IDEAS

Innovation & Development in Engineering and Applied Science

Journal homepage:
https://revistasojs.utn.edu.ec/index.php/ideas

Revisión sistemática de técnicas basadas en IA para la extracción de información de documentos no estructurados

Systematic review of artificial intelligence-based techniques for information extraction from unstructured documents

Johanna Guamán Granda[a], Pablo Landeta López[a], Zoila Irene Rubio Clavijo[b]

[a]Universidad Técnica del Norte, Ibarra, Ecuador.

[b]Unidad Educativa 17 de Julio, Ibarra, Ecuador.

Recibido/Received: 2026-01-14
Aceptado/Accepted: 2026-06-16
Publicado/Published: 2026-07-31
Autor de correspondencia: jaguamang1@utn.edu.ec
DOI: https://doi.org/10.53358/ideas.v8i2.1354
Copyright: CC BY 4.0

PALABRAS CLAVES:

Inteligencia artificial; OCR; Procesamiento del lenguaje natural; Extracción de información; Aprendizaje automático; Aprendizaje profundo; Documentos no estructurados; Revisión sistemática PRISMA.

RESUMEN

En la actualidad, muchas organizaciones gestionan o conservan sus documentos físicos en formato digital, guardando la información en archivos no estructurados como los PDF; sin embargo, al tratarse de documentos escaneados, la extracción y el análisis de la información resultan difíciles de obtener. Este artículo presenta una revisión sistemática que analiza técnicas basadas en inteligencia artificial aplicadas a la extracción de información en documentos no estructurados. Para ello, se empleó la metodología PRISMA, considerando estudios publicados entre 2020 y 2025 en bases de datos científicas como IEEE Xplore, SpringerLink, Google Scholar, ACM Digital Library y ScienceDirect. Como resultado del proceso de selección, se analizaron 40 estudios relacionados con el reconocimiento, procesamiento y análisis de documentos impresos, manuscritos y mixtos. Los hallazgos evidencian que las técnicas tradicionales de OCR han evolucionado hacia enfoques más avanzados basados en aprendizaje automático, aprendizaje profundo, procesamiento del lenguaje natural y modelos de lenguaje de gran escala. Entre las técnicas más utilizadas se identifican CNN, RNN, LSTM, BERT, modelos Transformers y LLMs, las cuales permiten mejorar tareas como reconocimiento de texto, clasificación documental, extracción de entidades y análisis semántico. Los resultados muestran que los enfoques basados en inteligencia artificial ofrecen ventajas frente a los métodos tradicionales, especialmente en la automatización del procesamiento documental y en la comprensión contextual de la información. Sin embargo, persisten limitaciones asociadas con la calidad de las imágenes, la variabilidad de formatos, el reconocimiento de escritura manuscrita y la falta de conjuntos de datos estandarizados. Se concluye que la integración de OCR, NLP, aprendizaje profundo y modelos generativos constituye una tendencia relevante para optimizar la extracción de información en documentos no estructurados. No obstante, se requieren futuras investigaciones orientadas a la evaluación comparativa de modelos, la mejora de métricas de desempeño y el desarrollo de marcos metodológicos reproducibles.

KEYWORDS:

Artificial intelligence; OCR; Natural language processing; Information extraction; Machine learning; Deep learning; Unstructured documents; PRISMA systematic review.

ABSTRACT

Currently, many organizations manage or store their physical documents in digital format, saving information in unstructured files such as PDFs; however, since these are scanned documents, extracting and analyzing the information is difficult. This article presents a systematic review that analyzes artificial intelligence–based techniques applied to information extraction in unstructured documents. For this purpose, the PRISMA methodology was used, considering studies published between 2020 and 2025 in scientific databases such as IEEE Xplore, SpringerLink, Google Scholar, ACM Digital Library, and ScienceDirect. As a result of the selection process, 40 studies related to the recognition, processing, and analysis of printed, handwritten, and mixed documents were analyzed. The findings show that traditional OCR techniques have evolved toward more advanced approaches based on machine learning, deep learning, natural language processing, and large language models. Among the most commonly used techniques are CNN, RNN, LSTM, BERT, Transformer models, and LLMs, which make it possible to improve tasks such as text recognition, document classification, entity extraction, and semantic analysis. The results show that artificial intelligence–based approaches offer advantages over traditional methods, especially in automating document processing and in the contextual understanding of information. However, limitations persist related to image quality, format variability, handwritten text recognition, and the lack of standardized datasets. It is concluded that the integration of OCR, NLP, deep learning, and generative models constitutes a relevant trend for optimizing information extraction in unstructured documents. Nevertheless, further research is required, focused on the comparative evaluation of models, the improvement of performance metrics, and the development of reproducible methodological frameworks.

Cómo citar

J. Guamán Granda, P. Landeta López, and Z. I. Rubio Clavijo, “Revisión sistemática de técnicas basadas en IA para la extracción de información de documentos no estructurados,” *Innovation & Development in Engineering and Applied Science*, vol. 8, no. 2, p. 29, 2026, **https://doi.org/10.53358/ideas.v8i2.1354**

1. Introducción

El crecimiento de documentos digitales no estructurados, especialmente en formato PDF, ha generado nuevos desafíos para la extracción, organización y análisis automatizado de información. Diversos estudios evidencian que el manejo y el procesamiento de documentos no estructurados representan un desafío significativo para las organizaciones, debido a la complejidad de los formatos, la presencia de texto, imágenes y tablas, y la alta dependencia de procesos manuales, lo que incrementa el riesgo de errores y los costos operativos [1, 2, 3].

Sin embargo, el procesamiento automático de documentos no estructurados continúa siendo un problema abierto, debido a la variabilidad en la calidad de las imágenes, la diversidad de formatos, la presencia de ruido, la complejidad del texto manuscrito y la limitada disponibilidad de conjuntos de datos estandarizados. Estas condiciones afectan la precisión de técnicas como OCR (Reconocimiento Óptico de Caracteres), aprendizaje automático, aprendizaje profundo y modelos de lenguaje aplicados a la extracción de información.

Además, la información contenida en este tipo de documentos tiene un amplio campo de aplicación en distintos sectores, lo que resalta su importancia estratégica. En este sentido, la transición hacia modelos de gestión documental digital no solo mejora la eficiencia operativa, sino que también contribuye a objetivos organizacionales y sostenibles [4, 5]. Por otra parte, la literatura señala que una adecuada gestión de la información contribuye significativamente a la competitividad organizacional, al facilitar la toma de decisiones y optimizar los procesos internos. Factores como la innovación, la agilidad organizacional y la capacidad de adaptación tecnológica son determinantes para generar ventajas competitivas sostenibles en entornos empresariales dinámicos [6, 7, 8].

No obstante, aunque estos enfoques destacan la importancia de la gestión de la información a nivel organizacional, presentan limitaciones cuando se aplican específicamente al procesamiento automatizado de documentos no estructurados, donde la complejidad del contenido exige soluciones más robustas basadas en inteligencia artificial.

Diversos estudios también evidencian que la inteligencia empresarial y la agilidad organizacional influyen positivamente en el rendimiento innovador de las organizaciones. Sin embargo, a pesar de estos avances, persisten desafíos importantes en la gestión eficiente de la información, especialmente cuando los datos se encuentran en formatos no estructurados como archivos PDF, lo que limita su aprovechamiento en procesos de toma de decisiones [9, 6].

Ahora bien, una gestión documental eficaz se consolida como un factor clave para mejorar la capacidad de adaptación organizacional frente a entornos dinámicos. En este sentido, la integración de la inteligencia artificial en los procesos de toma de decisiones está transformando la forma en que las organizaciones operan, permitiendo un análisis más eficiente de grandes volúmenes de información [6, 10].

Textos, imágenes y documentos escaneados son ejemplos de datos no estructurados que las soluciones basadas en Inteligencia Artificial (IA) pueden interpretar y clasificar con mayor precisión que los métodos tradicionales. Dado el volumen creciente de estos datos y la necesidad de aprovecharlos de forma efectiva, se requiere un enfoque basado en IA que permita procesarlos automáticamente y extraer información útil [11]. La Figura 1 presenta una visión general del problema asociado al procesamiento de documentos no estructurados. En ella se ilustran los diferentes tipos de documentos, incluyendo aquellos con texto impreso, manuscrito y datos mixtos, así como las principales dificultades que surgen en su tratamiento, tales como la baja calidad de las imágenes, la variabilidad en los estilos de escritura y la presencia de ruido en los datos.

Para obtener información precisa y útil de documentos extensos no estructurados, es necesario contar con un enfoque que permita resumir y filtrar los contenidos, eliminando datos irrelevantes y destacando lo esencial de manera rápida y automatizada [12]. La IA también contribuye a generar análisis más claros y comprensibles de este tipo de materiales, lo que facilita su incorporación en procesos clave de toma de decisiones dentro de las empresas [11].

A pesar de los avances en el uso de técnicas de inteligencia artificial, aún existen limitaciones importantes en la comprensión semántica y estructural de los documentos no estructurados. En este contexto, se hace necesaria una revisión sistemática de la literatura enfocada en las estrategias de extracción de información, que aborde de forma clara sus ventajas, limitaciones, clasificaciones y comparaciones. En la literatura actual se observa una carencia de análisis profundos sobre los conjuntos de datos de acceso público, así como sobre las técnicas de preprocesamiento y clasificación de datos. Asimismo, no se dispone de una evaluación integral de los métodos orientados a la extracción automática de información a partir de materiales no estructurados.

El objetivo de esta revisión sistemática es analizar, comparar y sintetizar las técnicas basadas en inteligencia artificial utilizadas para la extracción de información en documentos no estructurados, con el fin de identificar sus principales enfoques, ventajas, limitaciones, métricas de evaluación, tendencias emergentes y vacíos de investigación.

Con base en este objetivo, la presente investigación se orienta a responder las siguientes preguntas de investigación:

- **Q1.** ¿Cuáles son los enfoques basados en IA utilizados en el reconocimiento de datos de documentos no estructurados?
- **Q2.** ¿Cuáles son los dominios y aplicaciones que utilizan el procesamiento de datos no estructurados?
- **Q3.** ¿Cuáles son las diversas técnicas de preprocesamiento de datos asociadas con el procesamiento de documentos no estructurados?
- **Q4.** ¿Cuáles son los principales desafíos que enfrentan los algoritmos de extracción de información al procesar documentos no estructurados?

El informe está estructurado de la siguiente forma: en la sección I, aborda el estado del arte en el área de interés. En la sección II, se aplica la metodología PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [13], para la revisión de la literatura usando ecuaciones de estrategias y de búsqueda, como también criterios de inclusión y exclusión tal como se aplica en Monge-García et al [14]. En la sección III, se describe el proceso

de selección y análisis de estudios relevantes según criterios específicos. En la sección IV, se presenta un resumen de los resultados antes analizados, en el que se evalúan su eficacia, limitaciones y coherencia con estudios investigados. Asimismo, se adapta a una discusión crítica que relaciona los hallazgos con la literatura existente y plantea sus posibles implicaciones para futuras investigaciones. Finalmente, en la sección V, se exponen conclusiones dirigidas a ayudar a los investigadores que trabajan en esta área a crear métodos efectivos de extracción de información para documentos no estructurados.

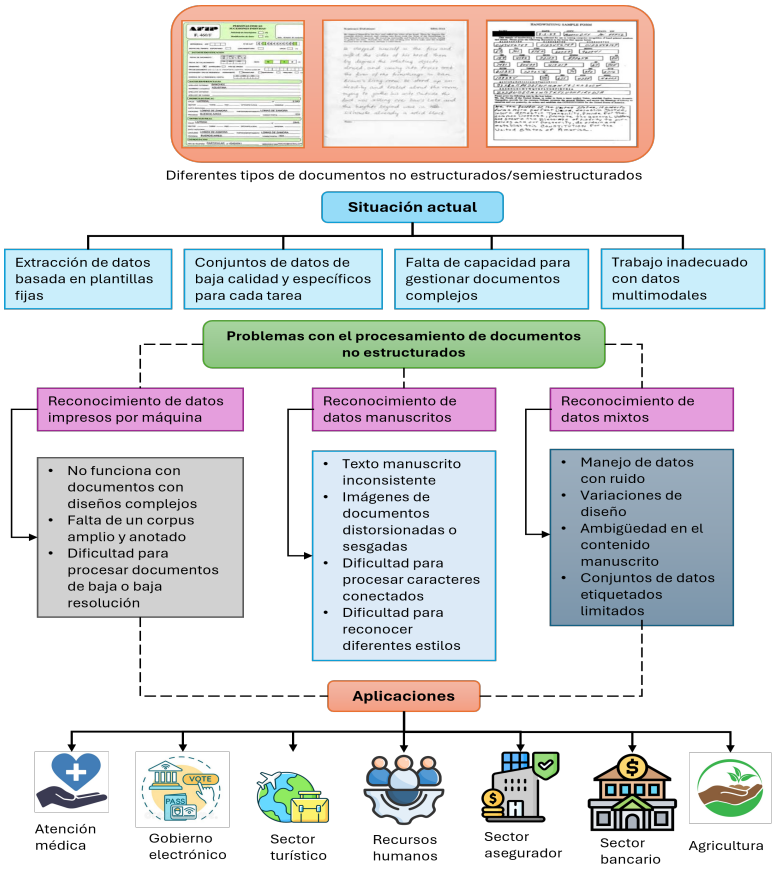

Figura 1: ¿Por qué el procesamiento de documentos no estructurados? [11]

Evolución de las técnicas utilizadas para el procesamiento de datos de texto mixto

Los métodos para la extracción automática de información a partir de documentos no estructurados han experimentado una evolución significativa a lo largo del tiempo, como se muestra en la Figura 2. En ella se observa el paso de técnicas tradicionales, como el OCR básico, hacia enfoques más avanzados que incorporan aprendizaje automático, aprendizaje profundo y automatización inteligente. A pesar de los avances, el reconocimiento de escritura manual (también conocido como OCR manuscrito o HTR) sigue siendo un reto técnico importante [15]. Lograr que el texto manuscrito sea legible por máquina es muy complicado debido a la gran variedad de estilos de escritura entre individuos y la baja calidad del texto manuscrito en comparación con el texto impreso. Los primeros estudios sobre el reconocimiento de escritura se remontan a las décadas de 1960 y 1970.

Sin embargo, en los años 80, la investigación en esta área disminuyó debido a los resultados poco satisfactorios obtenidos por los sistemas de entonces. En los años 90, comenzaron a desarrollarse algoritmos más eficaces, como los Modelos Ocultos de Markov (HMM), la programación dinámica para coincidencia de patrones, y las Redes Neuronales (NN), etc. A partir del año 2000, se incorporaron herramientas como léxicos, diccionarios y modelos lingüísticos, además de estrategias que combinaban múltiples sistemas de reconocimiento. En la década de 2010, se introdujeron robots de software con tecnología RPA (Automatización Robótica de Procesos) para procesar tanto texto manuscrito como impreso. Desde el año 2020 en adelante, el uso del OCR basado en IA evolucionó con la ayuda de técnicas de aprendizaje automático y aprendizaje profundo en esta área [11].

En cuanto al procesamiento de documentos impresos por máquina, anteriormente las empresas dependían

de personal para introducir datos de forma manual y gestionar los documentos en papel antes de que esa información pasara a los sistemas internos. El primer paso hacia la digitalización fue el uso del OCR, que permitía convertir documentos escaneados en archivos digitales. Las primeras versiones de OCR eran muy limitadas: solo podían reconocer una fuente a la vez y requerían imágenes específicas para cada carácter. A principios de los años 2000, se propuso el uso del "OCR omni-fuente" capaz de interpretar texto en múltiples tipografías.

Este avance se consolidó como un servicio en la nube, accesible desde computadoras y dispositivos móviles. Actualmente, gracias a APIs ofrecidas por diversos proveedores, es posible identificar con gran precisión una amplia variedad de caracteres y estilos de escritura. Con el progreso de la automatización, las empresas también comenzaron a incorporar soluciones de RPA para reemplazar tareas estructuradas, repetitivas y basadas en reglas, debido a que estos sistemas funcionan mediante lógica predefinida, ejecutando automáticamente procesos rutinarios. A lo largo del tiempo, la RPA ha pasado por distintas fases de evolución, desde sus versiones 1.0 hasta 3.0 [4]. Hoy en día, la automatización potenciada por IA emplea múltiples técnicas como Visión por Computadora, Procesamiento del Lenguaje Natural (NLP), aprendizaje automático (ML), aprendizaje profundo (DL) y análisis de texto para evaluar tanto datos estructurados como no estructurados. Gracias a estos avances, es posible clasificar, procesar y analizar grandes volúmenes de datos no estructurados, extrayendo información valiosa con mínima intervención humana [11].

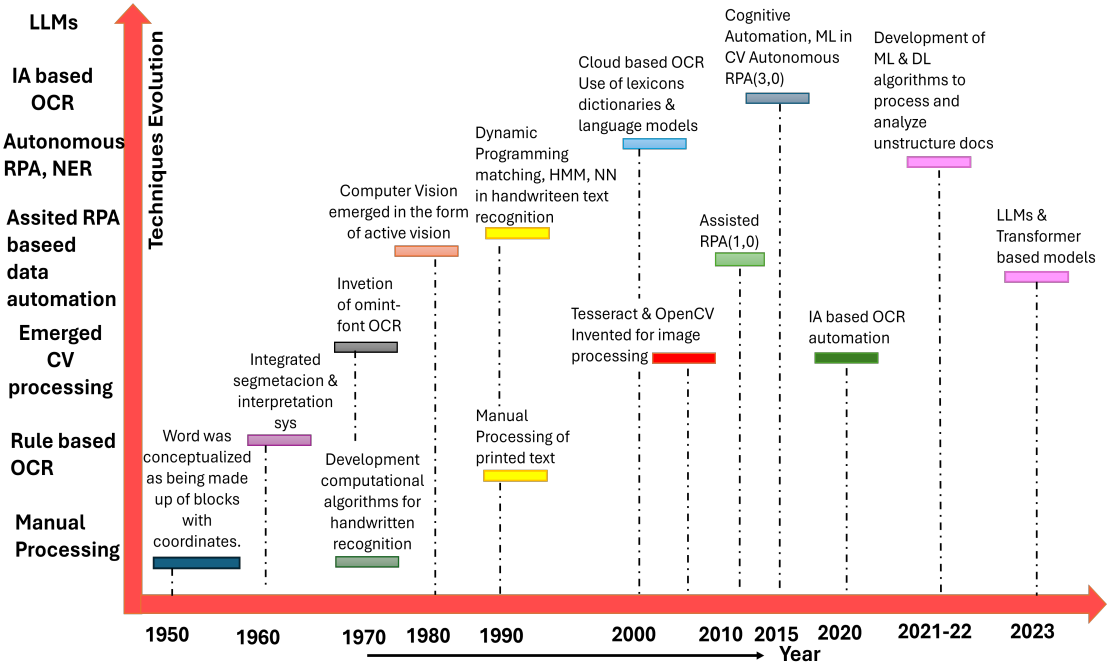

Figura 2: Evolución del procesamiento de documentos basado en texto mixto (impreso y manuscrito) [11].

Procesamiento de documentos no estructurados

A. Preparación y preprocesamiento de datos

La captura de una imagen de un documento no estructurado que contenga texto manuscrito o impreso es una parte fundamental de la fase de preparación de datos. Para ello, se requiere que la imagen esté en un formato específico, como PNG o JPEG, y puede obtenerse mediante cámaras digitales, escáneres u otros dispositivos adecuados. Posteriormente, en la etapa de digitalización, el documento físico se convierte en formato electrónico. Esto implica escanear el documento original para generar una imagen digital que pueda almacenarse y procesarse en un ordenador. Esta imagen es indispensable para iniciar el preprocesamiento [16], en la figura 3 se demuestra el procesamiento de documentos no estructurados, el proceso inicia con la adquisición del documento en formato digital, seguido de la extracción del texto mediante técnicas OCR.

Luego se aplican métodos de preprocesamiento con el fin de mejorar la calidad de la imagen, reduciendo el ruido y corrigiendo posibles distorsiones. A continuación, se realiza la extracción de características relevantes del texto, las cuales son utilizadas por modelos de aprendizaje automático y aprendizaje profundo para la clasificación y reconocimiento del contenido. Finalmente, el sistema transforma el texto en un formato digital estructurado y lleva

a cabo la extracción de información, permitiendo obtener datos relevantes y generar un resumen del documento procesado. Este flujo evidencia la importancia de integrar múltiples técnicas de inteligencia artificial para lograr un procesamiento eficiente y preciso de documentos no estructurados.

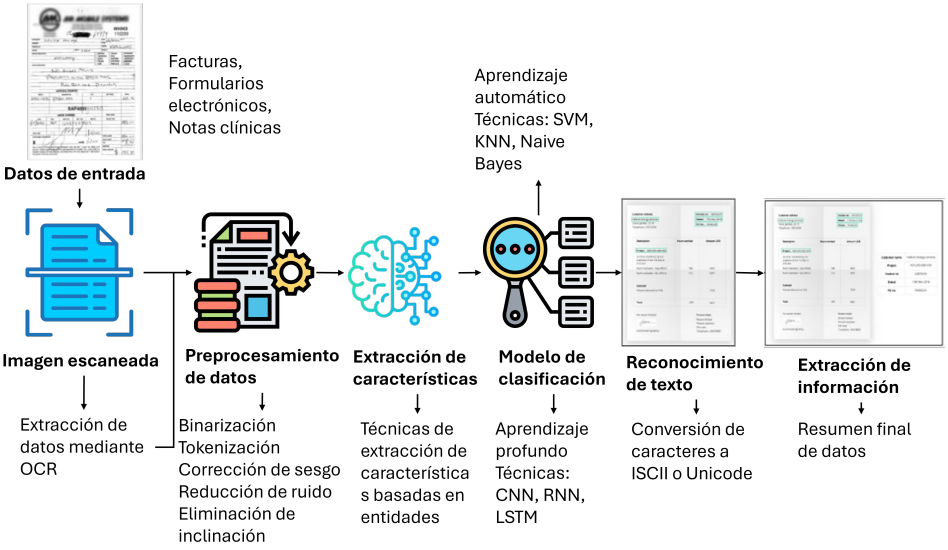

Figura 3: Reconocimiento de texto impreso y manuscrito a partir de documentos no estructurados [11].

En la Tabla 1 se distinguen cuatro etapas que generalmente un sistema OCR usa para el procesamiento de documentos no estructurados como preprocesamiento, segmentación, reconocimiento y posprocesamiento. La etapa de preprocesamiento busca optimizar la calidad de la imagen de entrada mediante técnicas como la binarización, el filtrado de ruido, la detección de bordes y las operaciones morfológicas, con el fin de mejorar la calidad de los datos para su posterior análisis. En la etapa de segmentación la imagen se subdivide en regiones o caracteres relevantes, lo que facilita un reconocimiento más preciso. En la etapa de reconocimiento se utilizan tanto algoritmos tradicionales de aprendizaje automático como enfoques avanzados basados en aprendizaje profundo, frecuentemente apoyados por recursos adicionales como diccionarios o modelos lingüísticos. Finalmente, el posprocesamiento se encarga de pulir los resultados obtenidos, corrigiendo posibles errores y ajustando el formato de salida para su uso práctico [17].

Tabla 1: Pasos de procesamiento de OCR

Técnica	**Ventajas**
Preprocesamiento [18]	Implica la aplicación de patrones de colores específicos en los exámenes para mejorar la detección de preguntas y respuestas antes de la binarización.
Preprocesamiento [19]	Algoritmo de Normalización MVSR (Media-Varianza-Softmax-Reescalado, Mean–Variance–Softmax–Rescale) entre las operaciones de escala de grises y binarización, las imágenes se normalizan utilizando la media global de píxeles y la desviación estándar.
Segmentación [20]	Utiliza el etiquetado de componentes conectados para identificar caracteres individuales.
Segmentación [21]	Emplea histogramas de proyección para la segmentación de líneas y palabras, principalmente para dividir líneas largas para el entrenamiento de modelos de redes neuronales recurrentes convolucionales (CRNN).
Reconocimiento [22]	Utiliza modelos Tesseract para los idiomas turco e inglés, con una mejora significativa en el modelo turco.
Reconocimiento [23]	Emplea un modelo LSTM entrenado y probado con un conjunto de datos de dos novelas otomanas.
Post-Procesamiento [24]	Se centra en mejorar el reconocimiento de matrículas en línea mediante un sistema en chip para matrículas turcas.

Continúa en la siguiente página...

Tabla 1(continuación)

Técnica	Ventajas
Post-Procesamiento [25]	Se centra en el uso de un modelo BiLSTM basado en caracteres para la corrección de errores.

B. Extracción de características

En cualquier técnica de reconocimiento, la selección de características representa una etapa clave. Estas deben ser fáciles de calcular y, al mismo tiempo fiables [11]. GloVe, TF-IDF, and Word2Vec son técnicas comunes de extracción de características empleadas en tareas de categorización y reconocimiento de texto. Para generar representaciones numéricas o vectoriales de las palabras, se puede recurrir a modelos como Global Vectors (GloVe) [4].

Mahadevkar et al. [11], manifiesta que el método TF-IDF es una técnica ampliamente utilizada para asignar pesos a las palabras según su relevancia dentro de un texto. Por otro lado, una de las formas más comunes de utilizar redes neuronales (neural networks) para generar incrustaciones de palabras a partir de grandes volúmenes de datos es Word2Vec, que puede emplear enfoques como el Continuous Bag of Words (CBOW). Para convertir texto en una matriz o vector de características, se requieren algoritmos específicos de extracción, proceso conocido generalmente como extracción de características de datos. En la Figura 4, se presentan los principales métodos de extracción de características utilizados en el procesamiento de texto, tales como Bag of Words, GloVe, TF-IDF, Word2Vec, CBOW y Globor Filter. Cada uno de estos enfoques permite representar la información textual en forma numérica, facilitando su análisis por parte de modelos de aprendizaje automático.

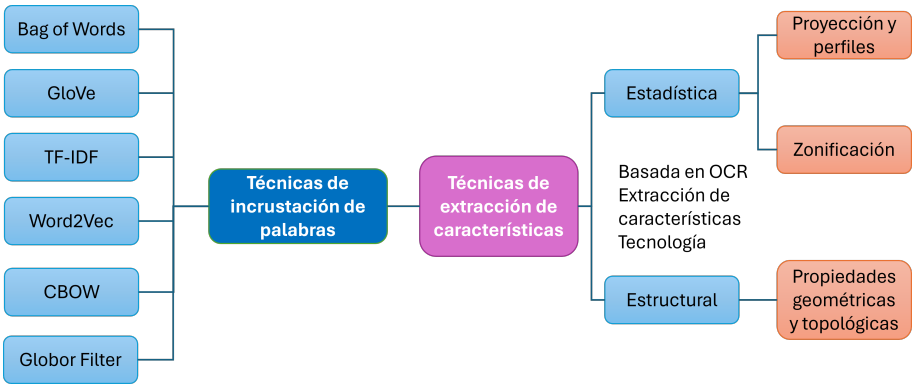

Figura 4: Técnicas de extracción de características [11]

En la fase de extracción de características se emplean diversos métodos orientados a transformar la información contenida en los documentos en representaciones que puedan ser procesadas por modelos de análisis y clasificación. Como se muestra en la Tabla 2, estas técnicas permiten convertir datos textuales y visuales en vectores o descriptores que facilitan el reconocimiento automático tanto de texto impreso como manuscrito, es decir estas técnicas pueden agruparse en enfoques basados en frecuencia, representación semántica y características estructurales. Métodos como Bag of Words y TF-IDF se centran en la frecuencia y relevancia de los términos, siendo útiles para tareas de clasificación básica; sin embargo, presentan limitaciones en la comprensión del contexto. Por otro lado, enfoques como GloVe y Word2Vec permiten capturar relaciones semánticas entre palabras, generando representaciones más ricas y adecuadas para el análisis de contenido.

Tabla 2: Técnicas de extracción de características

Técnicas	Descripción
Bag of Words (BoW)	Método que representa el texto mediante vectores dispersos según la presencia, ausencia o frecuencia de palabras, incluyendo variantes como unigramas, bigramas o trigramas.
GloVe (Global Vectors)	Técnica para generar representaciones vectoriales de palabras basada en estadísticas de coocurrencia. Captura relaciones semánticas globales y es útil en tareas de similitud y análisis léxico [26].
TF-IDF	Representación numérica que asigna pesos a palabras según su relevancia, combinando la frecuencia del término en un documento y su rareza en un corpus para identificar términos significativos.
Word2Vec	Modelo neuronal que aprende incrustaciones densas de palabras mediante arquitecturas como Skip-gram o CBOW, permitiendo capturar relaciones semánticas complejas.
Extracción de características estadísticas	Se basa en el análisis cuantitativo de las imágenes de texto. Una técnica común es la zonificación, que consiste en dividir la imagen en múltiples zonas para luego extraer características específicas de cada una, construyendo así un vector de características representativo. Otro método es el uso de proyecciones y perfiles [27].
Extracción de características estructurales	Capturan los rasgos geométricos del carácter, es decir que se analizan elementos como los trazos, líneas horizontales y verticales, puntos finales, intersecciones y bucles, que forman la estructura básica del carácter. Esta información permite entender la composición visual del signo, lo cual resulta esencial para diferenciar caracteres con formas complejas o similares entre sí.

C. Técnicas de clasificación de datos

En la etapa de clasificación se identifican los modelos más utilizados y efectivos para el reconocimiento de caracteres. Los enfoques más comúnmente utilizados por los investigadores para esta revisión son el aprendizaje automático y el aprendizaje profundo que se utilizan para clasificar texto impreso y manuscrito.

Algoritmos de aprendizaje automático

El aprendizaje automático es el proceso de extraer información directamente de una gran cantidad de instancias o muestras de entrenamiento utilizando técnicas computacionales. Estos algoritmos son capaces de ajustar sus modelos de forma adaptativa para encontrar soluciones óptimas, y tienden a mejorar su rendimiento conforme se les expone a una mayor cantidad de ejemplos durante el proceso de aprendizaje [28].

Los enfoques de aprendizaje automático más empleados para la clasificación de texto impreso y manuscrito en estudios de la literatura se presentan a continuación:

- Clasificador de Naive Bayes: este es un enfoque probabilístico para la clasificación. Para determinar la clase de datos nuevos o no reconocidos, aplica el teorema de Bayes de la probabilidad [29].
- Red Neuronal Artificial (ANN): este tipo de red ha mostrado una notable eficacia en la automatización del reconocimiento de texto y la extracción de datos, gracias a su capacidad para captar las relaciones entre letras y palabras. En el conjunto de datos del mundo real de pasaportes chinos y recibos médicos en [30], el reconocimiento de objetos se lleva a cabo utilizando redes neuronales convolucionales basadas en regiones (R-CNN).
- Máquina de Vectores de Soporte (SVM): es un modelo de aprendizaje supervisado ampliamente utilizado para tareas de clasificación y regresión. Aunque puede manejar datos no lineales proyectándolos en espacios de características de alta dimensión, su aplicación principal se centra en clasificar datos que son linealmente separables. Este enfoque maximiza la separación entre categorías, lo que mejora la robustez y precisión del modelo [31].
- K vecinos más cercanos (K-NN): este método clasifica los datos en función de la similitud entre elementos, asignando etiquetas según las características de los objetos más cercanos en el espacio de características. Es especialmente útil para tareas de reconocimiento de patrones en documentos escaneados. En investigaciones

que involucran alfabetos latinos y devanagari, K-NN se ha empleado eficazmente para distinguir entre letras mayúsculas y minúsculas del alfabeto latino, así como para identificar vocales y consonantes en devanagari [32].

- Árbol de decisión: es un clasificador supervisado que no depende de reglas predefinidas, sino que estructura el proceso de toma de decisiones a través de nodos, ramas y bordes dirigidos, similar a la forma de un árbol, mediante la poda, que consiste en eliminar ramas poco relevantes o redundantes, se reduce la complejidad del árbol, se evita el sobreajuste y se mejora la capacidad predictiva del modelo [33].

Algoritmos de aprendizaje profundo

Los clasificadores basados en aprendizaje profundo se emplean principalmente en contextos de aprendizaje no supervisado, aunque también pueden aplicarse en enfoques supervisados o semi-supervisados. Este tipo de algoritmos destaca por su capacidad para mejorar el rendimiento y la precisión del sistema, ya que aprenden y extraen características automáticamente a partir de los datos. Entre los modelos más comunes utilizados para la identificación y clasificación de texto, tanto manuscrito como impreso, son los siguientes:

- Red Neuronal Recurrente (RNN): las RNN son uno de los modelos de aprendizaje profundo más utilizados para resolver tareas de NER, debido a su capacidad para procesar secuencias de datos. En cada iteración, las RNN reutilizan todos sus pesos y parámetros, actualizando el estado oculto en función de la entrada actual y del estado anterior [11]. En el ámbito del reconocimiento de entidades biomédicas (BioNER), se han empleado RNN para identificar relaciones complejas entre entidades como proteínas, medicamentos, genes y enfermedades [27]. En el mismo estudio, se presenta un modelo llamado Red Neuronal Convolucional Recurrente (CRNN), que combina CNN y RNN para extraer texto de imágenes de resultados de laboratorios médicos. Esta herramienta está diseñada para facilitar a los profesionales de la salud la revisión de información clínica contenida en imágenes [27].
- LSTM y GRU: Las LSTM están formadas por unidades especiales llamadas celdas de memoria, que permiten almacenar y controlar el flujo de información a través de mecanismos llamados compuertas (de entrada, olvido y salida). La arquitectura de las Unidades Recurrentes con Compuertas (GRU) simplifican la arquitectura de las LSTM al combinar algunas de estas compuertas, lo que las hace más eficientes computacionalmente, aunque con un rendimiento comparable en muchas aplicaciones. Ambas utilizan un mecanismo de compuertas para realizar operaciones relacionadas con la memoria [11].
- Redes Neuronales Convolucionales (CNN): estas redes son ampliamente utilizadas en tareas de clasificación de imágenes y extracción de características, especialmente en el contexto del reconocimiento de documentos. En el estudio, las CNN se aplican a conjuntos de datos como MIDV500, MNIST y otros conjuntos personalizados, con el objetivo de identificar y detectar líneas de texto dentro de imágenes documentales [34], [35].
- BERT (Bidirectional Encoder Representations from Transformers): es un modelo preentrenado de redes neuronales diseñado específicamente para tareas de NLP, aunque también se ha aplicado en problemas de visión por computadora. BERT se entrena inicialmente con grandes volúmenes de texto no anotado, lo que lo convierte en una herramienta eficaz para el aprendizaje por transferencia [36].
- XLNet: es un modelo de lenguaje avanzado que combina elementos de los enfoques autoregresivos y autocodificadores [37]. A diferencia de BERT, que oculta palabras en la entrada y predice esas posiciones usando un contexto bidireccional, XLNet emplea un objetivo de permutación en el entrenamiento. Esto significa que el modelo aprende a predecir tokens considerando múltiples combinaciones posibles del orden de las palabras, lo que le permite capturar mejor las relaciones bidireccionales dentro del texto [11].
- U-Net: desarrollado por Olaf Ronneberger et al., el modelo U-Net fue ideado originalmente para la segmentación de imágenes biomédicas [37], aunque hoy en día se aplica ampliamente en múltiples áreas de visión por computadora. Su arquitectura se basa en una estructura en forma de “U”, compuesta por dos rutas paralelas: el codificador (encoder) y el decodificador (decoder).

D. Reconocimiento de datos (Data recognition)

Con el avance de la tecnología digital se puede almacenar y transferir datos en muchas industrias y prácticamente en todas las actividades cotidianas, la identificación de caracteres impresos por máquina y manuscritos se

ha convertido en un área de investigación importante, ya que existen documentos no estructurados que pueden estar impresos o manuscritos. La adquisición de imágenes, digitalización, preprocesamiento, segmentación, extracción de características y reconocimiento son todas etapas del proceso de reconocimiento de texto manuscrito. Mientras que el reconocimiento de texto impreso por máquina incluye pasos como la adquisición de la imagen del documento escaneado, el preprocesamiento de la imagen y la extracción de características, seguida de la clasificación y el reconocimiento de datos [38].

La fase de procesamiento de datos abarca diversas tareas fundamentales como la eliminación de ruido, corrección de inclinación, binarización, tokenización y stemming, que preparan el contenido textual para su análisis, como también para la extracción de características del documento, se aplican métodos como embeddings de palabras y Bag of Words, que convierten el texto en representaciones numéricas útiles para los modelos de clasificación. Posteriormente, se emplean algoritmos de aprendizaje automático como KNN, SVM y Naïve Bayes [31], junto con modelos de aprendizaje profundo como CNN, RNN, LSTM y GRU, para realizar tareas de clasificación y reconocimiento del contenido textual. Con el progreso de esta área, han surgido enfoques más sofisticados, como las Redes Neuronales Convolucionales Recurrentes (CRNN), métodos basados en transformadores, y combinaciones con técnicas avanzadas de aprendizaje como el metaaprendizaje, few-shot learning y el aprendizaje por transferencia entre otros. La Figura 4 presenta una visión general de estos métodos basados en IA aplicados al reconocimiento y extracción de datos en textos mixtos, detallando sus ventajas y limitaciones respectivas.

Enfoques basados en IA para el procesamiento de documentos no estructurados

En esta sección se describen distintos enfoques fundamentados en IA, analizando sus ventajas y desventajas, en la Figura 5 presenta los principales enfoques basados en inteligencia artificial aplicados al procesamiento de documentos no estructurados. Entre ellos se identifican técnicas como el OCR, RPA, la segmentación semántica, el reconocimiento de entidades nombradas (NER), los grafos de conocimiento y los LLM.

Estos enfoques permiten abordar diferentes etapas del procesamiento documental. Mientras que el OCR facilita la conversión de documentos escaneados en texto digital, técnicas como NER y los LLM permiten avanzar hacia una comprensión más profunda del contenido, identificando entidades, relaciones y significados dentro del texto. Por su parte, la segmentación semántica y los grafos de conocimiento contribuyen a organizar e interpretar la información extraída de documentos complejos.

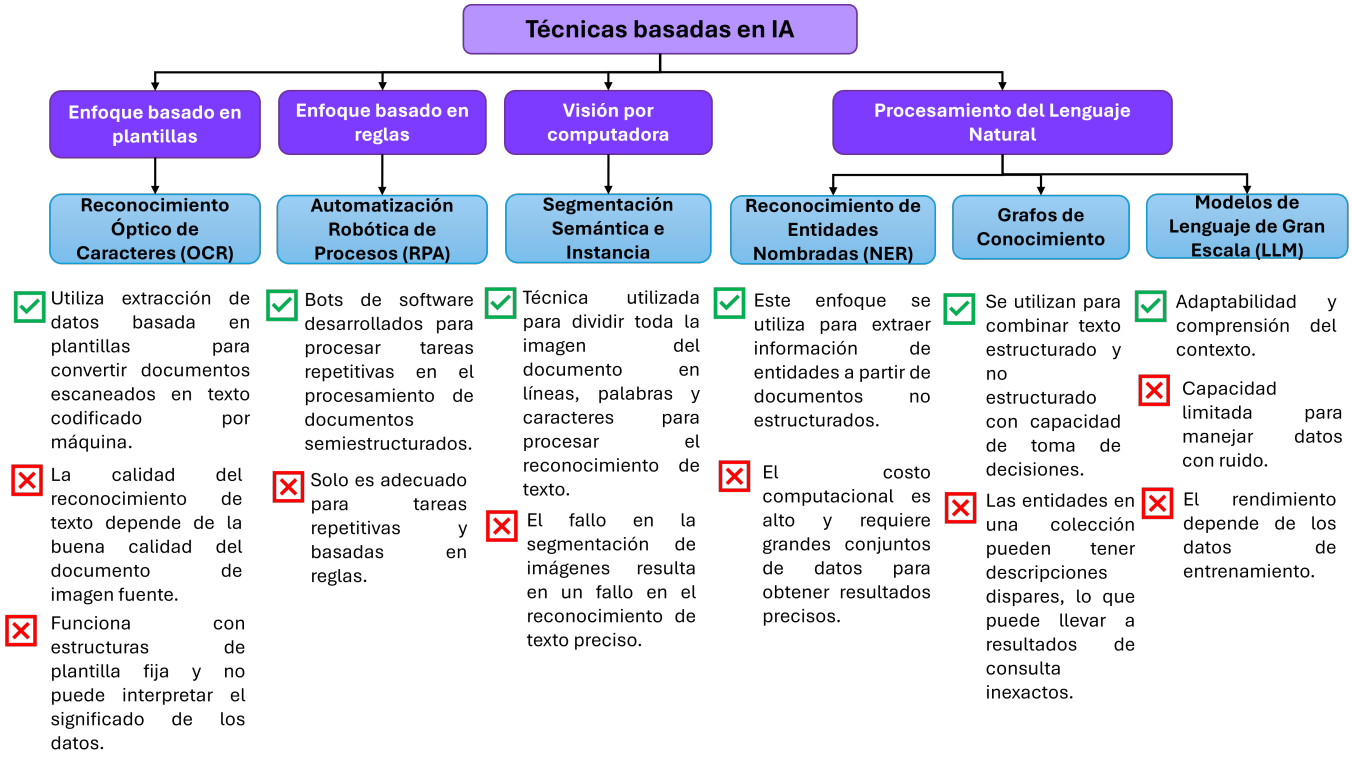

Figura 5: Enfoques de procesamiento de documentos no estructurados mediante IA [11].

La Tabla 3 presenta una comparación de los principales enfoques basados en IA utilizados para el procesamiento de documentos no estructurados. En ella se resumen distintos enfoques ampliamente empleadas en la literatura, como el OCR, la RPA, la segmentación semántica, el NER, los grafos de conocimiento y los Modelos de Lenguaje de Gran Escala (LLM). Para cada enfoque se describen sus características generales, así como sus principales ventajas y desventajas, permitiendo identificar sus capacidades, limitaciones y el contexto en el que resultan más adecuados.

Tabla 3: Descripción de diferentes enfoques de procesamiento de documentos no estructurados mediante IA

Enfoque	**Descripción**	**Ventajas**	**Desventajas**
Reconocimiento Óptico de Caracteres (OCR) [39]	El OCR se emplea para identificar el texto contenido dentro de una imagen, a menudo una copia escaneada de un documento impreso o manuscrito.	▪ Reduce errores en el reconocimiento de datos mediante una entrada eficiente. ▪ Convierte documentos escaneados en formatos editables (texto, Word).	▪ No puede comprender ni analizar los hechos. ▪ Se requiere una imagen fuente de alta calidad para lograr precisión. ▪ Los caracteres que no son texto no pueden ser recuperados.
Automatización Robótica de Procesos (RPA) [40]	El proceso de RPA implica identificar tareas adecuadas para la automatización, diseñar bots que repliquen acciones humanas, implementar la automatización basada en reglas, integrar fuentes de datos, gestionar errores y escalar con múltiples bots.	▪ Automatiza tareas mediante bots de software. ▪ Ejecuta procesos repetitivos, ahorrando tiempo y costos.	▪ Es difícil para la RPA manejar múltiples estilos y formatos de documentos no estructurados. ▪ Requiere el establecimiento de reglas específicas.
Segmentación semántica [41]	Para procesar documentos no estructurados basados en imágenes, la segmentación divide una imagen en segmentos que pueden estar escritos a mano o impresos en formato PDF como texto.	▪ Técnica de segmentación ampliamente utilizada. ▪ No requiere conocimiento visual previo de los objetos.	▪ La misma imagen puede ser segmentada de manera diferente por distintas personas. ▪ La segmentación depende de la percepción y el conocimiento visual humano.

Continúa en la siguiente página...

Tabla 3(continuación)

Enfoque	Descripción	Ventajas	Desventajas
Reconocimiento de Entidades Nombradas (NER) [42], [43]	Es una tecnología que ha avanzado significativamente en el procesamiento de datos en la era digital, particularmente dentro del NLP. NER identifica y clasifica entidades clave (personas, lugares, organizaciones) para mejorar la extracción y recuperación de información.	▪ No requiere plantillas ni reglas predefinidas; aprende a partir de los datos. ▪ Extrae automáticamente entidades de documentos no estructurados.	▪ Requiere grandes conjuntos de datos para un reconocimiento preciso. ▪ El costo computacional es elevado.
Grafos de Conocimiento [44]	La construcción de grafos de conocimiento es compleja cuando los datos no estructurados carecen de etiquetas u ontologías.	▪ Se utiliza para combinar datos tanto estructurados como no estructurados. ▪ Ayuda a las organizaciones a tomar decisiones más responsables y fundamentadas.	▪ Las entidades en una colección pueden tener descripciones dispares, lo que podría conducir a resultados de consulta inexactos.
Modelos de Lenguaje de Gran Escala (LLM) [45]	Los LLM requieren que los documentos (PDF o imágenes) se conviertan primero en texto mediante preprocesamiento.	▪ Permiten comprender el contexto y generar respuestas coherentes a partir de texto no estructurado.	▪ Los LLM tienen límites de tokens y requieren dividir documentos extensos en segmentos.

Ámbitos de aplicación

El reconocimiento y procesamiento de datos provenientes de documentos no estructurados esta en diferentes ámbitos como la gestión empresarial, el marketing, la salud, la educación, el sector financiero, la supervisión pública, gobierno gubernamental entre otros, por lo tanto, existe un potencial significativo en este campo para la extracción y el análisis de datos no estructurados, como se observa en la Tabla 4, en el ámbito de la salud, estas técnicas permiten analizar notas clínicas y apoyar la toma de decisiones médicas. En sectores como seguros y banca, facilitan la automatización de reclamaciones, la verificación de clientes y el análisis de documentos legales o financieros. De manera similar, en servicios gubernamentales y básicos, contribuyen a optimizar trámites administrativos, facturas y procesos internos, como también en documentos agrícolas y turísticos evidencia que estas técnicas también pueden aplicarse en dominios menos tradicionales, como la identificación de plagas, la gestión de reservas o la verificación de datos de pasajeros.

Tabla 4: Ámbitos de aplicación del procesamiento de documentos no estructurados

Ámbito	Aplicación principal
Salud [46]	Análisis de notas clínicas para mejorar la atención médica, optimizar procesos clínicos y apoyar la toma de decisiones mediante modelos basados en IA.
Seguros [47]	Automatización y análisis de reclamaciones para estandarizar procesos, agilizar decisiones y mejorar la orientación al cliente.
Banca [48]	Extracción de información de documentos legales y laborales para la verificación de clientes y el análisis de historiales financieros.

Continúa en la siguiente página...

Tabla 4(continuación)

Ámbito	**Aplicación principal**
Servicios básicos [49]	Digitalización y análisis de facturas para optimizar operaciones internas y mejorar la prestación de servicios a los usuarios.
Servicios gubernamentales [11]	Automatización de trámites administrativos como cambios de dirección y renovación de licencias mediante verificación de datos.
Documentos agrícolas [11]	Extracción de información estructurada a partir de documentos agrícolas para apoyar la identificación de plagas mediante modelos de lenguaje.
Sector turístico [11]	Extracción y verificación automática de información de reservas, boletos y datos de pasajeros a partir de documentos no estructurados.

2. Materiales y métodos

La presente investigación emplea un enfoque analítico, la cual ayudará a estudiar los elementos esenciales de lo investigado, con el fin de interpretar su estructura, significado y objetivo. La búsqueda bibliográfica se realizó siguiendo las fases de identificación, cribado, elegibilidad e inclusión propuestas por PRISMA [13]. El proceso consideró artículos publicados entre 2020 y 2025, escritos en inglés o español, disponibles en texto completo y relacionados con el uso de técnicas de inteligencia artificial para la extracción, reconocimiento, clasificación o análisis de información en documentos no estructurados. Las búsquedas se aplicaron en los campos título, resumen y palabras clave, utilizando operadores booleanos para garantizar la reproducibilidad del proceso.

2.1. Criterios de inclusión/exclusión:

Para garantizar la calidad y relevancia de los estudios seleccionados en la presente revisión sistemática, se definieron criterios de inclusión y exclusión alineados con el objetivo de investigación y la metodología PRISMA. Los criterios fueron aplicados de manera secuencial. En primer lugar, se filtraron los estudios por año de publicación. Posteriormente, se verificó que correspondieran a artículos científicos revisados por pares. En una tercera etapa, se evaluó la relación directa con el procesamiento de documentos no estructurados mediante técnicas de IA. Finalmente, se eliminaron duplicados, documentos sin acceso al texto completo y publicaciones que no estuvieran disponibles en inglés o español.

Los criterios de inclusión consideraron artículos científicos publicados en idioma inglés y español, dentro del periodo comprendido entre 2020 y 2025, que abordan específicamente el procesamiento de documentos no estructurados, en particular archivos PDF, mediante enfoques basados en inteligencia artificial. Además, se priorizaron estudios que presenten resultados relevantes en tareas como extracción de texto, clasificación, reconocimiento y análisis semántico.

Los criterios de exclusión permitieron descartar aquellos estudios que no se relacionan directamente con el tema de investigación, que pertenecen a otros campos de aplicación o que presentan baja relevancia científica. Asimismo, se excluyeron documentos duplicados, artículos incompletos y trabajos que no aportan evidencia suficiente para el análisis. Estos criterios se resumen en la Tabla 5, la cual presenta de manera estructurada las condiciones utilizadas para la selección de los estudios incluidos en la revisión.

Tabla 5: Criterios de Inclusión y Exclusión

Id	**Criterio de exclusión**	**Criterio de inclusión**
CIE1	Publicaciones anteriores a 2020	Periodo de publicación 2020–2025
CIE2	Estudios sin validación empírica o técnica	Artículos científicos y revisados por pares
CIE3	Documentos no académicos (blogs, notas de prensa)	Documentos que aborden la extracción o el análisis de texto en documentos no estructurados mediante técnicas de IA
CIE4	Trabajos duplicados o con acceso restringido	Publicaciones en inglés o español

2.2. Diagrama de flujo PRISMA

El proceso de selección de los estudios incluidos en esta SLR (Revisión Sistemática de la Literatura) fue documentado mediante un diagrama de flujo PRISMA, el cual describe cada etapa: desde la identificación inicial de 2075 registros, la eliminación de duplicados, la aplicación de criterios de inclusión y exclusión, hasta la selección final de 40 artículos.

Durante la fase de cribado, se aplicaron los criterios de inclusión y exclusión previamente definidos, descartando aquellos estudios que no estaban directamente relacionados con el objetivo de la investigación, que no correspondían al periodo de publicación establecido (2020–2025), que no estaban redactados en inglés o español, o que no abordaban el uso de IA en el procesamiento de documentos no estructurados. Asimismo, se excluyeron publicaciones sin revisión por pares, documentos con acceso restringido y trabajos con baja relevancia científica. Como resultado, el número de documentos se redujo a 168 artículos potencialmente relevantes.

En la etapa de evaluación de elegibilidad, se realizó una revisión detallada del texto completo de los 168 artículos, lo que permitió excluir 128 estudios adicionales que no cumplían con los criterios de calidad metodológica, presentaban información redundante o no aportaban resultados significativos para los objetivos de esta SLR.

Finalmente, tras completar este proceso riguroso de selección, se incluyeron 40 artículos que cumplieron con todos los criterios establecidos y fueron considerados para el análisis final. La Figura 6 presenta el diagrama de flujo correspondiente, proporcionando una visión clara y transparente del proceso de selección seguido en esta investigación.

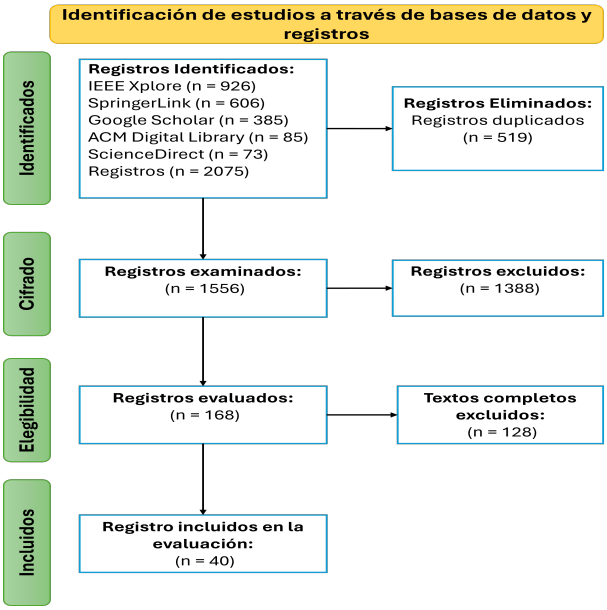

Figura 6: Diagrama de flujo PRISMA

2.3. Estrategias de búsqueda

La búsqueda bibliográfica se desarrolló en el periodo comprendido entre 2020 y 2025 empleando cinco destacadas bases de datos académicas: IEEE Xplore, Springer Link, Google Scholar, ACM Digital Library y Science Direct. La elección de estas plataformas se debe a su amplia cobertura temática y a su capacidad para incluir artículos científicos evaluados mediante revisión por pares.

La estrategia de búsqueda se estructuró a partir de una combinación de términos clave como “Artificial intelligence”, “large language model”, “text extraction”, “unstructured document processing”, “document management”, “handwritten text recognition”, “Mixed printed & handwritten text recognition”, “deep learning”, “AI”, “document automation”, “text analysis”, “machine learning”, “document processing”, “unstructured document processing”, “document digitization”, “neural networks”, “unstructured document processing”, “natural language processing”, “document understanding” y “printed and handwritten text recognition”. Una vez obtenido los términos relevantes se combinaron

con operadores booleanos (AND, OR) para garantizar una búsqueda exhaustiva.

Para garantizar una búsqueda exhaustiva y precisa se formularon ecuaciones de búsqueda específicas para cada una de las bases de datos seleccionadas, dichas ecuaciones integran los principales términos vinculados con técnicas de inteligencia artificial aplicadas a la extracción de información en documentos no estructurados, optimizando su alcance mediante el uso estratégico de operadores booleanos. En la Tabla 6 se presentan los ejemplos de estas ecuaciones en idioma inglés.

Tabla 6: Base de Datos científicas en línea

Base de datos académica	**Ecuaciones de búsqueda**
Google Scholar	(“Artificial intelligence” OR “large language model”) AND (“text extraction” OR “unstructured document processing”) AND (“document management” OR “handwritten text recognition”)
IEEE Xplore	(“Mixed printed & handwritten text recognition” AND “deep learning”) OR (“AI” AND “document automation”)
SpringerLink	(“text analysis” AND “machine learning”) AND (“document processing” OR “unstructured document processing”)
ScienceDirect	(“document digitization” AND “neural networks”) OR “unstructured document processing”)
ACM Digital Library	(“natural language processing” AND “document understanding” AND “printed and handwritten text recognition”)

2.4. Estudios encontrados

La Figura 7 muestra la distribución de los estudios identificados en las diferentes bases de datos consultadas, alcanzando un total de 2.075 registros iniciales antes de la aplicación de los criterios de inclusión y exclusión. Se observa una mayor concentración de estudios en IEEE Xplore, con 926 registros, seguida por SpringerLink con 606 y Google Scholar con 385, mientras que ACM Digital Library y ScienceDirect presentan una menor cantidad de resultados. Esta distribución evidencia una predominancia de fuentes especializadas en ingeniería y tecnologías de la información, lo cual es coherente con la naturaleza del tema de estudio. En particular, IEEE Xplore destaca como la principal fuente de publicaciones relevantes, lo que sugiere un alto nivel de producción científica en el área de procesamiento de documentos no estructurados dentro de esta plataforma.

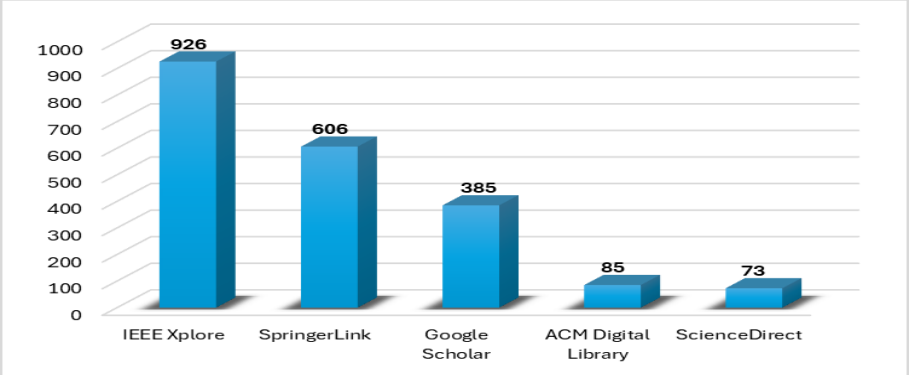

Figura 7: Exploración inicial de artículos

2.5. Selección de estudios

La búsqueda inicial arrojó un total de 2075 artículos. De estos, IEEE Xplore aportó 926, SpringerLink proporcionó 606, Google Scholar identificó 385, ACM Digital Library brindo 85 y ScienceDirect registró 73 documentos, aplicando de manera secuencial los criterios de inclusión y exclusión (CIE1–CIE4) definidos para esta Revisión Sistemática de la Literatura. CIE1 corresponde al número de estudios que fueron excluidos por no cumplir con el periodo

de publicación establecido (2020–2025), CIE2 representa los artículos descartados por no contar con validación empírica o técnica, o por no tratarse de publicaciones científicas revisadas por pares, CIE3 indica los documentos eliminados por no abordar de manera directa el uso de IA en la extracción o análisis de texto en documentos no estructurados, o por corresponder a fuentes no académicas. Finalmente, CIE4 recoge los estudios excluidos por duplicidad, acceso restringido al texto completo o por no estar redactados en inglés o español . Tras la aplicación progresiva de estos criterios, se obtuvo un conjunto final de 40 artículos, los cuales cumplieron con todos los requisitos metodológicos y fueron seleccionados para el análisis detallado en esta investigación en la Tabla 7 se muestra los resultados de la búsqueda de la información.

Este procedimiento permitió garantizar la trazabilidad del proceso de selección, desde la identificación inicial de estudios hasta la inclusión final de los artículos analizados. La aplicación secuencial de los criterios de inclusión y exclusión permitió reducir sesgos, eliminar documentos no pertinentes y asegurar que los estudios seleccionados aportaran evidencia relevante para responder las preguntas de investigación.

Tabla 7: Resultados de la búsqueda de información sobre criterios de inclusión y exclusión

Fuente	Total	CIE1	CIE2	CIE3	CIE4
IEEE Xplore	926	747	58	29	29
SpringerLink	606	331	50	6	6
Google Scholar	385	359	33	3	3
ACM Digital Library	85	65	4	0	0
ScienceDirect	73	54	23	2	2
Total	**2075**	**1556**	**168**	**40**	**40**

2.6. Resumen de los trabajos de la revisión

Mediante una metodología de cuatro etapas y criterios de selección, se realizó un proceso de filtrado de artículos académico. Se identificaron 2075 trabajos, con una mayor concentración en IEEE Xplore, seguido por SpringerLink y Google Scholar, como es en la Figura 8, se observa que la mayoría de los estudios se ubican en el nivel CIE1, lo que indica una alta cantidad de documentos iniciales sin aplicar criterios estrictos de selección. A medida que se aplican filtros más rigurosos, representados por los niveles CIE2, CIE3 y CIE4, el número de estudios se reduce progresivamente, evidenciando un proceso de depuración enfocado en la calidad y relevancia de la información.

Al aplicar todos los criterios de selección se obtuvo 40 papers. En su porcentaje más significativo tenemos 72.50 % de IEEE, así como SpringerLink 15 %, Google Scholar 7.50 %, ACM Digital Library 0.0 % y ScienceDirect 5 %. Esta distribución refleja la predominancia de fuentes especializadas en el área tecnológica y respalda la rigurosidad metodológica aplicada en la selección de estudios.

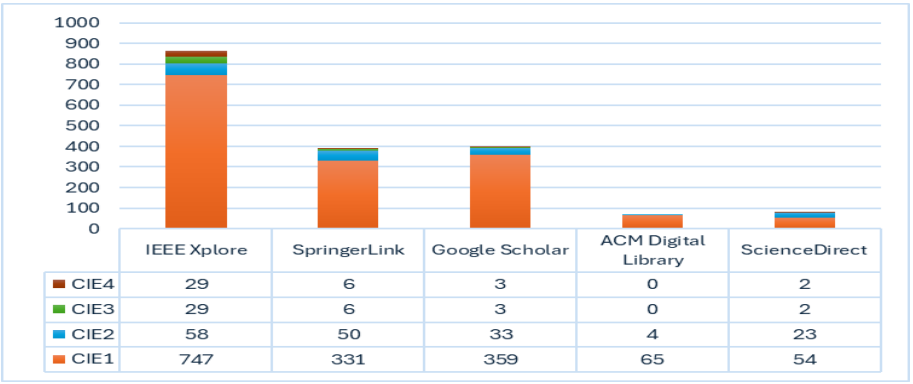

Figura 8: Resultado de la búsqueda según los criterios de selección

3. Resultados

Con el fin de analizar en profundidad los resultados obtenidos de la revisión sistemática, se presentan a continuación varias tablas comparativas que sintetizan las técnicas y enfoques utilizados en el reconocimiento y la extracción de información a partir de documentos no estructurados. Estas tablas permiten identificar patrones recurrentes, tecnologías predominantes, dominios de aplicación y desafíos técnicos reportados en la literatura reciente.

La Tabla 8 sintetiza los enfoques más utilizados en el procesamiento de documentos no estructurados donde se evidencia que el OCR continúa siendo una técnica fundamental, especialmente en tareas de digitalización y reconocimiento inicial de texto; sin embargo, presenta limitaciones en la comprensión del contenido semántico. En contraste, enfoques como NER y los LLM permiten avanzar hacia una interpretación más profunda de la información, facilitando la extracción de datos relevantes en dominios complejos como el sector legal, médico y científico, mientras que la segmentación semántica y los grafos de conocimiento permiten estructurar y organizar la información extraída, lo que resulta clave para el análisis de documentos con formatos heterogéneos. Por su parte, la RPA complementa estos enfoques al permitir la ejecución automatizada de tareas repetitivas en entornos organizacionales.

Tabla 8: Resumen de los enfoques y sus áreas de aplicación en el reconocimiento y la extracción de información en el procesamiento de documentos no estructurados

Enfoque utilizado	**Casos de uso de procesamiento de documentos no estructurados**
Reconocimiento Óptico de Caracteres (OCR) [4], [50], [41], [51], [52], [53], [54], [55], [56]	▪ Digitalización de documentos académicos. ▪ Reconocimiento de datos de recibos/facturas. ▪ Reconocimiento y clasificación de texto de documentos con imágenes escaneadas, tanto manuscritas como impresas. ▪ Reconocimiento de datos de documentos históricos. ▪ Extracción de datos de artículos científicos.
Automatización Robótica de Procesos (RPA) [4] , [57], [58]	▪ Automatización de registros bancarios y financieros. ▪ Automatización de historiales médicos de pacientes. ▪ Automatización de procesos de RR.HH. en una organización.
Reconocimiento de Entidades Nombradas (NER) [59], [60], [61], [62], [63] , [64]	▪ Extracción de datos de facturas ▪ Extracción de información de notas clínicas ▪ Extracción de datos de cláusulas legales ▪ Extracción de metadatos de artículos científicos
Segmentación semántica [41], [65], [66], [67], [68]	▪ Reconocimiento y extracción de datos de formularios ▪ Documentos basados en imágenes, texto e identificación de firmas
Grafos de Conocimiento [69], [44] , [70]	▪ Digitalización de registros clínicos ▪ Ciberseguridad para el conocimiento de ataques ▪ Análisis de redes ▪ Bioinformática

Continúa en la siguiente página...

Tabla 8(continuación)

Enfoque utilizado	Casos de uso de procesamiento de documentos no estructurados
Modelos de Lenguaje de Gran Escala (LLM) [45]	▪ Extracción de información de documentos agrícolas ▪ Comentarios de clientes

Como se observa en la Tabla 9, los sistemas de reconocimiento de texto manuscrito se aplican en diversos contextos, tales como facturas, formularios, documentos multilingües y registros escritos por múltiples usuarios. Se evidencia que los enfoques más utilizados combinan técnicas tradicionales como OCR con modelos basados en aprendizaje profundo, particularmente redes neuronales convolucionales (CNN).

Se identifica una tendencia hacia el uso de modelos más avanzados capaces de mejorar la precisión en el reconocimiento de texto manuscrito; sin embargo, estos sistemas aún enfrentan desafíos importantes. Entre las principales limitaciones se destacan la dificultad para reconocer escritura cursiva, la variabilidad en los estilos de escritura, la baja calidad de los datos de entrada y la dependencia de conjuntos de datos específicos para el entrenamiento de los modelos.

Asimismo, se observa que, aunque los modelos basados en aprendizaje profundo ofrecen mejores resultados en comparación con los métodos tradicionales, su rendimiento puede verse afectado por la falta de datos representativos y la complejidad del lenguaje, especialmente en casos como la escritura árabe o documentos con múltiples formatos.

Tabla 9: Resumen de sistemas de reconocimiento de texto manuscrito según áreas de aplicación, descripción, métodos utilizados y desafíos

Área de aplicación	Descripción	Técnicas o métodos utilizados	Desafíos
Facturas [22]	Se desarrolló un enfoque para identificar texto en facturas impresas.	OCR, Adaptive character recognition, Tesseract	Reconoce caracteres manuscritos de forma imprecisa.
Imágenes de formularios manuscritos [71]	Modelo propuesto basado en segmentación de imágenes para el reconocimiento de caracteres manuscritos.	CNN, Open CV, Neural Networks	El sistema no reconoce la escritura cursiva.
Imágenes de texto impreso/mecanografiado [72]	Transformer-based Visual Document Understanding (VDU)	Otsu's Algorithm for segmentation	Escritura cursiva árabe
Documentos escritos a mano por varios usuarios [73]	Técnica para clasificar texto manuscrito e impreso automáticamente en ICR desarrollada	OCR & ICR	La escritura manuscrita offline
Formularios rellenados a mano [16]	Convierte datos de formularios en texto procesable por computadora.	ANN, CNN, Pytessaract, Open CV2	No predice la casilla vacía sin texto escrito.
Reconocimiento de caracteres árabes [74]	Se desarrolló un modelo para reconocer caracteres árabes en documentos.	CNN–LSTM–CTC	La precisión se puede mejorar modificando el algoritmo CNN.
Formularios de reclamos de seguros de automóviles [38]	Método basado en anclajes y MCNN para digitalizar formularios manuscritos.	MCNN- Multichannel CNN, Anchor based object detection	Bajo rendimiento al entrenar solo con el conjunto de datos EMNIST.

Como se observa en la Tabla 10, los enfoques basados en inteligencia artificial, especialmente aquellos que emplean arquitecturas de aprendizaje profundo y modelos de lenguaje, presentan los mejores resultados en términos de precisión y capacidad de generalización. En particular, modelos como StrucTexT, BERT, RoBERTa y

combinación de técnicas basadas en CNN–LSTM alcanzan valores de accuracy y F1-score superiores al 90 % en diversos dominios.

Se evidencia que las técnicas tradicionales, como OCR, siguen siendo utilizadas como etapa inicial para la digitalización de texto; sin embargo, su rendimiento mejora significativamente cuando se combinan con modelos de aprendizaje profundo y métodos de extracción de características avanzados. Asimismo, los enfoques basados en visión por computadora, como YOLO y Faster R-CNN, resultan especialmente efectivos en el procesamiento de documentos visualmente complejos.

Por otro lado, se identifican variaciones en el rendimiento dependiendo del dominio de aplicación, siendo más altos en documentos estructurados, como facturas y formularios, y más desafiantes en contextos como el lenguaje biomédico o documentos multilingües. Además, algunos modelos presentan limitaciones asociadas a la disponibilidad de datos de entrenamiento y a la complejidad del lenguaje.

Se puede decir que el uso de integración de enfoques que combinan procesamiento de lenguaje natural, visión por computadora y aprendizaje profundo, lo que permite mejorar significativamente la precisión y eficiencia en la extracción de información a partir de documentos no estructurados.

La Tabla 11 presenta una síntesis de los enfoques utilizados en el procesamiento de documentos no estructurados, integrando las distintas etapas del flujo de trabajo, desde el preprocesamiento hasta los modelos de aprendizaje automático y profundo. Se observa que las técnicas de preprocesamiento, como la binarización, la eliminación de ruido, la normalización y la tokenización, constituyen una fase fundamental para mejorar la calidad de los datos antes de su análisis.

Tabla 10: reconocimiento y extracción de texto impreso por máquina en diferentes áreas de aplicación, con técnicas, métodos de extracción utilizados, dominio y rendimiento

Técnicas utilizadas	**Métodos de extracción de características utilizados**	**Dominio**	**Rendimiento**
StrucTexT [61]		Reconocimiento de facturas, documentos estructurados	Accuracy: 94–95 %
NER, Bi-LSTM-CRF, BERT [75]	Word2Vec, BERT	Análisis de textos biomédicos y clínicos	F-score: 88.80 %
YOLOv5s, PaddleOCR [76]	CNN	Procesamiento de documentos logísticos	Accuracy: 89.5 %, F1: 90.1 %
OCR, MSER, CNN, CRNN, LSTM [77]	MSER, CNN, LSTM	Documentos financieros, facturas de proveedores	Accuracy: 95 %
Bio-NER, RNN + CRF [78]	Word embeddings (Word2Vec, GloVe)	Historias clínicas electrónicas, texto biomédico	F1-score: 88.34 %
BERT, RoBERTa [79]		Información del contrato	F1-Score: 58 %
CNN, BiGRU, BiLSTM [80]	CNN, LSTM/GRU, embeddings Word2Vec y FastText	Salud / Clínico	F1: 0.88
BI-LSTM, CNN, hybrid [12]	Word2Vec	Secuencia de palabras	F1:0.90 Precision:0.90 Recall:0.89
CNN [81]	CNN	Análisis de Imágenes de Documentos (DIA)	CLMID Precision:89.78 F1-Score:0.89 ICDAR 2015 Precision:94.15 F1-Score:0.92
OCR, CRF, NER [82]		Informes de laboratorio	Accuracy: 0.93 F1-score: 0.86
Faster–RCNN [83]		Negocio financiero	Precision-66.80 %
SVM, LR,RF [84]	TF-IDF	Bases de datos NoSQL	SVM-80 % RF-78 %

Continúa en la siguiente página...

Tabla 10(continuación)

Técnicas utilizadas	Métodos de extracción de características utilizados	Dominio	Rendimiento
RPA,mBERT, XLMRO-BERTa, InfoXLM, Layout XLM [85]		Financiera	F1 Score-BRC-97 % BL-81 %
Shelf OCR engine, Bi-directional attention complementation mecha-nism(BiACM) [86]		Documentos impresos escanea-dos	FUNSD Precision—0.87 Recall—0.89 F1—0.88 RVL-CDIP Accuracy-96.17 %
Faster R-CNN [87]	Incrustación de imágenes y texto	Documentos impresos escanea-dos	FUNSD Precision—76 % Recall—81 % F1—79 % SROIE Precision—95 % Recall—95 % F1—95 %
Transformer-based [88] Visual Document Unders-tanding (VDU)		Documentos impresos, recibos, tickets, tarjetas de presentación	RVL-CDIP Accuracy- 95.30 %
Masked Detection Mo-deling (MDM) [89] Extension of masked vision-language modeling (MVLM) Visual Document Understanding (VDU)	Vinculación de Entidades mediante Asociación de Palabras Ancla	Extracción de información es-tructurada de documentos vi-sualmente enriquecidos	CORD-Accuracy—94 % FUNSD-Accuracy—84 %

En cuanto a la extracción de características, se evidencia el uso de métodos tanto tradicionales como basados en representaciones vectoriales, destacando técnicas como Bag of Words y Word2Vec, así como enfoques más avanzados como BERT. Estos métodos permiten transformar el contenido textual en representaciones que facilitan su procesamiento por modelos de aprendizaje.

Se identifica una clara predominancia de los enfoques de aprendizaje profundo, especialmente redes neuronales convolucionales (CNN), redes recurrentes (RNN), arquitecturas LSTM y modelos basados en transformadores como BERT, los cuales ofrecen mejores resultados en comparación con los métodos tradicionales de aprendizaje automático, como SVM y KNN. Esta tendencia se debe a la capacidad de los modelos profundos para capturar patrones complejos y relaciones semánticas en datos no estructurados.

Tabla 11: Descripción general de los enfoques utilizados en el procesamiento de documentos no estructurados, destacando las distintas técnicas aplicadas.

Tipo de información	Técnicas de preprocesamiento	Técnicas de extracción de características	Enfoques de aprendizaje automático	Enfoques de aprendizaje profundo
Texto impreso [12]	Binarización y Normalización	Word2Vec		CNN, Bi-LSTM
Texto manuscrito [38]	Eliminación de ruido e inclinación	Bolsa de palabras (BoW, Bag of words)		MCRNN

Continúa en la siguiente página...

Tabla 11(continuación)

Tipo de información	Técnicas de preprocesamiento	Técnicas de extracción de características	Enfoques de aprendizaje automático	Enfoques de aprendizaje profundo
Texto manuscrito [90]		CNN	SVM	CNN
Texto manuscrito [16]	Binarización, Eliminación de ruido, Alineación morfológica o angular			CNN
Texto manuscrito [91]	Binarización, Eliminación de ruido, Normalización		SVM	CNN
Texto impreso [92]	Tokenización, Eliminación de palabras vacías, Normalización	Bolsa de palabras (BoW, Bag of words)	KNN	
Texto manuscrito [93]	Eliminación de ruido, normalización		KNN	
Texto impreso [27]	Eliminación de ruido, normalización	BERT		RNN,Bi-LSTM,VGG16
Texto impreso y manuscrito [41]	Eliminación de ruido, normalización, eliminación de inclinación			MCRNN, Unet
Texto impreso [94]	Eliminación de ruido, normalización		SVM, KNN y SVM	CNN
Texto impreso [95]	Tokenización, normalización	BERT, Word2Vec	SVM	CRF, LSTM, BERT
Texto impreso [96]		Word2Vec		RNN, BERT
Texto impreso [35]		Word2Vec		CNN, RNN
Texto impreso [97]	Tokenización, normalización			LSTM, CRF
Texto impreso [98]	Tokenización	BERT	CRF	Bi-LSTM-CRF,BERT
Texto manuscrito [99]	Tokenización, eliminación de inclinación	BERT	CRF	CNN, LSTM
Texto impreso [100]	Eliminación de ruido	BERT	SVM	CNN, BERT, Bi-LSTM

4. Discusión

Q1. ¿Cuáles son los enfoques basados en IA utilizados en el reconocimiento de datos de documentos no estructurados?

Los documentos no estructurados representa un desafío considerable debido a su complejidad, costo, consumo de tiempo y propensión a errores. Estos documentos pueden presentarse en formatos impresos o manuscritos, puedes ser como formularios, documentos financieros, historiales médicos, registros académicos entre otros. A pesar de su estructura irregular, contienen información valiosa que puede aprovecharse mediante técnicas adecuadas de extracción de datos. Esta SLR analiza diversos enfoques utilizados para extraer información de documentos no estructurados, comparando sus ventajas y desventajas según el dominio y tipo de tarea a pesar de que estos documentos pueden tener estructuras muy variadas y complejas, no existe un método universal de extracción que funcione en todos los contextos, lo cual representa uno de los principales retos actuales en la investigación.

Los métodos tradicionales para la extracción de datos no estructurados se han basado principalmente en enfoques basados en reglas o plantillas, los cuales son lentos y costosos de implementar, por ello las organizaciones

han comenzado a buscar una solución donde se pueda usar la IA, que combinan algoritmos de CV, NLP y herramientas de automatización como RPA u OCR, logrando así procesos más eficientes y automatizados [101]. Ademas es necesario mencionar que los enfoques basados en reglas RPA, enfoques basados en plantillas, OCR, enfoques basados en PLN como NER, grafos de conocimiento y enfoques basados en segmentación semántica utilizados para el reconocimiento y la clasificación de datos mixtos se mencionan en la tabla 8. Modelos de ML y DL, junto con arquitecturas preentrenadas, han demostrado ser altamente efectivos en la automatización de la extracción de datos, especialmente en sectores como la banca, la salud y las finanzas, en las tablas 9, 10 muestran los trabajos existentes que usan dichos modelos. Un desafío clave en el uso de CV es extraer de forma automática información precisa a partir de documentos con estructuras no convencionales. Las técnicas automatizadas de extracción de información se utilizan en una variedad de dominios de aplicación con distintos objetivos [102].

Recientemente, los métodos de aprendizaje profundo han cobrado mayor relevancia, destacándose por su capacidad de aprender automáticamente representaciones complejas y realizar tareas completas de forma autónoma, desde la extracción de características hasta la clasificación final. No obstante, persisten desafíos, como la elección de la arquitectura de red más adecuada, la definición de la función de pérdida óptima y la necesidad de contar con grandes volúmenes de datos de entrenamiento.

Q2. ¿Cuáles son los dominios y aplicaciones que utilizan el procesamiento de datos no estructurados?

Actualmente, algunos sectores generan diariamente grandes volúmenes de documentos no estructurados por lo que se ha vuelto una necesidad clave para procesar esta información. En este estudio se detallan los distintos campos donde este tipo de procesamiento es fundamental, junto con una revisión de los trabajos de investigación existentes, en la tabla 9 se resume los estudios actuales desarrollados en diversos ámbitos de aplicación.

En el caso del sector financiero, se maneja una enorme cantidad de datos, que incluye información de clientes, registros de productos financieros, transacciones y datos externos provenientes de redes sociales o sitios web, todos ellos útiles para respaldar procesos de toma de decisiones, como también, en el ámbito de la salud, la mayor parte de la información se encuentra en formatos no estructurados, como historiales médicos, formularios de ingreso hospitalario, resúmenes de alta y prescripciones, tal como se menciona en la tabla 4. A pesar de la creciente necesidad, el reconocimiento de texto mixto en diferentes dominios sigue siendo un área poco explorada. Esto abre un amplio campo para la investigación en sectores como salud, finanzas, seguros, banca, mercadeo, mercados, educación, entre otros. Los estudios actuales aún no han propuesto un marco sólido o una solución eficaz para abordar la digitalización de documentos no estructurados y multimodales, lo que pone en evidencia el enorme potencial de investigación que permanece disponible para quienes trabajan en este campo.

Q3. ¿Cuáles son las diversas técnicas de preprocesamiento de datos asociadas con el procesamiento de documentos no estructurados?

Cada día se generan grandes volúmenes de documentos no estructurados en múltiples sectores, ya que si estos datos se recuperan y analizan adecuadamente, pueden ofrecer un valor significativo para las organizaciones, ayudándolas a tomar decisiones estratégicas basadas en información fundamentada. Tras la fase de recolección o generación de datos, se lleva a cabo una serie de operaciones de preprocesamiento que preparan el contenido para su análisis posterior, de las cuales entre las más comunes se encuentran la eliminación de palabras vacías (stopwords), la conversión a minúsculas, la supresión de caracteres no alfanuméricos, la eliminación de filas vacías y la corrección de palabras unidas o mal segmentadas. Asimismo, la falta de conjuntos de datos públicos y multilingües limita el desarrollo de estrategias de preprocesamiento robustas para documentos en idiomas distintos del inglés, especialmente el español, que es predominante en América Latina.

Un desafío actual en este campo es la creación de conjuntos de datos de alta calidad, diseñados mediante técnicas avanzadas de preprocesamiento y compartidos abiertamente con la comunidad investigadora. Este estudio ofrece un análisis detallado de las distintas estrategias de preprocesamiento, respaldado por literatura especializada. Dado el potencial de mejora en esta área, existen numerosas oportunidades para que investigadores impulsen el desarrollo de nuevas técnicas de preprocesamiento más eficaces, debido a que este proceso resulta esencial para manejar documentos no estructurados a gran escala, ya que al acceder a conjuntos de datos bien estructurados y limpios no solo mejora la calidad de los resultados, sino que también facilita el avance de investigaciones más precisas y escalables.

Los hallazgos de esta revisión sugieren que las organizaciones latinoamericanas que buscan implementar soluciones de extracción automática de información deben priorizar enfoques escalables, independientes de plantillas y con bajo costo computacional, así como fomentar la creación de repositorios abiertos de datos documentales.

La colaboración entre academias, sector público y privado emerge como un factor clave para reducir las brechas identificadas y acelerar la adopción de tecnologías basadas en IA para la gestión eficiente de documentos no estructurados.

Q4. ¿Cuáles son los principales desafíos que enfrentan los algoritmos de extracción de información al procesar documentos no estructurados?

Con base en las revisiones realizadas en este estudio se identificó que existen grandes problemas y desafíos en el procesamiento de documentos no estructurados, ya que las técnicas existentes no reconocen al 100 % los múltiples tipos de modalidades de datos en documentos no estructurados, dado que no existe muchas investigaciones sobre métodos basados en IA para extraer automáticamente la información de dichos documentos.

Resulta difícil extraer información de documentos escritos a mano porque se conoce que cada persona tiene su propio estilo de escritura y muchos de ellos contemplan diferentes tipos de estilo de escritura lo cual son con letra desordenada que hasta pueda dar textos confusos. En la investigación realizada se determinó que los algoritmos tienen mayor dificultad para identificar la letra por el reconocimiento de escritura, puede correr en el riesgo de interpretar en diferentes datos.

Las principales limitaciones identificadas en el procesamiento de documentos no estructurados son las siguientes:

- Reconocimiento de datos: El tratamiento de datos multimodales que incluyen texto manuscrito, texto impreso, símbolos, entre otros sigue siendo un desafío poco explorado, especialmente en documentos semiestructurados. Esta diversidad puede afectar la precisión al reconocer información en formularios u otros documentos.
- Ingreso manual de datos: Muchas industrias aún dependen de la digitación manual de textos manuscritos, debido a la falta de soluciones automatizadas eficaces para su digitalización.
- Problemas de orientación: Las tecnologías OCR suelen enfrentar dificultades cuando los documentos escaneados presentan inclinaciones o están mal orientados, lo que complica su correcta interpretación.
- Inconsistencias en el texto: Los textos manuscritos a menudo presentan palabras desordenadas y caligrafía irregular, lo que dificulta su lectura automatizada. Para mejorar la comprensión del contenido, es posible combinar técnicas OCR con algoritmos de PLN y ML.
- Documentos borrosos: Al capturar imágenes con cámaras móviles o escáner sea documentos como formularios médicos, solicitudes gubernamentales y otros formatos semiestructurados pueden presentar reflejos que comprometen la precisión en la extracción de datos [103].
- Datos no procesados: A diario se generan grandes volúmenes de datos no estructurados que no son analizados. Si se procesaran y resumieran correctamente, podrían aportar información valiosa para la toma de decisiones empresariales.

5. Conclusiones

Con el objetivo de identificar áreas con potencial de mejora, este SLR estudia los enfoques actuales utilizados para la extracción de información en el procesamiento de documentos no estructurados donde el método de búsqueda bibliográfica se basó en el enfoque PRISMA utilizando los lineamientos propuestos por [13]. Como resultado del proceso de selección que consideró criterios de inclusión, exclusión y evaluación de calidad se identificaron 40 artículos relevantes que abordan los temas centrales del estudio, dichos artículos se recopilaron de los últimos cinco años, lo cual refleja tanto el creciente interés como el desarrollo sostenido en esta área de investigación.

Los principales resultados de la revisión, se evidencian en las Tablas 8, 9,10 y 11, la cual muestran los enfoques más utilizados que corresponden a tecnologías como Reconocimiento óptico de caracteres (OCR), Automatización Robótica de Procesos (RPA), reconocimiento de entidades nombradas (NER), métodos de segmentación, aprendizaje automático, aprendizaje profundo y modelos de lenguaje a gran escala (LLM), entonces se puede decir que OCR continúa siendo la base para la digitalización de documentos, mientras que modelos más avanzados como CNN, LSTM, BERT y transformers permiten mejorar significativamente la precisión en tareas de reconocimiento, clasificación y análisis semántico, tal como se evidencia en diversos estudios [81, 75, 88].

En base en los resultados obtenidos, queda claro que estos métodos tienen amplias aplicaciones en los campos legal, financiero y médico, donde la automatización del procesamiento de documentos ayuda a mejorar la eficiencia operativa y la toma de decisiones [54, 77, 82, 22], siempre y cuando se considere que el rendimiento de estas soluciones depende de factores como la calidad de los datos, el tipo de documento y las técnicas empleadas, lo que se refleja en la variabilidad de métricas como accuracy, precision, recall y F1-score reportadas en la literatura.

El estudio destaca áreas clave que requieren más investigación, especialmente en tecnologías como OCR, RPA, NER, métodos de segmentación y modelos de lenguaje a gran escala, pero es importante aclarar que estas líneas de trabajo son fundamentales para avanzar hacia una extracción autónoma de información más robusta y eficiente, es decir que aún se requieren avances importantes para abordar con éxito la complejidad y variabilidad de los documentos no estructurados, fundamentalmente en escenarios con datos ruidosos, escritura manuscrita y estructuras heterogéneas.

Finalmente, se concluye que, aunque la automatización en la extracción de datos ha comenzado a implementarse en varios sectores, aún es necesario desarrollar enfoques más integrados, robustos y reproducibles, así como establecer métricas estandarizadas y conjuntos de datos públicos que permitan evaluar de manera consistente el desempeño de las técnicas propuestas.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Declaración de contribución de autores — CRediT

Conceptualización: J.G.G., P.L.L., Z.I.R.C.

Metodología: J.G.G., P.L.L.

Validación: P.L.L., Z.I.R.C.

Análisis formal: J.G.

Investigación: J.G.

Curación de datos: J.G.

Redacción — borrador original: J.G.

Redacción — revisión y edición: J.G.G., P.L.L., Z.I.R.C.

Visualización: J.G.G.

Supervisión: P.L.L., Z.I.R.C.

Administración de proyectos: J.G.G., P.L.L.

Financiación

No aplica.

Reseñas de autores

Johanna Guamán Granda

Es profesional en Tecnologías de la Información por la Universidad Técnica del Norte (UTN), Ecuador. Su formación e intereses académicos se orientan al área de las tecnologías de la información, inteligencia artificial y gestión documental. Sus líneas de interés incluyen inteligencia artificial generativa, reconocimiento óptico de caracteres (OCR), procesamiento del lenguaje natural y extracción automatizada de información de documentos no estructurados. Mantiene afiliación académica con la Universidad Técnica del Norte.

Correo: anavigail1993@gmail.com

ORCID: 0009-0008-6588-9812

Pablo Andrés Landeta López

Ingeniero en Sistemas Computacionales y Magíster en Gerencia de Sistemas, docente de la Universidad Técnica del Norte (UTN), Ecuador. Su actividad académica e investigativa se desarrolla en ciencias de la computación y tecnologías de la información. Sus líneas de interés incluyen inteligencia artificial, modelos de lenguaje de gran escala, procesamiento del lenguaje natural, computación en la nube, ingeniería y desarrollo de software, experimentación computacional y optimización. Es integrante del grupo de investigación GISIA de la Universidad Técnica del Norte.

Correo: palandeta@utn.edu.ec

ORCID: 0000-0002-2914-8696

Zoila Irene Rubio Clavijo

Profesional vinculada a la Unidad Educativa 17 de Julio, Ecuador, con interés en el ámbito educativo, las tecnologías de la información y su aplicación en procesos de investigación. Sus intereses académicos comprenden el uso de herramientas digitales, inteligencia artificial, procesamiento de información y aplicación de nuevas tecnologías en contextos educativos y de investigación.

Correo: irerubio12@outlook.es

ORCID: 0009-0007-7807-0825

Declaración sobre uso de Inteligencia Artificial

Sí se utilizó inteligencia artificial como apoyo técnico o complementario.

Se declaró su utilización para elaboración o redacción del manuscrito, corrección gramatical, sintáctica o estilística y sugerencias de estructura general del texto.

Herramienta empleada: ChatGPT.

Referencias

[1] S. Shilaskar, S. Chakole, A. Samarth, and P. Shejole, “GenAI based Data Extraction with Query-based Insights”, ESCI, 2025, doi: 10.1109/ESCI63694.2025.10987906.

[2] F. Shahid, M. H. Hsu, Y. C. Chang, and W. S. Jian, “Using Generative AI to Extract Structured Information from Free Text Pathology Reports", Journal of Medical Systems, vol. 49, no. 1, pp. 1-16, 2025, doi: 10.1007/s10916-025-02167-2. [Online]. Available: https://link.springer.com/article/10.1007/s10916-025-02167-2.

[3] M. K. Ugale, S. J. Patil, and V. B. Musande, “Document management system: A notion towards paperless office", ICISIM, vol. 2017-January, pp. 217-224, 2017, doi: 10.1109/ICISIM.2017.8122176.

[4] D. Baviskar, S. Ahirrao, V. Potdar, and K. Kotecha, “Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions", IEEE Access, vol. 9, pp. 72894-72936, 2021, doi: 10.1109/ACCESS.2021.3072900. [Online]. Available: https://ieeexplore.ieee.org/document/9402739.

[5] S. D. O. Jiménez, A. Z. Galindo, and G. V. Álvarez, “Paperless Office: a new proposal for organizations", Academic Journals Database, and Google Scholar, vol. 3, pp. 47-55, 2015. [Online]. Available: https://www.iiisci.org/journal/sci/FullText.asp?var=&id=HA544MP15.

[6] J. Vassallo, A. M. Lopez, and G. Urchipia, “Uso de IAG en Gestión Documental para Implementación de la Agilidad Organizacional: Estrategias de Agilidad Organizacional: Aplicación de la Inteligencia Artificial Generativa en la Gestión Documental", PODIUM, no. 47, pp. 37–58, 2025, doi: 10.31095/podium.2025.47.1249. [Online]. Available: https://revistas.uees.edu.ec/index.php/Podium/article/view/1249.

[7] S. Denning, “The age of agile : how smart companies are transforming the way work gets done", AMACOM, 2018. [Online]. Available: https://books.google.com/books/about/The_Age_of_Agile.html?hl=es&id=cGl9tAEACAAJ.

[8] S. M. Y. O. Alkaabi, N. S. B. Sohaimi, and A. B. Aminurraasyid, “The Effect of Technological Innovation and Knowledge Management Process on Organisational Agility: A Systematic Literature Review", vol. 14, no. 4, pp. 15121-15126, 2024, doi: 10.48084/ETASR.7691. [Online]. Available: https://etasr.com/index.php/ETASR/article/view/7691.

[9] N. M. Aljawarneh, “The Mediating Role of Organization Agility between Business Intelligence & Innovative Performance", vol. 13, no. 3, pp. 929-938, 2024, doi: 10.18576/JSAP/130307. [Online]. Available: https://www.naturalspublishing.com/Article.asp?ArtcID=28260.

[10] B. R. Kumar, D. Madhuri, and S. Bathina, “The Role of Artificial Intelligence in Decision-Making Processes", African Journal of Biological Sciences, vol. 6, pp. 6344-6362, 2024, doi: 10.33472/AFJBS.6.6.2024.6344-6362.

[11] S. V. Mahadevkar, S. Patil, K. Kotecha, L. W. Soong, and T. Choudhury, “Exploring AI-driven approaches for unstructured document analysis and future horizons", vol. 11, no. 1, pp. 92-, 2024, doi: 10.1186/S40537-024-00948-Z. [Online]. Available: https://link.springer.com/article/10.1186/s40537-024-00948-z.

[12] Jang, B., Kim, M., Harerimana, G., Kang, S.-u., Kim, J. W., “Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism", vol. 10, no. 17, 2020, doi: 10.3390/APP10175841. [Online]. Available: https://www.mdpi.com/2076-3417/10/17/5841.

[13] S. S. Serrano, I. P. Navarro, and M. D. González, “¿Cómo hacer una revisión sistemática siguiendo el protocolo PRISMA?: Usos y estrategias fundamentales para su aplicación en el ámbito educativo a través de un caso práctico", Bordón: Revista de pedagogía, vol. 74, no. 3, pp. 51-66, 2022, doi: 10.13042/Bordon.2022.95090. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=8583045&info=resumen&idioma=SPA.

[14] M. G. M. García, N. A. V. Valdiviezo, W. T. M. Carpio, and R. C. Guidek, “Calidad del servicio del sector comercial: Una revisión sistemática de la literatura a través del método PRISMA", Código Científico Revista de Investigación, vol. 6, no. E2, pp. 124-140, 2025, doi: 10.55813/GAEA/CCRI/V6/NE2/1019. [Online]. Available: https://revistacodigocientifico.itslosandes.net/index.php/1/article/view/1019/2085.

[15] N. Mehta and J. Doshi, “A Review of Handwritten Character Recognition", International Journal of Computer Applications, vol. 165, no. 4, pp. 37-40, 2017, doi: 10.5120/ijca2017913855.

[16] A. Shah, N. Doshi, J. Shah, K. Goel, and P. Raut, “Extraction of handwritten and printed text from a form", Journal of Physics: Conference Series, vol. 1831, no. 1, pp. 012029, 2021, doi: 10.1088/1742-6596/1831/1/012029. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1831/1/012029.

[17] M. G. Öztürk and D. Özkan Sahin and E. Kiliç, “Turkish Optical Character Recognition Under the Lens: A Systematic Review of Language-Specific Challenges, Dataset Scarcity, and Open-Source Limitations", IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3614147. [Online]. Available: https://ieeexplore.ieee.org/document/11177005.

[18] Y. Gedik and S. Solak and M. H. B. Uçar and K. Üniversitesi and B. S. M. Bölümü and S. S. K. Üniversitesi, “Görüntü İşleme Teknikleri ve Kelime Benzerliklerinin Ağırlıkları Algoritması Kullanılarak Geleneksel Sınavların Değerlendirilmesi", Acta Infologica, vol. 6, no. 1, pp. 127-139, 2022, doi: 10.26650/ACIN.1009226. [Online]. Available: https://dergipark.org.tr/tr/pub/acin/issue/69446/1009226.

[19] S. Yaman and Y. Erol, “MVSR Normalization Algorithm Method for Improving Vehicle License Plate Recognition", Turkish Journal of Science and Technology, vol. 18, no. 2, pp. 543–552, 2023, doi: 10.55525/tjst.1350368.

[20] R. U. Arslan, M. O. Íncetas, S. Dikici, A. Makalesi, and R. A. R. U. Arslan, “Effects of Character Recognition with Shell Histogram Method Using Plate Characters", Journal of Polytechnic, vol. 22, no. 4, pp. 1093-1099, 2019, doi: 10.2339/POLITEKNIK.593633. [Online]. Available: https://dergipark.org.tr/en/pub/politeknik/issue/49017/593633.

[21] E. F. B. Tasdemir, “Printed Ottoman text recognition using synthetic data and data augmentation", vol. 26, no. 3, pp. 273-287, 2023, doi: 10.1007/S10032-023-00436-9. [Online]. Available: https://link.springer.com/article/10.1007/s10032-023-00436-9.

[22] H. Arslan, “End to End Invoice Processing Application Based on Key Fields Extraction", vol. 10, pp. 78398-78413, 2022, doi: 10.1109/ACCESS.2022.3192828. [Online]. Available: https://ieeexplore.ieee.org/document/9834945.

[23] A. Akcan, E. F. B. Taşdemir, F. Kızılırmak, M. Kuru, F. Öncel and B. Yanıkoğlu, “Akis: Transcription Software of Ottoman-Turkish Texts Using Deep Learning", vol. 9, no. 2, pp. 43-48, 2022, doi: 10.2979/TUR.2022.A902198. [Online]. Available: https://muse.jhu.edu/pub/3/article/902198.

[24] F. Alim, E. Kavakli, S. B. Okcu, E. Dogan, and C. Cigla, “Simultaneous license plate recognition and face detection application at the edge", vol. 12438, pp. 232-243, 2023, doi: 10.1117/12.2649988. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12438/124380T/Simultaneous-license-plate-recognition-and-face-detection-application-at-the/10.1117/12.2649988.full.

[25] A. Kayabas, A. E. Topcu, and O. Kilic, “Ocr Error Correction Using BiLSTM", ICECET, 2021, doi: 10.1109/ICECET52533.2021.9698712. [Online]. Available: https://www.semanticscholar.org/paper/OCR-Error-Correction-Using-BiLSTM-Kayabas-Topcu/6acbd04bb7308f49225a077bdf4f17505ee30f4a.

[26] M. Zhu and J. M. Cole, “PDFDataExtractor: A Tool for Reading Scientific Text and Interpreting Metadata from the Typeset Literature in the Portable Document Format", vol. 62, no. 7, pp. 1633-1643, 2022, doi: 10.1021/ACS.JCIM.1C01198. [Online]. Available: /doi/pdf/10.1021/acs.jcim.1c01198?ref=article_openPDF.

[27] W. Xue, Q. Li, and Q. Xue, “Text Detection and Recognition for Images of Medical Laboratory Reports with a Deep Learning Approach", IEEE Access, vol. 8, pp. 407-416, 2020, doi: 10.1109/ACCESS.2019.2961964. [Online]. Available: https://ieeexplore.ieee.org/document/8941040.

[28] Z. Zaman, H. Mahdin, K. Hussain, and Atta-Ur-Rahman, “Information extraction from semi and unstructured data sources: a systematic literature review", ICIC Express Letters, vol. 14, no. 6, pp. 593–603, 2020, doi: 10.24507/icicel.14.06.593. [Online]. Available: https://doi.org/10.24507/icicel.14.06.593.

[29] J. Su, A. Sayyad, J. Shirabad, S. Matwin, and Y. Huang, “Discriminative Multinomial Naive Bayes for Text Classification", pp. 30–11, 2012. [Online]. Available: http://www.site.uottawa.ca/.

[30] W. Cao, C. Zhou, Y. Wu, Z. Ming, Z. Xu, and J. Zhang, “Research Progress of Zero-Shot Learning Beyond Computer Vision", vol. 12453 LNCS, pp. 538-551, 2020, doi: 10.1007/978-3-030-60239-0_36. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-60239-0_36.

[31] S. V. Mahadevkar, B. Khemani, S. Patil, K. Kotecha, D. R. Vora, A. Abraham, and L. A. Gabralla, “A Review on Machine Learning Styles in Computer Vision - Techniques and Future Directions", IEEE Access, vol. 10, pp. 107293-107329, 2022, doi: 10.1109/ACCESS.2022.3209825. [Online]. Available: https://ieeexplore.ieee.org/document/9903420.

[32] P. Sahare and S. B. Dhok, “Multilingual Character Segmentation and Recognition Schemes for Indian Document Images", IEEE Access, vol. 6, pp. 10603-10617, 2018, doi: 10.1109/ACCESS.2018.2795104. [Online]. Available: https://ieeexplore.ieee.org/document/8263187.

[33] S. Chowdhury and M. P. Schoen, “Research Paper Classification using Supervised Machine Learning Techniques", IETC, 2020, doi: 10.1109/IETC47856.2020.9249211. [Online]. Available: https://ieeexplore.ieee.org/document/9249211.

[34] S. Stewart and B. Barrett, “Document image page segmentation and character recognition as semantic segmentation", pp. 101-106, 2017, doi: 10.1145/3151509.3151518. [Online]. Available: /doi/pdf/10.1145/3151509.3151518?download=true.

[35] Y. S. Chernyshova, A. V. Sheshkus, and V. V. Arlazarov, “Two-Step CNN Framework for Text Line Recognition in Camera-Captured Images", IEEE Access, vol. 8, pp. 32587-32600, 2020, doi: 10.1109/ACCESS.2020.2974051. [Online]. Available: https://ieeexplore.ieee.org/document/8999509.

[36] S. Francis, J. V. Landeghem, M. F. Moens, S. Francis, J. V. Landeghem, and M. F. Moens, “Transfer Learning for Named Entity Recognition in Financial and Biomedical Documents", vol. 10, no. 8, 2019, doi: 10.3390/INFO10080248. [Online]. Available: https://www.mdpi.com/2078-2489/10/8/248.

[37] W. Weng and X. Zhu, “INet: Convolutional Networks for Biomedical Image Segmentation", IEEE Access, vol. 9, pp. 16591-16603, 2021, doi: 10.1109/ACCESS.2021.3053408. [Online]. Available: https://ieeexplore.ieee.org/document/9330594.

[38] A. Chiney, A. R. Paduri, N. Darapaneni, S. Kulkarni, M. Kadam, I. Kohli, and M. Subramaniyan, “Handwritten Data Digitization Using an Anchor based Multi-Channel CNN (MCCNN) Trained on a Hybrid Dataset (h-EH)", vol. 189, pp. 175-182, 2021, doi: 10.1016/J.PROCS.2021.05.095. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921012187?via%3Dihub.

[39] S. Desai and A. Singh, “Optical character recognition using template matching and back propagation algorithm", vol. 2016, 2016, doi: 10.1109/INVENTIVE.2016.7830161. [Online]. Available: https://ieeexplore.ieee.org/document/7830161.

[40] “IEEE Guide for Taxonomy for Intelligent Process Automation Product Features and Functionality", IEEE, 2019, doi: 10.1109/IEEESTD.2019.8764094. [Online]. Available: https://ieeexplore.ieee.org/document/8764094.

[41] S. Patil, V. Varadarajan, S. Mahadevkar, R. Athawade, L. Maheshwari, S. Kumbhare, Y. Garg, D. Dharrao, P. Kamat, K. Kotecha, S. Patil, V. Varadarajan, S. Mahadevkar, R. Athawade, L. Maheshwari, S. Kumbhare, Y. Garg, D. Dharrao, P. Kamat, and K. Kotecha, “Enhancing Optical Character Recognition on Images with Mixed Text Using Semantic Segmentation", vol. 11, no. 4, 2022, doi: 10.3390/JSAN11040063. [Online]. Available: https://www.mdpi.com/2224-2708/11/4/63.

[42] J. Luthra and P. Maheshwary, “A Survey of Advances in Text Summarization Methods", ACROSET , 2024, doi: 10.1109/ACROSET62108.2024.10743404. [Online]. Available: https://ieeexplore.ieee.org/document/10743404.

[43] W. Kuntichod, W. Lee, K. Srisomboon, L. Pipanmekaporn, and A. Prayote, “Boundary and Contextual Structure Extraction for Named Entity Recognition", IEEE Access, vol. 13, pp. 187785-187798, 2025, doi: 10.1109/ACCESS.2025.3626034. [Online]. Available: https://ieeexplore.ieee.org/document/11218889.

[44] G. Agrawal, Y. Deng, J. Park, H. Liu, and Y. C. Chen, “Building Knowledge Graphs from Unstructured Texts: Applications and Impact Analyses in Cybersecurity Education", vol. 13, no. 11, 2022, doi: 10.3390/INFO13110526. [Online]. Available: https://www.mdpi.com/2078-2489/13/11/526.

[45] R. Peng, K. Liu, P. Yang, U. Z. Yuan, and S. Li, “Embedding-based Retrieval with LLM for Effective Agriculture Information Extracting from Unstructured Data", 2023, doi: 10.48550/arXiv.2308.03107. [Online]. Available: https://arxiv.org/pdf/2308.03107.

[46] S. Gehrmann, F. Dernoncourt, Y. Li, E. T. Carlson, J. T. Wu, J. Welt, J. Foote, E. T. Moseley, D. W. Grant, P. D. Tyler, and L. A. Celi, “Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives", PLOS ONE, vol. 13, no. 2, pp. e0192360, 2018, doi: 10.1371/JOURNAL.PONE.0192360. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0192360.

[47] P. Shah, S. Joshi, and A. K. Pandey, “Legal clause extraction from contract using machine learning with heuristics improvement", ICCCA, 2018, doi: 10.1109/CCAA.2018.8777602. [Online]. Available: https://ieeexplore.ieee.org/document/8777602.

[48] H. Rigneault, N. G. Kumar, R. Cossart, D. Septier, G. B. Waslilewki, A. Kudlinski, and A. Kaszas, “Receipt Dataset for Fraud Detection", Focus on Microscopy, no. 1, pp. 255, 2017, doi: 10.34894/VQ1DJA. [Online]. Available: https://hal.science/hal-02316349.

[49] H. Sidhwa, S. Kulshrestha, S. Malhotra, and S. Virmani, “Text Extraction from Bills and Invoices", ICACCCN, pp. 564-568, 2018, doi: 10.1109/ICACCCN.2018.8748309. [Online]. Available: https://ieeexplore.ieee.org/document/8748309.

[50] J. Memon, M. Sami, R. A. Khan, and M. Uddin, “Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)", IEEE Access, vol. 8, pp. 142642-142668, 2020, doi: 10.1109/ACCESS.2020.3012542. [Online]. Available: https://ieeexplore.ieee.org/document/9151144.

[51] M. Dhouib, G. Bettaieb, and A. Shabou, “DocParser: End-to-end OCR-Free Information Extraction from Visually Rich Documents", vol. 14191 LNCS, pp. 155-172, 2023, doi: 10.1007/978-3-031-41734-4_10. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-41734-4_10.

[52] E. Meoded, “Handwritten Text Recognition of Historical Manuscripts Using Transformer-Based Models", 2025, doi: 10.48550/arXiv.2508.11499. [Online]. Available: https://arxiv.org/pdf/2508.11499v1.

[53] C. Qu, C. Liu, Y. Liu, X. Chen, D. Peng, F. Guo, and L. Jin, “Towards Robust Tampered Text Detection in Document Image: New Dataset and New Solution", pp. 5937-5946, 2023, doi: 10.1109/CVPR52729.2023.00575. [Online]. Available: https://ieeexplore.ieee.org/document/10204113.

[54] P. Darshni, B. S. Dhaliwal, R. Kumar, V. A. Balogun, S. Singh, and C. I. Pruncu, “Artificial neural network based character recognition using SciLab", vol. 82, no. 2, pp. 2517-2538, 2022, doi: 10.1007/S11042-022-13082-W. [Online]. Available: https://link.springer.com/article/10.1007/s11042-022-13082-w.

[55] D. Fleischhacker, R. Kern, and W. Göderle, “Enhancing OCR in historical documents with complex layouts through machine learning", vol. 26, no. 1, pp. 3-, 2025, doi: 10.1007/S00799-025-00413-Z. [Online]. Available: https://link.springer.com/article/10.1007/s00799-025-00413-z.

[56] V. K. Vijaya, “Invoice Data Extraction Using OCR Technique", Interantional Journal of Scientific Research in Engineering and Management, vol. 08, no. 04, pp. 1-5, 2024, doi: 10.55041/IJSREM29981.

[57] P. Martins, F. Sa, F. Morgado, and C. Cunha, “Using machine learning for cognitive Robotic Process Automation (RPA)", CISTI, vol. 2020-June, 2020, doi: 10.23919/CISTI49556.2020.9140440. [Online]. Available: https://ieeexplore.ieee.org/document/9140440.

[58] C. Vijai and M.S.R. Mariyappan, Robotic Process Automation (RPA) in Human Resource Functions", Advances In Management, vol. 16, no. 3, pp. 30-37, 2023, doi: 10.25303/1603AIM030037.

[59] D. Premasiri, T. Ranasinghe, R. Mitkov, M. E. Haj, and I. Frommholz, “Survey on legal information extraction: current status and open challenges", vol. 67, no. 12, pp. 11287-11358, 2025, doi: 10.1007/S10115-025-02600-5. [Online]. Available: https://link.springer.com/article/10.1007/s10115-025-02600-5.

[60] B. Aejas, A. Belhi, and A. Bouras, “Contract Clause Extraction Using Question- Answering Task", vol. 15436 LNCS, pp. 320-333, 2025, doi: 10.1007/978-981-96-0579-8_23. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-96-0579-8_23.

[61] Z. Li, W. Tian, C. Li, Y. Li, and H. Shi, “A Structured Recognition Method for Invoices Based on StrucTexT Model", vol. 13, no. 12, 2023, doi: 10.3390/APP13126946. [Online]. Available: https://www.mdpi.com/2076-3417/13/12/6946.

[62] S. Patel and D. Bhatt, “Abstractive Information Extraction from Scanned Invoices (AIESI) using End-to-end Sequential Approach", 2020, doi: 10.48550/arXiv.2009.05728. [Online]. Available: https://arxiv.org/pdf/2009.05728.

[63] A. O. Ojo, “A Review on the Effectiveness of Artificial Intelligence and Machine Learning on Cybersecurity", vol. 4, no. 1, pp. 104-111, 2025, doi: 10.60087/JKLST.V4.N1.011. [Online]. Available: https://jklst.org/index.php/home/article/view/v4.n1.011.

[64] Q. Chen, A. Rankine, Y. Peng, E. Aghaarabi, and Z. Lu, “Benchmarking Effectiveness and Efficiency of Deep Learning Models for Semantic Textual Similarity in the Clinical Domain: Validation Study", JMIR Medical Informatics, vol. 9, no. 12, pp. e27386, 2021, doi: 10.2196/27386. [Online]. Available: https://medinform.jmir.org/2021/12/e27386.

[65] X. H. Li, F. Yin, and C. L. Liu, “Docsam: Unified Document Image Segmentation via Query Decomposition and Heterogeneous Mixed Learning", pp. 15021-15032, 2025, doi: 10.1109/CVPR52734.2025.01399. [Online]. Available: https://arxiv.org/pdf/2504.04085.

[66] R. A. Vulpoi, A. Ciobanu, V. L. Drug, C. Mihai, O. B. Barboi, D. E. Floria, A. I. Coseru, A. Olteanu, V. Rosca, and M. Luca, “Deep Learning-Based Semantic Segmentation for Objective Colonoscopy Quality Assessment", Journal of Imaging, vol. 11, no. 3, 2025, doi: 10.3390/jimaging11030084. [Online]. Available: https://www.mdpi.com/2313-433X/11/3/84.

[67] A.Waly, B. Tarek, A. Feteha, R. Yehia, G. Amr, W. Gomaa, and A. Fares, “Invizo: Arabic Handwritten Document Optical Character Recognition Solution", 2025, doi: 10.48550/arXiv.2502.05277. [Online]. Available: https://arxiv.org/pdf/2502.05277v1.

[68] M. Sinthuja, C. G. Padubidri, G. S. Jayachandra, M. C. Teja, and G. S. P. Kumar, “Extraction of Text from Images Using Deep Learning", Procedia Computer Science, vol. 235, pp. 789-798, 2024, doi: 10.1016/J.PROCS.2024.04.075. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924007518.

[69] T. A. Moslmi, M. G. Ocana, A. L. Opdahl, and C. Veres, “Named Entity Extraction for Knowledge Graphs: A Literature Overview", IEEE Access, vol. 8, pp. 32862-32881, 2020, doi: 10.1109/ACCESS.2020.2973928. [Online]. Available: https://ieeexplore.ieee.org/document/8999622.

[70] V. T. Le, K. T. Trung, and V. T. Hoang, “A Comprehensive Review of Recent Deep Learning Techniques for Human Activity Recognition", vol. 2022, no. 1, pp. 8323962, 2022, doi: 10.1155/2022/8323962. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/8323962.

[71] R. C. Karpagalakshmi, S. Hegde, R. R. Sharma, S. Mahapatra, E. S. Leni, and A. Sungheetha, “Deep Learning-Based Recognition of Handwritten English Characters", Procedia Computer Science, vol. 258, pp. 1783-1792, 2025, doi: 10.1016/J.PROCS.2025.04.430. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050925015339.

[72] A. Qaroush, A. Awad, M. Modallal, and M. Ziq, “Segmentation-based, omnifont printed Arabic character recognition without font identification", vol. 34, no. 6, pp. 3025-3039, 2022, doi: 10.1016/J.JKSUCI.2020.10.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S131915782030481X.

[73] M. Kasthuri and V. Manoharan, “Handwritten character recognition and genetic algorithms", Handbook of Computational Sciences: A Multi and Interdisciplinary Approach, pp. 197-224, 2023, doi: 10.1002/9781119763468.CH10. [Online]. Available: https://ieeexplore.ieee.org/document/10953877.

[74] R. Najam, and S. Faizullah, “Analysis of Recent Deep Learning Techniques for Arabic Handwritten-Text OCR and Post-OCR Correction", vol. 13, no. 13, 2023, doi: 10.3390/APP13137568. [Online]. Available: https://www.mdpi.com/2076-3417/13/13/7568.

[75] R. M. R. Zavala and P. Martínez, “Analyzing transfer learning impact in biomedical cross-lingual named entity recognition and normalization", vol. 22, no. 1, pp. 601-, 2021, doi: 10.1186/S12859-021-04247-9. [Online]. Available: https://link.springer.com/article/10.1186/s12859-021-04247-9.

[76] K. Cui, Q. Xu, Y. Ding, J. Mei, Y. He, and H. Liu, “Optical Character Recognition Method Based on YOLO Positioning and Intersection Ratio Filtering", vol. 17, no. 8, 2025, doi: 10.3390/SYM17081198. [Online]. Available: https://www.mdpi.com/2073-8994/17/8/1198.

[77] S. Aladhadh, H. U. Rehman, A. M. Qamar, and R. U. Khan, “Recurrent Convolutional Neural Network MSER-Based Approach for Payable Document Processing", Computers, Materials & Continua, vol. 69, no. 3, pp. 3399-3411, 2021, doi: 10.32604/CMC.2021.018724. [Online]. Available: https://www.techscience.com/cmc/v69n3/44158.html.

[78] A. Dash, S. Darshana, D. K. Yadav, and V. Gupta, “A clinical named entity recognition model using pretrained word embedding and deep neural networks", vol. 10, pp. 100426, 2024, doi: 10.1016/J.DAJOUR.2024.100426. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772662224000304?via%3Dihub.

[79] T. Vidler, K. McGarry, and D. Baglee, “Text Mining Legal Documents for Clause Extraction", CSCE, pp. 1469-1476, 2023, doi: 10.1109/CSCE60160.2023.00243. [Online]. Available: https://ieeexplore.ieee.org/document/10487546.

[80] S. Khalafi, N. Ghadiri, and M. Moradi, “A hybrid deep learning approach for phenotype prediction from clinical notes", vol. 14, no. 4, pp. 4503-4513, 2023, doi: 10.1007/S12652-023-04568-Y. [Online]. Available: https://link.springer.com/article/10.1007/s12652-023-04568-y.

[81] S. Umer, R. Mondal, H. M. Pandey, and R. K. Rout, “Deep features based convolutional neural network model for text and non-text region segmentation from document images", vol. 113, pp. 107917, 2021, doi: 10.1016/J.ASOC.2021.107917. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1568494621008395?via%3Dihub.

[82] M. W. Ma, X. S. Gao, Z. Y. Zhang, S. Y. Shang, L. Jin, P. L. Liu, F. Lv, W. Ni, Y. C. Han, and H. Zong, “Extracting laboratory test information from paper-based reports", vol. 23, no. 1, pp. 251-, 2023, doi: 10.1186/S12911-023-02346-6. [Online]. Available: https://link.springer.com/article/10.1186/s12911-023-02346-6.

[83] L. Nicolaieff, M. M. Kandi, Y. Zegaoui, and C. Bortolaso, “Intelligent Document Processing with Small and Relevant Training Dataset", ISCV, 2022, doi: 10.1109/ISCV54655.2022.9806100. [Online]. Available: https://ieeexplore.ieee.org/document/9806100.

[84] B. Jose and S. Abraham, “Intelligent processing of unstructured textual data in document based NoSQL databases", Materials Today: Proceedings, vol. 80, pp. 1777-1785, 2023, doi: 10.1016/J.MATPR.2021.05.605. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2214785321042322?via%3Dihub.

[85] S. Cho, J. Moon, J. Bae, J. Kang, and S. Lee, “A Framework for Understanding Unstructured Financial Documents Using RPA and Multimodal Approach", vol. 12, no. 4, pp. 939, 2023, doi: 10.3390/ELECTRONICS12040939. [Online]. Available: https://www.mdpi.com/2079-9292/12/4/939/htm.

[86] J. Wang, L. Jin, and K. Ding, “LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding", vol. 1, pp. 7747-7757, 2022, doi: 10.18653/v1/2022.acl-long.534. [Online]. Available: https://arxiv.org/pdf/2202.13669.

[87] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, “LayoutLM: Pre-training of Text and Layout for Document Image Understanding", pp. 1192-1200, 2020, doi: 10.1145/3394486.3403172. [Online]. Available: https://dl.acm.org/doi/10.1145/3394486.3403172.

[88] G. Kim, T. Hong, M. Yim, J. Y. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, “OCR-free Document Understanding Transformer", vol. 13688 LNCS, pp. 498-517, 2021, doi: 10.1007/978-3-031-19815-1_29. [Online]. Available: https://arxiv.org/pdf/2111.15664.

[89] H. Liao, A. Roychowdhury, W. Li, A. Bansal, Y. Zhang, Z. Tu, R. K. Satzoda, R. Manmatha, and V. Mahadevan, “DocTr: Document Transformer for Structured Information Extraction in Documents", pp. 19527-19537, 2023, doi: 10.1109/ICCV51070.2023.01794. [Online]. Available: https://ieeexplore.ieee.org/document/10377890.

[90] V. Sujay, B. S. Lakshmi, T. Venkatesan, S. S. Ram, E. Sangeetha, and S. Satheeshkumar, “Synergizing SVM Classifier with Hybrid CNN Architecture for Handwritten Recognition", ESCI , 2024, doi: 10.1109/ESCI59607.2024.10497404. [Online]. Available: https://ieeexplore.ieee.org/document/10497404.

[91] M. Vafaie, J. Waitelonis, and H. Sack, “Improvements in Handwritten and Printed Text Separation in Historical Archival Documents", Archiving Conference, vol. 20, no. 1, pp. 36-41, 2023, doi: 10.2352/ISSN.2168-3204.2023.20.1.7. [Online]. Available: https://library.imaging.org/archiving/articles/20/1/7.

[92] D. Yan, K. Li, S. Gu, and L. Yang, “Network-Based Bag-of-Words Model for Text Classification", IEEE Access, vol. 8, pp. 82641-82652, 2020, doi: 10.1109/ACCESS.2020.2991074. [Online]. Available: https://ieeexplore.ieee.org/document/9079815.

[93] S. Hamida, B. Cherradi, O. E. Gannour, A. Raihani, and H. Ouajji, “Cursive Arabic handwritten word recognition system using majority voting and k-NN for feature descriptor selection", vol. 82, no. 26, pp. 40657-40681, 2023, doi: 10.1007/S11042-023-15167-6. [Online]. Available: https://link.springer.com/article/10.1007/s11042-023-15167-6.

[94] S. Mallappa, D. B.V., and G. Mukarambi, “Script Identification from Camera Captured Indian Document Images with CNN Model", ICTACT Journal on Soft Computing, vol. 14, no. 2, pp. 3232-3236, 2023, doi: 10.21917/IJSC.2023.0453.

[95] N. Perera, M. Dehmer, and F. E. Streib, “Named Entity Recognition and Relation Detection for Biomedical Information Extraction", vol. 8, pp. 519578, 2020, doi: 10.3389/fcell.2020.00673. [Online]. Available: www.frontiersin.org.

[96] J. M. Steinkamp, W. Bala, A. Sharma, and J. J. Kantrowitz, “Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes", vol. 102, pp. 103354, 2020, doi: 10.1016/J.JBI.2019.103354. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S153204641930276X?via%3Dihub.

[97] J. Jouffroy, S. F. Feldman, I. Lerner, B. Rance, A. Burgun, and A. Neuraz, “Hybrid Deep Learning for Medication-Related Information Extraction From Clinical Texts in French: MedExt Algorithm Development Study.", JMIR medical informatics, vol. 9, no. 3, pp. e17934, 2021, doi: 10.2196/17934. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/33724196.

[98] S. Gao, O. Kotevska, A. Sorokine, and J. B. Christian, “A pre-training and self-training approach for biomedical named entity recognition", vol. 16, no. 2, pp. e0246310, 2021, doi: 10.1371/JOURNAL.PONE.0246310. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0246310.

[99] S. Momeni and B. BabaAli, “A Transformer-based Approach for Arabic Offline Handwritten Text Recognition", Signal, Image and Video Processing, vol. 18, no. 4, pp. 3053-3062, 2023, doi: 10.1007/s11760-023-02970-9. [Online]. Available: https://arxiv.org/pdf/2307.15045.

[100] W. W. Chi, T. Y. Tang, N. M. Salleh, M. Mukred, H. Alsalman, and M. Zohaib, “Data Augmentation With Semantic Enrichment for Deep Learning Invoice Text Classification", IEEE Access, vol. 12, pp. 57326-57344, 2024, doi: 10.1109/ACCESS.2024.3387860. [Online]. Available: https://ieeexplore.ieee.org/document/10496671.

[101] C. Clausner, A. Antonacopoulos, and S. Pletschacher, “Efficient and effective OCR engine training", vol. 23, no. 1, pp. 73-88, 2019, doi: 10.1007/S10032-019-00347-8. [Online]. Available: https://link.springer.com/article/10.1007/s10032-019-00347-8.

[102] R. Pramanik and S. Bag, “Shape decomposition-based handwritten compound character recognition for Bangla OCR", vol. 50, pp. 123-134, 2018, doi: 10.1016/J.JVCIR.2017.11.016. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1047320317302250?via %3Dihub.

[103] A. Nikolaidis and C. Strouthopoulos, “Robust text extraction in mixed-type binary documents", MMSP, pp. 393-398, 2008, doi: 10.1109/MMSP.2008.4665110. [Online]. Available: https://ieeexplore.ieee.org/document/4665110.