UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE COMERCIO y TURISMO

TESIS DOCTORAL

Aplicación de técnicas de inteligencia artificial en la clasificación de eventos para el marketing de destinos turísticos

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADO POR

Miguel Camacho Ruiz

DIRECTORES

Ramón Alberto Carrasco González y Antonio LaTorre de la Fuente

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE COMERCIO y TURISMO

DOCTORADO INTERUNIVERSITARIO EN TURISMO

TESIS DOCTORAL

Aplicación de técnicas de inteligencia artificial en la clasificación de eventos para el marketing de destinos turísticos

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADO POR

Miguel Camacho Ruiz

DIRECTORES

Ramón Alberto Carrasco González y Antonio LaTorre de la Fuente

Dedicatoria

A Miguel, mi padre, que me enseñó a ser curioso.

A Kris, mi mujer, por apoyarme y aguantarme en todos mis empeños.

Agradecimientos

Un doctorado no es un trabajo individual, aunque el título se lo den solo a una persona. Aquí mencionaré solo a algunas de las personas a las que estoy agradecido, aunque hay muchas, muchas más.

En primer lugar, estoy sumamente agradecido a mis directores de tesis, Ramón Alberto Carrasco y Antonio LaTorre por su apoyo durante estos años. Sus consejos y enseñanzas me han servido para transitar este camino de forma satisfactoria. Mi agradecimiento también a mi tutor Adolfo Hernández Estrada por su apoyo y trabajo.

Este doctorado ha sido un desafío para mí, me ha costado compaginarlo con mi trabajo y otras pasiones. Por suerte, Ramón Alberto Carrasco me ha acompañado pacientemente, desde que me animó a enrolarme en el programa hasta su último día. Me ha inspirado, instruido y corregido y por ello, le estoy agradecido.

Quiero agradecer también al equipo académico de la Facultad de Comercio y Turismo que me han ayudado a gestionar mi formación, en especial a Manuel de la Calle, gracias al cual he encontrado formaciones fantásticas.

También quiero incluir en mi agradecimiento a Gema Fernández-Avilés por su contribución a mi investigación y sus valiosos consejos.

Me gustaría cerrar dando las gracias a mi familia y amigos. Son ellos los que han aguantado mi cansancio o frustración y mis explicaciones sobre temas extraños, aunque también serán ellos con quien celebre más especialmente el final de esta etapa. Gracias Kris por apoyarme en todo momento, por escucharme y por compartir este viaje. Gracias Mamen y Sergio por ayudarme a mantener la ilusión y por vuestra curiosidad genuina. Gracias Pedro por tu curiosidad y por dejarme contarte tantos detalles, me ha ayudado enormemente. Gracias Ana, madre, por tu apoyo y cariño. Y gracias a ti, Miguel, padre. De alguna forma esto lo hago por ti y gracias a ti.

Tabla de contenido

Agradecimientos ... 3

Índice de figuras ... 7

Índice de tablas ... 8

Índice de ecuaciones ... 8

Índice de acrónimos ... 9

Resumen ... 11

Abstract ... 12

1. Introducción ... 13

1.1 Contexto y definiciones ... 15

1.2 Justificación del trabajo ... 18

1.2.1 Impacto económico del evento turístico ... 18

1.2.2 Impacto del marketing centrado en eventos turísticos ... 19

1.3 Objetivos ... 21

1.3.1 Objetivo general ... 21

1.3.2 Objetivos específicos ... 21

1.3.3 Cuestiones de investigación ... 22

1.4 Estructura de la memoria ... 22

2. Estado del arte ... 24

2.1. Uso de eventos para la toma de decisiones en el sector turístico ... 24

2.2. Creación de catálogos de eventos a través de sistemas taxonómicos o tipológicos ... 27

3. Fundamentos metodológicos ... 37

3.1. Metodología CRISP-DM ... 37

3.1.1. Comprensión del negocio ... 38

3.1.2. Comprensión de los datos ... 39

3.1.3. Preparación de los datos ... 40

3.1.4. Modelado ... 42

3.1.5. Evaluación ... 43

3.1.6. Despliegue ... 44

3.2. Procesamiento del lenguaje natural 45
3.2.1. BERT 45
3.3. Clasificación multiclase 50
3.3.1. Regresión logística 50
3.3.2. Evaluación de modelos de clasificación multi-clase 51
4. Modelo de clasificación de eventos turísticos 55
4.1. Comprensión del negocio 56
4.1.1. Objetivos de negocio 56
4.1.2. Desafíos de la creación del modelo 57
4.1.3. Objetivos para la ciencia de datos del modelo 60
4.1.4. Beneficios potenciales del modelo 62
4.2. Comprensión de los datos 64
4.2.1. ¿Qué datos necesito? 65
4.2.2. Recolección de datos 67
4.2.3. Análisis exploratorio de datos y filtrado inicial 69
4.2.4. Control de calidad 74
4.2.5. Comprensión de datos para nuestro caso de uso 75
4.3. Preparación de los datos 76
4.4. Modelado 78
4.4.1. Fase de entrenamiento 83
4.4.2. Fase de clasificación 85
4.5. Evaluación 86
4.5.1. Desempeño del modelo en la clasificación 86
4.5.2. Robustez ante variaciones en los datos 91
4.5.3. Eficiencia y escalabilidad 94
4.5.4. Limitaciones y áreas de mejora 95
4.5.5. Evaluación para nuestro caso de uso 97
5. Conclusiones y trabajo futuro 99
5.1. Conclusiones generales 99
5.2. Limitaciones 102
5.3. Trabajo futuro 104

6. Anexos 106
Publicaciones en revistas 106
Tourism destination events classifier based on artificial intelligence techniques, 2023 106
Congresos 110
Bibliografía 113

Índice de figuras

Figura 1: Esquema conceptual para la clasificación automática de eventos turísticos 15
Figura 2: Ejemplo de una matriz de confusión para un problema de clasificación de 4 clases 52
Figura 3 Modelo propuesto basado en CRISP-DM 55
Figura 4: Taxonomía de eventos turísticos 67
Figura 5: Conteo de eventos por categoría 72
Figura 6: Histograma del conteo de eventos por número de palabras del título 73
Figura 7: Histograma del conteo de eventos por número de palabras de su texto 74
Figura 8: Ejemplo de textos mal formados en eventos 76
Figura 9: Flujo de datos de una secuencia de la fase de clasificación 85
Figura 10: Ejemplo de sentencias clasificadas con su valor predicho por el modelo y su verdad base 87
Figura 11: Matriz de confusión del conjunto de datos de prueba 88
Figura 12: Matriz de confusión del conjunto de datos de prueba con ratios en lugar de conteos 89
Figura 13: Todas las muestras con sentencias de 2 caracteres del conjunto de datos de test 91
Figura 14: Tasa de acierto vs. Conteo de palabras vs. soporte 92
Figura 15: Detalle de la Tasa de acierto vs. Conteo de palabras vs. soporte 93
Figura 16: Errores de clasificación con los momios de cada categoría 97

Índice de tablas

Tabla 1: Estudios sobre clasificación de eventos turísticos 36
Tabla 2: Conteo de eventos por taxonomía..... 72
Tabla 3: Mapeo de los identificadores de categorías..... 78
Tabla 4: Informe de clasificación del conjunto de datos de prueba..... 89
Tabla 5: Reporte de clasificación del conjunto de datos de prueba para eventos limitados a 20 palabras 93
Tabla 6: Resumen de equipos utilizados en el entrenamiento y clasificación y sus tiempos de procesado 95

Índice de ecuaciones

Ecuación 1: Descripción matemática del mecanismo de atención de un transformer 47
Ecuación 2: Atención multi-cabeza 47
Ecuación 3: Función logística..... 50
Ecuación 4: Logaritmo del ratio las probabilidades de pertenecer a una clase..... 50
Ecuación 5: Función logística multinomial 51
Ecuación 6: Fórmula de la exactitud..... 53
Ecuación 7: Fórmula de la sensibilidad..... 53
Ecuación 8: Fórmula de la precisión..... 53
Ecuación 9: Fórmula de la tasa de falsos positivos 53
Ecuación 10: Fórmula del F1-score 53

Índice de acrónimos

API: Application Programming Interface

AUC: Area Under the Curve

BERT: Bidirectional Encoder Representations from Transformers

CRISP-DM: Cross Industry Standard Process for Data Mining

CRM: Customer Relationship Management

CTR: Tasa de clics

FN: Falsos Negativos

FP: Falso Positivo

FPR: Tasa de Falsos Positivos

HTML: HyperText Markup Language

IA: Inteligencia Artificial

IoT: Internet de las cosas

JSON: JavaScript Object Notation

KDD: Knowledge Discovery in Databases

KNN: K-Nearest Neighbor

KPI: Key Performance Indicator

MICE: Meetings, Incentives, Conventions, and Exhibitions

MLM: Masked Language Modeling

OTA: Online Travel Agency

PLN: Procesamiento del Lenguaje Natural

RNN: Recurrent Neural Network

ROC: Receiver Operating Characteristic

ROI: Return of Investion

SEMMA: Sample, Explore, Modify, Model, Assess

TE: Tiempo de Entrenamiento

TI: Tiempo de Inferencia

TN: Verdaderos Negativos

TP: Verdadero Positivo

TPR: Tasa de Verdaderos Positivos

URL: Uniform Resource Locator

Resumen

Esta tesis se centra en cómo mejorar la eficacia del marketing turístico a través de la clasificación automática de eventos turísticos. En un contexto donde la personalización y la segmentación son claves para atraer y retener viajeros, disponer de catálogos de eventos organizados y coherentes es esencial para implementar estrategias de marketing digital más precisas, relevantes y alineadas con los intereses del cliente.

Sin embargo, actualmente existe una gran heterogeneidad en la forma en que los eventos son descritos y clasificados entre diferentes fuentes, idiomas y formatos. Esta falta de estandarización limita la capacidad de destinos, aerolíneas, agencias online (OTAs) y otros actores del sector turístico para presentar contenidos personalizados, generar ofertas relevantes o realizar análisis estratégicos sobre la demanda de eventos.

Esta tesis propone y desarrolla un sistema automático de clasificación de eventos turísticos basado en técnicas de procesamiento del lenguaje natural y aprendizaje automático. El modelo está entrenado con descripciones reales de eventos y utiliza un sistema taxonómico que permite clasificar los eventos en distintas categorías reutilizables y adaptables a múltiples entornos.

El modelo ha sido validado mediante un caso de uso simulado en una compañía aérea, demostrando su aplicabilidad comercial, escalabilidad y eficacia para enriquecer la experiencia del usuario, mejorar la conversión de campañas y facilitar la creación de catálogos dinámicos adaptados a segmentos específicos de mercado.

Los resultados muestran que el sistema es capaz de categorizar eventos de forma efectiva, lo que permite construir catálogos dado un sistema taxonómico. Esto genera beneficios tanto para los turistas, al facilitar la personalización de la oferta, como para actores del sector turístico, como destinos, agencias o aerolíneas, que pueden diseñar estrategias más precisas y eficientes.

Abstract

This thesis focuses on enhancing the effectiveness of tourism marketing through the automatic classification of touristic events. In a context where personalization and segmentation are key to attracting and retaining travelers, having organized and coherent event catalogs is essential for implementing more accurate, relevant, and customer-aligned digital marketing strategies.

However, there is currently significant heterogeneity in how events are described and classified across different sources, languages, and formats. This lack of standardization limits the ability of destinations, airlines, online travel agencies (OTAs), and other tourism stakeholders to present personalized content, generate relevant offers, or conduct strategic analyses of event demand.

This thesis proposes and develops an automatic event classification system based on natural language processing and machine learning techniques. The model is trained using real event descriptions and uses a taxonomic system that allows classifying events into reusable categories adaptable to multiple environments.

The model was validated through a simulated use case involving an airline, demonstrating its commercial applicability, scalability, and effectiveness in enhancing user experience, increasing campaign conversion, and supporting the creation of dynamic catalogs tailored to specific market segments.

The results show that the system can categorize events effectively, enabling the construction of catalogs within a taxonomic system. This provides benefits for tourists, by facilitating offer personalization, and for tourism sector stakeholders such as destinations, agencies, and airlines, who can design more precise and efficient marketing strategies.

1. Introducción

En la era digital actual, el turismo se ha convertido en un sector económico crucial a nivel global, caracterizado por su naturaleza dinámica y en constante evolución. Los eventos turísticos son un elemento fundamental que impulsa esta industria, no solo atrayendo visitantes, sino también contribuyendo significativamente a la identidad y atractivo de los destinos.

Los eventos turísticos desempeñan un papel vital en la toma de decisiones de los viajeros y en la gestión de destinos turísticos. Pueden aumentar el flujo de turistas, fomentar el gasto turístico y fortalecer la identidad del destino. Un estudio realizado por Gration et al. (2016) reveló que los eventos pueden actuar como catalizadores para el desarrollo económico y la regeneración urbana, además de mejorar la imagen del destino.

Para los turistas, la disponibilidad de eventos alineados con sus intereses puede ser un factor decisivo en la elección de un destino. Según una investigación de John L. Crompton (1997), los eventos satisfacen múltiples necesidades de los visitantes, incluyendo la socialización, el escape de la rutina diaria y el enriquecimiento cultural.

La industria turística ofrece a los viajeros experiencias, y los eventos turísticos constituyen una parte fundamental de esas experiencias. El crecimiento exponencial del turismo centrado en eventos requiere de un enfoque estructurado que permita comprender, categorizar y optimizar la promoción de los eventos. En este contexto, surge la necesidad de construir un catálogo normalizado de eventos que no solo facilite su gestión, sino que también responda a las necesidades, intereses y comportamientos del viajero, en línea con la filosofía del marketing centrado en el cliente. Esta perspectiva reconoce que, para influir en la decisión de viaje, es esencial ofrecer información personalizada, relevante y fácilmente accesible. Por tanto, contar con una clasificación coherente de eventos permite mejorar la experiencia del usuario en plataformas digitales y refuerza la eficacia de las estrategias de marketing. Esto implica tanto el reconocimiento de los distintos tipos de eventos (Getz & Page, 2014; Oklobdzija, 2015) como la comprensión de cómo herramientas como el marketing digital afectan el comportamiento del viajero (Zarotis, 2021).

La industria turística ofrece a los viajeros experiencias, y los eventos turísticos son una parte importante de esas experiencias. El crecimiento exponencial del turismo centrado en eventos requiere de un enfoque estructurado que permita comprender, categorizar y optimizar la promoción de los eventos. En este contexto, surge la necesidad de construir un catálogo normalizado de eventos que no solo facilite su gestión, sino que también responda a las necesidades, intereses y comportamientos

del viajero, en línea con la filosofía del marketing centrado en el cliente. Esta perspectiva reconoce que, para influir en la decisión de viaje, es esencial ofrecer información personalizada, relevante y fácilmente accesible. Por tanto, contar con una clasificación coherente de eventos permite mejorar la experiencia del usuario en plataformas digitales y refuerza la eficacia de las estrategias de marketing. Esto implica tanto el reconocimiento de los distintos tipos de eventos (Getz & Page, 2014; Oklobdzija, 2015) como la comprensión de cómo herramientas como el marketing digital afectan el comportamiento del viajero (Zarotis, 2021). En este trabajo exploraremos los métodos y técnicas que se han usado hasta el momento para lograr esa estandarización, desde propuestas taxonómicas para eventos turísticos hasta esfuerzos por clasificar dichos eventos.

A pesar de la importancia de la segmentación de eventos en la gestión de destinos turísticos, existe una carencia notable de una taxonomía estandarizada para clasificar estos eventos. La mayoría de los eventos turísticos se publicitan en internet a través de agregadores, afiliados o sitios de organizadores, pero carecen de un estándar ampliamente utilizado para sus taxonomías.

Esta falta de estandarización dificulta tanto la gestión eficiente por parte de los destinos como la búsqueda y selección por parte de los turistas. Un estudio de Getz (2008) propuso una clasificación de eventos planificados en el turismo, pero su implementación práctica ha sido limitada debido a la complejidad y diversidad de los eventos turísticos.

Un estudio de Gration et al. (2016) destacó la necesidad de sistemas de clasificación más sofisticados para eventos, subrayando la relevancia de esta investigación en múltiples sectores del turismo. Además, Laing (2018) señaló la importancia de la tecnología en la gestión de eventos y la necesidad de herramientas innovadoras para mejorar la experiencia del visitante.

Por otra parte, la inteligencia artificial (IA) ha revolucionado el sector turístico, introduciendo una nueva era de personalización y eficiencia en los servicios. Según los análisis de Lu et al. (2020), Álvarez-Carmona et al. (2022), Doborjeh et al. (2022), Dang & Nguyen (2023) y Padma & Nabi (2024), la IA se muestra como una poderosa herramienta para ayudar a solucionar muchos de los problemas a los que se enfrenta la industria del turismo. Esto es particularmente relevante para el turismo de eventos, donde las herramientas de IA pueden predecir la demanda futura de eventos y optimizar las estrategias de precios y marketing y, en definitiva, ayudar al viajero a tomar mejores decisiones y disfrutar aún más de su experiencia.

De forma resumida, el propósito principal de esta investigación es desarrollar un proceso innovador para la clasificación automática de una variedad ecléctica de

eventos turísticos utilizando una taxonomía jerárquica. Este enfoque busca crear un catálogo normalizado y universal de eventos a través de diferentes regiones geográficas. También queremos establecer una estructura coherente y reutilizable que facilite la comparación y búsqueda de eventos, independientemente de cómo estén definidos originalmente. Queremos implementar un modelo de clasificación automática basado en técnicas de aprendizaje automático y procesamiento del lenguaje natural (PLN). Por último, queremos validar su rendimiento en un escenario simulado de una compañía aérea para demostrar su eficacia, escalabilidad y aplicabilidad comercial. Un resumen visual puede verse en la Figura 1:

Figura 1: Esquema conceptual para la clasificación automática de eventos turísticos

1.1 Contexto y definiciones

Para comprender los objetivos planteados en esta investigación, es necesario definir algunos conceptos fundamentales que estructuran tanto el marco teórico como el diseño metodológico del trabajo. Esta sección tiene como propósito clarificar los términos clave y contextualizar su relevancia en el desarrollo de un sistema de clasificación automática de eventos turísticos. En esta sección encontramos solamente definiciones de los conceptos, aunque se desarrollan en el resto del cuerpo de esta tesis.

Por evento turístico se entiende cualquier acontecimiento planificado que constituye una motivación principal o complementaria para los desplazamientos turísticos. Estos eventos brindan a los turistas la oportunidad de participar en actividades únicas y memorables, mostrando con frecuencia el carácter distintivo de un destino. Los eventos pueden abarcar desde festivales locales hasta competiciones deportivas internacionales, y juegan un papel fundamental en la configuración de la percepción y experiencia que los turistas tienen de un lugar (Werner et al., 2020). Su diseño y ejecución requieren una comprensión profunda de las motivaciones, preferencias y del panorama general del turismo.

Los catálogos de eventos turísticos funcionan como listados de eventos, proporcionando a los turistas información esencial sobre fechas, ubicaciones, descripciones y otros detalles relevantes. Estos catálogos actúan como un centro de referencia para que los turistas descubran y se informen sobre los distintos eventos que tienen lugar en una región o destino específico. Al consolidar la información en un solo lugar, los catálogos simplifican el proceso de planificación y facilitan que los turistas encuentren eventos que se ajusten a sus intereses. La información que suelen incluir estos catálogos abarca programas de eventos, precios de entradas, detalles del recinto y datos de contacto, lo que permite a los turistas tomar decisiones informadas y planificar sus itinerarios en consecuencia (Alsahafi et al., 2023).

Un catálogo normalizado, por su parte, se refiere a aquel que sigue un conjunto común de criterios, categorías y estructuras, permitiendo la comparación y reutilización de la información independientemente del origen o formato de los datos. Esta normalización es fundamental para que múltiples actores —desde empresas tecnológicas hasta organismos turísticos— puedan integrar, compartir y explotar dicha información de manera eficiente.

Para construir dicho catálogo, es necesario establecer una taxonomía o una tipología. Una taxonomía es un sistema jerárquico de clasificación que agrupa los eventos en niveles progresivos de especificidad, desde categorías generales hasta subcategorías concretas. Una tipología es un método utilizado para clasificar objetos, personas o fenómenos en tipos discretos basados en características compartidas. A diferencia de la taxonomía, que busca crear un sistema de clasificación jerárquico, la tipología se centra en identificar patrones o comportamientos distintivos y agrupar los casos según sus similitudes. Este enfoque se utiliza frecuentemente en las ciencias sociales para simplificar fenómenos complejos y hacerlos más comprensibles.

Una categoría representa un grupo de eventos que comparten atributos comunes dentro de la taxonomía o tipología. La definición rigurosa y estable de estas categorías es clave para lograr consistencia, interoperabilidad entre sistemas y comprensión por parte de usuarios humanos y algoritmos.

A partir de estas definiciones, el objetivo general del trabajo, es decir, la creación de un sistema automatizado y escalable para clasificar eventos turísticos se desglosa en varios objetivos específicos que se tratarán en detalle más adelante. En esta sección de definiciones se destaca el primero de ellos por su carácter fundacional: el desarrollo de un catálogo normalizado de eventos turísticos que sea universal.

Este objetivo de universalidad implica que el sistema propuesto debe ser capaz de integrar eventos provenientes de distintas regiones geográficas, en múltiples idiomas, con distintas estructuras y matices culturales, sin perder consistencia ni precisión. Este objetivo no solo responde a la fragmentación del sector, sino también a la creciente demanda de soluciones globales por parte de plataformas digitales que operan en mercados internacionales. La multiculturalidad y la heterogeneidad de las fuentes exigen un modelo flexible pero riguroso, capaz de abstraer lo esencial del contenido de los eventos y traducirlo a una representación estructurada común.

Es esta necesidad de universalidad y estandarización la que impulsa la elección de tecnologías como el procesamiento del lenguaje natural (PLN) mediante modelos semánticos como BERT (*Bidirectional Encoder Representations from Transformers*), junto con técnicas de aprendizaje automático para la clasificación multiclase como la regresión logística.

PLN es el campo de conocimiento de la IA que estudia la forma en que las máquinas pueden comunicarse con las personas a través de lenguajes naturales como el español o el inglés.

BERT es un modelo de representación del lenguaje desarrollado por Google que se utiliza para extraer representaciones semánticas y contextualizadas de textos en lenguaje natural (Devlin et al., 2018). Así, BERT nos permite traducir texto natural a vectores matemáticos que nos permiten manipular la información de manera automática incluso cuando se encuentran expresados en distintos estilos, idiomas o niveles de formalidad. Estas representaciones son especialmente útiles para ser procesadas posteriormente por otros procesos como algoritmos de clasificación.

La regresión logística es un modelo estadístico utilizado para predecir la probabilidad de que una observación pertenezca a una determinada categoría dentro de una variable cualitativa o también llamada categórica (James et al., 2023). Por ejemplo, la probabilidad de que un evento turístico pertenezca a una categoría definida previamente en un sistema taxonómico.

La adopción de una metodología CRISP-DM (del inglés *Cross Industry Standard Process for Data Mining*) garantiza que el desarrollo del modelo esté alineado con buenas prácticas del ámbito de la ciencia de datos, desde la comprensión del negocio hasta su aplicación en un caso real.

1.2 Justificación del trabajo

En el contexto actual del turismo global, los eventos turísticos se han consolidado como elementos estratégicos clave tanto para la dinamización económica de los territorios como para la proyección de su imagen en un mercado cada vez más competitivo. Esta investigación parte de la premisa de que los eventos no solo generan ingresos inmediatos, sino que además poseen un efecto multiplicador en múltiples sectores de la economía y actúan como herramientas eficaces de marketing territorial.

En este sentido, la relevancia de desarrollar un modelo automatizado de clasificación de eventos turísticos radica en la necesidad urgente de sistematizar y optimizar la gestión de la información relacionada con dichos eventos, permitiendo su integración eficiente en plataformas digitales, motores de recomendación y estrategias promocionales. Esta necesidad se ve amplificada por la falta de una taxonomía estandarizada que permita categorizar los eventos de forma coherente a través de fuentes, regiones e idiomas.

La tesis se justifica, por tanto, en el potencial transformador que tiene una herramienta como la propuesta: una solución tecnológica que permite estructurar y escalar el uso de los eventos turísticos como palancas de desarrollo económico y de posicionamiento estratégico de destinos. Para sustentar esta afirmación, se abordan a continuación dos dimensiones fundamentales: el impacto económico de los eventos turísticos y su influencia en las estrategias de marketing y branding de destinos.

1.2.1 Impacto económico del evento turístico

Según el *World Travel and Tourism Council*, el turismo contribuyó con el 9.1% del Producto Interno Bruto (PIB) mundial en 2023, algo menos del 10.4% alcanzado en 2019, y a pesar de la disrupción generada por la pandemia de la COVID-19, las previsiones para los próximos años apuntan a una clara recuperación, en la que los eventos tendrán un efecto multiplicador (Alsahafi et al., 2023).

Los eventos turísticos son importantes motores económicos. Leng et al. (2021) demuestran cómo el análisis de datos de teléfonos móviles puede cuantificar los flujos turísticos y evaluar el impacto financiero directo e indirecto de los eventos. Estas iniciativas generan ingresos a través del gasto de los turistas en alojamiento, transporte, gastronomía y entretenimiento, y también estimulan inversiones en infraestructura y servicios relacionados. Si hablamos de mega eventos, al atraer grandes cantidades de visitantes, generan efectos multiplicadores que impactan positivamente la economía local (Oklobdzija, 2015). Getz & Page (2014) enfatizan que los eventos también pueden servir como palanca para la renovación urbana y el desarrollo de capacidades locales, incrementando el atractivo de las destinaciones en

el largo plazo. Zarotis (2021) subraya que los ingresos derivados de los eventos no se limitan al consumo inmediato, sino que también incluyen beneficios colaterales, como la promoción de la región y la atracción de futuras inversiones.

Es fácil observar el impacto de eventos turísticos cuando estos son muy grandes. Eventos como la Copa del Mundo FIFA o los juegos olímpicos inyectan ingresos relevantes en el entorno local e incluso en los países en los que se celebran. Estas fuentes de ingresos se generan con la venta de entradas, derechos de emisión o patrocinios deportivos (Bohlmann & van Heerden, 2008). Por ejemplo, se estima que el impacto de la Copa del Mundo de Qatar de 2022 dejó unos beneficios de 11 millardos de dólares, principalmente, fruto del incremento del beneficio turístico y de actividades comerciales (Özyeşil, 2023). Del mismo modo, se estima que el *World Youth Day* de 2013 en Río de Janeiro generó un impacto de 1.9 millardos de dólares (Monteiro & Marques, 2015). Se observa también que los eventos más pequeños dejan una impronta positiva en la economía local gracias al incremento de gasto de los turistas (Cudny & Paluch, 2024; Valentina Bartolic, 2020). Es cierto que los eventos muy grandes se caracterizan por una inversión previa significativa y todavía es objeto de estudio si los beneficios de esos eventos se mantienen en el largo plazo (Alalawneh et al., 2021; Barrios et al., 2016).

Otro ejemplo del impacto económico de la industria del evento turístico puede encontrarse en el llamado sector *MICE* (*Meetings, Incentives, Conventions, and Exhibitions* por sus siglas en inglés). El sector MICE representa una parte importante del turismo centrado en eventos. En concreto, la frecuente naturaleza internacional de los eventos MICE amplifica su impacto económico atrayendo participantes foráneos y generando beneficios más allá del mercado local (Smagina, 2017).

Más allá del beneficio directo, la industria del evento turístico impacta positivamente en industrias adyacentes como es la industria hotelera o la restauración. La industria del transporte también se ve favorecida por el turismo centrado en eventos (Leng et al., 2016). Además, el impacto sobre el empleo en estas regiones se extiende más allá de la duración del evento turístico en muchos casos, fortaleciendo sectores adyacentes al evento y por tanto la capacidad de generar eventos atractivos que impacten la economía local (Chen, 2024).

1.2.2 Impacto del marketing centrado en eventos turísticos

El evento turístico se debe entender como un producto/servicio orientado a la satisfacción del turista. Getz (2008) afirma que los eventos son un importante motivador del turismo, atrayendo a viajeros interesados en experiencias únicas y específicas y que satisfacen diversas necesidades de los turistas como entretenimiento, socialización, enriquecimiento cultural o desarrollo profesional

dependiendo del tipo de evento. Iliev (2020) también afirma que la oferta turística de un destino actúa como atracciones que motivan el viaje de los turistas, animando y dando vida a destinos, resorts, parques y espacios urbanos.

Según Getz & Page (2014) los eventos son instrumentales en los planes de desarrollo y marketing de la mayoría de los destinos turísticos. Son muy útiles además para alcanzar diferentes objetivos estratégicos tales como atraer turistas, especialmente en temporadas bajas cuando más falta hace un buen reclamo. También catalizan la renovación urbana y el desarrollo de la infraestructura. Los eventos turísticos también contribuyen muy positivamente al marketing del destino. Ayudan a animar otras atracciones o incluso áreas geográficas específicas y en última instancia promueven una imagen positiva del destino.

Los eventos turísticos son un reclamo que impactan en las estrategias de marketing turístico debido a su capacidad para generar experiencias memorables, atraer visitantes a un destino y fortalecer la identidad de los destinos. Desde mega eventos hasta eventos locales, éstos desempeñan un papel prominente en el desarrollo turístico. Y es que los eventos constituyen un elemento clave del sistema turístico, tanto en la fase motivacional del viajero como para mejorar la oferta de destinos mediante la animación de espacios urbanos y naturales (Getz, 2008). Eventos como los juegos olímpicos o festivales culturales o musicales icónicos pueden posicionar positivamente un destino, atrayendo turistas tanto nacionales como internacionales y estimulando la economía local. Además, estos eventos actúan como catalizadores para superar la estacionalidad y diversificar la base económica de las comunidades anfitrionas (Oklobdzija, 2015).

Los eventos se utilizan cada vez más como herramientas de branding para los destinos. Como señalan Zarotis (2021) y Oklobdzija (2015) , un evento exitoso puede crear una asociación positiva entre el destino y la experiencia vivida, mejorando su imagen y reputación. Los eventos de carácter distintivo, como los eventos de marca o los mega eventos, no solo atraen turistas, sino que también ayudan a reposicionar ciudades y regiones en mercados altamente competitivos. Zarotis destaca también que los eventos tienen un papel fundamental en la creación de experiencias que refuerzan la lealtad del turista hacia un destino. Al ofrecer vivencias únicas y emocionales, los eventos generan un impacto positivo en la percepción del visitante y fomentan visitas repetidas. Este enfoque experiencial permite a los destinos diferenciarse en un mercado saturado.

Los eventos turísticos son herramientas esenciales para el marketing turístico, capaces de transformar destinos, generar ingresos y crear experiencias inolvidables. El futuro del marketing turístico está intrínsecamente ligado a la capacidad de integrar los

eventos como parte central de las estrategias de desarrollo sostenible y branding de destinos.

1.3 Objetivos

1.3.1 Objetivo general

El objetivo general de este trabajo es desarrollar un proceso automatizado, escalable y estandarizado para la clasificación automática de eventos turísticos utilizando una taxonomía jerárquica, con el propósito de construir un catálogo estructurado de eventos que permita su aplicación práctica en plataformas digitales, motores de recomendación y herramientas de marketing turístico. Esta clasificación debe ser independiente del idioma y de la fuente de origen de los eventos, y su diseño debe estar alineado con las necesidades reales del sector turístico, en especial en entornos donde la búsqueda y la comparación de eventos por parte de los viajeros es un componente clave de la experiencia digital.

1.3.2 Objetivos específicos

Se proponen varios objetivos específicos más concretos que complementan el objetivo general:

- Desarrollar un catálogo estandarizado de eventos turísticos: Construir un sistema de clasificación que permita normalizar eventos turísticos provenientes de distintas regiones geográficas, independientemente de su fuente original.
- Establecer categorías consistentes y homogéneas para la clasificación de eventos: Diseñar un mecanismo que facilite la asignación coherente de categorías a los eventos, permitiendo su comparación, filtrado y búsqueda, sin depender de la terminología utilizada por las fuentes de origen.
- Implementar un modelo de clasificación basado en datos bajo la metodología CRISP-DM: Aplicar el enfoque estructurado de la metodología CRISP-DM para desarrollar un modelo de clasificación automática que asigne categorías jerárquicas a eventos turísticos, utilizando técnicas de aprendizaje automático y PLN.
- Validar el modelo mediante un caso de uso en el sector turístico: Aplicar el modelo en un entorno simulado de una compañía aérea para evaluar su eficacia, escalabilidad y viabilidad económica, demostrando su utilidad práctica para empresas que gestionan o distribuyen información sobre eventos turísticos en múltiples regiones.

1.3.3 Cuestiones de investigación

A partir del objetivo general y los objetivos específicos de esta investigación, se formalizan las siguientes cuestiones de investigación que guían el desarrollo del trabajo:

- ¿Cómo impactan los eventos turísticos en el sector turístico desde la perspectiva del marketing?
- ¿Es valiosa la clasificación de eventos turísticos para la creación de catálogos estructurados y reutilizables?
- ¿Qué beneficios económicos, de tiempo y esfuerzo pueden esperarse de un proceso automático de clasificación de eventos basado en técnicas de procesamiento del lenguaje natural?

Estas preguntas orientan tanto el desarrollo metodológico como la validación práctica del modelo propuesto, y se retoman en el capítulo de conclusiones para valorar los resultados obtenidos.

1.4 Estructura de la memoria

La presente tesis doctoral se estructura en 6 capítulos, cada uno de los cuales responde a una fase concreta del proceso de investigación:

En el capítulo 1 se presenta el contexto general del turismo de eventos, la motivación del estudio, la justificación económica y de marketing, así como los objetivos generales y específicos de la investigación. Además, se introducen conceptos clave como taxonomía, tipología y categoría, fundamentales para comprender el enfoque metodológico.

En el capítulo 2 se lleva a cabo una revisión de la literatura. Se analiza el estado del arte en relación con la clasificación de eventos turísticos, las taxonomías existentes, las aplicaciones de procesamiento de lenguaje natural en turismo y las metodologías de análisis de datos. Este capítulo permite identificar vacíos en la literatura y justificar la necesidad de un enfoque automatizado y universal.

En el capítulo 3 se describe detalladamente el marco metodológico adoptado, basado en el modelo CRISP-DM. Se explican las decisiones técnicas relativas a la recolección y preparación de datos, la selección de modelos (BERT y regresión logística), y los criterios de evaluación utilizados para validar los resultados.

En el capítulo 4 se desarrolla el modelo propuesto. Se presenta la implementación del sistema de clasificación automática, incluyendo la arquitectura del modelo, el preprocesamiento de textos, el entrenamiento del clasificador. Se muestran ejemplos representativos de eventos y su categorización dentro de la taxonomía propuesta. En

este capítulo también se lleva a cabo la validación y se presenta un caso de uso: Se evalúa el rendimiento del modelo mediante métricas estándar y se presenta su aplicación en un escenario realista: una compañía aérea que desea enriquecer la experiencia de sus clientes mediante la recomendación de eventos en sus destinos. Este caso ilustra la escalabilidad, adaptabilidad y utilidad comercial del sistema propuesto.

El capítulo 5 resume los principales hallazgos del estudio, se reflexiona sobre sus limitaciones y se proponen posibles mejoras y extensiones para investigaciones futuras en clasificación automática, taxonomías turísticas y personalización de la experiencia del viajero.

Por último, en un capítulo de anexos se referencian las publicaciones fruto de esta tesis doctoral.

2. Estado del arte

Tras establecer en la introducción la creciente importancia de los eventos turísticos en la dinamización del sector y la problemática derivada de la falta de una clasificación estandarizada para los mismos, este capítulo se adentra en el estado del arte para contextualizar la presente investigación dentro del panorama científico actual. Para ello, se ofrece una revisión crítica del estado del arte, en la que se exploran las principales contribuciones teóricas y tecnológicas relacionadas con la organización, clasificación y aprovechamiento de los eventos turísticos.

En primer lugar, se analizará cómo la IA y las tecnologías asociadas han sido aplicadas en el sector turístico para mejorar la toma de decisiones, optimizar servicios y personalizar la experiencia del viajero. A continuación, y de forma especialmente relevante para este trabajo, el foco se desplazará hacia los esfuerzos previos por crear catálogos o sistemas de organización de eventos turísticos examinando críticamente los diversos sistemas taxonómicos y tipológicos propuestos hasta la fecha.

Este análisis tiene un doble propósito: por un lado, comprender los marcos conceptuales y metodológicos empleados hasta la fecha para estructurar la heterogénea oferta de eventos; y por otro, identificar las principales limitaciones y vacíos de conocimiento de estos enfoques, especialmente en lo que respecta a la automatización, escalabilidad, universalidad y aplicabilidad de procesos de clasificación en contextos multiculturales y multilingües.

En última instancia, esta revisión permitirá justificar la necesidad del modelo propuesto en esta tesis y resaltar su carácter original como respuesta a los vacíos detectados en la literatura.

2.1. Uso de eventos para la toma de decisiones en el sector turístico

Hay trabajos muy relevantes que demuestran el creciente interés en la toma de decisiones en el sector turístico. Samala et al. (2022) analiza el impacto de la IA y la robótica en el sector del turismo. En su trabajo destaca el papel de la IA y la robótica para mejorar servicios y experiencias en el turismo, integrando tecnologías como agentes conversacionales, realidad virtual y traductores de idiomas. Aunque también menciona que estas tecnologías no pueden reemplazar completamente el toque humano, que es complementario. Entre otros hallazgos destacan los servicios automatizados que permiten personalizar experiencias basadas en comportamientos

e intereses de los usuarios, disminuyendo la necesidad de agentes de viajes tradicionales.

La IA permite a los especialistas en marketing automatizar procesos, mejorar la personalización y ofrecer información relevante y actualizada al público. Una de las técnicas usadas para alcanzar estos objetivos es la segmentación y clasificación automática. Algunos ejemplos que encontramos en la literatura y en el contexto del marketing, la IA se usaría para segmentar y personalizar campañas basadas en datos de comportamiento, creación de experiencias de marketing inmersivas mediante realidad virtual y mejora de la interacción con clientes a través de agentes conversacionales y asistentes de voz. Se menciona también en esta revisión la clasificación de destinos según preferencias, tiempo y tráfico en tiempo real a partir de imágenes con la tecnología de Sistema de Posicionamiento Visual o VPS por sus siglas en inglés.

Doborjeh et al. (2022) analiza 146 artículos sobre aplicaciones de la IA en el mundo del turismo y la industria hotelera. Este trabajo se centra en identificar cómo se han implementado eficientemente los métodos de IA, desde el modelado de datos para pronósticos de demanda y destinos turísticos, hasta patrones de comportamiento y mejoras en servicios al cliente. Entre las aplicaciones más relevantes para este trabajo se encuentran la predicción de la demanda turística, el análisis de patrones de comportamiento del turista usando datos de ubicación y la automatización de servicios como agentes conversacionales y asistentes virtuales.

Desde el punto de vista de la clasificación automática, se mencionan técnicas de clasificación usadas para identificar patrones en el comportamiento de los turistas y puntos de interés en destinos turísticos. Doborjeh et al. (2022) también mencionan, al igual que Samala et al. (2022), la clasificación de datos visuales y textuales para predecir intenciones de viaje o identificar puntos de atracción clave.

Como podemos ver, las técnicas de aprendizaje supervisado y en particular, la clasificación son temas centrales en la aplicación de la IA al turismo. Bi & Liu (2022) exploran en su artículo el diseño de una plataforma inteligente basada en el internet de las cosas (IoT) y aprendizaje automático para servicios turísticos. Esta plataforma se centra en predecir el comportamiento de los turistas, mejorar la toma de decisiones en viajes y proporcionar experiencias personalizadas utilizando datos recolectados a través de sensores, aplicaciones móviles y otras fuentes. En este trabajo se explora la clasificación de destinos turísticos entre deseables y no deseables por el viajero a partir del análisis y categorización de preferencias de los turistas para mejorar la experiencia en tiempo real. Para llevar a cabo la clasificación se emplea una variante del algoritmo *K-Nearest Neighbors* (KNN) o K-vecinos próximos, un algoritmo usado para regresión y clasificación. Así se logra dar una probabilidad de que el viajero elija

un destino procesando los datos de su historial de búsqueda, reservas y comentarios. Desde el punto de vista del marketing, esta técnica facilita la producción de recomendaciones personalizadas basadas en patrones de comportamiento detectados en el historial del usuario. También ofrece un análisis predictivo para identificar preferencias y optimizar campañas de promoción. En este artículo sin embargo encontramos una limitación: A pesar de que el *abstract* menciona que este método puede ayudar a los viajeros a elegir si ver o no una atracción turística, no se menciona en todo el cuerpo del artículo ninguna información sobre qué datos tenemos de la atracción ni como se usarían. Todos los datos disponibles con los que se trabaja son intrínsecos al viajero (sus búsquedas, su posición, sus comentarios, etc.), pero no se menciona ninguna información que pueda tener el catálogo de atracciones que utilizan ni como se ha categorizado.

Como ya se ha mencionado en la introducción, el PLN es un campo de la IA que estudia formas de analizar, estudiar y producir texto en lenguaje natural humano. Es una familia de tecnologías exitosas para llevar a cabo, entre otras cosas, análisis de sentimiento, que se traduce en herramientas para segmentar reseñas de usuarios y estudiar la calidad del servicio que ofrecen empresas basadas en turismo como hoteles. Es útil también para realizar análisis de marca a partir de textos que el usuario deja en redes sociales e internet. El PLN es muy utilizado también para crear sistemas de recomendación, extrayendo patrones de cuerpos de texto e incluso recibiendo peticiones como texto libre (Álvarez-Carmona et al., 2022).

Las aplicaciones basadas en el PLN son muy potentes para el marketing ya que pueden procesar información del texto libre que los viajeros dejan en internet. Liu et al. (2021) explora cómo las estrategias de marketing en redes sociales de marcas de lujo impactan en la interacción de los clientes utilizando análisis de *big data* y PLN. En este artículo se analizan 900 *tweets* relacionados con marcas de lujo y se estudia la interacción de los usuarios y las marcas, la personalización de mensajes y la relación entre usuarios y marcas. Este análisis ayuda a crear estrategias basadas en entretenimiento y tendencias que impulsan el compromiso del cliente y contenidos adaptados al contexto emocional del cliente, fomentando la lealtad a la marca. El estudio no obstante no usa modelos de PLN semánticos y se limita a estudiar patrones de aparición de palabras o expresiones y otras técnicas de análisis estadístico sobre el texto.

El análisis de textos es el fuerte de las herramientas basadas en PLN y se puede utilizar para caracterizar opiniones de clientes sobre nuestra marca. García-Pablos et al. (2018) propone el uso de métodos no supervisados para analizar aspectos y polaridades en reseñas hoteleras. El artículo presenta un sistema casi no supervisado para realizar análisis de sentimientos basado en aspectos de la experiencia que

requiere mínima intervención humana. El sistema propuesto es capaz de aproximar clasificación de temas hacia categorías definidas por el usuario o diferenciar entre términos descriptivos (términos de aspecto) y términos de opinión. Así se pueden extraer automáticamente aspectos como la categoría de comida, servicio o ambiente entre las reseñas de un restaurante u hotel, lo que supone un avance en clasificación de polaridad y aspectos de productos y servicios y posibilita estrategias de personalización basadas en análisis de sentimientos. Aunque el artículo hace aportaciones notables, no utiliza modelos semánticos. Además, la clasificación que proponen no es estrictamente supervisada y, por tanto, difícil de utilizar con una taxonomía, es decir, un sistema cerrado de categorías. Aunque el sistema podría clasificar opiniones sobre eventos según categorías específicas como ubicación, calidad del servicio o ambiente de eventos turísticos, no se puede aplicar a la creación de catálogos de eventos ni a la clasificación de eventos turísticos.

2.2. Creación de catálogos de eventos a través de sistemas taxonómicos o tipológicos

Una vez establecida la importancia del evento turístico tanto para el turista como para el destino, que se beneficia directamente de la actividad turística, vamos a estudiar los beneficios de categorizar los eventos turísticos.

Aplicado a destinos, permite una mejor planificación y gestión de los eventos por parte de los destinos turísticos. También ayuda a desarrollar estrategias de marketing más efectivas al poder segmentar y dirigirse a diferentes tipos de turistas según el tipo de evento. Además, facilita la evaluación del impacto económico, social y cultural de los diferentes tipos de eventos en un destino (McKercher, 2016). Tener los eventos de un destino clasificados también permite crear un portafolio equilibrado de eventos que maximice los beneficios para los destinos a lo largo del año, combatiendo más eficientemente la estacionalidad de muchos destinos turísticos. Ayuda a identificar qué tipos de eventos tienen mayor potencial turístico y cuales requieren mayor inversión o desarrollo. Por último, también permite comparar y analizar diferentes eventos de forma sistemática, identificando mejores prácticas y oportunidades de mejora (Getz & Page, 2014).

La clasificación de eventos turísticos y la creación de un catálogo de eventos también es muy útil para el turista, el consumidor de los eventos. Permite a los turistas identificar y seleccionar eventos que se alinean con sus intereses específicos y ayudan a comprender la diversidad de eventos disponibles en un destino (McKercher, 2016). También facilita la comprensión de los diferentes tipos de motivaciones para asistir a eventos como pueden ser motivaciones intrínsecas, es decir, de interés personal y motivaciones extrínsecas como pueden ser trabajo, contactos, etc. (Getz, 2008). Al

final, el turista puede planificar mejor su viaje en torno a eventos de su interés ya que puede identificar mejores eventos que pueden enriquecer su experiencia de viaje.

En esta sección nos vamos a centrar exclusivamente en aquellos trabajos en los que la propuesta de un sistema taxonómico o tipológico para eventos turísticos o la creación de catálogos de eventos y su clasificación automática sean puntos centrales u objetivos principales de investigación.

Los eventos turísticos pueden clasificarse también según su nivel de adopción tecnológico. Neuhofer et al. (2014) desarrolla un marco conceptual y empírico para comprender cómo la tecnología está transformando las experiencias turísticas. Propone una matriz de tipología que clasifica los eventos turísticos según la intensidad de co-creación entre la industria y el consumidor y tecnología, y establece una jerarquía de experiencias tecnológicas. Tanto la intensidad de co-creación como la intensidad de adopción tecnológica pueden tomar tres valores: bajo, medio y alto. Combinando estos tres valores de estas dos dimensiones obtenemos 9 categorías para clasificar eventos mejorados con tecnología:

- Las 5 categorías con una baja intensidad de co-creación o una baja intensidad tecnológica se clasifican como turismo tradicional.
- Experiencia mejorada con tecnología y co-creación media
- Experiencia mejorada con tecnología y alto nivel de co-creación
- Experiencia tecnológica co-creada
- Experiencia basada en tecnología con alto nivel de co-creación

A partir de estas categorías, se deduce una jerarquía de adopción tecnológica y co-creación del evento de 4 niveles:

- Experiencia convencional
- Experiencia asistida por la tecnología
- Experiencia mejorada con tecnología
- Experiencia basada en tecnología

A pesar de que el artículo es novedoso ya que es el primero en considerar el nivel de co-creación y adopción tecnológica de las experiencias turísticas en una taxonomía, tiene algunas limitaciones. En primer lugar, no es general para todas las experiencias turísticas, sino solo para aquellas que tienen una base tecnológica y algún grado de co-creación sin analizar formato, escala, audiencia, duración o propósito del evento. En segundo lugar, aunque se sugiere analizar el grado de tecnologías de la información y comunicación interactivas presentes en la experiencia para determinar la intensidad tecnológica del evento, no se propone un método o un criterio claro para clasificar eventos en la taxonomía que proponen.

El de Getz (2008) es un artículo fundacional en este campo que ha tenido una importante repercusión en el sector. En él se propone un marco para entender y crear valor sobre los eventos turísticos. Los eventos turísticos planificados son un fenómeno que ocurre en un tiempo y un espacio determinados y cubren múltiples propósitos. Bajo esta definición Getz propone una tipología de categorías principales basados en el formato del evento. Algunas son celebraciones públicas mientras que otras se centran en la competición, diversión, entretenimiento o socialización. Otros eventos pueden necesitar infraestructura apropiada para su celebración y otros son políticos o relacionados con el mundo corporativo. Así, la tipología propuesta tiene ocho tipos:

- Celebraciones culturales: Festivales, carnavales, conmemoraciones, eventos religiosos, etc.
- Políticas y estatales: Celebraciones reales, eventos políticos, visitas de personalidades, etc.
- Arte y entretenimiento: Conciertos, entregas de premios, etc.
- Negocio y comercio: Encuentros, convenciones, ferias, exposiciones, mercados, etc.
- Educativos y científicos: Conferencias, seminarios, etc.
- Competición deportiva: Tanto amateur como profesionales.
- Recreacional: Deportes o juegos por placer
- Eventos privados: Bodas, fiestas, encuentros sociales, etc.

En este artículo también se propone una clasificación basada en las estrategias de marketing y el desarrollo necesario para ofrecer líneas de productos o servicios para evaluar el valor que aportan. Esta aproximación tiene cuatro categorías:

- Megaeventos: Alta demanda turística y alto valor
- Eventos distintivos periódicos: Alta demanda turística y alto valor
- Eventos regionales: Demanda turística media, ya sean periódicos o puntuales
- Eventos locales: Con baja demanda turística, ya sean periódicos o puntuales e independientemente de su valor.

Así, este trabajo propone un sistema tipológico para los eventos turísticos, pero no propone un sistema taxonómico, es decir, empírico y basado en un conjunto de eventos y por tanto no está ligado a una metodología operativa. En este artículo no se aborda cómo llevar a la práctica su modelo de clasificación mediante herramientas automatizadas o digitales. Su enfoque se centra en la teoría, con una discusión extensa sobre las interrelaciones entre turismo y eventos, pero carece de ejemplos aplicables a la gestión del evento y no ofrece soluciones prácticas para la clasificación automatizada o integración de datos.

Oklobdzija (2015) utiliza la tipología descrita por Getz y enfatiza el valor estratégico de los eventos para el desarrollo turístico, usando enfoques como el modelo de portafolio de Getz para evaluar la contribución de eventos a los destinos turísticos y aportar un marco conceptual amplio y categorizaciones útiles para entender el impacto de los eventos en el turismo. El artículo de Oklobdžija ofrece una perspectiva estratégica y teórica valiosa, pero no propone soluciones tecnológicas u operativas aplicadas que podrían potenciar los modelos conceptuales que propone.

Como hemos mencionado ya, el turismo es una actividad altamente fragmentada y disjunta, difícil de compactar. Incluso en la investigación, diferentes autores usan los mismos términos para referirse a diferentes ideas y esto frena la disciplina que tiene carencias ontológicas, epistemológicas y metodológicas. Y una de las áreas en que se hace más patente esta desconexión es en la clasificación de productos turísticos. McKercher (2016) propone una taxonomía de evento turístico basada en una jerarquía taxonómica de producto en marketing. A través del análisis de decenas de trabajos científicos y utilizando un método fenético[1], McKercher propone una taxonomía jerárquica de productos turísticos basada en cinco familias de necesidades: Placer, Búsqueda Personal, Comprensión del Esfuerzo Humano, Naturaleza y Negocios. Utilizando su método fenético para avanzar de lo general a lo específico, de estas cinco familias de necesidades se desprenden 27 familias de productos y 90 clases de productos. Esta taxonomía busca resolver problemas históricos de fragmentación y falta de uniformidad en el campo de los estudios de turismo. Para entender mejor la taxonomía propuesta por McKercher vamos a desglosar la familia de necesidades "Placer". Dentro de esta familia de necesidades podemos encontrar varias familias de producto con sus clases de productos:

- Comida y bebida
    - Bebida
    - Comida
    - Híbrido
    - Aprendizaje
- Ocio
    - Compras
    - Visitas o exploración
    - Segundas viviendas
    - Fotografía
- Indulgencia
    - Sexo

[1] La misión de los métodos fenéticos es la creación de taxonomías atendiendo únicamente a la similitud de las muestras ignorando su origen o procedencia común.

    - Clubs, bares y discotecas
    - Turismo narcótico
    - Turismo de fiesta
    - Cleptoturismo (robar un adoquín de la gran muralla, por ejemplo)
- Eventos personales
    - Familia
    - Amigos
- Atracciones construidas
    - Apuestas
    - Construidas con un propósito turístico
- Deportes
    - Activo
    - Pasivo
- Recreación
    - Activa
    - Pasiva
    - Instalaciones recreativas

Y a su vez, cada una de estas clases de productos se desgranan en líneas de productos, tipos de productos y en última instancia, productos. Para la familia de productos "Bebida", podemos tener líneas de productos como "centrados en vino" o "destilerías". Estas a su vez tendrán tipos de productos como "visitas a bodegas", "tours vinícolas", "catas de vino", etc.

Este artículo proporciona una base para futuros debates académicos sobre cómo estructurar los productos turísticos de manera integral, introduce una taxonomía detallada que podría ser utilizada para mejorar la planificación estratégica de eventos turísticos e identifica limitaciones actuales de las clasificaciones previas, destacando la fragmentación del campo y la falta de un marco común. La propuesta es, además, taxonómica, es decir, empírica, lo que eleva aún más este trabajo tan relevante en el campo de la categorización de eventos turísticos. Sin embargo, a pesar de su enorme contribución, este artículo nos deja una brecha en la aplicabilidad práctica de estas taxonomías al no indicar ningún método o criterio específico para clasificar eventos turísticos. Además, la taxonomía propuesta no es exhaustiva; se reconoce que los productos categorizados son representativos, pero no cubren la totalidad de ofertas turísticas posibles. Además, al no ofrecer un criterio de clasificación, se deja en el aire la clasificación de productos que podrían pertenecer en varias categorías.

Hay también esfuerzos a la hora de clasificar eventos turísticos utilizando IA. Cepeda-Pacheco & Domingo (2022) propone un sistema de recomendación basado en redes de aprendizaje profundo para mejorar la experiencia del turista. El sistema de

recomendación propuesto utiliza datos del perfil del viajero recogidos en una encuesta con mil turistas y dan información sobre la edad, actividades turísticas a realizar o razón del viaje entre otras. El sistema también se alimenta de la posición del turista geolocalizado, temperatura en el destino, predicción del clima, etc. Con estas entradas su sistema es capaz de recomendar hasta cuarenta diferentes atracciones. Este artículo es un excelente ejemplo de cómo la IA y los datos de IoT pueden transformar la experiencia turística en tiempo real e introduce un modelo novedoso de recomendación en tiempo real para ciudades inteligentes, mejorando la experiencia turística personalizada. No obstante, tiene algunas limitaciones. En primer lugar, no ofrece un sistema taxonómico de eventos y por tanto no se pueden clasificar fuera del contexto del usuario ni se puede generar un catálogo universal de eventos. En segundo lugar, existe una dependencia de la tecnología que hace el problema difícil de escalar y estandarizar: todos los turistas deben estar geolocalizados, compartiendo datos y aportando información sobre su viaje. Por último, el artículo está enfocado en atracciones turísticas y actividades dentro del contexto de ciudades inteligentes, no aborda eventos específicos como festivales o megaeventos.

En la misma línea de recomendar actividades turísticas utilizando aprendizaje profundo y dispositivos propios del IoT se encuentra Gupta et al. (2024) que propone un sistema de recomendación que toma información del usuario tal como el perfil de sus acompañantes si los hay, el propósito de su viaje o características personales. Junto con información obtenida de dispositivos IoT e información contextual como el clima, el sistema es capaz de clasificar eventos turísticos en eventos que le pueden interesar al usuario y eventos que no le interesan. Aunque se ofrece una clasificación de eventos, esta depende del usuario y por tanto no es universal. Además, necesita de dispositivos IoT en los turistas, cuestionarios e información contextual por lo que no es fácilmente escalable.

Los eventos virtuales también han sido objeto de estudios tipológicos. Yung et al. (2022) desarrolla una tipología de eventos virtuales en turismo y sector hotelero. Para ello aúna 3 dimensiones del evento: virtualidad del entorno, localización y presencia social para proponer un cubo que clasifique todos los tipos de eventos virtuales. Así, la dimensión localización puede tomar dos valores: localización física o virtual. La dimensión presencia social puede tomar dos valores: alta o baja. La virtualidad del entorno puede tomar dos valores: Real o virtual. La tipología incluye ocho vértices que representan combinaciones únicas de estas dimensiones, permitiendo clasificar eventos desde videoconferencias básicas hasta experiencias completamente inmersivas en realidad virtual. Este artículo tiene un alto valor al introducir un marco conceptual para abordar la ambigüedad terminológica en eventos virtuales y proponer una tipología de eventos virtuales. No obstante, la implementación del cubo

no se aborda, dejando abierta la pregunta de cómo operacionalizar esta tipología en sistemas automatizados.

Algunos autores se centran en temáticas particulares para eventos turísticos. Así Iliev (2020) hace un estudio del turismo religioso. Entre otras ideas menciona una tipología de recursos religiosos tales como sitios arqueológicos, sitios funerarios, templos, montañas sagradas, etc. hasta completar 11 categorías. Diana et al. (2020) usa una tipología similar para turismo religioso y peregrinación destacando la diferencia entre estos: El turismo espiritual está enfocado en la experiencia espiritual, la autenticidad del lugar religioso y el progreso personal mientras que el turismo de peregrinación está enfocado en el viaje ritual a sitios considerados sagrados, con un componente religioso explícito. Aunque ambos trabajos combinan una revisión bibliográfica con un análisis conceptual para identificar subtipos y clasificaciones dentro del turismo religioso no aborda cómo operacionalizar las tipologías propuestas ni cómo lidiar con las intersecciones entre turismo espiritual y peregrinación que a menudo se solapan, lo que dificulta establecer una distinción clara entre ambos tipos.

Los eventos turísticos deportivos también han sido objeto de estudio y arrojan luz sobre cómo clasificar eventos turísticos, aunque sea circunscritos al ámbito del deporte. Bjeljac et al. (2017) propone un sistema basado en diferentes criterios para categorizar eventos deportivos en Serbia. Estos criterios comprenden la significancia social nacional, tradición, autonomía financiera o accesibilidad entre otros. A partir de puntuaciones obtenidas para los 74 deportes analizados obtenemos 4 grupos. A pesar de proporcionar un marco práctico para clasificar eventos deportivos en Serbia, destacando su potencial turístico e introducir criterios que combinan aspectos económicos, sociales, mediáticos y ambientales, este artículo tiene limitaciones a la hora de implementar su categorización. Su aplicabilidad global está limitada por la falta de integración tecnológica y su enfoque local y no aborda cómo integrar herramientas tecnológicas para clasificar o analizar grandes volúmenes de datos.

Siguiendo los esfuerzos por proponer taxonomías o tipologías de eventos turísticos, Kahn (2015) propone una tipología basada en tres dimensiones que pueden tomar dos valores cada una:

- Con ánimo de lucro / sin ánimo de lucro
- Centrados en un solo deporte / Centrados en varios deportes
- Recurrente en el mismo sitio / puntual

Las combinaciones de estas dimensiones dan lugar a una tipología con 7 tipos (teóricamente 8, pero este estudio no contempla eventos centrados en varios deportes, con ánimo de lucro y recurrentes ya que no se identifica ningún evento así). Este

artículo se basa en estudios de caso y análisis histórico para identificar patrones comunes en la organización de eventos deportivos y examina la influencia de factores como la infraestructura, la recurrencia y el financiamiento en el éxito y la sostenibilidad de los eventos. También ofrece una tipología práctica que puede informar decisiones estratégicas en la planificación y gestión de eventos deportivos y aborda la importancia de las dimensiones financieras, organizativas y operativas en la sostenibilidad de los eventos. No obstante, se limita a un enfoque descriptivo, sin explorar herramientas tecnológicas modernas para su aplicación práctica y no propone un método para clasificar eventos en la tipología propuesta.

Se puede consultar un resumen comparativo de los principales trabajos sobre clasificación de eventos turísticos analizados en la Tabla 1.

Ref.	Fundamentos	Aplicación	Limitaciones
(Getz, 2008)	Creación de tipología con ocho tipos.	Eventos turísticos	Carece de metodologías prácticas que aborden las necesidades actuales de clasificación automatizada. Su enfoque en la tipología y el modelo de portafolio es valioso, pero está limitado al contexto conceptual.
(Neuhofer et al., 2014)	Creación de una tipología con 9 categorías y una jerarquía de 4 niveles para eventos con cierto nivel de co-creación y tecnológico.	Eventos turísticos asistidos por tecnología	No propone una forma de clasificar eventos en su propia tipología o jerarquía.
(Kahn, 2015)	tipología de 7 tipos basada en factores financieros, multidisciplinarios y de frecuencia de eventos.	Eventos turísticos	No propone un método para clasificar eventos en la tipología propuesta
(Oklobdzija, 2015)	Aporta un marco conceptual amplio y categorizaciones útiles para entender el impacto de los eventos	Eventos turísticos con foco en estrategia	Carece de metodologías prácticas que aborden las necesidades actuales de clasificación automatizada.

	en el turismo centrándose en el plano de la estrategia y planificación de destinos.	y planificación	
(McKercher, 2016)	Introduce una taxonomía jerárquica detallada con 5 familias de necesidades de las que se desprenden 27 familias de productos y 90 clases de productos.	Eventos turísticos	No indicar ningún método o criterio específico para clasificar eventos turísticos
(Bjeljac et al., 2017)	4 categorías para eventos deportivos en Serbia	Eventos turísticos deportivos	No propone una implementación y su enfoque es local.
(Iliev, 2020)	tipología de 11 tipos centrada en el turismo religioso.	Eventos y sitios religiosos	No aborda la implementación de su método ni como clasificar eventos.
(Diana et al., 2020)	tipología de 11 tipos centrada en el turismo religioso.	Eventos y sitios religiosos	No aborda la implementación de su método ni como clasificar eventos.
(Bi & Liu, 2022)	Clasificación de eventos turísticos entre deseables por el turista o no deseables (sistema de recomendación).	Eventos turísticos en Smart cities	No ofrece un sistema taxonómico de eventos, dependencia tecnológica, depende de datos del usuario
(Yung et al., 2022)	Creación de una tipología de 8 tipos basada en la combinación de 3 variables binarias: virtualidad del entorno, localización y presencia social.	Eventos turísticos virtuales	No aborda la implementación de su método ni como clasificar eventos.
(Cepeda-Pacheco & Domingo, 2022)	Sistema de recomendación basado en IA para clasificar y recomendar atracciones turísticas y actividades personalizadas.	Atracciones y eventos turísticos en Smart cities	No ofrece un sistema taxonómico de eventos, dependencia tecnológica, dependiente de datos del usuario

(Gupta et al., 2024)	Sistema de recomendación basado en IA para clasificar y recomendar atracciones turísticas y actividades personalizadas.	Atracciones y eventos turísticos en Smart cities	No ofrece un sistema taxonómico de eventos, dependencia tecnológica, depende de datos del usuario

Tabla 1: Estudios sobre clasificación de eventos turísticos

3. Fundamentos metodológicos

Este capítulo presenta los fundamentos metodológicos sobre los que se construye la presente investigación. Se expone el enfoque adoptado para abordar el problema de clasificación automática de eventos turísticos, con énfasis en la aplicación de técnicas de IA y PLN dentro de un marco estructurado. En particular, se detalla el uso de la metodología CRISP-DM como guía para el desarrollo del modelo, abarcando desde la comprensión del negocio y los datos, hasta la preparación, modelado, evaluación y despliegue. Esta metodología proporciona una estructura sólida y flexible que permite garantizar la trazabilidad y la reproducibilidad del proceso.

A continuación, se profundizará en las tecnologías clave empleadas en el núcleo de nuestra solución: las técnicas de PLN, con especial énfasis en el modelo BERT, fundamental para extraer la representación semántica del texto libre de los eventos. Finalmente, se abordarán los conceptos de clasificación multiclase, incluyendo el modelo de regresión logística seleccionado para asignar las categorías taxonómicas y los métodos y métricas específicas para evaluar el desempeño y la fiabilidad de este tipo de clasificadores en tareas complejas. Estos fundamentos serán utilizados en la implementación práctica y la validación del modelo propuesto en el capítulo siguiente.

3.1. Metodología CRISP-DM

El modelo CRISP-DM es una metodología ampliamente aceptada en el ámbito de la ciencia de datos. Fue desarrollado en 1999 por un consorcio liderado por DaimlerChrysler, SPSS y NCR, con el objetivo de proporcionar un marco estándar para la realización de proyectos de ciencia de datos en diferentes industrias (Schröer et al., 2021; Shafique & Qaiser, 2014).

CRISP-DM aborda la necesidad de sistematizar y estandarizar los procesos de ciencia de datos, eliminando la dependencia de herramientas específicas y promoviendo la reutilización de prácticas exitosas. Este modelo se estructura en seis fases bien definidas e iterativas, permitiendo una gran flexibilidad para adaptarse a los contextos específicos de los proyectos (Azevedo & Santos, 2008; Schröer et al., 2021).

A diferencia de otros modelos como SEMMA (*Sample, Explore, Modify, Model, Assess*) y KDD (*Knowledge Discovery in Databases*), CRISP-DM ofrece un enfoque más integral al incorporar explícitamente las fases de comprensión del negocio y despliegue, elementos cruciales para la alineación con los objetivos empresariales y la sostenibilidad de los proyectos (Azevedo & Santos, 2008).

Como se ha comentado, CRISP-DM se caracteriza por tener seis fases:

1. Comprensión del negocio
2. Comprensión de los datos
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Despliegue

Estas fases no siguen un orden estricto y es común iterar por ellas a medida que se avanza en el proyecto. Por ejemplo, el resultado de la fase Evaluación puede llevarnos a revisitar la fase de Comprensión del Negocio si no hemos alcanzado los objetivos iniciales o si hemos descubierto un aspecto en que el algoritmo funciona particularmente bien. Del mismo modo, tras la fase de Modelado podríamos querer volver a la fase de Preparación de los Datos si creemos que podemos procesar los datos de manera diferente para mejorar el modelo.

Vamos ahora a estudiar las diferentes fases del modelo CRISP-DM.

3.1.1. Comprensión del negocio

En esta fase nos aseguramos de que el proyecto de ciencia de datos está alineado con los objetivos estratégicos de la organización. Comprende varias actividades como la definición de los objetivos de negocio, la traducción de objetivos empresariales al problema técnico de ciencia de datos y la planificación preliminar del proyecto.

Los objetivos de negocio han de estar claramente definidos con un objetivo principal en que no caben ambigüedades y varios objetivos secundarios que nos aporten contexto y beneficios adicionales al objetivo principal.

De estos objetivos normalmente se desprenden una serie de desafíos que debemos tener en cuenta a la hora de llevar a cabo el proyecto. Normalmente estos desafíos representan dificultades operativas a la hora de implementar soluciones, obtener datos, etc., o directamente representan gaps en el conocimiento que tenemos que superar.

Una vez tenemos claros los objetivos y desafíos debemos traducir nuestro problema de negocio a objetivos específicos y viables para la ciencia de datos. Por ejemplo, podríamos traducir el objetivo de “mejorar la retención de clientes” en un modelo de predicción del abandono.

Por último, se genera un plan detallado de la elaboración del proyecto que contempla recursos a utilizar, marcos temporales, limitaciones esperadas, etc.

3.1.2. Comprensión de los datos

Esta fase del modelo CRISP-DM se centra en identificar nuestras necesidades en términos de datos, recolectarlos, explorarlos y evaluar su calidad. El objetivo de esta fase es acabar teniendo todos los datos necesarios y con la calidad suficiente como para acometer nuestro objetivo de negocio en última instancia. Esta comprensión del dato permite identificar problemas potenciales y formular hipótesis sobre patrones subyacentes que pueden responder a las preguntas planteadas en la fase de Comprensión del Negocio (Wirth & Hipp, 2000).

El primer paso de esta fase consiste en identificar y recopilar los datos iniciales desde diversas fuentes. Estas pueden incluir bases de datos internas, archivos planos, APIs (Application Programming Interface) externas o incluso datos no estructurados como texto e imágenes. Una API es un conjunto de definiciones y protocolos que permite la comunicación entre aplicaciones de software. En el contexto de esta investigación, las APIs permiten acceder automáticamente a fuentes de datos externas (por ejemplo, catálogos de eventos en línea o plataformas turísticas), facilitando la extracción estructurada de información actualizada y evitando la recolección manual. Es esencial recopilar y generar también metadatos que, si bien no son necesarios estrictamente para resolver el problema de negocio, nos ayudan a lidiar con los datos en fases posteriores. Entre estos metadatos se encuentra la procedencia del dato, el formato y estructura, el momento en que fueron recopilados, etc. La recopilación de datos es a veces un proyecto en sí, sobre todo cuando los datos están dispersos, en múltiples fuentes, con múltiples formatos, idiomas, etc.

La exploración de datos implica analizar el contenido para obtener una comprensión básica y detectar problemas evidentes. Las técnicas empleadas en esta etapa incluyen análisis estadísticos descriptivos, visualización de datos e identificación de relaciones entre las variables entre otras técnicas de análisis estadístico. Al hacer un análisis estadístico descriptivo Se generan métricas como medias, medianas, desviaciones estándar, entre otras, para resumir las características principales de las variables. Herramientas como gráficos de dispersión, histogramas y diagramas de caja son útiles para identificar patrones iniciales y detectar valores atípicos. Métodos como la correlación y el análisis de covarianza permiten descubrir relaciones entre diferentes atributos. La exploración de datos nos arroja información muy valiosa sobre la riqueza y limitaciones de nuestros datos.

El último paso de la fase de Comprensión de los datos del modelo CRISP-DM es la evaluación de la calidad de los datos. El objetivo de esta actividad es identificar problemas que podrían comprometer la integridad del modelo y el análisis posterior (Schröer et al., 2021; Wirth & Hipp, 2000). Es en este paso donde se identifican huecos en nuestros datos en forma de valores faltantes o nulos que pueden deberse a errores

de captura o inconsistencias en los sistemas de origen. También se debe prestar especial atención a errores de formato como inconsistencias en las fechas, valores categóricos no contemplados o directamente, valores bien extraídos pero que no tienen sentido en nuestro contexto. Si bien no son errores propiamente dichos, es necesario identificar valores atípicos que se desvíen significativamente del rango esperado para detectar posibles errores o casos extremos relevantes. Asimismo, se debe evaluar si el conjunto de datos está equilibrado en cuanto a las clases objetivo o si presenta algún sesgo que podría distorsionar los resultados.

Wirth & Hipp (2000) y Shafique & Qaiser (2014) sugieren metodologías y herramientas aplicadas a esta fase del proyecto. Sugiere así el uso de herramientas como Python y algunas de sus librerías como matplotlib, seaborn o pandas para llevar a cabo el análisis exploratorio, limpieza de datos inicial y pruebas de consistencia.

La fase de Comprensión del dato está estrechamente conectada con la fase de Comprensión del negocio. La evaluación de la calidad de los datos y la comprensión de sus patrones iniciales pueden requerir iteraciones hacia atrás para redefinir objetivos o ajustar los alcances del proyecto (Wirth & Hipp, 2000) y (Shafique & Qaiser, 2014).

3.1.3. Preparación de los datos

La fase de Preparación de los Datos en el modelo CRISP-DM es esencial para transformar los datos a un formato adecuado para el modelado, asegurando que los problemas identificados en la etapa de comprensión de los datos se aborden eficazmente. Esta fase, iterativa y flexible, abarca diversas actividades que incluyen selección de datos, limpieza, transformación e integración de datos para optimizar su uso en las etapas posteriores del proceso de ciencia de datos. El propósito principal de esta fase es generar un conjunto de datos limpio, completo y estructurado, listo para alimentar los algoritmos de modelado. Esto incluye seleccionar variables relevantes, tratar valores faltantes y realizar transformaciones necesarias que faciliten el análisis.

La selección de datos implica elegir los registros, atributos y casos que sean relevantes para el problema de negocio y el objetivo de ciencia de datos. En esta fase se definen reglas claras para seleccionar datos útiles, como incluir solo registros completos o excluir variables con alta correlación para evitar problemas de multicolinealidad.

En muchas ocasiones los datos son valiosos, pero vienen expresados de una forma difícil de consumir para el modelo o representando una complejidad innecesaria. La limpieza es un paso muy común que lidia con datos inconsistentes, duplicados, parciales o mal formados. Ante estos problemas se llevan a cabo acciones como eliminar registros si los datos faltantes no impactan excesivamente la cantidad de

datos con los que contamos. También es común imputar valores faltantes con algún método estadístico como análisis de medias o medianas para reemplazar valores perdidos. En este paso se corrigen inconsistencias y se unifican formatos de fechas, se corrigen errores tipográficos, se unifican variables categóricas y se ajustan unidades inconsistentes. Por último, y a veces, más importante, también se eliminan duplicados para garantizar que cada registro sea único, especialmente crítico en bases de datos transaccionales.

En algunos de estos casos, se pueden crear variables derivadas transformando o enriqueciendo las variables con las que ya contamos, una técnica conocida como ingeniería de características. Algunas de las prácticas más usuales de la construcción de datos derivados son la creación de atributos a partir de otros preexistentes o el enriquecimiento de datos derivados. Un ejemplo de creación de datos derivados es calcular la edad a partir de la fecha de nacimiento o calcular el índice de masa corporal a partir de la altura y el peso. Un ejemplo de enriquecimiento es tomar el código postal de un conjunto de datos y el sueldo medio por código postal de otro para generar una variable que no existe en el conjunto de datos original.

La transformación de datos, otro de los pasos relevantes de esta fase, incluye adaptar los datos a formatos requeridos por los algoritmos de modelado y mejorar su distribución. Por ejemplo, a través de la normalización y el escalado. También es usual convertir variables categóricas a formatos numéricos con técnicas como la codificación binaria, una técnica que genera nuevas características binarias a partir de una característica categórica. Las técnicas de reducción de la dimensionalidad también son muy usadas tanto para seleccionar atributos relevantes para el modelo (como técnicas de clusterización) como para proyectar los datos a un espacio de dimensionalidad inferior para simplificar el modelo.

Por último, hablamos aquí de la integración de datos. La integración implica combinar datos de diferentes fuentes para crear un conjunto único y completo. Se pueden concatenar registros provenientes de múltiples bases de datos o sistemas para unificar datos. También para manejar discrepancias en registros duplicados o datos provenientes de distintas fuentes que se refieran a un mismo evento.

La preparación de los datos es una continuación lógica de la fase de la fase Comprensión del Dato, ya que utiliza los hallazgos sobre la calidad, la estructura y los problemas de los datos para implementar soluciones concretas. Esta fase también influye directamente en el éxito de la fase de modelado, ya que la calidad de los datos es uno de los factores más críticos para el rendimiento de los algoritmos de aprendizaje automático.

3.1.4. Modelado

La fase de modelado en el modelo CRISP-DM se centra en aplicar técnicas analíticas y de aprendizaje automático a los datos preparados para construir modelos que respondan a los objetivos definidos en las fases anteriores. En esta etapa, los datos son transformados en conocimiento accionable mediante la selección, entrenamiento y evaluación de modelos de aprendizaje automático.

La primera tarea dentro de la fase de modelado es la elección del modelo que se va a utilizar. La elección del modelo es un paso crítico que debe alinearse con los objetivos y la naturaleza del problema. Normalmente se usan modelos de aprendizaje automático como:

- Modelos supervisados: Como clasificación (árboles de decisión, regresión logística, redes neuronales) o regresión (lineal, KNN).
- Modelos no supervisados: Como *Clustering* o reducción de dimensionalidad.

Aunque los datos ya han sido preparados, en esta fase puede ser necesario realizar ajustes adicionales para satisfacer los requisitos específicos de los algoritmos seleccionados. Estas transformaciones incluyen por ejemplo la división en conjuntos de datos de entrenamiento y testeo o estrategias de muestreo, ya sean para sobremuestrear con algoritmos como SMOTE o submuestrear en problemas de desequilibrio de clases.

La generación de modelos implica entrenar el modelo con los datos de entrenamiento y ajustarlo para optimizar su rendimiento. Este suele necesitar cierta algoritmia para iterar por los datos u optimizar hiperparámetros del modelo seleccionado.

Una vez los modelos han sido entrenados son validados utilizando el conjunto de validación, y se analizan métricas clave. Estas métricas nos dan una idea de lo bien que el modelo ha encontrado patrones en los datos. En tareas de clasificación es muy común utilizar métricas como la precisión, la sensibilidad o el *F1-score*.

Por último, en caso de que sea posible, tenemos que añadir una capa de explicabilidad. Esto incluye identificar las variables más influyentes y validar si los patrones detectados tienen sentido desde una perspectiva de negocio.

El éxito de esta fase depende de la calidad de los datos obtenidos en Preparación de los Datos y de la comprensión del problema. Los resultados del modelado se retroalimentan en la fase de evaluación para validar su adecuación a los objetivos empresariales. Un modelo que no cumple con estos objetivos requerirá ajustes en el modelado o incluso iteraciones hacia fases anteriores.

3.1.5. Evaluación

En la fase de evaluación se validan los resultados obtenidos en la fase de modelado y se estudia su adecuación a los objetivos definidos en la fase de comprensión del negocio. Esta etapa implica un análisis exhaustivo del modelo, sus resultados y su aplicabilidad en el contexto empresarial antes de proceder a su despliegue. El propósito principal de la fase de Evaluación es determinar si el modelo construido es adecuado para resolver el problema planteado en los objetivos de negocio descritos en la sección 3.1.1, asegurando que los resultados sean relevantes, precisos y accionables. Además, esta fase busca identificar posibles limitaciones del modelo y definir mejoras necesarias antes de su implementación en un entorno real. Por último, nos ofrece mucha información para caracterizar el modelo y saber que esperar de él en términos de resultados y rendimiento.

En primera instancia, tenemos que evaluar resultados del modelo para comprobar que se alinean con los objetivos de negocio. Se analiza si los patrones identificados y las predicciones generadas tienen sentido desde el punto de vista del dominio del problema y se realiza una estimación del impacto esperado de implementar el modelo, como ahorro de costos o mejora en la eficiencia operativa.

La interpretación de los resultados es fundamental para garantizar que las decisiones basadas en el modelo sean confiables. En esta fase se identifican las variables más influyentes en el modelo, lo que puede proporcionar información adicional para el negocio. Se verifica que los supuestos implícitos en el modelo sean consistentes con el conocimiento del dominio y se generan gráficos como matrices de confusión, curvas ROC (*Receiver Operating Characteristic* o Característica Operativa del Receptor) o diagramas de importancia de características ayudan a interpretar los resultados del modelo de forma más intuitiva.

De especial relevancia es la evaluación detallada del desempeño técnico del modelo mediante métricas cuantitativas para evaluar cómo de bien el modelo desempeña su función. Destaca aquí la técnica de validación cruzada que asegura que el modelo es robusto y no está sobreajustado a un subconjunto de datos específico.

En la fase de evaluación podemos comprobar que cada paso tomado durante la fase de modelado sea consistente con las mejores prácticas y que no se hayan omitido aspectos clave. Este paso es instrumental y nos permite evaluar si se requieren ajustes en los datos, el modelo o los algoritmos utilizados. Recordemos que los pasos de CRISP-DM no son secuenciales y podemos volver a una fase anterior en un proceso iterativo y recurrente hasta alcanzar la solución adecuada. Si el modelo no cumple con los objetivos, puede ser necesario regresar a fases como preparación de los datos o modelado para realizar ajustes.

En última instancia, es en la fase de evaluación que se valida la idoneidad del modelo para su propósito final. Una evaluación bien ejecutada asegura que los resultados son precisos, relevantes y alineados con los objetivos empresariales, estableciendo una base sólida para el despliegue.

3.1.6. Despliegue

El objetivo de la fase Despliegue es garantizar que el modelo y los resultados obtenidos tengan un impacto real y tangible en la organización, contribuyendo a la solución del problema identificado en la fase de comprensión del negocio. Esto puede implicar generar informes, implementar soluciones tecnológicas o automatizar procesos analíticos. Es en esta etapa donde los resultados obtenidos y validados en fases anteriores se implementan en un entorno real, permitiendo que el conocimiento descubierto sea aprovechado para la toma de decisiones o la automatización de procesos. La fase de despliegue no se limita a la creación del modelo; también incluye su integración en los sistemas empresariales, su monitoreo y mantenimiento continuo.

Antes de implementar el modelo, es necesario desarrollar un plan detallado que contemple los recursos técnicos, humanos y organizativos necesarios para el despliegue, que puede incluir:

- Definición de objetivos de despliegue: Determinar qué acciones se tomarán con los resultados del modelo. Por ejemplo, enviar alertas automáticas, generar informes predictivos, integrar recomendaciones en un sistema CRM (Customer Relationship Management).
- Evaluación de infraestructura: Verificar si la organización cuenta con los sistemas y hardware adecuados para la implementación del modelo.
- Identificación de partes interesadas: Coordinar con los departamentos implicados, como ingeniería, operaciones y negocio.
- Plan de riesgos: Identificar posibles problemas durante el despliegue, como caídas de rendimiento o resistencia al cambio, y definir estrategias de mitigación.

No todos los despliegues implican implementaciones tecnológicas complejas. En algunos casos, los resultados se entregan en forma de informes analíticos que faciliten la toma de decisiones. Estos informes pueden ser estáticos, resultados consolidados presentados en tablas, gráficos y resúmenes ejecutivos, o pueden ser interactivos con ayuda de herramientas como *Tableau* o *Power BI* que permiten a los usuarios explorar los resultados y generar análisis personalizados.

Es en esta fase de despliegue donde el modelo se puede integrar en sistemas productivos que implica llevar el modelo a un entorno operativo donde pueda generar resultados en tiempo real o de forma automatizada.

Por último, el despliegue no termina con la implementación; es fundamental asegurar que el modelo funcione correctamente y continúe proporcionando resultados precisos con el tiempo. Para ello es necesario monitorizar el modelo y mantenerlo en el tiempo. Por ejemplo, el rendimiento del modelo puede degradarse si la distribución de datos cambia con el tiempo. Además de la monitorización, hay que planificar cómo se mantendrá el modelo con reentrenamientos periódicos, ajustes de parámetros u otras actualizaciones. Y toda esta monitorización y mantenimiento del modelo debe quedar documentada y reportada, registrando las actualizaciones y mejoras realizadas al modelo.

3.2. Procesamiento del lenguaje natural

3.2.1. BERT

BERT es un modelo de lenguaje de aprendizaje profundo que supuso un hito y representaba el estado del arte de los modelos de lenguaje cuando fue presentado en 2018 por Devlin et al. (2018). BERT es un modelo que brilla en tareas clásicas de PLN y flujos de trabajo de clasificación textual (Tenney et al., 2019). Empíricamente, BERT supera a los enfoques tradicionales de procesamiento de lenguaje natural (NLP) para la clasificación de texto en diferentes conjuntos de datos (González-Carvajal & Garrido-Merchán, 2020). Además, BERT también supera a algoritmos de clasificación (González-Carvajal & Garrido-Merchán, 2020) como los basados en *bag-of-words* o en tareas como la clasificación de reseñas de compras de Yelp (Bilal & Almazroi, 2023).

Como su nombre indica, BERT hace referencia a representaciones del codificador bidireccional a partir de la arquitectura *transformer*. En los siguientes párrafos vamos a desgranar a qué hace referencia cada una de las palabras que forman el acrónimo BERT[2]:

1. La palabra bidireccional en BERT hace referencia a su capacidad para analizar el contexto completo de una palabra en una oración, tanto hacia la izquierda como hacia la derecha. Esto es una innovación respecto a los modelos de lenguaje tradicionales basados en RNNs (redes neuronales recurrentes) y otros de procesamiento secuencial anteriores a BERT que procesaban el texto

[2] Como detalle curioso e incluso divertido, BERT también es el nombre de un conocido personaje de Barrio Sésamo que en español se llama Blas y tenía cabeza de limón, personaje que sirve para representar el modelo en gran parte de la literatura no científica que habla sobre BERT.

unidireccionalmente (leían de izquierda a derecha o de derecha a izquierda, pero no las dos a la vez). Así, BERT supera esta limitación utilizando atención bidireccional, un mecanismo por el que cada palabra en una oración se representa teniendo en cuenta todas las palabras del contexto, independientemente de su posición relativa. Esto se logra gracias al uso del mecanismo de *Masked Language Modeling* (MLM), donde se predicen palabras enmascaradas teniendo en cuenta su contexto completo.

2. La palabra transformer en BERT hace referencia a la arquitectura del mismo nombre, presentada en un artículo con gran impacto (Vaswani et al., 2017) que ya acumula más de 100.000 referencias en Google Scholar. Los transformers utilizan una estructura de codificador-decodificador, pero a diferencia de modelos anteriores, se basan completamente en mecanismos de atención en lugar de RNNs o convolucionales. El componente central de la arquitectura transformer es el mecanismo de atención. La arquitectura transformer también se caracteriza por tener dos piezas clave: el codificador y el decodificador. El codificador consiste en una pila de capas idénticas, cada una con dos subcapas: atención de múltiples cabezas y una red propagación hacia adelante. El decodificador es similar, pero agrega una tercera subcapa de atención sobre la salida del codificador.
3. La palabra codificador hace referencia a que de la arquitectura transformer, normalmente con un codificador y un decodificador, BERT hace uso solamente del primero. Un codificador mapea las secuencias de símbolos a la entrada del modelo a una secuencia de representaciones continuas, es decir, representaciones vectoriales de las palabras. Este codificador utiliza el mecanismo de atención para relacionar cada palabra con todas las demás de una misma secuencia de entrada, lo que permite capturar dependencias a largo plazo y relaciones contextuales.
4. La palabra representaciones hace referencia a las representaciones vectoriales que produce el codificador. A diferencia de modelos estáticos anteriores como *Word2Vec* o *GloVe*, BERT produce representaciones contextuales dinámicas, es decir, la representación de una palabra depende del contexto completo en el que aparece (Shen & Liu, 2021). Por ejemplo, la palabra "banco" se codificará de forma diferente si digo "estoy en un banco del parque" que si digo "voy al banco a ingresar dinero" que "estamos sobre un banco de peces". La palabra banco aquí viene caracterizada por su contexto, y su representación vectorial es por tanto semántica. Estas representaciones son utilizadas a menudo como entrada para tareas de PLN, como clasificación de texto, análisis de sentimientos, y respuesta a preguntas.

3.2.1.1. *Arquitectura de BERT*

BERT es un modelo abierto y por tanto tiene muchas variantes que emanan de la original, cada una con sus propias características.

En Devlin et al. (2018) proponen dos configuraciones distintas llamadas *BERT-base* y *BERT-large*. BERT-base es un modelo con 12 bloques de Transformers, 768 capas ocultas, 12 cabezas de auto-atención y 110 millones de parámetros. BERT-*large* es un modelo mayor que BERT-*base*, tiene mejor desempeño y es más lento. En el resto de esta sección, cuando hablamos simplemente de BERT nos estamos refiriendo a *BERT-base*.

BERT está compuesto por 6 capas idénticas. Cada una de estas capas está compuesta por un mecanismo de autoatención de múltiples cabezas y una red de propagación hacia adelante. BERT usa una conexión residual (He et al., 2016) que conecta las dos sub-capas seguido de una capa de normalización (Ba et al., 2016).

En relación con la auto-atención multicabeza, definimos el producto escalar de la atención como se muestra en la Ecuación 1:

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}$$

Ecuación 1: Descripción matemática del mecanismo de atención de un transformer

Dónde Q is la matriz de las consultas, K es la matriz de las claves, V es la matriz de los valores y d_k es la dimensión de las matrices Q y K. Ahora podemos definir la atención multicabeza como se muestra en la Ecuación 2:

$$MultiHead(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Concat(head_1, \dots, head_h)W^{\boldsymbol{o}}$$

$$where\ head_i = Attention\left(\boldsymbol{Q}\boldsymbol{W}_{\boldsymbol{i}}^{\boldsymbol{Q}}, \boldsymbol{K}\boldsymbol{W}_{\boldsymbol{i}}^{\boldsymbol{K}}, \boldsymbol{V}\boldsymbol{W}_{\boldsymbol{i}}^{\boldsymbol{V}}\right)$$

Ecuación 2: Atención multi-cabeza

La atención multi-cabeza consiste en proyectar las consultas, claves y valores h veces utilizando diferentes proyecciones lineales aprendidas hacia las dimensiones d_k y d_v (dimensión de la matriz de valores), respectivamente. Luego, en cada una de estas versiones proyectadas de las consultas, claves y valores, se aplica la función de atención de manera paralela, generando valores de salida de dimensión d_v. Finalmente, estos valores se concatenan y se proyectan, resultando en los valores finales (Vaswani et al., 2017).

3.2.1.2. *Pre-entrenamiento y ajuste fino*

BERT es un modelo pre-entrenado. Esto significa que BERT ya ha consumido grandes cantidades de texto para construir una representación general del lenguaje que puede después transferirse a tareas específicas como clasificación de texto. Este pre-

entrenamiento es no supervisado y se basa en dos tareas: Enmascaramiento y predicción de la próxima sentencia.

BERT recibe *tokens* a su entrada en lugar de palabras. Un token es la unidad mínima de significado extraída de un texto. Generalmente, un token corresponde a una palabra, aunque dependiendo del método de segmentación utilizado también puede representar un carácter, una sílaba o incluso una subpalabra.

Para entrenar el modelo bidireccional, se enmascaran u ocultan algunos tokens de las sentencias de entrada, sustituyendo su token original por el token reservado *[MASK]*. En (Devlin et al., 2018) un 15% de las palabras son enmascaradas, pero esto puede cambiar en otras implementaciones de BERT. Tras el procesamiento de la sentencia, los vectores ocultos correspondientes al token enmascarado entran a una función softmax sobre el vocabulario utilizado y se aplica retropropagación del error para disminuir el error cometido por el modelo. Esta tarea de enmascaramiento tiene un problema: el token [MASK] no está disponible durante el ajuste fino y por tanto existe una desconexión entre estos dos procesos. Para mitigar este efecto, en el pre-entrenamiento no siempre se enmascaran tokens sustituyéndolos por el token [MASK] sino que se usa una combinación de la sustitución del token original por [MASK], la sustitución por un token aleatorio y la sustitución por el token original.

La predicción de la próxima sentencia es una tarea en la que BERT aprende a predecir si una segunda oración sigue lógicamente a la primera. Esta tarea es relevante para aplicaciones como la clasificación de texto, donde las relaciones entre diferentes segmentos de texto pueden ser relevantes. Para llevar a cabo este entrenamiento se seleccionan pares de sentencias y se envían al modelo. En el 50% de los casos, la segunda sentencia es una continuación natural de la primera. En el resto de casos, la segunda sentencia es escogida al azar entre todas las sentencias del texto.

Para llevar a cabo estas tareas se han usado datos procedentes del *BookCorpus* (800 millones de palabras) y de Wikipedia (2.500 millones de palabras).

El ajuste fino de BERT implica adaptar el modelo pre-entrenado a una tarea utilizando un conjunto de datos específico. El ajuste fino en BERT es muy sencillo gracias al mecanismo de auto-atención: para cada tarea se envían las entradas y salidas esperadas al modelo y se ajustan todos los parámetros en cada iteración. Este paso es análogo a la predicción de la siguiente sentencia, solo que en lugar de tener dos sentencias podemos tener:

- Pares de sentencias parafraseadas
- Hipótesis y premisas emparejadas
- Preguntas y respuestas
- Textos y etiquetas para clasificación de texto

En concreto, para tareas de clasificación de texto como las llevadas a cabo en este trabajo se usan textos y etiquetas. Los textos de entrada se tokenizan utilizando el tokenizador *WordPiece* de BERT. Para tareas de clasificación se añaden tokens especiales al inicio y al final de cada sentencia. El token [CLS] se coloca al inicio de la sentencia y se utiliza como representación agregada de toda la secuencia. El token especial [SEP] se coloca al final para indicar que la sentencia ha acabado.

A la salida del modelo se toma la representación del token [CLS] agregada que tiene tantos valores como capas ocultas tiene el modelo generado, es decir, un vector de 768 valores que encapsulan el contenido del input adaptado para tareas de clasificación. Esta representación se pasa por una capa de clasificación específica de la tarea, que generalmente consiste en una capa totalmente conectada seguida de una función softmax para calcular las probabilidades de las clases. Durante el ajuste fino, todos los parámetros de BERT (no solo la capa de salida) se ajustan usando el conjunto de datos etiquetado. Se utiliza una función de pérdida de entropía para entrenar el modelo, ajustando las predicciones del modelo a las etiquetas reales.

3.2.1.3. WordPiece

Clásicamente, manejar vocabularios abiertos y palabras poco frecuentes ha sido un gran desafío. Los enfoques tradicionales usan vocabularios fijos, lo que lleva a problemas cuando tenemos palabras que no se encuentran en el vocabulario predefinido. Si se opta por una estrategia de codificación carácter a carácter tenemos otro conjunto de problemas para trabajar con el resultado que nos impiden hacer tareas de alto nivel como clasificación textual de forma efectiva. WordPiece es un modelo de tokenización presentado por primera vez en 2012 (Schuster, 2012) que divide las palabras en subunidades más pequeñas llamadas tokens, logrando un equilibrio entre modelos basados en caracteres y palabras completas (Wu et al., 2016).

Durante la tokenización, las palabras son segmentadas en sus componentes de subpalabras según el modelo WordPiece entrenado. Se agrega un símbolo especial "_" al inicio de cada subpalabra para identificar límites de palabras y facilitar la reconstrucción del texto original. Dado que el modelo incluye marcadores explícitos para los límites de palabras, es posible recuperar la secuencia original a partir de la secuencia de subpalabras sin pérdida de información.

Esto también aumenta enormemente la eficiencia de los modelos: Al combinar la granularidad de los caracteres con la eficiencia de los modelos basados en palabras, WordPiece reduce la longitud de las secuencias en comparación con enfoques basados únicamente en caracteres. Para lenguajes occidentales, el vocabulario se limita a entre 8,000 y 32,000 unidades de subpalabras, lo que mejora la velocidad de decodificación y reduce la complejidad computacional. El modelo WordPiece es

ahora un estándar en la tokenización moderna, adoptado ampliamente en tareas de PLN más allá de la traducción, como se evidencia en modelos como BERT.

3.3. Clasificación multiclase

Una parte central de este trabajo es la clasificación de eventos. La clasificación es una tarea común en estadística y aprendizaje automático, más frecuente que los problemas de regresión. Para abordar estos problemas de clasificación, se utilizan modelos basados en datos de entrenamiento que buscan predecir correctamente tanto datos conocidos como nuevos (James et al., 2023).

3.3.1. Regresión logística

La regresión logística es una técnica estadística utilizada para predecir variables categóricas dentro de un conjunto predefinido de categorías. Este modelo asigna probabilidades de que una observación pertenezca a cada una de las categorías consideradas a partir de las variables independientes de la observación. De esta forma, puede usarse en combinación con otras técnicas para extender sus propiedades naturales.

La regresión logística modela la probabilidad de que una observación pertenezca a una categoría específica en función de sus predictores usando una función logística como la mostrada en la Ecuación 3.

$$p(X) = \frac{1}{1 + e^{-(\boldsymbol{\beta} \cdot \boldsymbol{X})}}$$

Ecuación 3: Función logística

Donde X es el vector de las variables independientes de la observación y $\boldsymbol{\beta}$ es el vector de los coeficientes o pesos aprendidos por el modelo.

La regresión logística se extiende naturalmente para incluir múltiples predictores (X_1, X_2, …, X_p). La Ecuación 3 se puede reescribir como el logaritmo de las probabilidades como una escala intermedia para representar la relación lineal entre los predictores y la respuesta como muestra la Ecuación 4.

$$\log \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Ecuación 4: Logaritmo del ratio las probabilidades de pertenecer a una clase

La Ecuación 4 suele ser apropiada para la clasificación binomial, en la que solo tenemos dos clases. Así, p(X) es la probabilidad de pertenecer a una categoría y 1-p(X) la probabilidad de pertenecer a la categoría complementaria. Pero cuando la respuesta tiene más de dos clases se utiliza la regresión logística multinomial. Este enfoque generaliza la regresión logística binaria mediante un sistema de ecuaciones

que modelan las probabilidades para *K* categorías. Así, la probabilidad de que una observación caracterizada por su vector de variables independientes X pertenezca a la clase k se modela tal y como se muestra en la Ecuación 5.

$$p(Y = k \mid X) = \frac{e^{\beta_{k0}+\beta_{k1}X_1+\beta_{k2}X_2+\cdots+\beta_{kp}X_p}}{1 + \sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}X_1+\beta_{l2}X_2+\cdots+\beta_{lp}X_p}}$$

Ecuación 5: Función logística multinomial

Esta ecuación utiliza la función exponencial normalizada que se utiliza para mapear el vector de *k* dimensiones de valores arbitrarios en otro de valores reales en el rango [0, 1], como corresponde a su naturaleza probabilística.

3.3.2. Evaluación de modelos de clasificación multi-clase

La evaluación de modelos de clasificación multi-clase es un proceso clave para medir el desempeño de un modelo y garantizar que funcione de manera adecuada en la tarea de clasificar muestras en diferentes categorías. Para ello, se emplean herramientas como las métricas de evaluación (precisión, sensibilidad, F1-score, entre otras) o la matriz de confusión. Estas herramientas permiten no solo identificar los aciertos y errores globales del modelo, sino también profundizar en su comportamiento para cada clase individual, lo cual es fundamental para mejorar su desempeño en tareas específicas(Heydarian et al., 2022; Krstinić et al., 2020; Vujović, 2021).

En problemas de clasificación multi-clase, donde un modelo debe decidir entre tres o más categorías posibles, la evaluación es más compleja que en tareas binarias. No basta con medir cuántas predicciones son correctas; también es crucial entender cómo se comporta el modelo para cada clase y cómo maneja las confusiones entre clases similares.

Una de las herramientas más potentes para evaluar modelos de clasificación multi-clase es la matriz de confusión. La matriz de confusión organiza los resultados en una tabla que muestra los aciertos y errores del modelo para cada clase. Por ejemplo, si un evento real de la clase *A* se clasifica como *B*, esto se reflejará en la matriz. Esta herramienta es esencial para identificar patrones de error y confusiones frecuentes entre clases. En una matriz con varias clases como las de la Figura 1 podemos ver cuántas muestras de cada clase han sido clasificados correctamente o confundidas con otras clases. Por ejemplo, la primera columna de la matriz de confusión tiene un conteo para todas las muestras predichas como clase *A* que efectivamente son de la clase *A*, cuántas de las muestras predichas como clase *A* eran en realidad de la clase *B* y así sucesivamente.

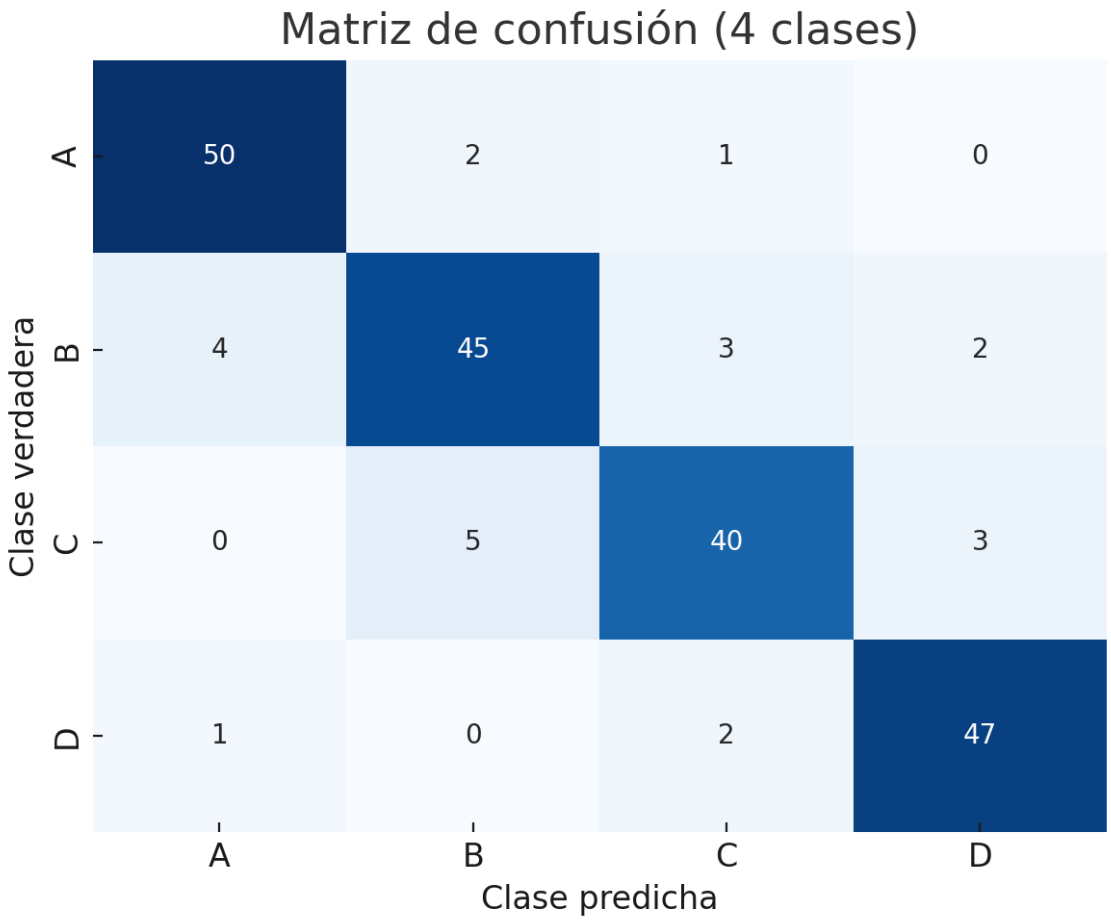

Figura 2: Ejemplo de una matriz de confusión para un problema de clasificación de 4 clases

De esta matriz de confusión emanan algunas definiciones. Por cada clase considerada, nuestras muestras pueden ser:

- TP (Verdadero Positivo): Son los casos en los que el modelo predice correctamente que una instancia pertenece a una clase específica.
- FP (Falso Positivo): Son los casos en los que el modelo predice incorrectamente que una instancia pertenece a una clase específica.
- TN (Verdaderos Negativos): Son los casos en los que el modelo predice correctamente que una instancia no pertenece a una clase específica.
- FN (Falsos Negativos): Son los casos en los que el modelo no detecta correctamente una clase específica.

Estas definiciones dependen de la clase que estemos considerando. Si una muestra de la clase *A* se clasifica como de la clase *C*, desde el punto de vista de la clase *A* es un FN. Sin embargo, desde el punto de vista de la clase *C*, es un FP. Para calcular métricas agnósticas de la clase considerada utilizando estas definiciones, se calculan métricas por cada clase y después se combinan para formar una métrica unificada. Estos

valores se pueden combinar de diferentes formas como una media simple o una media ponderada por el número de muestras de cada clase.

Una de las métricas más utilizadas es la exactitud, que como muestra la Ecuación 6 es el total de aciertos (TP y TN) entre el número total de muestras consideradas.

$$Exactitud = \frac{TP + TN}{TP + TN + FN + FP}$$

Ecuación 6: Fórmula de la exactitud

La sensibilidad o tasa de TP se calcula como el número de TP dividido entre el número de muestras positivas tal y como muestra la Ecuación 7. La precisión mide cuántas de las predicciones positivas realizadas por el modelo son correctas y se calcula tal y como muestra la Ecuación 8. La tasa de falsos positivos (FPR) es la proporción de los casos negativos que fueron clasificados erróneamente como positivos y se calcula tal y como muestra la Ecuación 9.

$$Sensibilidad = \frac{TP}{TP + FN}$$

Ecuación 7: Fórmula de la sensibilidad

$$Precisión = \frac{TP}{TP + FP}$$

Ecuación 8: Fórmula de la precisión

$$FPR = \frac{FP}{FP + TN}$$

Ecuación 9: Fórmula de la tasa de falsos positivos

La métrica *F1-score* combina precisión y sensibilidad en una sola métrica. Es la media armónica de ambas y se calcula tal y como muestra la Ecuación 10. Esta métrica es especialmente útil cuando el modelo necesita equilibrar precisión y sensibilidad como cuando tenemos clases desbalanceadas.

$$F1 - score = 2\frac{Precisión\ x\ Sensibilidad}{Precisión + Sensibilidad}$$

Ecuación 10: Fórmula del F1-score

El soporte es simplemente el número de instancias reales de cada categoría en el conjunto de datos. No es una métrica de desempeño, pero es fundamental para

contextualizar las demás métricas. Por ejemplo, si una categoría tiene un soporte bajo (pocos ejemplos), las métricas asociadas podrían ser menos fiables.

Estas métricas son esenciales porque proporcionan información más detallada que una métrica global como la precisión total del modelo. Por ejemplo:

- Un modelo con alta precisión, pero baja sensibilidad podría ser útil para evitar falsos positivos en aplicaciones críticas, pero podría ignorar muchos casos relevantes.
- El F1-score es clave para problemas con clases desbalanceadas, ya que penaliza tanto los falsos positivos como los negativos.
- El soporte contextualiza las métricas y ayuda a priorizar mejoras en categorías con mayor relevancia práctica.

La curva ROC y el área bajo la curva (AUC, por sus siglas en inglés) son herramientas utilizadas para evaluar el desempeño de un modelo de clasificación, especialmente en tareas donde es importante analizar la relación entre verdaderos positivos (TP) y falsos positivos (FP) a diferentes umbrales de decisión. Aunque se usan comúnmente en problemas de clasificación binaria, también se pueden adaptar para tareas multiclase.

La curva ROC es un gráfico que muestra el desempeño de un modelo al variar el umbral de decisión para clasificar una instancia como positiva. Este umbral determina qué probabilidad es suficiente para asignar una clase a una instancia. El eje X de la curva representa la Tasa de Falsos Positivos (FPR), mientras que el eje Y representa la Tasa de Verdaderos Positivos (TPR o Sensibilidad). Un modelo ideal se acercará al punto (0, 1), indicando una TPR alta (muchos aciertos) y una FPR baja (pocos errores).

El AUC es una métrica que resume el desempeño de la curva ROC en un solo valor. Es literalmente el área bajo la curva ROC y se calcula como una integral o sumando las áreas de los segmentos bajo la curva. Un AUC de 1.0 indica un modelo perfecto que clasifica todos los casos correctamente. Un AUC de 0.5 indica un modelo sin capacidad predictiva, equivalente a una clasificación aleatoria. Valores intermedios reflejan el equilibrio entre la sensibilidad y la tasa de falsos positivos del modelo. El AUC es muy útil ya que es independiente del umbral específico utilizado y proporciona una visión general del desempeño del modelo.

4. Modelo de clasificación de eventos turísticos

Se propone en esta sección la creación de un modelo general que clasifica eventos turísticos y se presenta un caso de uso de una aerolínea para validar su aplicación práctica.

Para llevar a cabo el modelo con el que evidenciamos que es posible la clasificación sistemática de eventos turísticos a partir de texto libre, usamos CRISP-DM. CRISP-DM ha sido una gran ayuda para estructurar el trabajo y nos permite iterar fácilmente. Los pasos de este CRISP-DM adaptado se muestran en la Figura 2.

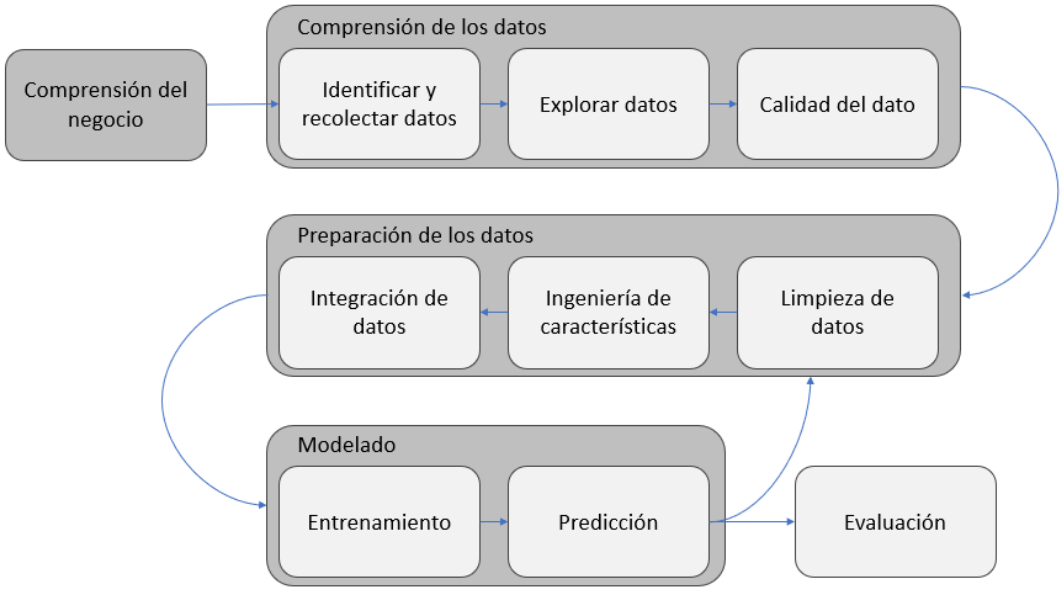

Figura 3 Modelo propuesto basado en CRISP-DM

Como hemos contado en el epígrafe 3.1, CRISP-DM es un modelo estándar que describe los pasos para llevar a buen puerto un proyecto de ciencia de datos. En este trabajo se ha utilizado el modelo CRISP-DM como marco metodológico general, pero con una ligera adaptación: se ha omitido la fase de despliegue, dado que se trata de una investigación científica y no de un proyecto de implementación comercial. Esta práctica es común en estudios académicos donde el énfasis está en la validación del modelo más que en su explotación productiva. CRISP-DM permite este tipo de adaptaciones, ya que no impone una estructura rígida y sus fases pueden ajustarse a los objetivos del estudio. Ejemplos relevantes en el ámbito del turismo que utilizan una versión adaptada del modelo incluyen los trabajos de Hamdan & Othman (2022), Andrews et al. (2019) y Yuensuk et al. (2022). En los dos primeros casos, al igual que en esta tesis, la fase de despliegue se omite o se menciona de forma superficial.

A continuación, se describe cada paso seguido hasta la evaluación de nuestro modelo.

4.1. Comprensión del negocio

4.1.1. Objetivos de negocio

El propósito de este trabajo desde el punto de vista de negocio es la generación de un modelo para la creación y normalización de catálogos de eventos turísticos como ya se ha mencionado en la sección 1.3. La motivación es clara: proveer a los diferentes actores del sector turístico de las herramientas y conocimiento necesarios para crear catálogos con el menor esfuerzo posible, trasladando este valor directamente al turista.

Como ya se ha visto en el capítulo introductorio y en el estado del arte, los eventos turísticos son productos o servicios diseñados para satisfacer múltiples necesidades del viajero, y actúan como fuertes motivadores del viaje. Para los destinos, los eventos son herramientas estratégicas que impulsan el desarrollo económico, la renovación urbana y la promoción turística. Clasificar los eventos turísticos permite a los destinos mejorar la planificación, segmentar audiencias, evaluar impactos y mitigar la estacionalidad. Para el turista, un catálogo estructurado facilita la identificación de eventos alineados con sus intereses y motivaciones, permitiendo una mejor planificación del viaje y una experiencia más enriquecedora.

Hay otros jugadores del sector turístico que también se benefician de la categorización de eventos turísticos como empresas de transporte (aerolíneas, cruceros, autobuses, etc.), ciudades inteligentes y OTAs (Online Travel Agency), principalmente a través de actividades relacionadas con el marketing. Estos jugadores pueden utilizar catálogos de eventos para mejorar la planificación del viaje del cliente final propiciando el uso de los servicios que ofrecen estos jugadores secundarios. Es fácil imaginar la situación en que un turista duda si comprar un billete de avión y la aerolínea le muestra un catálogo de eventos que puede encontrar en el destino, idealmente adaptado a las propias preferencias del turista, moviendo a este turista a la acción. También permiten otros beneficios como facilitar la creación de ofertas personalizadas que incluyan eventos relevantes para cada segmento del mercado. Como consecuencia directa la clasificación taxonómica facilita la promoción dirigida de eventos a audiencias específicas, aumentando la eficacia de las campañas de marketing, el desarrollo de nuevos productos turísticos basados en los tipos de eventos más populares o emergentes. Por último, hay beneficios indirectos: Los catálogos de productos normalizados permiten un análisis más detallado de la popularidad de eventos y los intereses de turistas de diferentes regiones lo que lleva a análisis estratégicos valiosos para la toma de decisiones. Este análisis también puede llevar a la optimización de recursos tanto en marketing como operativos y fomenta la

colaboración entre diferentes partes del ecosistema turístico permitiendo crear experiencias más completas e integradas gracias al marco común que representa el catálogo.

El problema de negocio que surge aquí es relevante: cómo podemos crear catálogos de eventos turísticos de forma estandarizada, eficiente y a bajo costo. Este es el problema de negocio en el que se centra este trabajo ya que como hemos visto en la revisión sistemática de la literatura en el capítulo 2, no es un problema resuelto y existe un gap de conocimiento entre la literatura y la técnica propuesta en este trabajo. Así, el objetivo de negocio que perseguimos es la creación de un método para crear catálogos que se pueda estandarizar, que sea rápido de crear y consultar y no implique grandes costes a los jugadores que deseen implementarlo.

Consideraremos el modelo sugerido exitoso si es capaz de clasificar grandes volúmenes de datos de naturaleza turística para la creación de un catálogo cometiendo errores razonables de una forma rápida y barata.

4.1.2. Desafíos de la creación del modelo

Aunque se hacen esfuerzos para la creación de catálogos de productos, la realidad es que es un esfuerzo ímprobo debido a la propia naturaleza de los eventos turísticos. Getz (2008) ya defiende que, para desarrollar el potencial de un destino turístico, los gestores de eventos turísticos deberían estar implicados en el proceso. Sin embargo, el turismo es una actividad fragmentada y disjunta (Benckendorff & Zehrer, 2013), (Echtner & Jamal, 1997), (Laws & Scott, 2015) y no es fácil llevar a cabo una adopción masiva de la segmentación de eventos. Ya Larsen & Mossberg (2007) habla del evento turístico como una ciencia relativa en la que la subjetividad del turista juega un papel fundamental y en la que se entrecruzan diferentes disciplinas como psicología, antropología social, estudios culturales, economía, marketing, etc. Hay muchos factores particulares de la creación de catálogos de actividades turísticas a escala que iremos desgranando en esta sección.

En primer lugar, los eventos turísticos son altamente descentralizados. Los promotores de eventos turísticos están especializados y son de muy diversa naturaleza. El turismo está caracterizado por la participación de una amplia gama de actores independientes, tales como promotores de eventos, agencias de viajes, operadores turísticos, gobiernos locales y empresas privadas. En este contexto, los organizadores de eventos no suelen formar parte de asociaciones o entidades unificadoras que funcionen como paraguas normativo. Esta ausencia de estructuras centralizadas o de un lobby unificado dificulta la estandarización de prácticas relacionadas con la publicación y categorización de eventos. Por ejemplo, mientras algunos eventos cuentan con descripciones exhaustivas y detalladas, otros apenas

proporcionan información básica. La falta de una estrategia de marketing común y de criterios consensuados para estructurar y presentar los datos genera inconsistencias significativas, complicando la creación de catálogos que sean coherentes y fáciles de usar.

También hay que tener en cuenta que hablamos de un sector enorme que engloba muchísimas actividades diferentes. Las actividades en las que un turista puede tener interés son increíblemente diversas y dependen mucho de la subjetividad del turista. Los eventos turísticos abarcan un espectro amplio y diverso que incluye festivales culturales, conciertos, ferias gastronómicas, competiciones deportivas, eventos religiosos, conferencias académicas y actividades recreativas, entre otros. Esta diversidad introduce desafíos significativos en la normalización de categorías, ya que los eventos pueden no ajustarse fácilmente a taxonomías predefinidas. Además, la clasificación adecuada de los eventos depende de su contexto cultural y social, lo que añade una capa de complejidad. Por ejemplo, un evento que en una región puede clasificarse como "religioso" podría tener un significado más amplio o diferente en otra cultura, como un festival comunitario.

Otra de las grandes complejidades de la situación actual de la industria del evento turístico es que las fuentes de información que contienen eventos turísticos son también increíblemente diversas. Los eventos turísticos se publican en una variedad de fuentes eclécticas, tales como sitios web oficiales de promotores, redes sociales, plataformas de venta de entradas, páginas de gobiernos locales, blogs de turismo y foros especializados. Estas fuentes difieren enormemente en términos de diseño, formato, profundidad de la información y objetivos. Si multiplicamos esta diversidad teniendo en cuenta fuentes de diferentes países o culturas en múltiples idiomas tenemos un crisol de fuentes heterogéneas de las que extraer información. Por ejemplo, un sitio oficial puede proporcionar información detallada del evento como nombre, descripción, horarios, geolocalización, precios, imágenes de calidad y otra información relevante. Sin embargo, una publicación del mismo evento en redes sociales puede limitarse a un nombre y una fecha, sin más detalles. Así otras fuentes como plataformas de venta de entradas pueden centrarse únicamente en aspectos comerciales omitiendo información contextual importante.

Otro desafío al que se enfrenta cualquiera que pretenda hacer un catálogo de eventos turísticos a escala es el crisol de condiciones particulares que tienen los eventos debidos a las barreras lingüísticas, regionales y culturales. La naturaleza regional del turismo implica la existencia de eventos en una variedad de idiomas. Solo la parte mecánica de los idiomas representa un obstáculo ya que nos encontramos con traducciones inconsistentes en la industria, múltiples alfabetos y sistemas de escritura y diferentes formas de expresar ideas. Pero si tenemos en cuenta la parte más cultural

y regional esta dificultad se multiplica. El lenguaje no es solo un medio de comunicación, sino también un reflejo del contexto cultural. Los significados, connotaciones y matices de ciertas palabras pueden variar considerablemente entre regiones, incluso dentro de un mismo idioma. Por ejemplo, términos como "feria" y sus traducciones pueden hacer alusión a eventos muy diferentes como eventos comerciales, recreativos, gastronómicos, culturales, etc. Estas diferencias dificultan la creación de taxonomías universales y obligan a los sistemas a adaptarse a múltiples interpretaciones de conceptos similares.

Algunos eventos turísticos están específicamente orientados a comunidades locales y se publican únicamente en el idioma predominante de la región. Esto puede limitar la visibilidad de dichos eventos a un público global. Por ejemplo, un festival de música en un pueblo de Japón puede ser anunciado exclusivamente en japonés, lo que hace que sea inaccesible para turistas internacionales sin conocimientos del idioma.

La multiculturalidad añade una capa de complejidad significativa a la compilación de catálogos de eventos turísticos, ya que cada cultura tiene formas únicas de concebir, organizar y promocionar eventos. Esto afecta tanto el contenido como la forma en que los eventos se perciben y clasifican. En un contexto multicultural, los eventos reflejan las tradiciones, valores y prácticas de sus comunidades de origen. Por ejemplo, las festividades religiosas son relevantes desde este punto de vista. Eventos como el Ramadán, el Día de los Muertos o Hanukkah tienen significados profundamente arraigados en sus respectivas culturas y pueden incluir actividades específicas difíciles de clasificar en taxonomías generales. O fuera de los eventos religiosos, eventos como ferias regionales, carnavales o mercados navideños son altamente específicos de ciertas culturas y pueden no tener equivalentes claros en otros contextos.

Otro factor que influye a la información a la que podemos acceder sobre eventos turísticos es la estrategia de marketing utilizada para promocionar eventos, que también depende de la cultura donde se organiza el evento. Así en algunos países los anuncios se enfocan en detalles logísticos (horarios, precios, ubicación), mientras que en otros se enfatiza el significado cultural o emocional del evento. Por ejemplo, un festival de música en Europa puede incluir listas detalladas de artistas, horarios y géneros, mientras que un evento similar en África podría anunciarse de manera más general, destacando el ambiente y la comunidad.

Si damos un paso atrás para contemplar con perspectiva la baraúnda de eventos asociados al sector turístico lo que nos encontramos es con una alta entropía, un caos formado por la conjunción de muchos factores: fragmentación de promotores, la definición amplia de evento turístico, el enorme número de eventos turísticos, la

naturaleza ecléctica de las fuentes de información, aspectos culturales, regionales y lingüísticos, diferentes idiomas, diversidad en el marketing del evento, etc. Esta situación actual del evento turístico lo hace muy difícil de categorizar por medios clásicos y es en el procesamiento automático de información donde debemos aproximarnos a una solución.

4.1.3. Objetivos para la ciencia de datos del modelo

El objetivo es obtener un modelo algorítmico que sea capaz de clasificar eventos turísticos atendiendo a una base taxonómica previamente definida. Esta clasificación ha de hacerse tomando como entrada únicamente el nombre y descripción del evento como texto libre. Obtener un modelo de clasificación funcional basado en texto libre es una forma de alcanzar el objetivo de negocio o al menos una prueba de concepto para asegurarnos de que vamos por el buen camino.

Aunque tenemos los datos y los medios para mejorar la clasificación usando piezas relevantes de información que trascienden el nombre y descripción del evento, evitamos este camino. La razón es que este método tiene vocación universal: Es nuestro deseo y esperanza que este trabajo pueda servir de base para que cualquier jugador del sector turístico pueda generar y normalizar su catálogo de eventos de la forma más sencilla posible. Y si hay algo que todas las colecciones de eventos turísticos tienen es el nombre y su descripción. Es más, en nuestro modelo la descripción puede dejarse vacía para tener en cuenta el caso más general posible.

Los objetivos específicos del proyecto para la ciencia de datos son múltiples:

- Demostrar la viabilidad del modelo: Validar que es posible realizar una clasificación automática de eventos turísticos con un modelo basado en IA, específicamente entrenado para procesar y categorizar texto libre. Este modelo debe ser capaz de asignar categorías taxonómicas previamente definidas, garantizando resultados consistentes y coherentes incluso en casos de datos incompletos o de baja calidad.
- Establecer un enfoque universal: Crear un modelo que no dependa de más datos adicionales para su uso que el nombre y la descripción del evento. Este diseño asegura que pueda aplicarse de manera generalizada, independientemente de la fuente de los datos, el idioma en que estén escritos o la región en la que se organicen los eventos. La capacidad de operar sin requisitos adicionales de datos lo convierte en una herramienta inclusiva y accesible para todos los actores del sector turístico.
- Medir la eficiencia en tareas de clasificación: Evaluar si el modelo puede procesar grandes volúmenes de datos de manera eficiente y en un tiempo razonable. Dado el alcance global del sector turístico, el modelo debe ser

escalable y capaz de manejar flujos de datos significativos sin perder precisión en la clasificación.

- Medir la calidad de la clasificación: Medir la precisión, la sensibilidad y el F1-score del modelo en diferentes categorías taxonómicas, incluyendo aquellas donde los eventos puedan presentar ambigüedades o superposiciones. Este análisis permitirá identificar las fortalezas y limitaciones del modelo, así como áreas potenciales de mejora.
- Probar la robustez del modelo: Asegurar que el modelo funcione adecuadamente incluso en condiciones adversas, como la ausencia de descripciones completas, textos mal estructurados, o en casos donde los eventos se encuentren en idiomas distintos al del conjunto de datos de entrenamiento principal. Este enfoque garantiza que el modelo sea resiliente y adaptable a diferentes escenarios reales.
- Sentar las bases para la normalización taxonómica: Generar un conjunto de resultados que puedan servir como referencia para normalizar catálogos de eventos turísticos a nivel global. Esto incluye validar si las categorías taxonómicas propuestas son adecuadas para representar la diversidad de eventos en el sector turístico y si el modelo puede ser una herramienta útil para crear catálogos uniformes y coherentes.
- Evaluar la aplicabilidad en el contexto empresarial: Determinar si los resultados del modelo son lo suficientemente sólidos como para ser adoptados por los jugadores clave del sector turístico, como destinos, operadores turísticos, aerolíneas y plataformas de reservas online. Esto incluye verificar que el modelo ofrezca beneficios tangibles, como la reducción de costos en la creación de catálogos y la mejora de la experiencia del usuario final.
- Identificar limitaciones y oportunidades de mejora: Analizar los errores y limitaciones del modelo para establecer futuras líneas de investigación. Por ejemplo, explorar la posibilidad de incorporar otras fuentes de información relevantes en iteraciones futuras, como imágenes, fechas o ubicaciones geográficas, sin comprometer la universalidad del método actual.

En última instancia se espera que el modelo propuesto sea capaz de clasificar eventos turísticos con precisión y eficiencia utilizando únicamente texto libre, estableciendo una base sólida para la creación de catálogos normalizados. De este modo, el modelo no solo respalda el objetivo de negocio general, sino que también demuestra el potencial de la IA como herramienta para resolver desafíos inherentes al sector turístico, como la fragmentación de eventos turísticos y su diversidad cultural y lingüística.

4.1.4. Beneficios potenciales del modelo

El modelo de clasificación automática de eventos turísticos considerado tiene el potencial de generar beneficios significativos para los diferentes actores del ecosistema turístico, desde destinos y operadores hasta los propios consumidores. Un buen catálogo de eventos turísticos puede integrarse en el ecosistema actual produciendo resultados positivos para la industria.

Uno de los beneficios más positivos del uso de catálogos normalizados es la creación de un estándar en el sector. La implementación de un modelo de clasificación universal fomenta la colaboración entre los diferentes actores del ecosistema turístico, permitiendo la integración de datos y experiencias. Un enfoque automatizado minimiza el tiempo y los recursos necesarios para clasificar eventos, especialmente en comparación con los métodos manuales.

Al proporcionar una base común para la gestión de eventos, el modelo abre la puerta a nuevas aplicaciones, como análisis predictivos, modelos de recomendación más avanzados o integración con ciudades inteligentes.

Pero este modelo también presenta beneficios específicos para destinos turísticos, turistas y proveedores.

4.1.4.1. Beneficios para los destinos turísticos

Los destinos turísticos se enfrentan al desafío de maximizar su atractivo y gestionar sus recursos de manera eficiente. Un modelo de clasificación automática permite mejorar la planificación y gestión de eventos dando una visión más estructurada de los eventos disponibles en su región geográfica. Facilita la identificación de brechas en ofertas de eventos y la implementación de estrategias para llenarlas, equilibrando la oferta durante todo el año y combatiendo la estacionalidad. También permite prever necesidades de infraestructura y servicios, como transporte, alojamiento y seguridad, en función de la naturaleza y popularidad de los eventos.

Desde el punto de vista del marketing de destino también es muy positivo: la segmentación de eventos basada en una taxonomía clara permite diseñar campañas de marketing más efectivas dirigidas a audiencias específicas, facilita la promoción cruzada de eventos relacionados (por ejemplo, un festival gastronómico vinculado a una feria de productos locales) y ayuda a destacar la identidad única del destino a través de eventos específicos que resalten su cultura y patrimonio.

Potencialmente, este modelo podría impactar económicamente en los destinos permitiendo medir de manera más precisa los beneficios económicos derivados de diferentes tipos de eventos. Por ejemplo, se podrían identificar las categorías con

mayor retorno de inversión, priorizando aquellos eventos que generan mayor gasto turístico o impacto social positivo.

La categorización sistemática puede ayudar a los destinos a mantener una oferta diversa y atractiva, asegurando que se cubran las necesidades e intereses de diferentes segmentos de turistas y puede reducir la redundancia en la oferta.

Por último, una taxonomía ampliamente usada por diferentes destinos permite a éstos comparar su desempeño con otros destinos de características similares, identificando mejores prácticas y áreas de mejora.

4.1.4.2. Beneficios para el turista

El modelo también tiene potencial beneficio para el consumidor final del evento turístico: el turista, facilitando la planificación de viajes y el descubrimiento de eventos relevantes.

La categorización clara y los catálogos organizados permiten a los turistas identificar eventos que se alineen con sus intereses específicos, incluso si los eventos provienen de múltiples fuentes heterogéneas. Así, se mejora la experiencia de usuario al reducir el tiempo y el esfuerzo necesario para buscar información dispersa.

La clasificación basada en taxonomías puede combinarse con herramientas de recomendación para sugerir eventos adaptados a las preferencias individuales del turista. De este modo, permite filtrar eventos por criterios como tipo, ubicación, fecha o idioma, ofreciendo un alto grado de personalización.

En general, un modelo que genere buenos catálogos de productos normalizados y organizados por categorías y niveles jerárquicos ayuda a los turistas a entender mejor la oferta cultural y social del destino, destacando aspectos únicos que podrían pasar desapercibidos.

4.1.4.3. Beneficios para los proveedores de servicios turísticos

Otros actores del sector, como aerolíneas, hoteles, agencias de viajes y OTAs, también se benefician significativamente de la normalización y clasificación de eventos.

El marketing de estos proveedores gira muy a menudo entorno al evento turístico: mostrar eventos relevantes durante el proceso de reserva puede influir en la decisión de compra de los turistas, especialmente para destinos emergentes o menos conocidos y facilita la creación de paquetes turísticos que incluyan eventos específicos como parte de la oferta. Así, uno de los beneficios potenciales de este modelo es ayudar al turista en la fase inspiracional de su viaje, lo que redunda directamente en un beneficio para el proveedor de servicios turísticos que le facilita el servicio. Incluso permite hacer marketing activamente al diseñar campañas de marketing dirigidas basadas en eventos destacados.

Para la industria turística, la información, en este caso de las preferencias de los turistas, es poder. Los catálogos estructurados generan datos valiosos sobre los intereses y comportamientos de los turistas, que pueden ser utilizados para desarrollar productos y servicios más alineados con sus necesidades. Combinado con otras técnicas, mejora la planificación operativa al predecir la demanda en función de eventos futuros.

De estos beneficios mencionados se desprende otro de los más deseados por los proveedores de servicios turísticos, el fomento de la fidelización. Ofrecer información sobre eventos relevantes mejora la experiencia del cliente, creando una percepción positiva de la marca y aumentando la probabilidad de repetición. Incluso puede ayudar al cliente a decidir usar solo proveedores que tengan información turística normalizada a fuerza de costumbre si ya la ha consumido varias veces con anterioridad.

Es en este contexto que se presenta un ejemplo de aplicación para una aerolínea que iremos desarrollando a lo largo de todo el capítulo. El supuesto es el de una aerolínea que opera en decenas de destinos internacionales y ve la oportunidad de reforzar su posicionamiento en el mercado a través de la promoción de eventos turísticos. Desde la perspectiva del marketing, el objetivo radica en desarrollar estrategias de segmentación y personalización que aumenten tanto la fidelización como las ventas de pasajes. Al incorporar un catálogo de eventos —clasificados según intereses como *música*, *deportes*, *arte* o *conferencias*— la aerolínea dispone de un canal adicional para dialogar con sus clientes potenciales, mostrándoles atractivas razones para volar hacia sus destinos.

Al mismo tiempo, la aerolínea busca fortalecer su propuesta de valor. Ofrecer un servicio de "inspiración al viaje" al exhibir eventos relevantes (festivales, ferias, exposiciones, etc.) permite diferenciarse de sus competidores que se limitan a la mera venta de billetes. Además, esta integración de contenidos se vincula con acciones de marketing multicanal (boletines, redes sociales, publicidad en buscadores), posibilitando dirigir campañas personalizadas a audiencias concretas: por ejemplo, usuarios que estén interesados en conciertos o en grandes citas deportivas.

4.2. Comprensión de los datos

La comprensión de los datos es el segundo paso de nuestra estrategia basada en CRISP-DM. En el caso de este trabajo, se enfoca en comprender la naturaleza, calidad y diversidad de los datos utilizados para entrenar y validar el modelo de clasificación automática de eventos turísticos. Dada la complejidad inherente de los eventos turísticos, esta sección aborda las características de las fuentes, las etapas de

recopilación, los desafíos asociados y las estrategias utilizadas para asegurar la calidad de los datos.

4.2.1. ¿Qué datos necesito?

El primer objetivo específico tal y como veíamos en la sección 4.1.3 es demostrar su viabilidad, es decir, que se puede crear un modelo de clasificación basado en técnicas de IA para clasificar eventos turísticos. El segundo objetivo específico es establecer un enfoque universal que no dependa de datos adicionales que el nombre y descripción del evento. Así nos aseguramos de que cualquier jugador del ecosistema turístico pueda utilizar este modelo con sus eventos, ya que lo único que se necesita es el texto libre del título del evento. Además, el modelo ha de ser robusto, así que algunos de estos textos estarán incompletos, mal estructurados y en múltiples idiomas.

Atendiendo al párrafo anterior queda claro que lo mínimo que necesitamos es un conjunto de datos con eventos turísticos que incluya el nombre del evento y la categoría a la que pertenece. Además, tras iterar en este proceso, añadimos una descripción del evento al modelo ya que solamente con el título los resultados son insatisfactorios.

El problema que surge aquí es que necesitamos las categorías de los eventos, las categorías que proponemos asociadas a eventos que encontramos en internet de fuentes que no conocen nuestra taxonomía. En otras palabras: Tenemos que generar estas categorías. Tenemos que resolver el problema de clasificación por otros medios que no son el aprendizaje supervisado para poder generar nuestro modelo. Tenemos que generar estas categorías sin utilizar IA siendo este un ejercicio fantástico para conocer el esfuerzo que tiene hacer un clasificador de eventos turísticos sin la ayuda de técnicas de IA.

Aunque el uso del algoritmo solo requiera de texto libre (nombre y una posible descripción) debemos obtener más información sobre los eventos para generar las categorías a las que pertenecen. Además, necesitamos definir muy bien las categorías posibles que pueden tener nuestros eventos.

Para obtener las categorías de los eventos vamos a apoyarnos en datos que suelen acompañar al evento y, por tanto, a recopilarlos. Estos datos son:

- Título: El título del evento, su nombre.
- Descripción: Una descripción del evento que información extra general sobre el evento.
- Geolocalización: la latitud y longitud del evento, es decir, dónde ocurre geográficamente.
- Lugar (opcional): la dirección o nombre del lugar donde se celebra el evento.

- Fuente: De dónde procede la información sobre el evento turístico recopilado. Puede venir de una API, de una red social, de una web en internet o de otras fuentes. En todo caso se recopilan los parámetros de la búsqueda. Si es una web, la URL (*Uniform Resource Locator*), si es una API, el contenido de la respuesta, y así con el resto de las fuentes.

Además, vamos a apoyarnos en una base de datos de lugares para complementar nuestra información. Así, en la fase de procesamiento, a partir de la geolocalización podremos determinar el lugar en el que se celebra un evento. Concluimos entonces que tendremos que obtener un conjunto de eventos que tengan al menos título, descripción (opcional), geolocalización, lugar(opcional), datos sobre su procedencia y, además, disponer de un conjunto de datos sobre lugares donde se pueden celebrar eventos.

Más allá de los datos que podemos recolectar necesitamos también un conjunto de categorías, una taxonomía dentro de la cual deben encajar nuestros datos. Para nuestro modelo hemos elegido una taxonomía con 6 categorías: *Artes escénicas*, *música*, *ferias y congresos*, *arte y cultura*, *deportes* y *otros eventos*. Esta elección responde a varios factores y es un equilibrio entre el esfuerzo de clasificar semiautomáticamente cientos de miles de eventos y tener un conjunto de datos relevante. Por un lado, estas categorías son recurrentes y representativas en los principales agregadores de eventos turísticos, lo que permite realizar una clasificación inicial semiautomática con un grado razonable de calidad y cobertura. Por otro lado, el uso de un sistema más granular o alternativo hubiera supuesto un esfuerzo mucho mayor en la fase de etiquetado manual, especialmente dado el gran volumen de datos manejado.

Aunque no se trata de una taxonomía canónica, cumple adecuadamente la función de estructurar el set de datos para tareas de clasificación automática, tal como se ha hecho en investigaciones anteriores con esquemas diseñados ad hoc (Gration et al., 2016; McKercher, 2016). Cabe señalar que, en el momento de definir esta taxonomía, no se preveían con claridad los posibles solapamientos conceptuales entre algunas categorías como música, artes escénicas o arte y cultura. Esta ambigüedad ha sido posteriormente observada en los resultados y es objeto de análisis y discusión en la sección 4.5. La taxonomía resultante se puede ver en la Figura 3.

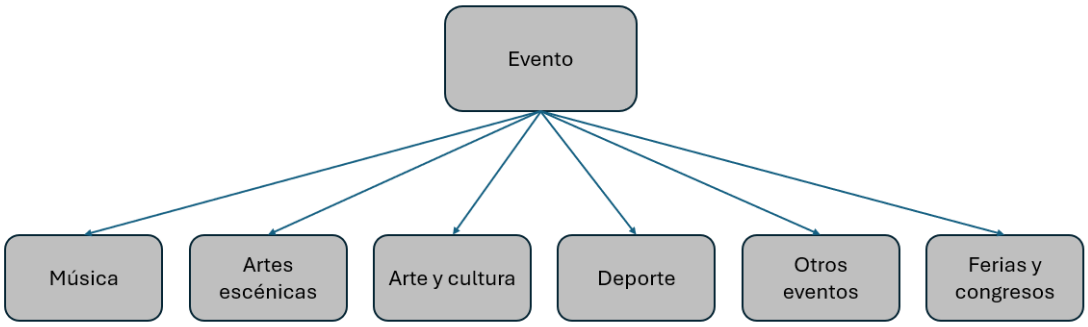

Figura 4: Taxonomía de eventos turísticos

4.2.2. Recolección de datos

Como hemos visto en la sección anterior tenemos dos conjuntos de datos que obtener: un conjunto de eventos turísticos y un conjunto de lugares donde ocurren eventos turísticos.

Para la recolección de datos de eventos turísticos se ha utilizado un software especializado en la adquisición de datos digitales de internet, lo que comúnmente se conoce en la industria como una araña (aprovechando el juego de palabras de la araña que recorre su red). Esta araña ha sido desarrollada en la industria específicamente para extraer información de eventos turísticos y está optimizada para este fin.

Esta araña es capaz de cargar el contenido de una web, navegar por ella y extraer tanto información estructurada como texto libre de sus fuentes. Al estar especializada en leer HTML es agnóstica del lenguaje en que venga la información, pudiendo obtener información en cualquier idioma. Esta araña es también capaz de obtener información directamente de una API, obteniendo información estructurada directamente.

Esta araña ha sido construida en lenguaje Python, diseñada para adaptarse a la heterogeneidad de fuentes y formatos presentes en el ecosistema de eventos turísticos. La araña emplea diversas librerías populares en el ámbito de la minería de datos en línea:

- Para la extracción de datos desde páginas estáticas o APIs REST, se utilizaron librerías como requests (Reitz, 2014) y lxml (Behnel, 2005), así como herramientas de manejo de datos estructurados como json (Python Software Foundation, 2024).
- Para fuentes más complejas o con elementos dinámicos, se emplearon Beautiful Soup (Richardson, 2019) y Selenium (Sharma, 2019), lo que permitió simular interacción con el navegador y acceder a los contenidos completos.

- El procesamiento y almacenamiento intermedio de los datos fue gestionado mediante pandas (The pandas development team, 2020), lo que facilitó la estructuración del dataset y su posterior limpieza y transformación.

Esta araña fue desarrollada con el objetivo de crear un conjunto de datos relevante, amplio y representativo. Su diseño modular ha permitido recolectar descripciones de eventos turísticos desde múltiples fuentes digitales como páginas oficiales, plataformas de agregación y redes sociales con un alto grado de control sobre el formato y la calidad de los datos.

Las fuentes de datos seleccionadas son muy amplias y se han escogido por la calidad de su información, por tener todas las piezas de datos que necesitamos y por mapearse bien con nuestras categorías. En total se han utilizado algo más de 1000 fuentes de datos en más de 30 países en 22 idiomas diferentes. Tras 6 años de adquisición de datos se han recopilado algo más de 700.000 eventos turísticos distribuidos por el globo con la calidad suficiente como para llevar a cabo este modelo.

La selección de fuentes de datos responde directamente a los requisitos definidos en la fase de comprensión del negocio descrita en la sección 4.1, en particular al caso de uso propuesto con una compañía aérea. Estas fuentes fueron seleccionadas porque representan destinos de interés comercial para dicha compañía, abarcando regiones en las que opera o desea ampliar su oferta. Esta cobertura geográfica amplia garantiza la diversidad cultural, idiomática y temática de los eventos, lo que permite evaluar el modelo de clasificación en un contexto realista y alineado con los objetivos de negocio.

El conjunto de datos de lugares susceptibles de tener eventos turísticos se ha construido a la vez que el conjunto de datos de eventos turísticos. Nunca se ha registrado un evento turístico sin un lugar asociado. Si un evento sucede en un lugar nunca registrado antes, el lugar se crea y se enriquece a través de una API externa, normalmente *Google Places*.

Los datos obtenidos son el resultado de extraer eventos turísticos con una araña periódicamente durante aproximadamente 6 años. Estos datos se han extraído y enriquecido con un proceso automático que ha conectado varias entidades relacionales de datos como son eventos turísticos, localizaciones en los que ocurren, traducciones, sesiones si se repiten en el tiempo, etc. A partir de estos datos enriquecidos se han codificado en formato JSON 8.48 Gigabytes de datos sobre eventos turísticos.

4.2.3. Análisis exploratorio de datos y filtrado inicial

El objetivo del análisis exploratorio de datos es analizar los datos disponibles para comprender las características, distribuciones y posibles problemas o limitaciones de los datos recopilados. En este análisis detectamos patrones y sesgos relevantes antes de la fase de modelado.

Los datos originales traen gran cantidad de información de la cual nosotros solo queremos un subconjunto. En nuestro conjunto de datos de eventos tenemos los siguientes campos:

- Traducciones: Tanto nombre como descripción se encuentran en múltiples idiomas. Los datos se recogen en el idioma original en que se encuentren en la fuente y después se traducen automáticamente hasta a 22 idiomas diferentes. El inglés siempre se encuentra presente tras la traducción de nombre y descripción. Debemos considerar que las traducciones no siempre son perfectas, no respetándose a veces nombres propios, la grafía del evento en su idioma original y otros efectos de la traducción automática.
- Fuente: Es el nombre de la fuente de la que se han extraído los eventos. Normalmente, el dominio del portal web en que vivía el evento. Estos portales web están compuestos por páginas de turismo de los destinos, proveedores de servicios como venta de entradas, gestores de eventos, etc.
- Geocodificación: Todos los eventos tienen latitud y longitud, toda la información está geolocalizada. Para eventos que no tienen un punto exacto y unívoco, como puede ser el evento "Semana Santa en Sevilla", se da una localización aproximada y relevante, como el centro de Sevilla o la catedral de Sevilla.
- Identificador de la región: Un identificador unívoco que identifica la región política en la que ocurre el evento.
- Fecha de inicio y final: Ambas referidas al evento. La fecha de inicio es la fecha en la que el evento da inicio en un lugar concreto y la fecha de final aquella en la que el evento se da por última vez. Algunos eventos, muy localizados en el tiempo tendrán la misma fecha de inicio y fin como, por ejemplo, un partido de fútbol que comienza y acaba el mismo día. Otros eventos comprenderán muchos días, como una obra teatral que se representa durante 3 meses en un teatro. Para que un evento se considere como único debe ocurrir en el mismo lugar. Así, un concierto de Joaquín Sabina en las Ventas de Madrid con tres sesiones (viernes, sábado y domingo) se considera un mismo evento, pero otro concierto de la misma gira en otra sala de conciertos se considera un evento diferente.

- Categorías: Todos los eventos del conjunto de datos considerado tienen un identificador taxonómico que hace referencia a la categoría y subcategoría a la que pertenece el evento. Una categoría puede ser *música* o *deporte*, por ejemplo. Una subcategoría puede ser *concierto* o *fútbol*. Cada subcategoría pertenece únicamente a una categoría padre.
- Identificación de revisión: La asignación de categoría a los eventos de nuestro conjunto de datos se han hecho usando múltiples técnicas, desde semisupervisadas hasta manuales. Este identificador nos dice si la asignación de categoría y subcategoría para un evento fue automática o manual (un humano leyó el evento, investigó y asignó manualmente el identificador de categoría y subcategoría).
- URL: La URL de la que se extrajo el evento. Las URL no solo nos sirven para tener trazabilidad de los eventos, sino que contienen información muy útil para clasificar automáticamente nuestros eventos. Por ejemplo, una ficticia URL de un evento que fuera "www.todosloseventos.com/conciertos/30568" nos dice que ese evento es un concierto (subcategoría) y, por tanto, *música* (categoría).
- Lugar: El nombre del sitio en que se celebra el evento como puede ser "*WiZink center*" o "Caja mágica".

Insistimos en este punto sobre la naturaleza de estos datos: No son crudos, tal como se extraen de las fuentes de datos, sino que ya han sido sometidos a un preprocesado para enriquecerlos. Es así por ejemplo como tenemos lugares sin duplicidad donde ocurren los eventos o directamente, un identificador de la categoría y subcategoría normalizadas de los datos.

El procesado de estos datos se hace con herramientas del ecosistema Python y constan de varios pasos. En primer lugar, todos los eventos incluyen un título y una descripción, recogidos en su idioma original y posteriormente traducidos automáticamente, asegurando siempre la presencia del inglés. Este paso se realiza para que todos los textos estén en el mismo idioma, facilitando las tareas de clasificación, aunque es importante señalar que las traducciones automáticas pueden introducir variaciones, especialmente en nombres propios o estructuras gramaticales particulares. Esta traducción se llevó a cabo con un servicio de traducción automática proporcionado por Google.

Junto a esta información textual, cada evento incorpora su fuente, entendida como el dominio web desde el que ha sido extraído (portales turísticos, plataformas de venta de entradas, sitios de organizadores de eventos, etc.), y un campo de URL que garantiza la trazabilidad del dato. Además, todos los eventos están geocodificados, es

decir, disponen de información de latitud y longitud. Esta información se extrae en el procesado de información de la página web del evento.

Con un sencillo análisis de nuestro conjunto de datos podemos extraer información sobre su morfología fácilmente. Tenemos un total de 10 columnas: fuente, lugar, descripción, fecha de inicio, fecha de final, identificador de revisión, URL, lugar, categoría, subcategoría y título. Nuestro conjunto de datos cuenta con 852.058 eventos diferentes, todos ellos con título, categoría y subcategoría. Adelantándonos un poco a la transformación de los datos y modelado sabemos que solo usaremos el título, descripción y categoría. Las demás columnas de este conjunto de datos tabular solo son usadas en esta sección para dar contexto.

Antes de continuar vamos a deduplicar nuestros datos. Consideraremos un evento único si tanto su título como descripción lo son. De forma implícita, su categoría y subcategoría serán las mismas. Así, deduplicamos eventos que vengan de diferentes fuentes u ocurran en diferentes lugares. Por ejemplo, un concierto de una gira que se llame "Dos pájaros de un tiro" y no disponga de descripción puede encontrarse varias veces en nuestro conjunto de datos por venir de varias fuentes diferentes o porque hemos considerado un evento diferente el concierto que ocurre en Madrid y el que ocurre en Barcelona. Sin embargo, para nuestro objetivo de proponer un clasificador de eventos turísticos que utilice exclusivamente el título y descripción del evento no tiene mucho sentido tener el mismo registro varias veces.

Así, antes del deduplicado, estos 852.058 eventos ocurren en 112.801 sitios distintos. Algunos de estos sitios son inválidos y no están curados. Así, el sitio en el que más eventos ocurren, 2647 eventos, es "*Please check meeting point*", un sitio claramente inválido fruto de un error en la adquisición de datos. También podemos observar que nuestros 852.058 eventos provienen de 425 fuentes diferentes, algunas de las cuales tienen varios portales diferentes de eventos y por tanto necesitan varias arañas (de ahí que el número de adquisitores de datos sea mucho mayor). Tras el deduplicado tenemos 491.385 eventos diferentes. Estos 491.385 eventos ocurren en 100.449 sitios distintos y provienen, al igual que antes del deduplicado, de 425 fuentes diferentes.

Las categorías de los eventos no están uniformemente distribuidas tal y como puede observarse en la Figura 4. Las necesidades de negocio a la hora de crear este conjunto de datos y a la hora de seleccionar las fuentes han hecho que tengamos un conjunto de datos pobremente balanceado. Se pueden consultar exactamente el número de eventos por categoría en la Tabla 2. El desbalanceo es evidente, hay más de 5 veces más eventos musicales que ferias y congresos, la primera categoría por número de eventos y la última respectivamente. Es de esperar un impacto en nuestro modelo. Normalmente los conjuntos de datos desbalanceados llevan a peores resultados en tareas de clasificación.

Categoría	Número de eventos
Música	182038
Artes escénicas	100765
Arte y cultura	74152
Deporte	57079
Otros eventos	44642
Ferias y congresos	32709

Tabla 2: Conteo de eventos por taxonomía

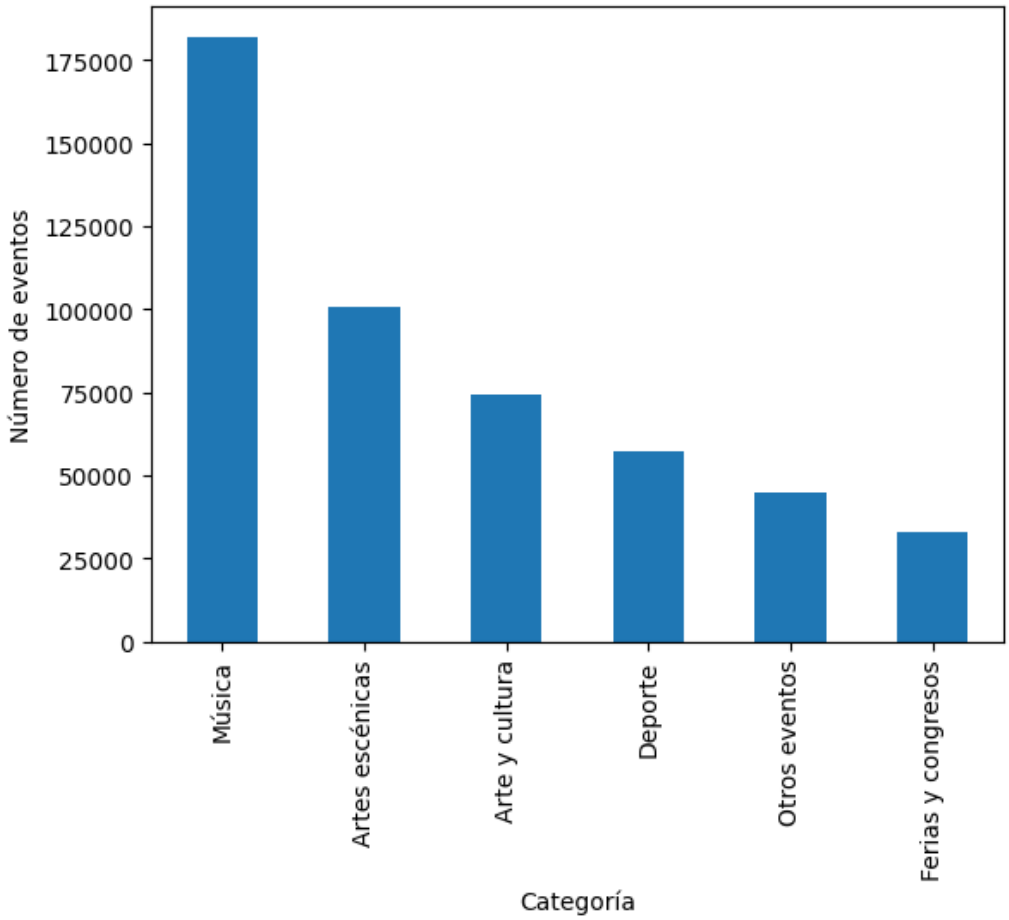

Figura 5: Conteo de eventos por categoría

Analizamos también la variable título, que es obligatoria en nuestro caso. Como podemos ver en la Figura 5, los títulos tienen en torno a 5 palabras, con una cola que se extiende hasta las 37 palabras, el máximo número en nuestro conjunto de datos.

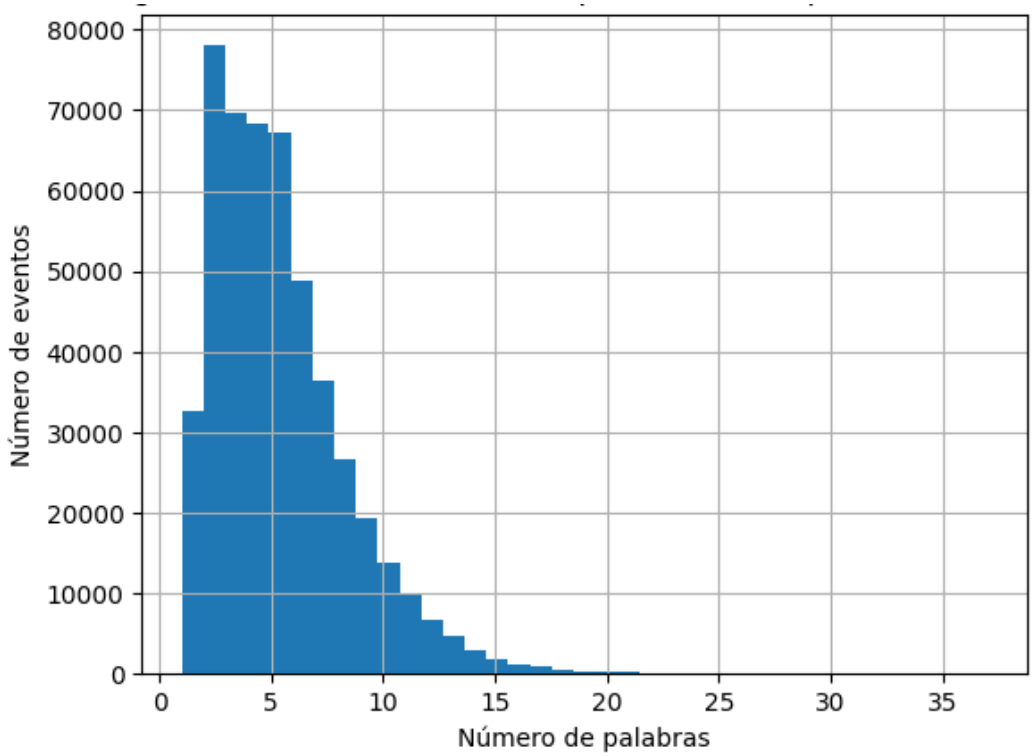

Figura 6: Histograma del conteo de eventos por número de palabras del título

Podemos hacer el mismo análisis con la variable descripción. Descripción en este caso sí que puede venir vacía y contener 0 palabras. De hecho, hay 126.015 eventos que no traen descripción alguna. Pero más notable aún es repetir este análisis con la concatenación del título y la descripción ya que este será el texto con el que alimentemos el modelo. Por limpieza y para mantener una estructura coherente, la concatenación se hace con un punto y espacio entre el título y la descripción. Llamaremos a esta concatenación de título, punto, espacio y descripción, el texto del evento. Como presenta la Figura 6, vemos que la gran mayoría de eventos tiene un número reducido de palabras, con un máximo en 3 palabras por evento. Para poner números y detalle a este máximo, observamos que 34.875 eventos tienen 3 palabras en su texto; el segundo valor más frecuente es 4 palabras con 17.447 eventos; le sigue 2 palabras con 16.116 eventos y a partir de ahí la larga cola de casos sigue hasta superar, en algunos eventos, las 500 palabras.

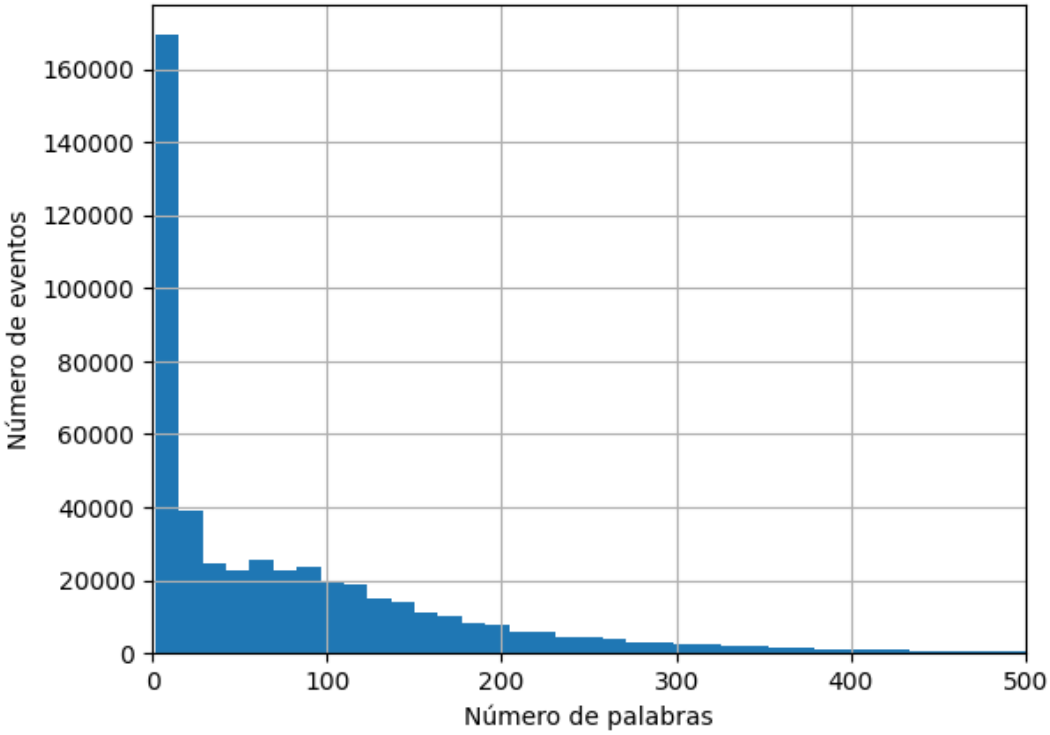

Figura 7: Histograma del conteo de eventos por número de palabras de su texto

Concluimos esta sección habiendo observado un fuerte desbalanceo en las categorías de los eventos. También hemos observado que hay un número relativamente bajo de palabras para describir un evento. Ambas circunstancias representan dificultades típicas en la clasificación de textos extraídos masivamente de internet, y nuestro modelo tendrá que lidiar con ello.

4.2.4. Control de calidad

Medir la calidad del texto libre procedente de múltiples fuentes, sobre diferentes temas en varios idiomas es una tarea difícil. Al tener un volumen de texto tan grande, las causas de la mala calidad en nuestro conjunto de datos se multiplican. Aun así, se ha llevado a cabo una concienzuda revisión de una parte relevante del conjunto de datos para identificar y de ser posible, reparar problemas en nuestro conjunto de datos.

En esa revisión manual de los datos se han encontrado bastantes palabras en idiomas diferentes al inglés. Incluso en alfabetos diferentes como el cirílico. Esto se puede deber a múltiples causas, pero la más probable es que el traductor automático ha decidido que es un nombre propio o una palabra que era mejor dejar en su idioma original. Otras causas se pueden deber a errores de traducción o palabras con errores tipográficos que al no ser reconocidas por el traductor se dejan en su forma original.

Otro de los problemas que se encuentran en el conjunto de datos es la inclusión de información de contacto o enlaces que no proveen información semántica y por tanto

son contraproducentes para el modelo. Estos datos suelen ser principalmente direcciones de correo electrónico o enlaces a gestión de entradas u otras operaciones relacionadas con el evento.

También se han encontrado en el conjunto de datos algunos caracteres que no están codificados en Unicode. Esto es negativo para la clasificación ya que nuestro modelo está preparado para funcionar solamente con textos codificados en Unicode y, por tanto, degrada la calidad de nuestro conjunto de datos.

Durante la fase de enriquecimiento del conjunto de datos, anterior a su uso en esta investigación, se implementó un proceso de control de calidad manual y periódico con el objetivo de mejorar la precisión de la categorización de eventos. Este proceso respondía a las limitaciones observadas en la asignación automática de categorías y subcategorías, especialmente en casos con descripciones poco claras, ambigüas o inusualmente breves.

Como parte de esta revisión, ciertos eventos fueron marcados como modificados manualmente cuando un revisor humano, tras analizar el título y la descripción, consideró que la categoría o subcategoría asignada originalmente no era adecuada. En total, se corrigieron manualmente 43.409 eventos. No obstante, no se dispone de registros que indiquen cuántos eventos fueron revisados en total ni cuáles eran las categorías erróneas que fueron sustituidas. A pesar de estas limitaciones, esta intervención permitió depurar significativamente el conjunto de datos y mejorar su calidad para fines de entrenamiento y evaluación del modelo.

Por ejemplo, un evento que tiene lugar en un estadio y ha sido extraído de una página especializada en eventos deportivos podría haber sido clasificado automáticamente como "deporte". Sin embargo, si el título del evento incluye palabras clave como "concierto" o menciona un artista musical, el sistema de revisión puede detectarlo y remitirlo a un revisor humano. Este, al identificar que se trata en realidad de un concierto celebrado en un estadio, puede corregir la clasificación, cambiando la categoría de "deporte" a "música".

4.2.5. Comprensión de datos para nuestro caso de uso

En el ejemplo que estamos usando en este capítulo sobre una aerolínea que decide desarrollar estrategias basadas en información de eventos turísticos, debemos tener en cuenta también la comprensión de los datos. Nuestra aerolínea de ejemplo tiene la intención de extraer valor del marketing asociado a la industria del evento turístico. Para diseñar campañas de marketing basadas en eventos, la aerolínea comienza por identificar qué tipo de información ayudará a perfilar mejor a su cliente objetivo. En primer lugar, se necesitan datos básicos (títulos y descripciones de los eventos, fechas, ubicaciones). Pero, desde un punto de vista publicitario, también resultan esenciales

aquellos indicadores que permitan medir la popularidad o relevancia de cada evento, como el número de menciones en redes sociales, valoraciones de usuarios o potencial de viralidad.

Al recopilar información proveniente de fuentes dispersas (portales especializados, agregadores de entradas, webs oficiales de organizadores), se deben contemplar los mercados a los que la aerolínea pretende dirigirse. Por ejemplo, para una campaña enfocada en el público "*millennial*", se priorizarán festivales de música electrónica y competiciones deportivas emergentes. En esta fase de comprensión de los datos, el equipo de marketing y el de análisis de datos colaboran para asegurar que el contenido recopilado sea lo suficientemente amplio y represente de manera realista el abanico de intereses de los clientes.

4.3. Preparación de los datos

En esta fase transformamos los datos para que el modelo pueda consumirlos correctamente y los limpiamos para en la medida de lo posible, elevar la calidad de los datos disponibles. En nuestro caso tenemos solamente texto libre y variables dependientes, así que las técnicas de preparación de datos serán predominantemente textuales. Y es que nuestro texto puede tener múltiples problemas derivados de la adquisición o el procesamiento posterior del conjunto de datos. La mayoría de estos problemas son causados por contaminación de HTML (*HyperText Markup Language*) o JSON (*JavaScript Object Notation*) durante el proceso de adquisición o por la inclusión de caracteres inválidos tal y como puede verse en la Figura 7.

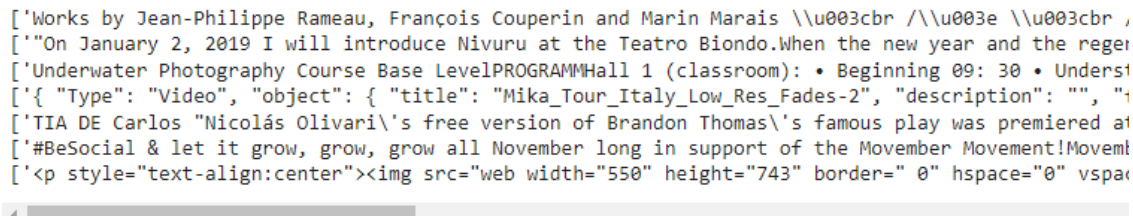

Figura 8: Ejemplo de textos mal formados en eventos

La preparación de datos es una etapa crítica para garantizar la calidad y eficacia del modelo de clasificación automática. En este trabajo, dicha preparación se ha realizado utilizando herramientas del ecosistema Python, especialmente diseñadas para la manipulación de datos estructurados y texto. Se han empleado librerías como pandas (The pandas development team, 2020) para la carga, filtrado y transformación de los datos tabulares; NumPy (Harris et al., 2020) para operaciones numéricas auxiliares; y módulos de manipulación de texto de Python (Van Rossum & others, 2020) como re, string, y unidecode, para normalizar, limpiar y simplificar los campos textuales.

El objetivo principal de esta fase es eliminar registros redundantes, incompletos o erróneos, y preparar las cadenas de texto para que puedan ser interpretadas

correctamente por los modelos de procesamiento del lenguaje natural. Por ejemplo, se eliminan caracteres especiales irrelevantes, se combinan título y descripción en un único campo y se transforman todas las cadenas a un formato homogéneo. Esto reduce el ruido del dataset y evita que el modelo aprenda patrones espurios o entrene con datos corruptos.

Ya hemos hecho un primer filtrado en la fase de comprensión de los datos: Hemos eliminado duplicados. Consideramos que dos eventos son duplicados si su nombre, descripción y categoría son idénticos, independientemente de la fuente de la que vengan o el lugar en el que ocurran. Esto evita tener la misma entrada para la fase de entrenamiento del modelo varias veces y por tanto evita la sobrerrepresentación no deseada de algunos eventos.

Comentamos en la fase de comprensión de los datos que hay textos con palabras que no están en inglés. Se ha decidido no tomar ninguna acción contra estos textos ya que se entiende que en la mayoría de los casos son nombres propios y palabras que tienen más sentido en su idioma original. Eso sí, tomaremos en cuenta esta circunstancia para seleccionar un modelo que lidie correctamente con estos textos y pueda gestionar texto multilingüe.

Mencionábamos también la aparición de información de contacto o enlaces web que nada aportaban a nuestro problema. Para eliminar estas piezas textuales hemos aplicado técnicas *regex* (de la contracción de *Regular Expression*, expresión regular en castellano) sobre el texto. Las expresiones regulares son patrones que identifican combinaciones de caracteres en una cadena de texto. En este caso se han usado para identificar enlaces y direcciones web y eliminarlas directamente del texto.

La sección 4.2 también nos alertaba sobre caracteres que no estaban codificados con Unicode, lo cual resulta muy nocivo para nuestro modelo. En este caso se ha optado por eliminar directamente todas aquellas palabras que contengan caracteres en una codificación diferente a Unicode.

Hay un último problema identificado con el texto derivado de errores de adquisición. Por ejemplo, antes del deduplicado, uno de los títulos más repetidos en nuestro catálogo era "// *PRIVATE EVENT*" con 134 ocurrencias, aún lejos del más repetido: "*Blue Man Group*" con 834 ocurrencias. Esto demuestra que hay eventos con nombres incorrectos en nuestro catálogo, pero se ha decidido no hacer nada al respecto y considerar estos errores de adquisición un problema inherente a la construcción de catálogos.

Más allá del texto, hay un pequeño problema en la codificación de las categorías que corresponden a cada evento: El modelo solo admite categorías secuenciales que comiencen en 0. Como nosotros consideramos 6 categorías, estas deben tener los

identificadores 0, 1, 2, 3, 4 y 5 en lugar del identificador con el que venían: 173, 268, 59, 2, 273 y 70. Para corregir esto sencillamente se hace un mapeo, una correspondencia biyectiva que conecte nuestros nuevos identificadores con los antiguos tal y como viene representado en la Tabla 3.

Identificador de la Antigua Categoría	Identificador de la Nueva Categoría	Categoría
173	0	Música
268	1	Artes escénicas
59	2	Arte y cultura
2	3	Deporte
273	4	Otros eventos
70	5	Ferias y congresos

Tabla 3: Mapeo de los identificadores de categorías

Por último, aplicamos otra transformación para que nuestro modelo pueda consumir correctamente la información. En nuestro caso vamos a usar un modelo basado en BERT que solo admite 512 tokens. Aunque no hay una correspondencia unívoca entre el número de tokens y palabras, el número de tokens es superior al de palabras para textos en inglés en una proporción aproximada de 4 tokens por cada 3 palabras. De los 512 tokens que admite el modelo, 2 se usan internamente en tareas de clasificación, así que nos quedan 510 tokens disponibles. Vamos a podar nuestros textos y ajustarlos a un máximo de 500 palabras para alimentar el modelo.

En este punto ya tenemos la información tratada y normalizada para que el modelo pueda consumirla. Pasamos pues a la fase de modelado, una de las más relevantes y si me permiten la licencia, de las más divertidas.

4.4. Modelado

El objetivo de la fase de modelado es obtener un modelo, valga la redundancia, que sea capaz de aceptar texto libre a su entrada y devuelva la categoría más probable a su salida según un sistema taxonómico predefinido.

Aunque normalmente hay más datos disponibles sobre eventos turísticos que su título y descripción textual, el modelo propuesto solo admite una cadena de texto a su entrada. Así cumplimos con el objetivo de negocio establecido en la sección 4.1.3 de tener un enfoque universal que no tome más datos que una cadena de texto.

Para construir este modelo que consume texto y devuelve una categoría debemos transformar el texto en una representación de su significado numérica que un ordenador pueda consumir. El siguiente paso es tomar esa representación numérica de un texto como entrada de un modelo clasificador y obtener la probabilidad de pertenecer a cada categoría y, por tanto, de su categoría más probable. Para el primer

paso, la transformación del texto en una representación numérica de su significado se utiliza un modelo basado en BERT. Para el segundo paso, la obtención de la categoría más probable de un evento, se usa un modelo basado en regresión logística.

BERT ha demostrado ser una de las tecnologías más efectivas para tareas de procesamiento del lenguaje natural como hemos visto en la sección 3.2.1. Es particularmente útil en contextos donde se requiere una comprensión profunda y contextual del texto, como en el caso de las descripciones de eventos turísticos. Su capacidad de generar representaciones semánticas bidireccionales y contextualizadas lo convierte en una herramienta ideal para capturar matices y relaciones complejas en el lenguaje.

En el contexto del análisis turístico, recientes estudios como el de Kirilenko & Stepchenkova (2025) destacan a BERT como uno de los modelos de lenguaje más avanzados para la identificación y modelado de tópicos, superando ampliamente a enfoques tradicionales como LDA en métricas de coherencia temática. En su estudio comparativo, BERT logró los mejores resultados tanto en calidad de temas como en facilidad de interpretación por expertos humanos. Esto se debe, en gran parte, a su capacidad para representar las palabras en función de su contexto completo, lo cual es especialmente útil cuando se trabaja con descripciones breves, ambiguas o multilingües, como ocurre habitualmente en el turismo de eventos. Además, el artículo enfatiza cómo BERT facilita el análisis semántico de contenido no estructurado, algo crucial cuando se extrae información de fuentes web heterogéneas y no estandarizadas. Por tanto, las funcionalidades de BERT no solo optimizan la tarea de clasificación automática, sino que también permiten construir modelos más robustos, interpretables y alineados con la realidad del lenguaje turístico actual.

BERT tiene muchas características relevantes para nuestro problema como también se ha visto en la sección 3.2.1. Su capacidad para trabajar con texto libre y obtener representaciones vectoriales significativas es una de las principales. Serán estas representaciones las que usaremos para clasificar y asignar categorías a los eventos que estos textos representan. Otra propiedad de los modelos BERT deseable es la bidireccionalidad. Debido a que las descripciones de eventos suelen ser complejas y algunos nombres necesitan de mucho contexto para entenderlos (como nombres de grupos teatrales o ferias marginales) necesitamos un modelo que extraiga de forma correcta todo ese contexto. Pensemos que un evento cuyo título sea “festival de primavera” necesita contexto para diferenciar si se trata de un evento cultural, gastronómico o religioso.

BERT es un modelo abierto que podemos entrenar nosotros mismos, así que podemos ajustarlo para tareas concretas como las que nos ocupa de clasificación siempre y cuando dispongamos de datos etiquetados como es nuestro caso. BERT también se

puede entrenar de forma no supervisada para mejorar la detección de patrones semánticos para mejorar la vectorización de textos. Adaptar el modelo mediante técnicas de ajuste fino es necesario para condicionarlo a nuestro problema y por tanto mejorar dramáticamente la efectividad de la clasificación. Además, BERT ofrece un buen equilibrio entre su capacidad para generar vectores representativos de texto, codificar los matices semánticos necesarios para diferenciar textos y clasificarlos correctamente y la capacidad computacional necesaria para correr el modelo. En nuestro caso el modelo usado puede correr sin mayores problemas en una gran proporción de ordenadores domésticos sin requerir el uso de BERT una gran inversión económica. Se ofrecerán datos sobre el hardware usado para la clasificación en la sección de evaluación del modelo de este mismo capítulo.

Otra de las fortalezas de BERT y que nos es útil para nuestro cometido es el soporte multilingüe. Algunas variantes de BERT permiten procesar datos textuales en más de 100 lenguas, lo que nos es muy útil a nosotros que tenemos eventos traducidos de 22 idiomas diferentes que tienen nombres e información aún en el idioma original.

BERT puede hacer uso de la técnica *transfer learning*, que nos permite aprovechar el conocimiento adquirido con potentes modelos de gran tamaño y transferirlo a modelos más pequeños y computacionalmente más baratos para resolver tareas específicas. Es gracias a este principio que seleccionamos el modelo derivado de BERT que vamos a usar en este trabajo: *DistilBERT* (Sanh et al., 2020). Este modelo es aproximadamente un 40% más pequeño que BERT mientras retiene un 90% de su capacidad, lo que lo hace especialmente adecuado para aplicaciones donde el volumen de datos es elevado o los recursos computacionales son limitados. Es por esto por lo que lo hemos seleccionado ya que entendemos que es más universal al necesitar máquinas más accesibles para funcionar.

Para este trabajo se ha obtenido el modelo preentrenado DistilBERT del repositorio de modelos de *Hugging Face*, una empresa que, entre otras cosas, gestiona modelos de IA. La variante específica seleccionada se llama *distilbert-base-multilingual-cased*. Ha sido entrenada en 104 idiomas y tiene 134 millones de parámetros, lo que garantiza una cobertura lingüística amplia y una adaptabilidad adecuada a los textos turísticos procedentes de fuentes internacionales. Su equilibrio entre rendimiento y eficiencia computacional lo convierte en una elección estratégica tanto para entornos de investigación como para futuras aplicaciones en el sector turístico. Además, este modelo necesita la mitad de tiempo que BERT base para llevar a cabo sus tareas.

Pasamos ahora a justificar el uso de un clasificador basado en regresión logística para clasificar nuestros eventos.

En primer lugar, como ya se vio en la sección 3.3.1, la regresión logística es simple y eficiente. Es rápida e interpretable. Además, las representaciones vectoriales generadas por BERT ya encapsulan información compleja y no lineal sobre las relaciones semánticas en los datos como se vio en la sección 3.2.1. Usar un clasificador sencillo como la regresión logística permite aprovechar directamente estas representaciones sin agregar complejidad innecesaria. Así, también se reduce el tiempo de entrenamiento en comparación con otros clasificadores más complejos como pueden ser aquellos basados en redes profundas.

Estudios recientes han demostrado que la combinación de modelos basados en BERT con clasificadores clásicos como la regresión logística puede ofrecer resultados de alta precisión, simplicidad interpretativa y eficiencia computacional. Por ejemplo, Hazim & Ata (2024) muestran que la regresión logística aplicada sobre embeddings de SBERT o RoBERTa alcanza rendimientos excepcionales, con la ventaja añadida de una mayor interpretabilidad y rapidez respecto a redes neuronales profundas. Otro ejemplo puede encontrarse en Guo et al. (2023), donde se clasifica texto para detectar spam con BERT y varios clasificadores, destacando la regresión logística por su calidad como clasificador.

La regresión logística aporta mucha información a la clasificación ya que obtenemos la probabilidad de un evento de pertenecer a cada una de las clases de nuestra taxonomía. Así podemos realizar la tarea trivial de seleccionar la taxonomía más probable. Pero también nos permite identificar casos dudosos en que las probabilidades de las categorías en cabeza sean muy parecidas y en general hacer un estudio estadístico mucho más completo. Incluso nos permite probar diferentes estrategias de asignación basada en probabilidades. Y no menos importante es que al tener salidas de probabilidades podemos obtener métricas de calidad como la curva ROC de forma nativa. Tenemos así otra de las propiedades que hacen a la regresión logística un buen clasificador para nuestro problema: La regresión logística es altamente interpretable, ya que permite analizar las contribuciones individuales de cada característica (en este caso, componentes del vector generado por BERT) a la probabilidad predicha.

La regresión logística brilla cuando se enfrenta a representaciones densas de alta dimensionalidad, es decir, vectores con muchas componentes de las cuales prácticamente todas ofrecen información relevante (en contraposición a vectores dispersos, en los que la mayoría de las posiciones son constantes o directamente 0). DistilBERT genera vectores densos de dimensionalidad 768, optimizados para capturar relaciones semánticas. La regresión logística puede mapear estos vectores directamente a categorías de manera efectiva. Como beneficio añadido, la simplicidad que la regresión logística aporta a este problema la hace más resistente al sobreajuste

al tener menos parámetros que clasificadores más complejos como aquellos basados en redes profundas. Esto es especialmente importante en problemas pobremente balanceados como el que tenemos nosotros.

La regresión logística también es computacionalmente eficiente y escalable. Dado que el conjunto de datos incluye cientos de miles de eventos, el clasificador debe ser lo suficientemente eficiente para entrenarse y predecir en un tiempo razonable, especialmente en aplicaciones prácticas como catálogos dinámicos de eventos turísticos.

Existen librerías para el lenguaje de programación Python muy bien implementadas y que nos ofrecen mucha libertad a la hora de parametrizar el clasificador, lo que facilita la creación de flujos de trabajo de clasificación.

Es por todo esto que se considera que la elección de un clasificador basado en regresión logística rápido, económico y fácil de mantener es coherente con el objetivo de ofrecer una solución escalable y de bajo costo que pueda ser adoptada ampliamente por actores del sector turístico, incluso aquellos con recursos técnicos limitados.

Tras la elección de estas dos tecnologías centrales en nuestro trabajo hemos de distinguir entre las fases de entrenamiento y clasificación del modelo. En la fase de entrenamiento el modelo consumirá texto asociado a eventos y sus categorías taxonómicas y nos devolverá una variante de BERT y de regresión logística con parámetros adaptados a la resolución de nuestro problema. En la fase de clasificación, el modelo con parámetros ya adaptados consumirá texto libre de eventos y devolverá la categoría a la que pertenecen.

Para la fase de entrenamiento se usa una parte representativa de nuestro conjunto de datos, en concreto un 75% de los datos. Esta selección se hace mediante un proceso de muestreo estratificado de tal modo que todas las categorías estén correctamente representadas. El 25% restante se utiliza para evaluar el modelo. Estos dos conjuntos de datos no deben tener solape ya que queremos evaluar cómo el modelo clasifica eventos que no haya visto nunca en la fase de entrenamiento, simulando así condiciones de su uso en producción.

Siguiendo nuestro ejemplo, para dar vida a la estrategia definida por nuestra aerolínea, se entrena un sistema de clasificación basado en técnicas de PLN como BERT y un algoritmo de aprendizaje automático supervisado que asigne automáticamente categorías relevantes a cada evento. Este paso es determinante para las estrategias de marketing digital, ya que la segmentación se vuelve mucho más precisa:

- Clasificación de eventos: Al etiquetar un concierto como "música electrónica" o "pop/rock", el equipo de marketing puede desplegar campañas específicas a los viajeros que mostraron afinidad por ese género musical en reservas pasadas o en interacciones con la web corporativa.
- Generación de audiencias personalizadas: Una vez catalogados, los eventos se integran con las plataformas de email marketing, redes sociales o sistemas de gestión de relaciones con el cliente (CRM). Esto permite crear audiencias basadas en el tipo de evento y lanzar anuncios altamente dirigidos, incrementando la tasa de conversión y la probabilidad de que el usuario reserve un vuelo.
- Sinergia con promociones cruzadas: Si la aerolínea colabora con patrocinadores o proveedores locales, puede ofrecer paquetes que incluyan vuelo, hospedaje y entradas a eventos. Gracias a la clasificación automática, esta coalescencia se realiza de forma escalable, fomentando un marketing relacional que va más allá de la simple venta de un billete de avión.

4.4.1. Fase de entrenamiento

El primer paso para disponer de un modelo que pueda clasificar eventos turísticos a partir de texto libre es entrenar el modelo. Para ello enviamos el texto junto a la categoría del evento al modelo con el objetivo de que aprenda los patrones semánticos de los textos y sea capaz de asignar categorías dado un texto en el futuro.

Los datos con los que contamos en esta fase son aquellos del set de entrenamiento, es decir, el 75% de los datos originales estratificados por categoría compuestos por texto libre y la categoría asignada al evento. El primer paso aquí es tokenizar todo el texto para obtener subpalabras con carga semántica. Para ello utilizamos un tokenizador preparado para funcionar con *distilBERT uncased*, el modelo basado en BERT que estamos utilizando. Además, es en esta fase cuando se añaden los dos tokens especiales necesarios para llevar a cabo tareas de clasificación: Uno que marca el comienzo de la frase y otro, el final. Es también en esta fase donde se truncan los tokens producidos para que el total de tokens enviados a la siguiente fase sean 512 como máximo; el máximo que admite el modelo son 510 tokens correspondientes al texto codificado y 2 correspondientes a los tokens especiales que el modelo necesita para llevar a cabo tareas de clasificación. La salida del tokenizador es doble: Por una parte, tenemos los tokens resultantes de codificar el texto libre, de longitud variable dependiendo del texto de entrada, y, por otra parte, una máscara que indica qué tokens son válidos y qué tokens son de relleno para llegar hasta los 512 que el modelo admite. Por ejemplo, podríamos codificar un texto de 75 palabras que resultara en 100 tokens. Se añaden los tokens especiales y acabaríamos con 102 tokens para nuestro

texto. Nuestra máscara será en este caso un vector de 512 posiciones donde las 102 primeras posiciones tienen el valor 1 y las restantes el valor 0.

El siguiente paso es la producción de vectores que encapsulen la información semántica del texto. Para ello sencillamente mandamos los tokens y la máscara obtenidos en el paso anterior al modelo distilBERT. Cada evento tokenizado produce un vector a la salida de distilBERT de un vector de 768 posiciones que codifican el texto original en un espacio algebraico semántico.

Este vector se envía entonces a un clasificador basado en regresión logística. El clasificador consume el vector de dimensión 768 y devuelve un vector de dimensión 6. Este vector de salida contiene las probabilidades de que el evento pertenezca a cada una de las categorías consideradas. Si la probabilidad más alta está en la primera posición diremos que la predicción de la categoría es *música*. Si la probabilidad mayor se encuentra en la segunda posición, consideraremos que la predicción de la categoría es *artes escénicas*. Y así sucesivamente.

Hasta ahora hemos hecho una operación en la que no hemos utilizado las categorías del conjunto de datos de entrenamiento y en la que la información ha sufrido diferentes transformaciones secuenciales, es decir, un proceso de propagación hacia adelante.

El vector de probabilidades obtenido del clasificador basado en regresión logística tiene, normalmente, todos sus valores por encima de 0. El modelo normalmente considera que la probabilidad de que un evento pertenezca a cualquier categoría puede ser baja, pero no nula. Es en este punto donde utilizamos las categorías de nuestro conjunto de entrenamiento, esas que presuponemos correctas, y las comparamos con el vector de probabilidades. El objetivo de esa comparación es una medida del error cometido en la clasificación, cuánto se aleja la predicción del modelo del valor real preasignado. Este error es enviado a distilBERT de nuevo para que modifique los pesos de sus parámetros de tal forma que se reduzca el error cometido. Al proceso de enviar el error cometido por el modelo para actualizar los pesos de las capas neuronales del mismo se le llama retropropagación del error.

Tras enviar nuestro conjunto de datos de entrenamiento al modelo hemos dado la oportunidad a cada evento de modificar el modelo para clasificar mejor los eventos turísticos. Pero normalmente no es suficiente con una sola pasada, o como se le llama en la literatura especializada, época. Tal y como se desprende de la lectura de la sección 3.2.1, lo ideal es utilizar entre 2 y 4 épocas. Tras repetir el proceso varias veces ya tendremos un modelo entrenado para clasificar textos en las categorías propuestas.

4.4.2. Fase de clasificación

La fase de clasificación es aquella en la que utilizamos el modelo para predecir categorías dado un texto a su entrada. Es en esta fase donde podemos alimentar al modelo con el conjunto de datos de prueba para obtener métricas sobre la calidad y eficiencia de la clasificación. Esta fase es muy parecida a la fase de propagación hacia adelante ya descrita por lo que no nos extenderemos con detalles.

Al igual que en la fase de propagación hacia adelante del apartado 4.4.1, empezamos tokenizando el texto de nuestro conjunto de datos de test. Utilizamos la salida del tokenizador, es decir, los tokens y la máscara para alimentar al modelo ya entrenado basado en distilBERT y producir vectores con información semántica. Estos vectores de dimensión 768 son una representación semántica de la secuencia textual de entrada. Estos vectores sirven de input para el clasificador basado en regresión logística que devuelve 6 valores con las probabilidades de pertenecer a cada categoría de nuestra taxonomía. Por último, y a diferencia del entrenamiento, aplicamos una función exponencial normalizada o softmax para colapsar el vector de probabilidades en el valor predicho del evento. Esta función transforma los valores obtenidos para cada clase en probabilidades normalizadas que suman 1, lo que permite interpretar la salida del modelo como la distribución de confianza sobre cada categoría. La función softmax es especialmente adecuada en tareas de clasificación multiclase con BERT como se comenta en la sección 3.2.1, como en este caso, ya que no solo permite identificar la clase más probable, sino también cuantificar el grado de certidumbre del modelo respecto a esa predicción. El flujo completo de la fase de clasificación puede verse en la Figura 8.

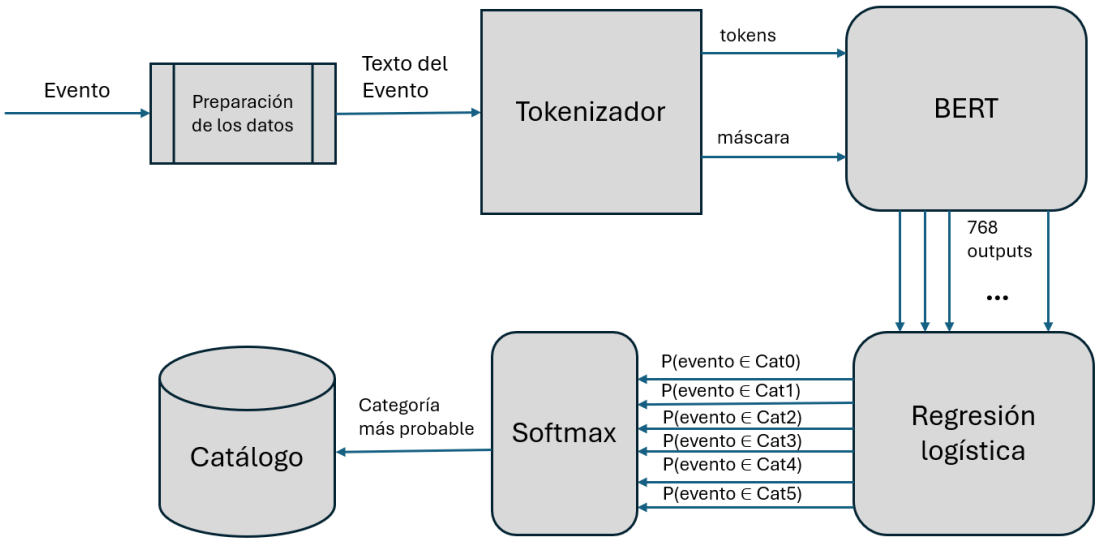

Figura 9: Flujo de datos de una secuencia de la fase de clasificación

Al final de esta fase disponemos de un valor categórico predicho para todos los eventos del conjunto de datos de test. Como también disponemos de la categorización predefinida de esos eventos podemos comparar cómo de bien ha clasificado el modelo, lo que llevaremos a cabo en la siguiente sección.

4.5. Evaluación

Es en esta fase de la metodología CRISP-DM cuando evaluamos el modelo, considerando sus resultados y contrastándolos con los objetivos propuestos. En esta sección vamos a evaluar diferentes aspectos del modelo planteando las siguientes secciones:

- Desempeño del modelo en la clasificación
- Robustez ante variaciones en los datos
- Eficiencia y escalabilidad
- Limitaciones y áreas de mejora
- Evaluación para nuestro caso de uso

4.5.1. Desempeño del modelo en la clasificación

El objetivo de esta sección es evaluar la capacidad del modelo propuesto para asignar correctamente las categorías taxonómicas.

En primer lugar, vamos a analizar la matriz de confusión tal y como se muestra en la Figura 10. Para construir la matriz de confusión sometemos las sentencias de nuestro conjunto de prueba al modelo, prediciendo la categoría de cada sentencia. Como conocemos las categorías originales, podemos comparar los resultados obtenidos en la predicción con la verdad base. En la Figura 9 puede verse un ejemplo de cómo se han clasificado algunas sentencias.

	predicción	valor verdadero	sentencia
0	5	5	Congreso Internacional El Triangulo de la Mujer. The International Congress celebrates its th Anniversary and will be celebrating it with a special event to be held on Saturday, July , from AM to P.M at the Sortis Hotel. This will be a conference where Panamanian warriors will meet in order to make known their concerns and opinions on different topics, such as fashion, skin care, cosmetics, exercises, among many other interesting topics. In this special anniversary edition, the Dominican psa...
1	1	1	December. December is a tragicomedy in one act. It takes place in a financial cave in downtown Buenos Aires as part of an economic debacle. The owner, Alfredo, keeps in captivity Pajarito, a kind of autistic who is a mathematical genius and can predict the value of the dollar. His secretary and lover, Selva, is his accomplice. However, she feels increasingly left out by Alfredo. The thirst for wealth, the obsession with social status and the family ghosts of the upper class collide with the ...
2	0	2	is a modern pop duo formed by Tuomas Kaitainen and Heikki Petrell. So far the future pop phenomenon have four hit singles and they are currently working on the debut album. is all about catchy pop melodies, electronic synth sounds and emotional lyrics. G Livelab . . at
3	3	3	Roskilde vs. Vejle BK. st Division Competition Date November Day Start
4	4	4	Burst My Bubble Tea Skip the long queues at bubble tea outlets and learn how to brew your very own cup of osmanthus honey bubble milk tea From preparing pearls to mixing ingredients, it will be a delicious hands on session where you can create your unique concoction
...	...	...	...
122842	3	3	HSBC UK Go Ride Scottish Cycling Cyclo Cross Training Cluster . The session is open to female riders born and male riders born and is designed for those who are already racing cyclocross, or who are planning to race this season for the first time. For those new to these sessions riders should expect a fun and intensive coaching session providing Technical and Tactical training opportunity. Working in small coaching groups with similar age participants we aim to help riders prepare for the de...
122843	3	3	th SAN SILVESTRE REQUENENSE. . Place of departure Avda. Arrabal. Distance . meters
122844	1	1	ERWAN QUESNEL DANS LA COMPLAINTE DU BIPO. One Man Show. Open, open the bird cage And come and sit in the head of Erwan Quesnel time for a show, just to understand the reasoning of a strange bird, the bipolarus schizophrenus . Specialist of the manic art, the author describes his psycho paranoid delusions in a single musical scene where the tragic mixes with the comic. Testimony, coming out, this show takes mental illness out of taboo and stigma. Music Laurent Zoppis. Staging Garoyan.
122845	4	0	Religious Festival in honor of the Holy Cross of de Motupe. Religious holiday, which highlights the pilgrimage for the transfer of the Cross from Zapotal to the chapel of Motupe where it is venerated.
122846	5	5	IX Municipal Conference of Culture. IX WORLD CULTURE CONFERENCE CULTURE AS A PUBLIC POLICY IN LONDON HISTORY AND POSSIBILITIES The Municipal Secretariat of Culture and the Municipal Council of Cultural Policy, announce the convening of the IX Conference of Culture of the City of Londrina which has as its theme CULTURE AS A PUBLIC POLICY IN LONDON HISTORY AND POSSIBILITIES . It will be held on September th and th, . For the discussion of the theme of the IX Conference of Culture, proposal rai...

Figura 10: Ejemplo de sentencias clasificadas con su valor predicho por el modelo y su verdad base

En el ejemplo, la sentencia con índice 1 se clasifica como clase 1, es decir, *artes escénicas*. La clase de la muestra original es 1 también, así que el modelo ha clasificado correctamente esta sentencia[3]. La sentencia 2 sin embargo está clasificada como *música*, mientras que la verdad base es *arte y cultura*. Esta sentencia está incorrectamente clasificada y representa un error de clasificación en nuestro modelo. Observando la sentencia vemos que está incompleta y probablemente se haya extraído de su fuente original incorrectamente. La sentencia habla de un grupo de pop al que no nombran, pero se da a entender que el evento es un acto relacionado con el grupo, no un concierto[4]. Extendiendo este análisis a las 122.847 sentencias en nuestro conjunto de test, podemos construir una matriz de confusión como la mostrada en la Figura 10.

[3] Por la curiosidad del lector, el texto alude a la obra de teatro Diciembre, de Alberto Maldonado, por lo que encaja perfectamente en la categoría de artes escénicas.

[4] El grupo musical al que hace alusión este evento turístico es *Kauriinmetsästäjät*, un nombre que se ha eliminado de la sentencia al contener caracteres prohibidos por nuestro modelo

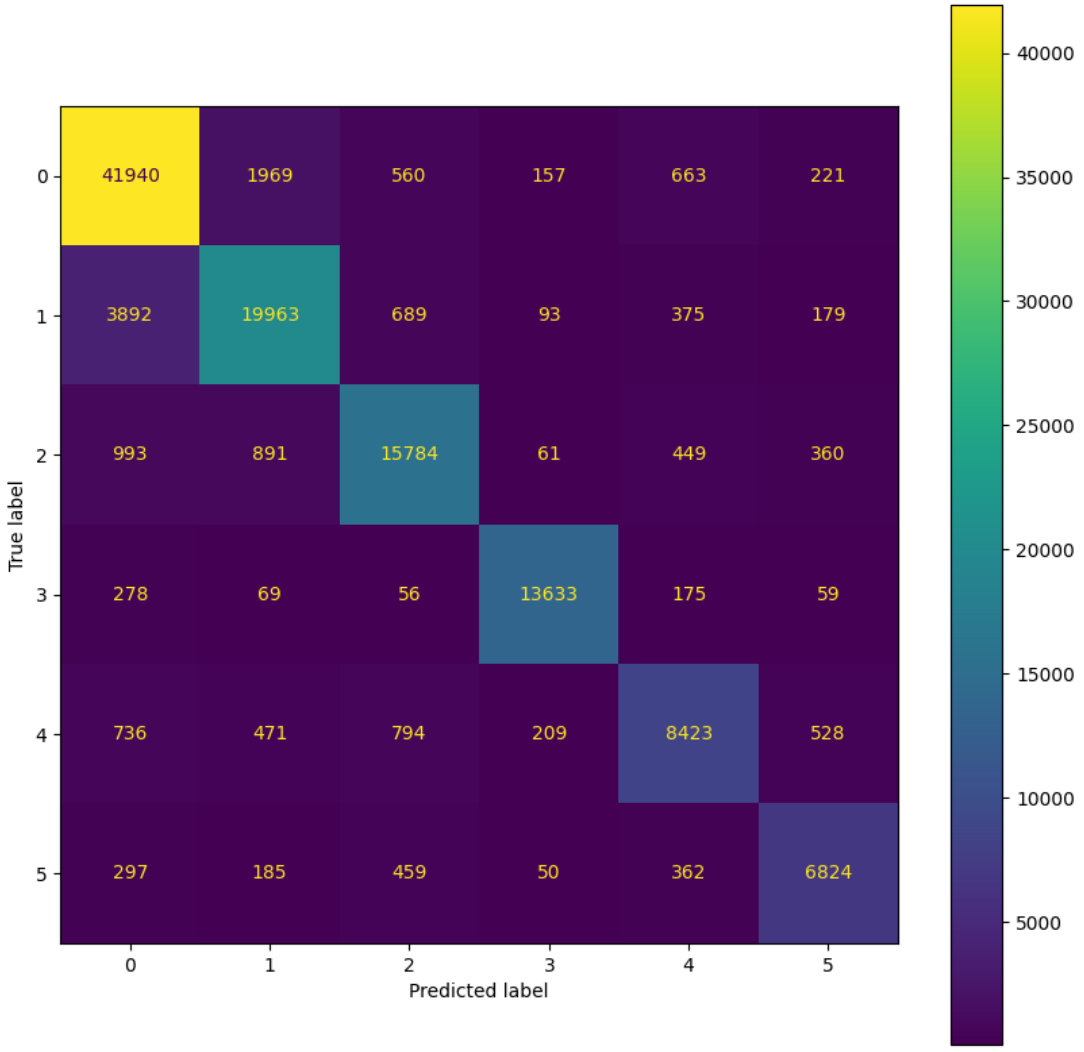

Figura 11: Matriz de confusión del conjunto de datos de prueba

Esta matriz de confusión nos cuenta que de las 45.510 sentencias que el conjunto de datos de prueba tenía de *música*, 41.940 se han clasificado correctamente como *música*, 1.969 se han clasificado erróneamente como *artes escénicas*, 560 se han clasificado erróneamente como *arte y cultura*, 157 se han clasificado erróneamente como *deporte*, 663 se han clasificado erróneamente como *otros eventos* y 221 se han clasificado erróneamente como *ferias y congresos*. Esto significa que para la clase *música*, se han clasificado correctamente un 92.15% de las sentencias por lo que podemos decir que la clase *música* tiene una sensibilidad del 92.15%. Podemos extender este análisis a las demás categorías y extraer una versión de la matriz de confusión donde en lugar del conteo de eventos clasificados por categorías, tengamos el porcentaje de eventos clasificados en cada categoría como se muestra en la Figura 11.

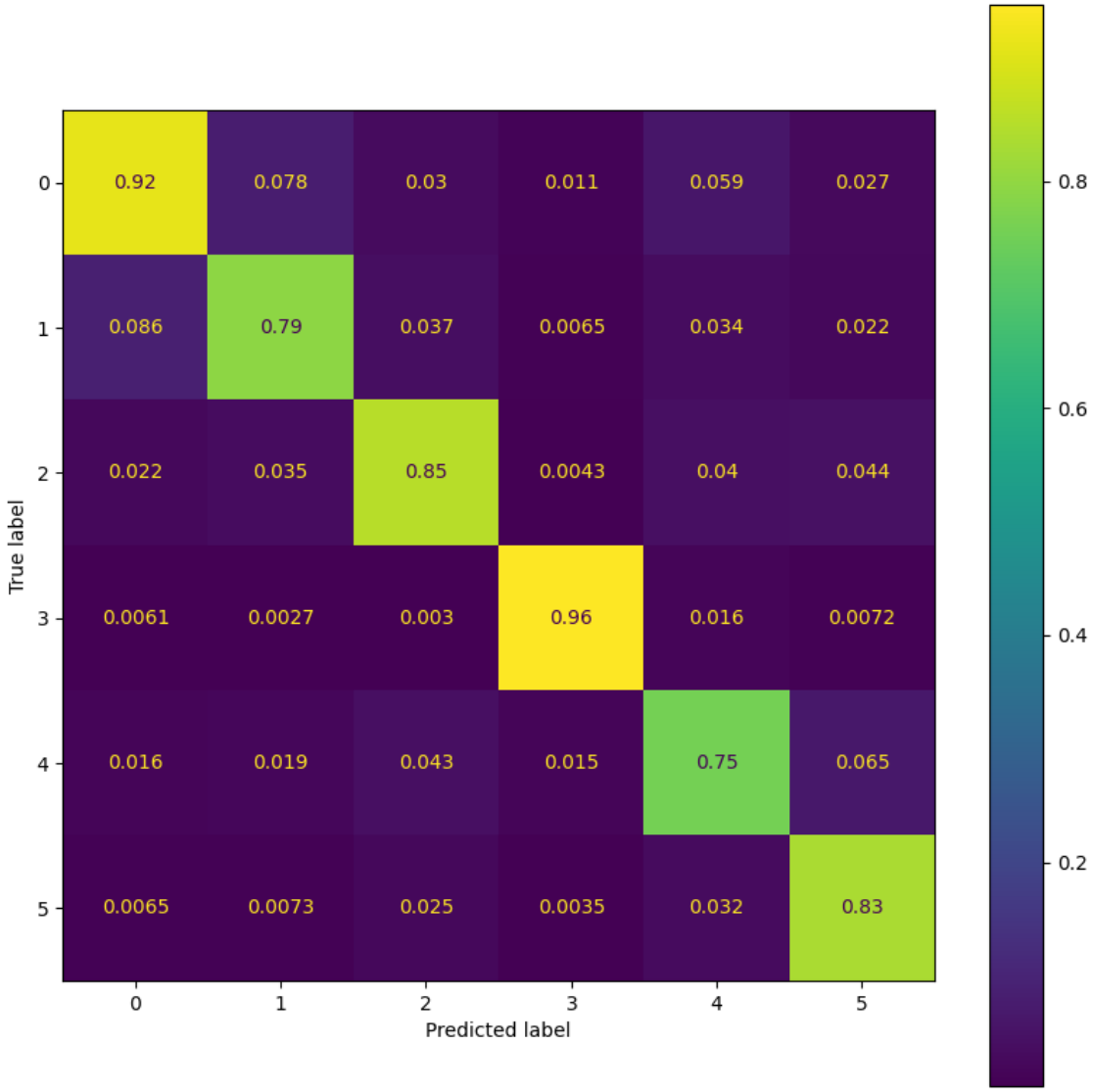

Figura 12: Matriz de confusión del conjunto de datos de prueba con ratios en lugar de conteos

Utilizando los datos de la matriz de confusión podemos generar un informe de clasificación como el descrito en la sección 2.2. El reporte se muestra en la Tabla 4.

Categoría	Precisión	Sensibilidad	F1-score	Soporte
0	0.87	0.92	0.9	45510
1	0.85	0.79	0.82	25191
2	0.86	0.85	0.86	18538
3	0.96	0.96	0.96	14270
4	0.81	0.75	0.78	11161
5	0.84	0.83	0.83	8177
Exactitud			0.87	122847
Media macro	0.86	0.85	0.86	122847
Media ponderada	0.87	0.87	0.87	122847

Tabla 4: Informe de clasificación del conjunto de datos de prueba

Este reporte de clasificación nos ofrece varias piezas de información. En primer lugar, nos da las precisiones, sensibilidades y métricas F1-score de todas las clases. Podemos observar que la clase con más precisión y sensibilidad, y por tanto *F1-score*, es la 3, es decir, *deporte*. Esto puede ser debido a que los títulos y descripciones de los eventos deportivos tienen un estilo y un vocabulario más consistente que en el resto de las categorías. También puede deberse a que la categoría *deporte* tiene menos solape con otras categorías. Como hemos visto en un ejemplo anterior, es fácil confundir *música* con *arte y cultura*, ya que hay un solape ahí. Sin embargo, es más difícil solapar eventos con *deporte*. Por otra parte, la categoría con menor precisión y sensibilidad (y por tanto F1-score) es la 4, es decir, *otros eventos*. Es muy razonable que la categoría más difícil de caracterizar por el modelo sea un cajón de sastre que puede encerrar eventos de muy diferente índole, desde eventos religiosos hasta turismo de naturaleza o de aventura. Por ejemplo, podemos observar el evento 122.845 de la Figura 9 cuyo título traducido es "Festival religioso en honor de la Sagrada Cruz de Motupe", considerado en el conjunto de datos como *música*, pero clasificado por el modelo como *otros eventos*, lo cual evidencia más una limitación en la taxonomía seleccionada que del modelo en sí.

Observamos cómo la categoría 1, es decir, *artes escénicas* es tras la de *otros eventos* la peor clasificada. Además, hay una diferencia notable entre la sensibilidad y la precisión. La precisión alta indica que de todas las predicciones de *arte y cultura* que ha hecho el modelo un 85% son acertadas, de *arte y cultura*. La sensibilidad baja indica que hay muchos eventos de *arte y cultura* que son etiquetados erróneamente como otras categorías. En la matriz de confusión podemos observar que hasta un 8.6% de los eventos de *arte y cultura* se confunden con los de la clase 0, *música*.

La categoría 5, *ferias y congresos*, tiene un F1-score de 0.83, un valor por debajo de la media. Probablemente hay factores positivos y negativos en la clasificación de esta clase. El hecho de que sea la que menor soporte tiene, solamente 8.177 muestras, puede indicar que el modelo no ha podido aprender todos los matices y patrones de este tipo de textos al no ser suficientes. Por otra parte, es muy posible que esta categoría tenga poco solape con otras y esté fraseada de una forma que permite al modelo clasificarla mejor.

Con todas las precisiones y sensibilidades de todas las clases podemos llegar a uno de los datos más relevantes del modelo: La exactitud. Nuestro modelo es capaz de categorizar correctamente un 87% de las sentencias de nuestro conjunto de datos de test. Esta exactitud es bastante alta y suficiente para utilizar un modelo de estas características en la industria y más si consideramos que un error del 13% es más que aceptable sin ayuda de ningún dato más que los estrictamente textuales con un conjunto de datos con errores e incompleto.

Otro detalle a tener en cuenta es el parecido entre la precisión y sensibilidad con el F1-score. Esto se debe a que, en media, la precisión y sensibilidad tienen valores parecidos, con lo que no acusamos fuertemente los efectos de un conjunto de datos desbalanceado en el que hay muchas más muestras en algunas clases, como se ve comparando la categoría *música* con 45.510 sentencias y la categoría *ferias y congresos* con 8.177 muestras.

4.5.2. Robustez ante variaciones en los datos

El propósito de esta sección es estudiar la capacidad del modelo para funcionar en condiciones adversas desde el punto de vista de los datos. En esta sección se va a realizar un pequeño análisis sobre cómo responde el modelo ante un empeoramiento de los datos utilizados.

El conjunto de datos usado para entrenar y validar el modelo es real, sin datos sintéticos por lo que arrastra una serie de problemas relacionados con la naturaleza ecléctica de las fuentes y problemas en el mecanismo de extracción de datos tal y como se describe en la sección 4.2. El resultado es que nuestro conjunto de datos es significativo en la industria, y otros conjuntos de datos generados por métodos similares tendrán problemas similares.

Los resultados expuestos han sido obtenidos con este conjunto de datos, por lo que el modelo ya tiene de forma intrínseca cierta robustez. Algunos ejemplos de datos mal formados se pueden ver en la Figura 12. En esta figura se ven todos los eventos con sentencias de solamente 2 caracteres que están clasificados incorrectamente. Es evidente que en estos casos no se puede inferir la categoría de un evento turístico solamente con los datos textuales obtenidos.

	0	1	2	3	4	5	predicción	valor verdadero	sentencia
11644	2.705065	2.717117	-1.669219	-1.725687	-2.456835	-3.636965	1	0	La
17625	2.705065	2.717117	-1.669219	-1.725687	-2.456835	-3.636965	1	0	La
34453	2.255208	2.459628	-1.701792	-1.841383	-2.273154	-2.546700	1	0	NA
36042	2.705065	2.717117	-1.669219	-1.725687	-2.456835	-3.636965	1	0	La
57604	2.705065	2.717117	-1.669219	-1.725687	-2.456835	-3.636965	1	0	La
65331	2.302232	0.647540	-1.402311	0.258222	-2.893871	-1.346876	0	1	Az
81631	2.705065	2.717117	-1.669219	-1.725687	-2.456835	-3.636965	1	0	La
104084	3.351726	1.148753	-1.057683	-2.172846	-2.430286	-2.688150	0	1	.
116608	1.446020	1.190131	-0.704095	-0.272177	-1.872320	-2.021883	0	1	of
118248	3.351726	1.148753	-1.057683	-2.172846	-2.430286	-2.688150	0	1	.
122477	2.662454	1.767454	-1.627249	-1.473963	-2.534469	-2.260213	0	1	Za

Figura 13: Todas las muestras con sentencias de 2 caracteres del conjunto de datos de test

Si generalizamos este razonamiento, podemos mostrar la relación entre la longitud de la sentencia en palabras y la exactitud de la clasificación. En la Figura 13 vemos cómo conforme sube el número de palabras y por tanto baja el soporte, la tasa de acierto se hace más errática.

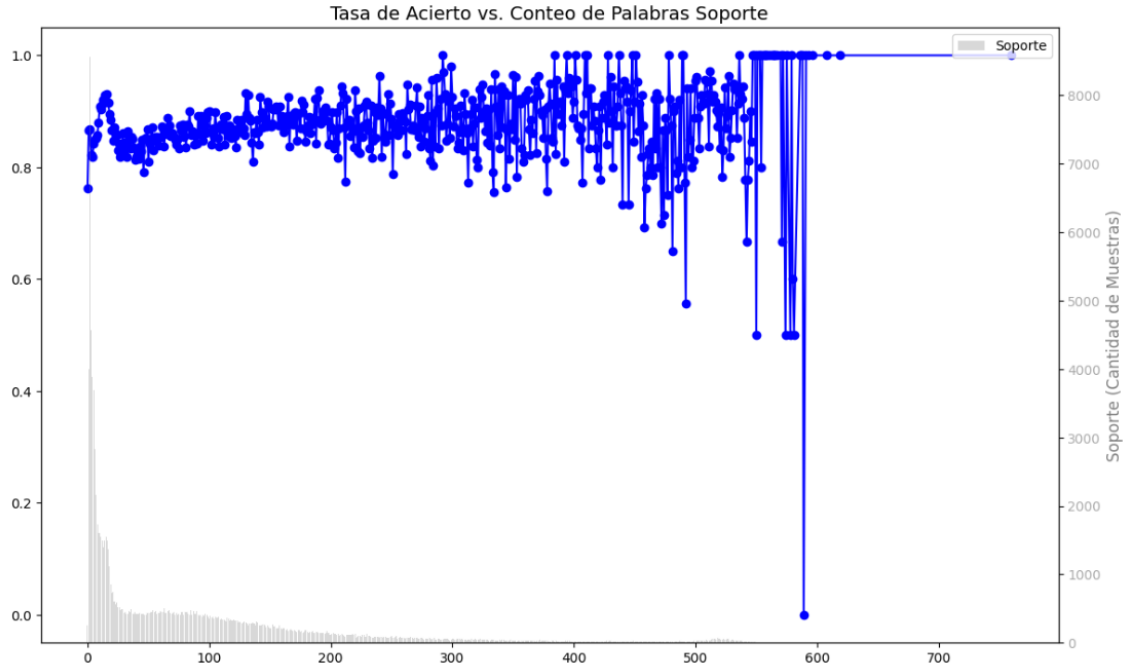

Figura 14: Tasa de acierto vs. Conteo de palabras vs. soporte

Como la mayoría de las muestras están concentradas en valores bajos de conteo de palabra, se muestra en la Figura 14 el detalle para un conteo de palabras bajo. Se puede observar que no necesariamente un menor número de palabras lleva asociado una peor predicción.

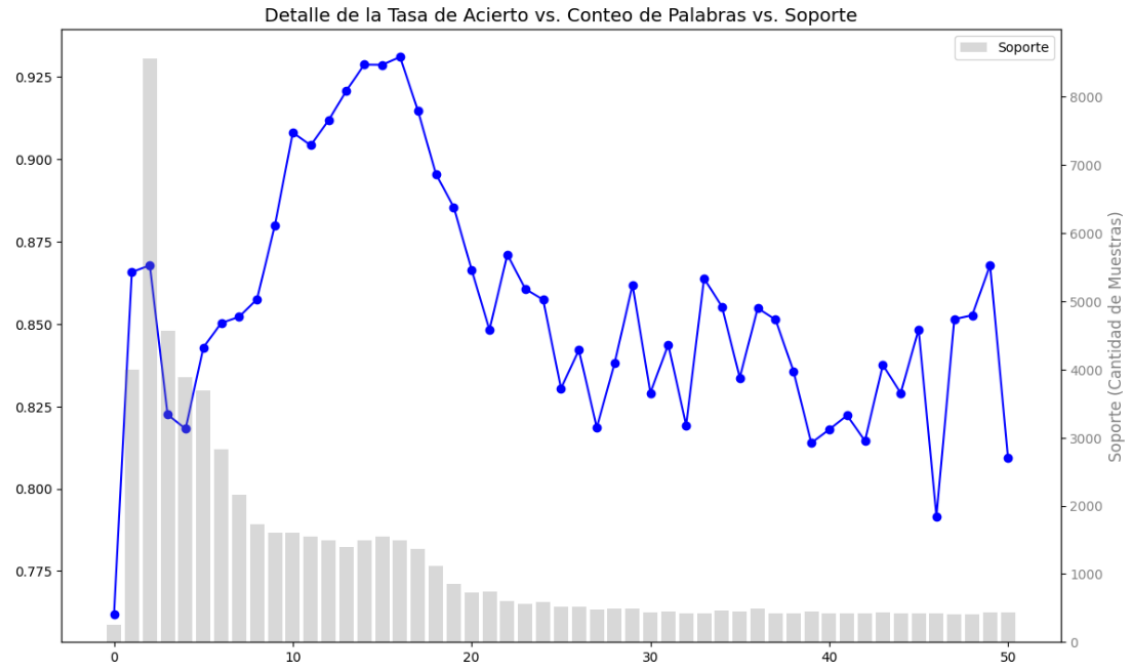

Figura 15: Detalle de la Tasa de acierto vs. Conteo de palabras vs. soporte

Aunque no necesariamente un menor número de palabras conlleva una peor predicción, esto solo aplica a textos autocontenidos que dan toda la información posible en sus títulos y descripciones. Si un evento solo necesita cuatro palabras para contar lo que quiere, como "Concierto Pavarotti en Ópera", un mayor número de palabras no habrían ayudado en la clasificación. Sin embargo, es lícito preguntarse qué ocurriría si el modelo recibe datos sesgados o cortados para clasificar, sentencias que no aportan todo el contexto porque están limitadas a un número de palabras. En la Tabla 5 puede verse el reporte de clasificación de nuestro modelo si en lugar de 510 palabras permitimos solamente un máximo de 20 palabras en cada evento: Las 20 primeras palabras de nuestra sentencia. El conjunto de prueba es el mismo analizado en la sección 4.5.1.

Categoría	Precisión	Sensibilidad	F1-score	Soporte
0	0.81	0.90	0.85	45510
1	0.77	0.69	0.73	25191
2	0.76	0.76	0.76	18538
3	0.95	0.94	0.94	14270
4	0.70	0.59	0.64	11161
5	0.72	0.72	0.72	8177
Exactitud			0.80	122847
Media macro	0.79	0.77	0.77	122847
Media ponderada	0.80	0.80	0.80	122847

Tabla 5: Reporte de clasificación del conjunto de datos de prueba para eventos limitados a 20 palabras

Como es razonable, al limitar el texto que describe cada evento a 20 palabras, la clasificación empeora. En este caso la exactitud ha bajado 7 puntos, desde el 0.87 al 0.8. Aun así, tomando solamente 20 palabras de eventos que pueden necesitar mucho más para explicarse o que pueden tener la información clave para la clasificación más allá de la palabra 20, el modelo acierta un 80% de las veces, lo que demuestra también cierta resiliencia a una posible degradación de los datos de entrada.

Por último, vamos a estudiar otra variación repitiendo el entrenamiento con sentencias recortadas. En este caso generamos un nuevo modelo que aprende de textos con una longitud máxima de 20 palabras y lo evaluamos con textos que tienen como máximo 20 palabras. Nótese que el reporte de clasificación de la Tabla 5 utiliza un modelo completo y se evalúa con textos recortados. Ahora vamos a utilizar un modelo recortado y a evaluarlo con datos recortados. El resultado es una exactitud de 0.83, algo mayor que en el caso anterior. Se observa así la importancia de mantener la consistencia entre la forma de los textos de entrenamiento y de inferencia y la resiliencia del modelo a fuertes cambios en los datos.

4.5.3. Eficiencia y escalabilidad

En esta sección vamos a examinar la capacidad del modelo para procesar grandes volúmenes de datos con medios y tiempos razonables. La clasificación se ha llevado a cabo en tres sistemas diferentes: Dos en la nube y uno físico. Esta elección responde a la necesidad de simular distintos escenarios de uso con diferentes capacidades computacionales y restricciones de recursos. Cómo sistema de computación en la nube se ha escogido *Kaggle* debido a que es gratuito y muy popular.

El primer sistema usa computación en la nube en la plataforma Kaggle. Kaggle (www.kaggle.com) es una conocida plataforma de ciencia de datos con un entorno de computación en la nube especializado en computar algoritmos basados en redes neuronales. Este sistema usa una *Intel Xeon 2.20 GHz* como unidad de procesamiento central y una *NVIDIA T4 x2* como unidad de procesamiento gráfico. En este sistema, logramos entrenar nuestro modelo basado en DistilBert con 368.538 eventos en 36.944 segundos, es decir, algo más de 10 horas. La inferencia del conjunto de datos de validación, tarea mucho menos compleja algorítmicamente que el entrenamiento, llevó al sistema 2.937 segundos.

El segundo sistema también utiliza computación en la nube en la plataforma Kaggle que usa una *Intel Xeon 2.20 GHz* como unidad de procesamiento central y una *Tesla P100* como unidad de procesamiento gráfico. En este escenario el entrenamiento se realiza en 18.607 segundos, es decir, algo más de 5 horas. La inferencia se llevó a cabo en 1.438 segundos.

El tercer sistema es un ordenador personal construido explícitamente para uso doméstico y computación de modelos profundos de PLN. Este ordenador dispone de una *Intel Core i9-13900K 3GHz* como unidad de procesamiento central y de una *NVIDIA RTX 4090* como unidad de procesamiento gráfico. En este escenario el entrenamiento se realiza en 3.444 segundos, es decir, casi una hora. La inferencia del conjunto de validación se llevó a cabo en 177 segundos.

En la Tabla 6 se puede ver el resumen de los sistemas utilizados junto con las métricas tiempo de entrenamiento (TE), tiempo de inferencia (TI), tiempo de entrenamiento por evento y tiempo de inferencia por evento.

Plataforma	CPU	GPU	Tiempo de entrenamiento	Tiempo de inferencia	TE / evento	TI / evento
Kaggle	*Intel Xeon 2.20 GHz*	*NVIDIA T4 x2*	36.944s	2.937s	100,244ms	23,9ms

Kaggle	*Intel Xeon 2.20 GHz*	*Tesla P100*	18.607s	1.438s	50,488ms	11,7ms
Ordenador personal	*Intel Core i9-13900K 3GHz*	*NVIDIA RTX 4090*	3.444s	177s	9,345ms	1,44ms

Tabla 6: Resumen de equipos utilizados en el entrenamiento y clasificación y sus tiempos de procesado

En el tiempo de escritura de esta tesis y desde hace años, Kaggle proporciona 30 horas por semana de uso de su plataforma con GPU de forma totalmente gratuita. También ofrece un uso sin GPU (solo CPU) de forma gratuita sin limitaciones temporales.

El conjunto de datos utilizado para este modelo contaba con 491.385 eventos turísticos y se ha entrenado y validado en menos de 30 horas en todos los escenarios probados, por lo que el coste de procesamiento es 0 con un amplio margen de incremento de datos. Además, el entrenamiento no es una tarea frecuente en la creación de catálogos de eventos, siendo razonable entrenarlo una vez y usarlo durante semanas o meses antes de volver a reentrenarlo. La inferencia es mucho más frecuente, pero también mucho más rápida por lo que se puede utilizar con cantidades enormes de datos antes de incurrir en coste alguno.

Se concluye esta sección observando que, en la práctica, la creación y utilización de este modelo no tiene coste alguno en plataformas como Kaggle, y se lleva a cabo en tiempos razonables de entrenamiento y en tiempos de inferencia compatibles con el tiempo real.

4.5.4. Limitaciones y áreas de mejora

En esta sección vamos a evidenciar algunos problemas observados durante la evaluación.

El primer problema observado es que hay eventos que no tienen la suficiente información para ser clasificados. Algunos casos son muy difíciles de solucionar, como aquellos en los que el texto sea vago o impreciso. Pero hay otros ejemplos que se solucionarían con un procesado de los datos más riguroso. Un ejemplo que destaca se muestra en la Figura 12. Estas sentencias tienen solamente dos caracteres, a veces solo signos de puntuación, claramente insuficiente para poder llevar a cabo una clasificación y sin duda perniciosas en la fase de entrenamiento del modelo. Una posible mejora podría ser el estudio y exclusión de patrones de texto que no contengan información sobre eventos turísticos como textos excesivamente cortos, compuestos en gran medida por caracteres no alfanuméricos, secuencias conocidas de error de servicios de consulta u otras casuísticas similares.

Se observa también que algunas secuencias pueden estar escritas en idiomas no soportados por el modelo utilizado de BERT y por tanto tendrán una baja probabilidad de ser clasificadas correctamente. Esto ocurre en casos marginales, ya que el conjunto de datos considerado está traducido al inglés y solo conserva en su idioma original nombres propios, errores de traducción o palabras no reconocidas por el traductor automático. Así, una posible mejora consiste en detectar el idioma de las palabras de la secuencia para ponerla en cuarentena o procesarla según se considere.

Otra observación es que muchos errores de clasificación vienen de una decisión de categoría muy disputada. Como puede observarse en la Figura 8, nuestro sistema devuelve una categoría de un array de probabilidades de categorías devueltas por la fase de regresión logística. Este array contiene los momios de las categorías: la categoría con el momio mayor es la más probable y la que se escoge como salida de nuestro sistema. Pero el sistema no nos da una medida de la seguridad que tiene en su decisión. En la Figura 15 se muestran algunos ejemplos de errores de clasificación y la usaremos para ilustrar clasificaciones en la que 2 momios sobresalen sobre los demás. Por ejemplo, en la segunda fila (con índice 25) se encuentra un evento cuyo mayor momio tiene un valor positivo y corresponde a la categoría *música*, su segundo mayor momio es positivo y corresponde a la categoría *artes escénicas*, y el resto de momios son negativos. Una interpretación de este array es que el sistema cree que el evento podría pertenecer a estas dos categorías, aunque con algo de mayor probabilidad en una de ellas, en este caso, la incorrecta. Esto también ocurre en la cuarta fila (índice 43) donde el sistema duda entre la categoría *artes escénicas* y *arte y cultura* con una diferencia en sus momios ínfima. Una posible mejora sería dotar al sistema de una salida que nos diera información sobre la seguridad del sistema en su decisión. De esta forma se podrían tomar mejores decisiones a largo plazo, poner en cuarentena esas decisiones o incluso preparar el sistema para una salida en la que un evento turístico pueda pertenecer a varias categorías a la vez.

	0	1	2	3	4	5	predicción	valor verdadero	sentencia	acierto	conteo_palabras
2	3.413687	2.035024	-0.211703	-3.599040	-3.118200	-2.936889	0	2	is a modern pop duo formed by Tuomas Kaitainen...	False	48
25	2.111954	1.868197	-1.873009	-0.157912	-2.504789	-3.516103	0	1	Bianca Del Rio. BALONEY HERE COMES BIANCA Not ...	False	146
35	1.054191	0.077457	2.895201	-2.612906	-0.652218	-2.761691	2	1	Reading for discouraged.	False	3
43	-3.123243	3.434931	3.500238	-3.810414	-1.167054	-2.359210	2	1	Arde Madrid Spanish Film Festival. The Spanish...	False	251
55	1.078479	3.443372	-2.607996	-1.480204	-2.235836	-3.327745	1	0	David Murray with Saul Williams. David Murray ...	False	10
...	...	...	...	...	...	...	...	...	...	...	...
122795	-2.473790	-1.166344	3.035485	-2.952158	0.107251	2.176878	2	5	Philosophy in the city. Women philosophers Wom...	False	35
122823	3.935503	0.423382	-2.078794	-0.186444	-2.539426	-3.908916	0	3	Struggle Dales.	False	2
122830	1.222371	0.719324	3.503039	-3.683103	-1.938773	-2.176426	2	1	CINEMA ROCK THE BEATLES. Rock Cinema presents ...	False	70
122832	-0.236203	1.552573	1.394564	-2.535288	-1.288704	-0.773592	1	2	Pompeu Fabra A Complete Language.	False	5
122845	1.296451	-2.549654	-1.067338	-3.274778	3.980515	-2.011300	4	0	Religious Festival in honor of the Holy Cross ...	False	34

Figura 16: Errores de clasificación con los momios de cada categoría

En la línea de razonamiento del párrafo anterior, otra posible mejora consiste en permitir al sistema clasificar un evento turístico en más de una categoría a la vez. Esto parece razonable ya que las categorías tienen cierto solape y algunos espectáculos pueden pertenecer a *música* y también a *artes escénicas*, por ejemplo. Una dificultad de esta mejora radica en la dificultad para medir el éxito del algoritmo: nuestro conjunto de datos de entrenamiento y validación tienen eventos que pertenecen a una sola categoría y por tanto es difícil realizar una comparación de resultados.

Otra mejora consistiría en seleccionar un sistema taxonómico menos ambiguo. El problema es que una taxonomía es empírica (a diferencia de una tipología) y se adapta a los eventos existentes con sus datos asociados, en este caso, las categorías de los eventos ya clasificados. Un cambio de sistema taxonómico nos obligaría a recategorizar cientos de miles de eventos, una tarea titánica.

Un último problema observado es el desbalanceo de las categorías. Como puede observarse en la Tabla 4, la categoría *música* tiene 45.510 eventos, más de cinco veces más que la categoría *ferias y congresos*, con 8.177 eventos. El desbalanceo supone un problema para los clasificadores como los regresores logísticos. Una posible mejora consistiría en adquirir más eventos de las clases menos representadas para obtener un conjunto de datos más uniforme.

4.5.5. Evaluación para nuestro caso de uso

En nuestro ejemplo, para medir la eficacia del clasificador, la aerolínea compara los resultados obtenidos por el modelo con un conjunto de eventos ya validados por expertos, verificando la concordancia entre categorías predichas y reales. No obstante, desde la óptica de marketing, la evaluación no se limita a la exactitud del modelo, sino que también se miden indicadores de desempeño en las posibles campañas promocionales:

- Tasa de apertura de correos: ¿Se incrementa cuando se muestran eventos clasificados bajo intereses detectados en la base de datos de clientes?
- Tasa de clics: (CTR por el término en inglés *Click-Through Rate*): ¿La clasificación precisa de eventos fomenta más clics en anuncios específicos sobre “conciertos de verano” o “ferias tecnológicas”?
- Conversión o reserva efectiva: ¿Cuántos usuarios terminan comprando el vuelo y/o las entradas al evento tras interactuar con la información segmentada?

Si las métricas de marketing reflejan mejoras significativas (por ejemplo, un CTR más elevado o un mayor promedio de billetes vendidos) se confirma la validez del modelo para estimular la venta de pasajes. Adicionalmente, el equipo de marketing revisa la

robustez de la clasificación ante descripciones poco convencionales o ante la introducción de nuevos géneros de eventos. Si el rendimiento se ve afectado, puede que sea necesario incorporar más muestras al conjunto de entrenamiento o desarrollar subcategorías específicas para capturar la riqueza de la oferta cultural de cada destino.

5. Conclusiones y trabajo futuro

Comenzaba este trabajo con varias preguntas fundamentales: ¿Cómo impactan los eventos turísticos al sector turístico desde el punto de vista del marketing? ¿Es valioso clasificar eventos turísticos para crear catálogos de eventos? ¿Qué podemos esperar en términos económicos, de tiempo y esfuerzo de un proceso automático de clasificación de eventos?

Las investigaciones presentadas a lo largo de esta tesis responden a la necesidad de contar con un proceso estandarizado y escalable para la clasificación automática de eventos turísticos, con un claro enfoque en la generación de valor en marketing. El objetivo se ha centrado, por un lado, en diseñar y validar una metodología basada en la combinación de CRISP-DM, técnicas de Big Data y algoritmos de PLN, y por otro, en demostrar cómo esta solución puede aplicarse de manera efectiva en un contexto de negocio real, utilizando como ejemplo el caso de una aerolínea que desea inspirar y atraer a sus potenciales clientes a través de un catálogo homogéneo de eventos.

A lo largo de la tesis, se han propuesto hipótesis relacionadas con la importancia de la estandarización de la clasificación de eventos, el impacto positivo en la promoción y segmentación de mercado y la viabilidad técnica de un sistema que funcione en múltiples idiomas y fuentes heterogéneas. En este capítulo se exponen las conclusiones generales y se sugieren diversas líneas de investigación y desarrollo que pueden profundizar y mejorar los resultados obtenidos.

5.1. Conclusiones generales

La primera pregunta, cómo impactan los eventos turísticos al sector turístico desde el punto de vista del marketing, ya está contestada en la literatura presentada, concluyendo que los eventos turísticos tienen un impacto muy positivo en el sector turístico.

Hay extensa literatura que confirma que los eventos turísticos representan un reclamo de marketing esencial para los destinos y las empresas turísticas. Su capacidad para generar experiencias memorables y fortalecer la imagen de los lugares anfitriones convierte a los eventos turísticos en un factor clave a la hora de posicionar y diferenciar la oferta en un mercado altamente competitivo. Además, el evento turístico puede integrarse en estrategias más amplias de comunicación que ayudan a segmentar mejor las audiencias y a superar la estacionalidad, ya que posibilitan la creación de campañas dirigidas a públicos específicos y la programación de eventos a lo largo de todo el año.

Por otro lado, el uso de un modelo de clasificación automático aporta orden y coherencia a la enorme diversidad de formatos y temáticas existentes, permitiendo unificar criterios y potenciar la eficiencia en la promoción cruzada. En consecuencia, destinos y proveedores como aerolíneas, hoteles o agencias, pueden concentrarse en diseñar y difundir ofertas más enfocadas, incrementando tanto el deseo de viajar del turista como la rentabilidad de sus acciones de marketing.

Así, este trabajo destaca la relevancia de los eventos como herramienta de branding, segmentación y fidelización, y cómo su normalización y clasificación contribuye de manera decisiva a potenciar el impacto positivo en la comercialización y desarrollo turístico.

Por otro lado, nos hacíamos la pregunta ¿Es valioso clasificar eventos turísticos para crear catálogos de eventos? En este trabajo se ha puesto de manifiesto que la clasificación automatizada de eventos aporta un gran valor a la industria turística al permitir la creación de catálogos unificados y coherentes. Dichos catálogos facilitan la gestión de grandes volúmenes de datos, agilizan la localización y la comparación de la oferta, y hacen posible el diseño de estrategias promocionales más precisas. Además, las soluciones basadas en taxonomías comunes ayudan a reducir la heterogeneidad que presentan las distintas fuentes y formatos de publicación, mejorando la experiencia del usuario, sea turista o agente de la industria al disponer de un sistema de categorías consistente.

¿Qué podemos esperar en términos económicos, de tiempo y esfuerzo de un proceso automático de clasificación de eventos turísticos? Para esta pregunta no se ha encontrado respuesta en la literatura y se ha diseñado y ejecutado un modelo para arrojar algunos resultados que expliquen el esfuerzo necesario para construir un clasificador de eventos y los resultados que podemos esperar de él. El objetivo de negocio de nuestro modelo era claro: Crear catálogos de eventos turísticos de forma estandarizada, eficiente y a bajo coste.

Desarrollamos a continuación cómo los objetivos específicos conectan con los resultados obtenidos al evaluar el modelo clasificador de eventos turísticos:

- Demostrar la viabilidad del modelo: El éxito del modelo propuesto demuestra su viabilidad al utilizar como única fuente de datos el nombre y la descripción del evento. La experiencia empírica evidencia que la IA, concretamente las técnicas de PLN, logra categorizar de forma consistente una alta diversidad de eventos, incluso en casos donde el texto suministrado es escaso o de calidad desigual. De esta manera, se confirma que la automatización de la clasificación es posible sin recurrir a campos adicionales (ubicación, fecha o información

multiestructurada), reduciendo notablemente las barreras de implementación y mejorando la accesibilidad de la solución.

- Establecer un enfoque universal: La adopción de un enfoque universal se ve reflejada en el hecho de que el modelo no depende de datos específicos más allá del texto libre, lo cual permite extenderlo prácticamente a cualquier región, idioma o sector del turismo. Este carácter amplio asegura que la creación de catálogos normalizados pueda llevarse a cabo con un esfuerzo mínimo de adaptación, favoreciendo tanto a actores con infraestructuras tecnológicas avanzadas como a aquellos que recién inician la digitalización de su oferta.
- Medir la eficiencia en tareas de clasificación: Los resultados muestran que el sistema alcanza tiempos de procesamiento razonables y mantiene un rendimiento estable conforme se incrementa el volumen de datos. En concreto se logran tiempos de inferencia de 1.44s por cada 1000 eventos utilizando un ordenador doméstico y tiempos de 11.7s por cada 1000 eventos en una plataforma en la nube gratuita. Esta capacidad para clasificar grandes cantidades de eventos de manera ágil supone un importante ahorro de recursos humanos y económicos. Además, la escalabilidad del enfoque es un factor determinante para sectores que manejan millones de registros, como aerolíneas u operadores turísticos con alcance global.
- Medir la calidad de la clasificación: Las métricas de precisión, sensibilidad y F1-score reflejan índices suficientemente altos para sostener la fiabilidad de la solución en la práctica empresarial. En concreto, se logra un F1-score, una precisión y una sensibilidad de 0.87, lo que indica que acertamos en un 87% de los eventos mientras apenas nos afecta el desbalanceo de clases. El análisis de categorías con mayor ambigüedad o solapamiento constata que el modelo se adapta correctamente a variaciones en la naturaleza del texto, respondiendo de forma robusta ante casos límites. Con ello, se afianza la idea de que, a pesar de la heterogeneidad de los eventos turísticos, el clasificador puede mantener un nivel de exactitud que satisfaga las exigencias comerciales y operativas de los usuarios finales.
- Probar la robustez del modelo: La robustez del modelo queda probada al someterlo a condiciones adversas, como la ausencia de descripciones detalladas, problemas de codificación o el uso de idiomas distintos al predominante en el conjunto de entrenamiento. Lejos de degradarse por completo, el sistema logró resultados aceptables y, en muchos casos, cercanos a su rendimiento óptimo. Esto demuestra que la solución no solo es eficiente en situaciones normales, sino que también es capaz de asimilar escenarios reales, donde la calidad y consistencia de los datos rara vez son ideales.

- Sentar las bases para la normalización taxonómica: El modelo contribuye a la construcción de catálogos de eventos estandarizados que agilizan la comparación, intercambio y análisis de datos entre diferentes actores de la industria. Esta propuesta de taxonomía, acompañada por la automatización de su aplicación, constituye un paso relevante hacia la convergencia de criterios en el sector turístico. Así, destinos turísticos, plataformas de reservas y proveedores de servicios disponen de un lenguaje común para etiquetar y describir su oferta, con las consecuentes mejoras en visibilidad y promoción. Sin embargo, la taxonomía utilizada en nuestro modelo tiene algunos problemas que desarrollaremos en la sección de limitaciones y trabajo futuro.
- Evaluar la aplicabilidad en el contexto empresarial: Se evidencia que la estrategia aquí planteada reduce significativamente los costes tradicionales ligados al clasificado manual. Además, abre oportunidades de monetización y optimización de la oferta: por un lado, facilita el diseño de productos orientados a nichos o segmentos muy concretos; por otro, posibilita la detección de vacíos u oportunidades en la programación de eventos. Para aerolíneas, agencias y OTAs, esto se traduce en un notable impulso para la diferenciación y el marketing personalizado, sumamente valorado en un mercado tan competitivo como el turístico.

Esta tesis representa una contribución significativa al estado del arte en la clasificación de eventos turísticos, abordando tres limitaciones detectadas en la literatura existente. En primer lugar, desde el punto de vista metodológico, numerosos trabajos carecen de una estructura formal o de una metodología replicable para la categorización de eventos, lo que limita su aplicabilidad y rigor científico. En segundo lugar, aunque algunos estudios han propuesto tipologías o sistemas de clasificación, no se identifican enfoques que permitan su implementación de forma automática, lo cual reduce su valor práctico en contextos reales de gran escala. En tercer lugar, las carencias metodológicas y de automatismos señaladas hacen que las aplicaciones encontradas en la literatura suelen estar circunscritas a ámbitos muy específicos o geográficamente limitados, sin una validación empírica amplia. Frente a estas limitaciones, el presente trabajo propone un modelo automatizado, basado en aprendizaje automático y procesamiento del lenguaje natural, que permite clasificar eventos turísticos de forma eficiente y escalable. Además, esta solución se valida a través de un caso de uso realista en el sector aéreo, lo que refuerza su aplicabilidad comercial y su potencial para integrarse en entornos digitales complejos.

5.2. Limitaciones

Pese a los resultados favorables demostrados sobre la eficacia de un modelo de clasificación automático para la creación de catálogos de eventos turísticos y su

contribución a la promoción de destinos, hay que señalar algunas limitaciones inherentes al trabajo realizado. Estas limitaciones delinean oportunidades de mejora y líneas futuras de desarrollo tanto en el ámbito metodológico como en el de la aplicación práctica.

La primera limitación es la dependencia de la calidad de los datos textuales. Aunque el enfoque se centra en el uso exclusivo del nombre y la descripción de cada evento, existen contextos en los que la información disponible es mínima o de baja calidad. Es frecuente hallar descripciones incompletas, ambiguas o escritas en varios idiomas sin la debida normalización. Esta realidad limita la precisión del modelo y, en consecuencia, la posibilidad de explotar el potencial de marketing que ofrecen los catálogos de eventos. Si bien se ha demostrado que el clasificador logra resultados notables incluso con datos adversos, la robustez podría verse afectada cuando se combinan deficiencias en la calidad de la descripción con la falta de traducciones fiables.

Una limitación importante de este trabajo es la taxonomía utilizada. El proceso de creación de catálogos depende en gran medida de la definición y estabilidad de la taxonomía de eventos. Sin embargo, el mercado turístico evoluciona rápidamente: surgen nuevas tendencias, formatos híbridos, categorías o subcategorías que no se contemplan en la taxonomía actual. Para mantener la relevancia de la propuesta, sería necesario reentrenar o ajustar el modelo cuando el listado de categorías sufra modificaciones, lo cual puede implicar costes adicionales y procesos iterativos de validación.

La taxonomía utilizada en conjunto con el modelo utilizado puede limitar la multidimensionalidad de un evento. Un evento puede asociarse simultáneamente a más de una categoría (por ejemplo, un concierto infantil). El modelo ha priorizado la asignación de una única etiqueta principal, una aproximación que reduce complejidad, pero puede perder matices relevantes para fines promocionales. En el contexto del marketing, donde la segmentación precisa es una de las claves para captar la atención del viajero, esta limitación puede conducir a soluciones parciales o menos efectivas si no se contemplan futuros desarrollos que permitan la asignación de eventos a varias categorías o se seleccionan taxonomías que separen mejor diferentes tipos de eventos turísticos.

Por último, a lo largo de este trabajo, se ha comprobado la utilidad del sistema de clasificación en la inspiración al viaje y en la segmentación del público objetivo. No obstante, no se han evaluado con detalle indicadores económicos que permitan estimar el retorno de la inversión (ROI) o la influencia de la clasificación en KPIs específicos de marketing (como el coste de adquisición de cliente, la tasa de conversión o la fidelización medida en reservas repetidas). La ausencia de este

análisis empírico deja espacio para mejorar la comprensión de cómo se materializa el impacto de la herramienta en la cuenta de resultados de las organizaciones turísticas.

5.3. Trabajo futuro

El siguiente paso de este trabajo sería extender el modelo a más piezas de información del evento turístico. Dado que muchas descripciones suelen ser breves o incompletas, se sugiere incorporar recursos como imágenes, vídeos, fuentes o metadatos geográficos que ofrezcan una visión más amplia de cada actividad. Con ello, el sistema de clasificación puede refinar su comprensión de la naturaleza de los eventos y ampliar el nivel de detalle de la taxonomía, de modo que se atiendan segmentos muy específicos de la audiencia o para enriquecer la calidad de los catálogos en general. Alternativamente, se pueden incluir procesos de enriquecimiento automático del texto, como la generación automática de descripciones a partir de imágenes, videos u otros formatos, integrando esa información en la entrada del modelo sin necesidad de modificar su arquitectura.

Otra mejora directa surge de la posibilidad de asignar varias categorías de forma simultánea. Muchos eventos combinan distintos atractivos, como espectáculos musicales y experiencias gastronómicas, o un componente virtual junto a otro presencial. Bajo este enfoque, la clasificación no quedaría restringida a una sola etiqueta, sino que capturaría la esencia híbrida de los actos. Como consecuencia, el público potencial podría segmentarse con mayor precisión, beneficiándose así las estrategias de promoción y venta relacionadas con el evento turístico.

La evolución constante del sector turístico motiva una tercera posible mejora: la renovación frecuente de la taxonomía de eventos. Para responder con agilidad a la aparición de formatos o tendencias emergentes, resulta aconsejable contar con un mecanismo que facilite la adaptación del sistema. En lugar de entrenar un modelo desde cero cada vez que surge una nueva categoría, podría desarrollarse un proceso incremental o activo que incorpore los cambios de manera fluida. Mantener una taxonomía actualizada de esta forma garantizaría la pertinencia de los catálogos.

Otra continuación directa de este trabajo es estudiar cómo la clasificación de eventos se refleja en indicadores de negocio. Sería valioso medir variables como el retorno de la inversión publicitaria o la tasa de conversión en reservas, a fin de comprobar la rentabilidad real de integrar esta herramienta en los procesos de promoción.

Otra mejora, y según se desprende de los resultados, menos importante, es la necesidad de escalar la infraestructura para manejar grandes volúmenes de datos en plazos muy cortos. Este punto implica evaluar opciones de computación distribuida u optimizaciones de hardware, a fin de que el modelo responda bien ante temporadas con elevada actividad turística. Es igualmente relevante llevar a cabo un seguimiento

continuado de los resultados, de manera que se identifique su comportamiento ante ciclos estacionales y posibles fluctuaciones en la disponibilidad de eventos.

Por último, un mayor nivel de colaboración entre destinos, proveedores y otros actores podría conducir a la creación de espacios de datos y criterios compartidos, mejorando así la competitividad y la eficiencia en el sector a escala global.

6. Anexos

Publicaciones en revistas

Tourism destination events classifier based on artificial intelligence techniques, 2023

Título: Tourism destination events classifier based on artificial intelligence techniques

Autores: Camacho-Ruiz, M., Carrasco, R. A., Fernández-Avilés, G., & LaTorre, A.

Año de publicación: 2023

DOI: 10.1016/J.ASOC.2023.110914

Link persistente: https://hdl.handle.net/20.500.14352/114236

Datos de la publicación

Título de la revista: Applied Soft Computing

Editorial: Elsevier Ltd

ISSN / ISBN: 1568-4946

Fecha: 2023-11-01

Impacto científico de la publicacion

- **Journal Impact Factor – JIF (JCR)**
  Año: 2023
  Factor de impacto de la revista: 7.2
  Factor de impacto sin autocitas: 6.6
  Cuartil mayor: Q1
  - *Área:* COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS *Cuartil:* Q1 (Decil 1) *Posición en el área:* 16/170 (SCIE)
  - *Área:* COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE *Cuartil:* Q1 *Posición en el área:* 27/197 (SCIE)
- **Journal Citation Indicator – JCI (JCR)**

  Año: 2023
  JCI de la revista: 1.48
  Cuartil mayor: Q1

- *Área:* COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS *Cuartil:* Q1 *Posición en el área:* 28/170
- *Área:* COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE *Cuartil:* Q1 *Posición en el área:* 30/198

- **Article Influence Score (JCR)**

  Año: 2023
  Article influence score: 1.282

- **SCImago Journal Rank**

  Año: 2023
  Impacto SJR de la revista: 1.843
  Cuartil mayor: Q1
  - *Área:* Software *Cuartil:* Q1 (Decil 1) *Posición en el área:* 46/497

- **Scopus CiteScore**

  Año: 2023
  CiteScore de la revista: 15.8
  - *Área:* Software *Posición:* 25 *Percentil:* 93

Citas

Fuente	Número de Citas	Fecha
Dimensions	**8**	**10-07-2025**
Scopus	**5**	**10-07-2025**
Web of Science	**2**	**09-05-2025**
Google Académico	**9**	**10-07-2025**

Citas normalizadas

- Dimensions: **Field Citation Ratio (FCR):** 5.1 (Compared to other publications in the same field, **this publication is highly cited** and has received approximately **5.09 times more citations** than average.)

Impacto social

Research Interest Score: 6.3 (This item's Research Interest Score is higher than 84% of research items published in 2023.)

Contribución a la ciencia abierta

Publicación en revista diamante

Presencia en:

- Directory of Open Access Journals (DOAJ)
- Open Policy Finder (Sherpa Romeo)

Applied Soft Computing Journal 140 (2023) 110914

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Tourism destination events classifier based on artificial intelligence techniques

Miguel Camacho-Ruiz [a], Ramón Alberto Carrasco [b], Gema Fernández-Avilés [c,*], Antonio LaTorre [d]

[a] Faculty of Commerce and Tourism Complutense, University of Madrid, 28223 Madrid, Spain
[b] Department of Marketing, Faculty of Statistics, Complutense, University of Madrid, 28040 Madrid, Spain
[c] Faculty of Legal and Social Sciences, Toledo, University of Castilla—La Mancha, Cobertizo de San Pedro Mártir, S/N. 45.071, Toledo, Spain
[d] Center for Computational Simulation, Universidad Politécnica de Madrid, 28660 Madrid, Spain

HIGHLIGHTS

- Computational techniques are used to classify tourism destination events.
- A Large Language Model (BERT) is used to get vectorial representations of events.
- A method to automatically classify events is proposed to easy the adoption of standards.
- There is great scope for extending this methodology to other applications.

GRAPHICAL ABSTRACT

ARTICLE INFO

Keywords:
Tourist destinations
Tourist events
Classification
CRISP-DM
Artificial intelligence

ABSTRACT

Identifying client needs to provide optimal services is crucial in tourist destination management. The events held in tourist destinations may help to meet those needs and thus contribute to tourist satisfaction. As with product management, the creation of hierarchical catalogs to classify those events can aid event management. The events that can be found on the internet are listed in dispersed, heterogeneous sources, which makes direct classification a difficult, time-consuming task. The main aim of this work is to create a novel process for automatically classifying an eclectic variety of tourist events using a hierarchical taxonomy, which can be applied to support tourist destination management. Leveraging data science methods such as CRISP-DM, supervised machine learning, and natural language processing techniques, the automatic classification process proposed here allows the creation of a normalized catalog across very different geographical regions. Therefore, we can build catalogs with consistent filters, allowing users to find events regardless of the event categories assigned at source, if any. This is very valuable for companies that offer this kind of information across multiple regions, such as airlines, travel

Congresos

Título del trabajo: Tourism destination events classifier based on AI techniques

Nombre del congreso: TMS 2022: Sustainability Challenges in Tourism, Hospitality and Management –

Tourism & Management Studies International Conferenc

Tipo evento: Congreso Internacional

Tipo de participación: Participativo - Ponencia oral (comunicación oral)

Ciudad de celebración: Olhao,Portugal,

Fecha de celebración: 16/11/2022

Fecha de finalización: 19/11/2022

Entidad organizadora: Universidade do Algarbe

Publicación: Proceedings of the TMS 2022: Sustainability Challenges in Tourism, Hospitality and Management – Tourism & Management Studies

International Conference. pp. 56 - 56.

ISBN: 978-989-9127-13-5

TMS ALGARVE 2022: Sustainability Challenges in Tourism, Hospitality and Management – Tourism & Management Studies International Conference

16 - 19 November - Olhão, Portugal

PROGRAMME AND ABSTRACTS

José António C. Santos, Margarida Custódio Santos, Alexandra Rodrigues Gonçalves and Miguel Ángel Solano-Sánchez (eds.)

Campus da Penha, Estrada da Penha
8005-139 Faro
PORTUGAL

ISSN: 978-989-9127-13-5

DOI: https://doi.org/10.34623/eryw-0423

Conference website: http://www.esght.ualg.pt/tms2022/index.php/tms2022/TMS2022
E-mail contact: mmsantos@ualg.pt

Tourism destination events classifier based on Artificial Intelligence techniques

Miguel **Camacho**
University of Madrid, mcamacho@atalayatech.com

Ramón Alberto **Carrasco**
University of Madrid, ramoncar@ucm.es

Gema Fernández-Avilés **Calderón**
University of Castilla-La Mancha, gema.FAviles@uclm.es

Antonio **LaTorre**
Universidad Politécnica de Madrid, a.latorre@upm.es

Identifying client needs to provide optimal services is crucial in touristic destination management. There are touristic events happening in touristic destinations that may cover those needs and hence, help satisfy tourists. Similar to what happens with product management, the creation of hierarchical catalogs to classify those events would help event management. Those events that can be found on the internet are described in desegregated and heterogeneous sources, which make direct classification a hard and time consuming task. The goal of this work is to create a process that automatically classifies touristic events of eclectic nature given a hierarchical taxonomy to help touristic destination management. This automatic taxonomization process allows the creation of a normalized catalog across very different geographical regions. Therefore, we can build catalogs with consistent filters to find events regardless of the source taxonomies if any. This is very valuable for companies that offer this kind of information across multiple regions such as airlines, OTAs or hotel chains and hence, valuable for the final user. Using a Data Science methodology such CRISP-DM, supervised automatic learning and natural language processing techniques, this work describes how we reached this goal using hundreds of thousands of events.

Keywords: Touristic destinations, touristic events, classification, CRISP-DM. [ID 412]

Meaningful experiences in tourism: A systematic review of psychological constructs

Ester **Câmara**
University of Algarve, a70232@ualg.pt

Margarida **Pocinho**
University of Madeir and CinTurs, mpocinho@staff.uma.pt

Dora **Agapito**
Faculty of Economics, University of Algarve and CinTurs, dlagapito@ualg.pt

Saúl Neves **Jesus**
University of Algarve and CinTurs, snjesus@ualg.pt

This systematic literature review aimed to answer the following questions: What is a meaningful tourist experience and its components associated with positive psychology, well-being, and mindfulness? How have these experiences been measured and defined? What are their psychological antecedents and consequences? The research protocol was composed by "Tourist experience"; "Meaningful"; "Memorable"; "Transformational"; "Authenticity"; "Extraordinary"; "Mindfulness". The chosen databases were Web of Science and SCOPUS. The inclusion criteria were: Peer-reviewed english articles; Inclusion of concepts from positive psychology; Studies developed on tourism context. The final sample was composed by 70 articles. The results revealed that the main elements assessing meaningful experiences are: Emotions (Positive/negative effects); Nature of the experience (e.g., Memorable tourism experience; delight consumer experiences; Rural tourism experiences); Psychological antecedents (e.g., Needs; Motivation; Familiarity); Well-being (e.g., Hedonia; Eudaimonia; Subjective well-being; Psychological well-being); Behavioral intentions (e.g., Revisit intentions; Positive word-of-mouth; Recommendation); Psychological outcomes (e.g., Memorability; Authenticity; Mindfulness). The antecedents were divided into: Personal (e.g., Search for spirituality; Development of self); Emotional (e.g., Positive sensations; Emotional regulation); Well-being (e.g., Eudaimonia; Hedonia); Behavioral (e.g., Behavioral intentions); Relational (e.g., Positive relationships). The outcomes had the same rational: Personal (e.g., Development of the self; Mindfulness); Well-being (e.g., Eudaimonia; Hedonia); Emotional (e.g., Positive, and negative emotions); Relational (e.g., Positive relationships); Behavioral (e.g., Behavioral intentions). The proposal represents the first work to address the conceptualization of meaningful tourism experiences, associated with positive psychology.

Keywords: Meaningful experiences, positive psychology, well-being, mindfulness, meaning. [ID 380]

Bibliografía

Alalawneh, M. M., Mammadov, J., & Alqasem, A. (2021). Nexus between FDI, infrastructure investment, tourism revenues, and economic growth: Mega event evidence. *Emerging Science Journal*, *5*(6), 953–963. https://doi.org/10.28991/esj-2021-01323

Alsahafi, R., Alzahrani, A., & Mehmood, R. (2023). Smarter Sustainable Tourism: Data-Driven Multi-Perspective Parameter Discovery for Autonomous Design and Operations. *Sustainability (Switzerland)*, *15*(5). https://doi.org/10.3390/su15054166

Álvarez-Carmona, M., Aranda, R., Rodríguez-Gonzalez, A. Y., Fajardo-Delgado, D., Sánchez, M. G., Pérez-Espinosa, H., Martínez-Miranda, J., Guerrero-Rodríguez, R., Bustio-Martínez, L., & Díaz-Pacheco, Á. (2022). Natural language processing applied to tourism research: A systematic review and future research directions. In *Journal of King Saud University - Computer and Information Sciences* (Vol. 34, Issue 10, pp. 10125–10144). King Saud bin Abdulaziz University. https://doi.org/10.1016/j.jksuci.2022.10.010

Andrews, R., Wynn, M. T., Vallmuur, K., Ter Hofstede, A. H. M., Bosley, E., Elcock, M., & Rashford, S. (2019). Leveraging data quality to better prepare for process mining: An approach illustrated through analysing road trauma pre-hospital retrieval and transport processes in Queensland. *International Journal of Environmental Research and Public Health*, *16*(7). https://doi.org/10.3390/ijerph16071138

Azevedo, A., & Santos, M. F. (2008). *KDD, semma and CRISP-DM: A parallel overview*. https://www.researchgate.net/publication/220969845

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer Normalization*. http://arxiv.org/abs/1607.06450

Barrios, D., Russell, S., Andrews, M., & School, H. K. (2016). *Bringing Home the Gold? A Review of the Economic Impact of Hosting Mega-Events*. www.hks.harvard.edu

Behnel, S. , F. M. , & B. (2005). *lxml: XML and HTML with Python*. https://github.com/lxml/lxml

Benckendorff, P., & Zehrer, A. (2013). A network analysis of tourism research. *Annals of Tourism Research*, *43*, 121–149. https://doi.org/10.1016/j.annals.2013.04.005

Bi, F., & Liu, H. (2022). Machine learning-based cloud IOT platform for intelligent tourism information services. *Eurasip Journal on Wireless Communications and Networking*, *2022*(1). https://doi.org/10.1186/s13638-022-02138-y

Bilal, M., & Almazroi, A. A. (2023). Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews. *Electronic Commerce Research*, *23*(4), 2737–2757. https://doi.org/10.1007/s10660-022-09560-w

Bjeljac, Z., Curcic, N., & Ivolga, A. (2017). Tourismological classification of sporting events. *Journal of the Geographical Institute Jovan Cvijic, SASA*, *67*(1), 53–67. https://doi.org/10.2298/ijgi1701053b

Bohlmann, H. R., & van Heerden, J. H. (2008). Predicting the economic impact of the 2010 FIFA World Cup on South Africa. *International Journal of Sport Management and Marketing*, *3*(4), 383–396. https://doi.org/10.1504/IJSMM.2008.017214

Cepeda-Pacheco, J. C., & Domingo, M. C. (2022). Deep learning and Internet of Things for tourist attraction recommendations in smart cities. *Neural Computing and Applications*, *34*(10), 7691–7709. https://doi.org/10.1007/s00521-021-06872-0

Chen, J. (2024). The economic benefits of hosting major football tournaments: UEFA European Championship and Copa America as examples. *SHS Web of Conferences*, *207*, 01008. https://doi.org/10.1051/shsconf/202420701008

Cudny, W., & Paluch, J. (2024). The Potential for Sustainable Tourism Development in Small-Scale Regions: A Case Study of Sulejów Municipality. *Journal of Environmental Management and Tourism*. https://doi.org/10.14505/jemt

Dang, T. D., & Nguyen, M. T. (2023). Systematic review and research agenda for the tourism and hospitality sector: co-creation of customer value in the digital age. *Future Business Journal*, *9*(1). https://doi.org/10.1186/s43093-023-00274-5

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

Diana, M., Cornelia, P., Tiberiu, I., Ramona, C., Loredana, V., & Ioan, P. (2020). *Spiritual tourism and pilgrimage tourism concepts and typology: Vol. XXII* (Issue 1).

Doborjeh, Z., Hemmington, N., Doborjeh, M., & Kasabov, N. (2022). Artificial intelligence: a systematic review of methods and applications in hospitality and tourism. *International Journal of Contemporary Hospitality Management*, *34*(3), 1154–1176. https://doi.org/10.1108/IJCHM-06-2021-0767

Echtner, C. M., & Jamal, T. B. (1997). THE DISCIPLINARY DILEMMA OF TOURISM STUDIES. In *Annals ofTowism Research* (Vol. 24, Issue 4).

García-Pablos, A., Cuadros, M., & Rigau, G. (2018). W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis. *Expert Systems with Applications*, *91*, 127–137. https://doi.org/10.1016/j.eswa.2017.08.049

Getz, D. (2008). Event tourism: Definition, evolution, and research. *Tourism Management*, *29*(3), 403–428. https://doi.org/10.1016/j.tourman.2007.07.017

Getz, D., & Page, S. J. (2014). Progress and prospects for event tourism research. In *Tourism Management* (Vol. 52, pp. 593–631). Elsevier Ltd. https://doi.org/10.1016/j.tourman.2015.03.007

González-Carvajal, S., & Garrido-Merchán, E. C. (2020). *Comparing BERT against traditional machine learning text classification*. https://doi.org/10.47852/bonviewJCCE3202838

Gration, D., Raciti, M., Getz, D., & Andersson, T. D. (2016). Resident valuation of planned events: An event portfolio pilot study. *Event Management*, *20*(4), 607–622. https://doi.org/10.3727/152599516X14745497664596

Guo, Y., Mustafaoglu, Z., & Koundal, D. (2023). Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. *Journal of Computational and Cognitive Engineering*, *2*(1), 5–9. https://doi.org/10.47852/bonviewJCCE2202192

Gupta, K., Kumar, V., Jain, A., Singh, P., Jain, A. K., & Prasad, M. S. R. (2024). Deep Learning Classifier to Recommend the Tourist Attraction in Smart Cities. *2024 2nd International Conference on Disruptive Technologies, ICDT 2024*, 1109–1115. https://doi.org/10.1109/ICDT61202.2024.10489419

Hamdan, I. Z. P., & Othman, M. (2022). Predicting Customer Loyalty Using Machine Learning for Hotel Industry. *Journal of Soft Computing and Data Mining*, *3*(2), 31–42. https://doi.org/10.30880/jscdm.2022.03.02.004

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. In *Nature* (Vol. 585, Issue 7825, pp. 357–362). Nature Research. https://doi.org/10.1038/s41586-020-2649-2

Hazim, L. R., & Ata, O. (2024). Textual Authenticity in the AI Era: Evaluating BERT and RoBERTa with Logistic Regression and Neural Networks for Text Classification. *2024 16th International Symposium on Electronics and Telecommunications, ISETC 2024 - Conference Proceedings*. https://doi.org/10.1109/ISETC63109.2024.10797291

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. http://image-net.org/challenges/LSVRC/2015/

Heydarian, M., Doyle, T. E., & Samavi, R. (2022). MLCM: Multi-Label Confusion Matrix. *IEEE Access*, *10*, 19083–19095. https://doi.org/10.1109/ACCESS.2022.3151048

Iliev, D. (2020). The evolution of religious tourism: Concept, segmentation and development of new identities. *Journal of Hospitality and Tourism Management*, *45*, 131–140. https://doi.org/10.1016/j.jhtm.2020.07.012

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning with Applications in Python*. Springer.

John L. Crompton, S. L. M. (1997). MOTIVES OF VISITORS ATTENDING FESTIVAL EVENTS. *Annals of Tourism Research*, *24*(2), 425–439.

Kahn, S. R. (2015). Routledge handbook of sports event management. *CHOICE: Current Reviews for Academic Libraries*, *53*(3), 460+. https://link.gale.com/apps/doc/A434319681/LitRC?u=anon~c7968ba2&sid=googleScholar&xid=c3e10f74

Kirilenko, A. P., & Stepchenkova, S. (2025). Facilitating topic modeling in tourism research:Comprehensive comparison of new AI technologies. *Tourism Management*, *106*. https://doi.org/10.1016/j.tourman.2024.105007

Krstinić, D., Braović, M., Šerić, L., & Božić-Štulić, D. (2020). *Multi-label Classifier Performance Evaluation with Confusion Matrix*. 01–14. https://doi.org/10.5121/csit.2020.100801

Laing, J. (2018). Festival and event tourism research: Current and future perspectives. *Tourism Management Perspectives*, *25*, 165–168. https://doi.org/10.1016/j.tmp.2017.11.024

Larsen, S., & Mossberg, L. (2007). Editorial: The Diversity of Tourist Experiences. *Scandinavian Journal of Hospitality and Tourism*, *7*(1), 1–6. https://doi.org/10.1080/15022250701225990

Laws, E., & Scott, N. (2015). Tourism research: building from other disciplines. *Tourism Recreation Research*, *40*(1), 48–58. https://doi.org/10.1080/02508281.2015.1005926

Leng, Y., Noriega, A., & Pentland, A. (2021). Tourism Event Analytics with Mobile Phone Data. *ACM/IMS Transactions on Data Science*, *2*(3), 1–22. https://doi.org/10.1145/3479975

Leng, Y., Noriega, A., Pentland, A. “Sandy,” Winder, I., Lutz, N., & Alonso, L. (2016). *Analysis of Tourism Dynamics and Special Events through Mobile Phone Metadata*. http://arxiv.org/abs/1610.08342

Liu, X., Shin, H., & Burns, A. C. (2021). Examining the impact of luxury brand’s social media marketing on customer engagement: Using big data analytics and natural language processing. *Journal of Business Research*, *125*, 815–826. https://doi.org/10.1016/j.jbusres.2019.04.042

Lu, S., Zhu, W., & Wei, J. (2020). Assessing the impacts of tourism events on city development in China: a perspective of event system. *Current Issues in Tourism*, *23*(12), 1528–1541. https://doi.org/10.1080/13683500.2019.1643828

McKercher, B. (2016). Towards a taxonomy of tourism products. *Tourism Management*, *54*, 196–208. https://doi.org/10.1016/j.tourman.2015.11.008

Monteiro, J. E. D., & Marques, O. R. B. (2015). A Jornada Mundial da Juventude 2013: os impactos econômicos dos gastos dos peregrinos na Cidade do Rio de Janeiro. *Tourism & Management Studies*, *11*(2), 71–77. https://doi.org/10.18089/tms.2015.11209

Neuhofer, B., Buhalis, D., & Ladkin, A. (2014). A Typology of Technology-Enhanced Tourism Experiences. *International Journal of Tourism Research*, *16*(4), 340–350. https://doi.org/10.1002/jtr.1958

Oklobdzija, S. (2015). The role of events in tourism development. *Bizinfo Blace*, *6*(2), 83–97. https://doi.org/10.5937/bizinfo1502083o

Özyeşil, M. , K. F. , T. H. , & Ç. M. (2023). The Impact of World Cup Organization on Country’s Economy Growth: The Case of Qatar Economy. *Nişantaşı Üniversitesi Sosyal Bilimler Dergisi*. https://doi.org/10.52122/nisantasisbd.1312631

Padma, S., & Nabi, T. (2024). *Artificial Intelligence’s Evolutionary Impact on the Tourism Sector* (pp. 118–146). https://doi.org/10.4018/979-8-3693-2432-5.ch007

Python Software Foundation. (2024). *json - JSON encoder and decoder*. https://docs.python.org/3/library/json.html

Reitz, K. , C. I. , P. N. (2014). *Requests: HTTP for Humans*. https://docs.python-requests.org/en/latest/

Richardson, L. (2019). *Beautiful Soup Documentation Release 4.4.0*. https://www.crummy.com/software/BeautifulSoup/

Samala, N., Katkam, B. S., Bellamkonda, R. S., & Rodriguez, R. V. (2022). Impact of AI and robotics in the tourism sector: a critical insight. *Journal of Tourism Futures*, *8*(1), 73–87. https://doi.org/10.1108/JTF-07-2019-0065

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. http://arxiv.org/abs/1910.01108

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Schuster, M. , & N. K. (2012). Japanese and korean voice search. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5460.

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). In *International Journal of Innovation and Scientific Research* (Vol. 12, Issue 1). http://www.ijisr.issr-journals.org/

Sharma, P. R. (2019). Selenium with Python: a Beginner's Guide. In *BPB Publications*. https://selenium.dev/documentation/

Shen, Y., & Liu, J. (2021). Comparison of Text Sentiment Analysis based on Bert and Word2vec. *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer, ICFTIC 2021*, 144–147. https://doi.org/10.1109/ICFTIC54370.2021.9647258

Smagina, N. (2017). The internationalization of the Meetings-, Incentives-,Conventions- and Exhibitions- (MICE) industry: Its influences on the actors in the tourism business activity. *Journal of Economics and Management*, *27*, 96–113. https://doi.org/10.22367/jem.2017.27.06

Tenney, I., Das, D., & Pavlick, E. (2019). *BERT Rediscovers the Classical NLP Pipeline*. https://github.com/

The pandas development team. (2020). *pandas: Python Data Analysis Library*. https://doi.org/10.5281/zenodo.3509134

Valentina Bartolic. (2020). *FESTIVALI, SPECIJALNI DOGAĐAJI I TURIZAM*. University of Pula.

Van Rossum, G., & others. (2020). The python library reference, release 3.8. 2. In *Python Software Foundation* (Vol. 16).

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*.

Vujović, Ž. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6), 599–606. https://doi.org/10.14569/IJACSA.2021.0120670

Werner, K., Griese, K. M., & Faatz, A. (2020). Value co-creation processes at sustainable music festivals: a grounded theory approach. *International Journal of Event and Festival Management*, *11*(1), 127–144. https://doi.org/10.1108/IJEFM-06-2019-0031

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. http://arxiv.org/abs/1609.08144

Yuensuk, T., Limpinan, P., Nuankaew, W. S., & Nuankaew, P. (2022). Information Systems for Cultural Tourism Management Using Text Analytics and Data Mining Techniques. *International Journal of Interactive Mobile Technologies*, *16*(9), 146–163. https://doi.org/10.3991/ijim.v16i09.30439

Yung, R., Le, T. H., Moyle, B., & Arcodia, C. (2022). Towards a typology of virtual events. In *Tourism Management* (Vol. 92). Elsevier Ltd. https://doi.org/10.1016/j.tourman.2022.104560

Zarotis, G. F. (2021). *Event Management and Marketing in Tourism*. https://doi.org/10.36348/gajhss.2021.v03i02.001