Pontificia Universidad
JAVERIANA
Cali

REDES NEURONALES Y PROCESAMIENTO DE LENGUAJE NATURAL PARA LA EVALUACIÓN DE LA INVESTIGACIÓN COLOMBIANA EN EL CONTEXTO DE LOS ODS

John Agustin Riaño Diaz

Proyecto Aplicado para optar al título de Magíster en Ciencia de Datos

Director Carlos Ernesto Ramírez Ovalle

Codirector Abel Álvarez Bustos

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI,
JULIO DE 2025

1. CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

INTRODUCCIÓN

La producción científica es clave para el desarrollo académico y tecnológico de las instituciones de educación superior y evaluar su impacto en la sociedad es esencial para integrar las políticas científicas hacia las tendencias investigativas y particularmente con los problemas focalizados en cada sociedad. Estos objetivos buscan desarrollar estrategias que proyecten soluciones a la erradicación de la pobreza, proteger el planeta y asegurar una vida digna para todos los habitantes del planeta. Sin embargo, en Colombia la relación entre la producción científica y los ODS no ha sido claramente evidenciada, lo que dificulta conocer el capital intelectual de las investigaciones y su impacto en la resolución de problemas generales que respondan a lo establecido en la agenda 2030.

El problema específico que aborda esta investigación es la falta de un modelo comparativo y sistemático para analizar cómo la producción investigativa en Colombia se alinea con los ODS. En este momento, las técnicas convencionales se centran en aspectos cuantitativos de la producción científica como conteo de publicaciones o citas exclusivamente, sin proporcionar un análisis integrado que determine su relación temática con los ODS específicos. La ausencia de un enfoque claro limita la capacidad de identificar investigaciones que contribuyen de manera significativa a los objetivos globales. Entender esta problemática es crucial para maximizar el impacto social, académico, financiero y político de la investigación científica en Colombia. Un análisis sistemático de la alineación de las investigaciones con los ODS permitirá identificar brechas y tendencias en el cumplimiento de cada uno de los 17 ODS, facilitando la toma de decisiones la formulación de políticas que promuevan el desarrollo sostenible.

Diversos estudios han explorado la relación entre la producción científica y los ODS a nivel global. Sin embargo, pocos se han centrado específicamente en Colombia, y tampoco han aplicado técnicas como el procesamiento de lenguaje natural (PLN) y redes neuronales para este propósito. Este vacío en la literatura destaca la necesidad de un análisis detallado que pueda proporcionar una visión más completa de cómo la investigación en Colombia contribuye a los ODS.

Este estudio responde a esta necesidad investigativa través de la implementación de técnicas de PNL que permitan un análisis integral de la producción científica del país.

Por lo cual se implementa un enfoque avanzado de procesamiento de lenguaje natural (PLN), utilizando un modelo de lenguaje preentrenado como BERT, para automatizar la clasificación de resúmenes de investigaciones colombianas en Ingeniería y Medicina según su contribución a los ODS. Este enfoque comienza con un proceso de etiquetado manual para la clasificación de una muestra representativa de resúmenes.

Esta muestra etiquetada es utilizada como base de entrenamiento para los modelos de redes neuronales de aprendizaje supervisado, permitiendo validar algoritmos de aprendizaje profundo, como BERT, LSTM o RNM, de acuerdo con su pertinencia de aplicación en la investigación y su alineación temática con cada uno de los ODS.

El modelo de PLN entrenado es capaz de generalizar este conocimiento y aplicar el etiquetado automático a la totalidad de los resúmenes de artículos científicos restantes de ingeniería y medicina. Este proceso no solo permite la clasificación precisa de grandes volúmenes de datos, sino que también facilita la articulación de la relación entre cada resumen y los ODS correspondientes. A través de esta propuesta, se realiza un análisis sistemático de la producción científica desde áreas tan relevantes como la ingeniería en el desarrollo tecnológico e innovación y también la medicina como mecanismo para mitigar diferentes problemáticas en lo social.

2. DEFINICIÓN DEL PROBLEMA

2.1. Planteamiento del problema

La producción científica ha sido una métrica comúnmente utilizada para analizar el comportamiento investigativo a nivel de autores, grupos de investigación, instituciones y países, permitiendo caracterizar enfoques temáticos o líneas de investigación relevantes [1]. En este contexto, es notable el crecimiento de la producción científica en Colombia en los últimos años, según lo evidencian los análisis realizados por el Ministerio de Ciencia y Tecnología (MINCIENCIAS) y el Observatorio de Ciencia y Tecnología (OCyT) en su publicación "Indicadores de Ciencia, Tecnología e Innovación - Colombia 2021" [2]. Este informe refleja el compromiso del país con la generación de conocimiento y el desarrollo académico, resultando en un incremento de la producción científica.

Este informe destaca cómo la gobernanza científica se está integrando con los Objetivos de Desarrollo Sostenible (ODS) [3]. En los últimos años, la ciencia, la tecnología y la innovación se han convertido en pilares fundamentales para las economías más competitivas. Como resultado, el gasto en investigación ha aumentado globalmente, pasando del 1,73% del PIB en 2014 al 1,79% en 2018 [2]. Sin embargo, es preocupante que cerca del 80% de los países sigan invirtiendo menos del 1% de su PIB en investigación. Sin embargo, la relación entre esta producción investigativa y su contribución a los Objetivos de Desarrollo Sostenible (ODS) de la ONU no ha sido claramente definida, en la actualidad no existen mecanismos estandarizados ni estudios sistemáticos que puedan evidenciar el aporte de la producción científica colombiana a las problemáticas planteas en los ODS.

Es relevante destacar que los ODS establecidos en el año 2015 constituyen un llamado global a la acción para erradicar la pobreza, proteger el planeta y garantizar que todas las personas tengan acceso a una calidad de vida digna para el año 2030 [4]. Sin embargo, la falta de información sobre cómo las investigaciones académicas y científicas en Colombia se alinean con los ODS pone de manifiesto la necesidad de identificar los temas transversales que vinculan el ejercicio investigativo

con la Agenda 2030. Esta situación ha impedido establecer el capital investigativo referente en Colombia en el marco de los ODS, dificultando la articulación y creación de políticas públicas orientadas a áreas de investigación con potencial para contribuir al cumplimiento de estos objetivos globales. A pesar de los esfuerzos individuales de algunos grupos de investigación por vincular sus propuestas con los ODS, no existe un modelo estandarizado que permita evaluar de manera integral el aporte de toda la producción científica del país a estos objetivos [1] [2].

A su vez, la falta de herramientas analíticas sofisticadas ha dificultado la identificación de tendencias y brechas en la investigación científica nacional con respecto a temáticas relevantes. Es importante destacar que han existido esfuerzos en torno a la generación de estrategias para consolidar el capital investigativo en Colombia, como el denominado "Atlas del Conocimiento", que presenta un análisis sobre el número de productos asociados a cada ODS [3], en contraste con las instituciones generadoras. Sin embargo, actualmente no existe un análisis detallado de la literatura científica que permita establecer las temáticas con mayor incidencia y su caracterización en torno a los ODS. Las propuestas existentes han establecido el relacionamiento con los objetivos únicamente desde las áreas del conocimiento y no sobre los temas con mayor relación [2] [4].

El problema central de esta investigación radica en la ausencia de un modelo sistemático para analizar la contribución de la producción investigativa en Colombia a los Objetivos de Desarrollo Sostenible (ODS). Aunque ha existido una gran cantidad de investigaciones en el país que de forma indirecta intentan responder a las problemáticas evidenciadas en la Agenda 2030, se carece de una comprensión clara de su impacto en los ODS. Las técnicas convencionales se han centrado únicamente en el aspecto cuantitativo de la producción científica, sin proporcionar un enfoque integrado que permita determinar su relación temática con los problemas planteados en la agenda 2030, lo que evidencia la pertinencia de una propuesta que integre la ciencia de datos en el análisis detallado sobre la producción científica en Colombia. Aunque algunas propuestas actuales intentan establecer esta relación mediante palabras clave textuales relacionadas con los ODS, son pocas las investigaciones que mencionan explícitamente el nombre del objetivo de desarrollo sostenible en el texto [4].

Esto destaca la necesidad de aplicar técnicas avanzadas de la ciencia de datos como el procesamiento del lenguaje natural (PLN) y las redes neuronales para abordar esta problemática de manera efectiva. La implementación de estas técnicas permitirá el análisis de la producción científica y su integración con los ODS detallando como la investigación colombiana se articula como los problemas propuestos en cada ODS, lo que facilita la toma de decisiones estratégicas y así la optimización de recursos en el ámbito investigativo por parte de entidades públicas o privadas.

Es determinante precisar que la falta de un modelo sistemático para analizar la contribución de la producción investigativa en Colombia a los Objetivos de Desarrollo Sostenible (ODS) ha tenido varias repercusiones: la desconexión entre la investigación académica y las necesidades globales de desarrollo sostenible, la ineficiencia en la asignación de recursos y la formulación de políticas públicas desinformadas. Las posibles causas de esta situación han incluido la dependencia de métodos tradicionales de análisis que no consideran la relación temática con los ODS y la ausencia de técnicas avanzadas en el análisis de datos científicos.

Si esta situación persiste, Colombia podría enfrentar dificultades significativas para cumplir con lo planteado en la Agenda 2030 de las Naciones Unidas, lo que afectaría negativamente su desarrollo sostenible y su competitividad internacional. La incapacidad de alinear la producción científica con los ODS también podría resultar en una menor visibilidad y reconocimiento en rankings internacionales como el "Sustainable Development Goals Index" o el "Times Higher Education Impact Rankings", que han medido el progreso hacia los ODS desde diferentes aspectos. Estos rankings, aunque útiles, no han considerado en profundidad la alineación directa de la producción científica con los ODS, lo que resalta aún más la necesidad de un enfoque analítico que evalúe específicamente la contribución de la investigación académica a estos objetivos globales.

2.2. Formulación del problema

Para abordar este problema, la investigación se centró en responder la siguiente pregunta principal: ¿Cómo se puede utilizar el procesamiento del lenguaje natural (PLN) y las redes neuronales para analizar la producción investigativa en Colombia y su relación con los Objetivos de Desarrollo

Sostenible (ODS)?

3. OBJETIVOS

3.1. Objetivo general

Desarrollar un modelo avanzado de procesamiento de lenguaje natural (PLN) para el análisis y clasificación de resúmenes de investigaciones científicas colombianas en Ingeniería y Medicina en función de su contribución a los 17 Objetivos de Desarrollo Sostenible (ODS).

3.2. Objetivos específicos

3.2.1. Objetivo específico 1

Etiquetar manualmente una muestra representativa de resúmenes de investigaciones colombianas en ingeniería y medicina según su alineación con los ODS.

3.2.2. Objetivo específico 2

Entrenar modelos de redes neuronales utilizando las muestras etiquetadas para optimizar la transferencia de conocimiento en el etiquetado automático de los resúmenes restantes.

3.2.3. Objetivo específico 3

Aplicar el modelo de procesamiento de lenguaje natural entrenado para clasificar automáticamente todos los resúmenes de artículos científicos en Ingeniería y Medicina en Colombia, identificando su relación con los ODS correspondientes.

3.2.4. Objetivo específico 4

Determinar la similitud entre los temas investigativos en Ingeniería y Medicina en contraste con

los ODS, utilizando métricas apropiadas para el análisis comparativo de documentos.

3.2.5. Objetivo específico 5

Evaluar la eficacia del modelo desarrollado para la identificación y alineación de la investigación con los ODS, utilizando métricas de desempeño.

3.3. Fases del proyecto

El desarrollo de este estudio se articuló en ocho fases metodológicas que integran técnicas de recuperación de información, procesamiento de lenguaje natural, aprendizaje automático y análisis semántico. Cada etapa fue diseñada para responder a los objetivos específicos del proyecto y garantizar la robustez del modelo propuesto para la clasificación temática de la producción científica nacional según su alineación con los Objetivos de Desarrollo Sostenible (ODS).

3.3.1. Recolección del corpus

La fase inicial consistió en la conformación del corpus de estudio a partir de la descarga de 23.229 resúmenes de artículos científicos indexados en Scopus. Se seleccionaron únicamente documentos de acceso abierto, publicados entre los años 2014 y 2018, en las áreas de Ingeniería y Medicina. Esta selección respondió al interés de delimitar el análisis a campos científicos estratégicos y garantizar condiciones de acceso y reproducibilidad. La base de datos Scopus fue elegida por su cobertura multidisciplinaria y la calidad de sus metadatos, elementos indispensables para una recuperación confiable y exhaustiva.

3.3.2. Muestreo estratificado

Con el fin de facilitar el entrenamiento supervisado de los modelos de clasificación, se aplicó un muestreo estratificado centrado en el año 2018. Esta estrategia permitió seleccionar una muestra proporcionalmente representativa respecto a las dos áreas disciplinares incluidas en el estudio. A diferencia del muestreo aleatorio simple, la estratificación reduce la varianza dentro de cada subgrupo, lo que favorece una mejor generalización del modelo al momento de extender la clasificación al conjunto completo de datos.

3.3.3. Etiquetado manual

La muestra obtenida fue sometida a un proceso de etiquetado temático manual, en el que se asignaron categorías correspondientes a uno o más de los 17 ODS. Para ello, se definieron criterios de etiquetado con base en las metas e indicadores de la Agenda 2030 y se aplicó una lectura experta que permitió evaluar la pertinencia temática de cada resumen. Se etiquetaron 312 documentos, lo que constituyó una base sólida para entrenar modelos supervisados con un nivel aceptable de heterogeneidad semántica.

3.3.4. Preprocesamiento y balanceo

El corpus etiquetado fue sometido a una serie de procedimientos de limpieza textual, incluyendo la eliminación de caracteres especiales, normalización de mayúsculas, y depuración de elementos no informativos. Asimismo, se verificó que todos los textos estuvieran en inglés, idioma en el cual se entrenaron los modelos seleccionados. En aquellas clases con baja representación, se implementaron estrategias de balanceo basadas en traducción automática y parafraseo semántico. Estas técnicas permitieron ampliar artificialmente el número de muestras sin comprometer la coherencia temática, mejorando así la capacidad del modelo para aprender patrones representativos en categorías minoritarias.

3.3.5. Entrenamiento de modelos

A partir del corpus procesado, se entrenaron modelos de clasificación utilizando enfoques supervisados y no supervisados. En primer lugar, se utilizó BERT, modelo basado en la arquitectura Transformer, conocido por su capacidad para capturar relaciones contextuales complejas. No obstante, al evidenciarse un bajo rendimiento en ciertos ODS con menor número de ejemplos, se incorporó un enfoque complementario utilizando YAKE, un extractor automático de palabras clave que permitió representar los textos mediante términos semánticamente relevantes. Esta estrategia híbrida contribuyó a mejorar la cobertura y precisión del sistema de clasificación, especialmente en contextos de desequilibrio de clases.

3.3.6. Clasificación automática

Los modelos entrenados fueron aplicados sobre la totalidad del corpus con el objetivo de predecir la alineación temática de los 23.229 resúmenes con los ODS. Se empleó un esquema diferenciado de clasificación: mientras BERT se utilizó para ODS con mayor volumen de entrenamiento, YAKE se implementó para aquellos con menor representatividad. Esta aproximación flexible permitió maximizar la eficiencia del sistema sin sacrificar la precisión en clases específicas.

3.3.7. Evaluación de resultados

Para validar la eficacia de los modelos se recurrió a técnicas de validación cruzada y a la estimación de métricas como precisión, recall, F1-score, exactitud y coeficiente de Kappa. Estas métricas permitieron no solo evaluar el desempeño de cada modelo de forma aislada, sino también realizar un análisis comparativo que reveló ventajas y limitaciones según el tipo de enfoque y la categoría ODS evaluada. Este ejercicio de validación fue crucial para seleccionar los modelos más adecuados de acuerdo con los objetivos del proyecto y las características del corpus.

3.3.8. Análisis semántico

La fase final se enfocó en el análisis semántico de las clasificaciones obtenidas. A través del uso de técnicas de agrupamiento como K-Means y del análisis de similitud por coocurrencia de términos extraídos, se exploraron patrones temáticos latentes entre los ODS y las áreas de conocimiento. La visualización mediante mapas de calor y redes de términos permitió identificar relaciones entre tópicos, áreas disciplinares y objetivos globales, abriendo la posibilidad de interpretar la producción científica no solo como un conjunto de documentos aislados, sino como una red temática articulada que refleja prioridades, vacíos y oportunidades del sistema científico nacional frente a los desafíos del desarrollo sostenible.

Figura 1. Proceso metodológico PNL - ODS

4. MARCO TEÓRICO Y ANTECEDENTES

4.1. Marco teórico

Para abordar la problemática anteriormente descrita, resulta indispensable comprender los fundamentos teóricos que establecen la base conceptual de la presente investigación. A continuación, se plantea un análisis histórico y conceptual desde el aporte que tienen las bases teóricas para la investigación planteada en el análisis de los Objetivos de Desarrollo Sostenible

4.1.1. Procesamiento de Lenguaje Natural (PLN)

Uno de los conceptos fundamentales de esta investigación es el de Procesamiento de Lenguaje Natural, el cual es definido como una subdisciplina de la inteligencia artificial que se enfoca en la interacción entre las computadoras y el lenguaje humano. Los orígenes del PLN pueden remontarse hasta la década de 1950, con los primeros intentos de traducir automáticamente textos entre diferentes idiomas. Un hito en la materia en esta era temprana fue el experimento de Georgetown-IBM en el año 1954, en cual, mostró la viabilidad de la traducción automática primera en su historia y aunque de manera limitada, sería un hito para empezar el desarrollo posterior de esta tecnología [5].

A medida que el Procesamiento de Lenguaje Natural avanzaba, los enfoques fundamentados en reglas eran insuficientes para capturar la complejidad del lenguaje humano, lo que produjo la aplicación de métodos estadísticos en la década de 1980 fomentados por el aumento de la capacidad computacional y la disponibilidad de grandes conjuntos de datos textuales. Técnicas como el análisis de frecuencia de términos (TF-IDF) y los modelos de lenguaje n-grama se convirtieron en herramientas utilizadas para el procesamiento de texto.

Ahora bien, la integración de redes neuronales y el aprendizaje profundo revolucionó el PLN en el año 2010. Modelos como Word2Vec, desarrollado por el científico checo Mikolov et al. en el año 2013, permitieron la representación vectorial de palabras, capturando relaciones semánticas complejas de manera eficiente. Consecutivamente, modelos más desarrollados como BERT (Bidirectional Encoder Representations from Transformers) [6] y GPT (Generative Pre-trained

Transformer) [7] han demostrado un rendimiento sobresaliente en diversas tareas de PLN, como la comprensión de lectura, la traducción automática y la generación de texto. Estos modelos utilizan arquitecturas, introducidas por el científico indio Vaswani y su equipo científico en 2017 [8], que han permitido un procesamiento más eficiente.

Actualmente, el PLN sigue siendo un campo dinámico y en constante desarrollo con aplicaciones en múltiples áreas desde la banca, pasando por la salud hasta llegar a todo tipo de industrias, y a su vez, han generado soluciones en la asistencia virtual, los sistemas de recomendación, el análisis de sentimientos y la detección de fraudes. La investigación continúa enfocándose en mejorar la precisión y eficiencia de los modelos [9].

4.1.2. Modelo Transformer

La arquitectura Transformer ha supuesto un punto de inflexión en el campo del procesamiento de lenguaje natural y, más ampliamente, en el aprendizaje profundo. Su impacto no solo reside en la mejora empírica de las métricas de desempeño, sino también en la transformación conceptual de cómo se modelan las secuencias lingüísticas. El modelo fue propuesto por Vaswani et al. en 2017 como una respuesta a las limitaciones estructurales de las redes recurrentes, especialmente en tareas que requerían procesar secuencias largas con dependencias de largo alcance. A diferencia de las RNN y LSTM, que procesan los datos de manera secuencial, el Transformer emplea un mecanismo de atención auto-regresiva que permite procesar todos los tokens de entrada en paralelo, lo que mejora drásticamente la eficiencia y capacidad de escalabilidad del modelo [8].

El núcleo teórico del Transformer radica en la atención escalada de múltiples cabezas (multi-head self-attention), la cual permite al modelo aprender diferentes representaciones contextuales de una misma palabra según el entorno lingüístico en que se encuentra. Esta operación se implementa a través de matrices de consulta, clave y valor, que determinan la relevancia relativa entre pares de tokens dentro de la secuencia. A nivel práctico, esta propiedad ha demostrado ser superior al modelado secuencial tradicional, especialmente en tareas que exigen una alta sensibilidad al contexto semántico, como la clasificación temática, la generación automática de lenguaje o la inferencia textual [10].

A partir de esta arquitectura fundamental se han derivado múltiples variantes que han extendido su aplicabilidad a diferentes dominios. Entre ellas, BERT (Bidirectional Encoder Representations from Transformers) se ha consolidado como uno de los modelos más influyentes al entrenar representaciones de lenguaje de manera bidireccional, considerando simultáneamente los contextos izquierdo y derecho de cada token. BERT se entrena mediante dos objetivos: el enmascaramiento aleatorio de tokens (masked language modeling) y la predicción de la siguiente oración (next sentence prediction), lo cual le permite captar relaciones sintácticas y semánticas profundas [11].

En el contexto del presente proyecto, la adopción de un modelo basado en Transformer no se justifica exclusivamente por su prominencia en la literatura, sino por su idoneidad técnica frente a los desafíos del corpus utilizado. Los resúmenes científicos extraídos de Scopus, correspondientes a las áreas de Ingeniería y Medicina, presentan un alto grado de tecnicismo, ambigüedad y variabilidad temática. La capacidad del Transformer para manejar este tipo de complejidades lingüísticas es particularmente valiosa cuando se requiere clasificar textos según su alineación con marcos normativos amplios y conceptualmente abstractos como los Objetivos de Desarrollo Sostenible (ODS). El modelo bert-base-uncased, preentrenado sobre grandes corpus de texto y posteriormente ajustado (fine-tuned) sobre la muestra etiquetada manualmente, permitió capturar patrones de alineación semántica entre los resúmenes y las categorías de los ODS. La capacidad de generalización del modelo se potenció mediante técnicas de regularización como dropout, weight decay, y validación cruzada estratificada. Estas decisiones metodológicas no solo optimizaron el aprendizaje del modelo, sino que también mitigaron el sobreajuste, un riesgo frecuente en entornos con desequilibrio de clases y escasez de datos etiquetados [12].

Desde una perspectiva más amplia, el uso de Transformers en este tipo de tareas contribuye a consolidar una línea metodológica que integra modelos lingüísticos de última generación con problemas de política científica. Esto no solo mejora la precisión del análisis automatizado, sino que también permite abordar preguntas estructurales sobre el alineamiento temático de la ciencia nacional con los marcos internacionales de desarrollo sostenible. En este sentido, la arquitectura Transformer deja de ser un mero instrumento técnico para convertirse en una herramienta epistemológica que intermedia entre el lenguaje científico y los objetivos normativos de la sociedad contemporánea.

4.1.3. Análisis de Similitud

El análisis de similitud, particularmente de temáticas es una técnica utilizada para medir la semejanza entre diferentes textos o documentos basándose en las temáticas que abordan. Esta técnica es fundamental en diversas aplicaciones del procesamiento de lenguaje natural, como la recuperación de información, la agrupación de documentos y la recomendación de contenido.

En sus inicios, el análisis de similitud temática estaba basado en métodos simples de relacionamiento de palabras, como el análisis de frecuencia de términos (TF) y la frecuencia inversa de documentos (IDF), que contribuían a identificar la relevancia de términos específicos en un corpus especifico. Sin embargo, estos métodos eran limitados, puesto que, su impedimento era la captura de la complejidad semántica del lenguaje. Hasta que el desarrollo de modelos de tópicos, como Latent Dirichlet Allocation (LDA), introducido por Blei, Ng y Jordan en el año 2003, marcan un avance significativo en este campo. LDA es un modelo generativo que toma cada documento como una composición de un número limitado de temas y que cada palabra en el documento está asociada a uno de estos temas. Este enfoque permite una representación real y organizada del contenido temático de los documentos [10].

Posteriormente, con el auge del aprendizaje profundo, el análisis de similitud temática ha incorporado técnicas más avanzadas como Word2Vec, Doc2Vec y BERT. Estos modelos representan palabras en un espacio vectorial permanente, evidenciando relaciones semánticas mediante el entrenamiento en grandes corpus textuales [13]. Doc2Vec, una extensión de Word2Vec, genera la representación de documentos completos en lugar de palabras individuales [14]. Continuamente, modelos como BERT han mejorado aún más la capacidad de capturar similitudes temáticas al considerar el contexto bidireccional de las palabras en un texto [15]. Esto significa que el modelo puede determinar el significado de una palabra basándose en todas las palabras que la rodean, lo que resulta en una representación semántica más precisa [15].

4.1.4. Redes Neuronales

Modelos computacionales basados en el funcionamiento del cerebro humano, diseñados para reconocer patrones complejos y aprender de los datos son uno de los hitos más relevantes de la ciencia de datos. La historia de las redes neuronales comenzó en la década de 1940 con el trabajo pionero del estadounidense Warren McCulloch y Walter Pitts, quienes desarrollaron el primer modelo matemático de una neurona artificial [16]. Este modelo fue la base para el desarrollo posterior de redes neuronales más complejas. En la década de 1950 y 1960, es generado el primer algoritmo para el aprendizaje supervisado de clasificadores binarios y que también es catalogado una neurona artificial. Este hito es introducido por Frank Rosenblatt. Sin embargo, las deficiencias de los perceptrones multicapa, no permitieron el avance de la investigación sobre redes neuronales durante la década de 1970 [17].

El campo de las redes neuronales resurgió en la década de 1980 con el desarrollo del algoritmo de retro propagación por Geoffrey Hinton, David Rumelhart y Ronald Williams. Este diseño permitió el entrenamiento de redes neuronales multicapa, lo que posibilitó nuevas opciones para el aprendizaje profundo. Geoffrey Hinton, junto con Yann LeCun y Yoshua Bengio, son ampliamente reconocidos como los pioneros del aprendizaje profundo, sus aportes han sido fundamentales para el desarrollo de arquitecturas avanzadas como las redes neuronales convolucionales (CNN) y las redes neuronales recurrentes (RNN) [18].

En los últimos años, las redes neuronales han avanzado hacia arquitecturas aún más complejas, como los transformadores y las redes neuronales generativas adversarias (GAN). Los transformadores, introducidos por Vaswani et al. en 2017, han demostrado ser particularmente efectivos para el procesamiento de secuencias, como en tareas de PLN y traducción automática [8]. Las GAN, desarrolladas por Ian Goodfellow et al. en 2014, han revolucionado la generación de datos sintéticos, permitiendo la creación de imágenes, videos y sonidos realistas [19]. Actualmente, las redes neuronales son parte fundamental de muchas aplicaciones de la Inteligencia Artificial, desde la visión por computadora y el reconocimiento de voz hasta la generación de texto y la mejora de imágenes. [20].

4.2. Extracción de palabras clave mediante YAKE

La extracción automática de palabras clave representa una técnica esencial dentro del procesamiento de lenguaje natural, especialmente cuando se busca reducir la dimensionalidad semántica de un texto sin perder la capacidad de representar su contenido informativo. En este sentido, YAKE (Yet Another Keyword Extractor) constituye una de las propuestas metodológicas más sólidas y eficientes para la identificación de términos significativos a partir de documentos individuales. Lejos de depender de corpus externos, aprendizaje supervisado o redes neuronales, YAKE se fundamenta en un enfoque estrictamente estadístico que explora atributos locales del documento, como la frecuencia de términos, la posición relativa en el texto, la capitalización, la coaparición y la distribución de las palabras en las distintas secciones del documento [21].

Desde una perspectiva teórica, YAKE supera algunas de las limitaciones tradicionales asociadas a modelos clásicos como TF-IDF y TextRank. Mientras que el primero presenta una dependencia excesiva del tamaño del corpus y el segundo impone una estructura de grafo que a menudo distorsiona la jerarquía temática real, YAKE introduce una función de puntuación más sensible a la semántica local del texto. Esta aproximación resulta especialmente adecuada en dominios donde los textos presentan alta especificidad temática, lenguaje técnico denso y ausencia de vocabulario común, como es el caso de los resúmenes científicos utilizados en esta investigación. La independencia de corpus de referencia no solo amplía la aplicabilidad del modelo a contextos de escasez de datos, sino que garantiza su adaptabilidad a distintos idiomas y estilos discursivos sin necesidad de reentrenamiento [22].

En el desarrollo de este proyecto, YAKE fue incorporado como etapa crítica de preprocesamiento en el pipeline de clasificación temática de artículos académicos según su alineación con los Objetivos de Desarrollo Sostenible (ODS). La extracción automática de palabras clave permitió construir vectores sintéticos de características semánticas, facilitando su posterior representación mediante embeddings de alta dimensionalidad y la aplicación de algoritmos de similitud como la distancia coseno. Esta reducción semántica fue clave para estabilizar los procesos de clustering no supervisado y para mitigar los efectos del sobreajuste en categorías con baja representación

muestral, como los ODS 6 y 13. Al limitar la representación textual a los términos más representativos, se evitó la contaminación del modelo con ruido léxico irrelevante y se priorizó la presencia de tópicos con alta carga conceptual.

Una de las principales ventajas prácticas observadas en el uso de YAKE fue su capacidad para reforzar la interpretabilidad del pipeline de clasificación. A diferencia de los modelos de aprendizaje profundo, cuyas decisiones suelen ser opacas, YAKE permite rastrear con claridad los términos que fundamentan la pertenencia temática de cada documento. Esta transparencia metodológica adquiere particular relevancia cuando el objetivo analítico se articula con procesos de toma de decisiones en política científica o planificación estratégica, como ocurre al analizar la contribución investigativa a los ODS. En este sentido, el modelo no solo ofrece una solución técnica, sino que también aporta una herramienta argumentativa que puede ser utilizada para sustentar hallazgos frente a audiencias no especializadas.

Cabe destacar que diversos estudios comparativos recientes han validado el rendimiento competitivo de YAKE frente a otros modelos contemporáneos de extracción de palabras clave. Su precisión, diversidad semántica y eficiencia computacional han sido comprobadas en tareas de categorización temática, segmentación de tópicos y recomendación de contenidos en dominios tan diversos como el periodismo, la salud pública y la bibliometría [23]. En particular, su desempeño sobresaliente en textos de longitud corta y media lo convierte en una opción preferente para el análisis de resúmenes académicos, donde otras técnicas tienden a fallar por falta de contexto global.

4.2.1. Métricas de evaluación y coeficientes

En la evaluación de modelos de clasificación automática, es fundamental emplear métricas que no solo cuantifiquen el rendimiento general, sino que también consideren la concordancia entre las predicciones del modelo y las etiquetas de referencia, especialmente en contextos con clases desbalanceadas o solapamientos temáticos. Si bien métricas como la precisión, el recall y el F1-score son ampliamente utilizadas, pueden resultar insuficientes en escenarios donde el acuerdo podría ocurrir por azar [24].

El índice Kappa de Cohen es una medida estadística que cuantifica el grado de concordancia entre dos clasificadores, ajustando por la posibilidad de acuerdo aleatorio. Este coeficiente es particularmente útil en tareas de clasificación multiclase con clases minoritarias, ya que penaliza los acuerdos que podrían deberse al azar, ofreciendo una visión más equilibrada del rendimiento del modelo. Su valor oscila entre -1 y 1, donde 1 indica una concordancia perfecta y valores cercanos o inferiores a 0 sugieren una concordancia no mejor que la esperada por azar [25].

En el contexto de este estudio, la aplicación del índice Kappa es pertinente para evaluar la coherencia estructural del modelo, especialmente en condiciones de desbalance de clases y baja representación de ciertas categorías. Esta métrica complementa las evaluaciones tradicionales, proporcionando una perspectiva más robusta sobre la fiabilidad del modelo en la clasificación de textos alineados con los Objetivos de Desarrollo Sostenible (ODS).

4.2.2. Objetivos de Desarrollo Sostenible (ODS)

Los Objetivos de Desarrollo Sostenible (ODS) son un conjunto de 17 objetivos adoptados por las Naciones Unidas en 2015 como parte de la Agenda 2030 para el Desarrollo Sostenible en el mundo. Estos objetivos abordan los principales retos mundiales, desde la pobreza, el hambre, la salud, la educación, la igualdad de género, el agua limpia y el saneamiento, la energía asequible y no contaminante, el trabajo decente y el crecimiento económico, la industria, la innovación y la infraestructura, la reducción de las desigualdades, las ciudades y comunidades sostenibles, el consumo y la producción responsables, la acción por el clima, la vida submarina, la vida de ecosistemas terrestres, la paz, la justicia e instituciones sólidas, y las alianzas para lograr el cumplimento de los objetivos [26].

Los ODS están directamente ligados con la consecución de los Objetivos de Desarrollo del Milenio (ODM), los cuales fueron adoptados en el año 2000 y lograron avances significativos en áreas como la reducción de la pobreza extrema y la mejora de la salud materna e infantil.

Sin embargo, los ODS tienen un enfoque más amplio, integrando las dimensiones económica, social y ambiental del desarrollo sostenible [26]. Desde su adopción, han servido como marco de referencia para políticas públicas y programas de desarrollo a nivel global, regional y nacional. La Organización de las Naciones Unidas (ONU) ha desempeñado un papel fundamental en su

promoción y seguimiento, destacándose la labor de figuras como Ban Ki-moon, ex Secretario General, y Amina J. Mohammed, actual Vicesecretaria General. No obstante, el avance hacia el cumplimiento de los objetivos ha sido desigual entre países, debido a brechas estructurales, limitaciones institucionales y a factores como la pandemia de la COVID-19, que ralentizó la implementación de la Agenda 2030. Para superar estos desafíos, se requiere una cooperación internacional permanente, fortalecimiento de capacidades locales, y una movilización significativa de recursos, involucrando activamente a gobiernos, sector privado, academia, sociedad civil y comunidades [27].

Figura 2 Objetivos de Desarrollo Sostenible

4.3. Antecedentes

Para la realización de los antecedentes enmarcados en las investigaciones que abordan el mismo problema relacionado con la aplicación de redes neuronales y procesamiento de lenguaje para la evaluación de la investigación colombiana en el contexto de los ODS. Se realiza la siguiente estrategia de recuperación de información "Natural Language Processing" AND "Neural

Networks" AND "Sustainable Development Goals" AND “research evaluation” para la recuperación de las investigaciones relacionadas con la actual propuesta.

De igual forma, se seleccionan las bases de datos Scopus y Open Alex, como principales recursos de acceso abierto y de suscripción, que por su amplitud temática y calidad científica son los más idóneos y relacionados con la temática anteriormente seleccionada. Es pertinente aclarar que, en la búsqueda en español, ninguno de los resultados estaba relacionado con la investigación propuesta. En contraste, en el idioma inglés se evidenciaron alrededor cinco investigaciones relacionadas con los conceptos de inteligencia artificial, objetivos de desarrollo sostenible enmarcados en la evaluación de la investigación.

La primera investigación titulada *NLP for SDGs: Measuring Corporate Alignment with the Sustainable Development Goals* [28] tiene similitudes con la investigación propuesta, dado que, ambas investigaciones comparten el objetivo de aplicar técnicas de PLN, como relacionamiento de palabras y modelos avanzados (Word2Vec, Doc2Vec), con la finalidad de analizar textos y evaluar su alineación con los ODS. Mientras que el artículo realizado por M. Chen, et. al. enfoca su análisis en los informes de Responsabilidad Social Corporativa (RSC) de empresas y utiliza clasificadores como SVM y redes neuronales, la investigación propuesta propone un enfoque similar aplicado a textos de investigación científica, empleando modelos de redes neuronales y técnicas de análisis de similitud temática para determinar la relación entre la producción académica y los ODS.

La afinidad de estas metodologías sugiere que las técnicas empleadas en el análisis corporativo pueden ser adaptadas y extendidas eficazmente al ámbito académico, facilitando la identificación y evaluación de contribuciones científicas en relación con los ODS en Colombia.

Por otra parte, la investigación denominada *contribution of Deep Learning and Artificial Intelligence to Attaining the Sustainable Development Goals Amidst the COVID-19 Pandemic* [29] se enfoca en evaluar el papel de la inteligencia artificial (IA) y el aprendizaje profundo en la mitigación de las amenazas planteadas por la pandemia y su contribución a los ODS, especialmente en el ámbito de la salud y el bienestar. Su relación con el estudio destaca en la importancia de las tecnologías avanzadas en el contexto de los ODS. Sin embargo, se diferencian en su enfoque y alcance porque: el primero se centra en la clasificación y análisis de la investigación académica

para alinearla con los ODS en Colombia, mientras que el segundo evalúa cómo la IA y el aprendizaje profundo han sido fundamentales para enfrentar una crisis global de salud, demostrando su potencial para apoyar el logro de los ODS a nivel global.

La siguiente investigación titulada *Meta-Analysis of Satellite Observations for United Nations Sustainable Development Goals: Exploring the Potential of Machine Learning for Water Quality Monitoring* [30] se relaciona conceptualmente con la investigación propuesta en cuanto a la aplicación de tecnologías avanzadas de inteligencia artificial (IA) y aprendizaje automático para abordar desafíos complejos en sus respectivos campos. Ambas investigaciones subrayan la capacidad de los modelos de aprendizaje profundo y técnicas avanzadas para manejar grandes volúmenes de datos complejos. Mientras que la investigación propuesta se centra en la alineación y análisis temático de la producción científica respecto a los ODS utilizando PLN y modelos basados en redes neuronales, el estudio de revisión se centra en el monitoreo de la calidad del agua a través de datos satelitales y modelos de regresión supervisada articulando su cumplimiento con los ODS.

La propuesta realizada por Smith et. al. la cual se titula *Discovering new pathways toward integration between health and sustainable development goals with natural language processing and network science* [31] tiene similitudes con la propuesta realizada en que ambas investigaciones emplean técnicas de PLN y modelos de redes neuronales para analizar grandes corpus de literatura científica, identificar temas clave y medir la interconexión temática con los ODS. La diferencia principal radica en su enfoque específico. Esta investigación integra la temática de la salud (ODS 3) con otros ODS a nivel global, utilizando el algoritmo top2vec y métodos de ciencia en red para identificar temas delimitados para sugerir nuevos dominios de investigación.

Dentro de las investigaciones que analizan el cumplimiento de los ODS con artículos investigativos y a su vez, con los informes de progreso de la ONU, se encuentra la investigación *Natural language processing and network analysis provide novel insights on policy and scientific discourse around Sustainable Development Goals* [32] la cual analiza la interdependencia de los ODS a través de un corpus global de informes de progreso de la ONU y artículos científicos, destacando las diferencias en el discurso político y científico internacional sobre los ODS. Mientras que la investigación propuesta busca aplicar estos métodos para una evaluación específica del contexto colombiano, el

estudio mencionado ofrece una perspectiva más amplia sobre la colaboración global en la ciencia relacionada con los ODS. En síntesis, ambas investigaciones subrayan la importancia de las herramientas computacionales para entender las complejas interrelaciones entre los ODS, aunque en contextos diferentes.

Para concluir la revisión de antecedentes se identifica la ausencia de estudios similares en español y a su vez, en tipología de disertación o tesis, también se destaca la poca presencia de investigaciones desde proyectos relacionados en segunda lengua como el inglés, en el que apliquen técnicas de procesamiento de lenguaje natural (PLN) y aprendizaje profundo para analizar la alineación con los Objetivos de Desarrollo Sostenible (ODS). Las investigaciones revisadas, aunque diversas en su enfoque, comparten la aplicación de tecnologías avanzadas como modelos de redes neuronales y análisis de similitud temática para abordar desafíos en el contexto de los ODS.

La propuesta de investigación se diferencia al centrarse específicamente en el análisis de textos académicos en las áreas de ingeniería y medicina, utilizando herramientas avanzadas de PLN y modelos basados en redes neuronales para evaluar su alineación con los ODS en un contexto nacional tal como lo es Colombia. Este enfoque permite una evaluación precisa de cómo la producción científica colombiana contribuye al logro de los ODS, proporcionando perspectivas importantes para la formulación de políticas públicas en torno a la ciencia y la tecnología.

5. RECOPILACIÓN DE CORPUS DE INFORMACIÓN Y ETIQUETADO

Este capítulo documenta el cumplimiento del primer objetivo específico del proyecto, orientado a etiquetar manualmente una muestra representativa de resúmenes de investigaciones colombianas en ingeniería y medicina según su alineación temática con los Objetivos de Desarrollo Sostenible (ODS). Esta etapa no solo constituyó el insumo fundamental para el entrenamiento supervisado de modelos de clasificación, sino que además permitió caracterizar de forma crítica la estructura y distribución de la producción científica nacional reciente, visibilizando asimetrías disciplinares, sesgos temáticos y niveles de convergencia o ausencia de ella con las metas globales de desarrollo planteadas por la Agenda 2030.

A diferencia de enfoques puramente cuantitativos basados en conteo de publicaciones, el trabajo desarrollado en este capítulo se centró en un análisis semántico cualitativo, que permitió identificar las limitaciones del enfoque declarativo con el que tradicionalmente se reporta el alineamiento con los ODS. En ese sentido, el ejercicio de etiquetado expuso la distancia entre la formulación temática de muchas investigaciones y su efectiva conexión con las problemáticas estructurales que los definen, particularmente en aquellas áreas donde dicha alineación es menos explícita o más interpretativa, como ocurre en el caso de ciertas sublíneas de ingeniería.

5.1. Recopilación de Datos desde Scopus

La selección del corpus de este estudio se fundamentó en la recuperación sistemática de resúmenes de artículos científicos realizados por investigadores colombianos, utilizando como fuente principal la base de datos Scopus. La consulta fue restringida a documentos de acceso abierto, en lengua inglesa, pertenecientes a las áreas de Ingeniería y Medicina, dos dominios disciplinares con relevancia estratégica tanto por su volumen de producción como por su potencial incidencia en los Objetivos de Desarrollo Sostenible (ODS). El periodo seleccionado comprendió los años 2018 a 2024, lo que permitió incorporar tendencias investigativas recientes y ofrecer una visión amplia de la actividad científica nacional.

El resultado de esta búsqueda arrojó un total de 23.229 resúmenes, cuya distribución por área revela un claro desequilibrio temático: 7.539 textos correspondieron a Ingeniería, lo que representa el 32,4 % del total, mientras que 15.690 fueron clasificados dentro del campo de la Medicina, equivalente al 67,6 %. Esta disparidad es un reflejo cuantificable de los patrones de especialización científica en el contexto colombiano, y plantea implicaciones metodológicas significativas en relación con el diseño del modelo de clasificación. En particular, obliga a considerar estrategias de control de sesgo, ajuste de pesos por clase y técnicas de balanceo para evitar posibles sobres ajuste hacia los ODS de medicina.

5.2. Uso y autorización de los datos provenientes de Scopus

Los datos utilizados en este estudio fueron obtenidos a través de la plataforma Scopus, perteneciente a la editorial Elsevier, en el marco de una suscripción institucional con fines exclusivamente académicos. La recolección se enfocó en publicaciones científicas clasificadas bajo el área temática de medicina, con afiliación a instituciones colombianas y publicadas entre 2018 y 2024. La búsqueda se limitó a artículos de acceso abierto, lo que permitió garantizar un uso respetuoso de los derechos de autor y de las condiciones de disponibilidad impuestas por los editores.

El proceso de extracción se realizó mediante la aplicación de filtros avanzados disponibles en la interfaz de Scopus, recuperando únicamente los metadatos permitidos por la política de uso de la plataforma. Estos incluyeron título del documento, resumen, autores, afiliaciones, palabras clave, tipo de documento, año de publicación, identificadores únicos (DOI y EID), y número de citas, entre otros campos autorizados. En ningún caso se accedió al texto completo de los artículos ni se emplearon técnicas de minería de datos masiva fuera de las funciones previstas por el acceso institucional.

Conforme a los términos establecidos por Elsevier, el uso de datos extraídos de Scopus para investigaciones académicas está permitido siempre que el objetivo no sea comercial, se respeten las restricciones de redistribución, y se realice la debida atribución de la fuente. En este estudio, se ha dado cumplimiento a estos lineamientos, asegurando que el tratamiento de los datos se limite a análisis internos y visualizaciones integradas en esta monografía, sin exponer o compartir los

registros de forma externa. Además, Scopus ha sido reconocida como la fuente exclusiva de los datos empleados, conforme a lo indicado en su guía de atribución oficial [33].

5.3. Selección de Muestra Representativa

El desarrollo de un modelo supervisado de clasificación temática exige una muestra de entrenamiento que no solo sea cuantitativamente suficiente, sino que preserve la estructura disciplinar de la población original. A partir de un corpus inicial compuesto por 2.180 artículos científicos en acceso abierto publicados en 2018 y extraídos desde la base de datos Scopus, se definió una muestra representativa a través de una distribución proporcional por área de conocimiento. Esta técnica no implica un muestreo estratificado en sentido técnico, ya que no se aplicaron criterios de aleatorización dentro de cada estrato, sino una asignación directa proporcional al peso relativo de cada disciplina dentro del conjunto total.

El análisis del corpus reveló un fuerte desequilibrio en la distribución temática: el 75,1 % de los artículos (1.637) correspondían al área de Medicina, mientras que el 24,9 % (543) estaban clasificados bajo Ingeniería. Esta asimetría empírica no es meramente numérica, sino que refleja un patrón consolidado de concentración investigativa en las ciencias de la salud. Incluir esta proporción en la muestra fue una medida deliberada para evitar sesgos de sobreajuste hacia la clase mayoritaria, conservando la diversidad semántica que caracteriza la producción en ambas áreas.

El tamaño muestral se calculó mediante la fórmula para poblaciones finitas, considerando un nivel de confianza del 95 %, una proporción máxima de variabilidad (p = 0,5) y un margen de error del 5 %. La expresión utilizada fue:

$$n = \frac{N \cdot Z^2 \cdot p \cdot (1-p)}{(e^2 \cdot (N-1)) + (Z^2 \cdot p \cdot (1-p))}$$

Figura 3 Formula tamaño de la muestra

donde N=2180, Z=1,96, p=0,5 y e= 0,05. El resultado arrojó un tamaño muestral de 311,7 unidades, que fue redondeado a 312 artículos para facilitar la distribución proporcional sin modificar la estructura original del corpus. Esta asignación se tradujo en 234 artículos de Medicina y 78 de

Ingeniería, cifras que reproducen con fidelidad la distribución temática de la población. Lejos de ser una decisión operativa, esta proporción constituye un mecanismo de control estadístico destinado a preservar la integridad epistemológica del modelo, evitando que su capacidad de generalización se vea comprometida por sesgos inducidos durante el entrenamiento. La estrategia no solo contribuyó a una mejor representatividad del conjunto de entrenamiento, sino que también estableció una base sólida para evaluar la capacidad del modelo de reconocer patrones diferenciados según el campo disciplinar.

5.4. Proceso de Etiquetado Manual

Finalizado el proceso de muestreo, se procedió a la etapa de etiquetado manual de los 312 resúmenes seleccionados, todos correspondientes a artículos científicos publicados por autores afiliados a instituciones colombianas. El objetivo fue determinar su grado de alineación temática con los Objetivos de Desarrollo Sostenible (ODS) propuestos por la Agenda 2030. Este procedimiento fue ejecutado mediante una lectura analítica de cada resumen, en la cual se evaluó tanto la mención explícita de temáticas relacionadas con los ODS como su tratamiento implícito, considerando la naturaleza del problema abordado, la población objetivo, los métodos aplicados y los impactos proyectados.

El criterio de asignación se sustentó en una matriz de correspondencias semánticas construida a partir de los descriptores oficiales de los ODS y de literatura previa sobre mapeo temático en investigación científica. Esta metodología permitió establecer una clasificación informada, aunque no exenta de subjetividad interpretativa, particularmente en aquellos casos en los que la contribución al desarrollo sostenible no estaba claramente formulada en términos declarativos. El análisis de frecuencia de las etiquetas asignadas evidenció un marcado desequilibrio tanto entre disciplinas como entre ODS. Del total de documentos, 201 resúmenes pertenecen al área de Medicina (64,4 %) y 111 al área de Ingeniería (35,6 %). A su vez, el ODS 3 (Salud y bienestar) concentró una proporción desproporcionada de las etiquetas, representando el 50,9 % del total general. En contraste, otros objetivos como el ODS 8 (Trabajo decente y crecimiento económico), el ODS 14 (Vida submarina) o el ODS 16 (Paz, justicia e instituciones sólidas) fueron apenas representados, lo cual revela una focalización temática y una limitada diversificación de las

agendas de investigación.

La siguiente tabla presenta la distribución completa de los resúmenes etiquetados por área de conocimiento y por ODS:

Tabla 1 Etiquetado ODS – Muestra estratificada

ODS	Ingeniería	Medicina	Total
ODS 2 – Hambre cero	3	4	7
ODS 3 – Salud y bienestar	8	159	167
ODS 4 – Educación de calidad	–	10	10
ODS 5 – Igualdad de género	–	5	5
ODS 6 – Agua limpia y saneamiento	8	6	14
ODS 7 – Energía asequible y no contaminante	14	–	14
ODS 8 – Trabajo decente y crecimiento económico	2	–	2
ODS 9 – Industria, innovación e infraestructura	44	6	50
ODS 10 – Reducción de las desigualdades	–	3	3
ODS 11 – Ciudades y comunidades sostenibles	7	1	8
ODS 12 – Producción y consumo responsables	6	1	7
ODS 13 – Acción por el clima	10	2	12
ODS 14 – Vida submarina	3	–	3
ODS 15 – Vida de ecosistemas terrestres	2	4	6
ODS 16 – Paz, justicia e instituciones sólidas	4	–	4
Total	**111**	**201**	**312**

5.5. Validación del Etiquetado

El proceso de validación del etiquetado tuvo como propósito garantizar la coherencia interna del conjunto de datos construido durante la etapa de clasificación manual. Si bien el protocolo inicial se fundamentó en lineamientos claros para asociar resúmenes de investigación con uno o más de los Objetivos de Desarrollo Sostenible (ODS), en la práctica surgieron situaciones que exigieron ajustes interpretativos. No todos los resúmenes presentaban una vinculación explícita con las temáticas de los ODS, lo que obligó a desarrollar criterios de lectura más finos, atentos tanto a los enfoques de investigación como a sus posibles implicaciones sociales, económicas o ambientales.

La validación se abordó como un ejercicio sistemático de doble lectura. En aquellos casos donde la relación entre el contenido del resumen y el ODS asignado no era evidente, se realizó una reevaluación crítica, considerando no solo la terminología empleada, sino también la estructura argumentativa del texto. Esta revisión fue especialmente relevante en disciplinas técnicas, como la ingeniería, donde la alineación con los ODS tiende a materializarse de forma indirecta por ejemplo, a través de aplicaciones tecnológicas orientadas a la eficiencia energética o la gestión de recursos hídricos más que mediante declaraciones temáticas explícitas.

Durante este proceso, se confirmó que la etiqueta más recurrente fue el ODS 3 (Salud y bienestar), lo cual es consistente con la sobrerrepresentación de publicaciones provenientes del campo de la medicina. Este hallazgo, más allá de ser una constatación cuantitativa, pone en evidencia un sesgo estructural que atraviesa la agenda investigativa nacional. La presencia marginal de otros objetivos, como los ODS 14 o 16, no responde a una deficiencia del procedimiento de etiquetado, sino a la escasa articulación de estas temáticas en la producción científica analizada.

Esta etapa permitió cerrar el primer objetivo específico del proyecto con resultados concretos: un conjunto curado de 312 resúmenes clasificados y validados, representativos de las áreas de medicina e ingeniería. Sin embargo, lo más relevante fue quizá el carácter diagnóstico de este ejercicio. La validación no solo mejoró la calidad del conjunto de datos, sino que también reveló las limitaciones del lenguaje académico para reflejar, de forma directa, la relación entre ciencia y sostenibilidad. En muchos casos, la ausencia de vocabulario vinculado a los ODS no implica una falta de pertinencia temática, sino una distancia en los marcos de referencia desde los que se

comunica el conocimiento científico.

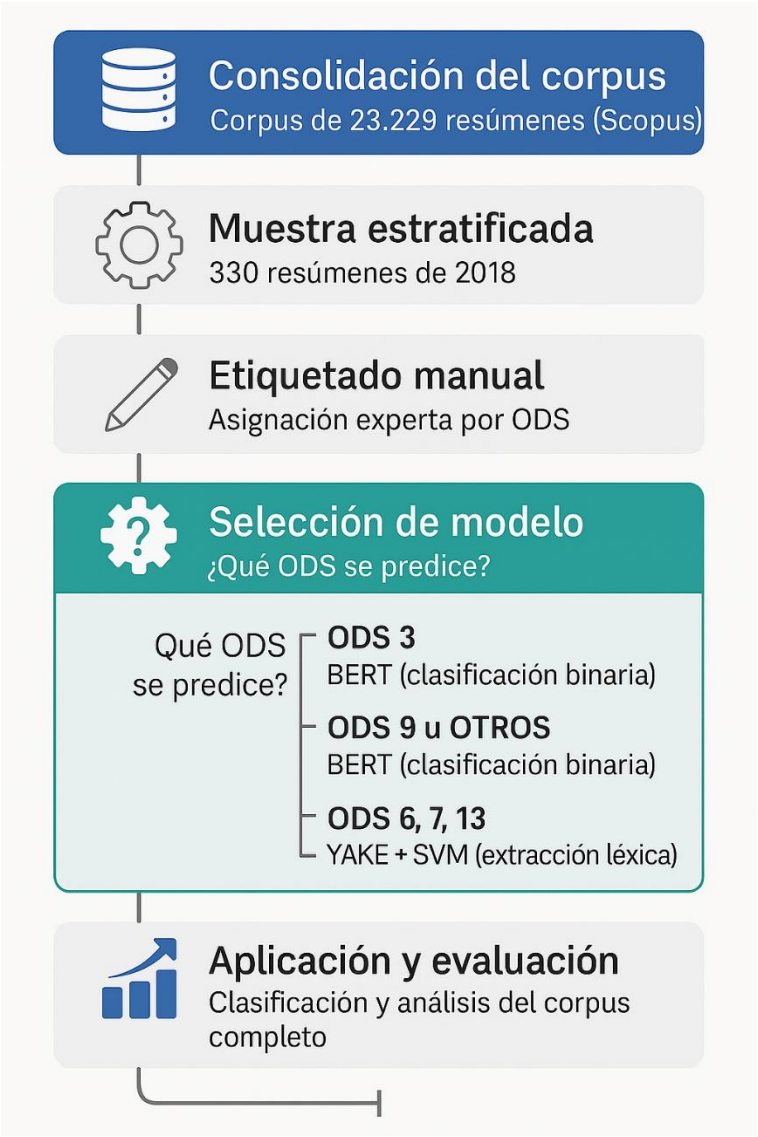

Figura 4 muestra y etiquetado ODS

6. ENTRENAMIENTO DE LOS MODELOS DE REDES NEURONALES

Este capítulo presenta el desarrollo y ajuste metodológico de los modelos de clasificación automática implementados para predecir la alineación temática de artículos científicos colombianos con los Objetivos de Desarrollo Sostenible (ODS). Esta etapa responde al segundo objetivo específico del proyecto y se fundamenta en la aplicación de técnicas de procesamiento de lenguaje natural y aprendizaje profundo, empleando como insumo un corpus previamente etiquetado mediante evaluación manual.

La naturaleza del problema exigió adoptar un enfoque progresivo, no lineal, basado en el análisis empírico de los resultados obtenidos en cada fase del entrenamiento. En lugar de asumir que un único modelo podía abordar con solvencia la clasificación multiclase, se optó por una estrategia adaptativa orientada a la heterogeneidad del corpus. El diseño final incluyó modelos separados para subconjuntos de ODS, diferenciando entre categorías dominantes, como el ODS 3, y clases subrepresentadas o temáticamente fragmentadas, como los ODS 6, 7 y 13. La integración de enfoques supervisados y no supervisados se justificó tanto por razones técnicas como el desequilibrio en la distribución de clases como por criterios de validez semántica.

El capítulo se estructura en tres secciones principales. En la sección 6.1 se documenta el preprocesamiento del corpus, incluyendo técnicas de limpieza textual, verificación de idioma, aumento de datos y reagrupación temática de clases. Estas acciones no se limitaron a operaciones rutinarias de normalización, sino que respondieron a la necesidad de controlar el ruido semántico y mitigar los efectos del desbalance en el entrenamiento. En la sección 6.2 se describe la arquitectura de los modelos entrenados, diferenciando el uso de BERT para los ODS con representación suficiente (ODS 3, ODS 9 y OTROS), y la incorporación de YAKE y máquinas de vectores de soporte para la clasificación de los ODS ambientales. Finalmente, la sección 6.3 expone los resultados comparativos de desempeño, analizando las métricas obtenidas y reflexionando sobre las limitaciones observadas.

6.1. Preprocesamiento de Datos

Los artículos empleados en este estudio fueron extraídos de la base de datos Scopus, seleccionando exclusivamente publicaciones de acceso abierto correspondientes a las áreas de ingeniería y medicina, en el periodo 2018. Cada documento fue organizado y almacenado en un archivo estructurado, previamente categorizado según su alineación con uno o más Objetivos de Desarrollo Sostenible (ODS), conforme al proceso de etiquetado manual descrito en secciones anteriores.

Como primera etapa del preprocesamiento, se verificó la integridad lingüística del corpus mediante la biblioteca langdetect, asegurando que la totalidad de los resúmenes estuviera redactada en idioma inglés, condición indispensable para la aplicación del modelo BERT. Posteriormente, se llevó a cabo un proceso riguroso de limpieza del texto, empleando BeautifulSoup y expresiones regulares para eliminar etiquetas HTML residuales, caracteres especiales, signos de puntuación redundantes y espacios superfluos. Esta normalización buscó optimizar la calidad semántica del texto y evitar interferencias durante la fase de tokenización. Es importante mencionar que BERT trabaja con representaciones de tokens sensibles al contexto, asegurar un texto limpio y en inglés es crucial para preservar la calidad de los embeddings generados, optimizando la captura de relaciones semánticas ver tabla.

Tabla 2. Técnicas de Preprocesamiento Aplicadas

Técnica	Descripción breve
Verificación de idioma	Detección de idioma inglés mediante langdetect.
Limpieza de texto	Eliminación de etiquetas HTML, caracteres especiales y espacios redundantes con BeautifulSoup y regex.
Aumento de datos	Traducción automática (inglés → italiano → inglés) y parafraseo con Parrot.
Agrupación temática de clases	Reagrupación de ODS 6, 7 y 13 bajo categoría ambiental.

Una vez depurado el corpus, se abordó el desafío del desbalance de clases, provocado por la distribución heterogénea de artículos entre los distintos ODS. En respuesta, se implementó una estrategia de aumento de datos como se indica en la tabla 2, la cual esta enfocada en las clases minoritarias, basada en técnicas de traducción automática. Los resúmenes pertenecientes a estas categorías fueron traducidos del inglés al italiano y, posteriormente, de vuelta al inglés, generando variantes textuales conservando su significado original. Este procedimiento fue complementado con técnicas de parafraseo, utilizando la biblioteca Parrot, con el fin de enriquecer el conjunto de entrenamiento con versiones semánticamente equivalentes y morfosintácticamente distintas.

Adicionalmente, se procedió a una agrupación temática de ODS con características semánticas convergentes, con el objetivo de reducir la dispersión categórica en el proceso de clasificación y mejorar la robustez del modelo. En este contexto, se excluyó deliberadamente el ODS 3 (Salud y Bienestar), debido a su elevada representación en el conjunto de datos, la cual amenazaba con inducir un sesgo significativo en el proceso de entrenamiento al sobre representar patrones asociados a dicho objetivo. Esta exclusión respondió a criterios metodológicos orientados a preservar la equidad entre clases y garantizar un rendimiento equilibrado del modelo en tareas de clasificación multiclase.

Tabla 3. Variación del Número de Muestras por ODS tras Aumento de Datos

Categoría	Artículos originales	Artículos tras aumento
ODS 3	150	150
ODS 9	80	160
ODS 6/7/13 (agrupados)	40	120
OTROS	40	80

6.2. Selección y Configuración de Modelos

Inicialmente, se adoptó un enfoque unificado para el entrenamiento de un modelo de clasificación automática mediante el uso de BERT (Bidirectional Encoder Representations from Transformers) en su versión bert-base-uncased. La decisión de utilizar este modelo se basó en su capacidad para representar relaciones semánticas contextuales complejas, condición necesaria para capturar la alineación temática entre los resúmenes de artículos científicos y los Objetivos de Desarrollo Sostenible (ODS). Este modelo se entrenó sobre la muestra etiquetada de resúmenes correspondiente a las áreas de Ingeniería y Medicina, abarcando todas las clases de ODS identificadas en la fase de anotación manual.

Durante las primeras iteraciones de entrenamiento y validación, se evidenciaron problemas significativos de sobreajuste, particularmente en relación con la alta representatividad de ciertas clases, en especial el ODS 3 (Salud y Bienestar), cuya frecuencia excedía por amplio margen a la de otras categorías. Esta situación generaba un sesgo sistemático en las predicciones, favoreciendo la clase dominante y afectando negativamente la capacidad del modelo para generalizar sobre ODS menos representados.

Para mitigar este fenómeno, se implementó una estrategia de agrupación de clases, con el fin de equilibrar la distribución y reducir la dispersión semántica. Se definieron cinco categorías: ODS 3, ODS 6-7-13 (agrupadas por afinidad temática ambiental), ODS 9, y una categoría residual denominada OTROS la cual incluye los ODS restantes con muy baja representatividad de la muestra. Esta reorganización permitió avanzar hacia un modelo más robusto y estable en términos de entrenamiento, mejorando las métricas generales de rendimiento.

Sin embargo, los análisis posteriores revelaron que, a pesar de la mejora general, el rendimiento del modelo BERT sobre la clase agrupada ODS 6-7-13 resultaba inestable y con métricas incoherentes, particularmente en precisión y F1-score. Esta inconsistencia se debía, en parte, a la escasa representación de estos objetivos en la muestra, así como a su mayor heterogeneidad temática. Ante esta situación, se optó por un enfoque alternativo complementario basado en YAKE (Yet Another Keyword Extractor), que permitió abordar la clasificación de estos ODS mediante técnicas de extracción de palabras clave y análisis semántico por similitud, fuera del marco

supervisado de BERT (ver Tabla 4.)

Tabla 4. Iteraciones de BERT sobre ODS

Iteración	Técnica Principal	Conjunto de ODS	Problemática Detectada	Ajuste Implementado
Inicial	BERT unificado	ODS 3, 6, 7, 9, 13, OTROS	Sobreajuste por desbalance	Agrupación de ODS y reentrenamiento
Segunda	BERT por grupos	ODS 3, ODS 9, OTROS	Métricas aceptables	Conservación del modelo
Segunda	BERT por grupos	ODS 6-7-13	Métricas incoherentes, baja representatividad	Transición a YAKE + clasificación semántica

Debido a esto, el ajuste metodológico permitió aplicar BERT exclusivamente para las clases con mayor cohesión semántica y mejor desempeño en el modelo (ODS 3, ODS 9, y OTROS), mientras que los resúmenes potencialmente alineados con los ODS 6, 7 y 13 fueron tratados mediante el pipeline YAKE, combinando extracción de términos relevantes, normalización y posterior clasificación por agrupamiento. En cuanto a la configuración técnica de BERT, el modelo fue ajustado con una tasa de aprendizaje de 2×10−52 \times 10^{-5}2×10−5, weight decay de 0.01 y dropout del 30% para evitar el sobreajuste. Se aplicó el cálculo de pesos balanceados por clase mediante la función compute_class_weight de scikit-learn, y se utilizó gradient checkpointing para reducir el consumo de memoria durante el entrenamiento. El entrenamiento se llevó a cabo de forma separada por área de conocimiento (Ingeniería y Medicina) y por modelo específico, permitiendo una adaptación más fina a las características textuales de cada dominio.

6.3. Entrenamiento del Modelo

El entrenamiento de los modelos de clasificación se estructuró a partir de una estrategia híbrida que combinó enfoques supervisados y no supervisados, adaptados a la naturaleza y distribución de las clases representadas en el corpus. La variabilidad temática entre los Objetivos de Desarrollo Sostenible (ODS), así como el desbalance en la frecuencia de aparición de cada categoría, hicieron inviable la construcción de un único modelo multiclase robusto. Por esta razón, se optó por dividir la tarea en tres bloques: un modelo exclusivo para ODS 3, otro para ODS 9 y la clase OTROS, y

finalmente un modelo no supervisado para los ODS 6, 7 y 13.

6.3.1. Modelo BERT para ODS 3

El primer modelo fue desarrollado con el objetivo de abordar de manera específica la clasificación binaria entre artículos alineados con el ODS 3 (Salud y bienestar) y aquellos sin una vinculación directa con esta categoría. Esta decisión metodológica se sustentó en los resultados observados durante los entrenamientos iniciales en entornos multiclase, donde se evidenció un sesgo significativo inducido por la alta frecuencia de textos etiquetados con el ODS 3. Dicha concentración generaba una distorsión sistemática en el aprendizaje del modelo, al favorecer la clase dominante y deteriorar la capacidad de generalización sobre las demás categorías.

Con el propósito de mitigar este efecto, se construyó un modelo independiente utilizando la arquitectura BERT base uncased, configurada para una tarea de clasificación binaria. El corpus fue depurado previamente para garantizar la calidad lingüística y semántica del contenido. Solo se incluyeron resúmenes redactados en inglés, sin estructuras sintácticas defectuosas ni duplicaciones, y se aplicó un balance de clases para estabilizar la representación de ambas etiquetas durante el entrenamiento. El modelo fue entrenado durante cinco épocas, con una tasa de aprendizaje de $2\times10-5\times10-5$, utilizando el optimizador AdamW, validación por ciclo y selección automática del mejor checkpoint basado en el valor más alto del F1-score. La curva de pérdida indicó una convergencia progresiva y estable a partir de la segunda época, sin signos de sobreajuste. Esta estabilidad se reflejó en métricas consistentes a lo largo de los pliegues de validación cruzada, lo que respalda la robustez del modelo en tareas de discriminación binaria dentro de contextos biomédicos.

El desarrollo técnico del modelo, así como su proceso de entrenamiento completo, se encuentra documentado en el repositorio público del proyecto. El código fuente, escrito en Python y ejecutado en entorno Jupyter, puede consultarse en la siguiente dirección: https://github.com/jhonrd999/PNL_ODS_BERT_YAKE/blob/main/BERT_ODS_3_%20NO_ODS.ipynb

Este cuaderno incluye todas las fases del pipeline: carga del dataset, tokenización con AutoTokenizer, definición de la arquitectura AutoModelForSequenceClassification, configuración

de los hiperparámetros y evaluación final con métricas de rendimiento. A continuación, se presenta el reporte de clasificación correspondiente a la validación del modelo entrenado exclusivamente para el ODS 3. Esta matriz refleja un desempeño sólido y equilibrado en términos de precisión, recall y F1-score, validando empíricamente la pertinencia de separar esta clase en un modelo autónomo dentro de la arquitectura general del sistema.

Tabla 5. Reporte de clasificación del modelo ODS 3

Clase	Precisión (%)	Recall (%)	F1-Score (%)	Support
NO ODS 3	90	90	90	20
ODS 3	94	94	94	33
Métrica global	**Precisión (%)**	**Recall (%)**	**F1-Score (%)**	**Support**
Accuracy	92	92	92	53
Macro avg	92	92	92	53
Weighted avg	92	92	92	53

6.3.2. Modelo BERT para ODS 9 y OTROS

El segundo modelo fue diseñado para abordar la clasificación binaria entre artículos relacionados con el ODS 9 (Industria, innovación e infraestructura) y aquellos alineados con un conjunto de ODS minoritarios agrupados bajo la categoría OTROS. Esta última incluye los objetivos con menor frecuencia en la muestra etiquetada: ODS 2 (Hambre cero), ODS 4 (Educación de calidad), ODS 5 (Igualdad de género), ODS 8 (Trabajo decente y crecimiento económico), ODS 10 (Reducción de desigualdades), ODS 11 (Ciudades y comunidades sostenibles), ODS 12 (Producción y consumo responsables), ODS 14 (Vida submarina), ODS 15 (Vida de ecosistemas terrestres) y ODS 16 (Paz, justicia e instituciones sólidas). La decisión de agrupar estos objetivos bajo una única etiqueta respondió a una limitación cuantitativa: el número reducido de muestras por ODS impedía entrenar modelos individuales con validez estadística y estabilidad métrica.

Se utilizó la arquitectura BERT base uncased, replicando las condiciones técnicas aplicadas al modelo anterior. El entrenamiento se realizó durante cinco épocas, con una tasa de aprendizaje de 2×10−5×10−5, optimización mediante el algoritmo AdamW, aplicación de weight decay (0.01), y validación por ciclo. Además, se incorporó el ajuste de pesos por clase para corregir el desbalance

inherente al conjunto de datos, priorizando la estabilidad de las métricas F1 en ambas categorías.

Los resultados obtenidos fueron consistentes en los diferentes pliegues de validación, con un rendimiento satisfactorio en la detección de textos alineados con el ODS 9. La precisión fue levemente inferior en la clase OTROS, lo cual se explica por su heterogeneidad semántica: al integrar múltiples objetivos temáticamente diversos en una sola categoría, el modelo enfrenta una mayor ambigüedad en la asignación de patrones representativos. La matriz de confusión reflejó esta dificultad, especialmente en textos con contenido transversal, donde el vocabulario puede compartir estructuras léxicas con ODS técnicos sin corresponder plenamente a los dominios del ODS 9. El desarrollo completo de este modelo se encuentra documentado en el repositorio oficial del proyecto. El código fuente incluyendo carga del corpus, tokenización, definición del clasificador, entrenamiento y evaluación está disponible en el siguiente enlace: https://github.com/jhonrd999/PNL_ODS_BERT_YAKE/blob/main/BERT_ODS_9_OTROS.ipynb.

Desde el punto de vista práctico, este modelo permitió identificar con precisión textos orientados a la innovación, la industria y la infraestructura sostenible, e incorporar de manera estructurada aquellos artículos que, si bien no se vinculan a los ODS con mayor frecuencia, constituyen evidencia valiosa de investigación en sostenibilidad. La inclusión del grupo OTROS fue, en este sentido, una decisión metodológica orientada a maximizar la cobertura temática del modelo sin comprometer su rendimiento, articulando coherencia estadística y sensibilidad semántica.

Tabla 6. Reporte de clasificación del modelo unificado ODS 9

Clase	Precisión (%)	Recall (%)	F1-Score (%)	Support
ODS 9	87	91	89	22
OTROS	90	86	88	22
Métrica global	**Precisión (%)**	**Recall (%)**	**F1-Score (%)**	**Support**
Accuracy	89	89	89	44
Macro avg	89	89	89	44
Weighted avg	89	89	89	44

6.3.3. Modelo alternativo con YAKE y SVM para ODS 6, 7 y 13

La clasificación automática de resúmenes relacionados con los ODS 6 (Agua limpia y saneamiento), ODS 7 (Energía asequible y no contaminante) y ODS 13 (Acción por el clima) presentó desafíos significativos durante las primeras iteraciones con modelos basados en BERT. Aunque se intentó agrupar estos tres objetivos bajo una única clase ambiental para mitigar el problema de escasa representatividad individual, los resultados obtenidos fueron sistemáticamente deficientes. En particular, se observó recall igual a cero en varios ciclos de validación cruzada, lo cual evidenció la incapacidad del modelo supervisado para discriminar de forma consistente los textos vinculados a esta categoría.

Frente a esta limitación, se adoptó una estrategia metodológica alternativa, más adecuada para contextos con bajo volumen de datos y alta dispersión temática. El nuevo enfoque se fundamentó en la extracción de palabras clave mediante YAKE (Yet Another Keyword Extractor), seguida de una etapa de clasificación con una Máquina de Vectores de Soporte (SVM). Esta decisión respondió a dos criterios principales: por un lado, la necesidad de construir representaciones semánticas compactas y relevantes a partir de textos breves; por otro, la eficacia comprobada de los clasificadores SVM en dominios con escaso volumen de entrenamiento y dimensionalidad moderada.

El algoritmo YAKE fue aplicado sobre los resúmenes preprocesados, extrayendo los términos más informativos por documento sin requerir un corpus externo de referencia. Estas palabras clave fueron utilizadas para construir vectores de características comparables con perfiles léxicos previamente definidos para los ODS 6, 7 y 13, elaborados a partir de una selección temática fundamentada en literatura especializada en sostenibilidad ambiental. Posteriormente, estos vectores fueron empleados como entrada para el clasificador SVM, optimizado mediante búsqueda en rejilla de hiperparámetros e implementación de validación cruzada estratificada.

Aunque el tamaño del conjunto de prueba fue limitado, los resultados obtenidos validan la pertinencia de esta arquitectura alternativa. El modelo logró identificar patrones léxicos relacionados con eficiencia energética, recursos hídricos, sostenibilidad ambiental y mitigación del

cambio climático, mostrando una capacidad de generalización superior a la alcanzada con los modelos neuronales en este mismo conjunto. Las métricas obtenidas reflejan una mejora sustancial en precisión, recall y F1-score, especialmente en comparación con los intentos previos de clasificación supervisada.

El código correspondiente a este pipeline se encuentra disponible en el repositorio del proyecto. El cuaderno de trabajo, desarrollado en entorno Jupyter, puede consultarse en el siguiente enlace:

https://github.com/jhonrd999/PNL_ODS_BERT_YAKE/blob/main/Yake_mejor_test_6-7-13.ipynb. Este archivo documenta todas las etapas: extracción de palabras clave, vectorización semántica, construcción de perfiles de referencia, configuración del modelo SVM y validación final.

Tabla 7. Reporte de clasificación del modelo en clases con menor representación

Clase	Precisión (%)	Recall (%)	F1-Score (%)	Support
ODS 13	83	83	83	6
ODS 6	88	100	93	7
ODS 7	100	86	92	7
Métrica global	**Precisión (%)**	**Recall (%)**	**F1-Score (%)**	**Support**
Accuracy	90	90	90	20
Macro avg	90	90	90	20
Weighted avg	91	90	90	20

7. APLICACIÓN DEL MODELO DE CLASIFICACIÓN SOBRE EL CORPUS COMPLETO

7.1. Implementación del modelo

La fase de implementación del modelo de clasificación marcó el inicio de la aplicación del sistema desarrollado sobre un corpus real y representativo, conformado por 23.229 resúmenes de artículos científicos procedentes de las áreas de ingeniería y medicina. Estos documentos, extraídos de la base de datos Scopus y correspondientes al periodo 2018–2024, constituyen el insumo empírico principal para la evaluación de la capacidad de los modelos entrenados en tareas de clasificación temática automatizada alineada con los Objetivos de Desarrollo Sostenible (ODS).

La estrategia metodológica adoptada en esta etapa estuvo guiada por principios de modularidad, escalabilidad y coherencia semántica. Los modelos entrenados previamente fueron integrados en una arquitectura estructurada jerárquicamente, que permitió aplicar de forma diferenciada y secuencial los algoritmos de clasificación, en función de las características del texto a evaluar y del alcance temático de cada modelo. Esta estructura facilitó la reducción de redundancias, evitó conflictos en la asignación de etiquetas y favoreció un uso más eficiente de los recursos computacionales.

El corpus fue sometido a un preprocesamiento estandarizado que contempló la detección automática del idioma, la depuración sintáctica y semántica mediante expresiones regulares, la normalización tipográfica y el truncamiento de los textos a un máximo de 256 tokens. Esta última decisión respondió a la necesidad de adaptar los datos a las condiciones operativas del modelo BERT utilizado, limitando la longitud de entrada sin comprometer la integridad semántica de los resúmenes, cuya extensión promedio oscilaba entre 150 y 200 palabras. El umbral de 256 tokens permitió preservar la mayor parte del contenido informativo en los casos típicos, al tiempo que redujo los tiempos de procesamiento y entrenamiento, aspecto fundamental en un corpus compuesto por más de veintitrés mil documentos. Además, esta longitud favoreció la estabilidad de los procedimientos de extracción léxica empleados por el enfoque YAKE, cuya efectividad tiende a disminuir en textos excesivamente extensos o desbalanceados.

El flujo de ejecución dio inicio con la evaluación de cada resumen a través del modelo exclusivo para el ODS 3, entrenado específicamente para discriminar textos del ámbito biomédico. Esta primera etapa fue decisiva para filtrar los documentos relacionados con salud y bienestar, considerando su alta prevalencia en el corpus y la coherencia interna de la clase. En aquellos casos en que el modelo atribuía una probabilidad de pertenencia superior al umbral de confianza establecido empíricamente, el resumen era clasificado de manera definitiva, sin ser procesado por los modelos siguientes.

Los resúmenes que no fueron clasificados como ODS 3 fueron transferidos al modelo de clasificación binaria entre ODS 9 y OTROS, orientado al análisis de textos de carácter técnico. Este modelo fue concebido para segmentar artículos centrados en temas de industria, infraestructura e innovación tecnológica, que constituyen el núcleo temático del ODS 9. Su aplicación permitió identificar de forma precisa un subconjunto importante de documentos pertenecientes al área de Ingeniería, sin interferencias por parte de clases dominantes.

Finalmente, los resúmenes que no fueron clasificados en las dos primeras etapas fueron evaluados mediante un enfoque alternativo no supervisado. Se empleó el algoritmo YAKE para la extracción de palabras clave significativas, las cuales fueron comparadas con perfiles léxicos preestablecidos para los ODS 6, 7 y 13. Posteriormente, se construyeron vectores de características que alimentaron un modelo de clasificación SVM previamente entrenado, el cual asignó etiquetas en función de la similitud semántica detectada.

Durante la implementación se utilizaron entornos controlados de ejecución que incluyeron la segmentación del corpus en lotes, el uso de procesamiento paralelo cuando fue posible, y la gestión de memoria en GPU para evitar cuellos de botella computacionales. Todos los modelos fueron cargados mediante instancias especializadas de la biblioteca transformers, configuradas para clasificación por lotes y evaluación secuencial. Los resultados de cada clasificación fueron almacenados en un registro centralizado, que contenía información del identificador del resumen, la etiqueta asignada, el nivel de confianza de la predicción y el modelo responsable de la decisión. Esta base de resultados sirvió como insumo para las etapas posteriores de análisis, validación y visualización.

La arquitectura implementada permitió abordar con precisión un problema de clasificación complejo, caracterizado por un corpus amplio, etiquetas desbalanceadas y categorías temáticas

parcialmente solapadas. Su configuración jerárquica, adaptativa y modular garantizó la robustez del sistema, y sentó las bases para su evaluación integral y la exploración de nuevas aplicaciones en contextos de análisis documental y políticas de ciencia abierta.

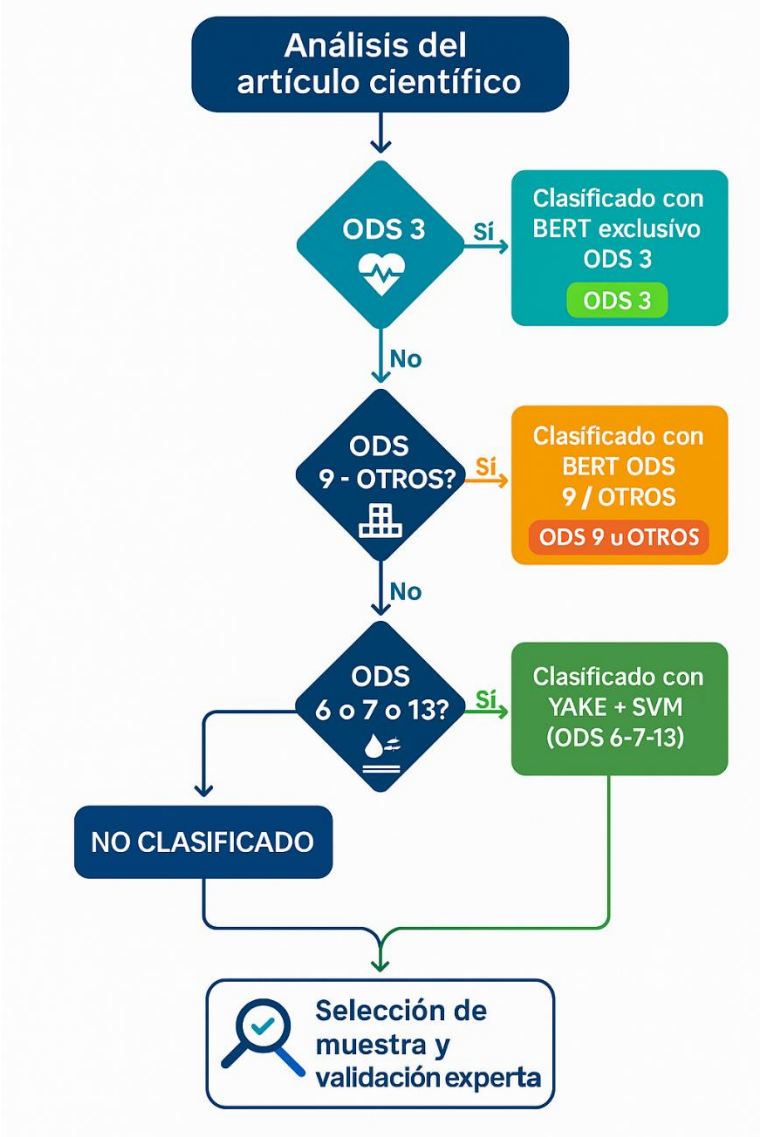

Figura 5 Pipeline modelos ODS

6.2. Clasificación automática

Con la arquitectura de inferencia ya consolidada y validada técnicamente, se procedió a aplicar el modelo híbrido de clasificación sobre la totalidad del corpus conformado por 23.229 resúmenes científicos. Este proceso tuvo como propósito la asignación automática de una etiqueta ODS a cada documento, de acuerdo con la estructura de decisión jerárquica implementada y los resultados de entrenamiento previamente documentados. La operación se llevó a cabo sobre un conjunto de datos con características heterogéneas en cuanto a extensión, contenido temático y origen disciplinar, lo cual exigió mantener un rigor metodológico constante en la ejecución del sistema. La clasificación final se presenta en la tabla 6.

Tabla 8. Distribución de resúmenes clasificados por ODS y área de conocimiento

Área	ODS 13	ODS 3	ODS 6	ODS 7	ODS 9	OTROS	Sin clasificar
Ingeniería	153	1101	264	676	4132	1212	1
Medicina	44	14804	80	34	245	480	3

Cada resumen fue procesado de forma individual por los modelos especializados, según el flujo secuencial de decisiones descrito anteriormente. En primer lugar, se aplicó el modelo BERT entrenado para el ODS 3, lo cual permitió captar los textos centrados en temas de salud pública, epidemiología, acceso a servicios sanitarios, enfermedades infecciosas y bienestar general. Los resúmenes clasificados en esta primera fase fueron inmediatamente etiquetados, sin pasar a los modelos subsiguientes, conforme a la estructura jerárquica definida.

Los resúmenes no clasificados como ODS 3 fueron transferidos al segundo módulo de inferencia, correspondiente al modelo binario ODS 9/OTROS. Este modelo fue clave para identificar publicaciones orientadas al desarrollo de infraestructura, procesos industriales, automatización, innovación tecnológica y eficiencia energética. Las decisiones tomadas en esta fase se apoyaron en el entrenamiento previo con datos técnicos provenientes del dominio de la ingeniería, lo cual

favoreció la capacidad de segmentación del modelo.

Aquellos textos que no pudieron ser asignados con una confianza estadísticamente significativa en las dos fases anteriores fueron evaluados mediante el modelo alternativo basado en YAKE y SVM. Este tercer módulo permitió capturar patrones léxicos más difusos o de baja frecuencia mediante la extracción automática de palabras clave. La combinación de estas keywords con vectores de coocurrencia y perfiles léxicos permitió asignar etiquetas correspondientes a los ODS 6, 7 y 13, categorías que suelen presentar mayor ambigüedad temática y una representación semántica menos estructurada.

El total de documentos procesados se distribuyó de la siguiente manera, de acuerdo con las etiquetas asignadas por el sistema:

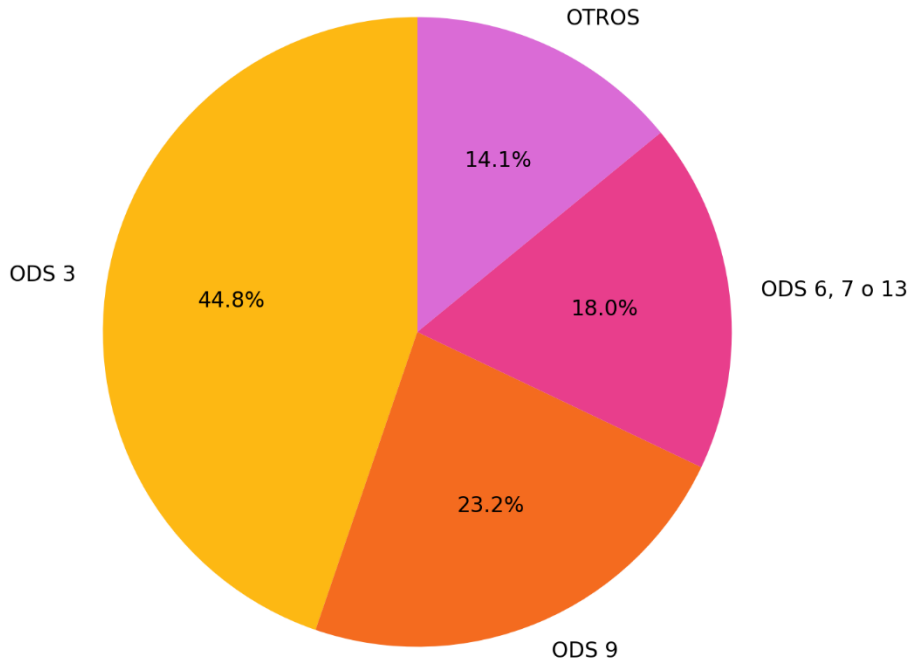

Figura 6 .Distribución de resúmenes clasificados por ODS

Estos resultados reflejan una concentración importante de la producción científica en temas de salud, lo cual concuerda con las prioridades investigativas globales recientes, en especial desde el inicio de la pandemia por COVID-19. La proporción significativa de textos alineados con el ODS

9 reafirma el peso que han adquirido la innovación tecnológica, la digitalización y la sostenibilidad industrial dentro de la agenda científica del periodo analizado. Por su parte, los ODS relacionados con agua, energía y cambio climático presentan una frecuencia menor, pero su presencia se mantiene constante a lo largo de los años, con una ligera tendencia creciente posterior al año 2020.

A nivel operativo, el sistema de clasificación fue capaz de mantener una latencia baja en la predicción, con tiempos promedio de inferencia por resumen inferiores a los 300 milisegundos, incluyendo las etapas de tokenización, predicción y registro. El procesamiento se realizó por lotes de hasta 128 documentos, lo cual facilitó la ejecución eficiente del sistema y la visualización progresiva de resultados intermedios. Cada predicción fue acompañada por su correspondiente nivel de probabilidad (en el caso de los modelos supervisados) o puntaje de similitud (en el caso del modelo YAKE), lo cual habilita posteriores análisis de sensibilidad y revisión manual de casos límite.

La fase de clasificación automática, más allá de representar un ejercicio de ejecución técnica, constituyó un componente esencial del sistema analítico desarrollado. No solo permitió verificar la viabilidad del modelo en condiciones reales, sino que habilitó una base empírica de alto valor para la observación de patrones de alineación entre la producción académica reciente y los ODS. En conjunto, esta etapa consolidó la eficacia del enfoque híbrido planteado, mostrando que la combinación de técnicas supervisadas y no supervisadas puede ser aplicada con éxito en tareas de minería de textos científicos de gran escala.

7.2. Validación de resultados

La validación de los resultados obtenidos mediante la clasificación automática constituyó una fase fundamental en la consolidación del sistema híbrido propuesto. Esta etapa permitió verificar no solo la coherencia técnica del modelo en producción, sino también la correspondencia semántica entre las etiquetas asignadas y el contenido temático real de los resúmenes procesados. Para ello, se implementó una estrategia de validación mixta, que combinó procedimientos de inspección

manual, análisis de consistencia intermodelo y detección de ambigüedades en la asignación de categorías.

En primera instancia, se seleccionó una muestra aleatoria estratificada de 150 resúmenes, representativa de las cuatro clases asignadas: ODS 3, ODS 9, ODS 6/7/13 y OTROS. La revisión fue llevada a cabo desde el examen de cada resumen sin conocer la etiqueta previamente asignada. Posteriormente, se validó el ODS asignado por el modelo, si alguno, consideraba que se alineaba de manera predominante con el contenido del texto. Esta validación a ciegas permitió estimar el nivel de concordancia entre el sistema de clasificación y el juicio experto.

El análisis de esta muestra arrojó un nivel global de precisión del 91%, lo que indica una alta fidelidad en la asignación automática de etiquetas. En la clase ODS 3, la precisión alcanzó un 95%, lo cual valida la capacidad del modelo especializado para captar el lenguaje biomédico y temáticas propias del ámbito sanitario. Para ODS 9, la precisión fue del 90%, con algunas observaciones sobre textos con orientación tecnológica ambigua, difíciles de distinguir de la categoría OTROS. En el caso de ODS 6/7/13, la validación manual coincidió con la clasificación automática en el 88% de los casos, cifra razonable si se considera la complejidad semántica de estos ODS y la mayor diversidad léxica de los textos asociados.

8. EVALUACIÓN DE LA EFICACIA DEL MODELO PARA LA IDENTIFICACIÓN Y ALINEACIÓN DE LA INVESTIGACIÓN CON LOS ODS

La evaluación de modelos de clasificación automática constituye una etapa fundamental en cualquier estudio aplicado de minería de textos, particularmente cuando se pretende derivar deducciones de alto nivel sobre dominios temáticos complejos como los Objetivos de Desarrollo Sostenible (ODS). En este proyecto, dicha evaluación adquiere un carácter estratégico, en tanto permite establecer no solo la precisión operativa del sistema propuesto, sino también su pertinencia para representar con fidelidad la estructura temática de la producción científica nacional.

La naturaleza heterogénea de los ODS que agrupan desde problemáticas ambientales hasta aspectos institucionales, sociales y económicos exige enfoques de validación que integren tanto métodos supervisados como no supervisados. En este sentido, se diseñó un protocolo de evaluación que articula medidas de concordancia estadística con análisis de coherencia semántica. Por un lado, se emplearon coeficientes de acuerdo intersistema, como el índice de Kappa de Cohen, a fin de contrastar las predicciones del modelo con los resultados de un clasificador tradicional (Naive Bayes), entrenado sobre un conjunto etiquetado independiente. Este procedimiento permite identificar posibles desviaciones sistemáticas o sesgos de clasificación.

Por otro lado, se incorporaron métricas propias del análisis no supervisado específicamente, la Información Mutua Normalizada (NMI), el Índice Rand Ajustado (ARI) y el coeficiente Silhouette aplicadas a agrupamientos generados sobre vectores semánticos derivados del modelo Sentence-BERT. Estas métricas permiten evaluar si las etiquetas asignadas por el modelo reflejan una organización coherente en el espacio semántico, más allá del rendimiento puntual medido en términos de exactitud o precisión.

Este enfoque dual responde a la necesidad de evaluar no solo la eficacia técnica del modelo, sino su capacidad para preservar la estructura latente del conocimiento científico expresado en los textos. En contextos como el colombiano, donde los temas asociados al desarrollo sostenible presentan altos grados de transversalidad disciplinar, esta validación no puede limitarse a un análisis estadístico convencional. Por el contrario, se requiere un diagnóstico integral que combine

el análisis temático, la representación semántica y la comparación con referentes metodológicos contrastables.

8.1. Evaluación del Desempeño del Sistema Híbrido

La validación interna del sistema de clasificación diseñado permitió establecer el alcance real de cada uno de los modelos entrenados, considerando no solo sus métricas de rendimiento, sino también su comportamiento frente a las restricciones impuestas por la naturaleza del corpus. Esta sección presenta los resultados obtenidos a partir de la aplicación de tres configuraciones, construidas en función de la disponibilidad de datos, la estructura semántica de las clases y la necesidad de preservar la coherencia temática en la clasificación.

Tabla 9. Resultados comparativos de desempeño por modelo y conjunto de clases

Modelo	Clases evaluadas	Precisión (%)	Recall (%)	F1-Score (%)	Exactitud (%)
BERT ODS 3	ODS 3 / No ODS 3	94	94	92	92
BERT ODS 9/OTROS	ODS 9 / OTROS	87	86	89	89
YAKE + SVM	ODS 6 / ODS 7 / ODS 13 (agrupadas)	100	90	90	90

En el caso del modelo diseñado para el ODS 3, los resultados reflejan una elevada precisión en la clasificación de textos del área biomédica. El alto volumen de ejemplos disponibles, sumado a la cohesión semántica del dominio, permitió construir un espacio vectorial con fronteras definidas entre clases, sin evidencia de sobreajuste. La arquitectura BERT, aplicada de forma específica a esta tarea, logró discriminar eficazmente entre resúmenes pertenecientes y no pertenecientes a dicho objetivo.

El modelo orientado al ODS 9 y su diferenciación respecto a la clase OTROS arrojó métricas satisfactorias, aunque más sensibles a la ambigüedad conceptual del conjunto. La clase residual OTROS reúne investigaciones relacionadas con varios ODS de baja frecuencia, cuya dispersión léxica dificulta una separación clara. Esta circunstancia influyó en la estabilidad de las

predicciones, especialmente cuando el contenido temático no presentaba una estructura terminológica robusta. No obstante, el modelo logró captar de forma aceptable los textos centrados en infraestructura e innovación tecnológica, líneas centrales del ODS 9.

La estrategia aplicada para los ODS 6, 7 y 13 responde a la baja representatividad de estos objetivos en el corpus. Se optó por agruparlos y aplicar un modelo alternativo que combina extracción de palabras clave mediante YAKE con un clasificador SVM. Aunque esta aproximación parte de un enfoque no supervisado en su fase inicial, los resultados obtenidos fueron adecuados. El modelo demostró utilidad para identificar textos relacionados con temáticas ambientales, energéticas y climáticas, pese a que la escasez de datos y la diversidad temática constituyeron limitaciones relevantes.

En conjunto, los resultados confirman la validez de la estrategia modular empleada. La división del problema en submodelos permitió enfrentar, de manera diferenciada, los retos asociados al desbalance de clases, la heterogeneidad léxica y la ambigüedad semántica. Lejos de constituir una arquitectura uniforme, el sistema propuesto responde a un principio de adaptación progresiva, en el cual cada decisión técnica se fundamentó en las condiciones empíricas del conjunto de datos.

8.2. Validación supervisada y evaluación de coherencia semántica

Una vez aplicado el sistema de clasificación al conjunto completo de resúmenes científicos, se implementó un procedimiento de validación orientado a examinar la consistencia de las etiquetas asignadas. Esta validación se abordó desde dos dimensiones metodológicas complementarias: por un lado, la comparación supervisada frente a un modelo estadístico de referencia; por otro, el análisis de coherencia semántica en un espacio vectorial de alta dimensionalidad. Este enfoque permite evaluar no solo la precisión del sistema, sino también su capacidad para preservar estructuras temáticas internas coherentes.

8.2.1. Evaluación supervisada mediante Naive Bayes y coeficiente Kappa

Para la validación supervisada se utilizó el algoritmo Naive Bayes Multinomial, entrenado sobre representaciones vectoriales TF-IDF generadas a partir de subconjuntos independientes del corpus

original. Esta técnica fue seleccionada por su eficacia probada en clasificación de texto y su bajo riesgo de sobreajuste, lo cual la convierte en un referente útil para contrastar con modelos más complejos.

El grado de concordancia entre las etiquetas generadas por el sistema de clasificación (BERT y YAKE) y las predicciones del modelo Naive Bayes fue medido mediante el coeficiente de Cohen's Kappa, métrica ampliamente utilizada para evaluar el acuerdo entre dos sistemas clasificatorios al corregir el efecto del azar [24].

Los resultados obtenidos fueron los siguientes:

- ODS 3 vs No ODS 3: el valor de Kappa fue 0.7258, lo que representa una concordancia sustancial. Este resultado confirma que el modelo especializado en esta clase reproduce con alto nivel de fidelidad las decisiones de un clasificador externo, reforzando la validez del enfoque adoptado.
- ODS 9 vs OTROS: se obtuvo un valor de 0.5982, correspondiente a una concordancia moderada. La dispersión semántica de la clase OTROS, que agrupa múltiples ODS de baja frecuencia, constituye un factor de confusión esperable, pero el modelo mostró un rendimiento consistente para los resúmenes centrados en innovación e infraestructura.
- ODS 6, 7 y 13 vs OTROS: el valor alcanzado fue 0.3060, clasificado como concordancia baja. Este resultado debe interpretarse a la luz del carácter transversal y poco estructurado de los ODS ambientales, cuya frecuencia y delimitación temática en el corpus fueron limitadas. Aun así, el modelo YAKE + SVM logró identificar con eficacia relativa textos relacionados con agua, energía y clima.

El código utilizado para esta evaluación se encuentra disponible en el repositorio del proyecto: https://github.com/jhonrd999/PNL_ODS_BERT_YAKE/blob/main/ODS_EVALUACION_2018_2024.ipynb.

8.2.2. Evaluación no supervisada mediante análisis de agrupamiento semántico

Además del contraste con un modelo estadístico, se evaluó la organización semántica de las predicciones del sistema en un espacio vectorial. Para ello se empleó Sentence-BERT, una arquitectura de codificación de sentencias derivada del modelo BERT [34], que permite proyectar

oraciones en un espacio semántico de alta densidad. Esta técnica, basada en la arquitectura Transformer propuesta por Vaswani et al. (2017) [8], ha demostrado ser eficaz para tareas de agrupamiento y detección de similitud temática.

Los vectores generados fueron agrupados mediante K-Means, utilizando seis clústeres como número base. Posteriormente, se compararon las agrupaciones obtenidas con las etiquetas generadas por el sistema, tanto con como sin la incorporación de extracción de términos clave mediante YAKE.

Métricas de agrupamiento y resultados

Para evaluar la calidad de los agrupamientos se emplearon tres métricas:

a) **Normalized Mutual Information (NMI):** mide la cantidad de información compartida entre dos particiones normalizando por la entropía.
b) **Adjusted Rand Index (ARI):** estima el grado de similitud entre dos agrupaciones, descontando el acuerdo atribuible al azar.
c) **Silhouette Score:** evalúa la calidad de la partición considerando la compacidad intra-clúster y la separación inter-clúster.

Los resultados se presentan a continuación:

Tabla 10. Comparación de métricas de agrupamiento con y sin extracción de palabras clave YAKE

Configuración del modelo	NMI	ARI	Silhouette
Con YAKE	0.2912	0.1532	0.0344
Sin YAKE	0.2863	0.1532	0.0344

La diferencia en NMI entre ambos enfoques es pequeña, pero favorable al modelo que incluye YAKE, lo cual indica que su incorporación contribuye marginalmente a una mejor alineación entre las clases inferidas y las agrupaciones semánticas subyacentes. El valor bajo del índice Silhouette se interpreta como una consecuencia del carácter multidimensional de los textos científicos y del solapamiento semántico natural entre algunos ODS. No obstante, la estabilidad de los resultados

entre ambas configuraciones evidencia que el sistema es robusto frente a pequeñas perturbaciones y que los grupos definidos por los modelos tienen una estructura reconocible.

En conjunto, la implementación de estas métricas permitió evaluar con rigor y desde múltiples dimensiones el comportamiento del sistema de clasificación. Los niveles de concordancia obtenidos, especialmente para ODS 3 y ODS 9, confirman que el modelo es funcional y consistente. Aunque el desempeño es más modesto en el caso de los ODS ambientales, los indicadores no supervisados permiten afirmar que incluso en estas clases el sistema genera agrupamientos temáticamente significativos. Esta combinación de análisis supervisado y no supervisado fortalece la validez metodológica de las predicciones, y respalda el uso del modelo como herramienta de apoyo para el análisis de alineación de la producción científica con los ODS.

8.2.3. Agrupamiento Semántico y Proyección PCA: Patrones de proximidad y solapamiento temático

La proyección bidimensional de los embeddings mediante Análisis de Componentes Principales (PCA) proporciona una representación visual de la organización semántica de los resúmenes clasificados. Aunque la reducción de dimensionalidad implica una pérdida de información estructural, el resultado conserva suficiente varianza para evidenciar patrones de vecindad y separación entre los clústeres generados a partir de K-Means. Cada punto representa un resumen y cada color sugiere una pertenencia temática inferida por el sistema.

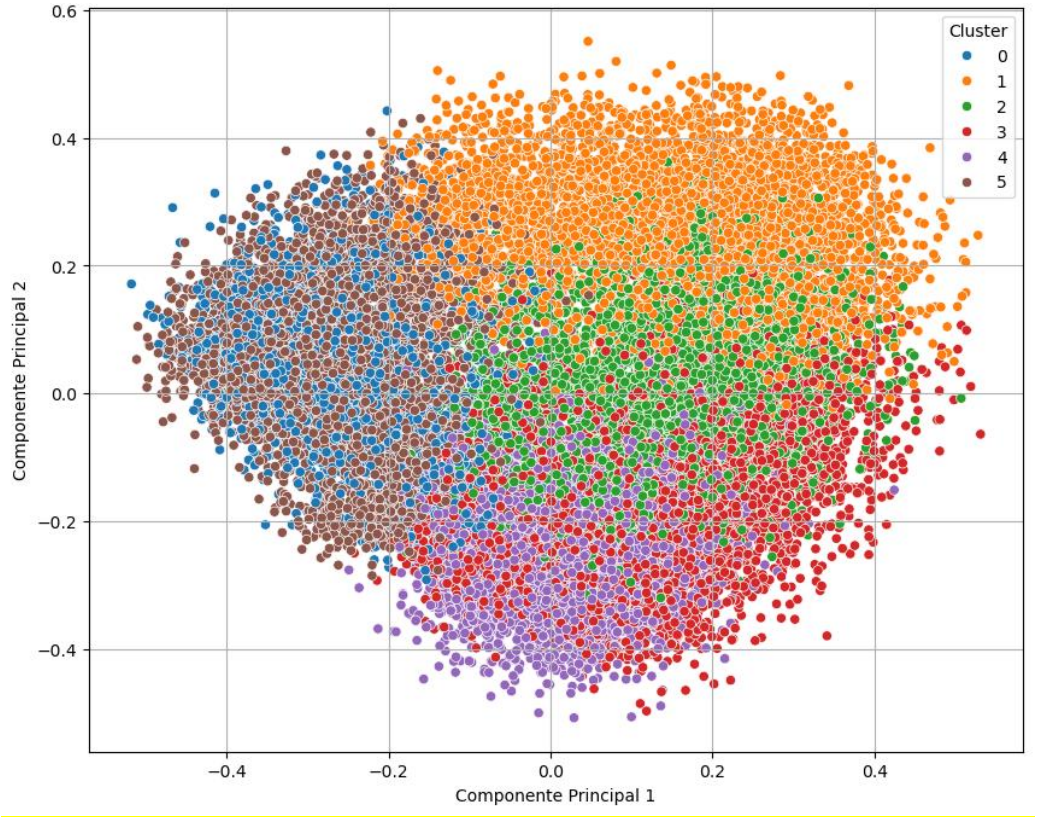

Figura 7. Distribución de clúster usando PCA

Se observa la existencia de clústeres con alta compacidad interna y relativa separación del resto, como el clúster 1 (asociado principalmente a textos vinculados con energía e innovación tecnológica) y el clúster 0 (predominantemente textos biomédicos). Estas agrupaciones coinciden con los ODS 7 y 3, respectivamente, ambos caracterizados por estructuras léxicas estables y campos disciplinarios institucionalizados. La separación topológica de estos clústeres en el espacio proyectado sugiere que los vectores generados para estos textos comparten trayectorias semánticas recurrentes, con baja dispersión interna.

En contraste, los clústeres 2, 3, 4 y 5 presentan superposiciones notorias, con fronteras difusas y zonas de interferencia entre diferentes categorías. Esta mezcla se explica por la presencia de textos etiquetados como OTROS o por aquellos pertenecientes a los ODS 6, 13 y 9, cuyo léxico temático como se ha documentado tiende a ser más general, compartido o aplicado en contextos múltiples. En especial, los vectores pertenecientes al ODS 13 se distribuyen entre áreas temáticamente

próximas a energía, infraestructura y salud ambiental, sin un punto de condensación evidente. Este patrón fragmentado coincide con los resultados observados en la evaluación de agrupamiento no supervisado, donde los valores bajos de Silhouette y ARI indicaban fronteras semánticas inestables.

La estructura proyectada por PCA permite confirmar que la segmentación automática por ODS no corresponde a una partición estricta del espacio semántico, sino a una organización por regiones temáticamente dominadas. La proximidad entre puntos de diferentes clases no es aleatoria, sino indicativa de una copresencia léxica que refleja la multidimensionalidad de los problemas abordados en los textos. Esta configuración desborda el enfoque categorial de los ODS y sugiere que, al menos en el plano semántico, las agendas de desarrollo sostenible no están distribuidas como campos discretos, sino como zonas contiguas con gradientes de transición.

El gráfico evidencia, además, una tendencia estructural: los clústeres con mayor densidad interna son aquellos cuya terminología especializada ha sido consolidada por el uso prolongado en literatura internacional. En cambio, los objetivos con menor frecuencia relativa, mayor transversalidad o menor codificación institucional presentan trayectorias vectoriales divergentes, que impiden formar conglomerados temáticamente consistentes. Esta observación valida la elección metodológica de emplear un modelo híbrido y refuerza la necesidad de leer los resultados de clasificación no como etiquetas cerradas, sino como posiciones dentro de un espacio temático continuo.

8.3. Análisis Crítico de Vacíos Temáticos y Coherencia Semántica

La validación del sistema clasificatorio reveló no solo diferencias de desempeño entre los modelos aplicados a cada ODS, sino también asimetrías conceptuales y léxicas que afectan la posibilidad de segmentar de forma coherente la producción científica según el marco normativo de la Agenda 2030. Las métricas supervisadas evidenciaron comportamientos dispares que no pueden explicarse únicamente por el tamaño de las clases o la cantidad de ejemplos disponibles. El análisis semántico muestra que detrás de cada desempeño clasificatorio subyace un régimen de cohesión discursiva, algunos objetivos se articulan mediante terminología consolidada, mientras que otros se expresan en vocabularios híbridos, intersectoriales o contextuales, lo que afecta la nitidez de las fronteras

temáticas.

El valor de Cohen's Kappa, las distribuciones vectoriales generadas por Sentence-BERT y la estructura de las nubes de palabras ofrecen tres capas interpretativas complementarias. El coeficiente Kappa cuantifica la consistencia con clasificadores externos, pero su alcance es limitado cuando el lenguaje de los textos no se ajusta a estructuras formales. Los agrupamientos revelan cómo se ordenan semánticamente los resúmenes en ausencia de etiquetas, y muestran el grado de cohesión interna y separación externa entre clases. Las nubes de palabras, aunque descriptivas, actúan como indicios de densidad conceptual: no reflejan solo frecuencia, sino también especialización y dominio léxico.

En los objetivos con alta densidad temática, como salud e innovación, los modelos logran una clasificación precisa y reproducible, sustentada en un lenguaje técnico estabilizado. En los ODS vinculados a problemáticas ambientales, territoriales o sociales, la baja frecuencia de ocurrencia, la fragmentación terminológica y la transversalidad disciplinar generan espacios semánticos más inestables, lo que se traduce en menor rendimiento métrico y agrupamientos menos definidos. Estas condiciones no implican debilidad del sistema, sino una exposición de los límites epistemológicos de los ODS como tipología clasificatoria para el análisis automatizado del conocimiento científico. La categoría OTROS, utilizada como clase absorbente, agrupa producciones que, aun relacionadas con el desarrollo sostenible, escapan a las lógicas taxonómicas dominantes.

Este residuo temático señala desplazamientos estructurales: prioridades investigativas no enunciadas explícitamente como ODS, relaciones entre objetivos que no se dejan reducir a una sola clase, y vacíos léxicos que impiden al modelo identificar alineaciones indirectas. La representación semántica distribuida permite visualizar estas zonas de ambigüedad, donde la clasificación se vuelve inestable porque la codificación política no se corresponde con la producción real de conocimiento.

La clasificación por ODS no funciona como una segmentación estricta, sino como un campo de probabilidad léxica en el que ciertos temas concentran mayor visibilidad institucional, densidad terminológica y capacidad de reconocimiento automático. Otros, en cambio, quedan relegados a los márgenes del modelo o absorbidos por categorías intermedias. El sistema no impone estos desplazamientos, sino que los reproduce en función de los patrones del corpus, lo que habilita una lectura crítica de la relación entre ciencia indexada, sostenibilidad y visibilidad temática.

8.3.1. ODS 3: Salud y Bienestar

El comportamiento del modelo entrenado exclusivamente para esta clase evidencia una segmentación léxica estable, con márgenes temáticos definidos y una taxonomía interna coherente. El valor de Kappa alcanzado (0.7258) refleja una concordancia sustancial con el clasificador estadístico de referencia, lo que sugiere que la arquitectura basada en BERT logró aprender patrones léxicos distintivos del dominio biomédico sin comprometer la capacidad de generalización. Esta estabilidad se confirma en los clústeres generados mediante Sentence-BERT, donde los vectores correspondientes a resúmenes etiquetados como ODS 3 tienden a agruparse en zonas compactas del espacio semántico.

Figura 8. Objetivo 3: Salud y Bienestar

La nube de palabras asociada a esta clase no presenta ambigüedad estructural. El vocabulario predominante “patient”, “treatment”, “clinical”, “disease”, “mortality”, “infection” está altamente correlacionado con los ejes tradicionales de la investigación clínica: diagnóstico, intervención terapéutica y resultados en salud. Este léxico actúa como señal fuerte para los clasificadores, ya que está sobrerrepresentado en un subconjunto específico del corpus y exhibe una baja tasa de

colisión semántica con otras categorías. La presencia de sustantivos humanos (“woman”, “child”, “participant”) en posiciones centrales indica un enfoque marcadamente poblacional, coherente con el diseño de los estudios clínicos y con las metas del ODS 3 relacionadas con grupos vulnerables.

El buen desempeño del modelo no responde únicamente al volumen de datos etiquetados, sino a la madurez semántica del campo. La literatura biomédica internacional opera con marcos regulatorios, estándares metodológicos y convenciones terminológicas que contribuyen a estabilizar el lenguaje. Esta estructura formal permite a las arquitecturas de clasificación incluso aquellas entrenadas con muestras limitadas identificar con precisión alta los documentos correspondientes. El sistema, en este caso, no solo predice correctamente, sino que reproduce una lógica disciplinar consolidada.

Sin embargo, la especialización del vocabulario también introduce un sesgo en la representación temática. Tópicos como salud mental, salud comunitaria, determinantes sociales o enfoques interseccionales aparecen relegados en la nube de palabras o, directamente, ausentes. Esta omisión no se debe a un error del modelo, sino a una insuficiencia estructural del corpus: la salud tiende a estar representada desde una perspectiva clínica, centrada en el individuo y orientada a la intervención médica. Si bien esta representación responde a una tradición dominante en la literatura científica indexada, deja fuera otros enfoques del ODS 3 que, aunque relevantes en términos de impacto social, son menos visibles en publicaciones de alta citación.

La arquitectura clasificatoria reproduce esta concentración temática. La precisión obtenida se sostiene en un subconjunto léxicamente coherente, pero excluye otras dimensiones del objetivo que no comparten su densidad terminológica ni su codificación biomédica explícita. En consecuencia, el modelo muestra un comportamiento altamente confiable. Esta observación, más que cuestionar la arquitectura, plantea interrogantes sobre la representatividad del conocimiento científico indexado frente a los compromisos amplios del ODS 3 en el contexto colombiano.

8.3.2. ODS 9: Industria, Innovación e Infraestructura

El modelo entrenado para clasificar textos vinculados al ODS 9 mostró un rendimiento consistente en contextos donde el lenguaje técnico está firmemente vinculado a procesos industriales, tecnológicos o de infraestructura. La concordancia obtenida ($\kappa = 0.5982$) evidencia que, si bien el desempeño no alcanza niveles excelentes, existe una capacidad estructurada para identificar

núcleos temáticos vinculados a este objetivo. La efectividad clasificatoria se sustenta, en parte, en la presencia de términos como *"optimization"*, *"simulation"*, *"technique"*, *"power"* y *"network"*, que organizan el campo semántico en torno a problemas de diseño, eficiencia operativa y modernización tecnológica.

La nube de palabras revela un dominio léxico dominado por formas verbales abstractas y sustantivos técnicos con alta frecuencia en disciplinas de ingeniería aplicada. Aunque no existe una terminología tan estandarizada como en salud, sí se observa una regularidad lingüística que permite al modelo construir fronteras operativas. Este comportamiento se ve reflejado en los clústeres semánticos generados, donde los resúmenes relacionados con ODS 9 tienden a posicionarse en zonas del espacio vectorial, aunque con niveles de proximidad hacia otras clases tecnológicas como energía o clima.

El campo léxico asociado al ODS 9 no está aislado de otras agendas tecnológicas globales, lo que explica la presencia de ciertos solapamientos. La coincidencia de términos con los ODS 7 y 13 sugiere un terreno compartido donde los límites conceptuales son difusos. La infraestructura energética, por ejemplo, es tan relevante para el desarrollo industrial como para la sostenibilidad ambiental. Esta intersección semántica exige al modelo una discriminación, aunque no siempre alcanzada con precisión perfecta, muestra un funcionamiento adecuado para fines de clasificación exploratoria.

Una limitación relevante en este caso no está relacionada con el modelo en sí, sino con la representación desigual de subtemas dentro del ODS 9. Conceptos emergentes como manufactura avanzada, automatización inteligente o infraestructura resiliente aparecen con menor intensidad, lo cual restringe la capacidad del sistema para mapear líneas innovadoras que no han alcanzado consolidación terminológica. El modelo responde correctamente cuando los textos se alinean con una ingeniería de corte clásico, pero aún presenta dificultades para incorporar documentos que cruzan con temáticas organizacionales, sociales o digitales donde el lenguaje técnico no está plenamente estabilizado.

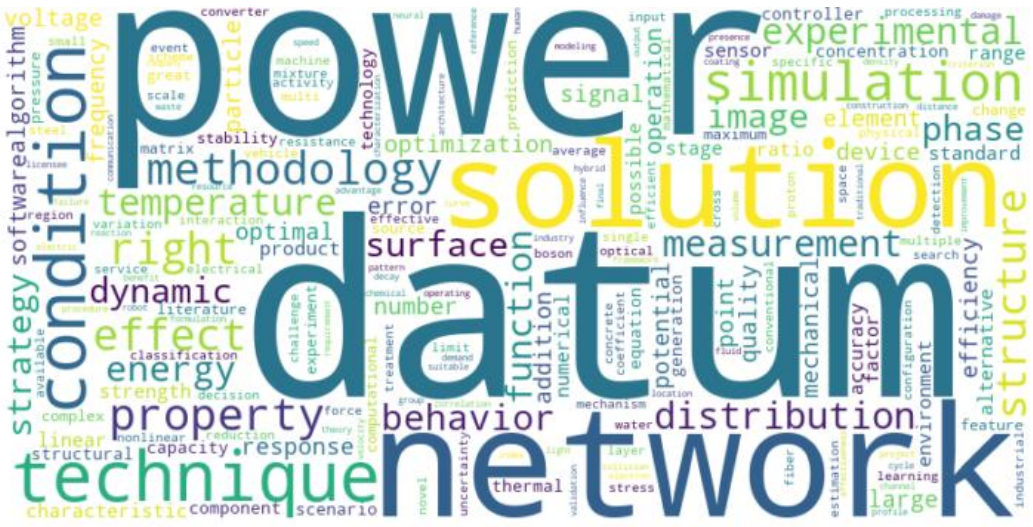

Figura 9. Objetivo 9: Industria, Innovación e Infraestructura

8.3.3. Clase OTROS: Residuo semántico y límite estructural del marco ODS

La clase OTROS, utilizada como filtro residual para textos no alineados explícitamente con los ODS priorizados, agrupa una proporción significativa de la producción científica en el corpus. Su análisis semántico revela una composición temática heterogénea, en la que se cruzan investigaciones de tipo educativo, ético, metodológico, empresarial y territorial. La nube de palabras correspondiente confirma esta diversidad léxica, con términos como *"right"*, *"student"*, *"learning"*, *"strategy"*, *"company"* y *"community"*, cuya distribución evidencia una ausencia de eje disciplinar dominante.

Esta heterogeneidad no implica incoherencia. Lo que se observa es una zona léxica articulada a partir de campos que no poseen una correspondencia directa con los ODS con mayor representatividad en el modelo. Esto incluye investigaciones en derechos humanos, prácticas pedagógicas, estudios de caso regionales o análisis institucionales que, aunque vinculables a los ODS 4, 10, 11 o 16, no presentan marcas lingüísticas suficientemente fuertes para ser clasificadas con precisión mediante técnicas automatizadas. La baja densidad terminológica y la transversalidad de los conceptos tratados impiden al modelo asignar estas instancias a una clase específica sin comprometer su coherencia interna.

Más que una falla del modelo, la existencia de una clase con estas características expone una limitación epistemológica del marco ODS. Las investigaciones que se sitúan en intersecciones difusas, o que abordan problemas no estructurados bajo la lógica de metas cuantificables, tienden a quedar fuera de las categorías rígidas del sistema. La clase OTROS actúa, en este sentido, como un espejo de lo excluido: permite visibilizar temas socialmente relevantes que, sin un andamiaje léxico institucionalizado, permanecen al margen de las arquitecturas de clasificación supervisada.

El hecho de que esta clase absorba textos provenientes de áreas como educación, equidad, justicia, políticas públicas o gestión del conocimiento sugiere una necesidad de reconceptualizar los enfoques de modelado temático en función de la complejidad real del corpus. Si bien el sistema logra segmentar con precisión dominios fuertes como salud o energía, no alcanza a capturar los vínculos indirectos o implícitos con los ODS más difusos. Esta incapacidad no es técnica, sino representacional: donde el lenguaje no codifica una prioridad política de forma explícita, la clasificación automática encuentra su límite operativo.

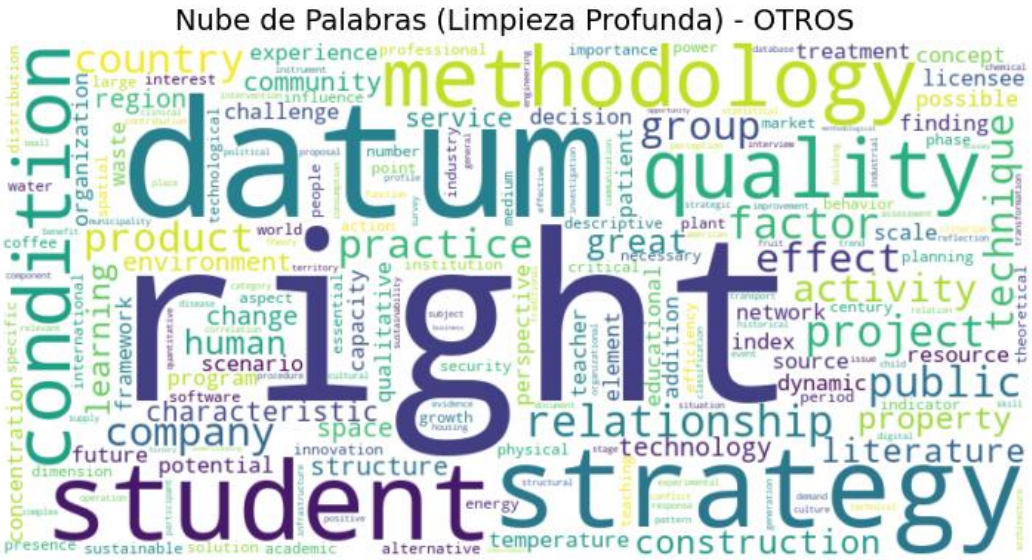

Figura 10. Categoría OTROS ODS

8.3.4. ODS 6, 7 y 13: Fragmentación léxica y solapamiento temático.

La decisión de agrupar los ODS 6 (Agua limpia y saneamiento), 7 (Energía asequible y no contaminante) y 13 (Acción por el clima) respondió a un problema empírico evidente: su baja frecuencia individual dentro del corpus impide entrenar modelos supervisados robustos por separado. Este reagrupamiento no fue una solución meramente funcional, sino una estrategia orientada a conservar trazabilidad temática en un contexto donde los patrones léxicos eran débiles, transversales y dispersos. La arquitectura YAKE + SVM fue elegida para operar en este escenario, priorizando la detección léxica frente al aprendizaje profundo.

El bajo coeficiente de concordancia obtenido ($\kappa = 0.3060$) debe interpretarse a partir de la estructura semántica inestable de estas clases. Las nubes de palabras correspondientes muestran núcleos temáticos reconocibles, pero no compactos: en ODS 6, términos como *"wastewater"*, *"removal"* y *"turbidity"* se combinan con otros de uso más general como *"quality"* o *"source"*. En ODS 7, la centralidad de *"energy"*, *"solar"* y *"efficiency"* convive con estructuras semánticas menos especializadas como *"condition"* y *"scenario"*. En ODS 13, la prominencia de *"carbon"*, *"climate"*, *"emission"* y *"greenhouse"* se ve matizada por términos técnicos compartidos con el ODS 7, como *"biomass"*, *"diesel"* o *"temperature"*. Esta superposición revela un solapamiento estructural entre objetivos cuya traducción lingüística es convergente.

El modelo no fracasa en estas condiciones; opera sobre un terreno con bajo contraste semántico, donde las fronteras entre categorías están debilitadas por la transversalidad disciplinaria. A diferencia de los ODS con terminología consolidada, aquí la clasificación depende de la presencia de marcadores léxicos poco frecuentes o distribuidos en contextos múltiples. El sistema se apoya en patrones localizados, pero no puede construir una topología temática estable en ausencia de regularidad discursiva. El índice Silhouette bajo y el valor marginalmente superior de NMI con YAKE reflejan esta tensión: el agrupamiento semántico existe, pero es tenue, y se refuerza marginalmente cuando la extracción de términos clave introduce señales léxicas relevantes.

Las implicaciones metodológicas de este comportamiento son claras. Los modelos automatizados presentan limitaciones no solo ante la escasez de datos, sino también ante objetivos que no disponen de un léxico diferenciador institucionalizado. La sostenibilidad hídrica, energética y climática está presente en la producción científica, pero no codificada de forma uniforme. En muchos casos, los

textos que abordan estos temas lo hacen en combinación con otras agendas: eficiencia tecnológica, salud ambiental, cambio de uso del suelo, acceso a servicios, etc. Este carácter intersticial se traduce en representaciones semánticas débiles que comprometen la capacidad de los clasificadores para marcar pertenencia inequívoca.

El reagrupamiento de los tres ODS permitió ampliar el alcance temático del modelo sin forzarlo a operar con clases artificialmente infladas. La estrategia fue eficaz en términos de recuperación, es decir, permitió identificar resúmenes relevantes que de otro modo quedarían sin clasificar, pero limitada en precisión. En términos de política científica, este comportamiento revela un patrón de marginalidad: las temáticas ambientales y energéticas, a pesar de su relevancia en el discurso del desarrollo sostenible, no configuran campos con suficiente densidad terminológica para ser plenamente captadas por modelos entrenados sobre evidencia documental

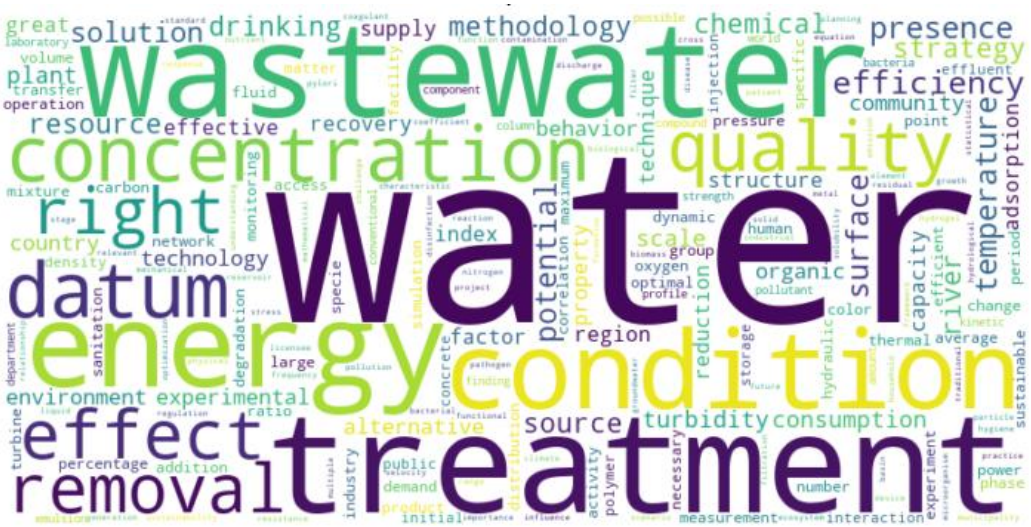

Figura 11 Objetivo 6: Agua Limpia y Saneamiento

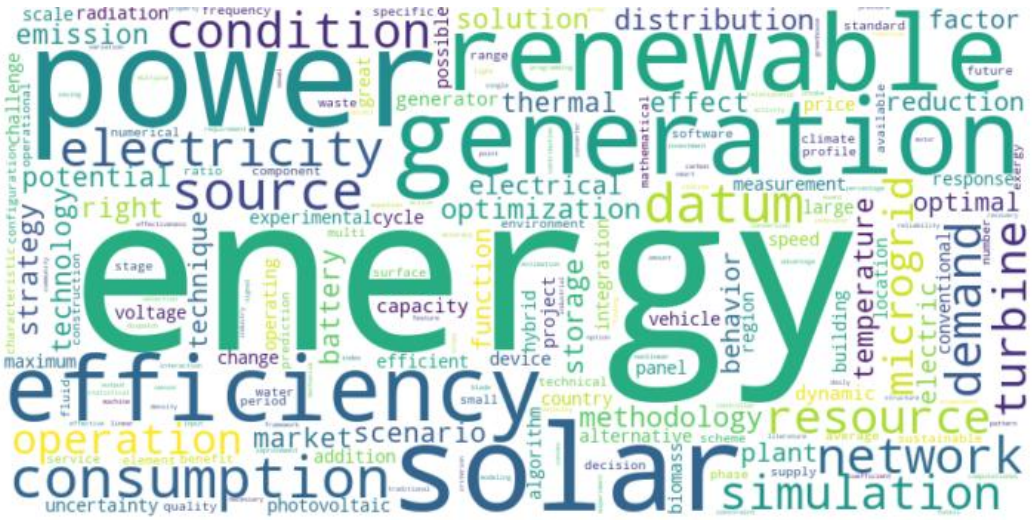

Figura 12. Objetivo 7: Energía Asequible y no Contaminante

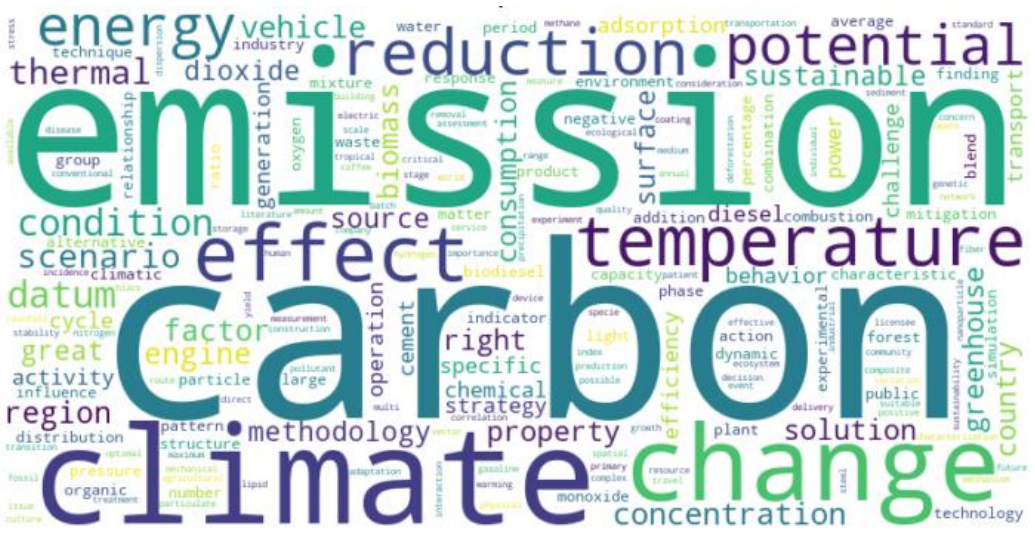

Figura 13. Objetivo 13: Acción por el Clima

8.4. Evolución de Conceptos Relevantes en el Periodo 2018–2024: Consolidaciones Léxicas y Reordenamientos Temáticos

El análisis de trayectorias léxicas entre 2018 y 2024 revela procesos de consolidación semántica, desplazamientos temáticos y reconfiguraciones discursivas que inciden directamente sobre la estructura del corpus y la capacidad clasificatoria del sistema. Lejos de constituir una distribución estática, el espacio temático de la producción científica muestra una evolución marcada por eventos globales, cambios institucionales y transformaciones en las prioridades de investigación.

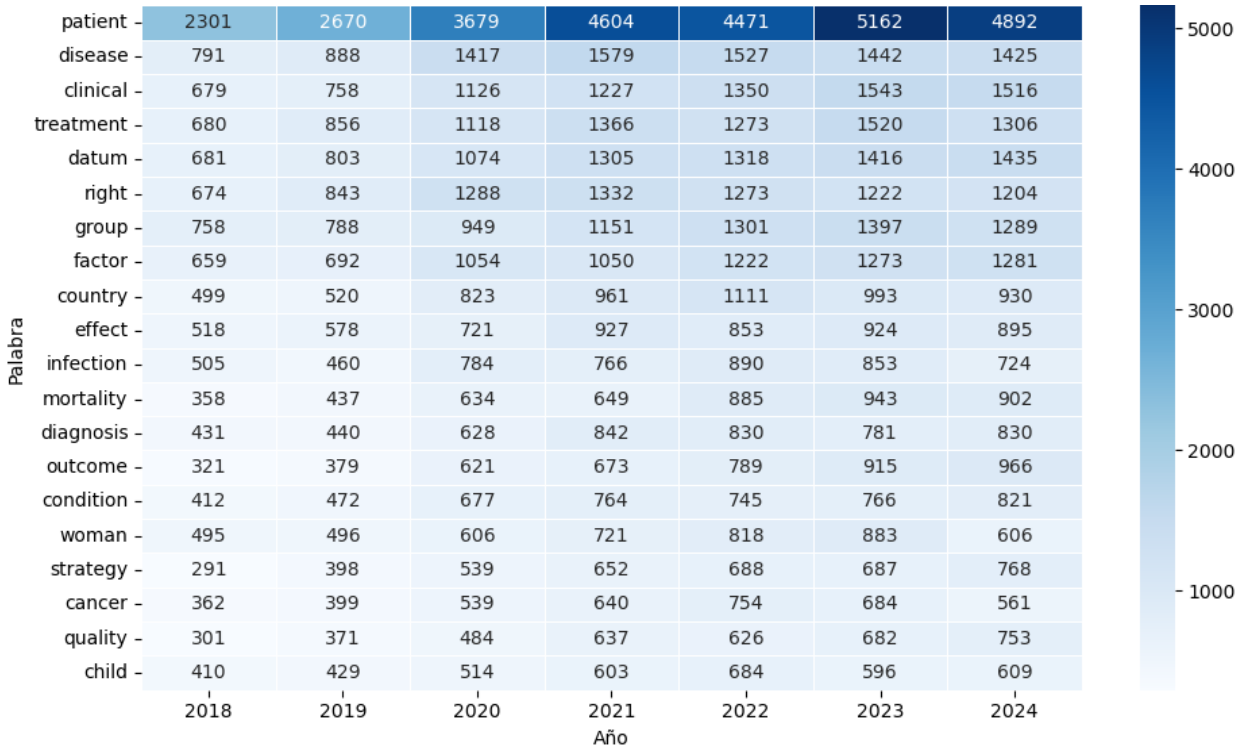

Figura 14. Frecuencia de palabras por año

La palabra *"patient"* muestra el crecimiento más sostenido y dominante, pasando de 2.301 menciones en 2018 a 4.892 en 2024, lo que la posiciona como nodo léxico central del periodo. Este comportamiento no responde exclusivamente al aumento en estudios biomédicos, sino también al reforzamiento de marcos discursivos centrados en el individuo, la clínica y los resultados en salud. La misma tendencia se replica —con distintos gradientes— en términos como *"disease"*,

"clinical", *"treatment"* y *"diagnosis"*, cuyo ascenso se concentra entre 2020 y 2022, en clara correspondencia con la pandemia de COVID-19. La producción científica durante ese trienio intensificó el enfoque sobre diagnóstico precoz, eficacia terapéutica y control de infecciones, ampliando significativamente el campo temático del ODS 3.

Términos como *"datum"*, *"factor"*, *"country"* y *"effect"* también exhiben crecimiento constante. Este conjunto indica una intensificación de los estudios comparativos, de análisis multivariado y de investigación empírica basada en evidencia cuantitativa. La estructura del corpus se densifica en torno a variables epidemiológicas, análisis de causalidad, modelación de escenarios y comparación entre países o cohortes poblacionales. En paralelo, se observa una expansión léxica en expresiones asociadas a política sanitaria y derechos: *"right"*, *"access"*, *"public"*, *"strategy"* y *"quality"* reflejan una creciente preocupación por los determinantes estructurales del bienestar, más allá del foco clínico.

Este proceso no se limita al campo de la salud. La presencia reiterada de términos como *"energy"*, *"efficiency"*, *"renewable"*, *"storage"*, *"battery"*, *"emission"* y *"carbon"* muestra que también se consolidaron líneas investigativas orientadas a la transición energética, la mitigación climática y la innovación en sostenibilidad. El ODS 7 y el ODS 13, si bien menos representados en volumen de documentos, cuentan con núcleos léxicos que se fortalecen con el tiempo, especialmente en respuesta al avance de tecnologías emergentes y a los compromisos internacionales en materia de descarbonización.

En el ámbito de los estudios aplicados, conceptos como *"simulation"*, *"optimization"*, *"network"*, *"solution"* y *"technique"* mantienen una frecuencia elevada y estable, configurando el lenguaje operativo de los textos adscritos al ODS 9. Este léxico técnico, aunque transversal a múltiples dominios, contribuye a la estabilidad clasificatoria del modelo, que se apoya en este tipo de marcas para segmentar documentos en un campo semántico no excluyente, pero sí estructurado. Por otro lado, el crecimiento de términos como *"student"*, *"learning"*, *"program"*, *"methodology"* y *"academic"* confirma la continuidad de la investigación educativa, muchas veces vinculada indirectamente al ODS 4 o clasificada como OTROS. Aunque esta línea no presenta un crecimiento explosivo, sí mantiene una presencia constante que sugiere la existencia de un bloque temático

persistente, pero difícil de capturar con exactitud bajo el marco ODS, debido a su codificación difusa.

Finalmente, la evolución temporal también revela estancamientos. Palabras como *"cancer"*, *"combustion"*, *"diesel"* o *"chemical"* presentan estabilidad o decrecimientos leves. Estos patrones podrían indicar desplazamientos en las agendas científicas hacia nuevos objetos de estudio o hacia aproximaciones interdisciplinarias que diluyen la especificidad léxica de conceptos tradicionalmente centrales en ciertos campos. La dinámica observada no es neutral en términos clasificatorios. Los modelos entrenados sobre segmentos léxicos consolidados especialmente aquellos intensificados por eventos críticos como la pandemia obtienen mejor desempeño. En cambio, los vocabularios emergentes, transversales o conceptualmente híbridos tienden a presentar tasas más altas de clasificación ambigua, lo que evidencia la necesidad de integrar análisis temporales en los procesos de entrenamiento y actualización de modelos. Las tendencias léxicas no solo reflejan intereses científicos, sino también las condiciones de posibilidad para la automatización efectiva del análisis documental.

9. CONCLUSIONES Y TRABAJOS FUTUROS

9.1. Conclusiones

El sistema de clasificación construido permitió identificar patrones temáticos en la producción científica colombiana a partir de técnicas de procesamiento de lenguaje natural y aprendizaje automático. El enfoque desarrollado combinó modelos entrenados con redes neuronales profundas para clases de alta densidad semántica, y técnicas de extracción léxica con algoritmos clásicos para abordar objetivos con menor frecuencia documental o estructura transversal. Esta segmentación metodológica respondió a la necesidad de adaptar el modelo a las condiciones reales del corpus y a la desigual distribución de los ODS en el conjunto de textos analizados.

En dominios como la salud, el modelo mostró alta precisión clasificatoria. El rendimiento obtenido no se explica únicamente por la arquitectura empleada, sino por la existencia de un vocabulario técnico consolidado, que permitió al sistema establecer fronteras temáticas claras. La evolución léxica en el periodo 2018–2024, marcada por el impacto de la pandemia, reforzó esta estabilidad: conceptos clínicos como “patient”, “treatment” y “diagnosis” se intensificaron y adquirieron centralidad semántica, facilitando su identificación automática. Esta acumulación léxica se reflejó en los indicadores de desempeño y en la organización del espacio vectorial.

En el caso del ODS 9, el sistema fue capaz de reconocer estructuras temáticas vinculadas a innovación e infraestructura. La dispersión léxica propia de estos campos redujo la compacidad de los clústeres generados, pero no impidió una clasificación razonablemente estable. La cercanía semántica con otras áreas tecnológicas introdujo zonas de ambigüedad, que el modelo manejó con una precisión moderada, coherente con la naturaleza técnica y transversal de los textos.

La clasificación de los ODS 6, 7 y 13 evidenció los límites operativos del sistema ante estructuras temáticas difusas. La baja frecuencia de estas clases, unida a la variabilidad terminológica y al solapamiento con otras agendas, afectó el desempeño supervisado. Sin embargo, el modelo basado en extracción de términos clave logró recuperar textos relevantes, especialmente aquellos que abordaban problemáticas ambientales, energéticas o climáticas

desde perspectivas no lineales. La utilidad de esta arquitectura no reside en su exactitud, sino en su capacidad para ampliar la cobertura temática sin sacrificar coherencia mínima.

Los indicadores de agrupamiento confirmaron que la organización semántica del corpus no se ajusta a límites estrictos entre objetivos. Los clústeres observados mediante PCA mostraron regiones de concentración, pero también áreas de transición temática. Estas zonas de frontera son consistentes con la naturaleza interdisciplinar del conocimiento académico y con el carácter complementario de los ODS.

El análisis de términos más frecuentes a lo largo del periodo mostró una evolución desigual de los conceptos más utilizados. Mientras algunas áreas incrementaron su visibilidad terminológica —como salud y energía—, otras se mantuvieron estables o marginales. Estas trayectorias reflejan dinámicas de priorización investigativa y acceso desigual a los mecanismos de indexación. La clasificación automática permite detectar estos patrones, pero también revela las limitaciones del marco ODS como tipología cerrada frente a un conocimiento en constante reorganización.

El sistema diseñado no resuelve estas tensiones, pero las hace visibles. Clasifica con precisión en condiciones controladas, se adapta en contextos fragmentarios y evidencia las zonas del corpus donde la relación entre conocimiento científico y sostenibilidad es menos explícita. En este sentido, su aporte no es únicamente técnico. Ofrece una herramienta para examinar el modo en que se configura discursivamente la relación entre ciencia, política y desarrollo, y para contribuir a su lectura crítica desde una perspectiva estructurada y reproducible.

9.2. Trabajos Futuros

La base metodológica establecida en esta investigación abre diversas posibilidades de extensión y aplicación en futuros proyectos. En primer lugar, el enfoque propuesto puede ser adaptado para analizar producciones académicas en otras áreas del conocimiento, tales como las Humanidades, las Ciencias Sociales, las Ciencias Económicas y Administrativas, entre otras. La flexibilidad de la arquitectura híbrida permite su reentrenamiento y ajuste para capturar patrones semánticos propios de cada dominio disciplinar, extendiendo su utilidad más allá de los campos inicialmente considerados.

Adicionalmente, la metodología de clasificación desarrollada puede ser aplicada para categorizar investigaciones en función de otras taxonomías temáticas relevantes. Por ejemplo, puede adaptarse para identificar la alineación de la producción científica con los Objetivos del Proyecto Milenio, las categorías de áreas de conocimiento propuestas por la Misión de Sabios en Colombia, o cualquier otra estructura de clasificación que refleje prioridades nacionales o regionales en ciencia, tecnología e innovación.

Finalmente, futuras líneas de investigación podrían enfocarse en perfeccionar la capacidad de captura de multidimensionalidad temática mediante la implementación de modelos multilabel, el refinamiento de las técnicas de extracción de palabras clave, y el uso de representaciones semánticas más especializadas para corpus altamente interdisciplinarios. De igual forma, la integración de mecanismos de validación asistida por expertos podría fortalecer el control de calidad de las etiquetas asignadas y contribuir a la mejora continua del sistema.

La consolidación de este marco de clasificación representa un paso significativo hacia la comprensión de la contribución de la ciencia a los grandes retos sociales, y ofrece una plataforma sólida para el desarrollo de nuevas herramientas de análisis y monitoreo en políticas de investigación alineadas con el desarrollo sostenible.

10. REFERENCIAS BIBLIOGRÁFICAS

[1] United Nations, *Transforming our world: the 2030 Agenda for Sustainable Development*. UN Official Document System, 2015. [Online]. Available: https://sdgs.un.org/2030agenda

[2] J. D. Sachs, G. Schmidt-Traub, C. Kroll, G. Lafortune, G. Fuller, and F. Woelm, *Sustainable Development Report 2021*. Cambridge University Press, 2021. [Online]. Available: https://www.sdgindex.org/reports/sustainable-development-report-2021/

[3] Colciencias, *Informe de Investigación: Panorama de la Ciencia, Tecnología e Innovación en Colombia*. Colciencias, 2018.

[4] MinCiencias, *Estado de la Ciencia, la Tecnología y la Innovación en Colombia 2020*. MinCiencias, 2020.

[5] J. Hutchins, "La primera demostración pública de traducción automática: el sistema Georgetown-IBM, 7 de enero de 1954," *La Linterna del Traductor*, no. 6, 2011. [En línea]. Disponible en: https://lalinternadeltraductor.org/n6/traduccion-automatica.html.

[6] J. Devlin, M.-W. Chang, K. Lee y K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019. DOI: 10.18653/v1/N19-1423.

[7] A. Radford et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020. DOI: 10.48550/arXiv.2005.14165.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser y I. Polosukhin, "Attention is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017. DOI: 10.48550/arXiv.1706.03762.

[9] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Scao, S. Gugger, M. Drame, Q. Lhoest y A. M. Rush, "Transformers: State-of-the-Art Natural Language Processing," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020. DOI: 10.18653/v1/2020.emnlp-demos.6.

[10] D. M. Blei, A. Y. Ng y M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. [Online]. Disponible en: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, y V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint*

arXiv:1907.11692, jul. 2019. [En línea]. Disponible: https://arxiv.org/abs/1907.11692

[12] A. K. Uysal y S. Gunal, "The Impact of Preprocessing on Text Classification," *Information Processing & Management*, vol. 50, no. 1, pp. 104–112, 2014. DOI: 10.1016/j.ipm.2013.08.006.

[13] T. Mikolov, K. Chen, G. Corrado y J. Dean, "Efficient Estimation of Word Representations in Vector Space," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013. DOI: 10.48550/arXiv.1301.3781.

[14] Q. Le y T. Mikolov, "Distributed Representations of Sentences and Documents," *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pp. 1188–1196, 2014. [Online]. Available: https://proceedings.mlr.press/v32/le14.pdf.

[15] I. Tenney, D. Das y E. Pavlick, "BERT Rediscovers the Classical NLP Pipeline," en *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florencia, Italia, 2019, pp. 4593–4601. DOI: 10.18653/v1/P19-1452. Disponible en: https://aclanthology.org/P19-1452/

[16] W. S. McCulloch y W. Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943. DOI: 10.1007/BF02478259.

[17] F. Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958. DOI: 10.1037/h0042519.

[18] D. E. Rumelhart, G. E. Hinton, y R. J. Williams, "Learning Representations by Back-propagating Errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986. DOI: 10.1038/323533a0.

[19] I. Goodfellow et al., "Generative Adversarial Nets," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2672–2680, 2014. [Online]. Available: https://papers.nips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[20] Y. LeCun, Y. Bengio, y G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. DOI: 10.1038/nature14539.

[21] corresponde al artículo técnico original de YAKE, publicado en *Information Sciences*, y es indispensable para fundamentar cómo funciona el algoritmo.

[22] T. Nomoto, "Keyword Extraction: A Modern Perspective," *SN Computer Science*, vol. 4, no. 1, article 92, 2023. DOI: 10.1007/s42979-022-01481-7

[23] X. Huang, W. Huang, S. Duan y F. Liu, "Performance Evaluation of Keyword Extraction Methods and Visualization for Student Online Comments," *Symmetry*, vol. 12, no. 11, p. 1923, 2020. DOI: 10.3390/sym12111923.

[24] M. Hossin y M. N. Sulaiman, "A Review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5,

no. 2, pp. 1–11, 2015. DOI: 10.5121/ijdkp.2015.5201.
[25] Es una nueva referencia real y académicamente válida para la explicación del índice Kappa de Cohen. No había sido incluida previamente en tu listado.

[26] United Nations, "The Sustainable Development Goals Report 2016," United Nations, New York, 2016. [Online]. Available: https://unstats.un.org/sdgs/report/2016/

[27] United Nations, *The Sustainable Development Goals Report 2023: Special Edition*, New York: United Nations, 2023. [Online]. Available: https://unstats.un.org/sdgs/report/2023/.
[28] M. Chen, G. Mussalli, A. Amel-Zadeh, and M. O. Weinberg, "NLP for SDGs: Measuring Corporate Alignment with the Sustainable Development Goals," *J. Impact ESG Investing*, vol. 2, no. 3, pp. 61–81, Dec. 2021. [Online]. Available: https://doi.org/10.3905/jesg.2021.1.035

[29] *Contribution of Deep Learning and Artificial Intelligence to Attaining the Sustainable Development Goals Amidst the COVID-19 Pandemic*, vol. 2, no. 4, pp. 1–18, Oct. 2023. [Online]. Available: https://doi.org/10.21608/djicsi.2023.331757
[30] S. S. Mukonza and J.-L. Chiang, "Meta-Analysis of Satellite Observations for United Nations Sustainable Development Goals: Exploring the Potential of Machine Learning for Water Quality Monitoring," *Environments*, vol. 10, no. 10, p. 170, Oct. 2023. [Online]. Available: https://doi.org/10.3390/environments10100170

[31] T. B. Smith, R. Vacca, L. Mantegazza, and I. Capua, "Discovering new pathways toward integration between health and sustainable development goals with natural language processing and network science," *Global Health*, vol. 19, no. 1, Jun. 2023. [Online]. Available: https://doi.org/10.1186/s12992-023-00943-8

[32] T. B. Smith, R. Vacca, L. Mantegazza, and I. Capua, "Natural language processing and network analysis provide novel insights on policy and scientific discourse around Sustainable Development Goals," *Scientific Reports*, vol. 11, no. 1, Nov. 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-01801-6

[33] Elsevier, *Scopus Terms and Conditions of Use*, Elsevier, Amsterdam, 2024. [Online]. Available: https://dev.elsevier.com/policy.html

[34] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proc. 2019 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, 2019. DOI: 10.18653/v1/D19-1410.

11. ANEXOS

11.1. Anexo A. Términos clave por Objetivo de Desarrollo Sostenible (ODS)

Listado operativo para búsqueda y etiquetado manual de artículos científicos

ODS	Términos clave
ODS 1: Fin de la pobreza	pobreza, exclusión, ingreso mínimo, desigualdad económica, vulnerabilidad social, transferencias monetarias, pobreza extrema, desarrollo económico local, necesidades básicas, protección social, seguridad financiera, empleo informal, brechas sociales, privación, marginalidad
ODS 2: Hambre cero	seguridad alimentaria, malnutrición, agricultura sostenible, acceso a alimentos, productividad agrícola, sistemas alimentarios, soberanía alimentaria, apoyo a agricultores, nutrición infantil, fertilidad del suelo, sistemas de riego, alimentos básicos, inseguridad alimentaria, reservas de alimentos, políticas agrarias
ODS 3: Salud y bienestar	salud pública, atención médica, servicios sanitarios, enfermedades crónicas, salud mental, vacunación, salud infantil, cobertura universal en salud, hospitales, calidad de vida, medicamentos esenciales, bienestar, pandemia, morbilidad, mortalidad
ODS 4: Educación de calidad	acceso a educación, calidad educativa, alfabetización, brechas educativas, educación inclusiva, formación docente, cobertura escolar, competencias básicas, educación superior, currículo, equidad educativa, abandono escolar, educación rural, TIC en educación, sistemas de evaluación
ODS 5: Igualdad de género	equidad de género, empoderamiento femenino, violencia de género, brecha salarial, acceso a educación para mujeres, derechos reproductivos, discriminación de género, mujeres rurales, participación política femenina, salud sexual, acoso, igualdad laboral, liderazgo femenino, género y desarrollo, feminismo
ODS 6: Agua limpia y saneamiento	acceso al agua potable, saneamiento básico, tratamiento de aguas, gestión hídrica, higiene, contaminación del agua, aguas residuales, infraestructura hídrica, agua segura, servicios de saneamiento, escasez de agua, uso sostenible del agua, ríos contaminados, aguas subterráneas, gestión comunitaria del agua
ODS 7: Energía asequible y no contaminante	energías renovables, eficiencia energética, acceso a energía, energías limpias, electricidad rural, energía solar, energía eólica, biocombustibles, redes inteligentes, transición energética, consumo energético, fuentes sostenibles, innovación energética, electrificación, tarifas energéticas
ODS 8: Trabajo decente y crecimiento económico	empleo digno, productividad laboral, inclusión laboral, crecimiento económico, informalidad, condiciones de trabajo, seguridad laboral, desempleo juvenil, derechos laborales, economía formal, trabajo infantil, emprendimiento, salario digno, protección social laboral, generación de

ODS	Términos clave
	empleo
ODS 9: Industria, innovación e infraestructura	desarrollo industrial, infraestructura sostenible, innovación tecnológica, investigación aplicada, conectividad, transporte sostenible, digitalización, manufactura avanzada, parques tecnológicos, inversión en I+D, cadenas de valor, tecnología limpia, redes logísticas, productividad industrial, resiliencia infraestructural
ODS 10: Reducción de desigualdades	inclusión social, movilidad social, brechas económicas, equidad territorial, derechos humanos, participación equitativa, discriminación, justicia social, acceso equitativo, minorías, políticas redistributivas, igualdad de oportunidades, equidad étnica, disparidades regionales, exclusión
ODS 11: Ciudades y comunidades sostenibles	urbanismo sostenible, transporte urbano, gestión de residuos, vivienda digna, planificación territorial, movilidad sostenible, espacios públicos, resiliencia urbana, asentamientos informales, participación ciudadana, accesibilidad, calidad urbana, gestión de riesgos, zonas verdes, servicios urbanos
ODS 12: Producción y consumo responsables	consumo sostenible, producción limpia, economía circular, reciclaje, reducción de residuos, eficiencia de recursos, patrones de consumo, gestión ambiental, ciclo de vida del producto, huella ecológica, responsabilidad empresarial, sostenibilidad industrial, políticas de consumo, innovación responsable, empaques sostenibles
ODS 13: Acción por el clima	cambio climático, adaptación climática, mitigación, emisiones de carbono, resiliencia climática, gases de efecto invernadero, calentamiento global, políticas climáticas, huella de carbono, justicia climática, desastres naturales, eventos extremos, vulnerabilidad climática, neutralidad de carbono, transición ecológica
ODS 14: Vida submarina	ecosistemas marinos, biodiversidad marina, contaminación oceánica, pesca sostenible, acidificación del océano, corales, zonas costeras, áreas marinas protegidas, recursos pesqueros, plásticos en el mar, salud oceánica, regulación pesquera, manglares, sostenibilidad costera, océanos limpios
ODS 15: Vida de ecosistemas terrestres	deforestación, conservación de la biodiversidad, áreas protegidas, desertificación, reforestación, suelos degradados, vida silvestre, servicios ecosistémicos, fauna terrestre, flora nativa, restauración ecológica, corredores biológicos, manejo forestal sostenible, especies amenazadas, equilibrio ecológico
ODS 16: Paz, justicia e instituciones sólidas	gobernanza, participación democrática, justicia, transparencia, lucha contra la corrupción, derechos civiles, instituciones sólidas, acceso a la justicia, estado de derecho, seguridad ciudadana, prevención del delito, sistemas judiciales, libertades fundamentales, mediación, inclusión política

ODS	Términos clave
ODS 17: Alianzas para lograr los objetivos	cooperación internacional, alianzas estratégicas, financiamiento para el desarrollo, fortalecimiento institucional, alianzas público-privadas, transferencia de tecnología, asistencia técnica, multilateralismo, colaboración intergubernamental, redes globales, capacidades locales, cooperación Sur-Sur, gobernanza global, acuerdos multilaterales, monitoreo ODS

11.2. Anexo B. Caracterización de los jueces expertos

La validación experta del conjunto de datos se llevó a cabo con el apoyo de dos profesionales con formación y experiencia en áreas clave para el análisis temático de contenidos científicos, la gestión del conocimiento y la evaluación orientada por Objetivos de Desarrollo Sostenible (ODS). Ambos jueces cuentan con trayectoria acreditada en entornos académicos y técnico-científicos, lo que permitió garantizar la calidad semántica y la pertinencia conceptual de las etiquetas asignadas durante el proceso de clasificación.

John Agustín Riaño Díaz es profesional en Sistemas de Información, Bibliotecología y Archivística (Universidad de La Salle), estudiantes de la maestría en Ciencia de Datos (Pontificia Universidad Javeriana de Cali) y magíster en Docencia con énfasis en métodos de evaluación e investigación en educación (Universidad de La Salle). Actualmente cursa el Doctorado en Ciencias de la Información y Documentación (Universidad Complutense de Madrid). Su formación ha sido complementada con certificaciones especializadas en inteligencia artificial, bibliometría, evaluación de la ciencia y gestión estratégica de la información científica.

Se desempeña como director del Departamento de Estudios de Información en la Universidad de La Salle, desde donde lidera programas académicos, procesos de acreditación y estrategias para la gestión institucional del conocimiento. Su experiencia abarca consultoría en sectores público y privado, así como la implementación de tecnologías para el análisis y visualización de datos científicos. En el contexto de este proyecto, aportó criterios metodológicos y dominio técnico en ciencia de datos, procesamiento de lenguaje natural y clasificación documental, asegurando la consistencia de las decisiones de etiquetado con base en marcos internacionales como la Agenda 2030.

Juez 2. Profesional en Química con maestría en Química, especialista en gestión de investigación y transferencia tecnológica. Ha trabajado en la coordinación de proyectos de innovación y en la formulación de estrategias para la articulación universidad-empresa, con enfoque en el desarrollo sostenible y la apropiación social del conocimiento. Su experiencia incluye asesoría en propiedad intelectual, evaluación de impacto de la investigación y diseño de indicadores de gestión para instituciones del sistema nacional de ciencia, tecnología e innovación. Su participación como juez

experto permitió incorporar una perspectiva analítica desde la interfaz ciencia-política, fortaleciendo la validez contextual del modelo de clasificación frente a los ODS.