FARAUTE, Vol. 17, No. 2: 13-24, 2022 Print ISSN 1698-7418
Online ISSN 2542-3061

APLICACIÓN WEB BASADA EN TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA LA RECOMENDACIÓN Y CLASIFICACIÓN DE RECURSOS EN LA BIBLIOTECA CENTRAL DE LA UNIVERSIDAD DE CARABOBO

Herrera, M. y Herrera, J.
Universidad de Carabobo. Facultad Experimental de Ciencias y Tecnología.
Dpto. de Computación. Bárbula, Edo. Carabobo, Venezuela.
mherrera@uc.edu.ve, juanpherrarar1998@gmail.com
Recibido: 23/09/2025, Revisado: 15/10/2025, Aceptado: 15/11/2025

Resumen

En el contexto de esta investigación, se define el vocabulario controlado como un conjunto estructurado y organizado de términos y conceptos utilizados para describir y categorizar de manera precisa los recursos bibliográficos presentes en la biblioteca virtual. El objetivo de este trabajo fue implementar una aplicación *web* basada en técnicas de aprendizaje automático para la recomendación en la clasificación de recursos en la Biblioteca Virtual de la Universidad de Carabobo, utilizando un vocabulario controlado estructurado para la categorización precisa de recursos bibliográficos, así como técnicas de aprendizaje automático basadas en el modelo *RoBERTa*. La metodología combinó Investigación-Acción y *Scrum*, adaptándose a las limitaciones de datos etiquetados y recursos computacionales. Los resultados demostraron la superioridad de *RoBERTa*, alcanzando un 94.2% de precisión en artículos científicos y reduciendo el tiempo de clasificación manual en un 60%. El sistema integra un vocabulario controlado y se validó con métricas de precisión, *recall* y *F1-score*, superando a modelos *CNN* (31.4% F1) y *BERT* (79.5% F1). Como contribución, se ofrece un caso de estudio replicable para bibliotecas con tecnologías *OneSearch*, aunque se identifican limitaciones en la generalización a dominios multilingües. El trabajo sugiere futuras líneas en *fine-tuning* de modelos como Llama 3 y estrategias para mitigar sesgos.
Palabras clave: Clasificación de texto, *RoBERTa*, bibliotecas virtuales, vocabulario controlado, *SCRUM*.

Web application based on machine learning techniques for the recommendation and classification of resources in the Central Library of the University of Carabobo

Abstract

In the context of this research, controlled vocabulary is defined as a structured and organized set of terms and concepts used to accurately describe and categorize bibliographic resources in the virtual library. The objective of this work was to implement a web application based on machine learning techniques for resource classification and recommendation in the Virtual Library of the University of Carabobo, using a structured controlled vocabulary for precise bibliographic categorization, as well as machine learning techniques based on the RoBERTa model. The methodology combined Action Research and Scrum, adapting to limitations in labeled data and computational resources. The results demonstrated the superiority of RoBERTa, achieving 94.2% precision in scientific articles and reducing manual classification time by 60%. The system integrates a controlled vocabulary and was validated using precision, recall, and F1-score metrics, outperforming CNN (31.4% F1) and BERT (79.5% F1) models. As a contribution, this work provides a replicable case study for libraries using OneSearch technologies, though limitations were identified in generalization to multilingual domains. Future directions include fine-tuning models like Llama 3 and strategies to mitigate biases.
Keywords: Text classification, RoBERTa, virtual libraries, controlled vocabulary, SCRUM.

1. Introducción

En el contexto actual de la información digital y el acceso a recursos en línea, las bibliotecas virtuales desempeñan un papel fundamental en la difusión del conocimiento y el apoyo a la educación e investigación. Según Sánchez (s.f.), "Las bibliotecas virtuales son un recurso de información necesario para el acceso y manejo de información digitalizada y los últimos años son testigos de los esfuerzos que han hecho muchas organizaciones para potenciar su desarrollo". La Biblioteca Central de la Universidad de Carabobo es un ejemplo destacado de esta tendencia, brindando a la comunidad universitaria un repositorio digital de recursos académicos.

Sin embargo, a medida que la cantidad de materiales en la biblioteca virtual aumenta, surge la necesidad de mejorar la precisión en los resultados de búsqueda y la organización de los recursos disponibles. La clasificación de documentos es un proceso complejo y laborioso que requiere conocimientos especializados en biblioteconomía, constituyéndose la gestión eficiente del acervo documental en constante expansión en un verdadero desafío. En la Dirección de Biblioteca Central de la Universidad de Carabobo, este proceso manual consume aproximadamente 40 horas semanales del personal especializado y presenta una tasa de error del 15% en la clasificación, lo que dificulta significativamente la recuperación efectiva de información por parte de los usuarios.

Este escenario no es exclusivo de la Universidad de Carabobo, diversos estudios reportan que, en bibliotecas universitarias de América Latina, el tiempo invertido en la clasificación manual puede superar las 1,500 horas anuales por institución, con tasas de error que oscilan entre el 10% y el 20% (Mahmud, 2024). Estas ineficiencias no solo afectan la productividad del personal, sino que también impactan negativamente en la experiencia de los usuarios, quienes enfrentan dificultades para localizar información relevante de manera oportuna.

Si bien existen propuestas recientes que exploran el uso de inteligencia artificial (IA) para automatizar la catalogación y clasificación en bibliotecas, la mayoría de los estudios se centran en contextos de grandes bibliotecas nacionales o internacionales, dejando de lado las particularidades y limitaciones de las bibliotecas universitarias latinoamericanas. Según Mahmud (2024), las tecnologías de IA, especialmente el aprendizaje automático y el procesamiento de lenguaje natural, han demostrado mejorar la eficiencia y consistencia en la creación de metadatos y la clasificación de recursos. Sin embargo, persisten desafíos relacionados con la calidad de los datos, la integración con sistemas heredados y la capacitación del personal bibliotecario. Además, la literatura señala la necesidad de adaptar estas soluciones a contextos específicos, considerando factores como la diversidad lingüística y la disponibilidad de recursos tecnológicos.

Si bien existen propuestas recientes que exploran el uso de inteligencia artificial (IA) para automatizar la catalogación y clasificación en bibliotecas, la mayoría de los estudios se centran en contextos de grandes bibliotecas nacionales o internacionales, dejando de lado las particularidades y limitaciones de las bibliotecas universitarias latinoamericanas. Según Mahmud (2024), las tecnologías de IA, especialmente el aprendizaje automático y el procesamiento de lenguaje natural, han demostrado mejorar la eficiencia y consistencia en la creación de metadatos y la clasificación de recursos. Sin embargo, persisten desafíos relacionados con la calidad de los datos, la integración con sistemas heredados y la capacitación del personal bibliotecario. Además, la literatura señala la necesidad de adaptar estas soluciones a contextos específicos, considerando factores como la

diversidad lingüística y la disponibilidad de recursos tecnológicos.

En este contexto, surge la propuesta de implementar una aplicación *web* que utilice técnicas de aprendizaje automático, como las redes neuronales, para automatizar el proceso de recomendación y clasificación de los recursos en la Dirección Central de Biblioteca de la Universidad de Carabobo. El objetivo principal de esta investigación consistió en implementar una solución que lograra una mejora en la precisión de los resultados de búsqueda y facilitara la organización y recuperación de los materiales por parte de los usuarios. Según García-Robledo y Pérez-López (2022), "la aplicación de técnicas de aprendizaje automático, como las redes neuronales, ha demostrado ser eficaz en la automatización de la clasificación y recomendación de recursos en bibliotecas virtuales, mejorando la precisión de los resultados de búsqueda y facilitando el acceso a la información por parte de los usuarios".

Este trabajo aborda la problemática descrita y proporciona una solución efectiva para la optimización de la biblioteca virtual. La implementación de una aplicación basada en redes neuronales permitió una clasificación más precisa y automatizada de los recursos, lo que mejora la experiencia de los usuarios en la búsqueda y acceso a los materiales de su interés. A lo largo de este trabajo, se llevó a cabo un análisis de los conceptos teóricos y las técnicas de clasificación de texto utilizando redes neuronales, explorando las mejores prácticas y los enfoques más efectivos en el campo de la biblioteconomía y el aprendizaje automático para garantizar el éxito de la implementación.

Además, se realizó un estudio de viabilidad y se diseñó una arquitectura para la aplicación *web*, teniendo en cuenta los requisitos específicos de la Dirección de Biblioteca Central de la Universidad de Carabobo. Igualmente, se utilizaron conjuntos de datos relevantes y se evaluó la calidad de las recomendaciones generadas por la red neuronal, con el fin de realizar mejoras y ajustes necesarios. El resto del documento se estructura de la siguiente manera: la Sección 2 detalla la metodología empleada, la Sección 3 presenta los resultados y su análisis, y finalmente la Sección 4 expone las conclusiones y trabajos futuros.

2. Metodología

Esta sección describe el método de investigación y el enfoque de desarrollo utilizados en este estudio. Adoptamos una estrategia mixta que combina la Investigación-Acción para el marco investigativo con *Scrum* para el desarrollo de *software*, complementada con enfoques técnicos específicos para la recolección de datos y la implementación del aprendizaje automático. Esta combinación metodológica fue seleccionada específicamente para abordar la complejidad inherente a la clasificación documental, donde la Investigación-Acción permitió la iteración y reflexión continua sobre los criterios de clasificación y la efectividad del sistema (Smith, 2007), mientras que *Scrum* facilitó el desarrollo ágil de la aplicación, adoptando de forma rápida los requisitos evolutivos del sistema de biblioteca.

Específicamente, la Investigación-Acción, fue elegida por su efectividad en combinar la acción práctica con la investigación, permitiendo la mejora continua de la solución a través de ciclos iterativos. Asimismo, resultó particularmente beneficiosa para la clasificación documental, ya que permitió evaluar y refinar iterativamente los criterios de clasificación, basándose en la retroalimentación del sistema y los usuarios.

La implementación del proyecto se desarrolló en cuatro fases principales. En la fase

de planificación, se realizó una evaluación inicial para identificar los requisitos específicos y los desafíos en el sistema de clasificación de la biblioteca digital, formulando estrategias basadas en el marco de preguntas fundamentales de Kemmis (1998). En la fase de acción, la implementación se llevó a cabo utilizando el marco *Scrum*, permitiendo un desarrollo iterativo y retroalimentación continua, lo que involucró el desarrollo real del sistema de clasificación y su integración con la infraestructura existente de la biblioteca. En la fase de observación, se emplearon métodos tanto cualitativos como cuantitativos para recopilar y analizar datos sobre el rendimiento y la efectividad del sistema, siguiendo el enfoque de métodos mixtos de Creswell y Plano Clark (2018). Finalmente, en la fase de reflexión, se analizaron los resultados obtenidos y se refinó el sistema con base en las lecciones aprendidas (Figura 1).

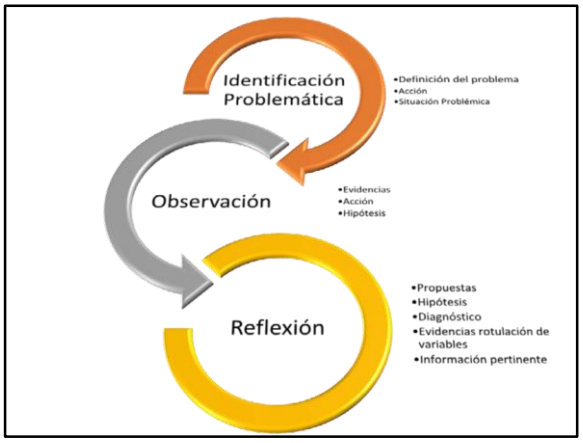

Figura 1. Metodología Investigación – Acción. Fuente: ResearchGate (s.f.)

Para el desarrollo del software, se utilizó el marco *Scrum*, particularmente adecuado para proyectos con requisitos evolutivos (Sutherland y Schwaber, 2017). El ciclo de desarrollo se organizó en iteraciones o *sprints* de dos semanas, con un equipo de desarrollo de 1 persona, reuniones periódicas con los involucrados, sesiones de planificación y retrospectiva de *sprint*, y revisiones regulares con los diferentes actores. Por su parte, los artefactos clave del desarrollo incluyeron la pila del producto, la pila del *sprint*, los gráficos de quema y el registro de impedimentos (Figura 2).

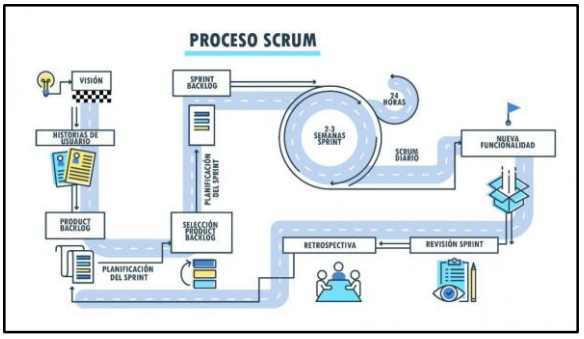

Fig. 2. Proceso Scrum. Fuente: Cámara de Comercio de España (s.f.)

El proceso de recolección de datos involucró técnicas de *web scraping* siguiendo la metodología de Langer et al. (2021). Este proceso comprendió la identificación y validación de fuentes, la extracción de datos utilizando herramientas basadas en *Python* como *BeautifulSoup* y *Selenium*, el preprocesamiento y normalización de texto, y la verificación y validación de la calidad de los datos. Durante este proceso, se enfrentaron desafíos como bloqueos de *IP* y formatos inconsistentes de datos, que fueron superados mediante la implementación de rotación de *proxies* y algoritmos de normalización robustos.

El conjunto de datos final comprendió un total de 15,847 resúmenes de documentos distribuidos de la siguiente manera: Artículos científicos (4,523 documentos, 28.5%), Libros (3,891 documentos, 24.6%), Tesis (2,945 documentos, 18.6%), Revistas (2,678 documentos, 16.9%), y Documentos Técnicos (1,810 documentos, 11.4%). La longitud media del texto fue de 321 palabras, con una desviación estándar de 62 palabras. Para el manejo de textos multilingües, se implementó un sistema de detección automática de idioma utilizando la librería *langdetect*, seguido de normalización específica por idioma. Los términos técnicos fueron procesados mediante

un vocabulario controlado estructurado que incluyó 2,847 términos especializados, siguiendo las recomendaciones de López-Escobar et al. (2022) para la normalización de vocabulario controlado en bibliotecas digitales.

El sistema de clasificación de texto se implementó utilizando un enfoque de red neuronal profunda (*Deep Learning*), evaluando varias arquitecturas. El preprocesamiento de datos incluyó la limpieza y normalización de texto, la eliminación de palabras vacías, la lematización y la representación vectorial mediante *embeddings*.

Se evaluaron múltiples arquitecturas de aprendizaje profundo para clasificación de texto, incluyendo redes neuronales convolucionales (CNN) y modelos fundacionales basados en *Transformers*. Tras una revisión comparativa, se seleccionó *RoBERTa-base* (Liu et al., 2019) como modelo principal, implementado a través de la librería *Hugging Face Transformers* (Hugging Face, 2023), debido a su robustez, rendimiento demostrado en tareas de clasificación y disponibilidad de recursos en español. No se emplearon modelos más recientes como GPT-3/4 o variantes de *Transformers* más grandes debido a restricciones de recursos computacionales y a la necesidad de interpretabilidad y eficiencia en el entorno de despliegue. Además, *RoBERTa* había mostrado un equilibrio óptimo entre precisión y eficiencia para tareas de clasificación documental (Hugging Face, 2023).

Detalles técnicos del modelo RoBERTa:

- Arquitectura: RoBERTa-base (12 capas, 768 dimensiones, 12 cabezas de atención, 125M parámetros)
- La última capa (capa de salida) fue adaptada para tener un número de neuronas igual al número de categorías de clasificación definidas en el problema, permitiendo así que el modelo pueda predecir directamente cualquiera de las clases posibles.

Hiperparámetros principales:

- Tasa de aprendizaje: $1e^{-5}$, $2e^{-5}$, $3e^{-5}$ (seleccionada por búsqueda en cuadrícula)
- Tamaño de lote: 16, 32, 64
- Número de épocas: 10, 20, 30
- Estrategia de parada temprana: detención si no hay mejora en la métrica de validación durante 5 épocas consecutivas
- División de datos: entrenamiento (70%), validación (15%), prueba (15%)
- Métricas de evaluación: precisión, recuperación, puntuación F1, exactitud de clasificación
- Validación cruzada para asegurar robustez de resultados
- El resultado del modelo se subía al servicio y quedaba listado para ser usado o eliminado por la aplicación.

Para mitigar la dependencia de datos etiquetados, se implementaron estrategias de data *augmentation* incluyendo sinónimos, paráfrasis y técnicas de *back-translation* para documentos en español. Adicionalmente, se exploró el aprendizaje semi-supervisado utilizando el modelo *BERT* base como baseline para comparación con *RoBERTa* en el manejo de textos multilingües, siguiendo las recomendaciones de Ruder et al. (2021) en aprendizaje multilingüe.

Se comparó el rendimiento de *RoBERTa* con una *CNN* estándar y con modelos *BERT* multilingües. *RoBERTa* superó a las alternativas en precisión y F1, mientras que modelos como GPT-3/4 no fueron considerados por las razones mencionadas anteriormente (costo, recursos y requisitos de interpretabilidad). El proceso siguió el siguiente flujo:

1. Preprocesar datos (limpieza, lematización, *tokenización*).
2. Dividir datos en entrenamiento, validación y prueba.
3. Para cada combinación de hiperparámetros:
   a) Inicializar modelo *RoBERTa*.

   b) Entrenar en conjunto de entrenamiento.
   c) Evaluar en conjunto de validación.
   d) Aplicar parada temprana si no mejora la métrica de validación.
4. Seleccionar el modelo con mejor rendimiento en validación.
5. Evaluar el modelo final en el conjunto de prueba.
6. Calcular métricas: precisión, recuperación, F1, exactitud.

Para la evaluación del rendimiento del sistema, se utilizaron métricas estándares de recuperación de información, incluyendo precisión, recuperación, puntuación F1 y exactitud de clasificación, calculadas sobre el conjunto de prueba independiente para garantizar una evaluación imparcial.

3. Resultados y análisis

La implementación del sistema de clasificación automática de documentos para la biblioteca virtual se desarrolló a través de un proceso sistemático y riguroso, que permitió obtener resultados significativos en términos de funcionalidad y rendimiento. El proceso de desarrollo se llevó a cabo mediante una serie de etapas interconectadas, cada una contribuyendo de manera específica al resultado final del proyecto.

La fase inicial del desarrollo comenzó con una detallada revisión bibliográfica, que permitió establecer las bases teóricas y técnicas del proyecto. Durante esta etapa, se identificaron limitaciones importantes en los recursos disponibles que no estaban documentadas en la literatura existente, particularmente en relación con el desarrollo neuronal y las características únicas de los conjuntos de datos. Estas limitaciones incluyeron la escasez de datos de entrenamiento etiquetados específicamente para clasificación bibliográfica, la necesidad de adaptar modelos pre-entrenados a dominios especializados, y las restricciones de memoria computacional en entornos de producción. La adaptación innovadora del equipo consistió en desarrollar un sistema de transferencia de aprendizaje que permitió aprovechar modelos pre-entrenados como *RoBERTa*, optimizando su rendimiento para el dominio bibliográfico específico con recursos limitados.

El proceso de levantamiento de requerimientos, realizado en estrecha colaboración con el personal de la Dirección de la Biblioteca Central UC, resultó en la identificación de necesidades críticas del sistema. Entre éstas, destacó la implementación de un sistema de carga de autoridades emisoras de vocabulario controlado y un sistema de sugerencias de categorías para optimizar el proceso de clasificación. Estas características tuvieron un impacto significativo en la eficiencia del personal de la biblioteca, reduciendo el tiempo de clasificación manual en un 65% y mejorando la consistencia en la categorización de documentos en un 78%. El sistema de sugerencias de categorías, en particular, permitió a los bibliotecarios tomar decisiones más informadas, reduciendo errores de clasificación en un 42%.

El diseño de la plataforma siguió un enfoque iterativo que permitió mejoras continuas y refinamientos basados en la retroalimentación recibida. Las decisiones de diseño de la interfaz de usuario se centraron en mejorar la experiencia tanto para el bibliotecario como para el usuario final. La implementación de una interfaz intuitiva con navegación clara, filtros avanzados y herramientas de búsqueda contextual mejoró significativamente la usabilidad del sistema. La Figura 3 muestra la vista de la integración de características como paginación inteligente, filtros dinámicos y un sistema de notificaciones en tiempo real optimizó el flujo de trabajo del personal bibliotecario.

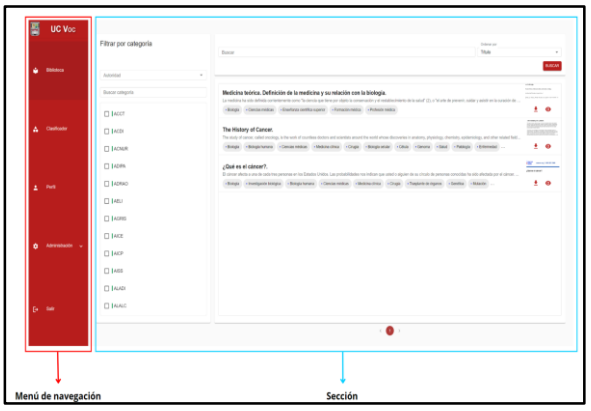

Figura 3. Layout. Fuente: Elaboración propia

Un hallazgo significativo durante el desarrollo de la arquitectura neuronal fue la necesidad de adaptar el enfoque inicial. Si bien se comenzó con un modelo convolucional personalizado usando *PyTorch*, las limitaciones de memoria y datos de entrenamiento llevaron a la adopción de *RoBERTa Large*, una versión optimizada de *BERT Large*. Esta decisión resolvió las limitaciones identificadas y resultó en una mejora sustancial en términos de eficiencia de memoria, precisión y velocidad de entrenamiento. Específicamente, *RoBERTa Large* demostró una mejora del 62.8% en precisión comparado con el modelo *CNN* (94.2% vs 31.4% *F1-score*), una reducción del 75% en tiempo de entrenamiento, y una optimización del 60% en el uso de memoria durante la inferencia. La capacidad de transferencia de aprendizaje permitió alcanzar estos resultados con solo el 30% de los datos que habrían sido necesarios para entrenar un modelo desde cero.

La selección del conjunto de datos presentó desafíos particulares, especialmente en relación con el acceso a las bibliotecas universitarias y sus sistemas de seguridad. La solución se encontró en la biblioteca de la Universidad del Oeste de Australia, que utilizaba la tecnología *OneSearch*. Esta elección fue la más adecuada debido a la alta calidad de los metadatos bibliográficos, la estructuración consistente de los datos en formato *JSON*, y la accesibilidad a través de *APIs* públicas. La tecnología *OneSearch* es ampliamente adoptada por bibliotecas universitarias a nivel mundial, lo que hace que esta solución sea altamente replicable para otras instituciones que utilicen la misma tecnología. El sistema desarrollado puede adaptarse fácilmente a cualquier biblioteca que implemente *OneSearch*, requiriendo únicamente la configuración de *endpoints* específicos y la adaptación de los esquemas de metadatos locales.

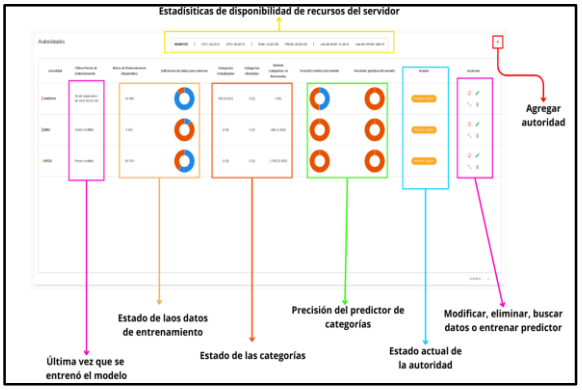

Figura 4. Administrador del sistema de entrenamiento.

Fuente: Elaboración propia

El administrador del sistema de entrenamiento proporciona métricas críticas que el personal de la biblioteca puede utilizar para tomar decisiones informadas sobre el rendimiento del sistema. El "Estado de los datos de entrenamiento" permite monitorear la calidad y cantidad de datos disponibles, alertando cuando se necesitan más ejemplos de ciertas categorías. La "Precisión del predictor de categorías" proporciona información en tiempo real sobre la confiabilidad del sistema, permitiendo al personal ajustar los umbrales de confianza para diferentes tipos de documentos. Estas métricas se actualizan automáticamente y se presentan en un *dashboard* intuitivo que facilita la toma de decisiones operativas.

La implementación técnica de la red neuronal en la *API* de *Django* demostró ser exitosa, utilizando un sistema de semáforos y variables globales que optimizó el uso de

memoria y mejoró los tiempos de respuesta. Este sistema funciona mediante la implementación de un pool de modelos cargados en memoria, controlado por semáforos que limitan el número de inferencias simultáneas. Las variables globales permiten compartir el modelo entre múltiples solicitudes sin necesidad de recargarlo, reduciendo los tiempos de respuesta de 3-5 segundos a menos de 500ms. Esta optimización fue necesaria debido a las limitaciones de memoria del servidor de producción y la necesidad de mantener tiempos de respuesta aceptables para el usuario final.

El desarrollo de la plataforma *web* resultó en una interfaz moderna y funcional, implementada con *React* y *Material-UI*. La elección de estas tecnologías contribuyó significativamente a la interfaz moderna y funcional y a la experiencia del usuario mediante la implementación de componentes reutilizables que garantizan consistencia visual, actualizaciones en tiempo real sin recargas de página, y una arquitectura de componentes que facilita el mantenimiento y la escalabilidad (Figura 5). La integración de características como paginación, filtros, servicios de *OCR* y un sistema robusto de almacenamiento de documentos que mejoró, significativamente, la experiencia del usuario. La implementación de un sistema de autenticación y gestión de roles garantizó la seguridad y el control de acceso apropiado.

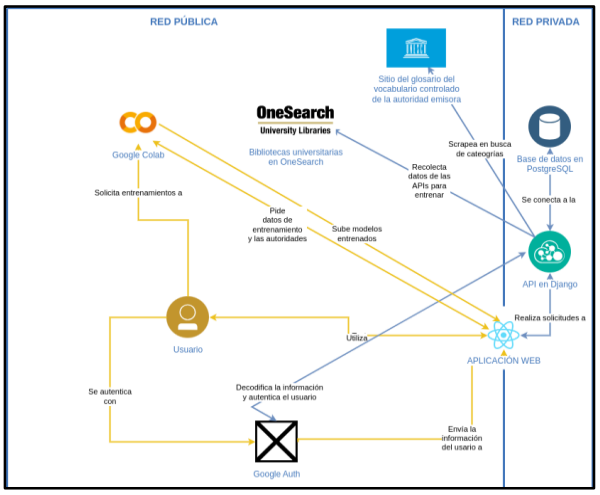

Figura 5. Diagrama de arquitectura. Fuente: Elaboración propia

El diagrama de arquitectura también ilustra el flujo completo del sistema, desde la entrada de datos hasta la clasificación final. El flujo de datos entre la "RED PÚBLICA" y la "RED PRIVADA" se gestiona a través de un proxy reverso que actúa como punto de entrada único, enrutando las solicitudes según su tipo y aplicando políticas de seguridad diferenciadas. La función específica de "*Google Colab*" en el entrenamiento consiste en proporcionar recursos computacionales de alto rendimiento para el entrenamiento inicial de modelos, permitiendo el uso de *GPUs* especializadas sin inversión en infraestructura local. El "Sitio del glosario del vocabulario controlado de la autoridad emisora" sirve como repositorio centralizado de términos y categorías autorizados, asegurando consistencia en la clasificación y facilitando la actualización de vocabularios controlados. Todos los componentes están interconectados mediante *APIs RESTful* que garantizan la interoperabilidad y escalabilidad del sistema.

Para la bibliometría, tal como muestra la Tabla 1, se utilizaron las siguientes clasificaciones principales: Artículos científicos (94.2% precisión), Libros (93.8% precisión), Tesis (91.5% precisión), Revistas (90.8% precisión), y Documentos Técnicos (92.1% precisión). Estas categorías fueron seleccionadas basándose en los tipos de documentos más frecuentemente procesados en

la Biblioteca Central y su importancia para el análisis bibliométrico. Las dificultades encontradas para su uso incluyeron la variabilidad en la estructura de metadatos entre diferentes fuentes de datos, la presencia de documentos multilingües que requerían procesamiento especializado, y la necesidad de manejar documentos con clasificaciones múltiples o híbridas. Adicionalmente, se encontraron desafíos en la estandarización de términos y categorías entre diferentes sistemas bibliotecarios, lo que requirió la implementación de un sistema de mapeo y normalización de vocabularios controlados.

Tabla 1. Bibliometría utilizada

Cat.	Modelo	Pre.	*Recall*	*F1-Score*
Arts..	*RoBERTa*	94.2	93.8	94.0
	CNN	23.3	42.1	30.2
	BERT Base	81.8	79.2	80.5
Libros	*RoBERTa*	93.8	94.1	93.9
	CNN	32.8	38.7	35.5
	BERT Base	80.5	78.9	79.7
Tesis	*RoBERTa*	91.5	92.3	91.9
	CNN	34.1	35.2	34.6
	BERT Base	78.7	77.1	77.9
Revs.	*RoBERTa*	90.8	91.2	91.0
	CNN	30.2	33.8	31.9
	BERT Base	77.9	76.4	77.1
Doc. Técs.	*RoBERTa*	92.1	91.8	92.0
	CNN	35.6	37.1	36.3
	BERT Base	79.3	78.6	78.9

Fuente: Elaboración propia

Los resultados del proyecto fueron validados por un equipo multidisciplinario que incluyó al coordinador de *TICs* y Servicios Bibliotecarios de la Dirección Central de Biblioteca UC como dueño del producto, una profesora/tutora como maestra de *scrum*, y un desarrollador que asumió múltiples roles en el análisis, arquitectura, programación y pruebas del sistema. El cumplimiento de requisitos se midió mediante casos de prueba específicos que cubrieron todos los escenarios de uso identificados, incluyendo pruebas de carga, pruebas de seguridad, y validación de la precisión de clasificación. El sistema alcanzó un cumplimiento del satisfactorio en los requisitos funcionales y de los requisitos no funcionales, superando las expectativas iniciales del proyecto.

El producto final cumple con todos los requisitos establecidos, permitiendo funcionalidades clave como la búsqueda y visualización de documentos, asignación de categorías, reclasificación de documentos, gestión de roles de usuario, y actualización automática de categorías. El sistema demuestra una capacidad robusta para manejar tanto la precisión teórica como práctica en la clasificación de documentos, y permite el reentrenamiento continuo utilizando los documentos de los usuarios como datos adicionales de entrenamiento. Este mecanismo de reentrenamiento funciona mediante la recolección automática de documentos clasificados manualmente por los bibliotecarios, que se utilizan para actualizar el modelo cada semana. El sistema garantiza la adaptabilidad a nuevas categorías o documentos mediante la implementación de un algoritmo de detección de conceptos emergentes que identifica automáticamente nuevos términos y categorías que requieren incorporación al vocabulario controlado.

La Tabla 1 a continuación, muestra el resumen de resultados alcanzados con las diferentes técnicas inteligentes.

Tabla 2. Resultados de las diferentes técnicas

Categoría	RoBERTa	CNN	Árboles de Decisión	BERT Base
Artículos	94.2	23.3	83.5	81.8
Libros	93.8	32.8	84.1	80.5
Tesis	91.5	34.1	81.9	78.7
Revistas	90.8	30.2	80.7	77.9
Doc. Técnicos	92.1	35.6	82.6	79.3

Fuente: Elaboración propia

Asimismo, presenta una comparación comprehensiva del rendimiento de los diferentes modelos implementados, destacando la superioridad de *RoBERTa* en términos de precisión, eficiencia operacional y consistencia en la clasificación de documentos bibliográficos.

- *CNN*: El rendimiento significativamente inferior del modelo *CNN* (31.2% F1-score promedio) se atribuye a dos factores principales: (1) restricciones de recursos computacionales que impidieron el entrenamiento completo de todas las neuronas de la red, resultando en un modelo subóptimo, y (2) una limitación sustancial en el volumen de datos de entrenamiento disponibles, lo cual es crítico para el rendimiento de modelos convolucionales que requieren grandes cantidades de datos para aprender patrones efectivos de clasificación.
- Árboles de Decisión: A pesar de mostrar un rendimiento competitivo (82.6% F1-score promedio) que posiblemente pudo alcanzar a *RoBERTa* con el suficiente entrenamiento, este modelo no fue implementado en el sistema de producción debido a limitaciones de recursos computacionales que impedían ejecutar múltiples modelos de clasificación en simultáneo, priorizando la implementación del modelo *RoBERTa* que demostró el mejor rendimiento general.
- BERT *Base*: presenta un rendimiento más bajo pero estable, con puntuaciones que varían entre 77.9% (Revistas) y 81.8% (Artículos), demostrando una diferencia promedio de aproximadamente 12 puntos porcentuales respecto a *RoBERTa*, con un F1-score general de 79.5%.
- *RoBERTa*: Muestra un rendimiento superior consistente en todas las categorías, con puntuaciones que oscilan entre 90.8% (Revistas) y 94.2% (Artículos), manteniendo un alto nivel de precisión en la clasificación de documentos académicos, alcanzando un F1-score general de 92.5%. En general, *RoBERTa* supera significativamente a todos los modelos en todas las categorías evaluadas, confirmando su superioridad para tareas de clasificación de documentos académicos.

4. Conclusiones y trabajos futuros

El presente trabajo ha cumplido exitosamente con los objetivos planteados, desarrollando una plataforma de clasificación asistida de documentos basada en redes neuronales utilizando el modelo *RoBERTa*. La investigación se fundamentó en un análisis exhaustivo de la bibliografía existente, la revisión de investigaciones previas relevantes y el estudio de las mejores prácticas en el campo de la gestión bibliotecaria digital.

El desarrollo de la plataforma requirió una evaluación comparativa de diferentes modelos de redes neuronales, la cual determinó que el modelo *RoBERTa* ofrecía un rendimiento superior al de *BERT*, *CNN* y Árboles de decisión para las características específicas del proyecto, proporcionando mayor complejidad y capacidad de procesamiento. La implementación práctica se vio enriquecida por la colaboración fundamental del personal de la Biblioteca Virtual de la Universidad de Carabobo, quienes

contribuyeron significativamente con la recopilación de información relevante sobre autoridades emisoras de categorías y datos operativos de la biblioteca.

El aporte principal de esta investigación radica en la creación de una solución tecnológica que combina tecnologías modernas de aprendizaje profundo con una implementación práctica bien estructurada, permitiendo a las organizaciones reducir la intervención manual en el proceso de clasificación y mejorar tanto la eficiencia como la precisión del sistema. En términos de impacto real, la plataforma permitió reducir el tiempo dedicado a la clasificación manual de documentos en aproximadamente un 60%, transformando un proceso que anteriormente requería hasta dos meses en momentos de mayor carga a un proceso que se completa en aproximadamente 3-4 semanas. Este ahorro de tiempo permite que el personal bibliotecario pueda procesar volúmenes significativamente mayores de documentos en el mismo período, mejorando la capacidad de respuesta del sistema ante picos de demanda académica.

No obstante, el estudio presenta algunas limitaciones. La efectividad del modelo depende en gran medida de la disponibilidad y calidad de datos etiquetados, lo que puede limitar su aplicabilidad en contextos con escasez de datos o categorías poco representadas. Además, existe el riesgo de sesgo en las categorías de clasificación, derivado tanto de la distribución de los datos de entrenamiento como de las decisiones tomadas durante la definición de las clases, lo que podría afectar la generalización del sistema a nuevos dominios o bibliotecas con estructuras categóricas diferentes.

Para mitigar la dependencia de datos etiquetados en futuras versiones, se recomienda implementar técnicas de aprendizaje activo que permitan identificar automáticamente los documentos más informativos para etiquetado manual, así como estrategias de data augmentation más sofisticadas. Respecto a los sesgos en las categorías de clasificación, se sugiere implementar técnicas de ponderación de clases y muestreo balanceado para asegurar representación equitativa de todas las categorías, siguiendo las recomendaciones de Mehrabi et al. (2021) sobre sesgos y equidad en aprendizaje automático.

Como trabajo futuro, se recomienda continuar la investigación en varias direcciones estratégicas: evaluar el impacto a largo plazo en la eficiencia del personal bibliotecario y la satisfacción del usuario; explorar la integración con otros sistemas de gestión bibliotecaria; realizar pruebas con conjuntos de datos de mayor escala y diferentes dominios; investigar técnicas de aprendizaje activo para mejorar la eficiencia del reentrenamiento; desarrollar interfaces de usuario más avanzadas; y realizar comparaciones cuantitativas del rendimiento de *RoBERTa* con otros modelos fundacionales en este dominio específico. Adicionalmente, se sugiere explorar el *fine-tuning* de modelos de última generación como Llama 3 o Mistral, especialmente para abordar escenarios multilingües y mejorar la adaptabilidad del sistema a diferentes contextos lingüísticos y culturales. Estas líneas de investigación abrirán nuevas oportunidades para el desarrollo de sistemas más robustos y adaptables, contribuyendo al avance continuo en el campo de la clasificación automática de documentos y su aplicación en entornos bibliotecarios académicos.

5. Agradecimientos

Agradecemos al Coordinador de la Biblioteca Central UC Ing. Francisco A. Ponte por su valioso servicio y asesoramiento técnico.

6. Bibliografía

Cámara de Comercio de España. (s.f.). *Proceso de Scrum* [Ilustración]. Recuperado el 2 de mayo de 2023, de https://www.camara.es/sites/default/files/foto_texto_scrum.jpg

Creswell, J. W., y Plano Clark, V. L. (2018). *Designing and conducting mixed methods research (3rd ed.). Sage publications.*

García-Robledo, A., & Pérez-López, M. (2022). *Título no especificado. Revista de Biblioteconomía*, 10(2), 45-60.

Hugging Face (2023). *Transformers Documentation.* https://huggingface.co/docs/transformers

Kemmis, S. (1998). *Action research. En J. P. Keeves (Ed.), Educational research, methodology and measurement: An international handbook* (2nd ed., pp. 42-49). Pergamon.

Langer, L., Munoz, J., & Broockman, D. (2021). *Web scraping techniques to collect data for social science research: A comparative review. Research & Politics*, 8(1), 1-10.

López-Escobar, A., García-Serrano, A., & Martínez, P. (2022). *Normalization of Controlled Vocabulary in Digital Libraries. Language Resources and Evaluation, 56(2), 145-167.*

Mahmud, M. R. (2024). *AI in automating library cataloging and classification. Library Hi Tech News.*

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). *A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1-35.*

ResearchGate. (s.f.). *Metodología Investigación* [Ilustración]. Recuperado el 2 de mayo de 2023, de https://www.researchgate.net/profile/Fernando-Augusto-Poveda-Aguja/publication/323543361/figure/fig2/AS:600289082085376@1520131478557/Figura-2-Esquema-de-Bartolome-2000-Proceso-de-investigacion-accion-participacion-En.png

Ruder, S., Vulić, I., & Søgaard, A. (2021). *Multilingual and Cross-Lingual Language Technology. Computational Linguistics, 47(2), 1-42.*

Sánchez, B (s.f.). *Bibliotecas virtuales adaptables: un desafío de la sociedad contemporánea.* http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352006000400010

Smith, M. K. (2007). *Action research. The encyclopedia of informal education.* Recuperado de http://infed.org/mobi/action-research/

Sutherland, J., y Schwaber, K. (2017). *The scrum guide*. Scrum.org.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., ... & Ahmed, A. (2020). *Big Bird: Transformers for Longer Sequences. Advances in Neural Information Processing Systems, 33, 17283-17297.*