PROCESAMIENTO DEL LENGUAJE NATURAL PARA EXTRACCIÓN DE INFORMACIÓN DE CELDAS SOLARES DE PEROVSKITA

JUAN PABLO GAMBOA DURÁN
VALERIA DE LOS ANGELES SARMIENTO CLAVIJO

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
BUCARAMANGA
2026

PROCESAMIENTO DEL LENGUAJE NATURAL PARA EXTRACCIÓN DE INFORMACIÓN DE CELDAS SOLARES DE PEROVSKITA

JUAN PABLO GAMBOA DURÁN
VALERIA DE LOS ANGELES SARMIENTO CLAVIJO

Trabajo de Grado para Optar al Título de
Ingenieros Electrónicos

Directora:
MÓNICA ANDREA BOTERO LONDOÑO
Doctora en Ciencias-Física
Codirector:
FRANKLIN ALEXANDER SEPÚLVEDA SEPÚLVEDA
Doctor en Ingeniería Electrónica

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
BUCARAMANGA
2026

AGRADECIMIENTOS

A todos aquellos que creyeron en mí y en mis ganas de hacer el mundo un lugar mejor,

A mi madre y mi padre, que me enseñaron a poner mi alma en lo que amo y nunca rendirme en ello,

A mi tía Luz Nayibe, que empapó mi vida con su amor al conocimiento,

A Juan Esteban, por quién siempre quiero ser mejor,

A Alejandro y Mayra, quienes nunca dudaron en ayudarme y me enseñaron a disfrutar el proceso,

A MinCiencias y su apoyo para con este proyecto,

Finalmente, a Dios, mi familia y amigos, por siempre sostener mi espalda, mi lápiz y mi corazón.

Con aprecio, **JUAN PABLO GAMBOA DURÁN**

AGRADECIMIENTOS

En este importante logro académico como ingeniera electrónica, expreso mi más profundo agradecimiento a todos quienes hicieron posible la obtención de este título con su acompañamiento y apoyo constante.

A Dios, por su guía e iluminación a lo largo de mi proceso formativo. A mi madre, por su amor, sus enseñanzas y por inculcarme el valor del estudio y la perseverancia. A mi familia, por su apoyo incondicional, destacando especialmente al Rolando Esteban Clavijo Arcos, así como a mis tías Margarita, Alicia y mi tío Leonardo, quienes fueron pilares fundamentales durante estos años.

A la Universidad Industrial de Santander, por abrirme sus puertas a dpesar de las dificultades que ello implicaba, y a sus docentes, por compartir sus conocimientos y contribuir a mi formación profesional.

A MinCiencias, por el apoyo brindado a este proyecto de investigación.

A mis amigos, por su compañía y apoyo durante este proceso.

Finalmente, a Colombia por acogerme y brindarme esta oportunidad, y a Migración Colombia por su gestión y apoyo en este proceso.

Con aprecio, **VALERIA DE LOS ANGELES SARMIENTO CLAVIJO**

CONTENIDO

pág.

LISTA DE FIGURAS

pág.

LISTA DE TABLAS

pág.

RESUMEN

TÍTULO: PROCESAMIENTO DEL LENGUAJE NATURAL PARA EXTRACCIÓN DE INFORMACIÓN DE CELDAS SOLARES DE PEROVSKITA [*]

AUTORES: JUAN PABLO GAMBOA DURÁN
VALERIA DE LOS ÁNGELES SARMIENTO CLAVIJO [**]

PALABRAS CLAVE: Procesamiento de lenguaje natural, celdas solares de perovskita, extracción de información, reconocimiento de entidades, BERT, minería de textos científicos.

DESCRIPCIÓN: En los últimos años, el crecimiento exponencial de la literatura científica sobre celdas solares de perovskita ha generado la necesidad de desarrollar herramientas que permitan extraer y estructurar información relevante de manera automática. En este contexto, el presente trabajo propone un sistema basado en técnicas de Procesamiento de Lenguaje Natural (PLN) para la extracción de parámetros fotovoltaicos clave a partir de artículos científicos en formato PDF.

La metodología implementada incluye la selección y adaptación de modelos de lenguaje preentrenados, específicamente BERT, RoBERTa y MatSciBERT, mediante técnicas de fine-tuning para la tarea de reconocimiento de entidades nombradas (NER). Adicionalmente, se emplea una estrategia basada en reglas heurísticas para establecer relaciones entre entidades, así como herramientas para la extracción de tablas, imágenes y metadatos. Como complemento, se desarrolla una interfaz web que permite la visualización y exploración de los datos obtenidos.

Los resultados evidencian un desempeño competitivo entre los modelos evaluados, destacándose BERT por su estabilidad y consistencia en validación cruzada, alcanzando un F1 de 0.93657. No obstante, se identifican limitaciones asociadas a la variabilidad en los formatos de los artículos y a la extracción de información desde tablas.

Finalmente, el sistema desarrollado demuestra la viabilidad de automatizar la estructuración de información científica en este dominio, constituyendo un aporte para la reducción de tiempos de análisis y el fortalecimiento de procesos de investigación en energías renovables.

---

[*] Trabajo de Grado

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Director: MÓNICA ANDREA BOTERO LONDOÑO

ABSTRACT

TITLE: NATURAL LANGUAGE PROCESSING FOR INFORMATION EXTRACTION FROM PEROVSKITE SOLAR CELLS [*]

AUTHORS: JUAN PABLO GAMBOA DURÁN
VALERIA DE LOS ANGELES SARMIENTO CLAVIJO [**]

Keywords: Natural language processing, perovskite solar cells, information extraction, named entity recognition, BERT, scientific text mining.

DESCRIPTION: In recent years, the exponential growth of scientific literature on perovskite solar cells has created the need for automated tools capable of extracting and structuring relevant information. In this context, this work proposes a system based on Natural Language Processing (NLP) techniques for extracting key photovoltaic parameters from scientific articles in PDF format.

The methodology includes the selection and adaptation of pretrained language models, specifically BERT, RoBERTa, and MatSciBERT, using fine-tuning techniques for the Named Entity Recognition (NER) task. Additionally, a heuristic-based approach is implemented to establish relationships between entities, along with tools for extracting tables, images, and metadata. As a complementary component, a web interface is developed to enable visualization and exploration of the extracted data.

The results show competitive performance among the evaluated models, with BERT standing out due to its stability and consistency in cross-validation, achieving an F1-score of 0.93657. However, limitations were identified related to the variability of article formats and the challenges associated with table information extraction.

Finally, the proposed system demonstrates the feasibility of automating the structuring of scientific information in this domain, contributing to the reduction of analysis time and supporting research processes in renewable energy.

---

[*] BSc Thesis

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Advisor: MÓNICA ANDREA BOTERO LONDOÑO

INTRODUCCIÓN

En los últimos años, la humanidad se ha enfrentado a la crisis del cambio climático. Las emisiones globales de CO alcanzaron un récord de 36,8 Gt en 2023, lo que impulsó la adición de 507 GW de nueva capacidad renovable, un 50 % más que en 2022 [1]. En este contexto, las energías renovables han tomado un papel central en la transición hacia un futuro más verde, y entre ellas, la energía solar se destaca como una de las más prometedoras por su disponibilidad, versatilidad y bajo impacto ambiental.

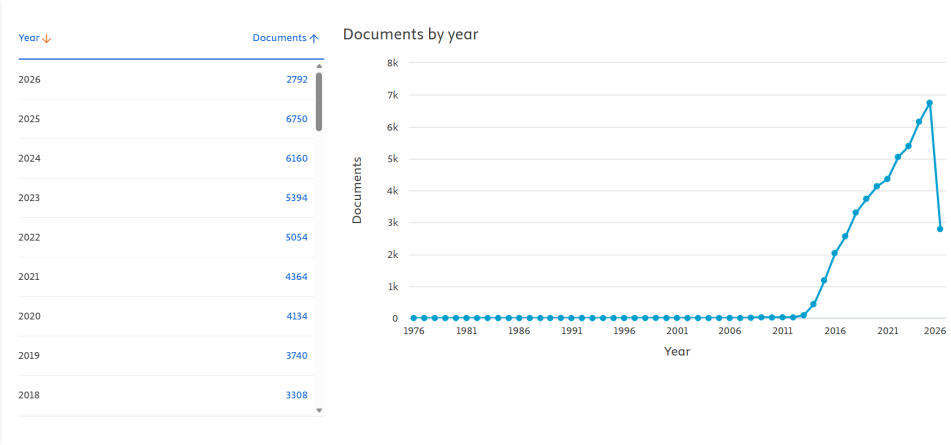
Sin embargo, el aprovechamiento eficiente de la energía solar aún enfrenta diversos desafíos técnicos y económicos. Es aquí donde entra en juego la innovación en materiales, y específicamente, la investigación sobre celdas solares de perovskita, una tecnología emergente que ha demostrado un rápido aumento en eficiencia de conversión energética, alcanzando valores superiores al 25 % en menos de una década [2]. El rápido avance en el campo de las celdas solares de perovskita ha dado lugar a un aumento significativo en la producción científica, como se puede observar en la base de datos Scopus bajo la búsqueda del término *perovskite solar cell* tal como se evidencia en la Figura. 0.1). Solo en 2025, se publicaron 6,750 documentos, un indicativo claro del creciente interés y progreso en las energías renovables y la tecnología fotovoltaica basada en perovskitas.

---

[1] INTERNATIONAL ENERGY AGENCY. *Electricity – Renewables 2023 – Analysis*. `https://www.iea.org/reports/renewables-2023/electricity`. [En línea]. Accedido: 18 de abril de 2025.

[2] T. XIE et al. «Creation of a structured solar cell material dataset and performance prediction using large language models». En: *Patterns* 5.5 (mayo de 2024), pág. 100955. DOI: 10.1016/j.patter.2024.100955.

Figura 0.1. Resultados de búsqueda en Scopus de documentos relacionados con ”perovskite AND solar AND cell” en el título, resumen o palabras clave.



Ante esta realidad, el presente proyecto propone una solución: utilizar modelos de lenguaje natural (NLP), específicamente modelos de lenguaje pre-entrenados del dominio científico como MatSciBERT [3], para automatizar la extracción de información clave contenida en artículos científicos sobre celdas solares de perovskita. Esta tarea, conocida como extracción de información (Information Extraction, IE), busca identificar de forma automática datos importantes como la eficiencia de conversión energética (PCE), el voltaje de circuito abierto (Voc), la corriente de cortocircuito (Jsc), entre otros. Como consecuencia, el uso de herramientas de procesamiento de lenguaje natural (PLN) adaptadas al dominio técnico representa una oportunidad para automatizar la extracción de dicha información desde artículos científicos.

En este contexto, los grupos de investigación GISEL y CEMOS, adscritos a la Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones (E3T) de la Universidad Industrial de Santander, podrán aprovechar esta tecnología para fortalecer sus

[3] T. GUPTA et al. «MatSciBERT: A materials domain language model for text mining and information extraction». En: *NPJ Comput. Mater.* 8.1 (mayo de 2022), pág. 102. DOI: 10.1038/s41524-022-00784-w.

líneas de investigación en dispositivos fotovoltaicos emergentes, reduciendo los tiempos de análisis, facilitando la toma de decisiones y potenciando el diseño de nuevos dispositivos. Sin embargo, para lograr un impacto real en estos procesos, es necesario superar obstáculos clave como: la dificultad en el acceso a la información y estructuración del conocimiento científico relevante, lo cual representa un cuello de botella para los investigadores y demás actores del ecosistema de energía solar fotovoltaica. Este es, precisamente, el problema específico que aborda este proyecto —no el desarrollo directo de nuevos materiales fotovoltaicos, sino la optimización del proceso de recolección de la información ya existente. Por ejemplo, el Perovskite Database Project requirió la intervención manual de 95 expertos para extraer datos de más de 15000 publicaciones especializadas, lo cual es un proceso costoso y propenso a errores [4].

A través de la aplicación de técnicas modernas de Procesamiento del Lenguaje Natural, se busca atacar una de las causas del estancamiento en la transferencia de conocimiento científico hacia la innovación tecnológica: la falta de herramientas automáticas, confiables y especializadas para convertir texto libre en información estructurada [3]. Este proyecto, por tanto, se sitúa en la intersección entre la inteligencia artificial, la ciencia de materiales y la sostenibilidad energética.

Más allá de las ventajas ya enunciadas, este proyecto representa un aporte para acelerar la investigación en celdas solares de perovskita al facilitar el acceso a datos experimentales previos; por otro, contribuye a democratizar el conocimiento científico, al convertir artículos complejos en información comprensible y reutilizable. En última instancia, esto puede traducirse en mayores avances en tecnologías limpias, mejor to-

---

[4] *The Wikipedia of perovskite solar cell research*. `https://www.pv-magazine.com/2021/12/28/the-wikipedia-of-perovskite-solar-cell-research/`. pv magazine International, [En línea]. Accedido: 18 de abril de 2025.

ma de decisiones por parte de investigadores e inversionistas, y un impulso hacia una economía energética más sostenible.

1. OBJETIVOS

1.1. OBJETIVO GENERAL

Implementar un método de extracción de información para celdas solares de perovskita a partir de artículos científicos mediante modelos grandes de lenguaje natural pre-entrenados.

1.2. OBJETIVOS ESPECÍFICOS

Adaptar un modelo de lenguaje pre-entrenado (LLM) para tareas de reconocimiento de entidades nombradas (NER) relevantes en el dominio de las celdas solares de perovskita, a través de técnicas de sintonización fina (fine-tuning).

Implementar una estrategia de detección de relaciones entre entidades basada en reglas heurísticas simples, con el objetivo de identificar asociaciones directas, sin requerir modelos adicionales de aprendizaje profundo.

Implementar un medio de visualización de los resultados como apoyo complementario en el desarrollo del proyecto.

2. ESTADO DEL ARTE

En los últimos años, el gran crecimiento de la literatura sobre celdas solares de perovskita (CSP) ha generado la necesidad de métodos automáticos que permitan transformar descripciones experimentales dispersas en conjuntos de datos estructurados. En este contexto, el procesamiento de lenguaje natural (PLN) se ha consolidado como una herramienta clave para extraer, normalizar y relacionar información técnica a partir de artículos científicos, informes y materiales suplementarios. Diversas revisiones en ciencia de materiales muestran que las técnicas modernas de PLN podrían permitir identificar composiciones, parámetros de procesado y propiedades físicas a una escala que sería inabordable mediante métodos manuales.[5] Estos trabajos también resaltan que, para ser realmente eficaces en dominios específicos como las perovskitas, los sistemas de PLN deben ser capaces de manejar una terminología altamente especializada y una presentación heterogénea de los datos en la literatura.

Los primeros esfuerzos a gran escala orientados específicamente a dispositivos fotovoltaicos demostraron que es posible construir bases de datos de celdas solares de manera casi completamente automatizada. Un ejemplo representativo es la base de datos de dispositivos de celdas solares sensibilizadas y de perovskita generada mediante *ChemDataExtractor*, donde se extraen de la literatura parámetros como la eficiencia de conversión, el voltaje de circuito abierto, la corriente de cortocircuito y el factor de llenado, junto con información sobre la arquitectura del dispositivo y los mate-

---

[5] J. H. LEE, M. LEE y K. MIN. «Natural Language Processing Techniques for Advancing Materials Discovery: A Short Review». En: *International Journal of Precision Engineering and Manufacturing-Green Technology* 10.5 (2023), págs. 1337-1349. DOI: 10.1007/s40684-023-00523-6.

riales de cada capa.[6] Este trabajo puso de manifiesto que las técnicas de PLN pueden reemplazar buena parte de la extracción manual de datos en este campo, pero también evidenció limitaciones relacionadas con la variabilidad en la manera de reportar resultados (abreviaturas, unidades, estilos de redacción y ubicación de la información en tablas o en el texto).

Para superar parte de estas limitaciones se han propuesto metodologías generales de extracción de propiedades de materiales que integran módulos de reconocimiento de entidades, extracción de relaciones y normalización de datos. Una de estas propuestas, aplicada inicialmente a grandes conjuntos de textos de polímeros, ilustra una arquitectura de PLN que identifica automáticamente materiales, procesos y propiedades, y los organiza en bases de datos reutilizables.[7] Aunque esta contribución no se centra exclusivamente en las perovskitas, su diseño modular resulta directamente transferible al caso de las CSP ya que se pueden adaptar los modelos de reconocimiento de entidades a la nomenclatura específica de estos materiales y ajustar las reglas de normalización a las magnitudes fotovoltaicas de interés.

En paralelo, se ha estudiado de forma cuantitativa cómo se presentan los datos de materiales en la literatura, utilizando herramientas de PLN para medir la frecuencia y la manera en que se reportan parámetros clave. Estos análisis revelan que, aun en campos de rápida evolución, los artículos suelen contener lagunas de información o reportes incompletos, y que los datos relevantes se distribuyen entre el cuerpo del

---

6 E. J. BEARD y J. M. COLE. «Perovskite- and Dye-Sensitized Solar-Cell Device Databases Auto-generated Using ChemDataExtractor». En: *Sci. Data* 9.1 (jun. de 2022), pág. 329. DOI: 10.1038/s41597-022-01355-w.

7 P. SHETTY et al. «A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing». En: *npj Computational Materials* 9.1 (2023), pág. 52. DOI: 10.1038/s41524-023-01003-w.

texto, las tablas y el material suplementario.[8] Para el área de las CSP, esto implica que los sistemas de PLN deben ser robustos frente a formatos dispares y capaces de combinar información procedente de múltiples secciones de un mismo documento, si se desea reconstruir de forma fiable las características completas de un dispositivo.

Más recientemente, se han desarrollado enfoques que combinan modelos de lenguaje de gran tamaño con técnicas de PLN clásicas para crear conjuntos de datos estructurados de celdas solares. En uno de estos trabajos se emplean modelos de lenguaje para ayudar en la extracción automática de información sobre materiales, arquitecturas y métricas de rendimiento, con el objetivo de construir un conjunto de datos amplio y balanceado para celdas solares y, a partir de él, entrenar modelos de predicción de eficiencia.[2] La combinación de PLN y modelos de lenguaje permite capturar dependencias complejas entre composición, condiciones de procesado y desempeño fotovoltaico. A nivel local, Jiménez desarrolló recientemente una herramienta basada en procesamiento de lenguaje natural para la extracción automática de parámetros de desempeño (PCE, J_{SC}, V_{OC}, FF) y entidades asociadas a las capas ETL, HTL y perovskita a partir de artículos en formato PDF, proponiendo además un esquema de etiquetado específico para este dominio [9]. Este método, el cual está basado en la librería *SpaCy*, fue luego mejorado en [10] y probado con los datos de celdas solares

---

[8] H. M. SAYEED et al. «NLP meets materials science: Quantifying the presentation of materials data in literature». En: *Matter* 7.3 (2024), págs. 723-727. DOI: 10.1016/j.matt.2023.12.032.

[9] M. Z. JIMÉNEZ DÍAZ. «Herramienta de software para la extracción automática de parámetros de desempeño a partir de publicaciones científicas de celdas solares de perovskita». Trabajo de grado. Bucaramanga, Colombia: Universidad Industrial de Santander, 2022.

[10] M. Z. JIMÉNEZ-DÍAZ, A. SEPÚLVEDA-SEPÚLVEDA y M. BOTERO-LONDOÑO. «Named Entity Recognition for Performance and Synthesis Information of Perovskite Solar Cells Using SpaCy». En: *Computational Science and Its Applications – ICCSA 2025*. Vol. 15649. Lecture Notes in Computer Science. Cham: Springer, 2025. DOI: 10.1007/978-3-031-96997-3_19.

publicados en [11].

Finalmente, se ha explorado el uso de modelos conversacionales de última generación, junto con estrategias de *prompt engineering*, para extraer datos de materiales con alta precisión a partir de artículos de investigación, incluidas configuraciones complejas de dispositivos de celdas solares. Estos enfoques formulan instrucciones detalladas a los modelos de lenguaje para que identifiquen en el texto y en las tablas la información relevante sobre composición la capa absorbente, tipo de arquitectura (*n-i-p* o *p-i-n*), capas de transporte y métricas fotovoltaicas, devolviéndola en formatos aprovechables para análisis posteriores.[12] Los resultados indican que, cuando se diseñan adecuadamente los *prompts* y se incluyen mecanismos de verificación, estos modelos pueden alcanzar una exactitud comparable o superior a la de las metodologías basadas únicamente en PLN tradicional, especialmente al interpretar tablas complejas y descripciones experimentales extensas, sin embargo, aún persiste una brecha importante, ya que las metodologías actuales no alcanzan una extracción completa, estandarizada y enfocada específicamente en este tipo de dispositivos, lo que limita el aprovechamiento pleno de la información disponible.[5]

---

[11] T. J. et al. JACOBSSON. «An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles». En: *Nature Energy* 7.1 (2022), págs. 107-115. DOI: 10.1038/s41560-021-00941-3.

[12] M. P. POLAK et al. «Extracting accurate materials data from research papers with conversational language models and prompt engineering». En: *Nature Communications* 15.1 (2024), pág. 1569. DOI: 10.1038/s41467-024-45914-8.

3. METODOLOGÍA

El proceso de extracción de información de celdas solares de perovskita por medio de procesamiento de lenguaje natural se separa en siete pasos, los cuales se desglosarán a continuación.

3.1. SELECCIÓN DE MODELOS

Como primer paso metodológico, se revisaron modelos para procesamiento de lenguaje natural basados en la arquitectura transformers para una mayor eficiencia. Teniendo esto en cuenta, se preseleccionaron 3 modelos.

3.1.1. BERT (*Bidirectional Encoder Representations from Transformers*) Es un modelo de representación del lenguaje basado en la arquitectura *Transformer*. Su entrenamiento sobre grandes corpus—entre ellos *BookCorpus* y *Wikipedia*—permite capturar relaciones sintácticas complejas y semánticas de alto nivel que facilitan su adaptación a tareas específicas mediante técnicas de *fine-tuning*[13]. Gracias a su mecanismo de atención bidireccional proveniente de la arquitectura *Transformer*, este modelo es adecuado para capturar relaciones contextuales, algo conveniente a la hora de recolectar datos para múltiples celdas solares dentro de un mismo artículo de investigación.

3.1.2. RoBERTa (*A Robustly Optimized BERT Pretraining Approach*) Es una reimplementación optimizada de BERT, el modelo introduce ajustes clave como entrenamientos más prolongados con lotes mayores, uso de un corpus sustancialmente am-

---

[13] Nadia Mushtaq GARDAZI et al. «BERT applications in natural language processing: a review». En: *Artificial Intelligence Review* 58 (2025). Accepted: 19 February 2025 / Published online: 15 March 2025, pág. 166. DOI: 10.1007/s10462-025-11162-5.

pliado, eliminación del objetivo de *Next Sentence Prediction* (NSP) y empleo de un esquema de enmascaramiento dinámico, lo cual permite alcanzar resultados de estado del arte en tareas como GLUE (*General Language Understanding Evaluation*), RACE (*ReAding Comprehension from Examinations*) y SQuAD (*Stanford Question Answering Dataset*)[14]. Estas estrategias de reajuste refuerzan las relaciones aprendidas y hacen que este modelo sea ventajoso en contextos de procesamiento de literatura ciencífica y técnica donde se requieren análisis más robustos, como lo es la literatura de CSP.

3.1.3. MatSciBERT Es un modelo de lenguaje especializado para el dominio de ciencia de materiales, basado en la arquitectura de BERT pero ajustado con texto técnico proveniente de artículos científicos y bases de datos del campo [3]. El enfoque de este modelo le permite capturar con mayor precisión el vocabulario, las entidades y las relaciones características de los dominios, pudiendo así generar representaciones contextuales más prácticas para el tipo de tareas a trabajar.

3.2. FINE-TUNING

Para adaptar los modelos preentrenados a la temática de este proyecto, se realiza un proceso denominado *Fine-Tunning* donde, haciendo uso de las etiquetas de Jiménez [9], y mediante aprendizaje supervisado, se adaptan sus capacidades a la extracción de información técnica de artículos científicos relacionados con CSP.

---

[14] Yinhan LIU et al. «RoBERTa: A Robustly Optimized BERT Pretraining Approach». En: *arXiv preprint arXiv:1907.11692* (2019). URL: https://arxiv.org/abs/1907.11692.

3.3. ESTRATEGIA DE VALIDACIÓN

Para evaluar el rendimiento de los modelo se emplea la técnica de validación cruzada, donde se divide el conjunto de datos en múltiples subconjuntos, permitiendo así entrenar y validar el modelo en diferentes particiones[15]. Para el caso puntual de este trabajo, se utiliza una partición inicial de datos en 90 % para entrenamiento (*train*) y 10 % para prueba (*test*) . Sobre el conjunto de entrenamiento se aplica validación cruzada tipo *k-fold* con $k = 5$,en la cual los datos se dividen en cinco subconjuntos. Como resultado de este procedimiento, y tomando en consideración las métricas obtenidas para el subconjunto de prueba del 10 % separado inicialmente, se determina el modelo a emplear.

3.4. PREPARACIÓN TEXTUAL

Después de la validación, se inicia el código principal de procesamiento y, con el fin de optimizar el rendimiento en la introducción de texto al modelo NER, se realiza una preparación textual mediante la creación y utilización de las funciones `clean_text(text)`, `chunk_text (text, chunk_size=400)` y `fix_spaces_in_number(s: str)`. Estas se encargan de eliminar saltos de línea y espacios extra, corregir caracteres especiales, dividir el texto en lotes de hasta 400 tokens, valor inferior al máximo de 512 admitido por el modelo [16]), así como corregir formatos numéricos mal escritos.

---

[15] SCIKIT-LEARN DEVELOPERS. *Cross-validation: evaluating estimator performance*. Consultado el 31 de marzo de 2026. s.f. URL: https://scikit-learn.org/stable/modules/cross_validation.html.

[16] HUGGING FACE. *BERT — Transformers Documentation*. Consultado el 31 de marzo de 2026. 2026. URL: https://huggingface.co/docs/transformers/model_doc/bert.

3.5. RECONOCIMIENTO DE ENTIDADES NOMBRADAS

Previo a iniciar la extracción del contenido, se realizan el reconocimiento, extracción y clasificación de las entidades específicas del texto no estructurado (*Named Entity Recognition* o *NER*)[17] utilizando, para este caso, etiquetas en formato *BIO* que sirven para para marcar si una palabra está fuera (*Outside*), al inicio (*Beginning*) o dentro (*Inside*) de una entidad nombrada[18], obteniendo 12 tipos de clases: ['ETL', 'FF', 'HTL', 'JSC', 'NFF', 'NJSC', 'NPCE', 'NVOC', 'PCE', 'PESK', 'PTF', 'VOC']y 25 etiquetas debido al formato BIO. Se selecciona así mismo el modelo con mejor rendimiento (*Bert*), llamádolo de la siguiente manera en `main()`:

```
MODEL_PATH = "/content/drive/MyDrive/MiModeloBERTGuardadoVC"
print(f"Directorio existe: {os.path.exists(MODEL_PATH)}")
tokenizer = AutoTokenizer.from_pretrained(MODEL_PATH, local_files_only=
    True)
model = AutoModelForTokenClassification.from_pretrained(MODEL_PATH,
    local_files_only=True)
ner_pipe = pipeline(
    "ner",
    model=model,
    tokenizer=tokenizer,
    aggregation_strategy="simple",
    batch_size=8)
```

---

[17] IBM. *What is Named Entity Recognition (NER)?* Consultado el 31 de marzo de 2026. s.f. URL: https://www.ibm.com/mx-es/think/topics/named-entity-recognition.

[18] Michael BRENNDOERFER. *BIO Tagging: Encoding Entity Boundaries for Sequence Labeling*. Consultado el 31 de marzo de 2026. mbrenndoerfer.com. 2025. URL: https://mbrenndoerfer.com/writing/bio-tagging-sequence-labeling-ner.

3.6. EXTRACCIÓN DE INFORMACIÓN

Combinando NER con reglas heurísticas, es decir, criterios prácticos basados en patrones y proximidad que permiten establecer relaciones entre entidades cercanas, se logra vincular las celdas con sus parámetros de desempeño: voltaje de circuito abierto (V_{OC}), densidad de corriente de cortocircuito (J_{SC}), factor de forma (FF) y eficiencia de conversión de potencia (PCE). Además, se incorpora información complementaria extraída de tablas e imágenes, con el objetivo de maximizar la cantidad de datos recopilados y construir un *dataset* estructurado a partir de los artículos procesados.

3.7. VISUALIZACIÓN DE INFORMACIÓN

Finalmente, se lleva a cabo el diseño y ejecución de una interfaz de visualización para los datos extraidos usando el modelo. Dicha interfaz se diseña haciendo uso *HTML* para su estructura general, *CSS* para su estructura visual y *JavaScript* para su funcionalidad y dinamismo. La aplicación se desarrolla en el entorno de *Visual Studio Code*, aprovechando herramientas tales como su autocompletado, su *Modo Agente* y su corrección sintáctica, previniendo así errores durante la ejecución.

4. RESULTADOS

4.1. SELECCIÓN DEL MODELO

A partir de las pruebas iniciales con validación cruzada, se obtuvieron los resultados mostrados en las Tablas 4.2 y 4.3, correspondientes al análisis por *folds* con incertidumbre y al promedio global (*cross-validation average*), respectivamente.

Tabla 4.1. Resultados de evaluación por fold (k-fold) para los modelos

Fold (k)	Modelo	Precisión	Recall	F1
k=1	BERT	0.9003	0.9558	0.9272
k=2	BERT	0.8801	0.9438	0.9108
k=3	BERT	0.9073	0.9545	0.9303
k=4	BERT	0.8936	0.9604	0.9258
k=5	BERT	0.9105	0.9456	0.9277
k=1	MatSciBERT	0.9049	0.9566	0.9300
k=2	MatSciBERT	0.8992	0.9593	0.9283
k=3	MatSciBERT	0.9256	0.9575	0.9413
k=4	MatSciBERT	0.9061	0.9612	0.9329
k=5	MatSciBERT	0.9106	0.9399	0.9250
k=1	RoBERTa	0.8979	0.9480	0.9222
k=2	RoBERTa	0.9092	0.9409	0.9247
k=3	RoBERTa	0.9039	0.9464	0.9247
k=4	RoBERTa	0.9152	0.9523	0.9334
k=5	RoBERTa	0.8898	0.9606	0.9238

A partir de los resultados obtenidos para cada validación k-fold por modelo, presentados en la Tabla 4.1, se calculó el promedio y la incertidumbre (SEM), cuyos valores se muestran en la Tabla 4.2.

Tabla 4.2. Resultados promedio con incertidumbre (SEM) para los modelos BERT, MatSciBERT y RoBERTa utilizando validación cruzada de cinco folds en la tarea de reconocimiento de entidades nombradas (NER)

Métrica	**BERT**	**MatSciBERT**	**RoBERTa**
Precisión	$0{,}8984 \pm 0{,}0054$	$0{,}9093 \pm 0{,}0045$	$0{,}9032 \pm 0{,}0044$
Recall	$0{,}9520 \pm 0{,}0032$	$0{,}9549 \pm 0{,}0038$	$0{,}9496 \pm 0{,}0033$
F1	$0{,}9244 \pm 0{,}0035$	$0{,}9315 \pm 0{,}0028$	$0{,}9258 \pm 0{,}0020$

Tabla 4.3. Métricas de evaluación obtenidas mediante validación cruzada promedio (k-fold) para los modelos BERT, MatSciBERT y RoBERTa después del proceso de ajuste fino

Métrica	**BERT**	**MatSciBERT**	**RoBERTa**
Precisión	0.9072	0.9027	0.8997
Recall	0.9679	0.9537	0.9646
F1	0.9366	0.9275	0.9310

Se observa un rendimiento similar entre los tres, donde los mejores resultados en la validación cruzada promedio y el error estándar de la media (SEM) para Eval_f1 por una leve variación los obtiene BERT; de igual manera se graficó la evolución de la función de pérdida durante el entrenamiento y validación (Ver Figura 4.1) observando como la perdida disminuye progresivamente y el comportamiento de las métricas de evaluación a lo largo de las épocas (Ver Figura 4.2) mejora gradualmente hasta alcanzar valores estables.

Figura 4.1. Curva de aprendizaje: pérdida de entrenamiento vs pérdida de validación

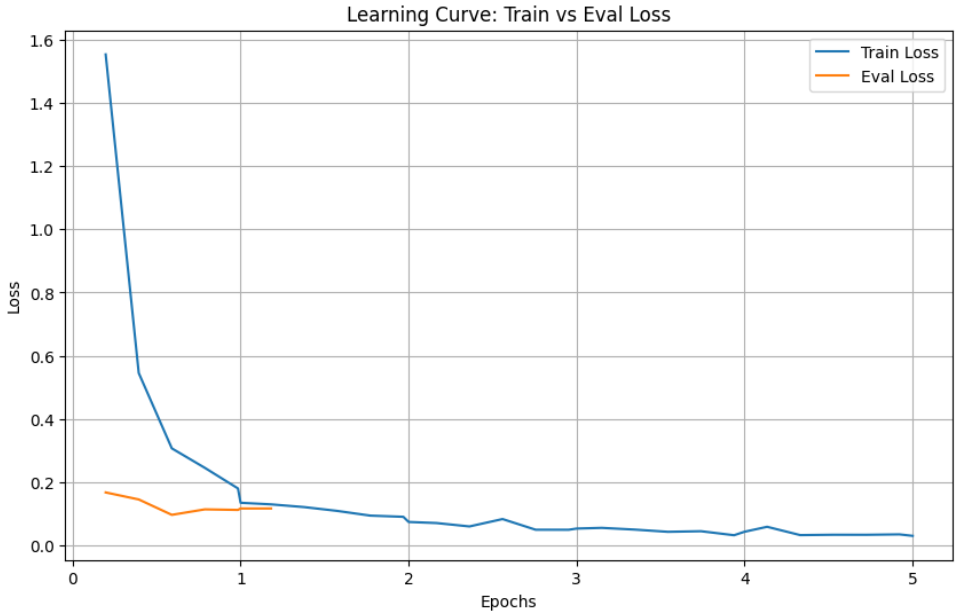
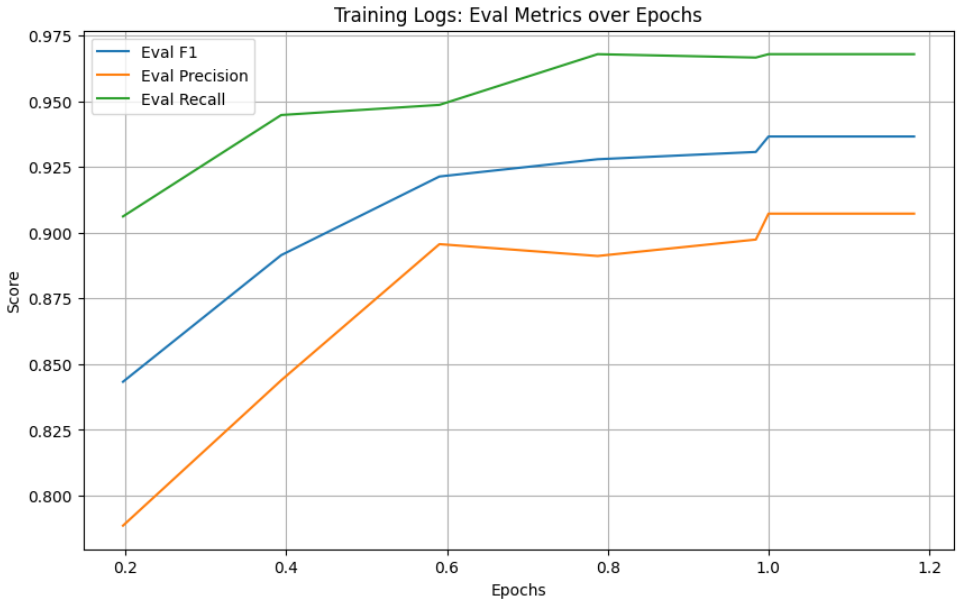


Figura 4.2. Registro de entrenamiento: métricas de evaluación a lo largo de las épocas



4.2. EXTRACCIÓN DE INFORMACIÓN

A partir del modelo seleccionado y por medio de una serie de reglas heurísticas con diferentes fines desde agrupar parámetros de una muestra a partir de proximidad (con una distancia máxima entre entidades de 35) hasta seleccionar los materiales y métodos utilizados en el artículo a través de la identificación de palabras clave, incluyendo la extracción de las imágenes y tablas se obtiene como consecuencia la tabla **??** donde se resumen los resultados de nuestro trabajo de grado que se almacenan en perovskite_dataset_fair.json para su posterior visualización.

Tabla 4.4. Procesos de extracción de información implementados y mecanismos de almacenamiento

Resultado	**Proceso**	**Herramienta**	**Almacenamiento**
Entidades (PCE, JSC, VOC, FF, PESK, HTL y ETL)	Extracción de entidades	`extract_entities(text, doi)`	`perovskite_dataset_fair.json`
Muestra por celda	Reglas heurísticas	`apply_heuristics(entities, text)`	`perovskite_dataset_fair.json`
Materiales (sustrato, perovskita, ETL, HTL y electrodo)	Entidades y reglas heurísticas	`extract_materials(text, entities)`	`perovskite_dataset_fair.json`
Métodos (síntesis y fabricación)	Reglas heurísticas	`extract_methods(text)`	`perovskite_dataset_fair.json`
Tablas	Table Transformer	`extract_tables(.)`	Carpeta Tablas
DOI	`re.search()`	`extract_doi(raw_text)`	`perovskite_dataset_fair.json`
Metadatos	PyPDF2	`extract_metadata(reader)`	`perovskite_dataset_fair.json`
Imágenes	pdfplumber	`extract_images_from_pdf(.)`	Carpeta Imágenes

4.3. VISUALIZACIÓN

Con la finalidad de conseguir una visualización más fluida, a la interfaz se le separó en 4 secciones de visualización: la vista por celda, donde se encuentra la información de cada celda individual obtenida; la vista matriz, donde se puede observar la información extraída de cada documento, tanto a nivel general como de celda individual; la vista imágenes, donde se tiene acceso a las imágenes de cada documento; y vista tablas con una funcionalidad similar a la anterior, pero para las tablas. La interfaz posee también una sección que solo se abre al seleccionar alguno de los valores dentro de la interfaz para revelar información adicional, por ejemplo, al seleccionar alguno de los valores en las vistas por celda o matriz, esta nueva ventana muestra la información del documento de donde se extrajo, junto con la demás información de esa celda (en caso de pertenecer a una celda específica), además también permite visualizar el contenido de las vistas imágenes y tablas de manera que no se saturen dichas secciones. Como un complemento, para ayudar con el filtrado de información, la interfaz posee una sección que permite reducir la búsqueda a una serie de valores o rangos deseados.

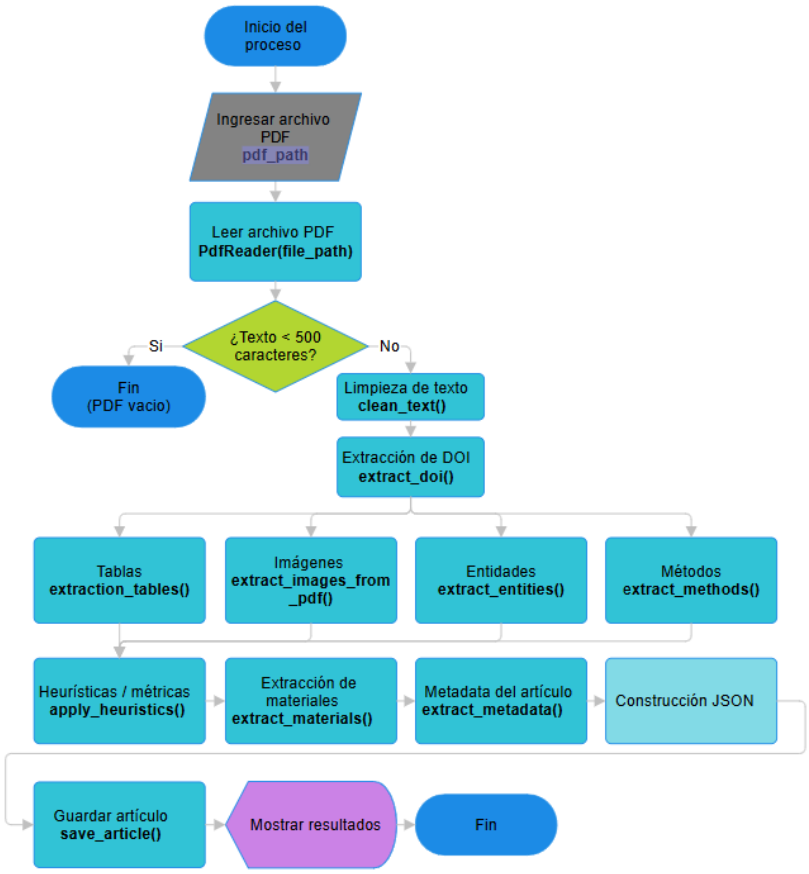
El funcionamiento de la aplicación es simple, utilizando el botón de *Cargar entidades (.json)*, se sube el archivo *.json* resultado del modelo, de manera similar con *Cargar carpeta Imágenes* y *Cargar carpeta Tablas*, se realiza el filtrado respectivo y, en caso de desear exportar la información filtrada, se presiona alguno de los dos botones de exportar, según el formato en que se desee. A la interfaz de visualización se puede acceder mediante: `https://gamjp.github.io/Perovskite-Visualizer/`

Adicionalmente, se puede encontrar el manual de uso con instrucciones más específica y mayor información acerca del uso de la aplicación en: `https://github.com/GamJP/Perovskite-Visualizer/blob/main/Manual_App_Visualizacion.pdf`

5. DISCUSIÓN

Con base en la metodología planteada y los resultados obtenidos, se propone el diagrama de flujo de la figura 5.1 como una síntesis del proceso general de extracción y estructuración de información a partir de documentos PDF. Este esquema resume las etapas principales del pipeline, aunque corresponde a una representación simplificada que omite funciones auxiliares de preprocesamiento, validación y control de errores.

Figura 5.1. Diagrama de flujo del proceso general de extracción de información desde documentos PDF. Se muestran las etapas principales del sistema, sin incluir funciones auxiliares de soporte.



A pesar de ello, el diagrama permite evidenciar la modularidad del sistema; sin embargo, los resultados indican que el rendimiento global no depende únicamente de esta arquitectura, sino principalmente del modelo de extracción empleado. En particular, su capacidad para identificar entidades y relaciones condiciona la calidad del resultado final. En este sentido, la separación del proceso en etapas (preprocesamiento, reconocimiento de entidades, aplicación de heurísticas y visualización) facilita tanto la interpretación del sistema como su posible extensión a otros dominios de aplicación.

Teniendo en cuenta lo anterior, se seleccionó el modelo BERT debido a su mayor estabilidad y desempeño global. En términos de validación cruzada (Ver Tabla 4.2), BERT presenta menor incertidumbre (SEM) en métricas relevantes como el F1-score, lo que indica una mayor estabilidad en el desempeño entre los distintos folds de validación (por ejemplo, F1 promedio de 0.9244). Esta estabilidad resulta especialmente relevante en aplicaciones donde la consistencia del modelo es prioritaria frente a pequeñas mejoras en métricas promedio.

Aunque MatSciBERT alcanza un F1 promedio ligeramente superior (0.9315), esta diferencia no resulta concluyente al considerar la variabilidad de los resultados. Además, en la evaluación final (Ver Tabla 4.3), BERT obtiene el mayor F1 (0.9366), superando a MatSciBERT (0.9275) y RoBERTa (0.9310), lo que evidencia un mejor balance entre precisión y recall, siendo este el criterio predominante en la evaluación del modelo. No obstante, la cercanía entre los resultados sugiere que el desempeño está fuertemente condicionado por el dominio y la calidad de los datos más que por la arquitectura específica del modelo.

A pesar del rendimiento del modelo, todavía se extraían valores que no correspondían a los correctos. Para mitigar este problema se aplicaron funciones auxiliares encarga-

das de limpiar, normalizar y estructurar la información; sin embargo, aún se presentan errores en ciertos casos. Este comportamiento puede atribuirse a la alta variabilidad en la forma en que los artículos científicos reportan la información, incluyendo diferencias en unidades, formatos y estilos de redacción, lo cual representa uno de los principales desafíos en tareas de PLN aplicadas a literatura científica.

En relación con la extracción de tablas, se presentaron inconvenientes debido a la gran diversidad de formatos presentes en los artículos. Inicialmente, se intentó aplicar reglas heurísticas, las cuales funcionaban correctamente solo en escenarios específicos, limitando su capacidad de generalización. Para abordar esta limitación, se utilizó Table Transformer, un modelo desarrollado por Microsoft que permite detectar y extraer tablas en documentos no estructurados, particularmente cuando estas se encuentran embebidas como imágenes [19]. Si bien esta solución mejora la capacidad de extracción, también introduce un mayor costo computacional y no garantiza resultados completamente precisos en todos los casos.

Por otra parte, el uso de reglas heurísticas para establecer relaciones entre entidades demuestra ser una estrategia efectiva en escenarios controlados; no obstante, su dependencia de patrones predefinidos y proximidad textual limita su desempeño en textos complejos o con estructuras no convencionales. Esto sugiere la necesidad de explorar enfoques más robustos, como el sintonizado fino de modelos grandes para la tarea de detección de relación entre entidades; pero, para ello se requiere contar con etiquetas apropiadas diseñadas manualmente.

Como consecuencia de lo anterior, surge la necesidad de comparar este trabajo con

---

[19] Brandon SMOCK, Rohith PESALA y Robin ABRAHAM. «PubTables-1M: Towards Comprehensive Table Extraction From Unstructured Documents». En: *arXiv preprint arXiv:2110.00061* (2021).

herramientas de inteligencia artificial como ChatGPT o Grok. Estos modelos han demostrado un mejor desempeño en la extracción de información científica, destacándose por su capacidad para interpretar el contexto y estructurar datos de manera flexible. Sin embargo, presentan limitaciones importantes, como su dependencia de prompts bien diseñados, la falta de reproducibilidad en algunos casos y la ausencia de mecanismos nativos para almacenamiento estructurado y visualización de resultados.

En contraste, el sistema propuesto en este trabajo ofrece ventajas como la generación automática de un dataset estructurado, la integración con una plataforma web de visualización y la capacidad de operar sin requerir prompts. Estas características lo convierten en una alternativa reproducible y orientada a aplicaciones prácticas, aunque aún existen oportunidades de mejora en términos de precisión, eficiencia y capacidad de generalización.

6. TRABAJOS FUTUROS

A partir de las limitaciones identificadas, se plantean diversas líneas de trabajo que permitirían mejorar y ampliar el alcance del sistema propuesto.

En primer lugar, se propone la incorporación de modelos de extracción de relaciones basados en aprendizaje profundo, con el fin de superar las limitaciones de las reglas heurísticas actuales y mejorar la precisión en la asociación entre entidades. Pero, antes de ello se ha de contar con un dataset para relación de entidades.

Finalmente, se propone la expansión del sistema hacia otros dominios de la ciencia de materiales, evaluando su capacidad de generalización y adaptabilidad a diferentes tipos de literatura científica.

7. CONCLUSIONES

El presente trabajo permitió cumplir con el objetivo de implementar un sistema basado en técnicas de Procesamiento de Lenguaje Natural (PLN) para la extracción automática de información relevante en artículos científicos sobre celdas solares de perovskita, logrando identificar y estructurar parámetros fotovoltaicos clave como Voc, Jsc, FF y PCE.

En relación con la adaptación de modelos de lenguaje, los resultados evidencian que BERT ofrece un mejor equilibrio entre desempeño y estabilidad frente a modelos como MatSciBERT y RoBERTa, destacándose por su menor variabilidad en validación cruzada y un F1 competitivo. Esto sugiere que, en este dominio, la consistencia del modelo puede ser más determinante que mejoras marginales en métricas promedio.

La implementación de reglas heurísticas permitió establecer relaciones entre entidades de manera efectiva; sin embargo, se identificaron limitaciones en escenarios donde la información se presenta de forma dispersa o en estructuras no convencionales, lo que impacta la precisión del sistema.

Asimismo, la integración de un sistema de visualización y la generación de un dataset estructurado constituyen un aporte relevante, al facilitar la exploración y reutilización de la información extraída a partir de documentos no estructurados.

BIBLIOGRAFÍA

BEARD, E. J. y J. M. COLE. «Perovskite- and Dye-Sensitized Solar-Cell Device Databases Auto-generated Using ChemDataExtractor». En: *Sci. Data* 9.1 (jun. de 2022), pág. 329. DOI: 10.1038/s41597-022-01355-w (vid. pág. 17).

BRENNDOERFER, Michael. *BIO Tagging: Encoding Entity Boundaries for Sequence Labeling*. Consultado el 31 de marzo de 2026. mbrenndoerfer.com. 2025. URL: https://mbrenndoerfer.com/writing/bio-tagging-sequence-labeling-ner (vid. pág. 23).

GARDAZI, Nadia Mushtaq et al. «BERT applications in natural language processing: a review». En: *Artificial Intelligence Review* 58 (2025). Accepted: 19 February 2025 / Published online: 15 March 2025, pág. 166. DOI: 10.1007/s10462-025-11162-5 (vid. pág. 20).

GUPTA, T. et al. «MatSciBERT: A materials domain language model for text mining and information extraction». En: *NPJ Comput. Mater.* 8.1 (mayo de 2022), pág. 102. DOI: 10.1038/s41524-022-00784-w (vid. págs. 12, 13, 21).

HUGGING FACE. *BERT — Transformers Documentation*. Consultado el 31 de marzo de 2026. 2026. URL: https://huggingface.co/docs/transformers/model_doc/bert (vid. pág. 22).

IBM. *What is Named Entity Recognition (NER)?* Consultado el 31 de marzo de 2026. s.f. URL: https://www.ibm.com/mx-es/think/topics/named-entity-recognition (vid. pág. 23).

INTERNATIONAL ENERGY AGENCY. *Electricity – Renewables 2023 – Analysis*. https://www.iea.org/reports/renewables-2023/electricity. [En línea]. Accedido: 18 de abril de 2025 (vid. pág. 11).

JACOBSSON, T. J. et al. «An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles». En: *Nature Energy* 7.1 (2022), págs. 107-115. DOI: 10.1038/s41560-021-00941-3 (vid. pág. 19).

JIMÉNEZ DÍAZ, M. Z. «Herramienta de software para la extracción automática de parámetros de desempeño a partir de publicaciones científicas de celdas solares de perovskita». Trabajo de grado. Bucaramanga, Colombia: Universidad Industrial de Santander, 2022 (vid. págs. 18, 21).

JIMÉNEZ-DÍAZ, M. Z., A. SEPÚLVEDA-SEPÚLVEDA y M. BOTERO-LONDOÑO. «Named Entity Recognition for Performance and Synthesis Information of Perovskite Solar Cells Using SpaCy». En: *Computational Science and Its Applications – ICCSA 2025*. Vol. 15649. Lecture Notes in Computer Science. Cham: Springer, 2025. DOI: 10.1007/978-3-031-96997-3_19 (vid. pág. 18).

LEE, J. H., M. LEE y K. MIN. «Natural Language Processing Techniques for Advancing Materials Discovery: A Short Review». En: *International Journal of Precision En-*

gineering and Manufacturing-Green Technology 10.5 (2023), págs. 1337-1349. DOI: 10.1007/s40684-023-00523-6 (vid. págs. 16, 19).

LIU, Yinhan et al. «RoBERTa: A Robustly Optimized BERT Pretraining Approach». En: *arXiv preprint arXiv:1907.11692* (2019). URL: https://arxiv.org/abs/1907.11692 (vid. pág. 21).

POLAK, M. P. et al. «Extracting accurate materials data from research papers with conversational language models and prompt engineering». En: *Nature Communications* 15.1 (2024), pág. 1569. DOI: 10.1038/s41467-024-45914-8 (vid. pág. 19).

SAYEED, H. M. et al. «NLP meets materials science: Quantifying the presentation of materials data in literature». En: *Matter* 7.3 (2024), págs. 723-727. DOI: 10.1016/j.matt.2023.12.032 (vid. pág. 18).

SCIKIT-LEARN DEVELOPERS. *Cross-validation: evaluating estimator performance*. Consultado el 31 de marzo de 2026. s.f. URL: https://scikit-learn.org/stable/modules/cross_validation.html (vid. pág. 22).

SHETTY, P. et al. «A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing». En: *npj Computational Materials* 9.1 (2023), pág. 52. DOI: 10.1038/s41524-023-01003-w (vid. pág. 17).

SMOCK, Brandon, Rohith PESALA y Robin ABRAHAM. «PubTables-1M: Towards Comprehensive Table Extraction From Unstructured Documents». En: *arXiv preprint arXiv:2110.00061* (2021) (vid. pág. 32).

The Wikipedia of perovskite solar cell research. https://www.pv-magazine.com/2021/12/28/the-wikipedia-of-perovskite-solar-cell-research/. pv magazine International, [En línea]. Accedido: 18 de abril de 2025 (vid. pág. 13).

XIE, T. et al. «Creation of a structured solar cell material dataset and performance prediction using large language models». En: *Patterns* 5.5 (mayo de 2024), pág. 100955. DOI: 10.1016/j.patter.2024.100955 (vid. págs. 11, 18).