence of it.

Methods: A named entity recognition model based on a recurrent neural network was developed based on an annotated corpus with affirmative, denied or uncertain metastasis mentions. This model was capable of automatically extracting distant metastasis findings and classifying each report at document level. The performance, measured in precision, completeness and F1-score, of this model was compared with a rule-based algorithm, which was used as a baseline.

Results: It is possible to detect distant metastases of prostate cancer, breast cancer and colorectal cancer in imaging and nuclear medicine reports using natural language processing methods. We were able to detect distant metastasis entities within the clinical text with a balanced mean performance of 0,856 measured in F1-score. In addition, documents were classified using deep learning with maximum F1-score performances of 0.90 for documents without distant metastases (M0) and 0.87 for documents with distant metastases (M1).

1.- INTRODUCCIÓN

1.1.- Antecedentes

Cáncer en Chile: Epidemiología y progresión a metástasis.

El cáncer es la primera causa de muerte en Chile desde el año 2019, superando por primera vez a las enfermedades cardiovasculares (1). Los tipos de cáncer con mayor incidencia y prevalencia en Chile durante el 2020 fueron el de próstata, colorrectal y de mama. En conjunto, estos tres cánceres alcanzan una incidencia de 19.707 casos y una prevalencia en un año de 16.622 personas. Además, el cáncer colorrectal y el cáncer de mama son los más mortales, junto al cáncer de pulmón (2).

La principal causa de muerte en pacientes con cáncer es la metástasis, proceso en el cual las células cancerosas de un tumor primario se diseminan a tejidos contiguos o a un órgano distal (3). Esta se produce a través del sistema circulatorio o linfático, invadiendo y proliferando en un nuevo tejido (4). La metástasis se da en etapas avanzadas de la enfermedad y produce al menos dos tercios de las muertes por cáncer (5).

Entre un 20% a un 30% de los pacientes con cáncer de mama progresan a metástasis (6). Se espera que 1 de cada 9 pacientes con cáncer de próstata la desarrolle (7). El 20% de los pacientes con cáncer colorrectal ya tienen metástasis al momento del diagnóstico (8) y se estima que un 70% la desarrollará en el transcurso de la enfermedad (9). A pesar de los avances en el desarrollo de nuevas terapias como la inmunoterapia y nanopartículas, las opciones curativas en la metástasis son reducidas (10,11).

Esto constituye un desafío no sólo en términos de demanda hospitalaria, diagnóstico y tratamiento, sino también en términos de registro y procesamiento eficiente de los datos para la gestión y priorización de los pacientes que esperan por una intervención o un tratamiento, con el fin proporcionar una atención clínica oportuna y pertinente.

La extracción de información asociada a metástasis de fuentes de datos como el registro clínico electrónico (RCE), permite entender mejor las condiciones de los pacientes y apoya el desarrollo de nuevas drogas y estudios clínicos (12).

Por otro lado, la extracción de información es un apoyo para la práctica cotidiana de la medicina, pues permite a los clínicos tenerla a disposición para tomar decisiones respecto a indicaciones médicas, coordinación del proceso de atención y la comunicación con el equipo de salud (13).

Clasificación de la metástasis

La clasificación TNM de tumores malignos es el estándar más utilizado para clasificar el estado de avance de un tumor. Esta clasificación también es llamada estadificación y permite planificar el tratamiento, entregar una idea del pronóstico del paciente, aportar la evaluación de los tratamientos, facilitar el intercambio de información y contribuir a la investigación y vigilancia epidemiológica en cáncer (14).

Según la clasificación TNM, la metástasis puede definirse como metástasis a linfonodos regionales (N) o como metástasis a distancia (M). Ésta última se define como aquella metástasis que invade órganos más allá de los límites anatómicos específicos para cada tipo de cáncer. La presencia de metástasis a distancia es descrita como M1 y la ausencia como M0. Existen subclasificaciones de M1 que dependen del tipo de cáncer (14).

Guías de Práctica Clínica y Rutas Clínicas

Las guías de práctica clínica son documentos de consenso que buscan optimizar el cuidado de los pacientes, están basadas en la evidencia científica y en la evaluación de beneficio-riesgo de las alternativas disponibles. Estas guías tienen el potencial de reducir la variabilidad en la práctica clínica, mejorar la calidad y seguridad asistencial, y además, entregan las recomendaciones para el mejor tratamiento y cuidado de la salud (15). En Chile, actualmente se utilizan para cáncer de próstata, cáncer colorrectal y cáncer de mama están descritas y definidas desde hace más de 5 años (16–18).

Al aterrizar las guías de práctica clínica en un contexto institucional, surgen las rutas clínicas. Estas rutas son herramientas guía para una atención de salud basada en la evidencia. Buscan traducir las guías de práctica clínica a los procesos de un determinado prestador. Una ruta clínica es un plan multidisciplinario que define un proceso clínico y estandariza el flujo de cuidado para un problema de salud específico en un grupo específico de la población (19), como por ejemplo, un paciente con cáncer de próstata.

En oncología son particularmente importantes porque permiten por un lado, mejorar la calidad de la atención médica y, por otro, controlar los costos de los tratamientos y generar prácticas que asuman mejor el riesgo financiero (20), garantizando recursos para una atención integral, oportuna y de mayor calidad.

Rutas Clínicas en la Fundación Arturo López Pérez (FALP)

La FALP es un centro oncológico de referencia, sin fines de lucro dedicado a diagnóstico, tratamiento, investigación y docencia. Cuenta con un modelo de salud asistencial, académico y de investigación, centrado en el paciente y su entorno y atiende pacientes desde los 15 años de edad de cualquier previsión de salud (21).

En esta institución se han desarrollado rutas clínicas para diferentes tipos de cáncer como los cáncer de próstata, colorrectal y de mama. Estas contemplan sub-rutas, definidas para cada caso, según los estadios del cáncer y otros factores de riesgo complementarios. Cada sub-ruta considera diferentes acciones diagnósticas y terapéuticas. Todos los pacientes con cáncer en la FALP deben ser clasificados según el estándar TNM de tumores malignos, lo que permite aportar información relevante para la definición del estadio clínico que definirá la sub-ruta y el tratamiento recomendado según la evidencia científica.

El estadio IV de un cáncer en la mayoría de las rutas clínicas corresponde a la presencia de metástasis a distancia y uno de los criterios que permite clasificar un cáncer en esa etapa es poseer uno o más exámenes imagenológicos o de medicina nuclear positivos para metástasis. La categorización de un paciente en este estado, es decir, el diagnóstico de metástasis permitirá gestionar la lista de espera de

consultas, intervenciones y procedimientos, priorizando a los pacientes según el requerimiento de su estado de salud.

Diagnóstico de metástasis

La pesquisa de metástasis a distancia se puede realizar mediante exámenes imagenológicos, por ejemplo, técnicas clásicas como la radiografía, tomografía computada (TC), la resonancia magnética (RM), el cintigrama Óseo (CO) y técnicas más complejas, por ejemplo , la medicina nuclear, como la tomografía por emisión de positrones con tomografía computada (PETCT en inglés), tomografía computarizada de emisión monofotónica (SPECT en inglés) y variantes de la RM, como la RM de cuerpo completo (11, 22–26).

En la literatura se destaca la TC, RM y PETCT, por tener una mayor especificidad y sensibilidad según el tipo de cáncer y de metástasis (23–26).

La mayoría de la información relativa a la metástasis de un cáncer se encuentra en reportes de imagenología y medicina nuclear (en adelante se agruparán ambas áreas en el concepto de imagenología), escritos por médicos radiólogos. Tradicionalmente, la información de estos reportes se encuentra en formato no estructurado, también llamado texto libre o texto clínico (27–29). Para permitir una mayor comprensión de los hallazgos, estos reportes se pueden escribir en formato semiestructurado (30).

Texto clínico en radiología y procesamiento de lenguaje natural

La interpretación del médico radiólogo sobre un examen imagenológico queda plasmada en los reportes en formato no estructurado o semiestructurado. Este formato permite al médico detallar o resumir sus ideas, otorgando un valor interpretativo y contextual. Si bien el vocabulario utilizado en este tipo de reportes es más acotado con relación a otros tipos de texto clínico, trabajar con ellos de manera eficiente es un desafío. La posibilidad de identificar rápida y automáticamente la información en este tipo de texto podría reducir la carga de trabajo de profesionales y técnicos encargados de procesos de registro y también, apoyar procesos de atención clínica e investigación (31).

Para que los sistemas puedan extraer información relevante del texto no estructurado o texto libre es necesario el uso de recursos lingüísticos, computacionales y estadísticos que permitan interpretar el lenguaje humano, también llamado lenguaje natural, con todas sus particularidades (29).

El procesamiento de lenguaje natural es el subcampo de las ciencias de la computación y la inteligencia artificial que busca entender y producir el lenguaje humano (32). Esta tiene por objetivo desarrollar máquinas que puedan desarrollar acciones que a un ser humano le requerirían inteligencia, es decir, que puedan replicar o simular nuestro comportamiento, por lo que el procesamiento del lenguaje natural es relevante para entender nuestra forma de comunicarnos.

Los métodos asociados al procesamiento del lenguaje natural pueden ser clasificados, por un lado, en aquellos basados en reglas, en los que un humano incorpora manualmente las reglas que el algoritmo debe cumplir para realizar una determinada tarea. Esto conlleva un alto costo en términos de trabajo y recursos, pues se debe controlar la tarea completamente. Por otro lado, están aquellos basados en aprendizaje de máquinas (*machine learning*), que corresponden a la aplicación de algoritmos para “enseñar” a las máquinas a aprender de patrones o ejemplos dados por humanos o de la propia experiencia en el procesamiento de los datos. En este tipo de modelos, se debe recurrir a la ingeniería de características, es decir, el humano debe procesar los datos disponibles y extraer características relevantes para el entrenamiento. Y finalmente, están los métodos asociados al aprendizaje profundo, un subcampo del aprendizaje de máquinas que se basa en el funcionamiento de redes neuronales artificiales, ordenadas en capas y en el que esta selección de características no es necesaria (33) .

Ejemplos de tareas modernas de PLN para textos de dominio general, son la traducción automática (*machine translation*), en la que los sistemas pueden generar oraciones en un idioma distinto al original; los sistemas de diálogo hablado y agentes de conversación, que reconocen el discurso humano en voz o texto y logran generar respuestas coherentes para una persona; la lectura automática (*machine reading*), que permite a los sistemas integrar y resumir información y, la minería de redes

sociales con un gran auge en los últimos años por la cantidad de información que contienen (32,34,35).

Según una revisión sistemática de Spasic y Nenadic (36), la gran mayoría de las aplicaciones desarrolladas específicamente para el procesamiento de texto clínico se asocian a la tarea de clasificación de texto, seguido por el reconocimiento de entidades nombradas y la extracción de información. Estas son las tareas principales que se abordarán en el presente estudio.

Clasificación de Texto Clínico

Es una de las tareas clásicas del procesamiento del lenguaje natural en el texto clínico. En ésta se busca asignar una clase predeterminada a un documento o parte de este. La clasificación se realiza tradicionalmente mediante métodos supervisados de aprendizaje de máquinas en los que se utiliza un conjunto de documentos a los que se asigna una clase de manera manual. Este conjunto de datos se utiliza como ejemplo para *entrenar* un algoritmo capaz de predecir la clase asignada en nuevos documentos.

Reconocimiento de Entidades Nombradas

El reconocimiento de entidades nombradas (*NER*) es una tarea del procesamiento del lenguaje natural que permite encontrar entidades o cadenas de caracteres relevantes en el texto y asignarles una etiqueta que permita identificarlas . Por ejemplo, en texto clínico se pueden identificar entidades de interés como nombres de enfermedades, medicamentos, diagnósticos, exámenes de laboratorio, codificación y estándares, entre otros (37). En el dominio del cáncer, el desarrollo de modelos que permitan identificar conceptos relevantes ha aumentado por los beneficios que trae para la gestión asistencial y la investigación clínica. Por ejemplo, la identificación de casos con cáncer en el registro clínico electrónico, etapificación del tumor de un paciente, la predicción de resultados de tratamiento y desarrollo de sistemas de soporte a la decisión clínica (38).

Los modelos de reconocimiento de entidades nombradas actuales, son desarrollados en base a redes neuronales. Ejemplos clásicos de estos modelos son las redes neuronales recurrentes y las redes neuronales convolucionales (37)

Aprendizaje Profundo

El estado del arte en modelos de clasificación de texto y reconocimiento de entidades nombradas se basa en el uso de aprendizaje profundo. El aprendizaje profundo es una rama del aprendizaje de máquinas en que el computador aprende sin que exista una selección manual de características por parte de un humano. Dada su simplicidad en términos de intervención manual, su procesamiento eficiente y los resultados que han obtenido en tareas de PLN, es que han adquirido rápidamente popularidad entre los investigadores (39). Estos modelos se construyen a partir de redes neuronales artificiales, inspiradas en las redes neuronales biológicas. Estas redes están compuestas por tres capas principales: la capa de input, capas intermedias, y la capa de salida. La información se propaga a través de estas capas utilizando funciones matemáticas, logrando así un resultado que puede ser, por ejemplo, la clasificación de una palabra o entidad. La primera capa de los modelos de aprendizaje profundo corresponde a una capa de redes neuronales más simples que permiten representar palabras vectorialmente a partir de cientos de características. Este proceso es llamado *embedding*. (33).

De los modelos de aprendizaje profundo aplicados al procesamiento del lenguaje natural, las arquitecturas basadas en redes neuronales recurrentes son las más utilizadas (31). Las redes neuronales recurrentes procesan la información de manera secuencial por lo que son ideales para el procesamiento de oraciones (secuencias de palabras). En este caso, cada neurona está vinculada de manera lineal a la otra, pasando el “mensaje” a la siguiente. En este tipo de redes neuronales el orden de las palabras es importante y puede variar su el significado de una oración. Este tipo de redes genera memoria, es decir, permite ir utilizando los cómputos previos en los cómputos posteriores que realiza el algoritmo.

Según la revisión metodológica realizada por Wu et al. (39), en texto clínico el 60,8% de las publicaciones que utilizaron aprendizaje profundo, utilizaron redes neuronales

recurrentes y las variantes más utilizadas fueron una red neuronal recurrente de memoria prolongada a corto plazo (*LSTM*), Unidad recurrente con compuerta (*GRU*) y la red neuronal *Vanilla*.

La ventaja de las redes neuronales recurrentes es que pueden retener información de datos procesados en el pasado. Pero junto con esto, son difíciles de entrenar, pues entre varias razones se destaca el problema del desvanecimiento del gradiente. Esto provoca que el modelo no pueda aprender la correlación entre eventos distantes (de dependencia de larga distancia) (40).

Las LSTM fueron una arquitectura introducida por Hochreiter (41), son un tipo de red neuronal recurrente que eran capaces de aprender recordando información en el tiempo. Cuentan típicamente con 3 compuertas: olvido, entrada y salida. La primera selecciona y almacena los elementos relevantes de los datos incorporados en el dataset de entrenamiento y "olvida" o les da menos valor a los datos menos relevantes. La segunda compuerta decide qué nueva información debe incorporarse a la red para actualizar la "memoria". Y la última compuerta utiliza la información previa para definir el *output* de la red neuronal. Este tipo de redes neuronales solucionan el problema de desvanecimiento del gradiente de las RNR porque introducen un estado separado de memoria agregada de toda la data ya procesada para así evitar los cambios bruscos (desvanecimiento o explosión).

En la revisión que realizó Névéol et al. (42), destaca que las tareas más importantes del procesamiento del lenguaje natural, como la clasificación y el reconocimiento de entidades nombradas, han sido ampliamente estudiadas pero, principalmente, para el idioma inglés. En ese sentido, para el idioma español se vislumbran grandes desafíos. Aún no se puede observar el impacto del procesamiento del lenguaje natural en la práctica clínica y la salud pública en países de habla hispana. Según esta revisión sistemática, el español fue el cuarto lenguaje con más artículos científicos encontrados asociados a procesamiento del lenguaje natural, detrás del francés, alemán y chino, pero aún muy por debajo del inglés.

A partir de la literatura revisada a la fecha, la detección de los hallazgos clínicos de metástasis a distancia no ha sido investigada en español, por lo que este proyecto

implica un aporte en la aplicación de modelos de aprendizaje de máquinas con aplicación en nuestro idioma.

1.2.- Problema a resolver

La metástasis es la etapa más avanzada del cáncer y ocurre cuando un tumor se disemina a otros órganos o tejidos. Su manejo es crítico y conlleva una terapia sistémica que permite destruir las células que han escapado al torrente sanguíneo o linfático. La mayoría de las veces, las terapias disponibles no permiten sanar el cáncer metastásico completamente. Cuando un paciente es asignado a una determinada ruta clínica, dependerá del estado o grado de avance del tumor si se puede gestionar una terapia curativa a pacientes con pronóstico favorable u otorgar un fin de vida digno a aquellos pacientes que por su pronóstico, necesiten de cuidados paliativos.

Para priorizar a los pacientes en lista de espera de atención, tratamiento o intervención quirúrgica, ya sea curativa o paliativa, es necesario desarrollar sistemas que permitan gestionar la demanda de atención. El problema, en el caso de la metástasis a distancia, es que la información en el registro clínico electrónico no se encuentra estandarizada, sino que está en formato de texto libre. Esto dificulta o hace menos eficiente el registro, extracción y uso de la información para la atención clínica, gestión asistencial o investigación.

Realizar esta tarea de manera manual consume mucho tiempo de trabajo y personal. Además, la estadificación TNM no es estática y cambia con el progreso de cada cáncer por lo que debe ser actualizada junto a los controles de cada paciente.

1.3.- Solución

Como solución a estos problemas, se propone el desarrollo de modelos de aprendizaje de máquinas que permitan, por un lado, extraer información de manera estandarizada en notación M, de los hallazgos de metástasis a distancia al interior de los reportes de exámenes imagenológicos en texto libre. Y por otro, clasificar según la presencia o ausencia de metástasis a distancia los reportes a nivel de documento para obtener una propuesta de clasificación M de un paciente, lo que conducirá a una gestión más eficiente de la lista de espera de atenciones e intervenciones y permitirá

mantener información estandarizada para estudios clínicos o epidemiológicos. Por ejemplo, conocer el tiempo de progresión de enfermedad, cuánto tiempo pasa desde que el paciente es diagnosticado hasta la progresión a metástasis en determinado cáncer y planificar la oferta y prestaciones de las instituciones en función de aquello.

Este trabajo sería el primero en extraer hallazgos clínicos asociados a la presencia o ausencia de metástasis a distancia en textos clínicos de dominio imagenológico en el idioma español. Por lo que permitiría avanzar, por una parte, en la adaptación a nuestro contexto de modelos de aprendizaje profundo que están en el estado del arte y que han sido probados para la extracción de otros hallazgos clínicos y clasificación de documentos radiológicos en idiomas como el inglés y por otro, aportar al conocimiento de las características del texto clínico en español en Chile.

1.4.- Trabajos relacionados

Los trabajos de revisión sistemática de Casey et al. (31) y Sorin et al. (33) dan cuenta de los trabajos realizados en procesamiento del lenguaje natural en reportes de imagenología recientemente. Miao et al. (43), por ejemplo, compararon tres metodologías de procesamiento del lenguaje natural para extraer la clasificación *BI-RADS* desde 540 reportes de ultrasonido de mama. Evaluaron un método basado en reglas, un método de campos aleatorios condicionales (*CRF*) y un modelo de redes neuronales recurrentes, logrando mejor rendimiento con el último (*F1-score* de 0,90). En este caso, al igual que en el problema a resolver en este proyecto, se utilizan reportes de imagenología y se evalúa una tarea de reconocimiento de entidades nombradas. Ahí extrajeron hallazgos clínicos que permitieran establecer un grado de malignidad del tumor. De las arquitecturas de redes neuronales recurrentes que utilizaron, la que mejor rendimiento tuvo fue el uso de redes neuronales recurrentes LSTM pero bidireccionales (Bi-LSTM), con representaciones vectoriales o embeddings a nivel de caracter.

En este artículo también se refieren a las mayores causas de errores en el reconocimiento de entidades nombradas. Entre ellas están, por ejemplo, el manejo de los modificadores (palabras que cambian el significado de una entidad), la

diversidad y cantidad de entidades por documento y la inconsistencia de la anotación manual.

Lee et al. (44) también evaluó redes neuronales recurrentes para automatizar la clasificación de texto clínico en fracturas y no-fracturas en 3,032 frases seleccionadas de reportes de radiología por un médico cirujano ortopédico. Se utilizó un modelo de LSTM de tres capas que mostró un *F1-score* de 0,967. En este caso, el problema tiene ciertas similitudes a este proyecto, puesto que se trata de texto clínico en reportes radiológicos, en que la idea es diferenciar la presencia o ausencia de un hallazgo clínico (fractura o no-fractura). Además del hallazgo específico, la diferencia con este trabajo radica en la longitud de los textos, puesto que en este trabajo se busca predecir la clase del documento completo.

En el ámbito de la clasificación de texto clínico radiológico, Chen et al. (45) clasificó reportes de TC según la existencia o ausencia de hallazgos de embolismo pulmonar, obteniendo un *F1-score* de 0,891. Esto lo hicieron con anotaciones a nivel de documento en tres categorías, utilizaron GloVe para obtener los embedding y redes neuronales convolucionales. Este artículo presenta una alternativa para la clasificación de documentos clínicos, problema a abordar en este proyecto de tesis, realizando una anotación a nivel de documento y además, menciona cómo la distancia entre palabras clave es uno de los varios problemas a tener en cuenta al trabajar con este tipo de redes. Lo que se podría solucionar, como fue mencionado previamente, al utilizar modelos de redes neuronales recurrentes como LSTM.

Lenain et al. (46) buscó clasificar TNM en reportes de anatomía patológica de pacientes con cáncer, entre ellos, de próstata y de mama. Para la clasificación de M obtuvieron 0,99 de F-score, utilizando métodos tradicionales de aprendizaje de máquinas como *SVM* y *Gradient Boosting.* Explican que este rendimiento puede ser explicado por el desbalance de clases, presente también en el dataset utilizado para este proyecto.

Banerjee et al. (47) demostró que los algoritmos basados en la atención sobrepasan en rendimiento a modelos de PLN tradicionales y a redes neuronales convolucionales y recurrentes para clasificar reportes de embolia pulmonar. Los modelos basados en

la atención emulan la capacidad que tenemos los humanos para poner énfasis en la información más relevante de toda la que estamos recibiendo. Los modelos computan los vectores que representan la información clave del texto.

Solarte Pabón et al. (48) desarrollaron un modelo basado en aprendizaje profundo que les permitió extraer la notación TNM de manera explícita (por ejemplo: ”pT2N0M0”) y otros hallazgos desde notas clínicas de pacientes con cáncer de pulmón en español. A diferencia de lo realizado en este artículo, en este trabajo no es posible extraer la notación TNM de manera explícita, pues no está escrita directamente en el texto. Esto produce que esta tarea deba desarrollarse entendiendo las entidades de metástasis a distancia como hallazgos clínicos, similares a las entidades definidas por ellos como “*Cancer entity*”. Para esto, entrenaron un modelo de Bi-LSTM. Este modelo cuenta con una capa de embeddings entrenada con textos científicos y artículos de internet relacionados a la salud, la capa Bi-LSTM que procesa la información en sentido derecho e inverso y las concatena en un vector y por último, una capa de Campos aleatorios condicionales (CRF en inglés), que entrega la mejor etiqueta para cada input u oración procesada.

El trabajo de Solarte Pabón entrega luces para trabajar un aspecto que podría ser problemático en la extracción de hallazgos clínicos, o metastásicos, en nuestro caso. Menciona que el modelo descrito anteriormente puede extraer las entidades relacionadas al cáncer de pulmón, pero que muchas de éstas podrían verse afectadas por modificadores como la negación y la incertidumbre. Para abordar aquello, utilizaron un modelo basado en reglas que les permitía detectar estos modificadores y asociarlos a las entidades detectadas. Al extraer todas las entidades nombradas sin filtrar, consiguieron un *F1-score* de 0,64, pero cuando se filtraron las entidades negadas y las inciertas, el *F1-score* aumentó a 0,89. Esto nos indica que la detección de negación e incertidumbre es un paso crucial en la extracción de información en esta especialidad del dominio médico.

Por otro lado, sus resultados dan cuenta de la necesidad de contar con embeddings de dominio clínico o dominio específico al tipo de problema a resolver, en comparación a embeddings de dominios generales. Por ejemplo, con el solo hecho de utilizar embeddings clínicos, aumentaron su *F1-score* de 0,76 a 0,85.

En relación a la identificación de metástasis específicamente, Groot et al. (49) y Senders et al. (50) buscaron cuantificar automáticamente las metástasis óseas en reportes de cintigrama óseo y resonancia magnética. Para esto, desarrollaron un modelo de clasificación binaria entre una metástasis única y metástasis múltiples, mediante modelos de regresión sobre la probabilidad a partir de los antecedentes clínicos de los pacientes de tener una u otra clase. En este sentido, nuestro trabajo abordará los hallazgos de metástasis a distancia de manera distinta, pues se los buscará mencionados explícitamente en el texto, de manera que éstos puedan ser revisados por los clínicos.

En las revisiones sistemáticas mencionadas con anterioridad, se destaca el incremento en el uso de aprendizaje profundo en tareas de procesamiento de lenguaje natural asociadas a la extracción o estructuración de texto libre. En ellas se menciona la utilidad del aprendizaje profundo en el etiquetado automático, clasificación de reportes, la generación de cohortes, entre otros, para distintas aplicaciones clínicas, como la clasificación de la información de enfermedades, la vigilancia y el aseguramiento de la calidad.

A pesar del aumento explosivo de la producción de modelos de aprendizaje profundo durante los últimos años, se destaca que aún los modelos de machine learning tradicionales son ampliamente utilizados y que, en muchos casos, se usan como línea de base para la evaluación del rendimiento de los nuevos modelos (31). Incluso, en el mismo documento se destaca que para la tarea de clasificación de texto, específicamente para el dominio de los textos radiológicos, los modelos más utilizados son *SVM*, regresión logística y *random forest*.

Los métodos convencionales de aprendizaje de máquinas han mantenido terreno porque tienen mejor interpretabilidad en comparación con los modelos de aprendizaje profundo. Entender las características que llevan a un modelo a predecir cierta clase u otra permite apoyar la toma decisiones clínicas (51) por lo cual, no es conveniente desechar su uso, aún con el avance hacia algoritmos más sofisticados de aprendizaje profundo en tareas como la clasificación de texto y el reconocimiento de entidades nombradas.

2.- HIPÓTESIS

El procesamiento de lenguaje natural permite desarrollar un sistema que detecte automáticamente la metástasis a distancia en reportes de imagenología en texto libre, con rendimiento medido en *F1-score* de al menos 0,8.

3.- OBJETIVOS

3.1.- Objetivo general

Determinar el rendimiento de modelos de aprendizaje de máquinas y aprendizaje profundo en la detección automática de metástasis a distancia descrita en reportes de imagenología y medicina nuclear en formato de texto libre de pacientes con cáncer de próstata, colorrectal y de mama.

3.2.- Objetivos específicos

1. Desarrollar un corpus anotado de las menciones de metástasis a distancia en reportes de exámenes imagenológicos de pacientes con cáncer de próstata, cáncer colorrectal y cáncer de mama.
2. Entrenar modelo de aprendizaje de máquinas para detectar automáticamente las menciones de metástasis a distancia al interior de reportes imagenológicos.
3. Desarrollar metodología para clasificar los reportes imagenológicos a nivel de documento.
4. Evaluar el rendimiento de los modelos versus el conjunto de datos de prueba.

4. MATERIALES Y MÉTODOS

4.1.- Desarrollo del corpus anotado.

Datos disponibles

Los datos de este proyecto fueron proporcionados por el Instituto Oncológico de la Fundación Arturo López Pérez, a través de la Unidad de Informática Médica y en el contexto de la aprobación de un proyecto clínico de investigación con la Universidad de Chile, por parte del comité ético-científico de la Fundación (Anexo 1). Desde la plataforma de gestión de rutas clínicas se obtuvieron 1.125 documentos, los que se distribuyen en reportes de imagenología y medicina nuclear. Específicamente, corresponden a reportes de PETCT, TC y RM de pacientes con cáncer colorrectal, de mama y de próstata. La distribución de los datos se describe en la Tabla 4.1

Tabla 4.1: Cantidad de reportes proporcionados según tipo de reporte y tipo de cáncer.

Tipo de Reporte	Cáncer de Próstata	Cáncer de Mama	Cáncer CR Colon y recto superior	Cáncer CR Recto Medio e Inferior	Total de documentos por tipo de reporte
PETCT	196	90	53	9	348
RM	62	209	62	23	356
TC	33	242	124	22	421
Total de documentos por tipo de cáncer	291	541	239	54	1.125

CR: Colorectal, **PETCT:** Tomografía por Emisión de Positrones – Tomografía Computada, **RM:** Resonancia Magnética, **TC:** Tomografía Computarizada.

Descripción de los Reportes de Imagenología y menciones de metástasis

La estructura, organización de la información y cantidad y tipos de menciones de metástasis se analizaron en una muestra de los reportes, los cuales se encuentran en formato de texto libre. Al interior de cada tipo de reporte se logran identificar

secciones que dividen los reportes para fines descriptivos y que dependiendo del examen, pueden estar definidas explícita o implícitamente en el texto. A continuación se detallan los principales aspectos identificados en cada tipo de reporte y se presentan ejemplos reales en las Figuras 4.1 a 4.3.

Tomografía por Emisión de Positrones – Tomografía Computada (PETCT)

Los reportes de PETCT son reportes semiestructurados, es decir, siguen una "plantilla" que los estructura, pero están en formato de texto libre. Incluyen un "título" e "introducción" donde se indica el tipo de examen y antecedentes del procedimiento. En ocasiones también se incorporan resultados anteriores o un resumen de antecedentes clínicos del paciente que sean relevantes para el estudio.

Luego se desarrolla el examen, donde se describe por zona anatómica los hallazgos. Se observan las secciones de "Cabeza y Cuello", "Tórax", "Abdomen y Pelvis" y "Músculo esquelético". El orden en que quedan almacenadas estas secciones en el reporte no es el mismo en todos los informes.

Tomografía Computada (TC)

Los reportes de TC varían en su longitud y complejidad, lo que depende del tipo del tejido analizado en cada TC. Debido a la ubicación de los tumores, se utilizaron en su mayoría TC de mama y abdomen y pelvis.

Los hallazgos se ordenan por sitio anatómico, pero su estructura no es tan definida como en los PETCT. Contienen el nombre del examen, observaciones del procedimiento, la descripción de los hallazgos observados y una conclusión o síntesis de lo observado.

Resonancia Magnética (RM)

Los reportes de RM tienen, en general, una menor longitud en comparación a los de PETCT y las TC. Incluyen un título del examen, observaciones del procedimiento y un breve resumen de antecedentes relevantes para el estudio. Se identifica además un apartado donde se relatan los hallazgos clínicos. Los informes de RM utilizados en este estudio son más bien focalizados en un sitio anatómico, por ejemplo, para un

cáncer de próstata se contaba con mayoría de RM de próstata o pelvis y para cáncer de mama, RM de mama o mama bilateral.

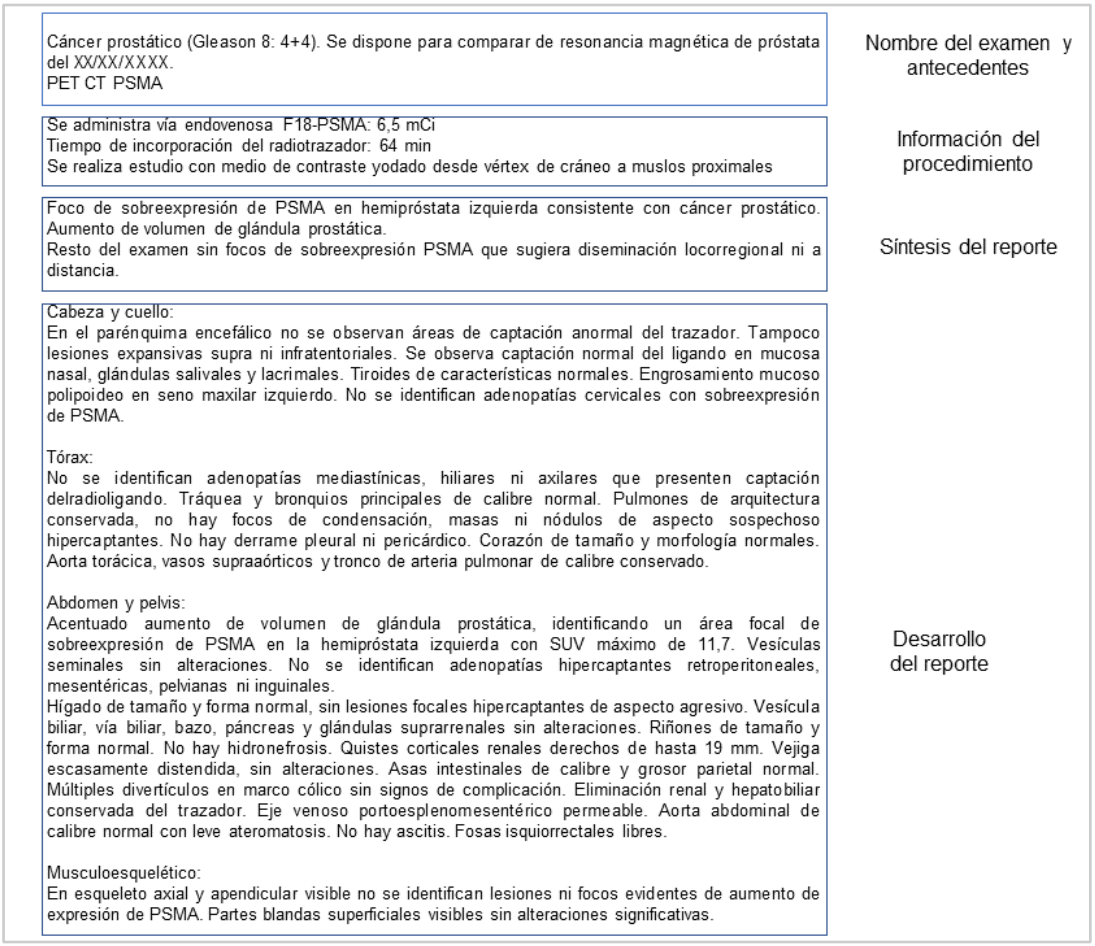

Figura 4.1. Ejemplo real de estructura de un reporte de PETCT, FALP. Cáncer de Próstata. Se enmascaró la fecha por confidencialidad.

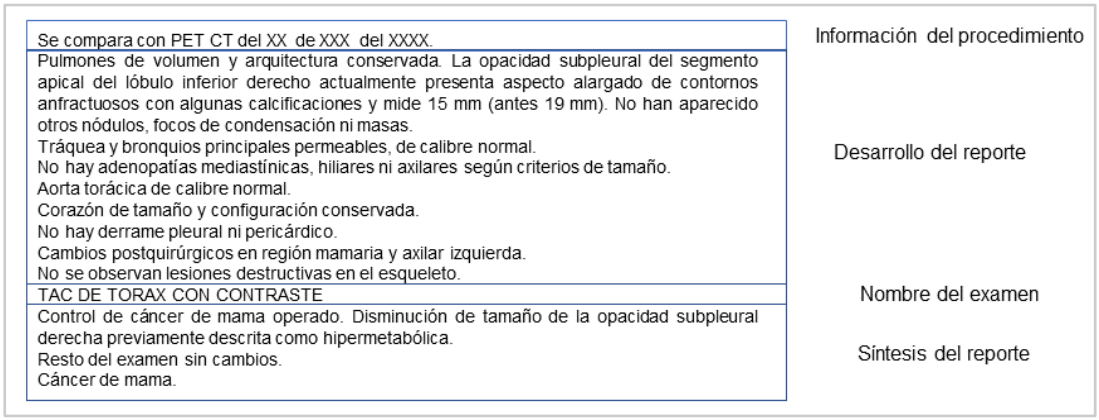

Figura 4.2. Ejemplo real de la estructura de un reporte de TC, FALP. Cáncer de mama.Se enmascaró la fecha por confidencialidad.

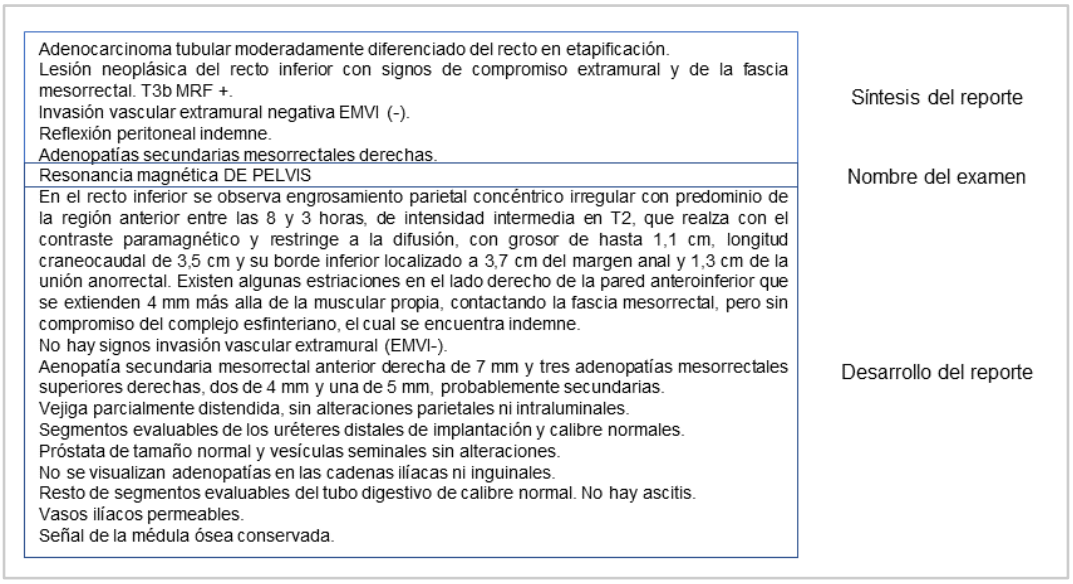

Figura 4.3. Ejemplo real de estructura de un reporte de RM, FALP. Cáncer colorrectal..

Hallazgos de Metástasis a Distancia

Los hallazgos de metástasis a distancia en el texto pueden catalogarse como afirmativos, inciertos o negados. Son afirmativos cuando se confirma la presencia de una lesión metastásica, son inciertos cuando el médico o la médica no puede entregar una información precisa o no tiene seguridad de lo que está viendo y por lo tanto, solicita más estudios. Por último, son negados cuando se descarta una enfermedad metastásica explícitamente.

Los y las médicas radiólogas rara vez describen los hallazgos metastásicos diciendo la palabra “metástasis” o derivados de esta raíz, de manera literal. La manera más frecuente de hacerlo es mediante la descripción de una lesión sospechosa ubicada en un órgano o tejido asociado a metástasis a distancia para ese cáncer, según la clasificación TNM de tumores malignos.

Este hallazgo se puede describir de distintas maneras y dependerá del examen, del médico e incluso de la institución, por ejemplo: “lesión hipermetabólica” y “foco hipermetabólico” en PETCT, “foco nodal”, “adenopatía” y “lesión focal” en TC.

Estas descripciones sólo corresponden a metástasis a distancia si están presentes en los órganos o ganglios linfáticos determinados por la clasificación TNM de tumores y son afirmativas, es decir, cuando el médico explícitamente así lo mencione. En la

mayoría de los casos, se describen menciones de metástasis a distancia negadas. Esto es cuando el o la médica descarta una lesión de estas características.

Además de la mención propiamente tal, se identificaron “moduladores” o palabras que cambian el sentido de las menciones, por ejemplo, “se descarta”, “sin evidencia de” o “sin”, para describir una mención negada, o por ejemplo, “de carácter indeterminado” e “indefinido” para menciones de metástasis a distancia ambiguas.

Proceso de Anotación

La anotación significa añadir manualmente una etiqueta a uno o más *tokens* para identificar a qué tipo de entidad representa*.* Un *token* es una serie de caracteres relevantes presentes en el texto. (29). Esta es la tarea más importante, pues el corpus anotado será utilizado para realizar el entrenamiento y validación del modelo.

La anotación fue realizada entre septiembre de 2021 y marzo 2022 por dos técnicas de enfermería de nivel superior, quienes se desempeñan en el Registro de Tumores de la FALP. Para realizar el etiquetado de los textos, se utilizó el software INCEpTION (52) alojado en el servidor del Grupo de Procesamiento de Lenguaje Natural del Centro de Modelamiento Matemático de la Universidad de Chile.[1]

Se capacitó a las anotadoras en el uso de la plataforma y se realizó una guía de anotaciones que permitió conducir las anotaciones de manera coherente y solucionar posibles discrepancias, presentando ejemplos e instrucciones. La guía se realizó mediante el software Annodoc (53) y está disponible en línea [2].

La anotación se realizó en tres etapas por tipo de reporte, con tres periodos de pre-anotación para cada tanda. La preanotación consistía en el etiquetado de un pequeño subset de documento, lo que permitía el entrenamiento en el proceso de anotación, la adecuación del análisis para cada tipo de examen, y la generación de dudas y consultas útiles para el mejoramiento de la guía de anotación.

[1] http://pln.cmm.uchile.cl/

[2] https://ahumadao.github.io/

Durante los periodos de pre-anotación, se realizaron reuniones semanales de evaluación y revisión de la anotación. En caso de no llegar a acuerdos o no contar con el conocimiento necesario para tomar una decisión respecto a una mención o la clasificación de un hallazgo, médicos y médicas especialistas en radiología y medicina nuclear, con expertise en la redacción e interpretación de este tipo exámenes, revisaron las frases y textos para apoyar la toma de decisiones respecto al etiquetado.

El proceso de anotación consideró la anotación a nivel de entidades, y la anotación a nivel de documento. La anotación de entidades se realizó para cada mención de metástasis a distancia afirmativa o negada que coincidiera con la notación de la clasificación TNM para tumores malignos (14). Además, entendiendo que los hallazgos en un reporte pueden tener un grado de incertidumbre, se añadió un atributo para este tipo de menciones.

En la anotación a nivel de documento, si un documento contenía al menos una mención afirmativa de metástasis a distancia, el documento completo se clasificaba como M1. Si no había menciones de metástasis a distancia afirmativas y había al menos una mención catalogada como M1 con incertidumbre, el documento completo se clasificaba como “M1 con incertidumbre”. Y, por último, si sólo había menciones de metástasis a distancia negadas en un mismo documento o no había menciones etiquetadas, se clasificaba como M0.

Acuerdo entre anotadoras

Para medir la calidad de las anotaciones se etiquetó en duplicado un 20% de los reportes, y se empleó el *F1-score,* según lo descrito por Dalianis (29), para determinar el grado de acuerdo entre las anotadoras al asignar una clase a las menciones.

El grado de acuerdo se midió de forma “estricta” y “relajada”. En la forma estricta, tanto los límites de la anotación como la clase asignada a la mención deben ser idénticos. En la relajada, los límites pueden no ser idénticos presentando algún grado de solapamiento, mientras que la clase asignada debe ser idéntica.

Curado de los textos anotados

Luego de medir el acuerdo entre anotadoras, se realizó el curado de los textos anotados. En esta etapa se corrige la pertinencia de la anotación, los límites y la clase asignada. Esto quiere decir que se revisó cada reporte anotado y se corrigieron las menciones que no se ajustaran a las reglas de la guía de anotación.

Esquema de anotación

El esquema de anotación incluye dos entidades y un atributo. Las entidades corresponden a las menciones de metástasis a distancia en dos casos particulares (en los ejemplos, las menciones están resaltadas en gris):

- M0: Cuando una mención de lesión metastásica a distancia se encuentra negada o es descartada, por ejemplo:

"no se identifican masas ni adenopatías hipermetabólicas".

- M1: Cuando se encuentra una mención afirmativa de una lesión metastásica a distancia, por ejemplo:

"se observan múltiples lesiones de carácter secundario en hígado".

El atributo de incertidumbre se utilizó solo para la mención M1. Este se añade cuando una mención de lesión metastásica a distancia tiene un carácter incierto, registrado explícitamente en el texto. Por ejemplo:

"se observan lesiones levemente hipermetabólicas de carácter indeterminado".

4.2.- Entrenar modelo de aprendizaje de máquinas para detectar automáticamente las menciones de metástasis a distancia al interior de reportes imagenológicos.

En la tarea de reconocimiento de entidades nombradas se empleó como línea de base, un algoritmo basado en reglas junto con un sistema de detección de negación. Adicionalmente, se implementó un modelo de aprendizaje profundo.

Para desarrollar estos modelos, primero se generó una partición del corpus anotado en un *subset* de entrenamiento (80% del corpus), uno de validación (10%) y otro de prueba (10%). El primero permitió el entrenamiento del modelo, el segundo permitió hacer los ajustes en el entrenamiento y el tercero se utilizó para la validación final del rendimiento de este. El conjunto de prueba no se utilizó en el entrenamiento de los modelos.

Los dos modelos se entrenaron para reconocer las entidades M1 (mención de metástasis afirmativa o incierta) y M0 (mención de metástasis negada). Adicionalmente, se entrenó un tercer modelo basado en aprendizaje profundo que aborda el problema de la incertidumbre como una entidad M1 con incertidumbre separada de la clase M1.

Algoritmo de NER basado en reglas

El algoritmo basado en reglas fue utilizado como línea de base para la evaluación del rendimiento del modelo de aprendizaje profundo. Este algoritmo se basó en la construcción de un lexicón con las palabras o frases asociadas a lesiones metastásicas. Posteriormente se realizó un emparejamiento o *match* de cada palabra/frase del lexicón con aquellas dentro de los reportes. Se asignó la clase a cada entidad emparejada según el sentido de la oración.

Construcción del lexicón

Para la construcción del lexicón, o listado de palabras que hacen referencia a metástasis a distancia, se utilizó las palabras que fueron etiquetadas por las anotadoras para el entrenamiento del modelo de aprendizaje profundo.

Para la construcción del lexicón se utilizaron las menciones anotadas en los subconjuntos de entrenamiento (80%) y validación (10%), y que se emplearon también en el entrenamiento de los modelos de aprendizaje profundo, lo que permitió la comparación de ambos modelos ante un mismo dataset. Las palabras y frases únicas se seleccionaron y se transformaron a minúscula para facilitar el emparejamiento.

Emparejamiento o *matching*

Para realizar el *match* entre las palabras del lexicón y el texto, se utilizó la librería spaCy (54). Con esta librería se dividió el texto de cada reporte en oraciones y en cada oración se identificó la palabra que coincidía exactamente con alguna del lexicón. Tanto el lexicón como el texto del reporte fueron procesados para que todas las letras estuvieran en minúscula y facilitar el emparejamiento. En caso de que hubiera dos palabras o frases que coincidieran con alguna del lexicón en la misma frase del texto, la librería mantenía aquella que tenía más caracteres. La duplicación podría haber provocado una sobreestimación de la clase de aquellas menciones.

Negación

Para definir si una mención de metástasis a distancia se encontraba negada, se incorporó un algoritmo basado en reglas, que se inspira en el algoritmo "NegEx" de Chapman et al. (55), llamado NegEx-MES (56) para negaciones en español. Este observa la palabra en el contexto de la oración y le asigna un sentido (oración afirmativa o negada).

Asignación de clase de las entidades

La asignación de una clase a las entidades reconocidas, se realizó en base al algoritmo NegEx-MES. Si una mención de metástasis era afirmativa, se catalogó como M1 y si era negativa, se catalogó como M0. Esto permitió obtener la proporción de entidades afirmativas y negadas por cada documento, en otras palabras, la proporción de menciones M1 y M0.

Modelo de NER basado en aprendizaje profundo

El modelo de aprendizaje profundo se implementó mediante un modelo basado en Redes Neuronales Recurrentes (RNR). En particular, se empleó la arquitectura propuesta inicialmente por Huang et al. (57)y optimizada por Lample et al. (58); una red neuronal recurrente bidireccional de gran memoria a corto plazo (Bi-LSTM) con una capa de decodificación basada en *conditional random fields (CRF)*, ver Figura

4.4. Se incorporaron embeddings pre-entrenados, a nivel de carácter y contextualizado, según lo descrito por Akbik et al. (59), ya que los *embeddings* contextualizados contribuyen a una mejora sustancial en la tarea de NER. En la capa de representación vectorial de las palabras fue concatenado un *embedding* contextual entrenado en un corpus clínico de Lista de Espera, desarrollado por Báez et al. (60), en conjunto con un *embedding* pre-entrenado, en el mismo corpus. Como estas representaciones se obtienen a partir de un modelo del lenguaje a nivel de carácter, los *embeddings* a ese nivel son útiles para reconocer palabras fuera del vocabulario o palabras mal escritas.

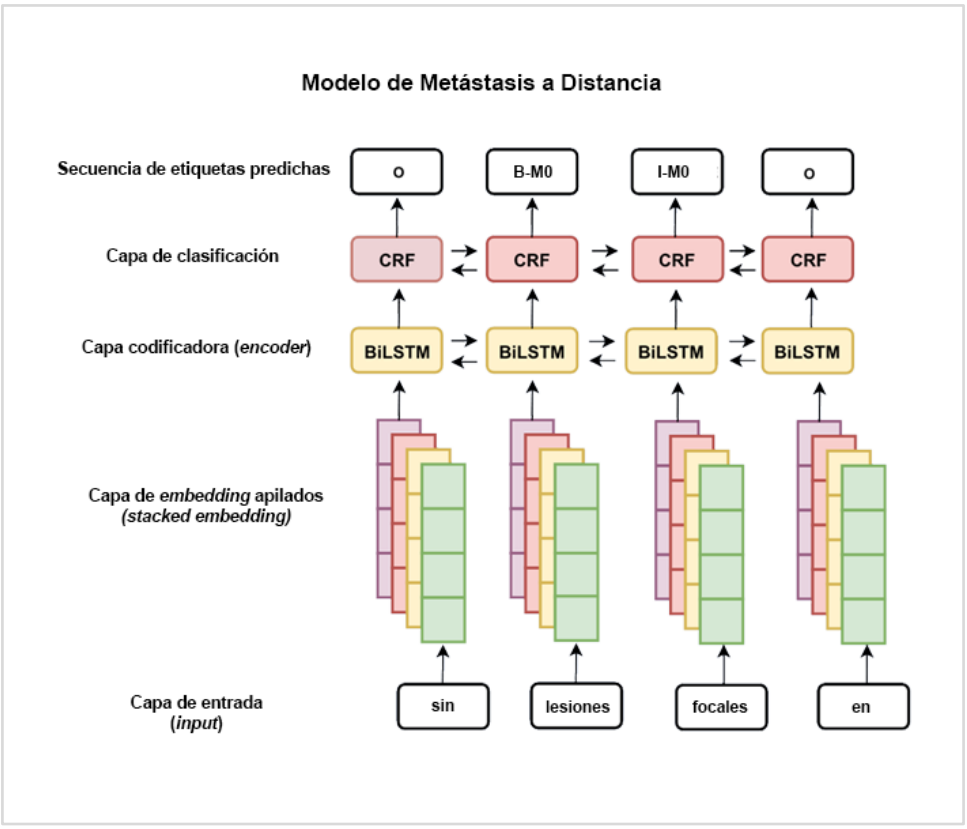

Figura 4.4. Representación del modelo Bi-LSTM con capa de clasificación de CRF, se muestra un ejemplo de identificación de una entidad M0. Adaptado de Villena et al. (61).

Para facilitar la implementación de la red neuronal recurrente BiLSTM, se utilizó la librería Flair (62), especialmente diseñada para problemas de clasificación y NER y que permite un manejo sencillo de *embedding*.

Esta arquitectura fue utilizada por Solarte Pabón et al. (48), que aborda un problema análogo, con una entidad definida como *“Cancer Entity”.* Este trabajo obtuvo mejores resultados con una capa de *embedding* previamente entrenado en corpus clínico.

A continuación, presentamos las razones para elegir esta arquitectura, y las ventajas que entrega el uso de redes neuronales recurrentes en nuestro trabajo:

- Este modelo le da importancia al orden de las palabras y su contexto y es efectivo analizando dependencias de larga distancia (33). Las dependencias de larga distancia son frecuentes en el texto clínico especializado, que tiende a tener oraciones más largas que el lenguaje común (63). Clínicamente, los hallazgos de metástasis a distancia dependen de cuál es el órgano involucrado según el cáncer primario. Es decir, la oración completa en la que se encuentra la mención cobra relevancia para predecir correctamente las entidades.

- En las menciones de metástasis no se registra explícitamente la palabra “metástasis” o derivados de la misma, sino que, se describen las lesiones con una alta variabilidad según el tipo de lesión, el órgano involucrado, el tipo de hallazgo por examen, las características de la señal observada por los médicos, más la variabilidad propia de la escritura de cada médico, lo que conlleva una dificultad para la implementación de un algoritmo basado en reglas. Esto explica por qué se utilizarán redes neuronales como modelo predilecto, evitando además la ingeniería de características.

- Con los reportes disponibles, se calculó obtener aproximadamente 17.000 a 20.000 oraciones, con aproximadamente 4.000 anotaciones, una cantidad de datos que permite entrenar un modelo de aprendizaje profundo, según se observa en el artículo de Solarte Pabón (48).

- El estado del arte en la tarea de NER es el empleo de redes neuronales. Específicamente, el modelo de BiLSTM-CRF, en conjunto con el uso de embeddings contextualizados fue el estado del arte en NER en el año 2018 (59).

- La clasificación de los hallazgos de metástasis como afirmativos, inciertos o negados, depende de su contexto. Si los moduladores se encuentren a una

distancia que excede la dependencia al contexto de los algoritmos empleados, se necesitan modelos más robustos, que cuenten con dependencias de larga distancia.

- Finalmente, existe experiencia en la aplicación de los modelos de aprendizaje profundo en el Grupo de Procesamiento de lenguaje natural del Centro de Modelamiento Matemático, con el cual se está realizando este trabajo de investigación. El modelo de BiLSTM-CRF ha sido evaluado en trabajos anteriores y en revisión. Se destacan dos artículos, el artículo en revisión de Villena et al. (61) donde se detectó entidades de morfología y topografía en reportes de anatomía patológica de FALP con F1-score de 0,857 y 0,895. Por otro lado, una investigación de Báez et al. (64) para la extracción de entidades anidadas en interconsultas en español, utilizando una capa de embeddings apilados, donde se conjugan embeddings de dominio general, a nivel de carácter y contextualizados.

Manejo del formato de los datos

La información anotada se extrajo desde INCEpTION (52) en formato JSON. Este formato incluye los metadatos, las anotaciones y el texto clínico. Posteriormente, las anotaciones se transformaron al formato *brat standoff* (65), que consiste en la lista de menciones asociadas a la posición en el texto y la clase asignada por las anotadoras, este último archivo es de tipo *.ann* (Figura 4.5*).*

```
T1    M0 128 162  adenopatías cervicales sospechosas
T2    M0 342 353  adenopatías
T3    M0 529 534  masas
T4    M0 839 872  lesiones focales hipermetabólicas
T5    M0 991 1010 focos hipercapantes
T6    M0 1392 1403      adenopatías
T7    M1 1704 1737      Lesiones blásticas hipercaptantes
T8    M1 1897 1932      lesiones osteoblásticas secundarias
```

Figura 4.5. Ejemplo de un documento anotado en formato .ann.

Finalmente, los archivos *.ann* fueron convertidos al formato estándar de NER, llamado CoNLL (66,67), mediante la librería *standoff2conll* (68). Este formato consiste en un archivo de texto, donde la primera columna corresponde a la división de todo el texto en tokens, con cada token por fila y cada oración separada en un salto de línea. La

segunda columna corresponde al tipo de entidad anotada o predicha con las etiquetas en formato BIO (Figura 4.6).

```
izquierda   O
,     O
secundarias O
.     O

No    O
se    O
identifican O
otras O
lesiones    B-M0
con   I-M0
sobreexpresión    I-M0
del   I-M0
PSMA  I-M0
.     O

Se    O
administra  O
```

Figura 4.6. Ejemplo de una documento anotado en formato CoNLL.

Definición de parámetros y entrenamiento del modelo.

Se documentaron 23 experimentos que consistían en el entrenamiento de modelos de aprendizaje profundo para la tarea de reconocimiento de entidades nombradas, los que consideraban combinaciones de los siguientes criterios: tipo de examen, tipo de cáncer, curado de los exámenes, tipo de *embeddings*, *learning rate* y número de épocas. Con estos experimentos se definieron los parámetros a utilizar en la validación cruzada.

Validación Cruzada

Se empleó la validación cruzada de *k* iteraciones con *k*=10. Este tipo de validación cruzada consta de 10 particiones iguales del dataset con las que se realizan 10 experimentos. En cada uno de ellos se utiliza una de las particiones como subconjunto de testeo y el resto, para entrenamiento del modelo. Finalmente se registraron y promediaron los rendimientos medidos en precisión, exhaustividad y *F1-score*.

Modelo de NER basado en aprendizaje profundo con tres clases

Para este modelo se realizó el mismo preprocesamiento y la validación cruzada de 10 iteraciones para definir el mejor modelo para el entrenamiento, tal como en el modelo para dos clases. Además, incorpora la clase asociada al atributo de incertidumbre, por lo tanto, las clases utilizadas para el entrenamiento fueron: M1, M0 y M1 incierto.

Análisis de errores

El análisis de errores del modelo de reconocimiento de entidades nombradas se basó en un trabajo propuesto por Nejadgholi et al. (69), donde se describen los distintos tipos de errores que pueden ocurrir para esta tarea. Estos se producen cuando hay diferencia entre la anotación y la predicción, es decir, cuando la clase de la entidad reconocida por el modelo no coincide en la clase etiquetada por humanos, con la posición o límites del carácter de inicio y término de la entidad en el texto (*span*, en inglés) o ambas.

Según lo anterior, se observan cinco tipos de error que se describen a continuación:

- **Error tipo 1: Falso positivo completo (FP).** Una entidad es reconocida por el modelo de NER pero no fue anotada por humanos.
- **Error tipo 2: Falso negativo completo (FN).** Una entidad anotada por humano no fue reconocida por el modelo de NER
- **Error tipo 3: Clase errónea, límites correctos**. Una entidad fue reconocida por el modelo de NER en la posición correcta dentro del texto, pero con una clase distinta a la anotada por humanos.
- **Error tipo 4: Clase errónea, límites sobrepuestos.** Una entidad fue reconocida por el modelo de NER, en una posición sobrepuesta y con una clase distinta a la anotación humana.

- **Error tipo 5: Clase correcta, límites sobrepuestos.** La clase de una entidad reconocida por el modelo de NER es igual a la anotación humana, pero en una posición sobrepuesta.

Anotación humana (entidad verdadera)	Se observan múltiples adenopatías hipercaptantes paravesicales izquierdas e ilíacas internas sospechosas de enfermedad secundaria.
Tipo de error	**Anotación por modelo de NER**
Tipo 1	Se observan múltiples adenopatías hipercaptantes paravesicales izquierdas e ilíacas internas sospechosas de enfermedad secundaria.
Tipo 2	Se observan múltiples adenopatías hipercaptantes paravesicales izquierdas e ilíacas internas sospechosas de enfermedad secundaria.
Tipo 3	Se observan múltiples adenopatías hipercaptantes paravesicales izquierdas e ilíacas internas sospechosas de enfermedad secundaria.
Tipo 4	Se observan múltiples adenopatías hipercaptantes paravesicales izquierdas e ilíacas internas sospechosas de enfermedad secundaria.
Tipo 5	Se observan múltiples adenopatías hipercaptantes paravesicales izquierdas e ilíacas internas sospechosas de enfermedad secundaria.

Tabla 4.2: Ejemplos para cada tipo de error en la predicción de entidades por un modelo de NER. Las clases según color son M1 (en gris claro) y M0 (en gris oscuro).

4.3 Desarrollar metodología para clasificar los reportes imagenológicos a nivel de documento.

La clasificación de los reportes a nivel de documento se realizó mediante dos aproximaciones, que dependen de la metodología utilizada para la tarea de NER. La primera para el algoritmo basado en reglas y la segunda, asociada a los modelos basados en aprendizaje profundo.

Clasificación a nivel de documento según algoritmo basado en reglas.

La clasificación a nivel de documento según el NER basado en reglas fue compleja, pues todos los reportes tenían al menos una mención afirmativa, por lo que, aplicar el flujo de decisión empleado para clasificar los reportes en el proceso de anotación no tenía sentido.

Para poder clasificar el documento, se utilizó la proporción de menciones afirmativas y negadas para un determinado documento, cuyo sentido fue predicho por el algoritmo NegEx MES. Debido a que tenemos una variable continua que debe ser clasificada en una variable categórica (M1 o M0), se utilizó la regresión logística para esa tarea. Como *input*, se incorporó la proporción de oraciones afirmativas y negadas del documento. Junto con la regresión logística, se evaluó el rendimiento de un modelo de máquina de soporte vectorial (*SVM*). Como *input* para ambos modelos se utilizó la proporción de oraciones afirmativas y negadas por documento.

Regresión logística

La regresión logística modela la probabilidad de que, dado un input numérico, un output corresponda a una clase o a otra. La probabilidad de el output corresponde a la combinación de variables independientes. Estas pueden ser binarias (1 o 0) o continuas (70).

Máquinas de soporte vectorial

Este modelo representa cada ejemplo en un espacio de dimensiones iguales al número de predictores o variables independientes, dividiendo cada espacio en dos subespacios. Una implementación básica de este modelo utiliza una línea recta para dividir el espacio, asegurando la maximización de la distancia entre las clases (70).

Clasificación a nivel de documento, modelos basados en aprendizaje profundo.

Para el modelo de dos clases, se utilizó un algoritmo de clasificación de cada reporte, a nivel de documento, basado en el razonamiento que hace el humano para realizar la anotación. Tal como se observa en la Figura 4.7, si en el texto se encontraba al menos una mención de M1, el reporte completo se clasificaba como M1. Si no había menciones de M1, el reporte se clasificaba como M0.

Para el modelo de tres clases se utilizó un flujo similar, pero incorporando la clase M1 incierto en la decisión (Figura 4.7) si en el texto se encontraba al menos una mención de M1, el reporte completo se clasificaba como M1. Si no había menciones de M1 y

se detectaba M1 incierto, el reporte se clasificaba como M1 incierto. Si no había menciones de M1 o M1 incierto, el reporte se clasificaba como M0.

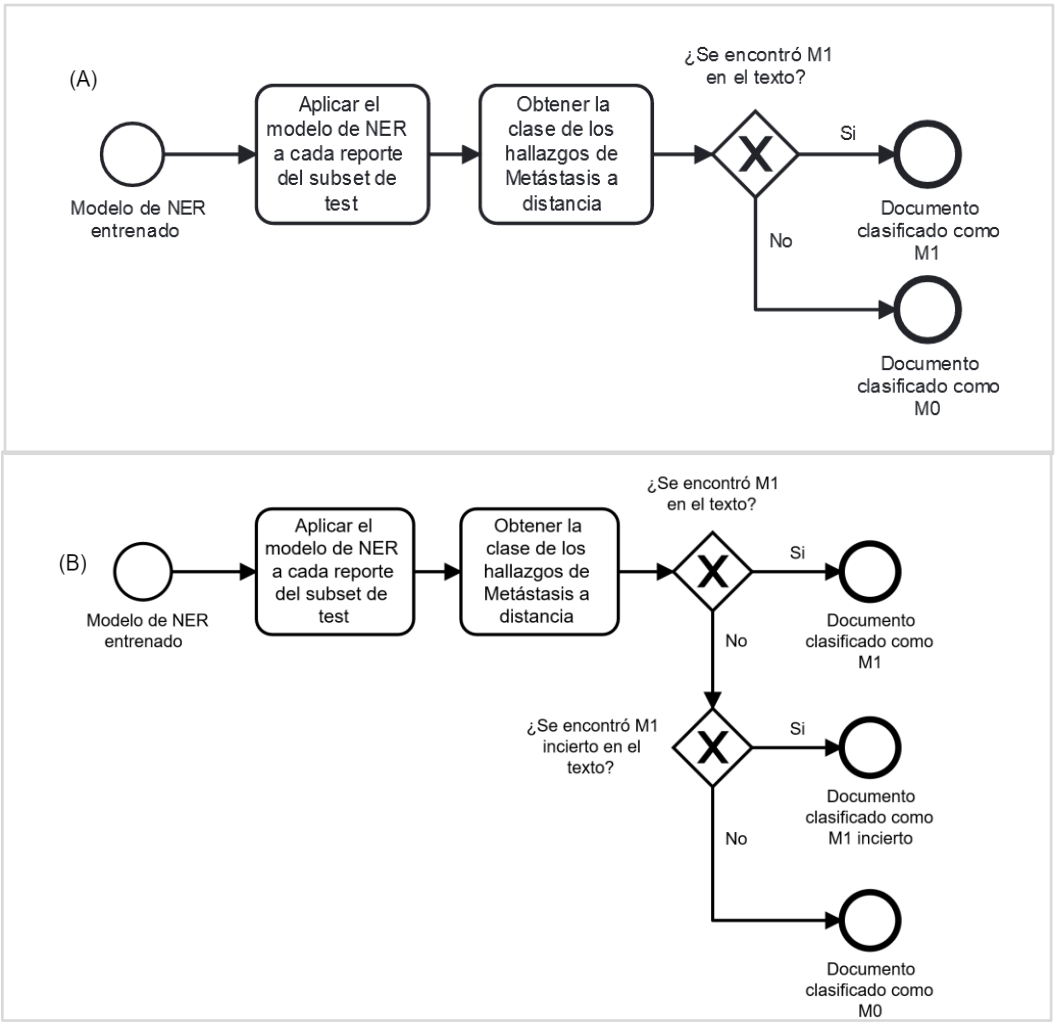

Figura 4.7*:* Flujo de clasificación a nivel de documentos en base al NER de metástasis a distancia. (A): modelo de dos clases. (B): modelo de tres clases.

Proporción de Metástasis

Debido a que con al menos una mención de metástasis a distancia afirmativa, ya sea M1 o M1 incierto, la clase del documento deja de ser M0, se introdujo una medida de proporción de hallazgos para metástasis a distancia que permite comparar la positividad de cada reporte, es decir, cuántos hallazgos tiene en función del total de anotaciones. Esta medida no es más que la proporción de entidades M1 o M1 inciertas respecto del total de entidades encontradas en un mismo reporte.

$$Proporción\ de\ metástasis = \ n° \ menciones \div total\ de\ menciones\ x\ 100$$

Con esto se puede diferenciar los reportes M1 o M1 incierto según su nivel de positividad de hallazgos de metástasis a distancia. Esto se basa en que no es lo mismo que un reporte tenga solo una mención M1 entre diez menciones totales (10%), que una que tenga cuatro menciones M1 entre ocho menciones totales (50%). En este caso, y asumiendo que el rendimiento de este modelo no es perfecto, el último caso tendría una mayor probabilidad de progresión a metástasis.

4.4. Evaluar el rendimiento de los modelos versus el conjunto de datos de prueba utilizando la métrica estándar micro *F1-score*.

Los algoritmos se evaluaron comparando las predicciones con las etiquetas reales del conjunto de testeo. Para poder realizar la comparación, se obtuvo la precisión, la exhaustividad y el micro *F1-score*.

- La precisión midió qué porcentaje de todos los reportes que fueron clasificados como M1 son realmente M1.
- La exhaustividad midió qué porcentaje de todos los reportes que efectivamente son M1 se logran identificar.
- El *F1-score* es una media armónica entre precisión y exhaustividad. Permitió obtener un valor de rendimiento general asociado a ambos indicadores. Es importante utilizar la versión micro *F1-score* por sobre el macro ya que existe un desbalance de las clases, y estas deben ser ponderadas según su frecuencia.

Además, para la tarea de clasificación a nivel de documento se incorporará el área bajo la curva ROC. Se espera que los modelos obtengan un *F1-score* similar a la literatura para realizar esta tarea, lo que corresponde a un *F1-score* mayor a 0,8.

5. RESULTADOS

5.1.- Desarrollo de un corpus anotado de las menciones de metástasis a distancia en reportes de exámenes imagenológicos de pacientes con cáncer de próstata, cáncer colorrectal y cáncer de mama.

Documentos anotados

De los 1.125 reportes proporcionados se anotaron 777 cuya distribución se detalla en la Tabla 5.1. Los 348 documentos restantes se descartaron debido a que durante la revisión preliminar se estableció que las resonancias magnéticas, en su gran mayoría, estaban enfocadas en el órgano del tumor primario y por lo tanto, no se podía evaluar metástasis a distancia. Además, otros exámenes tenían discordancias entre el tipo de reporte del documento y el contenido del texto, o no coincidieron con el tipo de reportes considerados en este estudio (como por ejemplo mamografías).

Tabla 5.1: Cantidad de reportes anotados según tipo de reporte y tipo de cáncer.

Tipo de Reporte	Cáncer de Próstata	Cáncer de Mama	Cáncer CR Colon y recto superior	Cáncer CR Recto Medio e Inferior	Total de Documentos
PETCT	193	86	53	8	340
RNM	48	12	35	9	104
TC	32	189	95	17	333
Total	273	287	183	34	777

Composición del dataset

La cantidad y proporción de las anotaciones por tipo de reporte se detalla en la Tabla 5.2. Se observa que la mayor cantidad de anotaciones fue obtenida desde los reportes de PETCT, con un 56,6% del total. Por el contrario, las RM solo representan un 3,45%.

Los PETCT poseen una extensión mayor que el resto de los exámenes, con alrededor de 550 tokens por cada documento y casi 13 menciones por reporte, mientras que se obtienen 9,4 anotaciones por examen desde TAC y 2,5 anotaciones desde RM.

Tabla 5.2: Cantidad y proporción de anotaciones y *tokens* según tipo de reporte.

	PETCT	TAC	RM	Total
N° de anotaciones (%)	4.405 (56,6)	3.120 (40,1)	264 (3,4)	7.789 (100,0)
Promedio de anotaciones por examen	12,95	9,4	2,53	10,02
N° Tokens	187.005	127.834	33.475	348.314
Promedio de tokens por examen	550	340	226	403

En la Figura 5.1 se observa la proporción total de menciones de metástasis a distancia, según clase. Se destaca un desbalance en las clases anotadas, con predominio de la clase M0.

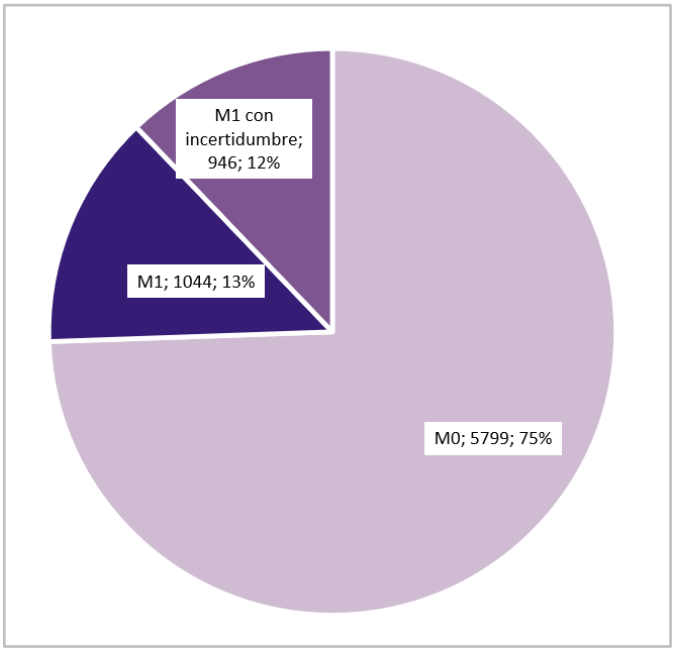

Figura 5.1. Cantidad y proporción de las menciones de metástasis a distancia anotadas según clase. Todos los tipos de cáncer y tipos de reporte.

Al analizar la distribución de entidades por tipo de reporte (Figura 5.2), se observa que la proporción de M0 es mayor en los TAC, con un 78% aproximadamente. En este mismo tipo de reporte se destaca una mayor proporción de entidades M1 con incertidumbre (14,8%). La mayor proporción de entidades M1 fue de 17,6% y se encontró en los reportes de PETCT.

Si se analiza la proporción de entidades según tipo de cáncer (Figura 5.3), se observa que casi un 20% de las entidades de cáncer de colon corresponde a M1, casi el doble de lo anotado para cáncer de mama. Debido a eso es que la proporción de entidades M0 en cáncer colorrectal es la más baja.

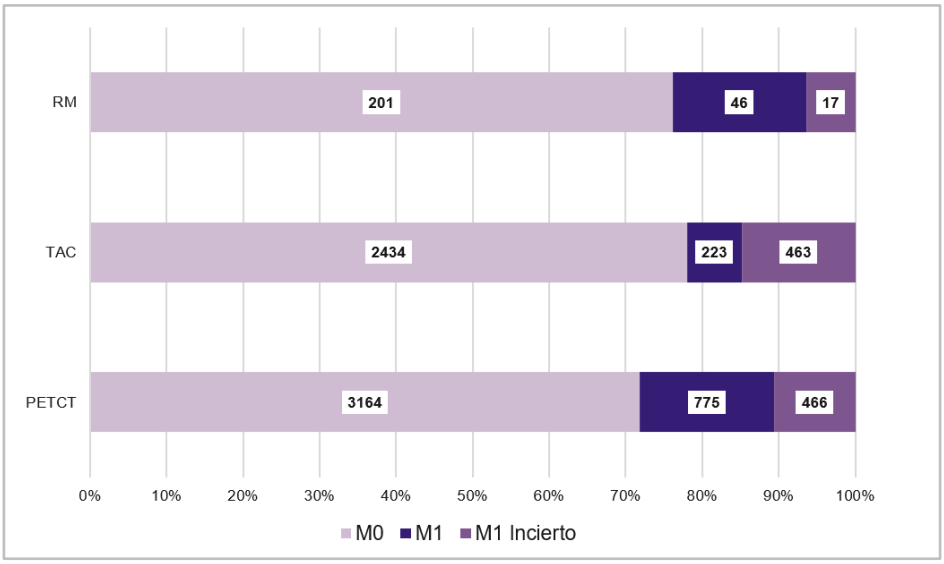

Figura 5.2 Cantidad y distribución de menciones anotadas de metástasis a distancia según clase y tipo de reporte (A) y clase y tipo de cáncer (B).

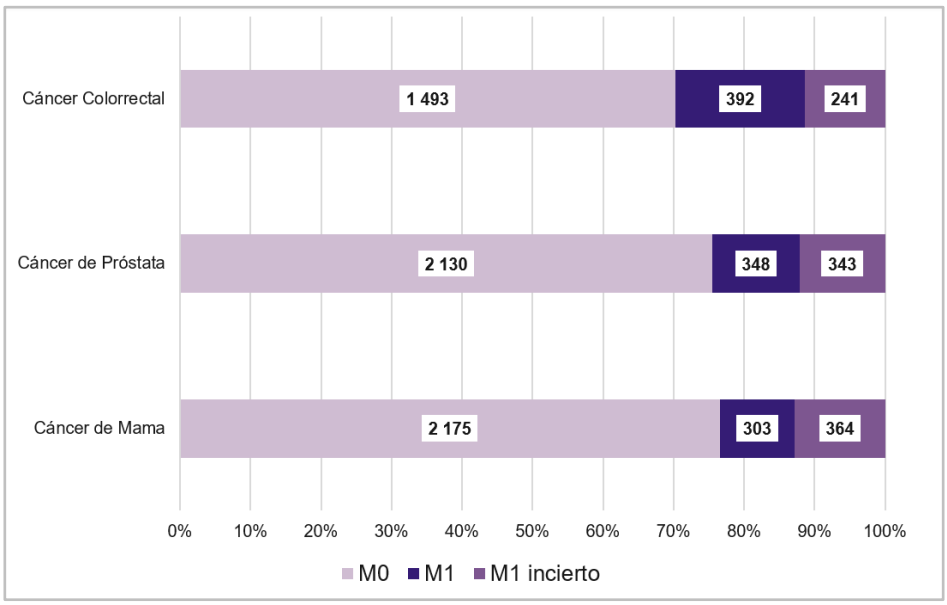

Figura 5.3 Cantidad y distribución de menciones anotadas de metástasis a distancia según clase y tipo de cáncer.

Evaluación del acuerdo entre-anotadoras

El acuerdo entre anotadoras, evaluado en F1-score, fue superior a 0,9, tanto en la medida estricta como en la relajada (Tabla 5.3). Se observa un aumento de esta métrica a medida que se avanzó en las rondas de anotación (Figura 5.4). Se puede inferir, que hubo mayor dificultad en el etiquetado de los documentos de PETCT y en aquellas entidades de clase M1, lo que se evidencia el menor acuerdo medido en F1-score.

Tabla 5.3: Evaluación del acuerdo entre anotadoras. PETCT y TC.

Tipo de coincidencia	Tipo de Reporte	Precisión	Exhaustividad	*F1-score*	Entidades testeadas (n)
Estricta	PETCT	0,90	0,91	0,90	292
	TAC	0,96	0,98	0,97	696
Relajada	PETCT	0,94	0,95	0,95	210
	TAC	0,97	0,98	0,98	692

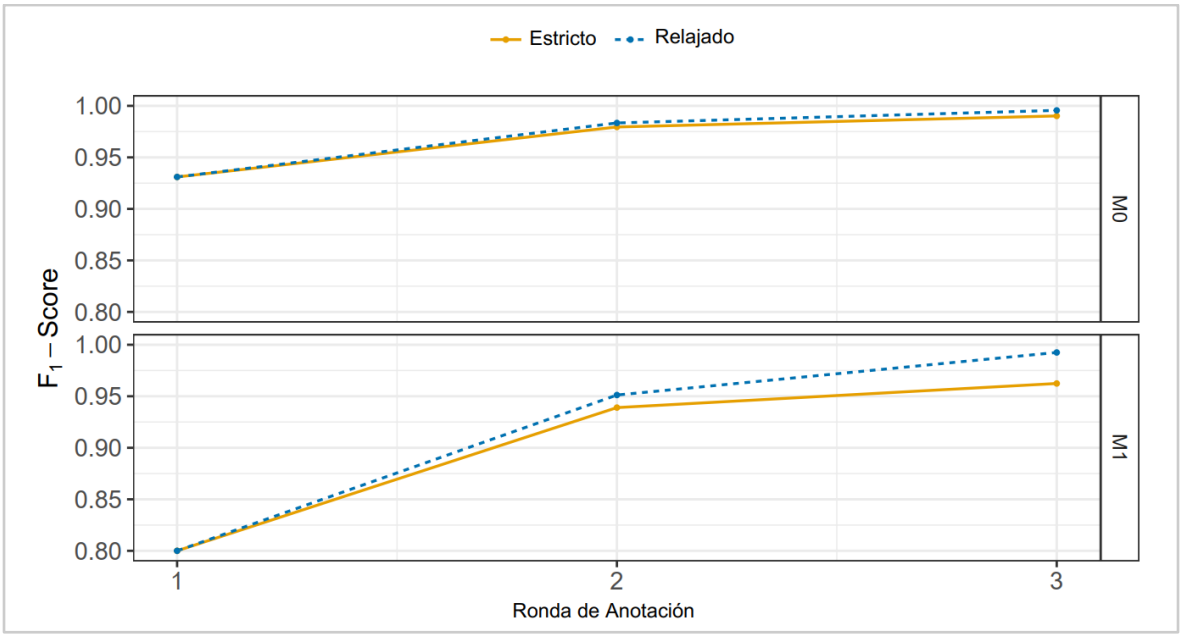

Figura 5.4. Evolución del acuerdo entre anotadoras medido en F1-score, según ronda de anotación y clase. Ronda 1: PETCT. Ronda 2: TC. Ronda 3: PETCT.

5.2.- Entrenamiento modelo de aprendizaje de máquinas para detectar automáticamente las menciones de metástasis a distancia al interior de reportes imagenológicos.

Modelo de NER basado en reglas

Este modelo se testeó en el mismo subconjunto de prueba del conjunto datos de entrenamiento del modelo de NER basado en aprendizaje profundo.

Rendimiento del NER basado en reglas

Los resultados del NER basado en reglas se detallan en la Tabla 5.4. Estos resultados corresponden al emparejamiento realizado con la librería spaCy de Python entre las menciones de metástasis a distancia contenidas en el lexicón y el texto de los reportes del conjunto de testeo. Como se puede observar, el modelo tiene un rendimiento pobre en el reconocimiento de las menciones M0 y M1, con *F1-scores* de 0,50 y 0,36 respectivamente.

Tabla 5.4. Rendimiento del modelo para la tarea de NER basada en reglas.

	Precisión	Exhaustividad	F1-Score	Entidades testeadas (n)
M0	0,54	0,47	0,50	567
M1	0,25	0,63	0,36	171
Media Micro	0,41	0,51	0,45	738
Media Macro	0,40	0,55	0,43	738
Media balanceada.	0,47	0,51	0,47	738

Modelo de NER basado en aprendizaje profundo con dos clases

Modelo escogido para la tarea de NER de dos clases

En la Tabla 5.8 se observan las métricas del modelo escogido para la tarea de NER basado en aprendizaje profundo. La decisión se basó en el modelo con mejor rendimiento en *F1-score* para la clase M1 y en el experimento que tenía una mejor media balanceada. Además, este experimento obtuvo la mayor media de precisión y

la segunda mayor media de exhaustividad. Al analizar por clase, este modelo obtuvo el mejor rendimiento en *F1-score*, la mayor precisión y la tercera mejor exhaustividad, para la clase M1.

Tabla 5.8. Rendimiento del modelo seleccionado para la tarea de NER de dos clases

	Precisión	Exhaustividad	*F1-score*	Entidades testeadas (n)
M0	0,8912	0,8959	0,8936	567
M1	0,7410	0,7193	0,7300	171
Media Micro	0,8573	0,8550	0,8562	738
Media Macro	0,8161	0,8076	0,8118	738
Media balanc.	0,8564	0,8550	0,8557	738

Validación cruzada con 10 iteraciones

Con base en los parámetros seleccionados, se realizó la validación cruzada con 10 iteraciones para evaluar la generalización del modelo. Los resultados de rendimiento medio se observan en la Tabla 5.9. A nivel general se destaca un mejor rendimiento para la clase M0, que supera en casi un 20% el rendimiento de la clase M1. Las métricas obtenidas para cada experimento de la validación cruzada se observan en las Tablas 5.10 y 5.11, paras la clase M0 y M1. Los resultados se encuentran ordenados en las Tablas, según el rendimiento en F1-score. Para la clase M0 los resultados mínimo y máximo en *F1-score* fueron de 0,8698 y 0,897. Para la clase M1, 0,625 y 0,730, con la que se observa una mayor diferencia

Este dataset tiene una proporción entre las clases M0 y M1 de 3:1, por lo tanto, la media que mejor se ajusta, para evaluar el rendimiento del modelo, es la media balanceada. Se destaca un rendimiento para el modelo de 0,8282 medido en *F1-score*, con una exhaustividad que supera levemente a la precisión.

Tabla 5.9. Medias de rendimiento de la validación cruzada para el modelo de NER de dos clases.

Medias de rendimiento para ambas clases (M0 y M1)			
	Precisión	Exhaustividad	*F1-score*
Media micro	0,8243	0,8414	0,8326
Media macro	0,7832	0,7872	0,7846
Media balanceada	0,8205	0,8365	0,8282
Medias de rendimiento por clase			
M0	0,864	0,8965	0,8799
M1	0,69763	0,6725	0,6843

Tabla 5.10. Rendimiento individual por experimento de la validación cruzada para la clase M0.

Id exp.	Precisión	Exhaustividad	*F1-score*	Entidades testeadas (n)
6	0,8867	0,9076	0,897	595
1	0,8912	0,8959	0,8936	567
3	0,8591	0,9084	0,8831	557
8	0,8607	0,9055	0,8825	614
2	0,8521	0,9064	0,8784	534
10	0,8577	0,895	0,876	505
9	0,8544	0,8985	0,8759	542
7	0,8538	0,897	0,8749	534
5	0,8625	0,8835	0,8729	575
4	0,8618	0,8678	0,8648	575

Tabla 5.11. Rendimiento individual por experimento de la validación cruzada para la clase M1.

Id exp.	Precisión	Exhaustividad	*F1-score*	Entidades testeadas (n)
1	0,741	0,7193	0,73	171
9	0,7123	0,729	0,7206	214
8	0,6868	0,7396	0,7123	169
7	0,725	0,6744	0,6988	172
3	0,7129	0,6804	0,6963	219
4	0,6777	0,681	0,6793	210
6	0,7049	0,6355	0,6684	203
2	0,697	0,6389	0,6667	180
10	0,6833	0,6119	0,6457	201
5	0,6354	0,615	0,625	187

Modelo de NER basado en aprendizaje profundo con tres clases

Modelo escogido para la tarea de NER de tres clases

Se utilizó los mismos parámetros de entrenamiento que para el modelo de dos clases y las mismas particiones de entrenamiento, validación y testeo con el objetivo de comparar los rendimientos contra el dataset empleado en el modelo de aprendizaje profundo de dos clases y el modelo basado en reglas. El rendimiento de este modelo en particular se detalla en la Tabla 5.12, donde se destaca una media balanceada superior a 0,82.

Tabla 5.12. Rendimiento del modelo entrenado para la tarea de NER de tres clases.

	Precisión	Exhaustividad	*F1-score*	Entidades testeadas (n)
M0	0.8787	0.8942	0.8864	567
M1 con incertidumbre	0.6625	0.5889	0.6235	90
M1	0.5604	0.6296	0.5930	81
Media Micro	0.8168	0.8279	0.8223	738
Media Macro	0.7005	0.7042	0.701	738
Media balanc.	0.8174	0.8279	0.8221	738

Validación cruzada con 10 iteraciones

Los resultados de la validación cruzada se muestran en la Tabla 5.13. Al igual que en el modelo de dos clases, la media balanceada se ajusta mucho mejor, con un rendimiento general para el modelo de tres clases aproximadamente 2% menor respecto al modelo de dos clases. Las entidades M1 fueron divididas entre ciertas e inciertas, con esto se observó una disminución del rendimiento en ambas clases.

En las Tablas 5.14 a 5.16 se detallan los rendimientos individuales de cada iteración de la validación cruzada, según la clase testeada. Se destaca la exhaustividad de la clase M0, llegando a niveles superiores a 0,91 en dos experimentos y en *F1-score* con máximos y mínimos de 0,8966 y 0,8611 respectivamente. Para la clase M1, el mejor rendimiento en *F1-score* fue de 0,68 y para la clase M1 con incertidumbre fue de 0,71.

Tabla 5.13. Medias de rendimiento de la validación cruzada para el modelo de NER de tres clases.

Medias de rendimiento para las tres clases (M0, M1 con incertidumbre, M1)			
	Precisión	Exhaustividad	*F1-score*
Media micro	0.8018	0.8204	0.8110
Media macro	0.7011	0.7016	0.6994
Media balanceada	0.8003	0.8204	0.8093
Medias de rendimiento por clase			
M0	0.8607	0.8955	0.8777
M1 con incertidumbre	0.6385	0.5999	0.6169
M1	0.6040	0.6096	0.6037

Tabla 5.14. Rendimiento individual por experimento de la validación cruzada para la clase M0.

Id exp.	Precisión	Exhaustividad	*F1-score*	Entidades testeadas (n)
6	0.8827	0.9109	0.8966	595
3	0.8696	0.9102	0.8895	557
1	0.8787	0.8942	0.8864	567
5	0.8697	0.8939	0.8816	575
8	0.8587	0.9007	0.8792	614
9	0.8502	0.9004	0.8746	542
2	0.8459	0.9045	0.8742	534
7	0.8546	0.8914	0.8726	534
10	0.8428	0.8812	0.8616	505
4	0.8545	0.8678	0.8611	575

Tabla 5.15. Rendimiento individual por experimento de la validación cruzada para la clase M1.

Id exp.	Precisión	Exhaustividad	*F1-score*	Entidades testeadas (n)
3	0.6818	0.687	0.6844	131
7	0.7262	0.5922	0.6524	103
8	0.5957	0.7089	0.6474	79
9	0.576	0.7059	0.6344	102
6	0.6211	0.602	0.6114	98
1	0.5604	0.6296	0.5930	81
2	0.6712	0.5213	0.5868	94
10	0.5632	0.5568	0.56	88
5	0.5098	0.5652	0.5361	92
4	0.5354	0.5271	0.5313	129

Tabla 5.16. Rendimiento individual por experimento de la validación cruzada para la clase M1 con incertidumbre.

Id exp.	Precisión	Exhaustividad	*F1-score*	Entidades testeadas (n)
4	0.6897	0.7407	0.7143	81
8	0.7024	0.6556	0.6782	90
9	0.7416	0.5893	0.6567	112
5	0.6224	0.6421	0.6321	95
1	0.6625	0.5889	0.6235	90
6	0.6522	0.5714	0.6091	105
2	0.5862	0.593	0.5896	86
10	0.625	0.531	0.5742	113
3	0.5974	0.5227	0.5576	88
7	0.5065	0.5652	0.5342	69

5.3 Desarrollo de una metodología para clasificar los reportes imagenológicos a nivel de documento.

Clasificación a nivel de documento a partir de NER basado en reglas.

Detección de entidades afirmativas y negadas

El algoritmo NegEx MES (56) permitió identificar un claro predominio de entidades afirmativas en los documentos clasificados manualmente como M1, con un promedio de 10, frente a un promedio de 4 para aquellos clasificados como M0. Respecto al promedio de entidades negadas, se observaron en promedio 8 menciones negadas por documento para ambas clases. La diferencia en la cantidad de entidades afirmativas y negadas es más evidente para los documentos clasificados manualmente como M0 (Figura 5.5).

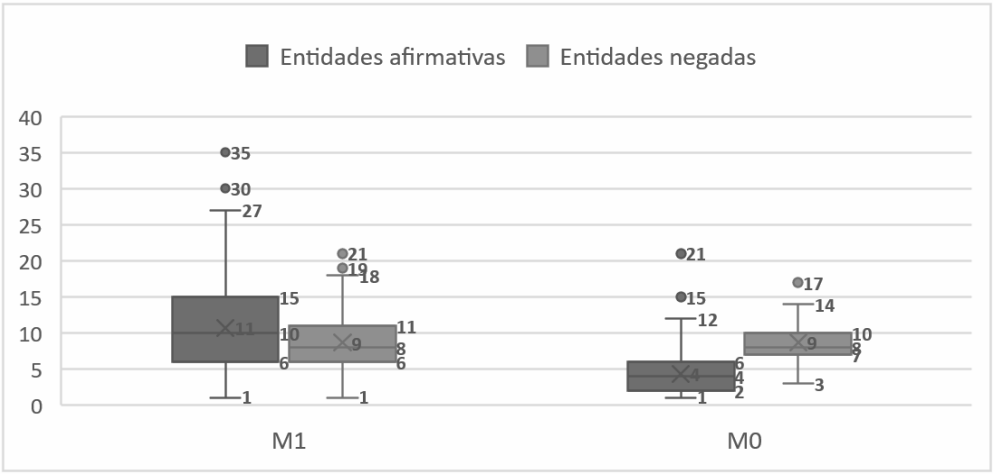

Figura 5.5. Distribución de entidades afirmativas y negadas por clase asignada por humano al documento. Algoritmo de NER basado en reglas, dos clases.

Regresión Logística

En la Figura 5.6 se presenta la matriz de confusión del modelo de regresión logística entrenado, empleando la cantidad de entidades afirmativas y negadas del punto anterior. Se observa una mayor cantidad de reportes clasificados como falsos positivos para metástasis a distancia, al contrario, solo hubo un reporte clasificado como falso negativo. En cuanto al rendimiento del modelo, se destaca la precisión de la clase M0 y la exhaustividad de la clase M1, con valores cercanos a 1 (Tabla 5.17). Lo anterior se ve reflejado en un área bajo la curva ROC de 0,80 (Figura 5.7).

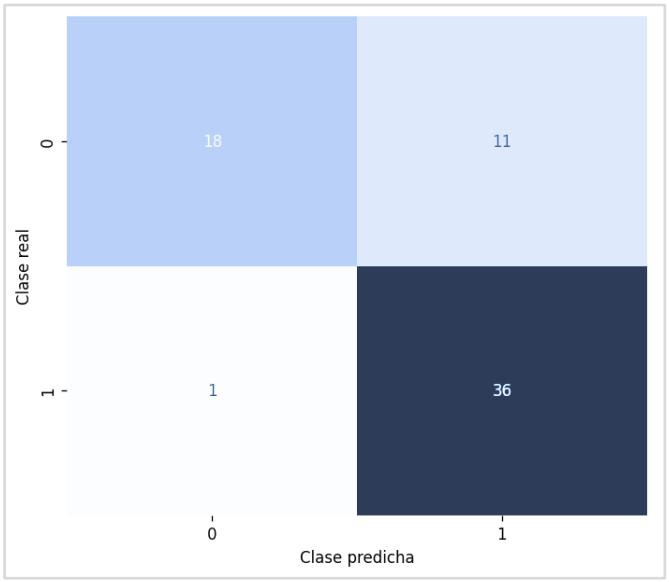

Figura 5.6. Matriz de confusión, modelo de regresión logística. Clasificación a nivel de documento según NER basado en reglas, dos clases. PETCT y CT. 0: M0, 1: M1.

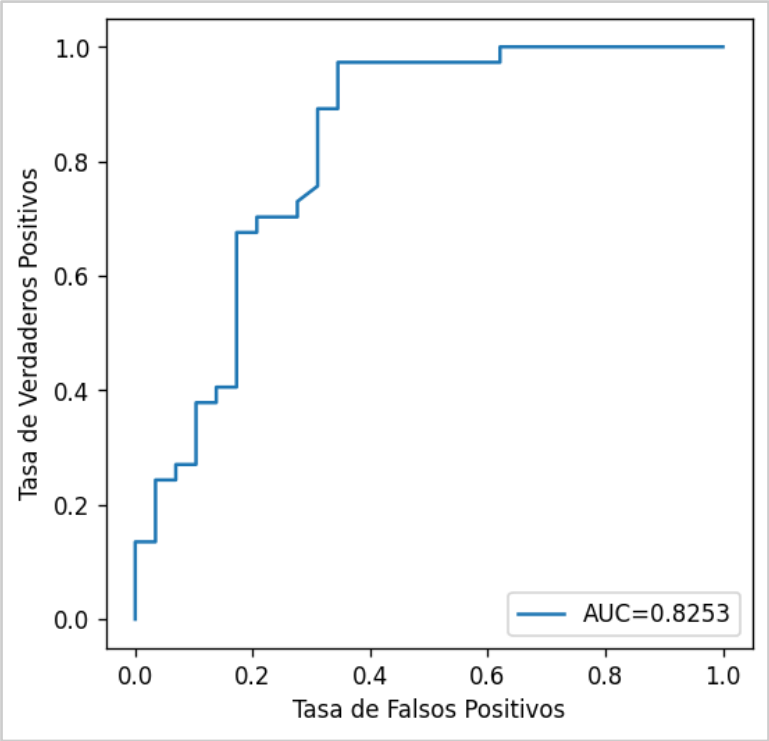

Figura 5.7. Área bajo la curva ROC (*AUC*), modelo de regresión logística. Clasificación a nivel de documento según NER basado en reglas, dos clases

Tabla 5.17. Rendimiento, modelo de regresión logística. Clasificación a nivel de documento según NER basado en reglas, dos clases

	Precisión	Exhaustividad	*F1-score*	Documentos testeados (n)
M0	0,95	0,62	0,75	29
M1	0,77	0,97	0.86	37
Media Micro			0.82	66
Media Macro	0,86	0,80	0,80	66
Media balanc.	0,85	0,82	0,81	66

Máquina de Vectores de Soporte

El rendimiento de este modelo se muestra en la matriz de confusión (Figura 5.8) y la Tabla de resultados de rendimiento (Tabla 5.18). Se observa una matriz de confusión más equilibrada entre falsos negativos y falsos positivos, con una leve disminución de los verdaderos positivos. Aun así, el área bajo la curva de la curva ROC es similar al modelo de regresión logística (Figura 5.9).

La media macro, que para comparar los modelos a nivel de documento cobra sentido, pues no existe desbalance de clases, es de 0,76, obteniendo 3 puntos menos que el modelo de regresión logística .

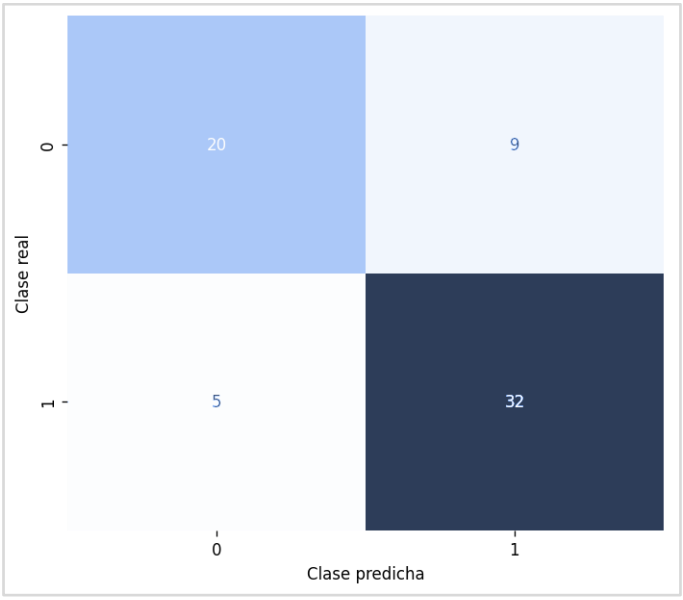

Figura 5.8. Matriz de confusión, modelo SVM. Clasificación a nivel de documento según NER basado en reglas, dos clases. PETCT y CT. 0 = M0, 1 = M1.

Tabla 5.18. Rendimiento, modelo SVM.
Clasificación a nivel de documento según NER basado en reglas, dos clases.

	Precisión	Exhaustividad	*F1-score*	Documentos testeados (n)
M0	0,80	0,69	0,74	29
M1	0,78	0,86	0.82	37
Media Micro			0,79	66
Media Macro	0,79	0,78	0,78	66
Media balanceada	0,79	0,79	0,79	66

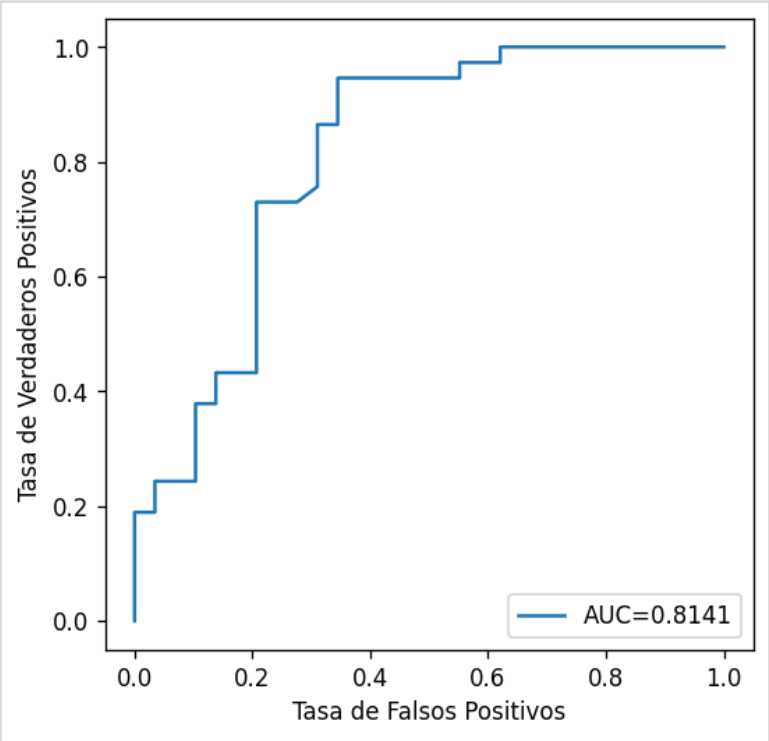

Figura 5.9. Área bajo la curva ROC (*AUC*), modelo de regresión logística. Clasificación a nivel de documento según NER basado en reglas, dos clases.

Clasificación a nivel de documento a partir de NER basado en aprendizaje profundo.

La clasificación a nivel de documento se realizó con base en las predicciones obtenidas a partir del modelo de NER. A continuación se presentan los resultados para los modelos de dos y tres clases (Tablas 5.19 y 5.20).

Clasificación a nivel de documento en base a modelo de NER de dos clases

Tabla 5.19. Rendimiento del algoritmo de clasificación del modelo de NER basado en aprendizaje profundo, con dos clases.

	Precisión	Exhaustividad	*F1-score*	Documentos testeados (n)
M0	0,7941	0,9310	0,8571	29
M1	0,9375	0,8108	0,8695	37

Clasificación a nivel de documento en base a modelo de NER de tres clases

Tabla 5.20. Rendimiento del algoritmo de clasificación del modelo de NER basado en aprendizaje profundo, con tres clases.

	Precisión	Exhaustividad	*F1-score*	Documentos testeados (n)
M0	0,8709	0,9319	0,9000	29
M1	0,7368	0,8750	0,8000	16
M1 con incertidumbre	0,8750	0,6667	0,7567	21

La división de la clase M1 permite gestionar la incertidumbre de los textos clínicos, detectando cuáles de ellos son interpretados como indeterminados o inciertos por el médico radiólogo. No obstante, el rendimiento disminuye entre 7 y 13 puntos porcentuales aproximadamente de *F1-score*. El mejor rendimiento se obtuvo para la clase M0 del modelo de tres clases, superando en casi un 5% al modelo de 2 clases.

Se destaca que para el modelo de 2 clases, la precisión baja 8 puntos aproximadamente.

Proporción de metástasis a distancia.

La proporción de metástasis a distancia para documentos anotados como M1 en los documentos clasificados como M1, tanto con dos como con tres clases, estuvo cercana al 30%. Los documentos clasificados como M1 incierto tuvieron un 17% aproximadamente (Tabla 5.21).

Se realizó la comparación de la proporción de metástasis a distancia en base a la clasificación asignada por el algoritmo versus la clase anotada con humanos (Figuras 5.10, 5.11 y 5.12). Se observa que los documentos clasificados por el algoritmo como M1, cuya clase anotada era M0 tuvieron una proporción de 21,43 y un 12,14% para los modelos de dos y tres clases, con un promedio de cantidad de menciones de 1 y 1.5 respectivamente. Este valor aumenta, cuando aquellos clasificados como M1 eran realmente M1, donde se obtuvo proporciones de 29,52 y 33,58% para dos y tres clases, respectivamente.

Los documentos clasificados como M1 incierto tienen un comportamiento similar, pues cuando el documento es clasificado como M1 incierto y el documento fue anotado como M1 incierto, la proporción de metástasis a distancia fue de un 2,46%, mientras que cuando el documento era M1 realmente, esta proporción, en promedio, sube a más de 15%. Lo que nos indica que cuándo se observa un mayor porcentaje de entidades M1 con incertidumbre, es posible que el documento sea M1 realmente.

Tabla 5.21. Puntuación de hallazgos de metástasis a distancia afirmativos del modelo de NER basado en aprendizaje profundo con dos y tres clases.

	Proporción media de entidades M1 (NER de 2 clases)	Proporción media de entidades M1 (NER de 3 clases)	Proporción media de entidades M1 con incertidumbre (NER de 3 clases)
Documentos anotados manualmente como M1	0,2902	0,306	0,0852
Documentos anotados manualmente como M1 con incertidumbre	-	-	0,1734

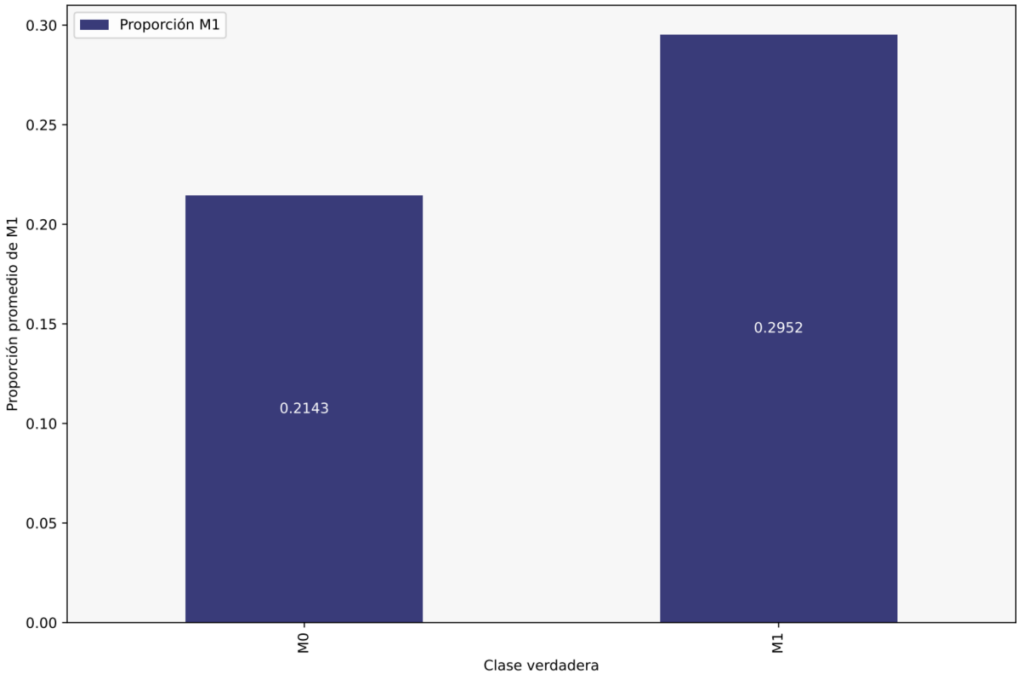

Figura 5.10. Media de la proporción de entidades M1 en documentos clasificados como M1 y M0, según clasificación real asignada por anotadores. Clasificación de documentos según NER basado en aprendizaje profundo, de dos clases.

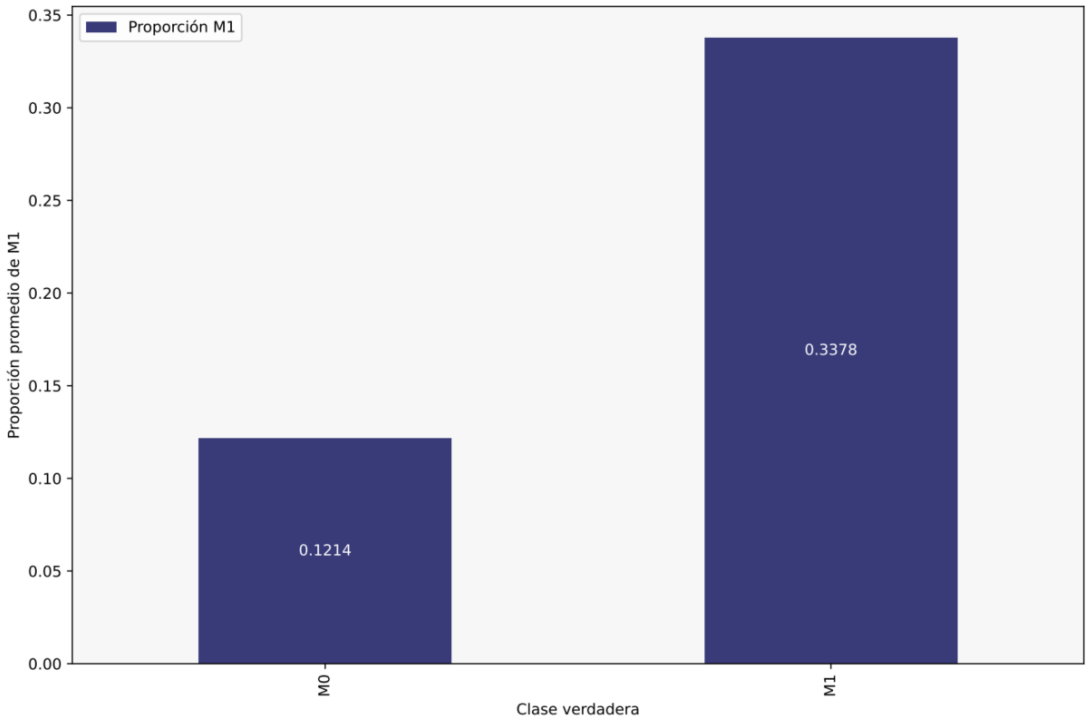

Figura 5.11. Media de la proporción de entidades M1 en documentos clasificados como M1, según clasificación real asignada por anotadores. Clasificación de documentos según NER basado en aprendizaje profundo, de tres clases.

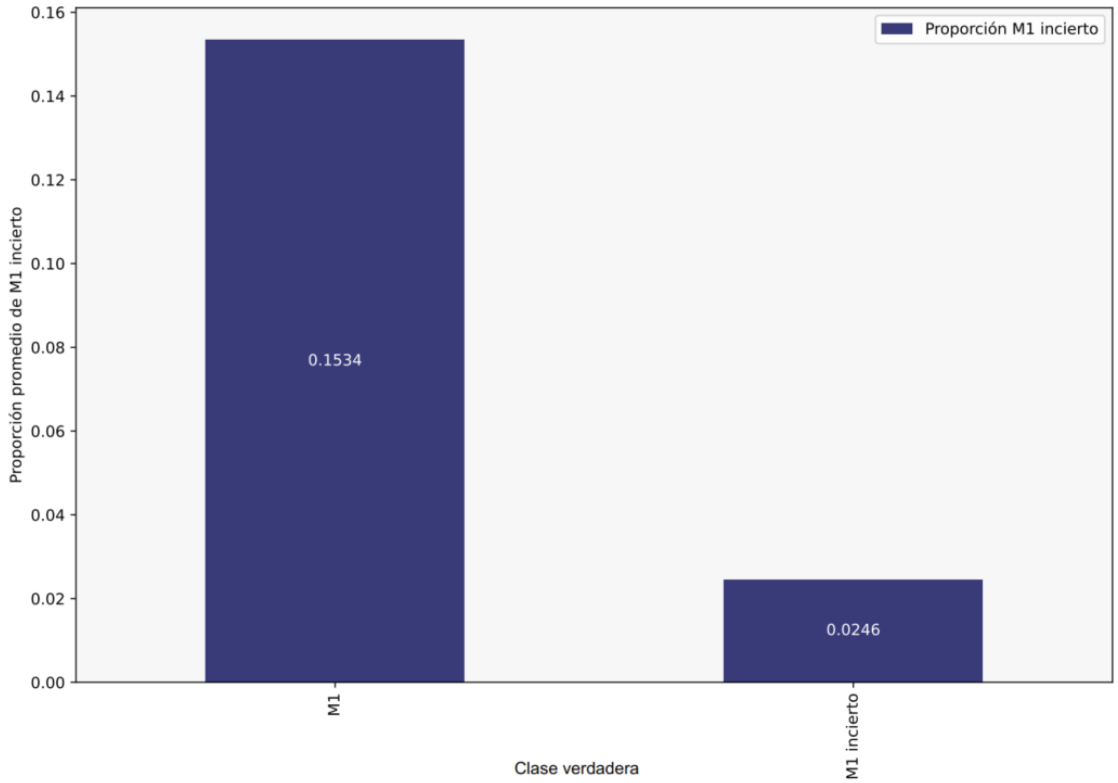

Figura 5.12. Media de la proporción de entidades M1 inciertas en documentos clasificados como M1 incierto, según clasificación real asignada por anotadores. Clasificación de documentos según NER basado en aprendizaje profundo, de tres clases.

5.4 Análisis de Errores

NER basado en Reglas

Este modelo identificó un total de 922 entidades, entre M1 y M0 (Figura 5.13), con una sobre detección en más del doble de entidades M1, en comparación con las entidades anotadas por humano.

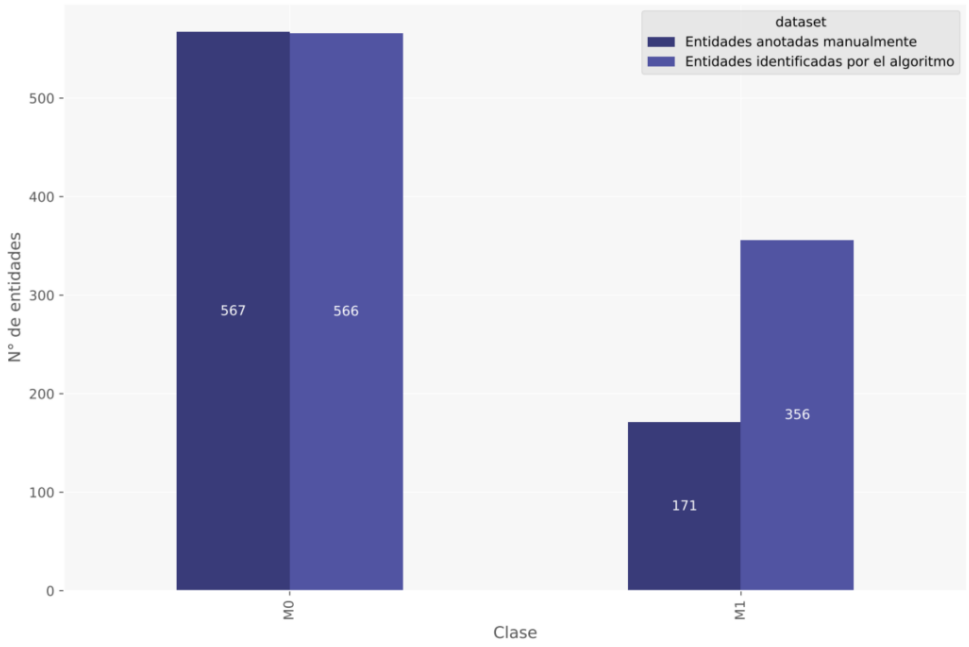

Figura 5.13. Cantidad de entidades según clase y conjunto de datos. Algoritmo de NER basado en reglas.

De las 922 entidades, solo 375 correspondían a verdaderos positivos. La mayor proporción de las 667 predicciones incorrectas corresponde a errores tipo 1 o falsos positivos (FP), con un 28,4% del total de predicciones (Figura 5.14). Esto explicaría la sobre detección de entidades M1, en comparación al conjunto de testeo. La cantidad y distribución dividida por clase, según el tipo de error, se observa en las Figuras 5.15 y 5.16 respectivamente.

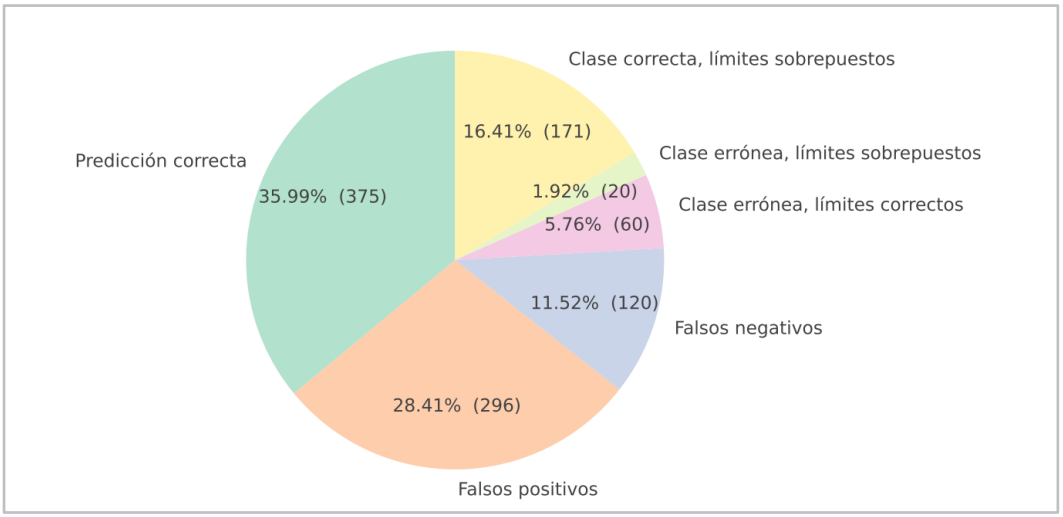

Figura 5.14. Proporción de entidades identificadas correcta e incorrectamente según tipo de error. Algoritmo de NER basado en reglas.

Los falsos negativos, es decir, aquellas entidades que no se pudieron detectar, corresponden a un 11.52% de las menciones con error.

Para las entidades M0, un 51.4% de los errores en la predicción corresponde al error tipo 5. Esto quiere decir que la clase se definió correctamente, pero los límites de la entidad (en términos de palabras) no coincidieron con lo anotado por humanos. Este tipo de error no representa mayor riesgo para la predicción ya que no cambia el resultado. Si bien las palabras no son exactamente las mismas, se identifica correctamente la entidad.

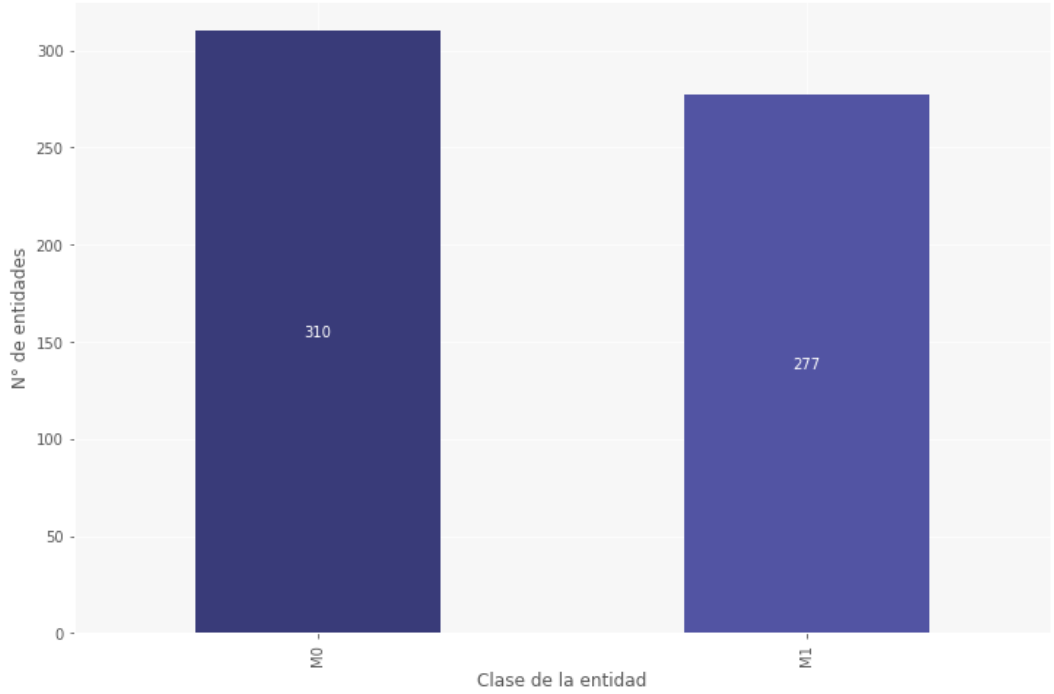

Figura 5.15. Cantidad de entidades identificadas incorrectamente según clase. Algoritmo de NER basado en reglas.

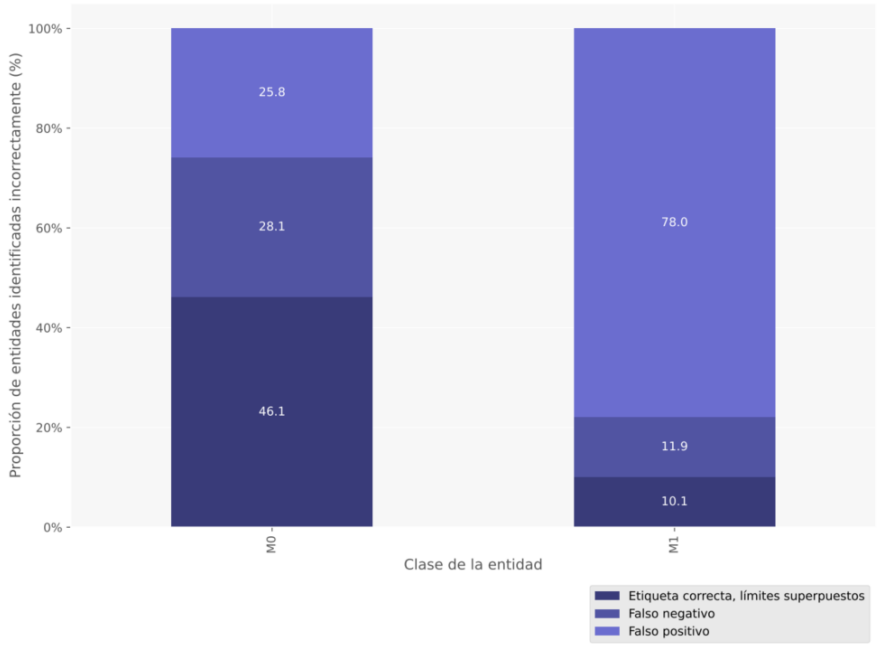

Figura 5.16. Proporción de entidades identificadas incorrectamente según clase y tipo de error. Algoritmo de NER basado en reglas.

Con respecto a las menciones identificadas con clase incorrecta, 76 entidades anotadas manualmente como M0 fueron identificadas como M1 por el modelo.

Mientras que solo 4 se identificaron como M0 cuando la clase verdadera era M1 (Figura 5.17). Al analizar en detalle las palabras de las entidades erróneamente identificadas como M1, se observa que la mayoría corresponden a “nódulos”, “adenopatías”, “quistes” y “linfonodos” (Tabla 5.22), palabras que en general no se escriben para “descartarlas”, es decir, que en general no están negadas y, por lo tanto, el método las identificó como M1.

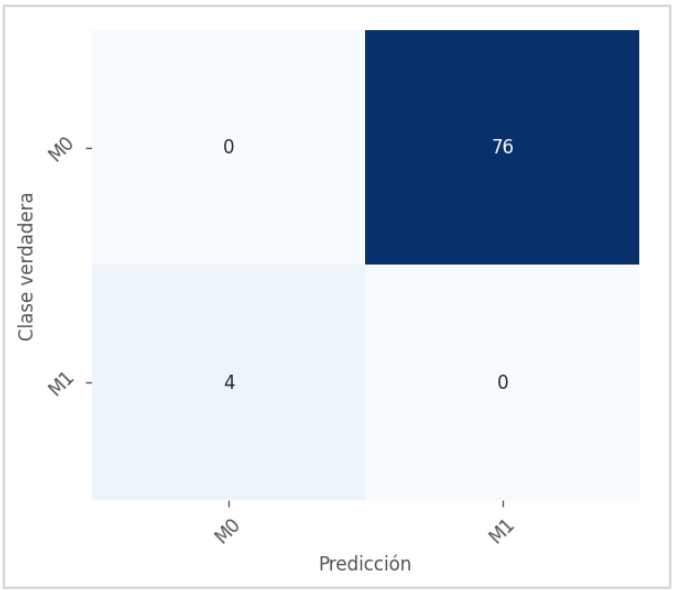

Figura 5.17. Matriz de confusión de entidades con clase errónea. Algoritmo de NER basado en reglas.

Tabla 5.22. Palabras identificadas como Falso Positivo y frecuencia. Algoritmo de NER basado en reglas.

Mención	Frecuencia (n)
“focos”	38
“adenopatías”	21
“linfonodos”	16
“sobreexpresión de psma”	15
“lesión”	14
“compromiso”	10
“ganglios”	10
“nódulo”	8
“masa”	7
“linfonodo”	6

NER de 2 clases basado en Aprendizaje Profundo.

El modelo de aprendizaje profundo, basado en la red neuronal Bi-LSTM con CRF, detectó la misma cantidad de entidades que las anotadas por humanos (Figura 5.18). Al analizar por clase, se observa una leve diferencia con el conjunto de testeo.

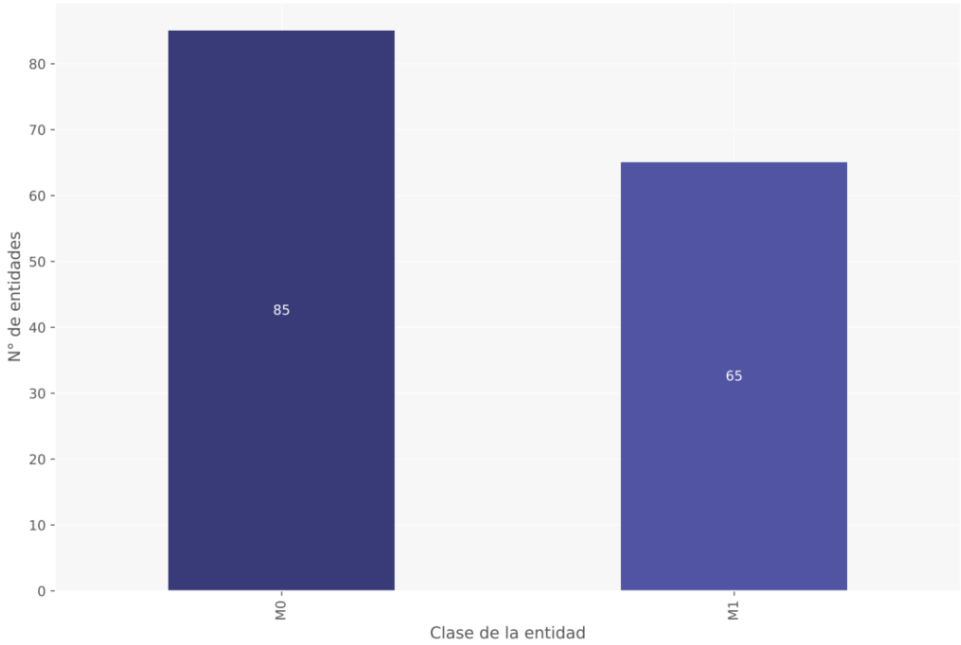

Figura 5.18. Cantidad de entidades según clase y conjunto de datos. Modelo de NER basado en Aprendizaje Profundo con dos clases.

Si bien la cantidad total es similar, casi el 80% de las entidades predichas corresponde a entidades identificadas correctamente. El 20% restante está dividido en 4 tipos de errores: tipos 1, 2 y 5 distribuidos de manera similar, y tipo 3 en menor proporción, con tan solo 1% de los errores (Figura 5.19). Los errores tipo 4 no ocurrieron en este modelo.

La cantidad de entidades erróneas por clase y su distribución por tipo de error se observan en las Figuras 5.20 y 5.21. La proporción de entidades identificadas incorrectamente para la clase M1, según el total de entidades del conjunto de testeo, fue 2,7 veces mayor que para la clase M0. Es decir, este modelo presenta un menor rendimiento para la clase M1.

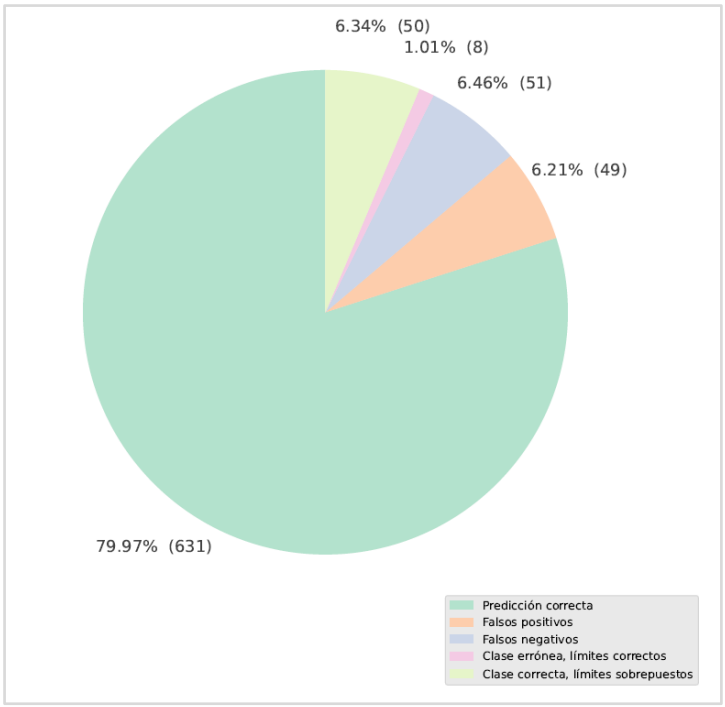

Figura 5.19. Cantidad de entidades correcta e incorrectamente según tipo de error. Modelo de NER basado en Aprendizaje Profundo con dos clases.

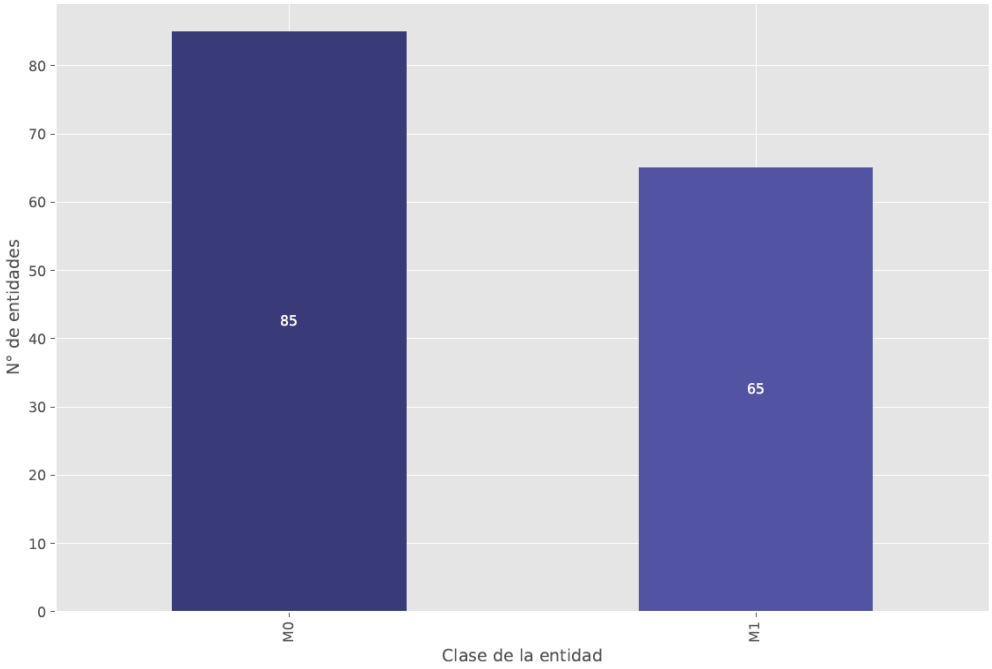

Figura 5.20. Cantidad de entidades identificadas incorrectamente según clase. Modelo de NER basado en Aprendizaje Profundo con dos clases.

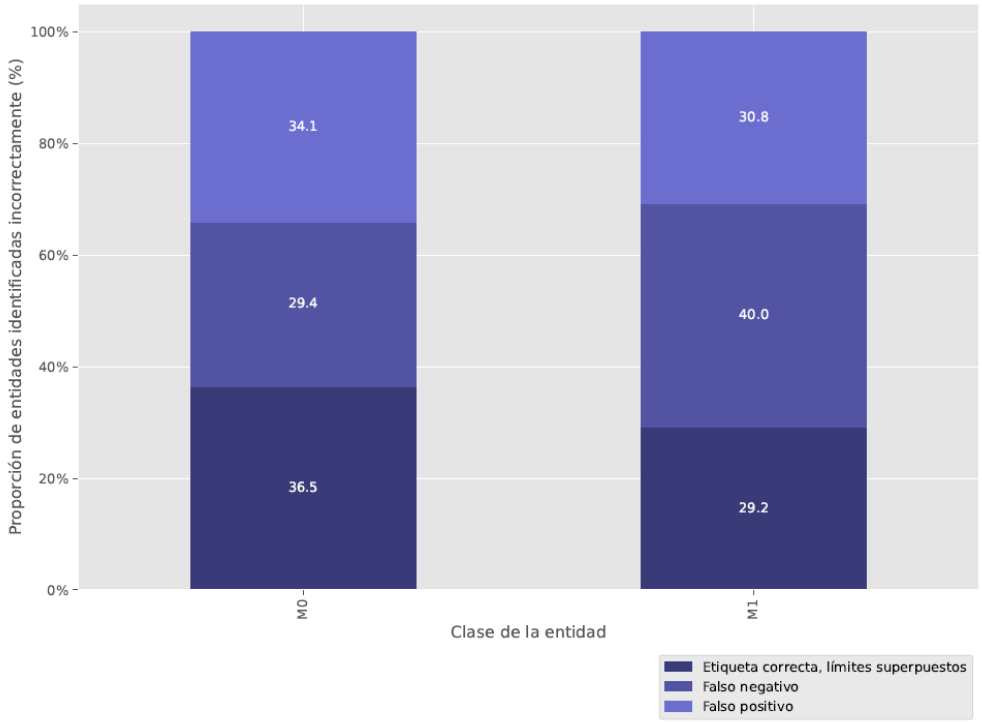

Figura 5.21. Proporción de entidades identificadas incorrectamente según clase y tipo de error. Modelo de NER basado en Aprendizaje Profundo con dos clases.

Aproximadamente un tercio de las entidades con error corresponden a errores tipo 5 (clase correcta, límites incorrectos), por lo que no es un error de relevancia, que vaya a cambiar la definición de la clase a nivel de documento que se realiza con el flujo de decisión a partir del NER. Los errores de relevancia en este caso, serían aquellos que cambian la clase de las menciones identificadas, es decir, que predicen clases erróneamente. El modelo de aprendizaje profundo no detectó errores tipo 4 y la cantidad de errores tipo 3 fue de un 1% de las entidades predichas, lo que corresponde a sólo 8 entidades.

El detalle de las menciones con clase incorrecta se puede observar en la Tabla 5.23, palabras similares y la cantidad de entidades con clase incorrecta según clase, en la matriz de confusión (Figura 5.22). Con esto, se releva la utilidad del NER de aprendizaje profundo para la definición de una clase a nivel de documento, un bajo nivel de error en la clasificación de las entidades produce un alto rendimiento en la definición de entidades a nivel de documento.

Tabla 5.23. Frecuencia de palabras identificadas con clase errónea. Algoritmo de NER basado en aprendizaje profundo con dos clases..

Mención	Frecuencia (n)
nódulo	2
Focos líticos	1
Linfonodos	1
Nódulo	1
Nódulos	1
lesiones blásticas	1
nódulos hipodensos	1

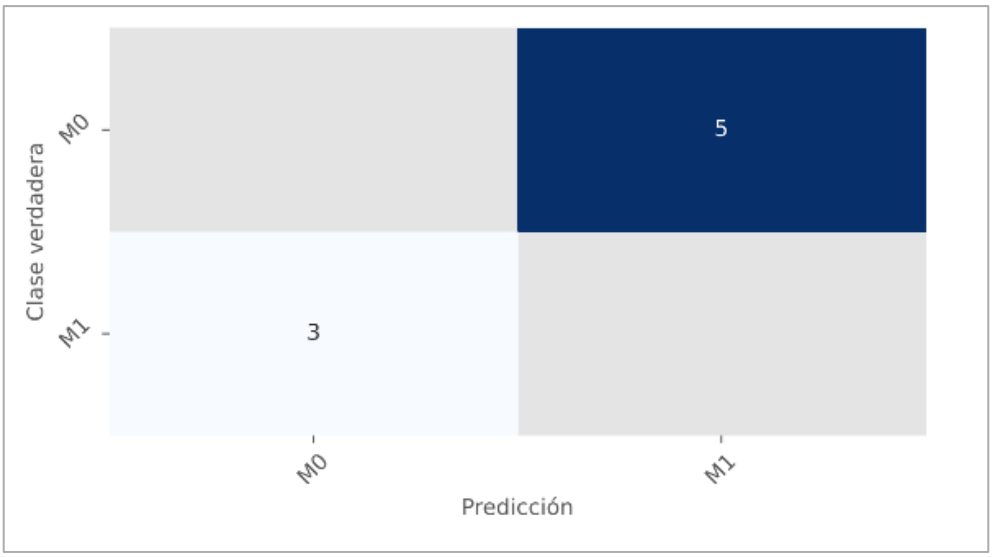

Figura 5.22. Matriz de confusión de entidades con clase errónea. Modelo de NER basado en aprendizaje profundo con dos clases.

NER de tres clases basado en aprendizaje profundo.

A pesar de que este modelo detectó una cantidad similar de menciones en comparación con el conjunto de testeo (Figura 5.23), la proporción de entidades predichas correctamente disminuyó, en cerca de un 3%, en comparación al modelo de NER de aprendizaje profundo con dos clases.

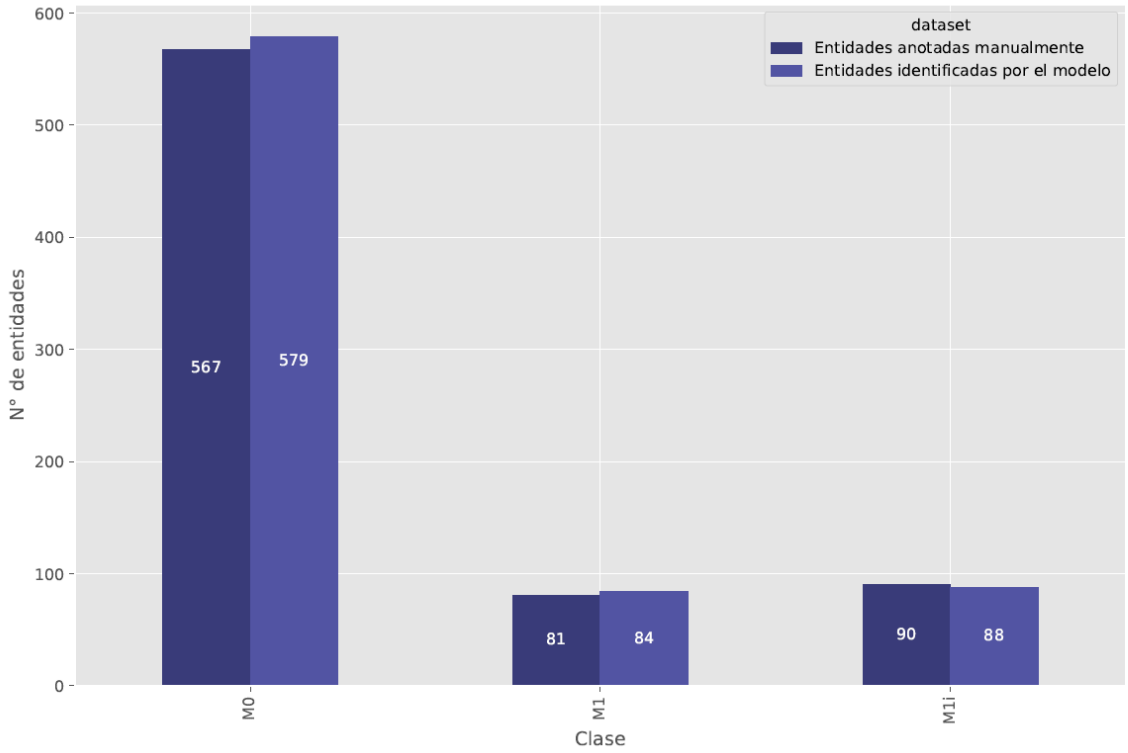

Figura 5.23. Cantidad de entidades según clase y conjunto de datos.
Algoritmo de NER basado en Aprendizaje Profundo con tres clases. **M1i**: M1 con incertidumbre.

Se observan dos diferencias importantes con el modelo de NER basado en aprendizaje profundo de de dos clases: primero, que los errores tipo 3 (clase errónea, límites correctos) aumentaron 4 veces, y segundo, que en este modelo se presentaron errores tipo 4 (Figura 5.24) que en el otro se produjeron.

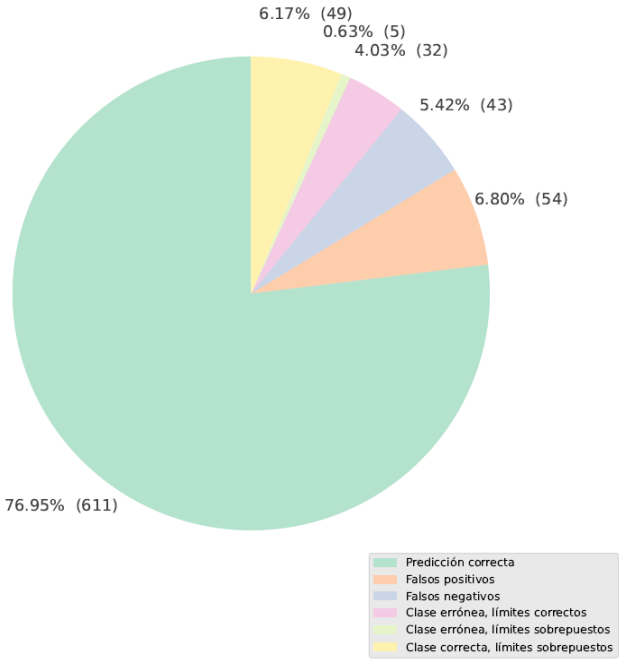

Figura 5.24. Cantidad de entidades correcta e incorrectamente según tipo de error. Modelo de NER basado en aprendizaje profundo con tresc lases.

Al analizar los resultados por clase, se observa la misma cantidad de entidades clasificadas erróneamente como M0, tanto en el modelo de NER de aprendizaje profundo de dos clases, como en el de tres clases. Las clases M1 y M1 incierto tuvieron un porcentaje de entidades identificadas erróneamente de 45,7% y 26,7%. En la Figura 5.26 se observa la proporción de errores por clase. Se destaca la proporción de falsos positivos en la clase M1, que asciende a un 43% aproximadamente. y de falsos negativos en la clase M1 incierto, con un 42% aprox. La clase M0 tuvo un mejor rendimiento contra el subset de testeo y eso se relaciona con la mayor cantidad de errores tipo 5, con identificación de la clase de manera correcta.

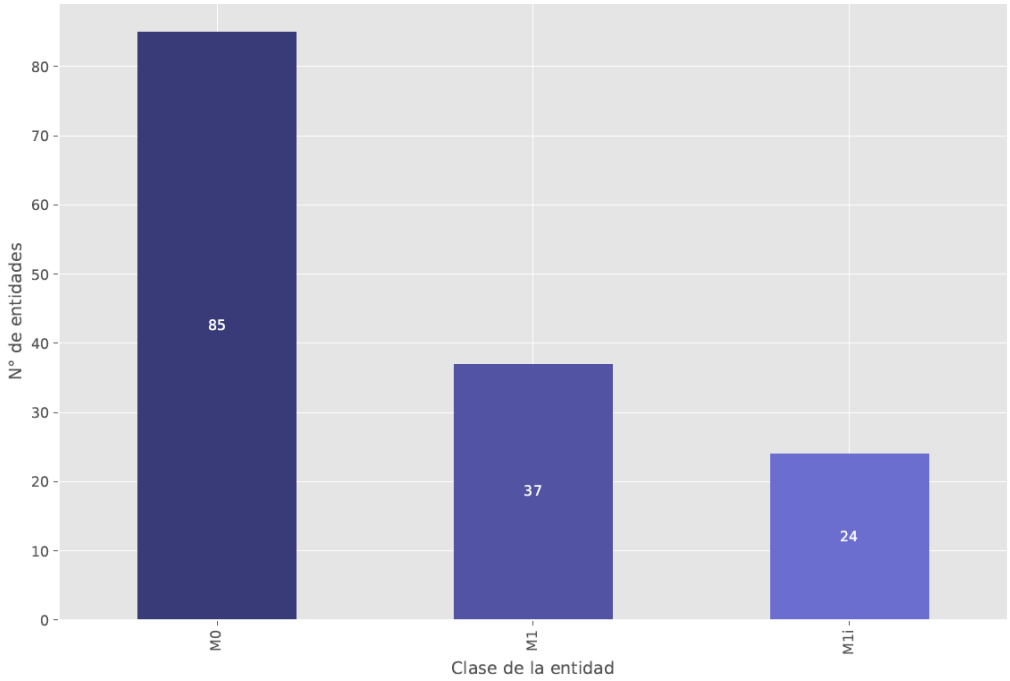

Figura 5.25. Cantidad de entidades identificadas incorrectamente según clase.
Modelo de NER basado en aprendizaje profundo con tres clases. **M1i**: M1 con incertidumbre.

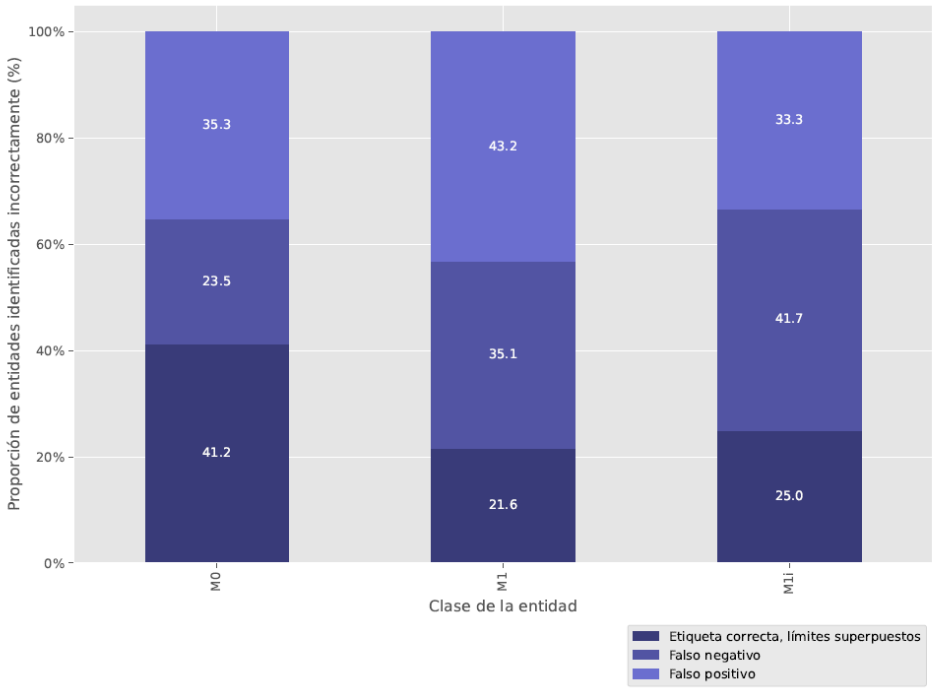

Figura 5.26. Proporción de entidades identificadas incorrectamente según clase y tipo de error.
Modelo de NER basado en aprendizaje profundo con tres clases. **M1i**: M1 con incertidumbre.

Al asignar una clase, la mayoría de los errores se produjeron cuando se involucra la clase M1 con incertidumbre (ver Figura 5.27). Es decir, se producen más entidades identificadas erróneamente entre M1 incierto y M1 o M0 que entre M1 y M0. Por lo que podemos suponer que el modelo tiende a confundir en mayor medida las menciones M1 con incertidumbre y en menor medida las M1 y M0.

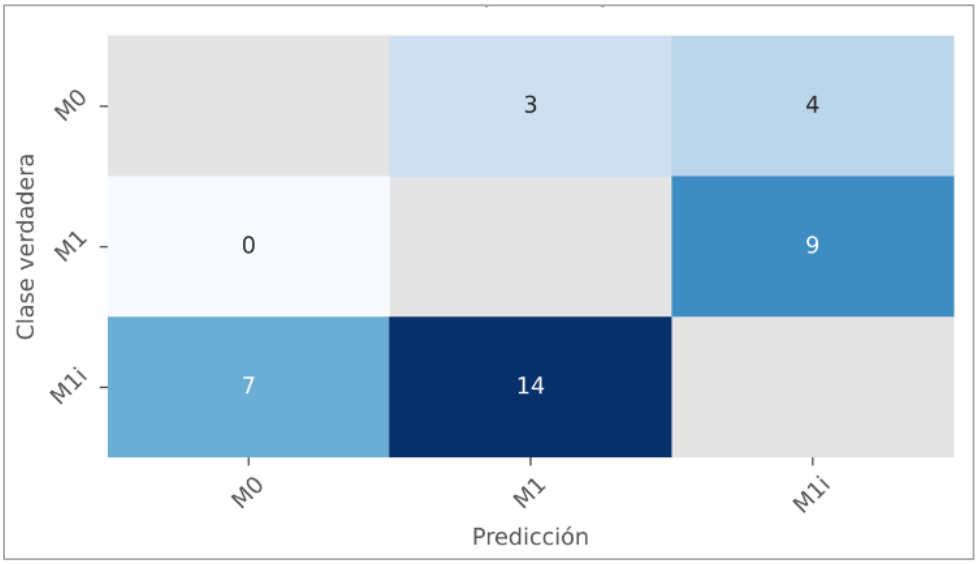

Figura 5.27. Matriz de confusión de entidades con clase errónea.
Modelo de NER basado en aprendizaje profundo con tres clases. **M1i**: M1 con incertidumbre.

5.5.- Evaluación del rendimiento de los modelos versus el conjunto de datos de prueba utilizando la métrica estándar *F1-score*.

El resumen de los resultados de rendimientos obtenidos para todos los modelos entrenados y evaluados en este trabajo se resumen en la Tabla 5.24 y 5.25. La hipótesis de este trabajo buscaba comprobar que el procesamiento del lenguaje natural permite la detección de hallazgos de metástasis a distancia con un *F1-score* de al menos 0,8.

Tabla 5.24. Resumen, evaluación de rendimiento para la tarea de NER de metástasis a distancia.

Clase M0				
Modelo	Precisión	Exhaustividad	*F1-score*	Entidades testeadas (n)
NER aprendizaje profundo, 2 clases.	0,89	0,89	0,89	567
NER basado en reglas, 2 clases.	0,54	0,47	0,50	567
NER aprendizaje profundo, 3 clases.	0,88	0,89	0,89	567
Clase M1				
NER aprendizaje profundo, 2 clases.	0,74	0,72	0,73	171
NER basado en reglas, 2 clases.	0,25	0,63	0,36	171
NER aprendizaje profundo, 3 clases.	0,56	0,62	0,59	81
Clase M1 con incertidumbre				
NER aprendizaje profundo, 3 clases.	0,66	0,59	0,62	90

Tabla 5.25. Resumen, evaluación de rendimiento para la tarea de clasificación a nivel de documento.

Clase M0				
Modelo	Precisión	Exhaustividad	*F1-score*	Documentos testeados (n)
NER aprendizaje profundo, 2 clases.	0,7941	0,9310	0,8571	29
NER basado en reglas, 2 clases, regresión logística.	0,95	0,62	0,75	29
NER basado en reglas, 2 clases, SVM.	0,80	0,69	0,74	29
NER aprendizaje profundo, 3 clases.	0,8709	0,9319	0,9000	29
Clase M1				
NER aprendizaje profundo, 2 clases.	0,9375	0,8108	0,8695	37
NER basado en reglas, 2 clases, regresión logística.	0,77	0,97	0,86	37
NER basado en reglas, 2 clases, SVM.	0,78	0,86	0,82	37
NER aprendizaje profundo, 3 clases.	0,7368	0,8750	0,8000	16
Clase M1 con incertidumbre				
NER aprendizaje profundo, 3 clases.	0,8750	0,6667	0,7567	21

6. DISCUSIÓN

El texto clínico es la principal forma de almacenamiento de texto en los sistemas de registro clínico electrónico debido a su flexibilidad y expresividad (71). La dificultad de su procesamiento, debido a la falta de estructura y estandarización, ha producido un aumento en el desarrollo de tecnologías que faciliten la extracción de información relevante del texto de reportes de imagenología (31). Este aumento se ha observado principalmente en los modelos de procesamiento de lenguaje natural basados en aprendizaje profundo, con una hegemonía de publicaciones científicas desarrolladas para textos escritos en inglés, que difícilmente se pueden aplicar en textos en español (29). La especificidad debido a las características de cada idioma y de cada dominio clínico da cuenta de la necesidad del análisis y procesamiento del lenguaje natural en nuestro idioma, así como también, la creación de corpus anotados en diferentes dominios clínicos para la aplicación del procesamiento del lenguaje natural en nuestros sistemas de información.

La utilidad de esta investigación radica en la priorización de los tratamientos curativos, con el fin de que, aquellos pacientes que tengan posibilidades de curación, sean intervenidos oportunamente. Esto es importante si se considera que, para el cáncer de colon, cuando un paciente llega a etapa metastásica, solo tiene un 10% de posibilidades de sobrevivir a los 5 años de diagnosticada la metástasis (72). En el caso del cáncer de mama, se calcula un 22% de tasa de sobrevida a 5 años (73). La tasa de sobrevida a 5 años de un cáncer de próstata localizado o regional es de casi un 100%, pero cuando éste evoluciona a metástasis, esa tasa disminuye a un 31% (74). Es en este último ejemplo donde se observa la relevancia de obtener la información de los reportes, estructurar el texto clínico y codificar la información en la ficha clínica y utilizarlos para la gestión de estos pacientes. Evitar que los pacientes lleguen a etapa de metástasis a distancia es fundamental y esta herramienta aporta a informar cuanto antes si un paciente presenta o no aquella condición.

Para el desarrollo de la tarea de NER y clasificación de metástasis a distancia se utilizó un conjunto de datos de 673 reportes, compuesto por TC y PETCT. Los reportes de RNM se descartaron, porque a pesar de que constituían el 13,4% del total de documentos, solo aportaron el 3,4% de las entidades anotadas, con un promedio

de 2,5 menciones por reporte. Esto se debe a que el tipo de RNM disponible para este trabajo fue, en su mayoría, focalizado en el órgano del tumor primario, por lo que en muchos de los casos, no fue posible evaluar metástasis a distancia. Además, cuando se incorporaron estos textos se presentó una disminución en el rendimiento de los modelos de aprendizaje profundo .

Según lo descrito por Casey et al. (31), nuestra cantidad de reportes estaba bajo la media de los reportes requeridos para técnicas como aprendizaje profundo, aprendizaje de máquinas o algoritmos basados en reglas: sobre 2.500 documentos. Si bien es un número más bajo, la cantidad de anotaciones obtenidas desde los documentos, relevante para la tarea del NER, es similar a otros trabajos con buenos resultados (48,75).

La anotación de los textos fue realizada por personal técnico de FALP. En un estudio de Xia y Yetisgen-Yikdiz (76) sobre la anotación de *corpus* clínicos, indican que los anotadores deben tener conocimiento en el dominio en el que se está trabajando y que en lo ideal, deben ser médicos expertos. En ese mismo estudio se menciona que el conocimiento médico no es suficiente, sino que debe ser acompañado del desarrollo de guías de anotación claras. La anotación por médicos expertos, especialistas radiólogos o en medicina nuclear en este caso, conlleva un presupuesto muy elevado, por lo que se optó por trabajar junto a técnicas en enfermería del registro hospitalario de tumores de FALP, es decir, con experiencia en la búsqueda de información clave en la ficha y reportes clínicos. Junto a esto, se desarrollaron sesiones de trabajo conjunto con médicos radiólogos para resolver dudas y disensos. El acuerdo entre-anotadoras fue satisfactorio, a medida que el proceso de anotación avanzaba, aumentó el nivel de acuerdo. Lo que da cuenta de la coordinación y coherencia entre las anotadoras, con base en las guías y las reuniones de anotación. Para trabajos futuros, se podría incorporar profesionales médicos en la confección de la guía de anotación y en el proceso de anotación, de esa manera se puede evaluar si es que aquello mejora el rendimiento del modelo de NER.

Según datos del registro hospitalario de tumores de FALP (77) , la proporción por año entre 2017 y 2019 de pacientes con diagnóstico inicial de cáncer en etapa IV o que están en etapa de metástasis a distancia, para el cáncer de mama, próstata y

colorrectal, fue en promedio de un 4,9%, 23,7% y 9,8% de pacientes, respectivamente. El año 2020, aunque a la fecha el registro está incompleto, la proporción aumenta a 5,6%, 27,0% y 19,6% para los mismos tipos de cáncer. Los documentos clasificados manualmente para este estudio tienen una mayor proporción de metástasis a distancia que lo captado por el registro hospitalario de tumores. La proporción de M1 para cáncer de mama fue de 25,1%, para cáncer próstata de 34,8% y para cáncer colorrectal de 42,4%. Esto se podría explicar porque los datos utilizados corresponden no solo a pacientes con diagnóstico inicial, sino que también de pacientes que ya contaban con diagnóstico de cáncer y se les está haciendo exámenes de seguimiento (a diferencia del registro hospitalario de tumores, que mide incidencia de cáncer).

A nivel de entidades anotadas, resultó un marcado desbalance de clases, con un 75% de entidades M0. Los reportes estudiados tienen la lógica del descarte de hallazgos, pues se describe en múltiples oportunidades, en un mismo reporte, cada uno de los hallazgos de metástasis a distancia que “no se observan”. Este desbalance podría explicar el rendimiento superior de la clase M0 en todos los modelos. Una de las estrategias que se proponen en este trabajo para mejorar el rendimiento de la clase M1 es aumentar la cantidad de entidades disponibles y por ende, de documentos, para llegar a números similares a los actuales para la clase M0.

En cuanto al algoritmo de NER de dos clases basado en reglas, que fue desarrollado como línea de base para la detección de las entidades de metástasis a distancia, éste no obtuvo resultados satisfactorios, con un *F1-score* de 0,36 para la clase M1, dado por una precisión de 0,25. Esto se debe a la gran cantidad de errores tipo 1 o falsos positivos. El criterio para definir una mención como positiva o negativa fue en base a la oración que la contenía, es decir: si el contexto es negado, la metástasis se descarta (entidad M0), pero si el contexto es afirmativo la entidad es M1. Como la clase de la mención depende del contexto (tipo de cáncer y sitio anatómico), el aumento de falsos positivos se debe a que se están identificando entidades erróneas, como lesiones asociadas a tumor primario o lesiones no-oncológicas o no-metastásicas. El modelo no es capaz de descartar este tipo de menciones, pues no asocia más información que la coincidencia de caracteres en el texto del reporte. Para un trabajo posterior, el algoritmo basado en reglas debe incluir información sobre la

parte del cuerpo y el tipo de cáncer. Debido al tiempo determinado para el desarrollo de este trabajo no fue posible incorporar esa información.

Otro aspecto a considerar y que no fue posible incluir en este trabajo, fue el procesamiento del lexicón para obtener las palabras lematizadas, con eso se podría disminuir la cantidad de falsos negativos que probablemente surgen debido a que no se produce un emparejamiento perfecto entre el texto y las palabras del lexicón. Esto sucede porque el texto clínico no es perfecto, existen faltas de ortografía, abreviaturas y omisión de caracteres en la escritura de los profesionales clínicos (29).

Los resultados de los experimentos realizados para definir los parámetros de entrenamiento de los modelos de NER basados en aprendizaje profundo, para la clase M0, muestran que el modelo con mejor rendimiento es aquel que se entrenó solo en TC y en todos los tipos de cáncer estudiados, obteniendo un *F1-score* de 0,92. Además, 2 de los 5 mejores resultados para esta clase se obtuvieron agrupando los conjuntos de datos por cáncer, para cáncer de próstata y mama. Para el cáncer colorrectal esto no sucedió. Una razón podría ser que este cáncer contaba con menos reportes disponibles. Por otro lado, 4 de los 5 mejores modelos fueron entrenados solo con reportes de PETCT y TAC, y el modelo que incorporó los reportes de RMN llegó solo a un *F1-score* de 0,86 (datos no reportados).

Para la detección de entidades M1 y M1 incierto también se observó que los modelos tendían a mejorar su rendimiento cuando se agruparon por tipo de reporte o cáncer. El problema de aquello es la baja cantidad de reportes disponibles para el entrenamiento. Como proyección, sería interesante evaluar con una mayor cantidad de reportes, cómo se comportan este tipo de modelos cuando se acota el objetivo del estudio, es decir, si se acota a un tipo de reporte o a un tipo de cáncer.

El modelo de NER basado en aprendizaje profundo de dos clases tuvo un mejor rendimiento que el algoritmo basado en reglas, la clase M0 obtuvo un rendimiento de 0,89 y la clase M1 un rendimiento de 0,73, ambos medidos en *F1-score*. Un tercio de las entidades con error fueron tipo 5, en ellas solo hubo diferencias en los límites de la mención, más no en la clase asignada. Es un error que no cambia el significado de la entidad, pero que da cuenta de ciertas falencias en la guía de anotación. Por

ejemplo, una instrucción más detallada respecto de los límites de las anotaciones, es decir, dónde empieza y dónde termina la entidad y qué elementos se debe o no incluir, un elemento a mejorar para una proyección de este trabajo.

Los otros dos tercios son falsos positivos y falsos negativos. Respecto de los falsos positivos, en ocasiones, los hallazgos en el órgano primario se describen de manera similar a los metastásicos, por lo que para evitar esta “confusión”, sería interesante incorporar la anotación de partes del cuerpo y de la enfermedad o tipo de cáncer. Aquello permitiría vincular las menciones en una misma oración para poder otorgar una clase, de manera similar a cómo se haría siguiendo el estándar de la clasificación TNM de tumores.

Los falsos negativos probablemente se produjeron debido a que se anotaron palabras que no tenían significancia por sí solas para metástasis a distancia, por ejemplo, “nódulos”, “compromiso”, “lesión”, “focos”, y que el modelo no reconoció en el contexto de la oración. Para trabajos posteriores, esto debería solucionarse desarrollando guías de anotación que induzcan a los anotadores a etiquetar palabras más específicas.

Al evaluar el modelo de NER basado en aprendizaje profundo, en tres clases, se obtuvo una media balanceada de 0,8093. La clase M0 tuvo un rendimiento similar al modelo de dos clases, basado en aprendizaje profundo, pero se observa una disminución en el rendimiento de la clase M1, bajando de 0,7300 a 0,5930 dado en mayor medida por una baja abrupta en la precisión. La clase M1 con incertidumbre tuvo un mejor rendimiento, pero con un *F1-score* que no supera el 0,63.

Para obtener las entidades M1 y M1 con incertidumbre, se tuvo que dividir el conjunto de entidades M1 del modelo de 2 clases, por lo que se contó con menos entidades de ejemplo para cada clase. Esto probablemente disminuyó el rendimiento, lo que puede ser corregido balanceando las clases, es decir, incorporando más ejemplos de entidades M1 y M1 con incertidumbre.

Cuando el modelo asignó incorrectamente la clase, la mayoría de los errores se produjeron al involucrar la clase M1 con incertidumbre. Por lo que, para apoyar esa decisión, se podría desarrollar un corpus anotado o utilizar un corpus ya disponible,

que permita etiquetar las palabras o segmentos que modifican el sentido de las menciones de metástasis a distancia, como por ejemplo: “indeterminado” o “incierto” para las oraciones inciertas y “no” o “sin” para las oraciones negadas. Con eso se podría filtrar las entidades para evitar la asignación errónea de clase por parte del modelo, tal como se describe en el estudio de Solarte-Pabón (48). Debido al tiempo y recurso humano que se requiere para un proceso de anotación, esto no se pudo llevar a cabo durante este trabajo.

La manera de describir los hallazgos de metástasis a distancia es similar para todas las clases. La diferencia entre una entidad M1 y M0 no es la entidad en sí misma, sino que son los elementos contextuales de la oración. Es decir, lo que cambia son las palabras o segmentos modificadores que acompañan la mención. Estas palabras no fueron anotadas, pero el modelo de Bi-LSTM con CRF permitió captarlas (con rendimientos de *F1-score* superiores a 0,7 para M0 y 0,8 para M1) de igual manera para poder asignar una clase a cada entidad.

La idea de haber desarrollado una tarea de NER, que posteriormente otorgó una clase a nivel de documento, tiene como objetivo poder dar cuenta del “razonamiento” del modelo para la clasificación. Los modelos de aprendizaje profundo funcionan como cajas negras y es muy difícil saber cómo el modelo está tomando la decisión por tal o cual clase para un documento. Por lo que la inclusión del NER da luces a los profesionales que revisan el registro clínico de por qué un documento fue clasificado con una determinada clase, y analizar por qué podría estar bien clasificado o si es necesario corregir para mejorar el rendimiento posterior del modelo, utilizando técnicas de aprendizaje activo (78).

Este modelo se utilizará para una gestión y priorización de pacientes para inicio de tratamiento, con el objetivo de otorgar atenciones oportunas y evitar que lleguen a etapas de metástasis muy avanzadas. Por lo que las métricas más relevantes para la tarea de NER, son las que permiten saber lo más exacta y precisamente qué pacientes son M0. En ese sentido, se deben disminuir las entidades con error tipo 1, de falso positivo, para la clase M1. Es decir, se debe aumentar la precisión de M1 en el modelo de NER. Esto es importante, porque según la metodología de clasificación

a nivel de documento de este trabajo, con solo una entidad M1 presente en el texto, cambia completamente la clase del reporte completo.

Respecto de la clasificación a partir del NER de aprendizaje profundo, se observa que aunque la tarea de detección de entidades al interior del texto clínico es mucho más difícil, especialmente para las entidades M1 y M1 incierto (*F1-score* < 0,8), fue posible utilizar los resultados del NER para clasificar el documento completo con rendimientos significativos, incluso con *F1-score* > 0,9. Esto se debe a que el NER es una tarea muy granular y requiere que cada mención sea detectada y se le asigne correctamente su clase. Para la clasificación de un documento, que puede tener muchas menciones (13 en promedio para los PETCT), basta con que alguna de esas menciones sea M1 y sea detectada correctamente para que el documento sea clasificado como M1. Esta condición es lo que mejora el rendimiento en la pesquisa de la clase global del reporte.

Para gestionar el error que se produce al clasificar reportes, se estableció que esta clasificación debe ir acompañada de la proporción de entidades de metástasis a distancia que corresponden para cada clase. Esto permitiría evaluar que, si un reporte tiene un porcentaje bajo de proporción de metástasis a distancia, se debe tener en cuenta de que este modelo podría no haber asignado una clase correcta a la entidad y por lo tanto, se está en presencia de un falso positivo, ya sea para M1 o para M1 con incertidumbre. Por otro lado, permite tener una noción a los equipos respecto del nivel de compromiso de la metástasis a distancia, pues no es lo mismo un reporte de un paciente con un 15% de menciones M1, versus un reporte de otro paciente con 60%. En este último caso, además de tener mayor certeza de la clasificación M1 a nivel de documento, se podría inferir que tiene un mayor compromiso metastásico, lo cual puede evaluarse en futuros trabajos.

La clasificación a partir de la metodología NER basada en reglas, se comporta de manera similar al utilizar regresión logística y SVM, pues la clase asignada a cada reporte va a depender de la cantidad de oraciones afirmativas y negativas, que se convierten finalmente en probabilidades. Eso explica que con un algoritmo para NER con rendimientos de *F1-score* de 0,5 y 0,36, se pueden obtener rendimientos de clasificación de un reporte superiores a 0,8 de *F1-score* en ambas clases.

Según la revisión de Casey et al. (31), los estudios realizados en reportes de imagenología para la aplicación clínica de información de enfermedad/clasificación, es decir, aquellos que buscaban clasificar la ocurrencia de una enfermedad o extraer información respecto de una enfermedad, tienen una media de rendimiento de 0,91 aproximadamente, con una mediana de alrededor de 0,86, lo que coincide con el rendimiento obtenido por los modelos de NER de aprendizaje profundo. Miao et al. (43), que buscaba extraer los hallazgos de BI-RADS de reportes de mamografía logró con una red neuronal recurrente un *F1-score* de 0,90, mientras que Yuan et al. (79), utilizando una red neuronal del tipo LSTM logró detectar y clasificar la progresión o regresión de nódulos pulmonares en texto radiológico. Por lo que el rendimiento de los modelos entrenados, principalmente, basados en aprendizaje profundo, se encuentran a la par de artículos de similares características en idioma inglés.

7. CONCLUSIÓN

La creación de sistemas de estandarización de información clínica son relevantes para la gestión asistencial, pues permiten extraer información contenida en texto libre de los sistemas de información y ponerla a disposición de los equipos clínicos. En ese sentido, el mantener un registro estandarizado del estado de un paciente en relación a metástasis a distancia, de la clasificación TNM, supone un apoyo para la priorización y gestión de tratamientos oportunos cuando existe la posibilidad y para la aplicación de terapias de cuidados paliativos y de fin de vida, cuando ya no es posible otorgar una cura.

Este trabajo permitió desarrollar un corpus anotado de metástasis a distancia, que contiene las entidades al interior del documento, como también, la clasificación a nivel de documento. Se incluyó las entidades M1 y M0, como entidades asociadas a la clasificación TNM, pero, para la gestión de la incertidumbre en los textos de imagenología, se incorporó un atributo que marcaba las entidades indeterminadas al interior de cada reporte. El proceso de anotación se desarrolló como se proyectaba, con un acuerdo entre anotadoras que fue mejorando con el avance de las fases de anotación. Es necesario volver a revisar las guías de anotación según los errores que comete el modelo al identificar automáticamente las menciones.

Si bien, el corpus anotado de metástasis permitió el desarrollo de este trabajo, es necesaria la expansión del mismo, en número y tipos de entidades anotadas. Se debe amplificar la cantidad de entidades M1 y M1 con incertidumbre, para balancear el corpus y volver a evaluar los rendimientos con una mayor cantidad de ejemplos. Además de continuar avanzando hacia las clasificaciones T, del tumor primario y N de metástasis a linfonodos regionales, de la clasificación TNM.

La tarea de NER supuso un desafío, pues el modelo de Bi-LSTM con CRF debía poder incorporar al análisis los elementos contextuales de las entidades de metástasis a distancia. Si bien, los rendimientos pueden mejorar, si queda claro que el aprendizaje profundo funciona para tareas de NER como esta. Además, su uso destaca por sobre los algoritmos basados en reglas, debido a que éstos carecen de escalabilidad y tienen un difícil manejo de lexicones cuando es necesario tener en consideración el contexto. Lo anterior se ve sostenido por los rendimientos comparados entre las tareas de NER de los modelos basados en aprendizaje profundo versus aquel basado en reglas. Buscar estrategias para maximizar la precisión de la clase M1 supone un desafío, pues la existencia de solo una entidad M1 etiquetada con error tipo 1, de falsos positivos, cambia la clase del documento completo.

La clasificación a nivel de documento fue factible y obtuvo rendimientos superiores a los esperados para todas las aproximaciones. Fue posible llegar a 0,90 de *F1-score* para la clase M0 y 0,87 para la clase M1.

Como conclusión podemos afirmar que la detección de hallazgos de metástasis a distancia en textos de TC y PETCT, de cáncer de próstata, colorrectal y mama, mediante el procesamiento del lenguaje es posible y factible de realizar para textos en español, con rendimientos comparables a la literatura y superiores a 0,8, comprobando la hipótesis planteada en este trabajo.

Referencias

1. Martínez-Sanguinetti MA, Leiva-Ordeñez AMaría, Petermann-Rocha Fanny, Celis-Morales Carlos. ¿Cómo ha cambiado el perfil epidemiológico en Chile en los últimos 10 años? Rev Médica Chile. 2021;149(1):147-58.
2. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. Global Cancer Observatory: Cancer Today [Internet]. Lyon, France: International Agency for Research on Cancer. 2021. Disponible en: https://gco.iarc.fr/today
3. Suhail Y, Cain MP, Vanaja K, Kurywchak PA, Levchenko A, Kalluri R, et al. Systems Biology of Cancer Metastasis. Cell Syst. 2019;9(2):109-27.
4. Seyfried TN, Huysentruyt LC. On the Origin of Cancer Metastasis. Cancer. 2013;18(1-2):43-73.
5. Dillekås H, Rogers MS, Straume O. Are 90% of deaths from cancer caused by metastases? Cancer Med. 2019;8(12):5574-6.
6. Pulido C, Vendrell I, Ferreira AR, Casimiro S, Mansinho A, Alho I, et al. Bone metastasis risk factors in breast cancer. Ecancermedicalscience. 2017;11:1-17.
7. Nafissi NN, Kosiorek HE, Butterfield RJ, Moore C, Ho T, Singh P, et al. Evolving Natural History of Metastatic Prostate Cancer. Cureus. 2020;12(11):10-5.
8. Riihimaki M, Hemminki A, Sundquist J, Hemminki K. Patterns of metastasis in colon and rectal cancer. Sci Rep. 2016;6:1-9.
9. Wang J, Li S, Liu Y, Zhang C, Li H, Lai B. Metastatic patterns and survival outcomes in patients with stage IV colon cancer: A population-based analysis. Cancer Med. 2020;9(1):361-73.
10. Zhang W, Wang F, Hu C, Zhou Y, Gao H, Hu J. The progress and perspective of nanoparticle-enabled tumor metastasis treatment. Acta Pharm Sin B. 2020;10(11):2037-53.
11. Zhou H, Dong D, Chen B, Fang M, Cheng Y, Gan Y, et al. Diagnosis of Distant Metastasis of Lung Cancer: Based on Clinical and Radiomic Features. Transl Oncol. 2018;11(1):31-6.
12. Griffith SD, Tucker M, Bowser B, Calkins G, Chang C hsu (Joe), Guardino E, et al. Generating Real-World Tumor Burden Endpoints from Electronic Health Record Data: Comparison of RECIST, Radiology-Anchored, and Clinician-Anchored Approaches for Abstracting Real-World Progression in Non-Small Cell Lung Cancer. Adv Ther. 2019;36(8):2122-36.
13. Asan O, Nattinger AB, Gurses AP, Tyszka JT, Yen TWF. Oncologists' Views Regarding the Role of Electronic Health Records in Care Coordination. JCO Clin Cancer Inform. 2018;(2):1-12.
14. UICC. What is the TNM cancer staging system? [Internet]. 2021. Disponible en: https://www.uicc.org/resources/tnm
15. Panteli D, Legido-Quigley H, Reichebner C, Ollenschläger G, Schäfer C, Busse R. Clinical Practice Guidelines as a quality strategy. En: Busse R, Klazinga N, Panteli D, Quentin W, editores. Improving healthcare quality in Europe: Characteristics, effectiveness and implementation of different strategies [Internet] [Internet]. Copenhagen: European Observatory on Health Systems and Policies; 2019. Disponible en: https://www.ncbi.nlm.nih.gov/books/NBK549283/
16. Ministerio de Salud de Chile. Guía clínica AUGE: Cáncer de próstata en personas de 15 años y más. Bibl Minist Salud. 2015;103-103.
17. Ministerio de Salud de Chile. Guías Clínicas AUGE Cáncer colorectal en personas de 15 años y más. Bibl Minist Salud. 2013;96-96.
18. Ministerio de Salud de Chile. Guías Clínicas AUGE Cáncer de Mama en personas de 15 años y más. Bibl Minist Salud. 2015;164-164.
19. Rotter T, Baatenburg de Jong R, Lacko Evans S, Ronellenfitsch U, Kinsman L. Clinical pathways as a quality strategy. En: Busse R, Klazinga N, Panteli D, Quentin W, editores. Improving healthcare quality in Europe: Characteristics, effectiveness and implementation of different strategies [Internet]. Copenhagen: European Observatory on Health Systems and Policies; 2019.
20. Chiang AC, Ellis P, Zon R. Perspectives on the Use of Clinical Pathways in Oncology Care. Am Soc Clin Oncol Educ Book. 2017;37:155-9.
21. Instituto Oncológico Fundación Arturo López Pérez. Modelo de Salud FALP [Internet]. Disponible en: https://www.institutoncologicofalp.cl/fundacion/modelo-de-salud-falp/
22. Jafari SH, Saadatpour Z, Salmaninejad A, Momeni F, Mokhtari M, Nahand JS, et al. Breast cancer diagnosis: Imaging techniques and biochemical markers. J Cell Physiol. 2018;233(7):5200-13.
23. Łukaszewski B, Nazar J, Goch M, Łukaszewska M, Stępiński A, Jurczyk MU. Diagnostic methods for detection of bone metastases. Wspolczesna Onkol. 2017;21(2):98-103.

24. Choi SH, Kim SY, Park SH, Kim KW, Lee JY, Lee SS, et al. Diagnostic performance of CT, gadoxetate disodium-enhanced MRI, and PET/CT for the diagnosis of colorectal liver metastasis: Systematic review and meta-analysis. J Magn Reson Imaging. 2018;47(5):1237-50.
25. Abrams-Pompe RS, Fanti S, Schoots IG, Moore CM, Turkbey B, Vickers AJ, et al. The Role of Magnetic Resonance Imaging and Positron Emission Tomography/Computed Tomography in the Primary Staging of Newly Diagnosed Prostate Cancer: A Systematic Review of the Literature. Eur Urol Oncol. 2021;4(3):370-95.
26. Pesapane F, Downey K, Rotili A, Cassano E, Koh DM. Imaging diagnosis of metastatic breast cancer. Insights Imaging. 2020;11(1).
27. Spandorfer A, Branch C, Sharma P, Sahbaee P, Schoepf UJ, Ravenel JG, et al. Deep learning to convert unstructured CT pulmonary angiography reports into structured reports. Eur Radiol Exp. 2019;3(1):1-8.
28. Roberts A. Language, Structure, and Reuse in the Electronic Health Record. AMA J Ethics. marzo de 2017;19(3):281-8.
29. Dalianis H. Clinical text mining: Secondary use of electronic patient records. Clinical Text Mining: Secondary Use of Electronic Patient Records. 2018. 1 p.
30. Buckley BW, Daly L, Allen GN, Ridge CA. Recall of structured radiology reports is significantly superior to that of unstructured reports. Br J Radiol. 2018;91(1083).
31. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak. 2021;21(1):1-18.
32. Hirschberg J, Manning CD. Advances in natural language processing. Science. 2015;349(6245):261-6.
33. Sorin V, Barash Y, Konen E, Klang E. Deep Learning for Natural Language Processing in Radiology—Fundamentals and a Systematic Review. J Am Coll Radiol. 2020;17(5):639-48.
34. Chen PH. Essential Elements of Natural Language Processing: What the Radiologist Should Know. Acad Radiol. 2020;27(1):6-12.
35. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: An introduction. J Am Med Inform Assoc. 2011;18(5):544-51.
36. Spasic I, Nenadic G. Clinical text data in machine learning: Systematic review. JMIR Med Inform. 2020;8(3).
37. Wu Y, Jiang M, Xu J, Zhi D, Xu H. Clinical Named Entity Recognition Using Deep Learning Models. AMIA Annu Symp Proc AMIA Symp. 2017;2017:1812-9.
38. Yim WW, Yetisgen M, Harris WP, Sharon WK. Natural Language Processing in Oncology Review. JAMA Oncol. 2016;2(6):797-804.
39. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: A methodical review. J Am Med Inform Assoc. 2020;27(3):457-70.
40. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. 30th Int Conf Mach Learn ICML 2013. 2013;(PART 3):2347-55.
41. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9(8):1735-80.
42. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: Opportunities and challenges. J Biomed Semant. 2018;9(1):1-13.
43. Miao S, Xu T, Wu Y, Xie H, Wang J, Jing S, et al. Extraction of BI-RADS findings from breast ultrasound reports in Chinese using deep learning approaches. Int J Med Inf. 2018;119(November 2017):17-21.
44. Lee C, Kim Y, Kim YS, Jang J. Automatic disease annotation from radiology reports using artificial intelligence implemented by a recurrent neural network. Am J Roentgenol. 2019;212(4):734-40.
45. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, et al. Deep learning to classify radiology Free-Text reports. Orig Res N Comput Appl Radiol. 2018;286(3):845-52.
46. Lenain R, Seneviratne MG, Bozkurt S, Blayney DW, Brooks JD, Hernandez-Boussard T. Machine learning approaches for extracting stage from pathology reports in prostate cancer. Stud Health Technol Inform. 2019;264:1522-3.
47. Banerjee I, Ling Y, Chen MC, Hasan SA, Langlotz CP, Moradzadeh N, et al. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artif Intell Med. 2019;97(November 2018):79-88.
48. Pabón OS, Torrente M, Provencio M, Rodríguez-Gonzalez A, Menasalvas E. Integrating speculation detection and deep learning to extract lung cancer diagnosis from clinical notes. Appl Sci Switz. 2021;11(2):1-21.

49. Groot OQ, Bongers MER, Karhade AV, Kapoor ND, Fenn BP, Kim J, et al. Natural language processing for automated quantification of bone metastases reported in free-text bone scintigraphy reports. Acta Oncol. 2020;59(12):1455-60.
50. Senders JT, Karhade AV, Cote DJ, Mehrtash A, Lamba N, DiRisio A, et al. Natural Language Processing for Automated Quantification of Brain Metastases Reported in Free-Text Radiology Reports. JCO Clin Cancer Inform. 2019;(3):1-9.
51. Shickel B, Tighe P, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques. Biomed Health Inform. 2018;
52. Klie JC, Bugert M, Boullosa B, de Castilho RE, Gurevych I. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. Proc Int Conf Comput Linguist. 2018;5-9.
53. Pyysalo S, Ginter F. Collaborative development of annotation guidelines with application to Universal Dependencies. Fifth Swed Lang Technol Conf SLTC 2014 Novemb 13-14 2014 Upps Swed. 2014;
54. Honnibal M, Montani I, Van Lendeghem S, Boyd A. spaCy: Industrial-strength Natural Language Processing in Python. 2020;
55. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34(5):301-10.
56. Sanamaría J. NegEx-MES. enero de 2019; Disponible en: https://doi.org/10.5281/zenodo.2542567#.YYCd46jg-oY.mendeley
57. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging. 2015; Disponible en: http://arxiv.org/abs/1508.01991
58. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. 2016 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol NAACL HLT 2016 - Proc Conf. 2016;260-70.
59. Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling. Proc 27th Int Conf Comput Linguist. 2018;1638-49.
60. Báez P, Villena F, Rojas M, Durán M, Dunstan J. The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish. 2020;(January):291-300.
61. Villena F, Báez P, Peñafiel S, Rojas M, Paredes I, Dunstan J. Automatic Support System for Tumor Coding in Pathology Reports in Spanish. SSRN Electron J [Internet]. 2021 [citado 3 de octubre de 2022]; Disponible en: https://www.ssrn.com/abstract=3982259
62. Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R. FLAIR: An easy-to-use framework for state-of-the-art NLP. NAACL HLT 2019 - 2019 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol - Proc Demonstr Sess. 2019;54-9.
63. Shen Y, Huang J, Zhang J, Yang M, Lei K. Discovering Medical Entity Relations from Texts using Dependency Information. IJCAI 2019 Workshop W21.
64. Báez P, Bravo-Marquez F, Dunstan J, Rojas M, Villena F. Automatic Extraction of Nested Entities in Clinical Referrals in Spanish. ACM Trans Comput Healthc. 31 de julio de 2022;3(3):1-22.
65. Stenetorp P, Pyysalo S, Topic G. Brat standoff format [Internet]. Disponible en: https://brat.nlplab.org/standoff.html
66. Chiarcos C, Ionov M, Glaser L, Christian F. Formal Data Structures for Tabular Formats in Language Technology. 2000;1-4.
67. SIGNLL. CoNLL: The SIGNLL Conference on Computational Natural Language Learning [Internet]. 2021. Disponible en: https://www.conll.org/
68. Pyysalo S. standoff2conll [Internet]. 2016. Disponible en: https://github.com/spyysalo/standoff2conll
69. Nejadgholi I, Fraser KC, de Bruijn B. Extensive Error Analysis and a Learning-Based Evaluation of Medical Entity Recognition Systems to Approximate User Experience. En: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing [Internet]. Online: Association for Computational Linguistics; 2020. p. 177-86. Disponible en: https://aclanthology.org/2020.bionlp-1.19
70. Hastie T, Tibshirani R, James G, Witten D. An introduction to statistical learning (2nd ed.). Springer Texts. 2021;102:618-618.
71. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc JAMIA. 2011;18(2):181—186.
72. Mody K, Baldeo C, Bekaii-Saab T. Antiangiogenic Therapy in Colorectal Cancer: Cancer J.

2018;24(4):165-70.
73. Sledge GW. Curing Metastatic Breast Cancer. J Oncol Pract. enero de 2016;12(1):6-10.
74. American Cancer Society. Cancer Facts & Figures 2022. 2022;80.
75. Gorinski PJ, Wu H, Grover C, Tobin R, Talbot C, Whalley H, et al. Named Entity Recognition for Electronic Health Records: A Comparison of Rule-based and Machine Learning Approaches. 2019;8.
76. Xia F, Yetisgen-Yildiz M. Clinical Corpus Annotation: Challenges and Strategies. :8.
77. Instituto Oncológico Fundación Arturo López Pérez. Registro Hospitalario de Tumores 2017-2020. 2022.
78. Ren P, Xiao Y, Chang X, Huang PY, Li Z, Gupta BB, et al. A Survey of Deep Active Learning. ACM Comput Surv [Internet]. octubre de 2021;54(9). Disponible en: https://doi.org/10.1145/3472291
79. Yuan J, Zhu H, Tahmasebi A. Classification of Pulmonary Nodular Findings based on Characterization of Change using Radiology Reports. 2019;10.

Anexo 1

Carta de Aprobación de Proyecto de Investigación Clínica FALP-Universidad de Chile.

Comité Ético Científico FALP
Manuel Montt 427 piso 4, Providencia RM
Tel. (56.2) 24457266 – (56.2) 24457349

ACTA DE APROBACION DE PROYECTOS CLÍNICOS

Con fecha 23 de Junio de 2021, el Comité Ético Científico (C.E.C.) de la Fundación Arturo López Pérez evaluó y aprobó el Protocolo de Investigación 2021-017-ITC-SIN-OTH, patrocinado por la Universidad de Chile en el ámbito de un estudio FONDECYT, titulado:

"Caracterización lingüística del texto clínico chileno: hacia una extracción automática de información".

y que conducirá como Co-Investigador el Dr. Inti Paredes en la Fundación Arturo López Pérez, ubicada en Rancagua 878, Providencia, RM, Chile.

Se evaluó y aprobó los siguientes documentos del Protocolo:

- Protocolo 2020_09_01_FormulacionPostEsp_2021_PBB
- Resumen ResumenPostEsp_PB

También se tomó conocimiento de los siguientes documentos:

- Carta de sometimiento
- Acta de aprobación del CEC Universidad de Chile 007 ACTA APROB. Proy. N 018-202 1 Sr. Pablo Baez 25-05-2021

Envío a usted la nómina de los miembros permanentes del Comité Ético Científico de la Fundación Arturo López Pérez que asistieron a la reunión de análisis del protocolo:

NOMBRES	CARGO	PROFESIÓN
María Verónica Anguita Mackay	Presidente	Teóloga
Gonzalo López Gaete	Vicepresidente	Abogado
Marina Nordiana Baruzzi	Secretaria Ejecutiva	Administrativa
Romina Vargas Aránguiz	Integrante	Enfermera
Marcela Penjean Rivera	Integrante	Kinesióloga
Francisco León Correa	Representante de la Comunidad	Académico

Andrea Vargas Godoy	Integrante	Médico

Veronica Anguita Mackay
Presidenta CEC FALP

COMITÉ ÉTICO CIENTÍFICO
23 JUN 2021
FUNDACION ARTURO LOPEZ PEREZ