Detección Automática de Patologías en Notas Clínicas en Español Combinando Modelos de Lenguaje y Ontologías Médicos

Automatic Pathology Detection in Spanish Clinical Notes Combining Language Models and Medical Ontologies

León-Paul Schaub-Torre[1,*,†], Pelayo Quirós[1,†] and Helena García-Mieres[1]

[1]CTIC Technology Centre. W3C Spain Office host, Ada Byron 39, Gijón, 33203, Asturias, Spain

Resumen

En este artículo presentamos un método híbrido para la detección automática de patologías dermatológicas en informes médicos. Usamos un modelo de lenguaje amplio en español combinado con ontologías médicas para predecir, dado un informe médico de primera cita o de seguimiento, la patología del paciente. Los resultados muestran que el tipo, la gravedad y el sitio en el cuerpo de una patología dermatológica, así como en qué orden tiene un modelo que aprender esas tres características, aumentan su precisión. El artículo presenta la demostración de resultados comparables al estado del arte de clasificación de textos médicos con una precisión de 0.84, micro y macro F1-score de 0.82 y 0.75, y deja a disposición de la comunidad tanto el método como el conjunto de datos utilizado.

Abstract

In this paper we present a hybrid method for the automatic detection of dermatological pathologies in medical reports. We use a large language model combined with medical ontologies to predict, given a first appointment or follow-up medical report, the pathology a person may suffer from. The results show that teaching the model to learn the type, severity and location on the body of a dermatological pathology as well as in which order it has to learn these three features significantly increases its accuracy. The article presents the demonstration of state-of-the-art results for classification of medical texts with a precision of 0.84, micro and macro F1-score of 0.82 and 0.75, and makes both the method and the dataset used available to the community.

Palabras clave

modelo de lenguaje, biomédico, ontología, método híbrido

Keywords

language model, biomedical, ontology, hybrid method

1. Introducción

La digitalización de informes médicos (EHR por *electronic health records*) es una iniciativa internacional que lleva décadas en desarrollo. Uno de los primeros protocolos de digitalización de los informes es el ISO TC 215[1] creado en 1998. Se trata de una norma cuyo objetivo es estandarizar la digitalización de los informes de más de 50 países, incluyendo España, tanto de tipo fotográfico (radiología, ecografía, etc.), como textual. Esto permite tener un contexto y un historial de cada paciente, así como facilitar los seguimientos [1]. Sin embargo, la aceleración de esta digitalización en los últimos 15 años, con la globalización de Internet a alta velocidad y de la capacidad de los servidores, ha provocado un crecimiento de la cantidad de datos. Es por eso por lo que el procesamiento del lenguaje natural (PLN) tiene gran potencial como herramienta de ayuda a los

SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.

*Corresponding author.

leon.schaub@fundacionctic.org (L. Schaub-Torre); pelayo.quiros@fundacionctic.org (P. Quirós); helenagmieres@gmail.com (H. García-Mieres)

0000-0002-0116-9698 (L. Schaub-Torre); 0000-0002-0500-9034 (P. Quirós); 0000-0002-2813-1737 (H. García-Mieres)

†These authors contributed equally.

[1]https://www.iso.org/committee/54960.html

médicos para facilitarles el trabajo de seguimiento de pacientes, al preanalizar los EHR [2] extraer entidades (NER) [3], o predecir las patologías que padece una persona [4]. En paralelo, los progresos en aprendizaje profundo [5] desde los años 2010 y los transformadores a partir de 2018 [6] han permitido la creación de modelos más precisos [7]. Combinando ambos avances, en los últimos años se han desarrollado modelos pre-entrenados de lenguaje especializados en el vocabulario médico, y ajustados (*fine-tuned*) para las aplicaciones mencionadas [8, 9, 10]. En lengua española y en cualquier otro idioma distinto del inglés [11] los recursos existentes son limitados, pero modelos como los desarrollados por [12, 13] consiguen resultados comparables a modelos en lengua inglesa para tareas de extracción de información, [14] como es el caso de para NER.

Sin embargo, son pocos los trabajos que se enfocan en la predicción de una enfermedad dentro de un informe clínico [15]. Existen trabajos de encaje léxico que han tenido éxito para conectar un informe y un concepto (por ejemplo, enfermedad o tipo de enfermedad) [16, 17], logrando superar a los trabajos de ontología y de semántica de la última década [18, 19]. Pese a ello, apenas existen corpus de referencia para tener a la vez informes médicos en español y la patología asociada, habiendo identificado como única referencia el corpus CARES [20], si bien está centrado en datos radiológicos. Tampoco se ha detectado un método del estado de la técnica que sea capaz de predecir a qué patología(s) corresponde un determinado informe médico.

Por otra parte, la motivación de este trabajo viene dada de no ser NER una tarea adaptada a nuestro problema por dos motivos:

1. No tenemos un conjunto de datos etiquetado en entidades nombradas (EN), lo cual supondría realizar una campaña de etiquetado y contar con un conocimiento experto del cual no disponemos.
2. Aunque tuviéramos ese conjunto etiquetado, un análisis cualitativo de los datos muestra que la presencia de ciertas EN no se corresponden con la patología que hemos de predecir. Por ejemplo, en el caso de sospechas, dudas o negaciones, el informe puede contener una EN como "Se sospecha una *queratosis* aparente que resulta ser un bulto maligno".

Buscamos resolver este problema, para lo cual presentamos un método híbrido que combina los transformadores con un modelo basado en RoBERTa [21] y ontologías [22]. Con este fin, creamos con ellos modelos en cascada: detectan el tipo (síntoma, proceso neoplásico, etc.), el sitio anatómico y la gravedad de la patología y un último modelo que, gracias a los anteriores, predice qué patología es. Los informes utilizados provienen de EHRs y son notas clínicas de pacientes escritas por médicos, que pueden ser de primera cita o un seguimiento. Cada informe tiene dos etiquetas asociadas: la patología y una codificación de la patología. Estos informes provienen de la unidad de dermatología de distintos centros de salud de España. Han sido anonimizados de manera semi automática con técnicas basadas en reglas simbólicas. Las contribuciones que realizamos con el trabajo actual son las siguientes:

- Un conjunto de datos anonimizado de EHR de dermatología en español, público y de libre acceso[2].
- Una nomenclatura de patologías dermatológicas que viene a enriquecer las ontologías y léxicos existentes.
- Un método híbrido basado en transformadores con ontología para la tarea de clasificación de las EHR con respecto a las patologías posibles.

Además de la introducción en la Sección 1, el artículo se divide en otras cuatro secciones. En la Sección 2 se presenta un estado de la cuestión donde resumimos tanto los métodos que se asemejan al nuestro como los recursos lingüísticos que existen. En la Sección 3 se proporciona una descripción de nuestra metodología y de la arquitectura final del modelo. La Sección 4 se centra en los resultados. La Sección 5 aborda la discusión, conclusiones, y los posibles trabajos futuros.

2. Estado de la cuestión

La minería de texto en informes clínicos es un campo importante del PLN desde hace años [23]. Sin embargo, la cantidad de trabajos relacionados con la extracción de información en el ámbito médico ha florecido con la expansión de los EHR [24].

En este sentido, a principios de los 2000 se solía utilizar una combinación de ontología y de web semántica para extraer nombres de enfermedades o de medicamentos [25, 26]. Además, las ontologías se utilizaban combinadas con algoritmos estadísticos [27].

La mayoría de los trabajos en este campo son en inglés, pero algunos como [3, 28] emplearon textos en español. A partir de 2010 se empezaron a usar redes neuronales como los LSTM [29], dado que tienen capacidad para retener relaciones entre las

[2] https://huggingface.co/fundacionctic/DermatES

frases. Estas redes se utilizaron para el NER médico tanto en lengua inglesa [30] como española [31].

Aun así, los trabajos que buscan asociar la totalidad del texto a un concepto de patología son escasos en comparación con los que aspiran a detectar las patologías nombradas.

En conjuntos de datos en inglés encontramos MIMIC [32] y MIMIC-III [33]. Por otro lado, en español está el conjunto privado [34], SPACCC (*Spanish Clinical Cases Corpus*) que no contiene etiquetado [35] o PharmacoNER [36] pero también para NER. El conjunto en español que más se asemeja al del presente trabajo es el CodiEsp [37] pero está etiquetado en diagnóstico y procedimiento. Para consultar conjuntos de datos existentes, se puede consultar la amplia lista creada por [38].

En general, siendo los datos de salud sensibles, y con la necesidad de cumplir el RGPD[3] [39], es imprescindible anonimizar los datos clínicos [40]. Como métodos explorados de anonimización, [41] usan BRAT [42] para anonimizarlos. [43] muestra que un modelo híbrido hace casi imposible la reidentificación de las personas. En conjunto con estos trabajos, nos inspiramos en [44] y MEDOCCAN [45] para anonimizar los nuestros.

Por otro lado, respecto a los métodos que clasifican el texto entero en vez de realizar NER, encontramos a [46], que emplea ontologías para predecir las enfermedades en textos clínicos. [47] aplica los transformadores inspirados en BERT [48] para crear encajes léxicos médicos. Los trabajos que ofrecen mejores resultados en términos de exactitud y precisión son los que combinan la potencia lingüística de los modelos de lengua masivos (LLM) y de las ontologías, por lo que nos inspiramos en estos para diseñar nuestro trabajo. Por ejemplo, [49] implementaron transformadores BERT y ontologías médicas, obteniendo los mejores resultados sobre MIMIC. Hasta donde alcanza nuestro conocimiento, no existe un trabajo en español que describa cómo predecir la patología de un paciente a partir del EHR textual utilizando estos métodos.

3. Metodología

En esta sección presentamos el conjunto de datos que creamos y procesamos, así como la técnica de anonimización que usamos para proteger la privacidad de su contenido. Finalmente, describimos los modelos que utilizamos además de nuestro método híbrido: transformador-ontología con modelos en cascada.

[3] https://www.hacienda.gob.es/es-ES/El%20Ministerio/Paginas/DPD/Normativa_PD.aspx

3.1. Descripción del conjunto de datos

Los datos utilizados se corresponden con notas clínicas de hospitales españoles con respecto a consultas de dermatología, tanto de primera consulta como de revisiones posteriores. Dichos datos vienen dados en ficheros individuales en formato HL7 (*Health Level 7*) [50], el cual se trata de un conjunto de estándares internacionales que permiten el intercambio, integración, compartición y recuperación de datos electrónicos de salud. Además, facilita que la comunicación entre diferentes sistemas sea más ágil y fiable. Nuestro corpus contiene 8881 informes y 173 patologías dermatológicas diferentes. Cada informe contiene una sola etiqueta de una patología dermatológica y estamos ante un caso de clasificación multiclase. Los informes clínicos ligados a 43 variables de diversa índole, incluidas el nombre y el código de la patología. Dado el objetivo planteado en este proyecto, se ha limitado dicho conjunto a dos variables de interés: el texto escrito por parte del facultativo sobre la consulta en lenguaje natural, y la variable que ofrece la patología diagnosticada al paciente dentro de la taxonomía considerada por el sistema de recogida. Un ejemplo del conjunto de datos está ilustrado en la Figura 1.

Teniendo en cuenta el número de patologías y su reparto desequilibrado (Figura 4 en Anexo B), intuíamos que incluso con nuestro método híbrido, muchas clases iban a ser obviadas durante el entrenamiento del modelo. Es por eso que estudiamos definir el umbral del mínimo de ejemplos por patología que maximice la precisión del modelo sin perder demasiadas patologías. Decidimos guardar las 25 patologías más representadas que corresponden a un mínimo de 61 ejemplos por categoría. En el anexo A.6 explicamos que ese umbral es el óptimo para conservar cierto número de categorías sin mermar la eficacia de los modelos y en el anexo.

El conjunto de datos no ha sido lematizado o pasado por otros preprocesamientos clásicos tal y como tampoco lo hicieron [51] explicando que los modelos de lenguaje hoy en día son capaces de procesar textos brutos.

3.2. Anonimización

A lo largo de este apartado se detalla el proceso de anonimizado del conjunto de datos. Los datos médicos pueden presentar información sensible que no ha de compartirse. Por esta razón, se ha procedido con diferentes fases semi-automatizadas que permitan enmascarar información de carácter privado. En un primer nivel, se ha procedido a eliminar todo aquel contenido numérico que aparezca en dichos

qa grados y faciales se aplica crioterapia aff signos inflamatorios en zona implantacion fronto parietal	Queratosis actínica
hoy revision sigue con multiples queratosis actinicas algunas hipertroficas tiene dermatitis seborreica	Queratosis actínica
consulta tfn ap biopsia escisional de piel con carcinoma basocelular bordes libres refiere cicatriz bien	Basocelular (ver carcinoma basocelular)
llama por tfno para actualizar orden de traslado porque le han llamado para revision en laser [Entidad]	Lupus eritematoso sistémico
nuevas lesiones en tronco y mmss superficiales exudativas dolorosas cultivo s aureus sensible amoxi clav	Ulcera sai
a similar colesterol en el limite y transas tb mejor placas muchas residuales y otras menos queratosicas	Psoriasis sai
relaciona con inico de toma de magnesio y antidepresivos no se ha aplicado ninguna crema en las lesiones	Granuloma anular
psoriasis en tratamiento con [Entidad] desde dic muy bien xerosis en palmas de manos resto sin lesiones	Psoriasis sai
dia postqx tiene una pequeña zona en el canto interno con necrosis que cureteamos el resto esta muy bien	Sin diagnostico
no signos de recidiva sobre cicatriz de epidermoide de labio inferior no adenopatias no lesiones nuevas	Carcinoma epidermoide (ver carcinoma espinocelular)

Figura 1: Ejemplo del conjunto de datos. A la izquierda, el informe de primera consulta o de seguimiento. A la derecha, la patología a predecir.

mensajes. Esto se debe a que dicha información está ligada a información sensible: fechas, años, edades o diferentes identificadores (DNI, identificador del paciente, etc.). Por otro lado, se ha abordado la detección de otros tipos de información sensible, que en lugar de eliminar como se ha hecho con los caracteres numéricos, se ha procedido con su enmascaramiento con la etiqueta "[Entidad]". En este caso, se han tratado entidades como nombres propios, apellidos, ciudades/localizaciones o nombres de hospitales. Para llevar a cabo este paso, se ha procedido con la identificación de diferentes fuentes externas que permitan localizar tal información, haciendo uso de las siguientes:

- Lista de apellidos y nombres de hombre y mujer más frecuentes en España, proporcionados por el INE (Instituto Nacional de Estadística)[4].
- Lista de palabras más frecuentes en el lenguaje español mediante recursos proporcionados por la Real Academia Española (RAE) ligadas al Corpus de Referencia del Español Actual (CREA)[5].
- Lista con las ciudades y hospitales más habituales dentro de la fuente de datos utilizada.

Así, el proceso de enmascaramiento se ha desarrollado del siguiente modo:

- Se han identificado todas aquellas apariciones de los nombres propios de hombre o mujer más frecuentes.
- Se han identificado todas aquellas apariciones de los apellidos más frecuentes como primer apellido.
- Entre dichos nombres y apellidos, se filtran aquellos que estén entre los términos más frecuentes para que no sean filtrados.
- Se añaden un total de 43 excepciones que puedan ser relevantes para el ámbito particular de la dermatología (*cabello*, *seco*, *benigno*, etc.).
- Se enmascaran patrones que han sido identificados como candidatos a contener información sensible (texto posterior a los términos *dr*, *dra*, *doctor*, *doctora*).

Con esto, se genera un conjunto de textos anonimizado y enmascarado, que se ha procedido a analizar para validar que la información está protegida. Para ello, se ha realizado una revisión manual por parte de dos revisores:

- Se ha seleccionado una muestra del conjunto de datos anonimizado del 10% del total.
- Se ha realizado dicha selección de forma estratificada con respecto a la categorización del texto original.
- Dicho conjunto se ha dividido a su vez en dos subconjuntos del mismo tamaño, donde cada uno de ellos tiene entradas únicas y entradas compartidas entre sí.

En la Figura 2 se presenta dicha partición generada, así como los tamaños exactos de cada uno de los conjuntos.

Así, se ha validado si el anonimizado de cada texto ha sido adecuado o si se ha equivocado protegiendo información innecesaria o no protegiendo información sensible.

Tras dicha revisión, se han identificado casos particulares a corregir y patrones generalizados. Se han corregido con iteraciones posteriores análogas para garantizar la correcta protección de dicha información.

A nivel de acuerdo interjueces, de las 112 observaciones comunes, los revisores solo han discrepado en 4. Dichos errores han sido analizados y subsanados en el proceso semi-automático previo para todo el conjunto.

[4] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177009&menu=resultados&idp=1254734710990

[5] https://corpus.rae.es/lfrecuencias.html

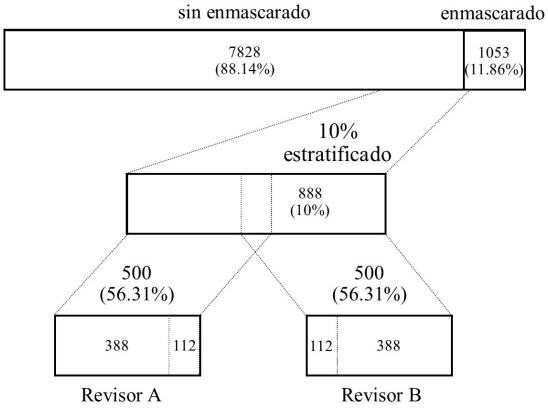


Figura 2: Representación gráfica de la partición generada para la validación de la anonimización realizada.

3.3. Modelo de lenguaje

Dada la potencia y la cantidad de datos necesarios para entrenar un modelo de lenguaje [52], la opción más eficiente para poder usar transformadores es realizar *fine-tuning*. El objetivo de esto es encontrar el modelo preentrenado que mejor se ajuste a nuestros datos y a la detección de una patología. Para este problema, elegimos el bsc-bio-ehr-es[6], un modelo preentrenado de lenguaje biomédico-clínico diseñado para el idioma español. El modelo ha sido preentrenado utilizando datos de textos biomédicos y clínicos en español para aprender patrones lingüísticos específicos. Se basa en la arquitectura de RoBERTa [53]. Sin embargo, los resultados (presentados en la Sección 4.3) evidenciaron la dificultad que tiene un único modelo para aprender a detectar enfermedades, por lo que enriquecimos el entrenamiento usando información de ontologías médicas y varios modelos en cascada, donde cada uno aprende informaciones específicas de lo datos.

3.4. Ontologías utilizadas y modelos en cascada

A lo largo de esta sección, abordamos en un primer nivel el tratamiento del desequilibrio de la variable objetivo, seguido por la revisión de las ontologías médicas y traducción aplicadas, con un apartado final centrado en los modelos en cascada propuestos.

[6] https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es

3.4.1. Tratamiento del desequilibrio de clases

El desequilibrio presente en los datos implica que la mayoría de las clases tienen una cobertura casi nula y provocan un sobreentrenamiento del modelo con las tres primeras clases. Para remediar este problema intentamos reducir la dimensionalidad no de manera matemática con PCA [54] o T-SNE [55], sino con modelos en cascada, cada uno tratando de resolver una tarea más simple y agregando su salida a la entrada del siguiente modelo, hasta ser capaz de predecir la patología exacta. Nos inspiramos en [56] y [57], quienes utilizaron este método para el NER y para el reconocimiento de voz, respectivamente. Por otro lado, en vez de usar métodos probabilísticos para reducir la variabilidad de las clases, introducimos determinismo en la arquitectura de nuestro método mediante ontologías que permiten extraer el tipo de patología, el sitio anatómico afectado, la gravedad o la intensidad. Esto nos permite reagrupar las patologías en relaciones semánticas mas genéricas que consiguen mejorar la precisión a la hora de predecir la patología en cada informe.

3.4.2. Ontologías médicas y traducción

Hubieron trabajos previos que combinaban aprendizaje automático y ontologías [58], pero dadas las características tanto lingüísticas como de dificultad de la tarea, nuestro método es original en términos de extracción de información y de combinación de modelos especializados. Aunque existen ontologías en lengua española como ONTERMET [59] o ECIEMAPS [60], tienen el inconveniente de ser demasiado especializadas o poco completas. Por este motivo decidimos traducir de forma automática [61] el nombre de las patologías de nuestro conjunto de etiquetas del español al inglés con la API Google Translate y usar ontologías médicas más generales y completas como UMLS [62], SNOMED [63], MedDRA [64] y HumanDO [65, 22].

Tras esto, se ha accedido a dicha información a través de las bibliotecas de Python *PyMedTermino* y *PyMedTermino2* [66], así como *medcat* [67] , diseñadas para acceder a estas ontologías. Con estas herramientas, identificamos la codificación correspondiente a cada ontología de las enfermedades analizadas de forma semisupervisada, revisando que la identificación se ajuste a la enfermedad real y no a posibles variaciones similares.

Analizando las características y metadatos de dichas ontologías, extrajimos varios metadatos relevantes. Primero, utilizando SNOMED es posible identificar diferentes sitios anatómicos de la patología a través del *finding site*. Luego, usando

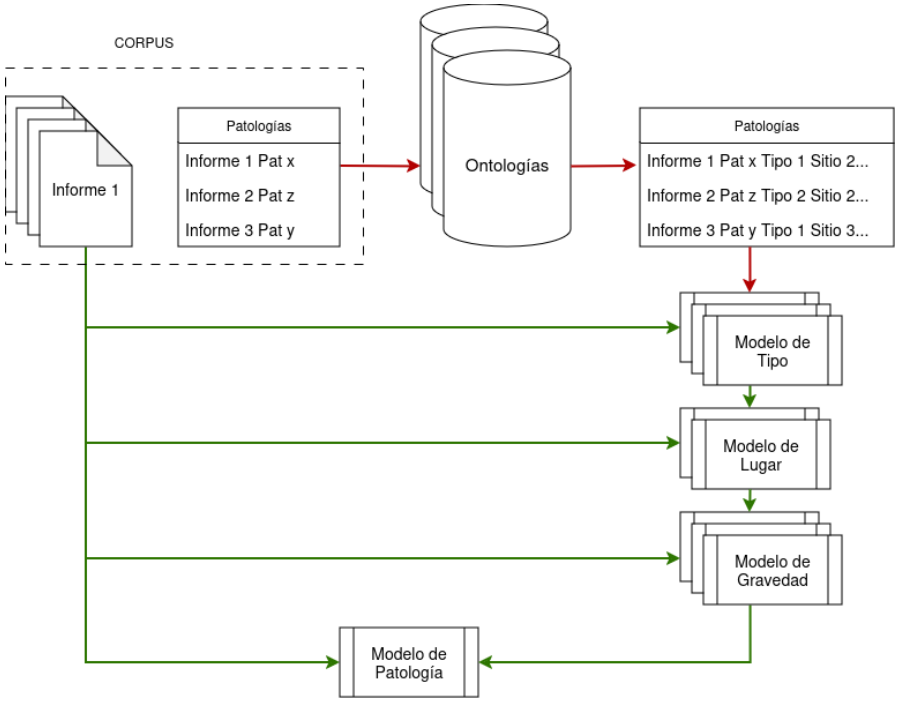


Figura 3: Arquitectura de nuestro método (en rojo las etapas solo de entrenamiento, en verde las de entrenamiento e inferencia).

de manera combinada UMLS, ICD-10 y MedDRA, extrajimos el tipo y la gravedad de la enfermedad. Para extraer estas características nos hemos inspirado en la clasificación de características predictivas dermatológicas propuesta en [68].

3.4.3. Modelos en cascada

Los modelos en cascada aprenden las relaciones mencionadas en la Tabla 1 (ampliada en la Tabla 8 del Anexo B).

Una vez que hemos extraído de los conceptos (las etiquetas traducidas) las relaciones comunes existentes en la ontología, entrenamos el bsc-bio-ehr-es para predecir cada una de ellas (ver Tabla 1). Cada relación predicha sirve para predecir la siguiente. En la Sección 4 mostramos en qué orden deben de ser aprendidas las relaciones. La Figura 3 ilustra nuestro método:

1. Extraemos del corpus el nombre de las patologías, las traducimos, las convertimos en conceptos y recuperamos las relaciones ligadas a esos conceptos dentro de las ontologías.
2. Por cada relación extraída, entrenamos un modelo. En modo entrenamiento, las relaciones son un oráculo. En modo inferencia, las relaciones son generadas por el modelo.
3. Cada modelo intermediario recibe como entrada los informes y una relación del modelo previo.
4. En cada etapa de la cascada, es decir cuando un modelo ha realizado su predicción, se descodifica cada una de ellas y se concatena con el informe inicial, y se vuelve a vectorizar el conjunto informe-relación predicha mediante el tokenizador del bsc-bio-ehr-es. En la ultima etapa, el vector de entrada contiene una representación del informe y de las tres relaciones.
5. Un modelo final aprende a predecir la patología a partir de los informes y de la salida del último modelo intermediario.

4. Experimentos y resultados

En esta sección describimos los resultados de nuestra arquitectura comparada con modelos de referencia, y evaluamos su rendimiento.

Tabla 1
Nomenclatura generada a partir de las características extraídas con *Pymedtermino* (muestra de las 5 enfermedades más frecuentes, versión completa con todas las enfermedades y frecuencias de aparición en la Tabla 8 del Anexo B)

Enfermedad	Tipo	Gravedad	Sitio
carcinoma de células basales	proceso neoplasico	importante	piel
psoriasis	proceso autoinmune	inofensivo	extremidades
nevus melanocítico	precancer	inofensivo	todo
acné	enfermedad	leve	todo
queratosis actínica	precancer	inofensivo	piel

4.1. Enriquecimiento con ontologías

Esta estrategia consiste en entrenar un modelo previo al de clasificación de informes: se trata de un modelo intermediario para aprender el tipo de enfermedad. Para eso, usamos UMLS [62] para recuperar los tipos de patología, SNOMED para el sitio anatómico y ICD10 para la gravedad[7]. Traducimos de manera automática el nombre de las enfermedades con Google Translate.

4.2. Modelos

Se han considerado diferentes enfoques para este modelado, tanto haciendo uso de técnicas basadas en transformadores, como modelos de aprendizaje supervisado de clasificación.

En cuanto a aquellos basados en transformadores, se han utilizado los siguientes:

- BETO [69]. Tal y como indican sus autores, se trata de un modelo BERT entrenado con un corpus en español, utilizando la técnica *Whole Word Masking*.
- *bsc-bio-ehr-es* [21]. Modelo generado por el BSC (*Barcelona Supercomuting Center*), utilizando de base un gran corpus de textos biomédicos en español.

Además, el segundo de estos modelos ha sido tratado a través del tuneado de hiperparámetros. Se especifican esos hiperparámetros y el material usado para entrenar los modelos en el Anexo A.3.

En lo que respecta a modelos clásicos de aprendizaje automático de clasificación, se ha recurrido a tres de los algoritmos más habituales: la regresión logística, las máquinas de vector soporte (SVM) y los *Random Forest*.

Se han propuesto dos enfoques:

- El modo **oráculo** (OR): el modelo[8] conoce el tipo, el sitio anatómico y la gravedad de la patología. El objetivo de este modo es demostrar la necesidad de información externa para realizar la tarea de clasificación.
- El modo **predictivo** (PR): el modelo final[9] debe predecir las tres características mencionadas, en el orden óptimo. Cada inferencia de una característica debe ayudar a la predicción de la siguiente. El objetivo de este modo es demostrar que nuestro modelo tiene una aplicación real y útil para la comunidad médica.

4.3. Resultados y evaluación

De cara a generar una comparativa de OR y PR, hemos seleccionado las siguientes métricas:

- Exactitud. Proporción de predicciones correctas sobre el total de predicciones realizadas.
- F1-score. Media armónica de la precisión y el *recall*. Para problemas multiclase, el F1-score puede calcularse de manera macro (media de f-score de cada clase) o micro (media entre falsos positivos, falsos negativos y reales positivos de todo el conjunto).
- Exactitud *top-k*. Proporción de casos en los cuales la clase verdadera está entre las k predicciones más probables del modelo. Esta métrica es útil cuando no solo importa la predicción más probable, sino también otras alternativas que el modelo considere razonables. Para todo este artículo, consideramos (k=2).
- F1-score *top-k*. Extensión del concepto de F1-score considerando las k clases más probables predichas por el modelo.

Inspirándonos en los trabajos de [70, 71] en aprendizaje automático y [72] en la tarea de POS-*tagging*,

[7] Se descartó MedDRA por no aportar más que las otras
[8] https://huggingface.co/fundacionctic/oracle-dermat
[9] https://huggingface.co/fundacionctic/predict-dermat

Tabla 2
Tabla con los resultados intermediarios para tipo (**t**), gravedad (**gr**) y sitio (**sit**) de cada enfermedad

Cat. info.	Prec.	Micro F1	Macro F1
t	0.57	0.56	0.38
gr	0.57	0.56	0.41
sit	0.68	0.67	0.59
t → **gr** → **sit**	0.70	0.68	0.58
gr → **t** → **sit**	0.62	0.61	0.51
sit → **gr** → **t**	**0.72**	**0.71**	**0.62**

consideramos que profesionales médicos usando nuestro modelo pueden ver más informativo tener no una sino dos predicciones (viendo la cantidad de clases posibles), al parecerse esto más al estilo de decisión natural humano, y dejando que sea el médico el que tenga el veredicto final de la enfermedad.

Estas métricas han sido obtenidas para todas las configuraciones contempladas en esta investigación. Las flechas corresponden a los diferentes modelos en cascada: cada característica predicha se convierte en característica conocida para el siguiente modelo.

4.3.1. Resultados modelos intermediarios

Exponemos en la Tabla 2 un resumen de los resultados obtenidos con el bsc-bio-ehr-es en PR sobre cada uno de los tres componentes del sistema de cascada: tipo, gravedad y sitio de la enfermedad. Los detalles de todos los resultados se pueden encontrar en el Anexo A.4.

La mejor combinación de categorías intermediarias en cascada parece ser primero predecir el sitio de la enfermedad, seguido por la gravedad y por fin el tipo. Observemos ahora si se refleja en la predicción final de la enfermedad.

4.3.2. Modelo final de predicción de enfermedad

En todas las combinaciones de experimentos, bien sea con modelos de aprendizaje automático clásicos o con transformadores ajustados, se trata de una tarea de clasificación supervisada multiclase monoetiqueta.

Observando la Tabla 3, se puede apreciar cómo los mejores resultados con respecto a las cuatro métricas se obtienen para el modelo basado en ontologías con todas las informaciones añadidas, existiendo una diferencia sustancial con el resto de opciones. En OR, cuando el modelo conoce las informaciones de las ontologías, los resultados superan el 0.84 de precisión absoluta y el 0.92 en precisión *top-k*. En PR, la ganancia también es significativa puesto que pasamos del 0.5 de precisión con el modelo "vanilla" a 0.66 con la mejor combinación de información. Es notable que la mejor combinación de características en cascada sea la de tipo seguido por sitio y gravedad, puesto que la gravedad solo tiene 4 variables, lo que nos hacía intuir que su aprendizaje sería más sencillo. Estos resultados confirman nuestras intuiciones: la necesidad de buscar informaciones externas para el entrenamiento del modelo y la necesidad de buscar en qué orden hay que aprender estas informaciones para optimizar el sistema final de clasificación, así como la eficacia del bsc-bio-ehr-es para nuestra tarea. En la Tabla 4 se presentan los mejores resultados en OR y PR para las 5 patologías más frecuentes.

4.4. Análisis de errores de los modelos

Tras la generación de estos modelos se ha procedido a realizar un análisis de errores de la mejor configuración PR y del mejor OR.

Por un lado, la poca precisión de los tres modelos clásicos de aprendizaje supervisado (regresión logística, SVM, *Random Forest*) tiene como posible causa la incapacidad de generalizar sobre etiquetas poco frecuentes. La principal limitación de los modelos clásicos de *machine learning* está en su falta de memoria de contexto: cuanto más largo sea un texto, más costoso es para la máquina recordar el contenido del principio. Las altas dimensiones de *embeddings* tampoco ayudan, puesto que la cantidad de variables a aprender es exponencial.

Examinando las categorías de enfermedad donde más erran los modelos, concluimos que las discrepancias son debidas a que las enfermedades confundidas comparten zonas del cuerpo afectadas similares, niveles de gravedad parecidos, y algunas descripciones de aspecto visual y síntomas compartidos. Por otro lado, a excepción de los cánceres, el modelo tiende a confundir enfermedades cuya apariencia física son las protuberancias. La confusión más frecuente es entre el carcinoma de células basales y de células escamosas, representando 844 de 2334 errores. Aunque pueden aparecer en cualquier parte del cuerpo, es más probable que estas enfermedades se desarrollen en áreas expuestas al sol, como la cabeza, el cuello y los brazos. La diferencia clave entre ellos es su gravedad, siendo el carcinoma escamoso más agresivo. Otra confusión frecuente del modelo es el acné con la queratosis seborreica con 325 errores. Estas enfermedades tienen similitudes en aspecto visual (protuberancias, enrojecimiento, picazón) y en lugares de aparición (cara y torso). Son dos condiciones dermatológicas que pueden parecerse en textos descriptivos, lo que

Tabla 3
Tabla con las métricas obtenidas para cada configuración considerada (AAC: aprendizaje automático clásico; TR: transformador; PR modo predictivo; OR modo oráculo; bsc bsc-bio-ehr-es)

Modelo	Precisión	Micro F1-sc.	Macro F1-sc	Prec. *top-k*	F1-sc. *top-k*
Regresión logística (AAC)	0.25	0.16	0.12	0.37	0.31
SVM (AAC)	0.263	0.13	0.14	0.39	0.29
Random Forest (AAC)	0.268	0.19	0.12	0.40	0.33
PR BETO (TR)	0.34	0.38	0.12	0.63	0.60
PR bsc (TR)	0.52	0.50	0.42	0.67	0.61
PR bsc (TR) + **gr** → **sit** → **t**	0.58	0.55	0.47	0.69	0.62
PR bsc (TR) + **sit** → **gr** → **t**	0.61	0.59	0.53	0.67	0.59
PR bsc (TR) + **t** → **gr** → **sit**	0.63	0.60	**0.54**	0.68	0.61
PR bsc (TR) + **t** → **sit** → **gr**	**0.66**	**0.61**	0.38	**0.71**	**0.65**
OR bsc (TR) + **t**	0.64	0.47	0.42	0.78	0.63
OR bsc (TR) + **gr**	0.55	0.53	0.36	0.69	0.60
OR bsc (TR) + **sit**	0.65	0.63	0.50	0.75	0.74
OR bsc (TR) + **sit** → **t**	0.77	0.76	0.66	0.87	0.87
OR bsc (TR) + **t** → **sit** → **gr**	**0.84**	**0.82**	**0.75**	**0.92**	**0.90**

Tabla 4
Tabla con las métricas (sin *top-k*) para las 5 patologías mas frecuentes

Enfermedad	F1 PR	F1 OR
acné	0.86	0.94
carc. cél. basales	0.70	0.92
psoriasis	0.81	0.87
nevus melanocítico	0.72	0.93
queratosis actínica	0.63	0.83

podría llevar a confusión en un modelo de lenguaje.

5. Discusión, conclusión y líneas de trabajo futuras

De manera general, los errores cometidos por el modelo pueden ser explicados, pese a la dificultad de la tarea (muchas enfermedades posibles, informes médicos que pueden ser tanto de primera cita o de seguimiento).

Debido a estos errores hemos decidido usar una métrica basada en el *top-k* en vez de la exactitud estricta. Si este sistema se utiliza en un entorno médico, entendemos que un profesional prefiere tener que elegir entre 2 posibles diagnósticos en vez de 25. Es también interesante mencionar que el mejor escenario en PR es cuando el modelo tiene que aprender tanto el sitio anatómico como el tipo y la gravedad de la enfermedad antes de adivinar esta última. Concluimos que un transformador necesita esa información externa para realizar la clasificación porque su modelo de lenguaje no es suficiente. En un primer momento, intuíamos que el modelo que aprende primero de la gravedad daría el mejor resultado, puesto que esta variable solo tiene 4 categorías (inofensiva, leve, moderada, extrema). Sin embargo, el mejor resultado lo da un modelo que aprende primero el tipo de enfermedad, seguido por el lugar y al final la gravedad de la enfermedad. Eso significa, y es lógico a nivel médico, que primero se detecta el tipo de patología, antes de intentar adivinar su gravedad.

Como conclusiones, nuestro trabajo presenta varias contribuciones importantes. Primero, proporcionamos un nuevo conjunto de datos de EHR en español de dermatología, anonimizado y etiquetado en patologías. Segundo, proponemos un método novedoso para predecir en EHR la patología que padece una persona. Tercero, este método es un sistema híbrido compuesto por varios modelos transformadores especializados puestos en cascada que usan como entrada además de los EHR, la salida del modelo precedente. El último modelo es el que predice la patología exacta. Los resultados muestran que los modelos en cascada son más eficaces que un solo modelo para distinguir enfermedades poco frecuentes. Eso significa que un solo modelo de extremo a extremo de transformadores no es suficiente para distinguir un concepto de patología entre varias decenas en un EHR, pero además que el uso de ontologías externas es necesario para que los transformadores aprendan conceptos intermediarios relacionados con la patología, de forma similar al

aprendizaje humano. Aun así, los resultados muestran un margen de mejora importante. Esto puede abordarse desde diferentes puntos de vista, como el etiquetado manual de los EHR en primera cita o cita de seguimiento, la búsqueda de nuevas relaciones ontológicas (como por ejemplo características de etiología más exhaustivas), o probar modelos diferentes para aprender cada característica.

Como posible línea de trabajo futuro proponemos automatizar el uso de ontologías mediante el RAG [73], como en BiomedRAG [74] así como el uso de NegEx [75] para evitar los falsos positivos, acompañado de nuestro sistema de modelos en cascada para eliminar el determinismo a la hora de encontrar relaciones de concepto para convertirse en un modelo de extremo a extremo.

Agradecimientos

Agradecimiento a la entidad SATEC por proporcionar el conjunto de datos, a la Fundación CTIC que puso a nuestra disposición todo el material y los recursos. Agradecemos a los revisores así como a Bea de Otto por la lectura y los comentarios pertinentes que mejoraron el articulo. Agradecemos y felicitamos al BSC por el modelo *bsc-bio-ehr-es*.

Referencias

[1] C. Hernández Salvador, Modelo de historia clínica electrónica para teleconsulta médica, 2004. URL: https://oa.upm.es/231/. doi:10.20868/UPM.thesis.231, unpublished.

[2] J. L. Hernández, Análisis y tipificación de errores lingüísticos para una propuesta de mejora de informes médicos en español, Procesamiento del Lenguaje Natural 70 (2023) 231–234. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6493.

[3] F. C. García, J. M. G. Hidalgo, M. de Buenaga Rodríguez, J. Mata, M. M. López, Acceso a la información bilingüe utilizando ontologías específicas del dominio biomédico, Procesamiento del Lenguaje Natural (2007) 107–117. URL: https://www.redalyc.org/articulo.oa?id=515751738012.

[4] S. Santiso, A. Casillas, A. Pérez, M. Oronoz, K. Gojenola, Document-level adverse drug reaction event extraction on electronic health records in Spanish, Procesamiento del Lenguaje Natural 56 (2016) 49–56. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5286.

[5] D. W. Otter, J. R. Medina, J. K. Kalita, A Survey of the Usages of Deep Learning for Natural Language Processing, IEEE Transactions on Neural Networks and Learning Systems 32 (2021) 604–624. doi:10.1109/TNNLS.2020.2979670.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[7] L. P. Schaub, M. M. Maestre, La inteligencia artificial como impulso del procesamiento del lenguaje natural: retos, fronteras y logros, Abaco: Revista de cultura y ciencias sociales 61 (2023) 66–81. URL: https://revista-abaco.es/la-inteligencia-artificial-como-impulso-del-procesamiento-del-lenguaje-natural-retos-fronteras-y-logros/.

[8] J. K. De Freitas, K. W. Johnson, E. Golden, G. N. Nadkarni, J. T. Dudley, E. P. Bottinger, B. S. Glicksberg, R. Miotto, Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records, Patterns (N. Y.) 2 (2021) 100337. doi:doi.org/10.1016/j.patter.2021.100337.

[9] T.-D. Le, R. Noumeir, J. Rambaud, G. Sans, P. Jouvet, Detecting of a Patient's Condition From Clinical Narratives Using Natural Language Representation, IEEE Open Journal of Engineering in Medicine and Biology 3 (2022) 142–149. doi:10.1109/OJEMB.2022.3209900.

[10] Y. Li, M. Mamouei, G. Salimi-Khorshidi, S. Rao, A. Hassaine, D. Canoy, T. Lukasiewicz, K. Rahimi, Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records, IEEE Journal of Biomedical and Health Informatics 27 (2023) 1106–1117. doi:10.1109/JBHI.2022.3224727.

[11] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical Natural Language Processing in languages other than English: opportunities and challenges, Journal of Biomedical Semantics 9 (2018) 12. URL: https://doi.org/10.1186/s13326-018-0179-8. doi:10.1186/s13326-018-0179-8.

[12] J. de la Rosa, E. G. Ponferrada, M. Romero,

P. Villegas, P. G. de Prado Salas, M. Grandury, BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[13] C. Aracena, N. Rodríguez, V. Rocco, J. Dunstan, Pre-trained language models in Spanish for health insurance coverage, in: T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, A. Rumshisky (Eds.), Proceedings of the 5th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 433–438. URL: https://aclanthology.org/2023.clinicalnlp-1.46. doi:10.18653/v1/2023.clinicalnlp-1.46.

[14] M. Rojas, J. Dunstan, F. Villena, Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing, in: T. Naumann, S. Bethard, K. Roberts, A. Rumshisky (Eds.), Proceedings of the 4th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Seattle, WA, 2022, pp. 87–92. URL: https://aclanthology.org/2022.clinicalnlp-1.9. doi:10.18653/v1/2022.clinicalnlp-1.9.

[15] J. Xu, X. Xi, J. Chen, V. S. Sheng, J. Ma, Z. Cui, A Survey of Deep Learning for Electronic Health Records, Applied Sciences 12 (2022). URL: https://www.mdpi.com/2076-3417/12/22/11709. doi:10.3390/app122211709.

[16] K. Araki, N. Matsumoto, K. Togo, N. Yonemoto, E. Ohki, L. Xu, Y. Hasegawa, D. Satoh, R. Takemoto, T. Miyazaki, Developing artificial intelligence models for extracting oncologic outcomes from japanese electronic health records, Adv. Ther. 40 (2023) 934–950. doi:doi.org/10.1007/s12325-022-02397-7.

[17] Z. Kraljevic, A. Shek, D. Bean, R. Bendayan, J. T. Teo, R. J. B. Dobson, MedGPT: Medical Concept Prediction from Clinical Narratives, CoRR abs/2107.03134 (2021). URL: https://arxiv.org/abs/2107.03134. arXiv:2107.03134.

[18] P. Haug, J. Ferraro, J. Holmén, X. Wu, K. Mynam, M. Ebert, N. Dean, J. Jones, An ontology-driven, diagnostic modeling system, Journal of the American Medical Informatics Association: JAMIA 20 (2013) 102–110. doi:10.1136/amiajnl-2012-001376.

[19] K. Buchan, M. Filannino, Özlem Uzuner, Automatic prediction of coronary artery disease from clinical narratives, Journal of Biomedical Informatics 72 (2017) 23–32. URL: https://www.sciencedirect.com/science/article/pii/S1532046417301466. doi:https://doi.org/10.1016/j.jbi.2017.06.019.

[20] M. Chizhikova, P. López-Úbeda, J. Collado-Montañez, T. Martín-Noguerol, M. C. Díaz-Galiano, A. Luna, L. A. Ureña-López, M. T. Martín-Valdivia, CARES: A Corpus for classification of Spanish Radiological reports, Computers in Biology and Medicine 154 (2023) 106581. URL: https://www.sciencedirect.com/science/article/pii/S001048252300046X. doi:https://doi.org/10.1016/j.compbiomed.2023.106581.

[21] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained Biomedical Language Models for Clinical NLP in Spanish, in: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: https://aclanthology.org/2022.bionlp-1.19. doi:10.18653/v1/2022.bionlp-1.19.

[22] L. M. Schriml, J. B. Munro, M. Schor, D. Olley, C. McCracken, V. Felix, J. A. Baron, R. Jackson, S. M. Bello, C. Bearer, otros, The human disease ontology 2022 update, Nucleic acids research 50 (2022) D1255–D1261. doi:10.1093/nar/gkab1063.

[23] A. T. McCray, J. L. Sponsler, B. Brylawski, A. C. Browne, The role of lexical knowledge in biomedical text understanding, in: Proceedings of the annual symposium on computer application in medical care, American Medical Informatics Association, 1987, pp. 103–107. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245098/.

[24] S. Doan, M. Conway, T. M. Phuong, L. Ohno-Machado, Natural language processing in biomedicine: a unified system architecture overview, Methods Mol Biol 1168 (2014) 275–294. doi:10.1007/978-1-4939-0847-9_16.

[25] P. Lambrix, Towards a semantic web for bioinformatics using ontology-based annotation, in: 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise (WETICE'05), IEEE, 2005, pp. 3–7. doi:10.1109/WETICE.2005.58.

[26] X. Jing, H. Min, Y. Gong, D. F. Sittig, P. Biondich, D. Robinson, T. Law, A. Wright, C. Nøhr, A. Faxvaag, L. Rennert, N. Hubig, R. Gimbel, A systematic review of ontology-based clinical decision support system rules: us-

age, management, and interoperability (2022). doi:10.1101/2022.05.11.22274984.

[27] G. J. Shannon, N. Rayapati, S. M. Corns, D. C. Wunsch, 2nd, Comparative study using inverse ontology cogency and alternatives for concept recognition in the annotated National Library of Medicine database, Neural Netw. 139 (2021) 86–104. doi:10.1016/j.neunet.2021.01.018.

[28] M. Romá-Ferri, OntoFIS: Tecnología ontológica en el dominio farmacoterapéutico, Ph.D. thesis, 2009. URL: https://www.researchgate.net/publication/265986206_OntoFIS_Tecnologia_ontologica_en_el_dominio_farmacoterapeutico.

[29] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780. URL: https://www.bioinf.jku.at/publications/older/2604.pdf.

[30] Y. Luo, Recurrent neural networks for classifying relations in clinical notes, Journal of biomedical informatics 72 (2017) 85–95. doi:doi.org/10.1016/j.jbi.2017.07.006.

[31] J. Giner Pérez de Lucía, Reconocimiento de entidades nombradas mediante técnicas de aprendizaje neuronal profundo en imágenes manuscritas, Ph.D. thesis, Universitat Politècnica de València, 2022. URL: http://hdl.handle.net/10251/185263.

[32] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, R. G. Mark, MIMIC-IV, a freely accessible electronic health record dataset, Scientific Data 10 (2023) 1. URL: https://doi.org/10.1038/s41597-022-01899-x. doi:10.1038/s41597-022-01899-x.

[33] A. Johnson, T. Pollard, R. Mark, MIMIC-III clinical database, 2023. doi:doi.org/10.13026/C2XW26.

[34] F. J. Moreno-Barea, H. Mesa, N. Ribelles, E. Alba, J. M. Jerez, Clinical Text Classification in Cancer Real-World Data in Spanish, in: I. Rojas, O. Valenzuela, F. Rojas Ruiz, L. J. Herrera, F. Ortuño (Eds.), Bioinformatics and Biomedical Engineering, Springer Nature Switzerland, Cham, 2023, pp. 482–496. doi:doi.org/10.1007/978-3-031-34953-9_38.

[35] A. Intxaurrondo, SPACCC, 2019. URL: https://doi.org/10.5281/zenodo.2560316. doi:10.5281/zenodo.2560316.

[36] A. Gonzalez-Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, M. Krallinger, PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track, in: K. Jin-Dong, N. Claire, B. Robert, D. Louise (Eds.), Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1–10. URL: https://aclanthology.org/D19-5701. doi:10.18653/v1/D19-5701.

[37] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2020. URL: https://ceur-ws.org/Vol-2696/paper_263.pdf.

[38] T. A. Koleck, C. Dreisbach, P. E. Bourne, S. Bakken, Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review, Journal of the American Medical Informatics Association 26 (2019) 364–379. doi:10.1093/jamia/ocy173.

[39] A.-J. Aberkane, G. Poels, S. V. Broucke, Exploring Automated GDPR-Compliance in Requirements Engineering: A Systematic Mapping Study, IEEE Access 9 (2021) 66542–66559. doi:10.1109/ACCESS.2021.3076921.

[40] A. Iglesias, E. Castro, R. Pérez, L. Castaño, P. Martínez, J. M. Gómez-Pérez, S. Kohler, R. Melero, Mostas: Un etiquetador morfosemántico, anonimizador y corrector de historiales clínicos, Procesamiento del lenguaje Natural 41 (2008) 299–300. URL: http://hdl.handle.net/10045/8615.

[41] T. Lordick, A. Hoch, B. Fransen, Anonymization of Electronic Health Care Records: The EHR Anonymizer, Springer International Publishing, Cham, 2022, pp. 485–499. URL: https://doi.org/10.1007/978-3-031-08411-9_18.

[42] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, brat: a Web-based Tool for NLP-Assisted Text Annotation, in: F. Segond (Ed.), Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Avignon, France, 2012, pp. 102–107. URL: https://aclanthology.org/E12-2021.

[43] S. Lima Lopez, N. Perez, L. García-Sardiña, M. Cuadros, HitzalMed: Anonymisation of clinical text in Spanish, in: Proceedings of the Twelfth Language Resources and Eval-

uation Conference, European Language Resources Association, Marseille, France, 2020, pp. 7038–7043. URL: https://aclanthology.org/2020.lrec-1.870.

[44] G. Francopoulo, L.-P. Schaub, Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP, in: workshop on Legal and Ethical Issues (Legal2020), LREC2020, ELRA, Marseille, France, 2020, pp. 9–14. URL: https://hal.science/hal-02939437.

[45] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodriguez, J. L. Martin, M. Villegas, M. Krallinger, Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results, in: IberLEF@ SEPLN, 2019, pp. 618–638. URL: https://ceur-ws.org/Vol-2421/MEDDOCAN_overview.pdf.

[46] G. S. Krishnan, S. Kamath S, Ontology-driven text feature modeling for disease prediction using unstructured radiological notes, Computación y Sistemas 23 (2019) 915–922. doi:doi.org/10.13053/cys-23-3-3238.

[47] O. B. Shoham, N. Rappoport, CPLLM: Clinical Prediction with Large Language Models, arXiv preprint arXiv:2309.11295 (2023). arXiv:2309.11295.

[48] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[49] X. Peng, G. Long, T. Shen, S. Wang, J. Jiang, Sequential Diagnosis Prediction with Transformer and Ontological Representation, in: 2021 IEEE International Conference on Data Mining (ICDM), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 489–498. URL: https://doi.ieeecomputersociety.org/10.1109/ICDM51629.2021.00060. doi:10.1109/ICDM51629.2021.00060.

[50] R. Saripalle, C. Runyan, M. Russell, Using HL7 FHIR to achieve interoperability in patient health record, J. Biomed. Inform. 94 (2019) 103188. doi:10.1016/j.jbi.2019.103188.

[51] P. Törnberg, How to use llms for text analysis, 2023. URL: https://arxiv.org/abs/2307.13106. arXiv:2307.13106.

[52] B. Kim, S. Cha, S. Park, J. Lee, S. Lee, S. Kang, J. So, K. Kim, J. Jung, J. Lee, S. Lee, Y. Paik, H. Kim, J. Kim, W. Lee, Y. Ro, Y. Cho, J. Kim, J. Song, J. Yu, S. Lee, J. Cho, K. Sohn, The breakthrough memory solutions for improved performance on llm inference, IEEE Micro 44 (2024) 40–48. doi:10.1109/MM.2024.3375352.

[53] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[54] M. A. H. Wadud, M. M. Kabir, M. F. Mridha, M. A. Ali, M. A. Hamid, M. M. Monowar, How can we manage offensive text in social media-a text classification approach using LSTM-BOOST, International Journal of Information Management Data Insights 2 (2022) 100095. doi:https://doi.org/10.1016/j.jjimei.2022.100095.

[55] G. Liu, M. Boyd, M. Yu, S. Z. Halim, N. Quddus, Identifying causality and contributory factors of pipeline incidents by employing natural language processing and text mining techniques, Process safety and environmental protection 152 (2021) 37–46. doi:doi.org/10.1016/j.psep.2021.05.036.

[56] M. Dinarelli, S. Rosset, Models Cascade for Tree-Structured Named Entity Detection, in: H. Wang, D. Yarowsky (Eds.), Proceedings of 5th International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing, Chiang Mai, Thailand, 2011, pp. 1269–1278. URL: https://aclanthology.org/I11-1142.

[57] S. Ghannay, A. Caubrière, S. Mdhaffar, G. Laperrière, B. Jabaian, Y. Estève, Where Are We in Semantic Concept Extraction for Spoken Language Understanding?, in: Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2021, p. 202–213. URL: https://doi.org/10.1007/978-3-030-87802-3_19. doi:10.1007/978-3-030-87802-3_19.

[58] S. Ghidalia, O. L. Narsis, A. Bertaux, C. Nicolle, Combining Machine Learning and Ontology: A Systematic Literature Review, 2024. arXiv:2401.07744.

[59] T. V. Vila, Ontoloxías e tradución biomédica: creación dunha base de coñecemento termi-

nolóxico sobre os erros innatos do metabolismo en francés e español, Ph.D. thesis, Universidade de Vigo, 2015. URL: https://infoling.org/informacion/T182.html.
[60] A. Villaplana, R. Martínez, S. Montalvo, Improving medical entity recognition in spanish by means of biomedical language models, Electronics 12 (2023). URL: https://www.mdpi.com/2079-9292/12/23/4872. doi:10.3390/electronics12234872.
[61] F. Stahlberg, Neural machine translation: A review, Journal of Artificial Intelligence Research 69 (2020) 343–418. doi:doi.org/10.1613/jair.1.12007.
[62] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic acids research 32 (2004) 267–270. doi:10.1093/nar/gkh061.
[63] K. A. Spackman, K. E. Campbell, R. A. Côté, SNOMED RT: a reference terminology for health care, in: Proceedings of the AMIA annual fall symposium, American Medical Informatics Association, 1997, p. 640. URL: https://pubmed.ncbi.nlm.nih.gov/9357704/.
[64] E. G. Brown, L. Wood, S. Wood, The medical dictionary for regulatory activities (MedDRA), Drug safety 20 (1999) 109–117. URL: https://bioportal.bioontology.org/ontologies/MEDDRA.
[65] L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, V. Felix, L. Jeng, C. Bearer, R. Lichenstein, et al., Human Disease Ontology 2018 update: classification, content and workflow expansion, Nucleic acids research 47 (2019) D955–D962. doi:10.1093/nar/gky1032.
[66] J.-B. Lamy, A. Venot, C. Duclos, PyMedTermino: an open-source generic API for advanced terminology services, Stud. Health Technol. Inform. 210 (2015) 924–928. URL: https://pubmed.ncbi.nlm.nih.gov/25991291/.
[67] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M. P. Richardson, R. Stewart, A. D. Shah, W. K. Wong, Z. Ibrahim, J. T. Teo, R. J. B. Dobson, Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit, Artif. Intell. Med. 117 (2021) 102083. doi:10.1016/j.artmed.2021.102083.
[68] H. M. Fisher, R. Hoehndorf, B. S. Bazelato, S. S. Dadras, L. E. King, Jr, G. V. Gkoutos, J. P. Sundberg, P. N. Schofield, DermO; an ontology for the description of dermatologic disease, J. Biomed. Semantics 7 (2016) 38. doi:doi.org/10.1186/s13326-016-0085-x.
[69] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, 2023. URL: https://arxiv.org/abs/2308.02976. arXiv:2308.02976.
[70] M. Lapin, M. Hein, B. Schiele, Top-k Multiclass SVM, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 28, Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/0336dcbab05b9d5ad24f4333c7658a0e-Paper.pdf.
[71] A. Sawada, E. Kaneko, K. Sagi, Trade-offs in Top-k Classification Accuracies on Losses for Deep Learning, CoRR abs/2007.15359 (2020). URL: https://arxiv.org/abs/2007.15359. arXiv:2007.15359.
[72] A. Ratnaparkhi, A Linear Observed Time Statistical Parser Based on Maximum Entropy Models, CoRR cmp-lg/9706014 (1997). URL: http://arxiv.org/abs/cmp-lg/9706014.
[73] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-Augmented Generation for Large Language Models: A Survey, 2024. arXiv:2312.10997.
[74] M. Li, H. Kilicoglu, H. Xu, R. Zhang, BiomedRAG: A Retrieval Augmented Large Language Model for Biomedicine, 2024. arXiv:2405.00465.
[75] G. Argüello-González, J. Aquino-Esperanza, D. Salvador, R. Bretón-Romero, C. Del Río-Bermudez, J. Tello, S. Menke, Negation recognition in clinical natural language processing using a combination of the NegEx algorithm and a convolutional neural network, BMC Medical Informatics and Decision Making 23 (2023) 216. URL: https://doi.org/10.1186/s12911-023-02301-5. doi:10.1186/s12911-023-02301-5.

A. Detalles experimentales

En este apartado, explicamos y describimos en detalles los resultados obtenidos y el procedimiento para reproducir el experimento.

A.1. Algoritmos de extracción de información de las ontologías

En el Algoritmo 1, explicamos el proceso de extracción de información de las ontologías para agregar

relaciones entre cada informe y la enfermedad asociada a éste.

1. Por cada enfermedad, la traducimos en inglés mediante la API de Google Translate.
2. Luego, por cada ontología en nuestra lista tenemos diferentes relaciones, por lo que es importante usarlas todas. A partir del nombre de la enfermedad traducida en inglés tenemos tres opciones:
   a) Si es SNOMED, extraemos el sitio del cuerpo y de la piel afectados.
   b) Si es UMLS, extraemos el por la enfermedad.
   c) Si es ICD10, extraemos la gravedad de la enfermedad mediante la interpretación de la información sobre si es una afección mayor o menor, y si conlleva morbilidad.

Algoritmo 1 Extracción de relaciones

```
load pymedtermino library
for diseases = 1, 2, ..., M do
    translate diseases with Google API
    for ontology = 1, 2, ..., N do
        if SNOMED then
            diseases.getType()
        else if UMLS then
            diseases.getLocation()
        else
            diseases.getAffection()
            if has(minor) then
                SetTo(light)
            else if has(major) then
                SetTo(important)
            else if has(morbidity) then
                SetTo(deadly)
            else
                SetTo(inoffensive)
            end if
        end if
    end for
end for
```

A.2. Algoritmo de los modelos en cascada

En este apartado, explicamos el funcionamiento de los modelos en cascada (Algoritmo 2):

1. Realizamos la combinación ordenada de las tres relaciones que tenemos, con paquetes de tamaño 1 o 3.
2. Por cada uno de esos paquetes, cargamos el dataset y lo tokenizamos.
3. Por cada elemento de cada paquete, entrenamos un modelo en modo supervisado para aprender ese elemento.
4. Una vez que cada modelo está entrenado, su predicción del elemento se concatena con la entrada original, y sirve de nueva entrada para entrenar el modelo a predecir el próximo elemento del paquete.
5. Entrenamos otro modelo con el próximo elemento del paquete y la nueva entrada.
6. Cuando todos los elementos han sido aprendido por los diferentes modelos en cascada, se entrena un último modelo con la entrada original enriquecida por la última salida (con todos los elementos predichos) para predecir la enfermedad.
7. Una vez que todos los paquetes están procesados, se comparan los modelos y se elige el mejor.

Algoritmo 2 Modelos en cascada

```
combi ← []
for iter1 = 1, 2, ... N do
    for iter2 = 1, 2, ... M do
        combi ← iter1, iter2
    end for
end for
for relations in combi do
    input ← string(medicalRecords)
    input.tokenize()
    for relation in relations do
        output ← model.train(relation)
        input ← input + output
    end for
    output ← Model.train(diseases)
end for
computeAccuracy()
getBestModel()
```

A.3. Configuración del entrenamiento

Se han identificado los siguientes parámetros del modelo como susceptibles de tratamiento:

- Tamaño del lote (*batch size*). Número de muestras de entrenamiento que se procesarán a través de la red en una sola iteración antes de que se actualicen los pesos del modelo.
- Tasa de aprendizaje (*learning rate*). Controla cuánto se ajustan los pesos del modelo en respuesta al error calculado en cada iteración del entrenamiento.

Tabla 5
Tabla con los resultados intermediarios para tipo, gravedad y sitio de cada enfermedad

Categoría de información	Precisión	Micro F1	Macro F1
t	0.57	0.56	0.38
gr	0.57	0.56	0.41
sit	0.68	0.67	0.59
gr → **t** → **sit**	0.62	0.61	0.51
gr → **sit** → **t**	0.66	0.66	0.58
t → **gr** → **sit**	0.70	0.68	0.58
t → **sit** → **gr**	0.70	0.68	0.57
sit → **t** → **gr**	0.69	0.67	0.58
sit → **gr** → **t**	**0.72**	**0.71**	**0.62**

- Número de épocas (*epochs*). Número de veces que el algoritmo de aprendizaje trabajará a través de todo el conjunto de datos de entrenamiento.

Para cada uno de estos hiperparámetros se han contemplado diferentes valores posibles y, tras un proceso de *grid search*, se ha determinado que los valores que mejores resultados proporcionan son *batch size* 64, *learning rate* 0.001 y *epochs* 10.

Los experimentos se llevaron a cabo mediante una NVIDIA GeForce RTX 2080 Ti 12GB y una NVIDIA RTX A6000 50GB. Usamos la bibioteca Pytorch 2.2.1 con CUDA 12.1. Cada entrenamiento duraba entre 3 y 5 minutos dados los pocos datos en el corpus. En total, contando el *grid search*, los experimentos necesitaron algo mas que 96 horas para obtener los resultados presentados en el artículo.

A.4. Resultados detallados de los modelos intermediarios

En esta sección, presentamos los resultados de varias combinaciones en la Tabla 5 que se llevaron a cabo para predecir la patología de la mejor manera posible. Tratándose de variaciones sin repetición, el numero total de combinaciones posibles es de la forma:

$$\sum_{k=1}^{n} V_n^k = \sum_{k=1}^{n} \frac{n!}{(n-k)!}.$$

Para $n = 3$, el número total de combinaciones (variaciones) de tamaños 1 a 3 sin repetición es:

$$\frac{3!}{(3-1)!} + \frac{3!}{(3-2)!} + \frac{3!}{(3-3)!} = 15.$$

A.5. Comparativa entre los resultados de los diferentes enfoques

Se ha llevado a cabo una comparativa del método OR completo con la información de las ontologías (modelo A) y el método basado en el modelo bsc-bio-ehr-es *vanilla* sin ninguna información añadida (modelo B) como referencia de los desarrollos presentados en esta investigación, haciendo especial hincapié en las mejoras que ha proporcionado el utilizar la visión *top-k*, y viendo en qué situaciones dichos cambios ha supuesto una mejora en los resultados. Esta comparativa es presentada en la Tabla 7.

Para el modelo A, las métricas de precisión y F1-score generadas sin el uso de *top-k* se corresponden con 0.82 y 0.73, las cuales se han visto mejoradas con el enfoque *top-k* pasando a valores 0.86 y 0.85. Del mismo modo, se ha observado como el modelo B pasa de valores 0.52 y 0.42 para dichas métricas a 0.67 y 0.61 con la inclusión del *top-k*. Si bien en el modelo A se aprecian mejoras evidentes, es en el caso B donde existe una mejora más sustancial.

Ilustrativamente, se presenta tabla 6 un análisis de en qué situaciones ambos modelos han mejorado por el uso del enfoque *top-k*. Con respecto al modelo A, los principales errores vienen de predecir "dermatitis atópica" donde la realidad se corresponde a "eccema" (17.65%) o "psoriasis" (9.8%). En el caso de "eccema", tiene sentido que se dé dicha confusión, dado que son dos términos que se usan a menudo indistintamente para referirse a la misma afección cutánea. La "dermatitis atópica" es el término más específico, mientras que el "eccema" es un término más general que abarca varios tipos de inflamación de la piel. En lo que respecta a la confusión con "psoriasis", pueden darse razones para ello, como la apariencia similar en ciertas situaciones (enrojecimiento, inflamación y picazón común, pudiendo

Tabla 6
Tabla con las métricas (sin *top-k*) para las 25 patologías mas frecuentes

Enfermedad	F1 *vanilla*	F1 PR	F1 OR
acné	0.43	0.86	0.94
carc. cél. basales	0.60	0.70	0.92
psoriasis	0.67	0.81	0.87
nevus melanocítico	0.52	0.72	0.93
queratosis actínica	0.49	0.63	0.83
carcinoma de células escamosas	0.43	0.52	0.86
eccema	0.45	0.59	0.62
rosácea	0.00	0.37	0.55
lentigo solar	0.54	0.65	0.97
liquen escleroatrófico	0.73	0.69	0.82
fibroma	0.57	0.50	0.87
llaga	0.64	0.69	0.78
melanoma	0.43	0.66	0.86
alopecia areata	0.74	0.80	0.80
dermatitis atópica	0.45	0.47	0.74
carcinoma espinocelular	0.40	0.60	0.91
queratosis seborreica	0.44	0.67	0.97
sin diagnostico	0.31	0.37	0.77
acné juvenil	0.04	0.00	0.63
verruga	0.57	0.82	0.98
urticaria crónica	0.22	0.71	0.87
hemangioma	0.86	0.77	0.97
nevus melanocítico atípico	0.00	0.00	0.91
dermatofibroma	0.53	0.53	0.95
ulcera	0.00	0.55	0.70

llegar a escamación en ambos casos), así como en cuanto a la localización corporal (ambas pueden aparecer en áreas como los codos, rodillas o cuero cabelludo).

En el caso del modelo B, la principal mejora se presenta en los casos en los que la enfermedad es "nevus melanocítico adquirido" y se predice "nevus melanocítico" (11.89%). Los nevus melanocíticos adquiridos y congénitos comparten varias características, incluyendo su apariencia, histología, genética y desarrollo. Sin embargo, se diferencian en el momento de su aparición, siendo los congénitos presentes al nacer o en las primeras semanas de vida, mientras que los adquiridos aparecen a lo largo de la vida. Esto ilustra cómo no solo el enfoque *top-k* mejora la precisión de los modelos, sino que algunos de los errores más comunes son entendibles dado el significado de los términos confundidos.

Cabe destacar también que el enfoque *top-k* permite en varios casos mejorar la predicción que incluye situaciones "sin diagnóstico", proporcionando dicha opción en casos donde el enfoque base no lo contempla. También es reseñable cómo en ambos métodos aparece un alto número de casos de confusión entre diagnósticos ligados a carcinoma ("carcinoma de células basales", "carcinoma de células escamosas") y queratosis ("queratosis actínica", "queratosis seborreica"). En este caso no son patologías semejantes, por lo que el incluir *top-k* proporcionaría al facultativo una alternativa sobre la que valorar cuál es la opción adecuada con su conocimiento experto.

A.6. Resultados del modelo con diferentes umbrales de frecuencia de cada patología

La Tabla 7 muestra los resultados tanto del método A como del B, y demuestra que el umbral de 61 ejemplos mínimo por categoría es el óptimo para guardar un máximo de categorías sin perder la eficacia de los modelos de clasificación.

B. Información complementaria

En esta sección se presenta información complementaria que sirve de apoyo a la comprensión de los desarrollos de este trabajo. En particular, se presenta la siguiente información:

Tabla 7
Métricas obtenidas con diferentes umbrales de ejemplos por patología A es el método con la información de las ontologías y B el método con el bsc-bio-ehr-es *vanilla*

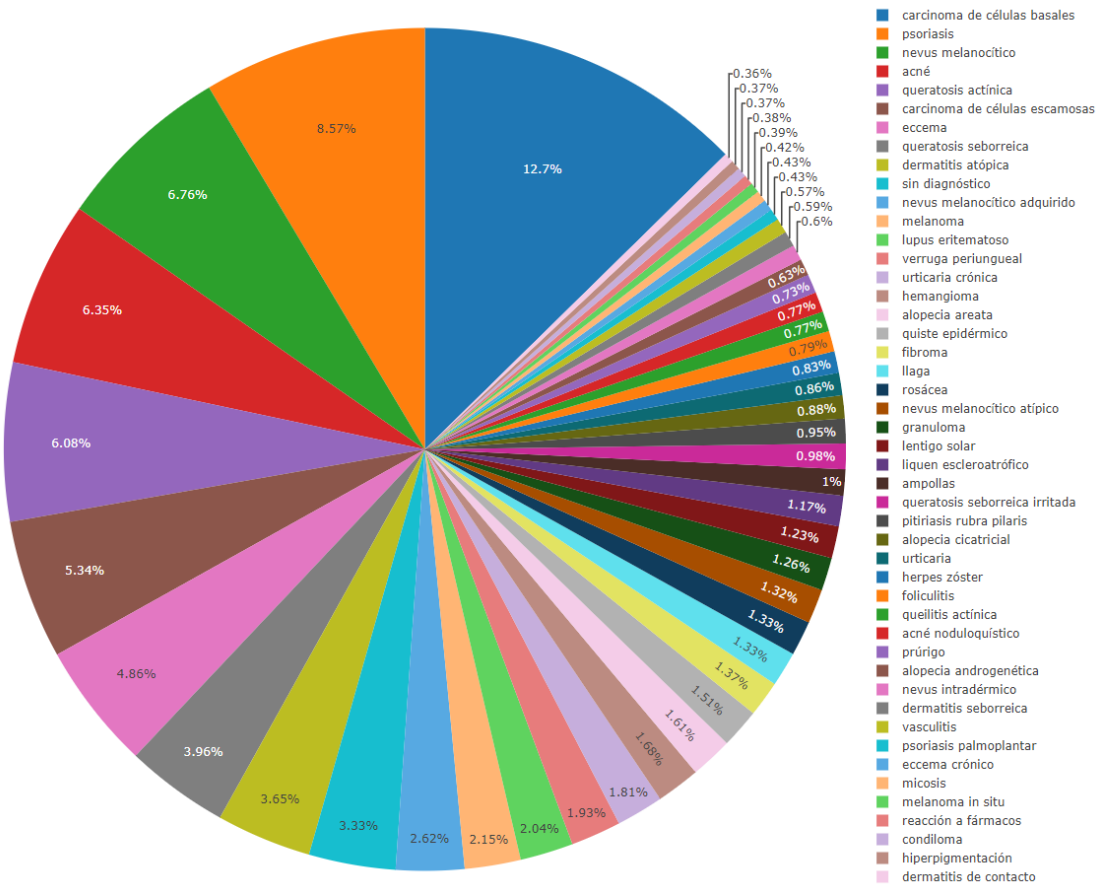
Modelo	Umbral	Num. clases	Prec.	Micro F1-sc.	Macro F1-sc	Prec. *top-k*	F1-sc. *top-k*
B	2	173	0.39	0.34	0.08	0.54	0.13
B	10	76	0.41	0.41	0.12	0.58	0.27
B	25	44	0.46	0.43	0.24	0.59	0.46
B	50	27	0.48	0.46	0.38	0.66	0.63
B (nuestro)	61	25	0.52	0.50	0.42	0.67	0.61
B	75	20	0.51	0.49	0.44	0.69	0.71
B	100	15	0.55	0.54	0.51	0.71	0.77
A	2	173	0.68	0.62	0.14	0.80	0.28
A	10	76	0.72	0.66	0.25	0.86	0.52
A	25	44	0.77	0.73	0.48	0.90	0.81
A	50	27	0.83	0.80	0.72	0.91	0.86
A (nuestro)	61	25	0.84	0.82	0.75	0.92	0.90
A	75	20	**0.87**	0.85	0.80	0.94	0.91
A	100	15	**0.87**	**0.87**	**0.85**	**0.96**	**0.92**

- Tabla 8. Versión ampliada de la Tabla 1, con la lista completa de enfermedades.
- Figura 4. Representación gráfica de la distribución de las enfermedades en el conjunto de datos.
- Figuras 5 y 6. Matrices de confusión de los métodos A y B presentados en el Anexo B, respectivamente.

Tabla 8
Nomenclatura generada a partir de las características extraídas con *Pymedtermino* con las enfermedades mas frecuentes y la cantidad de aparición

Enfermedad	Tipo	Gravedad	Sitio	Frecuencia
carcinoma de células basales	proceso neoplasico	importante	piel	1124
psoriasis	proceso autoinmune	inofensivo	extremidades	761
nevus melanocítico	precancer	inofensivo	todo	600
acné	enfermedad	leve	todo	564
queratosis actínica	precancer	inofensivo	piel	540
carcinoma de células escamosas	proceso neoplasico	extrema	piel	474
eccema	enfermedad	inofensivo	mano	432
queratosis seborreica	tumor benigno	inofensivo	piel	352
dermatitis atópica	enfermedad	inofensivo	articulaciones	324
sin diagnóstico	sin enfermedad	inofensivo	todo	296
nevus melanocítico adquirido	proceso neoplasico	inofensivo	extremidades	233
melanoma	proceso neoplasico	extrema	todo	191
lupus eritematoso	proceso autoinmune	extrema	tejido conectivo	181
verruga periungueal	infeccion	inofensivo	mano	171
urticaria crónica	sintoma	inofensivo	todo	161
hemangioma	tumor benigno	leve	todo	149
alopecia areata	proceso autoinmune	inofensivo	cabeza	143
quiste epidérmico	anormalidad	leve	cara	134
fibroma	tumor benigno	leve	pierna	122
llaga	sintoma	inofensivo	boca	118
rosácea	enfermedad	inofensivo	cara	118
nevus melanocítico atípico	proceso neoplasico	importante	torso	117
granuloma	infeccion	extrema	genitales	112
lentigo pielar	sindrome	leve	todo	109
liquen escleroatrófico	proceso autoinmune	leve	genitales	104
ampollas	sintoma	inofensivo	mano	89
queratosis seborreica irritada	enfermedad	inofensivo	todo	87
pitiriasis rubra pilaris	proceso autoinmune	leve	articulaciones	84
alopecia cicatricial	enfermedad	inofensivo	cabeza	78
urticaria	funcion patologica	leve	todo	76
herpes zóster	infeccion	importante	torso	74
foliculitis	enfermedad	inofensivo	cabeza	70
queilitis actínica	precancer	leve	boca	68
acné noduloquístico	infeccion	leve	cara	68
prúrigo	sintoma	inofensivo	cabeza	65
alopecia androgenética	enfermedad	inofensivo	cabeza	56
nevus intradérmico	precancer	inofensivo	piel	53
dermatitis seborreica	proceso autoinmune	inofensivo	cara	52
vasculitis	proceso autoinmune	extrema	articulaciones	51
psoriasis palmoplantar	enfermedad	leve	extremidades	38
eccema crónico	enfermedad	inofensivo	mano	38
micosis	infeccion	importante	todo	37
melanoma in situ	proceso neoplasico	inofensivo	todo	35
reacción a fármacos	envenenamiento	inofensivo	todo	34
condiloma	infeccion	leve	genitales	33
hiperpigmentación	anormalidad	inofensivo	todo	33
dermatitis de contacto	enfermedad	inofensivo	mano	32

Figura 4: Distribución de las enfermedades en el conjunto de datos generado.



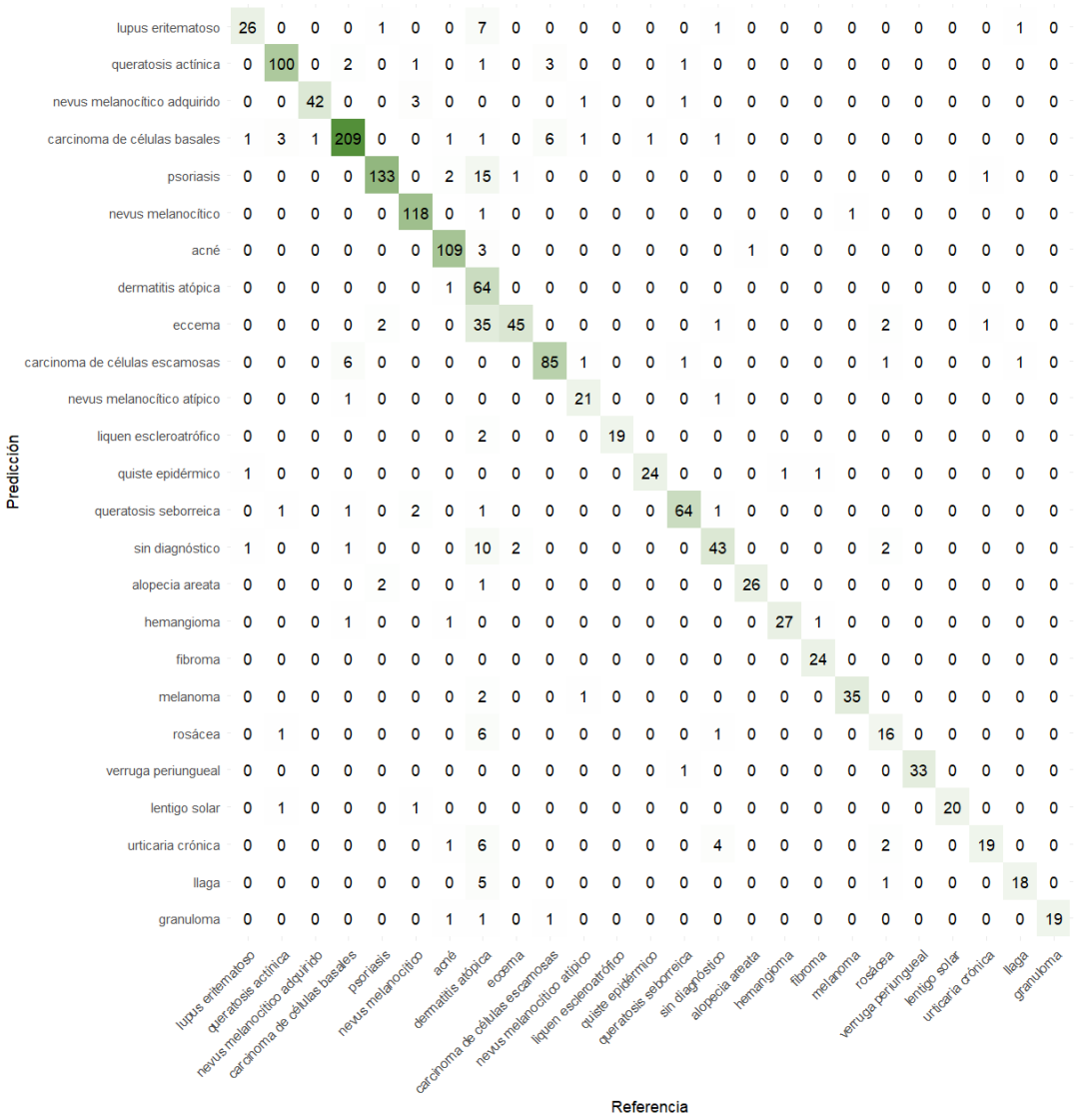


Figura 5: Matriz de confusión para modelo A.

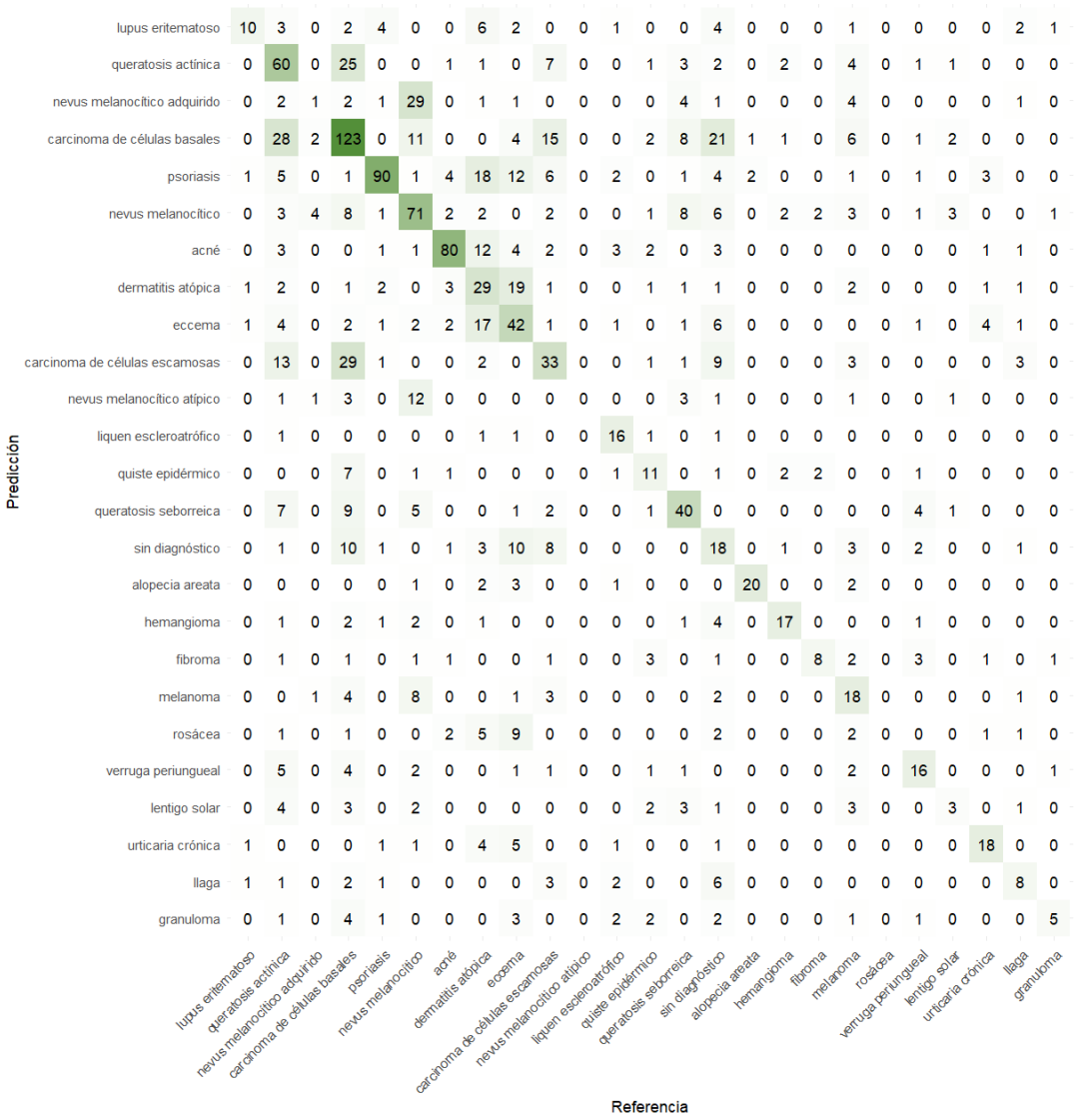


Figura 6: Matriz de confusión para modelo B.