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 :

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

4. Tarea de IA y algoritmo(s)

La tarea principal de IA en este proyecto es "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 pre-entrenados, como se indica en tu Word.

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 de las facturas serán sometidas a un pipeline de preprocesamiento para optimizar la extracción de texto. Esto incluye técnicas de binarización para mejorar el contraste, eliminación de ruido, corrección de inclinación (deskewing) y normalización de tamaño e iluminación.

Texto Extraído: Una vez que el modelo OCR extraiga el texto, se aplicarán técnicas de limpieza para normalizar los datos, como la eliminación de caracteres especiales irrelevantes, la corrección de errores comunes de OCR y la estructuración del texto para su posterior análisis por el modelo de lenguaje.

División de Datos El conjunto de datos de facturas se segmentará en tres subconjuntos para garantizar un entrenamiento y evaluación robustos del modelo:

70% para Entrenamiento: Utilizado para entrenar los modelos de extracción (CNN) y análisis (LLM).

15% para Validación: Empleado para el ajuste de hiperparámetros de los modelos durante la fase de entrenamiento, evitando el sobreajuste.

15% para Pruebas (Testing): Un conjunto de datos completamente nuevo para los modelos, reservado para la evaluación final y la medición del rendimiento real del prototipo.

Métricas de Evaluación Se utilizarán métricas específicas para evaluar cada componente del sistema:

Extracción de Texto (OCR): El rendimiento del modelo CNN se medirá utilizando la Tasa de Error de Caracteres (CER) y la Tasa de Error de Palabras (WER), comparando el texto extraído con una transcripción manual de referencia (ground truth).

Análisis y Clasificación de Campos: La capacidad del modelo de lenguaje para identificar y clasificar correctamente los campos de la factura (ej: NIT, fecha, valor total, ítems) se evaluará con las métricas de Precisión (Precision), Exhaustividad (Recall) y Puntuación F1 (F1-Score). El objetivo es superar un F1-Score de 0.85, como se indica en los objetivos del proyecto.

Líneas Base y Comparación Para cuantificar la mejora aportada por nuestros modelos, se establecerá una línea base (baseline) utilizando un motor de OCR de código abierto como Tesseract. El rendimiento de nuestro modelo CNN se comparará directamente con los resultados de esta línea base. Adicionalmente, la eficacia del LLM para el análisis de campos se contrastará con un enfoque más simple basado en expresiones regulares (regex).

6. Resultados esperados, ética y cronograma

Resultados Esperados e Hipótesis

Hipótesis 1 (Precisión Superior): Se espera que el modelo de extracción de texto basado en una Red Neuronal Convolucional (CNN) logre una Tasa de Error de Caracteres (CER) significativamente menor en comparación con la línea base de Tesseract en el conjunto de datos de prueba.

Hipótesis 2 (Clasificación Efectiva): Se plantea que el sistema integrado (CNN + LLM) alcanzará una puntuación F1 > 0.85 en la tarea de extracción y clasificación de los campos clave de las facturas.

Hipótesis 3 (Prototipo Funcional): Al finalizar el proyecto, se entregará un prototipo local capaz de procesar una imagen de factura y generar un documento final estructurado con los datos extraídos y validados.

Consideraciones Éticas y Mitigación de Riesgos

Privacidad de Datos: Las facturas contienen información sensible.

Mitigación: Se trabajará con datos anonimizados siempre que sea posible. Se establecerá una política estricta de manejo de datos, asegurando que el procesamiento se realice en un entorno seguro y que los datos no se expongan. Se respetará la privacidad mediante técnicas de desenfoque o recorte de información personal identificable en las fases de desarrollo.

Sesgos en el Modelo: El modelo podría tener un rendimiento deficiente en facturas con formatos o diseños poco comunes si el conjunto de datos no es diverso.

Mitigación: Es fundamental asegurar que el dataset de entrenamiento sea representativo de una amplia variedad de formatos de facturas. Se realizarán análisis periódicos para detectar y corregir posibles sesgos en el rendimiento del modelo.

Fiabilidad y Errores: La extracción automática no es infalible y un error (ej: en el valor total) puede tener consecuencias.

Mitigación: El sistema incluirá una puntuación de confianza para cada campo extraído, señalando al usuario los datos que pueden requerir una verificación manual. Para campos críticos, se podría implementar una lógica de validación cruzada.

Cronograma de Trabajo (6 Semanas)

# Cronograma de trabajo

|  |  |  |
| --- | --- | --- |
| Semana | Fechas Clave | Actividades Principales |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| 1 | 6 oct - 12 oct | Definición final del dataset y selección de arquitecturas (CNN, LLM). |
| 2 | 13 oct - 19 oct | Configuración del entorno (Python, NVIDIA) y pipeline de preprocesamiento. |
| 3 | 20 oct - 26 oct | Entrenamiento inicial del modelo OCR (CNN/CRNN) y pruebas preliminares. |
| 4 | 27 oct - 2 nov | Integración OCR + LLM. Ajuste fino del modelo de lenguaje. |
| 5 | 3 nov - 9 nov | Evaluación con métricas (F1, CER). Iteración y mejoras. |
| 6 | 10 nov | Documentación técnica, presentación y entrega del prototipo funcional. |

# Referencias

1. Do, T. H., & Le, T. (2022). Document Analysis and Recognition: A survey. arXiv preprint arXiv:2203.00339.
2. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
3. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, 86(11), 2278–2324.
4. Brown, T. B., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33.
5. https://marketingdigitalconsulting.com/beneficios-de-usar-asistentes-basados-en-ia/
6. https://www.tidio.com/es/blog/asistentes-inteligencia-artificial/
7. https://botpress.com/es/blog/what-is-an-ai-assistant
8. DIAN (2024). Guía de Facturación Electrónica y Automatización Contable en Colombia. https://www.dian.gov.co/
9. Portafolio (2024). Cómo la inteligencia artificial está transformando la contabilidad empresarial en Colombia. https://www.portafolio.co/
10. Actualícese (2025). Automatización contable y uso de IA en procesos de facturación electrónica. https://actualicese.com/