Proyecto: Extracción automática de información en facturas mediante IA

Universidad Jorge Tadeo Lozano
Curso: Inteligencia Artificial – Bogotá D.C., Colombia
Participantes:

Sergio Andrés Ramírez
Nelson David Rincon Osorio
Brayan Sebastian Garcia Cespedes

Octubre 2025

1 Resumen

Proponemos un sistema de extracción automática de información en facturas usando visión computacional y OCR sobre el dataset *High-Quality Invoice Images for OCR* (Kaggle, +1 000 imágenes). El modelo combinará CNNs para segmentación y detección de texto con LayoutLM/LLM (ChatGPT o Llama) para la clasificación semántica de campos como NIT, fecha y valor total. Se espera alcanzar un F1 ¿ 0.85 y una CER inferior a la de Tesseract, entregando un prototipo local funcional en 6 semanas. El desarrollo garantizará privacidad mediante anonimización y mitigará sesgos con datos diversos de facturas.

2 Problema local y motivación

En muchas empresas colombianas, los equipos contables deben asignar personal exclusivamente para el proceso de legalización y registro de facturas, lo que implica revisar y transcribir manualmente información como fechas, NIT, valores, conceptos y proveedores. Este procedimiento, además de ser repetitivo y propenso a errores humanos, consume una cantidad significativa de tiempo y recursos.

Estas problemáticas motivan el desarrollo de un sistema automatizado de extracción y validación de datos en facturas, que permita agilizar la legalización documental, reducir errores de digitación y optimizar la carga laboral del personal contable. De esta forma, los profesionales del área podrán concentrar sus esfuerzos en tareas de análisis, control y supervisión, que realmente aporten valor a la gestión financiera de la organización.

3 Dataset

Fuente y enlace; tamaño; variables; condiciones de uso/licencia; por qué es representativo.

Enlace:
`https://www.kaggle.com/datasets/osamahosamabdellatif/high-quality-invoice-images-for-ocr/suggestions`

La base de datos es representativa dado que tiene una amplia cantidad de facturas escaneadas, lo cual nos permite realizar el modelamiento y entrenamiento del programa que se busca construir.

4 Tarea de IA y algoritmo(s)

La tarea principal de IA en este proyecto es la "extracción automática de información" a partir de facturas escaneadas utilizando imágenes, es decir, una tarea de visión computacional combinada con procesamiento OCR (Reconocimiento Óptico de Caracteres). El sistema busca extraer campos clave como número de factura, fechas, montos y nombres de proveedores directamente de las imágenes de facturas digitales, como las proporcionadas por el dataset de Kaggle "High-Quality Invoice Images for OCR".

Tipo de tarea

- Es una tarea multimodal, porque requiere tanto el análisis de imágenes (visión) como el procesamiento y estructuración del texto extraído (texto).
- El objetivo es aplicar OCR para convertir imágenes de facturas en texto digital y luego clasificar, localizar y extraer los diferentes campos relevantes.

Algoritmos y modelos propuestos

1. **Modelos de Visión para OCR:**
   - Un enfoque principal es entrenar y/o hacer *fine-tuning* de una CNN (Red Neuronal Convolucional) sobre las imágenes del dataset para mejorar la detección y segmentación de las áreas relevantes de las facturas (cabeceras, totales, tablas de productos, etc.).
   - Para la extracción de texto, se utilizarán modelos como Tesseract OCR o alternativas modernas basadas en *deep learning* (por ejemplo, CRNN: Convolutional Recurrent Neural Network o Transformers como LayoutLMv3, diseñados para captar tanto texto como la disposición y relaciones espaciales en documentos).
2. **Modelos para análisis semántico:**
   - Luego de obtener el texto, se pueden aplicar algoritmos NLP (Procesamiento de Lenguaje Natural) para clasificar los campos, validar la coherencia e identificar errores o valores atípicos, usando modelos tipo Llama o ChatGPT preentrenados.

3. **Justificación técnica:**

   - Las CNNs se escogen por su rendimiento comprobado en tareas de segmentación y detección en imágenes.
   - Los modelos OCR mencionados son estándar de la industria y están optimizados para facturas y documentos escaneados.
   - LayoutLM y variantes de Transformers combinan visión y texto, aprovechando la disposición espacial del documento para mejorar la extracción por encima del OCR tradicional.
   - El análisis posterior con modelos NLP permite categorizar, filtrar y validar los datos extraídos, aumentando precisión y utilidad.

5 Metodología y evaluación

Para el desarrollo del proyecto, se seguirá una metodología estructurada que abarca desde el tratamiento inicial de los datos hasta la evaluación final del sistema integrado.

Preprocesamiento de datos:

- **Imágenes (Facturas):** Las imágenes serán sometidas a un pipeline de preprocesamiento para optimizar la extracción de texto (binarización, eliminación de ruido, corrección de inclinación y normalización).
- **Texto Extraído:** Una vez que el modelo OCR extraiga el texto, se aplicarán técnicas de limpieza para normalizar los datos y estructurarlos para el análisis posterior.

División de Datos:

- 70% para entrenamiento.
- 15% para validación.
- 15% para pruebas.

Métricas de Evaluación:

- **OCR:** CER (Tasa de Error de Caracteres) y WER (Tasa de Error de Palabras).
- **Clasificación:** Precisión, Recall y F1-Score. Objetivo: F1 ¿ 0.85.

Líneas Base y Comparación: Se establecerá una línea base con Tesseract OCR y se comparará con el rendimiento del modelo CNN. El LLM se contrastará con un enfoque basado en expresiones regulares.

6 Resultados esperados, ética y cronograma

Resultados esperados e hipótesis:

- **Hipótesis 1 (Precisión superior):** La CNN logrará una CER menor que Tesseract.
- **Hipótesis 2 (Clasificación efectiva):** El sistema CNN + LLM alcanzará F1 ¿ 0.85.
- **Hipótesis 3 (Prototipo funcional):** Se entregará un prototipo local capaz de procesar una imagen de factura y generar un documento estructurado.

Consideraciones éticas y mitigación de riesgos:

- **Privacidad de Datos:** Las facturas contienen información sensible.
  - **Mitigación:** Se trabajará con datos anonimizados y entornos seguros. Se aplicará desenfoque o recorte de información personal.
- **Sesgos en el Modelo:** El modelo podría tener bajo rendimiento en formatos no comunes.
  - **Mitigación:** Garantizar diversidad de datos y análisis periódicos para detectar sesgos.
- **Fiabilidad y Errores:** Un error en la extracción puede afectar resultados contables.
  - **Mitigación:** Se incluirá una puntuación de confianza y validación cruzada para campos críticos.

Cronograma (resumen):

- Semana 1: Definición y adquisición del dataset.
- Semana 2: Configuración del entorno y preprocesamiento.
- Semana 3: Entrenamiento inicial del modelo CNN.
- Semana 4: Integración OCR + LLM y ajuste fino.
- Semana 5: Evaluación cuantitativa.
- Semana 6: Documentación y entrega del prototipo.

References