Clasificación de Historias Clínicas Reales según CIE-10-ES para Localización de Neoplasias mediante Modelos Transformers

Alejandro Pascual-Mellado[1*], Fernando Gallego[1], Nuria Ribelles[2], José M. Jerez[1], y Francisco J. Moreno-Barea[1†]

[1]Departamento de Lenguajes y Ciencias de la Computación
Escuela Técnica Superior de Ingeniería Informática, Universidad de Málaga, Málaga, España

[2]Unidad de Gestión Clínica Intercentros de Oncología
Hospitales Universitarios Regional y Virgen de la Victoria, Málaga, España

{*ale.pas.mel@uma.es †fjmoreno@lcc.uma.es}

***Resumen*—La mayoría de la información clínica almacenada en los sistemas sanitarios españoles se encuentra como texto no estructurado en las historias clínicas electrónicas. La extracción automática de información valiosa contenida en estos documentos es una tarea crítica. Como información valiosa para las unidades de análisis clínicos de oncología, se encuentra la localización de la neoplasia que presenta un paciente. Esta localización, incluida en la categoría de la codificación CIE-10-ES, puede ser extraída de los textos mediante el procesamiento del lenguaje natural. Para ello, en este estudio hemos desarrollado metodologías basadas en el estado del arte del procesamiento del lenguaje natural, los modelos Transformers. Los resultados obtenidos muestran que la aplicación de estos modelos es de gran ayuda en esta tarea. En particular, el modelo RoBERTa-Base-Biomed obtuvo el mejor rendimiento, con un valor de 0.946 en porcentaje de aciertos, 0.920 en precisión, 0.898 en sensibilidad y 0.908 en F1-score, mostrando un gran rendimiento para la mayoría de las clases.**

***Palabras Clave*—Procesamiento de lenguaje natural, Transformers, Historia clínica electrónica, CIE-10-ES, Español**

I. Introducción

Los sistemas sanitarios públicos en España no sólo se siguen enfrentando a retos clásicos de sostenibilidad y mejora de la experiencia del paciente. Con el auge de la medicina personalizada y de precisión, la globalización y el uso de la inteligencia artificial (IA), y con una sociedad concienciada con la recogida generalizada y transparente de sus datos, los sistemas sanitarios se enfrentan además a la extracción de información para el análisis de datos del mundo real (*Real-World Data*, RWD). El RWD ha adquirido en los últimos años un creciente interés, siendo un enfoque complementario a los ensayos clínicos aleatorizados para obtener una mayor seguridad y eficacia en los estudios [1]. Por ello, es necesario que los RWD recogidos en las historias clínicas electrónicas (HCE), sean transformados en información útil que ayude a los clínicos en la toma de decisiones [2].

Sin embargo, el desarrollo e implantación gradual de sistemas sanitarios cuyo componente clave es la transformación de la información contenida en las HCEs presentan una serie de problemas que actualmente están siendo afrontados. Especialmente crítico es el hecho de que las HCEs almacenan información de naturaleza heterogénea. Los documentos incluyen comúnmente la información recogida en consulta junto con informes de analítica, anatomía patológica o de carácter radiológico. Por tanto, se convierten en campos textuales de naturaleza no estructurada, que contienen información clínica de varias fuentes [3]. Esta naturaleza no estructurada hace que la tarea de extracción automática de conceptos relevantes sea particularmente difícil, mientras que la extracción manual es no reutilizable y costosa.

El presente trabajo centra su atención en la oncología como ámbito clínico, y en la extracción de la localización de la neoplasia que presenta el paciente como tarea. Se plantea este problema concreto debido a que el grupo de Inteligencia Computacional en Biomedicina (ICB) lleva más de 15 años trabajando en el campo oncológico en estrecha colaboración con la Unidad de Gestión Clínica Intercentros de Oncología Médica de Málaga (UGCIO). En conjunto se desarrolló el sistema Galén [3], [4], un sistema informático integrado en los centros oncológicos de la provincia de Málaga, que recoge las HCEs y demás información clínica relevante, y los pone a disposición de la investigación clínica.

Un problema recurrente en las unidades oncológicas es la falta de tiempo del personal clínico para completar la información de los pacientes, incluyendo la localización de la neoplasia. Esta información se encuentra habitualmente presente en las HCEs en formato texto, pero comúnmente no se almacena en un campo electrónico específico. Sin embargo, en Galén se tiene asignado este tipo de información a las HCEs, lo que nos proporciona el entorno y los medios necesarios para desarrollar modelos de IA para la extracción automática.

Concretamente, la tarea de extracción automática de información se realiza mediante el procesamiento del lenguaje natural (*Natural Language Proccesing*, NLP). El NLP es una tecnología de IA capaz de interpretar, manipular y comprender el lenguaje humano, y cuya reciente difusión explosiva en la

sociedad se debe a los modelos aplicados en IA Generativa y chats conversacionales. En concreto, la tarea de extracción automática de la localización de la neoplasia es equiparable a la clasificación de textos. Esta tarea se define como la asignación de unidades textuales a una o varias categorías en función del contenido y la semántica presentes en el texto. Entre los enfoques de clasificación de texto se encuentran los métodos basados en reglas [5], en aprendizaje automático (ML) [6] y en aprendizaje profundo (DL) [7], [8]. En este último enfoque se encuentran las redes neuronales recurrentes (RNN) y los *Large Language Models* (LLM). Estos últimos conforman actualmente el estado del arte (SOTA) [9]. Estos grandes modelos lingüísticos preentrenados, basados en las arquitecturas Transformer [10], han superado la capacidad de los sistemas tradicionales para identificar elementos importantes, y han ganado una prominencia considerable en el campo biomédico [11], [12].

Considerando los aspectos anteriores, en este trabajo proponemos avanzar en la aplicación de modelos SOTA de NLP para la extracción automática de la localización de la neoplasia, conforme a un agrupamiento realizado del CIE-10-ES, a partir de HCEs escritas en español. Hemos experimentado con transformers entrenados con corpus multilingües y otros monolingües en español; además de modelos entrenados o afinados con corpus de propósito general, y otros adaptados al dominio biomédico o clínico. Hasta donde sabemos, este es el primer estudio que examina la aplicación de modelos basados en transformers para la extracción de información sobre la localización de la neoplasia a partir de HCEs reales en español.

II. Trabajos relacionados

Con la organización de diferentes retos y tareas compartidas de NLP, y la creciente adopción de HCE en los sistemas sanitarios de todo el mundo, los estudios de extracción automática de información han aumentado. Entre estos destaca la extracción de códigos ICD-10 (versión internacional del CIE-10) para codificar procedimientos y clasificar la mortalidad a partir de notas clínicas. Inicialmente, los modelos utilizados para esta tarea se basaban en modelos basados en reglas y aprendizaje automático [5], [13]. Recientemente, la irrupción de los modelos de DL ha llevado al desarrollo de modelos basados en RNN, con una mejor capacidad para considerar un mayor número de categorías simultáneas [14], incluyendo información oncológica [15]. A este respecto, mencionar el trabajo realizado por nuestro grupo ICB en una versión preliminar del problema tratado en el presente estudio [8], donde se demostró la capacidad de las RNN para extraer la información de la neoplasia teniendo en cuenta únicamente las 3 neoplasias con mayor incidencia: mama, pulmón y colón/recto.

Actualmente los modelos de DL basados en transformers establecen el SOTA en la tarea de extracción automática de códigos CIE-10 [16], [17], no solo en documentos en inglés, sino también en idiomas como el francés, italiano o portugués [18]–[20]. Cabe destacar un trabajo publicado recientemente [21], en el cual se realiza un entrenamiento de un gran modelo transformer en la tarea de reconocimiento de entidades, para localizar la topografía y la histología de neoplasias en textos clínicos en inglés, obteniendo resultados prometedores.

En español, varios proyectos se han esforzado por utilizar los datos de las HCEs no estructuradas. Sin embargo, debido a la limitada disponibilidad de corpus clínicos anotados, la tarea sigue siendo un reto. Podemos nombrar la tarea CodiEsp del CLEF eHealth 2020, que abordaba la codificación automática CIE-10 de términos referentes a diagnóstico y procedimientos en un corpus artificial de HCEs en español. En CodiEsp se examinaron diversas estrategias de aprendizaje profundo, obteniendo los modelos basados en transformers resultados prometedores [22], [23]. También se encuentra el reto CANTEMIST (*CANcer TExt MIning SharedTask*), cuyo objetivo era la extracción de conceptos de cáncer, centrándose en morfología tumoral, en registros médicos españoles artificiales [24]. Aquí destacamos los trabajos realizados por el grupo ICB, en los que se pre-entrenaron modelos transformer con un corpus general de Galén para abordar las tareas CodiEsp-D y CANTEMIST, obteniendo resultados SOTA.

III. Materiales de estudio

En esta sección se describe el corpus utilizado para realizar la extracción de la localización de la neoplasia que presenta un paciente. Como se ha mencionado con anterioridad, el equipo de investigación pudo obtener información de calidad de forma sencilla gracias a la disponibilidad del sistema Galén [4], el cuál recoge información de más de $60,000$ pacientes oncológicos de la UGCIO de Málaga. Concretamente, se seleccionó un corpus conformado por $23,704$ HCEs que contenían una neoplasia primaria asociada y más de 500 palabras. Cada documento incluía la información demográfica, la primera visita e información de los episodios y consultas restantes. Además, expertos autorizados des-identificaron los documentos para cumplir con la Ley Orgánica Española de Protección de Datos Personales y Garantía de Derechos Digitales (LOPD-GDD). Esto también permite asegurar la imposibilidad de que los modelos de NLP asocien nombres de clínicos o centros hospitalarios con ciertas neoplasias.

Siendo un aspecto importante de la práctica clínica la estandarización de conceptos y códigos, es recomendable utilizar el sistema del *International Statistical Classification of Diseases and Related Health Problems 10th edition* (ICD-10) [25]. Este sistema, con su equivalencia en español CIE-10-ES, incluye una codificación para la localización (topografía) y la histología (morfología) de tumores y neoplasias. Si bien cada posición dentro de la codificación otorga cierto grado de información, en el problema planteado únicamente se utilizan las posiciones referentes a la localización. En concreto, las correspondientes a tumores malignos de localización primaria cuyos códigos se encuentran comprendidos entre el C00 y el C97. Dada la distribución de grandes grupos de tumores del CIE-10-ES y la presencia en el sistema de información Galén, las categorías han sido agrupadas según lo descrito en la Tabla I. En esta se muestra cada código, la localización asignada al mismo y su presencia en el corpus seleccionado, el número absoluto (abs) de documentos y su frecuencia relativa (rel).

TABLA I
CÓDIGO, LOCALIZACIÓN ASOCIADA, Y NÚMERO Y FRECUENCIA DE DOCUMENTOS POR AGRUPACIÓN REALIZADA EN EL CORPUS.

Código	Localización	Presencia abs	Presencia rel
C15-C26	Órganos digestivos	1837	.0775
C18-C21	Colón, recto y ano	4020	.1696
C30-C39	Órganos respiratorios e intratorácicos	3371	.1422
C43-C44	Piel	661	.0279
C45-C49	Tejidos mesoteliales y tejidos blandos	1462	.0617
C50	Mama	6491	.2738
C51-C58	Órganos genitales femeninos	1335	.0563
C60-C63	Órganos genitales masculinos	1187	.0501
C64-C68	Vías urinarias	870	.0367
C81-C96	Tejido linfático y órg. hematopoyéticos	1133	.0478
SARCS	Sarcoma en huesos y tejidos blandos	592	.0250
OTROS	Otras localizaciones	745	.0314
Total		23704	

Es importante mencionar la división realizada entre las localizaciones de colon/recto/ano (C18-C21) y órganos digestivos (C15-C26), debido a la gran presencia de documentos asignados a los primeros y su importancia en la sociedad española [26]. También, mencionar que la categoría SARCS engloba los tumores en huesos y cartílagos articulares (C40-C41) y los sarcomas en tejidos blandos. Por último, se engloban otras localizaciones dentro de la categoría OTROS, debido al bajo número de documentos pertenecientes a estas en el corpus definido. Concretamente, OTROS engloba a tumores con código: C97, con una presencia relativa del 0.014 global; C76-C80, con presencia igual a 0.01; y C00-C14, C69-C72 y C73-C75, siendo la presencia conjunta de estas tres localizaciones del 0.007 en el corpus.

IV. MÉTODOS DE NLP

La metodología desarrollada para abordar el problema de clasificación está basada en las arquitecturas Transformer. Los transformers utilizan el mecanismo de autoatención multi-cabezal (*multi-head self-attention*) [10] para crear una representación numérica contextual de cada palabra de la entrada, así como para aumentar la eficiencia computacional mediante la paralelización. Los modelos transformers han obtenido una enorme popularidad debido especialmente al enfoque de aprendizaje por transferencia. Estos modelos pueden preentrenarse en corpus de dominio general y afinarse posteriormente en corpus de dominio específico para abordar una determinada tarea [27]. Se han obtenido resultados SOTA tanto en dominios biomédicos como clínicos empleando transformers en combinación con aprendizaje por transferencia [11], [28].

En este trabajo, al tratarse de la clasificación de textos médicos en español, hemos utilizado 6 modelos distintos basados en transformers que soportan el idioma español.

- XLM-R: Esta versión multilingüe de la arquitectura RoBERTa se preentrenó en un enorme corpus CommonCrawl de dominio general de 2.4TB en 100 idiomas [29], utilizando un gran vocabulario multilingüe de $\sim$250K tokens. Experimentamos con la versión Base ($\sim$277M de parámetros entrenables).
- XLM-R-Galén: Este modelo representa una versión específica del dominio de la arquitectura XLM-R Base. En concreto, se obtuvo realizando un preentrenamiento sobre un corpus de textos clínicos reales sin etiquetar extraídos de Galén [30], con el objetivo de adaptar el modelo a las particularidades del dominio clínico en español.
- RoBERTa-BNE: La versión en español de dominio general de la arquitectura RoBERTa se preentrenó en un corpus de 570GB obtenido de la Biblioteca Nacional de España (BNE) [31]. El modelo emplea un vocabulario español de $50,000$ tokens. Experimentamos con la versión Base ($\sim$124M de parámetros entrenables).
- RoBERTa-Bio: Este modelo basado en transformers constituye un modelo de lenguaje biomédico preentrenado para español [32]. Se obtuvo preentrenando la arquitectura RoBERTa Base desde cero en varios corpus biomédico-clínicos en español recogidos de recursos disponibles públicamente, así como en un corpus clínico del mundo real. El modelo utiliza vocabulario específico del dominio en español de $\sim$52K tokens.
- RoBERTa-Base-Biomed: Modelo de dominio biomédico preentrenado para español [33]. Este modelo basado en RoBERTa, se preentrenó en un corpus biomédico-clínico en español recogido de varias fuentes públicas. El modelo emplea un vocabulario español de $\sim$960M de tokens.
- BETO-Galen: Versión del modelo BETO [34] preentrenado sobre un corpus de textos clínicos reales sin etiquetar extraídos de Galén [30]. BETO-Galén utiliza la misma configuración y vocabulario que el BETO original.

Hemos desarrollado un enfoque integral para abordar el problema de la clasificación utilizando transformers, afinando los modelos sobre el corpus clínico obtenido de Galén. Como entrada de los modelos se proporciona cada secuencia de palabras de los documentos médicos tokenizada en una secuencia de subpalabras. Los resultados computados por los transformers a nivel de subpalabra pasan a alimentar una capa lineal de neuronas, las cuales se encargan de procesar esos resultados para proporcionar una de las categorías del problema. Durante el entrenamiento, se afina el modelo transformer y se entrena a la capa de neuronas para aumentar su rendimiento en la tarea de clasificación descrita. Una descripción visual de la metodología desarrollada se muestra en la Fig. 1.

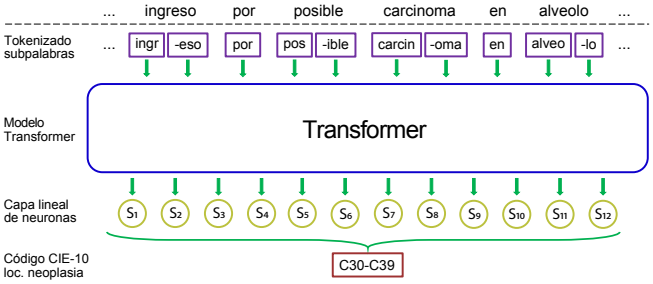

Fig. 1. Ilustración de la metodología basada en transformers incluyendo la capa lineal de neuronas para su aplicación en el problema de clasificación.

V. Experimentos y resultados

En esta sección se presenta la experimentación realizada con la metodología anteriormente explicada, así como los resultados obtenidos del estudio.

Para la experimentación, se realiza una división estratificada del corpus en conjuntos de entrenamiento, validación y test en 10 cajas, de las cuales 8 forman el conjunto de entrenamiento y las dos restantes forman el conjunto de validación y test respectivamente. Se realiza además un proceso de ajuste de hiperparámetros de los modelos, variando el tamaño de batch y la tasa de aprendizaje, utilizando los resultados obtenidos con el conjunto de validación. Durante el entrenamiento del modelo se observa además el rendimiento alcanzado en cada época con el conjunto de validación, con el propósito de realizar una parada temprana del entrenamiento. Por último, se realiza la inferencia sobre el conjunto de test y se obtienen las métricas tomadas como representativas del rendimiento honesto del modelo en cuestión.

Para evaluar la metodología desarrollada, se utilizan las métricas de porcentaje de aciertos o *accuracy* (Acc), precisión (P), sensibilidad o (*recall* (R), y F1-score (F1). La métrica más significativa para el estudio es la F1-score, la media armónica de la precisión y la sensibilidad. Esta proporciona una medida fiable del rendimiento alcanzado por los modelos en problemas en los que la sensibilidad es importante o existe un desbalanceo significativo de las clases a predecir.

TABLA II
Métricas de evaluación calculadas en el conjunto de test. Para cada modelo, se calculan las métricas de porcentaje de aciertos (Acc), precisión (P), sensitividad (R) y F1-score (F1).

Modelo	Acc	P	R	F1
XLM-R	.9240	.8774	.8787	.8776
XLM-R-Galen	.9350	.8894	.8949	.8913
RoBERTa-Base-bne	.9329	*.8972*	.8887	.8921
RoBERTa-Bio	*.9405*	.8950	**.9061**	*.8987*
RoBERTa-Base-Biomed	**.9456**	**.9204**	*.8982*	**.9080**
BETO-Galen	.9025	.8503	.8421	.8436

Se ha seguido el proceso de experimentación propuesto anteriormente, y en la Tabla II se muestran la evaluación obtenida por los modelos propuestos en el problema de clasificación. Los mejores valores alcanzados se muestran en negrita, mientras que los segundos mejores valores se muestran en cursiva. De los resultados mostrados en la Tabla II se puede concluir por un lado, un mayor rendimiento de los modelos preentrenados con información de carácter biomédico y clínico frente al modelo de dominio genérico (XLM-R), mejorando casi en un 3% el valor de F1-score. Por otro lado, los modelos RoBERTa obtienen los mejores resultados frente a dos modelos que fueron preentrenados en un corpus similar al corpus objetivo (modelos -Galén). Específicamente, el RoBERTa-Base-Biomed obtiene los mejores valores con 0.946 de porcentaje de aciertos y un 0.908 de F1-score.

Con el objetivo de realizar un análisis del rendimiento de RoBERTa-Base-Biomed, se muestran en la Tabla III los resultados obtenidos por el modelo para cada una 12 categorías

TABLA III
Métricas para cada grupo de códigos CIE-10-ES obtenidas por el modelo RoBERTa-Base-Biomed

Código	P	R	F1	Soporte
C15-C26	.8698	.9689	.9167	193
C18-C21	.9673	.9698	.9686	397
C30-C39	.9160	.9675	.9410	338
C43-C44	.9667	.9062	.9355	64
C45-C49	.9259	.8681	.8961	144
C50	.9984	.9923	.9954	649
C51-C58	.9362	.9565	.9462	138
C60-C63	.9573	.9492	.9532	118
C64-C68	.9268	.8539	.8889	89
C81-C96	.9640	.9386	.9511	114
SARCS	.8235	.7500	.7850	56
OTROS	.7931	.6571	.7188	70
Promedio	.9204	.8982	.9080	2370

CIE-10-ES agrupadas por cada clase concreta. Aparte de las métricas de evaluación, se muestra el número de instancias presentes en el conjunto de test (*Soporte*). El modelo obtiene un rendimiento mayor a 0.90 en F1-score para 8 de las 12 clases. El mejor resultado se obtiene en los documentos pertenecientes a C50 (localización mama) con un 0.995 de F1-score, siendo estos los más presentes. Las categorías con peor rendimiento son los sarcomas (SARCS) y los agrupados en OTROS, dos de las localizaciones menos presentes. Destacar el buen rendimiento en la categoría C43-C44 (localización piel) con un F1-score igual a 0.896, a pesar del reducido número de documentos presentes en el corpus.

La Tabla IV presenta una comparación de las métricas obtenidas de diferentes técnicas teniendo en cuenta únicamente las 3 neoplasias con mayor incidencia en el territorio español [26] (mama, pulmón y colorrectal). Los modelos ML, embedding y RNN son elegidos de nuestro trabajo anterior [8]. Estos modelos son entrenados siguiendo la experimentación del actual estudio, pues el corpus utilizado en [8] estaba ofuscado y no des-identificado. El modelo transformer elegido es RoBERTa-Base-Biomed, y se presentan los resultados con el modelo entrenado con las 12 clases (*), y entrenado específicamente para predecir únicamente las 3 clases principales. El rendimiento alcanzado por las otras metodologías de NLP supera el obtenido por los transformers. Concretamente, la red convolucional 2-BidirectionalGRU con un embedding fastText preentrenado (FT + C-2-BiGRU) obtiene el mejor valor de F1-score igual a 0.982, mientras que RoBERTa-Base-Biomed obtiene un valor de 0.9695 de F1-score.

TABLA IV
Métricas para la clasificación en localización de Mama/Pulmón/Colorrectal/Otros, comparando los mejores modelos de diferentes técnicas de NLP: ML, embedding y RNN.

		macro		
Técnica	Mejor Modelo	P	R	F1
ML	SVM	*.9760*	.9744	.9752
Embedding	fastText	.9758	*.9770*	*.9764*
RNN	FT + C-2-BiGRU	**.9816**	**.9800**	**.9808**
Transformer	RoBERTa-Base-Biomed	.9753	.9641	.9695
Transformer	RoBERTa-Base-Biomed*	.9645	.9725	.9683

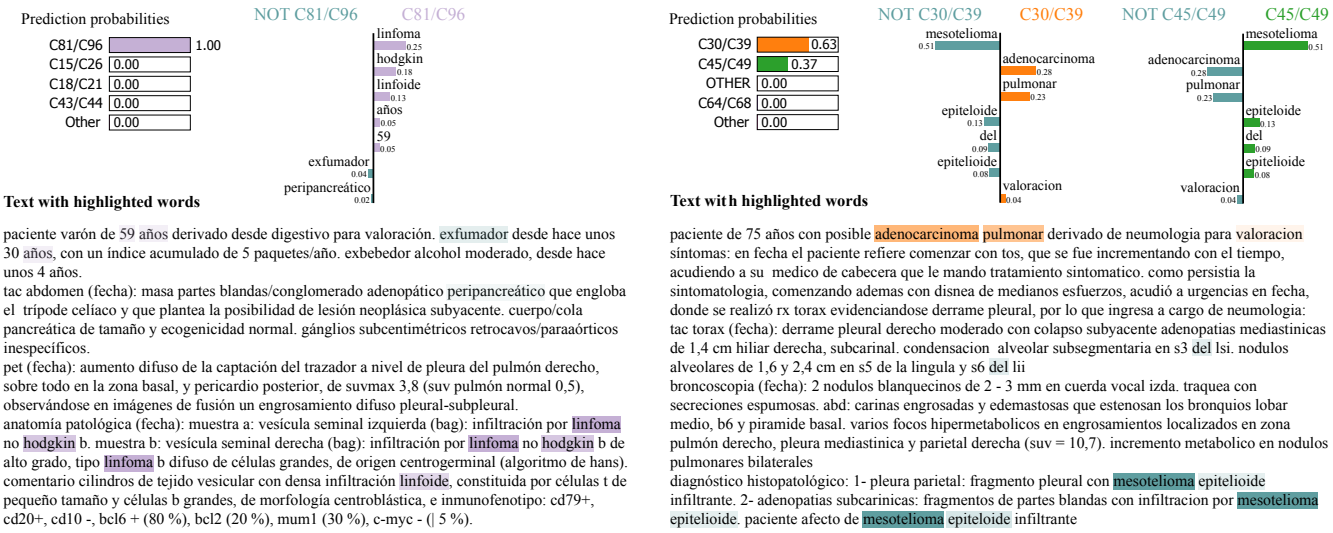

Fig. 2. Textos explicados con LIME y clasificados usando el modelo RoBERTa-Base-Biomed. A la izquierda un texto clasificado correctamente en C81-C96 (tejido linfático), y a la derecha un texto clasificado erróneamente en C30-C39 (órganos respiratorios) siendo su clase real C45-C49 (tejidos mesoteliales).

Finalmente, intentando obtener explicabilidad del funcionamiento de los modelos transformers, se emplea la técnica LIME (*Local Interpretable Model-Agnostic Explanations*) [35] . Esta técnica aproxima un modelo de aprendizaje de caja negra con un modelo local interpretable para explicar cada predicción individual. Para ello, modifica la entrada creando nuevas muestras e interpreta la contribución de las diferentes características a la salida. La Fig. 2 muestra la explicabilidad proporcionada por LIME para la clasificación realizada por el modelo RoBERTa-Base-Biomed con dos textos de ejemplo. Se muestran las probabilidades de pertenencia a la distintas clases predichas y la relevancia de hasta 7 palabras, y se acompaña del texto con las palabras relevantes destacadas.

El texto en la Fig. 2 (izquierda) ha sido clasificado correctamente en C81-C96 (tejido linfático), explicado con la aparición entre otros de la palabra "*linfoma*". Mientras, el texto en la Fig. 2 (derecha) es clasificado incorrectamente en C30-C39 (órganos respiratorios) por la aparición de "*adenocarcinoma pulmonar*", si bien detecta también la palabra "*mesotelioma*" que corresponde a su clase real C45-C49 (tejidos mesoteliales).

VI. Conclusión y trabajo futuro

El estudio se centra en la tarea crucial de extraer automáticamente la localización de neoplasias a partir de documentos clínicos reales del sistema de salud español. El corpus utilizado está compuesto por $23,704$ historiales clínicos electrónicos del servicio oncológico malagueño obtenidos del sistema Galén [4]. Para el proceso de extracción, se ha analizado el rendimiento de los modelos transformers, reconocidos por su eficacia en diferentes tareas de NLP. Destaca el rendimiento excepcional de los modelos basados en la arquitectura RoBERTa entrenados con textos de dominio clínico y biomédico, respecto a los modelos de dominio general y los entrenados con un corpus Galén similar. Específicamente, el modelo RoBERTa-Base-Biomed alcanzó una precisión, sensibilidad y F1-score macro-promediadas de 0.920, 0.898 y 0.908, respectivamente. Esto resultados refuerzan la utilidad de los transformers en la extracción de información clínica relevante de los registros clínicos.

Respecto al rendimiento analizado por cada categoría CIE-10-ES, destaca especialmente la clasificación del cáncer de mama (C50), con un F1-score igual a 0.995. Este rendimiento podría explicarse por su alta incidencia en el corpus, estando el modelo más expuesto a esta localización. Sin embargo, la categoría C43-C44 logró un F1-score de 0.936, a pesar de su menor representación en el corpus (2.8%). Esto sugiere una capacidad robusta del modelo transformer para extraer correctamente la localización incluso para neoplasias menos frecuentes. Ocho de las doce codificaciones superaron un valor de 0.9 en sensibilidad y F1-score, mostrando un rendimiento inferior para las categorías SARCS y OTROS. Estas incluyen algunas de las neoplasias más complejas de diagnosticar, como los sarcomas de tejidos blandos y los tumores germinales y de sitios desconocidos. Como se observó aplicando la técnica LIME, la presencia de palabras asociadas a otras localizaciones durante el proceso clínico puede explicar el inferior rendimiento en estas categorías.

Adicionalmente, se realizó una comparación del modelo transformer con otras técnicas de NLP en la tarea de clasificación de las tres principales localizaciones de neoplasias en España. Los modelos basados en RNN y ML obtuvieron mejores resultados, pero probablemente en tareas más complejas los transformers aporten un mayor rendimiento.

En futuros trabajos, se pretende investigar la extracción de la localización de la neoplasia con un mayor número de categorías, desgranando algunas de las agrupaciones realizadas para el presente trabajo. Por último, se intentará validar los modelos con corpus reales de otros hospitales españoles.

AGRADECIMIENTOS

Los autores agradecen el apoyo de la Agencia Estatal de Investigación bajo el proyecto PID2020-116898RB-I00/AEI/10.13039/501100011033, y del apoyo de Pfizer S.L., la Universidad de Málaga y de la Fundación General de UMA (UMA-FGUMA-Pfizer) mediante fondos privados para el contrato 807/47.6383.

REFERENCIAS

[1] H.-G. Eichler, F. Pignatti, B. Schwarzer-Daum, A. Hidalgo-Simon, I. Eichler, P. Arlett, A. Humphreys, S. Vamvakas, N. Brun, and G. Rasi, "Randomized controlled trials versus real world evidence: neither magic nor myth," *Clinical Pharmacology & Therapeutics*, vol. 109, no. 5, pp. 1212–1218, 2021.

[2] J. Concato and J. Corrigan-Curay, "Real-world evidence-where are we now?" *The New England journal of medicine*, vol. 386, no. 18, pp. 1680–1682, 2022.

[3] D. Urda, N. Ribelles, J. L. Subirats, L. Franco, E. Alba, and J. M. Jerez, "Addressing critical issues in the development of an oncology information system," *International journal of medical informatics*, vol. 82, no. 5, pp. 398–407, 2013.

[4] N. Ribelles, J. M. Jerez, D. Urda, J. L. Subirats, A. Márquez, C. Quero, and L. Franco, "Galén: Sistema de información para la gestión y coordinación de procesos en un servicio de oncología," *RevistaeSalud*, vol. 6, no. 21, pp. 1–12, 2010.

[5] S. V. Pakhomov, J. D. Buntrock, and C. G. Chute, "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques," *Journal of the American Medical Informatics Association*, vol. 13, no. 5, pp. 516–525, 2006.

[6] J. St-Maurice, M.-H. Kuo, and P. Gooch, "A proof of concept for assessing emergency room use with primary care data and natural language processing," *Methods of Information in Medicine*, vol. 52, no. 01, pp. 33–42, 2013.

[7] R. Wang, Z. Li, J. Cao, T. Chen, and L. Wang, "Convolutional recurrent neural networks for text classification," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–6.

[8] F. J. Moreno-Barea, H. Mesa, N. Ribelles, E. Alba, and J. M. Jerez, "Clinical text classification in cancer real-world data in spanish," in *International Work-Conference on Bioinformatics and Biomedical Engineering*. Springer, 2023, pp. 482–496.

[9] M. Khadhraoui, H. Bellaaj, M. B. Ammar, H. Hamam, and M. Jmaiel, "Survey of bert-base models for scientific text classification: Covid-19 case study," *Applied Sciences*, vol. 12, no. 6, p. 2891, 2022.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[12] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[13] M. Subotin and A. Davis, "A system for predicting icd-10-pcs codes from electronic health records," in *Proceedings of bionlp*, 2014, pp. 59–67.

[14] P.-F. Chen, S.-M. Wang, W.-C. Liao, L.-C. Kuo, K.-C. Chen, Y.-C. Lin, C.-Y. Yang, C.-H. Chiu, S.-C. Chang, F. Lai *et al.*, "Automatic icd-10 coding and training system: deep neural network based on supervised learning," *JMIR Medical Informatics*, vol. 9, no. 8, p. e23230, 2021.

[15] S. Baker, A. Korhonen, and S. Pyysalo, "Cancer hallmark text classification using convolutional neural networks," in *Proceedings of the 5th Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM)*, 2016, pp. 1–9.

[16] Z. Zhang, J. Liu, and N. Razavian, "Bert-xml: Large scale automated icd coding using bert pretraining," *arXiv preprint arXiv:2006.03685*, 2020.

[17] E. Al-Bashabsheh, A. Alaiad, M. Al-Ayyoub, O. Beni-Yonis, R. A. Zitar, and L. Abualigah, "Improving clinical documentation: automatic inference of icd-10 codes from patient notes using bert model," *The Journal of Supercomputing*, pp. 1–25, 2023.

[18] G. Bouzille and N. Grabar, "Supervised learning for the icd-10 coding of french clinical narratives," *Digital Personalized Health and Medicine: Proceedings of MIE 2020*, vol. 270, p. 427, 2020.

[19] S. Silvestri, F. Gargiulo, M. Ciampi, and G. De Pietro, "Exploit multilingual language model at scale for icd-10 clinical text classification," in *2020 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2020, pp. 1–7.

[20] A. D. Reys, D. Silva, D. Severo, S. Pedro, M. M. de Sousa e Sá, and G. A. Salgado, "Predicting multiple icd-10 codes from brazilian-portuguese clinical notes," in *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*. Springer, 2020, pp. 566–580.

[21] M. Chandrashekar, I. Lyngaas, H. A. Hanson, S. Gao, X.-C. Wu, and J. Gounley, "Path-bigbird: An ai-driven transformer approach to classification of cancer pathology reports," *JCO Clinical Cancer Informatics*, vol. 8, p. e2300148, 2024.

[22] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, and M. Krallinger, "Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020." *CLEF (Working Notes)*, vol. 2020, 2020.

[23] S. Amin, G. Neumann, K. Dunfield, A. Vechkaeva, K. A. Chapman, and M. K. Wixted, "Mlt-dfki at clef ehealth 2019: Multi-label classification of icd-10 codes with bert." in *CLEF (Working Notes)*, 2019, pp. 1–15.

[24] A. Miranda-Escalada, E. Farré, and M. Krallinger, "Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results." *IberLEF@ SEPLN*, pp. 303–323, 2020.

[25] W. H. Organization, *International Statistical Classification of Diseases and related health problems*. World Health Organization, 2004, vol. 3.

[26] S. E. de Oncología Médica (SEOM), "Las cifras del cáncer en españa 2022," 2022.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[28] G. López-García, J. M. Jerez, N. Ribelles, E. Alba, and F. J. Veredas, "Detection of Tumor Morphology Mentions in Clinical Reports in Spanish Using Transformers," in *Advances in Computational Intelligence*. Cham: Springer International Publishing, 2021, pp. 24–35.

[29] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Jul. 2020, pp. 8440–8451.

[30] G. López-García, J. M. Jerez, N. Ribelles, E. Alba, and F. J. Veredas, "Transformers for Clinical Coding in Spanish," *IEEE Access*, vol. 9, pp. 72 387–72 397, 2021.

[31] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Villegas, "MarIA: Spanish Language Models," *Procesamiento del Lenguaje Natural*, vol. 68, no. 0, pp. 39–60, 2022.

[32] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, and M. Villegas, "Pretrained biomedical language models for clinical NLP in Spanish," in *Proceedings of the 21st Workshop on Biomedical Language Processing*. Assoc. for Computational Linguistics, 2022, pp. 193–199.

[33] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, and M. Villegas, "Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario," 2021.

[34] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, "Spanish pre-trained bert model and evaluation data," *arXiv preprint arXiv:2308.02976*, 2023.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.