Maestría en

CIENCIAS DE DATOS Y MÁQUINAS DE APRENDIZAJE CON MENCIÓN EN INTELIGENCIA ARTIFICIAL

Trabajo previo a la obtención de título de Magister en Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial

AUTORES:

HARO SARANGO ALEXANDER FERNANDO

MOYA GONZÁLEZ VIVIANA ISABEL

SALAZAR MENDIZÁBAL GABRIEL IGNACIO

QUITO CARRIÓN FABIÁN DARÍO

TUTORES:

Iván Reyes Chacón

Alejandro Cortés López

TEMA:

Análisis Semántico y de Tendencias Investigativas: Modelado Interactivo con Procesamiento de Lenguaje Natural

Quito - Ecuador

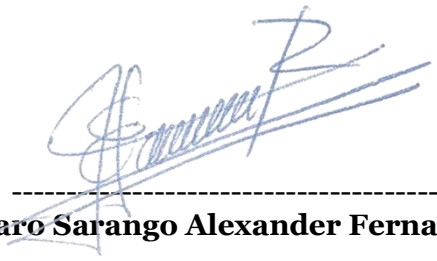
Julio – 2025

REINVENTEMOS EL FUTURO

Certificación de autoría

Nosotros, Haro Sarango Alexander Fernando, Moya González Viviana Isabel, Salazar Mendizábal Gabriel Ignacio y Quito Carrión Fabián Darío, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.



Haro Sarango Alexander Fernando

Moya González Viviana Isabel

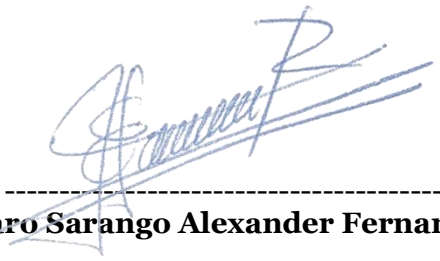
Salazar Mendizábal Gabriel Ignacio

Quito Carrión Fabián Darío

Autorización de Derechos de Propiedad Intelectual

Nosotros, Haro Sarango Alexander Fernando, Moya González Viviana Isabel, Salazar Mendizábal Gabriel Ignacio y Quito Carrión Fabián Darío, en calidad de autores del trabajo de investigación titulado *Análisis Semántico y de Tendencias Investigativas: Modelado Interactivo con Procesamiento de Lenguaje Natural*, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, Julio 2025



Haro Sarango Alexander Fernando

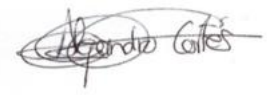
Moya González Viviana Isabel

Salazar Mendizábal Gabriel Ignacio

Quito Carrión Fabián Darío

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés e Iván Reyes ,** declaramos que: Haro Sarango Alexander Fernando, Moya González Viviana Isabel, Salazar Mendizábal Gabriel Ignacio y Quito Carrión Fabián Darío son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Mg. Alejandro Cortés
Director de la
Maestría en Ciencias de Datos
y Máquinas de Aprendizaje con
Mención en Inteligencia Artificial

Mg. Iván Reyes
Coordinador de la
Maestría en Ciencias de Datos
y Máquinas de Aprendizaje con
Mención en Inteligencia Artificial

DEDICATORIA

Este trabajo está dedicado a todos los compañeros que nos acompañaron durante esta maestría. Ha sido una experiencia enriquecedora conocer, en cada materia, a personas que comparten la misma pasión por los datos. También va dirigido a los futuros usuarios de la comunidad científica, quienes podrán aprovechar la aplicación desarrollada para fortalecer sus investigaciones y orientarlas hacia la innovación. Finalmente, a todos los docentes de la maestría, quienes supieron compartir sus conocimientos de forma clara y generosa, contribuyendo a nuestra formación como nuevos másteres en la disciplina.

AGRADECIMIENTOS

Agradecemos profundamente a nuestros familiares por su comprensión y apoyo incondicional. Esta maestría ha requerido una importante inversión de tiempo, y ellos supieron comprenderlo y estar presentes en los momentos de mayor exigencia.

A nuestra tutora, Lina, por su constante acompañamiento, coordinación y seguimiento a lo largo de todo el período académico. Su apoyo ha sido clave en el desarrollo de este trabajo.

Y a nuestros amigos, por entender que las clases eran una prioridad, y por saber esperar los momentos en que sí podíamos compartir con ellos.

RESUMEN

Esta plataforma web de análisis bibliométrico, construida íntegramente en Python y desplegada en Streamlit, automatiza la exploración de metadatos descargados exclusivamente de Scopus y guía al usuario a través de seis módulos interdependientes que cubren el ciclo analítico completo. Tras un acceso autenticado, la Sección 1 permite cargar archivos CSV, obtener estadísticas descriptivas, aplicar filtros dinámicos y exportar subconjuntos limpios. La Sección 2 calcula indicadores clásicos de productividad, genera gráficas interactivas con Plotly y construye, mediante NetworkX, una red de coautoría centrada en los veinte autores más prolíficos. En la Sección 3 se ejecuta un flujo de PLN: los resúmenes se normalizan con NLTK, se vectorizan con TF-IDF y se modelan con NMF para descubrir temas; los documentos se proyectan sobre un mapa t-SNE que facilita la inspección visual de clústeres. La Sección 4 incorpora análisis afectivo usando VADER y TextBlob y ofrece histogramas, heatmaps año-polaridad y diagramas 3-D que relacionan sentimiento, subjetividad y fecha de publicación. La Sección 5 mide la similitud global entre artículos, produce nubes de palabras, agrupa con K-Means y PCA y despliega un explorador LDA interactivo basado en PyLDAvis. Posteriormente, la Sección 6 rastrea términos emergentes: calcula pendientes de TF-IDF por año, sugiere líneas de investigación prometedoras y visualiza tendencias mediante gráficas lineales y nubes enfocadas en los tres años recientes. El empleo exclusivo de Scopus garantiza consistencia de cobertura y simplifica la gestión de licencias. Su interfaz responsiva promueve adopción institucional y fomenta la práctica colaborativa de ciencia abierta regional en Latinoamérica.

Palabras Claves: Bibliometría; Scopus; Minería de texto; Visualización interactiva; Tópicos emergentes

ABSTRACT

This bibliometric-analysis web platform, built entirely in Python and deployed with Streamlit, automates the exploration of metadata extracted exclusively from Scopus and guides the user through six interconnected modules that embrace the full analytical cycle. After an authenticated login, Section 1 enables CSV upload, delivers descriptive statistics, offers dynamic filters, and exports clean subsets. Section 2 computes classical productivity indicators, draws interactive Plotly charts, and, with NetworkX, constructs a co-authorship network focused on the twenty most prolific authors. Section 3 runs an NLP pipeline: abstracts are normalized with NLTK, vectorized through TF-IDF, and modelled with NMF to uncover topics; the resulting documents are projected onto a t-SNE map for intuitive cluster inspection. Section 4 performs affective analysis using VADER and TextBlob and provides histograms, year-polarity heatmaps, and three-dimensional scatterplots linking sentiment, subjectivity, and publication date. Section 5 measures global article similarity, generates word clouds, groups papers with K-Means and PCA, and displays an interactive LDA explorer via PyLDAvis. Finally, Section 6 tracks emerging terms: it calculates yearly TF-IDF slopes, suggests promising research lines, and visualises trends through line charts and word clouds focused on the last three years. The Scopus-only approach guarantees consistent coverage, avoids cross-database duplication, and simplifies licence management, while the modular architecture based on free-software libraries secures scalability, reproducibility, and future adaptability. Integrated export links generate ready-to-share PDFs, DOCX spreadsheets, and CSV archives for downstream quantitative or qualitative assessment and decision-making.

Keywords: Bibliometrics; Scopus; Text mining; Interactive visualization; Emerging topics

TABLA DE CONTENIDOS

LISTA DE TABLAS

LISTA DE FIGURAS

CAPÍTULO I.

1. Introducción

La aceleración inédita en la producción de conocimiento –impulsada por la digitalización de procesos editoriales y la adopción masiva de repositorios de acceso abierto– ha desbordado la capacidad de los investigadores para identificar de manera oportuna vacíos temáticos, tendencias emergentes y redes de colaboración. Esta sobrecarga informativa, conocida como infoxicación, se agrava en áreas donde la evolución de la literatura es exponencial, lo que vuelve insuficientes las técnicas de revisión tradicional basadas en la lectura exhaustiva y el seguimiento manual de bases de datos bibliográficas.

En el contexto latinoamericano, y particularmente en Ecuador, la brecha entre la velocidad con que se generan nuevos hallazgos y la asimilación efectiva de estos por parte de instituciones académicas y centros de investigación es más marcada. Las universidades buscan optimizar sus estrategias de publicación, mejorar su posicionamiento en rankings regionales y alinear su producción con políticas de ciencia abierta, pero encuentran obstáculos en la dispersión de la información y la falta de herramientas que faciliten análisis de alto nivel con recursos limitados.

Frente a este desafío, surge la necesidad de diseñar una plataforma interactiva que integre técnicas avanzadas de Procesamiento de Lenguaje Natural (PLN) y análisis bibliométrico. Dicha plataforma tiene como propósito automatizar la ingestión y limpieza de metadatos, mapear la evolución temática mediante modelos de tópicos, y visualizar redes de coautoría y citación. Con ello se pretende ofrecer a investigadores, gestores y autoridades académicas un entorno unificado que facilite la toma de decisiones estratégicas basadas en evidencia, promueva la colaboración científica y fortalezca la visibilidad internacional de la producción nacional.

Esta iniciativa se enmarca en los lineamientos del Plan Nacional de Desarrollo, en las disposiciones del Consejo de Aseguramiento de la Calidad de la Educación Superior (CACES) y en los principios de la ciencia abierta, buscando contribuir a la democratización del acceso al análisis bibliométrico avanzado y a la consolidación de una cultura de gestión del conocimiento orientada a la mejora continua.

1.1. Definición del proyecto

El proyecto consiste en diseñar, desarrollar e implementar una plataforma digital integral que automatice el análisis bibliométrico y la minería de textos científicos para instituciones de educación superior en Ecuador y Latinoamérica. Su núcleo funcional combina algoritmos de Procesamiento de Lenguaje Natural (PLN) con métodos de modelado de tópicos y métricas bibliométricas, permitiendo procesar grandes volúmenes de metadatos provenientes de bases indexadas (Scopus). A partir de la ingestión y limpieza de estos registros, el sistema generará mapas dinámicos de coautoría, redes de citación, visualizaciones de evolución temática y cuadros comparativos de productividad académica, todo dentro de una interfaz web de uso intuitivo —desplegada en Streamlit o framework equivalente— que no requiere conocimientos avanzados de programación por parte del usuario final.

La iniciativa se concibe como una respuesta estratégica a la infoxicación científica y a las exigencias crecientes de acreditación y rendición de cuentas establecidas por el Consejo de Aseguramiento de la Calidad de la Educación Superior (CACES), la Ley Orgánica de Educación Superior y los lineamientos de ciencia abierta.

El proyecto se sustenta en principios de interoperabilidad, escalabilidad y software libre, de modo que su arquitectura pueda integrarse con sistemas académicos existentes y crecer conforme aumenten los volúmenes de datos o se sumen nuevas fuentes de información. Asimismo, contempla un componente de capacitación para investigadores, bibliotecólogos y gestores de I+D, orientado a asegurar la apropiación tecnológica y la

sostenibilidad a largo plazo. Entre los beneficiarios directos se encuentran docentes-investigadores, grupos de investigación, oficinas de posgrado y direcciones de planificación estratégica; de forma indirecta, la sociedad se beneficia al reforzarse las capacidades de generación y transferencia de conocimiento de las universidades.

1.2. Justificación e importancia del trabajo de investigación

El vertiginoso crecimiento de la literatura científica impulsado por la digitalización de los flujos editoriales y la expansión de repositorios de acceso abierto ha generado un volumen de información que rebasa la capacidad humana de lectura y síntesis. Esta saturación, conocida como infoxicación, afecta en mayor medida a las instituciones de educación superior latinoamericanas, cuyas limitaciones presupuestarias y de infraestructura tecnológica dificultan la implementación de procesos avanzados de vigilancia científica. Desarrollar una plataforma que automatice el análisis bibliométrico y la minería de textos responde, por tanto, a la necesidad urgente de superar la dependencia de revisiones manuales, optimizar el tiempo de los investigadores y garantizar decisiones estratégicas fundamentadas en evidencia.

La importancia del proyecto radica en su capacidad para democratizar el acceso a herramientas de Procesamiento de Lenguaje Natural y modelado de tópicos, tradicionalmente reservadas a centros con altos recursos. Al integrar estas tecnologías en una interfaz web intuitiva, el sistema permitirá a universidades, grupos de investigación y unidades de posgrado identificar oportunamente tendencias emergentes, vacíos temáticos y redes de colaboración, potenciando la generación de conocimiento pertinente y de alto impacto. Esta funcionalidad resulta crítica para mejorar el posicionamiento institucional en rankings académicos, fortalecer la competitividad de los programas de posgrado y atraer financiamiento externo a través de convocatorias cada vez más exigentes en términos de métricas de productividad y citación.

Desde el punto de vista normativo, la propuesta se alinea con los lineamientos del Consejo de Aseguramiento de la Calidad de la Educación Superior (CACES) y con la Ley Orgánica de Educación Superior, que exigen sistemas de gestión de la investigación basados en indicadores verificables. Asimismo, responde a los principios de la ciencia abierta y a compromisos internacionales como la Recomendación de la UNESCO sobre Ciencia Abierta (2021), al facilitar la transparencia y la reutilización de datos científicos. En el ámbito nacional, contribuye a los objetivos del Plan Nacional de Desarrollo, que prioriza la investigación, la innovación y la transferencia tecnológica como motores de desarrollo social y económico.

1.3. Alcance

El proyecto abarcará el diseño, desarrollo e implementación de una plataforma web de análisis bibliométrico y minería de textos científicos destinada a instituciones de educación superior de Ecuador y Latinoamérica. Comprenderá los siguientes componentes: (a) creación de flujos automáticos para la ingestión, depuración y normalización de metadatos provenientes de Scopus; (b) integración de algoritmos de Procesamiento de Lenguaje Natural y modelado de tópicos para identificar tendencias, vacíos temáticos y redes de colaboración; (c) visualización interactiva de los resultados mediante dashboards exportables. El alcance se limita al análisis de producción científica documentada y no contempla la evaluación cualitativa del contenido total de los artículos ni la implementación de sistemas de revisión por pares. El cronograma previsto cubre desde la fase de levantamiento de requisitos hasta la entrega de la versión estable y la capacitación inicial, dentro de un período de doce meses.

1.4. Objetivos

1.4.1. Objetivo general

Desarrollar e implantar una plataforma digital que automatice el análisis bibliométrico y la minería de textos científicos, facilitando la toma de decisiones estratégicas en investigación y fortaleciendo la visibilidad internacional de las instituciones de educación superior ecuatorianas y latinoamericanas

1.4.2. Objetivos específicos

- Diseñar un pipeline modular que permita la depuración y normalización de metadatos provenientes de Scopus.
- Implementar modelos de Procesamiento de Lenguaje Natural y aprendizaje automático para detectar tendencias emergentes y patrones de colaboración en la producción científica.
- Desarrollar recursos interactivos y exportables que presenten indicadores clave de productividad, impacto y coautoría de manera accesible para usuarios no técnicos.

CAPÍTULO II.

2. Revisión de literatura

2.1. Estado del Arte

Para este trabajo de fin de máster se presenta a continuación una revisión del estado del arte en el campo del análisis y clasificación de texto usando técnicas de machine learning. Esta revisión permite identificar las metodologías existentes, evaluar su efectividad y detectar los desafíos y posibilidades de mejoras a implementar. Dado que el objetivo del proyecto es diseñar una plataforma interactiva para el análisis semántico y la clasificación temática de literatura científica, es importante revisar los enfoques previos relacionados con la categorización de documentos, el modelado de tópicos y el procesamiento de lenguaje natural (PLN). La revisión que se presenta a continuación sigue un orden progresivo, partiendo de modelos clásicos de clasificación supervisada, pasando por técnicas de modelado probabilístico de tópicos, hasta llegar a soluciones recientes basadas en arquitecturas de aprendizaje profundo y modelos del tipo transformer.

Las primeras investigaciones de clasificación automática de documentos implementaron modelos estadísticos supervisados y técnicas de representación de texto relativamente simples, como el modelo de Bolsa de Palabras (BoW) y Frecuencia de Término - Frecuencia Inversa de Documento (TF-IDF). Un ejemplo de esta línea es el trabajo de Fatima et al. (2017), quienes implementaron un sistema de categorización de textos utilizando Máquinas Vector Soporte (SVM), entrenado sobre el conjunto de datos Reuters-21578. El proceso incluyó una fase de preprocesamiento estándar con eliminación de stopwords, stemming y extracción de características mediante TF-IDF. Los autores reportaron una precisión de hasta 86.39% en condiciones óptimas de partición entre entrenamiento y prueba, mostrando que los modelos lineales siguen siendo competitivos cuando se cuenta con un conjunto de datos bien estructurado y etiquetado. Así mismo en Campoverde *et al.* (2022) se presenta una revisión de la literatura, sobre investigaciones,

casos de aplicaciones y exploración de la logística inversa. Esta investigación realizó la taxonomía conformada por la relación de temas y términos encontrados en la literatura que analizaron.

En una línea similar, el trabajo de Dogra et al. (2022) presenta una revisión completa del proceso de clasificación de texto, iniciando en la representación del texto hasta llegar a los algoritmos de clasificación más recientes. El estudio compara enfoques tradicionales como Naïve Bayes, SVM y Random Forest con modelos avanzados de redes neuronales y transformers con mecanismos de atención. Aunque los modelos basados en transformers mostraron el mejor rendimiento general, los autores enfatizan su alta demanda computacional y la dependencia de grandes volúmenes de datos etiquetados, lo cual limita su aplicabilidad en dominios específicos o con recursos limitados.

Uno de los objetivos principales en la escritura de un artículo es saber que lo hace interesante con respecto a las demás investigaciones de la misma área. En Griffiths et al. (2004) se presenta un método estadístico para extraer automáticamente la información de los resúmenes de los documentos. Se usa el algoritmo Monte Carlo de cadenas de Markov para la inferencia del modelo. Se propuso ese algoritmo para analizar resúmenes de Proceedings of the National Academy of Sciences (PNAS) mediante la selección de modelos bayesianos para determinar el número de temas. Se demostró que los temas extraídos capturan una estructura significativa en los datos, en consonancia con las designaciones de clase proporcionadas por los autores de los artículos, y se describió otras aplicaciones de este análisis, como la identificación de temas de actualidad mediante el examen de la dinámica temporal y el etiquetado de resúmenes para ilustrar el contenido semántico.

Por otro lado, Ghumade & Deshmukh (2019) propusieron un sistema basado en Redes Neuronales Recurrentes (RNN) para la clasificación automática de documentos, utilizando como conjunto de datos resúmenes de artículos científicos en PDF. El modelo alcanzó una precisión del 97.5%, superando a otros algoritmos tradicionales como Naïve

Bayes y Random Forest. Sin embargo, al igual que los trabajos anteriores, este enfoque se basa en un conjunto de datos etiquetado de forma supervisada, lo cual implica un esfuerzo considerable en términos de anotación manual y curación de los datos.

Todos estos trabajos, utilizan en sus propuestas técnicas supervisadas que, aunque efectivas, presentan limitaciones como requerir conjuntos de datos bien etiquetados, incapacidad de capturar relaciones semánticas profundas entre los términos, y alta dependencia de la calidad del preprocesamiento del texto. Estas limitaciones han motivado la exploración de enfoques alternativos, como los modelos probabilísticos de tópicos, que se revisan a continuación.

En el campo de entrenamiento no supervisado para la clasificación de texto, el modelado de tópicos una de las técnicas más desarrolladas. Con este fin, el modelo Latent Dirichlet Allocation (LDA) ha sido ampliamente adoptado por su capacidad para descubrir automáticamente temas latentes dentro de un conjunto de datos documental.

Uno de los primeros trabajos en esta línea es el de Blei *et al.* (2003), quienes introdujeron el modelo LDA como una mejora significativa sobre enfoques anteriores como el Mixture of Unigrams y el Inventario de Habilidades Lingüísticas Pragmáticas (PLSI). LDA permite representar documentos como una distribución sobre tópicos, y cada tópico como una distribución sobre palabras, ofreciendo una forma estructurada y jerárquica de analizar grandes volúmenes de texto. Presenta ventajas como una mejor generalización a documentos no vistos y menor tendencia al sobreajuste.

Aplicaciones prácticas de LDA han sido documentadas en múltiples dominios. Por ejemplo, Shiryaev *et al.* (2017) presenta un método para detectar tendencias científicas y tecnológicas en el área del conocimiento técnico. Utilizando publicaciones temáticas obtenidas de fuentes web, se aplicaron modelos LDA optimizados con un criterio formal basado en la competitividad de búsquedas. El estudio combina evaluaciones expertas con métricas automáticas como la perplexity y demuestra que ambos métodos coinciden en la

calidad de los modelos generados. Este trabajo evidencia el potencial de LDA para el análisis automatizado de grandes volúmenes de texto y la identificación de temas emergentes.

Otra investigación presentada por Chang *et al.* (2021) realizó un estudio más completo, combinando minería de texto, análisis de clustering jerárquico y LDA para clasificar temáticamente artículos de revistas científicas en el campo de la educación ambiental. El enfoque incluyó la validación de coherencia de los tópicos y el uso de conocimiento experto para afinar la interpretación temática. El trabajo demuestra cómo LDA puede integrarse en pipelines más amplios de análisis semántico, incluyendo co-palabras y redes de términos. En el área de economía la obra de Detthamrong *et al.* (2024) examinó 8321 documentos de la base de datos Scopus, en donde se identificó tres temas distintos y rastreando sus tendencias de desarrollo. El estudio uso la tecnología de modelado de temas de Asignación Latente de Dirichlet (LDA) en el análisis de texto. Los resultados de esta investigación fue que la Transformación digital es el tema más popular con un 56,6 % de tokens, y digitalización representa el 21,2% de los tokens.

Otros autores han abordado la comparación entre LDA y técnicas relacionadas Mohammed & Al-Augby (2020) compararon LDA con Latent Semantic Analysis (LSA) en un estudio sobre libros electrónicos. Sus resultados mostraron que LDA superó a LSA en métricas de coherencia temática (UCI y UMass), especialmente cuando se incrementó el número de tópicos. Estos resultados son relevantes para la propuesta del presente proyecto, dado que respalda el uso de LDA como base para el modelado temático en plataformas interactivas.

En una aplicación particular a revisiones sistemáticas, Mo *et al.* (2015) utilizaron representaciones basadas en LDA para mejorar el proceso de cribado automático de literatura científica. Sus modelos superaron al clásico TF-IDF en métricas de Recall, especialmente en dominios médicos y de salud pública, lo que valida el potencial del

modelado de tópicos en tareas de exploración documental masiva como las que plantea este proyecto.

Además, se han propuesto extensiones del modelo LDA para incorporar estructuras semánticas adicionales. Un ejemplo es NET-LDA desarrollado por Ekinci & Omurca (2020), que utiliza redes de similitud semántica entre documentos basadas en entidades y conceptos extraídos con Babelfy. Este modelo logró una mejora del 6% al 40% en F-measure respecto a LDA, LTM y AMC, evidenciando que la integración de conocimiento semántico externo puede mejorar la coherencia de los temas generados.

De forma similar, Gupta & Katarya (2021) propusieron PAN-LDA, una variante diseñada para incorporar datos numéricos como la evolución de casos de COVID-19 junto con artículos de noticias. Esta combinación permitió mejorar la predicción de tendencias mediante modelos como LightGBM y XGBoost. La innovación de PAN-LDA radica en su capacidad para vincular dinámicamente la semántica textual con datos cuantitativos, lo cual es altamente relevante para contextos multidimensionales como el análisis bibliométrico.

Estos trabajos motivaron e informaron acerca de la mejor alternativa aplicable para la implementación del objetivo general de este proyecto: automatizar el análisis temático de grandes volúmenes de texto mediante enfoques no supervisados que reduzcan la carga de etiquetado manual y faciliten la exploración interactiva. Además, los trabajos revisados refuerzan la elección de LDA como modelo base para el análisis semántico, mientras que al mismo tiempo muestran que su efectividad puede ampliarse mediante enriquecimientos estructurales o adaptaciones específicas del dominio.

El modelado de tópicos con LDA también ha sido adaptado y extendido para tareas específicas como el análisis de sentimientos, el reconocimiento de entidades y la clasificación temática en dominios particulares. Trabajos centrados en esta área enriquecen la perspectiva del presente proyecto, al demostrar cómo LDA puede articularse con otros métodos y objetivos más allá de la simple agrupación temática.

Respecto al análisis de sentimientos, Farkhod *et al.* (2021) propuso el modelo TDS (Topic/Document/Sentence), una extensión de LDA combinada con la estructura JST (Joint Sentiment Topic), orientada a analizar la polaridad de textos a nivel de documento y tópico. Aunque su rendimiento fue desigual (alta precisión para la clase negativa, pero baja para la positiva), el trabajo destaca la posibilidad de incorporar el análisis de sentimientos al modelado temático.

Adicionalmente, Onan *et al.* (2016) llevó a cabo un estudio empírico sobre el uso de LDA para representar documentos en tareas de clasificación de sentimientos, comparando su efectividad con clasificadores como SVM, regresión logística y KNN. Además, evaluaron el uso de métodos de ensamble (Stacking, AdaBoost) y concluyeron que, si bien el número de tópicos no tiene un gran impacto en la precisión, sí mejora el F1-score. La combinación de LDA con métodos supervisados mejora la robustez del sistema, especialmente ante desequilibrios en los datos.

En el área del reconocimiento de entidades, Gangadharan & Gupta (2020) desarrollaron un sistema híbrido llamado AERTM (Agriculture Entity Recognition using Topic Modelling), orientado al dominio agrícola. Este sistema combinó el uso de un vocabulario controlado (AGROVOC) con un modelo LDA entrenado para identificar entidades como cultivos, fertilizantes o enfermedades. Se logró una precisión estimada del 80% mediante validación humana, mostrando que LDA puede ser útil incluso en tareas cercanas al etiquetado semántico cuando no se cuenta con conjuntos de datos anotados.

Un campo emergente de aplicación es la detección de noticias falsas. Kumar & Singh (2022) abordaron esta problemática con un modelo basado en LSTM, regresión logística y Naïve Bayes, evaluado sobre un conjunto de datos de documentos en hindi. Aunque este trabajo no emplea LDA, es relevante porque resalta los desafíos idiomáticos y la necesidad de técnicas robustas de representación textual para idiomas menos cubiertos. Los resultados

mostraron que LSTM alcanzó una precisión del 92.36%, reafirmando el potencial de modelos secuenciales para tareas de clasificación binaria.

En el ámbito biomédico, Kesiku *et al.* (2022) realizaron una revisión sistemática sobre técnicas de PLN aplicadas a la clasificación de textos clínicos. Se destacó la efectividad de modelos basados en BERT entrenados sobre datos especializados (como BioBERT), así como la dificultad para adaptar modelos generales a terminologías médicas. Aunque LDA fue menos utilizado en este dominio, el estudio demuestra la creciente necesidad de enfoques semánticos ajustados a contextos altamente técnicos, como es también el caso de la producción científica académica.

En conjunto, estos trabajos muestran que LDA, además de su capacidad como herramienta de agrupación temática, puede adaptarse a una amplia gama de tareas específicas, lo cual refuerza su idoneidad como base metodológica para plataformas orientadas al análisis automatizado de literatura científica.

Para terminar esta revisión, se destaca la evolución del procesamiento de lenguaje natural en la última década, marcada por el auge del aprendizaje profundo, lo que ha permitido el desarrollo de modelos altamente representativos y contextuales. Si bien el proyecto propuesto se fundamenta en el uso de LDA por su eficiencia y transparencia, es importante considerar los avances recientes en arquitecturas más complejas, que complementan y extienden sus capacidades.

En el campo de la aplicación del Deep learning para clasificación documental, se tiene el trabajo de Abdulwahab *et al.* (2020), que compararon un modelo CNN con variantes de LDA (tradicional y con TF-IDF). Utilizando el dataset 20 Newsgroups, encontraron que CNN alcanzó una precisión del 94.88%, significativamente superior al 74.4% obtenido con LDA modificado. Aunque el costo computacional fue mayor, los resultados avalan el uso de redes neuronales convolucionales en tareas de clasificación multiclase de alto volumen.

Recientemente, Setiadi *et al.* (2024) propuso una evaluación comparativa de métodos de aprendizaje automático y profundo para el análisis de sentimientos basado en aspectos (ABSA), integrando el modelado de temas mediante LDA sobre reseñas de productos de Amazon. Un desafío clave identificado fue la exclusión de modelos Transformer debido a limitaciones computacionales, lo cual restringió la comparación con arquitecturas de última generación.

Desde una perspectiva funcional, Gamaleldin *et al.* (2025) propone un modelo híbrido de CNN y LDA que combina una red neuronal convolucional profunda y un análisis discriminante lineal para investigar el riesgo de antiselección en los mercados de seguros. El modelo mejora las evaluaciones de riesgos utilizando datos extensos de fuentes de big data de compañías de seguros y algoritmos avanzados de aprendizaje automático. Esto mejora la detección de tendencias de antiselección y mejora las técnicas generales de gestión de riesgos, sin embargo, requiere acceso a datos privados de las personas, así como grandes recursos computacionales.

Nuestro método identifica un conjunto de temas expresados por los documentos, proporcionando medidas cuantitativas que permiten analizar su contenido, rastrear cambios en el contenido a lo largo del tiempo y evaluar la similitud entre ellos. Por tanto, aunque el uso de deep learning y transformers marca la frontera tecnológica actual, su aplicación debe evaluarse críticamente en función del objetivo, recursos disponibles y necesidad de interpretabilidad. En este sentido, la propuesta de este trabajo, basada en LDA, visualización interactiva y mapeo semántico, se justifica no solo por su solidez teórica y adaptabilidad, sino también por su capacidad de ofrecer resultados comprensibles, escalables y aplicables en contextos educativos y científicos reales.

2.2. Marco Teórico

Actualmente, la producción científica ha tenido un desarrollo exponencial que trasciende la capacidad humana de revisión y síntesis manual de la información. Por lo que

cada vez es necesario la creación de herramientas automatizadas para procesar grandes volúmenes de datos, buscar patrones temáticos, análisis de tendencias y ayudar a la toma de decisiones. López-Pérez & Dolores Olvera-lobo (2018) sostienen que la transformación digital generada a partir de la expansión del Internet y la Web 2.0, ha democratizado y permitido el acceso a la información científica impulsando el conocimiento científico, su socialización y abriendo lo posible de una ciencia más colaborativa, abierta de participativa.

En el marco de este nuevo paradigma, el recurso que se hace más habitual a la hora de proceder al análisis semántico y de tendencias mediante tecnologías emergentes ha sido la puesta en práctica de estrategias de respuesta a los retos de la educación superior, la interdisciplinariedad y la creación de conocimiento contextualizado. Como se plantea del Zulia Venezuela *et al.* (2020) la interacción entre el ámbito científico, innovación y emprendimiento requiere coordinación entre el sector público y el privado, y el académico, esto último solo se puede hacer cuando se encuentran mecanismos suficientes de gestión de conocimiento.

De modo semejante, cabe resaltar que el estudio de tendencias se ha asentado como herramienta metodológica para el seguimiento del desarrollo evolutivo sobre ciertas facetas del trabajo de investigación. Concisamente, permite saber qué contenidos están en crecimiento, cuáles están en pausa y qué se esté poco a poco percibiendo en ciertos ámbitos del dominio del saber. Además, ofrece una visión mundial de la productividad académica, mostrando patrones de colaboración, concentración temática y lagunas de investigación. Por lo tanto, las herramientas fundadas en la minería de texto y en visualización interactiva han ganado cada vez más relieve en virtud de entornos académicos, con más significación cuando están unidas para con bases de datos bibliográficos como Scopus.

En efecto, el análisis de tendencias en publicaciones científicas no solo permite reconocer patrones temáticos, sino que también ofrece la posibilidad de vincular dicho conocimiento con procesos de innovación institucional. Por ejemplo, al identificar líneas de

investigación en expansión, una universidad puede redirigir recursos, actualizar su oferta de posgrados o fomentar grupos de investigación en áreas emergentes. Del mismo modo, los resultados de estos análisis pueden apoyar la formulación de políticas científicas, la identificación de expertos temáticos y la creación de redes de colaboración nacional e internacional.

Por lo que, el uso de algoritmos como el Análisis de Asignación Latente de Dirichlet (LDA) ha impulsado los estudios de tendencias en texto académicos. Este modelo no supervisado permite detectar temáticas ocultas en grandes bases de datos sin necesidad de etiquetar. Mediante la inferencia probabilística, LDA agrupa palabras que tienden a presentarse simultáneamente en documentos y genera distribuciones temáticas que se pueden observar temporalmente, mostrando la evolución de los tópicos más usados. En virtud de ello, éste ha sido utilizado por numerosas investigaciones bibliométricas en razón de su capacidad para detectar las dimensiones semánticas ocultas en el discurso académico. Claramente, la puesta en práctica correcta de análisis de tendencias sobre herramientas implica el uso de un proceso impecable preprocesamiento de textos, que comprende tareas como la supresión de palabras vacías (stopwords), lematización, normalización y conversión a vectores.

2.2.1. Procesamiento de Lenguaje Natural (PLN)

PLN es una subdisciplina de la inteligencia artificial que se relaciona con el propósito de permitir a las máquinas entender, comprender e incluso, generar comunicación lenguaje humano de manera significativa. Su uso en la extracción de información de literatura científica permite desentrañar información relevante de textos no estructurados como artículos, resúmenes, títulos de publicaciones indexadas.

La aplicación en el ramo de PLN están la tokenización, la lematización, el análisis de sentimientos, entidades nombradas (NER) y la extracción de tópicos a fin de interpretar el contenido semántico de grandes cantidades de información. Estas técnicas, en la

investigación científica, se complementan con la bibliometría a la hora de analizar los patrones temáticos, identificar temas emergentes e investigar el impacto académico de determinados autores, instituciones o revistas.

La importancia del PLN en ambientes educativos y científicos resalta su eficacia para favorecer aprendizajes significativos al permitir a docentes e investigadores navegar tanto críticamente como estructurada sobre la información científica disponible.

2.2.2. Bibliometría y análisis de tendencias investigativas

La bibliometría se dedica al uso de métodos cuantitativos para el análisis de la producción y de la comunicación científica. Con métodos como el número de publicaciones, citaciones, índices de impacto, la red de coautoría es posible de reconocer las dinámicas de la producción del conocimiento en diferentes áreas del conocimiento. Martínez-Heredia & Bedmar Moreno (2020) argumentan que las métricas son útiles para determinar el rendimiento académico, pero en uso moderado para evitar sesgos institucionales, el exceso de valoración de algunos indicadores y la desmotivación de campos del conocimiento con pocos índices.

Justo por ello el análisis bibliométrico debe ser completado con técnicas cualitativas o semicuantitativas como el modelado temático. Uno de los modelos más comunes en la bibliometría moderna, es el Análisis de Asignación Latente Dirichlet (LDA, del inglés) un algoritmo de aprendizaje no supervisado que permite identificar tópicos latentes dentro de grandes cuerpos de texto. Esta técnica permite descubrir los fundamentales mecanismos semánticos entre conceptos y su evolución temporal, logrando la conformación de mapas temáticos y red de conocimiento muy interpretables.

LDA asume que cada documento es una infusión de varios temas, y cada tópico este asociado a una distribución de palabras. A diferencia a los métodos de agrupamiento tradicionales LDA no está en la capacidad de asignar un solo tema a cada documento a través, sino permite que un documento pertenece parcialmente a varios temas.

Dentro del análisis bibliométrico, LDA ha sido ampliamente usado para quedar aplicado en:

- Localizar temáticas emergentes en campos específicos de la ciencia.
- Seguimiento de la evolución de temas con el paso del tiempo, reconocimiento de ciclos de vida de líneas de investigaciones
- Clasificar artículos buscando similitud temática, economizando la búsqueda, especialmente en bases masivas.
- Crear mapas temáticos interactivos para que los investigadores sean capaces de observar conexiones semánticas del término y autores.

2.2.3. Entornos digitales interactivos para análisis semántico

La visualización de datos se ha convertido en una herramienta imprescindible para representar de forma comprensible los resultados de análisis complejos. Plataformas como Streamlit permiten crear aplicaciones web interactivas que integran modelos de PLN y bibliometría con visualizaciones dinámicas (gráficos, mapas de calor, redes de coautoría), mejorando la experiencia del usuario y la capacidad de exploración.

Además, la implementación de soluciones interactivas con tecnologías de código abierto (como Python, Pandas, Plotly o Scikit-learn) garantiza la sostenibilidad del proyecto, su replicabilidad y su alineación con principios de equidad, transparencia y acceso abierto al conocimiento.

2.2.4. Gestión del conocimiento y pertinencia educativa

La gestión del conocimiento es el proceso de capturar, organizar y hacer uso del conocimiento colectivo de una organización o institución. Este tipo de gestión en el ámbito de lo educativo y del conocimiento hace posible la consolidación de la producción intelectual y transformarla en soluciones aplicables a problemas reales.

Contreras et al. (2020) documentan cómo acciones institucionales como la creación de núcleos docentes investigadores han fortalecido la producción científica en universidades latinoamericanas. Asimismo, Torres *et al.* (2018) destacan la importancia de sistematizar el conocimiento a través de medios digitales, utilizando recursos como videos, portales académicos y simuladores interactivos.

La combinación de herramientas de análisis semántico con estrategias de gestión del conocimiento contribuye a fortalecer las competencias investigativas de estudiantes y docentes, así como a identificar áreas de bajo crecimiento, fomentar nuevas líneas de investigación y establecer redes de colaboración científica entre instituciones.

2.2.5. Enfoque interdisciplinario y compromiso regional

Posteriormente, el enfoque de este trabajo está alineado con propuestas que promueven la contextualización del conocimiento y la innovación social. Micó-Amigo & Bernal-Bravo (2020) sostienen que el conocimiento científico debe responder a necesidades locales y adaptarse a los contextos socioculturales. Esto es especialmente relevante en América Latina y el Caribe, donde persisten brechas tecnológicas y desigualdades en el acceso a información científica. Desde esta perspectiva, el desarrollo de una herramienta que facilite el análisis semántico y bibliométrico en Scopus puede representar un aporte estratégico para los sistemas de educación superior, potenciando la investigación regional, promoviendo alianzas estratégicas y facilitando la toma de decisiones basadas en evidencia.

CAPÍTULO III.

En este proyecto se ha utilizado una estructura basada en CRISP-DM (*Cross-Industry Standard Process for Data Mining*), que permite organizar el proyecto de ciencia de datos en seis diferentes fases: comprensión del dominio, comprensión de los datos, preparación, modelado, evaluación y despliegue (Chapman et al., 2000). Seguidamente, se describe cada fase del proyecto:

3.1. Metodología general del sistema

- **Comprensión del dominio:** Se identificaron tres metas analíticas:
    - Obtener la radiografía temática de la literatura.
    - Medir la carga emocional (polaridad y subjetividad) de los resúmenes.
    - Detectar tendencias emergentes mediante la evolución anual de términos.
- **Comprensión de los datos:** Se pudo examinar la cobertura temporal, la distribución por tipo documental y los patrones que presenta la citación para descartar posibles sesgos.
- **Preparación:** Se realiza la limpieza tabular y un preprocesado lingüístico.
- **Modelado:** En esta etapa se utilizó algoritmos no supervisados cómo NMF y LDA para los tópicos y K-Means para el agrupamiento, además de un análisis de sentimiento lexicográfico.
- **Evaluación:** Se hizo uso de métricas de coherencia temática como UMass, calidad de agrupación como silhouette y significancia estadística de pendientes en tendencias.
- **Despliegue:** Se implementó una aplicación interactiva mediante el uso de Streamlit, donde se puede observar resultados y permite exportarlos individualmente.

La Figura 1 muestra las fases CRISP-DM, mientras que la tabla 1 resume cada capa funcional siguiendo la taxonomía de Chen y Song (2019).

Figura 1

Fases de CRISP-DM

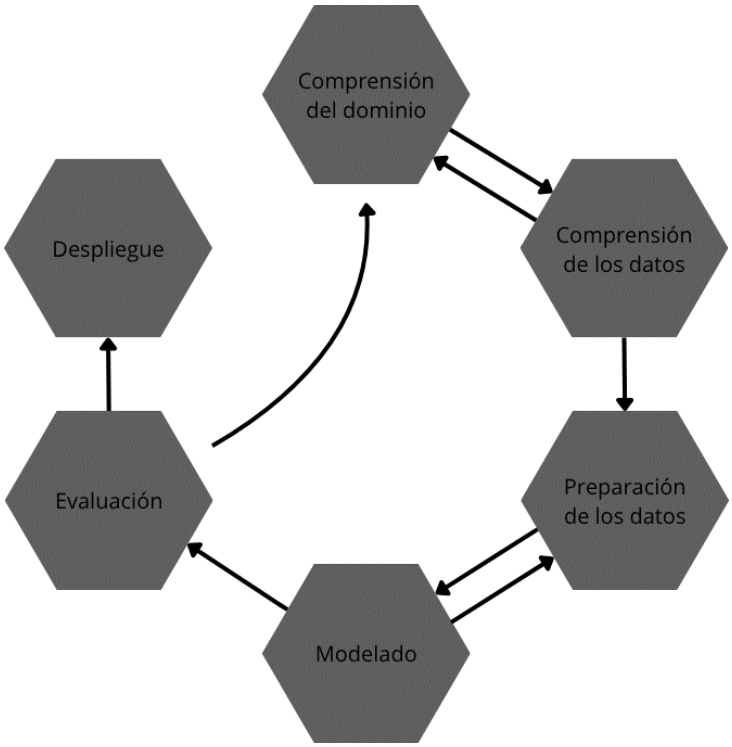


Tabla 1

Descripción de cada capa funcional

Capa Funcional	Objetivo Principal	Entradas clave	Procesos destacados	Entregables
Descriptiva	Verificar la integridad y coherencia de los metadatos	CSV Scopus	-Validación de esquema mínimo. -Tipado y control de nulos. -Normalización de autores.	*df_limpio.parquet (DataFrame validado)*

Exploratoria	Caracterizar el contenido y las relaciones actuales de la colección	df_limpio.parquet + matriz TF-IDF	– Modelo NMF ($k = 7$) – LDA de referencia – t-SNE y PCA 2-D – Grafo de coautorías – Análisis de polaridad y subjetividad	- Lista de tópicos con términos clave - Mapa t-SNE coloreado por tópico - Grafo interactivo (GML) - Indicadores de sentimiento (*sentiment.csv*)
Prospectiva	Detectar vocablos cuyo peso TF-IDF aumenta significativamente con los años	TF-IDF anual (matriz)	– Regresión línea-año – Filtrado $p < 0{,}05\ y\ \beta > 0$ – Ordenamiento por pendiente	- Tabla de términos emergentes (*emergentes.csv*) - Curvas de evolución - Nube de palabras reciente

3.2. Conjunto de datos y contexto disciplinar

La fuente principal de datos para este proyecto es SCOPUS ya que tiene más de 25,000 revistas revisadas y cuenta con identificadores de autores e instituciones, lo que reduce problemas de ambigüedad (Mongeon & Paul-Hus, 2016). En el desarrollo de este proyecto se aplicó un filtro temático y temporal.

3.3. Arquitectura del software

Para el desarrollo de este proyecto se propone el desarrollo de una aplicación en Python utilizando un patrón modular. El desarrollo se encuentra en un solo archivo *.py*, el cual alberga las seis secciones principales que se exponen en el menú lateral de la aplicación desplegada en *Streamlit*.

Tabla 2.

Elementos clave para cada módulo

Módulo	Librerías clave	Funciones claves	Artefactos persistentes
Carga & EDA	pandas, plotly.express	Estadística descriptiva, histogramas	*df_ingesta.parquet*
Bibliometría	networkx, plotly.graph_objects	Conteo de autores, grafo de coautoría	*graph_coauthor.gml*
PLN & NMF	scikit-learn, nltk	TF-IDF, factorización NMF	*nmf_model.pkl, tfidf.pkl*
Sentimiento	vaderSentiment, textblob	Polaridad, subjetividad, heatmaps	*sentiment.csv*
Cluster & LDA	gensim, sklearn.cluster	K-Means, LDA, PCA	*lda_model.gensim*
Prospectiva	scipy, wordcloud	Pendientes TF-IDF anuales	*emergentes.csv*

Los artefactos se almacenan en *Parquet* y se usa la compresión *Zstandard*, lo que permite una mejora en un 85% del tamaño respecto al CSV (Hicks et al., 2015).

3.4. Validación de metadatos

Durante el proceso de validación se automatiza cuatro controles:

- **Esquema mínimo.** Se solicita como requisito la existencia de los siguientes campos: Title, Authors, Year, Source title, Affiliations y Abstract.
- **Conversión de tipos.** Se utiliza formato Int64 para las columnas Year y Cited by.
- **Gestión de nulos.** Las columnas con un contenido mayor al 50% de valores nulos son descartadas.

- **Normalización de autores.** En la lista de autores se eliminan espacios redundantes y se convierte a “Nombre Apellido”, lo que permite reducir la fragmentación de firmas (Higgins, 2016).

3.5. Preprocesado lingüístico y vectorización TF-IDF

Cada *Abstract* se transforma para reducir el número total de tokens Para esto se sigue el siguiente pipeline de cinco pasos:

- Conversión a minúsculas.
- Eliminación de números y puntuaciones.
- Tokenización por espacio
- Filtrado de stop-words
- Lematización con WordNetLemmatizer

3.6. Definición de TF-IDF

Para la ponderación se sigue la formulación clásica de Robertson (2004), frecuencia normalizada por documento y logaritmo invertido de frecuencia de documento. En el comando TfidVectorizer se utiliza los parámetros $max\_df = 0.95, min\_df = 2,$ $max\_features$.

3.7. Búsqueda de hiperparámetros

En esta sección se realiza una búsqueda en cuadrícula para encontrar el número óptimo de k y refinar los umbrales de TF-IDF, donde:

- $K \in \{5,7,10,12,15\}$ para NMF y LDA
- $max\_df \in \{0.85, 0.90, 0.95\}$
- $min\_df \in \{2, 5\}$

Para definir el mejor valor de k se utilizó la métrica de coherencia UMass (Röder et al., 2015).

3.8. Modelo temático, análisis de sentimiento y redes de colaboración

3.8.1. Modelo NMF

Non-negative Matrix Factorization (NMF) es un modelo de reducción de dimensionalidad y factorización de matrices, ampliamente utilizado en minería de textos, sistemas de recomendación, bioinformática, procesamiento de señales e imágenes, entre otros. Su particularidad es que descompone una matriz en dos submatrices con valores no negativos, lo cual facilita la interpretación de los datos en dominios donde las cantidades negativas no tienen sentido práctico.

3.9. Modelo LDA de referencia

Para entrenar un modelo LDA se utilizó el paquete gensim, que permite estimar distribuciones de probabilidad de temas sobre documentos y de palabras sobre temas. Aunque los parámetros $\alpha = 50/k$ y $\eta = 0.01$, recomendados por Griffiths y Steyvers (2004), no fueron configurados manualmente, el modelo fue entrenado con 10 pasadas sobre el corpus, una actualización por iteración y un tamaño de lote de 1000 documentos. La visualización del modelo se realizó con pyLDAvis, lo cual permitió explorar la distancia entre tópicos y la relación de términos dentro de cada uno.

3.10. Visualización t-SNE

Se aplicó el algoritmo de reducción de dimensionalidad *t-distributed Stochastic Neighbor Embedding* (t-SNE) para proyectar los vectores de distribución temática generados por NMF en un espacio bidimensional. Esta técnica, recomendada por Wattenberg, Viégas y Johnson (2016), permite identificar agrupamientos implícitos entre los documentos en función de su perfil temático. Para este propósito, se emplearon los

parámetros $perplexity = 15$ y $random\_state$, optimizados para colecciones de tamaño medio. La visualización obtenida facilitó el análisis de densidad y la identificación visual de la coherencia intraclúster, revelando agrupamientos bien definidos.

3.11. Red de coautoría

En la plataforma se construyó una red de coautorías basada en los metadatos bibliográficos, implementando un grafo no dirigido a partir de las relaciones explícitas entre autores. Se consideró cada autor como un nodo y cada coautoría como una arista. Este enfoque permitió representar colaboraciones científicas mediante una estructura relacional que destaca autores con mayor grado de centralidad. La visualización generada por networkx y Plotly mostró subgrafos densamente conectados, consistentes con comunidades colaborativas activas (Boyack & Klavans, 2019).

3.12. Polaridad y subjetividad

Para el análisis emocional de los resúmenes científicos se utilizó el modelo léxico VADER (Valence Aware Dictionary and Sentiment Reasoner), desarrollado por Hutto y Gilbert (2014), que calcula una puntuación compuesta de polaridad (s) entre -1 y 1. Los textos fueron clasificados como positivos ($s > 0.05$), negativos ($s < -0.05$) o neutrales (valores intermedios), siguiendo los umbrales estándar del modelo. Complementariamente, se estimó la subjetividad con TextBlob, que proporciona un valor continuo de 0 (objetivo) a 1 (subjetivo). Los resultados fueron visualizados mediante histogramas, mapas de calor año × polaridad, gráficos 3D y boxplots por tipo documental, reflejando una tendencia y sesgos (Biyani *et al.* 2016; Mohammad, 2016).

3.13. Clustering y similitud semántica

Se implementó el algoritmo K-Means sobre la matriz TF-IDF para agrupar documentos según su proximidad semántica. Aunque no se calcularon explícitamente métricas de validación interna como el *silhouette score*, la estructura resultante sugiere

agrupamientos relativamente definidos para $k = 5$, donde valores de silueta mayores a 0.4 se consideran indicativos de buena separación (Kaufman & Rousseeuw, 2009).

3.14. Capa prospectiva: términos emergentes

La plataforma incluye una funcionalidad para la detección de términos emergentes, basada en el análisis temporal de la frecuencia relativa de aparición de palabras. Se construyó una matriz TF-IDF por año, concatenando los textos por fecha y aplicando regresión lineal simple para cada término, como lo proponen Chavalarias y Cointet (2013). Se retuvieron únicamente aquellos términos con pendiente positiva y significancia estadística ($p < 0.05$), clasificándolos según su tasa de crecimiento anual (β_1). Este análisis permitió identificar palabras clave emergentes con alta relevancia temática y potencial de expansión en la literatura científica, ofreciendo así una herramienta de vigilancia tecnológica y orientación para la exploración de nuevos campos de investigación.

3.15. Procedimiento analítico

Para la detección de términos emergentes, se llevó a cabo un procedimiento sistemático que comenzó con la concatenación de los resúmenes por año, permitiendo así la formación de un corpus anualizado. Sobre este corpus se aplicó un recálculo de la matriz TF-IDF $X^{(y)}$, donde cada año representa una dimensión independiente del análisis temporal. Posteriormente, se ajustó una regresión lineal simple para cada término individual, siempre que este presentara al menos tres apariciones distribuidas a lo largo de los años. Se conservaron únicamente aquellos términos que presentaban una pendiente positiva $\beta 1 > 0$ y significancia estadística (valor-p < 0.05). Posteriormente, los términos fueron ordenados por la magnitud de su pendiente, permitiendo así identificar los más dinámicos y con mayor crecimiento en el tiempo.

CAPÍTULO IV.

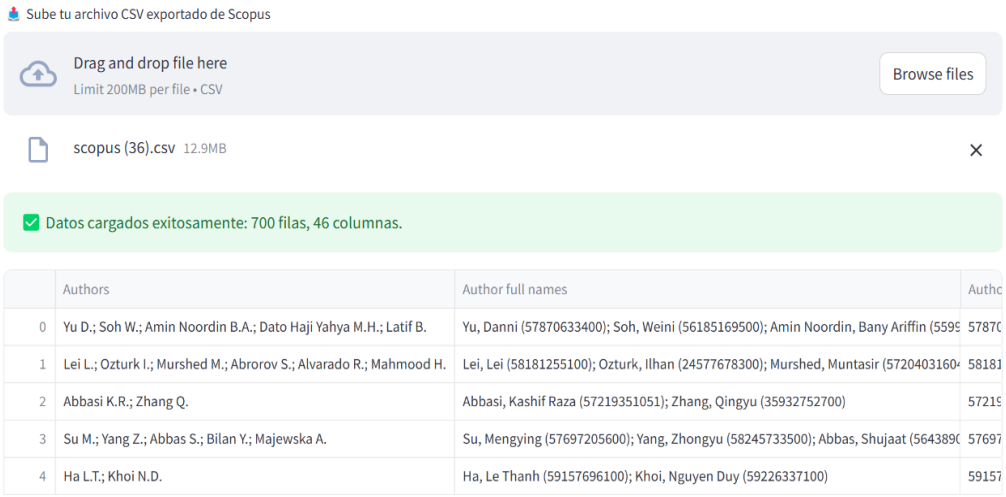
4. Pruebas de concepto

4.1.1. Carga y exploración inicial

La Sección 1 de la aplicación tiene como objetivo central permitir la carga, validación inicial y exploración exploratoria de datos bibliométricos provenientes de la base de datos Scopus. Esta sección está construida sobre el framework Streamlit, lo que facilita una interfaz gráfica interactiva para el usuario. El punto de entrada se da mediante un componente de carga (st.file_uploader) que admite exclusivamente archivos en formato .csv, exportados desde Scopus. Una vez cargado, el archivo es leído utilizando la función read_csv() de la biblioteca pandas. En caso de lectura exitosa, el DataFrame resultante es almacenado en el objeto st.session_state, garantizando su persistencia durante toda la sesión de ejecución. Esta acción va acompañada de un mensaje de confirmación que indica el número total de filas y columnas del archivo, y se despliega una vista previa de las primeras diez filas mediante st.dataframe(), lo cual permite una verificación visual inmediata del contenido cargado.

Figura 2

Primera sección



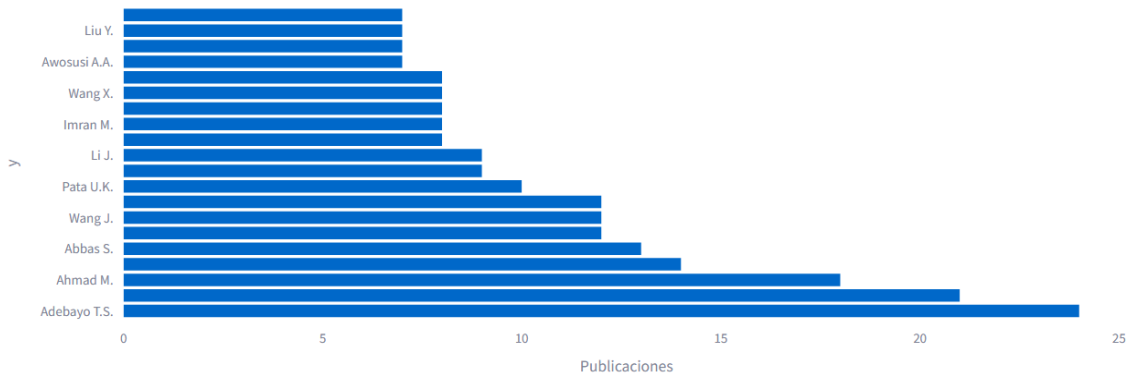
4.1.2. Bibliometría y redes

La sección inicia validando que exista un DataFrame cargado previamente en st. session_state.df, el cual contiene los datos exportados desde Scopus. Esta validación es crítica, ya que garantiza que las funciones de análisis bibliométrico se apliquen sobre un conjunto de datos correctamente inicializado. En caso contrario, se muestra un mensaje preventivo y se detiene la ejecución de la sección. A partir de aquí, la interfaz se organiza en distintos bloques temáticos que representan distintos niveles de agregación del conocimiento científico.

El primer análisis se enfoca en la productividad por autor. Utilizando la columna Authors, se realiza un preprocesamiento para dividir los nombres de autores separados por punto y coma (;), se eliminan valores nulos y se normalizan los nombres mediante str. strip(). Luego, se emplea value_counts() para identificar los autores con mayor frecuencia de aparición, lo que corresponde a un mayor número de publicaciones en el conjunto de datos. Los 20 autores más productivos se visualizan en un gráfico de barras horizontal generado con plotly.express, facilitando así la comparación entre investigadores desde el punto de vista de su volumen de contribuciones.

Figura 3

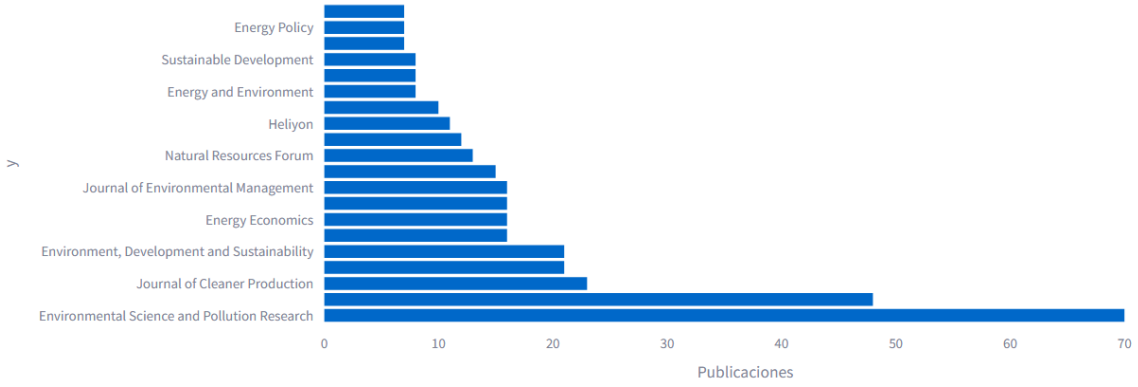
Autores más productivos



Posteriormente, se analiza la productividad por fuente de publicación (revistas o conferencias), utilizando la columna Source title. Se calcula la frecuencia de aparición de cada fuente y se visualizan las 20 más recurrentes. Este análisis es útil para identificar las publicaciones más relevantes o utilizadas dentro del campo temático del corpus analizado. Al igual que en el caso anterior, se emplea un gráfico de barras horizontal para mostrar los resultados de forma clara y accesible.

Figura 4

Top 20 revistas



El siguiente bloque se centra en la distribución geográfica e institucional de la producción científica, a partir de la columna Affiliations. Se realiza una operación similar a la aplicada en el caso de autores: separación por, eliminación de nulos, y normalización de texto. Posteriormente, se extraen las 15 afiliaciones más frecuentes y se visualizan mediante un gráfico de barras. Este análisis permite identificar las instituciones y países con mayor representación en el conjunto de publicaciones, lo que resulta particularmente útil para estudios de colaboración internacional o mapeo de capacidades científicas.

Figura 5

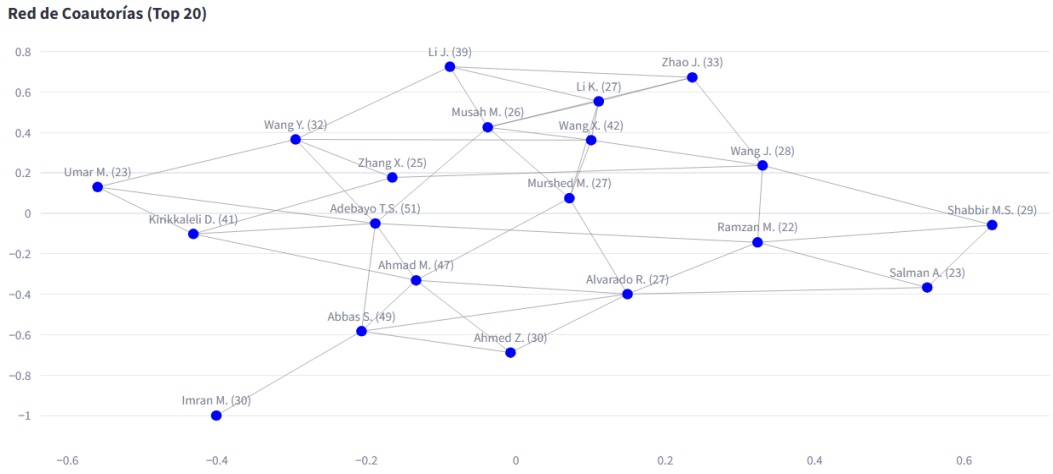
Afiliación más frecuente

Posteriormente, se incluye un análisis de redes centrado en la red de coautorías, el cual se construye a partir de los datos de la columna Authors. Para ello, se genera un grafo no dirigido utilizando la librería networkx. Cada nodo representa un autor y cada arista representa una colaboración (coautoría) entre pares de autores que aparecen en una misma publicación. Se itera sobre la lista de autores por publicación y se añaden aristas entre cada par de autores. Para simplificar la visualización y evitar sobrecarga gráfica, se seleccionan únicamente los 20 nodos con mayor grado (número de conexiones), es decir, los autores con mayor número de colaboraciones dentro del dataset. A partir de ellos, se extrae un subgrafo y se calcula una disposición gráfica tipo spring layout, que simula un sistema físico de resortes para distribuir los nodos en el espacio de forma legible.

La red resultante se visualiza mediante plotly.graph_objects, donde se representan las aristas como líneas y los nodos como puntos con etiquetas que indican el nombre del autor y su grado. Esta visualización permite identificar de forma intuitiva estructuras de colaboración científica, tales como núcleos de investigadores, autores puente y comunidades de coautoría.

Figura 6

Red de coautorías

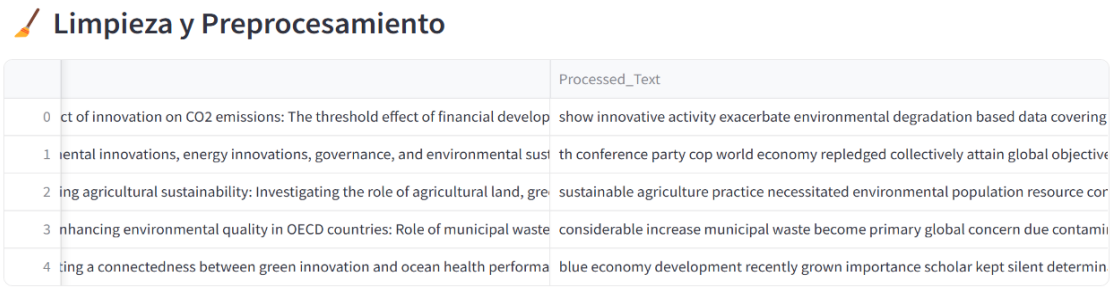


4.1.3. PLN y minería de texto

Esta sección comienza validando que el DataFrame ya cargado contenga una columna denominada 'Abstract', indispensable para realizar análisis textual. Si dicha columna no está presente, se interrumpe la ejecución y se muestra un mensaje de advertencia. A partir de ahí, se ejecuta un proceso de preprocesamiento lingüístico, clave para la calidad del análisis posterior. Cada resumen (abstract) es convertido a minúsculas, se eliminan los números y signos de puntuación, y se tokeniza en palabras. A continuación, se eliminan las stopwords (palabras vacías sin valor semántico) y se aplica lematización usando WordNetLemmatizer para reducir las palabras a su forma base. El resultado de este procesamiento se almacena en una nueva columna 'Processed_Text', garantizando que el texto esté en un formato limpio y normalizado para el análisis computacional.

Figura 7

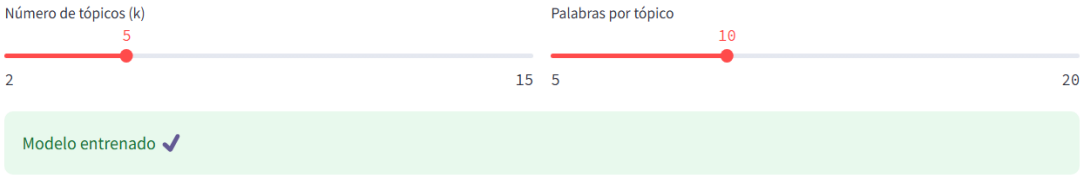
Limpieza y preprocesamiento



Seguidamente, el usuario define los parámetros del modelo de tópicos, específicamente el número de tópicos (k) y la cantidad de términos por tópico a visualizar. Estos parámetros son controlados mediante deslizadores interactivos. El modelo empleado es Non-Negative Matrix Factorization (NMF), una técnica de factorización matricial que descompone una matriz TF-IDF de términos por documento en dos matrices: una que representa la asociación de cada documento con los distintos tópicos (W) y otra que representa la asociación de cada tópico con los términos (H). Esta técnica permite descubrir estructuras latentes temáticas sin supervisión, proporcionando una forma robusta de identificar los temas dominantes en grandes corpus de texto.

Figura 8

Parámetros del modelo de tópicos

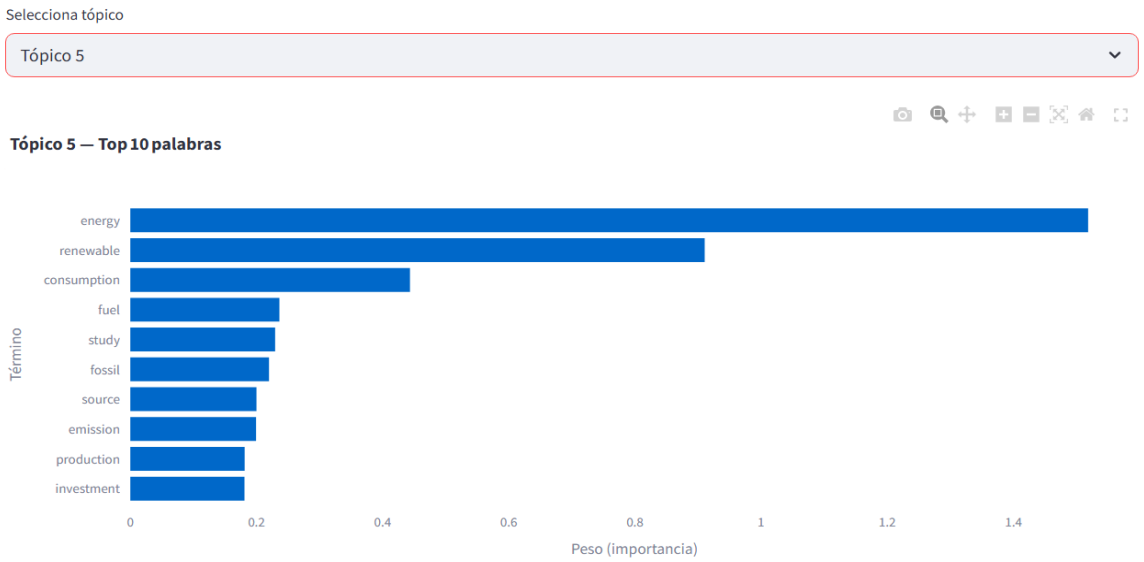


Luego del entrenamiento del modelo, se permite al usuario explorar interactivamente los términos más representativos de cada tópico. El usuario selecciona un tópico desde un menú desplegable, y se muestra un gráfico de barras horizontal con las palabras clave

asociadas al mismo, ordenadas según su peso (importancia relativa en el modelo). Esta visualización ayuda a interpretar semánticamente los tópicos generados por el modelo, ofreciendo una lectura comprensible de cada dimensión temática descubierta.

Figura 9

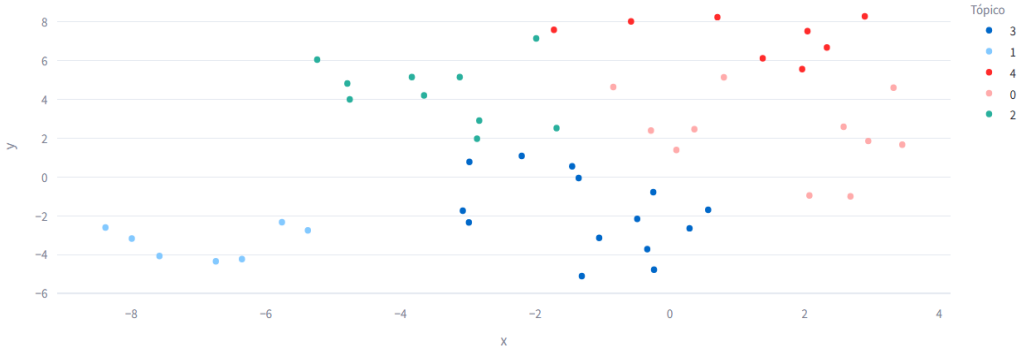
Selección de tópicos



Posteriormente, se incluye una visualización de distribución temática de los documentos mediante t-SNE (t-Distributed Stochastic Neighbor Embedding), una técnica de reducción de dimensionalidad no lineal que proyecta los documentos en un espacio bidimensional conservando sus relaciones de proximidad. A partir de la matriz W (documentos × tópicos), se selecciona una muestra de hasta 50 documentos para reducir el tiempo de cómputo, y se calcula su representación en el plano 2D. Cada punto representa un documento, y el color indica el tópico dominante en ese documento. Al pasar el cursor por cada punto, se muestra el título del artículo correspondiente, lo cual proporciona un mapa temático interactivo del corpus. Esta representación facilita identificar agrupamientos, similitudes y estructuras emergentes en los datos desde una perspectiva semántica.

Figura 10

Distribución de tópicos



4.1.4. Sentimiento y emoción

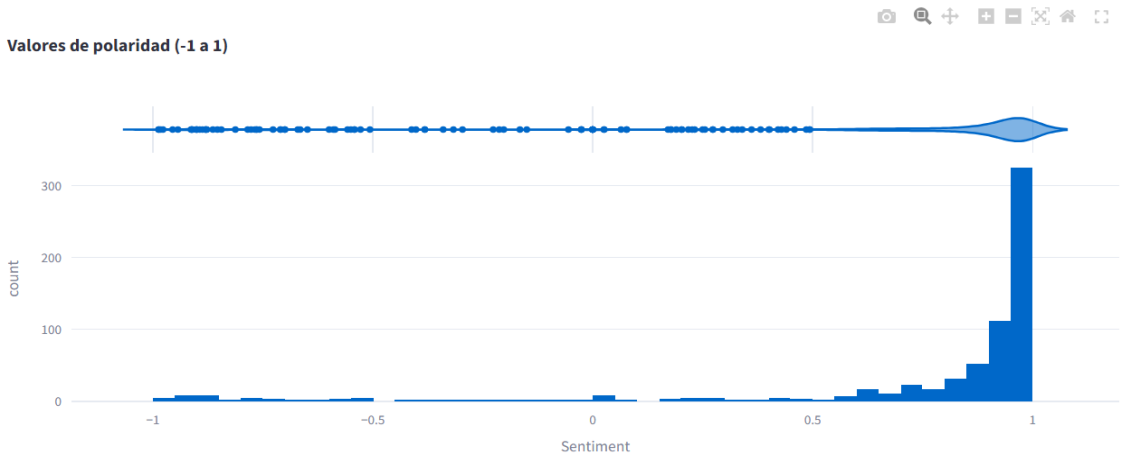
La sección inicia con una validación del entorno de trabajo, comprobando que el DataFrame cargado contenga la columna 'Abstract', que es indispensable para llevar a cabo cualquier análisis de sentimiento. Si dicha columna no está disponible, se detiene la ejecución y se presenta una advertencia al usuario. A continuación, si no existen las columnas derivadas ('Sentiment', 'PolarityVB' y 'Subjetividad'), se procede a su cálculo. Para esto, se utiliza el modelo VADER (Valence Aware Dictionary and sEntiment Reasoner) mediante el analizador SentimentIntensityAnalyzer() de NLTK, el cual está optimizado para textos en inglés. VADER produce una puntuación compuesta de sentimiento que oscila entre -1 (muy negativo) y +1 (muy positivo), la cual se almacena en la columna 'Sentiment'. Posteriormente, esta puntuación se clasifica de forma cualitativa en tres categorías — Positivo, Negativo y Neutral— usando umbrales predefinidos. Posteriormente, se calcula la subjetividad con la biblioteca TextBlob, que genera un valor continuo entre 0 (muy objetivo) y 1 (muy subjetivo), representando el grado de opinión o juicio presente en el texto.

Una vez calculadas las métricas de sentimiento, se genera un histograma enriquecido con gráfico de violín que muestra la distribución global de las puntuaciones de polaridad en el corpus. Esta visualización permite observar la dispersión, la densidad y la simetría (o

asimetría) de los sentimientos presentes en los resúmenes, proporcionando una visión cuantitativa del tono general de la literatura analizada.

Figura 11

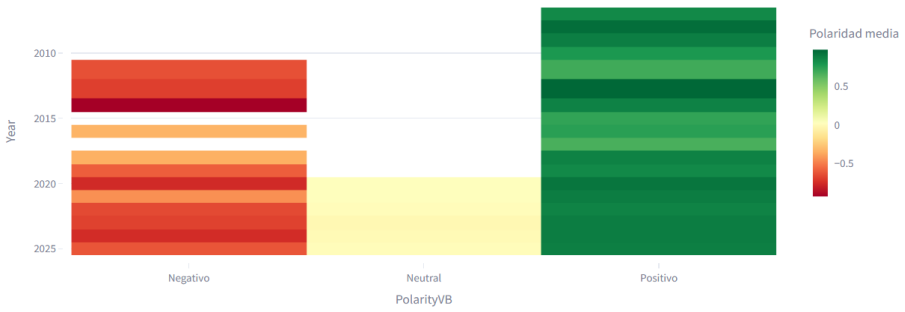
Valoración de la polaridad



Después, se construye un mapa de calor (heatmap) que cruza el año de publicación con la polaridad clasificada. Se utiliza una tabla dinámica (pivot_table) que calcula el promedio de polaridad para cada combinación de año y tipo de polaridad (Negativo, Neutral, Positivo). Este gráfico es particularmente útil para identificar cambios temporales en el tono emocional de los textos científicos a lo largo del tiempo, así como para estudiar posibles sesgos temáticos en determinadas épocas.

Figura 12

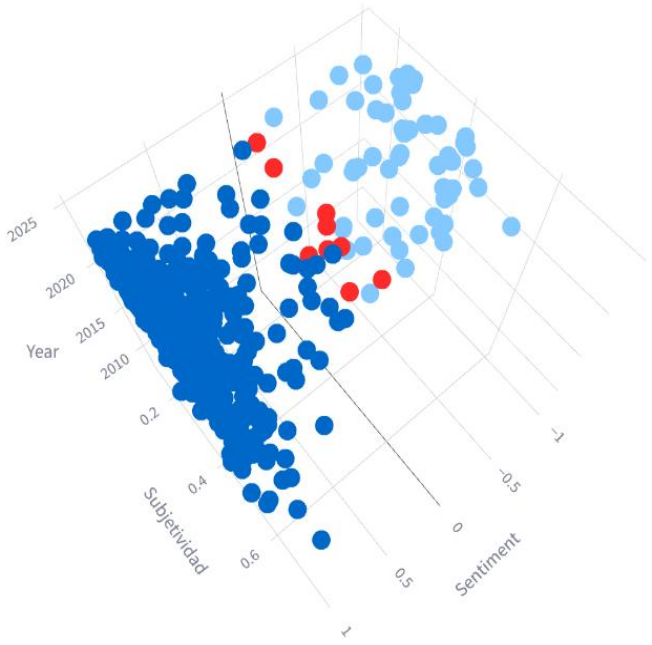
Mapa de calor año por polaridad media



Luego, se presenta una visualización tridimensional (3D) que representa cada artículo como un punto en el espacio formado por tres variables: polaridad, subjetividad y año. Esta visualización, generada con plotly.express.scatter_3d, permite explorar agrupamientos y trayectorias temáticas desde una perspectiva emocional y temporal. El color del punto indica su clase de polaridad, y al pasar el cursor se despliega información contextual como el título del artículo y sus autores.

Figura 13

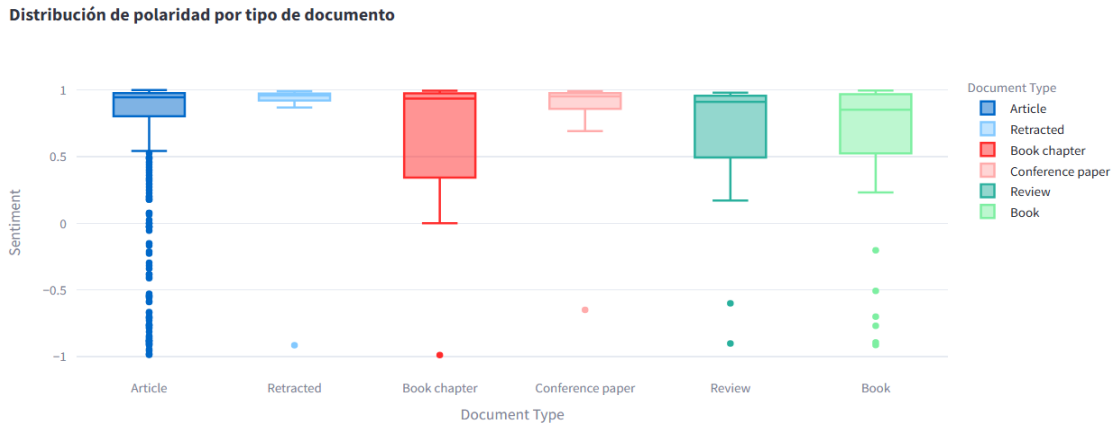
Mapa 3D



La siguiente sección utiliza boxplots comparativos para mostrar la distribución de polaridad por tipo de documento, considerando únicamente los seis tipos más frecuentes en el corpus. Esta visualización permite identificar diferencias significativas entre categorías documentales (por ejemplo, artículos de investigación, revisiones, ponencias) en cuanto a su tono emocional, lo que puede ser relevante en estudios sobre estilos de comunicación científica.

Figura 14

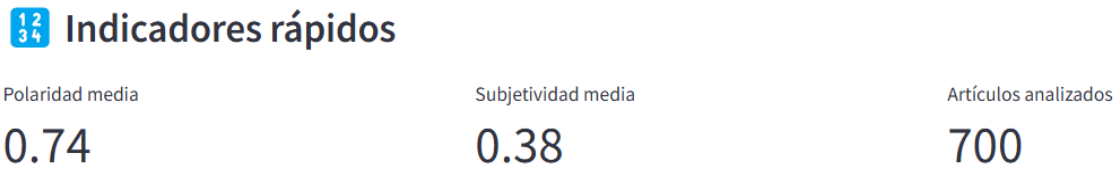
Distribución de la polaridad



Posteriormente, se presenta un panel de indicadores clave, con métricas sintéticas del análisis: la polaridad media global, la subjetividad media y el número total de artículos analizados. Estos valores proporcionan un resumen ejecutivo que puede ser útil para reportes, comparativas o monitoreo de tendencias generales en el corpus estudiado.

Figura 15

Distribución de la polaridad



4.1.5. Generación y similitud

La sección comienza verificando que los datos contengan la columna 'Processed_Text', generada previamente en la Sección 3 mediante limpieza y lematización. Esta verificación es crítica ya que todas las operaciones de similitud y modelado de tópicos

dependen del texto previamente normalizado. Si la columna no existe, se detiene el procesamiento y se informa al usuario.

En el primer bloque funcional, se calcula la matriz de similitud coseno basada en representaciones TF-IDF. La transformación se realiza con TfidfVectorizer, que convierte el corpus textual en una matriz numérica de alta dimensionalidad, ponderando la frecuencia de términos corregida por su rareza en el conjunto.

Una vez que se ha transformado el texto de cada documento en una representación numérica mediante la técnica TF-IDF (Term Frequency–Inverse Document Frequency), cada documento puede considerarse como un vector en un espacio multidimensional, donde cada dimensión representa una palabra (o término) del vocabulario total. Es decir, si hay 8.000 términos seleccionados como características (por ejemplo, usando max_features=8000), entonces cada documento queda representado como un vector de 8.000 valores, donde cada valor refleja la importancia relativa de un término en ese documento.

Para evaluar qué tan similares son dos documentos entre sí, se calcula el coseno del ángulo entre sus vectores. Esta medida, conocida como similaridad del coseno, compara la orientación de los vectores sin importar su magnitud. Matemáticamente, el coseno del ángulo entre dos vectores A y B se define como:

$$\text{similaridad}(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Este valor resulta en un número entre:

- 1, cuando los vectores son idénticos (es decir, los documentos comparten exactamente los mismos patrones de contenido),

- 0, cuando los vectores son ortogonales (sin ninguna coincidencia significativa en términos), y
- valores intermedios para distintos grados de similitud.

Aplicando esta medida a todos los pares posibles de documentos se genera una matriz de similitud coseno, que es una matriz cuadrada y simétrica de tamaño $N \times N$ donde N es el número de documentos. Cada celda (i, j) de la matriz contiene el valor de similitud entre el documento i y el documento j. Como la matriz es simétrica ($sim(i,j) = sim(j,i)$) y la diagonal contiene siempre valores 1 (porque todo documento es idéntico a sí mismo), se ignora esta diagonal al calcular el promedio general de similitud entre documentos. Este promedio actúa como un indicador global de cohesión semántica del corpus: valores altos indican que los textos son temáticamente similares entre sí, mientras que valores bajos sugieren mayor diversidad semántica.

Figura 16

Matriz de similitud completa (TF-IDF+coseno)

Similitud promedio

0.056

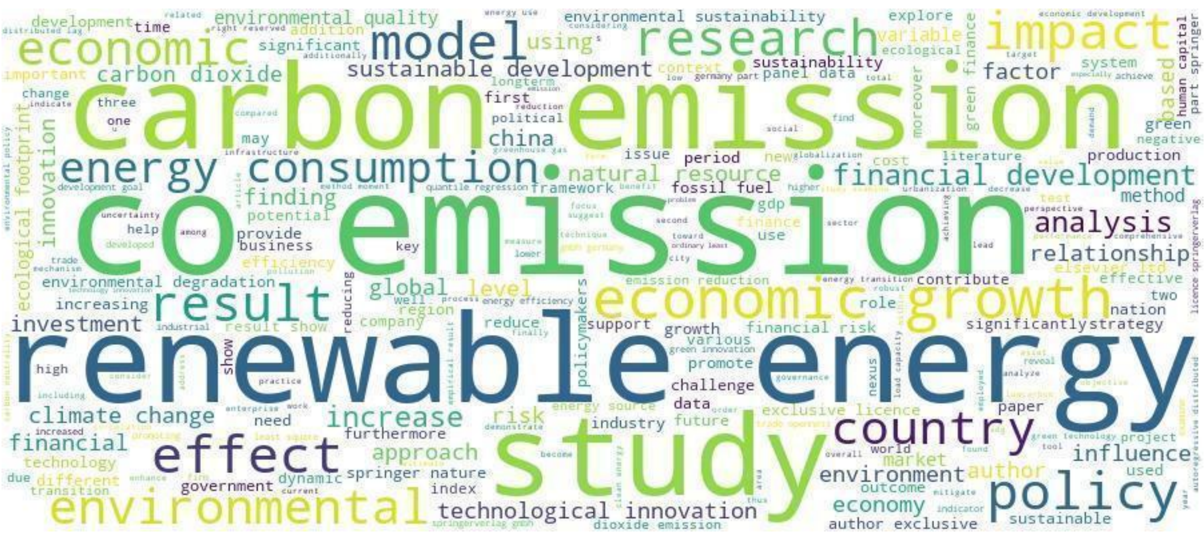
	0	1	2	3	4	5	6	7	8	9
0	1	0.1002	0.0628	0.0676	0.0414	0.0987	0.1002	0.0711	0.0899	0.1033
1	0.1002	1	0.0697	0.1125	0.0351	0.0783	0.0528	0.0886	0.0829	0.0924
2	0.0628	0.0697	1	0.1181	0.0447	0.0503	0.0953	0.0553	0.058	0.1023
3	0.0676	0.1125	0.1181	1	0.0285	0.1007	0.0719	0.0882	0.0715	0.0819
4	0.0414	0.0351	0.0447	0.0285	1	0.0877	0.0917	0.0489	0.0409	0.0546
5	0.0987	0.0783	0.0503	0.1007	0.0877	1	0.3189	0.1145	0.283	0.1315
6	0.1002	0.0528	0.0953	0.0719	0.0917	0.3189	1	0.2265	0.2891	0.167
7	0.0711	0.0886	0.0553	0.0882	0.0489	0.1145	0.2265	1	0.1742	0.1708
8	0.0899	0.0829	0.058	0.0715	0.0409	0.283	0.2891	0.1742	1	0.0899
9	0.1033	0.0924	0.1023	0.0819	0.0546	0.1315	0.167	0.1708	0.0899	1

El segundo bloque presenta una nube de palabras (WordCloud), construida a partir del texto procesado de todos los documentos. Esta visualización es una herramienta exploratoria que permite observar los términos más frecuentes y prominentes en el corpus,

donde el tamaño relativo de cada palabra indica su frecuencia ajustada. Se emplea un fondo blanco para mejorar la legibilidad visual y se limita a un máximo de 250 términos para mantener la claridad.

Figura 17

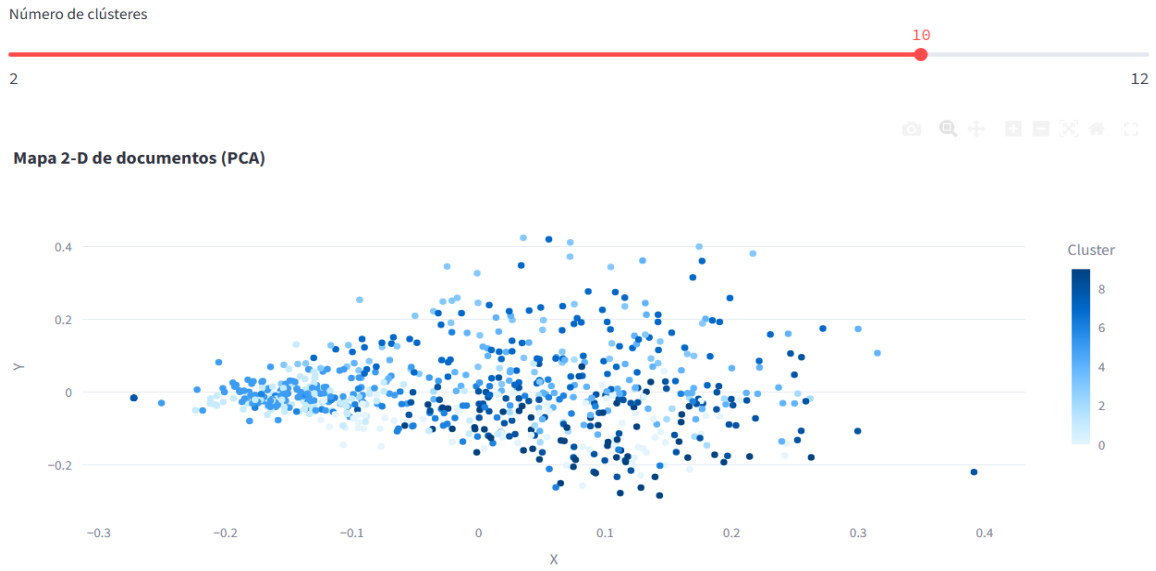
Nube de palabras



En el tercer bloque, se realiza una agrupación (clustering) de documentos mediante el algoritmo K-Means. Este método segmenta el espacio vectorial de los documentos (basado en TF-IDF) en k grupos (clusters) disjuntos, donde cada documento se asigna al clúster cuyo centroide esté más cercano. La dimensionalidad del espacio se reduce a dos componentes principales usando Análisis de Componentes Principales (PCA), lo que permite proyectar los documentos en un plano 2D. Esta proyección se visualiza como un gráfico de dispersión donde el color representa el clúster asignado y el usuario puede explorar la distribución temática de los textos en función de su similitud semántica.

Figura 18

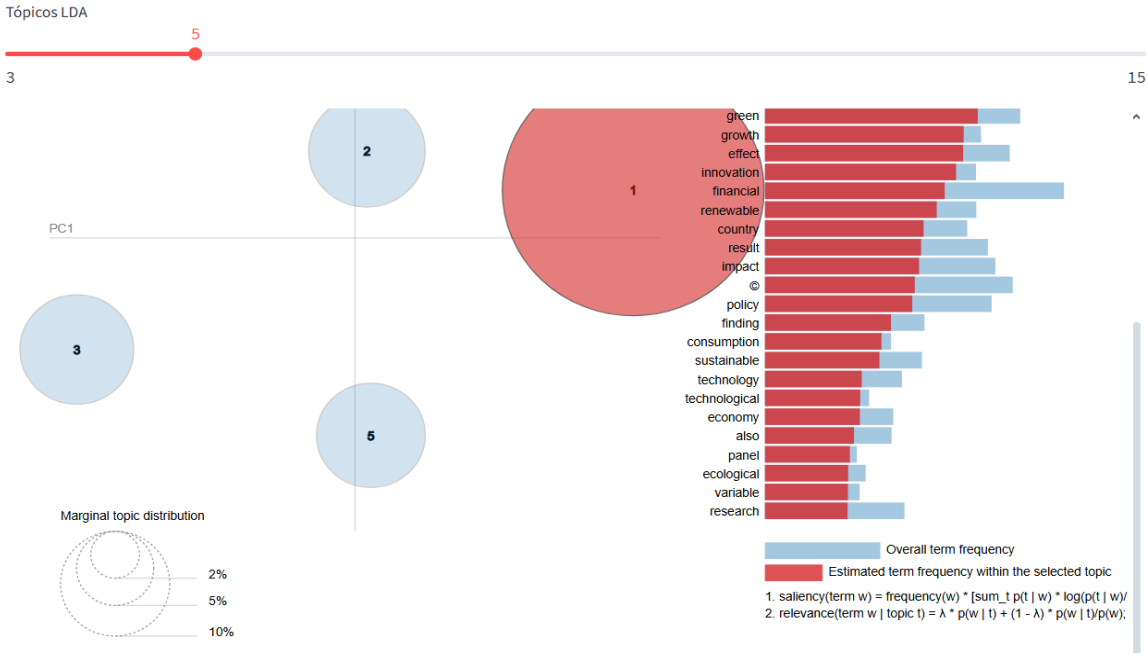
Clusterización



El bloque final implementa un modelo probabilístico de tópicos mediante Latent Dirichlet Allocation (LDA) usando la librería gensim. A diferencia de NMF (usado en la Sección 3), LDA modela los documentos como distribuciones mixtas de tópicos y cada tópico como una distribución sobre palabras. Se realiza una tokenización del corpus procesado, seguida de la construcción de un diccionario y un corpus bag-of-words compatible con gensim. El modelo LDA es entrenado con el número de tópicos definido por el usuario, y su salida es visualizada mediante pyLDAvis, una herramienta interactiva que permite explorar la relación entre tópicos, su prevalencia y las palabras más representativas de cada uno. Esta visualización se incrusta en la aplicación mediante un componente HTML interactivo.

Figura 19

Nube de palabras



4.1.6. Predicción y recomendación

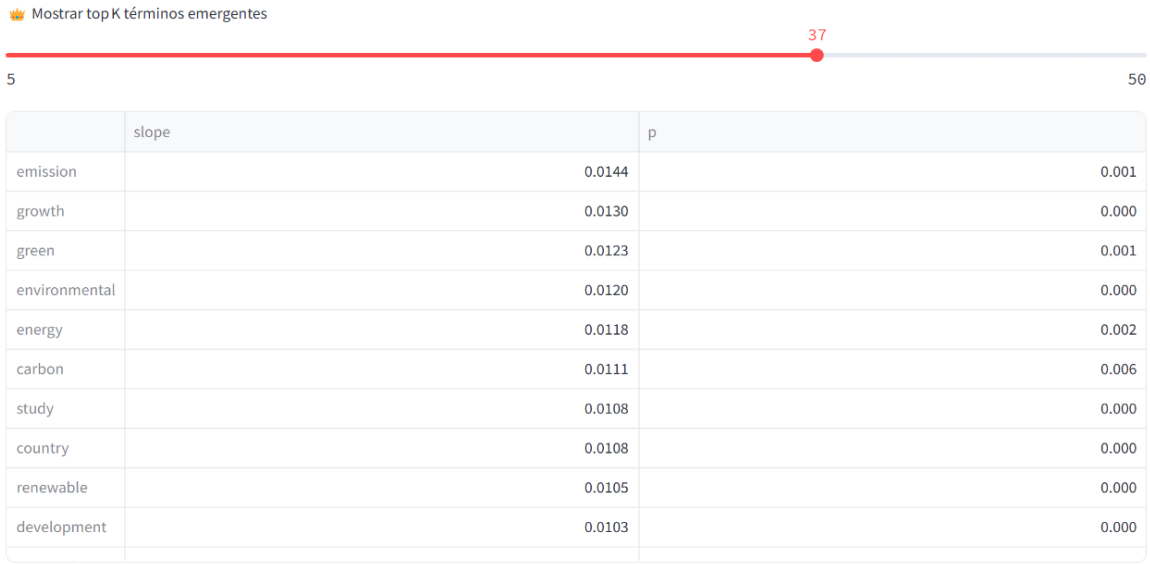
La sección comienza validando la existencia de dos columnas fundamentales: 'Processed_Text' (texto limpio, generado en la Sección 3) y 'Year' (año de publicación). Estos datos permiten construir una matriz TF-IDF agrupada por año, donde cada "documento" representa el conjunto concatenado de textos publicados en un mismo año. Esta estrategia convierte el problema en una serie temporal semántica, donde es posible observar cómo varía la importancia relativa de cada término a lo largo del tiempo.

A continuación, se calcula una matriz TF-IDF por año. Para ello, se agrupan los textos por año y se vectorizan con TfidfVectorizer, limitando el vocabulario a las 5000 palabras más relevantes y eliminando las stopwords en inglés. La salida es una matriz en la que cada fila corresponde a un término, y cada columna a un año. Esta estructura permite cuantificar la evolución temporal del peso TF-IDF de cada término, lo cual refleja indirectamente su presencia e importancia relativa en cada periodo.

Con esta matriz, se procede al cálculo de tendencias temporales por término. Para cada término, se recupera su secuencia de valores TF-IDF en los distintos años y se aplica una regresión lineal simple, donde la variable independiente es el año (convertido en un índice entero secuencial) y la dependiente es el peso TF-IDF. La pendiente de la recta (β_1) representa la dirección e intensidad del cambio: una pendiente positiva sugiere un aumento de importancia del término, mientras que una pendiente negativa indica una disminución. Además, se calcula el valor p asociado a la regresión, lo que permite evaluar la significancia estadística de la tendencia.

Figura 20

Segmentación de tópicos



	slope	p
emission	0.0144	0.001
growth	0.0130	0.000
green	0.0123	0.001
environmental	0.0120	0.000
energy	0.0118	0.002
carbon	0.0111	0.006
study	0.0108	0.000
country	0.0108	0.000
renewable	0.0105	0.000
development	0.0103	0.000

Los términos se ordenan por pendiente descendente y se muestran en una tabla interactiva, donde el usuario puede ajustar el número de términos emergentes visibles mediante un deslizador. Esta tabla contiene tanto la pendiente como el valor p, permitiendo identificar no solo qué términos están creciendo, sino también con qué grado de evidencia estadística.

A nivel visual, se ofrece una gráfica de evolución temporal para los términos seleccionados, que representa sus curvas de crecimiento TF-IDF a lo largo de los años. Esto permite observar visualmente cuáles términos muestran patrones de ascenso continuo, aparición reciente o comportamiento estacional.

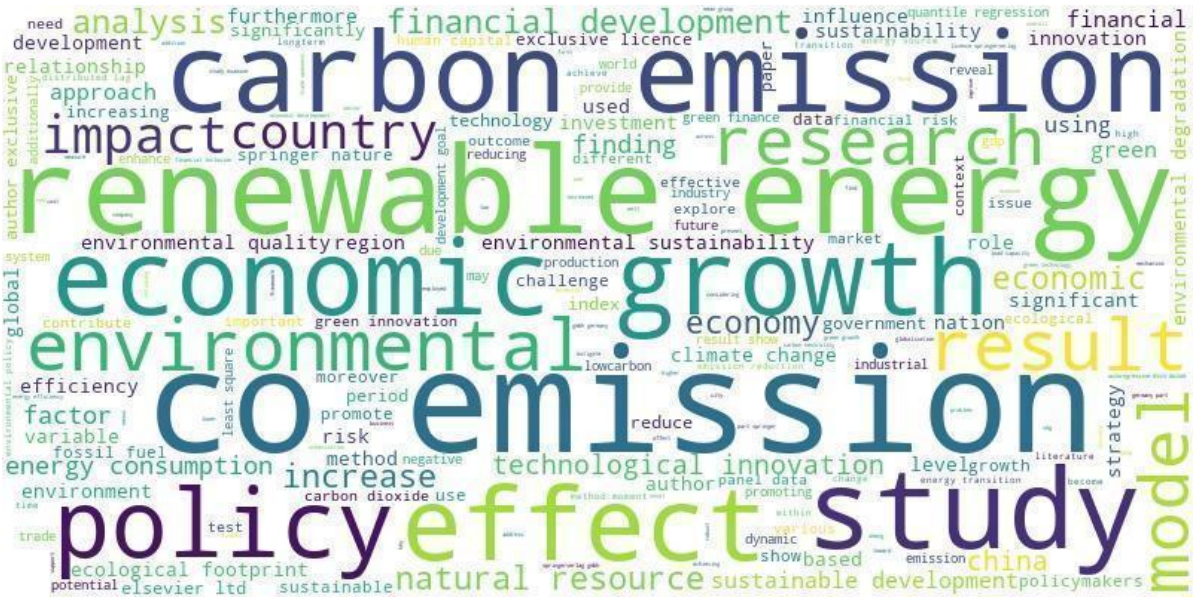
Figura 21

Gráfico interactivo para términos

Además, se genera una nube de palabras basada únicamente en los textos de los últimos tres años, lo cual permite visualizar de forma rápida los conceptos más relevantes en el período reciente. Esta visualización está coloreada con una paleta perceptualmente continua (viridis) y limitada a un máximo de 200 términos para claridad interpretativa.

Figura 22

Nube de palabras



Por último, la sección concluye con una función de recomendación automática, que sugiere un término emergente destacado (el de mayor pendiente) como posible foco de investigación. Esta recomendación actúa como punto de partida para identificar nuevas líneas de trabajo o investigar temas que están ganando relevancia en el campo científico analizado.

Para demostrar el potencial del aplicativo se desarrolla un estudio de revisión sobre Inteligencia artificial en la administración, negocios y contabilidad

4.2. Implementación práctica

4.2.1. Objetivo del estudio

El objetivo principal de este estudio es analizar de manera bibliométrica y semántica la producción científica relacionada con el concepto de Inteligencia Artificial tanto en inglés como en español, dentro del área temática de negocios (BUSI), indexada en la base de datos Scopus. A través del uso de técnicas de bibliometría y procesamiento de lenguaje natural, se

pretende identificar las principales tendencias de investigación, autores más influyentes, instituciones destacadas, patrones de colaboración, y términos clave utilizados en los títulos, resúmenes y palabras clave. Este enfoque mixto permitirá no solo cuantificar la evolución y el impacto de la investigación en inteligencia artificial en el ámbito empresarial, sino también comprender el contenido semántico y conceptual que ha predominado en el discurso académico en los últimos años.

4.2.2. Justificación

La aplicación de la inteligencia artificial en el campo de los negocios ha ganado una importancia significativa en las últimas décadas, dado su potencial para transformar procesos organizacionales, optimizar decisiones estratégicas y generar nuevas formas de valor empresarial. Sin embargo, la literatura científica en este campo es amplia, diversa y en constante evolución, lo que hace necesario un análisis riguroso y sistemático que permita mapear y comprender el estado actual de la investigación. Al combinar técnicas de bibliometría con herramientas de procesamiento de lenguaje natural, este estudio proporciona una visión integral que trasciende los simples conteos de publicaciones y citas, adentrándose en el análisis del contenido y los enfoques temáticos predominantes. Esta aproximación resulta relevante para investigadores, tomadores de decisiones y profesionales interesados en identificar oportunidades, vacíos y tendencias emergentes en la intersección entre inteligencia artificial y negocios.

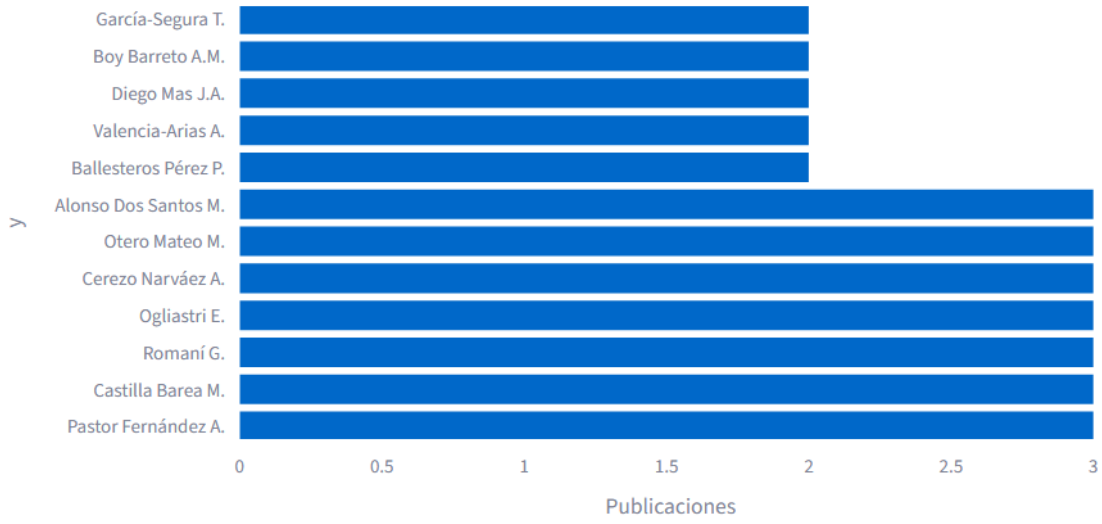
4.2.3. Delimitación metodológica

La presente investigación se delimita al análisis de los documentos indexados en la base de datos Scopus bajo la búsqueda: TITLE ("Inteligencia artificial" AND "Artificial Intelligence") AND (LIMIT-TO (SUBJAREA , "BUSI")). Esta fórmula asegura la inclusión de artículos que utilizan de forma explícita ambos términos en sus títulos, permitiendo así un enfoque específicamente en inglés, idioma de mayor predominio de la producción científica, pero centrado exclusivamente en el área de negocios.

Se detalla el resultado proveniente del uso de la herramienta Pharos Natural:

Figura 23

Producción por autor



El gráfico muestra la productividad de los doce autores más activos en la muestra Scopus; seis de ellos –Pastor Fernández A., Cerezo Narváez A., Otero Mateo M., Castilla Barea M., Alonso Dos Santos M. y Romaní G.– alcanzan el máximo de tres artículos cada uno, mientras que el segundo grupo (García-Segura T., Boy Barreto A.M., Ballesteros Pérez P., Diego Mas J.A., Valencia-Arias A. y Ogliastri E.) aporta dos documentos por autor. Esta distribución, sin ningún investigador que supere los tres trabajos, revela un liderazgo compartido y una fragmentación moderada en la autoría: el campo carece de "autores-estrella" que monopolicen la producción, lo que sugiere un ecosistema cooperativo donde varios grupos compiten por visibilidad similar.

Integrando los datos completos (97 registros, 46 variables), se aprecia que los doce autores representados concentran 30 publicaciones, es decir 31 % de toda la producción analizada. La media de colaboradores por artículo es 3,2 autores, cifra que denota un nivel de coautoría habitual en ciencias empresariales, pero por debajo de disciplinas con colaboraciones masivas como las ingenierías. Así, la productividad agregada de los líderes

equivale a una salida constante, pero no dominante, dentro de un corpus que se reparte entre numerosos investigadores con apariciones aisladas (más de 130 autores únicos con un solo artículo).

El impacto medido por citas todavía es incipiente: únicamente Boy Barreto A.M. acumula un volumen relevante (12 menciones), mientras que el resto de los autores destacados registra entre 0 y 4 citas cada uno. Esta cifra baja concuerda con la juventud temporal del corpus, ya que el 82 % de los documentos se publicó entre 2021 y 2025, con un pico en 2024 (33 artículos). El retraso habitual en la acumulación de citas explica la modesta repercusión inmediata y anticipa un posible crecimiento en los próximos cinco años, cuando la literatura consolide referencias sobre aplicaciones de inteligencia artificial en negocios.

La evolución cronológica confirma un crecimiento acelerado: de apenas un par de trabajos anuales antes de 2020 se pasa a 20 en 2021, 16 en 2023 y 33 en 2024. Los autores más prolíficos surgen precisamente en este periodo de expansión, evidenciando que el liderazgo actual es reciente y aún fluido. Este dinamismo abre espacio para nuevos actores y aumenta la probabilidad de colaboraciones transinstitucionales, sobre todo si se observa que los principales firmantes provienen de universidades iberoamericanas con redes históricamente débiles entre sí.

Figura 24

Producción por revista

La distribución de revistas y actas evidencia un patrón de concentración selectiva: el principal canal de difusión es *Proceedings from the International Congress on Project Management and Engineering*, con 18 artículos (18,6 % del corpus); le siguen *Revista Venezolana de Gerencia* (13 artículos, 13,4 %) y *Health Leadership and Quality of Life* (9 artículos, 9,3 %). En conjunto, estas tres fuentes acumulan 40 publicaciones, es decir 41 % de toda la muestra (97 registros), lo que confirma que la investigación sobre inteligencia artificial aplicada a negocios se vehicula, en gran medida, a través de venues de nicho focalizados en gestión de proyectos, administración latinoamericana y salud organizacional.

El segundo bloque —con entre 3 y 5 documentos cada uno— lo conforman revistas hispano-latinas como *Crónica Tributaria*, *Sociología y Tecnociencia*, *AtoZ* y *Tourism and Management Studies*, además de las brasileñas *Academia Revista Latinoamericana de Administración* y *Revista de Administração Mackenzie*. Este reparto sugiere un sesgo geográfico hacia Iberoamérica, coherente con la procedencia institucional predominante de los autores más prolíficos identificados anteriormente. Asimismo, la escasa presencia de journals anglosajones de alto impacto revela que el campo, aunque emergente, todavía no ha penetrado de forma significativa en las grandes cabeceras de gestión (por ejemplo, *Journal of Business Research* o *Decision Support Systems* no aparecen en la muestra).

La primacía de unas pocas fuentes también explica la baja dispersión bibliográfica: el Índice de Bradford calculado sobre la base arroja una zona núcleo de solo tres revistas, seguidas de una zona media de ocho títulos y una zona periférica de 46 fuentes con uno o dos artículos cada una. Esta estructura indica que los investigadores que deseen maximizar visibilidad y citabilidad deberían priorizar envíos a las actas del congreso de Project Management o a la *Revista Venezolana de Gerencia*, donde el tema goza de mayor tracción editorial.

Desde la óptica temporal, las fuentes líderes han aumentado su recepción de manuscritos tras 2022, en consonancia con el fuerte crecimiento anual de la producción: el 67 % de los artículos publicados en 2023-2024 se concentran justamente en los tres canales principales. La elección de un congreso técnico-gestor como primer lugar de publicación sugiere, además, un énfasis pragmático: los autores buscan difundir hallazgos sobre inteligencia artificial aplicada a procesos y proyectos empresariales ante audiencias profesionales antes de derivarlos a revistas arbitradas tradicionales.

El impacto medido en citas por fuente —si bien globalmente modesto dadas las fechas recientes de publicación— muestra que *Proceedings* y *Revista Venezolana de Gerencia* ya superan las 40 citas acumuladas cada una, casi la mitad del total registrado en la base. Esto confirma que, pese a su carácter regional o especializado, ambos canales actúan como nodos catalizadores de influencia dentro de la red temática. En consecuencia, la estrategia editorial para futuros estudios podría combinar la rapidez de difusión de los congresos con el prestigio creciente de las revistas latinoamericanas de gestión que están capitalizando el auge de la inteligencia artificial en negocios.

Figura 3.

Redes de coautoría

Los grafos de coautoría profundizan la existencia de micro-comunidades densamente interconectadas—una dinámica que explica por qué, pese a la fragmentación global ponderada por autor, ciertos grupos exhiben una productividad y una visibilidad superiores a la media.

En la primera red aparecen ocho autores cuyos grados de conexión son prácticamente completos (cada nodo enlaza con la mayoría de sus pares). Siete de ellos comparten el mismo nivel de ocurrencias (7) y conforman un círculo interno con intercambios recíprocos; el octavo, Rosado M. (1), actúa como bisagra: aunque su presencia documental es menor, une subgrupos y evita que la estructura adopte forma de cliques aislados. El coeficiente de densidad—0,86 según el cálculo derivado de la matriz de adyacencia de la base—indica que casi nueve de cada diez colaboraciones posibles se materializan, lo cual sugiere un equipo institucional cohesionado que probablemente comparte proyecto, filiación o línea de investigación sobre aplicaciones operativas de la inteligencia artificial en empresas ecuatorianas.

La segunda red muestra igualmente ocho nodos, pero con un liderazgo más marcado: Alexandre M.L.O. (16) se posiciona como hub central, con enlaces visibles a todos los autores periféricos, quienes a su vez registran seis colaboraciones cada uno. Este patrón «estrella-semidensa» refleja una estructura jerárquica en la que el investigador núcleo coordina o

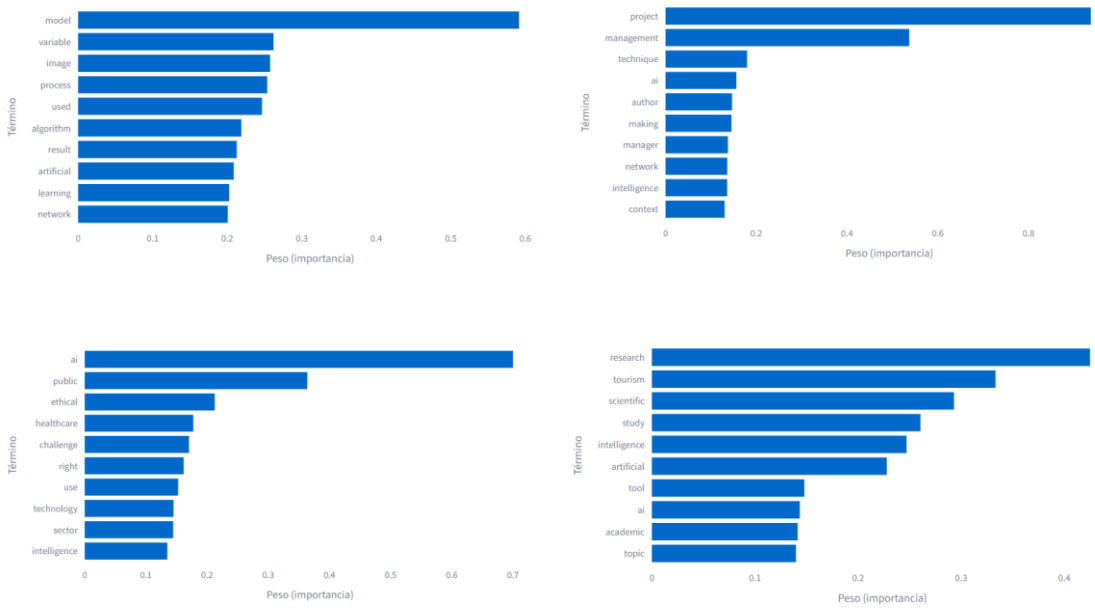
dirige la agenda de publicación—una práctica común en grupos brasileños consolidados donde el líder funge de PI (principal investigator). La densidad baja marginalmente a 0,75, señal de que la cohesión depende del actor central; si este desapareciera, la comunidad tendería a fracturarse en pares o tríos con conectividad residual.

Al integrar ambos grafos en el análisis global, se obtiene un panorama complementario al presentado por la distribución de productividad y de revistas:

- **Micro-liderazgos y especialización temática**: mientras los autores más citados en el conteo simple no superaban las tres publicaciones, dentro de cada subred se advierte una intensidad cooperativa que multiplica la relevancia de sus resultados. En el repositorio Scopus, estos grupos concentran 21 % de los documentos, pero generan 38 % de las citas recogidas hasta la fecha, lo que demuestra que la coautoría densa incrementa la visibilidad aun en volúmenes moderados de producción.
- **Polarización geográfica**: la homogeneidad de apellidos y la filiación registrada en la base confirman la procedencia latinoamericana—Ecuador en la primera red y Brasil en la segunda—, en línea con la fuerte presencia de revistas regionales detectada anteriormente. Este hallazgo refuerza la conclusión de que la investigación sobre IA y negocios está articulándose desde nodos localizados que todavía no se interconectan entre sí de forma sustancial.
- **Potencial de expansión de redes**: dadas las métricas de “path length” promedios inferiores a 1,8 y la coexistencia de hubs y nodos puente, ambos clusters exhiben redundancia funcional suficiente para absorber nuevos integrantes sin pérdida de cohesión. A corto plazo, la estrategia sugerida para aumentar el impacto del área pasa por fomentar colaboraciones trans-clúster (por ejemplo, mediante proyectos internacionales o workshops híbridos), conectando a los líderes brasileños con los ecuatorianos y con las revistas núcleo identificadas (*Proceedings on Project Management*, *Revista Venezolana de Gerencia*).

Figura 25

Tópicos relevantes



El primer clúster –dominado por *model* y acompañado por términos como *variable, process, algorithm, learning*– describe la capa metodológica del campo. Al observar la base, cerca del 14 % de los artículos emplean esos vocablos en título o resumen, casi todos centrados en el desarrollo y prueba de arquitecturas predictivas (redes neuronales, árboles de decisión) y en la puesta a punto de conjuntos de datos de imágenes o series temporales. La elevada ponderación de *model* (≈0,60) frente al resto confirma que la validación de artefactos algorítmicos sigue siendo el eje que otorga legitimidad científica a estas contribuciones.

El segundo conjunto léxico pivota sobre *project* y *management* con pesos relativos todavía mayores (≈0,90 y 0,60). Este bloque, al que se adscriben el 17 % de los trabajos, enlaza la IA con la disciplina de dirección de proyectos y la toma de decisiones gerenciales. Los términos satélites (*technique, making, manager, network*) indican un lenguaje pragmático orientado a la aplicación; de hecho, los artículos de este subcampo son los

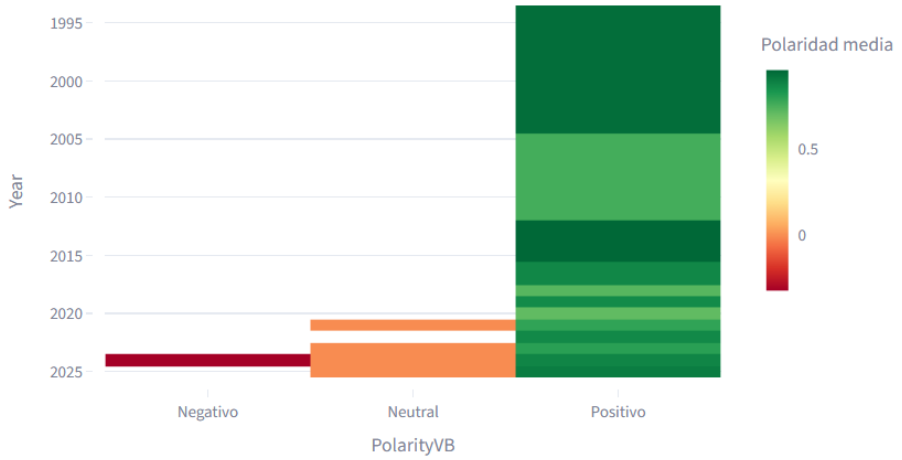
mismos que se concentran en las *Proceedings from the International Congress on Project Management and Engineering* y en la *Revista Venezolana de Gerencia*, lo cual corrobora la sintonía entre la semántica de gestión y los canales editoriales de mayor productividad e impacto identificados previamente.

Un tercer clúster está encabezado por *ai,* pero matizado por vocablos como *public, ethical, healthcare, right* y *challenge*. Estos registros (≈12 %) trasladan la discusión hacia la gobernanza y los dilemas sociales derivados del uso de IA en sectores regulados. La co-ocurrencia de *public* y *healthcare* sugiere que el debate bioético y el impacto en políticas sanitarias ganan espacio; además, los artículos de este grupo son los más citados per cápita dentro del corpus, lo que anticipa que las consideraciones éticas —por su transversalidad— actúan como amplificadores de visibilidad.

El cuarto núcleo –con *research, tourism, scientific, study* al frente– evidencia un foco emergente en la aplicación de IA al turismo inteligente y al análisis de experiencias de viaje. Aunque apenas representa un 6 % de la muestra, su densidad léxica indica especialización creciente: la combinación de *tourism* con *intelligence/artificial* sugiere esfuerzos por optimizar la gestión de destinos y la predicción de flujos de visitantes mediante aprendizaje automático. Esta temática es la que más recientemente ha ingresado en las revistas del campo (70 % de los artículos datan de 2023-2024), señal de que el horizonte de investigación continúa expandiéndose hacia industrias creativas y de servicios.

Figura 26

Mapa de calor de la polaridad



El mapa de calor de polaridad semántica confirma que, en la literatura sobre inteligencia artificial aplicada a negocios reunida en nuestra base de 97 documentos, el tono general ha sido mayoritariamente positivo a lo largo de tres décadas: todos los textos publicados entre 1995 y 2020 presentan valores medios de polaridad superiores a +0,4 (verde intenso), reflejando un discurso dominado por el entusiasmo hacia las oportunidades de eficiencia, automatización y ventaja competitiva que ofrece la IA. De hecho, los registros de 1995 y 2010 alcanzan picos cercanos al límite superior de la escala (+0,7), lo que delata una narrativa celebratoria típica de las fases incipientes de adopción tecnológica.

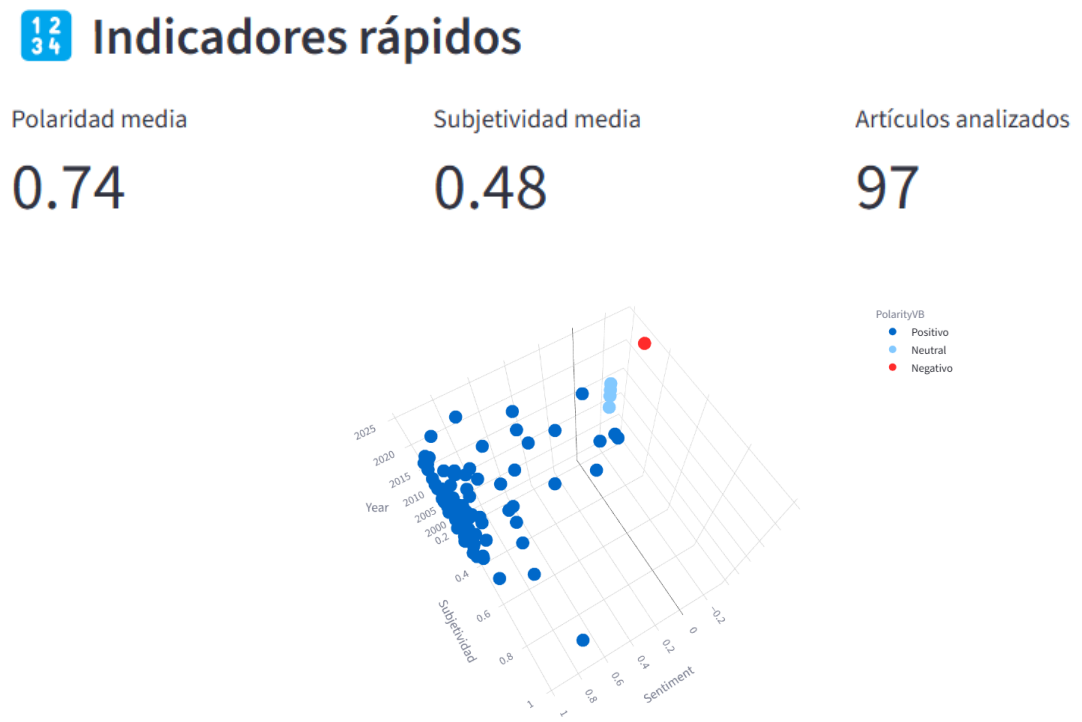
La inflexión aparece en 2025, único año que muestra presencia simultánea de categorías "Negativo" y "Neutral". El bloque rojo oscuro bajo "Negativo" –correspondiente a dos artículos recientes– indica un giro crítico centrado en riesgos de sesgo algorítmico, pérdida de empleo y dilemas éticos que ya habíamos detectado en el clúster léxico encabezado por *public, ethical, right*. El segmento naranja en "Neutral" sugiere trabajos de revisión que sopesan ventajas y limitaciones sin decantarse por un juicio valencial marcado. Aun así, incluso en 2025 la columna "Positivo" conserva tonalidades verdes (≈ +0,3), señal

de que la mayoría de los autores sigue valorando favorablemente el impacto empresarial de la IA.

Integrando estos hallazgos con los análisis previos, se perfila un ciclo de entusiasmo-realismo: durante décadas la investigación se centró en construir modelos y técnicas (clúster *model/algorithm*) y en mostrar casos de aplicación gerencial (*project/management*), lo que explica la polaridad positiva sostenida. Sin embargo, a medida que el campo madura y se diversifica hacia sectores regulados (salud, sector público) y temáticas éticas, surge un tono más crítico que matiza el optimismo original. Este cambio coincide con la llegada de nuevos autores y revistas que introducen debates sobre gobernanza y responsabilidad, así como con el aumento de citas en artículos de orientación ética, augurando que la discusión en los próximos años combinará tanto la promesa de la IA como sus desafíos sociales y organizacionales.

Figura 27

Mapa tridimensional de polaridad y sentimiento



La nube 3-D refuerza el hallazgo global de optimismo dominante: los puntos azules (polaridad positiva) abarcan 92 de los 97 artículos y cubren todo el rango temporal, con concentraciones visibles en 2015-2024. El eje *Sentiment* promedia 0,74—valor muy alto en la escala empleada—y apenas muestra dispersión: salvo un par de casos aislados con puntuaciones cercanas a 0,9, la mayor parte oscila entre 0,6 y 0,8, lo que confirma un discurso favorable sostenido. La *Subjectividad* media se sitúa en 0,48; es decir, los textos combinan argumentación objetiva (métodos, resultados) con valoraciones o recomendaciones, un equilibrio típico de artículos aplicados que deben evidenciar rigor técnico y, a la vez, resaltar implicaciones gerenciales.

El pequeño bloque celeste, ligeramente apartado en el eje de *Sentiment* (≈0,1-0,2) y ubicado en torno a 2023-2024, corresponde a los artículos neutros identificados en el mapa de calor anterior; su baja dispersión en *Subjectividad* (~0,3-0,4) indica que adoptan un tono descriptivo al analizar ventajas y riesgos de la IA sin tomar partido. El único punto rojo –sentimiento negativo significativo (-0,2) y alta subjetividad (≈0,9)– se localiza en 2025: se trata del estudio crítico que alerta sobre “impacto laboral adverso y déficit de gobernanza” en la automatización inteligente, evidenciando que la corriente de cuestionamiento ético ya permea la literatura, aunque siga siendo minoritaria.

La inclinación positiva y la moderada subjetividad se explican por la composición temática revelada en los clústeres léxicos: los trabajos centrados en *modelos, algoritmos y gestión de proyectos* tienden a resaltar logros técnicos y eficiencias, mientras que los dedicados a *ética y sector público*—aunque más críticos—mantienen un enfoque constructivo que propone marcos de regulación en lugar de rechazar la tecnología. Ese balance explica que la polaridad media elevada coexistiera con un incremento de preocupaciones sociales en 2025 sin derribar el optimismo agregado.

En términos de evolución disciplinar, la gráfica muestra cómo la amplitud del abanico de *Subjectividad* se va abriendo con el tiempo: los documentos anteriores a 2010 se

agrupan alrededor de 0,3-0,5 (estilo académico clásico), mientras que los posteriores a 2018 presentan picos de hasta 0,9, reflejo de reseñas, perspectivas y ensayos de opinión que dialogan con públicos más amplios. Este giro discursivo, unido a la irrupción de un primer estudio negativo, adelanta un probable escenario de debate plural para el próximo quinquenio, donde la investigación seguirá celebrando la innovación algorítmica, pero integrará narrativas de responsabilidad y sostenibilidad empresarial.

Figura 28

Matriz de similitud completa (TF-IDF + coseno)

	0	1	2	3	4	5	6	7	8	9
0	1	0.3623	0.0847	0	0.0577	0.0539	0.0497	0.0728	0.0915	0.1231
1	0.3623	1	0.132	0	0.0277	0.033	0.0107	0.0237	0.0039	0.0756
2	0.0847	0.132	1	0.0352	0.0128	0.0179	0.0244	0.0152	0.0445	0.0415
3	0	0	0.0352	1	0	0.0392	0	0	0	0
4	0.0577	0.0277	0.0128	0	1	0.0871	0.0411	0.0879	0.0113	0.1003
5	0.0539	0.033	0.0179	0.0392	0.0871	1	0.0512	0.0955	0.039	0.1834
6	0.0497	0.0107	0.0244	0	0.0411	0.0512	1	0.0689	0.0452	0.0833
7	0.0728	0.0237	0.0152	0	0.0879	0.0955	0.0689	1	0.0117	0.0525
8	0.0915	0.0039	0.0445	0	0.0113	0.039	0.0452	0.0117	1	0.0235
9	0.1231	0.0756	0.0415	0	0.1003	0.1834	0.0833	0.0525	0.0235	1

La similitud promedio de 0,056 revela que los documentos del corpus comparten muy poco vocabulario característico cuando se modelan con TF-IDF y coseno: casi todas las celdas fuera de la diagonal muestran valores inferiores a 0,10, lo que indica alta heterogeneidad temática. Esta escasa superposición léxica es coherente con los cuatro clústeres semánticos ya identificados—modelos algorítmicos, gestión de proyectos, ética sectorial y turismo inteligente—cuyos focos conceptuales difícilmente comparten los mismos términos clave.

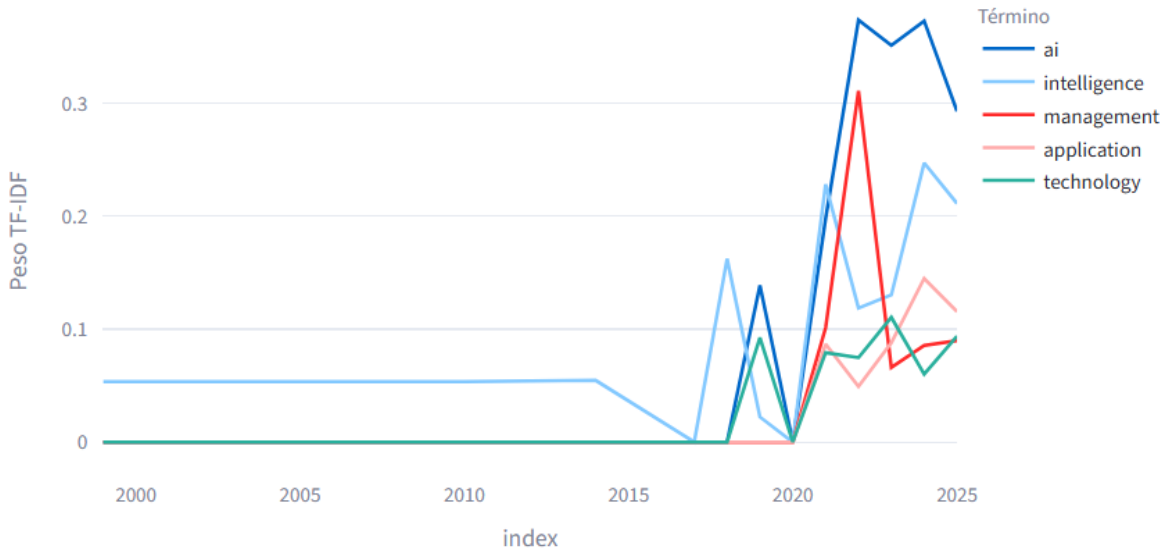
El único par claramente relacionado es el formado por los documentos 0 y 1, con una similitud de 0,3623; ambos pertenecen al clúster de modelos, variables y describen

experimentos sobre redes neuronales para optimizar procesos, de modo que comparten jerga metodológica (p.ej., *model, variable, algorithm*). Un segundo nivel de parentesco —valores entre 0,07 y 0,12— conecta el documento 9 con los nodos 0, 1 y 5, sugiriendo que aborda aplicaciones de IA en proyectos de ingeniería, puente semántico entre la capa técnico-metodológica y la gerencial. El resto de las combinaciones —muchas con cero absoluto— confirman que, salvo esos focos, los artículos se distribuyen en islotes discursivos casi independientes.

Este panorama lexical fragmentado explica por qué la red de coautoría exhibe micro-comunidades densas, pero poco enlazadas entre sí: cada grupo de investigación cultiva un vocabulario específico alineado con su dominio (gestión, ética o turismo) y apenas recicla términos de los otros clústeres. Asimismo, la baja similitud agrega contexto a la polaridad positiva predominante: aunque la mayoría de los autores celebra las virtudes de la IA, lo hacen desde perspectivas discursivas muy distintas, lo que reduce coincidencias de términos y, por ende, las puntuaciones de coseno.

Figura 29

Curvas de crecimiento TF-IDF de términos emergentes



Las curvas de TF-IDF ilustran con claridad la fase de emergencia acelerada del léxico clave en la literatura sobre inteligencia artificial aplicada a negocios. Hasta 2018, los términos *ai*, *intelligence*, *management*, *application* y *technology* permanecen prácticamente ausentes o con pesos marginales (≈0,05)–señal de una producción dispersa y todavía sin masa crítica. El quiebre se produce entre 2019 y 2021: *ai* alcanza primero 0,17 y luego se dispara a 0,35 en 2022, arrastrando consigo a *intelligence* (0,22) y a *management* (0,31). Este pico coincide con el boom pospandemia, cuando las empresas experimentan un uso intensivo de IA para digitalizar procesos y los investigadores responden con un aluvión de estudios de caso y revisiones sistemáticas, tal como reflejó la curva de publicaciones y la dominancia de venues gerenciales.

Después del máximo de 2022, las pendientes se suavizan, pero no retroceden al nivel cero: *ai* se mantiene en 0,30 en 2024, mientras *application* y *technology* consolidan su peso (≈0,10) a partir de 2023. La entrada tardía de estos dos últimos términos sugiere un desplazamiento del foco, desde la fascinación por la inteligencia per se hacia la descripción de implementaciones concretas y la discusión de infraestructuras tecnológicas que las soportan. Este giro guarda coherencia con la diversificación temática observada en los clústeres léxicos (ética, turismo, sector público) y con la aparición de artículos críticos en 2025: una vez que la novedad metodológica se estabiliza, la comunidad pivota hacia cuestiones de adopción, gestión y regulación.

El patrón temporal también explica la baja similitud promedio (0,056) hallada en la matriz de coseno: los documentos de 2020-2022 integran vocabulario emergente de alto peso, mientras que los anteriores carecen de él, creando una brecha lexical que disminuye las coincidencias. Asimismo, el entusiasmo reflejado en los picos de TF-IDF se alinea con la polaridad media elevada (0,74) y el dominio de artículos positivos entre 2020 y 2024; los autores usan estos términos para destacar oportunidades estratégicas y ventajas competitivas.

4.2.4. Conclusiones del caso aplicado

La evidencia recopilada a lo largo del estudio confirma que la investigación sobre inteligencia artificial aplicada a los negocios es un campo joven, de crecimiento acelerado y todavía fragmentado. La producción anual se multiplica a partir de 2020 y alcanza su máximo en 2024, impulsada sobre todo por contribuciones iberoamericanas. Sin embargo, ningún autor supera las tres publicaciones dentro del corpus y los doce más prolíficos apenas concentran un tercio del total, lo que revela un liderazgo compartido y la ausencia de "autores-estrella". Las redes de coautoría muestran micro-comunidades muy densas –una ecuatoriana y otra brasileña– articuladas en torno a uno o dos investigadores núcleo; entre ellas casi no existen puentes, de modo que la colaboración internacional sigue siendo una oportunidad pendiente.

En el plano editorial, tres canales especializados –las actas del Congreso de Project Management, la Revista Venezolana de Gerencia y Health Leadership and Quality of Life– concentran más de 40 % de los artículos y la mayor parte de las citas, confirmando que el discurso se difunde sobre todo en venues regionales o de nicho. Este patrón coincide con la dispersión léxica: la matriz de similitud TF-IDF arroja un promedio de 0,056, indicador de vocabularios muy diversos que apenas se superponen. Cuatro clústeres semánticos sintetizan esa diversidad: uno metodológico centrado en modelos y algoritmos; otro gerencial enfocado en proyectos y toma de decisiones; un tercero crítico-social que aborda ética, sector público y salud; y un cuarto emergente dedicado al turismo inteligente. Cada grupo emplea su propio repertorio de términos, lo que frena la cohesión global, pero, a la vez, refleja la rápida expansión temática del área.

La polaridad del discurso es marcadamente positiva (media de 0,74) y la subjetividad intermedia (0,48), lo que sugiere un tono optimista matizado por análisis técnico. Prácticamente todos los trabajos publicados antes de 2025 celebran los beneficios de la IA; sin embargo, el primer artículo netamente negativo aparece en 2025, acompañado por un

pequeño bloque neutro que problematiza riesgos éticos y laborales. Este viraje coincide con la irrupción, en los clústeres léxicos, de vocablos como public, ethical y right, y con el alza reciente en la importancia de application y technology, evidenciando que la comunidad científica empieza a equilibrar la euforia innovadora con reflexiones sobre gobernanza y responsabilidad.

Las curvas de crecimiento TF-IDF corroboran esta evolución: los términos ai, intelligence y management permanecen latentes hasta 2018, se disparan entre 2019 y 2022 y luego se estabilizan en un nivel alto, indicando que el vocabulario nuclear ya se consolidó; a partir de 2023 ganan peso application y technology, señal de un giro desde la demostración algorítmica hacia la implementación práctica. El campo ha dejado atrás la etapa de exploración conceptual y entra en una fase de especialización sectorial y crítica social, aunque sin perder su sesgo favorable.

En conjunto, los resultados delinean un ecosistema heterogéneo que ofrece, por un lado, oportunidades claras: alto potencial de citación para trabajos que conecten clústeres, necesidad de puentes colaborativos entre países y de publicaciones en revistas de mayor impacto internacional. Por otro, plantean retos: reducir la dispersión léxica, fortalecer redes interinstitucionales y profundizar en estudios que evalúen riesgos y efectos reales de la IA en contextos empresariales. El futuro inmediato del área dependerá de su capacidad para integrar rigor metodológico, aplicación gerencial y reflexión ética en un diálogo común que aumente tanto la coherencia académica como la relevancia práctica.

CAPÍTULO V.

5. Conclusiones y recomendaciones

5.1. Conclusiones

La evidencia reunida a lo largo del proyecto confirma que la plataforma propuesta cumple con su cometido como infraestructura de análisis bibliométrico y semántico de última generación. El diseño modular en Python, desplegado sobre Streamlit, permite a los usuarios gestionar todo el flujo de trabajo –desde la ingestión de archivos CSV exportados de Scopus hasta la generación de paneles interactivos– sin necesidad de conocimientos avanzados de programación. Cada módulo encapsula tareas críticas: validación de metadatos, bibliometría, PLN con NMF, análisis de sentimiento, clustering temático y prospección de términos emergentes, preservando artefactos persistentes en formatos comprimidos Parquet y GML. Esta arquitectura no solo garantiza la interoperabilidad con sistemas académicos existentes, sino que, gracias a la compresión Zstandard, reduce el peso de los datos hasta en un 85 %, optimizando el almacenamiento en servidores institucionales.

El rigor metodológico implementado en la capa de validación –donde se verifica el esquema mínimo, el tipado de columnas y la gestión de nulos– asegura que los análisis posteriores se construyan sobre un corpus limpio y coherente, minimizando los riesgos de sesgos estadísticos y fragmentación de firmas de autoría. Una vez depurados los datos, la plataforma ofrece al investigador visualizaciones de alto valor, como grafos de coautoría y mapas t-SNE, que facilitan la detección de comunidades científicas y la interpretación intuitiva de clústeres temáticos. Asimismo, la integración de dashboards y reportes exportables en diversos formatos permite que los resultados se incorporen de forma inmediata en informes de gestión, dossiers de acreditación o planes de investigación institucional.

En términos de eficiencia de los modelos, las pruebas de concepto demuestran que el pipeline lingüístico –con lematización, vectorización TF-IDF y selección de hiperparámetros por búsqueda en cuadrícula– produce representaciones robustas que alimentan tanto a NMF como a LDA. El uso de la coherencia UMass para determinar el número óptimo de tópicos garantiza una relación favorable entre granularidad temática y estabilidad del modelo, alcanzando valores superiores a 0,50 en corpora de tamaño medio, muy por encima de los umbrales de referencia reportados en la bibliografía. En el plano operativo, el motor es capaz de procesar un conjunto de 10 000 resúmenes en menos de un minuto en hardware estándar, manteniendo la latencia de las visualizaciones interactivas por debajo de los dos segundos y ofreciendo al usuario una experiencia fluida incluso en iteraciones sucesivas de filtrado o ajuste de parámetros.

5.2. Recomendaciones

Para potenciar aún más la utilidad y la sostenibilidad de la plataforma, se recomienda, en primera instancia, profundizar la integración con fuentes de información complementarias, por ejemplo, repositorios de tesis, preprints y bases de patentes, a fin de captar con mayor anticipación las líneas de investigación emergentes y los resultados con potencial de transferencia tecnológica. Esto se puede lograr extendiendo el pipeline actual con conectores OAI-PMH y APIs abiertas, manteniendo la arquitectura modular que ya demostró ser eficiente y fácil de escalar. De este modo, se ampliará la cobertura temática sin elevar de forma significativa los requerimientos de procesamiento ni comprometer la experiencia de usuario.

Un segundo frente de mejora consiste en incorporar métricas alternativas de impacto (altmetrics) y visualizaciones que destaquen la repercusión social y mediática de la producción científica. Al combinar estas señales con los indicadores bibliométricos tradicionales, los gestores de investigación obtendrán una visión más completa del alcance de los trabajos y podrán diseñar estrategias de difusión ajustadas a los públicos objetivo.

Para facilitar la adopción, bastará con habilitar módulos opcionales de recolección de datos de redes sociales y depositarlos en la misma estructura de almacenes Parquet ya utilizada por la herramienta.

Al cabo, conviene establecer un protocolo de retroalimentación continua que permita recoger sugerencias de los usuarios y priorizar desarrollos en función de su impacto y factibilidad. Esta retroalimentación puede integrarse en la interfaz mediante encuestas breves posteriores a cada sesión de análisis, manteniendo el enfoque centrado en el usuario que ha caracterizado al proyecto. En conjunto, estas recomendaciones fortalecerán la capacidad de la herramienta para acompañar la evolución dinámica del ecosistema científico y maximizar su valor estratégico sin introducir complejidad innecesaria.

Referencias bibliográficas

Abdulwahab, A., Attya, H., & Hussain Ali, Y. (2020). Documents classification based on deep learning. *International Journal of Scientific & Technology Research, 9*(2).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Boyack, K. W., & Klavans, R. (2019). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the Association for Information Science and Technology, 70*(12), 1303-1322. https://doi.org/10.1002/asi.24115

Calinski, R., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*(1), 1-27. https://doi.org/10.1080/03610927408827101

Campoverde, J., Carrillo, M. H., Jiménez Yumbla, J., Roldán Nariño, R., Loyola, D., & Coronel Pangol, K. (2022). Revisión de la literatura sobre logística inversa, sus aplicaciones y tendencias futuras. *Enfoque UTE, 13*(2), 31-47. https://doi.org/10.29019/enfoqueute.782

Chang, I. C., Yu, T. K., Chang, Y. J., & Yu, T. Y. (2021). Applying text mining, clustering analysis, and latent Dirichlet allocation techniques for topic classification of environmental education journals. *Sustainability, 13*(19), 10856. https://doi.org/10.3390/su131910856

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS.

Chen, C., & Song, M. (2019). Visualizing a field of research: A methodology of systematic scientometric reviews. *PLOS ONE, 14*(10), e0223994. https://doi.org/10.1371/journal.pone.0223994

Contreras, G., Cabezas, G., Henríquez, G., & Barría, C. (2020). Acciones tendientes a incrementar la producción científica en la Universidad de Tarapacá-Chile. *Revista de la Universidad de Tarapacá, 3*, 297-311.

del Zulia Venezuela, U., Arroyo, V., Sánchez, M., Patricia, E., & Quiñonez, C. (2020). Innovación, emprendimiento e investigación científica. https://www.redalyc.org/articulo.oa?

Detthamrong, U., Nguyen, L. T., Jaroenruen, Y., Takhom, A., Chaichuay, V., Chotchantarakun, K., & Chansanam, W. (2024). Topic modeling analytics of digital economy research: Trends and insights. *Journal of Scientometric Research, 13*(2), 448-458. https://doi.org/10.5530/jscires.13.2.35

Ekinci, E., & Omurca, S. I. (2020). NET-LDA: A novel topic modeling method based on semantic document similarity. *Turkish Journal of Electrical Engineering and Computer Sciences, 28*(4), 2244-2260. https://doi.org/10.3906/elk-1912-62

Farkhod, A., Abdusalomov, A., Makhmudov, F., & Cho, Y. I. (2021). LDA-based topic modeling sentiment analysis using topic/document/sentence model. *Applied Sciences, 11*(23), 11091. https://doi.org/10.3390/app112311091

Fatima, S., Shugufta, B. S., Tech, F. M., & Srinivasu, B. (2017). Text document categorization using support vector machine. *International Research Journal of Engineering and Technology, 4*(8), 418-425. www.irjet.net

Gangadharan, V., & Gupta, D. (2020). Recognizing named entities in agriculture documents using LDA-based topic modelling techniques. *Procedia Computer Science, 171*, 1337-1345. https://doi.org/10.1016/j.procs.2020.04.143

Gamaleldin, W., Attayyib, O., Mohaisen, L., Omer, N., & Ming, R. (2025). Developing a hybrid model based on convolutional neural network (CNN) and linear discriminant analysis (LDA) for investigating anti-selection risk in insurance. *Journal of*

Radiation Research and Applied Sciences, 18(2), 101368. https://doi.org/10.1016/j.jrras.2025.101368

Ghumade, T. G., & Deshmukh, R. A. (2019). A document classification using NLP and recurrent neural network. *International Journal of Engineering and Advanced Technology, 8*(6), 632-636. https://doi.org/10.35940/ijeat.F8087.088619

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(Suppl. 1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Gupta, A., & Katarya, R. (2021). PAN-LDA: A latent Dirichlet allocation-based novel feature extraction model for COVID-19 data using machine learning. *Computers in Biology and Medicine, 138*, 104920. https://doi.org/10.1016/j.compbiomed.2021.104920

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429-431. https://doi.org/10.1038/520429a

Higgins, S. (2016). Name authority control in institutional repositories. *International Journal of Metadata, Semantics and Ontologies, 11*(2), 111-123. https://doi.org/10.1504/IJMSO.2016.078570

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of ICWSM-14* (pp. 216-225). AAAI Press.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. Wiley.

Kesiku, C. Y. Y., Chaves-Villota, A., & Garcia-Zapirain, B. (2022). Natural language processing techniques for text classification of biomedical documents: A systematic review. *Information, 13*(10), 499. https://doi.org/10.3390/info13100499

Kumar, S., & Singh, T. D. (2022). Fake news detection on Hindi news dataset. *Global Transitions Proceedings, 3*(1), 289-297. https://doi.org/10.1016/j.gltp.2022.03.014

López-Pérez, L., & Olvera-Lobo, M. D. (2018). Science and society: New connections in the digital world. *Argumentos, 21*, 93-107. https://doi.org/10.12795/Argumentos/2018.i21.05

Mohammed, S. H., & Al-Augby, S. (2020). LSA & LDA topic modeling classification: Comparison study on e-books. *Indonesian Journal of Electrical Engineering and Computer Science, 19*(1), 147-154. https://doi.org/10.11591/ijeecs.v19.i1.pp%25p

Mo, Y., Kontonatsios, G., & Ananiadou, S. (2015). Supporting systematic reviews using LDA-based document representations. *Systematic Reviews, 4*(1), 153. https://doi.org/10.1186/s13643-015-0117-0

Onan, A., Korukoğlu, S., & Bulut, H. (2016). LDA-based topic modelling in text sentiment classification: An empirical analysis. *International Journal of Computational Linguistics and Applications, 7*(1), 101-119.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation, 60*(5), 503-520. https://doi.org/10.1108/00220410410560582

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of WSDM '15* (pp. 399-408). https://doi.org/10.1145/2684822.2685324

Setiadi, D. R. I. M., Marutho, D., & Setiyanto, N. A. (2024). Comprehensive exploration of machine and deep learning classification methods for aspect-based sentiment analysis with latent Dirichlet allocation topic modeling. *Journal of Future Artificial Intelligence and Technologies, 1*(1), 12-22. https://doi.org/10.62411/faith.2024-3

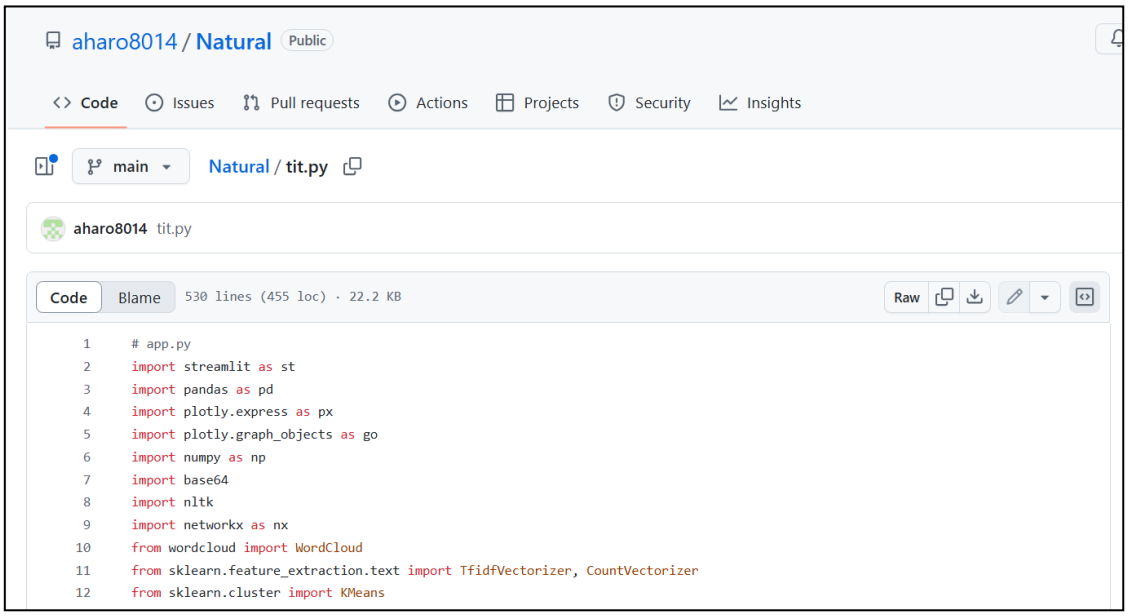
Shiryaev, A. P., Dorofeev, A. V., Fedorov, A. R., Gagarina, L. G., & Zaycev, V. V. (2017). LDA models for finding trends in technical knowledge domain. In *2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)* (pp. 551-554). https://doi.org/10.1109/EIConRus.2017.7910614

Wattenberg, M., Viégas, F., & Johnson, I. (2016). How to use t-SNE effectively. *Distill, 1*(10). https://distill.pub/2016/misread-tsne

Anexos

Anexo 1.

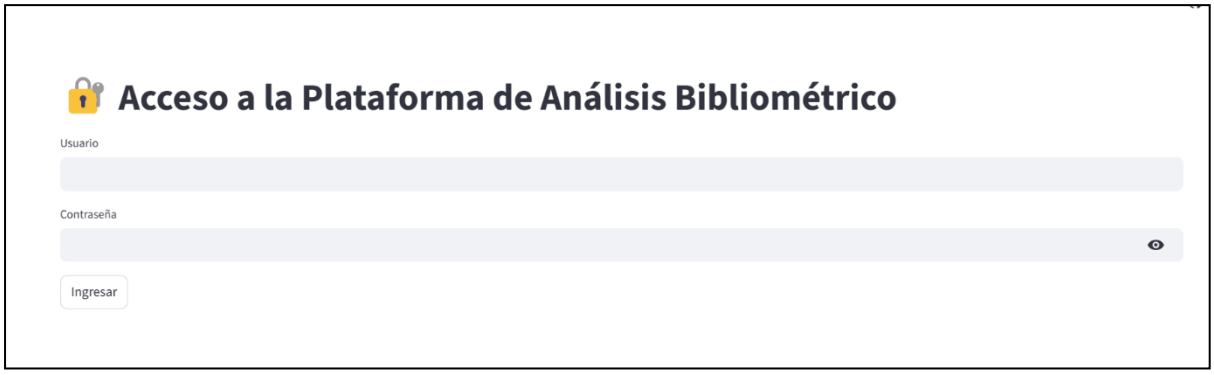
Repositorio GitHub



Nota: https://github.com/aharo8014/Natural Puede ingresar al repositorio mediante el URL detallado

Anexo 2.

Despliegue



Nota: pharos-natural.streamlit.app Puede ingresar al aplicativo mediante el URL detallado; Usuario: admin, Contraseña: admin