Desarrollo de un modelo de Procesamiento de Lenguaje Natural para clasificar posibles casos de cáncer en registros de defunción de Manizales y facilitar la captación del Registro Poblacional de Cáncer de Manizales

Trabajo de grado

María Fernanda Del corral Pérez

Juan Pablo Pájaro Hernández

Director

Pontificia Universidad Javeriana

Facultad de Ciencias

Pregrado en Ciencia de datos

Bogotá D.C

2026

Agradecimientos

Quiero agradecer, en primer lugar, a mi papá, Carlos Del Corral, q.e.p.d., a quien llevo en mi corazón, por el apoyo que me dio, por haber creído en mí y por haber acompañado siempre mis decisiones. Su confianza y amor siguen siendo una parte muy importante de este logro.

A mi mamá, Zulay Pérez, y a mi hermana, María José Del Corral, por su amor, compañía y apoyo durante todo este proceso. Gracias por estar siempre presentes y por ayudarme especialmente en la comprensión de la parte médica de este trabajo.

También agradezco a mi tutor, Juan Pablo Pájaro, por su acompañamiento, orientación y ayuda durante el desarrollo de este trabajo de grado.

Al equipo del Registro Poblacional de Cáncer de Manizales, a Alexandra Giraldo, Nelson Arias, Esther de Vries, Santiago Murillo y Santiago García, por su disposición y apoyo a lo largo del proyecto. Su acompañamiento fue fundamental para entender las necesidades reales del registro y orientar este trabajo hacia una utilidad práctica.

A todos mis amigos de la carrera, en especial a Juan Lizcano, David Santiago, Paula Bautista y Juan Arias, gracias por su compañía, amistad y apoyo durante estos años.

A mi novio, Felipe Rincón, gracias por su apoyo incondicional, por acompañarme en los momentos difíciles. Finalmente, agradezco a mis mascotas, Haru, Yippi y Maki, por acompañarme con su presencia y cariño durante muchas horas de trabajo.

A todos, gracias por hacer parte de este proceso y de la construcción de este trabajo de grado.

Resumen

El Registro Poblacional de Cáncer de Manizales enfrenta una alta exigencia operativa en la revisión de registros de defunción para identificar posibles casos registrables de cáncer. Esta labor, clave dentro del registro, se realiza sobre textos heterogéneos, con distinto nivel de detalle y ambigüedad clínica. Sin embargo, la revisión manual consume gran parte del tiempo del equipo disponible. Por ello, este estudio tuvo como objetivo diseñar y evaluar un modelo de procesamiento de lenguaje natural para clasificar registros de defunción según indicios de cáncer, con el fin de apoyar la revisión de potenciales casos por parte del registro.

La metodología siguió un enfoque de investigación basada en el diseño. Se compararon tres alternativas de clasificación construidas sobre una misma base de datos, limpiada y normalizada: un modelo basado en reglas, un modelo *Support Vector Machine* (SVM) jerárquico en dos etapas y un modelo basado en *Transformers.* La evaluación consideró exactitud, precisión, sensibilidad, F1 y especificidad. Además, se priorizó la reducción de falsos negativos: registros con indicios de cáncer clasificados erróneamente como ausencia de mención. Para el registro, este error es más grave que un falso positivo, ya que los casos positivos pasan por chequeos y codificación posterior, donde pueden corregirse errores.

Los resultados evidenciaron diferencias entre los modelos. El SVM alcanzó la mayor exactitud global (0,9875), superando al modelo basado en reglas (0,9525) y al basado en *Transformers* (0,9463). También presentó pocos errores críticos, con 2 casos, frente a 12 del modelo de reglas y 2 del modelo *Transformers*.

En conjunto, el SVM ofreció el mejor equilibrio entre desempeño global, estabilidad por clase y control de errores críticos, por lo que puede apoyar la revisión, reducir tiempos y favorecer una captación oportuna y consistente de casos potenciales.

Tabla de contenido

Agradecimientos 2

Resumen 3

Tabla de contenido 5

1. Introducción 8

1.1 Contexto: la importancia de los registros poblacionales de cáncer 8

1.2. Planteamiento del problema: registros de defunción y su oportunidad 9

1.3. Justificación: por qué un modelo de PLN es necesario 10

1.4. Pregunta de investigación 11

1.5. Objetivos 11

1.5.1 Objetivo general: 11

1.5.2. Objetivos específicos: 12

2. Marco teórico 13

2.1. Modelo basado en reglas 13

2.2. Modelo SVM 14

2.3. Modelo basado en *Transformers* 16

3. Revisión de la literatura 18

3.1. Resumen 18

3.2. Introducción 19

3.3. Antecedentes 19

3.4. Materiales y métodos 22

3.5. Resultados 25

3.6. Discusión 31

3.7. Conclusiones 32

4. Metodología 34

4.1 Enfoque metodológico 34

4.2. Fuente de datos y construcción de la muestra etiquetada 34

4.3. Selección de variables textuales y seudoanonimización 35

4.4. Limpieza, normalización y estandarización del texto 36

4.5. Construcción del texto de entrada 37

4.6. Desarrollo de los modelos 37

4.6.1. Modelo basado en reglas 38

4.6.2. Modelo SVM 39

4.6.3. Modelo basado en *Transformer*s 40

4.7. Estrategia de evaluación 41

4.8. Selección del modelo final 43

4.9. Consolidación del entregable para el RPCa-M 44

5. Resultados y análisis 46

5.1. Comparación de los flujos de entrenamiento 46

5.2. Criterios de decisión y configuración de los modelos 48

5.2.1. Modelo basado en reglas 48

5.2.2. Modelo SVM 49

5.2.3. Modelo basado en *Transformer*s 51

5.3. Resultados cuantitativos comparativos 52

5.3.1. Desempeño general 52

5.3.2. Desempeño por clase 54

5.3.3. Falsos 0 y recuperación de casos relevantes 55

5.4. Análisis comparativo final 59

5.5. Entrega del modelo y valoración del usuario final 61

6. Conclusiones 65

6.1. Respuesta a la pregunta de investigación 65

6.2. Principales hallazgos del estudio 66

6.3. Aporte del estudio al RPCa-M 67

6.4. Entregable al registro y valoración del resultado. 69

6.5. Limitaciones del estudio 70

6.6. Proyecciones de trabajo futuro 73

7. Anexos 75

8. Referencias 76

1. Introducción

1.1 Contexto: la importancia de los registros poblacionales de cáncer

Los registros de cáncer de base poblacional (RCBP) son sistemas que recopilan información de todos los nuevos casos de cáncer que ocurren dentro de una población definida geográficamente, con el objetivo de informar sobre la tasa de incidencia por cáncer en su población. La información sobre estos casos proviene de diferentes fuentes, como hospitales, laboratorios, clínicas privadas, instituciones prestadoras de salud (IPS) y certificados de defunción. La función de los RCBP es recopilar, organizar y analizar información sobre los casos de cáncer para generar estadísticas que sirven de base para la vigilancia, la investigación y la evaluación de políticas epidemiológicas y de salud pública (Bray, 2015; Universidad de Caldas, 2025).

Los registros deben transformar la información en variables núcleo estandarizadas que preserven significado clínico y comparabilidad. Entre ellas se encuentran la topografía, entendida como el sitio anatómico primario del tumor; la morfología, que describe el tipo histológico; el comportamiento, que indica la naturaleza biológica en términos de malignidad, tumor *in situ*, benignidad o incertidumbre; y el grado, que resume el nivel de diferenciación tumoral. Todas se codifican con la Clasificación Internacional de Enfermedades para Oncología (CIE-O) (Fritz et al., 2003; Parkin & Bray, 2009).

A partir de estas necesidades, y ya que una gran parte de la información se mantiene como texto no estructurado, se consideró el uso de Procesamiento de Lenguaje Natural (PLN) para apoyar la conversión del documento al dato estructurado en el registro (Universidad de Caldas, 2025).

1.2. Planteamiento del problema: registros de defunción y su oportunidad

Los registros de cáncer de base poblacional funcionan a través de sedes que cubren áreas geográficas bien definidas y cuentan con información sobre el número y la distribución de la población por edad y sexo. En Colombia, esta red incluye la sede de Manizales. El Registro Poblacional de Cáncer de Manizales (RPCa-M), adscrito al Instituto de Investigaciones en Salud de la Universidad de Caldas, forma parte de esta red y tiene la función de identificar, consolidar y depurar la información de cáncer en residentes del municipio, tanto en su área urbana como rural, dicho registro, inició sus actividades en el año 2001, en el marco de una iniciativa liderada por el Instituto Nacional de Cancerología (Arias-Ortiz & Rodríguez-Betancourt, 2022; Universidad de Caldas, 2025)

El RPCa-M recibe información estructurada y no estructurada, por parte de las Instituciones Prestadoras de Servicios de salud (IPS), incluyendo laboratorios de patología, centros de diagnóstico por imágenes, centros de endoscopia, clínicas y hospitales que prestan servicios de diagnóstico o de tratamiento en cáncer, además estadísticas vitales que incluyen la información de los certificados de defunción. Cuando se trata de defunciones, el registro aplica un flujo de trabajo que inicia con la revisión de la base de datos de defunciones para identificar qué personas fallecieron con o por cáncer. Esta base de datos compila la información recogida por los médicos al momento del fallecimiento de una persona y contiene información sobre la persona quien falleció (identificadores, edad, sexo, etc.) y sobe la causa subyacente que llevó a la muerte de la persona y otras condiciones de salud presentes al momento de la muerte. En un siguiente paso, una vez se haya identificado personas quienes fallecieron con o por cáncer, el registro debe verificar si la persona fallecida era residente de Manizales y revisar la presencia de cáncer al momento de la muerte. Si es necesario, completa la información faltante mediante consultas a fuentes externas

autorizadas. Solo cuando los campos están completos y verificados se realiza la codificación estandarizada y se integra el registro en la base de datos matriz, conservando la trazabilidad documental y los controles de calidad (Universidad de Caldas, 2025).

Para el RPCa-M, uno de los principales desafíos es la gran cantidad de registros de defunción que deben revisarse para identificar posibles casos de cáncer. En la actualidad, esta revisión se realiza de forma manual, a partir de textos con redacción y nivel de detalle variables, lo que prolonga el proceso y puede introducir diferencias en la interpretación. Por ello, se propuso una estrategia de Procesamiento de Lenguaje Natural (PLN) orientada a clasificar registros de defunción con indicios de cáncer, como apoyo a la revisión, con el fin de mejorar la oportunidad del proceso (Universidad de Caldas, 2025).

1.3. Justificación: por qué un modelo de PLN es necesario

El PLN puede contribuir a mejorar el proceso mediante el cual el RPCa-M revisa registros de defunción para identificar posibles casos de cáncer. Dado que una parte importante de esta información se encuentra en texto libre, el uso de PLN permitiría reconocer menciones relevantes en la causa de muerte y en campos relacionados, clasificar los registros según indicios de cáncer y apoyar una revisión más oportuna y consistente. En este contexto, su valor no radica solo en acelerar el proceso, sino también en reducir el riesgo de que registros con una posible mención de cáncer sean clasificados erróneamente como ausencia de mención y queden por fuera de la revisión posterior (Bray & Parkin, 2009; Universidad de Caldas, 2025).

Para el registro, este tipo de apoyo también puede favorecer una mejor priorización del trabajo humano cuando aumenta el volumen de información. En lugar de destinar el mismo esfuerzo a todos los registros, una clasificación asistida permitiría concentrar la revisión detallada

en aquellos casos con mayor probabilidad de corresponder a cáncer o con evidencia textual dudosa. Esto favorecería la captación de potenciales casos y reduciría el riesgo de omitir registros relevantes. Además, el desarrollo de este proyecto podría servir como antecedente técnico para futuros usos de PLN en otros flujos de datos del registro, como reportes de patología, pruebas de imágenes y otras notificaciones clínicas, aportando aprendizajes útiles en contextos similares. En ese sentido, el RPCa-M podría fortalecer su capacidad para generar información más oportuna y útil para la vigilancia epidemiológica y la toma de decisiones en salud pública (García et al., 2018; Universidad de Caldas, 2025; Van Eycken et al., 2025).

1.4. Pregunta de investigación

A partir de lo anterior, se plantea la siguiente pregunta de investigación:

¿Cuál modelo de procesamiento de lenguaje natural permite clasificar con mejor desempeño los registros de defunción de Manizales según indicios de cáncer, para apoyar la revisión de potenciales casos por parte del Registro Poblacional de Cáncer de Manizales?

1.5. Objetivos

1.5.1 Objetivo general:

Diseñar y evaluar un modelo de procesamiento de lenguaje natural para clasificar registros de defunción de Manizales según indicios de cáncer, con el fin de apoyar la revisión de potenciales casos por parte del Registro Poblacional de Cáncer de Manizales.

1.5.2. Objetivos específicos:

- Preparar y estructurar la base de registros de defunción que sería utilizada para el desarrollo y evaluación de los modelos.
- Implementar y ajustar distintos enfoques de procesamiento de lenguaje natural para la clasificación de registros con indicios de cáncer.
- Evaluar el desempeño de los modelos mediante exactitud, precisión, sensibilidad, F1 y especificidad, con énfasis en la reducción de falsos 0.
- Determinar el modelo con mejor desempeño y analizar su utilidad como apoyo al proceso de revisión del RPCa-M.

2. Marco teórico

En esta sección se presentan los fundamentos teóricos que sustentan los modelos de procesamiento de lenguaje natural considerados en el estudio y los conceptos necesarios para comprender su funcionamiento general.

2.1. Modelo basado en reglas

Los modelos basados en reglas son un enfoque de procesamiento de lenguaje natural que clasifica a partir de conocimiento previamente definido. Ese conocimiento puede expresarse mediante diccionarios, palabras clave, patrones léxicos y reglas de contexto. Su principal ventaja es la interpretabilidad, ya que permiten identificar con claridad qué parte del texto llevó al sistema a tomar una decisión. Por esta razón, siguen siendo útiles en el análisis de texto clínico, especialmente en contextos donde resulta importante rastrear la evidencia textual asociada a cada clasificación (Dan Jurafsky & James H. Martin, 2025; Hua Xu & Demner Fushman, 2024).

Dentro de este enfoque, QuickUMLS es una herramienta que permite identificar términos biomédicos en el texto y relacionarlos con conceptos del UMLS. El UMLS es un sistema terminológico que reúne vocabularios médicos y ayuda a reconocer que distintas expresiones pueden referirse a una misma idea clínica. Por su parte, medSpaCy es una biblioteca orientada al procesamiento de texto clínico, construida sobre spaCy, que facilita tareas como detección de entidades, análisis de contexto y normalización terminológica. En conjunto, estas herramientas ayudan a estructurar el texto clínico y a detectar evidencia relevante de manera explícita (Hua Xu & Demner Fushman, 2024).

De forma general, si x representa el texto de un registro, el sistema puede detectar un conjunto de menciones $M(x) = \{m_1, \dots, m_k\}$. Después, por medio de reglas de contexto, se determina cuáles de esas menciones constituyen evidencia útil. Ese subconjunto puede representarse como $E(x) \subseteq M(x)$. A partir de esta evidencia, la clase predicha puede expresarse como $\hat{y} = f(x)$, donde f es una función de decisión basada en reglas. En términos prácticos, la clasificación final depende de la evidencia textual identificada y de la forma en que esta es interpretada por el sistema (Dan Jurafsky & James H. Martin, 2025).

2.2. Modelo SVM

Las *Support Vector Machines* (SVM) son modelos supervisados de clasificación orientados a separar clases mediante una frontera de decisión. Su principio central consiste en identificar el hiperplano que divide de mejor manera los datos y que, al mismo tiempo, maximiza la separación entre las clases. Esta propiedad ha favorecido su uso en tareas de clasificación de texto, especialmente cuando los documentos se representan en espacios de alta dimensión (Dan Jurafsky & James H. Martin, 2025; James, Witten, Hastie, & Tibshirani, 2023).

Para aplicar una SVM al texto, resulta necesario transformar los documentos en una representación numérica. Una de las más utilizadas es TF-IDF, que asigna mayor peso a los términos frecuentes en un documento, pero no en todo el corpus. También pueden emplearse n-gramas, entendidos como secuencias contiguas de $n$elementos, ya sean palabras o caracteres, los cuales permiten capturar patrones locales del lenguaje. De igual manera, pueden utilizarse *word embeddings*, que corresponden a representaciones vectoriales densas en las que palabras con significados semejantes tienden a ubicarse próximas entre sí. Adicionalmente, la selección de la

configuración más adecuada del modelo puede realizarse mediante *grid search*, estrategia que evalúa distintas combinaciones de hiperparámetros para identificar aquella con mejor desempeño (Dan Jurafsky & James H. Martin, 2025; Hua Xu & Demner Fushman, 2024; James, Witten, Hastie, & Tibshirani, 2023).

Desde el punto de vista matemático, si cada documento se representa como un vector $x_i \in \mathbb{R}^p$ y cada etiqueta como $y_i \in \{-1,1\}$, una SVM lineal busca una función de decisión de la forma

$$f(x) = w^\top x + b,$$

donde w es el vector de pesos y b es el sesgo. La clase predicha depende del lado del hiperplano en que quede la observación. En la versión de margen suave, el problema se formula como

$$\min_{w,b,\xi} \ \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i$$

sujeto a

$$y_i(w^\top x_i + b) \geq 1 - \xi_i, \xi_i \geq 0.$$

En esta formulación, $\| w \|^2$controla el tamaño del margen, ξ_i permite admitir algunos errores y C regula el equilibrio entre separar mejor las clases y penalizar errores de clasificación. Esto ayuda a explicar por qué las SVM funcionan bien en problemas de texto con muchas variables y datos dispersos (James, Witten, Hastie, & Tibshirani, 2023).

2.3. Modelo basado en *Transformers*

Los *Transformers* son arquitecturas de aprendizaje profundo diseñadas para procesar secuencias de texto mediante mecanismos de atención. A diferencia de enfoques previos basados en procesamiento estrictamente secuencial, estas arquitecturas pueden modelar relaciones entre distintos fragmentos del texto de manera simultánea. Esto permite construir representaciones contextuales más ricas, en las que el significado de una palabra depende del contexto completo en el que aparece. Por esta razón, los *Transformers* se han consolidado como una de las bases principales del procesamiento de lenguaje natural contemporáneo (Hua Xu & Demner Fushman, 2024; Vaswani et al., 2023).

Dentro de esta familia, RoBERTa es una variante de los modelos *encoder* preentrenados, orientada a aprender representaciones profundas del lenguaje a partir de grandes volúmenes de texto. Posteriormente, estas representaciones pueden ajustarse a tareas específicas, como clasificación o reconocimiento de entidades. En el ámbito biomédico y clínico en español, el desarrollo de versiones especializadas de este tipo de modelos ha permitido mejorar el desempeño en tareas de clasificación y extracción de información relevante (García Subies et al., 2024).

En este contexto también se emplean componentes auxiliares que favorecen el entrenamiento y la decisión final del modelo. Por una parte, *WeightedRandomSampler* es una estrategia de muestreo ponderado que incrementa la probabilidad de seleccionar ejemplos de clases menos frecuentes durante el entrenamiento, lo que resulta útil cuando existe desbalance entre clases. Por otra, argmax corresponde a la operación que selecciona el índice del valor más alto en el vector de salida y, en problemas de clasificación multiclase, equivale a escoger la clase con mayor puntuación estimada.

La base matemática de los *Transformers* es la función de atención:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V,$$

donde Q representa las consultas, K las claves, V los valores y d_k la dimensión de las claves. Esta operación permite calcular cuánto debe atender cada elemento del texto a los demás elementos de la secuencia. Después de varias capas de transformación, el modelo genera una representación contextual del texto y produce un vector de puntaje $z = (z_1, \ldots, z_C)$. Esos puntajes pueden convertirse en probabilidades mediante

$$p(y = c \mid x) = \frac{e^{z_c}}{\sum_{j=1}^{C} e^{z_j}}.$$

La predicción final se obtiene como

$$\hat{y} = \arg\max_{c} \; p(y = c \mid x).$$

Esto significa que el modelo asigna la clase con mayor probabilidad estimada a partir de la representación contextual del texto (Hua Xu & Demner Fushman, 2024; Vaswani et al., 2023).

3. Revisión de la literatura

3.1. Resumen

Objetivo: Identificar las tareas de procesamiento de lenguaje natural clínico aplicadas en registros poblacionales de cáncer, así como los algoritmos, tipos de documento y procesos del registro en los que se han utilizado.

Métodos: Se realizó una revisión de alcance en Scopus, BVS, Web of Science y EBSCO. Se incluyeron estudios publicados entre 2015 y 2025 que aplicaran PLN clínico sobre texto en el flujo de trabajo de registros poblacionales de cáncer. Tras el cribado, 53 artículos fueron evaluados a texto completo y 48 se incluyeron en el análisis.

Resultados: Predominaron los estudios multitarea (28/48), los *pipelines* propios o adaptados localmente (42/48) y los objetivos de clasificación (47/48). Los documentos usados en los ejercicios más frecuentes fueron reportes patológicos (39), seguidos por notas clínicas (8) y reportes de imágenes (4); los certificados de defunción aparecieron en 2 estudios. En clasificación de texto, los algoritmos más frecuentes fueron SVM y *rule-based*, con 11 y 10 menciones cada uno.

Conclusión: La literatura se ha concentrado en apoyar decisiones operativas iniciales del registro mediante clasificación de texto y extracción de información. Sin embargo, predominan soluciones híbridas y adaptadas al contexto local. Además, los certificados de defunción fueron poco frecuentes en los estudios revisados, lo que muestra una oportunidad para desarrollar modelos de PLN aplicados a este tipo de documento.

3.2. Introducción

Los registros de cáncer de base poblacional son una herramienta fundamental para la vigilancia epidemiológica y para la producción de información útil en salud pública. Su funcionamiento depende de la recolección, revisión y organización de datos provenientes de distintas fuentes clínicas, como reportes de patología, certificados de defunción, notas médicas y otros documentos narrativos que contienen información relevante sobre los casos. Sin embargo, una parte importante de estos datos se encuentra en texto libre, lo que hace que su procesamiento requiera tiempo, esfuerzo y una revisión detallada por parte del personal del registro (Bray, 2015).

En este contexto, el procesamiento de lenguaje natural (PLN) clínico ha despertado interés como apoyo para transformar información textual en datos estructurados útiles para la operación del registro. Estas técnicas pueden facilitar la identificación de casos, la extracción de variables clínicas y la recuperación de información relevante a partir de documentos heterogéneos (Bray, 2015; Hua Xu & Demner Fushman, 2024). No obstante, su aplicación en registros poblacionales de cáncer ha sido diversa en términos de tareas analíticas, algoritmos utilizados y condiciones de implementación. Por ello, la pregunta que se busca responder en esta revisión de alcance es ¿Cuáles son las tareas de procesamiento de lenguaje natural clínico usadas en los registros poblacionales de cáncer?

3.3. Antecedentes

En esta revisión de alcance se analiza el uso de tareas de procesamiento de lenguaje natural (PLN) para transformar información textual procedente de los registros poblacionales de cáncer en datos estructurados útiles para la vigilancia epidemiológica y la gestión del registro.

Caracterizar qué tareas se han aplicado, cómo se han implementado y con qué propósitos permite comprender de qué manera el PLN clínico puede contribuir a reducir la carga de trabajo manual, mejorar la calidad del dato y favorecer la comparabilidad entre registros. Para organizar este análisis, la revisión se estructura a partir de una taxonomía que reúne las principales dimensiones conceptuales y metodológicas identificadas.

En este marco, las dimensiones de tarea analítica y algoritmo se definieron a partir de referentes consolidados en PLN biomédico y aprendizaje estadístico. Las tareas analíticas se toman de la clasificación propuesta en *Natural Language Processing in Biomedicine* e incluyen reconocimiento de entidades, extracción de relaciones, normalización de conceptos, clasificación de texto, recuperación de información, preguntas y respuestas, generación de texto y otras tareas reportadas en los estudios (Hua Xu & Demner Fushman, 2024). Por su parte, la dimensión de algoritmo se construye con base en los métodos de PLN descritos en *Natural Language Processing in Biomedicine* y se organiza siguiendo las bases conceptuales de *Speech and Language Processing* y *An Introduction to Statistical Learning* (Dan Jurafsky & James H. Martin, 2025; Hua Xu & Demner Fushman, 2024; James, Witten, Hastie, & Taylor, 2023). En esta categoría se consideran *Linear Regression, Logistic Regression, Support Vector Machine (SVM), Naive Bayes (NB), n-gram language models, Decision Trees, Random Forest (RF), Agglomerative Hierarchical Clustering (AHC), Word Embeddings, Hidden Markov Models (HMM), Conditional Random Fields (CRF), Hybrid BiLSTM–CRF models, Viterbi algorithm, Multilayer Perceptron/Deep Neural Network (MLP/DNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Attention mechanisms, Transformers, Inverted Index, Vector Space Model (VSM), TF–IDF, BM25, Greedy Decoding, Beam Search* y otros algoritmos.

Además, la dimensión de aplicación de PLN clínico incluye entornos como cTAKES, CLAMP, GATE, SciSpaCy, TextFlow, IAPPS Grid, OpenMinTed, MedSpaCy, pipeline propio y otras aplicaciones (Digan et al., 2021). A su vez, el enfoque se definió para esta revisión y distingue entre estudios de tarea única y de multitarea analítica. De manera complementaria, la topografía del cáncer se apoya en clasificaciones generales usadas en vigilancia oncológica y contempla tres categorías: todos los sitios como topografía, solo sólidos y solo hematológicos (Bray, 2015). Asimismo, el tipo de documento de donde sale el dato comprende certificado de defunción, reporte patológico, egresos hospitalarios, reporte de imágenes, reportes de laboratorio, autopsias, sistemas de notificación obligatoria, notas quirúrgicas, notas clínicas y datos administrativos o de seguros (Bray, 2015).

Finalmente, para ubicar los hallazgos dentro de la operación del registro, el proceso dentro del registro incluye identificar si un reporte menciona cáncer, extracción de elementos clave, identificación de la procedencia del caso, extracción de biomarcadores, identificación de recurrencia o metástasis, identificación de primarios múltiples, seguimiento y otros procesos (Hsu et al., 2024). Además, la cantidad de datos que ingresan se toma directamente de cada artículo y corresponde al número de datos reportados por los autores. Por último, el objetivo clínico contempla dos categorías: regresión y clasificación. La Tabla 1 presenta la taxonomía construida para analizar la revisión de la literatura a partir de los conceptos descritos.

Tabla 1: Taxonomía revisión de la literatura

Dimensión	**Características**				
Tarea analítica	Reconocimiento de entidades (NER)	Extracción de relaciones (RE)	Normalización de conceptos	Clasificación de texto	Recuperación de información
	Preguntas y respuestas (QA)	Generación de texto	Otro		

Algoritmo	Linear regression	Logistic regression	Support Vector Machine (SVM)	Naive Bayes (NB)	n-gram language models
	Decision trees	Random Forest (RF)	Agglomerative hierarchical clustering (AHC)	CHAR embeddings	Hidden Markov Models (HMM)
	Conditional Random Fields (CRF)	Hybrid BiLSTM–CRF models	Viterbi algorithm	Multilayer perceptron/deep neural network (MLP/DNN)	Convolutional Neural Networks (CNN)
	Recurrent Neural Networks (RNN)	Long Short-Term Memory (LSTM)	Bidirectional LSTM (BiLSTM)	Attention mechanisms	Transformers
	Inverted index	Vector space model (VSM)	TF–IDF	BM25	Greedy decoding
	Beam search	Otro			
Aplicación de PLN clínico	cTakes	CLAMP	GATE	ScispaCy	TextFlow
	LAPPS Grid	OpenMinTed	MedSpacy	pipeline propio	otro
Enfoque	Tarea única	Multitarea analítica			
Sitio del cáncer	Todos los sitios como topografía	Solo solidos	Solo hematológicos		
Tipo de documento de donde sale el dato	Certificado de defunción	Reporte patológico	Egresos hospitalarios	Reporte de imágenes	Reportes de laboratorio
	Autopsias	Sistemas de notificación obligatoria	Notas quirúrgicas	Notas clínicas	Datos administrativos o de seguros
Proceso dentro del registro	Identificar si un reporte menciona cáncer	Extracción de elementos clave	Identificación de la procedencia del caso	Extracción de información de biomarcadores	Identificación de la recurrencia o metástasis
	Identificación de primarios múltiples	Seguimiento	Otros		
Cantidad de datos que ingresan al modelo	Número de datos ingresados (según el artículo)				
Objetivo clínico	Regresión	Clasificación			

3.4. Materiales y métodos

Se realizó una búsqueda bibliográfica en Scopus, Biblioteca Virtual en Salud (BVS), Web of Science y EBSCO, mediante una ecuación de búsqueda en inglés, español y portugués que combinó términos de PLN clínico con términos asociados a registros de cáncer y excluyó expresiones relacionadas con registros hospitalarios. Adicionalmente, de forma complementaria, se incorporaron algunos artículos identificados mediante validación cruzada, con el fin de fortalecer la exhaustividad de la revisión.

Se incluyeron artículos publicados entre 2015 y 2025, identificados en las bases de datos como artículos científicos finalizados, orientados a registros poblacionales de cáncer o a procesos directamente relacionados con su operación, y que propusieran, desarrollaran o evaluaran aplicaciones de PLN clínico sobre texto. Además, los estudios debían aportar información suficiente para extraer al menos una tarea analítica, como reconocimiento de entidades, extracción de relaciones, normalización de conceptos, clasificación de texto, recuperación de información, preguntas y respuestas, generación de texto u otra tarea reportada; así como datos sobre el algoritmo utilizado, la aplicación o pipeline de PLN clínico, el enfoque del estudio, el sitio del cáncer, el tipo de documento analizado, el proceso del registro al que aportaba, la cantidad de datos ingresados al modelo y el tipo de predicción clínica, cuando esta información estuviera disponible. Se excluyeron estudios que no cumplieran con los metadatos mínimos para la extracción; que no estuvieran orientados a registros poblacionales de cáncer ni a procesos, documentos o variables útiles para estos registros; que no utilizaran texto clínico del registro o de fuentes clínicas relacionadas; que no aplicaran PLN sobre texto, como trabajos centrados únicamente en imágenes, señales o datos estructurados; que analizaran fuentes no clínicas sin vínculo clínico claro, como redes sociales, noticias o foros; que no incluyeran una tarea analítica identificable; o que se enfocaran exclusivamente en condiciones no neoplásicas, como diabetes o enfermedades cardiovasculares.

Después de eliminar duplicados y realizar el filtrado por título y resumen según los criterios de inclusión y exclusión, se identificaron 54 estudios potencialmente relevantes para evaluación a texto completo. Esta fase fue realizada por un grupo de 8 expertos, integrado por 4 profesionales del sector de registros poblacionales de cáncer y 4 profesionales de ingeniería con experiencia en PLN clínico. Para la revisión, el grupo se organizó en parejas conformadas por un experto de cada

perfil (RCBP e ingeniero), y entre estas parejas se llevó a cabo la evaluación de los textos completos y la extracción de la información. Finalmente, se incluyeron 48 estudios y se excluyeron 6 por no ajustarse al propósito de la revisión (ver Figura 1).

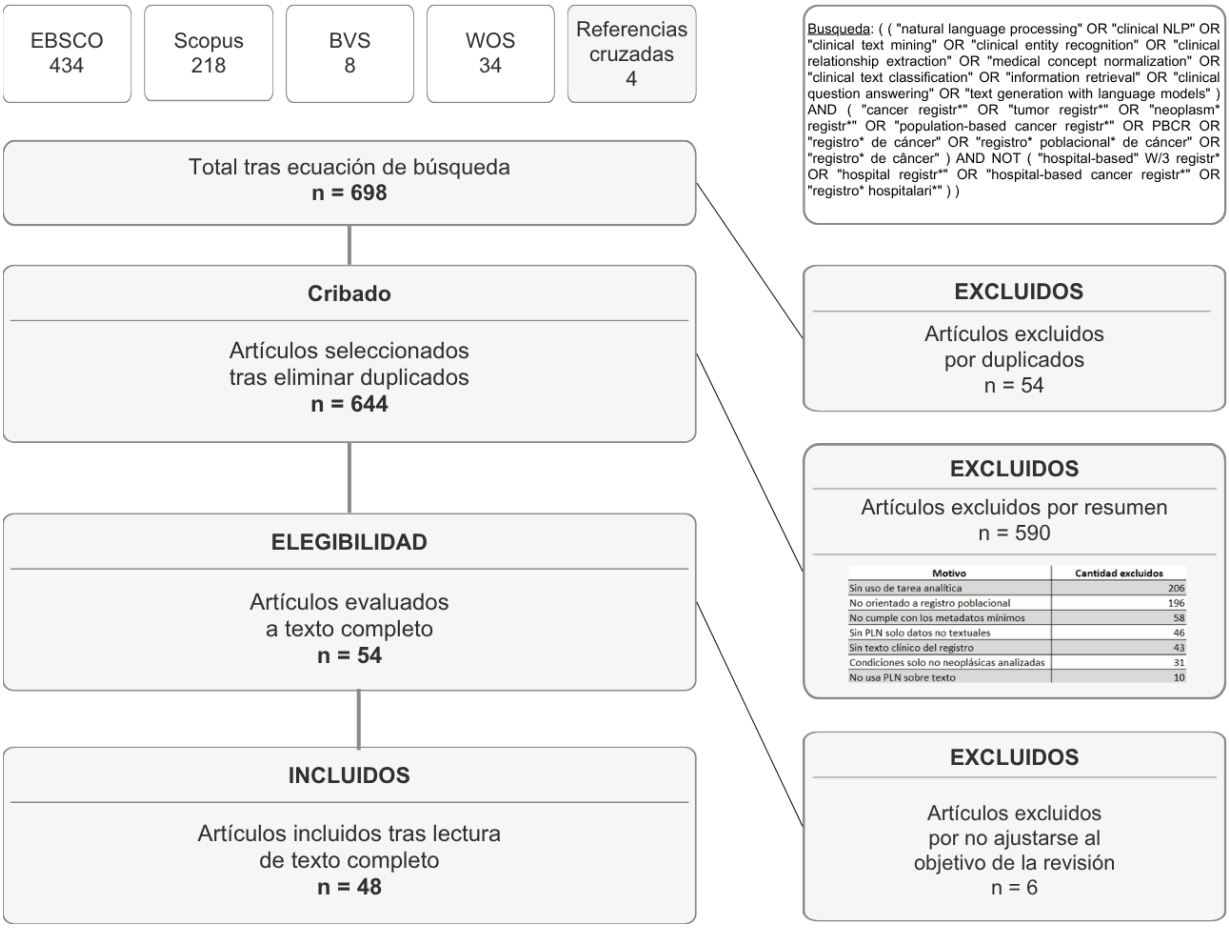

Figura 1. Diagrama de flujo PRISMA del proceso de selección de artículos

***Nota**. Elaboración propia a partir de la revisión de alcance.*

Los estudios excluidos en esta fase correspondieron, en su mayoría, a trabajos que, aunque utilizaban PLN en contexto oncológico, se enfocaban en seguimiento clínico, recurrencia, metástasis, biomarcadores o aspectos técnicos de infraestructura, sin centrarse en tareas analíticas aplicadas directamente a los procesos del registro poblacional de cáncer. La extracción de datos se realizó mediante una plantilla estructurada en Microsoft Excel. Para cada registro se consignaron

variables bibliográficas y, en los estudios incluidos, variables analíticas relacionadas con la taxonomía definida para la revisión. La base consolidada de artículos (*Articulos.xlsx*), presentada en el Anexo 1, se utilizó como insumo para el análisis descriptivo reportado en la sección de resultados.

3.5. Resultados

A partir de los 48 estudios incluidos, se observó que la literatura analizada fue diversa en sus enfoques, en la cantidad de datos utilizados y en los tipos de documentos estudiados, aunque también mostró algunos patrones comunes. Predominaron los estudios con enfoque de multitarea analítica (28/48), al igual que aquellos que desarrollaron *pipelines* propios o adaptados localmente (42/48). Además, 39 estudios reportaron el uso de algoritmos adicionales a los principales, lo que sugiere que fue frecuente combinar varias estrategias dentro de un mismo trabajo. En cuanto al alcance oncológico, hubo una distribución similar entre estudios orientados a todos los sitios topográficos (26/48) y estudios centrados en tumores sólidos (22/48). Estos últimos a veces incluyeron cualquier tumor sólido, y otras veces algún tipo o grupo de cánceres en particular. La predicción clínica estuvo dirigida casi por completo a tareas de clasificación (47/48), con una sola aplicación de regresión. Finalmente, el volumen de datos analizado fue muy variable, con tamaños entre 457(Xie et al., 2017) y 2,923,291(Peluso et al., 2024) registros o documentos, lo que muestra diferencias importantes en la escala de aplicación.

En cuanto al tipo de documento de donde se obtuvo la información, predominaron los reportes patológicos, con 39 menciones, seguidos por las notas clínicas (8) y los reportes de imágenes (4). Con menor frecuencia se identificaron datos administrativos o de seguros (3), sistemas de notificación obligatoria (2), reportes de laboratorio (2) y certificados de defunción (2),

así como autopsias, con una sola mención. La Figura 2 muestra la distribución del tipo de documento de donde se obtuvo el dato en los estudios incluidos.

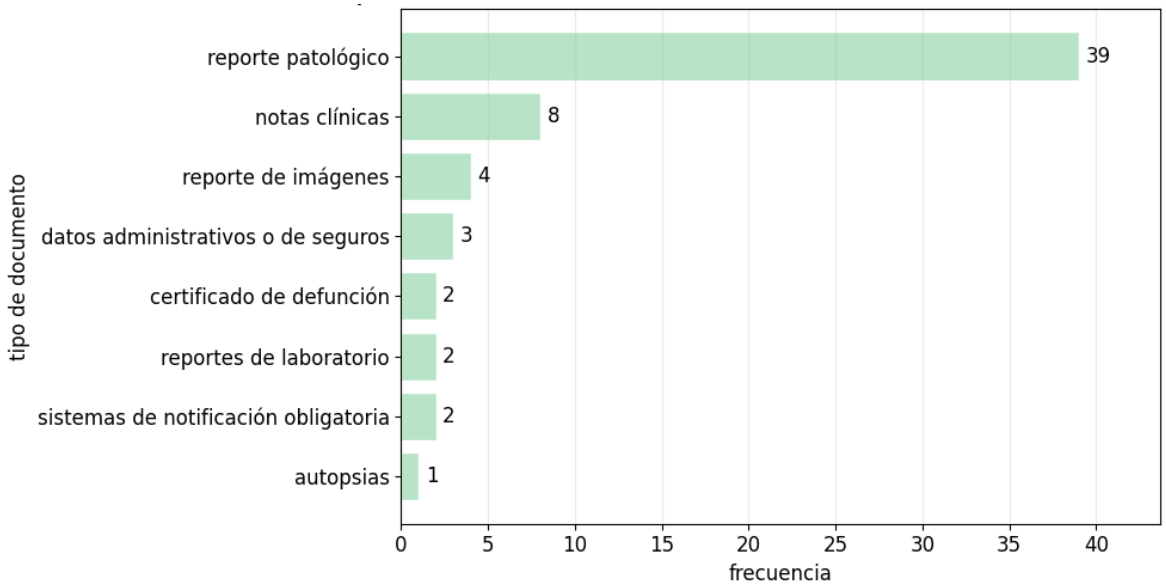

Figura 2: Tipo de documento de donde se obtuvo la información

***Nota**. Elaboración propia a partir de la revisión de alcance.*

Para la tarea de clasificación de texto, los algoritmos más frecuentes fueron *Support Vector Machine* (SVM), con 11 menciones, y los enfoques *rule-based*, con 10 menciones, seguidos por *Transformers* (8), *word embeddings* (8) y *regular expressions* (6). En reconocimiento de entidades (NER), predominaron los enfoques *rule-based* (8), seguidos por *regular expressions* (7), árboles de decisión (4), *Adversarial Domain Adaptation* (4) *y Conditional Random Fields* (CRF) (4). En normalización de conceptos, los enfoques *rule-based* fueron nuevamente los más frecuentes (7), seguidos por SVM (4); adicionalmente, se identificaron mecanismos de atención (2), *bag-of-words* (1) y *Conditional Random Fields* (CRF) (1). En recuperación de información, los algoritmos más frecuentes fueron *word embeddings* (3), seguidos por modelos de lenguaje basados en *n-gramas* (2) y enfoques *rule-based* (2), además de *Adversarial Domain Adaptation* (1) y un algoritmo NegEx modificado (1). En extracción de relaciones, se observaron principalmente enfoques *rule-*

based (2) y *Transformers* (2), seguidos por mecanismos de atención (1), mecanismos de atención basados en *self-attention* y *target-attention* (1) y *Conditional Random Fields* (CRF) (1). En conjunto, los estudios incluidos reportaron el uso concurrente de enfoques basados en reglas, modelos clásicos supervisados y métodos más recientes basados en arquitecturas profundas. En esta misma línea, (Viscosi et al., 2022) reportaron que LightGBM fue el algoritmo con mejor desempeño en su comparación de ocho modelos, con un F1 de 98.9%, mientras que (Schulz et al., 2021) informaron mejores resultados con una solución híbrida que combinó aprendizaje automático, mapeo terminológico y reglas. La figura 3 muestra la combinación entre tareas analíticas y algoritmos en los estudios de RCBP.

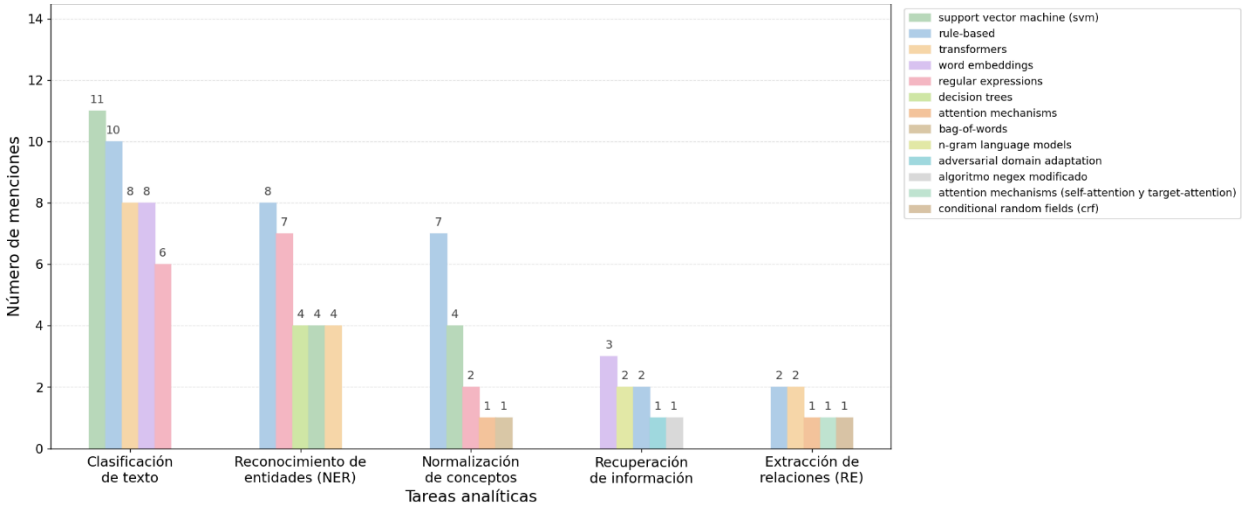

Figura 3: Grafica de top 5 algoritmos por tarea analítica de top 5 tareas analíticas

***Nota**. Elaboración propia a partir de la revisión de alcance.*

En cuanto a los procesos operativos del registro en los que se aplicó PLN clínico, se observó que la literatura se concentró principalmente en la extracción de elementos clave. El proceso más frecuente fue la extracción de elementos clave (33), seguido por la identificación de si un reporte menciona cáncer (20). Con menor frecuencia se encontraron la identificación de recurrencia o metástasis (6), otros procesos (3), la extracción de biomarcadores (5), la identificación de la procedencia del caso (4), el seguimiento (3) y la identificación de primarios

múltiples (1). En conjunto, esta distribución mostró una mayor concentración de estudios en etapas del flujo del registro relacionadas con la detección, priorización y estructuración de información oncológica relevante. En este contexto, se describieron desarrollos de PLN capaces de extraer términos, *n-grams*, códigos SNOMED CT y propiedades ICD-O-3 a partir de certificados de defunción, mientras que también se señaló que solo una fracción de los reportes histopatológicos contiene información útil para la evaluación oncológica del registro (Koopman et al., 2018; Viscosi et al., 2022).

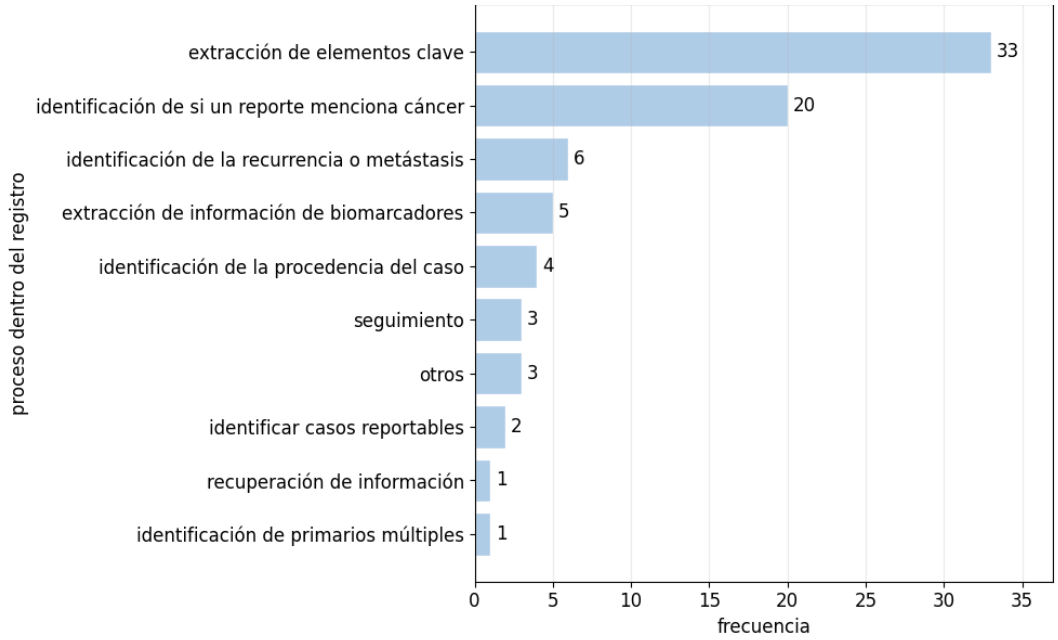

Figura 4: Frecuencia de proceso dentro del registro

***Nota**. Elaboración propia a partir de la revisión de alcance.*

Debe tenerse en cuenta que un mismo estudio podía aportar varias tareas analíticas dentro de un mismo proceso del registro. Adicionalmente, la categoría “otras tareas” agrupa tareas complementarias no incluidas entre las principales mostradas en la Figura 5. Bajo esta estructura de conteo, la clasificación de texto fue la tarea más frecuente en todos los grupos analizados. En la extracción de elementos clave, esta tarea registró 26 menciones, seguida por otras tareas (19), reconocimiento de entidades (17), normalización de conceptos (11), recuperación de información

(5) y extracción de relaciones (5). En la identificación de si un reporte menciona cáncer, la clasificación de texto también ocupó el primer lugar, con 20 menciones, seguida por normalización de conceptos (10), reconocimiento de entidades (9), otras tareas (8), recuperación de información (5) y extracción de relaciones (2). En la identificación de recurrencia o metástasis, la clasificación de texto volvió a ser la tarea predominante, con 6 menciones, seguida por normalización de conceptos (3), reconocimiento de entidades (3), recuperación de información (2), otras tareas (2) y extracción de relaciones (1). En los demás procesos dentro del registro, la clasificación de texto se mantuvo como la tarea más frecuente, con 17 menciones, seguida por reconocimiento de entidades (9), normalización de conceptos (9), otras tareas (9), extracción de relaciones (6) y recuperación de información (5). En conjunto, estos resultados muestran que la clasificación de texto operó como la tarea analítica más transversal, usualmente articulada con reconocimiento de entidades y normalización de conceptos. Este patrón es consistente con estudios orientados a extraer variables específicas desde documentos narrativos (Bozkurt et al., 2022; Hammami et al., 2021; Koopman et al., 2015).

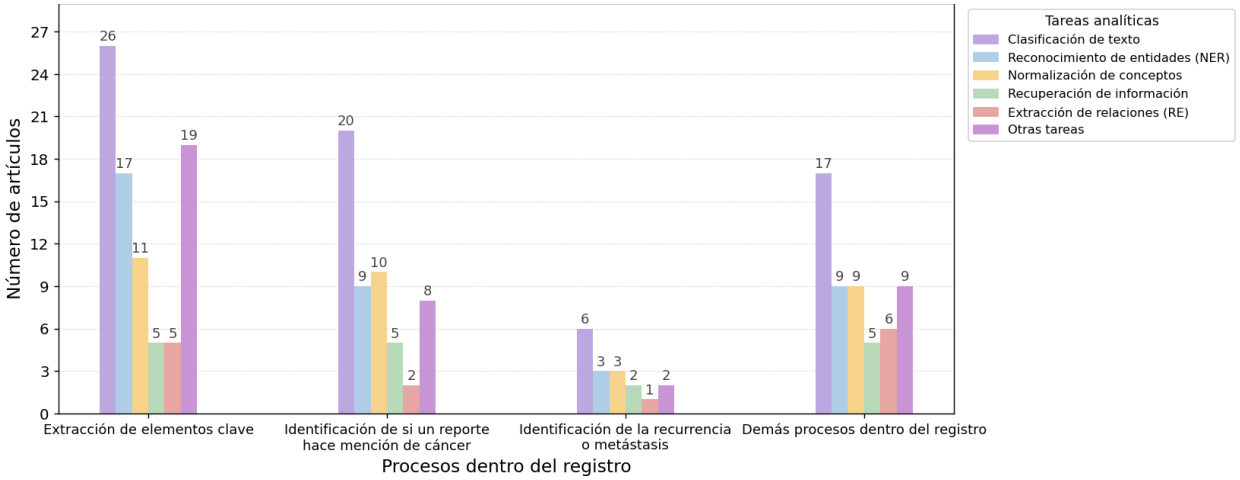

Figura 5: Top tareas analíticas por proceso dentro del registro

***Nota**. Elaboración propia a partir de la revisión de alcance.*

Las gráficas complementarias asociadas a los resultados de esta revisión se incluyen en el Anexo 2.

Además de las tareas, procesos y algoritmos identificados, los estudios incluidos reportaron de forma recurrente características del contexto de aplicación del PLN clínico en RCBP. Entre ellas, la heterogeneidad lingüística y documental fue uno de los hallazgos más frecuentes. Se describieron reportes en francés y neerlandés, en ocasiones combinados en un mismo documento, junto con variaciones en abreviaturas, estilos de escritura, puntuación y uso de mayúsculas (Pironet et al., 2021). De manera similar, varios estudios documentaron la necesidad de adaptar localmente herramientas y recursos lingüísticos. Se señaló que diversos sistemas de PLN desarrollados previamente estaban orientados a contextos angloparlantes y no se ajustaban al italiano (Viscosi et al., 2022); en la misma línea, se trabajó con reportes italianos (Hammami et al., 2021) y también se aportó evidencia en reportes de patología oncológica en español (Mendoza-Urbano et al., 2023). También se reportaron hallazgos de carácter operativo. En el Registro Integrado de Cáncer Catania–Messina–Enna, durante 2020 los operadores del registro realizaron 5,725 búsquedas en el portal web para revisar reportes de patología, de las cuales solo 1,127 fueron finalmente importadas, con una dedicación mínima estimada de 80.5 horas (Viscosi et al., 2022). Asimismo, en un estudio basado en historias clínicas electrónicas, más del 20% de los pacientes presentaba información de estadio faltante antes de aplicar el modelo, y el PLN permitió imputar categorías de estadio clínico T, N y M en 24%, 21% y 32% de los registros faltantes, respectivamente; para el estadio patológico, permitió imputar 71% de los T faltantes y 56% de los N faltantes (Bozkurt et al., 2022). Finalmente, varios estudios describieron limitaciones relacionadas con la evaluación y la calidad de los datos. Se reportaron discrepancias asociadas a diferencias en los flujos de trabajo entre sitios, desacuerdos entre contenido textual y contenido codificado, y manejo variable de

valores faltantes (Schulz et al., 2021); además, se señalaron posibles errores en el conjunto de referencia derivados de procesos manuales y automáticos de vinculación (Viscosi et al., 2022), así como una posible vulnerabilidad de privacidad por identificación de membresía en modelos de aprendizaje profundo entrenados con reportes clínicos (Alawad et al., 2020).

3.6. Discusión

Los hallazgos de esta revisión muestran que el uso de PLN clínico en RCBP se ha concentrado principalmente en tareas orientadas a identificar, clasificar y estructurar información relevante a partir de texto libre. En este contexto, la clasificación de texto fue la tarea más transversal, tanto por su frecuencia como por su presencia en distintos procesos del registro, especialmente en la extracción de elementos clave y en la identificación de reportes con mención de cáncer. Esto sugiere que, hasta ahora, el principal aporte del PLN clínico en este campo ha sido apoyar decisiones operativas iniciales del registro, como priorizar documentos, detectar casos potenciales y recuperar información útil para la abstracción oncológica (Bozkurt et al., 2022; Hammami et al., 2021; Koopman et al., 2018; Viscosi et al., 2022).

Otro hallazgo importante fue el predominio de soluciones híbridas y adaptadas localmente, en concordancia con el predominio de *pipelines* propios o adaptados al contexto local observado en los resultados. Los estudios revisados no mostraron un reemplazo completo de los enfoques basados en reglas por modelos más recientes, sino una coexistencia entre reglas, modelos clásicos supervisados y métodos de aprendizaje profundo, en línea con lo reportado en estudios que compararon o integraron diferentes estrategias metodológicas (Schulz et al., 2021; Viscosi et al., 2022). Este patrón parece estar relacionado con la heterogeneidad de los documentos clínicos, la

variación lingüística y las diferencias entre contextos institucionales, aspectos también descritos en reportes desarrollados en distintos idiomas y entornos de aplicación (Hammami et al., 2021; Mendoza-Urbano et al., 2023; Pironet et al., 2021).

En conjunto, estos resultados sugieren que el campo ha avanzado más en aplicaciones funcionales y específicas que en soluciones estandarizadas y fácilmente transferibles entre RCBP. Aunque el PLN clínico mostró utilidad para reducir carga manual, extraer variables y apoyar procesos del registro, persisten retos relacionados con la comparabilidad entre estudios, la validación en diferentes contextos, la calidad del dato y la adaptación multilingüe, limitaciones que también han sido señaladas en trabajos previos (Alawad et al., 2020; Schulz et al., 2021; Viscosi et al., 2022). En este sentido, futuras investigaciones podrían enfocarse en fortalecer la validación externa, mejorar la interoperabilidad de los *pipelines* y ampliar el desarrollo de aplicaciones en procesos menos explorados, como seguimiento, recurrencia, primarios múltiples y manejo de biomarcadores.

3.7. Conclusiones

Esta revisión de alcance muestra que las tareas de PLN clínico usadas en los RCBP se concentran principalmente en la clasificación de texto, el reconocimiento de entidades, la normalización de conceptos, la recuperación de información y, en menor medida, la extracción de relaciones. En conjunto, estas tareas se han aplicado sobre todo para identificar reportes relevantes, extraer elementos clave y estructurar información oncológica útil para la operación del registro.

La literatura revisada también muestra que su implementación ha dependido, en gran medida, de soluciones híbridas y adaptadas al contexto documental, lingüístico y operativo de cada

registro. Esto sugiere que el avance del campo requiere no solo el desarrollo de modelos más complejos, sino también el fortalecimiento de la validación, la comparabilidad entre estudios y la integración de estas herramientas en distintos entornos de registros.

Además, la baja frecuencia de estudios centrados en certificados de defunción muestra una oportunidad para desarrollar modelos de PLN aplicados a este tipo de documento, especialmente en contextos donde la revisión manual representa una carga operativa importante para los registros.

4. Metodología

En esta sección se describe el procedimiento metodológico que se siguió para preparar los datos de defunciones, desarrollar los modelos y evaluar su desempeño en la clasificación de registros de defunción con indicios de cáncer.

4.1 Enfoque metodológico

Este estudio se desarrolló con base en la investigación basada en diseño, ya que parte de una necesidad real del RPCa-M y propone como respuesta la construcción y evaluación de un artefacto computacional orientado a apoyar la revisión de registros de defunción. En este caso, el artefacto correspondió un modelo de PLN para clasificar registros según indicios de cáncer (González & Pomares, 2012).

De manera práctica, esta metodología se tradujo en cinco momentos principales: selección de la muestra etiquetada, seudoanonimización de la base, limpieza y normalización del texto, desarrollo de tres modelos de clasificación y comparación final de su desempeño. Todo el proceso se orientó por un mismo criterio: reducir al máximo los falsos 0, es decir, evitar que registros con mención clara o sospecha de cáncer fueran clasificados como ausencia de mención. Este criterio es coherente con la función del registro de captar de manera exhaustiva la totalidad de casos potencialmente registrables, y con la necesidad de mejorar la oportunidad y la calidad del proceso de revisión (Parkin & Bray, 2009; Universidad de Caldas, 2025).

4.2. Fuente de datos y construcción de la muestra etiquetada

La base de partida corresponde a una base cruda de 39.297 registros y 125 columnas. Antes del proceso de seudoanonimización, se seleccionó una muestra aleatoria de 4.000 registros, equivalente aproximadamente al 10% de la base total. Esta muestra se definió de manera que los diferentes grupos de edad (defunciones en menores de 19 años, y de 19 y más años) quedaran representados, con el fin de conservar una distribución etaria cercana a la de la base original.

Posteriormente, esta muestra fue revisada y etiquetada por empleados del registro de cáncer. A cada caso se le asignó una de tres clases: 0, cuando no había mención de cáncer; 1, cuando existió una mención clara de cáncer; y 2, cuando el texto sugería una mención probable o incierta. En términos operativos, la clase 2 correspondió a los casos ambiguos, es decir, aquellos registros en los que la evidencia textual no permitía una asignación suficientemente clara a las otras dos clases, pero que, precisamente por esa ambigüedad, debían mantenerse dentro del proceso de revisión para no descartar posibles casos relevantes. Esta base etiquetada fue la utilizada para el entrenamiento y la evaluación de los modelos. El uso de conjuntos anotados por expertos fue consistente con las prácticas descritas en el PLN clínico, donde la calidad del conjunto de referencia es fundamental para la validez del modelo (Hua Xu & Demner Fushman, 2024).

4.3. Selección de variables textuales y seudoanonimización

A partir de la base original se conservaron únicamente las variables textuales relevantes para la tarea de clasificación. En particular, se conservaron las columnas Diagnóstico A, Diagnóstico B, Diagnóstico C, Diagnóstico D, Otros Estados Patológicos y Otros Estados Patológicos 2, ya que en ellas se concentraba la evidencia clínica más útil para identificar presencia

o sospecha de cáncer. Esta decisión también permitió reducir ruido y asegurar que los tres modelos fueran comparados sobre el mismo insumo textual.

Después de esta selección se realizó la seudoanonimización de la base. Para ello, los identificadores directos se separaron de la base analítica y se reemplazaron por un identificador interno secuencial, ID_PSEUDO, conservando aparte una tabla llave. De esta forma, el procesamiento posterior se realizó sin exponer datos personales, pero manteniendo la posibilidad de trazabilidad técnica dentro del estudio. Este manejo fue consistente con los principios de calidad, comparabilidad y protección del dato que orientan el trabajo de los registros de cáncer y la reglamentación (habeas data) de Colombia (Bray & Parkin, 2009; Michalek et al., 2024).

4.4. Limpieza, normalización y estandarización del texto

Una vez construida la base analítica, se aplicó un proceso de limpieza y normalización sobre las columnas seleccionadas. El propósito de esta etapa fue corregir problemas de codificación, unificar variantes ortográficas y reducir diferencias superficiales de escritura que podían afectar el desempeño de los modelos. En el contexto del PLN clínico, este tipo de preparación es importante porque los textos médicos suelen contener abreviaturas, errores de digitación, caracteres dañados y variaciones terminológicas que dificultan la clasificación automática (Dan Jurafsky & James H. Martin, 2025; Hua Xu & Demner Fushman, 2024).

La normalización incluyó la corrección de tildes dañadas, caracteres mal codificados y errores ortográficos frecuentes. También se estandarizaron abreviaturas clínicas comunes. Por ejemplo, cuando aparecía la abreviatura “CA”, esta se transformaba a “cáncer”. De igual manera, variantes como “metastasis” se llevaron a “metástasis”, y expresiones equivalentes se unificaron

en una sola forma escrita. El objetivo no fue cambiar el sentido clínico del texto, sino dejar expresiones equivalentes escritas de manera consistente para facilitar su procesamiento posterior.

4.5. Construcción del texto de entrada

Después de la normalización, los campos seleccionados se consolidaron en una sola entrada textual por registro. Así, cada caso quedó representado por un texto unificado que reunía la información disponible en diagnósticos y estados patológicos. Esta decisión permitió trabajar con una misma unidad de análisis para los tres modelos y asegurar que las diferencias observadas entre ellos respondieran principalmente al enfoque analítico de cada alternativa y no a cambios en la estructura de entrada (Dan Jurafsky & James H. Martin, 2025; Hua Xu & Demner Fushman, 2024).

4.6. Desarrollo de los modelos

Este apartado presenta el desarrollo de los tres modelos de clasificación que se compararon en el estudio, todos construidos sobre una misma entrada textual y bajo un mismo criterio de evaluación. Con el fin de facilitar la trazabilidad técnica del desarrollo realizado, los enlaces a los repositorios correspondientes a los tres modelos comparados se incluyen en el Anexo 3.

La elección de estos tres enfoques se basó tanto en criterios técnicos como en los hallazgos identificados en la revisión de literatura. Los modelos basados en reglas permitieron contar con una alternativa interpretable, cercana a los flujos de decisión propios del registro. Por su parte, el modelo SVM se incluyó debido a su uso frecuente en tareas de clasificación de texto y a su buen desempeño con representaciones dispersas y de alta dimensión, como TF-IDF.

Finalmente, el modelo basado en Transformers se evaluó como una alternativa de aprendizaje profundo con capacidad para analizar el contexto y reconocer relaciones semánticas más complejas dentro del texto. Así, la comparación no se planteó como una selección arbitraria de algoritmos, sino como una evaluación aplicada de tres familias metodológicas relevantes dentro del procesamiento de lenguaje natural clínico.

Los tres modelos utilizaron como entrada las mismas variables textuales: Diagnóstico A, Diagnóstico B, Diagnóstico C, Diagnóstico D, Otros Estados Patológicos y Otros Estados Patológicos 2. Estas columnas fueron limpiadas, normalizadas y consolidadas en un único texto por registro. Esta decisión permitió que la comparación entre modelos dependiera principalmente del enfoque de clasificación utilizado, y no de diferencias en las variables de entrada.

La configuración específica de cada modelo, así como los criterios de decisión aplicados durante el ajuste, se presentan posteriormente en la sección 5.2.

4.6.1. Modelo basado en reglas

El primer enfoque correspondió a un modelo basado en reglas. Este modelo procesó el texto consolidado de cada registro mediante medSpaCy y QuickUMLS, complementados con reglas léxicas y contextuales para identificar menciones relacionadas con cáncer. Además de detectar términos biomédicos, el modelo revisó si esas menciones aparecían en contextos de negación, incertidumbre, antecedente familiar o antecedente personal. Este tipo de enfoque sigue siendo relevante en PLN clínico por su interpretabilidad y por su utilidad en tareas donde es importante rastrear la evidencia textual que lleva a una clasificación determinada (Dan Jurafsky & James H. Martin, 2025; Digan et al., 2021; Hua Xu & Demner Fushman, 2024).

A partir de esa evidencia, el sistema construyó un perfil de decisión por registro y asignó una clase final entre 0, 1 y 2. Para su desarrollo, la base etiquetada se dividió en entrenamiento, validación y prueba, con una distribución aproximada de 64 %, 16 % y 20 %, respectivamente. En este caso, el conjunto de entrenamiento no se utilizó para ajustar parámetros estadísticos, sino para construir la lógica heurística del modelo; el conjunto de validación permitió afinar prioridades entre reglas, revisar perfiles ambiguos y ajustar criterios de decisión; y el conjunto de prueba se reservó para estimar el desempeño final sobre datos no utilizados durante el ajuste. Esta lógica es coherente con aplicaciones basada en reglas previamente reportadas en documentos oncológicos y reportes de patología (Hammami et al., 2021).

4.6.2. Modelo SVM

El segundo enfoque correspondió a un modelo SVM jerárquico en dos etapas. Primero, el texto de cada registro se transformó en una representación numérica mediante TF-IDF. Sobre esa representación se entrenó un primer clasificador para diferenciar los casos 0 frente a los casos 1 o 2 agrupados. Luego, sobre los registros positivos, se entrenó un segundo clasificador para distinguir entre clase 1 y clase 2. Esta estructura respondió al criterio central del estudio, que fue reducir al máximo los falsos 0 (Dan Jurafsky & James H. Martin, 2025; James, Witten, Hastie, & Tibshirani, 2023).

La decisión de utilizar una estructura jerárquica se basó en la lógica operativa del registro. En una primera etapa, era necesario determinar si el registro debía mantenerse dentro del flujo de revisión por una posible mención de cáncer. Por esta razón, las clases 1 y 2 se agruparon como registros relevantes y se compararon frente a la clase 0.

Posteriormente, cuando un registro era identificado como potencialmente relevante, el segundo clasificador permitía distinguir entre una mención clara de cáncer y una mención probable o ambigua. Esta estrategia resultó más coherente con el objetivo de captación que una clasificación multiclase directa, ya que priorizó la reducción de falsos negativos en la clase 0 antes que la separación detallada entre las clases positivas.

Para este modelo, la base etiquetada se dividió en entrenamiento, validación y prueba, con una distribución aproximada de 64 %, 16 % y 20 %, respectivamente. Sobre el conjunto de entrenamiento se ajustó la representación TF-IDF y se entrenaron los clasificadores de ambas etapas. La selección de hiperparámetros se apoyó en GridSearchCV con validación cruzada estratificada, mientras que el conjunto de validación se utilizó para contrastar configuraciones candidatas, revisar los puntajes de decisión y afinar una lógica de decisión más conservadora en los casos dudosos. En la primera etapa, la búsqueda se orientó por sensibilidad, mientras que en la segunda se orientó por sensibilidad macro, con el fin de no desatender la clase minoritaria. Finalmente, el conjunto de prueba se mantuvo aparte para la evaluación final del modelo (Hammami et al., 2021; Koopman et al., 2018).

4.6.3. Modelo basado en *Transformer*s

El tercer enfoque correspondió a un modelo basado en *Transformer*s, específicamente un RoBERTa biomédico-clínico en español ajustado para clasificación multiclase. En este caso, cada registro fue tokenizado y truncado a una longitud máxima fija de 128 tokens, de modo que pudiera ser procesado por la arquitectura. Los *Transformer*s permitieron modelar relaciones contextuales dentro del texto y constituyeron una de las bases actuales del PLN clínico y biomédico (Dan Jurafsky & James H. Martin, 2025; Hua Xu & Demner Fushman, 2024; Vaswani et al., 2023).

Al igual que en los demás enfoques, la base etiquetada se dividió en entrenamiento, validación y prueba. Primero se separó un 20 % para prueba y, sobre el 80 % restante, se hizo una segunda partición para validación, lo que dio una distribución final aproximada de 64 % para entrenamiento, 16 % para validación y 20 % para prueba. Durante el ajuste fino se utilizaron pesos por clase y WeightedRandomSampler para mitigar el desbalance. Además, la decisión final no dependió solo del argmax de las probabilidades, sino también de umbrales calibrados sobre las salidas del modelo, con el fin de reducir falsos 0 (García Subies et al., 2024).

El fine-tuning se llevó a cabo con un número reducido de épocas, definido de acuerdo con el tamaño de la muestra, el desbalance entre las clases y la posibilidad de sobreajuste. En modelos de este tipo, extender demasiado el entrenamiento puede hacer que la red aprenda patrones muy específicos de los datos disponibles, especialmente cuando la muestra no es muy grande, en lugar de desarrollar una capacidad de generalización adecuada.

Por ello, la configuración del entrenamiento no se definió solo por las condiciones de infraestructura disponibles, sino también por una decisión metodológica. El propósito fue conservar un comportamiento estable durante la validación y disminuir el riesgo de que el modelo se ajustara en exceso a los datos de entrenamiento.

4.7. Estrategia de evaluación

La evaluación se realizó exclusivamente sobre la base etiquetada por expertos. En los tres enfoques se conservaron subconjuntos separados para entrenamiento, validación y prueba, con una distribución aproximada de 64 %, 16 % y 20 %, respectivamente. El conjunto de validación se

utilizó para ajustar componentes específicos de cada modelo, mientras que el conjunto de prueba se reservó exclusivamente para la evaluación final comparativa. El desempeño de los modelos se analizó mediante precisión, sensibilidad, F1, especificidad y exactitud global, complementados con matrices de confusión y revisión de errores. Estas métricas fueron consistentes con las prácticas habituales de evaluación en clasificación de texto y PLN clínico (Hua Xu & Demner Fushman, 2024; James, Witten, Hastie, & Tibshirani, 2023).

Sin embargo, el criterio principal de interpretación no fue la exactitud global, sino la capacidad de evitar falsos 0. Por ello, se revisó de manera prioritaria el comportamiento de los modelos en la detección de cualquier mención de cáncer, agrupando las clases 1 y 2 frente a la clase 0. Este criterio fue coherente con el propósito aplicado del estudio, ya que para el RPCa-M resultó más crítico no descartar registros potencialmente relevantes que maximizar una métrica global sin considerar ese riesgo, atendiendo al objetivo de la exhaustividad en la captación de casos (Bray & Parkin, 2009; Universidad de Caldas, 2025).

La validación realizada en este trabajo fue de tipo interna, debido a que los conjuntos de entrenamiento, validación y prueba se obtuvieron a partir de la misma base institucional. Si bien la división de los datos permitió evaluar el desempeño del modelo sobre registros que no fueron usados durante el entrenamiento, esta evaluación no puede considerarse una validación externa, temporal ni multicéntrica.

En otras palabras, el modelo no fue probado con información proveniente de otro periodo, de otro registro poblacional de cáncer o de una fuente documental independiente. Esta aclaración es relevante porque los resultados obtenidos muestran cómo se comportó el modelo dentro de las condiciones específicas de la muestra disponible. Sin embargo, no permiten asumir de manera

directa que tendría el mismo desempeño en otros escenarios o con datos construidos bajo condiciones diferentes.

Por esta razón, la evaluación debe entenderse como una primera validación del artefacto desarrollado para el RPCa-M. Para pensar en una implementación rutinaria, o en su posible uso en otros registros, sería necesario realizar pruebas adicionales con datos posteriores, bases de otros periodos o registros generados en contextos operativos distintos.

Estas nuevas validaciones permitirían revisar si el modelo mantiene un comportamiento estable frente a cambios en la forma de diligenciamiento, variaciones terminológicas, diferencias en los patrones de escritura y posibles modificaciones en la distribución de las clases.

4.8. Selección del modelo final

Los tres modelos se compararon bajo una misma lógica de entrada, una misma base etiquetada y un mismo criterio de evaluación. La selección del modelo final no dependió únicamente del mejor valor global de desempeño, sino del equilibrio entre comportamiento general, interpretabilidad y capacidad para reducir falsos 0. Bajo este criterio, se priorizó la alternativa con mayor utilidad para apoyar la revisión de posibles casos dentro del flujo de trabajo del RPCa-M, en línea con la necesidad de mejorar exhaustividad, oportunidad y consistencia en la revisión de registros de defunción (Bray & Parkin, 2009; Universidad de Caldas, 2025).

4.9. Consolidación del entregable para el RPCa-M

Como parte final de la metodología, el modelo SVM seleccionado fue organizado como un entregable local para el Registro Poblacional de Cáncer de Manizales. Para esto, se desarrolló una aplicación en Streamlit, con el fin de que el usuario final pudiera cargar una base de registros de defunción, aplicar el modelo y obtener una salida en Excel con la clasificación generada para cada registro.

El entregable incluyó las principales etapas del flujo desarrollado: el preprocesamiento de los datos, la construcción del texto de entrada, la aplicación del modelo seleccionado y, finalmente, la generación de una salida con la clasificación predicha para cada registro. Esta salida conserva el identificador seudoanonimizado, junto con el texto procesado o las variables necesarias para mantener la trazabilidad técnica del análisis. También incluye la clase predicha y, en los casos en que el modelo lo permite, el puntaje o la probabilidad asociada a la decisión.

Esta estructura facilita que el RPCa-M pueda revisar los registros priorizados dentro del flujo analítico, sin necesidad de exponer datos personales que no sean indispensables para el proceso. De esta manera, se mantiene un equilibrio entre la utilidad operativa del modelo y el manejo responsable de la información sensible.

Para apoyar la reproducibilidad técnica del trabajo, el código del experimento fue organizado en un repositorio de GitHub. Por el carácter sensible de la información utilizada, dicho repositorio no incluye los datos originales ni la tabla llave de seudoanonimización. Aun así, sí contiene la estructura general del flujo de procesamiento, los scripts o notebooks de entrenamiento, la configuración de los modelos, las dependencias principales y una descripción del formato esperado de entrada.

Con esto, otros investigadores pueden revisar la lógica del procedimiento, comprender cómo fue implementado y, si cuentan con datos autorizados, adaptarlo a sus propios contextos de análisis. El enlace actualizado del repositorio se presenta en el Anexo 3.

5. Resultados y análisis

En este capítulo se exponen los resultados obtenidos en los tres enfoques de clasificación y se analizan sus diferencias en función del objetivo del estudio.

5.1. Comparación de los flujos de entrenamiento

La figura 6 muestra que los tres modelos partieron de una misma muestra de 4000 registros de defunción, junto con sus respectivas etiquetas. A partir de esta base común, se organizaron los flujos de cada enfoque en cinco momentos: selección del modelo, ingeniería de características, división de los datos, entrenamiento y evaluación. Esta estructura permitió comparar los modelos bajo una misma tarea, aunque cada uno siguió una forma diferente de procesar el texto y generar la clasificación final.

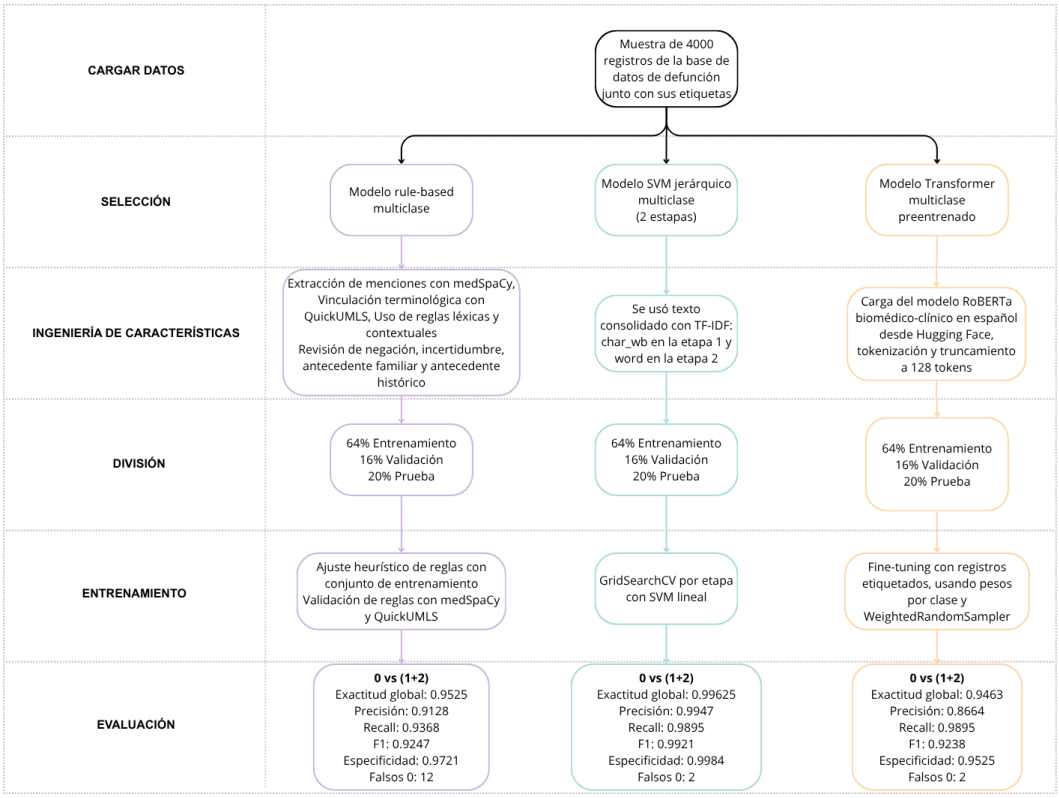

Figura 6: Flujo comparativo de entrenamiento y evaluación de los tres modelos de PLN aplicados a registros de defunción: reglas, SVM y Transformers.

Nota. *Elaboración propia.*

En el modelo basado en reglas, la clasificación se construyó a partir de la extracción de menciones con medSpaCy, la vinculación terminológica con QuickUMLS y el uso de reglas léxicas y contextuales. Además, se revisaron elementos como negación, incertidumbre, antecedente familiar y antecedente histórico. En este caso, el entrenamiento correspondió a un ajuste heurístico de reglas con el conjunto de entrenamiento, validado posteriormente con medSpaCy y QuickUMLS.

El modelo SVM se desarrolló como un clasificador jerárquico multiclase en dos etapas. Para representar el texto, se usó texto consolidado con TF-IDF de caracteres en la primera etapa y de palabras en la segunda. Posteriormente, se aplicó GridSearchCV por etapa con SVM lineal, manteniendo la misma división de 64% entrenamiento, 16% validación y 20% prueba.

En el caso del modelo *Transformer*, se partió de un modelo RoBERTa biomédico-clínico en español, preentrenado y cargado desde Hugging Face. Junto con el modelo, se utilizó su tokenizador para transformar los registros en secuencias de tokens, con truncamiento a 128 tokens. Luego, se realizó ajuste fino con los registros etiquetados, usando pesos por clase y WeightedRandomSampler para atender el desbalance de las categorías durante el entrenamiento.

Finalmente, los tres modelos fueron evaluados con las mismas métricas para la tarea principal de clasificación entre ausencia de mención de cáncer y cualquier mención de cáncer, es decir, 0 vs (1+2). Esto permitió comparar su desempeño en términos de exactitud global, precisión, sensibilidad, F1, especificidad y falsos 0.

5.2.Criterios de decisión y configuración de los modelos

Para complementar la interpretación de los resultados, esta sección describe los principales criterios de decisión y configuraciones utilizadas en cada enfoque. Además, los repositorios incluidos en el Anexo 3 permiten consultar la implementación de los tres modelos y revisar con mayor detalle los pasos seguidos para llegar a la clasificación final.

5.2.1. Modelo basado en reglas

La clase final no se obtuvo a partir de probabilidades entrenadas, sino de la combinación entre menciones detectadas, contexto clínico y perfiles de decisión. Primero, el sistema identificó términos y expresiones relacionadas con cáncer y luego revisó si esas menciones aparecían en contextos de negación, incertidumbre, antecedente familiar o antecedente personal. Con esa información, cada registro quedó resumido en un perfil de decisión. Esta lógica es coherente con los enfoques basados en reglas del PLN clínico, donde la decisión depende de evidencia textual explícita y de reglas de contexto, más que de un ajuste estadístico tradicional (Dan Jurafsky & James H. Martin, 2025; Hua Xu & Demner Fushman, 2024).

La selección de la clase siguió tres niveles. En primer lugar, algunos registros recibieron una asignación directa. Cuando no hubo evidencia positiva y sí hubo evidencia negada del caso, el sistema favoreció la clase 0. En cambio, cuando predominó la evidencia incierta o sospechosa, el modelo favoreció la clase 2. En segundo lugar, los registros que no quedaron resueltos de forma directa pasaron al mapeo perfil–clase. En tercer lugar, el código calculó puntajes heurísticos por clase para complementar la interpretación de la decisión final.

Los perfiles no resueltos por asignación directa siguieron una lógica por defecto. El perfil evidence_0 y los perfiles que comenzaban por negated_only__ favorecieron clase 0. Los perfiles

uncertain_only__ favorecieron clase 2. Los perfiles mixed__ favorecieron clase 2 cuando predominó la incertidumbre y clase 1 en los demás casos. Finalmente, los perfiles definite__ y weak__ favorecieron clase 1.

En este modelo no hubo hiperparámetros estadísticos en sentido estricto, sino parámetros heurísticos de decisión. Sobre el subconjunto destinado al desarrollo del modelo se definieron los criterios que guiaron la asignación final de clase. El análisis de contexto utilizó una ventana izquierda de 80 caracteres y una ventana local de 40 caracteres alrededor de la mención. Además, para favorecer la recuperación de la clase 2 se ajustaron parámetros como CLASS2_RECALL_BOOST = 2.40, CLASS1_DEFINITE_BOOST = 1.05, CLASS2_SELECTION_RATIO = 0.80 y CLASS2_DIRECT_EVIDENCE_MIN = 2. De esta forma, la validación no se orientó a optimizar parámetros estadísticos, sino a afinar la lógica heurística del modelo y a revisar su comportamiento frente a perfiles ambiguos o inciertos.

El código también incorporó una interpretación de confianza basada en el margen entre el puntaje más alto y el segundo puntaje más alto. Esa confianza no representó una probabilidad calibrada, sino una señal heurística de separación entre clases. El margen se interpretó como bajo cuando fue menor que 1.0, medio cuando estuvo entre 1.0 y menor que 2.5, y alto cuando fue mayor o igual a 2.5.

5.2.2. Modelo SVM

La clase final se obtuvo mediante un esquema jerárquico en dos etapas. En la primera, el clasificador separó los registros sin mención de cáncer de aquellos con cualquier mención o sospecha. En la segunda, solo sobre los casos positivos, un segundo clasificador distinguió entre clase 1 y clase 2. Esta organización permitió resolver primero la separación más importante del

problema y luego refinar la clasificación dentro del grupo positivo. Las SVM siguen siendo una alternativa sólida en clasificación de texto cuando se trabaja con representaciones dispersas y problemas de alta dimensionalidad (Dan Jurafsky & James H. Martin, 2025; James, Witten, Hastie, & Tibshirani, 2023).

En este modelo no se utilizó tokenización como etapa separada. La razón es que la representación del texto se construyó con TF-IDF, y este procedimiento ya realiza internamente la segmentación necesaria para convertir el texto en características. Tanto en la etapa 1 como en la etapa 2, la mejor configuración utilizó analyzer = 'char_wb', es decir, n-gramas de caracteres dentro de bordes de palabra. En consecuencia, una tokenización previa por palabras no solo era innecesaria, sino que tampoco correspondía a la forma en que el modelo estaba representando el texto.

En la etapa 1, la decisión inicial se basó en el signo del puntaje de decisión del clasificador. Sin embargo, el modelo no conservó automáticamente como clase 0 todos los casos negativos. Si un registro era enviado a 0, pero su puntaje quedaba por encima del umbral de reasignación, el sistema activaba una lógica más conservadora para evitar falsos 0. En la implementación final, ese umbral fue UMBRAL_REASIGNACION_ETAPA1 = -0.10. Durante la ejecución reportada por el código, este criterio llevó a reasignar 1 caso de 0 a 2. En la etapa 2, el segundo SVM distinguió entre clase 1 y clase 2. Aquí el puntaje se interpretó de forma directa: valores positivos favorecieron clase 2 y valores negativos favorecieron clase 1. Además, el código calculó un umbral de baja confianza a partir del percentil 20 del valor absoluto de los puntajes de entrenamiento. En la ejecución final, este umbral fue umbral_baja_confianza_etapa2 = 0.565496. Su función no fue forzar una reclasificación automática, sino señalar casos con baja separación entre clase 1 y clase 2.

En cuanto a los hiperparámetros, la etapa 1 se ajustó mediante GridSearchCV con validación cruzada estratificada de 5 particiones, usando como criterio principal sensibilidad. Los mejores hiperparámetros fueron: svm__C = 4.0, tfidf__analyzer = 'char_wb', tfidf__min_df = 2, tfidf__ngram_range = (3, 5) y tfidf__sublinear_tf = True, con class_weight = "balanced" y max_iter = 5000. La etapa 2 se ajustó con validación cruzada estratificada de 3 particiones y criterio de selección basado en sensibilidad macro. Los mejores hiperparámetros fueron: svm__C = 0.25, tfidf__analyzer = 'char_wb', tfidf__min_df = 2, tfidf__ngram_range = (4, 6) y tfidf__sublinear_tf = True, también con class_weight = "balanced" y max_iter = 5000.

5.2.3. Modelo basado en *Transformer*s

La clase final no se seleccionó únicamente por la probabilidad más alta. Aunque el modelo produjo probabilidades para las tres clases, la decisión final combinó esas probabilidades con un conjunto de umbrales calibrados en validación. Este tipo de ajuste es consistente con el uso actual de *Transformer*s clínicos, donde el ajuste fino suele complementarse con decisiones posteriores de calibración para adaptar la salida del modelo a tareas específicas del dominio biomédico (García Subies et al., 2024; Vaswani et al., 2023)

En términos generales, el sistema asignó clase 1 cuando la probabilidad de esa clase superó su umbral y además aventajó suficientemente a la clase 2. Asignó clase 2 cuando la probabilidad correspondiente superó su umbral o cuando existieron señales textuales de sospecha y la probabilidad conjunta de cualquier mención de cáncer fue suficiente. Asignó clase 0 cuando la probabilidad negativa superó su umbral sin activarse las reglas de las otras dos clases. Solo cuando ninguna de estas condiciones se cumplió, el modelo recurrió al argmax como criterio de respaldo.

Los umbrales activos en la ejecución final fueron t0 = 0.50, t1 = 0.40, t2 = 0.30, t_any = 0.35 y margin12 = 0.05. Estos valores permitieron ajustar la decisión final del modelo con énfasis en la reducción de falsos 0 y en la recuperación de registros con mención clara o probable de cáncer.

En cuanto a la configuración general, el modelo base fue PlanTL-GOB-ES/roberta-base-biomedical-clinical-es. El texto se procesó con una longitud máxima de 128 tokens, con una partición de 20% para prueba y una segunda partición sobre el conjunto restante para validación, lo que produjo una distribución aproximada de 64% para entrenamiento, 16% para validación y 20% para prueba. El ajuste fino se realizó durante 4 épocas, con learning_rate = 2e-5, weight_decay = 0.01, early_stopping_patience = 2, tamaño de lote de entrenamiento de 16 en GPU o 4 sin GPU, y tamaño de evaluación de 32 en GPU o 8 sin GPU. El mejor modelo se seleccionó usando eval_recall_macro como criterio principal. Además, se emplearon pesos por clase y WeightedRandomSampler, lo que hizo que el entrenamiento fuera más sensible al desbalance entre clases.

5.3. Resultados cuantitativos comparativos

Este apartado presenta de forma comparativa el desempeño de los tres modelos a partir de sus métricas globales, su comportamiento por clase y su capacidad para evitar falsos 0.

5.3.1. Desempeño general

Los tres modelos alcanzaron un desempeño global alto, pero con diferencias claras entre ellos. La Tabla 2 resume las métricas generales de cada modelo para la tarea principal del estudio:

distinguir entre ausencia de mención de cáncer, clase 0, y cualquier mención clara o probable de cáncer, clases 1 o 2.

Tabla 2: Comparación del desempeño general de los tres modelos para clasificar ausencia de mención de cáncer frente a cualquier mención clara o probable de cáncer en registros de defunción.

Modelo	Exactitud global	Precisión	Sensibilidad	F1	Especificidad
Reglas	0.9525	0.9128	0.9368	0.9247	0.9721
SVM	0.9875	0.9947	0.9895	0.9921	0.9984
Transformers	0.9463	0.8664	0.9895	0.9238	0.9525

Nota. *En esta comparación, la clase 0 corresponde a ausencia de mención de cáncer, mientras que las clases 1 y 2 se agrupan como cualquier mención clara o probable de cáncer.*

En conjunto, estos resultados muestran tres comportamientos distintos. El modelo basado en reglas mantuvo un desempeño favorable y estable en la detección de menciones claras, aunque con valores más moderados que el SVM. El modelo basado en *Transformer*s alcanzó una sensibilidad muy alto, igual al del SVM, lo que indica una alta capacidad para recuperar registros con cualquier mención de cáncer. Sin embargo, este desempeño estuvo acompañado por una menor precisión y especificidad. El modelo SVM, en cambio, combinó la mayor exactitud global con los mejores valores de precisión, F1 y especificidad, por lo que mostró el comportamiento más equilibrado en la tarea de distinguir entre ausencia de mención y cualquier indicio de cáncer.

Este resultado es importante porque el propósito del estudio no era únicamente obtener una exactitud alta, sino identificar registros con indicios de cáncer sin aumentar de forma excesiva los

errores de clasificación. Desde esta perspectiva, el SVM fue el modelo con el desempeño más sólido en la evaluación general.

5.3.2. Desempeño por clase

El análisis por clase permitió observar con mayor detalle el comportamiento de cada modelo frente a las tres categorías del estudio. La Tabla 3 presenta la precisión, sensibilidad y F1 para cada clase.

Tabla 3: Desempeño por clase de los tres modelos para clasificar registros de defunción según ausencia, mención clara o mención probable de cáncer.

Modelo	Clase	Precisión	Sensibilidad	F1
Reglas	0	0.9802	0.9721	0.9761
Reglas	1	0.8796	0.9231	0.9008
Reglas	2	0.2500	0.1250	0.1667
SVM	0	0.9967	0.9984	0.9975
SVM	1	0.9832	0.9670	0.9751
SVM	2	0.5000	0.6250	0.5556
Transformers	0	0.9966	0.9525	0.9740
Transformers	1	0.9396	0.9396	0.9396
Transformers	2	0.1429	0.6250	0.2326

Nota. *La clase 0 corresponde a ausencia de mención de cáncer; la clase 1, a mención clara de cáncer; y la clase 2, a mención probable o incierta de cáncer.*

Los resultados muestran que las clases 0 y 1 fueron las más estables en los tres modelos, mientras que la clase 2 fue la más difícil de clasificar. Esto indica que la mayor dificultad no estuvo en reconocer registros claramente negativos o visiblemente positivos, sino en clasificar registros con evidencia más ambigua o menos explícita de cáncer.

En esta comparación, el SVM presentó el comportamiento más equilibrado por clase. Mantuvo valores altos en las clases 0 y 1 y obtuvo el mejor F1 en la clase 2. El modelo basado en *Transformer*s también alcanzó una sensibilidad alta en la clase 2, pero con una precisión más baja. Esto indica que recuperó más casos ambiguos, aunque con mayor tendencia a clasificar como clase 2 registros que no necesariamente correspondían a esa categoría.

5.3.3. Falsos 0 y recuperación de casos relevantes

En el modelo basado en reglas, la matriz mostró que 11 registros de clase 1 y 1 registro de clase 2 fueron clasificados como 0. En total, esto equivale a 12 falsos 0. Además, en la clase 2 real, solo 1 caso fue detectado como 2, mientras que 7 no fueron detectados como 2; de esos 7, 1 quedó como 0 y 6 quedaron como 1. Esto indica que el modelo logró recuperar gran parte de los casos con mención de cáncer, pero tuvo más dificultad cuando la evidencia fue ambigua o incierta.

Tabla 4: Matriz de confusión multiclase del modelo basado en reglas en el conjunto de prueba para clasificar registros de defunción según indicios de cáncer.

Real \ Predicha	**0**	**1**	**2**
0	593	17	0
1	11	168	3
2	1	6	1

En el modelo SVM, la matriz de confusión mostró que 2 casos de la clase 1 fueron clasificados como 0 y que ningún caso de la clase 2 cayó en esa categoría. En otras palabras, de los 190 registros que realmente correspondían a las clases 1 o 2, solo 2 fueron enviados a la clase 0. Este resultado ubicó al SVM entre los modelos con menor cantidad de falsos 0 frente al criterio central del estudio.

Tabla 5: Matriz de confusión multiclase del modelo SVM en el conjunto de prueba para clasificar registros de defunción según indicios de cáncer.

Real \ Predicha	**0**	**1**	**2**
0	609	0	1
1	2	176	4
2	0	3	5

En el modelo basado en *Transformer*s, el comportamiento también fue favorable desde este criterio, ya que el modelo obtuvo una sensibilidad de 0.9895 para distinguir cualquier mención de cáncer frente a la clase 0. Esto confirma que logró retener una proporción muy alta de registros con mención real o probable de cáncer dentro del proceso de clasificación. Sin embargo, ese comportamiento estuvo acompañado por una precisión menor que la observada en el SVM, lo que indica una mayor tendencia a enviar hacia clases positivas algunos registros que en realidad pertenecían a la clase 0.

Tabla 6: Matriz de confusión multiclase del modelo basado en Transformers en el conjunto de prueba para clasificar registros de defunción según indicios de cáncer.

Real \ Predicha	**0**	**1**	**2**
0	581	8	21

1	2	171	9
2	0	3	5

Con el fin de resumir de manera directa este criterio de evaluación, la Tabla 7 presenta la cantidad de falsos 0 obtenida por cada modelo en el conjunto de prueba, es decir, los registros que realmente correspondían a las clases 1 o 2, pero fueron clasificados como 0. Esta representación resulta importante porque permite visualizar de forma comparativa el error más crítico para el propósito aplicado del estudio: dejar por fuera del proceso de revisión casos con mención clara o probable de cáncer. En este sentido, la tabla permite apreciar con mayor claridad la ventaja relativa del modelo SVM y del modelo basado en *Transformer*s, mientras que el modelo basado en reglas fue el que presentó la mayor cantidad de este tipo de errores.

Tabla 7: Falsos 0 por modelo en el conjunto de prueba

Modelo	Falsos 0 total	Falsos 0 clase_1	Falsos 0 clase_2
Reglas	12	11	1
SVM	2	2	0
Transformers	2	2	0

***Nota**. Los falsos 0 corresponden a registros que realmente pertenecían a las clases 1 o 2, pero fueron clasificados como clase 0, es decir, como ausencia de mención de cáncer.*

En conjunto, estos resultados muestran que los tres modelos fueron capaces de recuperar buena parte de los registros relevantes, pero no lo hicieron con la misma eficacia frente al criterio principal del estudio. El modelo basado en reglas alcanzó un desempeño general alto, aunque dejó más falsos 0. El modelo basado en *Transformer*s mostró una recuperación muy alta de casos con mención de cáncer. El SVM, por su parte, igualó el menor número de falsos 0 observado en el

estudio y, además, mantuvo el comportamiento más equilibrado entre recuperación de casos relevantes, estabilidad de clasificación y control de errores.

5.3.4. Análisis cualitativo de errores

Además de las métricas cuantitativas, se revisaron también los tipos de error más frecuentes que generaron los modelos, sobre todo pensando en el impacto que podrían tener dentro del flujo operativo del RPCa-M. En este punto, los falsos negativos fueron los errores más delicados, porque corresponden a registros donde sí había una mención clara o probable de cáncer, pero el modelo los clasificó como si no la tuvieran.

En la práctica, esto puede hacer que un caso que sí debería revisarse no sea priorizado dentro del registro. Por eso, más que buscar únicamente el mejor desempeño global, se consideró especialmente importante reducir este tipo de error al momento de seleccionar el modelo.

Los falsos positivos, por su parte, tienen otro tipo de efecto. En estos casos, el modelo identifica como posible mención de cáncer registros que realmente no corresponden a esa categoría. Esto puede aumentar el trabajo de revisión manual, pero su impacto es menor frente a los falsos negativos, ya que esos registros todavía pasan por etapas posteriores de verificación, depuración y codificación, donde la clasificación inicial puede corregirse.

La clase 2 fue particularmente difícil de clasificar. Esta clase incluye registros con evidencia textual probable, incompleta o ambigua, y por eso no siempre es sencillo separar una mención clara de cáncer de una mención dudosa. A veces la decisión depende de expresiones clínicas poco comunes, abreviaturas, errores de escritura, términos generales relacionados con malignidad o diagnósticos poco frecuentes.

Por esta razón, algunos errores no deben interpretarse únicamente como fallas del modelo. También muestran la dificultad propia del criterio clínico y registral que implica clasificar este tipo de registros.

5.4. Análisis comparativo final

En conjunto, los resultados muestran que los tres modelos fueron capaces de clasificar los registros con un desempeño general alto, pero no respondieron de la misma manera al criterio principal del estudio. Aunque los tres identificaron adecuadamente la mayoría de los casos, sus diferencias se hicieron más visibles cuando la comparación se centró en la reducción de falsos 0 y en el manejo de la clase 2, que fue la categoría más difícil en todos los enfoques.

El modelo basado en reglas mostró un comportamiento sólido en los casos más claros. Su desempeño fue alto en la clase 0 y también fue favorable en la clase 1. Además, tuvo la ventaja de ofrecer una decisión interpretable, ya que la clasificación dependió de menciones explícitas y de reglas de contexto que podían revisarse de forma directa. Sin embargo, este modelo presentó limitaciones importantes cuando la evidencia fue más ambigua. Esto se reflejó especialmente en la clase 2, donde obtuvo 0.1250 de sensibilidad y 0.1667 de F1, y también en la cantidad de falsos 0, que fue de 12, mayor que en los otros enfoques. En consecuencia, aunque fue una alternativa útil como línea base interpretable, no fue la opción que mejor respondió al criterio operativo del estudio.

El modelo SVM fue el que mostró el comportamiento más equilibrado. Alcanzó la mayor exactitud global, mantuvo resultados muy altos en las clases 0 y 1, y también logró el mejor F1 en

la clase 2. Además, se ubicó entre los modelos con menor cantidad de falsos 0, con solo 2 casos positivos reales clasificados como 0. Este resultado fue especialmente importante porque el propósito del estudio consistía en evitar que registros con mención clara o sospecha de cáncer quedaran fuera del proceso de revisión, por encima de una valoración centrada únicamente en el desempeño global.

El modelo basado en *Transformer*s mostró un comportamiento distinto. Su principal fortaleza estuvo en la recuperación de casos relevantes, como se observó en su sensibilidad para cualquier mención de cáncer frente a la clase 0. Esto indica que el modelo actuó de forma más conservadora y tendió a mantener dentro del proceso una proporción muy alta de registros con mención real o probable de cáncer. Sin embargo, esta ventaja estuvo acompañada por una menor precisión y por un comportamiento menos equilibrado entre clases. En particular, aunque logró recuperar parte de los casos de clase 2, lo hizo con una precisión baja, lo que muestra que esa sensibilidad adicional también incrementó la tendencia a llevar hacia clases positivas registros que realmente pertenecían a la clase 0.

Al comparar los tres enfoques, se observan entonces tres perfiles distintos. El modelo basado en reglas destacó por su interpretabilidad, pero mostró mayores limitaciones en casos ambiguos. El modelo basado en *Transformer*s priorizó la recuperación de casos potencialmente relevantes, aunque con una clasificación menos estricta. El modelo SVM, en cambio, logró el mejor equilibrio entre desempeño global, estabilidad por clase y reducción de falsos 0. Por esta razón, no solo presentó las mejores métricas generales, sino que también fue el que mejor respondió al criterio central del estudio.

Un hallazgo importante es que la clase 2 fue la más difícil para clasificar correctamente en los tres modelos. Esto muestra que la principal dificultad del problema no estuvo en distinguir

registros claramente negativos de registros claramente positivos, sino en clasificar aquellos casos donde la mención de cáncer era más incierta, menos explícita o más ambigua en el texto. Esta dificultad apareció de forma consistente en los tres enfoques, lo que refuerza la idea de que se trata de una complejidad propia del problema y no únicamente de una limitación particular de un modelo.

En síntesis, los resultados comparativos muestran que los tres modelos fueron útiles para la clasificación de registros de defunción según indicios de cáncer, pero con comportamientos distintos. El modelo basado en reglas aportó interpretabilidad y el modelo basado en *Transformer*s aportó sensibilidad. El modelo SVM, por su parte, presentó el mejor equilibrio entre desempeño global, estabilidad por clase y reducción de falsos 0. En este sentido, fue la alternativa más sólida para el propósito del estudio.

5.5. Entrega del modelo y valoración del usuario final

Como resultado aplicado del estudio, el modelo SVM seleccionado fue organizado como un entregable local para el RPCa-M. Inicialmente, el desarrollo quedó compuesto por el archivo de la aplicación, el archivo de requerimientos y un archivo comprimido con los modelos entrenados. Dentro de este último se incluyeron dos archivos en formato .joblib: modelo_etapa1.joblib, encargado de diferenciar entre ausencia y presencia de indicios de cáncer, y modelo_etapa2.joblib, orientado a distinguir entre las clases 1 y 2 cuando el registro fue identificado con posible mención de cáncer. También se incluyó un archivo metadata.json, utilizado para conservar parámetros necesarios para la ejecución del modelo.

Posteriormente, este desarrollo fue organizado en un repositorio de GitHub y empaquetado en Docker para facilitar su uso por parte del registro. Esta forma de entrega permitió que la aplicación pudiera ejecutarse de manera local desde el equipo del usuario final, sin depender de plataformas externas para procesar la información. Esto fue importante por el tipo de datos utilizados, ya que los registros de defunción contienen información sensible y requieren un manejo cuidadoso dentro del entorno del RPCa-M.

La aplicación permite cargar bases de registros de defunción, aplicar el modelo SVM jerárquico y descargar un archivo Excel con la predicción asignada a cada registro. En la salida se conserva la información necesaria para que el usuario pueda revisar los resultados y priorizar los registros clasificados como 1 o 2, es decir, aquellos con mención clara o probable de cáncer.

La Ilustración 1 presenta la vista general del despliegue local de la aplicación entregada al RPCa-M. En esta imagen se observa la interfaz principal de la herramienta, desde la cual el usuario puede cargar la base de registros de defunción, ejecutar la clasificación y descargar el archivo de resultados generado por el modelo.



Ilustración 1: Vista general del despliegue local del clasificador de registros de defunción entregado al RPCa-M.

Para valorar el entregable, se aplicó una encuesta breve a tres usuarios finales del registro, disponible en el Anexo 5. En las preguntas cerradas, los tres participantes respondieron “Totalmente de acuerdo” frente a la comprensión de las categorías 0, 1 y 2, la claridad de la información presentada, la coincidencia general entre el resultado de la aplicación y su criterio de revisión, la utilidad de la aplicación como apoyo, la reducción del tiempo de revisión y la identificación de registros que requerían una revisión más detallada.

En las respuestas abiertas, los usuarios señalaron que la aplicación era clara, accesible y útil para apoyar la revisión de registros de defunción. También destacaron que la herramienta permite revisar grandes volúmenes de información en menos tiempo y facilita la identificación de registros clasificados como 2, los cuales suelen ser más difíciles de verificar manualmente. Además, se resaltó que la aplicación puede servir para decidir qué registros revisar primero, identificar posibles casos relacionados con cáncer, detectar registros que generan duda y apoyar la revisión final sin reemplazar el criterio del revisor.

En conjunto, la valoración del usuario final fue favorable. Los resultados de la encuesta sugieren que el entregable fue comprendido por los usuarios, que la salida generada fue clara y que la aplicación puede apoyar el proceso de revisión del RPCa-M.

Adicionalmente, después de la entrega de la herramienta se realizó una revisión comparativa entre el etiquetado humano y la predicción generada por la aplicación en una base de prueba de 400 casos del registro. En esta revisión se identificaron 10 falsos negativos, asociados principalmente con diagnósticos o expresiones que no estaban presentes, o estaban poco representados, en los datos originales de entrenamiento. Entre estos casos se encontraron errores de escritura, caracteres especiales y términos menos frecuentes, como meningioma, mesotelioma, histiocitosis de células de Langerhans, gliomatosis cerebri y algunas menciones generales de

malignidad. Este hallazgo refuerza la necesidad de ampliar la base etiquetada con más ejemplos revisados por expertos, de modo que el modelo cuente con mayor variedad de expresiones clínicas para aprender y mejorar su capacidad de generalización.

También se observó que la herramienta captó tres casos relevantes: dos que no habían sido detectados previamente en el etiquetado humano y uno que, después de la revisión, fue considerado más cercano a clase 1 que a clase 2. Además, algunos registros positivos fueron clasificados como clase 2, aunque en el etiquetado humano correspondían a clase 1. Estos hallazgos se tomaron como insumo para futuras entregas, especialmente para mejorar la normalización del texto, ampliar el vocabulario clínico y ajustar la diferencia entre mención clara y probable de cáncer.

6. Conclusiones

Este capítulo presenta las conclusiones derivadas del desarrollo y evaluación de los modelos propuestos, a partir de los resultados obtenidos y de su interpretación frente al problema planteado. En esta sección se sintetizan los principales hallazgos del estudio, su aporte al RPCa-M y sus alcances para trabajos futuros.

6.1. Respuesta a la pregunta de investigación

Este estudio partió del problema que representa para el RPCa-M la revisión manual de registros de defunción redactados en texto libre, caracterizados por heterogeneidad, ambigüedad clínica y variabilidad en el nivel de detalle. Para responder a esta necesidad, se desarrolló una comparación aplicada entre tres enfoques de procesamiento de lenguaje natural, evaluados sobre una misma base de datos y bajo un criterio de selección orientado no solo al desempeño global, sino también a la reducción de errores críticos para el propósito del registro.

Los resultados obtenidos permiten concluir que el modelo SVM fue la alternativa que respondió de mejor manera al propósito del estudio. Entre los tres enfoques evaluados, este modelo mostró el mejor equilibrio entre desempeño global, estabilidad por clase y capacidad para reducir falsos 0, que constituyó el criterio más importante dentro de la evaluación.

Esta conclusión se sostiene en dos aspectos principales. En primer lugar, el modelo SVM alcanzó la mayor exactitud global entre las alternativas comparadas. En segundo lugar, presentó el mejor comportamiento frente al error más crítico del estudio, al minimizar con mayor claridad los casos en los que un registro con mención clara o probable de cáncer era clasificado como ausencia de mención. Este punto fue especialmente relevante porque, en el contexto del RPCa-M, resultó

más importante conservar dentro del proceso de revisión los registros potencialmente relevantes que maximizar una métrica aislada sin considerar ese riesgo.

Los resultados también mostraron que los tres modelos ofrecieron aportes distintos. El modelo basado en reglas destacó por su interpretabilidad y por su utilidad en los casos más claros. El modelo basado en *Transformer*s mostró una capacidad alta para recuperar casos relevantes. Sin embargo, el SVM fue el enfoque que integró de forma más consistente estas ventajas con un mejor control del error crítico. Por ello, fue el modelo que ofreció la respuesta más sólida dentro de la comparación realizada.

En conjunto, el estudio permitió establecer que sí es posible apoyar la revisión de registros de defunción mediante un modelo de procesamiento de lenguaje natural y que, entre las alternativas evaluadas, el SVM fue el enfoque con mayor utilidad para este propósito dentro del Registro Poblacional de Cáncer de Manizales.

6.2. Principales hallazgos del estudio

Uno de los principales hallazgos del estudio fue que los tres modelos evaluados alcanzaron un desempeño suficientemente alto para apoyar la clasificación de registros de defunción según indicios de cáncer. Esto muestra que la tarea sí puede abordarse mediante procesamiento de lenguaje natural y que, incluso con enfoques metodológicos distintos, es posible obtener resultados útiles para el contexto del registro.

El hallazgo más importante fue la diferencia entre desempeño general y utilidad aplicada. Aunque los tres modelos ofrecieron resultados favorables, no respondieron igual frente al criterio más relevante del estudio. El SVM fue el que mantuvo el comportamiento más consistente, porque

combinó mejor el desempeño global con la reducción de falsos 0. El modelo basado en reglas conservó como fortaleza principal la interpretabilidad, mientras que el modelo basado en *Transformer*s mostró una mayor sensibilidad para recuperar casos relevantes. Sin embargo, ninguno de estos dos enfoques logró un equilibrio tan sólido como el observado en el SVM.

Otro hallazgo importante fue la dificultad persistente de la clase 2. En los tres modelos, esta categoría presentó más problemas de clasificación que las clases 0 y 1. Esto indica que la mayor complejidad del problema no estuvo en reconocer registros claramente negativos o claramente positivos, sino en resolver aquellos casos en los que la mención de cáncer era menos explícita o más ambigua desde el punto de vista textual.

En conjunto, el estudio permitió dar cumplimiento a los objetivos específicos propuestos. Primero, se preparó y estructuró una base de registros de defunción adecuada para el desarrollo y evaluación de los modelos. Segundo, se implementaron y ajustaron tres enfoques de procesamiento de lenguaje natural con lógicas metodológicas distintas. Tercero, su desempeño fue comparado mediante métricas globales, comportamiento por clase y capacidad para reducir falsos 0. Finalmente, esta comparación permitió seleccionar el modelo SVM como la alternativa con mayor utilidad para apoyar la revisión de potenciales casos en el RPCa-M.

6.3. Aporte del estudio al RPCa-M

El principal aporte del estudio al RPCa-M fue demostrar que la revisión de registros de defunción puede apoyarse en un modelo de procesamiento de lenguaje natural construido sobre sus propias necesidades operativas y sobre una base textual real del registro. Esto permite pasar de

una revisión completamente manual a una estrategia en la que el modelo sirve como apoyo para identificar con mayor oportunidad los registros que contienen mención clara o probable de cáncer.

Este aporte es relevante porque el proceso de revisión de defunciones exige tiempo, atención sostenida y lectura detallada de textos heterogéneos, con distintos niveles de precisión clínica. En ese contexto, contar con una herramienta de apoyo puede ayudar a priorizar mejor los casos que merecen una revisión más cuidadosa y disminuir el riesgo de que registros potencialmente relevantes queden por fuera del proceso. El valor del estudio, entonces, no estuvo solo en comparar algoritmos, sino en mostrar una solución concreta para una necesidad real del registro.

Además, el trabajo deja una base metodológica útil para desarrollos posteriores dentro del mismo RPCa-M. La preparación de la muestra etiquetada, la estructuración del texto de entrada, la comparación entre enfoques y la definición de criterios de evaluación centrados en falsos 0 constituyen insumos que pueden aprovecharse en futuras aplicaciones. En ese sentido, el estudio no solo aporta un resultado puntual, sino también una ruta de trabajo que puede extenderse a otros flujos documentales del registro. Por ello, su aporte no se limita al modelo seleccionado, sino que incluye también una experiencia metodológica transferible dentro del contexto institucional del RPCa-M.

Desde una perspectiva académica, este aporte también se refleja en la aplicación comparativa de enfoques de procesamiento de lenguaje natural en un problema real de salud pública y registros poblacionales de cáncer. La contribución del estudio no está en proponer un algoritmo nuevo, sino en adaptar, comparar y evaluar métodos existentes bajo una necesidad concreta del RPCa-M, con una base textual propia y con criterios de evaluación definidos según el riesgo operativo de omitir posibles casos. En este sentido, el trabajo aporta evidencia

metodológica sobre el uso de modelos de PLN en registros de defunción, un tipo de documento que no aparece con tanta frecuencia en la literatura revisada. Además, plantea una ruta que puede servir como referencia para futuras investigaciones en PLN clínico aplicado a registros de cáncer, especialmente en contextos hispanohablantes.

6.4. Entregable al registro y valoración del resultado.

El entregable final permitió que el modelo seleccionado no quedara únicamente como un resultado de evaluación, sino como una herramienta aplicable dentro del contexto del RPCa-M.

Su organización como una aplicación local facilitó que la clasificación pudiera realizarse en el entorno del usuario final, manteniendo el control sobre la información y evitando la exposición de datos sensibles en plataformas externas. Esto fue importante porque el trabajo se desarrolló con registros de defunción, los cuales requieren un manejo cuidadoso dentro del proceso del registro.

Uno de los aportes más relevantes del entregable fue la reducción del tiempo necesario para obtener una primera clasificación de los registros. En una base cercana a 40.000 registros, la aplicación generó la clasificación completa en aproximadamente 2 minutos. Este resultado muestra que el modelo puede apoyar de manera práctica una etapa que, realizada manualmente, exige una carga considerable de tiempo y revisión por parte del personal del registro.

La valoración del usuario final también permitió observar que la herramienta fue comprendida y considerada útil como apoyo al proceso de revisión. Esta percepción es importante

porque el valor del modelo no depende solo de sus métricas, sino también de que pueda ser usado e interpretado por quienes participan directamente en la captación de casos.

En este sentido, el entregable no reemplaza la revisión experta ni los controles de calidad propios del registro. Su aporte consiste en ofrecer una primera clasificación que permite priorizar registros, orientar la revisión manual y disminuir el tiempo requerido para revisar grandes volúmenes de información. Por ello, el producto final representa un aporte práctico para el RPCa-M y una base para futuras mejoras o validaciones en condiciones rutinarias de uso.

6.5. Limitaciones del estudio

Este estudio presentó varias limitaciones que deben tenerse en cuenta al interpretar sus resultados. La primera está relacionada con la muestra etiquetada. Aunque esta fue suficiente para entrenar y comparar los tres modelos bajo una misma base, siguió siendo una fracción de la base total y, por tanto, no agotó toda la variabilidad posible en la forma como se diligencian los registros de defunción. Esto es especialmente importante en textos clínicos breves, donde pequeñas diferencias de redacción pueden cambiar el sentido de la información.

Una segunda limitación estuvo en el desbalance entre clases, particularmente en la clase 2, que representó la categoría más difícil en todos los enfoques. Su menor frecuencia redujo la cantidad de ejemplos disponibles para el aprendizaje y volvió más exigente su clasificación. Por esta razón, aunque el estudio permitió comparar los modelos y seleccionar la alternativa más adecuada, los resultados en esta clase deben interpretarse con cautela.

Otra limitación importante se relaciona con la interpretación del desempeño comparativo. Aunque el modelo SVM fue seleccionado como la alternativa con mejor equilibrio para este estudio, su superioridad debe entenderse dentro de las condiciones específicas de la base utilizada, la tarea de clasificación definida y el criterio de reducción de falsos 0. Por tanto, los resultados no implican que este enfoque sea necesariamente superior en cualquier contexto de PLN clínico, sino que fue el más adecuado para esta muestra, este tipo de documento y las necesidades operativas del RPCa-M. Además, la dificultad observada en la clase 2 muestra que los casos ambiguos siguen dependiendo de la calidad del texto original y de la interpretación experta, por lo que el modelo debe entenderse como una herramienta de apoyo y no como un mecanismo automático de decisión final

También debe tenerse en cuenta que la validación realizada fue interna. Es decir, los conjuntos de entrenamiento, validación y prueba salieron de la misma base institucional. Por esto, el modelo no se evaluó con datos de un periodo posterior, ni con información de otro registro poblacional de cáncer, ni con datos generados bajo condiciones operativas diferentes.

Así, aunque el modelo SVM mostró un desempeño alto, estos resultados deben entenderse dentro del contexto específico en el que fue evaluado. No significa necesariamente que el modelo vaya a comportarse de la misma manera en otros periodos, ciudades o registros. Para tener una evidencia más sólida sobre su capacidad de generalización, futuras investigaciones deberían incluir validaciones temporales, externas o multicéntricas.

En el caso del modelo basado en Transformers, el número de épocas de entrenamiento también debe interpretarse con cuidado. Cuando se hace fine-tuning con bases pequeñas y desbalanceadas, entrenar durante demasiado tiempo puede hacer que el modelo termine

memorizando patrones propios del conjunto de entrenamiento, en lugar de aprender relaciones que pueda aplicar a nuevos datos.

Por esta razón, la cantidad de épocas utilizada no debe entenderse solo como una restricción de infraestructura. También hizo parte de una decisión metodológica relacionada con el tamaño de la muestra, la estabilidad observada en la validación y la necesidad de reducir el riesgo de sobreajuste.

También debe considerarse que, aunque el certificado de defunción tiene un formato estándar, la calidad del diligenciamiento puede variar entre profesionales y contextos. En este sentido, el desempeño de la herramienta depende de la claridad, veracidad y completitud de los datos consignados en los campos de causa de muerte y estados patológicos relacionados. Por ello, aunque el desarrollo podría ser útil para otros registros de cáncer, su uso en otros contextos requeriría una validación previa con datos propios.

Además, la posibilidad de compartir e implementar el modelo presentó una dificultad adicional asociada a las condiciones reales de uso. Al trabajar con información sensible, fue necesario mantener medidas de protección sobre los datos y evitar cualquier exposición innecesaria de contenido clínico o identificable. Esto hizo que la transferencia del modelo no dependiera únicamente de su desempeño, sino también de su capacidad de adaptarse a un contexto operativo concreto.

Por último, el trabajo se concentró en la clasificación de registros según indicios de cáncer y no en una automatización completa del proceso registral. En consecuencia, el modelo debe entenderse como una herramienta de apoyo a la revisión, no como un reemplazo de la validación experta ni de los procedimientos de control de calidad que exige el registro.

6.6. Proyecciones de trabajo futuro

Una primera proyección consiste en ampliar y fortalecer la base etiquetada, con especial atención a los registros de la clase 2, que fueron los más difíciles de clasificar en los tres enfoques evaluados. Contar con más ejemplos revisados por expertos permitiría capturar mejor la variabilidad del lenguaje presente en los certificados de defunción y ofrecer más evidencia para entrenar modelos capaces de manejar con mayor precisión los casos ambiguos o inciertos.

De manera complementaria, se podrían probar enfoques híbridos que combinen la interpretabilidad de los modelos basados en reglas con la capacidad de aprendizaje de modelos supervisados como SVM o Transformers. Esto permitiría integrar reglas clínicas explícitas, vocabularios especializados y aprendizaje automático dentro de un mismo flujo de clasificación, con el fin de mejorar la detección de casos relevantes y reducir errores críticos.

También sería importante avanzar hacia una evaluación en condiciones de uso más cercanas al contexto real del RPCa-M. Esto permitiría observar con mayor claridad cómo se comporta el modelo como apoyo al flujo de revisión, qué tipo de errores persisten en la práctica y qué ajustes serían necesarios para facilitar su integración en la rutina del registro.

A partir de la valoración del usuario final y de la revisión posterior realizada sobre una base de prueba de 400 casos, también se identificaron posibles mejoras para futuras entregas de la aplicación. Entre ellas se encuentra la incorporación de filtros en el archivo de salida, con el fin de facilitar la revisión de los registros según la clasificación generada por el modelo. También podría explorarse la posibilidad de aceptar otros formatos de entrada, como archivos de texto, siempre que estos conserven la información necesaria para realizar la clasificación. Además, sería

importante ampliar el vocabulario clínico, fortalecer la normalización de errores de escritura y caracteres especiales, y ajustar la diferenciación entre registros con mención clara y probable de cáncer. Esto podría contribuir a reducir falsos negativos y mejorar la clasificación de diagnósticos menos frecuentes o escritos con variaciones.

Otra mejora futura sería ampliar la documentación técnica del entregable. Aunque la aplicación fue organizada para facilitar su uso local, sería útil incluir con mayor detalle los requerimientos mínimos de hardware y software, así como recomendaciones para su ejecución en diferentes equipos. Esto permitiría que el usuario final tenga mayor claridad sobre las condiciones necesarias para usar la herramienta de forma estable.

Finalmente, el trabajo deja abierta la posibilidad de extender esta línea a otros tipos de documentos y procesos del registro. La experiencia obtenida en la preparación del texto, en la comparación de modelos y en la priorización del error crítico puede servir como base para futuras aplicaciones de procesamiento de lenguaje natural en otras fuentes documentales del RPCa-M, siempre que estas se desarrollen de acuerdo con las condiciones de seguridad, disponibilidad de datos y capacidad técnica del entorno de uso.

7. Anexos

- Anexo 1. Base de datos de artículos de la revisión de la literatura

Disponible en: Articulos.xlsx

- Anexo 2. Gráficas complementarias de la revisión de la literatura

Disponible en: https://github.com/mafedcp65-netizen/Revision-de-la-literatura

- Anexo 3. Repositorios de los tres modelos comparados

Disponible en: https://github.com/mafedcp65-netizen/Trabajo-de-grado

- Anexo 4. Repositorio del entregable al registro de defunciones

Disponible en: https://github.com/mafedcp65-netizen/Codigo-registro

- Anexo 5. Encuesta sobre la aplicación entregada al registro

Disponible en: https://forms.cloud.microsoft/r/3BqXfNTjmh?origin=lprLink

8. Referencias

Alawad, M., Gao, S., Qiu, J. X., Yoon, H. J., Blair Christian, J., Penberthy, L., Mumphrey, B., Wu, X. C., Coyle, L., & Tourassi, G. (2020). Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. *Journal of the American Medical Informatics Association*, *27*(1), 89–98. https://doi.org/10.1093/jamia/ocz153

Arias-Ortiz, N., & Rodríguez-Betancourt, J. D. (2022). Trends in cancer incidence and mortality in Manizales, Colombia, 2008-2017. *Colombia Medica*, *53*(1). https://doi.org/10.25100/cm.v53i1.4920

Bozkurt, S., Magnani, C. J., Seneviratne, M. G., Brooks, J. D., & Hernandez-Boussard, T. (2022). Expanding the Secondary Use of Prostate Cancer Real World Data: Automated Classifiers for Clinical and Pathological Stage. *Frontiers in Digital Health*, *4*. https://doi.org/10.3389/fdgth.2022.793316

Bray, F., & Parkin, D. M. (2009). Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness. *European Journal of Cancer*, *45*(5), 747–755. https://doi.org/10.1016/j.ejca.2008.11.032

Bray, Freddie. (2015). *Planificación y desarrollo de registros de cáncer de base poblacional en los países de ingresos bajos y medios*. Agencia Internacional de Investigación sobre el Cáncer.

Dan Jurafsky, & James H. Martin. (2025). *Speech and Language Processing*. https://web.stanford.edu/~jurafsky/slp3/

Digan, W., Névéol, A., Neuraz, A., Wack, M., Baudoin, D., Burgun, A., & Rance, B. (2021). Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP

suites. *Journal of the American Medical Informatics Association*, *28*(3), 504–515. https://doi.org/10.1093/jamia/ocaa261

Fritz, A., Percy, C., Jack, A., Shanmugaratnam, K., Sobin, L., Parkin, D. M., & Whelan, S. (2003). *Clasificación Internacional de Enfermedades para Oncología Tercera edición Editores CIE-O*.

García, L. S., Bravo, L. E., Collazos, P., Ramírez, O., Carrascal, E., Nuñez, M., Portilla, N., & Millan, E. (2018). Cali cancer registry methods. *Colombia Medica*, *49*(1), 109–120. https://doi.org/10.25100/cm.v49i1.3853

García Subies, G., Barbero Jiménez, Á., & Martínez Fernández, P. (2024). A comparative analysis of Spanish Clinical encoder-based models on NER and classification tasks. *Journal of the American Medical Informatics Association*, *31*(9), 2137–2146. https://doi.org/10.1093/JAMIA/OCAE054

González, R., & Pomares, A. (2012). *La investigación científica basada en el diseño como eje de proyectos de investigación en ingeniería*. https://www.researchgate.net/publication/234660620_La_investigacion_cientifica_basada_en_el_diseno_como_eje_de_proyectos_de_investigacion_en_ingenieria

Hammami, L., Paglialonga, A., Pruneri, G., Torresani, M., Sant, M., Bono, C., Caiani, E. G., & Baili, P. (2021). Automated classification of cancer morphology from Italian pathology reports using Natural Language Processing techniques: A rule-based approach. *Journal of Biomedical Informatics*, *116*. https://doi.org/10.1016/j.jbi.2021.103712

Hsu, E., Hanson, H., Coyle, L., Stevens, J., Tourassi, G., & Penberthy, L. (2024). Machine learning and deep learning tools for the automated capture of cancer surveillance data. *JNCI Monographs*, *2024*(65), 145–151. https://doi.org/10.1093/jncimonographs/lgae018

Hua Xu, & Demner Fushman. (2024). *Natural Language Processing in Biomedicine*. https://link.springer.com/book/10.1007/978-3-031-55865-8

James, G., Witten, D., Hastie, T., & Taylor, J. (2023). *An introduction to statistical learning*.

Koopman, B., Zuccon, G., Nguyen, A., Bergheim, A., & Grayson, N. (2015). Automatic ICD-10 classification of cancers from free-text death certificates. *International Journal of Medical Informatics*, *84*(11), 956–965. https://doi.org/10.1016/j.ijmedinf.2015.08.004

Koopman, B., Zuccon, G., Nguyen, A., Bergheim, A., & Grayson, N. (2018). Extracting cancer mortality statistics from death certificates: A hybrid machine learning and rule-based approach for common and rare cancers. *Artificial Intelligence in Medicine*, *89*, 1–9. https://doi.org/10.1016/j.artmed.2018.04.011

Mendoza-Urbano, D. M., Garcia, J. F., Moreno, J. S., Bravo-Ocaña, J. C., Riascos, A. J., Harvey, A. Z., & Prada, S. I. (2023). Automated extraction of information from free text of Spanish oncology pathology reports. *Colombia Medica*, *54*(1). https://doi.org/10.25100/cm.v54i1.5300

Michalek, I. M., Martos, C., Caetano dos Santos, F. L., Giusti, F., Degerlund, H., Neamtiu, L., Taraszkiewicz, L., Van Eycken, L., & Visser, O. (2024). Advancing data collection and analysis: 2023 revised European Network of Cancer Registries recommendations for standard dataset. In *European Journal of Cancer* (Vol. 199). Elsevier Ltd. https://doi.org/10.1016/j.ejca.2024.113557

Parkin, D. M., & Bray, F. (2009). Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness. *European Journal of Cancer*, *45*(5), 756–764. https://doi.org/10.1016/J.EJCA.2008.11.033

Peluso, A., Danciu, I., Yoon, H. J., Yusof, J. M., Bhattacharya, T., Spannaus, A., Schaefferkoetter, N., Durbin, E. B., Wu, X. C., Stroup, A., Doherty, J., Schwartz, S., Wiggins, C., Coyle, L., Penberthy, L., Tourassi, G. D., & Gao, S. (2024). Deep learning uncertainty quantification for clinical text classification. *Journal of Biomedical Informatics*, *149*. https://doi.org/10.1016/j.jbi.2023.104576

Pironet, A., Poirel, H. A., Tambuyzer, T., De Schutter, H., van Walle, L., Mattheijssens, J., Henau, K., Van Eycken, L., & Van Damme, N. (2021). Machine Learning-Based Extraction of Breast Cancer Receptor Status From Bilingual Free-Text Pathology Reports. *Frontiers in Digital Health*, *3*. https://doi.org/10.3389/fdgth.2021.692077

Schulz, S., Fix, S., & Hartz, T. (2021). *Comparative evaluation of automated information extraction from pathology reports in three German cancer registries* (Vol. 17, Number 1)-10.3205/mibe000215.

Universidad de Caldas. (2025). Desarrollo de un asistente virtual basado en Lenguaje Natural e Inteligencia artificial para la optimización del procesamiento de datos en el Registro Poblacional de Cáncer de Manizales. *Universidad de Caldas*, 1–45.

Van Eycken, L. J., Giannopoulos, E., Tittenbrun, Z., Piñeros, M., Mery, L., Rahal, R., Helliwell, T. R., Kohler, B. A., Aitken, J. F., Rous, B. A., O'Sullivan, B., Johnson, S. E., Gospodarowicz, M. K., & Brierley, J. D. (2025). Future of population-based cancer registries: A global

perspective—A survey of population-based cancer registries. *International Journal of Cancer*. https://doi.org/10.1002/ijc.35516

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2023). *Attention Is All You Need*. https://arxiv.org/pdf/1706.03762

Viscosi, C., Fidelbo, P., Benedetto, A., Varvarà, M., & Ferrante, M. (2022). Selection of diagnosis with oncologic relevance information from histopathology free text reports: A machine learning approach. *International Journal of Medical Informatics*, *160*. https://doi.org/10.1016/j.ijmedinf.2022.104714

Xie, F., Lee, J., Munoz-Plaza, C., Hahn, E., & Chen, W. (2017). Application of text information extraction system for real-time cancer case identification in an integrated healthcare organization. *Journal of Pathology Informatics*, *8*(1). https://doi.org/10.4103/jpi.jpi_55_17

Docker Inc. (2026). *Docker documentation*. https://docs.docker.com/

Eyre, H., Chapman, A. B., Peterson, K. S., Shi, J., Alba, P. R., Jones, M. M., Box, T. L., DuVall, S. L., & Patterson, O. V. (2021). Launching into clinical space with medspaCy: A new clinical text processing toolkit in Python. *AMIA Annual Symposium Proceedings, 2021*, 438–447.

Soldaini, L., & Goharian, N. (2016). QuickUMLS: A fast, unsupervised approach for medical concept extraction. En *Proceedings of the 2nd SIGIR Workshop on Medical Information Retrieval (MedIR)*, 1–4.