

Alineación temática entre publicaciones científicas en ingeniería y Scopus: una evaluación basada en Latent Dirichlet Allocation

Thematic Alignment between Scientific Publications in Engineering and Scopus: An Evaluation Based on Latent Dirichlet Allocation

JUAN DANIEL TRIANA MANJARRÉS¹ - WILMER PINEDA-RÍOS² -
ANA MARÍA GÓMEZ LAMUS³

1. Magíster en Ciencia de Datos. Universidad Escuela Colombiana de Ingeniería Julio Garavito. juan.triana-m@escuelaing.edu.co, ORCID: <https://orcid.org/0009-0006-2217-5065>.
2. PhD(c) en Ciencias-Estadística. Universidad Escuela Colombiana de Ingeniería Julio Garavito. wilmer.pineda@escuelaing.edu.co, ORCID: <https://orcid.org/0000-0001-7774-951X>.
3. Magíster en Estadística. Universidad Escuela Colombiana de Ingeniería Julio Garavito. ana.gomez@escuelaing.edu.co, ORCID: <https://orcid.org/0000-0002-9714-6780>.

Recibido: 15/10/2025 Aceptado: 02/11/2025

Disponible en http://www.escuelaing.edu.co/es/publicaciones_revista
<http://revistas.escuelaing.edu.co/index.php/reci>

Resumen

En este artículo se analiza la alineación temática entre las publicaciones científicas de la Universidad Escuela Colombiana de Ingeniería Julio Garavito y las publicaciones indexadas en Scopus en el campo de la ingeniería. Para esto, se recopilaron 694 artículos del repositorio institucional y 1310 artículos de Scopus, correspondientes al periodo 2020-2024, mediante técnicas de *web scraping* con BeautifulSoup y automatización con Selenium. Los textos se preprocesaron mediante tokenización, eliminación de *stopwords*, lematización y traducción al inglés. Posteriormente, se entrenó un modelo de Latent Dirichlet Allocation (LDA), seleccionando cuatro temas como cantidad óptima con base en la métrica de perplexidad.

Los resultados revelan una convergencia parcial entre ambas fuentes: los temas de la universidad en áreas de tecnología y materiales presentan similitudes con los de Scopus (similitud coseno máxima del 20,14 %), mientras que se identificaron diferencias en enfoques como ingeniería de estructuras y transporte (universidad) frente a educación y procesos industriales (Scopus). Estos hallazgos ofrecen un diagnóstico cuantitativo útil para orientar las líneas de investigación institucional hacia los estándares nacionales e internacionales.

En el presente estudio se evalúan la alineación temática entre la producción científica de la Universidad Escuela Colombiana de Ingeniería Julio Garavito y la bibliografía indexada en Scopus en el campo de la ingeniería. Se conformó un corpus de 694 artículos institucionales y 1310 registros de Scopus (2020-2024), procesados mediante técnicas estándar de procesamiento de lenguaje natural. Se aplicó el modelo Latent Dirichlet Allocation (LDA) a ambos corpus y se identificaron

cuatro temas latentes en cada uno. La similitud coseno entre las distribuciones de temas reveló una convergencia parcial —más marcada en aspectos de tecnología (20,14 %)—, junto a divergencias significativas: la universidad presenta mayor actividad en ingeniería de estructuras y transporte, mientras que en Scopus predominan temas de ciencia de materiales, educación y procesos energéticos.

Estos resultados constituyen una línea base empírica para orientar la agenda investigativa institucional hacia las tendencias científicas del campo.

Palabras claves: Latent Dirichlet Allocation, modelado de temas, procesamiento de lenguaje natural, bibliometría, producción científica.

Abstract

This study examines the thematic alignment between the scientific publications of the Escuela Colombiana de Ingeniería Julio Garavito and Scopus-indexed publications in the field of engineering. A total of 694 articles from the institutional repository and 1,310 articles from Scopus, spanning the 2020-2024 period, were collected using web scraping techniques with BeautifulSoup and browser automation with Selenium. Text preprocessing included tokenization, stopword removal, lemmatization, and translation into English. A Latent Dirichlet Allocation (LDA) model was then trained, with four topics selected as the optimal number based on the perplexity metric.

The results indicate partial thematic convergence: university topics related to technology and materials showed similarities with Scopus topics (maximum cosine similarity of 20.14%), while notable differen-

ces emerged in areas such as structural engineering and transportation (university) versus education and industrial processes (Scopus). These findings provide a quantitative assessment that can guide institutional research priorities toward national and international standards.

This study assesses the thematic alignment between the scientific output of the Escuela Colombiana de Ingeniería Julio Garavito and Scopus-indexed literature in the field of engineering. A corpus of 694 institutional articles and 1,310 Scopus records (2020-2024) was compiled and processed using standard natural language processing techniques. Latent Dirichlet Allocation (LDA) was applied to both corpora, identifying four latent topics in each. Cosine similarity between topic distributions revealed partial convergence — most notably in technology-related themes (20.14%) alongside significant divergences: the university shows greater activity in structural engineering and transportation, while Scopus is dominated by materials science, education, and energy processes.

These results provide an empirical baseline for aligning the institutional research agenda with prevailing scientific trends in the field.

Keywords: Latent Dirichlet Allocation, topic modeling, natural language processing, bibliometrics, scientific production.

INTRODUCCIÓN

En Colombia, la inversión en investigación y desarrollo ha crecido de manera sostenida en la última década (Marroquín y Ríos, 2012), lo que ha impulsado un aumento en la producción científica de las instituciones de educación superior (IES). No obstante, un mayor volumen de publicaciones no garantiza que estas se alineen con las tendencias temáticas reconocidas en la comunidad científica internacional. Evaluar dicha correspondencia es fundamental para que las universidades orienten estratégicamente sus esfuerzos investigativos (González et al., 2020).

La Universidad Escuela Colombiana de Ingeniería Julio Garavito es una institución de referencia en formación e investigación en ingeniería en Colombia. A pesar de su trayectoria, no se cuenta con un análisis sistemático que cuantifique en qué medida su producción científica se alinea con los temas predominantes en bases de datos de referencia como Scopus. Esta ausencia de evidencia empírica dificulta la identificación de fortalezas y brechas en la agenda investigativa institucional.

En el presente estudio se aborda este problema aplicando el modelo Latent Dirichlet Allocation (LDA) (Blei et al., 2003) a un corpus de publicaciones de la universidad y de Scopus, con el objetivo de determinar el grado de alineación temática entre ambas fuentes mediante la comparación de temas latentes y métricas de similitud coseno.

La alineación de las publicaciones académicas con los estándares y temas reconocidos en los ámbitos nacional e internacional es un factor clave para garantizar la relevancia y el impacto de la investigación universitaria (Aria y Cuccurullo, 2017). Evaluar la correspondencia entre las investigaciones de una institución y los criterios de repositorios de referencia como Scopus permite comprender su posicionamiento en la comunidad académica global y detectar oportunidades de fortalecimiento.

La Universidad Escuela Colombiana de Ingeniería Julio Garavito es una institución de referencia en la formación e investigación en ingeniería en Colombia. Sin embargo, no se ha hecho un análisis sistemático que determine en qué medida sus publicaciones están alineadas con las tendencias temáticas reconocidas en el campo de la ingeniería a escala nacional. Este vacío representa una oportunidad para identificar fortalezas, así como áreas en las que la producción científica se podría orientar hacia temas de mayor impacto.

Para abordar este problema, en el presente estudio se emplea el modelo Latent Dirichlet Allocation (LDA), una técnica de minería de texto que permite identificar temas latentes en colecciones de documentos (Blei et al., 2003). Mediante la comparación de los temas extraídos de las publicaciones institucionales con aquellos provenientes de Scopus, se busca ofrecer un diagnóstico cuantitativo de la convergencia temática entre ambas fuentes.

El objetivo de este artículo es determinar el grado de alineación temática entre las publicaciones científicas de la Universidad Escuela Colombiana de Ingeniería Julio Garavito y las publicaciones indexadas en Scopus en el campo de la ingeniería, utilizando modelado de temas con LDA y métricas de similitud como herramientas de evaluación.

ESTADO DEL ARTE

Indicadores bibliométricos y sus limitaciones

Los indicadores bibliométricos —como el factor de impacto y el índice H— son herramientas ampliamente utilizadas para evaluar la calidad y visibilidad de las publicaciones científicas (González et al., 2020). Sin embargo, estos indicadores miden el impacto cuantitativo de las publicaciones sin captar su dimensión temática; es decir, permiten saber cuánto se publica y dónde, pero no sobre qué. Esta limitación ha motivado el desarrollo

de enfoques complementarios, basados en el análisis del contenido textual de las publicaciones.

Modelado de temas en el análisis de bibliografía científica

El modelado de temas permite identificar patrones temáticos subyacentes en grandes colecciones de texto (DiMaggio et al., 2013). Entre las técnicas disponibles, Latent Dirichlet Allocation (LDA) se ha consolidado como un referente para este propósito (Blei et al., 2003). LDA es un modelo probabilístico generativo que asume que cada documento es una mezcla de temas, y cada tema, una distribución sobre palabras. Se ha aplicado con éxito en la exploración de bibliografía científica (Jelodar et al., 2017; Asmussen y Møller, 2019), en revisiones sistemáticas (Grisales y Figueroa, 2022) y en el análisis de producción académica por disciplinas (Kozłowski et al., 2021).

Existen alternativas como Latent Semantic Analysis (Deerwester et al., 1990), Non-negative Matrix Factorization (Hofmann, 1999) y modelos basados en *embeddings* como BERT (Ambalavanan y Devarakonda, 2020). No obstante, se seleccionó LDA para este estudio, dado que su naturaleza probabilística permite interpretar directamente las distribuciones temáticas, se ha validado ampliamente en contextos bibliométricos y su costo computacional es adecuado para corpus de tamaño moderado.

Antecedentes en evaluación temática de producción científica

En varios estudios se ha empleado modelado de temas para diagnosticar la producción científica institucional o regional. Kozłowski et al. (2021) aplicaron LDA al análisis de patrones en comercio internacional, demostrando la capacidad del modelo para identificar estructuras temáticas en corpus especializados. Grisales y Figueroa (2022) utilizaron LDA para mapear el rol del aprendizaje automático en revisiones sistemáticas, identificando temas claves y autores relevantes. En el contexto colombiano, Rubiano (2022) aplicó el modelo al análisis de contenido en redes sociales, validando su utilidad para corpus en español.

Sin embargo, no se identificaron estudios previos que comparan sistemáticamente la producción científica de una universidad colombiana con las tendencias

temáticas de bases indexadas como Scopus, empleando modelado de temas.

MARCO TEÓRICO: FUNDAMENTOS DEL MODELO LDA

El modelo LDA se fundamenta en la distribución de Dirichlet, una distribución de probabilidad multivariada sobre vectores de probabilidades $(\theta_1, \dots, \theta_m)$ con $\theta_i \geq 0$ y $\sum \theta_i = 1$, parametrizada por un vector de concentración $\alpha = (\alpha_1, \dots, \alpha_m)$ con $\alpha_i > 0$ (Blei et al., 2003).

En el marco bayesiano, esta distribución actúa como prior conjugada de la distribución multinomial, lo que permite obtener distribuciones posteriores en forma analítica. En LDA, la Dirichlet se utiliza para modelar dos distribuciones claves: la proporción de temas en cada documento y la proporción de palabras en cada tema.

Proceso generativo

LDA es un modelo generativo probabilístico. Para esto, Blei, Ng y Jordan (2003) definen un corpus como una colección de texto con entidades como palabras. La idea básica es que los documentos se representan como mezclas aleatorias de temas latentes, donde cada tema se caracteriza por una distribución de palabras. LDA asume el siguiente proceso generativo para cada documento w en un corpus D :

1. Se selecciona el número de palabras N a partir de una distribución Poisson(ξ).
2. Se selecciona una distribución de temas θ a partir de una distribución Dir(α).
3. Para cada una de las N palabras:
 - a) Se elige un tema z_n de la distribución multinomial (θ).
 - b) Se elige una palabra w_n de la distribución $p(w_n | z_n, \beta)$, condicionada al tema seleccionado.

De esta manera, cada documento se representa como una mezcla de temas latentes, y cada tema como una distribución sobre el vocabulario del corpus.

Inferencia

El cálculo exacto de la distribución posterior en LDA es difícil debido al acoplamiento entre los parámetros

θ y β (Blei et al., 2003). Para resolver esto, se recurre a métodos de inferencia aproximada. La inferencia variacional busca una distribución candidata más simple que aproxime la posterior, minimizando la divergencia de Kullback-Leibler. El muestreo de Gibbs es una alternativa basada en simulación. En este estudio se empleó la implementación de Gensim en Python, que utiliza inferencia variacional.

Supuestos y limitaciones del modelo

LDA opera con dos supuestos fundamentales: independencia entre documentos y una representación de bolsa de palabras que ignora el orden léxico. Esto implica que el modelo no capta relaciones semánticas complejas ni dependencias sintácticas. Adicionalmente, el número de temas se debe definir *a priori*, lo cual requiere una métrica de selección —en este caso, la perplejidad—. Estas limitaciones se consideran en la interpretación de los resultados presentados más adelante, en la sección “Resultados y discusión”.

METODOLOGÍA

El proceso metodológico se organizó en cuatro etapas: 1) recopilación de datos; 2) preprocesamiento de texto; 3) selección del número de temas, y 4) entrenamiento del modelo LDA.

Recopilación de datos

Se recopilieron artículos de dos fuentes para el periodo 2020-2024:

- **Repositorio institucional.** Se extrajeron 694 artículos (tesis de grado y trabajos dirigidos) del repositorio virtual de la Universidad Escuela Colombiana de Ingeniería (<https://repositorio.escuelaing.edu.co>), mediante *web scraping* con la librería BeautifulSoup en Python.
- **Scopus.** Se extrajeron 1310 artículos indexados en Scopus, filtrados por el área temática de ingeniería y por afiliación a instituciones colombianas. Dado que Scopus requiere autenticación, se implementó un proceso de automatización con Selenium para simular la navegación, aplicar los filtros correspondientes y exportar los registros en formato CSV.

El corpus resultante comprende los títulos y resúmenes de los artículos de ambas fuentes. Si bien este alcance no incluye el texto completo, los resúmenes constituyen representaciones condensadas del contenido temático y se han empleado en estudios similares de modelado de temas (Asmussen y Møller, 2019).

Preprocesamiento de texto

El preprocesamiento incluyó las siguientes etapas secuenciales:

- **Tokenización.** Segmentación del texto en unidades léxicas individuales.
- **Eliminación de stopwords.** Remoción de palabras funcionales, incluyendo listas genéricas del idioma y términos específicos del dominio añadidos tras inspección manual del corpus (por ejemplo, “university”, “study”, “paper”).
- **Lematización.** Reducción de las palabras a su forma base para consolidar variantes morfológicas.
- **Traducción.** Los textos en español se tradujeron al inglés mediante la herramienta Google Translate API, para unificar el idioma del corpus y permitir la comparabilidad entre fuentes. Se reconoce que la traducción automática puede introducir sesgos léxicos que inciden en la representación de los temas.

Selección del número de temas

El número óptimo de temas se determinó mediante la métrica de perplejidad, que evalúa la capacidad predictiva del modelo sobre datos no observados: valores menores indican un mejor ajuste. Se entrenaron modelos LDA con un rango de dos a seis temas y se seleccionó el valor que minimizó la perplejidad (figura 1). En ambos corpus, el número óptimo resultó ser cuatro temas. Cabe señalar que se utilizó únicamente la perplejidad como criterio de selección; métricas complementarias como la coherencia de temas (Cv o CNPMI) podrían ofrecer una validación adicional y se consideran como trabajo futuro.

La perplejidad se define así:

$$\text{perplejidad} (D_{\text{test}}) = \exp \left\{ - \frac{\sum_{d=1}^M \log p(w_d)}{\sum_{d=1}^M N_d} \right\}$$

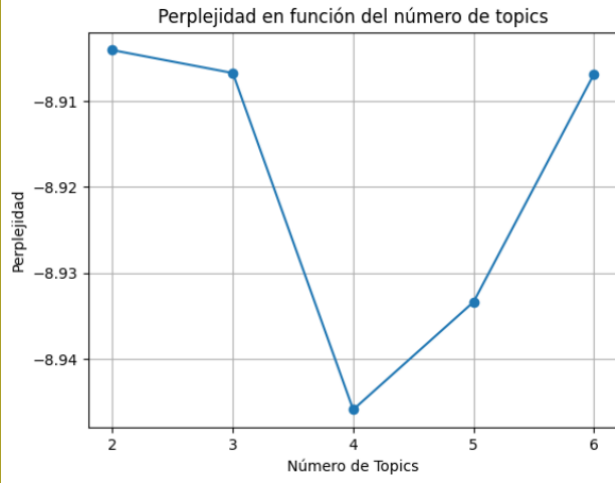


Figura 1. Perplejidad aplicada al corpus de la universidad
Fuente: Elaboración propia.

Entrenamiento del modelo y métricas de comparación

Se entrenaron dos modelos LDA independientes —uno para cada corpus—, utilizando la librería Gensim en Python. Para evaluar la alineación temática, se calculó la similitud coseno entre los vectores de distribución de palabras de los temas de cada modelo. La similitud coseno mide el coseno del ángulo entre dos vectores, tomando valores entre 0 (ninguna similitud) y 1 (similitud total). Esta métrica permite cuantificar el grado de correspondencia temática entre los temas de la universidad y los de Scopus.

La similitud del coseno se define como:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

RESULTADOS Y DISCUSIÓN

Descripción del corpus

En las tablas siguientes se presenta la distribución temporal de los artículos recopilados (tablas 1 y 2).

Tabla 1

Artículos del repositorio de la universidad por año

Año	2020	2021	2022	2023	2024
Artículos	103	227	162	175	27

Tabla 2

Artículos de Scopus por año

Año	2020	2021	2022	2023	2024
Artículos	327	323	277	270	99

Se observa una producción relativamente estable en ambas fuentes entre 2020 y 2023. La disminución en 2024 corresponde a un año con datos parciales al momento de la extracción, por lo que no refleja una tendencia decreciente. Cabe señalar que el corpus de Scopus es aproximadamente el doble del institucional, lo cual se explica por la mayor amplitud de la base indexada al incluir publicaciones de múltiples instituciones colombianas en el área de ingeniería.

Selección del número de temas

Se evaluaron modelos LDA con un rango de dos a seis temas. La Figura 2 muestra la curva de perplejidad para el corpus de la universidad, cuyo valor mínimo se alcanzó en cuatro temas. Este mismo resultado se obtuvo para el corpus de Scopus, por lo que se adoptó $k = 4$ como parámetro para ambos modelos. El rango evaluado fue acotado (2 a 6), debido al tamaño moderado de los corpus; con conjuntos de datos más amplios, un rango mayor podría revelar estructuras temáticas adicionales.

Temas identificados

El modelo LDA permitió identificar cuatro temas en cada corpus, etiquetados a partir de las palabras con mayor peso en cada distribución (figura 3):



Figura 3
Temas
identificados y
etiquetados en
cada corpus

Temas de la universidad

– Tema 0: Terapia y atención médica

Este tema refleja una línea de investigación orientada hacia la salud desde la ingeniería, abarcando áreas como la ingeniería biomédica, el diseño de dispositivos terapéuticos y la optimización de procesos de atención clínica. Su presencia en una institución de ingeniería sugiere un enfoque interdisciplinario que busca aplicar métodos cuantitativos y tecnológicos a la solución de problemas en el sector de la salud, como la rehabilitación, la instrumentación médica y el análisis de datos clínicos.

– Tema 1: Ingeniería de estructuras

Constituye un eje temático tradicional y consolidado dentro de la ingeniería civil. Este tema agrupa investigaciones relacionadas con el análisis, diseño y comportamiento de estructuras, incluyendo materiales de construcción, resistencia sísmica y modelado estructural. Su prominencia es coherente con la trayectoria de la Universidad Escuela Colombiana de Ingeniería, cuya formación en ingeniería civil ha sido uno de sus pilares académicos.

– Tema 2: Tecnología y servicios

Este tema abarca investigaciones vinculadas al desarrollo tecnológico y su aplicación en la prestación de servicios, incluyendo sistemas de información, transformación digital, automatización de procesos y soluciones basadas en *software*. Representa el tema con mayor convergencia frente a Scopus (similitud

coseno del 20,14 %), lo que indica que la universidad está produciendo conocimiento alineado con las tendencias globales en esta área.

– Tema 3: Transporte e infraestructura

Agrupa publicaciones enfocadas en la planificación, diseño y gestión de sistemas de transporte e infraestructura vial. Este tema responde a necesidades específicas del contexto colombiano, donde los desafíos en movilidad urbana, logística y desarrollo de infraestructura son prioritarios. Su presencia diferenciada respecto a Scopus evidencia una producción científica orientada a problemas locales y regionales.

Temas de Scopus

– Tema 0: Materiales y propiedades

Este tema concentra la investigación internacional en ciencia e ingeniería de materiales, incluyendo el estudio de propiedades mecánicas, térmicas y químicas de materiales avanzados, como compuestos, nanomateriales y aleaciones. Refleja una tendencia global hacia la innovación en materiales como motor de avances en múltiples ramas de la ingeniería, desde la manufactura hasta la biomedicina.

– Tema 1: Educación y aprendizaje

Representa una línea temática centrada en la investigación sobre pedagogía en ingeniería, metodologías de enseñanza-aprendizaje, uso de tecnologías educa-

tivas y evaluación de competencias. Su presencia en Scopus evidencia un creciente interés de la comunidad científica internacional por mejorar la formación de ingenieros, un área en la que la universidad podría fortalecer su producción investigativa.

– Tema 2: Tecnología y aplicaciones

Agrupar publicaciones sobre el desarrollo y la implementación de soluciones tecnológicas en diversos dominios, incluyendo inteligencia artificial, internet de las cosas (IoT), computación en la nube y sistemas inteligentes. Este tema es el que presenta mayor similitud con el tema 2 de la universidad (“Tecnología y servicios”), confirmando que existe un espacio de convergencia temática entre la producción institucional y las tendencias internacionales.

– Tema 3: Energía y procesos

Este tema refleja la relevancia global de la investigación en sistemas energéticos, eficiencia de procesos industriales, energías renovables y sostenibilidad. Su presencia dominante en Scopus y su baja correspondencia con los temas de la universidad representa una oportunidad para que la institución explore y desarrolle líneas de investigación en transición energética y procesos sostenibles.

Los temas de la universidad reflejan un perfil investigativo con énfasis en ingeniería civil e infraestructura, complementado por un componente interdisciplinario en salud (tema 0) que probablemente corresponde a trabajos de grado con aplicaciones en bioingeniería o gestión hospitalaria. Los temas de Scopus, en cambio, muestran un perfil orientado hacia ciencia de materiales, procesos energéticos y educación en ingeniería, áreas que concentran una proporción significativa de la producción científica indexada en el ámbito nacional.

Análisis de similitud

A renglón seguido se presenta la matriz de similitud coseno entre los temas de ambos modelos. Los valores oscilan entre 0 (sin similitud) y 1 (similitud total), expresados en porcentaje (tabla 3).

Tabla 3

Similitud coseno entre temas LDA de la universidad y Scopus

	Scopus T0	Scopus T1	Scopus T2	Scopus 3
Univ. T0	19,23 %	14,92 %	18,15 %	18,74 %
Univ.T1	11,65 %	12,27 %	15,80 %	13,25 %
Univ. T2	15,77 %	13,89 %	20,14 %	15,73 %
Univ. T3	15,81 %	15,96 %	17,55 %	17,09 %

La mayor similitud se observa entre el tema 2 de la universidad (“Tecnología y servicios”) y el tema 2 de Scopus (“Tecnología y aplicaciones”), con un 20,14 %. Esto indica que el área de tecnología es aquella en la que la producción institucional tiene mayor correspondencia con las tendencias nacionales indexadas. Le sigue la correspondencia entre los temas 0 (19,23 %), que si bien comparten vocabulario, representan temas distintos (salud vs. materiales), lo que sugiere una coincidencia léxica más que temática.

En contraste, el tema 1 de la universidad (“Ingeniería de estructuras”) presenta los valores más bajos de similitud con todos los temas de Scopus (11,65 % – 15,80 %), lo que indica que esta línea de investigación institucional tiene menor representación en la bibliografía indexada nacional. Esto no implica necesariamente una debilidad, sino que puede reflejar un nicho de especialización institucional con menor presencia en Scopus.

Implicaciones para la agenda investigativa

Estos resultados permiten identificar tres escenarios para la planificación investigativa de la universidad:

- Áreas de fortaleza alineada.** Tecnología y aplicaciones, donde la universidad ya muestra convergencia con Scopus y podría potenciar su visibilidad internacional.
- Áreas de oportunidad.** Ciencia de materiales, procesos energéticos y educación en ingeniería, donde Scopus muestra actividad significativa pero la producción institucional es limitada.
- Áreas de especialización local.** Ingeniería de estructuras y transporte, donde la universidad tiene un perfil propio que responde a prioridades del contexto colombiano, aunque con menor presencia en la bibliografía indexada.

CONCLUSIONES

En este estudio se aplicó el modelo LDA para evaluar cuantitativamente la alineación temática entre la producción científica de la Universidad Escuela Colombiana de Ingeniería Julio Garavito y las publicaciones indexadas en Scopus en el campo de la ingeniería. Se identificaron cuatro temas latentes en cada corpus, con una convergencia parcial más marcada en el área de tecnología (similitud coseno del 20,14 %) y divergencias en áreas como ingeniería de estructuras y transporte.

Los hallazgos permiten concluir que la universidad presenta un perfil investigativo con fortalezas en tecnología y un componente interdisciplinario en bioingeniería y gestión hospitalaria, mientras que áreas como ciencia de materiales, energía y educación en ingeniería representan oportunidades de crecimiento en línea con las tendencias de la producción científica nacional. Estos resultados constituyen una línea base empírica que puede servir como insumo para la planificación estratégica de las líneas de investigación institucional.

En el estudio se presentan limitaciones que se deben considerar: a) el corpus se limitó a títulos y resúmenes, lo que podría no captar la totalidad del contenido temático; b) la traducción automática mediante Google Translate API puede introducir sesgos léxicos; c) el modelo LDA opera con supuestos de bolsa de palabras e independencia entre documentos, y d) se utilizó únicamente la perplejidad como criterio de selección del número de temas.

Como líneas de trabajo futuro, se recomienda explorar modelos complementarios como BERT o NMF que capten relaciones semánticas más complejas, ampliar las fuentes de comparación, incluyendo bases como Web of Science, incorporar análisis temporales de la evolución de temas y extender el corpus al texto completo de las publicaciones.

REFERENCIAS

Ambalavanan, A. K., y Devarakonda, M. V. (2020). Using the contextual language model BERT for multi-criteria classification of scientific articles. *Journal of Biomedical Informatics*, 112, 103578. <https://doi.org/10.1016/j.jbi.2020.103578>

Aria, M., y Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959-975. <https://doi.org/10.1016/j.joi.2017.08.007>

Assmussen, C. B., y Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature

review. *Journal of Big Data*, 6(1), 1-18. <https://doi.org/10.1186/s40537-019-0255-7>

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

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., y Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407. [https://doi.org/10.1002\(SICI\)1097-4571\(199009\)41:6<391::AID-AS1>3.0.CO;2-9](https://doi.org/10.1002(SICI)1097-4571(199009)41:6<391::AID-AS1>3.0.CO;2-9)

DiMaggio, P., Nag, M., y Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), 570-606. <https://doi.org/10.1016/j.poetic.2013.08.004>

Egger, R., Kroner, M., y Stöckl, A. (2022). Web scraping. En R. Egger (ed.), *Applied data science in tourism* (pp. 67-82). Springer. https://doi.org/10.1007/978-3-030-88389-8_5

González, J. L., Castro, A. R., Mesa, M. L., y Maya, C. J. (2020). Producción científica en América Latina y el Caribe en el periodo 1996-2019. *Revista Cubana de Medicina Militar*, 49(4).

Grisales Aguirre, C. J., y Figueroa Vallejo, A. M. (2022). Modelado de temas aplicado al análisis del papel del aprendizaje automático en revisiones sistemáticas. *Revista de Investigación, Desarrollo e Innovación*, 12(2).

Hofmann, T. (1999). Probabilistic latent semantic indexing. *En Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 50-57). ACM. <https://doi.org/10.1145/312624.312649>

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., y Zhao, L. (2017). Latent Dirichlet Allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78, 15169-15211. <https://doi.org/10.1007/s11042-018-6894-4>

Kozłowski, D., Semeshenko, V., y Molinari, A. (2021). Latent Dirichlet Allocation model for world trade analysis. *PLOS ONE*, 16(2), e0245393. <https://doi.org/10.1371/journal.pone.0245393>

Marroquín Arreola, J., y Ríos Bolívar, H. (2012). Inversión en investigación y crecimiento económico: un análisis empírico desde la perspectiva de los modelos de I+D. *Investigación Económica*, LXXI(282), 15-33.

Nigam, H., y Biswas, P. (2021). Web scraping: from tools to related legislation and implementation using Python. En J. S. Raj, A. M. Ilyasu, R. Bestak y Z. A. Baig (eds.), *Innovative Data Communication Technologies and Application* (pp. 149-164). Springer. https://doi.org/10.1007/978-981-15-9651-3_14

Oyewole, A. T., Adeoye, O. B., Addy, W. A., Okoye, C. C., Ofodile, O. C., y Ugochukwu, C. E. (2024). Automating financial reporting with natural language processing: a review and case analysis. *World Journal of Advanced Research and Reviews*, 21(03), 575-589. <https://doi.org/10.30574/wjarr.2024.21.3.0700>

Rubiano, M. A. (2022). *Análisis de temas utilizando Twitter: una aplicación del modelo LDA al caso colombiano* [trabajo de grado]. Universidad Santo Tomás.

Universidad Javeriana (2022). *Inversión en Investigación y Desarrollo (I+D): evidencia a través de las patentes*. Laboratorio de Economía de la Educación.

Wasserman, M. (2000). *Sobre la importancia de investigar en Colombia, un país subdesarrollado*. Diálogos Globales de la Exposición.