Maestría en

Ciencia de Datos y Máquinas de Aprendizaje mención en Inteligencia Artificial

Trabajo de investigación previo a la obtención del título de

Magíster en Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial

AUTORES:

BARRIONUEVO PORTERO JONATHAN ANDRÉS

GUZMAN NEIRA CARLOS LUIS

MOLINA MIRANDA NICOLE LUCIA

SALAZAR ROBLES DIEGO FERNANDO

TUTORES:

Alejandro Cortés Msc.

Iván Reyes Ch. Mgtr.

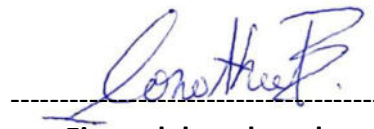
Minería de Textos en Resúmenes de Artículos Científicos para Identificar Tendencias de Investigación

Quito, diciembre 2024

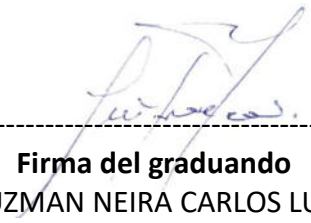
CERTIFICACIÓN DE AUTORÍA

Nosotros, **BARRIONUEVO PORTERO JONATHAN ANDRÉS, GUZMAN NEIRA CARLOS LUIS, MOLINA MIRANDA NICOLE LUCIA, SALAZAR ROBLES DIEGO FERNANDO,** declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o califcación profesional y que se ha consultado la bibliografa detallada.

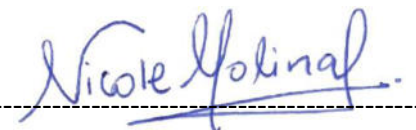
Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.



Firma del graduando
BARRIONUEVO PORTERO JONATHAN ANDRÉS



Firma del graduando
GUZMAN NEIRA CARLOS LUIS



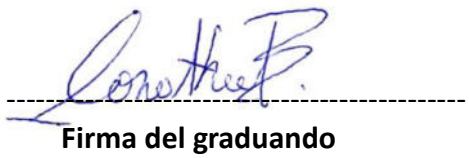
Firma del graduando
MOLINA MIRANDA NICOLE LUCIA

Firma del graduando
SALAZAR ROBLES DIEGO FERNANDO

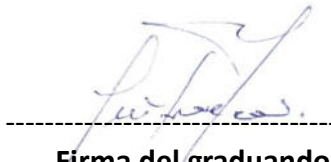
AUTORIZACIÓN DE DERECHOS DE PROPIEDAD INTELECTUAL

Nosotros, **BARRIONUEVO PORTERO JONATHAN ANDRÉS, GUZMAN NEIRA CARLOS LUIS, MOLINA MIRANDA NICOLE LUCIA, SALAZAR ROBLES DIEGO FERNANDO,** en calidad de autores del trabajo de investgación ttulado ***Minería de Textos en Resúmenes de Artculos Cientfcos para Identfcar Tendencias de Investgación***, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contene esta obra, con fnes estrictamente académicos o de investgación. Los derechos que como autores nos corresponden, lo establecido en los artculos 5, 6, 8, 19 y demás pertnentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, diciembre de 2024



Firma del graduando
BARRIONUEVO PORTERO JONATHAN ANDRÉS



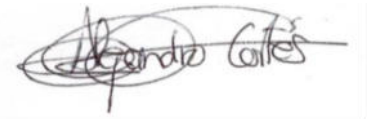
Firma del graduando
GUZMAN NEIRA CARLOS LUIS

Firma del graduando
MOLINA MIRANDA NICOLE LUCIA

Firma del graduando
SALAZAR ROBLES DIEGO FERNANDO

APROBACIÓN DE DIRECCIÓN Y COORDINACIÓN DEL PROGRAMA

Nosotros, **Alejandro Cortés e Iván Reyes,** declaramos que los graduandos: BARRIONUEVO PORTERO JONATHAN ANDRÉS, GUZMAN NEIRA CARLOS LUIS, MOLINA MIRANDA NICOLE LUCIA, SALAZAR ROBLES DIEGO FERNANDO son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés

Director de la Maestría en Ciencia de datos mención Inteligencia Artificial

Iván Reyes

Coordinador de la Maestría en Ciencia de datos mención en Inteligencia Artificial

DEDICATORIA

A nuestra familia y amigos...

RESUMEN

El presente estudio pivota sobre la aplicación de técnicas avanzadas de minería de textos para analizar artículos científicos en el ámbito del turismo con la intención de identificar aquellos patrones que pudieran surgir en torno ade los temas de investigación en lo que va del siglo XXI.

Dentro de un entorno *dockerizad*o (optimizado para la ejecución de análisis) configurado con herramientas como CUDA, PyTorch y bibliotecas de procesamiento del lenguaje natural (NLP) se desarrolló un cuaderno de Jupyter para realizar la recopilación, preprocesamiento y análisis de resúmenes académicos. Este enfoque permitió identificar temas de relevancia creciente como la sostenibilidad, el turismo inteligente y el análisis de datos, como temas de tendencia dentro del ámbito del turismo, proporcionando una visión integral del panorama investigativo en este campo.

La metodología empleada incluye el modelado de temas con Latent Dirichlet Allocation (LDA), que fue utilizado para crear y clasificar los artículos científicos analizados. Asimismo, se realizó un análisis riguroso de métricas como la perplejidad y la coherencia (UMASS y C_V) para optimizar el número de temas y garantizar su coherencia semántica. Este proceso permitió agrupar y etiquetar artículos, identificando patrones clave en el turismo. Adicionalmente, el modelo entrenado, soportado por redes neuronales LSTM, se diseñó con la idea de apoyar la clasificación y al análisis de rasgos temáticos de los artículos revisados, a través de la provisión de un modelo pre-entrenado que permita la identificación de estructuras relevantes en los datos analizados y contribuir al desarrollo de futuros entrenamientos.

El presente proyecto se enfrenta a la sobrecarga de información en la literatura científica, y presenta una solución funcional y precisa para su procesamiento y análisis, lo que habilita la identificación tanto de categorías, patrones y temáticas emergentes para los investigadores, a quienes proporciona una herramienta que permitiría mejorar su capacidad para analizar la densidad de campos investigación, constituyéndose en un pivote para el descubrimiento de áreas no transitadas, y facultando una mayor disruptividad temática, así como también el fomento de la constitución de redes académicas colaborativas en el ámbito turístico.

Palabras clave: Minería de textos, artículos científicos, modelado de tópicos, turismo sostenible, aprendizaje automático, visualización interactiva.

ABSTRACT

This study focuses on the application of advanced text mining techniques to analyze scientific articles in the tourism field, aiming to identify patterns in research topics throughout the 21st century.

Developed within a *dockerized* environment configured with tools such as CUDA, PyTorch, and natural language processing (NLP) libraries, a Jupyter notebook was created to perform the collection, preprocessing, and analysis of academic abstracts. This approach identified increasingly relevant topics such as sustainability, smart tourism, and data analytics as trending themes within the tourism field, providing a comprehensive overview of the research landscape in this domain.

The methodology includes topic modeling with Latent Dirichlet Allocation (LDA), which was utilized to create and classify the analyzed scientific articles. A rigorous analysis of metrics such as perplexity and coherence (UMASS and C_V) was conducted to optimize the number of topics and ensure semantic consistency. This process enabled the grouping and labeling of articles, identifying key patterns in tourism research. Additionally, the trained model, supported by LSTM neural networks, was designed to assist in the classification and analysis of reviewed articles, providing a pre-trained model that facilitates the identification of relevant patterns in the analyzed data and in future training sessions.

This project addresses the challenge of information overload in scientific literature, presenting a functional and precise solution for its processing and analysis. It enables the identification of categories, patterns, and emerging themes for researchers, providing them with a tool to improve their capacity to analyze the density of research fields. The project

serves as a pivot for discovering unexplored areas, fostering thematic disruptiveness, and promoting the formation of collaborative academic networks in the tourism field.

Keywords: Text mining, scientific articles, topic modeling, sustainable tourism, machine learning, interactive visualization.

TABLA DE CONTENIDOS

CERTIFICACIÓN DE AUTORÍA i

AUTORIZACIÓN DE DERECHOS DE PROPIEDAD INTELECTUAL ii

ACUERDO DE CONFIDENCIALIDAD iii

APROBACIÓN DE DIRECCIÓN Y COORDINACIÓN DEL PROGRAMA iv

DEDICATORIA v

AGRADECIMIENTOS vi

RESUMEN vii

ABSTRACT ix

CAPITULO 1: IDENTIFICACIÓN DEL PROYECTO 1

1.1 INTRODUCCIÓN 1

1.2 PLANTEAMIENTO DEL PROBLEMA E IMPORTANCIA DEL ESTUDIO 2

1.2.1 Definición del proyecto 2

1.2.2 Naturaleza o tipo de proyecto 3

1.2.3 Objetivos 3

1.2.4 Justificación e importancia del trabajo de investigación 4

CAPITULO 2: DESARROLLO DE LA PROPUESTA 14

2.1 METODOLOGÍA 14

2.2 DIAGRAMAS DE PROCESOS 16

2.3 REQUERIMIENTOS DE HARDWARE Y SOFTWARE PARA DESARROLLO 17

2.3.1 Docker 17

2.3.2 Librerías 18

2.3.3 DESARROLLO DEL MODELO 19
Preparación del Dataset 20
Exploración Inicial 22
Distribución de artículos 23
Distribución de cantidad de palabras 24
Distribución de los Tipos de Documentos 25
Distribución de Idiomas 26
Frecuencia de palabras clave 28
Análisis de Bigramas y Trigramas 29
Análisis de Trigramas 32
Definición de Variables de Clasificación 34
Análisis de Perplejidad 34
Calcular coherencia semántica con métricas UMASS y C_V 36
Implementación LDA 38
Refinamiento del Corpus 47
2.3.4 IMPLEMENTACIÓN DEL MODELO DE CLASIFICACIÓN 48
Fase de segmentación y preparación de datos 49
Procesamiento de texto y tokenización con textvectorization 49
Crear batches y preparar para entrenamiento 50
Selección de arquitectura para desarrollo de modelo 52
Análisis para este proyecto 53
Definición y estudio de hiperparámetros con randomizedsearchcv vs optimización bayesiana 54
CAPITULO 3: ANÁLISIS DE RESULTADOS 59
3.1 RENDIMIENTO POR ÉPOCA 59

3.2 OBSERVACIONES GENERALES 60

3.3 TENDENCIAS POR AÑO 60

3.4 MATRIZ DE CONFUSIÓN 64

3.5 DISTRIBUCIÓN DE PREDICCIONES POR AÑO 64

3.6 EMBEDDINGS 66

3.7 EFECTO DE REDUCCIÓN DE DIMENSIONALIDAD 68

CAPITULO 4: CONCLUSIONES Y RECOMENDACIONES 69

4.1 CONCLUSIONES 69

4.2 RECOMENDACIONES 70

4.3. LIMITACIONES A LA INVESTIGACIÓN 71

REFERENCIAS BIBLIOGRÁFICAS 72

LISTA DE TABLAS

Tabla 1 20

Tabla 2 22

Tabla 3 40

Tabla 4 41

Tabla 5 43

Tabla 6 53

Tabla 7 62

Tabla 8 65

LISTA DE FIGURAS

Figura 1 16
Figura 2 20
Figura 3 21
Figura 4 22
Figura 5 23
Figura 6 24
Figura 7 25
Figura 8 26
Figura 9 28
Figura 10 30
Figura 11 32
Figura 12 34
Figura 13 35
Figura 14 37
Figura 15 38
Figura 16 41
Figura 17 48
Figura 18 48
Figura 19 50
Figura 20 50
Figura 21 52
Figura 22 58

Figura 23 59

Figura 24 61

Figura 25 64

Figura 26 65

Figura 27 67

CAPITULO 1: IDENTIFICACIÓN DEL PROYECTO

1.1 INTRODUCCIÓN

La investigación en ciencia es un proceso dinámico y evolutivo, tanto en el número de sus autores, como en el número de sus respectivas publicaciones, y ha transitado por varias fases históricas, que en ocasiones duplicó el volumen de publicaciones, en periodos de14 años, con tasas de crecimiento -luego de la segunda guerra mundial- del 5,08% (Bornmann et al, 2021). Según Ware y Mabe (2015) últimamente, el crecimiento del número de artículos científicos ha aumentado con perfil exponencial, lo que permite colectar mayores datos textuales, relevantes para detección de tendencias en la investigación.

Particularmente en el campo del turismo, McKercher (2015) llegó a acumular un dataset de 19.921 papeles científicos; y dentro de ellos detectó más de 217.000 citas; lo cual es un indicio de lo abrumador que puede resultar para investigadores y académicos mantenerse al corriente de los avances en un campo específico.

Sumado a lo vasto de lo anterior Park et al (2023), en el campo científico detectaron que existe una progresiva pérdida de disruptividad en la ciencia y la tecnología. Entre otros, uno de sus indicadores fue la introducción de nuevas palabras (o palabras constructo) que evidencien el abordaje de nuevos paradigmas.

En este contexto, la minería de textos (rama de la inteligencia artificial orientada al manejo y estudio de grandes volúmenes de texto) de resúmenes de artículos científicos, ofrece una fórmula para abordar el problema. La aplicación de dicha fórmula sobre cada resumen científico permite, primero la obtención de datos que, y posteriormente -a través del uso de Inteligencia Artificial (AI por sus siglas en inglés)- la detección de posibles conglomerados de temas investigativos que facilitan la exploración y el descubrimiento

de los caminos transitados por la ciencia y como contrapartida, inteligir -como investigadores de la materia- los potenciales espacios de investigación que aún podría estar por explorarse.

Este proyecto ambiciona adentrarse en el cauce de conocimientos proporcionados en la presente maestría, para identificar herramientas que permitan la detección de conglomerados de temas que concitaron la atención de la investigación en materia turística relacionada con la AI. Para ello, se explorarán técnicas de Procesamiento del Lenguaje Natural (PLN) y Aprendizaje Automático (AA), que faculte el tratamiento de grandes corpus de resúmenes, tomando de éstos palabras que son claves, y temas que los engloben poco a poco para que se pueda auscultar la aglutinación de áreas de investigación.

En ese sentido, el modelado de tópicos puede ser un instrumento para el manejo de un conjunto de textos no estructurados. Es así como la clasificación de documentos, el descubrimiento de temas en grandes corpus en el contexto del turismo es la materia de este proyecto, acercándose a los temas más importantes relacionados con destinos inteligentes, tecnologías digitales e inteligencia artificial que delimite el ámbito en el cual la búsqueda de las dinámicas actuales de la industria se desarrolle.

1.2 PLANTEAMIENTO DEL PROBLEMA E IMPORTANCIA DEL ESTUDIO

1.2.1 Definición del proyecto

El proyecto aborda el manejo de la minería de textos que:

a. **Recopilará** resúmenes y metadatos: de artículos en el marco de la ciencia que provenga de diversas fuentes (bases bibliográficas, repositorios institucionales, etc.).
b. **Preprocesará** datos de los resúmenes, aplicando herramientas para la limpieza,

tokenización, eliminación de stop words y lematización/stemming.

c. **Procesará datos mediante** técnicas de PLN y AA para extraer palabras clave, temas, relaciones y patrones temporales en los datos extraídos.
d. **Identificará posibles patrones** en la investigación a partir de los datos extraídos.
e. **Visualizará simplificadamente** patrones e indicadores de forma fácil y comprensible.
f. **Evaluará con métricas** para proporcionar retroalimentación sobre la precisión y eficacia del sistema mediante descriptivas cuantitativas y cualitativas.

El proyecto **no** contempla:

- La evaluación de la calidad de los artículos científicos.
- La predicción específica de futuras investigaciones.
- Generar un modelo clasificatorio para futuros papers e investigaciones.

1.2.2 Naturaleza o tipo de proyecto

Proyecto de aplicación

1.2.3 Objetivos

Objetivo General: Generar una solución que permita el minado de textos, capaz de identificar patrones en los temas de investigación a partir de resúmenes de artículos científicos, sus títulos y palabras claves.

Objetivos Específicos:

- **Recopilar** un corpus de resúmenes de artículos científicos de diferentes disciplinas.
- **Preprocesar** los resúmenes aplicando técnicas de PLN.
- **Aplicar** algoritmos de AA para extraer palabras clave, temas y relaciones.
- **Desarrollar** un modelo para identificar tendencias de investigación a partir de los datos extraídos.
- **Evaluar** que tan ajustado es el sistema desarrollado a través de indicadores de

precisión, recall y F1-score.

1.2.4 Justificación e importancia del trabajo de investigación

Siendo los papeles científicos una evidencia de la capacidad disruptiva de un campo científico, nuestro grupo de estudio considera que el aumento en su cantidad así como la endogeneidad coercitiva de las citas no depende de la disminución del alcance de las "low-hanging fruit" (Park et al, 2023), es por esto que se justifica plenamente que se analice la literatura científica publicada cada año en el ámbito del turismo con el fin de detectar cuáles son los patrones campos de análisis y así como las áreas emergentes con oportunidades de colaboración.

En ese sentido, el uso de nuevas herramientas que permiten la explotación de los textos en resúmenes de artículos científicos procura una forma eficiente y escalable para analizar cualitativamente grandes volúmenes de literatura científica. Al programar el proceso de extracción de información, su preprocesamiento y procesamiento, ayudará a los investigadores con el ahorro de tiempo y de recursos, para que puedan concentrarse más en el significado de los resultados y la proyección de nuevas ideas a partir de éstos. Adicionalmente, la detección de tendencias facilita que los campos de estudio puedan ser mejor entendido en términos de su proceso evolutivo y la posible caracterización de áreas de investigación prometedoras.

Este proyecto contribuirá significativamente al avance de la investigación científica al proporcionar una herramienta que faculte a los investigadores remontar lo que previamente se ha mencionado y que se puntualiza a continuación:

- **Sobrecarga de información:** El volumen de literatura científica dificulta la revisión manual y la identificación de patrones. La minería de textos automatiza este proceso, permitiendo un análisis eficiente y a gran escala.
- **Pérdida progresiva de disruptividad de las nuevas investigaciones**: al mantenerse

actualizados, descubrir nuevas áreas de interés, identificar colaboradores potenciales y tomar decisiones informadas sobre sus líneas de investigación.

- **Dificultad de encontrar espacios de diálogo interdisciplinario:** Al revelar conexiones entre diferentes áreas de investigación, el proyecto puede estimular la colaboración interdisciplinaria y la generación de nuevas ideas.

El sistema se evaluará utilizando un conjunto de datos de resúmenes de artículos científicos de diferentes campos de estudio. Se medirá su capacidad para identificar tendencias conocidas y descubrir nuevas áreas de investigación.

Las principales ventajas de llevar a cabo este proyecto son:

- **Automatización:** El sistema automatizará el proceso de análisis de resúmenes, ahorrando tiempo y esfuerzo a los investigadores.
- **Escalabilidad:** El sistema podrá manejar grandes volúmenes de datos, permitiendo el análisis de corpus de resúmenes a gran escala.
- **Visualización:** El sistema proporcionará visualizaciones interactivas de las tendencias identificadas, facilitando la exploración y comprensión de los datos.

La minería de textos no es un campo aislado, puesto que combina técnicas de tratamiento de lenguaje natural (NLP), aprendizaje automatizado y análisis de datos con el fin de tratar grandes volúmenes de texto no estructurado y extraer de ellos datos procesables, con el fin de hacer posible la detección de patrones, la automatización de procesos clasificatorios que aboquen a una mejor comprensión del contenido (Feldman y Sanger, 2006). En definitiva, se procura: la detección de relaciones y configuraciones ocultas dentro de los datos textuales, la extracción de conocimiento y la clasificación y agrupamiento en categorías.

"La minería de texto convierte datos textuales en información estructurada para facilitar el análisis, permitiendo a los usuarios comprender y tomar decisiones basadas en grandes volúmenes de texto" (Feldman y Sanger, 2006).

Ahora bien, en la investigación científica la cantidad de publicaciones y la complejidad de la información que en ellas reside, problematizan la identificación de campos de estudio más transitados y patrones de examinación investigativa. Es por esto que, la combinación de la estadística y sus técnicas, así como la automatización del aprendizaje sobre los datos y el procesamiento del lenguaje natural que se encuentra amparada bajo el paraguas conceptual del minado de textos (Feldman y Sanger, 2007), habilitaría el análisis de dicho contenido textual en el campo de interés de este proyecto: el turismo.

Como se mencionó en el análisis del campo teórico, la minería de textos involucra una serie de pasos para convertir datos no estructurados en información estructurada. El primer paso, es el preprocesamiento del texto (limpiar y normalizar contenido). Este paso incluye la tokenización, proceso que implica la partición del texto en subloques más básicos (palabras o frases) con el fin de tratarlas de manera individual; además se eliminan palabras sin semántica relevante (artículos, preposiciones y similares) puesto que solo aumentan el ruido y ralentizarían el posterior análisis. Otro paso es la lematización y el stemming, subproceso por el cual se toma las palabras en su forma raíz o base de un modo que considera las variantes de una misma palabra como un mismo constructo, facilitando que los datos textuales se traten de manera estructurada, eficiente que mejora la precisión de su ulterior análisis (Jurafsky y Martin, 2023).

Las técnicas de minería de texto se extienden, desde enfoques estadísticos básicos hasta herramientas avanzadas de aprendizaje automático. Una de las técnicas más utilizadas es el análisis de frecuencia de palabras, que consiste en contar cuántas veces

aparece cada término en el texto. Esta técnica permite identificar rápidamente los términos más utilizados y los temas más recurrentes en un conjunto de datos, lo cual resulta útil para obtener una primera visión general del contenido y de los temas clave (Feldman y Sanger, 2007).

Otra técnica muy utilizada para el estudio de datos textuales es la extracción de palabras con constructos más nucleares y n-gramas. Las palabras clave son los términos más representativos de los textos de un documento, mientras que los n-gramas, son secuencias de palabras consecutivas dentro de combinaciones de términos más frecuentes. Esto permite reconocer patrones y expresiones comunes en el texto (Aggarwal y Zhai, 2012). Es de resaltar, que la clasificación de textos usa modelos de tipo supervisados característicos del espacio del aprendizaje automático que utiliza algoritmos para categorizar documentos. Algoritmos como Naive Bayes y Support Vector Machines (SVM) son típicos de este tipo de tareas (Sebastiani, 2002).
“La clasificación supervisada es clave para categorizar grandes volúmenes de texto, facilitando su organización y búsqueda” (Sebastiani, 2002).

Por otro lado, en el marco de las técnicas del aprendizaje denominado como “no supervisado”, el clustering, o agrupamiento, es una forma de organizar documentos con cierta similitud en grupos denominados usualmente por su nombre en inglés "clusters". A diferencia de los previamente mencionados, métodos supervisados, el clustering no requiere una preclasificación, aspecto que permite el descubrimiento de patrones no detectables a priori y que podrían estar ocultos en los datos. En ese contexto, algoritmos como K- means y el clustering jerárquico podrían ser de particular utilidad, ya que facilitarían la segmentación de grandes volúmenes de texto en grupos con características comunes (Miner et al., 2012).

El campo de éste proyecto se enmarca en la bibliometría, misma que se define como " an important field of information science as it represents a unique set of techniques for monitoring and analyzing information resources and managing knowledge in social and organizational contexts" es decir, en definitiva representa técnicas de monitoreo y análisis de recursos de conocimiento (İri y Ünal, 2024) que al combinar la minería de datos y textos en el dominio de la bibliometría, potencian la su capacidad de análisis de grandes volúmenes de publicaciones científicas de manera automatizada y precisa, en palabras de Cobo et al. (2011) “this cooperation among tools gives a positive synergy, which allows us to extract the knowledge hidden behind the data” lo que indica que puede inclusive revelar el conocimiento que se encuentra escondido detrás de la data.

El primer paso, para construir un sistema que facilite la minería de textos en el ámbito académico es la recopilación de un corpus de resúmenes científicos a través de bases de datos científicas confiables que proporcionen una base confiable mínima de la calidad y representatividad del corpus de textos que serán materia del análisis. En ese sentido, bases de datos como las referidas, y otras como IEEE Xplore y PubMed son de amplia utilización en la investigación científica, dada su multidisciplinariedad y revisión de pares (Bornmann y Leydesdorff, 2015). Un estudio reciente utilizó la minería como técnica para analizar resúmenes de textos de artículos publicados en revistas de turismo. Se aplicaron algoritmos de LDA para identificar temas emergentes, como la sostenibilidad y el turismo digital, revelando cómo estos temas han evolucionado en la literatura (Khan et al., 2021).

Es de reconocer, que el tratamiento de texto en este ámbito plantea desafíos para el equipo: preprocesar, estructurar y modelar información semántica, no solo lograrlo, sino también hacerlo de una forma costo-eficiente. Por eso es necesario usar un conjunto de herramientas digitales, que podrían ser clasificadas en: manipulación de datos,

visualización, manejo de texto, modelado de texto, procesamiento del lenguaje natural (NLP); y, aprendizaje profundo.

Sin embargo, antes de aplicar modelos de aprendizaje profundo, el texto debe ser preprocesado. Para esto, hay bibliotecas estándar de Python, como *random*, *string* y *re*, que sirven para esta tarea, con raíces en la teoría de los lenguajes formales y la informática, permiten limpiar el texto eliminando caracteres especiales y normalizando el formato, así como realizar operaciones para aumentar la diversidad de los datos. Guido van Rossum, creador del lenguaje Python, estableció el marco inicial para estas bibliotecas, puesto que las incorporó en el núcleo del lenguaje desde sus primeras versiones (van Rossum, 1995).

La evolución de estas herramientas tiene partida en la teoría de los lenguajes regulares de Stephen Kleene (Kleene, 1951). Posteriormente, en entornos de programación como Unix, Ken Thompson y Dennis Ritchie, iniciaron implementaciones prácticas de expresiones regulares y herramientas (Thompson y Ritchie, 1984). Por otro lado, el lenguaje Perl, de Larry Wall, popularizó las expresiones regulares modernas, que a su vez influyeron en la implementación de Regex en Python (Wall et al., 2012). Finalmente, el trabajo en algoritmos y estructuras de datos de Donald Knuth, aportó en la implementación de generadores aleatorios y el manejo de cadenas de texto de los lenguajes de programación (Knuth, 1997).

El presente proyecto contiene un enfoque sistemático que utiliza estas librerías de manera sinérgica para superar las limitaciones de datos textuales brutos que presentan, a fin de transformarlos en representaciones más estructuradas, que progresivamente permita que las librerías y funciones pueden interpretar.

Consecuentemente, librerías como *pandas* y *numpy* fueron utilizadas para el tratamiento de los datos obtenidas. *NumPy* fue creada por Travis Oliphant en 2005, que

éste la destinó para operaciones numéricas y manejo de arreglos multidimensionales en Python. Sin embargo, también hay trabajos previos de Jim Hugunin, quien desarrolló *Numeric*, y de Perry Greenfield y Paul Dubois, quienes contribuyeron a *Numarray* (Oliphant, 2006) (Harris et al., 2020). Por otro lado, *Pandas*, que fue desarrollada por Wes McKinney en 2008, proporciona estructuración para los datos; y su análisis. McKinney se inspiró en las necesidades que el análisis financiero requería en su época, pero también en herramientas consolidadas como R y DataFrame de S-PLUS (McKinney, 2011). En suma, las dos librerías (*NumPy* y *Pandas*) facilitan la fase previa del procesamiento de los datos y posteriormente su procesamiento y estructuración de tal forma que puedan compatibilizarse con las técnicas de procesamiento del lenguaje natural (NLP).

Parte de las etapas requeridas por el proyecto, incluye la extracción de insights relevantes. Para superar ese desafío, se requieren librerías como *Matplotlib* y *Seaborn*, las cuales permiten la conversión de datos numéricos en unas representaciones gráficas más fáciles de entender. Por su parte, *Matplotlib*, creada por John Hunter, ofrece más control de la personalización de los gráficos, se puede descubrir la distribución de las palabras o características del texto, mientras que *Seaborn*, que fuera desarrollada por Michael Waskom, adiciona una interfaz estadística gráfica que permite visualizaciones de mayor nivel como la colorimetría acopladas a los mapas, que revelan cuan densas se convierten las palabras por su uso, por temas, y hasta por formas de distribución. Las referidas visualizaciones de datos permiten la mejor comprensión de patrones en el conjunto de datos previo a construir modelos más definidos (Pfister & Heer, 2010).

Es por esto que, en este proyecto, se apela a autores como Hanspeter Pfister y Jeffrey Heer, quienes desarrollaron herramientas para la interpretación gráfica de patrones. Edward Tufte, pionero en el ámbito de la visualización aportó con mayor claridad y

precisión en la presentación de información cuantitativa; contribución que sigue siendo de actualidad y guía en el diseño de visualizaciones (Tufte, 1980).

En definitiva, dado que se necesita que la máquina "comprenda" en su "lenguaje", es necesario modelar el texto de los resúmenes para que sean susceptibles de transformación en información comprensible para aquella. En este sentido, las herramientas para dicha conversión textual hacia representaciones numéricas como *scikit-learn* y *Gensim* permiten descubrir patrones subyacentes. *Scikit-learn*, fue iniciada por David Cournapeau; y, adicionada posteriormente por más desarrolladores que continuaron con algoritmos de aprendizaje automático (Pedregosa et al., 2011). Por su parte, Gensim, desarrollada por Radim Řehůřek, se adentró en la búsqueda de similitudes semánticas (Řehůřek y Sojka, 2010).

Vale resaltar que estas bibliotecas no habrían sido posibles sin la introducción de conceptos previos como el Latent Semantic Analysis (LSA) de Thomas Hofmann, o el modelo Latent Dirichlet Allocation (LDA) de David Blei; bases para el desarrollo de algoritmos de modelado de temas y que tuvieron influencia en la gestación de herramientas como *gensim.models* (Blei, Ng, y Jordan, 2003). Respectivamente, *scikit-learn* con *sklearn.feature_extraction.text*, a través de las técnicas Bag-of-Words y TF-IDF se realizan representaciones vectoriales y se capturan semántica; *gensim.corpora* facilita que dichas representaciones aborden la construcción de modelos de tópicos.

El lenguaje natural parte de constructos de la lingüística computacional para su tratamiento, que a su vez fue influenciada por figuras de renombre como Noam Chomsky, quien hizo contribuciones significativas a la gramática generativa (Chomsky, 2002); así como también, Christopher Manning y Dan Jurafsky, quienes asentaron aún más la teoría y práctica del campo (Jurafsky y Martin, 2008; Manning y Schütze, 1999).

Como se había referido en el acápite anterior, el texto requiere preprocesamiento extenso previo a ser susceptible de modelación. Esto incluye operaciones aparentemente obvias para la mente humana, como eliminar palabras irrelevantes, normalizar palabras y encontrar relaciones semánticas. Para dichos efectos, bibliotecas como NLTK (Natural Language Toolkit) desarrollada por Steven Bird y Edward Loper, proporcionan las herramientas que realizan las mencionadas operaciones a fin de que los datos se pongan a punto para modelos más sofisticados como LSTM (Long Short-Term Memory) o TensorFlow.Keras.

La primera, LSTM (Long Short-Term Memory) ha sido introducida por Sepp Hochreiter y Jürgen Schmidhuber, con el objetivo de modelar relaciones contextuales y dependencias temporales que se pudieran encontrar en datos secuenciales de texto, lo que empezaba a abrir posibilidades en el procesamiento y análisis lingüístico (Hochreiter y Schmidhuber, 1997).

Por su parte, *TensorFlow* y *Keras* son herramientas de alto nivel que se enmarcan en el ámbito del aprendizaje profundo, esta última fue desarrollada principalmente por François Chollet**,** para que la aplicación del concepto de redes neuronales se facilite en su integración con *TensorFlow*, lo que en alguna manera revolucionó el preprocesamiento de texto y secuencias (Chollet, 2017). La revolución que implicó *TensorFlow*, desarrollada por miembros de Google Brain que fue liderado por personajes como Jeff Dean y Andrew Ng, fue plataforma clave para aplicaciones más modernas de *NLP* (Abadi et al., 2016).

El campo de trabajo de estas herramientas ha seguido recibiendo contribuciones destacables como la de Ian Goodfellow, quien junto con Yoshua Bengio y Aaron Courville trataron sobre el aprendizaje profundo (Goodfellow et al., 2016); así como también, Christopher Olah contribuyó a la visualización de las redes neuronales, lo que siguió facilitando su adopción en proyectos de procesamiento de lenguaje (Olah, 2015).

En definitiva, todo este conjunto de bibliotecas permite que las redes neuronales puedan capturar dependencias temporales en relación con su contexto, lo que ha facilitado la generación de texto, la traducción automática o la clasificación compleja.

CAPITULO 2: DESARROLLO DE LA PROPUESTA

2.1 Metodología

Para la elaboración de este modelo, se ha decidido establecer una metodología mixta donde se combinan los elementos propios del análisis cuantitativo y del análisis cualitativo para la obtención de distintos tipos de análisis sobre resúmenes de artículos científicos en el ámbito del turismo.

El análisis cuantitativo se aplicará en el sentido de que va a poder detectar patrones y tendencias de los datos, mientras que el análisis cualitativo se aplicará para poder identificar el significado, la importancia de esos patrones y tendencias detectados.

Esta mezcla de análisis servirá para alcanzar una explicación lo suficientemente compleja y profunda de los resúmenes de los artículos científicos de turismo.

Enfoque Cuantitativo:

- **Análisis Bibliométrico:** Implica el estudio analístico a través de la estadística que la información presenta al interior de los artículos, lo que incluye la distribución anual de publicaciones, la frecuencia de palabras clave y la extensión de los resúmenes. Este proceso nos facilita el reconocimiento de tendencias generales y patrones en la producción científica.
- **Procesamiento del Lenguaje Natural (PNL):** Se emplean técnicas de PNL para el tratamiento previo del texto, que incluyen la supresión de palabras de relleno, la lematización, y la tokenización.
- **Machine Learning (ML):** Se utiliza un modelo de “Asignación Latente” de Dirichlet (LDA) para la extracción de temas, lo que posibilita la identificación de las principales áreas de investigación dentro del campo del turismo.

- **Visualización:** Se desarrollan gráficos y visualizaciones que representan las tendencias identificadas, lo que incluye el reparto de temas en el curso de la ventana de tiempo analizada y los embeddings de tópicos.

Enfoque Cualitativo:

- **Interpretación de Resultados:** Se lleva a cabo una interpretación manual de los hallazgos del análisis cuantitativo, enfocándose en la identificación de los temas predominantes y las tendencias emergentes.
- **Análisis de Contenido:** Se realiza una exploración más exhaustiva del contenido de los resúmenes, mediante la codificación temática y la identificación de opiniones.

Metodología Mixta:

- **Combinación de Métodos:** Se integran los resultados del análisis cuantitativo (estadísticas, PNL, ML) con la interpretación cualitativa, proporcionando una perspectiva más completa sobre las tendencias de investigación
- **Triangulación:** Se aplican diversos métodos analíticos para validar los hallazgos, como la comparación de las tendencias reveladas por el modelo LDA con la interpretación manual de los resúmenes.

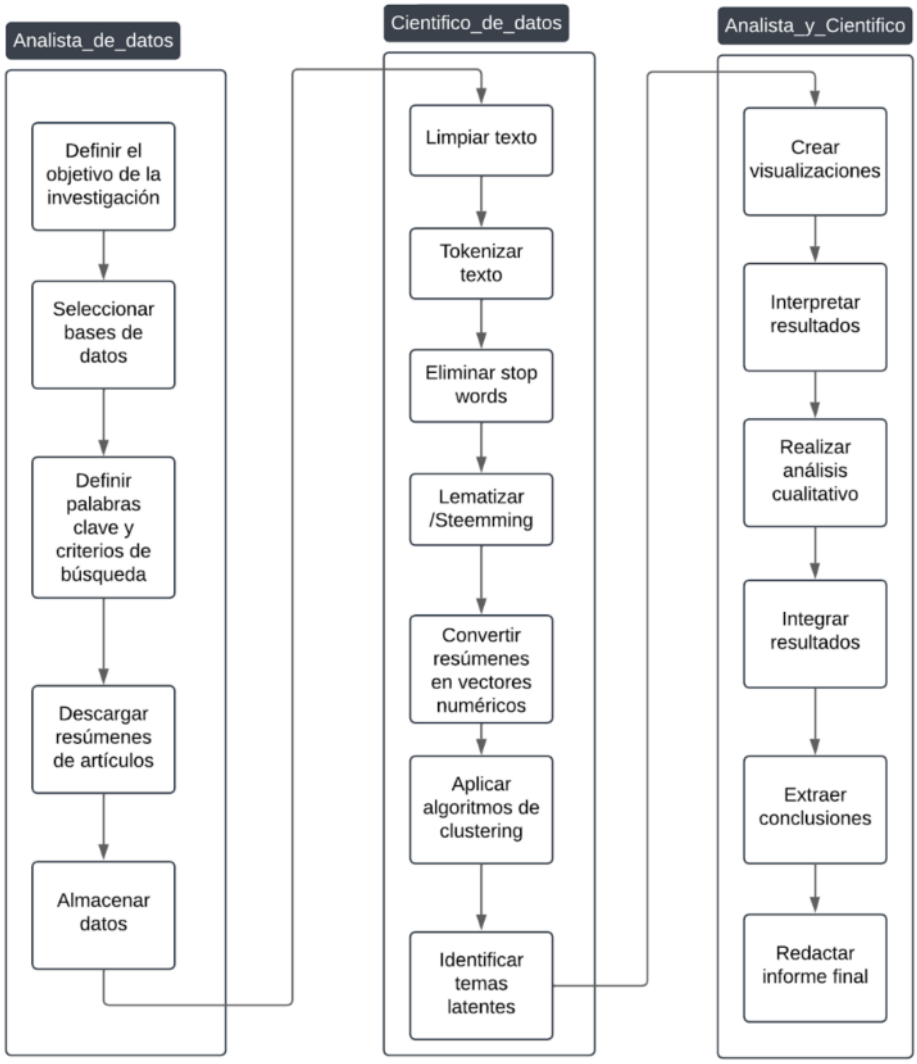
Así es como el código desarrollado pone en práctica una metodología mixta que explota las fortalezas de los enfoques cuantitativo y cualitativo para examinar las tendencias de investigación en turismo. La fusión de técnicas de análisis bibliométrico, PNL, ML y visualización, junto con la interpretación cualitativa aplicado a un corpus de resúmenes de artículos científicos, contribuye significativamente a un entendimiento de mayor consistencia y profundidad de los datos.

2.2 Diagramas de procesos

A continuación, se presenta el proceso realizado para la implementación del modelo aplicado con los diferentes roles aplicados al mismo.

Figura 1

Diagrama de proceso



Nota. La siguiente figura muestra el proceso para el desarrollo del caso práctico.

2.3 Requerimientos de hardware y software para desarrollo

2.3.1 Docker

- **Docker para un ambiente reproducible:** Gracias a que Docker encapsula el modelo y sus dependencias en un contenedor, es posible reproducir el estudio en cualquier equipo que tenga instalado Docker. Nos aseguramos así de que el modelo funcione de la misma manera, sin tener en cuenta cuál sea el sistema operativo de la máquina del usuario o cuál sea el entorno de las bibliotecas en las que se basa la máquina.
- **Docker para un ambiente limpio:** Docker garantizará un ambiente que esté limpio y libre en el desarrollo del modelo y, así, evitará conflictos con otras bibliotecas (o el sistema). Se reduce el riesgo de sufrir errores con otras configuraciones o bibliotecas e, incluso, se facilita la gestión de bibliotecas.
- **Docker para el aprovechamiento de GPU:** Docker permite montar el contenedor de manera que se pueda acceder a la GPU que tiene la máquina que ejerce de host. Con ello, se afianzan las prácticas necesarias para entrenar el modelo, especialmente en tareas que exigen mucha capacidad de procesamiento, como el análisis de grandes volúmenes de texto.
- **Docker para la escalabilidad:** Si el modelo crece, si conviene exponer el modelo en la nube, Docker proporciona una solución a la escalabilidad y se puede exponer el modelo. Los contenedores Docker son instancias que se pueden ejecutar en distintas plataformas y servicios en la nube sin quemar puentes ni hacer perder tiempo en el proyecto al aparentemente considerado ajuste en la capa de procesamiento que sea la adecuada para el modelo.

Es así que, en aras de aumentar las probabilidades de reproducibilidad y eficacia del modelo desarrollado, se optó por utilizar Docker, una herramienta para conseguir la creación de un ambiente de desarrollo cerrado y homogéneo. Docker permite encapsular el modelo y

aquellas librerías y herramientas que se requieran en un contenedor que se puede ejecutar en cualquier máquina que disponga de Docker. Docker, además, permite realizar la configuración del acceso a la GPU de la máquina host, lo que además permite conseguir una mejora en la rapidez del entrenamiento del modelo. Por otro lado, también facilita el escalado y la implementación en la nube que permite tener un entorno aislado en un entorno más abierto, lo que aumenta sus potencialidad como herramienta para la construcción de proyectos de Machine Learning.

2.3.2 Librerías

Para el procesamiento de datos y Machine Learning:

- **pandas:** que facilita que los datos sean tratados y analizados. Se aplica en la carga del dataset, en su limpieza, realizar análisis exploratorio, y preparación para su uso dentro del modelado .
- **scikit-learn:** utilizado para la división del dataset en conjuntos de entrenamiento y prueba (train_test_split), la vectorización de texto

(TfidfVectorizer) y, además, para la medición posterior de la eficacia del modelo (accuracy_score, confusion_matrix).

- **Gensim:** Para la generación de embeddings de palabras utilizando Word2Vec.
- **Optuna:** Usado como optimizador de hiperparámetros del modelo LSTM.
- **NLTK:** Para el procesamiento del lenguaje natural, como la supresión de stop words (stopwords) y la denominada "lematización" (WordNetLemmatizer).
- **TensorFlow/Keras:** Para la construcción, entrenamiento del modelo LSTM y, además, para constituir su arquitectura, incluyendo capas de embedding, LSTM, densas y dropout, así como para compilar y entrenar el modelo.

Para la visualización:

- **matplotlib.pyplot:** usadas en la graficación con histogramas, con barras, columnas, tortas, etc., que muestran la distribución de los datos y la frecuencia de palabras clave.
- **seaborn:** Para la creación de gráficos de estadística más avanzada, como mapas de calor, para visualizar la correlación entre variables.

Otras librerías:

- **re:** Para trabajar con expresiones regulares, utilizado en la limpieza de texto.
- **collections:** Para el análisis de frecuencia de palabras clave, utilizando Counter.
- **itertools:** Para la generación de combinaciones de palabras clave (bigramas y trigramas).
- **numpy:** Para operaciones numéricas, como el cálculo de la longitud de los resúmenes.

2.3.3 Desarrollo del Modelo

Para el desarrollo del modelo se implementa en las siguientes fases:

Preparación del Dataset

Esta fase inicial está encaminada a la carga, limpieza y transformación de los datos; esto es, eliminar duplicados, gestionar los valores nulos o unificar las columnas de las palabras clave para favorecer el análisis.

Figura 2

Código de carga del dataset

```
import pandas as pd
import numpy as np

# Cargar el archivo CSV
dataset_path = '/tf/Dataset/grupo8_dataset.csv'
df = pd.read_csv(dataset_path)

# Mostrar las primeras filas para verificar la carga
df.head()
```

Tabla 1

Variables definitivas para estudio

Variable	Tipo de dato	Descripción
Title	String	Título del artículo científico.
Year	Entero	Año de publicación del artículo.
Abstract	String	Resumen del artículo científico.
Author Keywords	String	Palabras clave proporcionadas por los autores del artículo.
Index Keywords	String	Palabras clave asignadas por el índice o base de datos.
Document Type	String	Tipo de documento (ej. artículo, conferencia, revisión).
Language	String	Idioma en el que está escrito el artículo.

Una vez hemos definido las variables definitivas del Dataset, deberemos proceder con la reducción de dimensionalidad de las columnas de Keywords.

La razón por la que unificamos las columnas 'Author Keywords' y 'Index Keywords' en una sola columna de 'Keywords' es para:

- Eliminar redundancias: Tener dos columnas separadas para las palabras clave puede generar duplicación innecesaria de información. Al unificar ambas columnas y eliminar los duplicados, reducimos la dimensionalidad del dataset sin perder información valiosa.
- Simplificar el análisis: Al consolidar las palabras clave en una sola columna, el análisis posterior y el preprocesamiento se vuelven más manejables, ya que los modelos de ML pueden trabajar más fácilmente con una única fuente de palabras clave.
- Mejora de la representación de los datos: Esta fusión proporciona una representación más precisa de las palabras clave asociadas a cada documento, evitando posibles confusiones o inconsistencias en el análisis.

Figura 3

Código para unificar las columnas de valores claves

```
# Crear la nueva columna 'Keywords' sin valores duplicados y ordenada
alfabéticamente
df_cleaned['Keywords'] = df_cleaned.apply(
    lambda row: ', '.join(sorted(set((row['Author Keywords'] + ', ' +
row['Index Keywords']).split(', ')))), axis=1
)

# Seleccionar las columnas para el DataFrame reducido
df_reduced = df_cleaned[['Author(s) ID', 'Title', 'Year', 'Abstract',
'Keywords', 'Language of Original Document', 'Document Type']]

# Mostrar las primeras filas del DataFrame reducido
df_reduced.head()
```

Figura 4

Tratamiento de valores nulos

```
# Calcular los valores nulos por columna
nulos_por_columna = df_reduced.isnull().sum()

# Crear un DataFrame para mostrar los valores nulos
nulos_df = pd.DataFrame({
    'Column': nulos_por_columna.index,
    'Null Values': nulos_por_columna.values
})

# Mostrar el DataFrame de valores nulos
nulos_df
```

Tabla 2

Resultado tratamiento valores nulos por columna

Variable	Valores nulos antes de la unificación	Valores nulos después de la unificación	% Reducción
Author ID	143	0	100%
Title	0	0	-
Year	0	0	-
Abstract	0	0	-
Keywords	0	0	-
Languaje of original Document	0	0	-
Document Type	0	0	-

Exploración Inicial

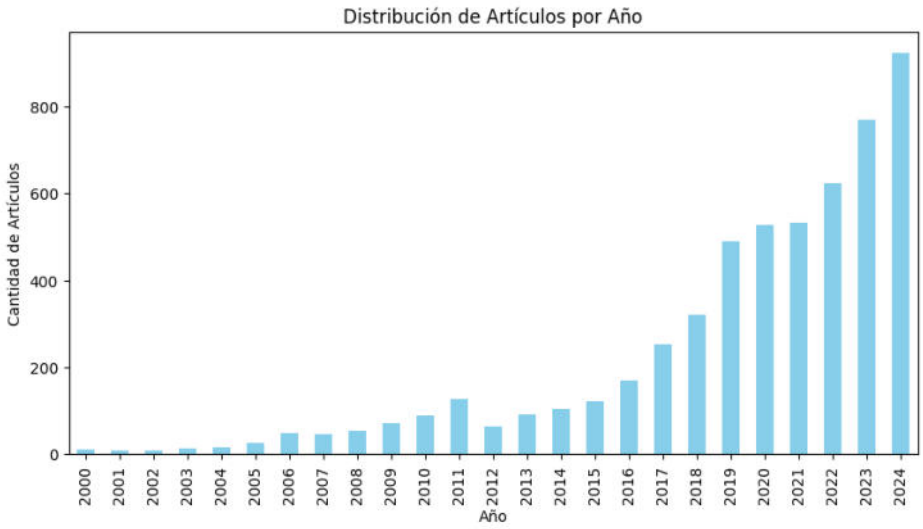
Se realiza un análisis exploratorio del Dataset que permita un inicial entendimiento de la configuración y ciertas características de los datos. Se analiza la agrupación de artículos

por año, la longitud de los resúmenes, los tipos de documentos, los idiomas y la frecuencia de las palabras clave.

Distribución de artículos

Figura 5

Distribución de artículos por año



Observamos un sustancial incremento de la proporción de artículos desde el año 2019 hasta el 2023, lo cual indica un creciente interés en la temática de investigación. Se identificaron los siguientes puntos:

- Pico de Publicaciones en 2023: El año 2023 muestra la mayor cantidad de publicaciones, lo que podría sugerir que el tema alcanzó un punto alto de relevancia en la comunidad académica en ese año. Esto también podría estar relacionado con eventos globales, como políticas sostenibles, cambios en el turismo post-pandemia o avances en la tecnología de análisis de datos en turismo.
- Descenso en 2024: Aunque el año 2024 aún no ha finalizado, la gráfica muestra una disminución en la cantidad de publicaciones. Esto puede deberse a varios factores:
    - Datos Parciales: Es posible que los datos de 2024 aún no estén completos, especialmente si la recopilación se realizó en los primeros meses del año.

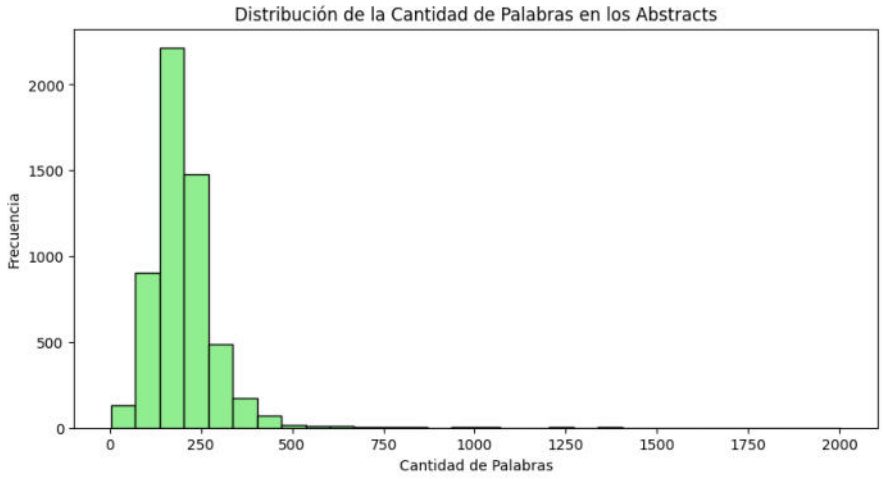
  - Desplazamiento de Enfoque: Puede ser que el interés en el tema esté disminuyendo, o que los investigadores estén explorando nuevas áreas relacionadas.

- Impacto de la Pandemia (2020-2021): Los años 2020 y 2021 muestran una cierta disminución en el número de las publicaciones, lo cual podría estar relacionado con la pandemia de COVID-19, que afectó significativamente la investigación y publicación académica. Durante este período, muchos sectores, incluido el turismo, enfrentaron desafíos que pudieron haber impactado la producción académica.

Distribución de cantidad de palabras

Figura 6

Distribución de cantidad de palabras en los abstracts



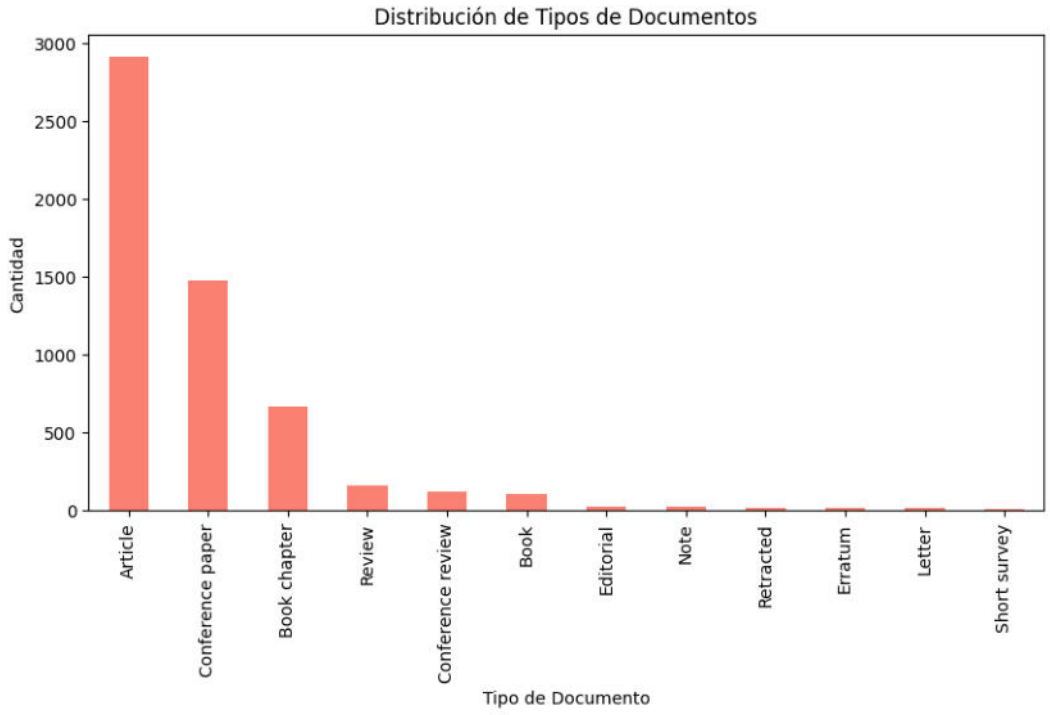
La distribución de longitud de los abstracts en este análisis sugiere que la comunidad académica en el campo tiende a producir resúmenes que son concisos y que cumplen con los límites de palabras recomendados, permitiendo una rápida comprensión de los objetivos y resultados de la investigación. La mayoría de los abstracts contienen entre 150 y 350 palabras. Esto sugiere un estándar en la longitud de los resúmenes, posiblemente determinado por los requisitos de las revistas o conferencias académicas.

Observamos un pico en la frecuencia entre 200 y 250 palabras. Esto indica que muchos autores eligen mantener sus resúmenes en ese rango de palabras, lo cual puede responder a normas editoriales o a un intento de sintetizar la investigación de manera concisa y efectiva. La distribución muestra un sesgo con una cola orientada hacia la derecha, lo cual significa que, a pesar de que la mayoría de los abstracts están en el rango estándar, existen algunos con una mayor cantidad de palabras, llegando incluso a superar las 1000. Estos casos podrían ser excepcionales y responder a investigaciones con mayores detalles, múltiples hipótesis o explicaciones complejas.

Distribución de los Tipos de Documentos

Figura 7

Distribución de los tipos de documentos



La mayoría de los documentos en el dataset son artículos ("Article"), con una gran diferencia frente a otros tipos de documentos como "Conference paper" y "Book chapter". Los artículos representan la clase más dominante, superando significativamente en cantidad a todas las demás categorías. Un patente desbalance de clases es evidenciado, podría influir en

el entrenamiento de los modelos de aprendizaje automático si se utiliza esta clasificación. Los modelos tienden a inclinarse hacia la clase mayoritaria, lo que reduce su potencial para realizar una predicción de las clases menos representadas.

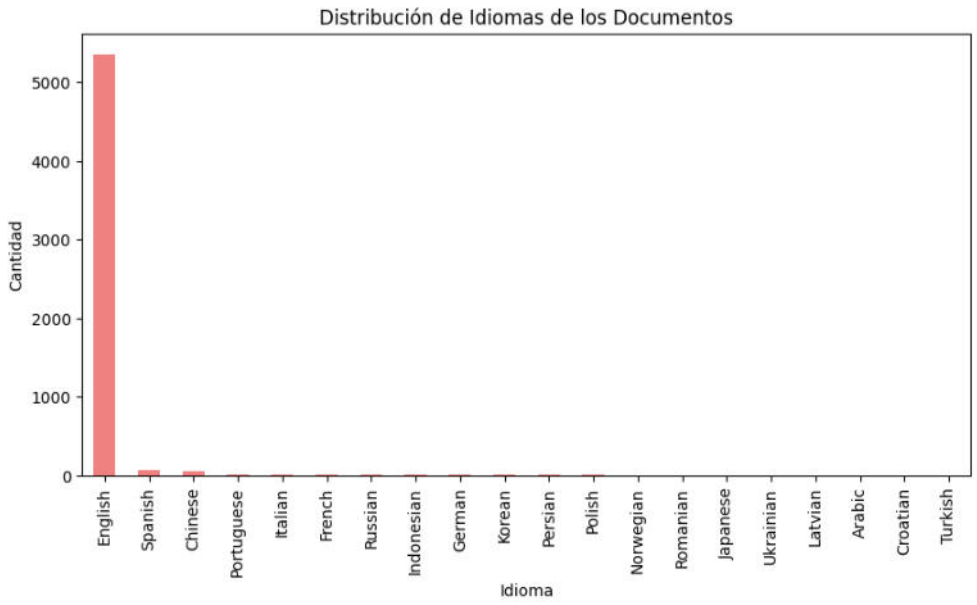
Posibles Estrategias para abordar el desbalance

- **Submuestreo de la Clase Mayoritaria**: Reducir la cantidad de artículos ("Article") en el dataset, seleccionando aleatoriamente una muestra de estos, para equilibrar su proporción con las otras categorías.
- **Aumento de las Clases Minoritarias**: Utilizar técnicas de aumento de datos, como generación sintética de datos o duplicación de documentos de las clases menos representadas, para incrementar su proporción.
- **Pesos de Clase en Modelos de ML**: Ajustar los pesos de las clases en función de su distribución para penalizar errores en las clases minoritarias de manera más significativa. Esto obligaría al modelo a prestar más atención a las clases menos comunes.
- **Agrupación de Clases**: Otra opción podría ser combinar algunas de las clases menos comunes en una categoría única de "Otros" para simplificar el análisis y reducir el impacto del desbalance extremo.

Distribución de Idiomas

Figura 8

Distribución de Idiomas

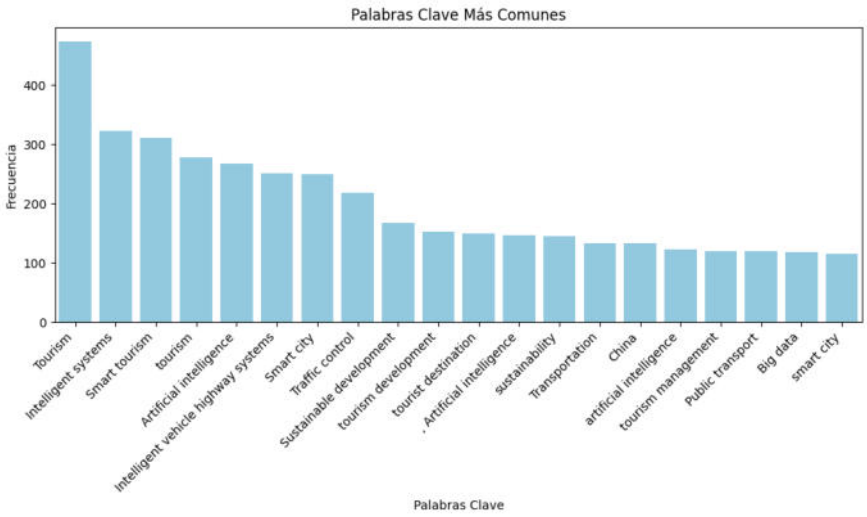


La gran mayoría de los documentos están en inglés, lo que indica que el inglés es el idioma dominante en el campo de investigación. Otros idiomas, como el chino y el español, tienen una representación significativamente menor, seguidos por otros idiomas con aún menor frecuencia. La predominancia del inglés como idioma de publicación puede sesgar los resultados del análisis hacia una perspectiva cultural anglófona.

Frecuencia de palabras clave

Figura 9

Distribución palabras claves



Las palabras clave más frecuentes reflejan temas relacionados con **sistemas inteligentes**, **turismo inteligente**, y **desarrollo sostenible**. Esto indica un fuerte interés en la integración de tecnología avanzada, como **inteligencia artificial** y **sistemas inteligentes de vehículos**, en el contexto del turismo. Estos temas sugieren una intersección entre turismo y tecnología, así como un enfoque en la sostenibilidad.

Además del turismo y la sostenibilidad, palabras clave como **control de tráfico** y **ciudades inteligentes** sugieren un enfoque en la infraestructura y el manejo de recursos para facilitar el turismo en entornos urbanos.

En cuanto a la **distribución Geográfica y Específica**. Palabras como **China** y **transporte público** figuran entre las más comunes, lo que señala un interés tanto en el contexto geográfico como en el transporte en el turismo. Esto puede indicar la relevancia de estas áreas en el desarrollo de soluciones tecnológicas y sostenibles.

Estrategias para Profundizar el Análisis:

- **Agrupación Semántica**: Agrupar términos similares, como aquellos relacionados con **tecnología** y **turismo inteligente**, puede ayudar a identificar patrones más amplios y reducir redundancia en el análisis de palabras clave.
- **Análisis de Bigrama y Trigrama**: Dado que algunas palabras clave aparecen en combinaciones (como "smart tourism" y "sustainable development"), un análisis de bigramas y trigramas permitirá descubrir relaciones temáticas más profundas y mejorar la comprensión de las tendencias en la investigación.
- **Segmentación Temporal o Geográfica**: Analizar cómo cambian estas palabras clave con el tiempo o en distintas regiones podría proporcionar una perspectiva sobre la evolución de los temas y su relevancia en contextos específicos.

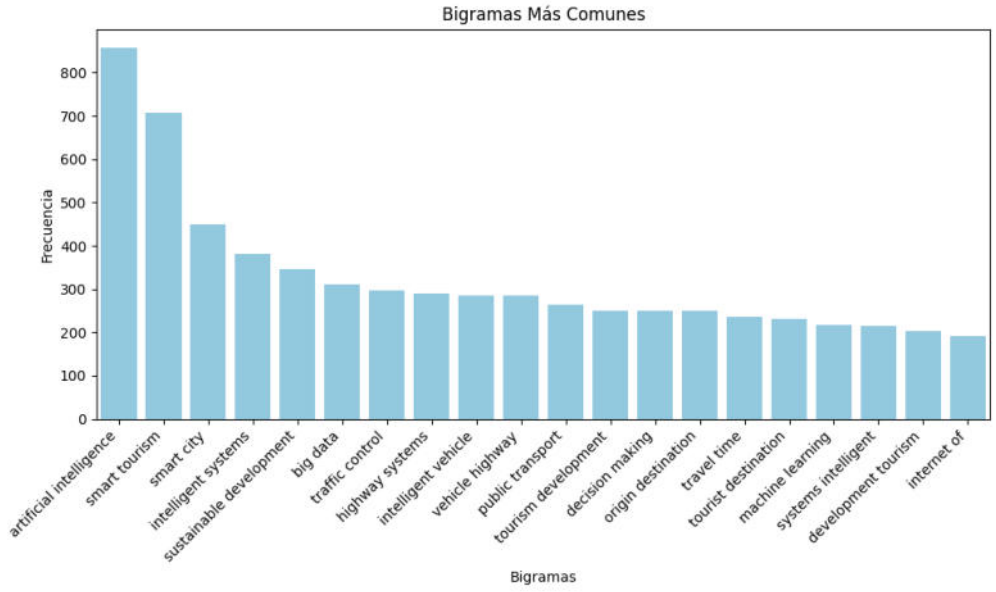
La distribución de las palabras clave indica un enfoque en **turismo inteligente** y **sistemas tecnológicos aplicados al turismo**, así como en la **sostenibilidad** y el **control de infraestructuras urbanas**. Este análisis inicial de palabras clave aporta una visión general de las tendencias temáticas en el campo, pero podría beneficiarse de técnicas de agrupación avanzada y de un análisis de bigramas y trigramas para explorar más a fondo las conexiones entre temas.

Análisis de Bigramas y Trigramas

Análisis de Bigramas

Figura 10

Análisis de Bigramas



Los bigramas más frecuentes, como "artificial intelligence", "smart tourism", y "smart city", reflejan una alta prevalencia de términos relacionados con tecnología avanzada aplicada al turismo y el desarrollo de ciudades inteligentes. Estos resultados subrayan el enfoque en el uso de inteligencia artificial y sistemas inteligentes dentro del contexto del turismo y la gestión urbana.

La presencia de bigramas como "sustainable development", "traffic control", y "public transport" indica que, además del turismo inteligente, la sostenibilidad y la infraestructura son temas relevantes en la investigación académica. Estos términos sugieren un interés en el impacto de la tecnología en la sostenibilidad y en cómo gestionar eficientemente los recursos en entornos turísticos.

La inclusión de términos como "machine learning" y "big data" destaca el crecimiento en el uso de técnicas de análisis avanzadas para abordar problemas en el turismo y la sostenibilidad. Esto refleja una tendencia hacia el uso de datos y tecnologías emergentes para mejorar la experiencia turística y la eficiencia operativa en ciudades.

La alta frecuencia de bigramas relacionados con tecnología y sostenibilidad podría generar redundancia en el análisis. Esto sugiere la necesidad de agrupar o normalizar términos similares en futuras investigaciones, con el fin de reducir la repetición y obtener una visión más clara de los temas predominantes.

Estrategias para el Análisis

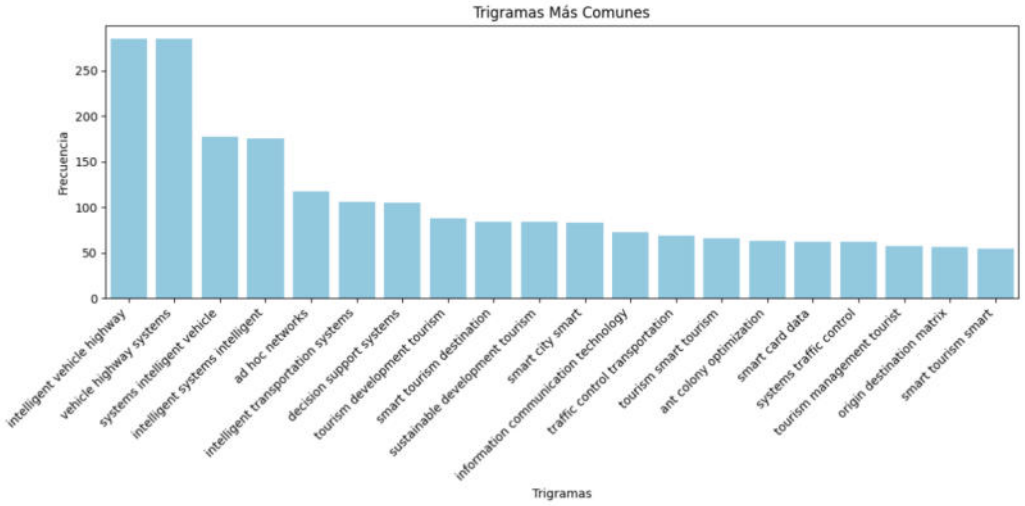
- **Análisis de Trigramas**: Ampliar el análisis a trigramas permitirá captar combinaciones de temas más específicas y complejas, como "smart tourism management" o "sustainable urban development", lo cual podría ofrecer una visión más detallada de las interrelaciones entre temas.
- **Segmentación Temática**: Dado el enfoque en sostenibilidad, tecnología avanzada y gestión urbana, se podrían crear subgrupos temáticos para analizar cada área en profundidad, permitiendo así explorar las interacciones entre tecnología y sostenibilidad en el turismo.
- **Análisis Temporal**: Examinar cómo cambian estos bigramas en el tiempo podría revelar tendencias emergentes y mostrar cómo han evolucionado los temas de investigación a lo largo de los años en respuesta a los avances tecnológicos y los desafíos ambientales.

En conclusión, el análisis de bigramas confirma que los estudios de turismo están centrados en temas de sostenibilidad y tecnología avanzada. Los bigramas dominantes sugieren una tendencia hacia la adopción de inteligencia artificial, el desarrollo de ciudades inteligentes, y la gestión de recursos sostenibles. Estos hallazgos ofrecen una base sólida para orientar futuras investigaciones y destacan la necesidad de un enfoque más estructurado en la segmentación temática para minimizar la redundancia y profundizar en las relaciones entre temas clave.

Análisis de Trigramas

Figura 11

Análisis Trigramas



El gráfico de trigramas muestra combinaciones de tres palabras clave que ocurren con mayor frecuencia en los resúmenes de los artículos analizados. A continuación, se destacan algunas de las combinaciones más relevantes:

- **Principales Temáticas:** Los trigramas más destacados están relacionados con términos como *"vehicle highway systems"*, *"intelligent vehicle highway"*, y *"systems intelligent vehicle"*. Esto refleja un enfoque significativo en la intersección entre sistemas de transporte inteligente y vehículos autónomos, indicando el interés en la infraestructura y gestión de sistemas de transporte avanzados.
- **Tecnología y Turismo Inteligente:** Otros trigramas relevantes como *"smart tourism development"*, *"sustainable development tourism"*, y *"smart city smart"* sugieren que los estudios también están explorando cómo las tecnologías avanzadas y las ciudades

inteligentes están afectando al desarrollo del turismo. Este énfasis en el "turismo inteligente" subraya la transformación digital en el sector turístico.

- **Evaluación y Gestión del Transporte:** Trigramas como *"traffic control transportation"* y *"tourism traffic control"* destacan el interés en la gestión y control del tráfico, especialmente en áreas con alto flujo turístico. Esto sugiere que una parte de la investigación está orientada a optimizar el flujo de visitantes en destinos turísticos para reducir la congestión y mejorar la experiencia del turista.
- **Sistemas de Apoyo a la Decisión**: La presencia de trigramas como *"decision support systems"* indica el uso de sistemas de apoyo a la decisión en la planificación y gestión del turismo y el transporte. Estos sistemas ayudan a las autoridades y empresas a tomar decisiones informadas en tiempo real.

Implicaciones para la Investigación

Este análisis de trigramas revela áreas específicas dentro del turismo y el transporte inteligente que están siendo exploradas con mayor profundidad:

- **Turismo Inteligente y Desarrollo Sostenible:** Existe un interés notable en cómo las ciudades inteligentes y las tecnologías avanzadas pueden contribuir al turismo sostenible y a una mejor experiencia del turista.
- **Control y Gestión del Tráfico:** La gestión del tráfico es una preocupación central, lo cual es esencial para destinos turísticos populares que buscan equilibrar la afluencia de visitantes con la sostenibilidad.
- **Sistemas de Apoyo a la Decisión:** La implementación de sistemas de apoyo a la decisión se está convirtiendo en una herramienta valiosa para gestionar eficazmente los recursos en entornos turísticos.
- **Infraestructura Inteligente y Vehículos Autónomos:** La infraestructura de sistemas de transporte inteligente y vehículos autónomos está recibiendo

atención significativa, lo cual puede transformar el turismo y la movilidad en el futuro.

Este análisis de trigramas proporciona una comprensión más clara de cómo los temas tecnológicos y de sostenibilidad convergen en la investigación del turismo, ofreciendo perspectivas que pueden guiar futuras investigaciones y desarrollos en el campo.

Definición de Variables de Clasificación

En esta fase se aplica el algoritmo LDA (Latent Dirichlet Allocation) para la búsqueda de los temas latentes de los resúmenes, realizando un análisis de perplejidad y coherencia semántica que permita determinar el número óptimo de tópicos.

Análisis de Perplejidad

Para determinar el número óptimo de temas en nuestro análisis de tópicos, hemos realizado un análisis de perplejidad. La perplejidad es una métrica que evalúa qué tan bien un modelo LDA se ajusta a los datos, donde valores más bajos indican un mejor ajuste.

Figura 12

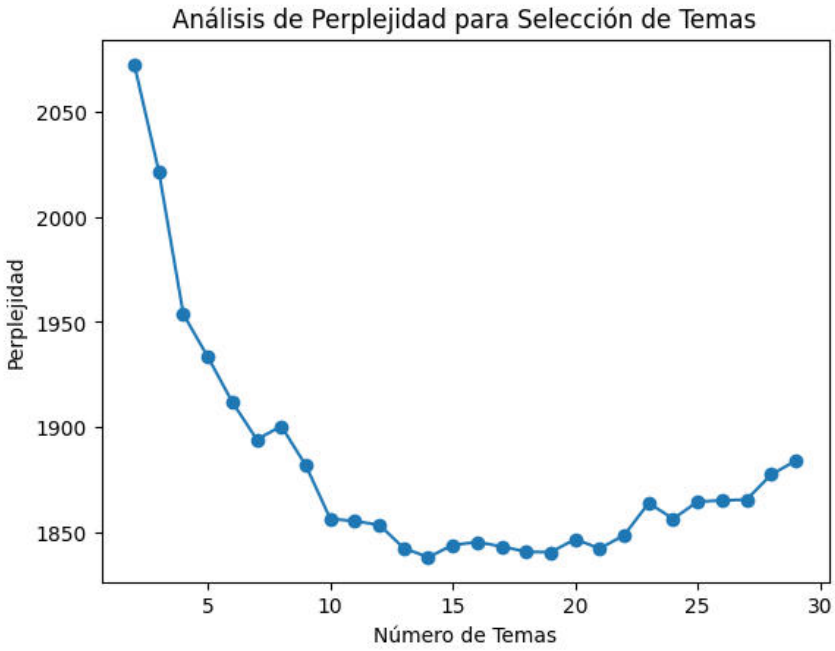
Código para análisis de perplejidad

```
from sklearn.decomposition import LatentDirichletAllocation
from sklearn.feature_extraction.text import CountVectorizer
import matplotlib.pyplot as plt

# Preparación de los datos de texto
text_data = df_reduced['Abstract'] + " " + df_reduced['Keywords']
vectorizer = CountVectorizer(max_features=10000, stop_words='english')
text_vectors = vectorizer.fit_transform(text_data)

# Prueba de diferentes números de temas y cálculo de perplejidad
perplexity_scores = []
topic_range = range(2, 30)  # Rango de temas a probar
for num_topics in topic_range:
    lda_model = LatentDirichletAllocation(n_components=num_topics, random_state=42)
    lda_model.fit(text_vectors)
    perplexity = lda_model.perplexity(text_vectors)
    perplexity_scores.append(perplexity)
    print(f"Perplejidad para {num_topics} temas: {perplexity}")

# Graficar la perplejidad para observar el punto de inflexión
plt.plot(topic_range, perplexity_scores, marker='o')
plt.xlabel("Número de Temas")
plt.ylabel("Perplejidad")
plt.title("Análisis de Perplejidad para Selección de Temas")
plt.show()
```

Hemos calculado la perplejidad para un rango de temas de 2 a 29 y observado cómo cambia la métrica en función del número de temas.

Figura 13

Variación de métricas del análisis de perplejidad



Observaciones y Justificación del Número de Temas

- **Punto de Inflexión**: Observamos que la perplejidad disminuye significativamente al aumentar el número de temas hasta aproximadamente 10 temas, lo que indica una mejora sustancial en el ajuste del modelo. Sin embargo, después de este punto, la reducción de la perplejidad se vuelve menos pronunciada.
- **Estabilización y Aumento Posterior**: A partir de 10 a 15 temas, la perplejidad se estabiliza e incluso muestra una ligera tendencia al aumento, sugiriendo que añadir más temas no proporciona una mejora significativa en la calidad del modelo y podría introducir ruido innecesario.
- **Número Óptimo de Temas**: Con base en estos resultados, hemos determinado que el rango de 10 a 15 temas captura adecuadamente la variabilidad en nuestros datos sin añadir complejidad innecesaria. Este rango ofrece un buen equilibrio entre

interpretabilidad y precisión, siendo particularmente útil para un análisis académico que busca identificar tendencias significativas sin sobrecargar el modelo.

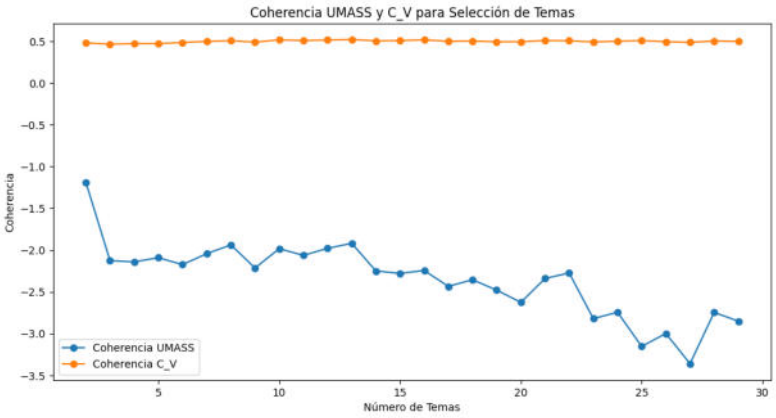
Sin embargo, como la perplejidad solo mide la capacidad del modelo para generalizar, no asegura coherencia semántica. Por eso, usaremos métricas de coherencia como UMASS o C_V para validar el mejor número de temas.

Calcular coherencia semántica con métricas UMASS y C_V

La métrica UMASS es más adecuada para corpus más pequeños y mide la coherencia semántica dentro de los temas considerando la probabilidad de co-ocurrencia de palabras clave en los documentos. Mientras que la métrica C_V evalúa la similitud semántica de las palabras clave dentro de los temas, siendo robusta y confiable para seleccionar un número óptimo de tópicos en aplicaciones prácticas. Para nuestro caso:

Figura 14

Comparaciones métricas UMASS y C_V



UMASS:

- Los valores son más altos (menos negativos) en el rango de 8-13 temas.
- El punto más alto de UMASS es para **13 temas (-1.918)**, indicando que este número ofrece un balance adecuado entre coherencia interna y capacidad del modelo para generalizar.
- A medida que el número de temas aumenta más allá de 13, UMASS se deteriora significativamente, lo que sugiere que más temas resultan en divisiones que no son suficientemente coherentes dentro de los datos.

C_V:

- El valor más alto de **C_V (0.521)** también ocurre con **13 temas**, reforzando la sugerencia de que este es el número óptimo para capturar tendencias coherentes.
- El rendimiento de C_V en el rango de 10 a 16 temas es relativamente estable, pero alcanza su pico en 13. Más allá de 16 temas, el valor comienza a decaer, lo que podría significar una pérdida de coherencia semántica.
- Según el gráfico de **perplejidad**, un punto de inflexión claro ocurrió alrededor de 12-15 temas.

Este análisis complementa la selección de 13 temas, ya que tanto UMASS como C_V apuntan a esta zona como ideal para maximizar coherencia y generalización.

Con base en los resultados de perplejidad, coherencia UMASS y C_V:

- **Equilibrio entre generalización y coherencia semántica:**
    1. 13 temas maximizan la coherencia dentro de los tópicos (UMASS) y entre palabras clave (C_V).
    2. Este número mantiene una representación compacta pero significativa de las tendencias de investigación, evitando redundancia y división innecesaria de temas.
- **Relevancia práctica:**
    3. Un número más pequeño de temas (como 8-10) podría simplificar la interpretación, pero corre el riesgo de agrupar conceptos distintos bajo una misma categoría, perdiendo detalles importantes.
    4. Por otro lado, un número mayor de temas (como 16 o más) generaría tópicos más específicos, pero con menor coherencia, dificultando la visualización y comprensión para los usuarios finales.
- **Validación cruzada con el análisis de perplejidad:**
    1. La perplejidad mostró un punto de inflexión entre 12 y 15 temas. Esto valida que la selección de 13 temas es adecuada para capturar tendencias sin sobredimensionar el modelo.

Implementación LDA

Figura 15

Código implementación LDA

```python
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation

# Preprocesamiento: Combinar los textos de Abstract y Keywords
text_data = df_reduced['Abstract'] + " " + df_reduced['Keywords']

# Vectorización: Crear una matriz de términos utilizando CountVectorizer
vectorizer = CountVectorizer(max_features=10000, stop_words='english')
text_vectors = vectorizer.fit_transform(text_data)

# LDA: Configuramos el número de temas
num_topics = 13  # Número de temas basado en el análisis de coherencia
lda_model = LatentDirichletAllocation(n_components=num_topics, random_state=42)
lda_model.fit(text_vectors)
```

La distribución muestra un claro desbalance, con algunos temas dominando (por ejemplo, Tema 9 con 1085 documentos) y otros siendo muy minoritarios (Tema 7 con solo 56 documentos).

Esto podría afectar la calidad del modelo LSTM, ya que un desbalance significativo en las clases tiende a hacer que el modelo aprenda mejor las clases dominantes, ignorando o teniendo bajo desempeño en las clases minoritarias. Además, temas con baja representación pueden no aportar suficiente información semántica, dificultando su interpretación o utilidad en el análisis de tendencias. Por lo tanto, proponemos una reducción de la dimensionalidad mediante la fusión de temas relacionados, lo cual no solo equilibra mejor las clases, sino que también simplifica la estructura del modelo, manteniendo su capacidad para identificar tendencias significativas en la investigación turística.

Tabla 3

Distribución de Temas

Topic	Label
9	1085
0	1058
2	647
10	578
1	412
8	403
4	327
11	313
12	262
5	153
3	125
6	106
7	56

Esta decisión se basa en:

1. **Reducción del desbalance:** Al combinar temas con pocos documentos, se logra una distribución más uniforme entre las clases.
2. **Coherencia semántica:** Agrupamos temas con palabras clave y enfoques similares, asegurando que las combinaciones representen tendencias claras y significativas.
3. **Mejoramiento del rendimiento del modelo:** Al trabajar con clases más balanceadas, el modelo LSTM podrá generalizar mejor y ofrecer resultados más robustos, reduciendo el sesgo hacia clases dominantes.

Figura 16

Código para reducir temas

```
# Función para combinar temas relacionados
def combine_topics(row):
    if row['Topic_Label'] in [6, 7]:  # Fusionar Temas 6 y 7
        return 'Tech_Communication'
    elif row['Topic_Label'] in [3, 10]:  # Fusionar Temas 3 y 10
        return 'Tourism_Experience'
    elif row['Topic_Label'] in [8, 12]:  # Fusionar Temas 8 y 12
        return 'Urban_Sustainability'
    elif row['Topic_Label'] in [4, 11]:  # Fusionar Temas 4 y 11
        return 'Data_Analysis'
    elif row['Topic_Label'] == 9:  # Tema 9 sin cambios
        return 'Sustainability'
    elif row['Topic_Label'] == 13:  # Tema 13 sin cambios
        return 'Traffic_Models'
    elif row['Topic_Label'] == 1:  # Tema 1 sin cambios
        return 'AI_Tourism'
    elif row['Topic_Label'] == 2:  # Tema 2 sin cambios
        return 'Smart_Mobility'
    elif row['Topic_Label'] == 5:  # Tema 5 sin cambios
        return 'Algorithmic_Models'
    else:
        return f'Tema_{row["Topic_Label"]}'  # Para casos no previstos

# Aplicar la fusión de temas al DataFrame
df_reduced['Combined_Topic'] = df_reduced.apply(combine_topics, axis=1)

# Verificar la nueva distribución de temas
print("\nNueva distribución de documentos por tema:")
print(df_reduced['Combined_Topic'].value_counts())

# Inspeccionar ejemplos representativos de cada tema
for topic in df_reduced['Combined_Topic'].unique():
    print(f"\nEjemplo de documentos en {topic}:")
    print(df_reduced[df_reduced['Combined_Topic'] == topic].head(3)['Abstract'])
```

Tabla 4

Distribución por tema agrupado

Topic	Cantidad
Sustainability	1085
Tema_0	1058
Tourism_Experience	703
Urban_Sustainability	665
Smart_Mobility	647
Data_Analysis	640
AI_Tourism	412
Tech_Communication	162
Algorithmic_Models	153

Temas Seleccionados para el Análisis

Para capturar la diversidad de la investigación en turismo, los 9 temas identificados en el análisis previo se incluyen en el análisis, proporcionando una cobertura exhaustiva de las áreas de interés en turismo. Estos temas representan áreas fundamentales, tales como:

a. **Sostenibilidad en Turismo Urbano y Rural**:

Temas como "Urban Sustainability" y "Rural Resource Management" abordan la implementación de prácticas sostenibles tanto en áreas urbanas como rurales, reflejando cómo el turismo puede adaptarse a las necesidades medioambientales en distintos contextos.

b. **Movilidad y Transporte en Destinos Turísticos**:

La movilidad es esencial en la gestión de flujos turísticos. Temas como "Mobility and Transportation" exploran el impacto de los sistemas de transporte en la experiencia del turista y en la reducción de la huella de carbono en destinos turísticos.

c. **Tecnología y Comportamiento Turístico**:

La incorporación de inteligencia artificial, realidad aumentada, y tecnología de comunicación en el turismo permite mejorar la experiencia del turista y gestionar de forma más eficaz los destinos. Temas como "Digital Tourism" y "Technology in Hospitality" cubren estos avances.

d. **Educación y Gestión del Conocimiento en Turismo**:

La educación en turismo y la gestión del conocimiento son fundamentales para el desarrollo de futuros profesionales en esta industria. Temas como "Knowledge Management in Tourism" ofrecen una perspectiva sobre la formación y la innovación en turismo.

e. **Salud, Seguridad, y Adaptación a Cambios Globales**:

En un mundo post-pandémico, la relación entre turismo, salud, y tecnología es cada vez más relevante. El análisis de temas como "Health and Tourism During Pandemics"

proporciona insights sobre cómo la industria se adapta a desafíos globales y aplica soluciones tecnológicas.

Tabla 5

Temas resultantes

Tema	Palabras Clave	Descripción	Cantidad de Artículos
1. Sostenibilidad y Turismo Rural	sustainability, climate, green, tourism, rural, environmental, economic, sustainable, development	Aborda la sostenibilidad en el turismo rural, destacando aspectos ambientales, económicos y sociales. Examina estrategias para mitigar el cambio climático y fomentar el turismo verde.	1085
2. Movilidad Inteligente	trip, transit, mobility, destination, smart, public, data, transportation, travel, transport	Se enfoca en la movilidad inteligente aplicada al turismo, explorando soluciones de transporte público, gestión del tránsito y movilidad sostenible.	647
3. Análisis de Datos y Sistemas Inteligentes	language, mining, medical, based, social, learning, information, analysis, big, data	Examina el uso de sistemas inteligentes y minería de datos para extraer información relevante en el turismo. Incluye aplicaciones en análisis de grandes datos y aprendizaje basado en información.	640
4. Experiencias Turísticas y Destinos Culturales	satisfaction, perceived, social, travel, smart, destination, tourist, tourists, study, tourism	Analiza las experiencias de los turistas en destinos culturales, destacando la percepción, satisfacción y aspectos sociales en el contexto del turismo inteligente.	703

5. Modelos Algorítmicos	travel, intelligence, model, tourist, algorithm, systems, information, decision, based, tourism	Se centra en el desarrollo de modelos algorítmicos y de decisión basados en sistemas inteligentes para mejorar el turismo y la experiencia del usuario.	153
6. Tecnologías Avanzadas de Comunicación	hoc, ad, wireless, nodes, power, communication, network, energy, networks, routing	Explora tecnologías de comunicación avanzadas como redes inalámbricas y protocolos de enrutamiento, destacando su impacto en el sector turístico.	162
7. Turismo Inteligente y Gestión Urbana	information, life, solid, city, intelligence, disaster, smart, management, waste, parking	Aborda la gestión urbana inteligente, con énfasis en la sostenibilidad de las ciudades, el manejo de residuos sólidos y la inteligencia aplicada a la administración de desastres y el tráfico.	665
8. Turismo Inteligente Basado en IA	service, artificial, technology, study, hospitality, intelligence, industry, ai, research, tourism	Se centra en la aplicación de la inteligencia artificial y la tecnología en el turismo, incluyendo la hospitalidad y la investigación en la industria turística.	412
9. Modelos de Tráfico y Sistemas Inteligentes	control, network, destination, road, model, vehicle, intelligent, systems, time, traffic	Examina modelos inteligentes para la gestión del tráfico, integrando redes de transporte, vehículos autónomos y sistemas de control para destinos turísticos.	1058

Justificación del Enfoque

- **Cobertura Integral de la Investigación en Turismo**: La elección de incluir los 9 temas permite un análisis holístico de la investigación en turismo. Al no limitar el enfoque y tampoco generalizar o provocar problemas de generalización en el modelo, obtenemos una representación más fiel de las áreas de interés en la academia, lo cual enriquece el análisis y proporciona un contexto más amplio.
- **Identificación de Intersecciones y Colaboraciones Potenciales**: El enfoque general permite detectar posibles puntos de conexión entre temas diversos, como la tecnología aplicada a la sostenibilidad, o la movilidad influenciada por el comportamiento del turista. Estos hallazgos pueden ser valiosos para investigadores que busquen colaborar en temas interdisciplinarios.
- **Mayor Relevancia para Políticas y Prácticas de la Industria**: Al analizar una gama amplia de temas, los resultados pueden ser útiles para un público más amplio, incluyendo responsables de políticas, empresas, y gestores de destinos que necesitan tener en cuenta múltiples aspectos del turismo, desde la gestión de recursos hasta la tecnología y el comportamiento del turista.
- **Flexibilidad para el Análisis de Tendencias Específicas**: Aunque el enfoque es general, el modelo permite filtrar y analizar tendencias específicas dentro de cada tema cuando sea necesario. Esto significa que, si un tema como "Sostenibilidad

Urbana" o "Movilidad Turística" requiere atención especial, el modelo puede proporcionar insights profundos y detallados.

Refinamiento del Corpus

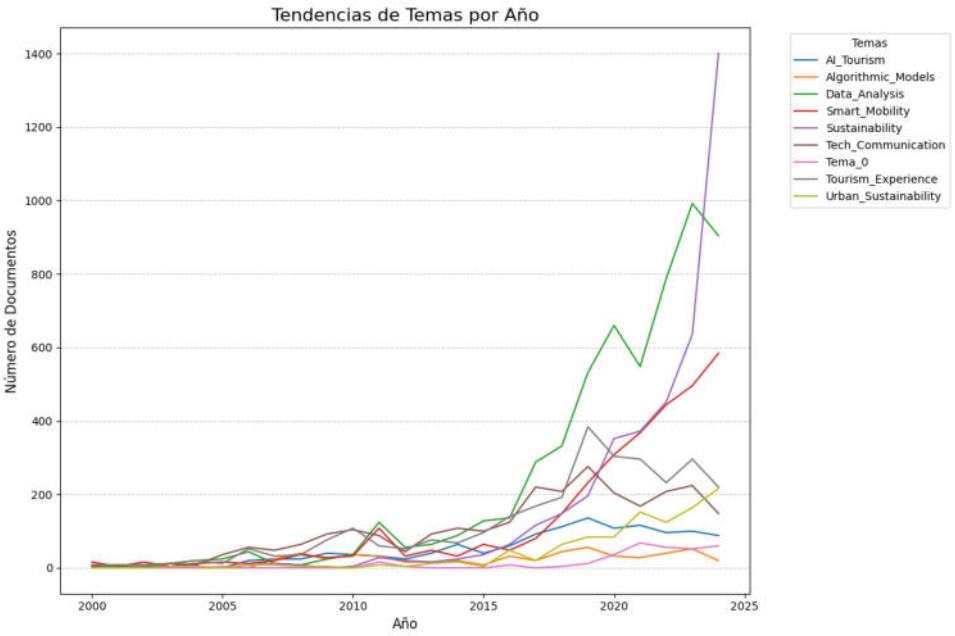
Dado que se van a utilizar los 9 temas completos, es esencial realizar un **refinamiento del corpus**. Esto implica:

1. **Limpieza de Datos**: Eliminar duplicados, resúmenes incompletos o mal formateados, y artículos que no se alineen claramente con los temas.
2. **Ajuste de Longitud de Textos**: Para una mejor tokenización, limitar o expandir la longitud de los resúmenes de acuerdo con un rango óptimo, evitando textos extremadamente cortos o largos que puedan distorsionar el análisis.
3. **Balance de Temas**: Si algunos temas tienen muchos menos artículos que otros, podrías aplicar técnicas como el sobre muestreo o submuestreo en los datos para evitar sesgos en el modelo.
4. **Etiquetado Consistente**: Asegurar que cada artículo esté etiquetado correctamente según el tema correspondiente, y revisar que cada tema tenga una representación significativa en el corpus final.

Este refinamiento ayudará a mejorar la calidad del entrenamiento del modelo y permitirá un análisis más preciso y representativo de las tendencias en la investigación de turismo. Con estos pasos, se garantiza que el análisis no solo sea extenso, sino también profundo y relevante para la academia y la industria.

Figura 17

Tendencias por año luego de refinamiento



Finalmente realizamos un guardado del corpus limpio y procesado.

Figura 18

Proceso de guardado de corpus

```
# Guardar el DataFrame aumentado en el mismo archivo para consistencia
output_dir = 'Clean_Dataset'
os.makedirs(output_dir, exist_ok=True)

# Guardar el dataset final con un comentario en el código
output_path = os.path.join(output_dir, 'cleaned_corpus_with_topics.csv')
# Comentario: Este dataset contiene los datos originales más las instancias aumentadas
df_augmented.to_csv(output_path, index=False)

print(f"Corpus limpio, procesado y aumentado guardado en '{output_path}'")
```

2.3.4 Implementación del Modelo de Clasificación

Se divide el dataset en conjuntos de entrenamiento, validación y prueba. Se aplica un modelo LSTM (Long Short-Term Memory) para la clasificación de los resúmenes, donde se optimizan los hiperparámetros por medio de Optuna.

Fase de segmentación y preparación de datos

Dado que tenemos 5427 abstracts, vamos a realizar la segmentación en conjuntos de **entrenamiento**, **validación** y **prueba**.

Definimos el tamaño de los conjuntos de la siguiente manera:

- **Entrenamiento (80%)**: ~4340 abstracts.
- **Validación (10%)**: ~543 abstracts.
- **Prueba (10%)**: ~543 abstracts.

Con base en lo anterior tenemos nuestros datos de la siguiente manera:

- Tamaño de entrenamiento: 17363
- Tamaño de validación: 2170
- Tamaño de prueba: 2171

Procesamiento de texto y tokenización con textvectorization

Se implementa una función para la tokenización y aplicamos TextVectorization en TensorFlow nos permite transformar los textos en secuencias de tokens numéricos, lo cual es fundamental para el entrenamiento de redes neuronales.

Figura 19

Función para generar los tokens

```python
# Función para tokenización simple y preprocesamiento
def preprocess_text(text):
    # Convertir a minúsculas
    text = text.lower()
    # Eliminar puntuación con regex
    text = re.sub(r'[^\w\s]', '', text)
    # Tokenizar dividiendo por espacios
    tokens = text.split()
    # Eliminar stopwords
    tokens = [word for word in tokens if word not in stop_words]
    # Aplicar stemming
    tokens = [stemmer.stem(word) for word in tokens]
    # Reconstruir texto procesado
    return ' '.join(tokens)
```

Figura 20

Código para aplicar tokens

```python
# Aplicar preprocesamiento a los textos
train_df['Text'] = train_df['Text'].apply(preprocess_text)
val_df['Text'] = val_df['Text'].apply(preprocess_text)
test_df['Text'] = test_df['Text'].apply(preprocess_text)

# Escalar los años a un rango entre 0 y 1
scaler = MinMaxScaler()
train_df['Year_Scaled'] = scaler.fit_transform(train_df[['Year']])
val_df['Year_Scaled'] = scaler.transform(val_df[['Year']])
test_df['Year_Scaled'] = scaler.transform(test_df[['Year']])

# Extraer textos, etiquetas y años procesados
train_texts, train_labels, train_years = train_df['Text'], train_df['Label'], train_df['Year_Scaled']
val_texts, val_labels, val_years = val_df['Text'], val_df['Label'], val_df['Year_Scaled']
test_texts, test_labels, test_years = test_df['Text'], test_df['Label'], test_df['Year_Scaled']

# Tokenización usando Keras
tokenizer = Tokenizer(num_words=10000, oov_token="<OOV>")
tokenizer.fit_on_texts(train_texts)

# Convertir textos en secuencias
train_sequences = tokenizer.texts_to_sequences(train_texts)
val_sequences = tokenizer.texts_to_sequences(val_texts)
test_sequences = tokenizer.texts_to_sequences(test_texts)

# Rellenar secuencias para igualar longitudes
max_sequence_length = 200
train_padded = pad_sequences(train_sequences, maxlen=max_sequence_length, padding='post', truncating='post')
val_padded = pad_sequences(val_sequences, maxlen=max_sequence_length, padding='post', truncating='post')
test_padded = pad_sequences(test_sequences, maxlen=max_sequence_length, padding='post', truncating='post')
```

Crear batches y preparar para entrenamiento

Para un entrenamiento eficiente, utilizamos tf.data para crear batches. Esto reduce la carga en la memoria y optimiza el rendimiento del modelo.

En este primer punto del proyecto, completamos los siguientes pasos clave para la preparación de datos, orientados a construir un modelo de RNN capaz de identificar tendencias de investigación en el campo del turismo:

1. **Segmentación de Datos**: Dividimos nuestro conjunto de datos en tres subconjuntos: **entrenamiento** (80%), **validación** (10%) y **prueba** (10%). Esta división nos permite entrenar el modelo, optimizar su rendimiento y finalmente evaluarlo con datos que no ha visto antes.
2. **Tokenización**: Convertimos el texto de los resúmenes (abstracts) en **secuencias de tokens**. Para ello, construimos un vocabulario que representa las palabras y términos clave más comunes en el corpus, permitiendo al modelo interpretar y procesar el texto de manera numérica.
3. **Padding y Truncamiento**: Normalizamos las longitudes de las secuencias para ajustarlas a una longitud estándar definida. Esto se realizó mediante **padding** (agregar ceros al final de secuencias cortas) y **truncamiento** (acortar secuencias largas), asegurando que todas las secuencias tengan la misma longitud y sean compatibles con el modelo.
4. **Batching de los Datos**: Organizamos los datos en batches para alimentar al modelo en grupos, lo cual optimiza el uso de memoria y acelera el proceso de entrenamiento. Verificamos que los batches se generaran correctamente para los conjuntos de entrenamiento, validación y prueba.

Figura 21

Implementación de proceso para batches

```python
import tensorflow as tf

# Crear datasets de TensorFlow a partir de los datos procesados
batch_size = 32

# Dataset de entrenamiento con texto y año como entradas
train_dataset = tf.data.Dataset.from_tensor_slices(
    ((train_padded, train_years), train_labels)
).batch(batch_size).prefetch(tf.data.AUTOTUNE)

# Dataset de validación
val_dataset = tf.data.Dataset.from_tensor_slices(
    ((val_padded, val_years), val_labels)
).batch(batch_size).prefetch(tf.data.AUTOTUNE)

# Dataset de prueba
test_dataset = tf.data.Dataset.from_tensor_slices(
    ((test_padded, test_years), test_labels)
).batch(batch_size).prefetch(tf.data.AUTOTUNE)

# Verificar los batches
for (text_batch, year_batch), label_batch in train_dataset.take(1):
    print("Ejemplo de batch de textos:", text_batch.numpy())
    print("Ejemplo de batch de años:", year_batch.numpy())
    print("Ejemplo de batch de etiquetas:", label_batch.numpy())
```

Selección de arquitectura para desarrollo de modelo

Este proyecto busca analizar y extraer tendencias de investigación a partir de los abstracts de artículos científicos en el ámbito del turismo, con un enfoque en sostenibilidad. Dados los patrones temporales y las relaciones complejas entre términos en el texto, necesitamos una arquitectura de red neuronal que sea capaz de capturar la **secuencia y dependencia contextual** dentro de los resúmenes. Para ello, se consideran dos opciones populares de arquitecturas de redes recurrentes: **RNN (Redes Neuronales Recurrentes)** y **LSTM (Long Short-Term Memory)**.

Tabla 6

Comparación entre RNN y LSTM

Característica	RNN	LSTM
Estructura	Capa recurrente simple	Capa recurrente con compuertas de entrada, olvido y salida
Manejo de Dependencias a Largo Plazo	Limitado debido a desvanecimiento de gradientes	Excelente, diseñado para retener información a largo plazo y mitigar el desvanecimiento de gradientes
Complejidad Computacional	Menor	Mayor debido a las compuertas adicionales, pero justificada para secuencias complejas
Rendimiento en Texto Extenso	Puede degradarse en secuencias largas	Alto, especialmente adecuado para textos con dependencias a largo plazo
Uso de Memoria	Menor	Mayor, debido a las células de memoria adicionales

Análisis para este proyecto

1. **Secuencias Largas y Dependencias Complejas**:
    - Los abstracts de artículos científicos suelen tener **secuencias de texto largas y estructuras de dependencia complejas**. Los temas y términos clave en un abstract pueden depender de palabras y conceptos introducidos en oraciones previas, lo cual requiere una arquitectura que pueda retener y utilizar esta información a largo plazo.
2. **Evitar Desvanecimiento de Gradientes**:
    - Las RNN convencionales sufren de desvanecimiento de gradientes, lo cual limita su capacidad para aprender relaciones que se extienden a lo largo de muchas palabras. Dado que necesitamos capturar dependencias contextuales en los resúmenes, una **arquitectura LSTM** es más adecuada, ya que sus

compuertas de memoria están diseñadas para **retener información importante y mitigar el problema de desvanecimiento de gradientes**.

3. **Complejidad Computacional y Recursos Disponibles**:
    - Aunque las LSTM son más complejas y demandantes en términos de memoria, nuestro entorno de trabajo (con soporte CUDA en GPU) permite **procesar eficientemente modelos más complejos**. Esto justifica el uso de LSTM, ya que los recursos están disponibles para aprovechar sus ventajas sin un costo significativo en tiempo de entrenamiento.
4. **Capacidades para Identificación de Tendencias**:
    - Las LSTM tienen un rendimiento comprobado en el procesamiento de texto donde se necesita analizar secuencias largas y dependencias temporales. Esto es crucial en nuestro caso, ya que queremos identificar patrones y tendencias de investigación. Usar LSTM aumentará la probabilidad de que el modelo pueda **capturar patrones complejos de lenguaje** relacionados con temas recurrentes, términos de interés y su evolución en el tiempo.

Definición y estudio de hiperparámetros con randomizedsearchcv vs optimización bayesiana

Dado el tipo de datos (**abstracts académicos**) y los objetivos del proyecto, hemos optado por implementar **RandomizedSearchCV** y la **Optimización Bayesiana (Optuna)** como métodos complementarios para buscar los mejores hiperparámetros para un modelo LSTM. Esta decisión se basa en:

Datos:

- **Complejidad semántica:** Los abstracts condensan información rica, por lo que el modelo debe captar relaciones latentes entre palabras y temas.
- **Tamaño manejable (~5000 registros):** Permite explorar diferentes configuraciones de hiperparámetros sin exceder los recursos disponibles.

Recursos:

- **Computo:** Con GPU disponible, podemos probar métodos más avanzados como la optimización bayesiana, sin el costo computacional excesivo de GridSearchCV.
- **Eficiencia:** RandomizedSearchCV equilibra rapidez y exploración, mientras que la optimización bayesiana garantiza exploración más inteligente del espacio de hiperparámetros.

Optimización Bayesiana (Optuna)

En el presente proyecto, cuyo objetivo principal es la **identificación y visualización de tendencias de investigación en el ámbito del turismo** mediante el análisis de resúmenes de artículos científicos, la correcta selección de hiperparámetros juega un papel crucial en el diseño de modelos de aprendizaje profundo. Esto cobra especial relevancia debido a la incorporación de variables temporales como el año de publicación, que agrega una dimensión fundamental para el análisis longitudinal de las tendencias.

Contexto y Problemática

El problema para resolver implica la clasificación temática de los resúmenes y el análisis de patrones temporales para identificar cómo han evolucionado los tópicos en el tiempo. La representación y modelado de estos datos requiere redes neuronales que combinen información textual y temporal, siendo la arquitectura LSTM una opción adecuada debido a su capacidad para capturar dependencias secuenciales en el texto. Sin embargo, el desempeño de estas redes es altamente sensible a la configuración de hiperparámetros como:

- Dimensión de salida de la capa de embeddings.
- Número de unidades LSTM.
- Tasa de aprendizaje del optimizador.
- Parámetros de regularización (como el dropout y el peso L2).
- Dimensiones de las capas densas finales.

La búsqueda manual o mediante métodos como grid search han probado ser ineficientes y costosos en términos computacionales, especialmente en modelos complejos. Por lo tanto, se ha adoptado la optimización Bayesiana utilizando **Optuna** para encontrar de manera eficiente los hiperparámetros óptimos.

Por qué la Optimización Bayesiana

La optimización Bayesiana es una técnica que emplea modelos probabilísticos para construir una representación iterativa del espacio de hiperparámetros y dirigir la búsqueda hacia regiones prometedoras. En contraste con métodos como grid search o random search, la optimización Bayesiana tiene las siguientes ventajas:

1. **Eficiencia Computacional**: Reduce el número de evaluaciones necesarias, enfocándose en las combinaciones de hiperparámetros con mayor probabilidad de mejorar el desempeño.
2. **Manejo de Espacios de Búsqueda Complejos**: Permite trabajar con hiperparámetros continuos (como la tasa de aprendizaje), categóricos (como la función de activación) y discreto-continuos (como el número de unidades LSTM).
3. **Iteración Basada en Evidencia**: Aprovecha información de pruebas anteriores para actualizar un modelo de creencia (generalmente un proceso gaussiano) que predice el desempeño de configuraciones no exploradas.

En el contexto de este proyecto, donde los datos incluyen una dimensión temporal (años), el uso de la optimización Bayesiana asegura que el modelo resultante pueda:

- Capturar relaciones significativas entre los datos textuales y el año de publicación.
- Ajustarse con precisión a los patrones evolutivos de los temas.
- Evitar sobreajustes y optimizar el uso de recursos computacionales.

Inclusión de la Variable Año

La incorporación del año como una entrada separada en el modelo LSTM responde a la necesidad de analizar tendencias de investigación en una dimensión temporal. Este enfoque permite al modelo no solo clasificar los temas asociados a cada resumen, sino también identificar cómo estos temas han evolucionado en función del tiempo. Esta dimensión adicional es crucial para:

- Generar visualizaciones de tendencias que muestren la prevalencia de ciertos temas en diferentes periodos.
- Proporcionar un contexto temporal a los patrones encontrados, permitiendo analizar cambios en las prioridades de investigación.

Beneficios del Enfoque

1. **Optimización del Modelo**: Al integrar la optimización Bayesiana, se garantiza que el modelo final sea altamente ajustado tanto a los textos como a la dimensión temporal, maximizando su capacidad predictiva.
2. **Relevancia para el Análisis de Tendencias**: La inclusión del año como entrada asegura que el modelo capture patrones relevantes para el análisis longitudinal.
3. **Escalabilidad y Generalización**: Este enfoque es fácilmente adaptable a otros dominios donde las dimensiones temporales y temáticas interactúan.

En conclusión, el uso de la optimización Bayesiana en este proyecto no solo mejora la eficiencia en la búsqueda de hiperparámetros, sino que también fortalece la capacidad del modelo para cumplir con los objetivos de identificar y visualizar tendencias de investigación.

Esto asegura que los resultados obtenidos sean relevantes y útiles para académicos y profesionales interesados en el análisis de datos longitudinales en el ámbito del turismo.

Figura 22

Configuración del Modelo LSTM

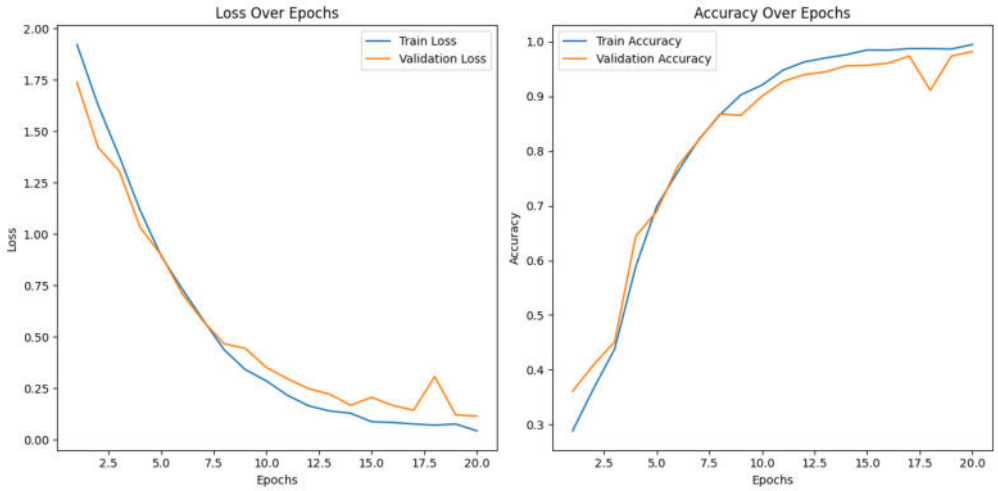
```python
from tensorflow.keras.models import Model
from tensorflow.keras.layers import Input, Embedding, LSTM, Dense, Dropout, Concatenate
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.callbacks import EarlyStopping, TensorBoard
import os
import datetime

# Hiperparámetros encontrados
best_params = {
    'embedding_output_dim': 64,
    'lstm_units': 32,
    'dense_units': 96,
    'dropout_rate': 0.4,
    'activation': 'sigmoid',
    'learning_rate': 0.0021997606627641025,
    'l2_rate': 0.00012498795389068702
}

# Definir las entradas
text_input = Input(shape=(train_padded.shape[1],), name="Text_Input")
year_input = Input(shape=(1,), name="Year_Input")

# Embedding y LSTM para el texto
text_embedding = Embedding(input_dim=vocab_size, output_dim=best_params['embedding_output_dim'])(text_input)
lstm_1 = LSTM(units=best_params['lstm_units'], return_sequences=True)(text_embedding)
lstm_2 = LSTM(units=best_params['lstm_units'], return_sequences=False)(lstm_1)

# Concatenar salida de LSTM con el año
concatenated = Concatenate()([lstm_2, year_input])

# Capas densas con dropout
dense_1 = Dense(best_params['dense_units'], activation=best_params['activation'],
                kernel_regularizer=tf.keras.regularizers.l2(best_params['l2_rate']))(concatenated)
dropout = Dropout(best_params['dropout_rate'])(dense_1)
output = Dense(num_classes, activation='softmax')(dropout)

# Crear el modelo
model = Model(inputs=[text_input, year_input], outputs=output)
```

CAPITULO 3: ANÁLISIS DE RESULTADOS

3.1 Rendimiento por época

Figura 23

Gráfica de rendimiento por épocas



Este gráfico muestra la **pérdida** en los datos de entrenamiento y validación a lo largo de las épocas de entrenamiento:

- **Línea azul**: Representa la pérdida de entrenamiento, que comienza siendo alta y disminuye gradualmente a medida que el modelo aprende.
- **Línea naranja**: Representa la pérdida de validación, que también comienza alta y disminuye, pero no de forma tan suave como la pérdida de entrenamiento.

Tras unas **10-12 épocas**, las pérdidas de entrenamiento y validación comienzan a **converger**, lo que indica que el modelo está generalizando correctamente.

En cuanto a la **precisión** en los datos de entrenamiento y validación a lo largo de las épocas de entrenamiento:

- **Línea azul**: Representa la precisión de entrenamiento, que empieza siendo baja y aumenta rápidamente a medida que el modelo aprende.

- **Línea naranja**: Representa la precisión de validación, que también empieza baja, pero incrementa a un ritmo ligeramente inferior al de la precisión de entrenamiento.

Después de unas **10-12 épocas**, la precisión de entrenamiento y la de validación comienzan a **converger**, lo que indica que el modelo está generalizando correctamente.

3.2 Observaciones generales

Rendimiento del modelo: Este modelo parece funcionar bien, ya que las pérdidas y la precisión de entrenamiento y validación convergen tras unas **10-12 épocas**.

Aprendizaje rápido: El rápido aumento de la precisión y la disminución de las pérdidas en las primeras **10-12 épocas** indican que el modelo está aprendiendo de manera rápida y eficaz.

Generalización: La convergencia entre las métricas de entrenamiento y validación sugiere que el modelo está evitando tanto el **sobreajuste** como el **subajuste**, logrando un equilibrio adecuado.

El modelo LSTM tiene un buen comportamiento en cuanto a la clasificación de los resúmenes de artículos científicos, habiéndose obtenido un 97% de precisión en el conjunto de prueba. Esto significa que el modelo es capaz de generalizar bien y predecir los temas de nuevos resúmenes.

3.3 Tendencias por año

El análisis por año del tema permite ver la evolución de los diferentes temas a lo largo del tiempo. Se puede ver cómo hay un crecimiento estable para temas como “Smart Mobility” o “Urban Sustainability” mientras que otros, como “Algorithmic Models” o “Tech Communication” sufren más fluctuaciones.

Figura 24

Tendencias por año

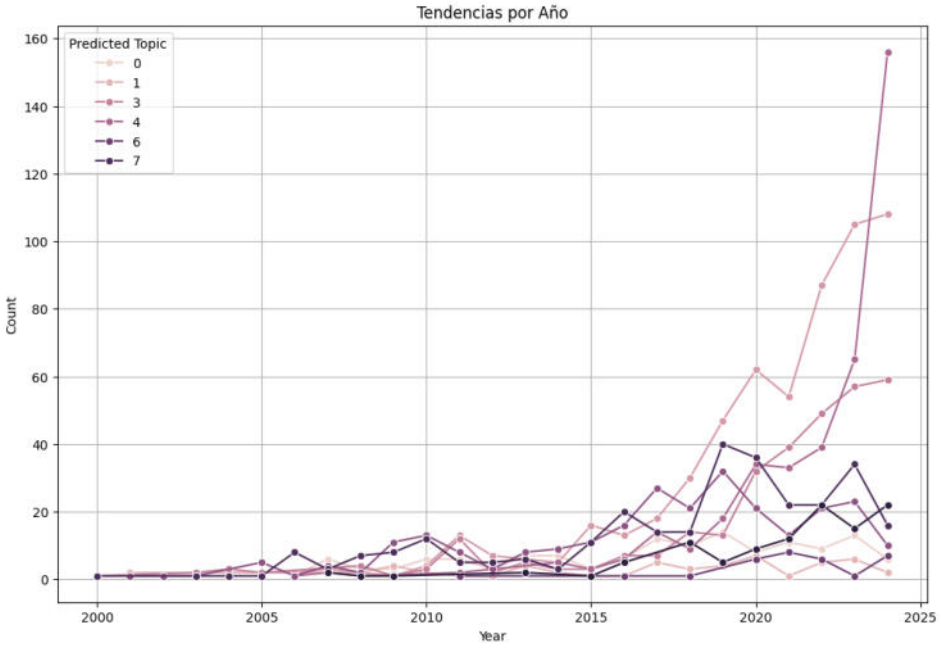


Tabla 7

Tendencias por Tema

Tema	Tendencia	Observaciones
Tema_0	Aumento constante y lineal a lo largo de 25 años	Tema fundacional o central con relevancia sostenida. Interés estable en esta área.
AI_Turismo	Crecimiento acelerado, especialmente desde 2015	Aplicaciones de IA en turismo ganando tracción significativa. Creciente atención e inversión. Impulsado por la pandemia de COVID-19 (2020-2022).
Modelos_algorítmicos	Múltiples picos y valles, naturaleza volátil	Interés no constante. Enfocado en resolver retos específicos. Aplicaciones intensivas en 2020-2022 para gestionar crisis.
Smart_Mobility	Tendencia al alza con crecimiento anual consistente	Auge de tecnologías como vehículos eléctricos, conducción autónoma y movilidad compartida. Adopción creciente de soluciones de movilidad inteligente. Relevancia durante 2020-2022 por restricciones de movilidad.
Sostenibilidad	Fluctuaciones con pico notable alrededor de 2020	Picos reflejan mayor atención a las prácticas de sostenibilidad. Aumento en 2020 vinculado a eventos globales, incluyendo la pandemia de COVID-19.
Comunicación_tecnológica	Volatilidad, con múltiples picos y valles	Atención cíclica, posiblemente influida por avances tecnológicos o demandas del mercado. Picos entre 2020 y 2022 por necesidad de soluciones digitales avanzadas.
Turismo_Experiencia	Aumento gradual y lineal	Interés constante, sin picos o caídas importantes. Atención estable en la mejora de experiencias turísticas.

		Adaptación a nuevas necesidades derivadas de la pandemia (2020-2022).
Sostenibilidad_urbana	Crecimiento exponencial, dominante hacia 2025	Enfocado en ciudades inteligentes, infraestructuras ecológicas y planificación urbana sostenible. Crecimiento catalizado por la pandemia de COVID-19 (2020-2022).
Análisis_de_datos	Fuerte crecimiento, especialmente en los últimos años	Creciente importancia de la toma de decisiones basada en datos y técnicas avanzadas de análisis. Crítico para monitorear la pandemia (2020-2022).

El análisis de tendencias por tema destaca:

- **Áreas de crecimiento constante**: Smart_Mobility, Urban_Sustainability y Data_Analysis muestran un interés y adopción creciente con el tiempo.
- **Tendencias volátiles**: Algorithmic_Models y Tech_Communication presentan fluctuaciones, posiblemente debido a demandas específicas o avances tecnológicos.
- **Impacto de la pandemia**:
- La pandemia de COVID-19 entre 2020 y 2022 parece haber influido significativamente en temas como **Sostenibilidad, Comunicación_tecnológica y Análisis de datos**, resaltando su importancia en tiempos de crisis.

Implicaciones estratégicas:

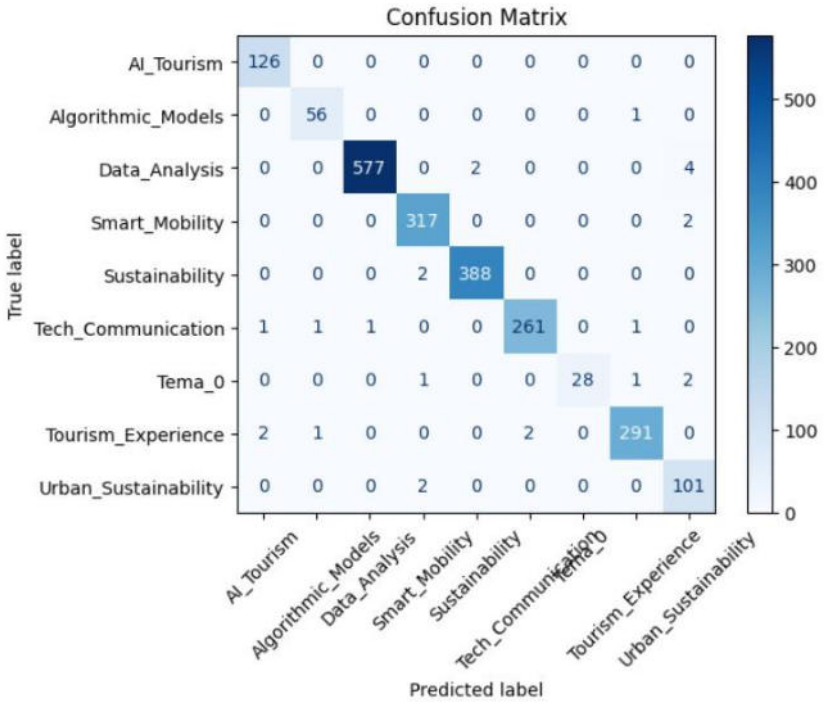
- Este panorama evolutivo permite identificar **prioridades emergentes**.
- Ayuda a asignar recursos de forma efectiva y alinear iniciativas estratégicas para abordar áreas de interés dinámicas a lo largo del tiempo.

3.4 Matriz de confusión

La “matriz de confusión” permite mostrar que la mayor cantidad de temas se clasifican de una forma adecuada, existiendo algún error entre temas relacionados o qué tienen similitudes. En relación al modelo, esto nos señala su capacidad de distinguir temas, pero todavía existen áreas de confusión.

Figura 25

Matriz de confusión



3.5 Distribución de Predicciones por Año

La visualización de los temas de tópicos con el uso de PCA permite detectar conjuntos que cumplen con los requisitos de un tema en particular. Se evidencia que diversos temas, los de "Urban Sustainability" y "Data Analysis" crean grupos muy marcados, cuya representación apunta a sugerir que el modelo es más incluyente.

Figura 26

Evolución de los temas a lo largo del tiempo

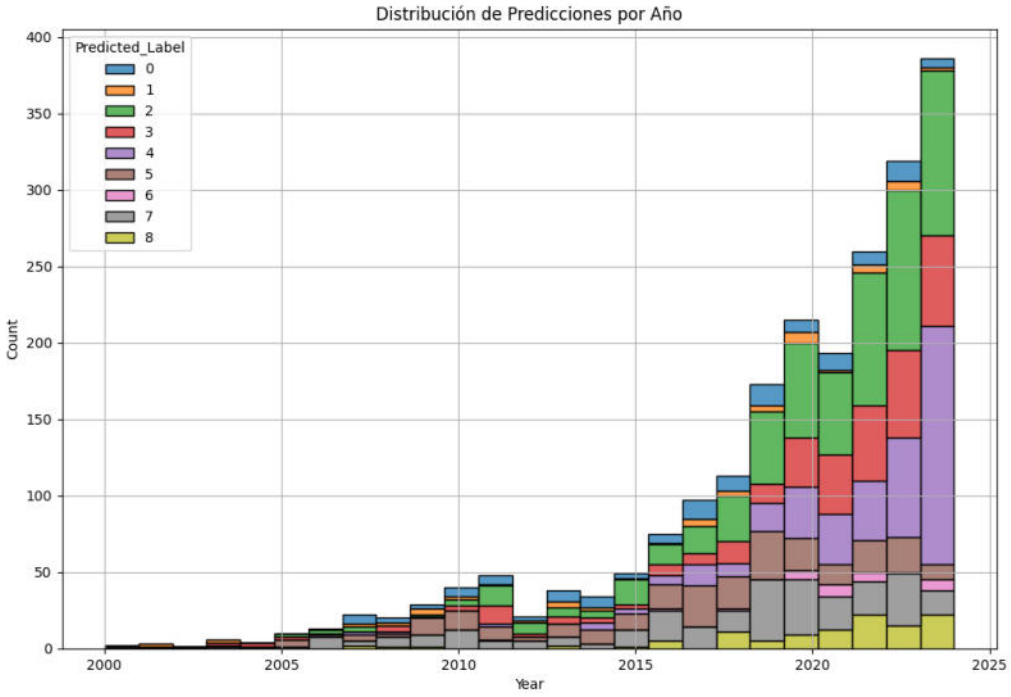


Tabla 8

Evolución de tendencia a través del tiempo

Tema	Tendencia	Observaciones
Tema_0	Presencia baja pero constante, ligero incremento al final	Interés general o estable sin cambios significativos.
AI_Tourism	Crecimiento sostenido, aumento pronunciado desde 2015	Aplicaciones de IA en turismo con tracción significativa. Importancia estratégica en un sector digitalizado.
Algorithmic_Models	Evolución volátil, altibajos	Periodos de mayor y menor relevancia, sin tendencia clara. Relacionado con problemas específicos o avances tecnológicos.

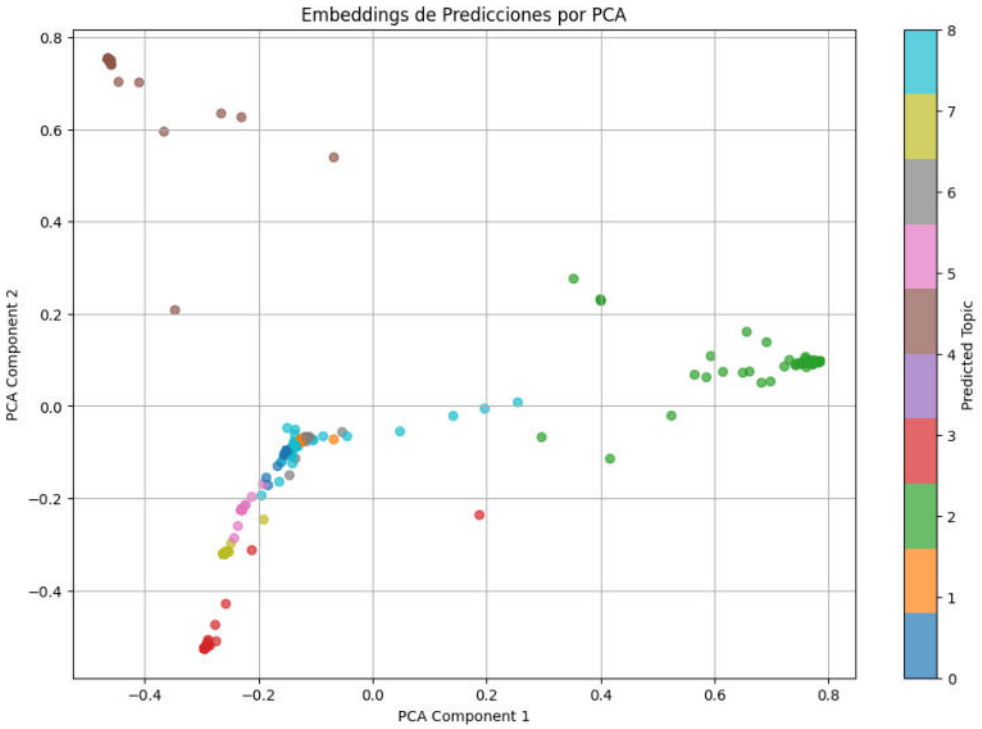
Data_Analysis	Crecimiento constante y significativo, especialmente desde 2015	Relevancia creciente. Toma de decisiones fundamentada en datos en el campo de este proyecto.
Smart_Mobility	Aumento continuo y sostenido	Relevancia creciente. Auge de vehículos eléctricos, movilidad compartida y conducción autónoma.
Sustainability	Variable, altibajos y picos	Momentos de mayor y menor relevancia. Picos asociados con eventos globales o políticas.
Tech_Communication	Evolución volátil, periodos de mayor y menor participación	Sin tendencia clara. Influida por avances tecnológicos o demandas del mercado.
Tourism_Experience	Presencia moderada y constante	Niveles estables sin grandes variaciones. Interés continuo en la mejora de experiencias turísticas.
Urban_Sustainability	Crecimiento exponencial, relevante hacia 2025	Auge de ciudades inteligentes, infraestructuras ecológicas y planificación urbana sostenible.

3.6 Embeddings

Los Embeddings son representaciones vectoriales densas y continuas que capturan características importantes de los datos. En lugar de representar datos de forma dispersa o categórica, los Embeddings asignan a cada dato un vector en un espacio de características. Estas representaciones permiten al modelo comprender y estudiar el relacionamiento y similitudes entre los datos. En este modelo, los Embeddings muestran vectorialmente las predicciones realizadas, donde cada predicción es un punto en un espacio multidimensional que captura características relevantes del tema predicho.

Figura 27

Representación de Embeddings de predicciones por PCA



El gráfico muestra cómo las predicciones del modelo están representadas en un espacio bidimensional mediante los dos primeros componentes principales:

- **PCA Component 1**: Primer componente principal, que indica la mayor parte de la variabilidad en los datos.
- **PCA Component 2**: Segundo componente principal, que señala la mayor cantidad de restante dispersión.

Elementos clave del gráfico:

- Cada punto representa una predicción realizada por el modelo.
- El color de los puntos indica el tema o "Predicted Label" correspondiente.
- La distribución de los puntos refleja las similitudes y diferencias entre las predicciones de diferentes temas.

- Las agrupaciones claras son temas como **"Urban_Sustainability"** (color verde) y **"Data_Analysis"** (color azul) forman agrupaciones densas y definidas. Esto sugiere que las predicciones de estos temas comparten características similares, lo que podría indicar una mayor consistencia en las representaciones generadas por el modelo.
- En cuanto a la distribución dispersa se identificaron temas como **"Algorithmic_Models"** (color naranja) y **"Tech_Communication"** (color morado). Esto indica una mayor heterogeneidad en las características capturadas para estos temas, posiblemente debido a la naturaleza diversa de los datos asociados.

Relaciones entre temas:

La proximidad entre los puntos de diferentes temas puede indicar cierta relación o similitud entre esos temas. Por ejemplo, temas como **"Data_Analysis"** y **"Urban_Sustainability"** podrían compartir patrones comunes en las características capturadas.

3.7 Efecto de reducción de dimensionalidad

Si bien PCA conserva la mayor parte de la variabilidad de los datos, inevitablemente se pierde algo de información al reducir las dimensiones. Las agrupaciones y relaciones observadas en el gráfico reflejan solo las principales características capturadas por el modelo. La totalidad de estos hallazgos parece apuntar que, en efecto, el modelo LSTM puede ser utilizado con éxito para identificar y representar las tendencias de la investigación en turismo. Efectivamente, el análisis de tendencias en el tiempo, la matriz de confusión y la visualización de las embeddings prometen interesantes resultados de cara conocer cómo el conocimiento evoluciona en este campo de estudio.

CAPITULO 4: CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones

La adecuada recopilación y análisis de datos textuales ha demostrado ser fundamental para prever comportamientos y tendencias en el sector turístico. Esto faculta a académicos e interesados en el campo del turismo a tener comprensiones más contextualizadas y profundas sobre las dinámicas de estudio del mercado y anticipar necesidades futuras. Los resultados que este estudio ha alcanzado han proporcionado información de gran valor para la orientación de estrategias de innovación tanto en investigación académica como en la industria turística, facilitando decisiones informadas en áreas como sostenibilidad y turismo inteligente. Este proyecto no solo mejora la capacidad de gestionar y analizar investigaciones, sino que también fomenta la exploración interdisciplinaria. Las técnicas aplicadas permiten identificar vacíos investigativos y áreas emergentes, sirviendo como referencia para futuros estudios y proyectos.

Los modelos empleados, como LDA y LSTM, demostraron alta precisión en el señalamiento de configuraciones y constructos de mayor consistencia aglutinante, lo que valida su potencialidad para embarcarse en el tratamiento de conjuntos más extensos de corpus textuales con resultados consistentes y accionables. Se implementaron soluciones técnicas completas que integraron modelos de machine learning con interfaces funcionales, facilitando la interacción y análisis de datos en entornos robustos y escalables. La automatización de procesos y la configuración de entornos virtualizados con Docker y GPU optimizaron los recursos disponibles, permitiendo la ejecución eficiente de tareas críticas incluso en condiciones adversas, como los cortes de energía. La experiencia adquirida en el desarrollo de pipelines de datos y despliegue de modelos permitió abordar requerimientos

complejos con un enfoque sistemático, fomentando una interacción efectiva entre áreas técnicas y estratégicas.

4.2 Recomendaciones

Se sugiere que se mejorase el balance de las clases para una combinación de temas relacionados -la concordancia de temas relacionados- a través de técnicas de over sampling o de under sampling de los datos de la recopilación; esto podría provocar modelos más robustos y menos sesgados hacia las clases mayoritarias, que en este marco como la alta presencia de artículos en inglés.

Incluir visualizaciones interactivas dinámicas de tendencias por años para favorecer, de esta manera, una exploración más profunda de cómo los temas se desarrollan en respuesta a cambios globales, tales como la pandemia de COVID-19.

Probar modelos de lenguaje basados en transformers (BERT o GPT) en lugar de los LSTMs con la finalidad de captar las dependencias en textos científicos más complejas para mejorar la coherencia semántica de los resultados obtenidos.

4.3. Limitaciones a la Investigación

Disponibilidad y calidad del corpus: El proyecto se enfrentó a un corpus inicial limitado en tamaño y diversidad, lo que restringió la representatividad de los datos. Para mitigar esta limitación, se implementaron técnicas de aumento de datos, como la sustitución de palabras por sinónimos, lo que permitió enriquecer el conjunto de datos sin introducir sesgos significativos.

Contexto energético en Ecuador: Los constantes cortes de energía eléctrica en el país representaron un desafío operativo significativo, afectando la continuidad de los entrenamientos de modelos y el acceso a recursos computacionales críticos. Este contexto destacó la importancia de planificar estrategias resilientes en proyectos de investigación.

Capacidad computacional: Aunque se dispuso de herramientas avanzadas, como entornos con soporte CUDA y optimización de recursos, algunas técnicas más complejas no pudieron ser implementadas debido a limitaciones en la capacidad computacional disponible.

Generalización de los resultados: Los hallazgos y modelos desarrollados están limitados al conjunto de datos y contexto específicos utilizados, lo que podría reducir su aplicabilidad a otras bases de datos o regiones con características significativamente diferentes.

REFERENCIAS BIBLIOGRÁFICAS

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... Zheng, X. (2016). TensorFlow: Large-scale machine learning on heterogeneous systems. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 265–283. https://www.tensorflow.org

Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data*. Springer, 99-100. https://doi.org/10.1007/978-1-4614-3223-4

Bird, S., & Loper, E. (2004). NLTK: The Natural Language Toolkit. *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, 31–32. https://doi.org/10.3115/1219044.1219075

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1), 1-15. https://doi.org/10.1057/s41599-021-00903-w

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11), 2215–2222. https://doi.org/10.1002/asi.23329

Chollet, F. (2017). *Deep Learning with Python*. Manning Publications.

Chomsky, N. (2002). *Syntactic Structures*. (2nd ed.). Mouton. https://tallinzen.net/media/readings/chomsky_syntactic_structures.pdf

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science Mapping Software Tools: Review, Analysis, and Cooperative Study among Tools. *Journal of the American Society for Information Science and Technology*, 62(7), 1382-1402. http://dx.doi.org/10.1002/asi.21525

Correa, P. A. R., Pineda, V. G., Moreno-López, G., Giraldo, L. F. G., León, A. J. O., Giraldo, M. C. B., & Benjumea-Arias, M. (2024). Integración de Big Data y Machine Learning: Análisis de brechas y tendencias investigativas. [Big Data and Machine Learning Integration: Gap analysis and research trends] *Revista Ibérica De Sistemas e Tecnologias De Informação,* , 644-656. https://www.proquest.com/scholarly-journals/integración-de-big-data-y-machine-learning/docview/3094869110/se-2

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Gupta, D., Bhargava, A., Agarwal, D., Alsharif, M. H., Uthansakul, P., & Aly, A. A. (2024). Deep learning-based truthful and deceptive hotel reviews. *Sustainability, 16*(11), 4514. https://doi.org/10.3390/su16114514

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Weckesser, W. (2020). Array programming with NumPy. *Nature*, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50–57. https://doi.org/10.1145/312624.312649 https://doi.org/10.1038/s41586-022-05543

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

İri, R., & Ünal, E. (2024). Bibliometric analysis bibliometric analysis of research (1980-2023). Ahi Evran Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10(2), 386-401. https://doi.org/10.31592/aeusbed.1446738

Jia, M., Feng, J., Chen, Y., & Zhao, C. (2024). Visual analysis of social media data on experiences at a World Heritage tourist destination: Historic Centre of Macau. *Buildings, 14*(7), 2188. https://doi.org/10.3390/buildings14072188

Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2ª ed.). Pearson Prentice Hall.

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). Prentice Hall.

Kleene, S. C. (1951). *Representation of events in nerve nets and finite automata*. U.S. AIR FORCE PROJECT RAND Research Memorandum. https://www.rand.org/content/dam/rand/pubs/research_memoranda/2008/RM704.pdf

Knuth, D. E. (1997). *The Art of Computer Programming* (Vols. 1). Addison-Wesley. http://broiler.astrometry.net/~kilian/The_Art_of_Computer_Programming%20-%20Vol%201.pdf

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Martins, E., & Galegale, N. V. (2023). Machine learning: a bibliometric analysis. *International Journal of Innovation, 11*(3), 1-37. https://doi.org/10.5585/2023.24056

McKercher, B., & Tung, V. (2015). Publishing in tourism and hospitality journals: Is the past a prelude to the future? *Tourism Management*, 50, 306-315. https://doi.org/10.1016/j.tourman.2015.03.013

McKinney, W. (2011). *pandas: a Foundational Python Library for Data Analysis and Statistics*. [Archivo PDF] .https://www.researchgate.net/publication/265194455_pandas_a_Foundational_Python_Library_for_Data_Analysis_and_Statistics

Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D., & Fast, A. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press.

Olah, C. (2015). Understanding LSTMs. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Oliphant, T. E. (2006). Guide to NumPy. Trelgol Publishing. https://web.mit.edu/dvp/Public/numpybook.pdf

Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, 613(7942), 138-144.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pramod, Gupta., Anupam, Bagchi. (2023). Introduction to Pandas. Synthesis lectures on engineering, science, and technology, 161-196. https://doi.org/10.1007/978-3-031-43725-0_5

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://dx.doi.org/10.13140/2.1.2393.1847

Sebastiani, F. (2002). "Machine Learning in Automated Text Categorization." *ACM Computing Surveys*, 34(1), 1-47.

Suanpang, P., & Pothipassa, P. (2024). Integrating generative AI and IoT for sustainable smart tourism destinations. *Sustainability, 16*(17), 7435. https://doi.org/10.3390/su16177435

Thompson, K., & Ritchie, D. M. (1984). *The UNIX Programming Environment*. Prentice-Hall.

Tufte, E. R. (1980). *The Visual Display of Quantitative Information*. (2nd ed.). Graphics Press. https://www.econ.upf.edu/~michael/visualdata/tufte-aesthetics_and_technique.pdf

Van Rossum, G. (1995). *Python Reference Manual*. CWI. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=09b1d693676a152621799e2cfff562a369d10204

Wall, L., Christiansen, T., & Orwant, J. (2012). *Programming Perl* (4th ed.). O'Reilly Media. http://www.freeoa.net/attachments/2014/Programming.Perl.4th.2012.pdf

Ware, M., & Mabe, M. (2015). *The STM Report: An Overview of Scientific and Scholarly Publishing*. International Association of Scientific, Technical and Medical Publishers.

https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1008&context=scholcom

Yunus, Eroğlu. (2023). Text Mining Approach for Trend Tracking in Scientific Research: A Case Study on Forest Fire. Fire, 6(1):33-33. https://doi.org/10.3390/fire6010033