PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Reconocimiento de texto en manuscritos históricos peruanos utilizando modelos mixtos

Trabajo de investigación para obtener el grado académico de Maestra en Informática que presenta:

Luz Silvana Tarazona Cruz

Asesor:

Pablo Alejandro Fonseca Arroyo

Lima, 2024

Informe de Similitud

Yo, Pablo Alejandro FONSECA ARROYO, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor de el trabajo de investigación titulado *Reconocimiento de texto en manuscritos históricos peruanos utilizando modelos mixtos* de la autora Luz Silvana TARAZONA CRUZ, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 09%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 11/04/2024.
- He revisado con detalle dicho reporte y la tesis de investigación, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha:

San Miguel, 11 de Abril de 2024.

Apellidos y nombres del asesor / de la asesora: Fonseca Arroyo, Pablo Alejandro	
DNI: 44695174	Firma
ORCID: 0000-0002-0208-2842	

RESUMEN

El presente trabajo de investigación aborda la tarea del reconocimiento automático de texto escrito a mano (handwritten text recognition - HTR, por sus siglas en inglés) en los manuscritos históricos de autores peruanos, que están bajo la custodia de la Biblioteca Nacional del Perú (BNP), enfrentando diversas dificultades como la variabilidad caligráfica, el deterioro del papel, entre otras. Para esta tarea, se emplearon modelos de reconocimiento de imágenes preentrenados en otros idiomas disponibles en la plataforma de código abierto denominado OCR4all. Se entrenaron tres modelos utilizando el conjunto de datos SPA-Sentences, conjunto que consta de imágenes y traducciones de aproximadamente 13,000 oraciones en idioma español, logrando una tasa de error de caracteres (character error rate - CER) promedio de 4.11% en el conjunto de validación.

Posteriormente, este modelo elaborado se aplica en los manuscritos históricos peruanos, obteniendo una tasa de error promedio de 9.39%. El CER obtenido, ligeramente menor en comparación con el conjunto de datos SPA-Sentences utilizado en la etapa de entrenamiento, es atribuible a las diferencias en la calidad de las imágenes, así como en las características propias de los manuscritos.

Este trabajo y el enfoque desarrollado en él demuestran la utilidad de los modelos de reconocimiento de imágenes preentrenados para abordar la tarea de HTR en manuscritos históricos, y se identifican áreas para futuras mejoras, como la optimización de la calidad de las imágenes, la diversidad del conjunto de datos y la exploración de modelos avanzados con la arquitectura Transformer.

Palabras clave — OCR4all, reconocimiento de texto escrito a mano, modelos mixtos, manuscritos

ABSTRACT

This paper addresses the task of automatic handwritting text recognition (HTR) in historical manuscripts of Peruvian authors held by the National Library of Peru, facing various challenges such as calligraphy variability, paper deterioration, among others. To achieve this, pre-trained mixed models in other languages available on the open-source OCR platform called OCR4all were employed. Three models were trained using the SPA-Sentences dataset, which consists of a written Spanish collection of approximately 13,000 sentences, achieving an average Character Error Rate (CER) of 4.11% on the validation set.

Subsequently, this developed model is applied to the Peruvian historical manuscripts, obtaining an average error rate of 9.39%. Although this value indicates slightly lower accuracy compared to processing the SPA-Sentences dataset, this is attributed to differences in the image's quality, as well as the specific characteristics of the manuscripts.

This approach demonstrates the usefulness of pre-trained models for addressing HTR in historical manuscripts in historical manuscripts, and areas for future improvements are identified, such as optimizing image quality, dataset diversity, and exploring advanced models with Transformer architecture.

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

SECCIÓN I

INTRODUCCIÓN

En los últimos años, las bibliotecas, los museos y otras instituciones del patrimonio cultural han escaneado un volumen cada vez mayor de sus archivos de documentos históricos. Esto ha intensificado la necesidad de transcribir el texto completo de estos documentos de archivo. En el caso de la Biblioteca Nacional del Perú (BNP), que custodia los documentos históricos peruanos, el proceso de transcripción de estos documentos, se realiza con el apoyo de profesionales, especialistas y público en general, quienes de manera colaborativa y a través de una plataforma digital de la entidad, realizan la transcripción de manuscritos antiguos mediante eventos como TranscriptónBNP, que luego son incluidos en un repositorio digital de libre acceso para el alcance de los ciudadanos[1]. Actualmente cuentan con documentos históricos desde los años 1543 al 1943 (siglos XVI, XVII, XVIII, XIX y XX), los cuales son puestos gradualmente a disposición de los ciudadanos en la URL https://memoriamanuscrita.bnp.gob.pe/.

Como precisa C. Reul et al. [2], el reconocimiento óptico de caracteres (OCR) en impresiones históricas es una tarea desafiante, principalmente debido a la complejidad del diseño, la tipografía muy variada, la falta de fuentes antiguas computarizadas, entre otros. Sin embargo, con los grandes avances realizados en el área del OCR histórico, se cuenta con herramientas potentes de código abierto para el reconocimiento de texto de documentos impresos y manuscritos antiguos, tal como es el caso del software de OCR denominado OCR4all. Esta herramienta hace uso del motor OCR Calamari [8] dentro de su flujo de trabajo semiautomático y utiliza redes neuronales profundas con procedimientos de preentrenamiento y votación, y cuenta con una amplia variedad de configuraciones. Cuenta con modelos que están preentrenados para el reconocimiento de texto impreso y manuscrito en varios idiomas, como inglés, alemán, español, francés, italiano, entre otros, que sirven de base para entrenar conjuntos de datos propios que son ajustados (fine-tuning en inglés) para adaptarse al tipo de documentos que se están procesando.

El conjunto de datos públicos SPA-Sentences, disponible para fines de investigación a través del repositorio institucional de la Universidad Politécnica de Valencia[2], está conformado por textos manuscritos en español moderno para el entrenamiento y evaluación de sistemas de reconocimiento de escritura en lengua española. Como indica S. Boquera et al. [1], el corpus consta de frases manuscritas extraídas de 1,617 formularios producidos por el mismo número de escritores y cuenta con un amplio repertorio de estilos de escritura. Los ficheros del corpus comprenden las imágenes escaneadas de los formularios, así como información de su segmentación en líneas y su transcripción manualmente supervisada. Este conjunto de datos puede ser

[1] https://memoriamanuscrita.bnp.gob.pe/
[2] https://aplicat.upv.es/exploraupv/ficha-tecnologia/patente_software/27402?busqueda=spa-sentences

utilizado para realizar pruebas de rendimiento y comparaciones precisas con otros sistemas de reconocimiento de escritura a mano.

Por lo tanto, el presente trabajo propone hacer uso de la plataforma OCR4all y sus modelos mixtos para entrenar las imágenes de texto en español y sus respectivas traducciones contenidas en el conjunto de datos SPA-Sentences; y con ello, generar un modelo personalizado de aprendizaje supervisado que permita realizar el reconocimiento y traducción automático con el mínimo error posible de los textos históricos digitalizados de autores peruanos y en resguardo de la BNP.

El documento está estructurado de la siguiente manera: en primer lugar, la sección II proporciona una descripción general de los trabajos importantes relacionadas con modelos OCR relevantes para el presente trabajo. En la sección III, se describe los métodos utilizados. A continuación, en la sección IV, se realiza varios experimentos y se detallan los resultados antes de que la sección V concluya el documento al resumir las ideas y señalar los futuros trabajos relacionados a los manuscritos peruanos. Se incluye un apéndice que describe en detalle las configuraciones utilizadas en los experimentos.

SECCIÓN II

TRABAJOS RELACIONADOS

El reconocimiento de texto en manuscritos es un proceso que implica convertir imágenes de texto escrito a mano en texto digital legible por computadora. Este proceso generalmente comprende pasos como la obtención de una imagen, seguida de su preprocesamiento para mejorar la calidad, eliminar el ruido y distorsiones, y preparar la imagen para la extracción de características. Posteriormente, se realiza el reconocimiento de caracteres utilizando técnicas de aprendizaje, como son las redes neuronales convolucionales (CNN) y redes neuronales recurrentes (RNN) y su extensión, las LSTM (Long Short-Term Memory), y por último puede utilizarse el postprocesamiento para la corrección de errores.

Inicialmente, las CNN, como destaca LeCun et al [9], emplean el aprendizaje basado en gradientes con el algoritmo de retropropagación (backpropagation) para el reconocimiento de caracteres manuscritos. Además, introduce el concepto de redes de transformadores de grafos (GTNs), siendo en su momento un nuevo enfoque para entrenar sistemas de reconocimiento de documentos complejos y multimodulares.

Por otro lado, las RNN, que incluyen una entrada, una capa oculta recurrente y una salida, mantiene la memoria a lo largo del tiempo y pueden procesar secuencias de longitud variable. Este concepto se extiende al LSTM, que puede almacenar conocimiento de estados previos y es eficiente para la predicción de datos secuenciales de series temporales [10].

Adicionalmente, la extensión de una RNN regular a una red neuronal recurrente bidireccional (BRNN), como señala Schuster [6], permite procesar información en ambas direcciones temporales, lo que significa que consideran tanto la información pasada como la futura al predecir un elemento en la secuencia, muestra mejoras en el rendimiento de tareas de regresión y clasificación, tanto con datos artificiales como reales.

Los modelos mixtos son aquellos que están entrenados con una variedad de libros, manuscritos, tipos de impresión y estilos de escritura. Se utilizan para mejorar la precisión y el rendimiento del reconocimiento de texto al aprovechar las características de diversos idiomas y estilos de escritura. La idea es que cada modelo tiene sus propias fortalezas y debilidades, y al combinarlos, se pueden obtener mejores resultados que con un solo modelo.

C. Reul et al. [3] informó sobre experimentos usando modelos mixtos en un corpus de 35 manuscritos alemanes medievales con cerca de 12,500 líneas de texto para dos estilos de escritura a mano ampliamente utilizados como son gótica y bastarda (gothic y bastard respectivamente en inglés). Realizó la evaluación inicial de los modelos mixtos en cuatro manuscritos lo que resultó en una tasa de error de

caracteres (CER) promedio del 6,22%, y al entrenarlos con 2, 4 y 32 páginas, respectivamente, la CER se redujo a 3,27%, 2,58% y 1,65%. Concluyen que, si bien es muy importante el reconocimiento en el dominio y el entrenamiento de modelos orientado a un cierto tipo de material para producir los mejores resultados posibles, el trabajo con los modelos mixtos perfectamente adecuados y el uso de modelos previamente entrenados conduce a tasas de error significativamente más bajas que el empezar desde cero.

García [4] realizó la experimentación en imágenes de ediciones de Arnao Guillén de Brocar del proyecto BECLaR del siglo XV transcritas con ayuda de Transkribus y OCR4all. El estudio se centró en el uso de herramientas de reconocimiento óptico de caracteres (OCR) para transcribir textos impresos históricos en español, y luego crea un modelo de red neuronal con los resultados para mejorar aún más la precisión de la transcripción.

Es así como el presente trabajo se enfoca en el uso de modelos mixtos dentro de un software de reconocimiento óptico de caracteres (OCR), a disposición de una gama de usuarios tales como los investigadores, académicos, estudiantes y cualquier persona interesada en acceder y explorar los manuscritos históricos disponibles en la BNP. C. Reul et al. destaca a OCR4all [2] como una herramienta que permite a los usuarios procesar documentos históricos con un nivel de esfuerzo manejable. El software de código abierto integra diferentes herramientas en una interfaz de usuario unificada, lo que permite realizar el flujo de trabajo integral y semiautomático. Comienza con el preprocesamiento de las imágenes de los documentos (Preprocessing), seguido de la segmentación del diseño (Region Segmentation, hecho con la herramienta semiautomática de código abierto LAREX[3]), la extracción de regiones de diseño clasificadas y la segmentación de línea (Line Segmentation), el reconocimiento de texto para reconocer los caracteres individuales (Recognition) y finaliza con la corrección del producto final textual para mejorar su legibilidad (Ground Truth Production), todo ello desarrollando modelos de OCR adaptados a textos específicos.

Los autores C. Reul et al. evaluaron el rendimiento de OCR4all en un conjunto de datos de documentos históricos en varios idiomas, incluyendo el latín, alemán, francés y neerlandés, con énfasis en el alemán. Este conjunto de datos abarcaba libros impresos en el siglo XV, periódicos impresos en el siglo XIX y cartas manuscritas en el siglo XVIII. Los resultados muestran que OCR4all es capaz de procesar estos documentos con una tasa de error de caracteres (CER) inferior al 1%, incluso en documentos con diseños complejos y textos degradados. OCR4all está disponible gratuitamente para el público en general en la plataforma GitHub en la URL https://github.com/OCR4all.

En cuanto a la segmentación, se usa LAREX (Layout Analysis and Region Extraction) [5] que es una herramienta semiautomática de código abierto que permite estructurar

[3] https://github.com/chreul/LAREX

y segmentar adecuadamente las regiones y líneas de texto de las imágenes preprocesadas. Se integra como un submódulo en la plataforma OCR4all y permite la edición integral de las imágenes y sus traducciones manuales generando archivos en formato PageXML.

Asimismo, OCR4all cuenta con un motor de reconocimiento de caracteres individuales en las imágenes de documentos denominado Calamari [8]. Este se basa en redes neuronales recurrentes (RNN) implementadas en la plataforma de código abierto Tensorflow. Calamari admite técnicas modernas como el preentrenamiento y la votación para mejorar la precisión y el rendimiento de los modelos. El software es compatible con arquitecturas de red personalizadas construidas con capas de redes neuronales convolucionales (CNN) y de memoria a largo/corto plazo (LSTM). Los modelos se entrenan utilizando el algoritmo de clasificación temporal conexionista (CTC) de Graves et al. (2006).

Finalmente, en los últimos años Vaswani et al., [7] publicó el modelo Transformer, que revolucionó el campo del procesamiento del lenguaje natural y la visión por computadora ya que aborda el problema de la traducción automática, y que elimina las redes neuronales recurrentes (RNN) y convolucionales (CNN) en favor de una arquitectura basada en atención. En su arquitectura detalla que la estructura del Transformer, consta de codificadores y decodificadores. Los codificadores capturan información en la entrada, mientras que los decodificadores generan la salida traducida. Si bien el presente trabajo no incluye el modelo, sería interesante ampliarlo e integrarlo con el OCR4all en trabajos futuros.

SECCIÓN III

MÉTODOS

A. Arquitectura

La herramienta OCR4all cuenta con los siguientes modelos utilizados para el reconocimiento:

- Predeterminada (default): La estructura de red predeterminada de Calamari original que se compone de dos redes neuronales convolucionales (CNN). La primera CNN consta de 40 capas de características y la segunda de 60, con un tamaño de agrupación de 3x3 respectivamente, cada una seguida de una capa de agrupación máxima de 2x2 y una capa de red LSTM de 200 nodos ocultos, utilizando una tasa de abandono (dropout) de 0.5. (ver Figura 1).
- deep3: Esta configuración presenta una estructura de red más profunda y amplía la red predeterminada (default), agregando otra capa convolucional adicional y dos LSTM más a la arquitectura básica.
- htr+ (Handwritten Text Recognition+): Esta configuración representa una adaptación de la estructura de red estándar de la plataforma Transkribus, con variaciones más complejas en cuanto a tamaños de filtro u otras características. Se basa en una combinación de técnicas de aprendizaje profundo, como las redes neuronales convolucionales (CNN) y las redes neuronales recurrentes (RNN), específicamente las redes LSTM (Long Short-Term Memory).

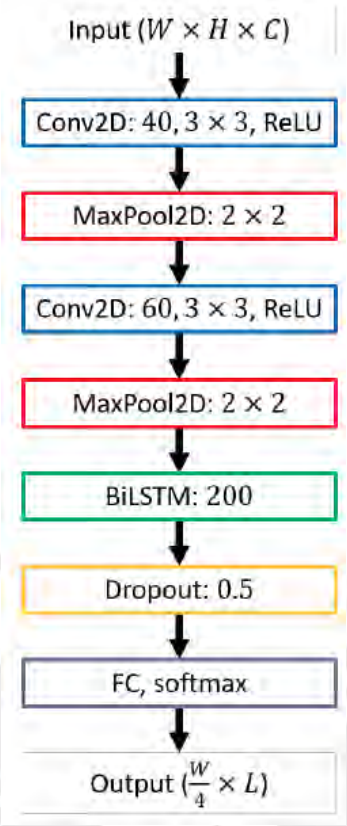


Figura 1: Arquitectura de red predeterminada de Calamari

"La imagen de entrada (ancho W, altura H y canales de color C) pasa a la primera capa convolucional. Al aplicar 40 operaciones de filtro con un tamaño de kernel de 3 x 3 se obtienen 40 mapas de características que luego se reducen mediante una operación de agrupación máxima con un tamaño de agrupación de 2 x 2. Los dos pasos se repiten, utilizándose en esta oportunidad 60 mapas de características durante la convolución. A continuación, los mapas de características se concatenan verticalmente y se pasan a una LSTM bidireccional con 200 nodos ocultos. Se introduce una capa de eliminación con una tasa de eliminación de 0.5 para reducir el efecto del sobreajuste. Finalmente, para cada posición horizontal, una capa totalmente conectada con softmax conduce a la matriz de probabilidad de salida final con W/4 columnas (el ancho original de la imagen de entrada se reduce en un factor 4 debido a las dos operaciones de agrupación) x L filas, con L representando el número de etiquetas, que es el tamaño del alfabeto más la etiqueta en blanco."[2]

B. Conjunto de datos

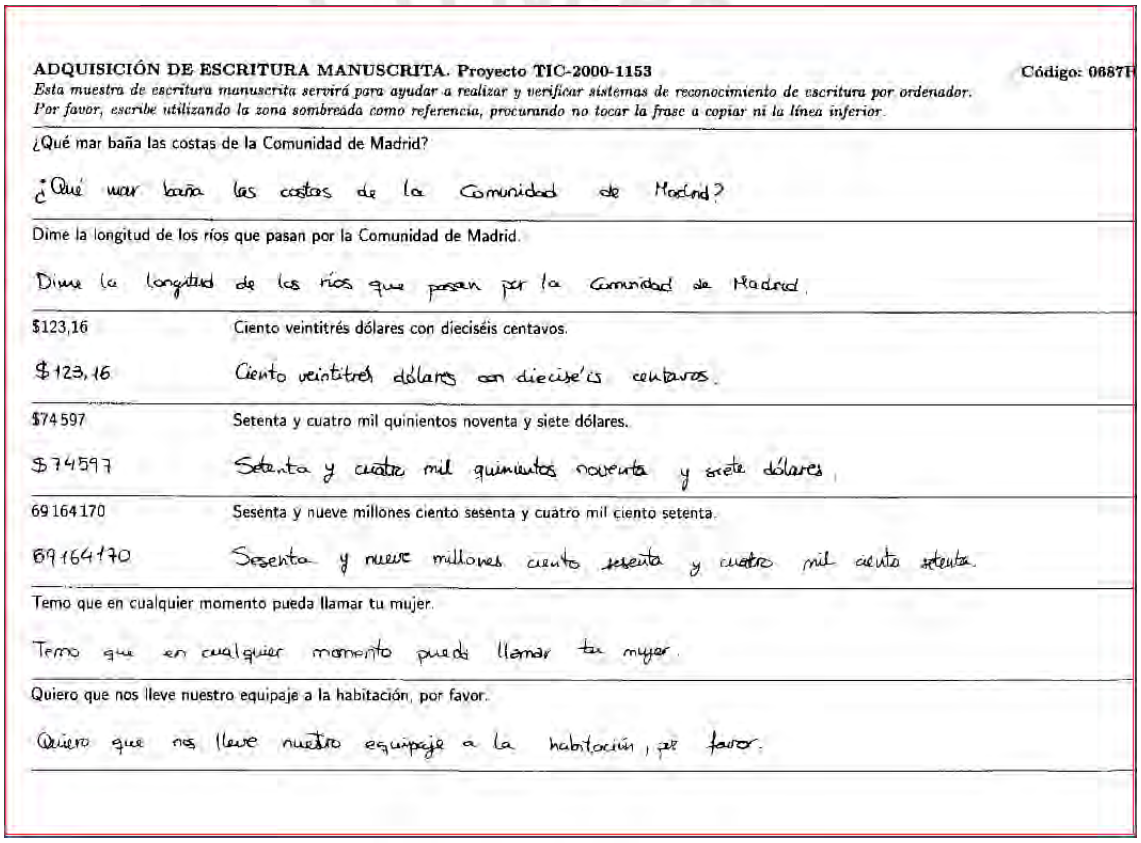
Se ha considerado dos conjuntos de datos, SPA-Sentences para contar con una base de datos en español donde se entrenan diversos modelos preentrenados en otros idiomas, dentro de la herramienta OCR4all, y los manuscritos destacados del BNP con los que se entrena para reforzar el contexto en el procesamiento de los manuscritos con OCR4all.

- SPA-Sentences:
  El conjunto de datos públicos SPA-Sentences, accesible para fines de investigación, está conformado por textos manuscritos en español moderno dividido en cinco particiones (numeradas de P00 a P04). Estos textos provienen de 1,617 formularios generados por igual número de escritores, totalizando 13,691 oraciones y alrededor de 100,000 instancias de palabras, con un vocabulario de 3,288 palabras dentro de

la colección [1]. Cuenta con dos tipos de formas: formas verticales y horizontales con un promedio de 70 palabras manuscritas por formulario en ambos casos.

Este conjunto de datos incluye las imágenes de los manuscritos en formato PNG, junto con las coordenadas de segmentación de línea y sus transcripciones en formato XML. Se realizó una adaptación de los metadatos de los archivos originales para agregar la segmentación de las regiones de cada hoja y sus coordenadas de puntos de referencia (ver Figura 2 y 3).

Este conjunto de datos se seleccionó por su variedad de estilos de manuscritos y se entrenó con modelos en otros idiomas disponibles en la herramienta OCR4all.



ADQUISICIÓN DE ESCRITURA MANUSCRITA. Proyecto TIC-2000-1153 Código: 0687F

Esta muestra de escritura manuscrita servirá para ayudar a realizar y verificar sistemas de reconocimiento de escritura por ordenador.
Por favor, escribe utilizando la zona sombreada como referencia, procurando no tocar la frase a copiar ni la línea inferior.

¿Qué mar baña las costas de la Comunidad de Madrid?

¿Qué mar baña las costas de la Comunidad de Madrid?

Dime la longitud de los ríos que pasan por la Comunidad de Madrid.

Dime la longitud de los ríos que pasan por la Comunidad de Madrid.

$123,16	Ciento veintitrés dólares con dieciséis centavos.
$123,16	Ciento veintitrés dólares con dieciséis centavos.
$74 597	Setenta y cuatro mil quinientos noventa y siete dólares.
$74597	Setenta y cuatro mil quinientos noventa y siete dólares.
69 164 170	Sesenta y nueve millones ciento sesenta y cuatro mil ciento setenta.
69164170	Sesenta y nueve millones ciento sesenta y cuatro mil ciento setenta.

Temo que en cualquier momento pueda llamar tu mujer.

Temo que en cualquier momento pueda llamar tu mujer.

Quiero que nos lleve nuestro equipaje a la habitación, por favor.

Quiero que nos lleve nuestro equipaje a la habitación, por favor.

Figura 2: Ejemplo del formulario contenido en el conjunto de datos SPA-Sentences

```
<PcGts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15
 http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15/pagecontent.xsd">
  <Metadata>
    <Creator>kraken</Creator>
    <Created>2023-09-07T01:55:47</Created>
    <LastChange>2023-09-08T23:53:23</LastChange>
  </Metadata>
  <Page imageFilename="0013.bin.png" imageHeight="2352" imageWidth="3201">
    <TextRegion id="r0" type="paragraph">
      <Coords points="1,1 3198,1 3198,2348 1,2348"/>
      <TextLine id="l0">
        <Coords points="112,478 2084,478 2084,554 112,554"/>
        <Baseline points="112,478 2084,478"/>
        <TextEquiv index="0">
          <Unicode>Dime el nombre de los ríos que desembocan en el océano Atlántico .</Unicode>
        </TextEquiv>
      </TextLine>
      <TextLine id="l1">
        <Coords points="115,736 2281,736 2281,807 115,807"/>
        <Baseline points="115,736 2281,736"/>
        <TextEquiv index="0">
          <Unicode>Dime el nombre de los ríos que pasan por la Comunidad de Madrid .</Unicode>
        </TextEquiv>
      </TextLine>
      <TextLine id="l2">
        <Coords points="96,993 2206,993 2206,1068 96,1068"/>
        <Baseline points="96,993 2206,993"/>
        <TextEquiv index="0">
          <Unicode>$794.50 Setecientos noventa y cuatro dólares con cincuenta centavos .</Unicode>
        </TextEquiv>
      </TextLine>
      <TextLine id="l3">
        <Coords points="107,1252 1735,1252 1735,1317 107,1317"/>
        <Baseline points="107,1252 1735,1252"/>
        <TextEquiv index="0">
          <Unicode>16440€ Dieciséis mil cuatrocientos cuarenta euros .</Unicode>
        </TextEquiv>
      </TextLine>
      <TextLine id="l4">
        <Coords points="99,1517 2597,1517 2597,1589 99,1589"/>
        <Baseline points="99,1517 2597,1517"/>
        <TextEquiv index="0">
          <Unicode>23987800 Veintitrés millones novecientos ochenta y siete mil ochocientos .</Unicode>
        </TextEquiv>
      </TextLine>
```

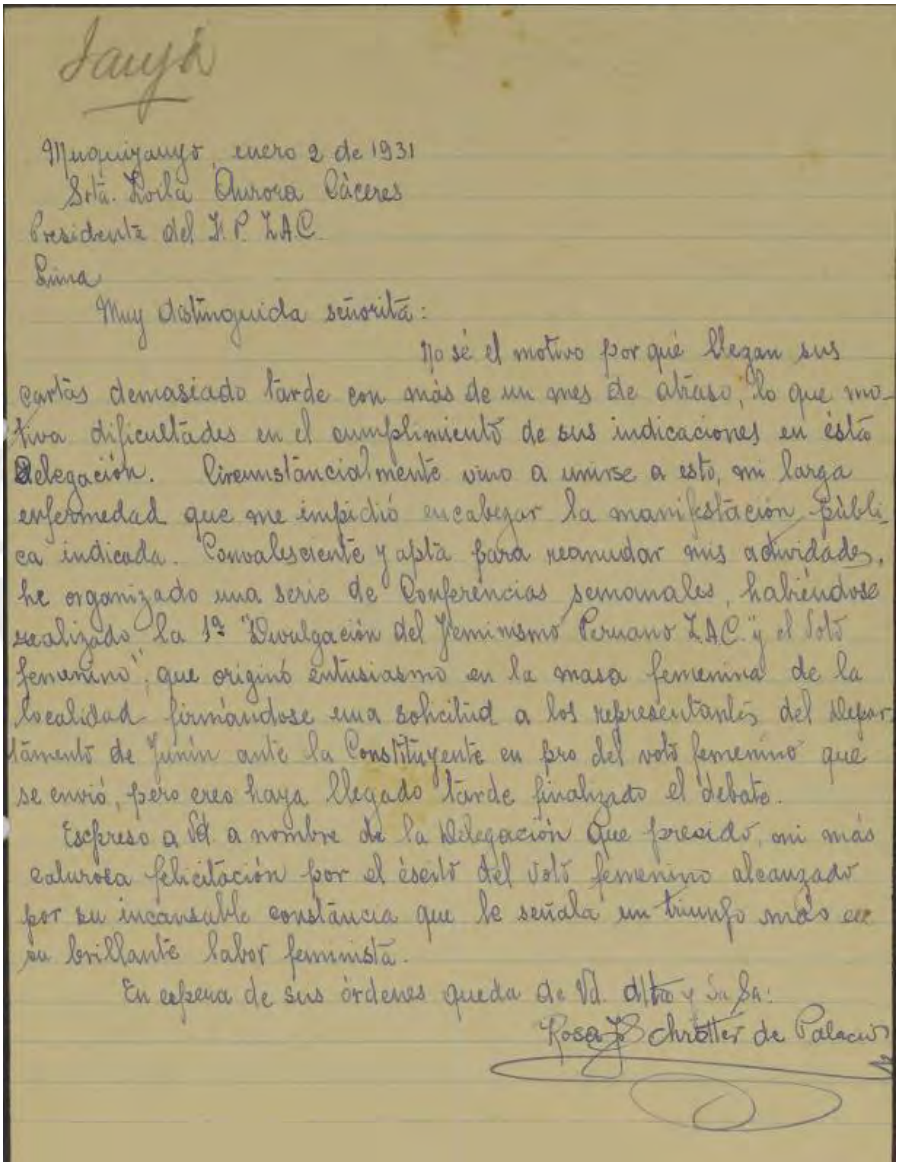

Figura 3: Archivo XML adaptado correspondiente al formulario escaneado

- Manuscritos:

  Para llevar a cabo los experimentos, en primer lugar, se recolectaron las imágenes de los manuscritos que se analizarían. Para ello, se revisó la información disponible en la página web de la BNP, donde se identificaron un total de 1284 manuscritos que forman parte del patrimonio bibliográfico documental. De este conjunto, 96 manuscritos pertenecen al siglo XVI, 51 documentos son del siglo XVII, 60 del siglo XVIII, 901 del siglo XIX y 176 del siglo XX. Los manuscritos abarcan una variedad de autores peruanos, y para el presente documento se consideraron aquellos que datan del año 1543 hasta 1941.

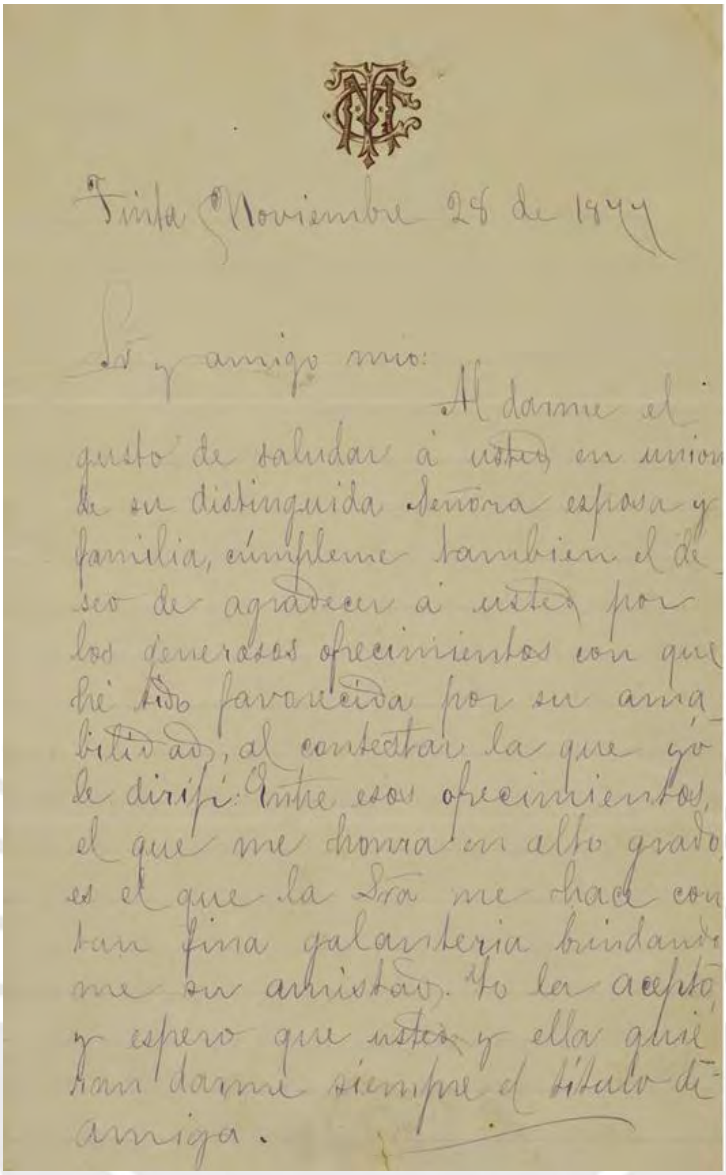
  Considerando que los manuscritos presentan las dificultades en cuanto a la calidad de la imagen como la borrosidad, puntos por pulgada diverso u otros, lo que dificulta la extracción de caracteres de forma precisa, así como la variabilidad en la escritura por parte de sus autores, ruidos, manchas o deterioro del papel a lo largo del tiempo, los que pueden interferir con el proceso de reconocimiento de caracteres. Además, se observa un espaciado irregular entre palabras y el uso de ligaduras, abreviaturas, dígrafos y la presencia de caracteres no usados en las tipografías actuales como s longa, r rotunda, entre otros.

Si bien cada autor maneja de forma general una caligrafía y estilo únicos, se estableció como criterio de selección que la escritura sea legible, lo que permite que las letras del texto pueden ser de cualquier tipo y tamaño. En consecuencia, los manuscritos elegidos corresponden a testimonios, cartas, interrogatorios y/o correspondencias que contienen información de eventos y diversas situaciones de la época. En total, se seleccionaron 14 manuscritos que abarcan 74 páginas y contienen aproximadamente 1400 líneas de texto, los cuales fueron procesadas utilizando la herramienta OCR4all.



Huancayo, enero 2 de 1931
Srta. Dña Aurora Cáceres
Presidenta del C. P. F. A. C.
Lima
Muy distinguida señorita:
No sé el motivo por qué llegan sus cartas demasiado tarde con más de un mes de atraso, lo que motiva dificultades en el cumplimiento de sus indicaciones en esta Delegación. Circunstancialmente vino a unirse a esto, mi larga enfermedad que me impidió encabezar la manifestación pública indicada. Convaleciente y apta para reanudar mis actividades, he organizado una serie de Conferencias semanales, habiéndose realizado la 1ª "Divulgación del Feminismo Peruano F.A.C." y el Voto femenino", que originó entusiasmo en la masa femenina de la localidad firmándose una solicitud a los representantes del Departamento de Junín ante la Constituyente en pro del voto femenino que se envió, pero creo haya llegado tarde finalizado el debate.
Expreso a Vd. a nombre de la Delegación que presido, mi más calurosa felicitación por el éxito del voto femenino alcanzado por su incansable constancia que le señala un triunfo más en su brillante labor feminista.
En espera de sus órdenes queda de Vd. atta y Sa Sa:
Rosa Schutter de Palacios

Figura 4: Manuscrito digitalizado del año 1931 que resguarda la Biblioteca Nacional del Perú



Tinta 8 Noviembre 26 de 1944

Sr y amigo mio:

Al darme el gusto de saludar á usted en union de su distinguida Señora esposa y familia, cúmpleme tambien el deseo de agradecer á usted por los generosos ofrecimientos con que fuí favorecida por su amabilísimo, al contestar la que yo le dirijí. Entre esos ofrecimientos, el que me honra en alto grado, es el que la Sra me hace con tan fina galantería brindándome su amistad. Yo la acepto, y espero que usted y ella quieran darme siempre el título de amiga.

Figura 5: Manuscritos digitalizado del año 1876 que resguarda la Biblioteca Nacional del Perú

C. Flujo de trabajo en la herramienta OCR4all

A continuación, se presenta una síntesis del flujo de trabajo general, siguiendo los pasos principales mostrados en la Figura 6, que facilita la experimentación con los modelos específicos de la herramienta y, posteriormente el entrenamiento de modelos ajustados para obtener mejores resultados en el reconocimiento de los manuscritos. Las configuraciones implementadas se detallan en el apéndice.

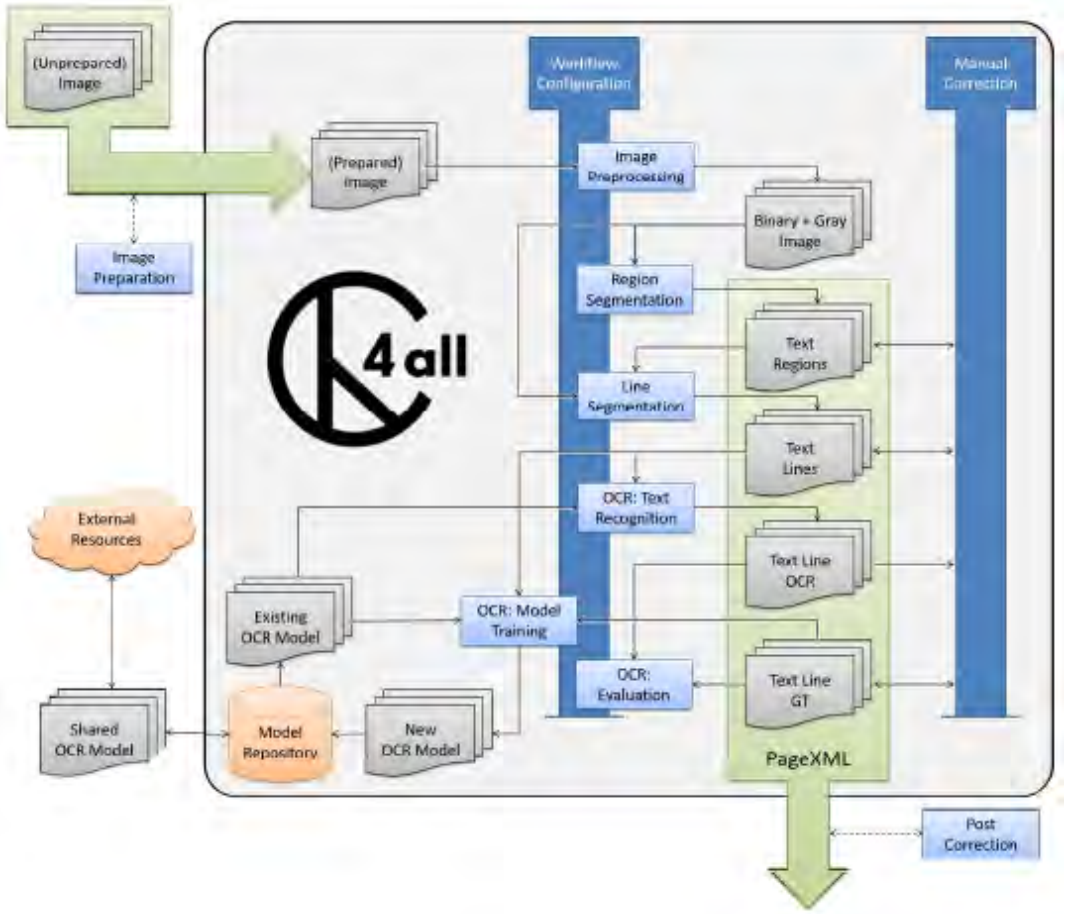


Figura 6: Diagrama del flujo de trabajo en la plataforma OCR4all

- *Preprocesamiento de imágenes:* Las imágenes de los manuscritos se procesan para mejorar la calidad y la legibilidad. En esta experimentación sólo se realizó el proceso de binarización para convertir la imagen en blanco y negro. No obstante, la herramienta ofrece la posibilidad de realizar la normalización del tamaño de la imagen, la corrección de la orientación y/o la eliminación de ruido.

- *Segmentación:* Las CNN se utilizan para extraer características importantes de las imágenes de los manuscritos. Estas características pueden incluir bordes, texturas y patrones relevantes que ayudan a distinguir entre diferentes letras y palabras.

- *Entrenamiento:*

  Seleccionando el mejor modelo inicial:
  - Se realizó una predicción utilizando diversos modelos preentrenados en otros idiomas. Luego, se seleccionaron los tres con la menor cantidad de errores (Tabla 1) al ser evaluados con una imagen del alfabeto manuscrito compuesta por 14 caracteres (Figura 7).

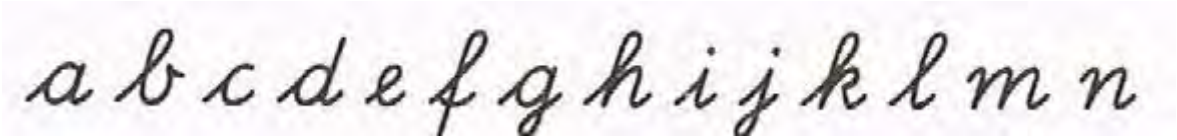


Figura 7: Alfabeto evaluado para la selección de los modelos iniciales.

Modelo	Resultado	Tasa de error
deep3_fraktur-hist/2	aFiePo hipkilmn	57.14% (8 errorese 14 caracteres en total)
deep3_fraktur19/2	abedelakikemn.	50.00% (7 errores, 14 caracteres en total)
deep3_lsh4/2	abcdelqHijAlmn'	28.57% (4 errores, 14 caracteres en total)

Tabla 1 Tasa de error de los modelos seleccionados

- A continuación, se detalla un comparativo de los modelos de reconocimiento de texto según su idioma y estilo de escritura (Tabla 2).

Modelos	Idioma	Estilo de escritura
deep3_fraktur19/2	alemán	Fraktur moderno (siglo XX)
deep3_lsh4/2	latín	Escritura histórica latín
deep3_fraktur-hist/2	alemán	Escritura histórica en fraktur (siglos XVIII-XIX)

Tabla 2 Idioma y estilo de escritura de los modelos seleccionados

Entrenamiento con las imágenes de los manuscritos del conjunto de datos SPA-Sentences:

- Tras determinar los mejores modelos con una menor tasa de error, se procedió a evaluar el conjunto de datos SPA-Sentences. Inicialmente se evaluaron 324 páginas de las 1617 disponibles, lo que representa el 25% del total, debido a demoras en el tiempo de procesamiento. De la evaluación realizada, se determinó que el modelo deep3_fraktur-hist/2 mostró una tasa de Error de Reconocimiento de Caracteres (CER) del 4.62%, la más baja en comparación con los otros modelos evaluados con el mismo número de páginas.

- Una vez identificado el modelo más preciso, con una tasa de error menor, se efectuó una segunda evaluación sobre el conjunto de datos completo. En este escenario, el modelo logró un CER de 4.11%, lo que sugiere una mejora en la precisión a medida que aumenta la cantidad de imágenes procesadas. El modelo

ajustado se denominó como “modelo_SPA”, con el cual se llevó a cabo el reconocimiento de los manuscritos pertenecientes a la BNP.

Modelos preentrenados	Número de páginas	Tasa de error de caracteres (CER)	Nuevo modelo
deep3_fraktur19/2	324	4.74%	
deep3_lsh4/2	324	4.77%	
deep3_fraktur-hist/2	324	4.62%	
deep3_fraktur-hist/2	1617	4.11%	modelo_SPA

Tabla 3 Rendimiento de los modelos evaluados en el conjunto de datos SPA-Sentences

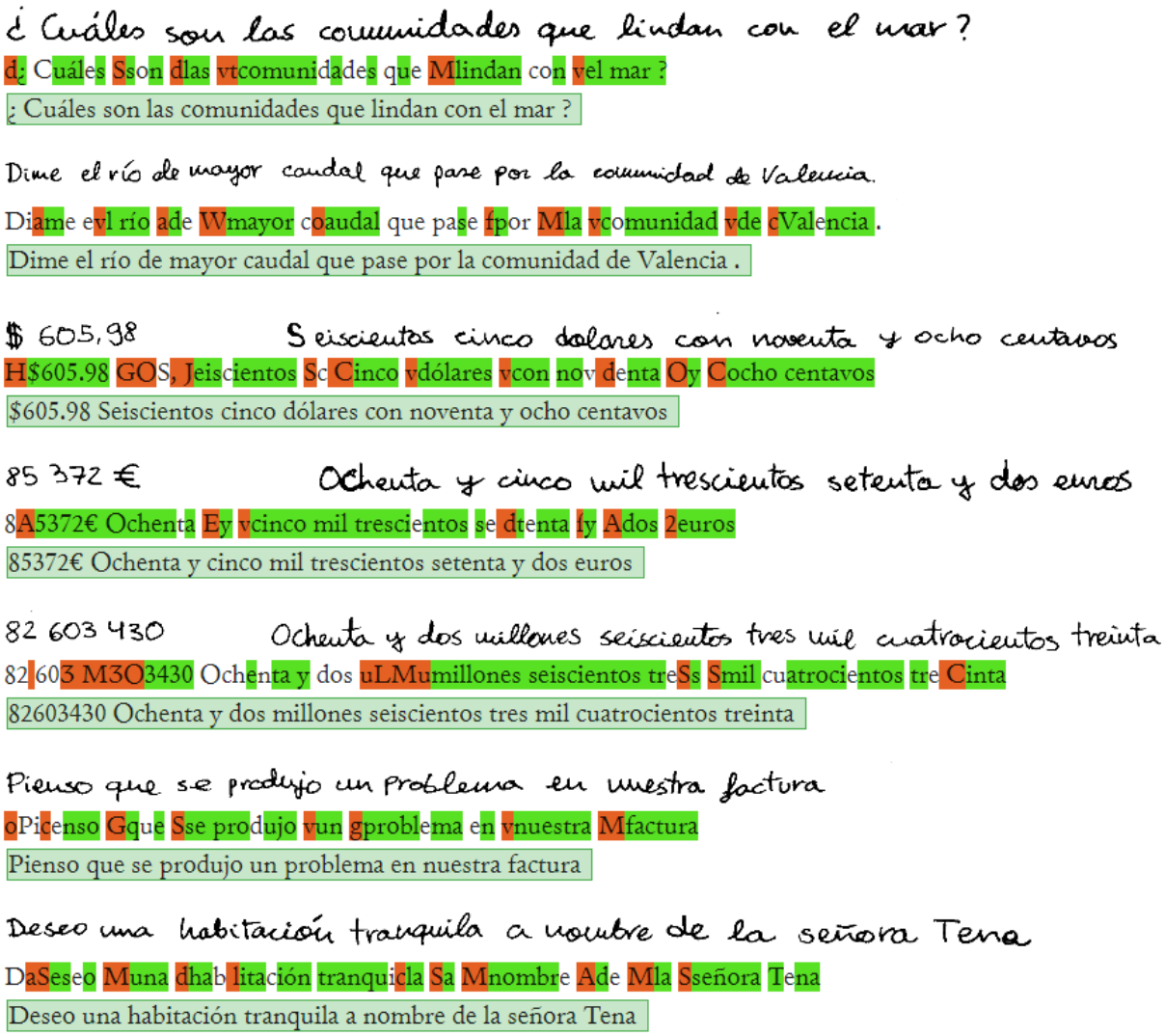


Figura 8: Resultado del reconocimiento de texto del conjunto de datos SPA-Sentences

Entrenamiento a los manuscritos de la BNP:

- Se realizó el entrenamiento, donde en el primer escenario, el modelo “modelo_SPA” fue evaluado en un conjunto de 4 manuscritos, abarcando un total de 22 páginas, y logró una tasa

de CER del 15.63%. En el segundo escenario, se evaluó en un conjunto de 14 manuscritos, con un total de 74 páginas, obteniendo una tasa de CER del 9.39%. Estos manuscritos pertenecen a una autora peruana de finales del siglo XIX.

Modelo preentrenado	Cantidad manuscritos	Número de páginas	Tasa de error de caracteres (CER)
modelo_SPA	4	22	15.63%
	14	74	9.39%

Tabla 4 Rendimiento del modelo preentrenado "modelo_SPA" en los manuscritos de la BNP

Figura 9: Resultado del reconocimiento de texto de manuscritos de autores peruanos

- *Postprocesamiento:* Es un paso adicional para mejorar la precisión del reconocimiento. Esto puede incluir técnicas como el uso de diccionarios para corregir errores ortográficos o la incorporación de información contextual para mejorar la coherencia del texto reconocido; sin embargo, para el presente trabajo no fue necesario implementar este paso.

SECCIÓN IV

RESULTADOS

El objetivo principal del presente trabajo fue generar un nuevo modelo de aprendizaje supervisado para el reconocimiento de textos en manuscritos históricos peruanos, utilizando modelos mixtos.

Para ello, se realizaron diversas experimentaciones utilizando tanto del conjunto de datos SPA-Sentences como los propios manuscritos. Se aprovecharon los modelos preentrenados disponibles en otros idiomas que están integrados en la herramienta OCR4all, además se generaron nuevos modelos para los próximos entrenamientos.

La evaluación de desempeño de los modelos se basó en el CER (Character Error Rate), una métrica estándar para evaluar la precisión del reconocimiento óptico de caracteres (OCR) en la transcripción de manuscritos u otros tipos de texto. Cuanto menor sea el valor del CER, mayor será la precisión del sistema OCR.

Se evaluaron tres modelos preentrenados: deep3_fraktur19/2, deep3_lsh4/2 y deep3_fraktur-hist/2. El modelo deep3_fraktur-hist/2, que emplea una escritura histórica en fraktur (siglos XVIII-XIX), obtuvo el menor CER (4.62%) en una muestra del 25% del conjunto de datos SPA-Sentences. Este modelo mixto se entrenó con la totalidad el conjunto de datos SPA-Sentences, lo que resultó en un nuevo modelo denominado “modelo_SPA” con un CER de 4.11%, evidenciando una mejora con una mayor cantidad de imágenes.

Posteriormente, se empleó el modelo generado “modelo_SPA”, para procesar las imágenes de los manuscritos de la BNP, obteniendo una tasa de error de 9.39%. Este valor refleja una precisión ligeramente menor en comparación con el procesamiento del conjunto de datos de SPA-Sentences, atribuible a diferencias en la calidad y características de los manuscritos en comparación con los datos de entrenamiento.

Estos resultados indican que el nuevo modelo generado “modelo_SPA” basado en el modelo preentrenado deep3_fraktur-hist/2, es prometedor para el reconocimiento de textos en manuscritos históricos peruanos, y que el rendimiento puede mejorar con un entrenamiento con un número mayor de imágenes de manuscritos agrupados por autores y épocas.

CONCLUSIONES

- La disponibilidad de los modelos mixtos preentrenados para el reconocimiento automático de caracteres manuscritos reduce significativamente el esfuerzo requerido para realizar transcripciones manuales desde cero, ya que han sido entrenados previamente en una amplia variedad de libros y composiciones tipográficas, lo que les permite reconocer una amplia gama de estilos de escritura y formatos de texto.

- El uso de modelos entrenados en el mismo idioma que los documentos que van a ser transcritos mejora la precisión del reconocimiento de texto manuscrito al identificar letras y palabras en cada segmento de línea.

- La calidad de las imágenes es fundamental para realizar el reconocimiento del texto. Estas deben ser legibles y evitar la borrosidad o la falta de nitidez, ya que esto afectará directamente a la efectividad del OCR.

- Este proyecto planteó proporcionar un modelo ajustado preentrenado que permita brindar una solución al desafío de reconocer automáticamente la información de los manuscritos por medio de las redes neuronales obteniéndose un CER de 9.39% sobre manuscritos pertenecientes a autores peruanos de finales del siglo XIX, custodiados por la BNP.

TRABAJOS FUTUROS

- *Ampliar el conjunto de datos de español escrito a mano*. Si bien el conjunto de datos SPA-Sentences brinda una línea base y es un buen comienzo para establecer un conjunto de datos estándar para la tarea de reconocimiento de texto escrito a mano fuera de línea en un corpus español, carece de contextualización y no incluye formas escritas con grafías antiguas. Se podría replicar la dinámica efectuada para la generación del conjunto de datos mencionado considerando un contexto determinado definido en conjunto con representantes de la BNP.

- *Contar con imágenes de calidad.* La disponibilidad de manuscritos digitalizados nítidos es crucial para realizar el reconocimiento de los textos. Por lo tanto, se podría coordinar con representantes de la BNP para obtener acceso a manuscritos adicionales de alta calidad que puedan ser incluidos en las pruebas.

- *Optimizar la configuración del servidor.* El uso de GPU podría permitir realizar nuevas experimentaciones, como el uso con transformaciones de degradación de imagen, la exploración de redes neuronales más profundas con nuevas configuraciones o la aplicación de técnicas de eliminación de ruido para mejorar la calidad de las imágenes al eliminar pequeñas imperfecciones, como manchas en los escaneos.

- *Incorporar modelos en base a la arquitectura Transformer*. Se podría implementar modelos de última generación los cuales cuentan con modelos preentrenados específicamente diseñados para la tarea de reconocimiento de manuscritos.

REFERENCIAS BIBLIOGRÁFICAS

[1] España-Boquera, S., & Castro-Bleda, M. J. (2022). A Spanish dataset for reproducible benchmarked offline handwriting recognition. *Language Resources and Evaluation*, 1-14.

[2] Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., ... & Puppe, F. (2019). OCR4all—An open-source tool providing a (semi-) automatic OCR workflow for historical printings. *Applied Sciences, 9(22)*, 4853.

[3] Reul, C., Tomasek, S., Langhanki, F., & Springmann, U. (2022, May). Open Source Handwritten Text Recognition on Medieval Manuscripts Using Mixed Models and Document-Specific Finetuning. In *International Workshop on Document Analysis Systems* (pp. 414-428). Cham: Springer International Publishing.

[4] García, M. A. (2022). Las ediciones de Arnao Guillén de Brocar de BECLaR transcritas con ayuda de Transkribus y OCR4all: creación de un modelo para la red neuronal y posible explotación de los resultados. *Historias Fingidas,* 151-173.

[5] Reul, C., Springmann, U., & Puppe, F. (2017, June). *Larex: A semi-automatic open-source tool for layout analysis and region extraction on early printed books. In Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage* (pp. 137-142).

[6] Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673-2681.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[8] Wick, C., Reul, C., & Puppe, F. (2018). Calamari-a high-performance tensorflow-based deep learning package for optical character recognition. *arXiv preprint arXiv:1807.02004*.

[9] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[10] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306.

APÉNDICE

Configuraciones y flujo de trabajo realizadas en la herramienta OCR4all

1. Conjunto de datos SPA-Sentences:
    - Configuración y preparación de imágenes
        - Se estableció la configuración inicial del conjunto de archivos creando carpetas con proyectos dentro de la carpeta de datos, con los formularios de los corpus divididos en 5 particiones (P00 al P04).
        - Se crearon 2 proyectos, uno con el 25% y otro con el 100% del total de los formularios: SPA_Sentences con 1,617 imágenes y SPA_Manuscritos_25% con 324 imágenes.
        - Se copiaron los archivos en formato PNG en el espacio de trabajo del proyecto.
        - Se aprobaron los ajustes propuestos para generar archivos con un formato estándar, asignando un identificador de página individual y correlativo a cada imagen (por ejemplo: el archivo original h000_0.png se etiquetó con el identificador 0001.png, h000_1.png como 0002.png y así sucesivamente).
    - Preprocesamiento:
        - Se ejecutó esta etapa con la configuración predeterminada, generando imágenes binarias y en escala de grises imágenes necesarias para una exitosa segmentación y reconocimiento OCR posterior.
    - Segmentación:
        - Se adaptó el archivo XML original para considerar la página completa como un único segmento de texto continuo para cada imagen.
    - Entrenamiento:
        - Los experimentos se realizaron principalmente con los parámetros predeterminados en la "configuración general" y la "configuración avanzada".
        - Se realizaron diversos entrenamientos en base a los modelos disponibles en otros idiomas y usando la estructura de red: *deep3* con la siguiente configuración*: cnn=40:3x3, pool=2x2, cnn=60:3x3, pool=2x2, cnn=120:3x3, lstm=200, lstm=200, lstm=200, dropout=0.5*
        - Se configuró la lista blanca de caracteres (whitelist of characters) considerando el alfabeto para idioma castellano, caracteres especiales y tildes

        configuración:

        ABCDEFGHIJKLMNÑOPQRSTUVWXYZabcdefghijklmnñopqrstuvw01 23456789/().:?áéíóúÁÉÍÓÚ€$Üü<>
        - Se estableció el criterio de detención anticipada (early stopping) para finalizar el entrenamiento si el CER de validación no mejoraba en tres ocasiones consecutivas.

- Con respecto a otros (hiper) parámetros se mantuvieron los valores predeterminados de Calamari, siguiendo las mejores prácticas establecidas.
- Cada nuevo modelo generado con su correspondiente CER se almacenó para su uso en el entrenamiento de los manuscritos.

- Los resultados obtenidos con los tres modelos seleccionados se muestran en la tabla 3.

2. Manuscritos:

- Configuración y preparación de imágenes:
  - Se realizó la configuración inicial del proyecto “Manuscritos” incluyendo la creación de las subcarpetas input y processing.
  - Se copiaron los manuscritos originales digitalizados en el formato PDF en el espacio de trabajo designado para el proyecto.
  - Se cargó el proyecto en la plataforma OCR4all y se aceptaron los ajustes recomendados para la conversión automática de los archivos PDF al formato PNG, utilizando el valor predeterminado de 300 dpi. Esta acción resultó en la generación de imágenes individuales para cada página de los documentos originales; de esta manera, el archivo PDF “4000003983.15.pdf”, que constaba de 8 páginas, fue convertido en imágenes independientes etiquetadas secuencialmente como 0001.png, 0002.png, hasta el 008.png.

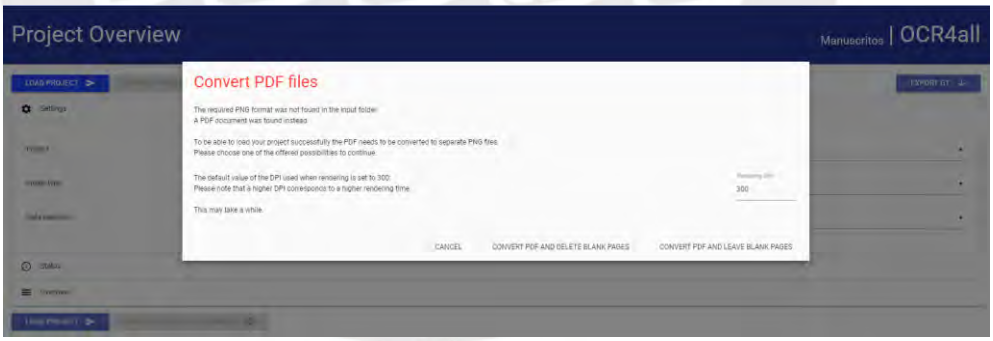


Figura 10: Carga de archivos en la herramienta

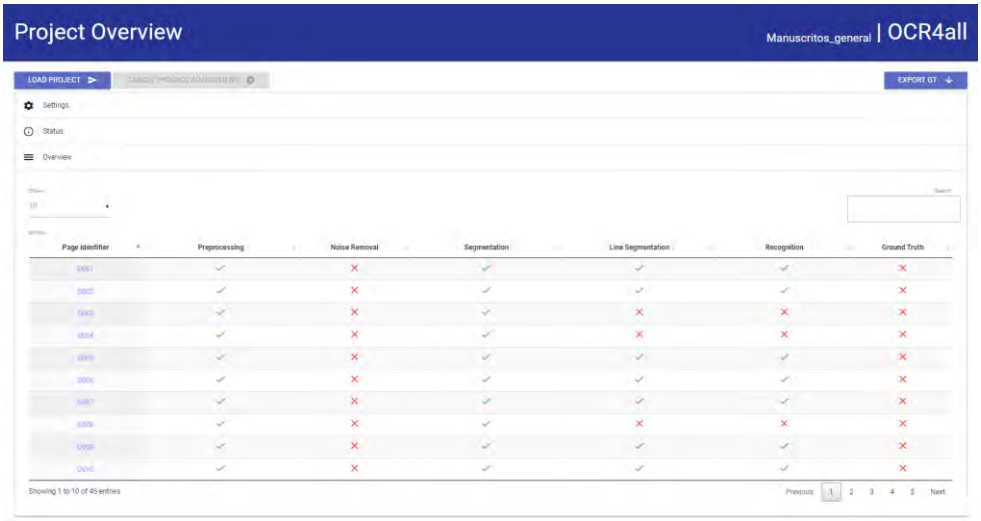


Figura 11: Visualización de archivos individuales cargados al nuevo proyecto

- Preprocesamiento:
  - Se realizó con la configuración predeterminada y sin cambios ("configuración general" y "configuración avanzada"), con lo que se crearon imágenes binarias y en escala de grises imágenes necesarias para una exitosa segmentación y reconocimiento OCR posterior.

- Segmentación:
  - Inicialmente se realizó la segmentación de las regiones para contar con la información de la posición, tipo y orden de lectura de las imágenes utilizando la segmentación Larex.
  - Seguidamente, se realizó la segmentación de líneas de las regiones de textos de las imágenes preprocesadas usando la configuración por defecto ("configuración general" y "configuración avanzada") aplicando la herramienta Kraken, integrada en la plataforma OCR4all.
  - Aun cuando se obtuvo una segmentación inicial estructurada para los manuscritos, fue necesario realizar segmentaciones manuales para la distinción entre texto y no texto, y sus especificaciones adicionales, esto también incluye información sobre el orden de lectura de la página, es decir, el orden de lectura y orden de uso de los elementos de diseño disponibles.
  - Por cada página actualizada se realizó la aprobación manual del resultado de manera individual.

- Reconocimiento de caracteres (OCR)
  - Una vez que las imágenes estuvieran listas, se realizaron varias experimentaciones:
    1. Reconocimiento óptico de caracteres con modelos mixtos preentrenados en otros idiomas y están disponibles automáticamente al crear la imagen de Docker.

2. Reconocimiento óptico de caracteres con los nuevos modelos creados en base al corpus (SPA-sentences)
3. Reconocimiento óptico de caracteres después de realizar transcripciones manuales en las páginas iniciales y usando los nuevos modelos creados en base al corpus (SPA-sentences)

- Transcripción manual:
  - En este paso se realizó la transcripción y/o corrección manual de las líneas mediante la aplicación LAREX.
  - Para este conjunto de datos se ha transcrito manualmente más de 1400 líneas de texto.

- Entrenamiento
  - El experimento se realizó principalmente con los parámetros predeterminados en la "configuración general" y la "configuración avanzada".
  - Se utilizó la estructura de red: *deep3* con la siguiente configuración: *cnn=40:3x3, pool=2x2, cnn=60:3x3, pool=2x2, cnn=120:3x3, lstm=200, lstm=200, lstm=200, dropout=0.5*
  - Para todos los casos, se configuró la lista blanca de caracteres con el alfabeto para idioma castellano, los caracteres especiales y tildes:

configuración:

ABCDEFGHIJKLMNÑOPQRSTUVWXYZabcdefghijklmnñopqrstuvw0123456789/().:?áéíóúÁÉÍÓÚ€$Üü<>

  - Se determinó el criterio de detención anticipada (early stopping) para que el entrenamiento se detiene si el CER de validación no mejoró tres veces consecutivas.
  - Respecto a otros (hiper) parámetros, se optó por los valores predeterminados de Calamari, siguiendo las mejores prácticas establecidas.
  - Los resultados obtenidos en las diferentes iteraciones del procesamiento de los manuscritos se detallan en la tabla 4.
  - El modelo generado (“modelo_SPA”) se almacenó para futuras pruebas con manuscritos adicionales.