Reducción del diagnóstico tardío de la infección por VIH aplicando técnicas de Procesamiento del Lenguaje Natural

Trabajo de Fin de Máster

Rodrigo Morales Sánchez

Trabajo de investigación para el
Máster Universitario en Tecnologías del Lenguaje

Universidad Nacional de Educación a Distancia

Dirigido por

Dra. Raquel Martínez Unanue

Dra. Soto Montalvo Herranz

Septiembre 2023

A Raquel y a Soto por su ayuda, su paciencia,
y por confiar en mí para este proyecto.

A mi familia por su infinito amor y por dármelo todo.

Y a Laura por sostenerme y no dejarme caer.

Resumen

La mejora en el diagnóstico de la infección por VIH es un tema vital para avanzar en el control de la epidemia. En España se estima que un 10% de pacientes infectados desconocen su estatus serológico. Además, alrededor del 50% de los nuevos diagnósticos cada año son detectados de manera tardía.

Se han realizado varios estudios sobre estrategias de cribado y se reconoce como la mejor alternativa en relación coste-eficiencia, aquella basada en indicadores. Sin embargo, mientras el cribado en algunos ámbitos como las urgencias está bastante implantado, en otros como la hospitalización no ha recibido suficiente atención. Además, existen barreras por parte de los profesionales para la realización de serología de VIH como el tiempo o la falta de percepción de riesgo.

En este trabajo se presentan dos propuestas para la ayuda en el diagnóstico basado en el uso de la información de la historia clínica electrónica de los pacientes, utilizando técnicas de aprendizaje automático y procesamiento del lenguaje natural. La primera consiste en una propuesta no supervisada basada en conocimiento experto a partir de una serie de indicadores. Y la segunda, una propuesta supervisada de clasificación binaria.

No es mucha la literatura existente sobre la predicción de VIH utilizando técnicas de Procesamiento del Lenguaje Natural, por ello, este trabajo se postula como un enfoque novedoso en el que ambas propuestas consiguen unos resultados prometedores. La propuesta supervisada destaca por sus mejores resultados, mientras que la propuesta no supervisada tiene un interesante potencial por su mejor explicabilidad.

Abstract

Improving the diagnosis of HIV infection is a crucial issue in order to make progress in controlling the epidemic. In Spain, it is estimated that 10% of infected patients do not know their serological status. In addition, around 50% of new diagnoses each year are detected late.

Several studies have been carried out on screening strategies and indicator-based screening is recognized as the best option in terms of cost-effectiveness. However, while screening in some settings such as the emergency department is well established, in others such as hospitalization it has not received sufficient attention. In addition, there are barriers for professionals to perform HIV serology, such as time or lack of risk perception.

This paper presents two proposals for diagnostic assistance based on the use of information from patients' electronic medical records, using machine learning and natural language processing techniques. The first consists of an unsupervised proposal based on expert knowledge from a series of indicators. The second is a supervised binary classification approach.

The existing literature on HIV prediction using Natural Language Processing techniques is not extensive, therefore, this work is postulated as a novel approach in which both proposals achieve promising results. The supervised approach stands out for its better results, while the unsupervised approach has an interesting potential for its better explainability.

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

1.1 Motivación

La infección por VIH es un problema de salud pública y de salud individual muy relevante. Actualmente existe la posibilidad de disminuir y controlar la transmisión de esta enfermedad si todas las personas infectadas estuvieran diagnosticadas (supieran que tienen el virus), se trataran con fármacos antirretrovirales y consiguieran tener el virus no detectable en la sangre. Esta situación de virus indetectable en sangre hace que las personas infectadas no transmitan el virus y de esta forma se interrumpa la cadena de transmisión. A pesar de ello, hoy día todavía hay un porcentaje de personas con la infección VIH que desconocen su estado.

El objetivo de ONUSIDA (Programa Conjunto de las Naciones Unidas sobre el VIH/SIDA)[1] para 2030 es conseguir la relación 95-95-95 (95% de personas con VIH diagnosticadas, 95% en tratamiento antirretroviral y 95% con carga viral suprimida). Se trata de una estrategia clave para frenar la epidemia y su eventual erradicación. La mejora en el diagnóstico de la infección por VIH es un tema vital para avanzar en el control de la epidemia y en la consecución del objetivo de la ONU. El hecho de tener pacientes con infección no diagnosticados conduce a una mayor transmisión de la infección, diagnósticos tardíos, aparición de enfermedades oportunistas, mayor morbimortalidad, peor recuperación inmunológica y, por ende, mayores costes sanitarios.

En España, a fecha de 2020 no se había conseguido aún el objetivo mar-

[1] ONUSIDA. Consultado el 26 de abril de 2023.

cado del 90% de personas con infección por VIH que conocen su serología, que se estimaba en el 87% (DCVIHT, 2021). Actualmente la tasa global de nuevos diagnósticos de VIH en España está en niveles similares a los de otros países de Europa Occidental, pero superior a la media de la Unión Europea. En el año 2021 (último año con datos actualizados) se notificaron 2.789 nuevos diagnósticos de VIH, lo que supone una tasa de 5,89/100.000 habitantes. De estos nuevos diagnósticos, el 49,8% presentaban un diagnóstico tardío (DT), siendo mayor en mujeres, 54,4%. Se considera DT cuando el primer recuento de linfocitos CD4 tiene una cifra inferior a 350 células/μl (DCVIHT, 2022). Disminuir el DT de la infección por VIH es uno de los principales retos de la respuesta a la epidemia del VIH. La población y los profesionales sanitarios todavía no son plenamente conscientes de que cualquier persona que realice prácticas de riesgo es vulnerable al VIH, y de que es importante diagnosticar la infección lo antes posible (DCVIHT, 2022). En un estudio realizado en el Hospital Universitario Fundación Alcorcón (HUFA) para evaluar el hábito de solicitar la serología según los riesgos e indicadores de infección, se objetivó que tan solo el 78% de los médicos la solicitaría a todo paciente que declarase promiscuidad sexual, y sólo el 24% de los profesionales consideró adecuado realizar cribado a todos los pacientes que acudiesen al hospital (Crespillo et al., 2015). Los resultados de este trabajo muestran la falta concienciación de los profesionales que puede ser extensiva a otros centros. Existe, por tanto, margen de mejora en la realización de cribado por parte de los profesionales, en particular en pacientes que presentan algunas características clínicas o epidemiológicas asociadas con la infección por VIH.

Hay varias propuestas para realizar el cribado de la infección por VIH: una estrategia universal (*opt-out*) o bien una estrategia basada en condiciones que sugieren la realización de la prueba, llamados indicadores (*indicator-condition*, IC, en inglés). Se sabe que el diagnóstico rutinario de VIH de los pacientes con IC aumenta la detección de la infección. Por otro lado, se ha objetivado que en el año anterior al diagnóstico de la infección por VIH, la mitad de los pacientes tienen algún contacto con el sistema sanitario y a pesar de tener IC no se les realiza la prueba de VIH. Se considera que el punto de corte de prevalencia de un IC para que la realización del cribado de VIH sea coste-efectivo es 0,1%. Es decir, indicadores que aparecen en una proporción de al menos el 0,1% de los pacientes con VIH. Este dintel se con-

firmó en el estudio HIDES (*HIV Indicator Diseases Across Europe Study*) (Sullivan et al., 2013) y es el que posteriormente adoptó el Ministerio de Sanidad en 2014. Los costes de realizar cribado son 15 veces mayores si se realiza de forma universal que si se hace guiado por indicadores (Menacho et al., 2013). Por ello, las recomendaciones nacionales e internacionales son las de realizar cribado siguiendo IC para hacer sostenible la intervención.

Existen distintas barreras para solicitar la serología VIH. Por parte del paciente destacan baja percepción de riesgo, miedo al diagnóstico y sus implicaciones o dificultad de acceder a los servicios sanitarios (Montoy, Dow, y Kaplan, 2016); por parte del personal sanitario, la falta de tiempo, cargas de trabajo elevadas, falta de formación específica, falta de percepción del riesgo u otras prioridades sanitarias. Es por tanto relevante la búsqueda de nuevos instrumentos que puedan ayudar a los profesionales a tomar decisiones, optimizando así las estrategias de cribado.

La historia clínica electrónica proporciona un recurso valioso para la detección de pacientes con riesgo de infección por VIH. El desarrollo de una herramienta basada en el análisis de posibles indicadores de riesgo en estas historias, podría ser una solución eficaz que, incorporada a la práctica habitual, podría aliviar las dificultades de los sanitarios. Las nuevas técnicas de aprendizaje automático y procesamiento del lenguaje natural se postulan como una solución prometedora para este reto.

1.2 Objetivos

El objetivo de este trabajo es desarrollar un sistema para ayudar en la detección temprana de casos de riesgo de infección por VIH. Para ello, se propone la utilización y comparación de varios modelos y técnicas dentro del Procesamiento del Lenguaje Natural y el Aprendizaje Automático.

A lo largo de esta memoria se irá profundizando acerca de las distintas propuestas y técnicas utilizadas, así como de su evaluación correspondiente. Los objetivos específicos que busca conseguir este trabajo son los siguientes:

- Revisar el estado del arte en Procesamiento del Lenguaje Natural dentro del campo clínico, enfocado a la predicción de infección por VIH.
- Desarrollar una propuesta no supervisada capaz de detectar casos de riesgo de infección por VIH a partir de una serie de indicadores proporcionados por médicos especialistas.

- Desarrollo de una propuesta supervisada basada en la clasificación mediante algoritmos de *Machine Learning*.
- Evaluar las distintas propuestas, y compararlas entre sí en busca de la mejor solución posible.
- Realizar un análisis completo de errores y localizar los puntos de posible mejora de cara al futuro.

1.3 Estructura del documento

Este documento se estructura en varios capítulos destinados a tratar distintos aspectos relacionados con el trabajo realizado. La estructura y una pequeña descripción de los capítulos se detallan a continuación.

Capítulo 2. Estado del arte. En este capítulo se describen con mayor detalle las disciplinas que nos ocupan, detallando de lo que tratan y su estado actual. Se realiza una pequeña descripción de las disciplinas del *Machine Learning* y el Procesamiento del Lenguaje Natural, centrando esta última sobre todo en su subrama biomédica (BioNLP). Se realiza una revisión detallada del tema que nos acontece como es el estudio del VIH y su relación con los dos temas anteriores, y las técnicas actuales más utilizadas, así como sus debilidades. Por último, se realiza un análisis crítico, detectando posibles problemas a mejorar, como punto de partida del trabajo.

Capítulo 3. Marco experimental. Este capítulo engloba la descripción del problema a abordar, el análisis del conjunto de datos utilizado, y la metodología y métricas de evaluación utilizadas en los capítulos siguientes.

Capítulo 4. Herramientas y recursos. Recopilación de las herramientas y recursos utilizados en el desarrollo del trabajo y las distintas propuestas.

Capítulo 5. Propuesta no supervisada. Primera propuesta del trabajo. Esta hace uso de una serie de indicadores ofrecidos por expertos y algoritmos basados en reglas para solucionar el problema.

Capítulo 6. Propuesta supervisada. La segunda propuesta, trata el problema como un clásico problema de clasificación con aprendizaje automático, comparando algoritmos tradicionales y *transformers* para buscar la mejor solución.

Capítulo 7. Discusión. Se analizan y discuten en profundidad los resultados obtenidos en las evaluaciones presentadas en ambas propuestas mediante una comparativa. Además, se incluye un análisis de errores de las estrategias y decisiones tomadas a lo largo del trabajo.

Capítulo 8. Conclusiones y trabajo futuro. Este último capítulo recopila las diferentes conclusiones extraídas del trabajo realizado y propone algunas líneas de trabajo futuro.

Capítulo 2

Estado del arte y preliminares

Este trabajo hace uso de distintos conceptos agrupados dentro del campo de la Inteligencia Artificial (IA) y el del Procesamiento del Lenguaje Natural (PLN). Más concretamente, dentro de la IA nos enfocaremos en la rama del aprendizaje automático, y dentro del PLN, en la rama del procesamiento de textos biomédicos, en especial de historias clínicas y el reconocimiento de entidades.

También se dará una visión acerca del estado del arte con respecto al diagnóstico y detección de riesgo de VIH, enfocado en su relación con diversas técnicas informáticas e IA, haciendo especial hincapié en el uso del PLN en esta materia.

2.1 Aprendizaje Automático

El aprendizaje automático (del inglés *Machine Learning*, ML) es un subcampo de las ciencias de la computación y una rama de la IA, cuyo objetivo es el de desarrollar algoritmos que mejoren automáticamente a través del aprendizaje (Mitchell, 1997). El aprendizaje se entiende como la capacidad de hacer predicciones o tomar decisiones sin haber sido programado exactamente para esa tarea (Russell y Norvig, 2021).

Estos algoritmos se denominan "modelos", pues construyen modelos matemáticos que son capaces de inferir patrones o relaciones en los datos y, gracias a la estadística, hacer predicciones sobre nuevos datos nunca vis-

tos. Normalmente el aprendizaje se basa en el "ajuste" de una serie de parámetros numéricos que se van optimizando a través de un proceso iterativo (Nadkarni, Ohno-Machado, y Chapman, 2011).

El ML puede a su vez dividirse en dos ramas principales, diferenciadas por el tipo de datos que se utilizan en cada una. Estas son el aprendizaje no supervisado y supervisado. Este trabajo presenta una propuesta supervisada y otra no supervisada, basada en conocimiento experto. No obstante, en ambas propuestas se hace uso de herramientas entrenadas mediante aprendizaje supervisado.

2.1.1 Aprendizaje supervisado

El aprendizaje supervisado se caracteriza por el uso de datos etiquetados, es decir, los datos de entrenamiento consisten en pares de objetos, una componente de entrada, y una componente (etiqueta) con el resultado deseado. Los modelos se entrenan iterativamente sobre estos pares de datos hasta que optimizan la precisión de una función que les permite predecir la salida de nuevos datos nunca vistos. Este tipo de algoritmos se utilizan sobre todo en tareas de clasificación y regresión.

En este tipo de algoritmos puede producirse un problema llamado sobreajuste (*overfitting*). Este problema se da cuando un modelo se ajusta de manera casi perfecta a los datos de entrenamiento, pero no tiene la capacidad de predecir ("generalizar") sobre nuevos datos. Suele aparecer cuando el modelo es demasiado complejo, o el número de datos de entrenamiento no es suficiente. Para minimizar el riesgo de sobreajuste se utilizan técnicas como la validación cruzada (*cross-validation*).

A continuación se explicarán las bases teóricas de los modelos utilizados en la realización del trabajo. Los detalles concretos de la utilización de cada uno se presentan en los capítulos 5 y 6.

2.1.1.1 Máquinas de Vectores de Soporte (SVM)

Las Máquinas de Vectores de Soporte, en inglés *Support-Vector Machines* (SVM) (Cortes y Vapnik, 1995), son un conjunto de algoritmos de aprendizaje supervisado especialmente relacionados con clasificación binaria. Formalmente, una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de gran dimensionalidad para dividir este en subespacios que definen las clases correspondientes, normalmente dos.

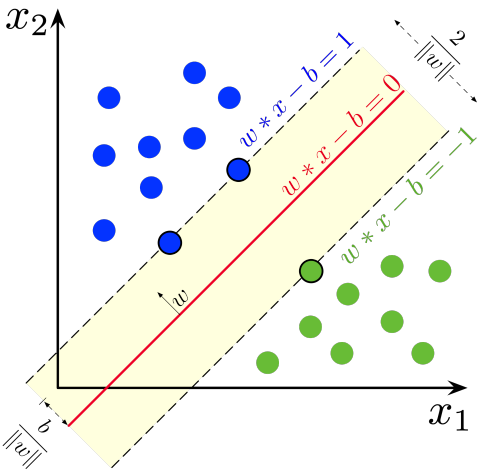

Figura 2.1: Clasificación binaria con SVM. Un hiperplano, línea roja, separa ambas clases, diferenciadas en verde y azul. Los puntos más cercanos al hiperplano forman dos vectores paralelos a este, llamados vectores de soporte, indicados en los mismos colores que su respectiva clase.

La característica fundamental de estos algoritmos reside en el concepto de "separación óptima", que consiste en buscar que el hiperplano tenga la máxima distancia posible (margen) con los puntos más cercanos a él. Al vector formado por los puntos más cercanos al hiperplano, y paralelo a este, se le llama vector de soporte. La Figura 2.1 muestra el hiperplano (en rojo) y los vectores de soporte (azúl y verde) que separan dos clases[1].

La razón para elegir este modelo es que se trata de un algoritmo sólido en cuanto a clasificación y tiene un coste computacional de entrenamiento no demasiado alto. El algoritmo utilizado es la función `SVC`[2] de la librería `sklearn`.

2.1.1.2 Modelo de Máxima Entropía

Los modelos clasificadores de máxima entropía, también llamados modelos de regresión logística multinomial, son un tipo de análisis de regresión para predecir el resultado de una variable categórica en función de una serie de variables predictoras. Es uno de los modelos más utilizados en problemas de clasificación en PLN porque no asume la independencia estadística de las variables predictoras (Malouf, 2002).

Asumiendo una variable categórica binaria y_i. Se calcula la probabilidad

[1] De Larhmam - Trabajo propio, CC BY-SA 4.0, wikimedia.org
[2] sklearn.SVM.SVC. Consultado el 4 de julio de 2023.

de predecir la clase positiva $P(y_i = 1|X_i)$ como:

$$\hat{p}(X_i) = \text{expit}(X_i w + w_0) = \frac{1}{1 + \exp(-X_i w - w_0)} \tag{2.1}$$

La optimización del modelo consiste en la minimización de la siguiente función de coste:

$$\min_w C \sum_{i=1}^{n} (-y_i \log(\hat{p}(X_i)) - (1 - y_i) \log(1 - \hat{p}(X_i))) + r(w) \tag{2.2}$$

2.1.1.3 Aprendizaje profundo

El aprendizaje profundo, en inglés *Deep Learning* (DL), es un conjunto de algoritmos de ML que, haciendo uso de redes de neuronas (RNN), intentan modelar abstracciones de alto nivel en datos. Este modelado se realiza mediante la extracción y asimilación de información (características o *features* en inglés) dentro de estos datos. A este proceso se le conoce como aprendizaje de características o *Feature Learning* (FL) (Bengio, Courville, y Vincent, 2013). El nombre "profundo" viene del gran número de capas de neuronas que poseen estas redes. El uso de múltiples capas permite el aprendizaje progresivo de características cada vez más abstractas de los datos. Alcanzar estos niveles de abstracción permite al sistema discriminar la información irrelevante y descubrir patrones ocultos (Goodfellow, Bengio, y Courville, 2016).

Dentro del aprendizaje profundo existen multitud de modelos dependiendo del tipo de RNN utilizadas. Destacan las redes convolucionales (*Convolutional Neural Networks*, CNN), muy usadas en tareas de visión por computador (Lecun et al., 1998) gracias a tener una organización jerárquica, localidad e invarianza espacial que les permite identificar objetos aunque estén rotados, escalados o transformados. Estas propiedades también son muy útiles en tareas de PLN (Chen, 2015) por su capacidad para extraer las características más relevantes de un grupo de palabras, o características morfológicas a nivel de carácter, por ejemplo.

Las redes recurrentes (*Recurrent Neural Networks*, RNN) (Elman, 1990) son un tipo de redes enfocadas en extraer información a lo largo de una secuencia. Su arquitectura se vale de una ventana deslizante que recorre la secuencia en ciclos. En cada ciclo se procesa una parte de la información y, además, se recibe la información previamente procesada. Esto permite

que la red tenga conocimiento del contexto anterior para tomar nuevas decisiones. Esta naturaleza secuencial las hace muy útiles a la hora de trabajar con texto. Existen multitud de variantes dependiendo de la cantidad de contexto que pueden procesar (*Long-Short Term Memory*, LSTM) (Hochreiter y Schmidhuber, 1997) o si tienen en cuenta tanto el contexto anterior como el posterior (*RNN bidireccionales*, biRNN) (Schuster y Paliwal, 1997). No obstante, estas redes no son capaces de almacenar el contexto de forma infinita, lo que hace que parte de la información se vaya perdiendo conforme avanza la secuencia. Además, su naturaleza recurrente dificulta su paralelización y las hace muy lentas.

Los *transformers* (Vaswani et al., 2017) son modelos de DL que incorporan mecanismos de atención (*self-attention*). La arquitectura del transformer se divide en dos partes: *encoder* y *decoder*. El primero, recibe la secuencia de entrada y genera una representación de alto nivel que, puede pasar directamente al *decoder*, o a otras capas adicionales para distintas tareas (clasificación o predicción, entre otras). El segundo, recibe la salida del *encoder* o alguna de las otras capas y genera una salida. El mecanismo de atención permite que el *decoder* pueda tener en cuenta toda la secuencia de entrada en cada paso de salida, para así decidir a qué partes de la entrada presta más atención en cada paso, asignando distintos pesos en función de la importancia que le da. Al depender cada elemento del resto, se le llama representación contextualizada. En la Figura 2.2 se muestra la arquitectura "básica" de un transformer, siendo la parte izquierda la correspondiente al *encoder*, y la derecha al *decoder* (Vaswani et al., 2017).

Estos bloques que forman la arquitectura se pueden utilizar de distintas maneras. Esta configuración depende del objetivo que se quiera conseguir. Una configuración con sólo *encoder*, llamados modelos discriminativos, está centrada en representar texto y es adecuada para tareas enfocadas en entender frases. A partir de estas representaciones se pueden hacer otras tareas como clasificar textos o extraer información. El modelo BERT (*Bidireccional Encoder Representations from Transformers*) (Devlin et al., 2018) es el modelo más representativo. La configuración de sólo *decoder*, llamados modelos generadores, sirven para generar texto, y el modelo más conocido es GPT (Radford et al., 2018). Por último, los que utilizan ambas partes son llamados modelos *sequence-to-sequence* (seq-to-seq), capaces de generar texto a partir de otro, y donde destaca el modelo T5 (Raffel et al., 2020), sobre todo

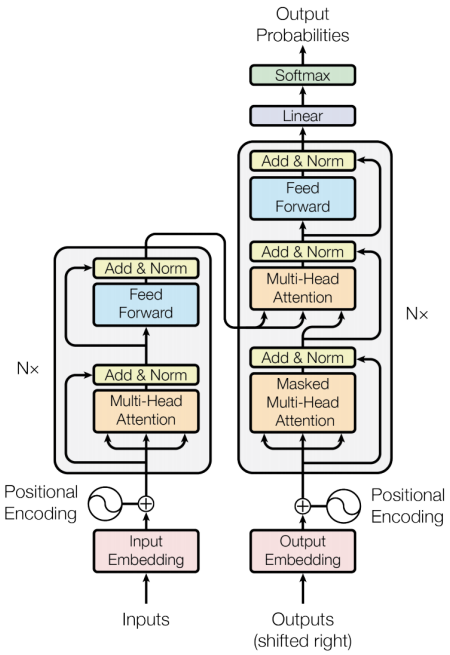

Figura 2.2: Arquitectura del modelo *transformer*. A la izquierda la entrada y el *encoder*, y a la derecha el *decoder* y la salida (Vaswani et al., 2017).

en tareas de traducción.

2.2 Procesamiento del lenguaje en textos biomédicos

El Procesamiento del Lenguaje Natural es una disciplina dentro de las ciencias de la computación y la lingüística, que se ocupa de la investigación de mecanismos para la comunicación humano-máquina por medio del lenguaje natural, y de modelar sistemas capaces de comprender el lenguaje para realizar tareas complejas (traducción, generación de resúmenes, extracción de información, etc.) (Cortez Vásquez et al., 2009).

Como en muchas otras disciplinas, el inmenso aumento en el número de documentos relacionados con la biomedicina, ya sean publicaciones o historias clínicas, obliga al desarrollo de sistemas capaces de procesar y aprovechar toda esta información. La mayoría de esta información se encuentra en forma de texto libre y requiere de su estructuración para poder ser utilizada. Por esto, la investigación en PLN en biomedicina ha crecido de

forma exponencial en las últimas décadas, convirtiéndose en una de las ramas más importantes dentro del PLN (Friedman, Rindflesch, y Corn, 2013).

Para promover la investigación, desde finales de los 90 se engloba esta colaboración entre biomedicina y PLN en un nueva área de investigación conocida como BioPLN, con el objetivo de desarrollar métodos de PLN que impulsen y ayuden en la investigación biomédica (Huang y Lu, 2016). A lo largo de estos años se han organizado un gran número de tareas o *workshops* dentro de conferencias enfocados a hacer avanzar esta colaboración. Unas de las primeras fueron las tareas de genómica de la ACM KDD Cup 2002[3] o la TREC 2003[4].

Estas primeras tareas se centraban en la recuperación y clasificación de documentos, conocido como recuperación de información (en inglés *Information Retrieval*, IR) dentro de la comunidad de PLN. En 2004, surgen las primeras tareas de detección de entidades biomédicas: la tarea de detección de genes en BioCreative I[5], y la tarea de JNLPBA (Collier et al., 2004), que incluye varios tipos de entidades (proteínas, ADN, ARN y tipos de células). El Reconocimiento de Entidades Nombradas (*Named Entity Recognition*, NER, en inglés), es una de las tareas clave dentro del BioNLP y será abordada con más profundidad en la sección 2.2.2.

Uno de los principales problemas con los que se encuentra la investigación con datos clínicos reside en la inaccesibilidad a gran cantidad de estos datos, siendo este el gran cuello de botella en cuanto al desarrollo de BioPLN (Friedman, Rindflesch, y Corn, 2013). Por razones de seguridad y privacidad, estos datos requieren ser anonimizados. Generalmente, sólo los investigadores afiliados con un centro médico pueden acceder a este tipo de información clínica, pero dichos documentos son difíciles de ser compartidos al resto de la comunidad a menos que pasen por determinados controles. Esto hace que la anonimización sea una de las tareas claves dentro del BioNLP. Uno de los modelos de anonimización más famosos es el *MITRE Identification Scrubber Toolkit* (MIST) (Aberdeen et al., 2010), que automáticamente se encarga de detectar información sensible y anonimizarla.

En textos en español, la tarea MEDDOCAN: Medical Document A-

[3] KDD Cup 2002: Biomed document; plus gene role classification. Consultado el 10 de mayo de 2023.
[4] TREC 2003 Genomics Track. Consultado el 10 de mayo de 2023.
[5] BioCreative I. Consultado el 10 de mayo de 2023.

nonymization Track[6], dentro del contexto de la IberLef 2019, proponía el desarrollo de modelos para anonimización a partir de un corpus de 1000 casos clínicos obtenidos manualmente de SciELO (Scientific Electronic Library Online)[7], una biblioteca electrónica que recoge artículos científicos de América Latina, Sudáfrica y España.

Otro problema importante, aunque este no es único del BioPLN, es la dificultad para obtener grandes cantidades de datos médicos etiquetados, sobretodo debido al alto coste que supone la anotación experta por parte de médicos. En este contexto, el *Transfer Learning* (aprendizaje transferido) se coloca como la mejor solución para este problema, pues permite obtener resultados muy buenos en tareas específicas a partir de conocimiento ya aprendido por otro modelo, lo que ahorra gran parte del aprendizaje y, por tanto, tiempo y dinero. El modelo más representativo dentro de este dominio fue el primer modelo BioBERT (Lee et al., 2020), que supuso el empujón inicial para el desarrollo de grandes modelos de lenguaje enfocados a BioPLN.

En el caso del español, cabe destacar la iniciativa gubernamental llamada Plan de Impulso de las Tecnologías del Lenguaje (Plan TL)[8], con el objetivo de fomentar el desarrollo del PLN en español y las lenguas cooficiales de España. Desde esta iniciativa se ha promovido el desarrollo de multitud de conjuntos de datos y modelos de *transformers*, algunos específicos del dominio biomédico[9]. Algunos de estos modelos han sido utilizados en este trabajo y serán comentados en mayor profundidad posteriormente (Sección 4.1.1).

2.2.1 Procesamiento de historias clínicas

La historia clínica electrónica (EHR, del inglés *Electronic Health Record*), es el registro unificado y personal, en el que se archiva de forma electrónica toda la información referente a un paciente y a su atención médica (Gérvas y Fernández, 2000). Esta historia contiene información de toda índole, datos médicos recogidos durante toda la vida del paciente, registros de la situación y evaluación clínica en todo proceso asistencial; además de datos sociales como datos demográficos, consumo de sustancias u orientación sexual. Esta

[6] MEDDOCAN. Consultado el 4 de mayo de 2023.
[7] SciELO.org. Consultado el 4 de mayo de 2023.
[8] Plan de Impulso a las Tecnologías del Lenguaje. Consultado el 4 de mayo de 2023.
[9] PlanTL-GOB-ES. *Hugging Face*. Consultado el 4 de mayo de 2023.

cantidad y variedad de información hacen de los EHR una fuente de información de gran valor para la investigación médica. A este tipo de uso, más allá de la propia asistencia médica, se le conoce como uso secundario de los EHR (Pomares-Quimbaya, Kreuzthaler, y Schulz, 2019).

Esta variedad en la información se extrapola también a los tipos de datos almacenados, que pueden ser clasificados en estructurados y no estructurados. Los datos estructurados consisten en diagnósticos, medicamentos, y pruebas de laboratorio en forma de valores numéricos o categóricos. En cambio, los datos no estructurados están en forma de texto libre y suelen representar el 80% de los datos totales del EHR, como pueden ser notas clínicas o informes de alta (Xiao, Choi, y Sun, 2018). En este tipo de datos es donde el PLN entra en escena.

Los EHR presentan una serie de problemas y desafíos como son la privacidad y la falta de anotación. No obstante, en este caso, estos problemas se magnifican porque las distintos datos en un EHR pueden haber sido escritos por varios médicos a lo largo de distintos años, presentando una gran variabilidad en su calidad, y haciéndolos más difíciles de anotar (Li et al., 2022).

Las tareas de clasificación y predicción en los EHR son fundamentales para procesar rápidamente miles de textos de gran tamaño como apoyo a la toma de decisiones clínicas, investigación y optimización de procesos. Estas tareas además engloban varias subtareas, donde cada una juega un papel muy importante en el correcto procesamiento de estos documentos. Por ejemplo, la anonimización de la que se hablaba en la sección 2.2 puede tratarse como una tarea de clasificación. A continuación, repasaremos otras de las más relevantes:

- **Clasificación de textos médicos.** Esta clasificación puede tener fines muy diversos. Por ejemplo, (Marafino et al., 2014) es capaz de, usando un clasificador SVM y n-gramas, clasificar las notas en varios tipos de procedimientos y enfermedades. Otros, como (Yao, Mao, y Luo, 2019), utilizan una combinación de características basadas en reglas y DL para clasificar enfermedades. También hay trabajos más específicos, como (Valmianski et al., 2019), que intentan clasificar las notas según la motivación del paciente para acudir a consulta (*chief complaint*).

- **Segmentación.** Aunque existen estándares de EHR, no existe una

unificación total de criterios, lo que hace que el formato utilizado dependa de las propias instituciones médicas. Esto hace que, por ejemplo, existan notas que pueden estar o no explícitamente divididas en secciones (motivación del paciente, anamnesis, analíticas, etc.). La tarea de segmentación se refiere a la labor de identificar automáticamente estas secciones, estén o no explicitadas dentro de un EHR. (Badjatiya et al., 2018) utiliza mecanismos de atención y redes convolucionales para clasificar de forma binaria si una frase era o no el inicio de una sección. (Goenaga et al., 2021) compara tres métodos distintos, uno basado en reglas, un perceptrón y un algoritmo de DL, para identificar secciones en informes de alta con estándar HL7[10]. Sus resultados, sobre todo con DL son muy buenos, pero su conclusión evidencia que la generalización en este tipo de sistemas es muy complicada, habiendo grandes diferencias incluso entre hospitales. La tarea ClinAIS[11] dentro de IberLEF 2023[12], es la primera tarea en español que propone un dataset para identificar distintas secciones dentro de notas clínicas.

- **Desambiguación de significado**. Esta tarea consiste en asignar el correcto significado a una palabra dado su contexto. Es muy útil cuando hay palabras que tienen diversos significados o cuando hay conceptos que se pueden escribir de varias maneras (Wang et al., 2018). Suele ser una parte importante en tareas de *Entity Linking* (EL), consistentes en relacionar una entidad con un concepto dentro de una base de conocimiento (*knowledge base*, KG). Uno de los ejemplos más relevantes es el modelo deepBioWSD (Pesaranghader et al., 2019), que utiliza embeddings extraídos de UMLS con una red BiLSTM para solucionar ambigüedades. La desambiguación de abreviaturas es un caso especial dentro de la desambiguación. En esta, destaca en español especialmente la tarea de evaluación BARR (*Biomedical Abbreviation Recognition and Resolution*), desarrollada en los años 2017[13] y 2018[14] en el marco de IberEval, dentro de la SEPLN. Esta tarea propone la identificación y resolución de acrónimos en un corpus compuesto

[10] HL7 (en inglés). Consultado el 12 de mayo de 2023.
[11] ClinAIS. Consultado el 11 de mayo de 2023.
[12] IberLEF 2023. Consultado el 11 de mayo de 2023.
[13] IberEval2017 - Biomedical Abbreviation Recognition and Resolution (BARR) (en inglés). Consultado el 30 de mayo de 2023.
[14] IberEval2018 - Biomedical Abbreviation Recognition and Resolution 2nd Edition (BARR2) (en inglés). Consultado el 30 de mayo de 2023.

por resúmenes y estudios clínicos de diversas fuentes. Especialmente enfocado en la resolución de acrónimos en EHR, destaca el corpus IULA-SCR-ABB de (Aguado y Bel Rafecas, 2022).

- **Codificación médica.** Esta tarea intenta mapear una pieza de texto de un EHR a un sistema estándar de codificación, normalmente siendo este la Clasificación Internacional de Enfermedades (en inglés *International Classifcation of Diseases*, ICD). Estos códigos se organizan en dos categorías, diagnósticos y procedimientos, y contienen códigos para todo tipo de conceptos médicos, enfermedades, síntomas, circunstancias sociales, etc. Tienen una gran importancia a la hora de ayudar a los médicos en tareas de diagnóstico o elaboración de estadísticas, entre otras. No obstante, la asignación manual de este tipo de códigos es una labor compleja, ya que las definiciones de estos conceptos son muy específicas y detalladas. Por ello, el desarrollo de codificadores automáticos se ha convertido en una gran necesidad para ayudar en esta tarea. Debido al inmenso número de códigos que existen, este proceso puede ser visto como una tarea de Clasificación Multi-etiqueta Extrema (en inglés *Extreme Multi-label Classification*, XMLC). Este es un caso particularmente difícil de clasificación donde el sistema debe ser capaz de encontrar un subgrupo de clases que asignar al documento, entre un grupo extremadamente grande de categorías (Liu et al., 2017). El primer trabajo en tratar esta tarea como un problema de XMLC fue (Almagro et al., 2020).

2.2.2 Reconocimiento de Entidades Nombradas

El Reconocimiento de Entidades Nombradas (*Named Entity Recognition*, NER, en inglés), es una tarea dentro de la extracción de información (*Information Extraction,* IE, en inglés) encargada de localizar y categorizar partes de un texto. Estas "entidades nombradas" son una palabra o conjunto de palabras que representan un concepto perteneciente a una categoría predefinida, como pueden ser una persona, un lugar, una fecha, etc. (Chinchor y Robinson, 1997).

Formalmente, dada un secuencia de *tokens* $s = \langle w_1, w_2, \cdots, w_N \rangle$, el sistema NER obtiene como resultado una lista de tripletas de la forma $\langle I_s, I_e, t \rangle$, donde cada elemento es una entidad de la secuencia s, siendo $I_s \in [1, N]$ e $I_e \in [1, N]$ los índices inicial y final donde se localiza la entidad dentro del

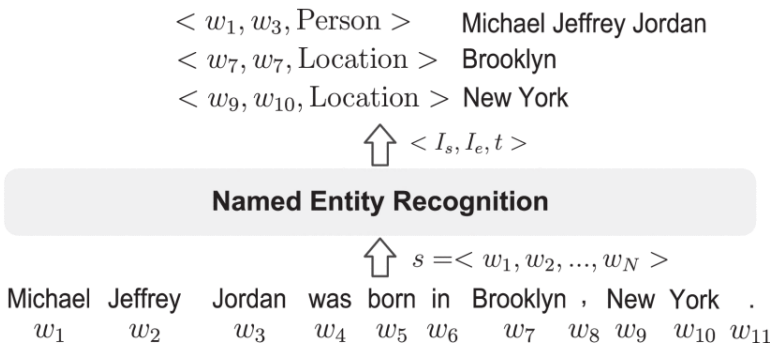

Figura 2.3: Ejemplo del proceso de anotación de un sistema NER.

texto, y t el tipo de entidad dentro del conjunto de entidades predefinidas. La Figura 2.3 muestra un ejemplo de la formalización de un sistema NER reconociendo una frase (Li et al., 2020).

En el caso de que estas entidades estén formadas por más de una palabra, suele utilizarse como estándar la notación IOB (*Inside-Outside-Beginning*) para representar estas entidades gracias a una serie de etiquetas. Estas indican los límites de la entidad y son: *B-entity* para indicar la palabra inicial, *I-entity* para las palabras siguientes que forman parte de la entidad, y *O* para el resto de palabras que no pertenecen a ninguna entidad (Ramshaw y Marcus, 1999).

En el ámbito biomédico, el NER se suele utilizar para identificar entidades como nombres de enfermedades, fármacos, partes anatómicas o genes, entre otros. De forma similar a otras tareas de PLN, existen gran cantidad de recursos para entrenar sistemas NER biomédicos, la mayoría derivados y promovidos en conferencias, como decíamos en la Sección 2.2. Desde las primeras tareas descritas anteriormente, el NER ha tomado una relevancia esencial, siendo clave en el soporte de tareas más complejas.

Centrándonos en la detección de enfermedades, existen varios corpus en español que contienen anotaciones referentes a estas, cada uno con sus propias características. El corpus IxaMed-GS (Oronoz et al., 2015) recopila una colección de EHR con anotaciones de enfermedades y fármacos, relacionándolos en caso de indicar un posible efecto adverso. Otros como el Chilean Waiting List Corpus (Báez et al., 2020) o CT-EBM-SP (Campillos-Llanos et al., 2021) contienen anotaciones de enfermedades, procedimientos o partes del cuerpo, entre otras.

En el contexto de este trabajo, la extracción de entidades nos permite

obtener un resumen de las historias clínicas en forma de entidades concretas. Por ejemplo, en un contexto ideal, la extracción de entidades de enfermedades de un paciente nos permite tener todo el historial de enfermedades de este, y probablemente en orden cronológico. Si comparáramos estas entidades con el listado de posibles indicadores de una enfermedad, podríamos identificar si existe un posible riesgo de padecer esa enfermedad.

2.2.2.1 Métodos NER

Dentro de la tarea de NER, existen tres enfoques principales: métodos basados en reglas, basados en *machine learning* y basados en *deep learning*.

- **Métodos basados en reglas.** Estos sistemas se basan en reglas diseñadas a mano, aunque también hay métodos de generación de reglas automáticos a partir de ciertos parámetros. Las reglas suelen diseñarse a partir de una nomenclatura o diccionario específico dentro de un dominio y de patrones sintáctico-léxicos, como FACILE (Black, Rinaldi, y Mowatt, 1998) o SRA (Aone et al., 1998). Al ser reglas muy específicas para un dominio, funcionan muy bien cuando tienen un léxico muy exhaustivo, pero empeoran su calidad cuando los diccionarios en los que se basan no son especialmente completos, o cuando se transfieren a otros dominios.

- **Métodos basados en ML.** Como se ha explicado en la sección 2.1 existen los sistemas de ML no supervisados y supervisados.

  - **Métodos no supervisados.** En relación a los primeros, el método más utilizado es el *clustering*. Los sistemas NER basados en esta técnica extraen las entidades agrupando los elementos basándose en su similitud. En el dominio biomédico, (Zhang y Elhadad, 2013) recurren a estadísticas como TF-IDF o vectores de contexto, y a la terminología del dominio para extraer las entidades, superando a métodos anteriores.
  - **Métodos supervisados.** Estos sistemas convierten el NER en una tarea de clasificación multi-clase o de etiquetado de secuencias. En la actualidad, estos métodos destacan por encima de los no supervisados por su capacidad para extraer patrones en los datos. Los sistemas más utilizados son modelos de Markov (*Hidden Markov Models*, HMM, en inglés) (Morwal, Jahan, y Chopra,

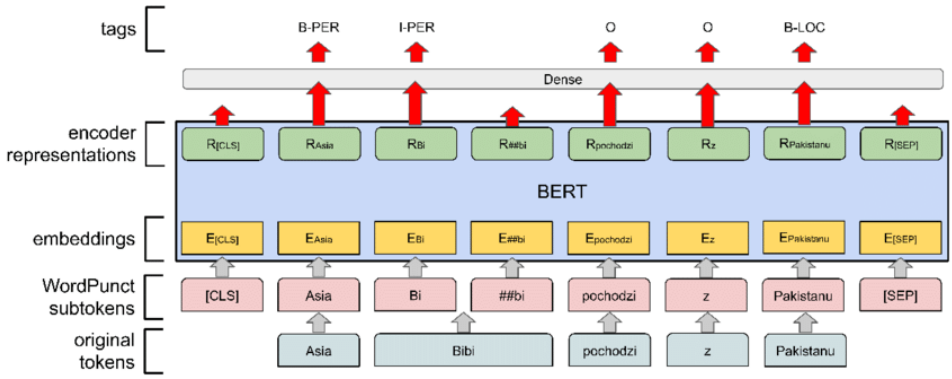

Figura 2.4: Arquitectura de transformer (BERT) para la tarea de NER (Arkhipov et al., 2019).

2012), modelos de campo aleatorio condicional (*Conditional Random Fields*, CRF, en inglés) (Lafferty, McCallum, y Pereira, 2001) y SVM.

- **Métodos basados en DL.** En los últimos años, la tendencia ha sido el uso de modelos de DL, sobre todo redes LSTM bidireccionales (Bi-LSTM), combinados con CRF (BiLSTM-CRF) (Cho y Lee, 2019). La aparición de los *transformers* también ha dado muy buenos resultados en la tarea NER. En la Figura 2.4 se muestra un ejemplo de la arquitectura de estos sistemas, una capa de *embeddings*, seguida de una capa de *encoding* posicional y, por último, una capa donde se calculan las probabilidades de que una palabra pertenezca a una entidad (Arkhipov et al., 2019). Esta última capa puede ser una capa densa como en la figura, o un modelo CRF que calcule la probabilidad final.

2.2.2.2 Normalización de entidades

La normalización de entidades (*Entity Normalization*, EN, en inglés), es una subtarea que suele acompañar al NER en multitud de tareas de evaluación dentro de BioNLP. Este es un problema concreto dentro de la codificación médica, en el que las entidades son relacionadas con códigos únicos dentro algún sistema estandarizado de terminología clínica. También puede ser interpretado como una tarea de EL, en la que se relaciona una entidad con un nodo en una base de conocimiento.

Entre estas terminologías destacan los códigos ICD-10, mencionados anteriormente, o SNOMED-CT[15] (*Systematized Nomenclature of Medicine – Clinical Terms*). SNOMED-CT proporciona una ontología de más de 350.000 conceptos jerarquizados y relacionados semánticamente. Dentro del gran abanico de estándares, destaca UMLS[16] (*Unified Medical Language System*) como el compendio unificado de la mayoría de estos vocabularios.

En los últimos años se ha fomentado la aparición de varias tareas de evaluación de NER enfocadas a la detección y normalización de enfermedades. Tareas como DisTEMIST[17] o CodiESP[18], son dos ejemplos de este tipo de tareas, normalizando enfermedades con SNOMED-CT y ICD-10, respectivamente.

Los métodos utilizados para solucionar esta tarea van muy ligados a los comentados en la tarea de NER. La mayoría de tareas de evaluación con EN, incluyen un gazetteer, o lexicón, específico de la nomenclatura utilizada, por lo que las soluciones basadas en reglas o diccionarios tienen mucha presencia. Tareas de evaluación como DisTEMIST incluyen este tipo de gazetteer[19], donde se recogen todos (o la mayoría) de conceptos dentro de SNOMED-CT, en este caso, incluyendo sinónimos y acrónimos, y un código único que los unifica.

Como en otras muchas cuestiones, los métodos basados en DL han ocupado gran parte de la atención recientemente dentro de la tarea de EL. La arquitectura general de estos acercamientos (Sevgili et al., 2022), consiste en utilizar un método de desambiguación que ayude a identificar una serie de posibles "candidatos" a partir de una entidad. Es decir, extraer de un KG la lista de posibles entidades normalizadas. Posteriormente, codificar tanto las entidades como los candidatos, y compararlos utilizando una medida de similitud.

2.3 Detección de VIH

Como se ha comentado en el Sección 1.1, la detección temprana de VIH es el auténtico reto al que se enfrentan los países más desarrollados en relación

[15] SNOMED International. Consultado el 29 de junio de 2023.
[16] Unified Medical Language System (UMLS). Consultado el 29 de junio de 2023.
[17] DisTEMIST. Consultado el 19 de mayo de 2023.
[18] CodiEsp. Consultado el 19 de mayo de 2023.
[19] DisTEMIST gazetteer. Consultado el 19 de mayo de 2023.

a la enfermedad, debido a su importancia en la disminución de la transmisibilidad y, sobre todo, en la reducción de coste sanitarios. Es más, existen evidencias acerca de que el inicio prematuro de la terapia antirretroviral (TAR) reduce el riesgo de SIDA (Sanders et al., 2005).

Existen varios trabajos previos que utilizan la historia clínica de los pacientes para ampliar la población a la que realizar pruebas de VIH en función de su riesgo de padecer la enfermedad. En (Felsen et al., 2017) se propone un sistema de alerta basado en reglas donde surgía un aviso si un paciente nunca se había realizado un test, o si, después de un test negativo, aparecían en su expediente algún posible factor de alto riesgo identificado por un código ICD9-CM (infecciones de transmisión sexual, hepatitis B o C, entre otras). Esta ayuda electrónica supuso un alto incremento de diagnósticos mientras estuvo en ejecución, pasando los nuevos diagnósticos de 8.2/100,000 hospitalizados a 37.0/100,000.

En España, en Atención Primaria en Barcelona, se desarrolló una herramienta similar para alertar a pacientes que presentaran un factor indicador de riesgo, de que se realizaran una prueba de VIH. Este recordatorio fue el factor más importante asociado a la solicitud de test mientras estuvo activo (Redondo et al., 2019).

En 2020 se publicó un Documento de Consenso (DC) (González del Castillo et al., 2020), donde se establecen una serie de recomendaciones para la práctica del cribado dirigido de VIH en los servicios de urgencias hospitalarias (SUH). Estas recomendaciones se basan en la necesidad de realizar un cribado si aparecen una de las seis circunstancias clínicas que ellos determinan como claves por su alta prevalencia en pacientes positivos, y la frecuencia con las que son atendidas en urgencias. Entre estas se encuentran enfermedades como infecciones de transmisión sexual (ETS) o neumonía adquirida en la comunidad. En (Miró i Andreu et al., 2023), se prueba que la puesta en marcha de esta estrategia es posible y eficiente.

En (Salmerón-Béliz et al., 2022) estudian "oportunidades perdidas" en visitas de pacientes al SUH en 27 hospitales españoles de 7 comunidades autónomas. Estas se refieren a oportunidades en las que el paciente acudió a urgencias, previas al diagnóstico de la infección, en las que el paciente presentaba indicios de una posible infección, pero no se realizó una prueba para conocer su estado. Se revisaron todas las consultas en los 5 años previos, y el 16,3% fueron una oportunidad perdida. De los pacientes con al menos

una visita a urgencias en esos años, el 45% acudió al menos una vez por una condición asociada al VIH.

2.3.1 Modelos de predicción de VIH

La labor de intentar predecir el riesgo de padecer una enfermedad ha sido siempre uno de los problemas que más interés ha despertado en la comunidad científica (Fieggen et al., 2022). En especial cuando se trata de enfermedades infecciosas como el VIH o COVID-19 (Agrebi y Larbi, 2020; Chiu et al., 2022). En relación al VIH existen numerosos trabajos que han planteado soluciones para tratar de identificar pacientes en riesgo de padecer la enfermedad.

Al principio, los pacientes eran clasificados sólo utilizando factores de riesgo "tradicionales", es decir, conocidos y estudiados. Después, surgieron propuestas que calculaban una "puntuación de riesgo" usando regresión logística, con pruebas en distintos grupos de riesgo entre los que se incluyen hombres que tienen sexo con hombres (HSH, o en inglés: MSM de *men-who-have-sex-with-men*) o parejas serodiscordantes (sólo una persona de la pareja es positiva). Más recientemente, gracias al uso de ML hay propuestas que intentan cuantificar las complejas relaciones entre distintos factores de riesgo. En (Balzer et al., 2019) se compararon estas tres aproximaciones y su conclusión fue que el modelo de ML mejoraba a los otros dos en términos de eficiencia y sensibilidad.

En (Krakower et al., 2019) se compararon unos 40 algoritmos y más de 100 variables extraídas de EHR para predecir qué pacientes podrían tener VIH. Su AUC de 0.86 fue una mejora importante comparado con aproximaciones previas. El estudio fue incluso puesto en marcha en un centro médico de Boston donde encontraron que la capacidad predictiva del modelo era menor que en sus experimentos.

En un estudio paralelo, (Marcus et al., 2019) desarrolló un algoritmo similar para identificar posibles candidatos para PrEP entre 3.7 millones de pacientes. La PrEP (profilaxis prexposición)[20] es una estrategia de prevención mediante medicamentos que reduce el riesgo de infección por VIH. El algoritmo final, también un modelo LASSO, tenía 44 variables predictoras extraídas de EHR entre las que se incluían variables de distintos dominios, por ejemplo el código postal. El sistema superaba en AUC y sensibilidad a

[20] Acerca de la PrEP. *CDC*. Consultado el 15 de septiembre de 2023.

otros algoritmos que solamente utilizaban datos acerca de enfermedades de trasmisión sexual y la orientación sexual de los pacientes.

El estudio de "oportunidades perdidas", se refiere a la identificación de posibles indicadores o predictores asociados a un posible diagnóstico temprano de la infección para evitar un DT. (Weissman et al., 2021) utiliza técnicas de ML para intentar encontrar estos predictores. Las conclusiones de este trabajo recalcan el potencial de estas técnicas en el estudio de la detección temprana de la infección, así como la relevancia de las pruebas en los entornos de urgencias como clave para esta.

2.3.2 PLN y VIH

Como se ha visto anteriormente, la gran mayoría de los trabajos alrededor de la predicción de riesgo de VIH se han centrado mayormente en datos estructurados extraídos de los EHR. No obstante, hay factores de riesgo muy importantes para la evaluación del VIH que no se encuentran en estos datos estructurados, sino que se encuentran en forma de texto libre en notas clínicas de los pacientes. Estos pueden ser el uso de drogas, ambiente familiar, comportamientos sexuales, etc.

(Feller et al., 2018) fue el primer trabajo en predicción de VIH en utilizar PLN para analizar la historia clínica de los pacientes y utilizar esta información para enriquecer sus modelos. Su análisis de texto tiene dos partes principales: identificación de palabras clave (*keyword extraction*) y modelado de temas (*topic modelling*). La primera tiene como objetivo identificar palabras potencialmente valiosas para representar la nota clínica mediante TF-IDF. La segunda, utiliza un algoritmo LDA (*Latent Dirichlet Allocation*)(Blei, Ng, y Jordan, 2003) para clasificar las notas en una serie de temas. El algoritmo recibe como entrada una serie de notas y devuelve K grupos (*clusters*), que representan la distribución de las palabras en las notas.

En el paper comparan tres modelos: (i) utilizando únicamente información estructurada (información demográfica, resultados de pruebas); (ii) la información estructurada en combinación con los temas identificados en el *topic modelling*; (iii) la información estructurada combinada con las palabras clave. Los resultados muestran que los dos sistemas que añaden información extraída del texto libre mejoran significativamente al sistema que no la utiliza, siendo ligeramente superior el sistema (iii). Cabe decir que el

procesamiento del lenguaje en este trabajo es bastante básico, por ejemplo, no tiene en cuenta elementos como la negación, la variabilidad léxica, ni identifican conceptos clínicos estandarizados, como pueden ser conceptos de UMLS.

(Araujo et al., 2022) analizan los datos extraídos de la cohorte CoRIS[21], que incluyen datos rutinarios de pacientes infectados de VIH, mediante el uso de reglas de asociación (en inglés *association rules*, AR) y algoritmos de ML. Las AR (Agrawal, Imieliński, y Swami, 1993), son un método de minería de datos que pretende descubrir patrones de co-ocurrencia entre elementos de una base de datos. El objetivo del estudio es descubrir relaciones poco exploradas entre enfermedades asociadas al VIH usando estas reglas.

Para descubrir estas relaciones, se valen de un método semi-supervisado llamado EXTRAE (Sánchez-de Madariaga et al., 2022), que con una mínima cantidad de reglas (los mejores resultados que obtienen utilizan 35 reglas anotadas para entrenar) obtiene grandes resultados, siendo capaces de identificar reglas apoyadas por expertos en la literatura, pero poco conocidas.

2.4 Valoración del estado del arte e identificación de la línea de trabajo

Tras la realización del estudio del estado del arte se ha comprobado que no hay una gran cantidad de trabajos acerca de la detección de riesgo de infección por VIH utilizando PLN. La mayoría de trabajos que abordan el riesgo de infección por VIH lo hacen desde el punto de vista de la ciencia de datos, basándose especialmente en pruebas de laboratorio, y sin prestar atención a otro tipo de indicadores de riesgo que suelen estar presentes en forma de texto libre en las historias clínicas.

Apenas (Feller et al., 2018) aborda este problema desde la óptica del lenguaje natural, demostrando que el uso de esta información no estructurada mejoraba el diagnóstico. No obstante, su trabajo tiene mucho espacio de mejora, como es el hecho de tener en cuenta la negación, o utilizar conceptos clínicos estandarizados para apoyar sus decisiones.

El hecho de trabajar con datos en español también presenta dificultades. Tampoco se han encontrado trabajos en español que aborden la predicción de infección por VIH. (Araujo et al., 2022) es el único ejemplo encontrado

[21] CoRIS. Consultado el 30 de junio de 2023.

que relaciona PLN y VIH en español pero sin enfocarse en la predicción.

Por estos motivos, se identifican dos problemas abiertos en la literatura que serán abordados en este trabajo. El primero es el uso de PLN para el análisis de historias clínicas, utilizando técnicas y recursos diferentes a los explorados anteriormente en busca de una nueva solución. Y el segundo, la realización del trabajo completamente en español, tanto en los datos utilizados, modelos de lenguaje u otros recursos.

Se incidirá de forma especial en las mejoras propuestas en (Feller et al., 2018), la estandarización de los conceptos clínicos, y la identificación de la negación. Para ello, se realizará el entrenamiento de un sistema NER capaz de identificar conceptos clínicos y su sentido afirmativo o negativo dentro de la historia clínica. De forma posterior, se normalizarán las entidades resultantes para ser estandarizadas.

Capítulo 3

Marco Experimental

En este capítulo se describe en profundidad el marco experimental utilizado para llevar a cabo el trabajo. Dentro de este marco se realiza el análisis del conjunto de datos utilizado y la explicación de la metodología de evaluación.

El problema a resolver se puede interpretar como un problema de clasificación binaria (VIH positivo o negativo) en el que categorizar una serie de notas clínicas en base al riesgo de los pacientes de estar infectados de VIH.

3.1 Análisis del conjunto de datos

A partir de un sólo conjunto de datos, se utilizan en el trabajo dos representaciones distintas del mismo. Por un lado, el propio conjunto de datos, consistente en una colección de historias clínicas (corpus), y por otro, el conjunto de entidades, referentes a enfermedades y síntomas, extraídas del mismo mediante el sistema NER explicado en la Sección 4.1. A continuación, se realiza un análisis diferenciado de ambos conjuntos para estudiarlos con más detalle.

3.1.1 Corpus

El conjunto de datos utilizado se trata de una recopilación de 406 notas clínicas, de las que 100 pertenecen a pacientes diagnosticados con VIH (VIH+), y 306 a pacientes sin VIH (VIH-). Es un corpus claramente desbalanceado, puesto que en el contexto real (sanos/infectados) también lo es, aunque en mayor medida.

El corpus se compone de una colección de notas seleccionadas de los

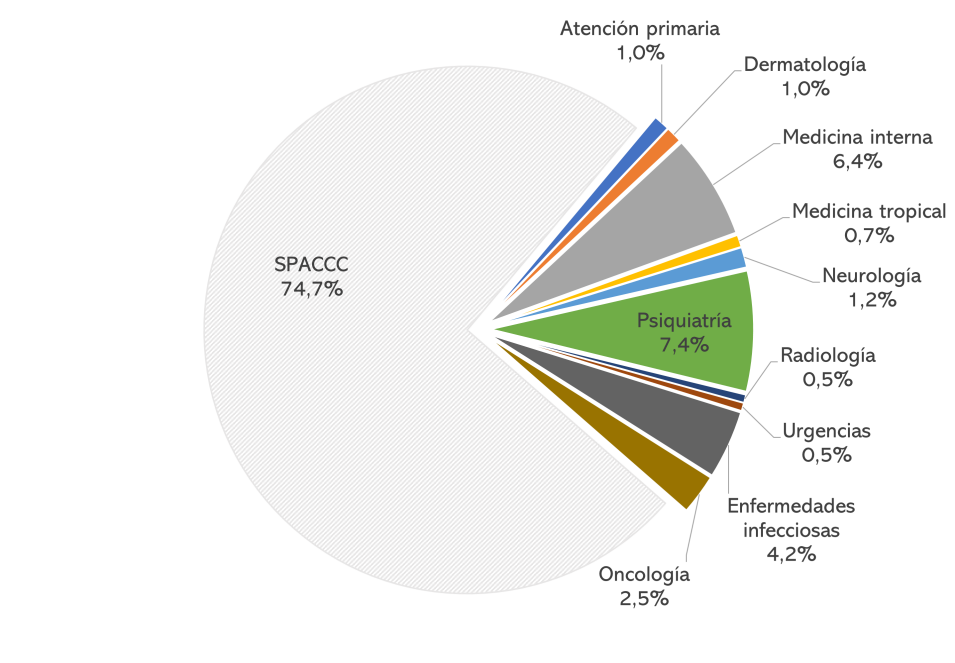

Figura 3.1: Procedencia y distribución de las notas del corpus utilizado. El gráfico y los porcentajes representan la distribución de las notas con respecto al total.

corpus de SPACCC (Intxaurrondo, 2018) y MEDDOPROF (Lima-López et al., 2021), y reanotadas manualmente con las clases VIH+ o VIH-, según estas notas pertenezcan a pacientes infectados con el virus o no, respectivamente. Estas notas son de procedencia heterogéneas en relación a la especialidad médica a la que refieren. Las procedentes de MEDDOPROF, incluyen este tipo de información, teniendo notas procedentes de medicina interna o psiquiatría, como ejemplo de las dos más numerosas. En cambio, la mayoría, pertenecientes a SPACCC, no vienen especificadas de origen. La Figura 3.1 recopila todas las especialidades que aparecen en el corpus, acompañadas del porcentaje que suponen con respecto al total de notas. En la figura, SPACCC hace referencia a las notas pertenecientes a este corpus, sin procedencia concreta.

A continuación, se va a proceder a realizar un estudio de la frecuencia de las palabras en el corpus según la clase (VIH+ o VIH-) a la que pertenezcan. Estas palabras han sido sometidas a un preprocesamiento que nos facilite su estudio. Este consistió en convertir todo el texto a minúsculas, eliminar signos de puntuación, eliminación de *stopwords* (palabras muy repetidas y carentes de significado propio que no dan información), y lematización,

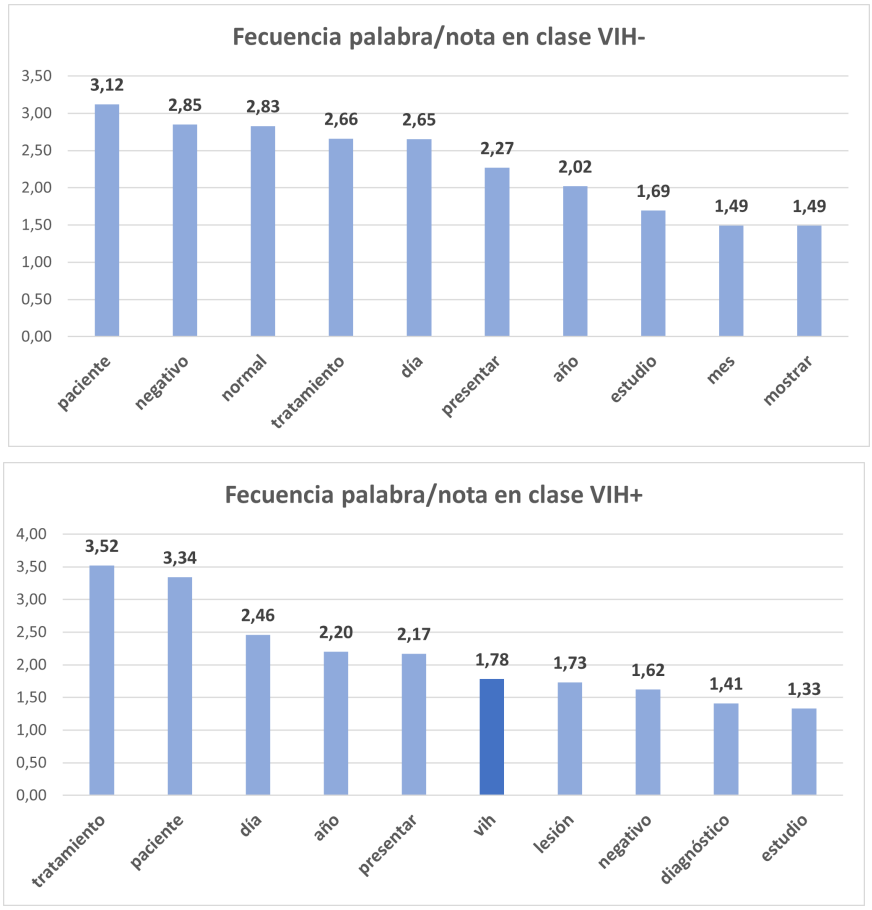

Figura 3.2: Estudio de frecuencia palabra/nota en notas con VIH negativo (VIH-) y positivo (VIH+). En azul más oscuro se indican las palabras relacionadas con la infección por VIH.

que consiste en convertir una palabra a su "lema" para agruparlas más fácilmente. El lema es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra, por ejemplo, el lema de un verbo suele ser su infinitivo.

El análisis de las notas negativas y positivas se presenta en la Figura 3.2, mostrando las 10 palabras más frecuentes por nota en el corpus, separado por la clase a la que pertenecen las notas. Como se ha comentado anteriormente, las palabras aparecen en forma de lemas. En ambas, vemos como predominan palabras que fácilmente se podrían relacionar con una nota clínica como "paciente" o "tratamiento". Como diferencias, más como cu-

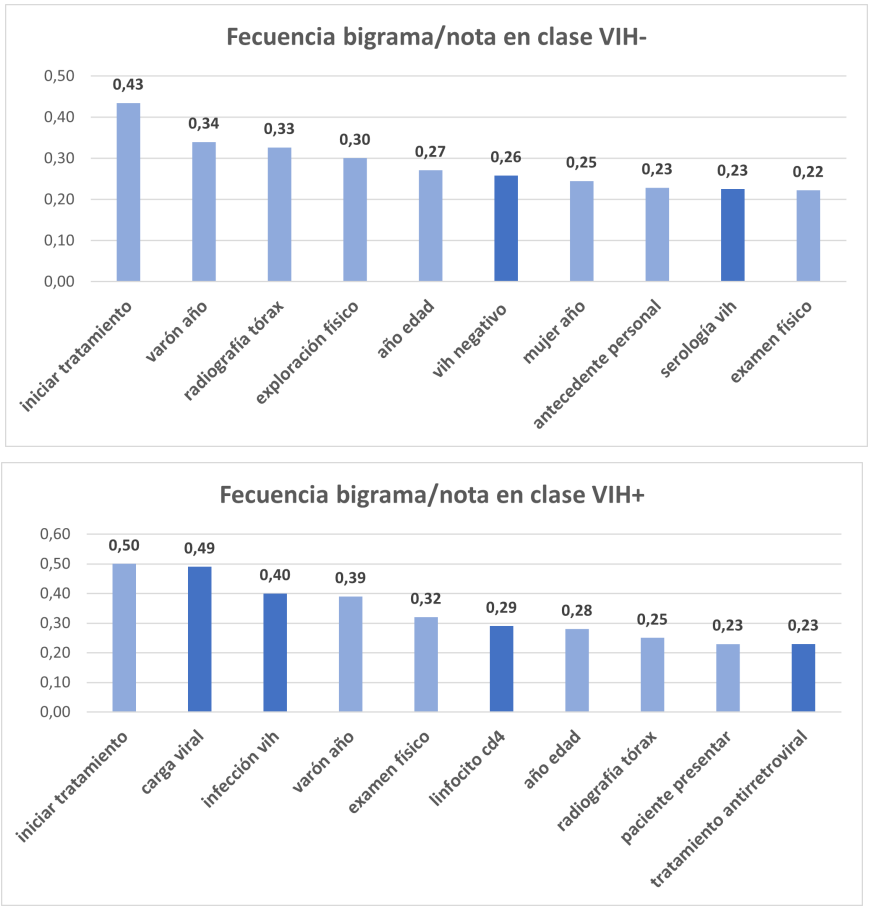

Figura 3.3: Estudio de frecuencia bigrama/nota en notas con VIH negativo (VIH-) y positivo (VIH+). En azul más oscuro se indican las palabras relacionadas con la infección por VIH.

riosidad que como algo relevante, vemos como "negativo" predomina mucho más en las notas negativas (2.85), que en las positivas (1.62). Cabe destacar la frecuencia de la palabra "VIH" en las notas positivas, apareciendo más de una vez por nota, algo entendible en un paciente infectado.

Además, se ha realizado otro análisis de frecuencias, pero, en este caso, estudiando pares de palabras (bigramas), para identificar algunas relaciones relevantes que se den en el texto. En la Figura 3.3, en referencia a las notas negativas, aparecen los pares "vih negativo" y "serología vih".

En el caso de la clase positiva, destacan los pares que hacen referencia a una infección por VIH: "carga viral", "linfocito cd4", "tratamiento an-

tirretroviral" y, por supuesto, "infección vih". Esto nos indica, que en un paciente infectado, los aspectos relacionados con el VIH ocupan la mayor parte de la información de las notas.

Como detalles a resaltar en ambas figuras, destaca la presencia de una prueba médica como es "radiografía tórax". Esto se debe a que esta prueba es una de las primeras que se les realiza a los pacientes sospechosos de padecer VIH (Estrada Chacón et al., 2002). Otro detalle sería la frecuencia que presenta el par "varón año", haciendo referencia a partículas que se repiten en muchas notas como "paciente varón de xx años". Destaca esta especialmente, pues es más frecuente que "mujer año" en ambos casos, pero aún más en el caso VIH+, donde este par tiene un 0,16 de frecuencia de aparición, e indicando la mayor presencia de esta infección por el virus en hombres.

3.1.2 Entidades

Durante la mayor parte de la experimentación del trabajo, sobre todo en la propuesta no supervisada (Capítulo 5), se han utilizado las entidades referentes a enfermedades extraídas por el sistema NER, en lugar de todo el contenido de las notas. Por tanto, también se va a realizar un análisis de las características que tienen estas entidades.

El corpus de entidades extraídas se compone de 12.421 entidades, de las cuales 6.192, algo menos de la mitad, son únicas. Las notas positivas tienen de media 27,8 entidades, mientras que las negativas tienen 31,5.

En este caso, se ha obviado el análisis de palabras individuales, pues la información que ofrecían no era tan diferente a la que daban las propias notas. En cambio, el análisis de los bigramas sí que ofrece algunos datos a resaltar. Ambos análisis se encuentran en la Figura 3.4.

En estas figuras, se ha querido resaltar en amarillo bigramas referentes a enfermedades o síntomas relacionados con la misma. Estos pueden ser síntomas comunes en ambas clases, como "pérdida (de) peso". Pero también, hay enfermedades mucho más relacionadas con el VIH, y sólo presentes en la clase positiva, como "candidiasis orofaríngea", "toxoplasmosis cerebral" o "tuberculosis pulmonar".

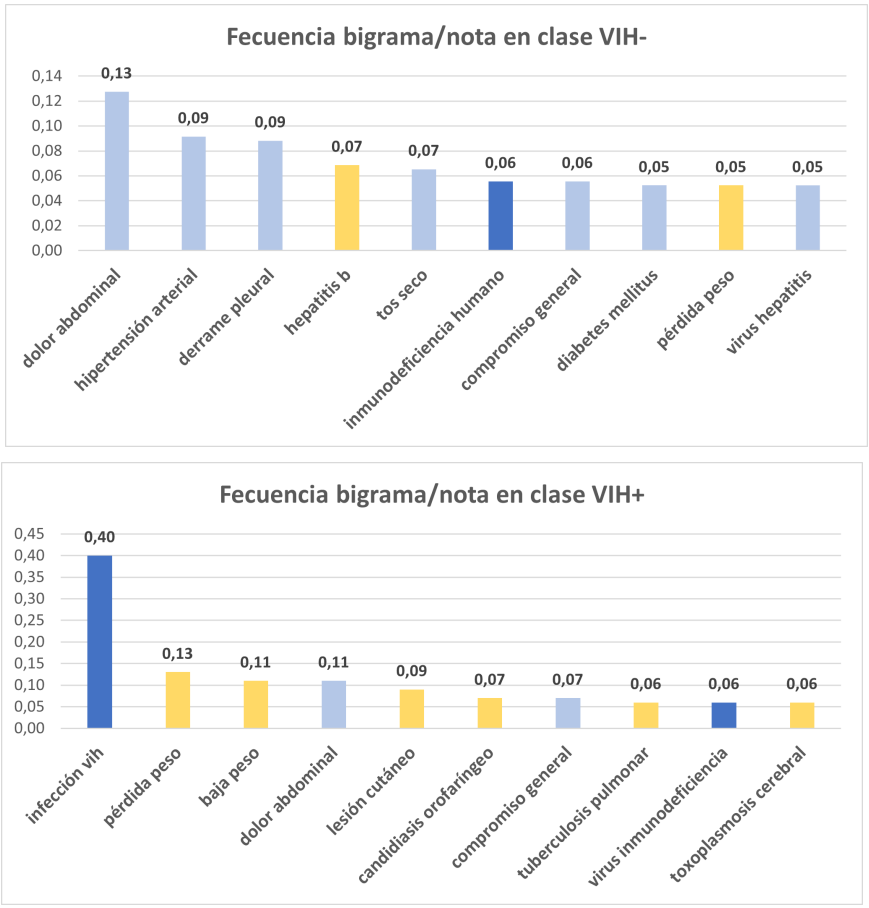

Figura 3.4: Estudio de frecuencia bigrama/nota en el conjunto de entidades con VIH negativo (VIH-) y positivo (VIH+). En azul más oscuro se marcan los bigramas referente a la infección por VIH, y en amarillo los referentes a enfermedades o síntomas relacionados con esta.

3.2 Metodología de evaluación

La salida de los sistemas propuestos consiste en una salida binaria, siendo 1 (positivo) si el sistema en cuestión clasifica una nota como que esta presenta riesgo de VIH, y 0 (negativo) en caso contrario. Al saber la verdadera clasificación de las notas, se comparan los resultados, y se valora esta clasificación como verdadera o falsa. Por último, se calculan las métricas descritas más adelante, y se realizan las comparaciones y valoraciones entre los experimentos y propuestas para elegir la mejor.

Habitualmente, en la literatura, para este tipo de problemas se utilizan

las métricas de precisión, sensibilidad y valor-F. La primera, mide la proporción de resultados clasificados correctamente. La segunda, también llamada exhaustividad o *recall*, mide la probabilidad de un resultado de ser correcto. Se decidió escoger el nombre de sensibilidad, pues parece el más acertado para el propósito que se le quiere dar al trabajo. Y la tercera métrica, es la media armónica que combina estos valores y sirve como valor único ponderado de ambos.

Estas métricas se sirven de los términos "verdadero positivo" (TP, del inglés *true positive*), "verdadero negativo" (TN, del inglés *true negative*), "falso positivo" (FP), y "falso negativo" (FN) para evaluar el resultado de la clasificación. Un VP significa que una persona enferma ha sido diagnosticada como tal, mientras que un FP significa que una persona sana ha sido clasificada erróneamente como enferma, lo contrario respectivamente para los casos de VN y FN.

Según estos términos, la precisión y la sensibilidad se definen en la Función 3.1 como:

$$Precision = \frac{VP}{VP + FP} \qquad Sensibilidad = \frac{VP}{VP + FN} \tag{3.1}$$

Mientras que el valor-F se define en la Función 3.2:

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Sensibilidad}{(\beta^2 \cdot Precision) + Sensibilidad} \tag{3.2}$$

siendo β un número real que pondera la importancia de cada métrica. En nuestro caso, el valor es 2, dando más valor a la sensibilidad que a la precisión. La elección del 2 como β radica en ser el valor comúnmente utilizado cuando se quiere dar más importancia a la sensibilidad frente a la precisión.

Además de conseguir una buena capacidad predictiva, nuestro trabajo persigue reducir el número de diagnósticos tardíos, por lo que se valora maximizar la sensibilidad del estimador, intentando rebajar lo máximo posible los FN, aunque sin descuidar la precisión.

En la Tabla 3.1 se muestra un ejemplo de las tablas que se van a utilizar a lo largo del trabajo para resumir los resultados de cada experimento. La primera columna contiene información variable que depende del experimento. Esta información puede ser alguna estrategia que se haya utilizado en ese experimento o unos tipos de datos concretos. La segunda columna, recopila el número de notas positivas correctamente clasificadas en la primera

Estrategias o Datos	VIH+	VIH-	Precisión	Sensibilidad	valor-F
	TP	FN			
	FP	TN			

Tabla 3.1: Tabla de ejemplo de evaluación.

fila (TP), y erróneamente clasificadas en la segunda (FP). Mientras que la tercera recopila el número de notas negativas erróneamente clasificadas en la primera fila (FN), y correctamente clasificadas en la segunda (TN). La últimas tres columnas contienen los valores de precisión, sensibilidad y el valor-F de cada experimento.

Capítulo 4

Herramientas y recursos

Este capítulo sirve para unificar y describir en profundidad las herramientas y materiales más relevantes que han sido utilizados en ambas propuestas.

4.1 Sistema NER

En esta sección, explicamos el desarrollo del sistema NER entrenado para el trabajo. Este sistema es capaz de identificar enfermedades y síntomas en notas clínicas, y reconocer si estas menciones están afirmadas o negadas. La arquitectura consiste en dos modelos NER independientes, uno de detección de enfermedades y síntomas, y otro de detección de negación, que actúan en paralelo. Ambos identifican sus entidades correspondientes y después se ponen en común para clasificar las entidades de enfermedades según su naturaleza positiva o negativa. El primer modelo, el modelo NER de enfermedades y síntomas, será nombrado como NER de enfermedades de ahora en adelante para acortar.

La arquitectura del sistema NER se muestra en la Figura 4.1. Se toma un texto, normalmente dividido en oraciones y se introduce al sistema. La primera parte, el NER de negación, detecta las partes del texto que implican un sentido negativo. "NEG" hace referencia al elemento que induce esta negación, y "NSCO" al alcance (*scope*) de la negación, que se entiende como la parte afectada por la negación anterior. La segunda parte, el NER de enfermedades, identifica que elementos en el texto se corresponden con una enfermedad, codificada como "DISO". Finalmente, el sistema compara si alguna de estas enfermedades forma parte de la negación, y de ser así las identifica como negativas ("NEG"). En caso contrario, se identifican como

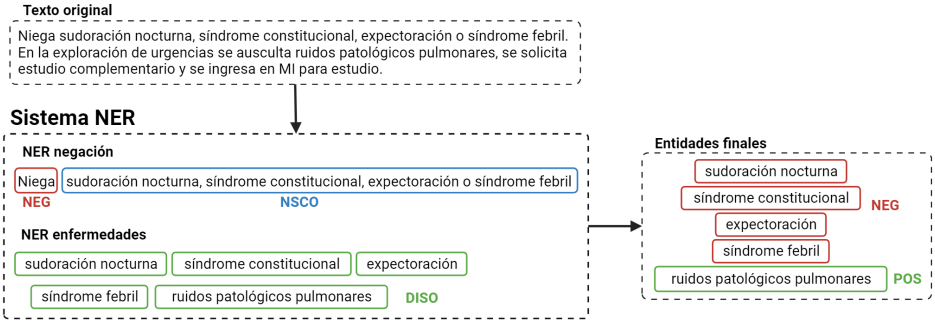

Figura 4.1: Arquitectura del sistema NER.

positivas ("POS"). La salida del sistema devuelve el conjunto de entidades referentes a enfermedades, acompañadas de una etiqueta que indica si son negativas o positivas.

Para construir este sistema necesitamos tres elementos fundamentales: un modelo adecuado, y dos corpora, cada uno especializado en su tarea, para entrenar con ellos. A continuación, exploraremos las decisiones tomadas para elegir este modelo y los corpus utilizados. Después, explicaremos como ha sido el entrenamiento y mostraremos la evaluación de los sistemas entrenados.

4.1.1 Modelos

La selección de un modelo preentrenado determina en gran medida el rendimiento de un sistema NER. La decisión puede depender de varios factores, entre los que destacan el idioma en el que ha sido entrenado y el tipo de datos. En nuestro caso, al trabajar con notas clínicas en español, necesitamos que el modelo haya sido entrenado con textos en español y, preferiblemente, utilizando notas clínicas, o al menos datos clínicos o del dominio biomédico.

Actualmente, el entrenamiento de grandes modelos en español ha sufrido un gran empujón gracias sobre todo al PlanTL[1], como se comentó anteriormente en la Sección 2.2. Antes de la irrupción de éstos ya se trabajaba con grandes modelos en español, como el modelo multilingüe mBERT (Devlin et al., 2018), o el modelo BETO (Cañete et al., 2020), aunque ambos modelos no pertenecen a un dominio específico, siendo de ámbito general.

Para nuestro interés, el PlanTL ha impulsado el entrenamiento de mode-

[1] Plan de Impulso a las Tecnologías del Lenguaje. Consultado el 4 de mayo de 2023.

los específicos para el dominio biomédico. Más concretamente, en (Carrino et al., 2022) se presentan los dos modelos más grandes entrenados en español en este dominio. Entre los corpus biomédicos con lo que fueron entrenados destacan el *Spanish Biomedical Crawled Corpus*[2] (Carrino et al., 2021), artículos de SciELO o resúmenes de PubMed, entre otros, suponiendo un total de 1,1 millardos de tokens de 2,5 millones de documentos.

De estos dos hay uno que sobresale por encima del otro para nuestro caso concreto, el modelo `bsc-bio-ehr-es`[3], que, además de haber sido entrenado con este corpus inmenso, ha sido entrenado con más de 514.000 informes y notas clínicas reales, que suponen 95 millones más de tokens. Además en (Carrino et al., 2022) se utiliza este modelo específicamente para el entrenamiento en varias tareas NER[4] [5], consiguiendo grandes resultados, por lo que parece un buen modelo sobre el que construir un sistema de este tipo.

4.1.2 Datasets

A continuación, se describen los datasets utilizados para el desarrollo de los modelos NER de detección de enfermedades y detección de la negación.

4.1.2.1 CT-EBM-SP

El dataset elegido para desarrollar el modelo NER de detección de enfermedades fue el dataset abierto CT-EBM-SP (Campillos-Llanos et al., 2021). Este dataset se compone de una colección de textos de *abstracts* de revistas y retrospectivas de ensayos clínicos publicados en PubMed y el repositorio SciELO, y anuncios de ensayos clínicos recopilados de EudraCT[6]. Las entidades anotadas pertenecen a cuatro grupos semánticos extraídos de UMLS (Bodenreider, 2004) referentes a patologías (codificadas como DISO), entidades anatómicas (ANAT), sustancias bioquímicas y farmacológicas (CHEM) y diagnósticos, procedimientos y test de laboratorio (PROC).

La razón principal de la elección de este dataset sobre otros que reconocen entidades clínicas, en concreto enfermedades, es que las entidades de este sistema se basan en UMLS. Utilizar la anotación de UMLS nos permite unificar criterios, tener una base muy fiable sobre la que trabajar y, además,

[2] Spanish Biomedical Crawled Corpus. Consultado el 23 de mayo de 2023.
[3] bsc-bio-ehr-es. *Hugging Face*. Consultado el 23 de mayo de 2023.
[4] bsc-bio-ehr-es-pharmacorner. *Hugging Face*. Consultado el 23 de mayo de 2023.
[5] bsc-bio-ehr-es-cantemist. *Hugging Face*. Consultado el 23 de mayo de 2023.
[6] EudraCT. Consultado el 8 de mayo de 2023.

ofrece sinergias con otras herramientas utilizadas en el trabajo, y que se comentan en la Sección 4.2.

Otra razón reside en el hecho de tener entidades separadas por categorías. Esto permite que las anotaciones sean más "concentradas", es decir, son entidades más cortas y concisas. En experimentos con otros datasets como DisTEMIST (Miranda-Escalada et al., 2022), nos dimos cuenta de que las entidades extraídas eran muy largas y complejas, haciendo esto que fueran más difíciles de normalizar y relacionar con conceptos de SNOMED-CT, en este caso. Por poner un ejemplo, en pruebas con DisTEMIST (más concretamente el sistema entrenado para la tarea por parte del grupo SINAI[7]) obteníamos entidades de enfermedades tales como "calcificaciones vasculares en ambos hipocondrios y en pelvis", mientras que nuestro sistema devuelve "calcificaciones vasculares" como enfermedad, e "hipocondrios" y "pelvis" como entidades anatómicas. Esto no critica la calidad de los propios corpus o sistemas, pero, entidades como las del segundo caso son más útiles para lo que se busca en este trabajo.

4.1.2.2 NUBes

En un dominio como el biomédico, identificar el correcto significado de una frase puede ser crucial para diferenciar entre si un paciente sufre una enfermedad o no. La detección de la negación es un problema complejo, sobre todo porque la propia negación no es un fenómeno trivial, sino que involucra aspectos sintácticos, morfológicos y semánticos. Esta complejidad hace que no existan una gran cantidad de corpus que anoten la negación en el dominio biomédico y en español. Más concretamente, existen ocho datasets en español con anotaciones de negación, de los cuales seis pertenecen al dominio biomédico, y de los que sólo los corpus IULA Spanish Clinical Record Corpus (IULA-SCRC) (Marimon, Vivaldi, y Bel Rafecas, 2017) y NUBes (Lima Lopez et al., 2020) están disponible de forma pública.

El dataset NUBes es el corpus biomédico más grande anotado con negación y especulación en español. Este e IULA+, versión de IULA-SCRC reanotado según los criterios de anotación de NUBes, añadiendo también especulación (Lima Lopez et al., 2020), fueron los corpus elegidos para entrenar nuestro sistema NER de detección de negación y especulación.

En detalle, este dataset se compone de 29,682 frases, de las cuales el

[7] chizhikchi/Spanish_disease_finder. *Hugging Face*. Consultado el 24 de mayo de 2023.

25.5% de ellas incluyen negación, y el 7.5% especulación. El corpus distingue entre cuatro tipos de entidades, que se pueden agrupar en dos tipos generales: señal, el elemento de la frase que modifica su significado, y alcance (*scope*), la parte de la frase afectada por esa señal. La negación y la especulación tienen su señal correspondiente, NEG y UNC, respectivamente, y un alcance, NSCO y USCO, para un total de cuatro tipos de entidades.

4.1.3 Entrenamiento y evaluación

En esta sección, se tratarán las decisiones de entrenamiento de cada uno de los modelos y se expondrán los datos de evaluación sobre sus respectivos conjuntos de test.

Ambos modelos han sido entrenados siguiendo la misma metodología, utilizando las herramientas del *framework* Flair (Akbik et al., 2019), con las que podemos hacer uso del método FLERT (Schweter y Akbik, 2020). Este, a diferencia de otras aproximaciones al NER que sólo hacen uso de representaciones a nivel de frase (*sentence-level*), demostró que este tipo de representación no es suficiente para obtener todo el contexto de una palabra, proponiendo un método que fuera capaz de incluir información a nivel de documento (*document-level*) en la representación. Para esto, añaden a la representación de la frase actual, información de la anterior y de la siguiente.

Los hiper-parámetros utilizados en el entrenamiento también fueron los mismos para ambos sistemas: 10 épocas, un tamaño de lote de 8 y una tasa de aprendizaje de $5e-5$. El tamaño de contexto, es decir, la cantidad de frase anterior y siguiente utilizada para entrenar fue de 64 tokens.

Las Tablas 4.1 y 4.2, recogen la evaluación sobre el conjunto de test tras el entrenamiento de ambos modelos. Las métricas de evaluación son las mismas que las descritas en la Sección 3.2. Las dos primeras columnas hacen referencia a la precisión y *recall* (sensibilidad), la tercera es el valor-F, pero en este caso $\beta = 1$, y soporte, hace referencia al número de entidades de cada clase en el conjunto de test de cada corpus.

En ambos casos, se puede ver que los resultados son bastante buenos, superando o acercándose, salvo en pocas ocasiones, al 0,9 en todas las clases. Las clases que salen peor paradas tienen mucha menor presencia en el dataset, lo que las acaba penalizando.

No obstante, al centrarse el trabajo en enfermedades, sólo se va a atender a tres de las clases presentes. En el caso de CT-EBM-SP, sólo se van a

Por clase:	precisión	recall	valor-F	soporte
PROC	0.8581	0.8811	0.8695	3364
DISO	0.8911	0.8908	0.8910	2472
CHEM	0.9091	0.9073	0.9082	1565
ANAT	0.8082	0.7468	0.7763	316
micro avg	0.8770	0.8840	0.8805	
macro avg	0.8666	0.8565	0.8612	7717
weighted avg	0.8770	0.8840	0.8804	

Tabla 4.1: Evaluación sobre el conjunto de test del sistema NER entrenado con el corpus CT-EBM-SP para la detección de enfermedades y otras entidades clínicas.

Por clase:	precisión	recall	valor-F	soporte
NEG	0.9579	0.9592	0.9586	1423
NSCO	0.8777	0.9042	0.8907	1325
USCO	0.7195	0.7825	0.7497	400
UNC	0.8405	0.8825	0.8610	400
micro avg	0.8859	0.9101	0.8978	
macro avg	0.8489	0.8821	0.8650	3548
weighted avg	0.8878	0.9101	0.8987	

Tabla 4.2: Evaluación sobre el conjunto de test del sistema NER entrenado con el corpus NuBES para la detección de negación y especulación.

identificar enfermedades (DISO), y en el caso de NuBES, sólo se va a prestar atención a la negación (NED y NSCO). Por lo que el rendimiento del sistema, se podría decir que está cerca o por encima del 0,9.

4.2 Otras herramientas

En esta sección, se hace referencia a otras herramientas más difíciles de organizar en una única categoría, por lo que se agrupan juntas a continuación.

4.2.1 HESML

HESML (Lastra-Díaz, Lara-Clares, y Garcia-Serrano, 2022) es una librería de software Java que implementa medidas de similitud semántica basadas en ontologías y modelos de *Information Content* (IC) basados en ontologías como WordNet[8], SNOMED-CT o MeSH[9]. Este recurso sirve como plataforma

[8] WordNet. *Princeton University*. Consultado el 27 de julio de 2023.
[9] MeSH. *National Library of Medicine (NIH)*. Consultado el 27 de julio de 2023.

de experimentación para medir similitud entre palabras/conceptos, especialmente adecuado para grandes experimentos por su eficiencia y escalabilidad.

La librería implementa un gran abanico de algoritmos para medir distancias semánticas, de los cuales se han escogido dos:

- **Medida Pedersen** (Pedersen et al., 2007). Esta medida utiliza vectores de contexto para representar dos conceptos. La medida de similitud se corresponde con el coseno del ángulo entre ellos. Es decir, la similitud entre dos conceptos c_1 y c_2 se calcula como:

$$Sim(c_1, c_2) = \frac{\vec{v_1} \cdot \vec{v_2}}{|v_1| \cdot |v_2|}$$

  donde $\vec{v_1}$ y $\vec{v_2}$ son los vectores de contexto correspondientes a c_1 y c_2, respectivamente.

- **Medida Nguyen and Al-Mubaid** (Nguyen y Al-Mubaid, 2006). Medida entre dos conceptos utilizando el LCS (Least Common Subsume) y el camino más corto entre ellos. Se denomina LCS al ancestro común más cercano entre dos conceptos. Formalmente, la similitud entre dos conceptos c_1 y c_2 se calcula como:

$$Sim(c_1, c_2) = \log_2 \left([L(c_1, c_2) - 1] \times [CSpec(c_1, c_2)] + 2 \right)$$

  donde:

$$CSpec(c_1, c_2) = D - depth\left(L(c_1, c_2)\right)$$

  siendo $L(c_1, c_2)$ el camino más corto entre c_1 y c_2, $depth\left(L(c_1, c_2)\right)$ es la profundidad de $L(c_1, c_2)$ contando los nodos, y D la profundidad máxima dentro de la taxonomía.

4.2.2 MedLexSp

MedLexSp (Medical Lexicon for Spanish)[10] es un lexicón unificado de términos médicos en español con información lingüística y semántica (Campillos-Llanos, 2019). Contiene una serie de términos médicos mapeados a sus respectivos términos de UMLS y CUIs. En la Figura 4.2 se muestra un ejemplo del lexicón con distintas entradas[11]. Estas entradas se distinguen por un

[10] MedLexSp. *DIGITAL.CSIC*. Consultado el 24 de mayo de 2023.
[11] NLPMedTerm. Consultado el 24 de mayo de 2023.

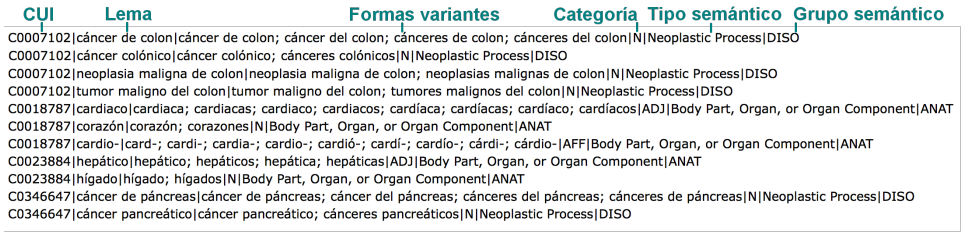

Figura 4.2: Ejemplo de entradas en MedLexSp. Cada entrada en el lexicón pertenece a uno de los posibles nombres de un concepto de UMLS. Las entradas se distinguen por un código único CUI, un nombre (lema), posibles variantes de este (formas variantes), categoría (léxica), tipo semántico y grupo semántico al que pertenecen.

código único CUI, un nombre (lema), posibles variantes de este (formas variantes), categoría léxica del concepto, tipo semántico dentro de UMLS y grupo semántico al que pertenecen según la misma clasificación que en la Sección 4.1.2.1.

Capítulo 5

Propuesta no supervisada

En este capítulo se presentan varias propuestas no supervisadas para clasificar notas clínicas como posible riesgo de VIH o no. Para ello nos valemos del conocimiento experto ofrecido por médicos especialistas en enfermedades infecciosas, combinando la representación basada en entidades explicada en la Sección 3.1.2 y algoritmos basados en reglas. Estas propuestas serán evaluadas y comparadas en busca de la mejor solución posible.

5.1 Introducción

Para esta primera propuesta partimos del conocimiento experto de una serie de médicos para identificar determinados criterios o indicadores en las notas clínicas, que denominaremos a partir de aquí IC, que nos pueden hacer sospechar de una posible infección por VIH. Entre estos criterios se incluyen enfermedades definitorias e indicadoras de VIH, enfermedades o condiciones en las que se espera cierta prevalencia de infección por VIH, síntomas de infección aguda por VIH, enfermedades de transmisión sexual (ETS), procedencia de países con más prevalencia de infección por VIH, categorías de edad según prevalencia, resultados de laboratorio con alteraciones presentes en la infección por VIH, comportamientos de riesgo, y embarazo, pues un embarazo implica siempre la realización de una prueba serológica.

Ante esta cantidad de indicadores, hemos querido fijar los criterios a tener en cuenta en el trabajo sólo en las enfermedades y síntomas, obviando el resto. La decisión se basa en abordar en un principio un problema más concreto, reducido y sencillo de evaluar. Las posibles consecuencias de esta decisión serán abordadas con mayor profundidad en el Capítulo 7.

Las enfermedades se dividen en tres subgrupos que se definen como:

- **Enfermedades definitorias:** enfermedades que aparecen al final de la infección por VIH y que implican un diagnóstico tardío de la enfermedad. Es "obligatorio" que todo paciente con estas enfermedades tenga realizada la serología de VIH. Como ejemplo, entre estas podemos encontrar la neumonía por *Pneumocystis jirovecii*, que en el periodo 2012-2021 es la enfermedad definitoria de sida más frecuente (30,8% de los casos) (DCVIHT, 2022).

- **Enfermedades indicadoras:** los pacientes con estas enfermedades tiene una mayor probabilidad de tener VIH, aunque las enfermedades pueden ser por causas diferentes al VIH.

- **Otras enfermedades:** enfermedades en las que el VIH tiene una prevalencia $> 0,1\%$. Es decir, enfermedades que aparecen en una proporción mayor del 0.1% de los pacientes con VIH. La inclusión de estas enfermedades da la oportunidad de detectar a pacientes en estadios más precoces de la infección.

Por su parte, los síntomas estudiados se refieren a síntomas que aparecen en caso de una infección aguda por VIH. Es decir, son síntomas cuya existencia en un paciente aumenta la probabilidad de infección, y cuya inexistencia la disminuye, cada uno de ellos en mayor o menor medida. En caso de que estos síntomas aparezcan, puede servir de apoyo para la decisión a elegir por el sistema.

En el Anexo A se presentan las tablas con todas las enfermedades y síntomas aportados por los médicos que se han tenido en cuenta en este trabajo. Acompañando a cada concepto, se incluyen las valoraciones numéricas dadas por éstos en función de su sensibilidad y especificidad en una posible infección por VIH. Este valor es la media de las valoraciones dadas por 5 doctores, puntuando cada concepto con un número real entre 0 y 5, ambos incluidos.

La arquitectura general del sistema desarrollado para la propuesta no supervisada se muestra en la Figura 5.1 y consiste en lo siguiente:

1. Extraer todas las entidades afirmativas referentes a enfermedades y síntomas utilizando el sistema NER entrenado con anterioridad, explicado con detalle en la Sección 4.1. Las entidades negativas no son tenidas en cuenta en esta propuesta.

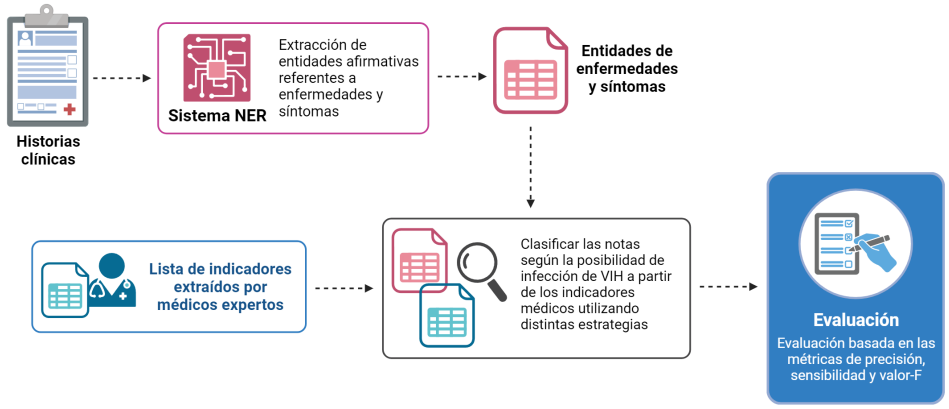

Figura 5.1: Arquitectura de la propuesta no supervisada.

2. A partir de los IC y utilizando distintas estrategias, clasificar las notas de forma binaria en relación a la posibilidad de infección por VIH o no.

3. Evaluación de la clasificación obtenida utilizando la metodología explicada en la Sección 3.2.

Propuestas

A continuación, detallaremos las distintas propuestas. La primera (Sección 5.2), consiste en una propuesta "Baseline" que nos sirva para tener un punto de partida sobre el que comparar. La segunda (Sección 5.3), propone dos aproximaciones para solucionar el problema de la normalización de entidades comentado en la Sección 2.2.2.2, utilizando el lexicón MedLexSp. Y la tercera (Sección 5.4), propone una solución basada en la similitud semántica entre términos dentro de la ontología de SNOMED-CT, utilizando la librería HESML (Lastra-Díaz, Lara-Clares, y Garcia-Serrano, 2022).

Estrategias

A lo largo de las tres propuestas utilizaremos dos estrategias distintas. A la primera la denominaremos estrategia "Coincidencia". En esta, una nota se clasifica como positiva si el nombre de cualquiera de las entidades extraídas coincide con alguno de los IC. Para la segunda, además de los nombres, contamos con unos valores (puntuación) asignados por los médicos a cada una de ellas en función de su sensibilidad y especificidad. Estos valores van del 0 al 5, y aunque podrían ser discutibles, nos permiten tener un dato fiable

Texto	Estrategia (Puntuación)	Clasificación
"La serología para toxoplasmosis fue positiva... inició tratamiento empírico como una **toxoplasmosis cerebral**..."	Coincidencia	VIH+
	Pesos (4,8)	
"Sin presentar mejoría consultó nuevamente; se diagnosticó **enfermedad pélvica inflamatoria**, la cual se manejó con antibióticos"	Coincidencia	VIH+
	Pesos (3,3)	VIH-

Tabla 5.1: Ejemplo de clasificación con la propuesta no supervisada, separando las dos estrategias utilizadas. La estrategia "Pesos" se acompaña de la puntuación asociada a la entidad identificada en negrita en el texto. La clasificación es VIH positivo (VIH+) y VIH negativo (VIH-).

con el que medir la relevancia de cada uno de los criterios. Para tener en cuenta estos valores, planteamos una nueva estrategia denominada "Pesos". Cada nota se evalúa mediante una puntuación. Esta empieza en 0, y por cada entidad extraída de ella que se corresponda con uno de los IC (misma estrategia que la anterior), se sumará su valor a la puntuación de la nota. Finalmente, superar un umbral dado determinará si la nota se clasifica como positiva o no.

En la Tabla 5.1 se muestra un ejemplo de ambas estrategias. La primera columna contiene un texto de muestra con las entidades extraídas en negrita. En la segunda columna se indica la estrategia con la que ha sido clasificada cada uno, con la puntuación atribuida a la nota en la estrategia pesos. En la última, la clasificación resultante, VIH positivo (VIH+) y VIH negativo (VIH-). En el caso de la estrategia "Pesos", el umbral escogido es 4,2, decisión que se explicará en los experimentos siguientes.

5.2 Propuesta "Baseline"

Este primer sistema simplemente comprueba si alguna de las entidades extraídas por nuestro sistema NER coincide con el nombre de alguno de los indicadores (enfermedades o síntomas). Se realizan experimentos utilizando distintos elementos de preprocesado: convertir a minúsculas, eliminación de símbolos, tildes, caracteres especiales, *stopwords* y *stemming* (convertir las palabras a su raíz). A la combinación de todas las operaciones la denominaremos "preprocesamiento completo" de aquí en adelante. Además, experimentaremos teniendo en cuenta sólo las enfermedades, en conjunto o separadas por grupo, sólo los síntomas, o combinando enfermedades y síntomas, para comprobar su relevancia en la clasificación.

Estrategia	VIH+	VIH-	Precisión	Sensibilidad	valor-F
sin preprocesado	20 13	80 293	0,606	0,200	0,231
cm	29 18	71 288	**0,617**	0,290	0,324
cm + sc&sp	31 25	69 281	0,554	**0,310**	**0,340**
preprocesado completo (cm + sc&sp + stm)	31 26	69 280	0,544	**0,310**	**0,340**

Tabla 5.2: Resultados del sistema propuesto como Baseline, sólo teniendo en cuenta las enfermedades indicadoras, utilizando la estrategia "Coincidencia". Cada fila indica las operaciones de preprocesamiento realizadas: convertir a minúsculas (cm), quitar caracteres especiales y stopwords (sc&sp), y preprocesamiento completo, que añade stemming (stm).

5.2.1 Experimentos de preprocesado

Las primeras pruebas van en la dirección de decidir el mejor preprocesado para utilizar en el resto de los experimentos. En este primer experimento sólo tenemos en cuenta a las enfermedades para la clasificación, y se utiliza la estrategia "Coincidencia". Los resultados se presentan en la Tabla 5.2. Primero, podemos apreciar que, a medida que se añaden elementos de preprocesado, aumenta el número de notas en las que aparece una entidad dentro de los IC (notas clasificadas como positivas, TP y FP), pasando de 33 (20+13) sin preprocesado, a 57 (31+26) cuando se realizan todas las operaciones. La proporción de enfermedades identificadas en las notas positivas (31%), es muy superior a la de las notas negativas (0,085%).

A medida que aumentan las entidades identificadas como IC, aumenta la sensibilidad y disminuyen los FN. Esta mayor sensibilidad, provoca también que la precisión vaya disminuyendo. El mejor valor-F se obtiene con el preprocesado completo con y sin *stemming*, con una diferencia apenas anecdótica en la precisión.

El siguiente experimento es igual que el anterior, pero teniendo en cuenta esta vez sólo los síntomas. Los resultados se pueden ver en la Tabla 5.3. En este caso no hay ninguna variación en los resultados entre las entidades con y sin preprocesado, excepto cuando interviene el *stemming*. El preprocesado sin *stemming* en este caso no es tan relevante como en la Tabla 5.2, porque las entidades de síntomas son mucho más cortas y fáciles de identificar, y además no suelen escribirse en mayúsculas (a diferencia de muchos nombres de enfermedades), por eso estas operaciones no surgen efecto.

Estrategia	VIH+	VIH-	Precisión	Sensibilidad	valor-F
sin preprocesado	36 90	64 216	0,286	0,360	0,342
cm	36 90	64 216	0,286	0,360	0,342
cm+sc&sp	36 90	64 216	**0,286**	0,360	0,342
preprocesado completo (cm+sc&sp+stm)	39 112	61 194	0,258	**0,390**	**0,354**

Tabla 5.3: Resultados del sistema propuesto como Baseline, sólo teniendo en cuenta los síntomas indicadores, y utilizando la estrategia "Coincidencia". Cada fila indica las operaciones de preprocesamiento realizadas: convertir a minúsculas (cm), quitar caracteres especiales y stopwords (sc&sp), y preprocesamiento completo, que añade stemming (stm).

No obstante, el *stemming* sí es relevante, pues muchos de estos síntomas se escriben indistintamente en plural o singular en multitud de casos (e.g. adenopatía/adenopatías).

Viendo los resultados, se opta por escoger el preprocesamiento completo como la combinación de operaciones a utilizar en los siguientes experimentos. Mientras que en los resultados con enfermedades no era diferencial el *stemming*, en el caso de los síntomas sí, por lo que se escoge como estándar de ahora en adelante.

5.2.2 Experimentos sobre los IC

Una vez decidido el preprocesamiento, para el tercer experimento se quería comprobar cuál era la incidencia de cada grupo de enfermedades (definitorias, indicadoras y otras) por separado. La Tabla 5.4 recoge los resultados de este experimento, donde el número de notas con enfermedades definitorias (23+7=30) dobla al de los otros dos grupos de enfermedades. Esto puede deberse a varios factores, o bien el sistema NER es capaz de identificar mejor esas enfermedades en concreto, o bien los datos contienen muchas más de estas que del resto. Esto puede verse en el análisis del conjunto de datos de la Sección 3.1.2, donde dos de los bigramas más relevantes ("tuberculosis pulmonar" y "toxoplasmosis cerebral"), pertenecen a este grupo.

Diferenciando por clases, podemos ver que las enfermedades definitorias e indicadoras están mucho más presentes en las notas positivas que en las negativas, así como ocurre al contrario en el caso de las "otras". Lo primero es natural, pues las enfermedades definitorias son enfermedades que apare-

Datos	VIH+	VIH-	Precisión	Sensibilidad	valor-F
Enfermedades definitorias	23 7	77 299	**0,767**	0,230	0,267
Enfermedades indicadoras	12 7	88 299	0,632	0,120	0,143
Otras enfermedades	2 13	98 293	0,133	0,020	0,024
Enfermedades + síntomas	56 129	44 177	0,303	**0,560**	**0,479**

Tabla 5.4: Resultados del sistema propuesto como Baseline, sólo teniendo en cuenta las enfermedades dentro de los IC, separadas en sus grupos correspondientes, y utilizando la estrategia "Coincidencia". Cada fila es un experimento distinto según el grupo de enfermedades estudiado.

cen en estados avanzados de la infección, y las indicadoras tienen una alta presencia en personas infectadas. Pero lo segundo es especialmente reseñable pues, la inclusión de este último grupo de enfermedades responde al hecho de intentar detectar pacientes en estadios más precoces de la infección. Esto será comentado más a fondo en el Capítulo 7.

Por último, se añade en la Tabla 5.4 un apartado para recoger de forma conjunta cómo sería la clasificación si se combinasen enfermedades y síntomas. Esto nos indica que tenemos un total de 185 (56 + 129) notas que contienen alguna entidad (enfermedad o síntoma) dentro de los IC. La combinación provoca un aumento importante de la sensibilidad, aunque con una reducción muy significativa de la precisión. No obstante, esta sensibilidad tan alta provoca también un valor-F muy elevado, quedando cerca del 0,5.

Esta reducción en la precisión se debe al gran número de notas negativas que presentan algún síntoma. Esto es normal, pues los síntomas son indicadores de la infección, pero no únicos de esta. Por tanto, es comprensible que en unas notas de pacientes que acuden a consulta aparezcan síntomas tan comunes como fiebre o diarrea.

Se puede apreciar también, en conjunto con los datos de las Tablas 5.2 y 5.3, que, mientras que el aporte a la clasificación positiva es parejo entre enfermedades y síntomas, en el caso negativo es muy superior la aportación de los síntomas.

Los resultados de los experimentos con la estrategia "Pesos" están recogidos en la Tabla 5.5. Para asignar pesos, las entidades han sido preprocesadas completamente, y el umbral elegido es de 4,2. Este valor corresponde a la

Datos	VIH+	VIH-	Precisión	Sensibilidad	valor-F
Enfermedades	25 8	75 298	**0,758**	0,250	0,289
Síntomas	9 28	91 278	0,243	0,090	0,103
Enfermedades + síntomas	31 41	69 265	0,431	**0,310**	**0,328**

Tabla 5.5: Resultados del sistema propuesto como Baseline, teniendo en cuenta enfermedades y síntomas, utilizando la estrategia "Pesos". Cada fila indica un experimento distinto según si son enfermedades o síntomas, combinándolos en la última.

puntuación mínima adjudicada por los médicos a una de las enfermedades definitorias ("neumonía recurrente", concretamente), y sabiendo que estas enfermedades implican un DT, se escoge este valor como umbral, también en futuros experimentos.

En los experimentos se puede apreciar la relevancia muy superior que tienen las enfermedades sobre los síntomas. De los 31 TP de la Tabla 5.2, 25 de ellos superan el umbral en este caso, mientras que de los 39 TP de la Tabla 5.3 sólo suman lo suficiente 9 de ellos.

Se puede ya afirmar que el enfoque de los pesos cumple su función de expresar la distinta relevancia de los criterios, permitiendo una clasificación más acertada en referencia a los niveles de importancia dados en el conocimiento experto. No obstante, en esta prueba baseline, al detectar tan pocas entidades este filtro extra resta de mucha sensibilidad al sistema.

Finalmente, se procede a elegir el mejor experimento *baseline* que posteriormente será comparado con el resto de mejores propuestas. El mejor resultado es el conseguido en la Tabla 5.4 con la estrategia "Coincidencia", combinando enfermedades y síntomas, con una precisión de 0,303, una sensibilidad de 0,560 y un valor-F de 0,479.

5.3 Propuesta con normalización basada en distancia entre palabras

Para esta segunda propuesta no supervisada se abordará el problema de la normalización utilizando un enfoque basado en reglas. Durante la experimentación, nos dimos cuenta de que varias de las entidades extraídas por el sistema NER no eran identificadas como IC por no tener la capacidad

Datos	VIH+	VIH-	Precisión	Sensibilidad	valor-F
Enfermedades	41 36	59 270	**0,532**	0,410	0,430
Síntomas	64 185	36 121	0,257	0,640	0,493
Enfermedades + Síntomas	76 198	24 108	0,277	**0,760**	**0,564**

Tabla 5.6: Resultados del sistema propuesto con normalización, teniendo en cuenta enfermedades y síntomas, utilizando la estrategia "Coincidencia". Cada fila indica un experimento distinto según si los IC utilizados son enfermedades o síntomas, combinándolos en la última.

de normalizar las entidades, ya sea porque estas contienen alguna falta ortográfica, se puedan escribir de varias maneras, o porque contienen alguna palabra dentro, como puede ser un adjetivo (e.g. enfermedad leve, enfermedad previa), y no permite que el sistema identifique correctamente la entidad.

Revisando la literatura, varias tareas de evaluación de NER y normalización de entidades, y las herramientas disponibles, no encontramos ningún método que solucionase todos los problemas encontrados de una manera completa, sin requerir un trabajo extra, difícilmente realizable por falta de tiempo. En este momento, nos encontramos con el lexicón MedLexSp, del que se habla en más en profundidad en la Sección 4.2.2. Este nos permite disponer de una gran colección de términos de UMLS, con multitud de sinónimos, acrónimos y variaciones, y sus respectivos códigos CUI (*Concept Unique Identifier*), que permiten unificarlos. Además, incluye un lematizador especialmente entrenado en terminología de UMLS, muy útil para la normalización. Sabiendo que las soluciones basadas en reglas pueden ser buenas soluciones en un dominio controlado, se decidió optar por esta vía.

El primer experimento realizado consiste en normalizar cada una de las entidades de las que disponemos y relacionarla con su respectivo CUI. Para ello, utilizamos el preprocesado completo y, además, añadimos la lematización que nos ofrece MedLexSp. De las 12.421 entidades extraídas, quedan después de la normalización, 7.257 entidades enlazadas con sus CUI. Eso quiere decir que se pierden en este proceso el 41,6% de las entidades, un número muy relevante. De estas, 793 entidades se identifican dentro de los IC.

Los resultados se recogen en la Tabla 5.6. La estrategia utilizada para

el experimento en este caso es la estrategia "Coincidencia". Como se puede observar en la misma, el número de total de notas con entidades dentro de los criterios aumenta en gran medida, alcanzando las 276 (76+200) notas en el último experimento. Eso quiere decir que, aunque se han perdido muchas entidades, las sospechas acerca de la normalización eran ciertas y se estaban obviando muchas entidades relevantes.

En comparación con el baseline, se puede ver que, en el caso de las enfermedades, el valor-F es superior a causa de una mayor sensibilidad del sistema. Un aumento de 0,088 en el valor-F y de 0,1 en la sensibilidad. El aumento en la detección de síntomas, aún más apreciable que en el caso de las enfermedades, hace que la clasificación conjunta tenga mejores resultados.

5.3.1 Distancia Levenshtein

No obstante, la propuesta de normalización anterior sólo arregla el problema de identificar distintas variantes de un mismo concepto, pero todavía queda algún problema en relación con faltas ortográficas o la aparición de palabras extra en las entidades. En (Rojas et al., 2022) se propone el uso de la distancia Levenshtein (Levenshtein, 1966) para normalizar las entidades. Esta distancia es un algoritmo clásico dentro del PLN, y calcula el número mínimo de operaciones de edición requeridas para transformar una cadena de caracteres en otra, por eso también se le conoce como "distancia de edición". Estas operaciones son a nivel de carácter y son tres: eliminación, inserción y sustitución. Todas suelen tener el mismo valor, excepto en una variante llamada Indel, donde se penaliza la sustitución valiendo el doble.

El problema es que el algoritmo es costoso, pues este debe de aplicarse un número $n \times m$ de veces, siendo n el tamaño del diccionario de referencia, en este caso MedLexSp, y m el número de entidades. Teniendo en cuenta los órdenes de magnitud de ambos, la operación sería imposible. No obstante, se decidió hacer una pequeña selección de conceptos del lexicón para tener sólo en cuenta las enfermedades y síntomas de los IC. Además, nos percatamos de que varias de las enfermedades de los criterios no se encontraban dentro del lexicón, por lo que algunas fueron añadidas a mano. Estas enfermedades añadidas se pueden encontrar en el Anexo B. La selección final consiste en 551 entradas (369 de enfermedades y 182 síntomas) de MedLexSp, con un total de 1.345 variantes.

Para implementar la distancia Levenshtein nos hemos valido de la librería

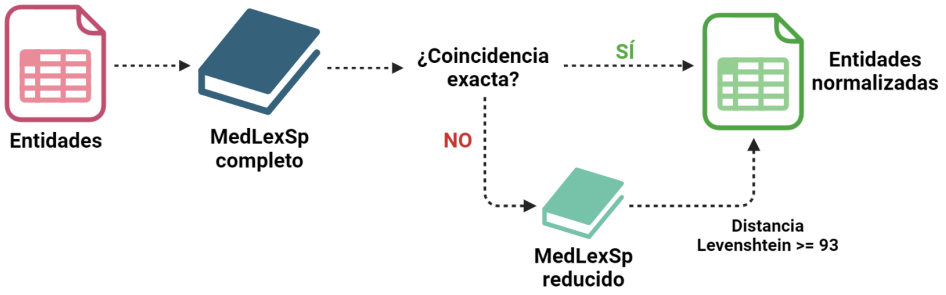

Figura 5.2: Proceso de normalización con la distancia Levenshtein.

`rapidfuzz`[1]. La función exacta aplicada se llama `WRatio`[2], y calcula la distancia Indel[3] entre dos entidades ponderando entre el número de palabras que coincidan entre ellas y la posición de estas. Esta selección de funciones son las que mejores resultados nos han ofrecido en las pruebas realizadas.

Para normalizar las entidades, primero, son sometidas a un preprocesamiento completo y son comparadas con todo el lexicón de MedLexSp, como en la primera normalización. En caso de que una entidad no tenga una coincidencia exacta, es comparada utilizando Levenshtein con el lexicón reducido en busca de la mayor semejanza posible. Si esta coincidencia supera un umbral, es escogida como válida. Este proceso se explica de manera gráfica en la Figura 5.2.

Entre los ejemplos de entidades que se consiguen normalizar en este caso destacan: entidades con adjetivos calificativos (e.g. "neumonía grave por P. jirovecii" tiene un 95 de distancia con "neumonía por P. jirovecii" después de preprocesarlas) o entidades con faltas de ortografía (e.g. "jirovesi" o "jiroveci" con "jirovecii", o "carinni" con "carinii").

Otros ejemplos de distancias serían "citomegalovirus" (normalmente referente a una prueba química) y "citomegalovirosis" (infección por citomegalovirus) tienen un distancia de 87,5, o "neumonía extrahospitalaria" y "neumonía intrahospitalaria" con un 92,31. Estableciendo el umbral de corte en 93, se pierden algunas normalizaciones a priori interesantes (e.g.

[1] RapidFuzz documentation. Consultado el 22 de mayo de 2023.
[2] RapidFuzz.Fuzz.WRatio. Consultado el 3 de julio de 2023.
[3] RapidFuzz.Indel. Consultado el 22 de mayo de 2023.

Datos	VIH+	VIH-	Precisión	Sensibilidad	valor-F
Enfermedades	45 45	55 261	**0,500**	0,450	0,459
Síntomas	64 186	36 120	0,256	0,640	0,492
Enfermedades + Síntomas	76 201	24 105	0,274	**0,760**	**0,561**

Tabla 5.7: Resultados de la propuesta con normalización de entidades, utilizando Levenshtein. Esta prueba sólo tiene en cuenta las enfermedades, utilizando la metodología "Coincidencia". El valor de umbral indica la distancia mínima entre los conceptos en cada experimento.

"adenopatía" tiene una distancia de 90 con "adenopatía periférica"), pero también se evitan los errores anteriores. Esto será abordado con más atención en la Sección 7.2.2.

La colección final contiene 7.304 entidades. Esto se corresponde con un aumento de 47 entidades con respecto a las 7.257 de la primera normalización, de las cuáles 840 son identificadas como entidades dentro de los IC. El aumento no es muy significativo en cantidad total, pero todas las entidades añadidas intervienen en la clasificación, lo que las hace importantes.

Resultados

Los resultados de la Tabla 5.7 recogen los experimentos realizados con los datos normalizados teniendo en cuenta enfermedades, síntomas y ambos en combinación, utilizando la estrategia "Coincidencia". Comparando con los resultados de la Tabla 5.6, la única mejora se da en el grupo de enfermedades, donde se identifican 4 notas positivas más. Esto no es baladí, pues persigue uno de los principales objetivos del trabajo que es disminuir los FN, por tanto, la estrategia se puede considerar exitosa.

El mayor cambio está en la clasificación errónea de las notas negativas, que hace bajar ligeramente la precisión en todos los grupos. Esto provoca que disminuya ligeramente el valor-F con respecto a la normalización anterior en el caso de la combinación de enfermedades y síntomas. Además, se puede apreciar que esta normalización no afecta en gran medida a los síntomas, manteniéndose igual.

El siguiente experimento, recogido en la Tabla 5.8, separa las enfermedades en grupos para comprobar cómo ha afectado la normalización en

Datos	VIH+	VIH-	Precisión	Sensibilidad	Valor-F
Enfermedades definitorias	38 17	62 289	**0,691**	**0,380**	**0,418**
Enfermedades indicadoras	21 11	79 295	0,656	0,210	0,243
Otras enfermedades	2 19	98 287	0,095	0,020	0,024

Tabla 5.8: Resultados de la propuesta con normalización de entidades utilizando la distancia Levenshtein. Esta prueba separa las enfermedades en grupos, y utiliza la metodología "Coincidencia".

Datos	VIH+	VIH-	Precisión	Sensibilidad	Valor-F
Enfermedades	36 18	64 288	**0,667**	0,360	0,396
Síntomas	41 97	59 209	0,297	0,410	0,381
Enfermedades + Síntomas	60 114	40 192	0,345	**0,600**	**0,523**

Tabla 5.9: Resultados de la propuesta con normalización de entidades, utilizando Levenshtein. Esta prueba tiene en cuenta todos los IC, utilizando la metodología "Pesos".

cada uno. Lo primero a destacar es que la distribución de entidades por grupo se mantiene con respecto a la propuesta Baseline. El número de entidades crece en todos los grupos, demostrando que la normalización es clave para una correcta detección de las enfermedades. Destaca otra vez, aún más esta vez por crecer en número, la superioridad de las "Otras enfermedades" en las notas negativas.

También como en la propuesta Baseline, se quería comprobar la influencia de los pesos en esta clasificación, sobre todo por el hecho de tener más entidades con las que trabajar en este caso. En la Tabla 5.9 se encuentran los resultados de estos experimentos, donde cada fila indica el grupo de enfermedades que se ha tenido en cuenta.

5.4 Propuesta basada en distancia semántica

La última propuesta propone utilizar la distancia semántica entre términos de una ontología. La distancia semántica se puede medir de varias maneras, pero en este caso se va a medir utilizando una ontología como es SNOMED-CT. La distancia semántica en una ontología mide la similitud entre dos

términos mediante la cercanía que estos tengan dentro de una jerarquía. La razón de esto viene de aprovechar los términos de UMLS obtenidos mediante la normalización con MedLexSp, para encontrar posibles relaciones entre las entidades y los IC dentro de SNOMED-CT. Es decir, en lugar de centrarnos sólo en las menciones exactas a los IC, encontrar otras posibles entidades que por cercanía semántica, nos den una posible relación con estos, que nos haga sospechar de una infección. Para esto, nos hemos valido de la librería HESML, explicada más en profundidad en la Sección 4.2.1.

Los datos utilizados son los mismos que en la normalización con Levenshtein. Aunque los resultados conseguidos agregando Levenshtein no hayan sido mejores que en la normalización "simple", el mayor número de entidades lo hace más interesante para estos experimentos.

Para comparar, se utilizan dos métricas distintas para medir la similitud semántica basadas en ontologías. A continuación, se explicarán las particularidades de cada una, y se recopilarán y comentarán los respectivos experimentos realizados con estas medidas.

5.4.1 Distancia Pedersen

La primera prueba utiliza la distancia Pedersen, explicada en la Sección 4.2.1. En esta, los valores van de 0 a 1, siendo 0 nada de coincidencia y 1 la máxima coincidencia entre conceptos. Cabe decir, que, en nuestras pruebas, mientras que entre el 0 y el 0,5 hay un rango de valores muy amplio, entre el 0,5 y el 1 no lo hay.

Como ejemplos de distancias podemos encontrar que 0,5 marca la distancia entre un concepto hijo más concreto, y un concepto padre general. Como ejemplo, tenemos "cuadro febril" con "fiebre", el mismo valor que hay entre "faringitis aguda" y "faringitis", o entre "lesiones cutáneas" y "angiomatosis bacilar" (infección cutánea causada por bacterias[4]).

El siguiente escalón es 0,33 y marca relaciones más lejanas entre términos con un ancestro común, como la relación de "neumonía" con "tuberculosis pulmonar", ambos con "neumonitis" como ancestro común.

Resultados

Los resultados en la Tabla 5.10 reúnen los resultados de tres experimentos realizados sólo con enfermedades en distintos umbrales, siguiendo la

[4] Angiomatosis bacilar. *Manual MSD*. Consultado el 30 de junio de 2023.

Umbral	VIH+	VIH-	Precisión	Sensibilidad	valor-F
0.3	85 237	15 69	0,264	**0,850**	0,589
0.5	75 100	25 206	0,429	0,750	**0,652**
1	45 45	55 261	**0,500**	0,450	0,459

Tabla 5.10: Resultados del experimento dentro de la propuesta basada en distancia semántica, utilizando la distancia Pedersen. Este experimento sólo tiene en cuenta las enfermedades, utilizando la estrategia "Coincidencia". El umbral es el valor que define la distancia mínima entre los conceptos, a mayor umbral, más coincidencia.

Datos (Umbral)	VIH+	VIH-	Precisión	Sensibilidad	valor-F
Enfermedades (0.3)	70 137	30 169	0,338	0,700	0,577
Enfermedades (0.5)	65 53	35 253	**0,551**	0,650	**0,627**
Enfermedades + Síntomas (0.3)	85 249	15 57	0,254	**0,850**	0,579
Enfermedades + Síntomas (0.5)	81 178	19 128	0,313	0,810	0,615

Tabla 5.11: Resultados del experimento dentro de la propuesta basada en distancia semántica, utilizando la distancia Pedersen. Este experimento tiene en cuenta todos los IC, utilizando la estrategia "Pesos". El umbral es el valor que define la distancia mínima entre los conceptos, a mayor umbral, más coincidencia.

metodología "Coincidencia". Los valores más interesantes se encuentran en el umbral 0,5, que aúna una sensibilidad muy alta, y una precisión bastante alta. Esto hace que se obtenga el mejor valor-F hasta este momento, con un 0,640.

Respecto a los otros dos umbrales, vemos como el 0,3 es un umbral demasiado bajo. El crecimiento de la sensibilidad sí repercute esta vez en la precisión, haciendo que sea muy baja, y provoque un menor valor-F. El umbral 1, en cambio, consigue valores similares a los obtenidos en experimentos de normalización anteriores, por lo que aporta poca nueva información.

En la Tabla 5.11 se reúnen los mejores resultados de los experimentos realizados tanto con enfermedades como con síntomas, utilizando la metodología "Pesos". Los umbrales utilizados se indican entre paréntesis en la primera columna al lado del grupo de enfermedades consideradas. Los umbrales que no aparecen en este caso se debe a la poca información que

ofrecen, el umbral 1 en enfermedades devuelve un resultado parecido al de la Tabla 5.10 y, en el caso de umbral 0,3 y enfermedades y síntomas, se devuelven como positivas casi el 90% de las entidades.

Los experimentos en el nivel 0,5 obtienen muy buenos resultados en general, aunque no se consigue mejorar el mejor resultado anterior. En el caso de sólo enfermedades, se consigue aumentar bastante la precisión, pero el ligero descenso en la sensibilidad provoca a un valor-F menor con respecto a la estrategia "Coincidencia". Al sumarle los síntomas, se devuelve un buen resultado, aunque la sensibilidad también desciende comparado con el caso de las enfermedades.

5.4.2 Distancia Nguyen and Al-Mubaid

Como segunda prueba utilizamos la distancia Nguyen and Al-Mubaid, explicada en la Sección 4.2.1. En este caso, la distancia entre conceptos se mide de forma negativa, la coincidencia exacta sigue siendo 1, pero el resto de valores intermedios van del 0 hacia abajo, siendo -2,18 el valor mínimo.

No obstante, el rango de números intermedios entre el 0 y el -2,18 es mucho más amplio que en la distancia Pedersen, consiguiendo una granularidad mucho más fina. Por hacer la comparación con Pedersen, en este caso "faringitis aguda" y "faringitis" tienen una similitud de -0,282, existiendo también valores más altos y, por tanto, más cercanos, como "cuadro febril" y "fiebre" con un -0,199.

Cabe decir que esta mayor granularidad tiene un coste. Mientras que la ejecución del programa con la distancia Pedersen tarda unos 3 minutos en calcular la similitud entre las más de 7.000 entidades y todos los IC, la ejecución de la distancia Nguyen and Al-Mubaid tarda alrededor de 3 horas, 48 minutos. Esto puede ser un punto muy importante a la hora de realizar una experimentación a mayor escala.

Resultados

La Tabla 5.12 agrupa los resultados de los experimentos con sólo enfermedades, utilizando la distancia Nguyen and Al-Mubaid, y la estrategia "Coincidencia". La primera columna sigue marcando el umbral mínimo elegido para escoger las entidades. Observando los resultados, se puede decir que el umbral de -0,3 se correspondería, por similitud en sus estadísticas, con el umbral 0,5 de los experimentos con Pedersen. En este caso, la métrica

Umbral	VIH+	VIH-	Precisión	Sensibilidad	valor-F
-0.3	75 94	25 212	0,444	**0,750**	**0,659**
-0.25	72 78	28 228	0,480	0,720	0,655
-0.2	68 57	32 249	**0,544**	0,680	0,648
1	45 45	55 261	0,500	0,450	0,459

Tabla 5.12: Resultados del experimento dentro de la propuesta basada en distancia semántica, utilizando la distancia Nguyen and Al-Mubaid. Este experimento sólo tiene en cuenta las enfermedades, utilizando la estrategia "Coincidencia". El umbral es el valor que define la distancia mínima entre los conceptos, a mayor umbral, más coincidencia.

Datos (Umbral)	VIH+	VIH-	Precisión	Sensibilidad	valor-F
Enfermedades (-0.3)	65 54	35 252	0,546	0,650	**0,627**
Enfermedades (-0.2)	63 42	37 264	**0,600**	0,630	0,624
Enfermedades + síntomas (-0.3)	78 156	22 150	0,335	**0,780**	0,616
Enfermedades + síntomas (-0.2)	76 133	24 173	0,364	0,760	0,624

Tabla 5.13: Resultados del experimento dentro de la propuesta basada en distancia semántica, utilizando la distancia Nguyen and Al-Mubaid. Este experimento tiene en cuenta todos los IC, utilizando la estrategia "Pesos". El umbral es el valor que define la distancia mínima entre los conceptos, a mayor umbral, más coincidencia.

es más precisa, y consigue aumentar la precisión sin que baje la sensibilidad, obteniendo un valor-F de 0,659, siendo este el mejor resultado obtenido en cualquier experimento.

Podemos comprobar como la sensibilidad se mantiene bastante alta en todos los umbrales (obviando el caso de máxima coincidencia). En el valor -0,2, se consigue incluso una precisión por encima del 0,5, valor que mejora al utilizar los pesos, como se puede ver en la Tabla 5.13. Se repite, como en casos anteriores, que los síntomas añaden bastante sensibilidad al sistema, pero la caída en la precisión provoca que el valor-F no resalte por encima de los mejores resultados obtenidos.

Propuesta	Características	VIH+	VIH-	Precisión	Sensibilidad	Valor-F
Baseline	Enfermedades + Síntomas	56 129	44 177	0,303	0,560	0,479
Distancia Levenshtein	Enfermedades + Síntomas	76 198	24 108	0,277	**0,760**	0,564
Pedersen	Enfermedades / Umbral 0,5	75 100	25 206	0,429	0,750	0,652
AlMubaid & Nyungen	Enfermedades / Umbral -0,3	75 94	25 212	**0,444**	0,750	**0,659**

Tabla 5.14: Comparación de los mejores resultados obtenidos en los experimentos de las propuestas no supervisadas. La primera columna contiene la propuesta a la que pertenece, y la segunda el tipo de datos tenidos en cuenta y características concretas del experimento. El resto contienen las métricas de evaluación.

5.5 Conclusiones

Este apartado sirve como recopilación de las conclusiones obtenidas tras toda la experimentación realizada en la propuesta no supervisada.

La conclusión principal que se puede sacar de esta propuesta, es que el sistema NER condiciona gran parte del rendimiento de la misma. La propuesta depende casi en su totalidad de una correcta extracción y, sobre todo, normalización de las entidades. Por eso, las distintas aproximaciones que se han ido proponiendo, intentaban encontrar mejores maneras para lidiar con este problema. Los errores o problemas con el propio sistema NER serán discutidos de manera concienzuda en la Sección 7.2.

Se pueden extraer varias conclusiones a partir de toda la experimentación, y la tabla comparativa (Tabla 5.14) que acompaña esta sección:

1. Las distintas propuestas que siguen al "Baseline" cumplen el propósito de mejorar la propuesta inicial. Además, se identifican más entidades, y se identifican mejor. El hecho de normalizarlas y asignarles un código dentro de un estándar médico, como es UMLS, otorga a las entidades un poso de veracidad extra.

2. Se puede concluir que, la adición de los síntomas indicadores en la clasificación, necesita un refinamiento mayor. La adición de los síntomas como indicadores puede dar una información complementaria muy valiosa, pero, en algunas de las estrategias planteadas, sólo consiguen empeorar la clasificación al perderse mucha precisión. Esto es muy visible en la Tabla 5.14, comparando la precisión entre las dos primeras

filas y el resto.

3. El uso de Levenshtein, junto con el lexicón reducido, ayuda en la normalización de entidades, siendo eficaz con las faltas de ortografía y en algunos casos con adjetivos calificativos. La mejora no es muy grande (apenas unas 50 entidades), por el hecho de evitar errores en la normalización, pero es un sistema eficaz en esos casos.

4. Los mejores resultados se obtienen en la aproximación basada en distancia semántica por varias razones. Esta propuesta permite ir más allá de una coincidencia exacta, entidades que antes no eran útiles, ahora pueden tener alguna conexión con un IC que las haga potencialmente relevantes para la clasificación. Esas conexiones deberían de ser exploradas más a fondo, pero, a priori, la propuesta es positiva.

5. Finalmente, respecto a los dos mejores resultados de la Tabla 5.14. Ambos son muy similares, con una ligera mayor precisión en el caso de la distancia Al-Mubaid & Nyungen. Pero, como se ha comentado anteriormente, la diferencia en tiempo de computación entre ambas es muy significativa. En términos absolutos, la propuesta con Al-Mubaid & Nyungen, supera a la de Pedersen y a todo el resto, pero el factor del tiempo de computación debería de ser tomado en consideración para futuros experimentos.

Capítulo 6

Propuesta supervisada

Este capítulo recopila varios experimentos realizados con algoritmos supervisados para clasificar notas clínicas con posible riesgo de VIH o no.

6.1 Descripción

La propuesta en este caso se toma como un problema de clasificación supervisada clásica. Los modelos utilizados entrenan con una parte de los datos, y son evaluados con la parte restante. Estos modelos son 4: un clasificador SVM, un clasificador probabilístico de máxima entropía, y dos *transformers*, uno entrenado en un dominio general, y otro entrenado sobre el dominio biomédico.

Además, se hacen pruebas con las dos representaciones comentadas en la Sección 3.1: todo el texto de las notas clínicas, y el conjunto de entidades de enfermedades extraídas de las mismas. Esta segunda representación, nos permitirá hacer una comparación más directa entre esta propuesta y la no supervisada, al trabajar con los mismos datos.

También se realizaron pruebas con las entidades utilizadas en los experimentos con Levenshtein y HESML, pero la falta de datos (apenas algo más de 1,000 entidades en el mejor caso), con algunas notas incluso sin entidades, hacía muy difícil un correcto entrenamiento de este tipo de algoritmos.

6.1.1 Algoritmos clásicos

En primer lugar, se procede al preprocesado de los datos. En el caso de los algoritmos clásicos, es el mismo que el preprocesado completo utilizado en

la propuesta no supervisada. Es decir, convertir a minúsculas, eliminación de caracteres no alfanuméricos, tildes y *stopwords*, y *stemming*.

Una vez se han preprocesado los datos, es necesario aplicar técnicas de extracción de características, ya que los modelos de ML necesitan recibir como entrada vectores de características. En este caso se utilizan dos técnicas clásicas: codificación *one-hot* y TF-IDF (*Term Frequency-Inverse Document Frequency*). La primera, es una técnica de representación categórica en la que cada palabra está representada por un vector del tamaño del vocabulario, donde todos los elementos tienen un valor 0, a excepción del elemento correspondiente a la palabra, que tiene valor 1. El segundo, mide la importancia que tiene una palabra para un documento en una colección. Para ello se vale de la frecuencia con la que aparece una palabra en un documento, y en la colección completa, reduciendo el valor de las palabras muy comunes, y dando valor a palabras menos frecuentes.

Los vectores resultantes tienen un tamaño de 11.133 palabras en el caso de la colección de notas, y 3.914 en el de las entidades. No obstante, estos vectores presentan una dispersión muy elevada, pues de entre todas estas palabras, sólo hay un elemento que no es 0. Para reducir este problema, se realiza un proceso de selección de características, para identificar las características más importantes y reducir la dimensionalidad del problema. Para ello se realiza un test *chi-squared*, un test estadístico que mide la correlación entre una característica y su clase correspondiente. Gracias a este método, se disminuye la dimensionalidad hasta las 2.000 y 500 características para las notas y entidades respectivamente.

6.1.1.1 Máquinas de Vectores de Soporte (SVM)

La elección de los mejores hiperparámetros se realiza mediante un algoritmo que compara varias combinaciones y se queda finalmente con la de mejor rendimiento. Los hiperparámetros escogidos finalmente son:

- **Función de regularización**: añade un parámetro de penalización al modelo a medida que aumenta su complejidad para evitar un sobreajuste. El algoritmo utilizado en concreto siempre implementa la función de regularización L2. Esta función se suma a la función de

coste que se muestra en la Función 6.1.

$$J(\beta) + J(\beta) + C\sum_{j=1}^{p}\beta_j^2 \tag{6.1}$$

- **Parámetro C**: es el parámetro de regularización, y determina el grado de optimización que tiene que cumplir el modelo. El número seleccionado en este caso es 1.000.

- **Función de kernel**: la función kernel utilizada es la función de base radial (Radial Basis Function, RBF), pues es la más comúnmente utilizada en este tipo de algoritmo (Chang et al., 2010). El coeficiente γ escogido es de 0,0001.

6.1.1.2 Modelo de Máxima Entropía

Los mejores hiperparámetros escogidos en este caso son los siguientes:

- **Función de regularización**: se corresponde con el término $r(w)$ y tiene la misma finalidad que en el modelo SVM. Al igual que en este, la función escogida es la función de penalización L2, que toma la forma:

$$r(w) = \frac{1}{2}w^T w \tag{6.2}$$

- **Parámetro C**: este parámetro también tiene la misma finalidad que en el caso anterior, y el valor escogido en este caso es 1.

6.1.2 Transformers

El segundo enfoque, consiste en el reajuste (*fine-tuning*) de un modelo preentrenado, para ajustarse a la tarea de clasificación propuesta. El modelo recibe una secuencia de tokens (normalmente palabras, aunque en algunos casos pueden ser a nivel más bajo como subpalabras o caracteres) perteneciente a un texto, y devuelve su clasificación.

Los dos modelos preentrenados escogidos para la tarea son los entrenados en el marco del PlanTL, y ya han sido mencionados con anterioridad en la Sección 4.1.1. Estos son: el modelo `roberta-base-bne` para el modelo de dominio general, y el modelo `bsc-ehr-es` para el modelo de dominio clínico.

En este caso, el preprocesado es casi el mismo que en el anterior, pero se obvia la parte de eliminación de tildes y *stemming*, para asemejar los datos con los que originalmente fueron utilizados en el entrenamiento de los modelos. Otra diferencia es que los transformers no hacen uso de las mismas técnicas de extracción de características. Los datos se codifican a partir de los textos en crudo por el propio transformer en un proceso conocido como tokenización. En estos modelos, el vector resultante tiene 768 dimensiones para cada token de los 512 que admite el modelo como entrada.

Ambos entrenamientos han sido realizados con los mismos hiperparámetros para hacerlos comparables. Debido a la poca cantidad de datos, no fueron necesarios largos entrenamientos, consiguiendo converger muy pronto. Los modelos se entrenaron durante 5 épocas, con un tamaño de lote de 8 y una tasa de aprendizaje de $2e-5$.

6.2 Evaluación

Ante la reducida cantidad de datos de los que se dispone, se ha utilizado validación cruzada para evaluar los experimentos a fin de evitar un sobreajuste de los modelos. En concreto, el método de validación cruzada utilizado se conoce como k-fold. El conjunto se divide el k particiones iguales, en este caso 5. El modelo utiliza $k-1$ de las particiones para entrenar, y 1 sobre la que evaluar. Esto se repite un número k de veces y se calcula la media de los resultados obtenidos. Estas particiones han sido las mismas en todos los experimentos para que estos sean comparables.

Como primeros resultados, en la Tabla 6.1 se presentan los mejores resultados de los experimentos realizados con los algoritmos de clasificación "clásicos", SVM y de Máxima Entropía (max-ent). Estos experimentos han sido realizados sobre el corpus de notas, preprocesadas previamente, y utilizando las técnicas de extracción y selección de características explicadas antes. Las columnas de la tabla se refieren al modelo utilizado, la técnica de extracción de características y las medidas de evaluación explicadas en las Sección 3.1. En el caso de la Tabla 6.2, se recogen los resultados utilizando la representación de sólo entidades en esta ocasión, siguiendo el mismo procedimiento que con la primera.

Como se puede ver en ambas tablas, el modelo de máxima entropía es superior, aunque con poca diferencia, en el caso de la representación de todo

Modelo	Características	Precisión	Sensibilidad	valor-F
SVM	One-Hot	0,930	0,810	0,831
	TF-IDF	0,917	0,680	0,717
max-ent	One-Hot	0,953	**0,810**	**0,835**
	TF-IDF	**0,969**	0,670	0,714

Tabla 6.1: Resultados propuesta supervisada sobre la representación de todo el texto de las notas clínicas con los algoritmos clásicos SVM y métodos probabilísticos de Máxima Entropía (max-ent).

Modelo	Características	Precisión	Sensibilidad	valor-F
SVM	One-Hot	0,879	**0,800**	**0,815**
	TF-IDF	**0,968**	0,650	0,696
max-ent	One-Hot	0,890	0,720	0,749
	TF-IDF	0,967	0,630	0,677

Tabla 6.2: Resultados propuesta supervisada sobre la representación de sólo entidades con los algoritmos clásicos SVM y métodos probabilísticos de Máxima Entropía (max-ent).

el texto, mientras que el modelo SVM sí es bastante superior en el caso de la representación basada en entidades.

Donde sí se aprecia una superioridad clara, es en la comparación de resultados entre las aproximaciones One-Hot y TF-IDF, destacando por mucho la primera. La principal razón de esto recae sobre todo en la cantidad de datos utilizados. Creemos que una mayor cantidad de datos podría hacer que el uso de TF-IDF pudiera destacar por encima de la codificación binaria.

En la Tabla 6.3, se agrupan los resultados de los experimentos realizados con los transformers para ambas representaciones. La primera columna indica el nombre del modelo utilizado, la segunda la colección de datos, y las siguientes, las métricas como en las tablas anteriores.

Los resultados indican que el modelo entrenado a partir del `bsc-bio-ehr-es`, es decir, el modelo de dominio clínico, ofrece unos resultados muy superiores a los del modelo general, y muy por encima de todo el resto de resultados obtenidos a lo largo del trabajo, con una sensibilidad de 0,930 y una precisión de 0,834, lo que devuelve un valor-F de 0,909.

Esta superioridad se estrecha en los experimentos con entidades. Sigue por delante el modelo clínico, pero vemos que ha sufrido una gran caída en sus estadísticas, a raíz de la menor cantidad de datos. Sorprende la mejora de rendimiento en el modelo de dominio general con la entidades con res-

Modelo	Datos	Precisión	Sensibilidad	Valor-F
roberta-base-bne	Notas	0,714	0,630	0,645
	Entidades	0,805	0,680	0,702
bsc-bio-ehr-es	Notas	**0,834**	**0,930**	**0,909**
	Entidades	0,770	0,770	0,770

Tabla 6.3: Resultados propuesta supervisada sobre el corpus de notas y entidades con *transformers*.

Modelo	Datos	Precisión	Sensibilidad	valor-F
max-ent One-Hot	Notas	**0,953**	0,810	0,831
bsc-bio-ehr-es		0,834	**0,930**	**0,909**
SVM One-Hot	Entidades	0,879	0,800	0,815
bsc-bio-ehr-es		0,770	0,770	0,770

Tabla 6.4: Comparación de los mejores resultados obtenidos durante la experimentación de la propuesta supervisada. La primera columna contiene el modelo utilizado (acompañado de la representación vectorial en los algoritmos clásicos), la representación de los datos utilizada en el experimento, y las métricas de evaluación.

pecto al entrenamiento con las notas, siendo bastante superior en todas las métricas. Una de las razones de esto, puede ser el hecho de tener términos más "comunes". Al utilizar sólo entidades de enfermedades, se elimina una gran cantidad de términos químicos, información de pruebas serológicas, etc. Es posible que la eliminación de estos términos, aún más alejados del dominio general que las enfermedades, provoque esta mejora en el rendimiento del sistema.

6.3 Conclusiones

Se pueden extraer varias conclusiones acerca de la propuesta supervisada. Para ello, se acompaña esta sección de la Tabla 6.4, que contiene los mejores resultados obtenidos durante la experimentación.

1. Se puede apreciar claramente en la Tabla 6.4, que el mejor resultado,

con bastante diferencia, es el obtenido en el entrenamiento con el *transformer* especializado `bsc-bio-ehr-es`, utilizando todo el texto de las notas con un valor-F de 0,909.

2. No obstante, los algoritmos clásicos ofrecen muy buenos resultados, especialmente a la hora de trabajar con la representación de las entidades en comparación con el *transformer*. Esto se debe especialmente al número de datos con el que se ha trabajado. Los modelos de lenguaje necesitan de una gran cantidad de datos, mientras que los algoritmos clásicos funcionan mejor en situaciones con menos volumen de información. De ahí la superioridad de estos en el caso de las entidades.

3. Como detalle, en términos de tiempo y espacio, los algoritmos clásicos son mucho más rápidos y ligeros que los modelos de lenguaje. En un ámbito profesional, estas cualidades pueden ser tan importantes como la cantidad de datos que se estén utilizando.

Capítulo 7

Discusión

Este capítulo analiza y discute en profundidad los resultados obtenidos en la evaluación presentada en los capítulos anteriores. Se analizarán ambas propuestas, con sus métodos correspondientes, y se evaluarán los errores obtenidos para preparar mejoras en trabajos futuros.

7.1 Comparativa de propuestas

En esta sección, realizaremos la comparativa entre ambas propuestas, partiendo de los resultados obtenidos durante la experimentación.

En términos absolutos, los datos recogidos en la Tabla 7.1 indican que la propuesta supervisada con el modelo especializado en el dominio clínico `bsc-bio-ehr-es`, supera a todas las demás propuestas en cuanto a la clasificación de las notas. La mejor propuesta no supervisada, incluso se ve superada por el mismo modelo entrenado sólo con la representación de las entidades extraídas. Esto continúa dando muestras del gran potencial de los *transformers* en este tipo de tareas, destacando especialmente en esta por el reducido número de datos disponibles.

La propuesta no supervisada más destacada es la propuesta basada en distancia semántica, más concretamente la métrica Al-Mubaid & Nyungen, utilizando la distancia entre conceptos dentro de la ontología de SNOMED-CT. Como se comentaba en la Sección 5.5, esta propuesta es capaz de aprovechar las entidades más allá de la sola coincidencia, buscando relaciones "ocultas" entre las entidades y los indicadores. A pesar de no obtener los mejores resultados, creemos que esta aproximación tiene un gran potencial de mejora.

Propuesta	Características	Precisión	Sensibilidad	Valor-F
No Supervisada Distancia semántica	Al-Mubaid (-0.3) Sólo enfermedades	0,444	0,750	0,659
Supervisada Transformers `bsc-bio-ehr-es`	Notas	**0,834**	**0,930**	**0,909**
	Entidades	0,770	0,770	0,770

Tabla 7.1: Tabla comparativa de los mejores resultados de ambas propuestas (supervisada y no supervisada). La primera columna contiene los datos referentes a la propuesta, la segunda las características del experimento, y las tres siguientes, las métricas habituales.

Aparte de los resultados, la propuesta no supervisada ofrece otra serie de ventajas. Entre estas está la no necesidad de entrenamiento ni de datos anotados. Aunque el uso de HESML con esta métrica en concreto sea algo costosa.

No obstante, la ventaja más importante es la explicabilidad, que es uno de los problemas más cruciales que tienen los *transformers*, como modelos de caja de negra que son. El sistema no supervisado puede apoyar su decisión, correcta o errónea, en una serie de indicadores que han sido detectados y valorados previamente por expertos. En un ámbito profesional, esto sería clave para que el profesional sanitario tuviese la máxima información de cara a tomar una decisión. Por otro lado, debido al gran potencial que tienen este tipo de modelos, es conveniente trabajar con el objetivo de mejorar la explicabilidad de los mismos.

7.2 Análisis de errores

En esta sección se identificarán los errores que llevan a un mal rendimiento de las soluciones propuestas. Este análisis se centrará en los errores del sistema NER y de la propuesta no supervisada.

7.2.1 Errores en el sistema NER

Primero, se identificarán algunos de los problemas del sistema NER, es decir, la parte de identificación y extracción de entidades, tanto en la detección de enfermedades, como en la detección de la negación.

El error más problemático que presenta el sistema surge a la hora de lidiar

con una concentración grande de entidades. Por ejemplo, cuando aparecen muchas enfermedades a modo de pruebas serológicas:

> "... serologías de virus hepatotropos: **VHA**, **VHB**, **VHC**, **VHE**, **CMV**, **EB**, **VVZ**, HHV 6-8 que fueron IgG e IgM negativas"
>
> "Serologías (**Borrelia**, **Brucella**, **toxoplasma**, **rubéola**, **lúes**, **VEB**, **CMV**, **VHS**, **VIH**, **Varicela-Zoster**, **sarampión** y **parotiditis**) en sangre y LCR negativas"

En negrita se encuentran todas las enfermedades detectadas. En estos casos, el sistema falla a la hora de identificar el *scope* de la negación al ser muy extenso, lo que hace que todas estas entidades sean interpretadas por el sistema como enfermedades que tiene el paciente. Cabe decir, que la identificación de la negación en el resto de casos, es bastante satisfactoria, con errores puntuales.

Otro de los problemas identificados acerca del sistema NER, es la imposibilidad del mismo para detectar entidades referentes a términos en latín. Esto es un problema, pues muchos de los términos médicos referentes a enfermedades están en latín. Por ejemplo, el sistema es capaz de identificar "neumonía por Pneumocystis jiroveci", pero incapaz de identificar "Pneumocystis jiroveci" fuera de su condición de adjetivo acompañando a palabras como "neumonía" o "infección". Ocurre lo mismo con nombres de hongos como "Candida albicans".

7.2.2 Errores en la propuesta no supervisada

Aquí, se identificarán los principales errores cometidos en la experimentación con la propuesta no supervisada. En este punto intervienen tanto la selección de los IC y datos utilizados, como la normalización de entidades o las estrategias utilizadas para la clasificación de las notas.

Datos

Comenzaremos hablando de los datos. Los datos contienen una gran concentración de los indicadores estudiados, lo que los hace muy útiles para la experimentación.

Como se ha comentado en el análisis del conjunto de datos de la Sección 3.1, los notas han sido reanotadas según el paciente está infectado de VIH

o no. Una de las problemáticas de los datos, concretamente de los datos positivos, radica en eso mismo, que los pacientes ya están infectados, y, por tanto, las enfermedades y síntomas que presentan son de pacientes ya infectados. Esta es una de las razones por las que las enfermedades definitorias tienen un gran presencia, porque son enfermedades que aparecen sobre todo en estadios avanzados de la infección.

Las notas, como se comentó en la Sección 3.1.1, son muy heterogéneas. Además de tener distintas procedencias, algo positivo porque se ven varias combinaciones de síntomas o enfermedades diferentes, también son muy distintas en longitud y en formato.

Algunas de las notas son muy narrativas y extensas, y otras son más concisas y se limitan al diagnóstico y los hechos. Esto hace que en muchas notas aparezcan una gran cantidad de menciones a enfermedades que después no tienen ninguna importancia en el diagnóstico y pueden provocar mucho ruido.

Indicadores

En cuanto a la selección de los indicadores. Creemos que la selección de enfermedades y síntomas como indicadores para estudiar el posible riesgo de infección por VIH ha ofrecido resultados aceptables. Los mejores resultados, dan la razón a la propuesta, demostrando la gran incidencia de muchas de estas enfermedades en los casos positivos, cuya correcta detección era el mayor objetivo del trabajo. Esta concentración de entidades, se ve también en el propio estudio de los datos de la Sección 3.1.2.

No obstante, es cierto que acotar el trabajo sólo al estudio de enfermedades y síntomas, limita en parte el alcance del mismo. Por ejemplo, se decidió dejar fuera un indicador tan importante como son las enfermedades de transmisión sexual (ETS). Estas enfermedades aparecen, sobre todo en las notas negativas, como parte de pruebas serológicas, con el problema que antes se ha comentado que esto conlleva. Por tanto, se decidieron desechar en pos de tener una cantidad de entidades más manejable. Tampoco se han tenido en cuenta datos numéricos como analíticas o recuentos de linfocitos CD4 con el objetivo de centrar el trabajo sólo en el estudio del lenguaje natural y datos no estructurados.

Por último, la lista de indicadores es limitada. Los IC dados por los médicos expertos sirven como pautas. Por ejemplo, “tuberculosis extrapul-

monar" es un abanico de enfermedades muy grande: "tuberculosis ganglionar", "ósea", "intestinal", etc. Esto ocurre con muchos de los IC y provoca que puedan aparecer enfermedades en las notas que son indicadoras, pero que no son posibles de identificar como tales porque no están entre las estudiadas. La actualización y perfeccionamiento de esta lista es uno de los puntos claves para el futuro.

Normalización

Un problema clave, como se ha comentado a lo largo de todo el trabajo es la normalización de las entidades. Algunos de los problemas ya han sido comentados brevemente en las conclusiones de la propuesta no supervisada (Sección 5.5). Sin duda, esta es la parte más complicada del trabajo realizado.

La normalización de entidades está lejos de ser perfecta. Partiendo de 12.421 entidades, se consiguen normalizar, en el mejor de los casos, 7.340, perdiéndose un 41,6% de las mismas. No obstante, el sistema NER tampoco es perfecto y en ocasiones extrae entidades incorrectamente etiquetadas que la normalización consigue evitar.

La estrategia de normalización utilizada hace uso de un lexicón que, aunque muy completo y extenso, no puede contener todas las enfermedades que existen, y mucho menos todas sus variantes. El lexicón es una gran herramienta, pero claramente es limitado al igual que la lista de IC, y hace que varias entidades, a priori importantes, queden fuera de la clasificación por no estar presentes en él. En varias ocasiones se han añadido nuevas variantes no presentes en el lexicón a partir de términos encontrados en las propias notas.

En cuanto a las técnicas utilizadas para preprocesar las entidades y compararlas con el lexicón, también tienen espacio para la mejora. En ocasiones, el *stemming* entra en conflicto con las propias entidades, pues la raíz de algunos términos clínicos pueden ser la misma y hacer errar al sistema. Igual que alguna corrección en la eliminación de *stopwords*. Por ejemplo, en "Hepatitis A", la "A" era eliminada, o en "hepatitis no infecciosa", el "no" corría la misma suerte.

Con respecto a la utilización de Levenshtein, surge el mismo problema, no hay una solución perfecta. En ocasiones, se quiere reducir el umbral para añadir el mayor número de entidades posibles, pero eso hace que también

aumenten las posibilidades de error y entren entidades incorrectamente normalizadas.

En este caso, se ha querido priorizar el evitar fallos y no introducir entidades que no están verdaderamente presentes en las notas, pues eso sí se catalogó como una línea roja. Esto provoca que varias entidades correctas también queden fuera, por lo que habrá que estudiar mejores métodos para futuras aproximaciones.

Propuestas

Por último, se abordará de manera más general las decisiones tomadas durante el desarrollo de la propuesta no supervisada, así como la propia arquitectura de la misma.

Anteriormente, ya se han comentado errores relacionados con la extracción, identificación y normalización de entidades. Esta serie de errores se van propagando, añadiendo ruido, y, afectando a la predicción final. No obstante, a continuación identificaremos algunas decisiones que podrían mejorar el sistema y mitigar este ruido.

Por ejemplo, se podría haber realizado un estudio previo de las notas antes de la extracción de entidades. Las entidades se extraen sin ningún otro tipo de información, es decir, a parte de la negación no se estudia nada más. En el sistema actual, de una frase como "vacunado de hepatitis B", se extraería la entidad "hepatitis B" y se clasificaría la nota como positiva. Sería conveniente realizar un procesado previo que captara estos detalles, así como uno capaz de diferenciar las distintas partes de una nota (analíticas, diagnóstico, etc.), que podría aliviar algunos de los problemas que se comentaban en la parte de errores del sistema NER. Idealmente, incluso un sistema NER capaz de capturar este tipo de características.

La especulación también se decidió dejar fuera para no añadir más complejidad a la toma de decisiones, pero sería conveniente tenerla en cuenta. Por ejemplo, no es lo mismo decir "posible infección de sífilis" que "infección de sífilis", y con el sistema actual esto sería identificado como afirmativo.

Con respecto a las dos estrategias utilizadas. Rápido se detectó que la estrategia "Coincidencia" no era adecuada para clasificar las notas utilizando los síntomas. La cualidad de los síntomas son que aparecen en caso de infección aguda, pero no todos son tan específicos, y eso hace que aparezcan en la mayoría de las notas (en el último conjunto utilizado, al menos un

síntoma aparece en 250 de 360 notas). No obstante, en la estrategia "Pesos", aunque con amplio rango de mejora también, proporcionan una información más adecuada para la clasificación. En caso de que acompañen a una enfermedad, añaden razones para una posible infección, y si aparecen varios en conjunto, también pueden ser motivo de valoración. También merecen una revisión el umbral utilizado en esa estrategia, así como los valores atribuidos a las enfermedades.

Por último, queda abordar la propuesta que mejores resultados ha ofrecido dentro de la propuesta no supervisada, la propuesta basada en distancia semántica. La principal ventaja que tiene esta propuesta, y que ya se ha comentado, es ser capaz de ampliar los indicadores más allá de la simple coincidencia con las entidades. Esto sería similar a ampliar la propia lista de indicadores, o de usar un sistema de reglas más complejo.

7.3 Otros comentarios

En esta última sección, se quiere prestar atención a algunos detalles encontrados durante toda la experimentación y que no han tenido cabida en las secciones anteriores, pero que a priori son interesantes de analizar.

Como se ha podido comprobar durante la experimentación de la propuesta no supervisada (Capítulo 5), los mejores experimentos mostraban una sensibilidad alta a costa de una precisión menor, teniendo una precisión bastante baja en la mayoría de estos. Esto quiere decir que muchas notas negativas eran clasificadas como positivas por el hecho de presentar una entidad dentro de los IC. No obstante, lejos de ser un problema, podemos interpretar estos "fallos" de otra manera.

El hecho de que estas notas estén anotadas como negativas, viene de que hay un prueba realizada que así lo indica. Como se ha explicado en la motivación (Sección 1.1), estas pruebas no se suelen realizar sin sospecha, y aún con pruebas, hay incluso casos en las que no se realizan.

Aunque en la parte de la clasificación se identifiquen como errores, es valioso saber que la propuesta es capaz de identificar posibles indicadores por los que se pudo haber realizado la prueba en esos casos. Por ejemplo, en la Tabla 5.7, se identifican 201 de 306 notas con alguna enfermedad o síntoma dentro de los indicadores, más concretamente, 45 con alguna enfermedad, y 186 con algún síntoma. Y como se ha comentado en el análisis de errores,

hay algunos indicadores que no se están teniendo en cuenta, por lo que estos números sólo pueden crecer.

Algo parecido ocurre en el caso de la Tabla 5.4, donde comentábamos que era curioso que se identificaran más enfermedades dentro del grupo de las "otras" en las notas negativas que en las positivas. Aún sabiendo que estas enfermedades tienen relación no muy alta con la infección por VIH, es valioso ver casos en los que la aparición de estos indicadores más "ignorados" ha provocado que se realice una prueba.

Conocemos que estas notas son negativas, pero en el caso de ser positivas, podrían haber sido casos en los que se evitara un diagnóstico tardío, que es el mayor objetivo que se persigue en este trabajo.

Queda para un trabajo futuro el comprobar si la propuesta es efectivamente capaz de identificar el riesgo de infección previo a la prueba. Utilizando historias clínicas completas que vayan en orden cronológico, podríamos ser capaces de identificar el punto en el que habría sido adecuado realizar una prueba a partir de los indicadores.

Capítulo 8

Conclusiones y trabajo futuro

Este último capítulo recopila las conclusiones finales extraídas del trabajo realizado. Para acabar, se proponen algunas líneas de trabajo futuro a partir de los problemas y puntos de mejora identificadas en el capítulo anterior.

8.1 Conclusiones

La reducción del diagnóstico tardío de la infección por VIH, es uno de los principales problemas a los que se enfrentan los países desarrollados en la lucha contra esta enfermedad. Este trabajo es una primera aproximación para el desarrollo de un sistema capaz de reducir este diagnóstico tardío, proponiendo un enfoque novedoso mediante la explotación de historias clínicas utilizando únicamente técnicas de procesamiento del lenguaje natural.

Como principal aportación del trabajo, se han presentado dos propuestas, una supervisada y otra no supervisada, cada una con sus puntos fuertes y débiles, ambas ofreciendo resultados interesantes de estudiar. Este trabajo se sitúa como uno de los primeros en utilizar el PLN para la tarea de la predicción de infección por VIH en español.

Como conclusión general del trabajo, podemos decir que los resultados obtenidos por ambas propuestas son prometedores. La propuesta supervisada ofrece unas métricas muy elevadas, dando muestras del gran potencial de los *transformers* en tareas de clasificación de texto. En el caso de la propuesta no supervisada, aún con unas métricas sustancialmente más

reducidas, destaca por ofrecer una solución, sin necesidad de entrenamiento, solamente basada en conocimiento experto, y con un alto nivel de explicabilidad, clave en el contexto clínico en el que nos encontramos. En ambas se han comparado varios de modelos y técnicas distintas, intentando realizar una investigación exhaustiva con el fin de ir mejorando paso a paso.

En el caso de la propuesta no supervisada, el proceso de investigación se ha explicado al detalle. Primero, se estableció una propuesta que sirviera como baseline. Se identificaron los problemas que estaban apareciendo en el reconocimiento de entidades, y se puso como solución la normalización de las entidades, mejorando los resultados. Por último, se añadió una capa más, implicando el uso de la distancia semántica entre términos en una ontología como SNOMED-CT, consiguiendo con esta aproximación los mejores resultados.

La propuesta supervisada siguió los mismos principios. Se partió de los modelos y el uso de técnicas más "sencillas", para acabar utilizando los modelos más complejos disponibles para la tarea. En este caso, los *transformers* obtienen los mejores resultados, con un corpus disponible muy reducido, dando muestra de las grandes capacidades que tienen este tipo de herramientas.

A la evaluación y comparación de las propuestas siguió un análisis de errores, que va desgranando las decisiones tomadas y los errores encontrados, en otra muestra del espíritu de mejora continua que tiene este trabajo.

Entre las contribuciones del trabajo se encuentra también el desarrollo de un sistema NER capaz de identificar distintas entidades clínicas y clasificarlas según su sentido afirmativo o negativo, utilizando modelos de lenguaje y conjuntos de datos disponibles en el estado del arte. Además, se añade un proceso de normalización de entidades permitiendo su identificación dentro del estándar UMLS y dotando de mayor rigor a la propuesta. Mencionar por último la dificultad de haber trabajado únicamente en español, con la falta de referencias existentes.

No hay duda de que queda mucho trabajo para perfeccionar las propuestas y obtener una mejor solución, especialmente una que pueda ser útil en un contexto profesional. Los primeros pasos son positivos y se seguirá trabajando en ello.

8.2 Trabajo futuro

A partir del análisis de errores, realizado en la Sección 7.2, hemos podido identificar una serie de posibles mejoras que, aún estando varias ya identificadas en esa sección, serán recopiladas aquí.

El primer objetivo a futuro es evaluar estas propuestas en datos reales de hospitales, a fin de descubrir si son eficaces fuera del contexto "artificial" de este trabajo. Este es uno de los grandes problemas que muestran varios de los estudios observados durante la recopilación del estado del arte y, por tanto, es clave analizarlo.

Sería necesario ampliar y refinar la lista de indicadores. Aunque la lista con la que se ha trabajado incluye la mayoría de indicadores importantes reconocidos por los expertos, existen otros síntomas y enfermedades también estudiados como indicadores que no figuraban en esta. Incluso se podrían tener en cuenta enfermedades o síntomas menos conocidos, pero respaldados con estudios, como algunos de los identificados en (Araujo et al., 2022), y que pueden ser potencialmente valiosos para una detección temprana de la infección.

Además de ampliar la lista de enfermedades y síntomas, es necesario incorporar otros indicadores como las ETS, la edad del paciente, el sexo y otras circunstancias socio-demográficas que amplíen la información de la que dispone el sistema. Así como incorporar también datos estructurados provenientes de pruebas serológicas, analíticas, etc, que enriquezcan la información del paciente. Cabe decir que esta última parte se podrá realizar siempre que se cuente con el visto bueno del comité de ética de la institución que facilite esta información y de acuerdo a la ley de protección de datos.

La investigación en todos los campos, especialmente en el PLN, no para de avanzar, por lo que hay que seguir investigando y probando nuevas herramientas y posibles soluciones que nos ayuden. Por ejemplo, encontrar mejores datasets con los que entrenar el sistema NER, nuevos modelos o técnicas para mejorar la propia identificación de entidades y su normalización.

Sería también interesante proponer un sistema de reglas más complejo que sea capaz de identificar enfermedades que, junto con ciertos síntomas, enfermedades o la ausencia de estos, pueden implicar o no un mayor riesgo de infección.

Por último, transportar el sistema no supervisado a otros problemas

relacionados, como identificar posibles casos de "oportunidades perdidas". Es decir, a partir de notas clínicas ordenadas cronológicamente, identificar si hubo una fecha en la que al paciente se le podría haber realizado una prueba serológica previa al momento en el que se le realizó. Sería también interesante probar la solución supervisada en este caso, aunque eso requeriría de un proceso extra de anotación.

Bibliografía

Bibliografía

[Aberdeen et al.2010] Aberdeen, John, Samuel Bayer, Reyyan Yeniterzi, Ben Wellner, Cheryl Clark, David Hanauer, Bradley Malin, y Lynette Hirschman. 2010. The mitre identification scrubber toolkit: design, training, and assessment. *International journal of medical informatics*, 79(12):849–859.

[Agrawal, Imieliński, y Swami1993] Agrawal, Rakesh, Tomasz Imieliński, y Arun Swami. 1993. Mining association rules between sets of items in large databases. En *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, páginas 207–216.

[Agrebi y Larbi2020] Agrebi, Said y Anis Larbi. 2020. Use of artificial intelligence in infectious diseases. En *Artificial intelligence in precision health*. Elsevier, páginas 415–438.

[Aguado y Bel Rafecas2022] Aguado, Mercedes y Núria Bel Rafecas. 2022. A corpus of spanish clinical records annotated for abbreviation identification. *Procesamiento del Lenguaje Natural. 2022;(68): 99-109.*

[Akbik et al.2019] Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, y Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, páginas 54–59, Minneapolis, Minnesota, Junio. Association for Computational Linguistics.

[Almagro et al.2020] Almagro, Mario, Raquel Martínez Unanue, Víctor Fresno, y Soto Montalvo. 2020. Icd-10 coding of spanish electronic

discharge summaries: An extreme classification problem. *IEEE Access*, 8:100073–100083.

[Aone et al.1998] Aone, Chinatsu, Lauren Halverson, Tom Hampton, y Mila Ramos-Santacruz. 1998. Sra: Description of the ie2 system used for muc-7. En *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

[Araujo et al.2022] Araujo, Lourdes, Juan Martinez-Romo, Otilia Bisbal, y Ricardo Sanchez-de Madariaga. 2022. Discovering hiv related information by means of association rules and machine learning. *Scientific Reports*, 12(1):18208.

[Arkhipov et al.2019] Arkhipov, Mikhail, Maria Trofimova, Yurii Kuratov, y Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. En *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, páginas 89–93.

[Badjatiya et al.2018] Badjatiya, Pinkesh, Litton J Kurisinkel, Manish Gupta, y Vasudeva Varma. 2018. Attention-based neural text segmentation. En *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, páginas 180–193. Springer.

[Báez et al.2020] Báez, Pablo, Fabián Villena, Matías Rojas, Manuel Durán, y Jocelyn Dunstan. 2020. The chilean waiting list corpus: a new resource for clinical named entity recognition in spanish. En *Proceedings of the 3rd clinical natural language processing workshop*, páginas 291–300.

[Balzer et al.2019] Balzer, Laura B, Diane V Havlir, Moses R Kamya, Gabriel Chamie, Edwin D Charlebois, Tamara D Clark, Catherine A Koss, Dalsone Kwarisiima, James Ayieko, Norton Sang, Jane Kabami, Mucunguzi Atukunda, Vivek Jain, Carol S Camlin, Craig R Cohen, Elizabeth A Bukusi, Mark Van Der Laan, y Maya L Petersen. 2019. Machine learning to identify persons at high-risk of human immunodeficiency virus acquisition in rural kenya and uganda. *Clinical Infectious Diseases*, 71(9):2326–2333, 11.

[Bengio, Courville, y Vincent2013] Bengio, Yoshua, Aaron Courville, y Pascal Vincent. 2013. Representation learning: A review and new perspec-

tives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

[Black, Rinaldi, y Mowatt1998] Black, William J, Fabio Rinaldi, y David Mowatt. 1998. Facile: Description of the ne system used for muc-7. En *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

[Blei, Ng, y Jordan2003] Blei, David M, Andrew Y Ng, y Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Bodenreider2004] Bodenreider, Olivier. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

[Campillos-Llanos2019] Campillos-Llanos, Leonardo. 2019. First steps towards building a medical lexicon for spanish with linguistic and semantic information. En *Proc. of BioNLP 2019*, August 1st.

[Campillos-Llanos et al.2021] Campillos-Llanos, Leonardo, Ana Valverde-Mateos, Adrián Capllonch-Carrión, y Antonio Moreno-Sandoval. 2021. A clinical trials corpus annotated with umls entities to enhance the access to evidence-based medicine. *BMC medical informatics and decision making*, 21(1):1–19.

[Cañete et al.2020] Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, y Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020(2020):1–10.

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

[Carrino et al.2022] Carrino, Casimiro Pio, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, y Marta Villegas. 2022. Pretrained biomedical language models for clinical nlp in spanish. En

Proceedings of the 21st Workshop on Biomedical Language Processing, páginas 193–199.

[Chang et al.2010] Chang, Yin-Wen, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, y Chih-Jen Lin. 2010. Training and testing low-degree polynomial data mappings via linear svm. *Journal of Machine Learning Research*, 11(48):1471–1490.

[Chen2015] Chen, Yahui. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

[Chinchor y Robinson1997] Chinchor, N. y P. Robinson. 1997. Muc-7 named entity task definition. En *Proceedings of the 7th Conference on Message Understanding*, volumen 29, páginas 1–21.

[Chiu et al.2022] Chiu, Han-Yi Robert, Chun-Kai Hwang, Shey-Ying Chen, Fuh-Yuan Shih, Hsieh-Cheng Han, Chwan-Chuen King, John Reuben Gilbert, Cheng-Chung Fang, y Yen-Jen Oyang. 2022. Machine learning for emerging infectious disease field responses. *Scientific Reports*, 12(1):328.

[Cho y Lee2019] Cho, Hyejin y Hyunju Lee. 2019. Biomedical named entity recognition using deep neural networks with contextual information. *BMC bioinformatics*, 20:1–11.

[Collier et al.2004] Collier, Nigel, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, y Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. En *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, páginas 73–78, Geneva, Switzerland, Agosto 28th and 29th. COLING.

[Cortes y Vapnik1995] Cortes, Corinna y Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

[Cortez Vásquez et al.2009] Cortez Vásquez, Augusto, Hugo Vega huerta, Jaime Pariona Quispe, y Ana Maria Huayna. 2009. Procesamiento de lenguaje natural. *Revista de investigación de Sistemas e Informática*, 6(2):45–54, dic.

[Crespillo et al.2015] Crespillo, Clara, C.M. Rodriguez, Juan Emilio Losa, Rosa Escudero, R García, Rafael Hervás Gómez, L Moreno, César .A.J. Henríquez, y María Velasco. 2015. Evolución en el hábito de solicitud de vih en un hospital general en tres años. ¿hacia dónde vamos? En *XIX Congreso SEIMC*.

[DCVIHT2021] DCVIHT. 2021. Plan de prevención y control de la infección por el VIH y las ITS 2021-2030 en España. Informe técnico, División de control de VIH, ITS, Hepatitis virales y tuberculosis. Ministerio de Sanidad/Centro Nacional de Epidemiología. Instituto de Salud Carlos III, Diciembre.

[DCVIHT2022] DCVIHT. 2022. Vigilancia Epidemiológica del VIH y sida en España 2021: Sistema de Información sobre Nuevos Diagnósticos de VIH y Registro Nacional de Casos de Sida. Informe técnico, Centro Nacional de Epidemiología. Instituto de Salud Carlos III/División de control de VIH, ITS, Hepatitis virales y tuberculosis. Ministerio de Sanidad, Noviembre.

[Devlin et al.2018] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, y Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Elman1990] Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

[Estrada Chacón et al.2002] Estrada Chacón, Ulises, Juan Francisco Bandera Tirado, Daniel Portela Ramírez, y Sorgalim Benavides García. 2002. Alteraciones radiológicas en pacientes VIH con infección respiratoria aguda. *Revista Cubana de Medicina*, 41, 12.

[Feller et al.2018] Feller, Daniel J, Jason Zucker, Michael T Yin, Peter Gordon, y Noémie Elhadad. 2018. Using clinical notes and natural language processing for automated hiv risk assessment. *Journal of acquired immune deficiency syndromes (1999)*, 77(2):160.

[Felsen et al.2017] Felsen, Uriel R, Chinazo O Cunningham, Moonseong Heo, Donna C Futterman, Jeffrey M Weiss, y Barry S Zingman. 2017. An expanded hiv testing strategy leveraging the electronic medical record

uncovers undiagnosed infection among hospitalized patients. *Journal of acquired immune deficiency syndromes (1999)*, 75(1):27.

[Fieggen et al.2022] Fieggen, Joshua, Eli Smith, Lovkesh Arora, y Bradley Segal. 2022. The role of machine learning in hiv risk prediction. *Frontiers in Reproductive Health*, 4.

[Friedman, Rindflesch, y Corn2013] Friedman, Carol, Thomas C Rindflesch, y Milton Corn. 2013. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *Journal of biomedical informatics*, 46(5):765–773.

[Gérvas y Fernández2000] Gérvas, Juan y Mercedes Pérez Fernández. 2000. La historia clínica electrónica en atención primaria. fundamento clínico, teórico y práctico. *SEMERGEN-Medicina de Familia*, 26(1):17–32.

[Goenaga et al.2021] Goenaga, Iakes, Xabier Lahuerta, Aitziber Atutxa, y Koldo Gojenola. 2021. A section identification tool: towards hl7 cda/ccr standardization in spanish discharge summaries. *Journal of Biomedical Informatics*, 121:103875.

[González del Castillo et al.2020] González del Castillo, Juan, Guillermo Burillo-Putze, Alfonso Cabello, Adrián Curran, Eissa Jaloud Saavedra, Pierre Malchair, María José Marchena, Òscar Miró, Alberto Pizarro, Cesar Sotomayor, y others. 2020. Recomendaciones dirigidas a los servicios de urgencias para el diagnóstico precoz de pacientes con sospecha de infección por vih y su derivación para estudio y seguimiento. *Emergencias*, 32(6).

[Goodfellow, Bengio, y Courville2016] Goodfellow, Ian, Yoshua Bengio, y Aaron Courville. 2016. *Deep learning*. MIT press.

[Hochreiter y Schmidhuber1997] Hochreiter, Sepp y Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Huang y Lu2016] Huang, Chung-Chi y Zhiyong Lu. 2016. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics*, 17(1):132–144.

[Intxaurrondo2018] Intxaurrondo, Ander. 2018. Spaccc, Nov.

[Krakower et al.2019] Krakower, Douglas S, Susan Gruber, Katherine Hsu, John T Menchaca, Judith C Maro, Benjamin A Kruskal, Ira B Wilson, Kenneth H Mayer, y Michael Klompas. 2019. Development and validation of an automated hiv prediction algorithm to identify candidates for pre-exposure prophylaxis: a modelling study. *The Lancet HIV*, 6(10):e696–e704.

[Lafferty, McCallum, y Pereira2001] Lafferty, John, Andrew McCallum, y Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[Lastra-Díaz, Lara-Clares, y Garcia-Serrano2022] Lastra-Díaz, Juan J, Alicia Lara-Clares, y Ana Garcia-Serrano. 2022. Hesml: a real-time semantic measures library for the biomedical domain with a reproducible survey. *BMC bioinformatics*, 23(1):23.

[Lecun et al.1998] Lecun, Y., L. Bottou, Y. Bengio, y P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[Lee et al.2020] Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, y Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

[Levenshtein1966] Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. En *Soviet physics doklady*, volumen 10, páginas 707–710. Soviet Union.

[Li et al.2022] Li, Irene, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlalı, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, y others. 2022. Neural natural language processing for unstructured data in electronic health records: A review. *Computer Science Review*, 46:100511.

[Li et al.2020] Li, Jing, Aixin Sun, Jianglei Han, y Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

[Lima Lopez et al.2020] Lima Lopez, Salvador, Naiara Perez, Montse Cuadros, y German Rigau. 2020. NUBes: A corpus of negation and

uncertainty in Spanish clinical texts. En *Proceedings of the Twelfth Language Resources and Evaluation Conference*, páginas 5772–5781, Marseille, France, Mayo. European Language Resources Association.

[Lima-López et al.2021] Lima-López, Salvador, Eulàlia Farré-Maduell, Antonio Miranda-Escalada, Vicent Brivá-Iglesias, y Martin Krallinger. 2021. Nlp applied to occupational health: Meddoprof shared task at iberlef 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. *Procesamiento del Lenguaje Natural*, 67:243–256.

[Liu et al.2017] Liu, Jingzhou, Wei-Cheng Chang, Yuexin Wu, y Yiming Yang. 2017. Deep learning for extreme multi-label text classification. En *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, páginas 115–124.

[Malouf2002] Malouf, Robert. 2002. A comparison of algorithms for maximum entropy parameter estimation. En *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

[Marafino et al.2014] Marafino, Ben J, Jason M Davies, Naomi S Bardach, Mitzi L Dean, y R Adams Dudley. 2014. N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *Journal of the American Medical Informatics Association*, 21(5):871–875, 04.

[Marcus et al.2019] Marcus, Julia L, Leo B Hurley, Douglas S Krakower, Stacey Alexeeff, Michael J Silverberg, y Jonathan E Volk. 2019. Use of electronic health record data and machine learning to identify candidates for hiv pre-exposure prophylaxis: a modelling study. *The lancet HIV*, 6(10):e688–e695.

[Marimon, Vivaldi, y Bel Rafecas2017] Marimon, Montserrat, Jorge Vivaldi, y Núria Bel Rafecas. 2017. Annotation of negation in the iula spanish clinical record corpus. *Blanco E, Morante R, Saurí R, editors. SemBEaR 2017. Computational Semantics Beyond Events and Roles; 2017 Apr 4; Valencia, Spain. Stroudsburg (PA): ACL; 2017. p. 43-52.*

[Menacho et al.2013] Menacho, I, E Sequeira, M Muns, O Barba, L Leal, T Clusa, E Fernandez, L Moreno, D Raben, J Lundgren, y others.

2013. Comparison of two hiv testing strategies in primary care centres: indicator-condition-guided testing vs. testing of those with non-indicator conditions. *HIV medicine*, 14:33–37.

[Miranda-Escalada et al.2022] Miranda-Escalada, Antonio, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, y Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. En *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

[Miró i Andreu et al.2023] Miró i Andreu, Òscar, Emília Miró, Miriam Ejarque Carbó, Mireia Saura, Alexis Rebollo, Rocío De Paz, Josep Maria Guardiola, Alejandro Smithson Amat, Daniel Iturriza, Cristina Ramió Lluch, y others. 2023. Detección en urgencias de infección por vih en pacientes que consultan por condiciones potencialmente relacionadas con infección oculta: resultados iniciales del programa" urgències vihgila". *Revista Espanola de Quimioterapia, 2023, vol. 36, num. 2, p. 169-179.*

[Mitchell1997] Mitchell, Tom Michael. 1997. *Machine learning*. McGraw-Hill.

[Montoy, Dow, y Kaplan2016] Montoy, Juan Carlos C., William H. Dow, y Beth C. Kaplan. 2016. Patient choice in opt-in, active choice, and opt-out hiv screening: randomized clinical trial. *BMJ*, 352.

[Morwal, Jahan, y Chopra2012] Morwal, Sudha, Nusrat Jahan, y Deepti Chopra. 2012. Named entity recognition using hidden markov model (hmm). *International Journal on Natural Language Computing (IJNLC) Vol*, 1.

[Nadkarni, Ohno-Machado, y Chapman2011] Nadkarni, Prakash M, Lucila Ohno-Machado, y Wendy W Chapman. 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551.

[Nguyen y Al-Mubaid2006] Nguyen, Hoa A y Hoa Al-Mubaid. 2006. New ontology-based semantic similarity measure for the biomedical domain. En *2006 IEEE International Conference on Granular Computing*, páginas 623–628. IEEE.

[Oronoz et al.2015] Oronoz, Maite, Koldo Gojenola, Alicia Pérez, Arantza Díaz De Ilarraza, y Arantza Casillas. 2015. On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. *Journal of biomedical informatics*, 56:318–332.

[Pedersen et al.2007] Pedersen, Ted, Serguei VS Pakhomov, Siddharth Patwardhan, y Christopher G Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299.

[Pesaranghader et al.2019] Pesaranghader, Ahmad, Stan Matwin, Marina Sokolova, y Ali Pesaranghader. 2019. deepbiowsd: effective deep neural word sense disambiguation of biomedical text data. *Journal of the American Medical Informatics Association*, 26(5):438–446.

[Pomares-Quimbaya, Kreuzthaler, y Schulz2019] Pomares-Quimbaya, Alexandra, Markus Kreuzthaler, y Stefan Schulz. 2019. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. *BMC Medical Research Methodology*, 19(1):155, Julio.

[Radford et al.2018] Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, y others. 2018. Improving language understanding by generative pre-training.

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

[Ramshaw y Marcus1999] Ramshaw, Lance A y Mitchell P Marcus. 1999. Text chunking using transformation-based learning. *Natural language processing using very large corpora*, páginas 157–176.

[Redondo et al.2019] Redondo, Laia Cayuelas, Marina Ruíz, Belchin Kostov, Ethel Sequeira, Pablo Noguera, Maria Alba Herrero, Ignacio Menacho, Olga Barba, Thaïs Clusa, Benet Rifa, y others. 2019. Indicator condition-guided hiv testing with an electronic prompt in primary healthcare: a before and after evaluation of an intervention. *Sexually Transmitted Infections*, 95(4):238–243.

[Rojas et al.2022] Rojas, Matías, Jose Barros, Mauricio Araneda, y Jocelyn Dunstan. 2022. Flert-matcher: A two-step approach for clinical named entity recognition and normalization.

[Russell y Norvig2021] Russell, Stuart J. y Peter Norvig. 2021. *Artificial Intelligence: A Modern Approach*. Pearson, 4 edición.

[Salmerón-Béliz et al.2022] Salmerón-Béliz, Octavio José, Elia Pérez-Fernández, Oscar Miró, Manuel Salido-Mota, Verónica Diez-Diez, Manuel Gil-Mosquera, Neus Robert-Boter, María Arranz-Betegón, Carmen Navarro-Bustos, José María Guardiola-Tey, y Juan González del Castillo. 2022. Evaluation of emergency department visits prior to an hiv diagnosis: Missed opportunities. *Enfermedades infecciosas y microbiologia clinica (English ed.)*.

[Sánchez-de Madariaga et al.2022] Sánchez-de Madariaga, Ricardo, Juan Martinez-Romo, José Miguel Cantero Escribano, y Lourdes Araujo. 2022. Semi-supervised incremental learning with few examples for discovering medical association rules. *BMC Medical Informatics and Decision Making*, 22(1):1–11.

[Sanders et al.2005] Sanders, Gillian D., Ahmed M. Bayoumi, Vandana Sundaram, S. Pinar Bilir, Christopher P. Neukermans, Chara E. Rydzak, Lena R. Douglass, Laura C. Lazzeroni, Mark Holodniy, y Douglas K. Owens. 2005. Cost-effectiveness of screening for hiv in the era of highly active antiretroviral therapy. *New England Journal of Medicine*, 352(6):570 – 585.

[Schuster y Paliwal1997] Schuster, Mike y Kuldip Paliwal. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45:2673 – 2681, 12.

[Schweter y Akbik2020] Schweter, Stefan y Alan Akbik. 2020. Flert: Document-level features for named entity recognition. *arXiv preprint arXiv:2011.06993*.

[Sevgili et al.2022] Sevgili, Özge, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, y Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

[Sullivan et al.2013] Sullivan, Ann K, Dorthe Raben, Joanne Reekie, Michael Rayment, Amanda Mocroft, Stefan Esser, Agathe Leon, Josip Begovac, Kees Brinkman, Robert Zangerle, y others. 2013. Feasibility and effectiveness of indicator condition-guided testing for hiv: results from hides i (hiv indicator diseases across europe study). *PloS one*, 8(1):e52845.

[Valmianski et al.2019] Valmianski, Ilya, Caleb Goodwin, Ian M Finn, Naqi Khan, y Daniel S Zisook. 2019. Evaluating robustness of language models for chief complaint extraction from patient-generated text. *arXiv preprint arXiv:1911.06915*.

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

[Wang et al.2018] Wang, Yue, Kai Zheng, Hua Xu, y Qiaozhu Mei. 2018. Interactive medical word sense disambiguation through informed learning. *Journal of the American medical informatics association*, 25(7):800–808.

[Weissman et al.2021] Weissman, Sharon, Xueying Yang, Jiajia Zhang, Shujie Chen, Bankole Olatosi, y Xiaoming Li. 2021. Using a machine learning approach to explore predictors of health care visits as missed opportunities for hiv diagnosis. *AIDS (London, England)*, 35(Suppl 1):S7.

[Xiao, Choi, y Sun2018] Xiao, Cao, Edward Choi, y Jimeng Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428.

[Yao, Mao, y Luo2019] Yao, Liang, Chengsheng Mao, y Yuan Luo. 2019. Clinical text classification with rule-based features and knowledge-guided

convolutional neural networks. *BMC medical informatics and decision making*, 19(3):31–39.

[Zhang y Elhadad2013] Zhang, Shaodian y Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, 46(6):1088–1098.

Anexo A

Enfermedades y síntomas indicadores

A.1 Enfermedades definitorias

Nombre	**Puntuación**
Neumonía recurrente ($\geq$ 2 NAC bacterianas durante 12 meses)	4,2
Bacteriemia recurrente por Salmonella	4,5
Tuberculosis pulmonar o extrapulmonar (ganglionar, ósea, intestinal, diseminada, meníngea, cerebral)	4,5
Tuberculosis miliar	4,8
Micobacterias atípicas diseminadas o extrapulmonares (*Mycobacterium avium*, *Mycobacterium kansasii*)	4,8
Candidiasis esofágica, bronquial, traqueal o pulmonar	4,5
Neumonía por *Pneumocystis jirovecii*, *Pneumocystis carinii*	5
Histoplasmosis extrapulmonar	4,8
Coccidioidomicosis extrapulmonar	4,7
Criptococosis extrapulmonar	4,8
Criptosporidiosis	4,8
Infecciones por virus del herpes simple: ulceras crónicas, bronquitis, neumonía o esofagitis	4,3
Infecciones por citomegalovirus (retinitis por CMV, hepática, esplénica, ganglionar)	4,7
Toxoplasmosis cerebral	4,8
Leucoencefalopatía multifocal progresiva	4,7
Sarcoma de Kaposi	5
Linfomas (de Burkitt, cerebral primario, inmunoblásticos)	4,8

Sigue en la página siguiente.

Nombre	Puntuación
Carcinoma de cérvix uterino invasivo	4,5
Encefalopatía asociada al VIH	5
Síndrome de caquexia progresiva por VIH	5

Tabla A.1: Lista de enfermedades definitorias de VIH, con las puntuaciones medias otorgadas por los médicos expertos según su sensibilidad y especificidad en una posible infección. La puntuación es un número real entre 0 y 5, siendo 0 el mínimo y 5 el máximo.

A.2 Enfermedades indicadoras

Nombre	Puntuación
Angiomatosis bacilar	3,7
Candidiasis orofaringea, muguet	3,5
Candidiasis vulvovaginal; persistente, frecuente, o que no responde al tratamiento	3,2
Displasia cervical (moderada o severa)/carcinoma cervical in situ	3,7
Síndrome constitucional, fiebre persistente (38.5°C) y/o diarrea crónica de >1 mes de duración (COMBINACIÓN)	4
Leucoplasia oral vellosa	3,8
Herpes zoster, al menos 2 episodios distintos o más de un dermatoma	3,7
Púrpura trombocitopénica idiopática	3
Listeriosis	2,8
Enfermedad pélvica inflamatoria, abscesos tuboováricos	3,3
Neuropatía periférica	3

Tabla A.2: Lista de enfermedades indicadoras de VIH, con las puntuaciones medias otorgadas por los médicos expertos según su sensibilidad y especificidad en una posible infección. La puntuación es un número real entre 0 y 5, siendo 0 el mínimo y 5 el máximo.

A.3 Otras enfermedades indicadoras

Nombre	Puntuación
Cáncer de pulmón primario	2,5
Meningitis linfocítica	3,2
Psoriasis grave o atípica	2,8
Síndrome de Guillain-Barré	2,3
Mononeuritis	2,2
Demencia subcortical	2,5
Esclerosis múltiple	2,5
Insuficiencia renal crónica idiopática	2
Hepatitis A	2,7
Neumonía adquirida en la comunidad	2,5
Dermatitis atópica	2,2

Tabla A.3: Lista de enfermedades en las que el VIH tiene una prevalencia $> 0,1\%$, con las puntuaciones medias otorgadas por los médicos expertos según su sensibilidad y especificidad en una posible infección. La puntuación es un número real entre 0 y 5, siendo 0 el mínimo y 5 el máximo.

A.4 Síntomas

Nombre	Puntuación
Úlceras mucocutáneas: orales, labiales, yugales, bucales, faríngeas, anales, genitales	3,7
Exantema/Rash/Erupción cutánea	2,8
Mialgias y artralgias	1,7
Anorexia, pérdida de peso injustificada	2,7
Fiebre	2,8
Manifestaciones graves a nivel del sistema nervioso central/ Meningitis/Encefalitis	3,5
Fatiga, malestar, astenia	2
Cefalea	1,7
Linfadenopatía periférica/Adenopatías	3,5
Faringitis	2,5
Alteraciones gastrointestinales, diarrea	2,8
Mononucleosis/Síndrome mononucleósido	4,8

Tabla A.4: Lista de síntomas indicadores de la infección por VIH, con las puntuaciones medias otorgadas por los médicos expertos según su sensibilidad y especificidad en una posible infección. La puntuación es un número real entre 0 y 5, siendo 0 el mínimo y 5 el máximo.

Anexo B

Enfermedades y síntomas añadidos a MedLexSp

B.1 Enfermedades y síntomas añadidos a MedLexSp

Nombre	CUI	Grupo
Micobacteriosis atípica	C0026919	Definitoria
Candidiasis pulmonar	C0153251	Definitoria
Encefalitis por citomegalovirus	C0238097	Definitoria
Tuberculosis cerebral	C0275909	Definitoria
Hepatitis por citomegalovirus	C0276252	Definitoria
Mononucleosis por citomegalovirus	C0276254	Definitoria
Encefalopatía asociada al vih	C0276548	Definitoria
Coccidioidomicosis extrapulmonar/diseminada	C0276667	Definitoria
Neumonía por herpes simple	C0339975	Definitoria
Colitis por citomegalovirus	C0341335	Definitoria
Micobacteriosis diseminada	C0343431	Definitoria
Esofagitis por herpes simple	C0343570	Definitoria
Candidiasis traqueal	C0343862	Definitoria
Criptococosis extrapulmonar/diseminada	C0343890	Definitoria
Tuberculosis extrapulmonar	C0679362	Definitoria
Neumonía recurrente	C0694550	Definitoria
Infección micobacteriana atípica diseminada	C0694566	Definitoria

Sigue en la página siguiente.

Nombre	**CUI**	**Grupo**
Infección causada por mycobacterium avium	C1299472	Definitoria
Candidiasis bronquial	C1536274	Definitoria
Infección diseminada por mycobacterium kansasii	C1827561	Definitoria
Bacterinimia causada por Salmonella	C2584752	Definitoria
Tuberculosis ósea	C3203357	Definitoria
Bronquitis por herpes simple	C5442294	Definitoria
Candidiasis orofaríngea-esofágica	C0343859	Indicadora
Candidiasis vulvovaginal recurrente	C4023003	Indicadora
Cáncer de pulmón primario	C4273669	Otra
Úlcera labial	C0267033	Síntoma
Placa eritematosa	C0332477	Síntoma
Úlcera mucocutánea	C0521478	Síntoma
Úlcera faríngea	C1290332	Síntoma

Tabla B.1: Lista de enfermedades y síntomas dentro de los IC añadidos a MedLexSp que no estaban en el lexicón original.