Estudios turísticos N°230
julio - diciembre 2025 | Pp. 121-137
DOI: https://doi.org.10.61520/et.2302025.1263
eISSN: 3020-6723

POTENCIANDO LA INVESTIGACIÓN EN TURISMO CON CIENCIA DE DATOS: UNA BREVE GUÍA DE HERRAMIENTAS CON EL LENGUAJE DE PROGRAMACIÓN PYTHON

ENHANCING TOURISM RESEARCH WITH DATA SCIENCE: A BRIEF TOOLKIT WITH THE PYTHON PROGRAMMING LANGUAGE

Maria Fernanda Bernal Salazar*, Elisa Baraibar-Diez y Jesús Collado Agudo*****
*Universidad de Cantabria | maria-fernanda.bernal@alumnos.unican.es | https://orcid.org/0000-0002-6069-082X
**Universidad de Cantabria | elisa.baraibar@unican.es | https://orcid.org/0000-0003-4677-3255
***Universidad de Cantabria | jesus.collado@unican.es | https://orcid.org/0000-0002-4152-4439

ENTREGADO: 30/06/2025 ACEPTADO: 03/11/2025

Resumen: La ciencia de datos es un campo transversal que potencia diversas disciplinas y transforma la forma de generar y analizar conocimiento. Su aplicación como marco metodológico ofrece grandes oportunidades, pero también desafíos para los investigadores. En el ámbito del turismo, aún se requieren mayores esfuerzos para su adopción plena. Este trabajo destaca, por un lado, la relevancia de la ciencia de datos en la investigación turística y, por otro, presenta una guía básica de herramientas, modelos y bibliotecas de Python para investigadores que se inician en este campo, incluyendo aplicaciones específicas para el análisis de reseñas en TripAdvisor.

Palabras clave: Ciencia de datos, investigación en turismo, *big data*, contenido generado por el usuario, UGC

Abstract: Data science is a cross-cutting field that enhances various disciplines and transforms the way knowledge is generated and analyzed. Its application as a methodological framework offers significant opportunities, but also poses challenges for researchers. In the field of tourism, further efforts are still needed to achieve full adoption. This work highlights, on the one hand, the relevance of data science in tourism research and, on the other, provides a basic guide to Python tools, models, and libraries for researchers entering this field, including specific applications for analyzing reviews on TripAdvisor.

Keywords: Data science, tourism research, big data, User Generated Content, UGC

1. INTRODUCCIÓN Y REVISIÓN DE LA LITERATURA

La expansión de internet y las nuevas tecnologías ha transformado sectores como el turismo, generando grandes volúmenes de datos (*big data*) a partir de dispositivos inteligentes, redes sociales y plataformas digitales (Pereira-Moliner et al., 2024). Este fenómeno ha revolucionado la gestión de destinos y empresas turísticas, promoviendo la personalización y un enfoque centrado en el cliente, planteando importantes desafíos analíticos (Li & Law, 2020; Mariani et al., 2018). Por su parte, la investigación científica se ve impulsada por las capacidades computacionales que facilitan el análisis del *big data,* lo que exige formar a los futuros investigadores tanto en sus áreas específicas como en ciencia de datos (Pennington et al., 2020). Esta última se integra transversalmente en diversas disciplinas, redefiniendo la forma de generar conocimiento.

La literatura reciente evidencia un creciente interés por la aplicación de la ciencia de datos y el *big data* en el turismo (António & Rita, 2023; Cai et al., 2024; Mariani & Baggio, 2022), como también lo reflejan los *Virtual Special Issues* del *IJCHM* dedicados a *Artificial Intelligence (AI) in Hospitality and Tourism* y *Big Data in Hospitality and Tourism* (International Journal of Contemporary Hospitality Management, n.d.-a, n.d.-b). La incorporación de estas

metodologías aporta un valor significativo a la investigación turística y tiene importantes implicaciones prácticas para la gestión de destinos y el marketing del sector. Entre sus principales aplicaciones destacan el seguimiento de patrones de movilidad de los turistas, la evaluación de aspectos psicográficos (sentimientos, percepciones y emociones), la predicción de la demanda y la personalización de servicios (Cai et al., 2024; Egger, 2022; Xu et al., 2019).

La industria turística genera tres grandes fuentes de datos susceptibles de ser analizadas con estas metodologías: contenido generado por el usuario (User-Generated Content, UGC, por sus siglas en inglés), dispositivos (GPS, móviles) y transacciones (búsquedas, reservas) (Li et al., 2018). En particular, las reseñas online constituyen una de las formas más distinguidas de UGC y son fundamentales para el *marketing* turístico y la investigación (D'Acunto et al., 2020). Estas influyen en las decisiones de los turistas, con un impacto directo en los resultados económicos de las empresas y destinos (Bigné et al., 2019; Liu & Park, 2015; Xiang & Gretzel, 2010).

Los investigadores han comenzado a aplicar técnicas de procesamiento del lenguaje natural (PLN) (Guerrero-Rodriguez et al., 2023), *machine learning* (Cervera et al., 2024; Mor et al., 2023) y redes neuronales (Liang et al., 2024)en estudios turísticos. Por ejemplo, Lalicic et al. (2021) analizan la imagen online de un destino turístico a partir reseñas online. D'Acunto et al. (2024) identifican el perfil sociodemográfico de los turistas que publican reseñas ambientales. Meneghini & Tuzzi (2025) utilizan modelado de temas y el modelo Mistral-7b para evaluar la orientación inclusiva de proyectos turísticos. Saoualih et al. (2025) analizan reseñas negativas sobre 37 museos marroquíes para identificar brechas experienciales mediante análisis de sentimientos y modelado de temas (BERTopic).

A pesar de estos avances, persiste una adopción limitada de estas metodologías en la investigación turística, lo que indica que su potencial aún no ha sido plenamente aprovechado (Egger, 2022; Köseoglu et al., 2020; Mariani & Baggio, 2022). Según Mariani et al. (2018), en comparación con otras disciplinas académicas, los investigadores en turismo pueden presentar cierta resistencia a emplear nuevos enfoques cuantitativos avanzados y algoritmos computacionales, ya sea por falta de experiencia en lenguajes de programación, o por escasez en el acceso a hardware y software especializado. Asimismo, la propia naturaleza heterogénea y no estructurada de los datos turísticos —texto, imágenes, videos y redes sociales— puede generar confusión e incertidumbre durante el proceso analítico (Egger, 2022). Cabe destacar, además, que la implementación de estas metodologías plantea desafíos relacionados con la privacidad y seguridad de los datos, los posibles sesgos algorítmicos y la necesidad de competencias técnicas especializadas, aspectos que deben abordarse críticamente para promover un uso responsable de la ciencia de datos en el turismo.

Este trabajo busca reducir la brecha metodológica identificada, ofreciendo una guía práctica sobre las principales herramientas, modelos y bibliotecas del lenguaje de programación Python aplicables a la investigación turística. Asimismo, se presentan aplicaciones centradas en el análisis de UGC, con el propósito de ilustrar su potencial

para la investigación y la gestión del sector. De esta forma, se enfatiza la necesidad de un enfoque multidisciplinario que integre la ciencia de datos como marco metodológico en los estudios turísticos.

2. METODOLOGÍA

Esta guía presenta herramientas de ciencia de datos aplicables a la investigación turística. La ciencia de datos es una disciplina interdisciplinaria que combina estadística, machine learning, minería de datos y análisis de información para descubrir patrones y generar conocimiento a partir de grandes volúmenes de datos (George et al., 2016). Su proceso comprende cinco fases: recopilación, preprocesamiento, exploración, modelado e interpretación de resultados (Gandomi & Haider, 2015; Manning et al., 2009; Ramasamy et al., 2023; Varian, 2014). Se emplea Python por su facilidad de uso, versatilidad y amplia adopción en la comunidad científica. Para cada fase se presentan las herramientas, modelos y bibliotecas más utilizadas. Los ejemplos de código están diseñados para ejecutarse en Google Colab (ver Anexo 1), una plataforma gratuita en la nube que permite combinar código, texto y visualizaciones sin requerir instalación local.

2.1. Recopilación de datos

Los datos se clasifican en tres tipos según su estructura (Gandomi & Haider, 2015): estructurados, con formato definido y fácilmente organizables (ej. bases de datos, hojas de cálculo); no estructurados, sin formato fijo y más complejos de analizar (ej. texto, imágenes, videos, redes sociales); y semiestructurados, con cierta organización, pero sin seguir un modelo rígido (ej. metadatos, archivos XML o JSON).

Las estrategias de recopilación varían según la fuente. Para datos internos o estructurados, se usan extracciones directas o APIs[1] específicas. Para datos web se aplica *web scraping*, técnica que automatiza la extracción de información desde sitios web, transformando contenido no estructurado en datos organizados (Mitchell, 2018). Una herramienta accesible para extracción de datos es Octoparse, que permite recopilar información de forma automatizada y estructurada mediante una interfaz gráfica, ideal para usuarios sin experiencia en programación. Su versión gratuita permite hasta 10 tareas y 50 000 datos mensuales. Alternativamente, Python ofrece mayor flexibilidad mediante bibliotecas como BeautifulSoup o Selenium, aunque requiere conocimientos técnicos.

2.2. Preprocesamiento de datos

Una vez recopilados los datos turísticos, es necesario realizar un preprocesamiento para garantizar la calidad y utilidad en análisis posteriores. Los datos extraídos frecuentemente contienen errores, inconsistencias o formatos inadecuados (García et al., 2015). En esta fase se limpian, transforman y estructuran los datos para facilitar su análisis. La limpieza corrige problemas comunes como valores faltantes, atípicos o errores de formato. La transformación adapta los datos, por ejemplo, normalizando variables o codificando categorías. La estructuración organiza la información en formatos coherentes, como tablas, y verifica su consistencia.

En el caso de datos textuales, como reseñas turísticas, se aplican técnicas de preprocesamiento del lenguaje natural (Uysal & Gunal, 2014), que incluyen la tokenización (dividir el texto en unidades significativas), eliminación de palabras vacías (como artículos o preposiciones), conversión a minúsculas (para unificar términos) y lematización (reducir palabras a su forma base), lo que permite optimizar tareas como clasificación o análisis de sentimientos. La Tabla 1 presenta algunas de las principales bibliotecas de Python para preprocesamiento y procesamiento de datos, incluyendo aplicaciones en PLN.

Tabla 1. Bibliotecas de Python para preprocesamiento y procesamiento de datos

USO	NOMBRE	ENLACE	DESCRIPCIÓN
Procesamiento de datos	pandas	https://pandas.pydata.org/	Biblioteca para manipulación y análisis de datos; permite trabajar con tablas y series temporales.
	NumPy	https://numpy.org/	Biblioteca para operaciones matemáticas con arrays multidimensionales.
Procesamiento de texto	re	https://docs.python.org/3/library/re.html	Módulo para procesar texto mediante expresiones regulares; útil para limpieza y búsqueda de patrones.
	PyPDF2	https://pypdf2.readthedocs.io/en/3.x/	Biblioteca para extraer texto y manipular archivos PDF, como fusionar o dividir documentos.
Manipulación de archivos y sistema operativo	os	https://docs.python.org/3/library/os.html	Módulo que permite interactuar con el sistema operativo, gestionando archivos, directorios y rutas.
Procesamiento de lenguaje natural (PLN)	NLTK (*Natural Language Toolkit*)	https://www.nltk.org/	Biblioteca para procesamiento básico de lenguaje natural, como tokenización, lematización y análisis gramatical.
	spaCy	https://spacy.io/	Biblioteca optimizada para PLN a gran escala; incluye análisis sintáctico, reconocimiento de entidades y procesamiento avanzado.

Fuente: Elaboración propia.

2.3. Exploración de datos

La tercera fase busca comprender de manera detallada la naturaleza y características de los datos. Es clave iniciar con una exploración y visualización de los datos antes de aplicar

técnicas de ciencia de datos o estadísticas complejas. El análisis exploratorio de datos (AED) (también conocido como Exploratory Data Analysis, EDA, por sus siglas en inglés) se basa en cálculos aritméticos y gráficos básicos. El objetivo es poder comprender la distribución de los datos, su tamaño, su dispersión y las posibles relaciones existentes entre ellos. Esto facilita la identificación de patrones, tendencias y anomalías que podrían influir en la selección y aplicación posterior de métodos avanzados.

2.4. Modelado y análisis

En esta fase se aplican técnicas estadísticas y algoritmos de *machine learning* para extraer patrones y conocimiento a partir de los datos preparados. Según George et al. (2016), los principales desafíos se relacionan con el análisis de conjuntos con numerosas variables y con el manejo de grandes volúmenes de información. Para el primero, se sugieren métodos como regresión (PLS, Ridge, Lasso), análisis de componentes principales y árboles de regresión. Para grandes volúmenes, se proponen técnicas como paralelización, *bootstrapping* y análisis secuencial. También se aplican el modelado de temas, análisis semántico, métricas basadas en entropía y aprendizaje profundo *(deep learning)*. La Tabla 2 sintetiza algunas de las principales técnicas de ciencia de datos aplicables en esta fase. Esta guía se centra en métodos fundamentales para investigadores que se inician en el análisis de datos, por lo que no incluye técnicas de aprendizaje profundo.

Tabla 2. Principales técnicas de ciencia de datos aplicables en la investigación sobre turismo

NOMBRE		DESCRIPCIÓN
Minería de texto y procesamiento de lenguaje natural (PLN)	Preprocesamiento lingüístico	Transformación y normalización de texto mediante técnicas como la tokenización, lematización y eliminación de stopwords. Estas técnicas estandarizan y preparan el texto para su análisis, facilitando su estructuración y gestión.
	Extracción de información	Técnicas para identificar patrones y conocimiento en grandes volúmenes de texto, permitiendo extraer información útil desde datos no estructurados.
	Minería de opiniones	Aplicación específica de la minería de texto centrada en extraer y analizar emociones, sentimientos y opiniones expresadas en texto. Entre sus técnicas principales están: análisis de sentimientos y emociones, detección de polaridad y extracción de aspectos.
	Análisis de frecuencia	Cálculo de frecuencias para identificar términos clave y patrones recurrentes en el texto.

NOMBRE		DESCRIPCIÓN
Machine Learning	Aprendizaje no supervisado	Modelado de temas: identificación automática de temas y patrones en grandes volúmenes de texto mediante algoritmos probabilísticos.
	Aprendizaje supervisado	Modelos de predicción: métodos de predicción que incluyen regresión lineal y modelos de clasificación ordinal (ologit y gologit), que permiten predecir valores continuos o categorías ordinales a partir de relaciones entre variables.
		Clasificación de texto: métodos para categorizar automáticamente documentos en clases predefinidas o emergentes, utilizando modelos preentrenados. El enfoque Zero-shot learning permite a un modelo de IA clasificar conceptos sin haber sido entrenado explícitamente en ellos.

Fuente: Elaboración propia.

Asimismo, existe una plataforma de código abierto llamada Hugging Face, útil para investigadores que se inician en esta línea de investigación. Esta plataforma alberga modelos preentrenados, conjuntos de datos y recursos educativos sobre inteligencia artificial (IA), siendo reconocida como uno de los repositorios de IA más grandes a nivel global. Su biblioteca Transformers ofrece una infraestructura en Python para utilizar modelos avanzados como BERT, GPT o RoBERTa mediante una API unificada, aplicable en tareas de lenguaje natural, visión, audio y proyectos multimodales. Además, incorpora *pipelines* que automatizan tareas comunes con una interfaz simplificada, facilitando su uso tanto para principiantes como para implementaciones rápidas. La Tabla 3 muestra tres bibliotecas clave que pueden emplearse durante esta fase.

Tabla 3. Bibliotecas de Python útiles durante el modelado y análisis de los datos

NOMBRE	ENLACE	DESCRIPCIÓN
Scikit-learn	https://scikit-learn.org/	Biblioteca que contiene herramientas y algoritmos eficientes para tareas como de clasificación, regresión, clustering (agrupación), reducción de dimensionalidad y selección de modelos.
Statsmodels	https://www.statsmodels.org/	Biblioteca para análisis estadístico, econométrico y modelado de series temporales, útil también para explorar y visualizar datos.
Transformers	https://huggingface.co/	Biblioteca de Hugging Face que permite usar y ajustar modelos de lenguaje preentrenados basados en Transformers para aplicaciones específicas.

Fuente: Elaboración propia.

2.5. Interpretación de resultados

En esta fase final se interpretan los resultados mediante la evaluación del rendimiento de los modelos, el análisis de las métricas obtenidas y la visualización clara de los hallazgos. George et al. (2016) puntualizan que todo el proceso de ciencia de datos debe estar acompañado de rigurosos procesos de validación que incluyan la validación cruzada, muestras de retención y experimentos de campo. Asimismo, destacan la importancia de utilizar múltiples enfoques para asegurar la robustez de los resultados y verificar la causalidad, no solo las correlaciones, combinando adecuadamente métodos estadísticos tradicionales con nuevas técnicas de análisis de big data para obtener resultados significativos y confiables.

Por otro lado, Garijo et al. (2014) señalan que, la presentación de los resultados es tan importante como su generación. Por tanto, los investigadores pueden utilizar visualizaciones (tablas, gráficos, archivos, etc.) para dar a conocer sus hallazgos y tomar decisiones clave. La Tabla 4 presenta dos bibliotecas relevantes para la visualización de datos en Python.

Tabla 4. Principales bibliotecas de Python para la visualización de datos

USO	NOMBRE	ENLACE	DESCRIPCIÓN
Visualización de resultados	Matplotlib	https://matplotlib.org/	Biblioteca fundamental para la visualización de datos que permite generar gráficos de alta calidad con gran flexibilidad y personalización.
	Seaborn	https://seaborn.pydata.org/	Biblioteca de visualización de estadística avanzada basada en Matplotlib. Brinda una interfaz de alto nivel para generar gráficos estadísticos atractivos e informativos.

Fuente: Elaboración propia.

3. RESULTADOS

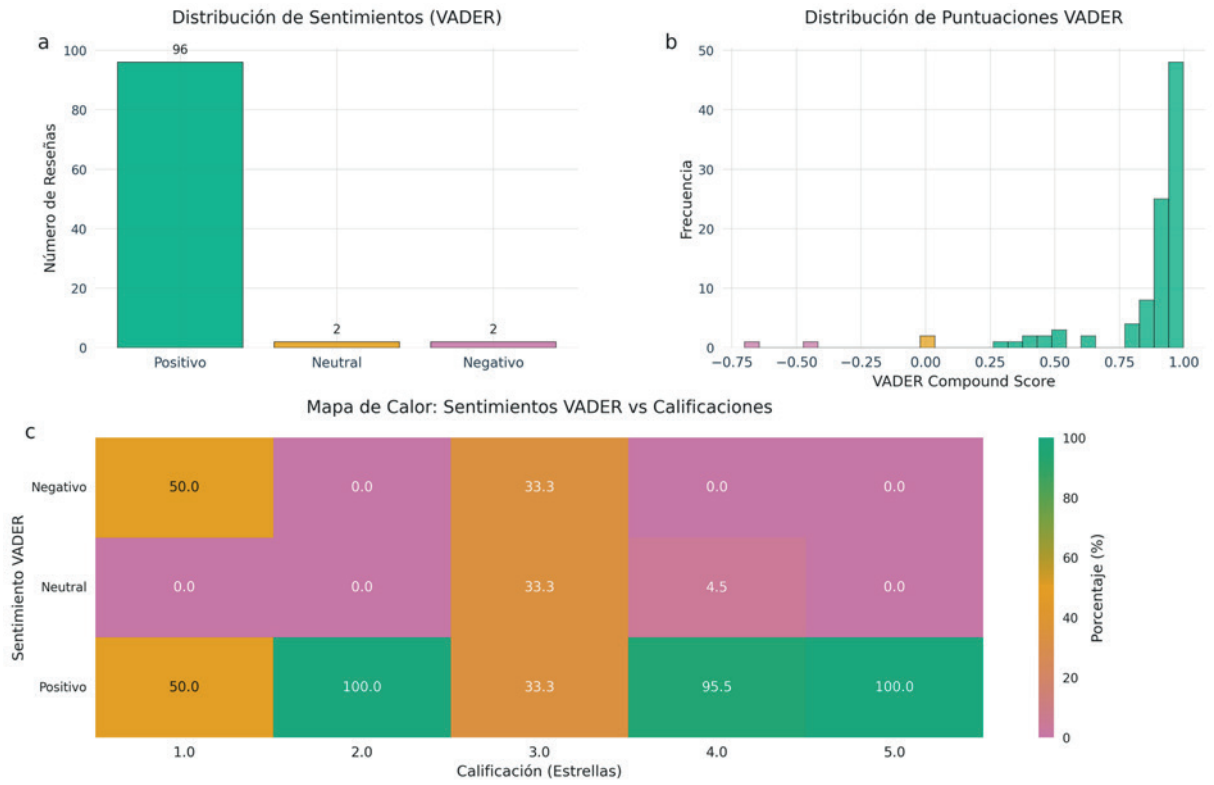
Esta sección presenta la aplicación práctica de tres técnicas de ciencia de datos al UGC: análisis de sentimientos, clasificación de texto y modelado de temas. La base de datos está compuesta por 5304 reseñas de TripAdvisor correspondientes a ocho hoteles de Andalucía (España). Los datos fueron recopilados mediante la herramienta de *web scraping* Octoparse en marzo de 2024. Para cada reseña se extrajo el texto del comentario, la nacionalidad del turista, la fecha de estancia, el nombre del hotel y la calificación otorgada. Con el fin de garantizar los principios éticos y la protección de datos personales, se omitió toda información sensible, incluyendo nombres, fotografías e identificadores de usuario.

El repositorio de GitHub (ver Anexo 1) incluye el *notebook* ejecutable, los datos en formato Excel y un archivo «requirements.txt» con las versiones exactas de las bibliotecas utilizadas, lo que facilita la verificación y réplica del análisis. El código está diseñado para Google Colab, incluye comentarios explicativos y cubre las fases de preprocesamiento, exploración de datos y modelado. Cabe destacar que el preprocesamiento se realizó sobre el conjunto completo de datos; sin embargo, para las fases analíticas posteriores se emplearon muestras aleatorias con semilla fija, a fin de facilitar la replicabilidad y optimizar los recursos computacionales. No obstante, los resultados pueden mostrar ligeras variaciones debido a la naturaleza probabilística de los algoritmos empleados.

3.1. Análisis de sentimientos

El análisis de sentimientos permite evaluar automáticamente la polaridad (positiva, negativa o neutral) de las opiniones expresadas en las reseñas turísticas. Para este estudio se empleó VADER (Valence Aware Dictionary and sEntiment Reasoner), implementado en la biblioteca NLTK de Python mediante la clase SentimentIntensityAnalyzer. VADER es un analizador de sentimientos optimizado para texto de redes sociales, basado en un diccionario léxico que asigna intensidades emocionales a las palabras.

Ilustración 1. Análisis completo de sentimientos con VADER



Fuente: Elaboración propia.

Dado que el modelo fue diseñado específicamente para inglés y su diccionario contiene palabras con intensidades emocionales definidas en ese idioma, las reseñas fueron

traducidas automáticamente desde el español antes del procesamiento. Del conjunto total de 5304 reseñas recopiladas, se seleccionó una muestra aleatoria de 100 comentarios para el análisis, garantizando consistencia mediante una semilla fija.

La Ilustración 1 muestra los resultados del análisis de sentimientos con VADER. La Ilustración 1a presenta la distribución de sentimientos, evidenciando una predominancia de comentarios positivos (96 %), frente a un número marginal de reseñas neutrales (2 %) y negativas (2 %), lo que refleja una percepción general muy favorable hacia los hoteles analizados. La Ilustración 1b muestra la distribución de las puntuaciones compuestas (*compound scores*), donde la mayoría de los valores se concentran cerca de 1, confirmando el sesgo positivo general. Por último, la Ilustración 1c representa un mapa de calor que relaciona los sentimientos VADER con las calificaciones otorgadas. Se observa una correspondencia directa entre los sentimientos positivos y las valoraciones más altas (4-5 estrellas), mientras que los pocos comentarios negativos se asocian con puntuaciones bajas (1-2 estrellas).

Los ejemplos completos de comentarios, junto con sus traducciones y puntuaciones de sentimiento obtenidas con VADER, se incluyen en el *notebook* asociado (ver Anexo 1). Este material permite reproducir el análisis paso a paso y observar cómo el modelo asigna las diferentes polaridades a las reseñas turísticas.

3.2. Clasificación de texto

Este enfoque permite identificar automáticamente las categorías más representativas dentro de las reseñas turísticas, sin requerir un conjunto de entrenamiento previo. Para ello se aplicó un modelo de aprendizaje *zero-shot*, que posibilita categorizar textos en clases predefinidas sin haber sido entrenado explícitamente para ellas (Alhoshan et al., 2023).

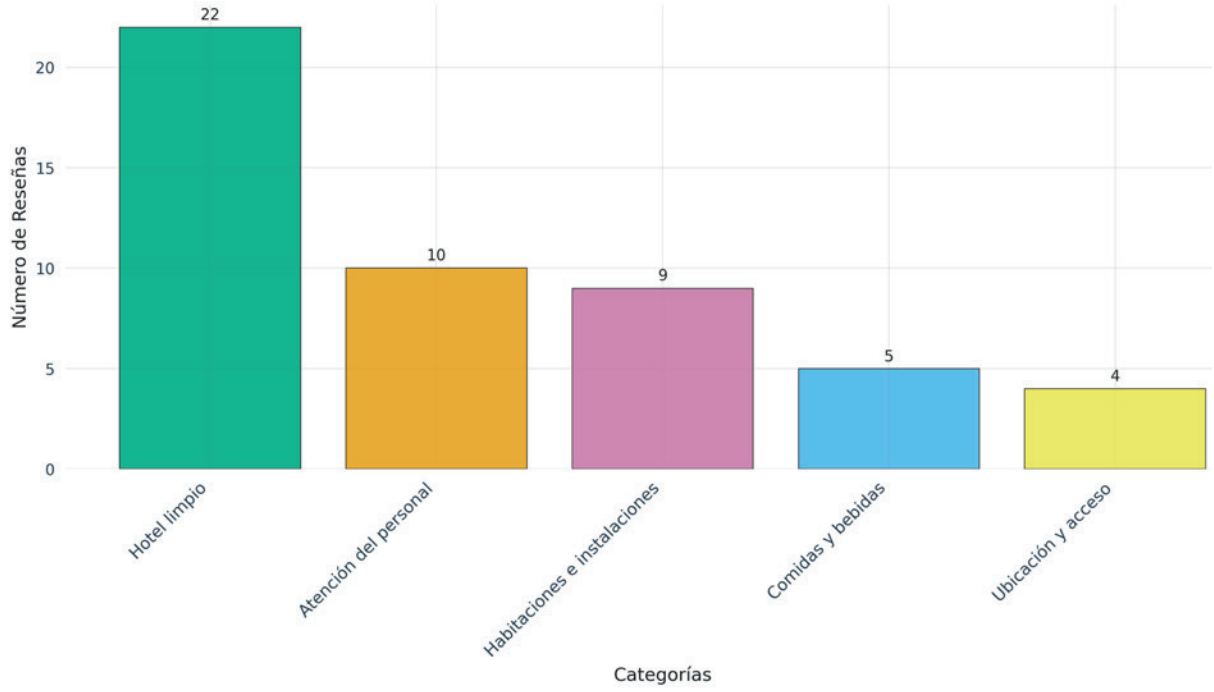
La implementación se realizó mediante el modelo «bert-base-spanish-wwm-cased-xnli», desarrollado por la empresa Recognai (2021), disponible en la plataforma Hugging Face y basado en la arquitectura BERT (Bidirectional Encoder Representations from Transformers)[2] (Cañete et al., 2020). Este modelo, entrenado con el conjunto multilingüe XNLI *(*Cross-lingual Natural Language Inference), está optimizado para tareas de inferencia y clasificación de lenguaje natural en español.

Para el análisis se eligió una muestra aleatoria de 50 reseñas y se definieron siete categorías para las reseñas hoteleras: hotel limpio; atención del personal; habitaciones e instalaciones; comidas y bebidas; ubicación y acceso; precio y opinión general. El modelo clasifica cada reseña asignando probabilidades a cada categoría, donde la suma total es 1.

La Ilustración 2 presenta la distribución de clasificaciones obtenidas para las 50 reseñas analizadas. Los resultados muestran que Hotel limpio es la categoría más identificada con 22 reseñas, seguida de Atención del personal (10) y Habitaciones e instalaciones (9). Las categorías Comidas y bebidas y Ubicación y acceso presentan menor frecuencia (5 y 4 reseñas, respectivamente). Estos resultados sugieren que los aspectos vinculados con la higiene, el confort y la atención del personal son los elementos más destacados

en las valoraciones de los turistas. Los ejemplos detallados de clasificación individual, incluyendo las probabilidades específicas asignadas a cada categoría, se encuentran disponibles en el *notebook* de análisis para una comprensión más profunda del proceso de clasificación.

Ilustración 2. Resultados de la clasificación de texto de reseñas turísticas mediante el modelo BETO (BERT español)



Fuente: Elaboración propia.

3.3. Modelado de temas

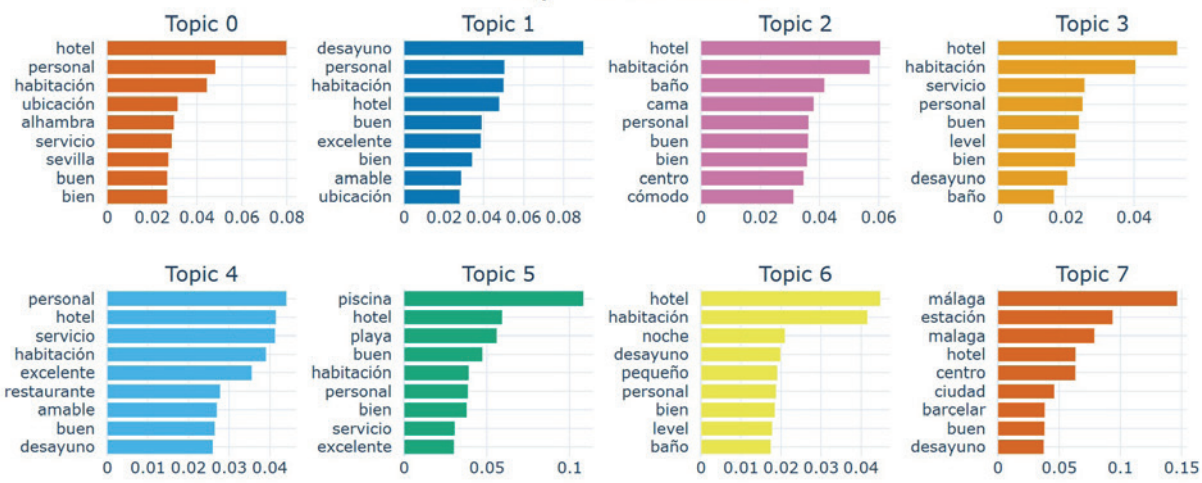
Es una técnica que permite descubrir automáticamente los temas presentes en grandes volúmenes de texto. Para este análisis se empleó BERTopic, una técnica avanzada que combina representaciones semánticas generadas por modelos Transformers (como BERT) con algoritmos de agrupamiento y el método c-TF-IDF, lo que permite identificar temas coherentes y contextualmente relevantes dentro del contenido textual (Grootendorst, 2021).

Para el análisis se seleccionó una muestra aleatoria de 2000 reseñas del conjunto original de datos. El modelo BERTopic se configuró con parámetros específicos de reducción de dimensionalidad mediante UMAP, agrupamiento K-Means y lenguaje establecido en español. Esta configuración permitió obtener ocho temas principales relacionados con distintos aspectos de la experiencia hotelera.

La Ilustración 3 muestra las palabras clave más representativas de cada tema, observándose la aparición de conceptos vinculados con la atención del personal, la

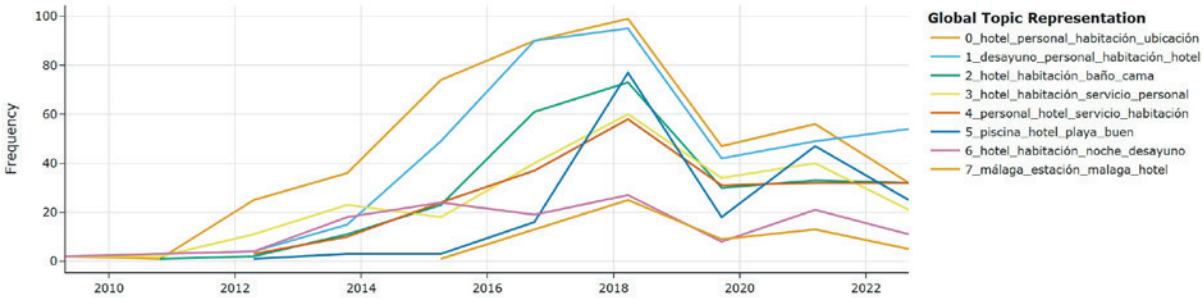
limpieza, las instalaciones y la ubicación, lo que coincide con las categorías identificadas en el análisis de clasificación. La Ilustración 4 presenta la evolución temporal de los temas, donde se aprecia un incremento sostenido en la frecuencia de menciones entre 2014 y 2018, seguido de una ligera reducción posterior, lo que podría reflejar cambios en el volumen de reseñas o en los intereses de los turistas a lo largo del tiempo. Finalmente, la Ilustración 5 muestra el dendrograma jerárquico que agrupa los temas más cercanos semánticamente, destacando la relación entre aquellos centrados en la experiencia del personal y los vinculados con las habitaciones y servicios. Los detalles del código, así como las visualizaciones interactivas y la asignación de temas a cada reseña, se incluyen en el notebook correspondiente (ver Anexo 1).

Ilustración 3. Temas identificados por BERTopic y sus palabras clave asociadas



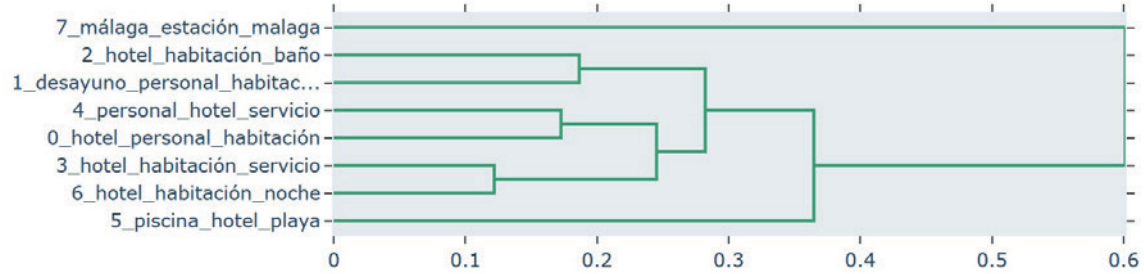
Fuente: Elaboración propia a partir de visualizaciones generadas con BERTopic.

Ilustración 4. Evolución temporal de los temas identificados por BERTopic.



Fuente: Elaboración propia a partir de visualizaciones generadas con BERTopic.

Ilustración 5. Agrupación jerárquica de los temas identificados por BERTopic



Fuente: Elaboración propia a partir de visualizaciones generadas con BERTopic.

4. DISCUSIÓN Y CONCLUSIONES

Los resultados obtenidos evidencian el potencial de las técnicas de ciencia de datos para analizar UGC en el ámbito turístico. En primer lugar, el análisis de sentimientos mostró una alta coherencia entre la polaridad textual y las calificaciones otorgadas, lo que sugiere que las reseñas reflejan de manera fiable la satisfacción de los turistas. En segundo lugar, el enfoque *zero-shot* demostró su utilidad práctica al permitir la categorización automática de reseñas sin requerir entrenamiento previo, facilitando el análisis eficiente de grandes volúmenes de comentarios. En tercer lugar, el modelado de temas con BERTopic permitió identificar estructuras semánticas latentes y patrones de discurso, ofreciendo una comprensión más profunda de las dimensiones que conforman la experiencia turística. Estas técnicas proporcionan una aproximación complementaria y consistente para explorar percepciones y comportamientos en entornos digitales.

Para maximizar el impacto de esta guía, se recomienda un enfoque progresivo comenzando con técnicas básicas como análisis de sentimientos antes de abordar métodos más complejos. Es fundamental mantener consideraciones éticas rigurosas, especialmente en la recolección y análisis de datos personales. La documentación detallada de cada fase metodológica asegura la replicabilidad y transparencia del proceso. Dado el carácter iterativo del análisis de datos, se recomienda conservar tanto ensayos exitosos como fallidos, y realizar pruebas piloto con conjuntos de datos reducidos antes de implementaciones a gran escala.

4.1. Implicaciones teóricas

Esta guía práctica contribuye a reducir la brecha metodológica identificada en la literatura turística (Egger, 2022; Köseoglu et al., 2020; Mariani et al., 2018; Mariani & Baggio, 2022), al ofrecer un marco claro y accesible para aplicar técnicas de ciencia de datos. Su principal aporte radica en facilitar la comprensión y adopción de metodologías avanzadas, promoviendo el uso de técnicas de PLN y modelos de aprendizaje profundo como BERT. Estos enfoques ya han sido aplicados en investigaciones recientes sobre análisis de reseñas turísticas (Chai et al., 2021; Rey-Moreno et al., 2023; Zhang et al., 2023), demostrando su potencial para clasificar, extraer patrones y analizar grandes volúmenes de texto.

Asimismo, el trabajo enriquece la literatura sobre el análisis del UGC, un campo que ha evolucionado desde enfoques descriptivos hacia métodos computacionales más sofisticados, como el análisis de sentimientos, la identificación de temas y la exploración de relaciones semánticas (Cai et al., 2024). Este enfoque permite aprovechar de forma más eficiente la información disponible en la red para la investigación turística, dado que el UGC constituye una fuente dinámica y continua de datos que facilita la identificación de tendencias, percepciones y patrones de comportamiento de los turistas, aspectos que resultan difíciles de observar mediante métodos de investigación tradicionales (Daugherty et al., 2008; León et al., 2025; Lu & Stepchenkova, 2015).

4.2. Implicaciones prácticas

El marco metodológico propuesto ofrece aplicaciones concretas para distintos ámbitos del sector turístico, facilitando la transformación de datos en conocimiento útil para la toma de decisiones. En *marketing* turístico, permite la segmentación automatizada de mercados, el seguimiento en tiempo real de la reputación de marca mediante análisis de sentimientos, la detección temprana de crisis reputacionales y la personalización de campañas publicitarias a partir de las percepciones expresadas en el UGC.

En cuanto a la planificación y gestión de destinos, el uso de técnicas de ciencia de datos contribuye a la gestión del *overtourism* mediante el análisis de patrones temporales y espaciales, la identificación de nuevas atracciones potenciales, la mejora de infraestructuras y servicios públicos a partir del *feedback* de los visitantes, y la optimización de recursos turísticos a través de modelos predictivos de demanda. Desde una perspectiva de política pública, estas metodologías permiten el diseño y evaluación de estrategias turísticas basadas en evidencia empírica, así como la medición del impacto de las políticas mediante análisis longitudinales de sentimientos y temas.

4.3. Limitaciones de la investigación

Este trabajo presenta varias limitaciones que deben considerarse para futuros desarrollos. Primero, la dependencia de herramientas y modelos externos como Hugging Face o Octoparse implica riesgos de disponibilidad a largo plazo, ya que estos servicios podrían modificar sus APIs o condiciones de acceso. Se recomienda que los investigadores, cuando sea posible y las licencias lo permitan, mantengan copias locales de los modelos utilizados. Segundo, aunque se proporciona un archivo «requirements.txt» con versiones específicas, la rápida evolución de las bibliotecas de Python puede generar incompatibilidades futuras. Tercero, los ejemplos presentados se limitan a reseñas en español de hoteles andaluces, lo que puede limitar la generalización a otros contextos geográficos o tipos de establecimientos turísticos.

4.4. Futuras líneas de investigación

Se proponen varias direcciones para futuras investigaciones. Primero, la integración de herramientas de inteligencia artificial generativa (como ChatGPT, Claude o Gemini) en el marco metodológico propuesto representa una oportunidad prometedora para automatizar tareas de preprocesamiento, interpretación de patrones, facilitar la traducción

y adaptación cultural de análisis multilingües. Asimismo, la aplicación de estos modelos podría fortalecer los procesos de segmentación de mercados, análisis de reputación de marca y planificación territorial mediante la simulación de escenarios turísticos.

Segundo, el desarrollo de modelos específicos para el dominio turístico, entrenados con corpus especializados, podría mejorar la precisión de las técnicas presentadas. Tercero, la incorporación de análisis multimodal (texto, imágenes, videos) ampliaría significativamente las capacidades analíticas del marco propuesto. Finalmente, el desarrollo de *interfaces* gráficas intuitivas podría facilitar aún más la adopción de estas herramientas por parte de investigadores sin formación técnica avanzada.

5. NOTAS

1) Una API (Interfaz de Programación de Aplicaciones, *Application Programming Interface*, por sus siglas en inglés) es un conjunto de mecanismos, protocolos y herramientas que permite que diferentes aplicaciones, sistemas o servicios se comuniquen entre sí de manera estructurada.

2) BERT es un modelo de lenguaje desarrollado por Google en el año 2018 que emplea la arquitectura de *Transformers* (Devlin et al., 2018). Los modelos *Transformers* son redes neuronales que son capaces de comprender el significado contextual al examinar cómo se relacionan entre sí las palabras en una secuencia. Utilizan técnicas matemáticas como la atención propia para identificar estas conexiones y capturar el contexto de la información (Vaswani et al., 2017).

6. REFERENCIAS

Alhoshan, W., Ferrari, A., & Zhao, L. (2023). Zero-shot learning for requirements classification: An exploratory study. *Information and Software Technology*, *159*, 107202. https://doi.org/10.1016/J.INFSOF.2023.107202

António, N., & Rita, P. (2023). Twenty-two years of International Journal of Hospitality Management: A bibliometric analysis 2000-2021. *International Journal of Hospitality Management*, *114*. https://doi.org/10.1016/j.ijhm.2023.103578

Bigné, E., Oltra, E., & Andreu, L. (2019). Harnessing stakeholder input on Twitter: A case study of short breaks in Spanish tourist cities. *Tourism Management*, *71*, 490-503. https://doi.org/10.1016/J.TOURMAN.2018.10.013

Cai, Y., Li, G., Wen, L., & Liu, C. (2024). Intellectual landscape and emerging trends of big data research in hospitality and tourism: A scientometric analysis. *International Journal of Hospitality Management*, *117*, 103633. https://doi.org/10.1016/J.IJHM.2023.103633

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish Pre-trained BERT Model and Evaluation Data. *ArXiv*. https://doi.org/10.48550/arXiv.2308.02976

Cervera, D. de J., de Esteban Curiel, J., & Pérez-Bustamante Yábar, D. C. (2024). Machine Learning for short-term property rental pricing based on seasonality and proximity to food establishments. *British Food Journal*, *126*(13), 332-352. https://doi.org/10.1108/BFJ-07-2023-0634

Chai, C., Song, Y., & Qin, Z. (2021). A Thousand Words Express a Common Idea? Understanding International Tourists' Reviews of Mt. Huangshan, China, through a Deep Learning Approach. *Land*, *10*(6), 549. https://doi.org/10.3390/LAND10060549

D'Acunto, D., Filieri, R., & Amato, S. (2024). Who is sharing green eWOM? Big data evidence from the travel and tourism industry. *Journal of Sustainable Tourism*, 1–23. https://doi.org/10.1080/09669582.2024.2328103

D'Acunto, D., Tuan, A., Dalli, D., Viglia, G., & Okumus, F. (2020). Do consumers care about CSR in their online reviews? An empirical analysis. *International Journal of Hospitality Management*, *85*. https://doi.org/10.1016/j.ijhm.2019.102342

Daugherty, T., Eastin, M. S., & Bright, L. (2008). Exploring Consumer Motivations for Creating User-Generated Content. *Journal of Interactive Advertising*, *8*(2), 16–25. https://doi.org/10.1080/15252019.2008.10722139

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*. https://doi.org/10.48550/arXiv.1810.04805

Egger, R. (2022). *Tourism on the verge. Applied data science in tourism: Interdisciplinary approaches, methodologies, and applications*. Springer.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/J.IJINFOMGT.2014.10.007

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining*. Springer Cham. https://doi.org/10.1007/978-3-319-10247-4

Garijo, D., Alper, P., Belhajjame, K., Corcho, O., Gil, Y., & Goble, C. (2014). Common motifs in scientific workflows: An empirical analysis. *Future Generation Computer Systems*, *36*, 338–351. https://doi.org/10.1016/J.FUTURE.2013.09.018

George, G., Osinga, E., Lavie, D., & Scott, B. (2016). Big Data and Data Science Methods for Management Research. *Academy of Management Journal*, *59*(5), 1493–1507. https://doi.org/10.5465/AMJ.2016.4005

Grootendorst, M. (2021). *MaartenGr/BERTopic: Fix embedding parameter*. https://doi.org/10.5281/ZENODO.4430182

Guerrero-Rodriguez, R., Álvarez-Carmona, M., Aranda, R., & López-Monroy, A. P. (2023). Studying Online Travel Reviews related to tourist attractions using NLP methods: the case of Guanajuato, Mexico. *Current Issues in Tourism*, *26*(2), 289–304. https://doi.org/10.1080/13683500.2021.2007227

International Journal of Contemporary Hospitality Management. (n.d.-a). *Virtual special issue: Artificial intelligence (AI) in hospitality and tourism*. Emerald Publishing. Retrieved October 27, 2025, from https://www.emeraldgrouppublishing.com/journal/ijchm/virtual-special-issue-artificial-intelligence-ai-hospitality-and-tourism

International Journal of Contemporary Hospitality Management. (n.d.-b). *Virtual special issue: Big data in hospitality and tourism*. Emerald Publishing. Retrieved October 27, 2025, from https://www.emeraldgrouppublishing.com/journal/ijchm/virtual-special-issue-big-data-hospitality-and-tourism

Köseoglu, M. A., Mehraliyev, F., Altin, M., & Okumus, F. (2020). Competitor intelligence and analysis (CIA) model and online reviews: integrating big data text mining with network analysis for strategic analysis. *Tourism Review*, *76*(3), 529–552. https://doi.org/10.1108/TR-10-2019-0406

Lalicic, L., Marine-Roig, E., Ferrer-Rosell, B., & Martin-Fuentes, E. (2021). Destination image analytics for tourism design: An approach through Airbnb reviews. *Annals of Tourism Research*, *86*, 103100. https://doi.org/10.1016/J.ANNALS.2020.103100

León, C. J., Suárez-Rojas, C., Cazorla-Artiles, J. M., & González Hernández, M. M. (2025). Satisfaction and sustainability concerns in whale-watching tourism: A user-generated content model. *Tourism Management*, *106*, 105019. https://doi.org/10.1016/J.TOURMAN.2024.105019

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, *68*, 301–323. https://doi.org/10.1016/J.TOURMAN.2018.03.009

Li, X., & Law, R. (2020). Network analysis of big data research in tourism. *Tourism Management Perspectives*, *33*, 100608. https://doi.org/10.1016/J.TMP.2019.100608

Liang, X., Li, X., Shu, L., Wang, X., & Luo, P. (2024). Tourism demand forecasting using graph neural network. *Current Issues in Tourism*. https://doi.org/10.1080/13683500.2024.2320851

Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management*, *47*, 140–151. https://doi.org/10.1016/J.TOURMAN.2014.09.020

Lu, W., & Stepchenkova, S. (2015). User-Generated Content as a Research Mode in Tourism and Hospitality Applications: Topics, Methods, and Software. *Journal of Hospitality Marketing & Management*, *24*(2), 119–154. https://doi.org/10.1080/19368623.2014.907758

Manning, C. D., Raghavan, P., & Schütze, H. (2009). Introduction to Information Retrieval. In *Introduction to Information Retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Mariani, M., & Baggio, R. (2022). Big data and analytics in hospitality and tourism: a systematic literature review. *International Journal*

of Contemporary Hospitality Management, *34*(1), 231–278. https://doi.org/10.1108/IJCHM-03-2021-0301

Mariani, M., Baggio, R., Fuchs, M., & Höepken, W. (2018). Business intelligence and big data in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management*, *30*(12), 3514-3554. https://doi.org/10.1108/IJCHM-07-2017-0461

Meneghini, A., & Tuzzi, A. (2025). Can tourism be a mean for promoting inclusive development? A textual analysis of funded projects in Greece. *Current Issues in Tourism*. https://doi.org/10.1080/13683500.2025.2485381

Mitchell, R. (2018). *Web scraping with Python: Collecting more data from the modern web*. O'Reilly Media, Inc.

Mor, M., Dalyot, S., & Ram, Y. (2023). Who is a tourist? Classifying international urban tourists using machine learning. *Tourism Management*, *95*, 104689. https://doi.org/10.1016/J.TOURMAN.2022.104689

Pennington, D., Ebert-Uphoff, I., Freed, N., Martin, J., & Pierce, S. A. (2020). Bridging sustainability science, earth science, and data science through interdisciplinary education. *Sustainability Science*, *15*(2), 647–661. https://doi.org/10.1007/S11625-019-00735-3

Pereira-Moliner, J., Villar-García, M., Molina-Azorín, J. F., Tarí, J. J., López-Gamero, M. D., & Pertusa-Ortega, E. M. (2024). Using tourism intelligence and big data to explain flight searches for tourist destinations: The case of the Costa Blanca (Spain). *Tourism Management Perspectives*, *51*, 101243. https://doi.org/10.1016/J.TMP.2024.101243

Ramasamy, D., Sarasua, C., Bacchelli, A., & Bernstein, A. (2023). Workflow analysis of data science code in public GitHub repositories. *Empirical Software Engineering*, *28*(1), 1–47. https://doi.org/10.1007/S10664-022-10229-Z

Recognai. (2021). *Model: bert-base-spanish-wwm-cased-xnli*. https://huggingface.co/Recognai/bert-base-spanish-wwm-cased-xnli

Rey-Moreno, M., Sánchez-Franco, M. J., De La, M., & Rey-Tienda, S. (2023). Examining transaction-specific satisfaction and trust in Airbnb and hotels. An application of BERTopic and Zero-shot text classification. *Tourism & Management Studies, 19*(2), 21-37. https://doi.org/10.18089/tms.2023.190202

Saoualih, A., Shen, S., Safaa, L., & Su, Y. (2025). A thematic investigation of tourist experiential gaps in Moroccan cultural heritage museums using a sentiment-guided BERTopic text mining approach. *Tourism Recreation Research*. https://doi.org/10.1080/02508281.2025.2538250

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104–112. https://doi.org/10.1016/J.IPM.2013.08.006

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/JEP.28.2.3

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv*, *2017-December*, 5999-6009. https://doi.org/10.48550/arXiv.1706.03762

Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism Management*, *31*(2), 179-188. https://doi.org/10.1016/J.TOURMAN.2009.02.016

Xu, F., Nash, N., & Whitmarsh, L. (2019). Big data or small data? A methodological review of sustainable tourism. *Journal of Sustainable Tourism*, *28*(2), 144-163. https://doi.org/10.1080/09669582.2019.1631318

Zhang, H., Liu, R., & Egger, R. (2023). Unlocking Uniqueness: Analyzing Online Reviews of Airbnb Experiences Using BERT-based Models. *Journal of Travel Research*. https://doi.org/10.1177/00472875231197381

Agradecimientos

Este trabajo forma parte del Proyecto de Investigación «TED2021-131314B-I00 SOSTENIBILIDAD CORPORATIVA Y TURISMO INTELIGENTE EN COMUNIDADES RURALES: INFLUENCIA EN EL DESARROLLO ECONÓMICO Y SOCIAL DEL TERRITORIO», financiado por MICIU/AEI/10.13039/501100011033 y por la Unión Europea "NextGenerationEU"/PRTR.

ANEXOS

Anexo 1. Repositorio de GitHub del estudio.

El repositorio se encuentra disponible en: https://github.com/marib5635-web/tourism-data-science.git. Contiene todos los recursos necesarios para replicar los análisis presentados en este estudio, incluyendo los archivos de código en formato .ipynb, los conjuntos de datos en .xlsx y el archivo requirements.txt con las versiones exactas de las bibliotecas utilizadas. El *notebook* está diseñado para ejecutarse directamente en Google Colab, facilitando el acceso a investigadores y profesionales sin necesidad de configuraciones locales. Cada sección del código está comentada para guiar al usuario a lo largo de las etapas del flujo de trabajo: preprocesamiento, análisis de sentimientos, clasificación de texto y modelado de temas.

ENSAYO