Trabajo de Fin de Grado

Uso de modelos del lenguaje para analizar la imagen de hoteles en internet

Grado en Ciencia e Ingeniería de Datos

Raúl García Nuez

TUTORIZADO POR:
José Javier Lorenzo Navarro
Sergio Moreno Gil

Julio/2025

SOLICITUD DE DEFENSA DE TRABAJO DE FIN DE TÍTULO

D/Dª Raúl García Nuez, autor/a del Trabajo de Fin de Título “Uso de modelos del lenguaje para analizar la imagen de hoteles en internet”, correspondiente a la titulación Grado en Ciencia e Ingeniería de Datos, plan 40,
en colaboración con la empresa/proyecto (indicar en su caso) _____.

S O L I C I T A

que se inicie el procedimiento de defensa del mismo, para lo que se adjunta la documentación requerida, haciendo constar que

[X] se autoriza / [ ] no se autoriza, la grabación en audio de la exposición y turno de preguntas.

Asimismo, con respecto al registro de la propiedad intelectual/industrial del TFT, declara que:

[ ] Se ha iniciado o hay intención de iniciarlo (defensa no pública).
[X] No está previsto.

Y para que así conste firma la presente. (fecha en firma electrónica)

El/la estudiante

Fdo.: _____

A rellenar y firmar **obligatoriamente** por el/la/los/las tutor/a/es/as.
En relación a la presente solicitud, se informa: (firmar donde corresponda)

Positivamente (en caso de detección de copia, esta firma quedará invalidada)	Negativamente (la justificación en caso de informe negativo deberá incluirse en el TFT05)
Fdo.: _____	Fdo.: _____

DIRECTOR DE LA ESCUELA DE INGENIERÍA INFORMÁTICA

Agradecimientos

A mi tutor Javier Lorenzo por el apoyo y todo el tiempo dedicado.

A mi familia cercana por confiar en mí y apoyarme en todo el proceso.

A mi pareja por ayudarme en todo momento que lo he necesitado.

A mis amigos, por aguantarme en todo este proceso y obligarme a no desconcentrarme y seguir adelante.

Resumen

Para realizar este Trabajo de Fin de Grado se han combinado modelos avanzados del lenguaje junto con modelos multimodales basados en técnicas “Visual Question Answering”. El objetivo principal es evaluar en qué medida la imagen que proyectan los hoteles españoles a través de su listado oficial de servicios, descripciones comerciales e imágenes publicadas en diversas plataformas online coincide con la identidad que desean transmitir y con la percepción real que tienen sus clientes, reflejada en sus reseñas y comentarios.

El análisis se ha llevado a cabo sobre un conjunto representativo de hoteles españoles situados en distintas ciudades turísticas, estableciendo una metodología replicable para detectar posibles discrepancias o coherencias entre el discurso promocional de los hoteles y las opiniones reales de los usuarios. Asimismo, se ha desarrollado una plataforma web interactiva que facilita la consulta rápida de las valoraciones y alineamientos, contribuyendo así a una mejor gestión estratégica basada en datos.

Abstract

This Final Degree Project combines advanced language models with multimodal models that answer questions about images. The goal is to see how well Spanish hotels' official service lists, marketing texts, and online photos match the image they want to project—and whether that matches what guests actually feel, as shown in their reviews.

We looked at a representative group of Spanish hotels in different tourist cities and designed a repeatable method to spot any gaps or alignments between the hotels' promotional messages and real guest opinions. We also built an interactive web platform that lets users quickly explore these findings, helping hotels make more informed, data-driven decisions.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Turismo a nivel mundial

El turismo hoy en día se posiciona como uno de los grandes motores de la economía mundial, por ejemplo en 2024 generó 10,9 billones de dólares, lo que es equivalente a un 10 % del PIB global, y sostuvo más de 357 millones de empleos, lo que se traduce como uno de cada diez puestos de trabajo del planeta [1]. Ese mismo año se registraron 1,4 mil millones de llegadas internacionales, rozando la marca histórica de 2019, es decir, recuperando valores prepandemia [2]. Por otro lado, los organismos sectoriales pronostican que en 2025 la aportación del turismo al PIB global ascenderá hasta los 11,7 billones de dólares (10,3 % del PIB) y el empleo generado por este sector alcanzará los 371 millones de puestos, gracias a la creación de aproximadamente 14 millones de empleos en un año [3].

La relevancia actual del turismo a nivel mundial también se mide en su capacidad para activar otros sectores y cadenas de valor dentro de las cuales se encuentran el transporte, construcción, agricultura, tecnología y servicios culturales. El turismo ha sido un catalizador de inversión en infraestructuras físicas y digitales, ya que se han adoptado soluciones innovadoras como la inteligencia artificial para la gestión de la demanda o la construcción de alojamientos con consumo de energía casi nula, así como la transición hacia combustibles sostenibles. Por todo ello, el turismo se convierte también en un laboratorio de innovación verde y digital [4]. Con estos fundamentos y la evolución actual del sector, el turismo apunta a batir récords los próximos años y se prevé que lo haga de manera sostenible y social. En relación a ello, se debe resaltar que también se ha posicionado como un sector con muchas oportunidades laborales para jóvenes y mujeres en países de todos los niveles de renta [5].

1.2. Turismo a nivel nacional

España mantiene una posición de liderazgo mundial en el sector turístico. Según la Estadística de Movimientos Turísticos en Fronteras (FRONTUR), el país alcanzó en 2024 un récord

histórico de 93,8 millones de turistas internacionales, un 10,1 % más que en 2023 y casi dos puntos por encima de 2019 [6]. Este dato se traduce en 2,7 millones de empleos directos en el sector, consolidando de esta manera al turismo como una pieza esencial en la economía española [7]

La fortaleza del turismo en España no solo se explica por el volumen de visitantes, sino también por el aumento del gasto de los turistas, debido a que el desembolso medio por viajero aumentó hasta 1.441€, un 5,9 % más que el año anterior, elevando la factura total del sector a 126.282 millones de € [8]. Gran parte de este impulso viene dado por la diversificación geográfica y de producto que existe en el territorio nacional. Las comunidades autónomas que más concentran este turismo son Cataluña, Andalucía y Baleares, ya que poseen más de un 55 % de las llegadas, mientras que comunidades como la Valenciana, Canarias o Madrid han reforzado su oferta en los últimos años, con turismo gastronómico, rural y de relajación, reduciendo la dependencia estacional. Todo ello sitúa a España como uno de los países con mejor desempeño competitivo en el Travel & Tourism Development Index 2024 del Foro Económico Mundial, reforzando su capacidad para atraer turistas, por tanto, inversión [4]

1.3. Infraestructura Nacional

El sistema turístico español se apoya en una combinación de infraestructuras de transporte, una gran diversidad geográfica y una oferta de alojamiento variada que permite saciar casi cualquier motivación de viaje. En primer lugar, desde el punto de vista de la movilidad, el transporte aéreo continúa siendo el gran conector internacional; los aeropuertos de la red Aena movieron 309,3 millones de pasajeros en 2024, consiguiendo de esta manera un récord absoluto y un 9,2 % más que el año anterior [9]. En el ámbito terrestre, la ampliación de la red de trenes de alta velocidad ha situado a los trenes como una alternativa eficiente y sostenible: Renfe y los nuevos operadores superaron los 537 millones de viajeros [10]. Por otro lado, también se encuentra el auge de los cruceros, recalando en España con más de 12,8 millones de turistas [11]. Estos métodos de transporte completan un ecosistema logístico que facilita los flujos masivos de turistas.

Sobre esta estructura de conexiones entre destinos, se despliega una gran variedad de productos turísticos en España. En esta variedad sigue liderando el turismo de sol y playa gracias a los litorales mediterráneo y atlántico (Canarias, Baleares, Comunidad Valenciana o Andalucía). Por otro lado, el turismo urbano-cultural de ciudades como Barcelona, Madrid, Sevilla o Valencia gana peso por la densidad patrimonial, la agenda de eventos y la oferta gastronómica. Paralelamente, el turismo rural y de naturaleza, impulsado por la pandemia y la creciente sensibilidad ambiental, continúa mostrando una evolución positiva. En abril de 2025, las pernoctaciones en casas rurales aumentaron un 14,4 % y en campings un 13,5 % con respecto al mismo mes del año anterior [12], multiplicando de esta manera la actividad en comarcas de interior y parques nacionales. A estos tipos de turismo se le añaden el aumento de turismos específicos como el enoturismo en La Rioja y Ribera del Duero, astroturismo en Sierra Morena, cicloturismo en Vías Verdes y religiosos como el Camino de Santiago que también mezcla deporte y cultura. También aporta el turismo MICE (reuniones, incentivos

y congresos) ya que está aumentando notablemente en ciudades como Madrid, Barcelona, Málaga o Valencia. Mientras que eventos deportivos globales, como la Fórmula 1 que ampliará su número de fechas en España en 2026, la Copa Davis, así como otros eventos como fútbol o baloncesto refuerzan la proyección internacional de los destinos españoles.

Para poder recibir la demanda turística del país, se cuenta con más de 17.000 hoteles y 1,8 millones de plazas regladas, a los que se suman apartamentos turísticos, campings, hostales, albergues, paradores históricos y casas rurales. Una encuesta realizada por el INE muestra que, sólo en abril de 2025, las pernoctaciones extrahoteleras repuntaron un 19,4 % interanual [12] confirmando la diversificación de los alojamientos más allá de los hoteles. Esta creciente diversidad de establecimientos turísticos subraya la necesidad de herramientas inteligentes capaces de auditar, de manera homogénea, la información visual y textual que cada alojamiento proyecta en la web.

1.4. Digitalización y plataformas de reserva

La digitalización ha cambiado de manera radical la manera en la que se selecciona, se busca y se contratan los diferentes alojamientos en España. Hoy en día el 88 % de los residentes declaran haber reservado alguno de los componentes del viaje exclusivamente por internet [13]. El programa estatal de Destinos Turísticos Inteligentes ha aportado 225 millones de euros a más de 500 ciudades para dotarlas de big data y analítica predictiva, de modo que la toma de decisiones se apoya cada vez más en las nuevas tecnologías [14]. Este ecosistema digital ha derivado en la extinción práctica de la reserva presencial: la Encuesta de gasto turístico revela que solo un 6 % de los visitantes internacionales utiliza ya la agencia física como canal principal, frente a un 78 % que contrata por web o app y un 16 % que lo hace desde metabuscadores o redes sociales [8].

En esta nueva realidad, las agencias on-line han adquirido mucho protagonismo. La Comisión Nacional de los Mercados y la Competencia documenta que Booking.com concentra entre el 70 % y el 90 % de las reservas hoteleras intermediadas, muy por delante de Expedia y Airbnb [15].

Esta transformación digital se percibe también dentro de los propios establecimientos, ocho de cada diez hoteles emplean ya programas en la nube para publicitarse y vender sus habitaciones en varias páginas web a la vez; seis de cada diez automatizan la fijación de sus precios con programas de aprendizaje automático que aprenden de la demanda y uno de cada tres hoteles permite abrir la puerta con el móvil o hacer el check-in con reconocimiento facial. Estas mejoras, según HotelTechReport, recortan los costes alrededor de un 12 % al año y aumentan la satisfacción de los clientes entre dos y cuatro puntos [16]. No obstante, la dependencia tecnológica actual también tiene riesgos, ya que si un hotel no utiliza los planes promocionales de plataformas como Booking.com, puede perder visibilidad en las plataformas y, en consecuencia, clientes. Para compensar esto, muchos establecimientos se ven obligados a igualar precios con otros canales o a ofrecer extras gratuitos para no quedar atrás.

Por tanto, las plataformas de reserva como Booking.com se han convertido en “escaparates”

digitales donde las imágenes, las descripciones y las opiniones de otros huéspedes ayudan a realizar la elección final de compra de muchos usuarios.

1.5. Información de los alojamientos en internet

En relación a las imágenes, la vista es el primer filtro del humano, y el más importante, ya que en lo primero que se fija un usuario al entrar a la web de un hotel es en la imagen que aparece en la portada de la misma. Un estudio publicado en "Humanities & Social Sciences Communications" confirma que la calidad estética (iluminación, ángulo, presencia humana o paleta de color) explica buena parte del "clic-through" y de la intención de reserva [17]. Por ello, Booking.com prioriza automáticamente las propiedades que superan cierto umbral de resolución y variedad de fotos, ya que se ha demostrado que las páginas con imágenes cuidadas obtienen más conversiones.

Por otro lado y en relación a las descripciones, tras la primera impresión visual, la descripción aclara toda aquella información que no se ve y reduce la incertidumbre en el usuario. Un análisis con numerosas fichas de hoteles mostró que si un hotel posee descripciones claras, consistentes y con un vocabulario adecuado, puede elevar la tasa de reserva frente a otros hoteles con textos genéricos o desactualizados [18]. Cuando la redacción es ambigua o demasiado escueta, el usuario tiende a abandonar el proceso de compra y estudia más opciones.

Por último, los comentarios constituyen el tercer pilar. Ya que según SiteMunder, el 81 % de los viajeros lee siempre las reseñas antes de reservar una habitación y el 52 % nunca reservaría un hotel sin valoraciones [19]. Además, un estudio realizado recientemente analizó 111.000 comentarios y demostró que el estilo lingüístico y las emociones transmitidas en el mismo influyen de manera significativa en la conversión real de clientes [20]. Por eso, Booking.com, de manera automática, destaca las reseñas más útiles y premia a los alojamientos que responden con rapidez, ya que la relación e interacción hotel-huésped refuerza la credibilidad.

1.6. Imagen proyectada en internet

Relacionado con el apartado anterior, la imagen proyectada que un hotel trata de exponer en internet se define como "la combinación de mensajes e impresiones que se emiten deliberadamente sobre un destino (o establecimiento) y que se divulga al mercado objetivo con el fin de despertar conciencia e interés"[21]. Trasladado al ámbito de los alojamientos turísticos, el concepto hace referencia al conjunto de fotografías, textos, símbolos y discursos promocionales que el propio hotel expone en su web o en webs como Booking.com o Tripadvisor para moldear la percepción previa de los clientes potenciales.

Esta imagen proyectada ejerce una influencia directa sobre las expectativas del huésped [22]. Cuando la experiencia real vivida por los clientes dista de lo presentado por el alojamiento de manera online se produce un impacto negativo en la satisfacción del usuario, en su intención de retorno al establecimiento y en la reputación digital del mismo. Existe un estudio que

demuestra que los hoteles con mayor discrepancia entre las imágenes subidas por el hotel y las imágenes capturadas y publicadas por los clientes reciben un 18 % más de reseñas negativas y sufren descensos considerables en la nota media [23].

1.7. Motivación

Este Trabajo de Fin de Grado nace de la realidad de que, en la era digital en la que vivimos actualmente, la reputación de un hotel se construye y se deteriora en internet a gran velocidad. Las plataformas de reserva, redes sociales y metabuscadores han convertido las imágenes, las descripciones y las reseñas de los hoteles en el principal escaparate del establecimiento. No obstante, estos contenidos pueden omitir de manera regular servicios clave del hotel; esto puede generar en el cliente una duda a la hora de reservar, ya que no se refleja la experiencia real que se va a vivir en el alojamiento. Detectar estas carencias y alinear la narrativa digital con la realidad operativa de los hoteles es esencial para que el hotel proyecte una imagen sólida y atractiva en la web.

Por tanto, la motivación principal es doble; por un lado, es esencial proporcionar a los hoteles una metodología que les permita comprobar de forma sistemática y automatizada qué servicios no están presentes en su información expuesta en internet. Y por otro lado, intentar mejorar la transparencia y la reputación del hotel, para que el cliente no obtenga la visión de que el establecimiento ofrece servicios que no brinda realmente.

Por último, este proyecto pretende ayudar a los hoteles a aumentar la conversión de reservas y evitar reclamaciones posteriores, automatizando la detección de los servicios más “criticados” del hotel. De esta manera, el TFG aspira a tener un impacto notable en la estrategia comercial de los hoteles.

1.8. Objetivos

El objetivo, por tanto, consiste en analizar si la imagen que proyectan los hoteles a través de su listado de servicios e imágenes está alineada con la percepción que tienen los clientes en sus valoraciones, utilizando modelos de lenguaje. Este objetivo principal se puede desglosar en los siguientes subobjetivos.

- Adquisición de los datos desde las páginas de los hoteles y de reserva.
- Revisión exhaustiva de la literatura sobre modelos de lenguaje (texto) y modelos multimodales que incluyan visión.
- Procesamiento del conjunto de datos para su limpieza y adecuación a los modelos.
- Implementación y, si procede, entrenamiento de modelos para analizar descripciones, opiniones e imágenes.
- Evaluación del rendimiento del sistema desarrollado mediante métricas estándar.

- Análisis de los resultados obtenidos y propuesta de mejoras o futuras líneas de investigación.

Capítulo 2

Aportaciones del trabajo, Competencias y ODS

2.1. Principales aportaciones

Desde el punto de vista técnico y científico, este Trabajo de Fin de Grado desarrolla una metodología que combina "web scraping", visión por computador y procesamiento del lenguaje natural para revisar y ayudar a mejorar la presencia online de los hoteles. El resultado es un modelo híbrido capaz de identificar de manera automática los servicios presentes en imágenes y descripciones, así como las prestaciones detectadas en los comentarios para su posterior análisis, asignando un índice de conformidad, lo que contribuye a la línea emergente de la aplicación de la inteligencia artificial al marketing hotelero.

En el ámbito empresarial, se aporta una herramienta práctica para la dirección de los hoteles, para automatizar la detección de los servicios menos valorados del alojamiento y su presencia en la información expuesta en internet. De este modo, facilita la toma de decisiones basada en evidencias, optimiza la gestión de contenido en los diferentes canales de reserva y reduce los costes asociados a reclamaciones o correcciones posteriores al poder detectar carencias de manera automática.

Finalmente, desde la perspectiva del cliente y la reputación de la marca, el trabajo contribuye a elevar la transparencia del establecimiento y a alinear las expectativas del huésped con la experiencia real que encontrará en su alojamiento. Al identificar y corregir discrepancias en la comunicación online, se refuerza la confianza entre el usuario y el establecimiento.

Con ello, el TFG no solo ofrece un avance tecnológico, sino que genera un impacto real en la competitividad del sector hotelero y en la calidad de la experiencia turística.

2.2. Competencias

Las competencias relacionadas con el proyecto en cuestión, según la Memoria del Plan de Estudios 40 para el Grado en Ciencia e Ingeniería de Datos [24] son las siguientes:

- **ED1:** Capacidad para conocer los fundamentos, paradigmas y técnicas propias de los sistemas inteligentes y analizar, diseñar y construir sistemas, servicios y aplicaciones informáticas que utilicen dichas técnicas en cualquier ámbito de aplicación
- **ED2:** Capacidad para adquirir, obtener, formalizar y representar el conocimiento humano en una forma computable para la resolución de problemas mediante un sistema informático en cualquier ámbito de aplicación, particularmente los relacionados con aspectos de computación, percepción y actuación en ambientes o entornos inteligentes.
- **ED4:** Capacidad de identificar y analizar problemas y diseñar, desarrollar, implementar, verificar y documentar soluciones software en ámbitos de aplicación de la Inteligencia Artificial en Ciencia e Ingeniería de Datos.

2.3. Alineamiento con los objetivos de desarrollo sostenible

Tabla 2.1: Grado de relación del TFG con los objetivos de desarrollo sostenible.

ODS	**Grado de relación**			
	0 No procede	1 Bajo	2 Medio	3 Alto
1 Fin de la Pobreza	X			
2 Hambre cero				
3 Salud y Bienestar	X			
4 Educación de calidad	X			
5 Igualdad de género	X			
6 Agua limpia y saneamiento	X			
7 Energía asequible y no contaminante	X			
8 Trabajo decente y crecimiento económico				X
9 Industria, Innovación e Infraestructuras			X	
10 Reducción de las desigualdades	X			
11 Ciudades y comunidades sostenibles			X	
12 Producción y consumo sostenibles	X			
13 Acción por el clima	X			
14 Vida submarina	X			
15 Vida de ecosistemas terrestres	X			
16 Paz, justicia e instituciones sólidas	X			
17 Alianzas para lograr objetivos	X			

- **ODS 8:** El proyecto impulsa la competitividad en el sector hotelero mediante inteligencia artificial, ayudando a mejorar la oferta digital de los hoteles y fomentando de esa manera el crecimiento económico.

- **ODS 9:** Este proyecto, promueve el uso de tecnologías avanzadas en el sector turístico, impulsando la innovación digital y la modernización de procesos para mejorar la imagen de los hoteles.

- **ODS 11:** Contribuye a un turismo más responsable, ayudando a que la oferta digital sea coherente con la experiencia real. Y es extendible tanto para ciudades como comunidades.

Capítulo 3

Estado actual y métodos

3.1. Estado del Arte

La transformación digital del sector hotelero ha redefinido la manera de gestionar las estrategias comerciales y comunicativas, posicionando la imagen digital en un escalón clave para influir en las decisiones de reserva de los clientes. Este nuevo contexto presenta importantes desafíos en la gestión coherente de la comunicación online, considerando las expectativas generadas en plataformas digitales y la correlación pertinente con la experiencia real de los huéspedes.

3.1.1. La importancia de la imagen digital en las decisiones de los clientes

La imagen digital que los establecimientos turísticos proyectan en internet se ha convertido hoy en día en un elemento crucial en la toma de decisiones de los usuarios. Según un estudio reciente publicado por TripAdvisor [25], el impacto visual generado por las imágenes de los hoteles es uno de los factores más decisivos a la hora de captar clientes potenciales. Este estudio destaca que las fotografías claras, detalladas y profesionales tienen una influencia directa en la percepción que tienen los clientes sobre el valor del hotel, incrementando de esta manera la confianza del usuario y aumentando las probabilidades de reserva.

Investigaciones adicionales refuerzan estos resultados, como por ejemplo, el estudio realizado por Expedia [26], que revela que mostrar múltiples imágenes de alta calidad con iluminación natural y perspectivas atractivas incrementa las emociones positivas de los usuarios. Debido a que estos factores visuales incrementan la sensación de transparencia y credibilidad del hotel.

Por otro lado, un estudio [17] menciona que la representación visual no solo cumple una función estética, sino también emocional y psicológica, al permitir a los usuarios anticipar o

imaginarse cómo será la experiencia en el hotel. Este hecho facilita el reducir incertidumbres y ayuda al cliente a realizar una toma de decisiones más ágil y segura.

3.1.2. Problema del desajuste entre la imagen proyectada y la experiencia real

Vistos los numerosos y evidentes beneficios de proyectar una imagen digital atractiva, varios estudios han advertido sobre los diferentes riesgos de no hacer una gestión adecuada de la coherencia entre la imagen mostrada online y la experiencia real de los clientes. Este desajuste, denominado frecuentemente como "expectation vs reality", es crítico en el sector turístico y hotelero, dado que la expectativa creada en internet puede diferir de la experiencia real vivida por los clientes en su estancia.

Zhang et al. [27] realizaron una profunda investigación sobre la relación entre las imágenes y la expectativa que generan y la posterior satisfacción de los clientes en su experiencia. Este estudio reveló que, cuando existe una discrepancia notable entre las expectativas visuales creadas por parte del hotel y la realidad, los niveles de satisfacción disminuyen de manera notable, generando emociones negativas y numerosas valoraciones desfavorables. Estas situaciones pueden derivar en pérdidas económicas y de reputación difíciles de gestionar, especialmente en plataformas como Booking en donde las opiniones de los usuarios tienen gran visibilidad.

En línea con estos hallazgos, un estudio [17] demostró que varias características de las imágenes, como iluminación, esquema de colores y presencia humana incrementan significativamente la atención visual y, con ella, la intención de reserva. Las fotografías que combinan características como luz diurna, tonos neutrales, y algún elemento humano elevan la probabilidad de éxito un 14 % frente a imágenes genéricas. Por otro lado, imágenes oscuras, con colores estridentes o sin presencia humana reducen la intención de compra hasta en un 9 %.

3.1.3. Necesidad de metodologías automatizadas para validar la coherencia

La problemática descrita con anterioridad ha generado un consenso entre investigadores y profesionales del sector turístico acerca de la necesidad de desarrollar soluciones más sistemáticas y automatizadas para evaluar la coherencia entre la imagen digital proyectada por los hoteles y la experiencia real percibida por sus clientes.

En los artículos [28] y [29] se destaca que el sector hotelero necesita avanzar hacia métodos analíticos más sofisticados, capaces de integrar grandes volúmenes de datos tanto textuales como visuales provenientes de plataformas digitales, tanto propias como de motores de búsqueda tales como "Booking" o "TripAdvisor". Estos métodos permitirían una evaluación continua y objetiva del desempeño comunicativo de los hoteles, proporcionando información valiosa sobre posibles áreas de mejora en la comunicación digital.

La Inteligencia artificial se ha postulado en los últimos años como la alternativa más prometedora para abordar este reto. Las técnicas basadas en IA, particularmente aquellas relacionadas con el procesamiento del lenguaje natural y el análisis de contenido visual, ofrecen herramientas potentes para automatizar la identificación de inconsistencias en la comunicación digital de los hoteles. Por ejemplo, modelos de lenguaje avanzado, tales como "BERT" [30] o "GPT" [31], permiten la extracción automática de información relevante existente en textos descriptivos y reseñas de clientes. Por su parte, los modelos multimodales basados en Visual Question Answering (VQA), como "BLIP-2" [32] o CLIP [33] permiten analizar simultáneamente textos e imágenes, identificando de esta manera discrepancias entre lo que los hoteles muestran visualmente y lo que describen textualmente.

Estos métodos automatizados no solo mejoran significativamente la eficiencia respecto a los análisis manuales tradicionales, sino que también permiten detectar patrones recurrentes. Por ejemplo, en [34] se desarrolló un modelo de aprendizaje automático profundo capaz de clasificar automáticamente imágenes de baños de hoteles como satisfactorias o insatisfactorias, basándose en más de 11.000 imágenes reales compartidas por usuarios en plataformas de reseñas. Este enfoque demuestra cómo la inteligencia artificial puede identificar discrepancias entre la imagen promocional y la realidad percibida por los clientes, facilitando una evaluación objetiva.

3.1.4. Evolución de los modelos del lenguaje y Visual Question Answering (VQA)

La rápida evolución de los modelos del lenguaje ha ampliado de forma notable las posibilidades de análisis semántico y multimodal en el ámbito turístico. El punto de inflexión en este ámbito llegó con "BERT", cuyo entrenamiento bidireccional permitió extraer aspectos y sentimientos a gran escala en miles de reseñas hoteleras [30]. Posteriormente, la llegada de los "transformers" generativos de gran tamaño como "GPT-3.5" o "GPT-4" posibilitó resumir hilos de opiniones, detectar incoherencias textuales y generar respuestas de manera personalizada con una precisión parecida a la humana [35].

En paralelo a estos avances se han desarrollado otros en modelos multimodales para el tratado de imágenes. CLIP demostró que es posible alinear texto con imágenes en un espacio vectorial compartido, pudiendo etiquetar imágenes de manera automática [33]. Este tipo de modelos han seguido avanzando hasta la actualidad donde existen modelos más desarrollados como "BLIP-2" [36], "Flamingo" [37] y "LLaVA" [38] que integran "visual question answering", subtitulado automático y diálogo contextual, con tasas de acierto superiores al 80 %. Gracias a ello, hoy en día es posible pasar una imagen a los modelos y realizar preguntas, como "existen vistas al mar en esta imagen" y obtener una respuesta automática que lo verifique.

La industria hotelera ya ha puesto en uso estas capacidades. Por ejemplo, "Homes & Villas by Marriott Bonvoy" está probando un buscador que acepta preguntas en lenguaje natural y devuelve resultados entre más de 140.000 propiedades; para ello, está haciendo uso de un modelo multimodal que cruza texto e imágenes para conseguir el resultado más preciso posible [39]. Por otro lado, "IHG Hotels & Resorts" lanzará este 2025, junto a "Google Cloud Vertex

AI" y los modelos "Gemini", un planificador con el que será posible conversar y que resumirá opciones de vuelo, hotel y ocio a partir de las preferencias indicadas por el usuario [40]. Por su parte, la empresa "Accor" emplea redes que combinan puntuaciones de satisfacción y fotografías reales para identificar de manera automática qué recursos visuales utilizados en las imágenes de promoción de los hoteles impulsan la intención de reserva por parte de los clientes [41]. Estos casos demuestran que la IA multimodal y los grandes "transformers" ya no son prototipos, sino modelos útiles para mejorar la búsqueda, la personalización y la conversión en el sector.

3.2. Modelos utilizados

3.2.1. BERT (Bidirectional Encoder Representations from Transformers

"BERT" [30] es un modelo desarrollado por el equipo de investigación de "Google AI Language" y se presentó en octubre de 2018. Desde su aparición, se ha convertido en uno de los avances más importantes en el área del procesamiento del lenguaje natural (PLN), revolucionando por completo la forma en la que se entrenan y utilizan modelos lingüísticos. Hasta este avance, las representaciones de palabras y textos estaban generalmente basadas en modelos unidireccionales (que procesan el texto de izquierda a derecha o viceversa), limitando notablemente la comprensión del contexto completo de cada palabra.

Antes de "BERT", los modelos predominantes eran las redes neuronales recurrentes (RNN, LSTM, GRU) y modelos como "ELMo" (Embeddings from Language Models) [42] y "GPT" (Generative Pretrained Transformer). Mientras "ELMo" empleaba redes "LSTM" bidireccionales relativamente lentas y difíciles de entrenar a gran escala, "GPT" utilizaba "Transformers" con un procesamiento exclusivamente unidireccional de izquierda a derecha, lo que limitaba su capacidad para representar adecuadamente palabras en contextos amplios. "BERT" combina lo mejor de ambos enfoques, ya que adopta la arquitectura de un modelo "Transformer" por su capacidad de procesar secuencias largas y complejas de texto, pero incluyendo la característica de procesar contextos bidireccionales mediante un mecanismo de entrenamiento basado en dos tareas simultáneas. Obteniendo de esta manera su gran capacidad de captar simultáneamente el contexto de cada palabra en ambas direcciones, lo que lo hace más eficiente y preciso en una gran cantidad de tareas.

- **Funcionamiento**
  En relación al funcionamiento interno de "BERT", este modelo recibe las frases y las pasa por un proceso inicial de tokenización usando un algoritmo llamado "WordPiece". Este método divide palabras poco frecuentes o compuestas en subpalabras o subunidades, reduciendo de esta manera el vocabulario a alrededor de 30.000 tokens.

  Cada token de entrada recibe la suma de tres tipos diferentes de representaciones (embeddings):

  – **Token embeddings:** Representan la identidad léxica del token.

  - **Positional embeddings:** Indican la posición exacta del token en la secuencia, crucial para que la arquitectura "Transformer" pueda comprender la estructura secuencial del texto, ya que no posee recurrencia.
  - **Segment embeddings:** Indican a qué frase pertenece cada token, especialmente útil para tareas que implican pares de oraciones. El primer token de cada secuencia es un token especial llamado [CLS] que contiene información global de toda la secuencia, en relación a las frases, dentro de la misma secuencia, estas se separan por el token especial [SEP].

- **Arquitectura**
  "BERT" también, como se dijo con anterioridad, posee una arquitectura basada en el encoder del "Transformer", que destaca por su capacidad de atender simultáneamente a todos los tokens de una secuencia, logrando captar relaciones contextuales complejas.

  Cada una de las capas que componen el encoder de "BERT" procesa la información mediante mecanismos secuenciales definidos. Concretamente, cada capa comienza aplicando un mecanismo llamado autoatención multi-cabeza. Este permite que cada token dentro de la frase reciba una representación específica que tiene en cuenta la relevancia de dicho token en relación a todos los demás tokens presentes en la secuencia, independientemente de la distancia entre ellos. Posteriormente, y para mantener la estabilidad durante el aprendizaje y asegurar una transmisión fluida de la información a través de las capas, el modelo incorpora técnicas de normalización junto a conexiones residuales. Esto facilita que el gradiente pueda fluir sin desvanecerse entre las capas, lo que acelera y estabiliza el entrenamiento del modelo.

  Para finalizar con la estructura de "BERT" se debe resaltar que cada capa del encoder contiene una red neuronal "feed-forward" de dos capas completamente conectadas, que se encargan de enriquecer y transformar las representaciones internas obtenidas tras el mecanismo de autoatención. Esta red "feed-forward" utiliza funciones de activación "GELU"(Gaussian Error Linear Unit), que son realmente eficaces para permitir al modelo captar relaciones más complejas dentro del lenguaje natural. Esta estructura se repite en cada una de las capas del encoder, refinando progresivamente la representación contextual de cada palabra en el texto.

- **Pre-entrenamiento**
  En esta fase se utilizan dos tareas auto-supervisadas muy importantes para el éxito de "BERT":

  - **Masked Language Modelling (MLM):** En esta tarea se selecciona al azar un 15 % de los tokens en cada secuencia. De los tokens seleccionados, un 80 % se sustituyen por el token especial [MASK], un 10 % por un token aleatorio, y otro 10 % de deja intacto. El modelo trata de predecir cuál es el token original en cada posición oculta, esto obliga al modelo a utilizar intensamente el contexto en cada una de las direcciones.
  - **Next Sentence Prediction (NSP):** Durante el pre-entrenamiento, el modelo recibe pares de frases y debe decidir si la segunda frase es realmente la consecutiva

a la primera en el texto original o una frase tomada de otra parte del corpus de manera aleatoria. Cada una de las dos opciones son pasadas al modelo por igual, es decir un 50 % de las ocasiones cada una. Esto ayuda a BERT a desarrollar una capacidad explícita para entender la coherencia textual y las relaciones semánticas entre frases.

- **Ejemplo de funcionamiento**

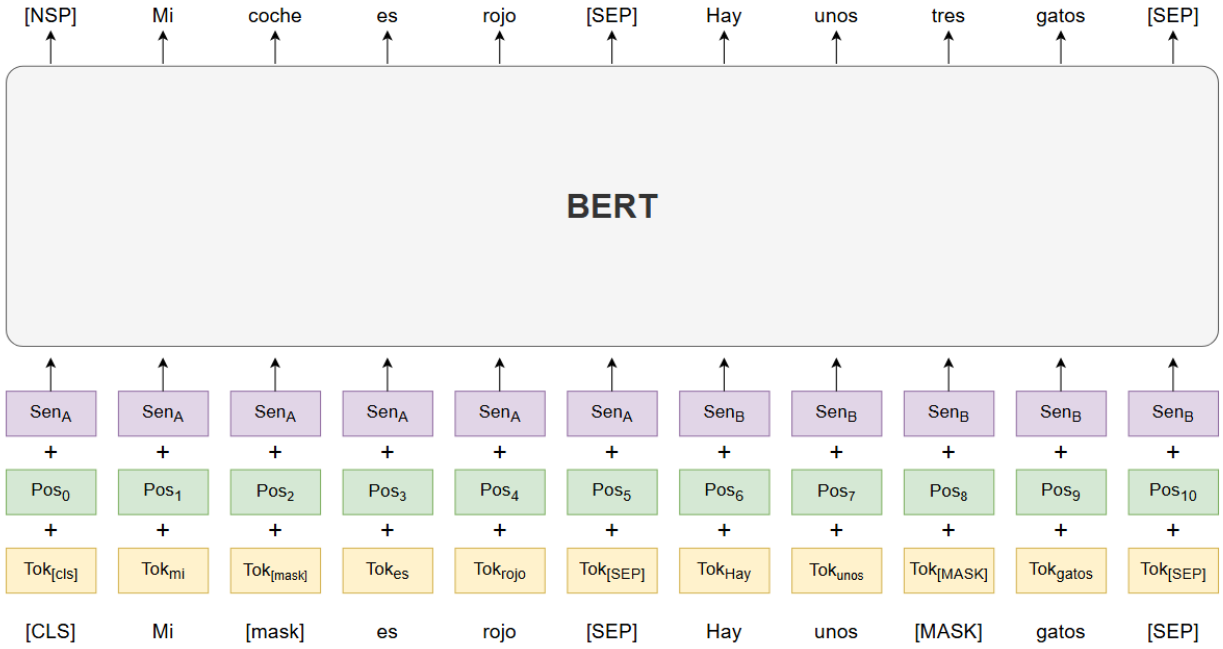


Figura 3.1: Ejemplo de funcionamiento de BERT

Fuente: Modelos del lenguaje basados en redes neuronales artificiales. [43].

En la figura 3.1 se puede observar un ejemplo de funcionamiento del modelo BERT.

- **Entrenamiento** El entrenamiento original de "BERT" se realizó utilizando dos grandes fuentes de datos textuales:
  - **BooksCorpus:** 800 millones de palabras extraídas de libros.
  - **Wikipedia en inglés:** 2500 millones de palabras
- **Utilidades de BERT**
  Desde su introducción, "BERT" ha sido una herramienta imprescindible en una gran variedad de aplicaciones en procesamiento del lenguaje natural:
  - **Clasificación de textos y análisis de sentimientos**
  - **Respuesta automática de preguntas**
  - **Búsqueda y recuperación de información semántica**

  - **Etiquetado de secuencias**
  - **Sistemas conversacionales y chatbots**

3.2.2. RoBERTa (Robustly Optimized BERT Approach)

"RoBERTa" [44] es un modelo del lenguaje creado por "Facebook AI" y presentado en julio de 2019. El desarrollo de este modelo se realizó con la finalidad de optimizar y mejorar el rendimiento obtenido por "BERT" en diferentes tareas del procesamiento del lenguaje natural. "RoBERTa" utiliza la misma arquitectura básica que "BERT", que incluye el encoder del modelo "Transformer"; no obstante, introduce modificaciones importantes en los procedimientos y técnicas de pre-entrenamiento para conseguir representaciones contextuales más robustas y eficaces.

El equipo de "RoBERTa", observó que la tarea de "Next Sentence Prediction" (NSP) no aportaba mejoras significativas en muchos escenarios y que la estrategia fija de enmascaramiento de tokens podría limitar el potencial del modelo. Por tanto, "RoBERTa" nace con la intención de superar esas limitaciones mediante diversos ajustes en la configuración del entrenamiento, permitiéndole alcanzar un rendimiento superior.

- **Funcionamiento**
  RoBERTa recibe el texto y, al igual que BERT, lo procesa inicialmente mediante un algoritmo de tokenización basado en subpalabras, en este caso, se utiliza **"Byte Pair Encoding"**. Este método divide palabras poco frecuentes o compuestas en unidades más pequeñas y comunes, facilitando un vocabulario compacto y manejable por el modelo.

  Cada token generado por esta tokenización también se representa mediante "embeddings", que son vectores numéricos capaces de capturar información léxica y semántica detallada. De igual manera, "RoBERTa" utiliza:

  - **Token embeddings**
  - **Positional embeddings**
  - **Segment embeddings**

- **Arquitectura**
  En relación a la arquitectura, "RoBERTa" utiliza la misma estructura que "BERT", basada en el encoder de "Transformer", destacando por su mecanismo de atención simultánea a todos los "tokens" de la secuencia. Donde en cada capa se ejecutan secuencialmente las mismas tres operaciones: autoatención multi-cabeza, normalización y conexiones residuales; por último, redes neuronales "feed-forward".

- **Pre-entrenamiento**
  El mayor cambio introducido por el modelo en cuestión respecto a "BERT" se encuentra en esta fase. En lugar de utilizar las dos tareas simultáneas de "BERT" (MLM y NSP), "RoBERTa" simplifica y mejora el entrenamiento en los siguientes puntos:

  - **Eliminación de la tarea Next Sentence Prediction:** este paso fue eliminado debido a que, el equipo desarrollador de "RoBERTa" observo que esta tarea no era particularmente útil, y en algunos casos, perjudicial para el rendimiento del modelo.
  - **Dynamic Masking:** "RoBERTa" utiliza un método de enmascaramiento dinámico. En cada iteración del entrenamiento, diversos tokens son seleccionados aleatoriamente para ser enmascarados. Este cambio evita que el modelo memorice patrones específicos, mejorando su capacidad de generalización.

- **Entrenamiento**
  El entrenamiento de "RoBERTa" fue más intensivo que el de "BERT" original; en él se destacan los siguientes aspectos:
  - **Corpus utilizado:** 160 GB provenientes de diversas fuentes, notablemente más extenso y variado que el utilizado por "BERT"
  - **Aumento en duración y capacidad de entrenamiento:** se utilizaron más pasos de entrenamiento y secuencias más largas, con lotes significativamente mayores.

3.2.3. Modelos RoBERTa especializados en español: PlanTL-GOB-ES (RoBERTa-base-BNE y RoBERTa-large-BNE)

Dentro de la iniciativa Plan del Lenguaje (PlanTL) [45], impulsada por la Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA) del Gobierno de España, se han desarrollado varios modelos específicos basados en "RoBERTa" para cubrir las diferentes necesidades del procesamiento del lenguaje natural en español. Entre ellos, destacan los modelos **"RoBERTa-base-BNE"** y **"RoBERTa-large-BNE"**, dos variantes entrenadas íntegramente sobre un corpus extenso y representativo del español actual procedente de la Biblioteca Nacional de España (BNE).

Estos modelos fueron creados con el objetivo de proporcionar modelos lingüísticos avanzados en español, capaces de mejorar el rendimiento en tareas propias del PLN en castellano, respecto a modelos multilingües o modelos originalmente entrenados en inglés, como es el caso de "BERT" o "RoBERTa" original.

- **Características comunes**
  Ambas variantes (base y large) del modelo "RoBERTa-BNE‘" utilizan la arquitectura básica del "encoder" de "Transformer" empleada originalmente por "RoBERTa". Cada token en estos modelos es representado mediante "embeddings" generados a partir del método de tokenización **"Byte Pair Encoding"** (BPE), adaptado especialmente para capturar de manera eficiente y precisa el vocabulario y las particularidades léxicas del idioma español.

- **Descripción específica de cada variante**

  - **RoBERTa-base-BNE:** Esta variante posee una arquitectura compuesta por 12 capas de atención, con una dimensionalidad de 768 unidades y 12 cabezas de atención por capa, sumando aproximadamente 125 millones de parámetros. Gracias a sus menores requerimientos de memoria y tiempo de inferencia, es especialmente adecuado para implementaciones que no poseen una gran cantidad de recursos sin sacrificar de manera signiticativa la precisión en tareas generales del PLN.
  - **RoBERTa-large-BNE:** Esta variante presenta una arquitectura más profunda y robusta, con 24 capas, cada una con una dimensionalidad de 1024 unidades y 16 cabezas de atención por capa, esta estructura suma aproximadamente 355 millones de parámetros. Esta configuración es más costosa que la anterior, permitiendo representar contextos más complejos con mayor precisión, alcanzando mejores resultados en tareas avanzadas y especializadas del PLN.

- **Corpus de entrenamiento**
  El corpus empleado para entrenar estos modelos proviene de la Biblioteca Nacional de España (BNE), una fuente extensa y variada de textos digitalizados en español. Este corpus incluye libros completos digitalizados, publicaciones periódicas históricas y actuales, manuscritos, ensayos académicos y literatura científica, cubriendo diferentes registros lingüísticos, épocas históricas, ámbitos temáticos y estilos. El peso total del corpus empleado en el entrenamiento de este modelo supera los 500 GB de texto en castellano, proporcionando una amplia base para el entrenamiento.

- **Importancia y contribución al PLN en español**
  El desarrollo de estas variantes especializadas de "RoBERTa" representa un paso clave hacia una tecnología lingüística avanzada en castellano. Los modelos en cuestión, al estar íntegramente entrenados en textos españoles, ofrecen una gran ventaja en términos de precisión, robustez y adaptabilidad lingüística frente a modelos multilingües más generales o adaptados a otros idiomas.

3.2.4. CLIP

"CLIP" (Contrastive Language-Image Pre-training) es un modelo multimodal desarrollado por "OpenAI" [33] presentado en enero de 2021. La gran innovación de este modelo reside en su capacidad para relacionar imágenes con lenguaje natural de manera eficiente y precisa, mediante un método de aprendizaje denominado contrastivo. Este tipo de aprendizaje es una forma intermedia entre el supervisado y el no supervisado, que pretende aprender representaciones eficientes de los datos centrándose en diferenciar entre pares de ejemplos similares y disímiles. Gracias a ello, se permite generar representaciones semánticas compartidas entre textos e imágenes, facilitando tareas avanzadas que implican ambos tipos de datos.

Antes del surgimiento de "CLIP", los modelos visuales estaban generalmente entrenados mediante tareas supervisadas, dependientes de conjuntos específicos de imágenes etiquetadas de manera manual. Esto limitaba la capacidad de los modelos para generalizar más allá de las categorías específicas aprendidas, además de ser costoso debido al tiempo empleado en la anotación manual de las imágenes. "CLIP", no obstante, utiliza un enfoque distinto,

aprendiendo directamente de imágenes acompañadas de sus correspondientes descripciones textuales extraídas masivamente de internet, sin necesidad de una supervisión específica y costosa.

- **Funcionamiento del modelo CLIP**
  "CLIP" opera de manera simultánea sobre dos tipos de entrada: imágenes y textos. Su objetivo es aprender un espacio compartido donde las imágenes y sus respectivas descripciones textuales sean representadas mediante vectores semánticamente alineados. El funcionamiento es el siguiente:

  – **Encoder visual:** Las imágenes se procesan utilizando un "encoder" visual basado en arquitecturas como "ResNet" [46] o "Vision Transformer" (ViT) [47]. Este "encoder" transforma cada imagen en un vector numérico ("embedding visual") que captura la información semántica clave contenida en la imagen.

  – **Encoder textual:** Los textos asociados a las imágenes se procesan mediante un encoder basado en la arquitectura "Transformer". Este "encoder" textual convierte cada texto descriptivo en un vector numérico ("embedding textual"), capturando su contenido semántico.

  – **Espacio de representaciones compartidas:** Durante el entrenamiento, "CLIP" ajusta ambos encoders para maximizar la similitud entre el vector generado por una imagen y el vector generado por su correspondiente descripción textual, mientras que minimiza simultáneamente la similitud con vectores generados por textos no relacionados. Este proceso permite al modelo aprender de manera efectiva relaciones semánticas entre texto e imagen.

- **Proceso de entrenamiento**
  El entrenamiento de "CLIP" se basa exclusivamente en una tarea contrastiva a gran escala, usando pares de imágenes y sus correspondientes textos descriptivos. En concreto, "CLIP" fue entrenado sobre aproximadamente 400 millones de pares texto-imagen provenientes de diversas fuentes abiertas en la web, permitiendo que el modelo adquiriera una representación semántica robusta.

  Durante el entrenamiento, "CLIP" optimiza una función de pérdida contrastiva, que impulsa al modelo a asociar estrechamente las imágenes con sus textos correctos y separarlas de textos que no se corresponden. Esta estrategia permite obtener representaciones semánticas potentes y transferibles, capaces de generalizar en múltiples tareas y escenarios sin necesidad de realizar un entrenamiento adicional específico.

- **Arquitecturas principales utilizadas por CLIP**
  "CLIP" puede construirse sobre diferentes arquitecturas neuronales según el "encoder" visual utilizado. Las dos configuraciones más comunes son:

  – **CLIP-ResNet:** Utiliza variantes adaptadas de la arquitectura "CNN ResNet" como "encoder" visual, aprovechando su capacidad efectiva para extraer información visual en imágenes.

  – **ViT:** Emplea la arquitectura "Vision Transformer", basada en autoatención, que

divide la imagen en pequeños fragmentos y procesa estas secuencias mediante atención multi-cabeza. Esta configuración mostró mayor flexibilidad en la representación visual, especialmente efectiva en tareas que requieren una comprensión profunda del contexto visual.

- **Aplicaciones prácticas de CLIP**
  Gracias a su entrenamiento y sus representaciones flexibles, "CLIP" ha mostrado un gran desempeño en multitud de tareas que requieren de un modelo de capacidad multimodal:
  - **Clasificación visual ("zero-shot"):** Permite clasificar imágenes usando simplemente una descripción visual, sin necesidad de entrenamiento previo en dicho problema.
  - **Recuperación y búsqueda visual por texto:** Facilita la búsqueda de imágenes utilizando consultas de lenguaje natural.
  - **Generación automatizada de etiquetas visuales:** Permite describir automáticamente imágenes complejas mediante textos precisos, siendo especialmente útil en etiquetado automático de contenidos multimedia.
  - **Moderación y filtrado de contenidos visuales:** Puede detectar automáticamente imágenes inapropiadas o sensibles usando instrucciones textuales.
  - **Análisis visual en redes sociales:** Clasificación y etiquetado automático en redes sociales sin necesidad de anotaciones manuales adicionales.
- **Limitaciones**
  Aunque "CLIP" ofrece ventajas y aplicaciones prácticas, frente a otros modelos, es importante destacar ciertas limitaciones que se deben considerar al utilizar este modelo:
  - **Sesgos y calidad de los datos en la web:** El entrenamiento utiliza datos masivos de internet, permitiendo de esta manera que el modelo pueda introducir sesgos y estereotipos presentes en los textos o imágenes en línea.
  - **Capacidad limitada para detalles específicos:** Aunque "CLIP" es utilizado para representar categorías generales y semánticas, puede presentar limitaciones a la hora de distinguir detalles muy finos si no están adecuadamente representados en el entrenamiento.

3.2.5. BLIP-1 (Bootstrapped Language-Image Pre-training)

BLIP-1 [48] es un modelo que aparece a comienzos de 2022 como respuesta al reto que primaba en la investigación de modelos multimodales: aprovechar la abundancia de texto-imagen en la web sin depender de manera directa de anotaciones manuales ni sacrificar la capacidad generativa.

Hasta la salida de este modelo, el panorama estaba comprendido por numerosos modelos entre los que destacaban los siguientes tipos:

- **Modelos contrastivos "puros":** este tipo de modelos alinean la imagen y el lenguaje de manera exitosa, pero no saben producir texto coherente de manera efectiva, entre ellos destacan modelos como "CLIP" explicado anteriormente o "ALIGN".
- **Modelos de captioning supervisado:** en este caso, los modelos de esta categoría sí generan descripciones, pero limitados a unos pocos datos.

"BLIP-1" propone una nueva vertiente intermedia, en la que se parte de un conjunto escueto de imágenes que poseen una descripción fiable, y aprende simultáneamente a emparejar cada foto con el texto correcto, a verificar si un par texto-imagen tiene sentido y a redactar una descripción fiable de cada imagen. Una vez dominadas estas tres tareas, el modelo utiliza su propio generador de descripciones para subtitular millones de imágenes sin etiqueta y añade ese material nuevo a su entrenamiento. Se combina lo mejor de los métodos contrastivos (flexibles y funcionan con categorías muy variadas) con la capacidad de los modelos generativos para crear lenguaje detallado, todo ello sin depender de etiquetado de datos manual.

El impacto de "BLIP-1" se percibe en dos dimensiones principales:

- **Democratización de los datos:** al auto-anotar cientos de millones de imágenes, reduce el coste de creación de un corpus de datos de calidad.
- **Unificación de tareas:** ofrece un núcleo único que sirve para recuperar imágenes con lenguaje natural como para generar descripciones fluidas o responder preguntas visuales sin ajustes posteriores.

"BLIP-1" representa para los modelos VQA (Visual Question Answering) un punto de inflexión al igual que BERT en el ámbito del procesamiento del lenguaje natural. Debido a que supuso un cambio de paradigma que pasa de la supervisión laboriosa a la explotación a gran escala de datos abundantes.

- **Funcionamiento**
  Cuando se le entrega una imagen, una frase, o ambas cosas, BLIP-1 sigue siempre los tres mismos pasos; no obstante, en tareas de búsqueda solo utiliza los dos primeros, mientras que en "captioning" o VQA también se hace uso del tercero.
  - **Codificación dual:** En este primer paso, BLIP-1 convierte la información visual y la textual en vectores densos de idéntica dimensión.
    - **Imagen (embedding visual):** La imagen se segmenta en parches por ejemplo de 16x16 píxeles, y cada uno de ellos se aplana y proyecta mediante una capa linea a un espacio de dimensión $\boldsymbol{D}$. A estos vectores-parche se les añade una representación vectorial posicional para conservar su lugar dentro de la cuadrícula, y se le añade al principio un token especial *[CLS]* que actúa como representante global. La secuencia resultante se procesa mediante un *Vision Transformer* (ViT), donde cada bloque de auto-atención multi-cabezal permite que un parche consulte a cualquiera de los demás, permitiendo que este aprenda texturas, bordes locales, disposiciones globales o contrastes de color. Una vez recorridos $\boldsymbol{N}$ de estos bloques, se extrae el vector asociado *[CLS]*, se normaliza a norma unidad y de esta manera se crea el "embedding" visual

$\mathbf{d}_v \in \mathbb{R}^D$ que compacta en una única representación toda la escena.

- **Texto ("embedding" textual):** La frase entrante se fragmenta en subpalabras usando *Byte-Pair Encoding*, lo que mantiene un vocabulario compacto y descompone términos en unidades frecuentes, Cada sub-token obtiene un vector de entrada, se añaden "embeddings" posicionales para indicar su orden, y se inserta un tokken *[CLS]* al principio de la secuencia. Esta cadena atraviesa un "Transformer encoder" lingüístico idéntico a "ViT", con capas de auto-atención multi-cabeza y "feed-forward" con activación *GELU* que permite que cada palabra pueda "atender" a todo el contexto, que incluye dependencias gramaticales, matices semánticos y negaciones. Al final del proceso, el estado del token *[CLS]* se extrae y se normaliza generando de esta manera el "embedding" textual $\mathbf{d}_t \in \mathbb{R}^D$ que resume el significado completo de la oración en un único vector.

Gracias a ello, los "embeddings" están representados en un mismo espacio vectorial y poseen el mismo tamaño.

– **Alineación y verificación:**

- **Alineación mediante pérdida contrastiva (ITC):** Después de obtener los "embeddings" normalizados de imagen $\mathbf{d}_v$ y de texto $\mathbf{d}_t$, ambos pasan por proyecciones lineales que los sitúan en un espacio semántico unificado. En dicho espacio se aplica la pérdida de *Image-Text Contrastive (ITC)*, esta consiste en que en cada lote de entrenamiento, cada par correcto texto e imagen se trata como positivo y todos los demás pares dentro del lote como negativos. Para ello, se calcula la similitud coseno de cada uno de los vectores $\mathbf{d}_v$ y $\mathbf{d}_t$ y usando un *"softmax" cruzado*, el modelo aprende a acercar los emparejamientos correctos y a dispersar los incorrectos. Este mecanismo asegura que, a nivel global, una imagen de específica esté mucho más cercana de su descripción que de cualquier otra frase del lote, estableciendo una alineación semántica.

- **Verificación con atención cruzada (ITM):** Complementariamente, BLIP-1 introduce una *Image-Text Matching (ITM)* que afina el emparejamiento para mejorar la captación de detalles finos. Aquí el texto vuelve a procesarse en un encoder lingüístico similar al inicial, pero enriquecido con atención cruzada al embedding visual $\mathbf{d}_v$, donde cada capa de *self-attention* incorpora consultas que "preguntan" directamente a la representación de la imagen. La salida posteriormente, entra en una cabeza de clasificación binaria que determina si la frase describe con precisión la escena mostrada. Durante el entrenamiento, los negativos más complicados, es decir, aquellos similares a la respuesta correcta, pero con alguna discrepancia como "gato negro" y "gato blanco" se utilizan para forzar al modelo a distinguir atributos como color, cantidad o posición. Gracias a ello, BLIP-1 no solo sabe qué pares son conceptualmente afines, sino también si coinciden en algunos de sus matices.

– **Generación condicionada (LM) (opcional):** Cuando la tarea a realizar, exige devolver texto, por ejemplo, subtitular una imagen o responder a una pregunta

visual, BLIP-1 activa un pequeño decodificador autoregresivo. Este componente recibe como contexto el embedding visual $\mathbf{d}_v$ ya alineado y arranca con un token *[BOS]*. A cada paso se aplican dos atenciones: en primer lugar *causal self-attention*, que solo mira a los tokens ya emitidos para mantener un orden lógico de la frase, y en segundo lugar *cross-attention* al vector visual, garantizando de esta manera que cada palabra generada siga anclada a la escena. El proceso se repite hasta que produce el token *[EOS]* o alcanza una longitud límite. Gracias a la rica representación compartida aprendida en las fases anteriores, el modelo posee la capacidad de redactar descripciones o respuestas detalladas, sin necesidad de un corpus de ajuste adicional.

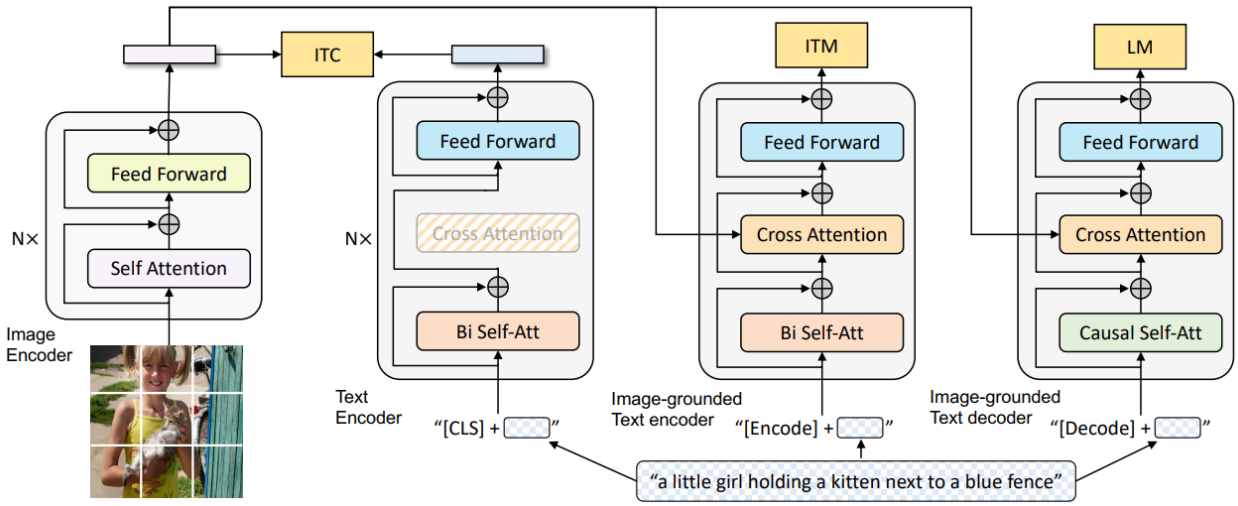


Figura 3.2: Arquitectura y funcionamiento BLIP-1

Fuente: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation [48].

- **Entrenamiento**
  El entrenamiento de BLIP-1 consta de tres fases claramente diferenciadas en la figura 3.2:

  – **Fase semilla:** Se parte de un conjunto de aproximadamente 14 millones de pares imagen-descripción fiables, de diferentes conjuntos como COCO, Visual Genome, Conceptual Captions y SBU. Gracias a ello, el modelo aprende tres objetivos a la vez, explicados con anterioridad: ITC (contrastivo), ITM (matching), LM(captioning).

  – **Auto-anotación masiva:** Con el captioner (modelo generador de subtítulos) ya entrenado, BLIP-1 subtitula más de 100 millones de imágenes de LAION que carecían de texto, creando un corpus sintético mucho mayor.

  – **Fase ampliada:** Al final se obtienen cerca de 129 millones de pares imagen-descripción entre los que se encuentran los pares humanos y los auto-generados. Con ello se vuelve a entrenar todo el modelo desde cero.

Al terminar este proceso, BLIP-1 dispone de un espacio multimodal compacto, un verificador de coherencia y un captioner capaz de describir o responder sobre imágenes nunca vistas, todo ello sin depender de nuevas anotaciones manuales.

- **Utilidades**
  - Búsqueda y recomendación "zero-shot"
  - Subtitulación automática
  - Visual Question Answering (VQA)
  - Detección de incoherencias imagen-texto
  - Filtrado y moderación de contenidos
  - Re-ranking de reseñas con evidencias visuales
- **Limitaciones de BLIP-1**
  - Ruido heredado del web Scrapping
  - Precisión limitada en detalles muy finos
  - Resolución fija de entrada
  - Coste computacional
  - Dependencia del inglés

3.2.6. BLIP-2

Tras el éxito de BLIP-1 en demostrar que un modelo único puede alinear, verificar y describir imágenes sin grandes corpus etiquetados manualmente, surgieron dos frentes donde el modelo tenía una gran capacidad de mejora: el coste computacional y la capacidad de razonamiento lingüístico en tareas complejas. BLIP-2 se crea en 2023 para responder a estos retos.

La idea clave fue desacoplar los dos mundos, es decir, aprovechar encoders visuales muy potentes ya entrenados como ViT-G o EVA y dejarlos congelados, por otro lado, conectar esos embeddings a un LLM externo como OPT, Vicuna, FLAN-T5, etc, que ya domina la comprensión y generación del lenguaje.

Por tanto, BLIP-2 se construye para heredar la cobertura semántica de BLIP-1 pero añadirle la potencia de grandes modelos de lenguaje sin generar el coste computacional de re-entrenar cientos de millones de parámetros visuales, allanando así el camino a aplicaciones multimodales más rápidas y con respuestas elaboradas.

- **Funcionamiento**
  El modelo consta de tres bloques independientes que encajan entre sí, en primer lugar un encoder visual ya pulido, en segundo lugar un puente ligero que aprende en pocas GPU-horas y un gran modelo del lenguaje que no se toca.

- **Encoder de imagen congelado:** Se parte de un *Vision Transformer* o un *EVA-CLIP* ya entrenado sobre cientos de millones de imágenes. La salida no es un vector único, sino una malla de embeddings, uno por parche, que capturan texturas, bordes y disposición de objetos. Ese encoder no vuelve a entrenarse; se usa tal cual, igual que un detector de características en visión clásica.
- **Q-Former:** Entre la malla visual y el mundo textual se coloca un pequeño Transformer con pocas capas y un número grande de tokens de consulta aprendibles, por ejemplo 32. Cada uno de esos tokens "pregunta" a la representación visual mediante atención cruzada y regresa con la información que necesita. El resultado es un lote compacto de vectores que contienen la esencia de la escena, ya resumido y libre de ruido. A continuación, una proyección lineal ajusta su dimensión al formato que el modelo de lenguaje espera. Sólo este puente y su proyección se entrenan, con lo que el coste de ajuste se reduce de cientos a unos millones de parámetros.
- **Inyección en el LLM congelado:** Los vectores del Q-former se introducen como contexto en un LLM ya existente.
  - Si el LLM es de tipo decodificador (OPT, Vicuna, GPT-J) los vectores se concatenan al comienzo del prompt y el generador continúa palabra a palabra.
  - Si el LLM es encoder-decoder (FLAN-T5, Gemini Pro), los vectores se colocan como prefijo del encoder y el decoder toma la representación interna y genera la salida.

  En ninguno de los dos casos se modifican los pesos del LLM, por tanto, se preserva íntegro su conocimiento gramatical.

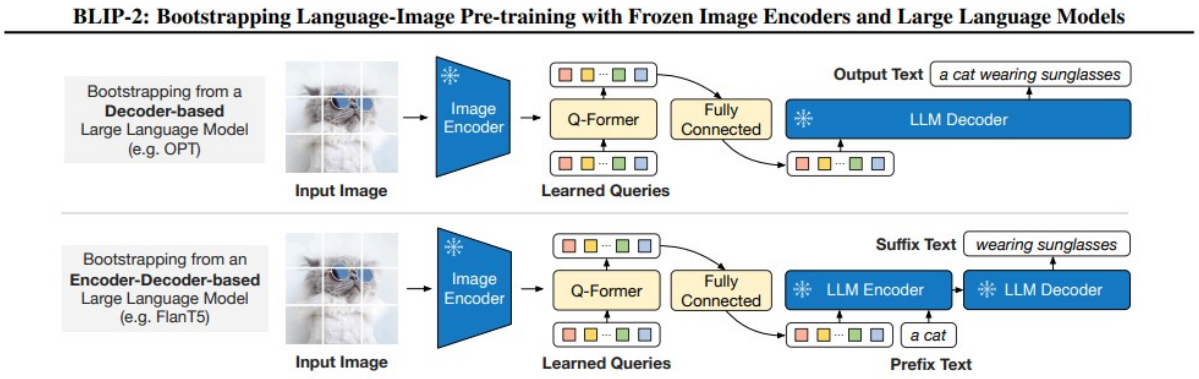


Figura 3.3: Estructura de BLIP-2

Fuente: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models [32].

En la figura 3.3 se puede observar como funciona el modelo explicado y se muestran sus diferentes partes.

- **Entrenamiento**
  El entrenamiento de BLIP-2 se basa en el principio de congelar lo más caro de entrenar y pulir lo ligero. Ya que ni el encoder visual ni el gran modelo de lenguaje reciben gradientes; todo el esfuerzo se concentra en el puente que los conecta (Q-Former + proyección). El proceso se ejecuta en dos fases encadenadas.
  - **Fase de alineación visual-lingüística**
    - **Datos:** Se emplea el mismo conjunto masivo de pares imagen-texto que BLIP-1 generó de aproximadamente 129 millones de muestras, mezclado con COCO y Visual Genome para mantener ejemplos humanos con una alta confianza.
    - **Objetivos:** Existen dos objetivos claros:
      - **Contrastivo (ITC):** Acerca la imagen a su texto correcto y la aleja de todas las demás
      - **Matching binario (ITM):** El Q-Former relee el caption con atención cruzada al embedding visual y un clasificador decide si es el texto correcto o no.
    - **Actualización:** Solo se optimizan los parámetros internos del Q-Former y la proyección lineal. El entrenamiento dura unas 60.000 iteraciones con lotes de 1024 pares.

    Al finalizar esta fase, las consultas del Q-Former ya producen vectores que reflejan con precisión la escena y se alinean semánticamente con la frase correspondiente.
  - **Fase de acoplamiento con el LLM:** Con el puente ya alineado a la imagen, al LLM congelado le llegan los vectores generados en el anterior paso para ello se hace los siguiente:

    - **Preparación del promt:** Existen aproximadamente 32 vectores del Q-Former, y estos se insertan como prefijo de embeddings delante del texto de entrada o como bloque inicial del encoder en un modelo encoder-decoder.
    - **Pérdida de modelado de lenguaje:** El LLM predice la salida completa o la continuación solicitada, usando entropía cruzada normal. El gradiente sólo se propaga hacia el Q-Former y la proyección; los pesos del LLM se mantienen intactos.
    - **Estrategia de curriculum learning progresivo:** Durante las primeras épocas sólo se reconstruyen captions breves, después se introducen preguntas libres y diálogos cortos para que el LLM practique respuestas condicionales más largas.
  - **Ajustes ligeros por tarea:** Para VQA, diálogo visual o introducción multimodal, se aplican unas pocas miles de interacciones adicionales sobre conjuntos como VQAv2, GQA o LLaVA-Instruct.

- **Utilidades**
  - Chat e instrucción multimodal
  - Visual Question Answering de nivel avanzado
  - Generación y reescritura de captions
  - Búsqueda semántica "zero-shot"
  - Anotación o autocompletado de datos
  - Evaluación de coherencia imagen-texto

- **Limitaciones**
  - Dependencia de modelos externos congelados
  - Resolución fija y pérdida de detalle fino
  - Doble arrastre de sesgos (visual + lingüístico)
  - Alto consumo de VRAM en generación
  - Alucinaciones o invención de contenido
  - Precisión numérica limitada en conteos

Capítulo 4

Desarrollo

4.1. Metodología

Para realizar este proyecto se ha seleccionado la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), un marco contrastado en el campo de la ciencia de datos que combina estructura y flexibilidad. Su ciclo iterativo permite regresar a etapas previas cada vez que aparecen nuevos hallazgos o es necesario revisar pasos anteriores, una característica esencial en este proyecto, debido a que los modelos de visión por computador y PLN se ajustan varias veces hasta reflejar de manera fiable los servicios del establecimiento. Además, CRISP-DM arranca con una fase de comprensión del negocio, facilitando así que todo el trabajo técnico se ajuste lo máximo posible a la meta empresarial designada. A día de hoy, sigue siendo la metodología más usada en analítica y minería de datos [49].

Esta metodología consta de seis fases:

1. **Comprensión del negocio:** En esta fase se parte de la necesidad o problema planteado por el negocio y se traduce a objetivos concretos. También se incluye la lectura de numerosos artículos para entender lo máximo posible el problema planteado.
2. **Comprensión de los datos:** Una vez claros los requisitos del negocio, se localizan las fuentes de datos, o se recolectan, y se exploran estadísticamente distribuciones, correlaciones y anomalías con el objetivo principal de comprender los datos.
3. **Preparación de los datos:** Esta fase transforma los datos brutos en un conjunto listo para modelar. Incluye selección de atributos relevantes, limpieza de outliers, creación de variables derivadas útiles para el posterior modelado, divisiones en train y test.
4. **Modelado:** Con los datos ya preparados se eligen técnicas de minería de datos o machine learning adecuadas. Se definen experimentos, se ajustan hiperparámetros y se evalúa el desempeño mediante métricas apropiadas.
5. **Evaluación:** Más allá de las métricas, en esta fase se verifica la validez del modelo frente a los requisitos del negocio y se selecciona el más apto. Se comprueba si las ganancias

justifican los costes de implantación del modelo, la sensibilidad sobre posibles cambios en los datos. En esta fase, a menudo surgen hallazgos que obligan a volver a fases anteriores.

6. **Despliegue:** La fase final consiste en realizar una aplicación, un informe interactivo, una API o otras formas diferentes de mostrar la utilidad de proyecto.

En la ilustración 4.1 se muestran las diferentes fases de la metodología.

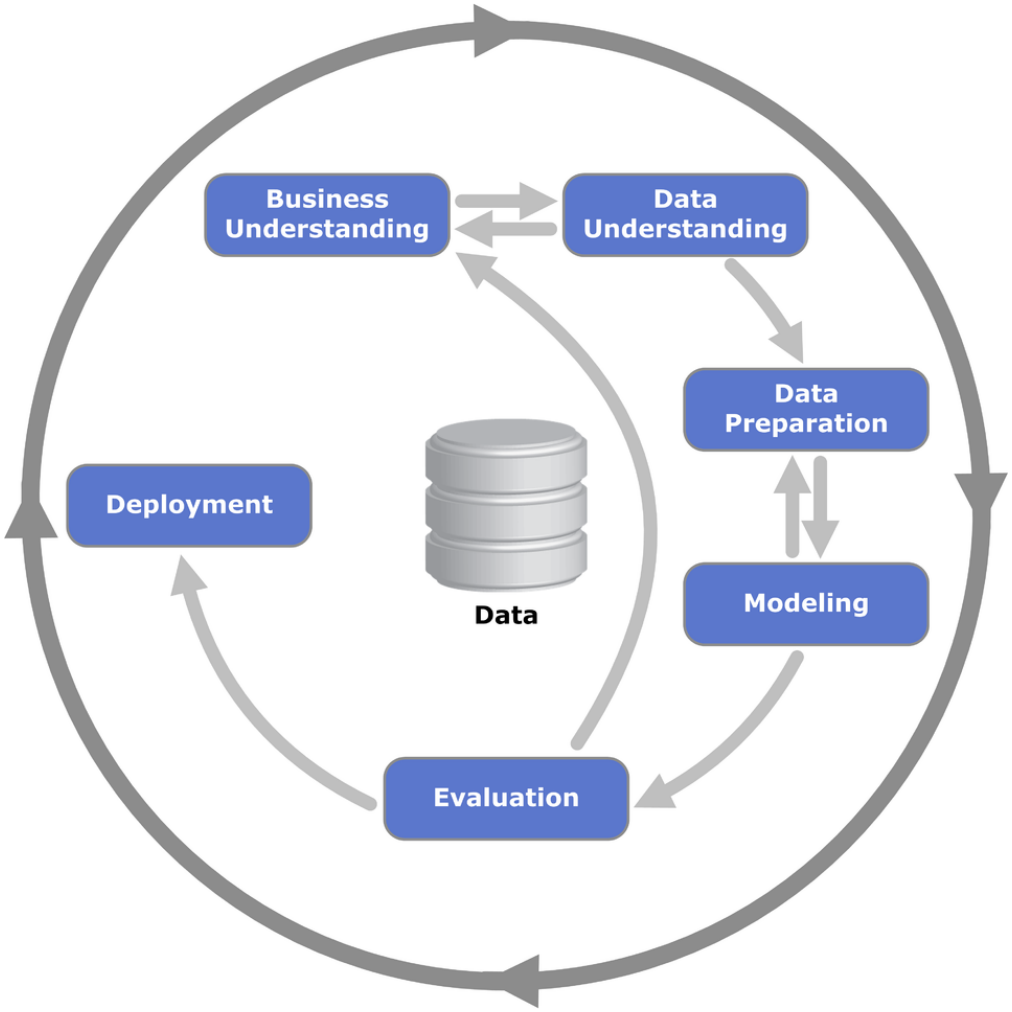


Figura 4.1: Diagrama de la metodología CRISP-DM

Fuente: Archivo:CRISP-DM Process Diagram.png [50].

4.2. Herramientas

- **Entorno de desarrollo integrado:** El entorno de desarrollo utilizado es Visual Studio Code, debido al equilibrio que posee entre potencia y ligereza. Su integración nativa con Git permite gestionar ramas y "pull requests" sin abandonar el editor. Además, es un editor de código gratuito, multiplataforma y altamente personalizable.
- **Lenguaje de programación:** El proyecto se desarrolló en Python ya que es el lenguaje dominante en ciencia de datos, dispone de un ecosistema de librerías contrastado y actualizado, una curva de aprendizaje suave y una gran comunidad para facilitar la resolución de dudas. La versión seleccionada fue la 3.10 y se ha instalado dentro de un entorno virtual alojado en una carpeta env, gracias a ello el intérprete y las librerías están aisladas del resto del sistema, evitando de esta manera problemas de versiones. Todas las dependencias se registran en un archivo requirements.txt, para reproducir el conjunto de librerías con sus respectivas versiones al completo en cualquier dispositivo.
- **Cuaderno interactivo:** Para el trabajo exploratorio, se ha optado por utilizar Jupyter Notebook, ya que es de código abierto y permite reunir en un solo documento, código ejecutable, texto explicativo en formato Markdown o LaTeX y elementos multimedia como gráficos o tabla. Este formato facilita la observación ya que cada celda puede ejecutarse de manera independiente y los resultados aparecen debajo de cada una, esto es imprescindible a la hora de comprobar rápidamente variantes de modelos, diferentes métricas o gráficas.
- **Control de versiones:** Para gestionar de forma eficiente el control de versiones del proyecto, se ha utilizado Git como sistema de control distribuido y GitHub como plataforma de alojamiento remoto del repositorio. Esta combinación ha permitido llevar un seguimiento detallado de los cambios realizados a lo largo del desarrollo, facilitando la organización del código, la recuperación de versiones anteriores y la colaboración en distintas fases del trabajo. Además, el uso de ramas ha permitido separar funcionalidades y pruebas sin comprometer la estabilidad del proyecto principal.
- **Uso de modelos preentrenados con Hugging Face:** Para la implementación de modelos del lenguaje y visión por computador, se ha empleado la plataforma *Hugging Face*, un ecosistema de código abierto ampliamente utilizado en la comunidad científica y de ingeniería de datos. Gracias a su librería *transformers*, se han podido cargar y adaptar modelos preentrenados como BERT, RoBERTa, CLIP o BLIP-2, reduciendo considerablemente el tiempo de desarrollo y entrenamiento. Además, mediante la librería *datasets* y el acceso a su hub de modelos, se ha facilitado la reutilización de recursos y la integración directa con otros módulos del proyecto.
- **Uso de archivos JSON:** Para almacenar y estructurar la información de los hoteles, sus descripciones y comentarios, se han utilizado archivos en formato JSON (JavaScript Objetc Notation). Este formato ligero y fácilmente interpretable por máquinas ha permitido mantener la jerarquía de los datos de forma organizada y ha facilitado el acceso a ellos de manera eficiente para su posterior análisis y tratamiento.

- **Uso de archivos YAML:** Los archivos YAML (YAML Ain´t Markup Language) se han empleado para definir flujos de trabajo estructurados, como los pasos condicionales para la detección de servicios en imágenes. Su sintaxis sencilla y legible los convierte en una opción adecuada para representar configuraciones complejas de forma clara y editable.
- **Librerías utilizadas:**
  - **Playwright.sync-api:** Esta librería permite controlar navegadores de forma programática para automatizar la navegación, clics y descargas en páginas web dinámicas.
  - **Request:** Librería ligera para hacer peticiones HTTP. Se emplea para tareas donde no es necesario simular el navegador completo, como la descarga directa de ciertos recursos.
  - **BeautifulSoup (bs4):** Herramienta para analizar HTML y XML para extraer información estructurada de las páginas obtenidas con las librerías anteriores.
  - **Pandas:** Librería fundamental para la manipulación de datos estructurados. Utilizada para cargar, filtrar, transformar y exportar datos tabulares en diferentes formatos.
  - **PyTorch:** Framework de deep learning utilizado para trabajar con tensores y entrenar modelos.
  - **Transformers (HuggingFace):** Librería utilizada para cargar y ejecutar modelos preentrenados. Permite tanto el procesamiento del input como la generación de salidas en tareas como “captioning” o “visual question answering”.

4.3. Comprensión del negocio

En esta fase inicial del proyecto, el foco estuvo en entender con profundidad el problema a abordar y traducirlo en objetivos concretos y abarcables desde el punto de vista técnico. Para ello, se llevó a cabo una revisión detallada de bibliografía relacionada con el sector hotelero, la percepción del cliente en entornos digitales y el papel de la inteligencia artificial en el análisis de imagen y texto. Además, se mantuvieron varias reuniones para contrastar la viabilidad del enfoque y aclarar diferentes conceptos clave. Estas conversaciones fueron útiles para acotar el problema desde una perspectiva práctica y establecer los límites y prioridades del proyecto.

4.4. Comprensión de los datos

4.4.1. Justificación del Scraping

La premisa central del proyecto consiste en analizar la imagen que proyectan los hoteles en internet a través de sus recursos digitales, tales como descripciones detalladas, listados de servicios ofrecidos, imágenes representativas y la percepción real manifestada por los clientes mediante las reseñas en plataformas especializadas. Para poder llevar a cabo un análisis significativo y robusto, es imprescindible obtener acceso a fuentes heterogéneas y completas que proporcionen:

- **Información objetiva del hotel:** Incluye la descripción oficial del alojamiento, la lista completa de servicios disponibles, así como imágenes del propio establecimiento.
- **Información subjetiva de los clientes:** Incluye comentarios positivos y negativos de usuarios reales que han experimentado los servicios del hotel.

Estas fuentes de información están habitualmente disponibles en plataformas especializadas de reserva hotelera, como Booking.com o Expedia. Sin embargo, dichas plataformas rara vez proporcionan un acceso masivo o sistemático mediante interfaces oficiales (APIs), y en los casos en los que están disponibles estas interfaces, suelen implicar elevados costes o restricciones que limitan de manera considerable la viabilidad para desarrollar investigaciones académicas. Ante esta situación, se ha decidido emplear un enfoque basado en *web scraping*, una técnica automatizada que permite la extracción estructurada de información a partir de contenidos HTML publicados abiertamente en internet. No obstante, esta práctica debe realizarse de manera responsable siguiendo estrictamente las siguientes consideraciones técnicas y éticas:

- **Políticas de uso razonable:** Se ha adoptado una política técnica de contención y distribución temporal de solicitudes, aplicando mecanismos como esperas aleatorias entre peticiones, rotación periódica de IP y simulación de patrones naturales de navegación humana. Esto se ha llevado a cabo, para minimizar el impacto sobre los servidores y evitar la detección automática como bot.
- **Uso estrictamente académico:** Toda la información obtenida está destinada únicamente al análisis académico, con fines únicamente de investigación, y sin ningún propósito comercial ni redistribución de los datos capturados.

Estas precauciones garantizan no solo la ética del proceso, sino también la sostenibilidad técnica, permitiendo obtener un conjunto de datos suficiente en calidad y cantidad para sustentar robustamente el posterior análisis y modelado planteado en el proyecto.

4.4.2. Metodología detallada del scraping

El proceso de scraping se ha estructurado en varias etapas diferenciadas, cada una con objetivos específicos que garantizan la calidad y completitud de los datos recopilados.

- **Construcción de URLs de búsqueda:** Para iniciar el scraping, se definió un módulo específico *url_builder*, encargado de generar dinámicamente las URLs adecuadas para cada destino turístico y período temporal de interés. Este módulo se basa en parámetros configurables, permitiendo así flexibilidad y precisión en la búsqueda inicial, para de esta manera poder conseguir la información de muchos hoteles sin la necesidad de escribir el nombre del alojamiento en concreto. Los parámetros utilizados son los siguientes:
  - **`checkin`**: fecha de entrada al alojamiento (formato `YYYY-MM-DD`). Permite definir desde qué día se desea iniciar la estancia.
  - **`checkout`:** fecha de salida (también en formato `YYYY-MM-DD`). Determina la duración total de la reserva.
  - **`dest_type`:** tipo de destino. En este caso, se utiliza el valor **`city`** para realizar búsquedas dentro de ciudades concretas.
  - **`group_adults`:** número de adultos para la reserva. Se establece por defecto en 2, simulando el caso más frecuente de pareja viajera.
  - **`group_children`:** número de niños. Se fija en 0 para simplificar los resultados y evitar configuraciones adicionales.
  - **`no_rooms`:** número de habitaciones deseadas. Se establece en 1 para simular una reserva sencilla y común.
  - **`order`:** criterio de ordenación de los resultados. El valor `popularity` indica que se priorizan los alojamientos más populares según el algoritmo interno de la plataforma.
  - **`city`:** término de búsqueda, que corresponde con el nombre de la ciudad introducida por el usuario.

  Las fechas se van cambiando debido a que existen hoteles que no poseen algunas fechas disponibles. Las ciudades también se van cambiando entre las 6 elegidas para este proyecto que son las siguientes:
  - Asturias
  - Barcelona
  - Gran Canaria
  - Ibiza
  - Madrid
  - Tenerife

- **Captura y extracción del contenido HTML:** El núcleo del scraping está gestionado por el módulo *scraper.py*, que utiliza la biblioteca *Playwright* para simular un navegador web real en modo "headless". Esta herramienta es capaz de cargar completamente páginas generadas por JavaScript, necesarias en portales como Booking.com. La metodología en este módulo incluye:
  - Acceso inicial mediante URL previamente construida.
  - Scroll automático y pausas estratégicas para permitir la carga completa de contenido dinámico
  - Apertura automatizada de galerías mediante la interacción con botones específicos identificados por selectores CSS.
  - Extracción del contenido HTML final, que es posteriormente procesado mediante *BeautifilSoup* y expresiones regulares para obtener datos estructurados (nombre, descripción, servicios e imágenes)

  Además, para optimizar la extracción y asegurar que el contenido sea accesible y completo, se implementaron varias estrategias adicionales en este módulo:
  - **Gestión proactiva del banner de cookies:** Debido a que los banners emergentes de cookies pueden interferir en la navegación automática, se incluyó una rutina específica para detectarlos y cerrarlos de manera automática. Esta rutina utiliza selectores CSS para localizar y cerrar el banner, evitando posibles errores y facilitando la interacción posterior con los elementos relevantes.
  - **Intentos múltiples para acceder a la galería de imágenes:** Se implementó una lógica para la interacción con botones específicos que abren la galería de imágenes del hotel. La metodología se basa en hacer clic secuencialmente en todos los botones disponibles con clases CSS concretas hasta que se logra abrir la galería. Posteriormente, se verifica la aparición efectiva del elemento HTML que corresponde a la galería, asegurando así la captura de todas las imágenes correspondientes al hotel.
  - **Fallback para la extracción de imágenes:** En caso de que no se pueda abrir la galería de imágenes de un hotel después de múltiples intentos, se activa un mecanismo de extracción alternativa, donde se obtienen imágenes directamente visibles en la página principal del establecimiento. Este mecanismo asegura que siempre se capturen imágenes.
  - **Extracción de comentarios:** Los comentarios de usuarios se obtienen mediante solicitudes adicionales personalizadas a páginas específicas de valoraciones. Se aplica una búsqueda avanzada para obtener tanto valoraciones positivas como negativas, lo que facilitará el análisis de opinión en etapas posteriores.
  - **Guardado de datos estructurados:** Tras completar la extracción de información de cada hotel (nombre, descripción, lista de servicios y comentarios), los datos son encapsulados en un diccionario estructurado en formato clave-valor. Este diccionario se guarda en un archivo JSON llamado `hotels.json`, lo que permite

mantener una representación legible y manipulable de todos los hoteles procesados. Esta estrategia facilita la carga posterior para análisis, modelado o visualización.

- **Verificación de duplicados:** Antes de almacenar los datos de un nuevo hotel, se comprueba si ya existe en el fichero `hotels.json`. Para ello, se busca si el nombre del hotel ya está presente entre las entradas almacenadas. En caso afirmativo, se omite el scraping de ese hotel y se continúa con el siguiente. Esta verificación evita realizar trabajo redundante, optimiza el tiempo de ejecución y previene la generación de entradas duplicadas en el conjunto de datos final.

Estas técnicas garantizan que la información obtenida sea detallada y representativa, proporcionando un conjunto de datos completo para el análisis posterior.

– **Programación del scraper automático**
El módulo *scheduler.py* permite definir tareas recurrentes con una frecuencia determinada; en el caso de este proyecto, se ha ido variando con valores comprendidos entre 5 y 10 minutos. El objetivo principal de este módulo es automatizar la recolección progresiva de datos, gestionando internamente la persistencia del índice del hotel actual mediante el archivo *last_index.json*. Así, en cada ejecución, el scraper continúa justo donde se quedó en la ejecución anterior; esta característica permite recorrer listas extensas de hoteles sin repetir pasos previos.

– **main.py** Este archivo actúa como punto de entrada del sistema de scraping. Proporcionando una interfaz de línea de comandos para que el usuario ajuste los parámetros indicados con anterioridad.

Con esta información, el script construye la URL de búsqueda utilizando el módulo *url_builder.py*, y lanza el proceso automatizado gestionado por *scheduler.py*. Este diseño permite una alta flexibilidad, pudiendo adaptar el scraper a cualquier ciudad o fecha sin necesidad de modificar directamente el código.

4.4.3. Arquitectura de almacenamiento de datos

La organización y almacenamiento de los datos obtenidos mediante el scraping se estructuran en dos partes claramente diferenciadas, permitiendo un acceso eficiente para su posterior análisis.

- **Almacenamiento de imágenes:** Cada destino turístico analizado tiene asignada una subcarpeta dentro del directorio principal `data/images`. A su vez, cada una de estas subcarpetas agrupa en su interior otra subcarpeta por cada alojamiento de este destino y dentro del mismo, las imágenes.

  Las imágenes están nombradas siguiendo un patrón numérico secuencial, permitiendo identificar fácilmente las imágenes.

  Ejemplo visual de la estructura:

```
data/
  images/
      Asturias/
         Hotel_X/
             image_1.jpg
             image_2.jpg
      Gran Canaria/
          Hotel_Y/
              image_1.jpg
              image_2.jpg
```

- **Almacenamiento textual en formato JSON:** Toda la información textual de los hoteles se guarda de forma estructurada en un único archivo JSON denominado `hotels.json`. Este archivo contiene una lista de objetos, donde cada objeto representa un hotel y está compuesto por los siguientes campos:
  - **`name`:** Nombre del alojamiento.
  - **`description`:** Descripción textual del establecimiento.
  - **`services`:** Lista de servicios oficiales del hotel.
  - **`comments`:** Lista de opiniones de clientes. Cada elemento incluye dos campos: `positive` y `negative`.

  Esto permite una organización clara y compacta del conjunto de datos, facilitando su lectura, análisis, visualización y acceso posterior. Un único archivo agrupa todos los hoteles, simplificando tareas como la carga masiva de datos o su posterior transformación para modelos de lenguaje y visualización.

  Ejemplo simplificado de estructura en JSON:

```
{
  "name": "Hotel X",
  "description": "Hotel X posee habitaciones dobles,
  con camas muy cómodas",
  "services": [
    "Piscinas",
    "Restaurante o Buffet o Bar",
    "Área de Playa"
  ],
  "comments": [
    {
      "positive": "Excelente comida y trato del personal.",
      "negative": "La habitación no estaba al nivel esperado."
    },
    ...
  ]
}
```

4.4.4. Análisis exploratorio de datos

Para comprender los datos obtenidos mediante el proceso de scraping, se realizó un breve análisis exploratorio utilizando técnicas gráficas para identificar patrones relevantes y evaluar la calidad del conjunto de datos.

- **Conteo de imágenes, servicios y hoteles:** En primer lugar, se realizó un análisis exploratorio para cuantificar el alcance del *dataset*: se determinaron el número total de hoteles recopilados, la variedad de servicios únicos identificados y el volumen de imágenes disponibles. Estos valores sirven como referencia cuantitativa para los análisis descriptivos y los futuros modelos de aprendizaje automático y se muestran en la tabla 4.1

Tabla 4.1: Resumen cuantitativo del *dataset*

Métrica	**Cantidad**
Hoteles	550
Servicios únicos	412
Imágenes	19 933

- **Relación entre el número de imágenes y el número de servicios:** Se analizó mediante un `scatterplot` la relación existente entre la cantidad de imágenes disponible para cada hotel y el número de servicios que ofrece. Este análisis permite evaluar si los alojamientos que más imágenes poseen también tienden a enumerar un mayor número de servicios.

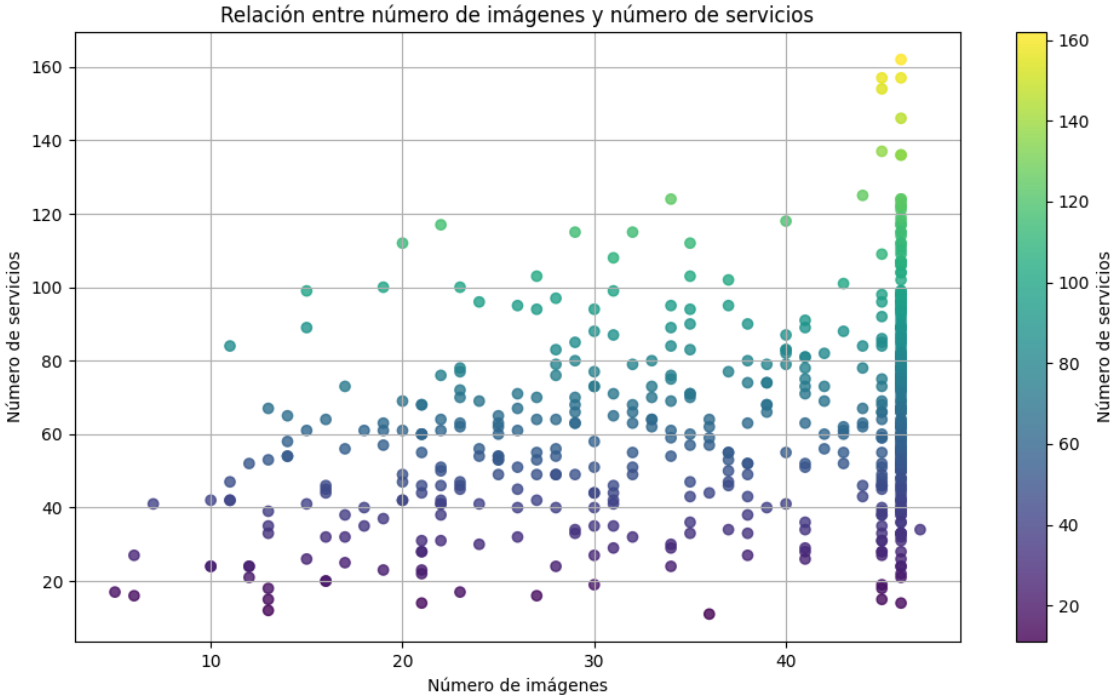


Figura 4.2: Relación entre número de servicios e imágenes

En la figura 4.2 se observa una tendencia positiva, aunque también se destaca una considerable variabilidad. Algunos hoteles con muchas imágenes poseen un número reducido de servicios, mientras que otros con menos imágenes ofrecen un mayor número de servicios. Esto indica una relación no lineal. Sin embargo, los hoteles con mayor número de servicios poseen 40 imágenes o más. También se muestra un límite superior en 46 imágenes, ya que es el número máximo de imágenes permitidas en la plataforma "Booking.com".

- **Distribución del número de imágenes por destino:** La distribución del número de imágenes por hotel se estudió mediante gráficos de caja, segmentados por destino turístico.

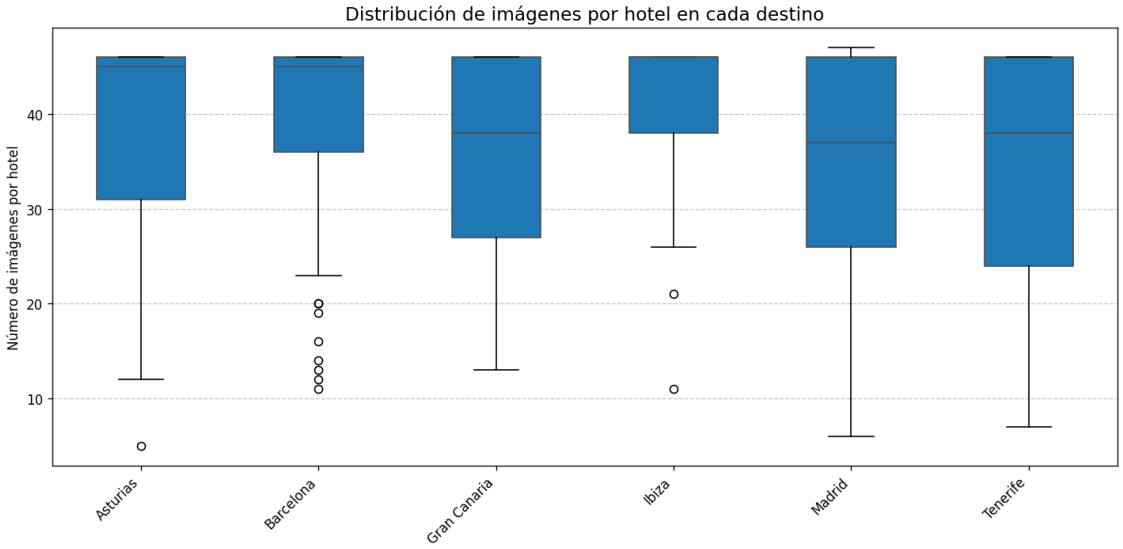


Figura 4.3: Número de imágenes por destino

En la figura 4.3 se observan diferencias entre los diferentes destinos, en primer lugar, Asturias y Madrid muestran una mayor variabilidad, con hoteles que abarcan un rango amplio de imágenes. Por otro lado, Barcelona muestra varios valores atípicos, indicando una concentración de hoteles que ofrecen menos imágenes que el promedio del destino.

- **Número de hoteles por destino:** Se ha realizado una gráfica para visualizar el número de hoteles descargados por cada destino.

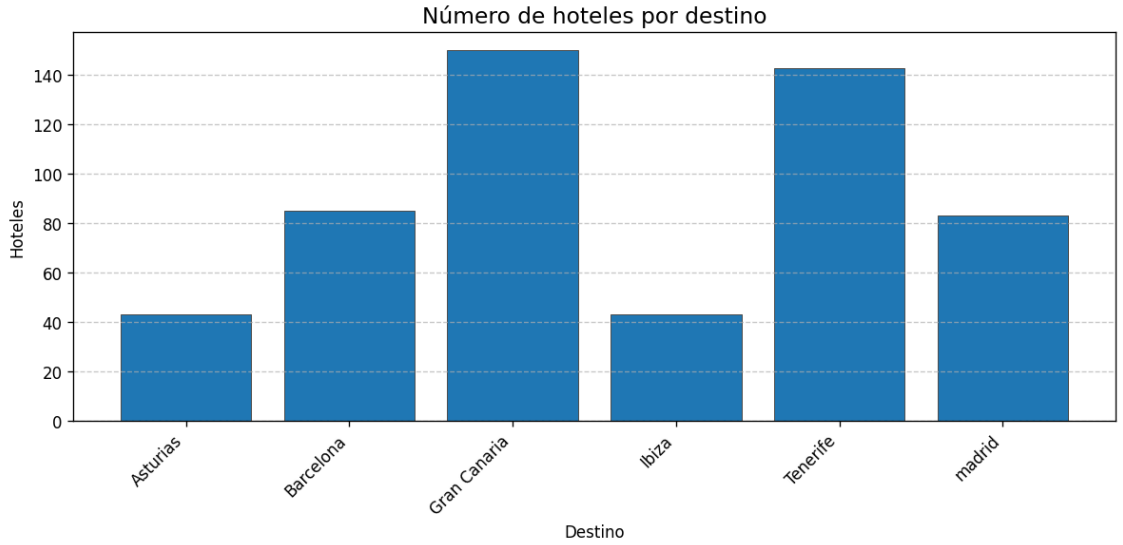


Figura 4.4: Número de hoteles por destino

En la figura 4.4, se puede observar que los destinos canarios son los que concentran mayor cantidad de alojamientos, mientras que Asturias e Ibiza tienen menos hoteles listados.

- **Longitud de las descripciones:** También se ha realizado un estudio sobre la longitud de las descripciones, para ver si los hoteles son más propensos a generar descripciones amplias o escuetas.

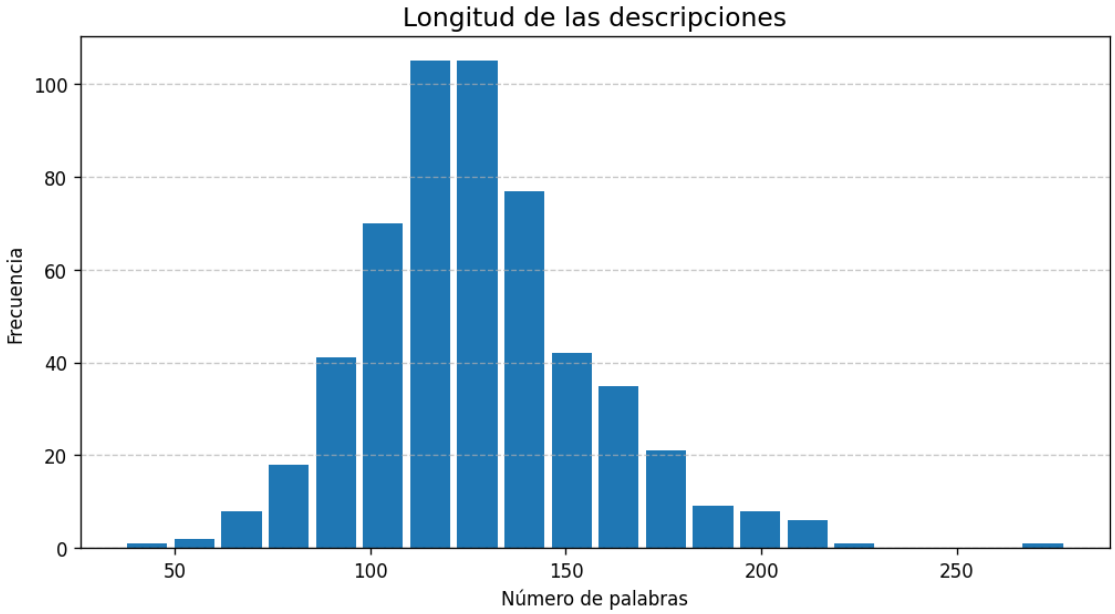


Figura 4.5: Longitud de las descripciones

En la figura 4.5 se puede observar la distribución del número de palabras en las descripciones, muestra una concentración alrededor de 100 a 150 palabras, lo que indica un estándar generalizado en cuanto a la cantidad de información que los alojamientos consideran oportuno para describirse de manera clara. Un pequeño número de hoteles se desvía de esta tendencia tanto por arriba de la concentración como por los valores menores.

- **Frecuencia de servicios ofrecidos:** En la figura 4.6 se muestra una gráfica que da información sobre los servicios más comunes en los hoteles de los que se posee información.

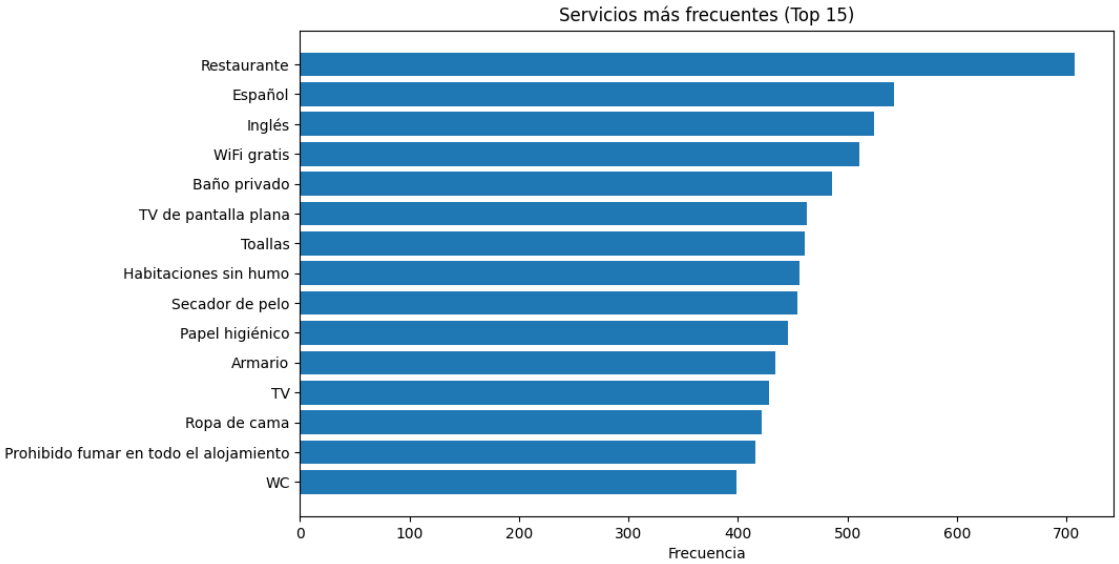


Figura 4.6: Servicios más comunes

Se puede ver que al identificar los servicios más frecuentes en los hoteles analizados, se observa que los elementos que más destacan en los alojamientos son aquellos más básicos como restaurantes, conexión WiFi, baño y TV.

- **Distribución del número de servicios por hotel:** Por último, se ha graficado la cantidad de servicios ofrecidos por los hoteles.

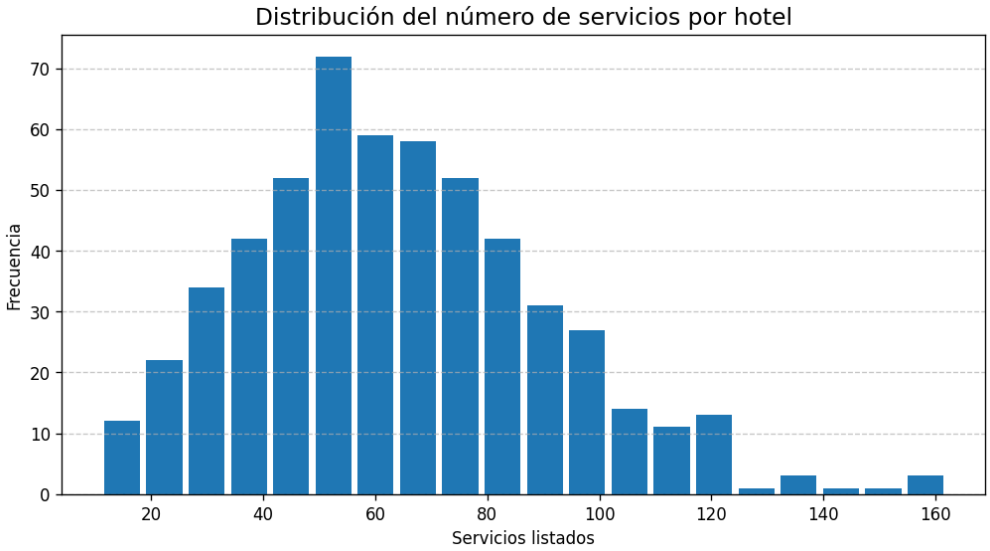


Figura 4.7: Distribución de servicios

En la figura 4.7 se muestra la cantidad de servicios ofrecidos por hotel. Esta figura presenta una distribución parecida a la normal, con la mayoría de los alojamientos comprendidos en el rango de 40 a 80 servicios diferentes. Esta distribución muestra que hay pocos establecimientos que se desvíen significativamente del resto.

Este análisis preliminar no solo permite detectar posibles errores, sino que también proporciona contexto útil para la futura interpretación de los resultados de los modelos utilizados.

4.5. Preparación de los datos

El proceso de preparación de datos para este proyecto ha sido conciso gracias al enfoque adoptado desde la obtención inicial. Debido a que, al llevar a cabo personalmente la obtención mediante scraping, se procuró obtener datos lo más limpios y estructurados posible desde el origen, minimizando de esta manera la necesidad de ajustes posteriores.

4.5.1. Metodología de limpieza y normalización de datos textuales

La limpieza y normalización se llevaron a cabo en un notebook específico denominado `limpieza.ipynb` mediante Python y bibliotecas estándar como Pandas. Los pasos seguidos fueron los siguientes:

- **Carga inicial de datos:** Los datos se cargan directamente desde `hotels.json` y se seleccionan los servicios.
- **Eliminación de duplicados:** Se identificaron y eliminaron entradas duplicadas en los servicios.
- **Normalización textual:** Se aplicó una conversión general a minúsculas y se eliminaron posibles caracteres especiales. Para posteriormente, guardarlos en formato `CSV` junto a la frecuencia de aparición de cada uno de los servicios.

4.5.2. Categorización mediante JSON

Para organizar y categorizar adecuadamente los servicios recopilados, se utilizó un archivo `categorization_services.json` previamente definido, que contiene correspondencias entre todos los servicios que aparecen en el `CSV` y 23 categorías generales.

- **Categorías generales:** Estas categorías generales, fueron seleccionadas cuidadosamente buscando cubrir todas las dimensiones relevantes presentes en la información inicial de todos los alojamientos, permitiendo una agrupación semántica coherente que facilitará su manejo, análisis e interpretación futura. Estas categorías se muestran en la tabla 4.2:

Tabla 4.2: Servicios ofrecidos

Categoría (ES)	**Categoría (EN)**	**Descripción**
Dormitorio	Bedroom	Camas, mobiliario de dormitorio
Salón/Zona de Estar	Living Room	Zona con sofás, sillas y mobiliario de salón
Mobiliario de Interiores	Indoor Furniture	Sofás, mobiliario de comedor y aire acondicionado
Entretenimiento y Multimedia	Entertainment and Multimedia	Televisión, pantallas y equipos de sonido
Baño/Aseo	Bathroom/Toilet	Bathroom or a toilet
Cocina	Kitchen	Zona de cocina integrada, electrodomésticos y utensilios de preparación
Restaurante o Buffet o Bar	Restaurant or Buffet or Bar	Restaurante, buffet y bar
Área de Playa	Beach Area	Acceso a la playa o playa cercana
Piscinas	Swimming Pools	Piscinas principales y zonas especiales de piscina
Deportes de Interior	Indoor Sports	Deportes de mesa y bolos
Deportes en la Naturaleza	Outdoor/Nature Sports	Deportes de raqueta, minigolf, deportes acuáticos, deportes de invierno y senderismo
Fitness, Spa y Belleza	Fitness, Spa and Beauty Services	Entrenamiento físico, clases, spa, relajación y servicios de belleza
Chillout y Relax	Chillout and Relax	Zonas de relajación y descanso, con tumbonas y mobiliario cómodo
Entretenimiento y Eventos	Events and Entertainment	Juegos en interiores, espectáculos en vivo, música y eventos sociales
Transporte y Accesibilidad	Transportation and Accessibility	Servicios de movilidad, infraestructura para coches eléctricos y accesibilidad para personas con movilidad reducida y parking
Servicios Adicionales	Additional Services	Lavandería, limpieza de ropa, utensilios para planchado, snacks, tienda interna y servicios privados en la habitación
Recepción y Atención al Huésped	Reception and Guest Services	Recepción y atención personalizada a los huéspedes
Terraza	Terraces	Espacios al aire libre diseñados para el descanso y la socialización, equipados con mobiliario cómodo y vistas atractivas.
Instalaciones para Negocios	Business and Meeting Facilities	Espacios destinados para reuniones, conferencias y actividades de negocios.
Actividades Culturales y Excursiones	Cultural Activities and Tours	Experiencias culturales, visitas guiadas y excursiones para explorar la cultura local.
Vistas al mar	Sea views	Vistas al mar o a la playa
Parque infantil	Playground	Zona de juegos para niños al aire libre
Zona de juegos	Play area	Zona de juegos como ping pong, billar etc

- **Asignación de servicios:** Posteriormente, se procesaron todos los hoteles, utilizando este `JSON` de categorización, para asignar unicamente a los hoteles los servicios pertinentes.
- **Guardado posterior:** Por último, se modifica el archivo `hotels.json`, para actualizarlo con los nuevos servicios de cada hotel.

4.5.3. Preprocesamiento visual

Para algunos de los modelos de procesamiento de imágenes, se optimizó el proceso de inferencia para reducir el uso de memoria y acelerar el cómputo. Para ello, se utilizaron varias estrategias:

- **Redimensionado proporcional:** Todas las imágenes se reescalan para que no excedan los 512 píxeles. Esto reduce el tamaño del tensor de entrada sin perder la estructura visual necesaria.
- **Conversión a RGB:** Todas las imágenes se transforman al espacio de color RGB para garantizar compatibilidad con los modelos.

4.6. Modelado

En primer lugar, se debe resaltar que todas las pruebas y el modelado se han realizado en un equipo compuesto por los siguientes componentes:

- **Procesador:** "AMD Ryzen 7 5700X" de 8 núcleos.
- **Memoria RAM:** El equipo dispone de 32 GB de memoria RAM DDR4.
- **Tarjeta gráfica:** La tarjeta gráfica utilizada fue una "NVIDIA Geforce RTX 4060 Ti" de 8 GB de VRAM.

El proceso de modelado desarrollado en este proyecto se ha estructurado en tres fases diferenciadas, cada una centrada en un tipo específico de dato obtenido a través del scraping: imágenes, descripciones y comentarios. Esta separación permite aplicar estrategias y modelos adaptados a la naturaleza particular de cada una de las fuentes, intentando maximizar la precisión y la escalabilidad.

- En la primera fase, se abordó el modelado de imágenes, utilizando modelos multimodales preentrenados para detectar la presencia visual de servicios ofertados por los hoteles.
- La segunda fase se centró en el análisis de las descripciones, aplicando modelos del lenguaje para identificar y clasificar automáticamente los servicios mencionados de forma explícita o implícita en los datos.
- Por último, la tercera fase se dedicó al tratamiento de comentarios de clientes, donde se realizó una detección de servicios y un análisis de sentimientos.

4.6.1. Modelado de imágenes

– **Justificación de la tarea y modelos seleccionados**
El objetivo principal de esta fase es detectar visualmente los servicios que ofrecen los hoteles, a partir de las imágenes extraídas mediante `scraping` de la plataforma de *Booking.com*. Esta tarea es importante en el proyecto, ya que permite validar si la imagen que proyectan los hoteles en su información visual está alineada con los servicios que indican en esta plataforma. Dado que muchos servicios se muestran de manera explícita (piscinas, terrazas, gimnasio...), la clasificación visual resulta útil para comprobar la coherencia.

Para abordar esta tarea, se ha decidido utilizar modelos del tipo *Visual Question Answering* (VQA), capaces de responder a preguntas en lenguaje natural condicionadas por el contenido visual de una imagen.

– **Flujo jerárquico en YAML**
Este flujo se ha utilizado en todos los modelos menos en CLIP ya que no era necesario. Para estructurar de forma lógica la clasificación de los servicios se diseñó un archivo específico denominado `flujo.yml`, que contiene una secuencia de pasos condicionales. Cada uno de estos pasos representa una unidad lógica de decisión que el modelo debe resolver mediante una pregunta formulada en lenguaje natural. Este enfoque se inspira en árboles de decisión clásicos, pero adaptado a un contexto multimodal donde la entrada no es una variable estructurada, sino una imagen. Cada paso del flujo incluye cuatro elementos fundamentales:

- Una pregunta en inglés directamente asociada a una categoría visual concreta. Ejemplo: “Is there a swimming pool in the image?”.
- Un conjunto de respuestas válidas habitualmente “yes” y “no”, aunque se adaptan según el tipo de pregunta.
- Un diccionario `next_step`, que determina cuál es el siguiente paso del flujo dependiendo de la respuesta obtenida.
- Una posible asignación directa de categoría, a través del campo `assign_category`, cuando se detecta de forma inequívoca la presencia de un servicio.

El flujo comienza siempre con una pregunta inicial: “¿La imagen representa un espacio interior o exterior?” A partir de esta bifurcación inicial, se recorren rutas distintas que guían al modelo hacia las preguntas posteriores. Por ejemplo, si se detecta que la imagen es exterior, el flujo dirige al modelo hacia pasos relacionados con “Piscinas”, “Área de playa”, etc. En cambio, si se detecta un entorno interior, el flujo conduce hacia “Dormitorio”, “Baño/Aseo” y demás categorías relacionadas. No obstante, hay servicios incluidos en las dos rutas, como puede ser “Restaurante, Buffet o Bar”, “Entretenimiento y Eventos”, etc.

Este diseño aporta varias ventajas clave:

- Minimiza el número total de preguntas por imagen, al evitar formular una cuestión

por servicio.

- Facilita la depuración y trazabilidad del razonamiento, ya que cada imagen sigue un camino en concreto.
- Permite escalabilidad, debido a que, se pueden añadir nuevas categorías o reformulas preguntas sin necesidad de reentrenar modelos o cambiar código.

Por otro lado, se tomó la decisión de formular todas las preguntas en inglés, dado que los modelos a probar fueron entrenados mayoritariamente sobre conjuntos de datos anglófonos como ”COCO“, ”Visual Genome“ o “Conceptual Captions”. En pruebas previas, se observó que realizar las preguntas en castellano degradaba significativamente la calidad de las respuestas. El uso de inglés maximiza, por tanto, la probabilidad de obtener respuestas precisas.

A continuación, se muestra un ejemplo del flujo:

```
- step: "1"
  question: "Does the image show an indoor or outdoor space?"
  valid_answers: ["indoor", "outdoor"]
  next_step:
    indoor: "2_indoor_context"
    outdoor: "2_outdoor_context"

- step: "2_indoor_context"
  question: "Is there a visible bed?"
  valid_answers: ["yes", "no"]
  next_step:
    yes: "assign_bedroom"
    no: "2_indoor_terrace_check"

- step: "assign_bedroom"
  assign_category: "Dormitorio"
  next_step:
    default: "2_indoor_terrace_check"
```

Y por último, se muestra un ejemplo del funcionamiento del flujo:

```
=== Iniciando flujo para imagen: image_1.jpg ===
Paso '1': Pregunta: Does the image show an indoor or outdoor space?
Please answer with: indoor, outdoor
Respuesta cruda: outdoor
next_map: {'indoor': '2_indoor_context',
'outdoor': '2_outdoor_context'}
Paso siguiente: 2_outdoor_context
Paso '2_outdoor_context':
Pregunta: Is there a swimming pool visible?
Please answer with: yes, no
```

```
Respuesta cruda: yes
next_map: {True: 'assign_pools', False: '3_outdoor_beach'}
...
...
Categorías asignadas para image_1.jpg: {'Piscinas'}
```

– **CLIP ViT-B/32 sin flujo**
Con el objetivo de establecer un punto de partida rápido y eficiente, se decidió emplear el modelo `CLIP ViT-B/32`, desarrollado por OpenAI, como base para la detección visual de servicios en alojamientos turísticos. Este modelo no utiliza lógica condicional, es decir, no se utilizará el flujo de preguntas explicado con anterioridad. Debido a que se basa en el cálculo de similitud semántica entre imágenes y textos, lo que permite determinar qué conceptos son detectables directamente a partir del contenido visual de cada imagen.

- **Carga de datos:** En primer lugar, se comenzó con la recopilación automática de imágenes correspondientes a una muestra aleatoria de 30 hoteles, seleccionados de entre los disponibles en el conjunto completo. Para representar los servicios a detectar, se partió del archivo `services_translation`, que contiene la lista completa de los 23 servicios definidos en la tabla 4.1, traducidos al inglés. Esta traducción es necesaria ya que `CLIP` ha sido entrenado casi en su totalidad con datos anglófonos, y su rendimiento disminuye considerablemente con datos en castellano.

- **Inferencia por similitud texto-imagen:** Para cada imagen, utilizando `CLIP` se calculó la distribución de probabilidad condicional sobre los textos, dado el "embedding visual" extraído de la imagen. Internamente, el modelo proyecta tanto la imagen como el texto en un espacio vectorial conjunto y computa la similitud coseno entre el "embedding visual" y cada uno de los "embeddings textuales". Esta similitud se transforma posteriormente en una distribución de probabilidad mediante la aplicación de una función "softmax" sobre las puntuaciones obtenidas. El resultado es un vector de 23 probabilidades, uno por cada servicio, que refleja qué tan probable es que cada clase esté presente en la imagen. Para tomar una decisión binaria, se fijó un umbral de corte de 0,20 de modo que aquellas categorías cuya probabilidad es superior a ese valor, son declaradas como presentes en la imagen, mientras que las demás quedan descartadas. Este umbral fue seleccionado de manera empírica, después de realizar numerosas pruebas, buscando un equilibrio razonable entre precisión (proporción de aciertos entre las predicciones positivas realizadas) y sensibilidad (proporción de aciertos sobre el total de casos reales positivos).

- **Almacenamiento de resultados:** Cada detección positiva fue registrada en un archivo de resultados en formato `CSV (resultados_clip.csv)`, en el que se anotaron los siguientes campos para cada predicción:

  - **Modelo:** indicando que se trata de CLIP.

  - **Ciudad:** nombre de la ciudad donde se ubica el hotel.

- **Hotel:** nombre del alojamiento.
- **Imagen:** nombre del servicio detectado

Este archivo se diseñó para ser fácilmente legible, para calcular posteriormente la precisión y la sensibilidad

- **Evaluación cuantitativa del rendimiento**
  El uso de `CLIP` como modelo sin flujo permitió una evaluación inicial rápida en el conjunto de imágenes. Tuvo una capacidad relativamente alta para detectar servicios visualmente obvios como "Piscina", "Área de playa", "Restaurante, Buffet o Bar". No obstante, se identificaron limitaciones claras:
  - La precisión cualitativa estimada fue del 60 % aproximadamente, evaluada a partir de una revisión manual de los servicios asignados en un subconjunto de imágenes.
  - `CLIP` mostró un rendimiento aceptable para categorías como las nombradas con anterioridad, sin embargo, su desempeño fue mucho más pobre en categorías contextuales o abstractas como "Servicios Adicionales", "Recepción y atención al huésped", "Zona de juegos".

  Por consecuencia, el modelo `CLIP ViT-B/32` proporcionó un buen punto de partida. Su principal virtud fue la rapidez, ya que hizo el procesamiento de los 20 hoteles en 11 minutos y 27 segundos, en el equipo explicado con anterioridad, gracias a la posibilidad de realizar inferencias en lote. Sin embargo, su falta de capacidad para clasificar los servicios en las imágenes motivó la decisión de buscar soluciones más sofisticadas.

– **BLIP-1 VQA Base**
  La primera prueba exhaustiva se llevó a cabo con **BLIP VQA Base**, el modelo original de la "familia" BLIP entrenado para contestar preguntas en lenguaje natural sobre imágenes. Este experimento se realizó para tener una primera toma de contacto con esta familia de modelos, ya que es el menos costoso computacionalmente (214 millones de parámetros) y ofrece una comprensión semántica muy superior a la que se obtiene con modelos como *CLIP*.

  - **Protocolo de inferencia**
    Para realizar este ensayo con `BLIP-1`, se utilizó el flujo explicado anteriormente, sin alterar el preprocesado ni la lógica de ramificación. Cada imagen entra en el modelo como tensor RGB redimensionado a 512 píxeles como máximo, y se formatea en `FP16` para reducir el consumo de `VRAM`. El modelo procesa una única imagen por vez, es decir, `batch = 1`, para no romper la lógica condicional. El modelo genera respuestas limitadas a 30 `tokens`. Este límite, fue suficiente para las respuestas generadas por el modelo, evitando de esta manera, desbordar el contexto interno del decodificador. Las respuestas generadas por `BLIP-1` se sometían a una normalización que comprendía desde conversión a minúsculas, eliminación de signos de puntuación y compactación de espacios. A continuación esa respuesta se compara, con las respuestas válidas especificadas en cada paso del

YAML. Este mecanismo permitió traducir expresiones como “yes, of course”, “no, obvious” a los tokens canónicos “yes” y “no”, minimizando falsas transiciones en el árbol de decisión.

- **Evaluación cuantitativa del modelo**
  Para evaluar el rendimiento del modelo BLIP-1 se llevó a cabo una validación manual sobre el mismo subconjunto de 20 hoteles, escogido en la evaluación de CLIP, para asegurar una medición justa entre modelos. Este proceso de evaluación, se hizo de esta manera ante la imposibilidad de realizar un etiquetado manual de las imágenes en los plazos del proyecto. La metodología empleada consistió en ejecutar el flujo de preguntas para cada imagen con el modelo en cuestión, y registrar las categorías asignadas automáticamente. A continuación, se procedió a una revisión manual de cada imagen, contrastando si los servicios asignados por el modelo estaban efectivamente representados en la imagen. Por ejemplo, si se había detectado la categoría “Piscinas” pero la imagen mostraba un baño interior, se consideraba una predicción incorrecta. En cambio, si se había asignado “Restaurante, Buffet o Bar”, y se observaba una mesa con platos en un salón, se contaba como acierto.

  Partiendo de esta revisión, imagen por imagen, se calcularon dos métricas principales precisión y sensibilidad. La precisión se define como la proporción de categorías asignada por el modelo que fueron consideradas correctas tras la validación manual. Por otro lado, la sensibilidad se define como la proporción de categorías realmente visibles en las imágenes que fueron detectadas de forma correcta por el modelo.

  El modelo BLIP-1 obtuvo una precisión estimada del 68 %, lo que se define como que de cada 100 predicciones realizadas por el modelo en torno a 68 eran efectivamente correctas. Por su parte, la sensibilidad obtuvo un valor notablemente menos, entorno al 60 %, reflejando que el modelo omitía frecuentemente servicios que estaban presentes en la imagen.

  Estos resultados aunque no son perfectos, se han considerado satisfactorios para una primera fase exploratoria con un modelo ligero, antes de probar los modelos más potentes de la familia BLIP.

– **BLIP-2 OPT-2.7B**
Una vez establecida una línea base con BLIP-1, se decidió probar una arquitectura más potente y moderna: BLIP-2 OPT-2.7B. Este modelo combina un codificador visual avanzado con un modelo de lenguaje OPT de 2.7 mil millones de parámetros, ofreciendo una mayor capacidad de razonamiento semántico que BLIP-1, a cambio de un mayor consumo computacional.

- **Proceso de inferencia**
  El proceso de inferencia se diseñó para reutilizar íntegramente el mismo “pipeline” lógico y visual aplicado a BLIP-1. Por ello, se utilizó el mismo procesamiento visual, y el flujo condicional establecido en flujo.yml sin modificar el orden, las preguntas, ni las rutas de ramificación. Al igual que con BLIP-1, el paso de

las imágenes por el modelo se realizo de forma secuencial con *batch = 1*, para no romper la lógica del flujo condicional. También, se normalizaron las respuestas de la misma manera (pasadas a minúsculas, sin tildes ni puntuación, y sin espacios redundantes). Cada imagen producía, como resultado, un conjunto de categorías binarias, en caso de que el servicio esté presente en la imagen se asignará el valor 1, y en el caso negativo valor 0. Cada salida al igual que en los modelos anteriores, quedaba registrada en un CSV con una fila por imagen y 23 columnas (una por servicio visual), junto a los campos ciudad, hotel y número de imagen.

- **Evaluación cuantitativa del modelo**
  Al igual, que con los modelos probados con anterioridad, no se poseía un conjunto de datos etiquetados, por lo que se optó nuevamente por una evaluación visual manual. Se procesaron las imágenes de los mismos 20 hoteles de prueba. Y se utilizaron las mismas métricas, evaluadas de igual manera. Las métricas estimadas fueron las siguientes: la precisión mejoró notablemente, estableciéndose en torno al 81 %, mientras que la sensibilidad se quedó en el 76 %.

- **Análisis general**
  El modelo `BLIP-2 OPT-2.7B` se comportó como una solución muy equilibrada, con mejores resultados que `BLIP-1` pero no mucho coste computacional. No obstante, se va a proceder a probar con una versión mejorada de `BLIP-2`, para comprobar si compensa los resultados con el tiempo de computo.

– **BLIP-2 FLAN-T5-XL**
El último modelo evaluado en esta fase de clasificación visual fue `BLIP-2 FLAN-T5-XL`, una de las variantes más potentes de la familia `BLIP-2`. Este modelo incorpora un decodificador basado en `FLAN-T5`, una arquitectura entrenada específicamente para seguir instrucciones y resolver tareas complejas de razonamiento en lenguaje natural. Su selección respondió a una doble motivación: en primer lugar, evaluar el impacto de un decodificador más competente que `OPT` sobre el mismo flujo visual, y por el otro lado, analizar si el mayor coste computacional se justificaba en términos de precisión.

- **Proceso de inferencia**
  Todo el proceso de inferencia se diseñó para que fuera exactamente idéntico al utilizado con `BLIP-2-OPT-2.7B`, asegurando una comparación directa.

- **Rendimiento y evaluación cualitativa**
  Gracias a la capacidad superior de `FLAN-T5` para interpretar instrucciones complejas y manejar preguntas condicionales con mayor precisión semántica, este modelo ofreció los siguientes resultados: en relación a la precisión el modelo otorgó en torno al 86 %, mientras que la sensibilidad subió hasta un 80 %. Estas cifras reflejan una mejora generalizada respecto a los modelos anteriores, especialmente en categorías difíciles o ambiguas.

4.6.2. Modelado de descripciones

– **Introducción y motivación**
La imagen que los hoteles proyectan en internet no sólo depende de sus fotografías, sino también de las descripciones textuales que acompañan a cada alojamiento. Analizar estos textos permite verificar si los servicios que el establecimiento declara ofrecer, coinciden con lo que el usuario podría esperar tras leer la descripción. Para abordar esta tarea se creo un conjunto de datos supervisado y se entrenaron varios modelos de lenguaje con un enfoque multi-etiqueta.

– **Etiquetado manual y partición del conjunto de datos**
Para entrenar modelos supervisados, fue necesario contar con un conjunto de descripciones anotado. El proceso de etiquetado se realizó íntegramente de forma manual, sin intervención de herramientas automáticas ni más revisores:

- **Lectura individual:** Se leyeron las 550 descripciones una por una.
- **Asignación de etiquetas:** Tras la lectura, se marcó en un archivo `CSV` si la descripción mencionaba de forma literal o claramente implícita cada uno de los 23 servicios definidos en la fase de preparación. El `CSV` resultante contiene 25 columnas, una para el nombre del hotel y otra para la descripción, las 23 restante son columnas binarias (0/1) correspondientes a los servicios.
- **Verificación rápida:** Una vez completado el etiquetado, se revisaron aleatoriamente 50 registros para comprobar consistencia.

Con la matriz ya construida, se generó un split 80/20:

- 440 descripciones para entrenamiento.
- 110 descripciones para prueba.

La estratificación se hizo con "train_test_split" de scikit-learn usando la frecuencia combinada de etiquetas para mantener la misma proporción de servicios en ambos subconjuntos.

– **Estrategia de aumento de datos**
Para ampliar la variedad léxica y mitigar el desequilibrio entre categorías, se diseñó un proceso de "data augmentation" textual. El objetivo era generar versiones alternativas de cada descripción sin alterar el significado esencial, incrementando de esta manera la exposición del modelo a construcciones sintácticas diversas y sinónimos relevantes. El flujo de aumento de datos se articuló en tres etapas:

- **"Back-translation" controlada:** Cada descripción se tradujo al inglés y se volvió a traducir al español. Solo se conservaron las reformulaciones cuya similitud léxica (porcentaje de palabras exactamente iguales que comparten dos textos) superaba el 70 % respecto al texto original. De esta forma se garantizaba fidelidad semántica.
- **Paráfrasis léxica con sinónimos:** Posteriormente, sobre estos ejemplos, se

aplicó un reemplazo de sinónimos extraídos de "WordNet" y validados con representaciones vectoriales "word2vec". El algoritmo sustituyó adjetivos o sustantivos no críticos, preservando términos relevantes para la detección de servicios como piscina, spa, gimnasio, etc. para no introducir ruido.

- **Ajuste de longitud:** Finalmente, se generó una versión abreviada mediante la eliminación aleatoria de una oración no principal, identificada con "spacy". El objetivo era exponer al modelo a "inputs" más cortos sin perder la esencia descriptiva.

En consecuencia, cada descripción original produjo dos variantes sintéticas: una procedente de la traducción y otra de la variante léxica abreviada. El tamaño efectivo del corpus de entrenamiento aumentó de 440 ejemplos a 1320, equilibrando mejor la presencia de categorías minoritarias.

– **RoBERTa-base sin fine-tuning**
Como línea de referencia mínima se utilizó "RoBERTa-base" tal y como se publica en "Hugging Face", es decir, sin realizar ningún tipo de reentrenamiento sobre el corpus de hoteles. El objetivo era medir hasta qué punto un modelo lingüístico, entrenado sobre grandes corpus genéricos, es capaz de reconocer menciones explícitas de servicios mediante "zero-shot inference"

- **Configuración y método**
Antes de lanzar la inferencia se tradujeron al inglés todas las descripciones que estaban originalmente en español, para ello se utilizó el modelo "Helsinki-NLP/opus-mt-es-en". Por otro lado, no se utilizó ningún tipo de aumento de datos en esta fase, ya que la finalidad era medir el rendimiento del modelo de manera "cruda" sobre texto limpio y homogéneo. Una vez traducidas y normalizadas, cada descripción se introdujo en el "pipeline de clasificación zero-shot", configurado con "RoBERTa-base" como modelo subyacente y el modo "multi_label=True". El procedimiento interno del "pipeline" es el siguiente:

  - **Plantillas de hipótesis:** Para cada una de las 23 etiquetas de servicio, se genera una frase hipótesis del tipo "This text is about label".

  - **Parejas premisa-hipótesis:** El modelo recibe, por un lado, la descripción completa como premisa y, por otro lado, cada hipótesis de forma independiente. De esta manera, se crean 23 parejas (premisa + hipótesis) por descripción.

  - **Cálculo de la implicación:** RoBERTa-base genera, para cada una de las parejas premisa-hipótesis, tres puntuaciones diferentes denominadas "logits" que corresponden a los estados de "contradicción", "neutralidad" e "implicación". El "pipeline" se queda con el "logit" de implicación, lo pasa por una función "softmax" y lo normaliza para obtener una probabilidad comprendida entre 0 y 1.

  - **Decisión multi-etiqueta:** Estas 23 probabilidades se comparan con un umbral escogido de manera arbitraria, después de diferentes experimentos, con un valor de 0,45. Si la probabilidad otorgada en el paso anterior a una pareja

supera el umbral, se marca la etiqueta como presente, en el caso contrario, se marca como ausente.

Este enfoque permite asignar varias etiquetas simultáneas a un mismo texto, sin necesidad de reentrenar el modelo.

- **Rendimiento y evaluación cualitativa**
  Aplicado al conjunto de prueba, de manera automática se calcularon varias métricas:
  - **Precisión:** La precisión (medida de cuántas de las etiquetas que el modelo predijo como positivas estaban realmente presentes) otorgada por el modelo fue de un 45 % (valor ponderado según la frecuencia de cada clase; weighted avg).
  - **Sensibilidad:** La sensibilidad (proporción de servicios realmente presentes que el modelo fue capaz de detectar) fue de un 78 % (valor ponderado según la frecuencia de cada clase; weighted avg).
  - **F1-micro:** El "F1-micro" (media armónica de precisión y sensibilidad agregada sobre todas las clases, ponderada por el número total de etiquetas) fue de un 54 % (valor ponderado según la frecuencia de cada clase).

  Estas cifras indican que, sin entrenamiento específico, el modelo acierta aproximadamente la mitad, para dejar escapar esa misma medida de los servicios que aparecen en los textos.
- **Punto de partida**
  Estas métricas no son suficientes para obtener un resultado satisfactorio, por ello se ha decidido avanzar al siguiente paso, que será realizar fine-tuning.

– **RoBERTa-large con fine-tuning** Una vez probado el modelo sin reentrenar, se decidió afinarlo de forma supervisada para evaluar la mejora que aporta el "fine-tuning". Se escogió RoBERTa-large porque, dentro de la familia `RoBERTa`, representa la versión de referencia: supera a BERT al prescindir de los "tokens" segmentadores, utilizar máscaras dinámicas y entrenarse con lotes considerablemente mayores, lo que se traduce en una comprensión contextual más estable.

- **Configuración detallada de entrenamiento**
  Para lograr un equilibrio entre rendimiento y coste computacional se fijaron los siguientes hiperparámetros:
  - **Épocas:** se han realizado 10 épocas, estas han sido suficientes para que la curva de validación convergiera sin riesgo de sobre-ajuste.
  - **Tasa de aprendizaje inicial:** 2×10^{-5} con calentamiento lineal del 10 % de los pasos totales.
  - **Weight-decay:** 0,01 replicando la receta el "paper" original [44]

- **Early-stopping:** se ha utilizado una paciencia de 3 épocas sobre "f1_macro" para detener el proceso si no había mejora real.
- **Regularización:** dropout del 10 % en la capa de clasificación y un recorte del gradiente a 1,0.

- **Convergencia y estabilidad**
  Durante las 10 épocas de ajuste, el valor de la pérdida binaria descendió con rapidez en las dos primeras pasadas por los datos y después se estabilizando, bajando poco a poco entre las épocas. La pérdida de validación se mantuvo en todo momento por debajo de la de entrenamiento, con una diferencia máxima de 0,015 puntos, señal de que el modelo no llegó a sobre-ajustar el corpus. La supervisión intermedia también reveló que los gradientes permanecieron en rangos moderados gracias al recorte del gradiente a 1, lo que previno explosiones numéricas y permitió mantener estable el aprendizaje.
- **Rendimiento y evaluación cualitativa**
  Tras finalizar el ajuste, se evaluó el rendimiento sobre las descripciones del conjunto de prueba. Las métricas agregadas fueron:
  - **Precisión:** la precisión obtenida fue del 74 %
  - **Sensibilidad:** la sensibilidad fue del 72 %
  - **F1-micro:** la media armónica entre las dos métricas anteriores, resultaron en un 70 % de "f1-micro", ponderada por el número total de etiquetas, es decir, favorece clases frecuentes.
  - **F1-macro:** esta métrica hace referencia al promedio simple de "f1" por categoría. Y el resultado fue del 61 %.

  La diferencia entre los valores de "F1-micro" y "F1-macro" pone de manifiesto que el modelo mantiene un rendimiento elevado en aquellas categorías más frecuentes, mientras que sigue encontrando dificultades con los servicios menos representados en el conjunto de entrenamiento. A pesar de ello, la mejora observada con respecto al modelo sin entrenamiento previo es significativa, lo que confirma que el proceso de "fine-tuning" no solo es útil, sino claramente necesario para alcanzar un rendimiento competitivo en todas las clases.

– **Motivación del uso de modelos en español**
Aunque los modelos anteriores, se construyeron en torno a descripciones uniformadas en inglés, lo cierto es que el 100 % de los textos utilizados se encontraban en español. Por ello, se consideró fundamental realizar experimentos con modelos nativos en castellano, con el objetivo de preservar al máximo el contexto lingüístico original. Traducir automáticamente textos turísticos puede acarrear pérdidas de matices importantes, como pueden ser topónimos, expresiones culturales, etc. Estos matices, un modelo preentrenado únicamente en inglés difícilmente podría entenderlos a la perfección. En este sentido, se decidió incorporar modelos como RoBERTa-BNE-base y RoBERTa-BNE-large, ambos preentrenados sobre grandes corpus en español, por instituciones como

la “Biblioteca Nacional de España o el “Barcelona Supercomputing Center”. Su entrenamiento específico permite al modelo captar referencias locales y construcciones semánticas propias del idioma, lo que resulta especialmente útil en textos redactados por hoteles canarios o peninsulares.

A pesar de haber trabajado directamente con texto en castellano, se mantuvo la estrategia de aumento de datos por traducción controlada. Es decir, se aplicó una traducción al inglés seguida de una re-traducción al español, para generar reformulaciones sintácticas diversas, sin alternar el contenido semántico. Esta técnica de aumento de datos no tiene como objetivo cambiar el idioma de trabajo, sino aumentar la variedad expresiva a partir del texto original. De hecho, se ha comprobado que esta diversidad mejora especialmente el rendimiento en categorías poco frecuentes.

– **RoBERTa-BNE-base**
Tras los buenos resultados obtenidos con RoBERTa en inglés, se decidió entrenar una versión en castellano que partiera de un preentrenamiento adaptado al contexto lingüístico del proyecto. Para ello se eligió “RoBERTa-BNE base”. Este modelo utiliza la misma arquitectura que RoBERTa con 125 millones de parámetros. Por tanto, la configuración de entrenamiento, fue muy parecida al modelo explicado con anterioridad.

  - **Configuración de entrenamiento**
    El “fine-tuning” sobre “RoBERTa-BNE-base” se ha llevado a cabo con los mismos parámetros que el modelo entrenado con anterioridad “RoBERTa-large”.
  - **Resultados cuantitativos**
    La evaluación sobre el conjunto de prueba mostró los siguientes resultados:
    - **Precisión:** 81 %.
    - **Sensibilidad:** 82,0 %.
    - **“F1-micro”:** 81 %
    - **“F1-macro”:** 75 %

    Este rendimiento posiciona a “RoBERTa-BNE base” como un modelo útil y utilizable.

No obstante, se ha decidido probar con un modelo más potente, para mejorar el rendimiento.

– **RoBERTa-BNE-large**
Tras los buenos resultados obtenidos con RoBERTa-BNE base, se decidió entrenar la versión “RoBERTa-BNE large”, que cuenta con 355 millones de parámetros. Esta versión también fue desarrollada por la “Biblioteca Nacional de España”, utilizando el mismo corpus pero con una mayor capacidad. El objetivo principal de este experimento era comprobar si el aumento de capacidad del modelo mejoraba de manera rentable el rendimiento. Se mantuvo el mismo “pipeline” de preprocesamiento, división de datos y aumento de datos que en el modelo español, descrito anteriormente.

- **Configuración de entrenamiento**
  El modelo fue entrenado con las mismas condiciones que los dos modelos anteriores.
- **Rendimiento y evaluación cuantitativa:** Se utilizó el mismo conjunto de prueba que con los modelos anteriores, para que la comparación fuera justa y los resultados fueron los siguientes:
  - **Precisión:** 84 %
  - **Sensibilidad:** 88 %
  - **F1-micro:** 85 %
  - **F1-macro:** 75 %

  Los resultados en relación al rendimiento mejoran notablemente respecto al modelo anterior. Por ello, se ha decidido no seguir realizando experimentos.

4.6.3. Modelado de comentarios de clientes

A diferencia de las descripciones o las imágenes, los comentarios de los clientes plantean un reto adicional, debido a que su redacción es informal, muchas veces estas reseñas poseen faltas de ortografía, abreviaciones, expresiones coloquiales e incluso palabras propias del español hablado. Por ello, se optó por no traducir los comentarios al inglés, ya que se perdería una parte importante del contexto semántico y cultural. En lugar de ello, se entrenaron dos modelos en castellano: "RoBERTa-BNE base" y "RoBERTa-BNE large" explicados con anterioridad.

- **Proceso de etiquetado semi-supervisado**
  Antes de entrenar los modelos, fue necesario construir un conjunto de entrenamiento con etiquetas. Dado el gran volumen de comentarios disponibles y su naturaleza, se optó por una estrategia semi-supervisada. Para ello, se llevó a cabo una lectura intensiva de miles de comentarios, identificando patrones léxicos y términos clave asociados a cada uno de los 23 servicios definidos en el proyecto. A partir de este análisis se construyó un diccionario de palabras clave por categoría denominada "word_map.json", que asocia a cada servicio un conjunto de expresiones representativas, por ejemplo, para el servicio "piscinas" se han elegido las siguientes palabras: "piscina", "piscinas", "piscina cubierta", "piscina al aire libre", "piscina climatizada", "piscina infinita". Este diccionario sirvió como base para realizar un etiquetado de forma automática, para todos los comentarios extraídos de "Booking.com". Cada comentario se analiza individualmente, en busca de coincidencias con el "JSON", y se le asignan las categorías correspondientes en el caso de detectar un término relevante.

  Este procedimiento permitió generar un conjunto de datos, con etiquetado automático asistido, mucho más eficiente que la anotación completamente manual de más de 19.000 comentarios. Sin embargo, puede introducir cierto ruido, pero se comprobó que la calidad era suficientemente buena, como para realizar entrenamientos efectivos.

Por otra parte, para aquellos comentarios que no hacían alusión directa a ningún servicio en concreto pero incluían la palabra "todo", se optó por asignar todos los servicios existentes para el hotel en el dataset principal "hotels.json". Una lógica similar se aplicó cuando en el comentario se encuentra la palabra "instalaciones", no obstante, para este caso se definió un subconjunto de servicios considerados habitualmente como parte de las instalaciones físicas del establecimiento, debido a que esta palabra no engloba características, por ejemplo "Deportes en la naturaleza", y los servicios de esta lista se comparaban con los existentes en el "JSON" que contiene los datos de los hoteles. Estas medidas evitaron perder información con este tipo de comentarios.

- **Entrenamiento** Con el conjunto de datos semi-etiquetado, se procedió a entrenar dos modelos distintos:
  - **RoBERTa-BNE base:** arquitectura más ligera.
  - **RoBERTa-BNE large:** modelo más completo y profundo, con mayor representación contextual.

  En relación al "pipeline" utilizado, fue diferente que con las descripciones hoteleras. El lenguaje utilizado en los comentarios es significativamente más informal y heterogéneo, con una alta presencia de errores ortográficos, expresiones coloquiales y referencias implícitas. Por este motivo, no se aplicó el mismo flujo de preparación ni las técnicas de aumento de datos empleadas en la sección de descripciones. Con el conjunto de datos etiquetados, se procedió a entrenar los dos modelos nombrados, no sin antes dividir el conjunto de datos en entrenamiento y prueba, en un 80 % y un 20 %, con una arquitectura multi-etiqueta y funciones de pérdida basadas en entropía binaria. El entrenamiento se realizó desde cero, sin aprovechar pasos previos de los modelos entrenados sobre descripciones, debido a la diferencia sustancial en estilo y estructura de los textos. Cada comentario fue tratado como una instancia independiente, sin necesidad de traducción ni reformulación, aprovechando el potencial de los modelos.

- **Rendimiento y evaluación cuantitativa**
  La fase de evaluación se llevó a cabo sobre un conjunto amplio de comentarios previamente etiquetados.
  - **RoBERTa-BNE base:** El modelo base, registró los siguientes resultados:
    - **Precisión ("weighted"):** 89 %
    - **Sensibilidad ("weighted"):** 94 %
    - **"F1-score" ("weighted"):** 91 %

    A pesar de que la versión base entrega métricas sólidas se ha decidido experimentar con la versión más completa de este modelo:
  - **RoBERTa-BNE large:** El modelo base, registró los siguientes resultados:
    - **Precisión ("weighted"):** 89 %
    - **Sensibilidad ("weighted"):** 95 %

- **"F1-score" ("weighted"):** 91 %

Con estos resultados, el modelo más completo confirma que no es necesario seguir con la experimentación de modelos en esta fase.

- **Análisis de sentimientos**
  También se implementó un análisis de sentimientos, que utiliza el analizador en español de **"pysentimiento"** para clasificar cada comentario como positivo, neutro o negativo. Se debe resaltar, respecto a esta clasificación, que los comentarios durante el proceso de "scraping", se consiguieron con la etiqueta "positiva" y "negativa"; no obstante, estas etiquetas originales se descartaron para este proceso, debido a que los usuarios de las plataformas online, marcaban erróneamente su sentimiento, por ejemplo, seleccionaban el comentario positivo, pero lo acababan con numerosas quejas sobre el establecimiento. Una vez etiquetados los comentarios, se transforma la salida en una métrica numérica comprensible tanto a nivel global como por servicio, el flujo utilizado para ello fue el siguiente:

  - **Predicción del sentimiento:** cada reseña pasa por el modelo de análisis de sentimientos que etiqueta el modelo como se dijo anteriormente, en positivo, neutro o negativo, y se calcula cuánta confianza tiene en su decisión. Solo se toman en cuenta las clasificaciones de confianza media-alta, las más dudosas se dejan como neutras para no sesgar el resultado.
  - **Nota global (0-10):** en este paso se calcula la media de esos valores y se convierte a la escala 0-10 (-1 pasa a 0, y +1 pasa a 10). El resultado es una única nota que resume cómo se sienten, en general, los clientes sobre el hotel.
  - **Desglose por servicio:** como cada comentario ya estaba asociado a los servicios que menciona, se repite el mismo cálculo pero por categoría. De esta forma se ve con rapidez qué áreas reciben mejor o peor valoración.

  En conclusión, este pipeline convierte opiniones desestructuradas en las siguientes métricas claras:

  - **Nota global del hotel.**
  - **Nota detallada por servicio.**

4.7. Evaluación

Una vez completada la fase de modelado para cada uno de los tres ejes del proyecto, se procede a presentar y analizar los resultados obtenidos en cada uno de ellos. Esta sección tiene como objetivo principal comparar de manera cuantitativa el rendimiento de los distintos modelos aplicados, así como valorar su idoneidad en función del tipo de datos, el coste computacional y la precisión alcanzada en tareas de clasificación de servicios hoteleros. Para cada modalidad se han empleado métricas estándar como precisión, sensibilidad y "F1-score" en sus variantes

micro, macro y balanceada, también se han medido los tiempos de inferencia. El análisis, al igual que el modelado, se ha estructurado en tres bloques:

- **Clasificación visual:** evaluación de los modelos "CLIP", "BLIP-1", "BLIP-2-OPT" y "BLIP-2-FLAN-TL X5", midiendo su capacidad para identificar servicios directamente representados en imágenes.
- **Clasificación textual en descripciones:** análisis de la eficacia de modelos del lenguaje ("RoBERTa", y variantes "BNE") sobre un conjunto etiquetado manualmente de más de 500 descripciones.
- **Clasificación a partir de comentarios:** evaluación de modelos entrenados sobre un conjunto semi-supervisado generado mediante reglas lingüísticas.

A lo largo de esta sección se compararán los resultados obtenidos, destacando los puntos fuertes y las limitaciones de cada enfoque, y justificando la elección de los modelos finales utilizados para cada tarea. En primer lugar, se procederá a analizar los resultados relacionados con las imágenes:

4.7.1. Evaluación de la clasificación de imágenes

Para evaluar la capacidad de detección visual de servicios hoteleros, se analizaron los 4 modelos nombrados con anterioridad. Todos ellos, menos "'CLIP", integraron el mismo flujo de inferencia basado en "YAML" y procesaron un conjunto homogéneo de imágenes correspondientes a 20 hoteles, seleccionados de forma aleatoria, pero siendo el mismo conjunto para todos los modelos. Debido a la inexistencia de un "dataset" de imágenes etiquetadas de forma manual, se optó por una validación visual manual. Tras ejecutar el "pipeline" con cada modelo, se revisó el resultado almacenado como un archivo csv, contrastando las categorías asignadas de manera individual. Esta evaluación permitió estimar de forma robusta dos métricas fundamentales.

- **Precisión:** proporción de categorías asignadas que fueron correctas (verdaderos positivos sobre predicciones positivas).
- **Sensibilidad:** proporción de categorías visuales presentes que fueron correctamente identificadas por el modelo.

Adicionalmente, se recogió el tiempo total de ejecución de cada modelo sobre el conjunto completo de 20 hoteles (612 imágenes) en el equipo nombrado anteriormente con una gráfica "NVIDIA RTX 4060 TI" de 8GB de VRAM, con el objetivo también de valorar la viabilidad práctica en términos de coste computacional. Los resultados obtenidos se muestran en la siguiente tabla:

Tal y como se aprecia, "CLIP" fue el modelo más rápido en procesar las imágenes, completando la inferencia en menos de 12 minutos. Sin embargo, su precisión y sensibilidad fueron las más bajas, lo que evidencia su debilidad en tareas que requieren razonamiento e identificación de elementos visuales menos explícitos.

Modelo	Precisión (%)	Sensibilidad (%)	Tiempo (min)
CLIP ViT-B/32	60	51	**11.4**
BLIP-1 VQA Base	68	60	43
BLIP-2 OPT-2.7B	81	76	196
BLIP-2 FLAN-T5-XL	**86**	**80**	217

Tabla 4.3: Resultados de precisión, sensibilidad y tiempo de ejecución.

El modelo "BLIP-1" ofreció una mejora notable respecto a "CLIP", con una precisión del 68 %. Esta mejora se debe a que "BLIP-1" está específicamente entrenado en tareas de "Visual Question Answering", lo que le permite interpretar mejor la imagen y asociar preguntas a respuestas semánticamente coherentes. No obstante, su capacidad es limitada en servicios ambiguos o parcialmente visibles.

Los mayores avances se lograron con las variantes de "BLIP-2". "BLIP-2 OPT-2.7B" alcanzó un 81 % de precisión y un 76 % de sensibilidad, lo que representa una mejora significativa respecto a los modelos analizados con anterioridad. Su arquitectura avanzada y el uso de un decodificador "OPT" permitió identificar de manera más precisa múltiples servicios. Además, su tiempo de inferencia (196 minutos para 20 hoteles) lo convierte en una opción razonable en términos de coste-beneficio.

Finalmente, el modelo "BLIP-2 FLAN-T5-XL" se posicionó como el mejor en rendimiento puro, alcanzando un 86 % de precisión y un 80 % de sensibilidad. Su decodificador basado en "FLAN-T5" mostró una compresión excelente de las instrucciones condicionales del flujo, mejorando especialmente en categorías difíciles. Sin embargo, su coste computacional fue el más elevado, con un tiempo de ejecución de 217 minutos.

Teniendo en cuenta el equilibrio entre precisión, capacidad de razonamiento y cobertura de servicios visuales, se seleccionó finalmente "BLIP-2 FLAN-T5-XL" como modelo para la clasificación de imágenes. Aunque su tiempo de inferencia fue el más elevado, los resultados obtenidos justifican plenamente esta inversión computacional, debido a que superó a todas las demás alternativas tanto en aciertos directos como en la detección de servicios complejos.

Por tanto, pese al mayor coste computacional, se considera que "BLIP-2 FLAN-T5-XL" ofrece la mejor relación calidad-rendimiento dentro del marco del proyecto. Y su adopción garantiza una mayor coherencia en la identificación de servicios visuales, reforzando de esta manera la precisión global del proyecto.

4.7.2. Evaluación de la clasificación de descripciones

Para evaluar la capacidad de los modelos lingüísticos para detectar menciones de servicios hoteleros en texto libre, se analizaron los cuatro modelos descritos en el modelado. Todos ellos procesaron el mismo conjunto de prueba para evaluar el rendimiento, garantizando de esta manera la comparabilidad entre modelos. Este conjunto de prueba, constaba del 20 % de las muestras totales que eran 550, es decir, 110 muestras. A diferencia del experimento de

la fase anterior de imágenes, no fue necesaria una revisión humana posterior; las métricas se calcularon automáticamente, comparando las etiquetas predichas con las etiquetas reales. Se midieron tres indicadores fundamentales:

- Precisión
- Sensibilidad
- "F1-micro"

Los resultados obtenidos se muestran en la siguiente tabla:

Modelo	**Prec(%)**	**Sens(%)**	**F1-micro(%)**	**Tiempo ent.**	**Tiempo eti.**
RoBERTa-base (zero-shot)	45	78	54	0 m	**9 s**
RoBERTa-large + FT	74	72	70	190 m	15 s
RoBERTa-BNE-base + FT	81	82	81	145 m	12 s
RoBERTa-BNE-large + FT	**84**	**88**	**85**	**227 m**	17 s

Tabla 4.4: Resultados de precisión, sensibilidad, F1-micro, tiempo de entrenamiento y tiempo de etiquetado para la clasificación de descripciones de hoteles.

Tal y como se aprecia, "RoBERTa-base sin entrenamiento" ofrece la línea de base mínima, ya que no requiere de entrenamiento y etiqueta las descripciones en apenas 9 segundos, pero su precisión del 45 % revela que solo identifica la mitad de los servicios. La traducción previa al inglés elimina expresiones culturales claves, de modo que el modelo confunde servicios sutiles y tiende a sobrepredecir los más evidentes.

"RoBERTa-large" con "fine-tuning" supone un salto cualitativo respecto a la versión "zero-shot" de "RoBERTa-base": tras algo más de 3 horas de entrenamiento, su "F1-micro" sube en 16 puntos, alcanzando un 70 %. Aún así, la obligación de traducir las descripciones al inglés sigue lastrando el resultado, ya que se omiten matices culturales y se observa un claro rendimiento desigual con los modelos entrenados en castellano.

En relación a los modelos nativos en castellano, la mejora es inmediata. "RoBERTa-BNE-base" necesita un tiempo de entrenamiento moderado y eleva notablemente la precisión y la sensibilidad hasta el 81 % y 82 % respectivamente, consiguiendo un "F1-micro" del 81 %. Al estar preentrenado con grandes corpus en español, interpreta expresiones propias del entorno turístico nacional sin recurrir a traducciones, lo que reduce la ambigüedad y mejora la detección de servicios menos habituales. El etiquetado posterior (15 segundos) se mantiene dentro de márgenes operativos exigentes, aunque el modelo todavía muestra cierta variabilidad en frases muy largas o cargadas de detalles.

Por su parte, "RoBERTa-BNE-large" es el más sólido del conjunto. Tras un entrenamiento de unas cuatro horas, alcanza un 84 % de precisión y un 88 % de sensibilidad, con un "F1-micro" del 85 %. Al trabajar directamente en español y disponer de mayor capacidad, capta relaciones complejas entre servicios y mantiene un equilibrio estable entre clases comunes y poco representadas. El coste computacional es mayor, pero el etiquetado sigue siendo ágil (17 segundos) y el rendimiento adicional justifica la inversión.

Por tanto, se llega a la conclusión de que los modelos entrenados en español superan en rendimiento a sus homólogos basados en traducción, porque preservan el contexto original, evitando pérdidas semánticas y reconocen mejor los términos hoteleros. Dentro de ellos, "RoBERTa-BNE-large" combina la mayor cobertura con una precisión elevada, por lo que se adopta como solución definitiva para la clasificación textual del proyecto.

4.7.3. Evaluación de los modelos de comentarios

Para valorar la capacidad de los modelos lingüísticos en un contexto ruidoso como es el de los comentarios de los usuarios, se utilizó el conjunto de datos etiquetado de manera semi-automática, este conjunto se dividió en entrenamiento y prueba, dejando un 20 % para este último. Este porcentaje de datos se utilizó para evaluar el rendimiento en los dos modelos experimentados.

Modelo	**Prec (%)**	**Sensibilidad (%)**	**F1-micro**	**Tiempo ent. (min)**
RoBERTa-BNE base	89	94	91 %	**125**
RoBERTa-BNE large	**89**	**95**	**91 %**	260

Tabla 4.5: Métricas de desempeño y tiempo de entrenamiento para los modelos de análisis de comentarios.

Tal y como se observa en la tabla, el modelo "RoBERTa-BNE-base" constituye un primer punto de referencia del funcionamiento de esta familia de modelos, para el análisis de comentarios informales. Tras algo más de dos horas de ajuste, este modelo llegó a unos resultados de un 89 % de precisión y un 94 % de sensibilidad, con un "F1-micro" de un 91 %. Estas cifras confirman que un modelo de tamaño medio es capaz de absorber correctamente la ortografía irregular y el tono coloquial de las reseñas. Sin embargo, el modelo tiende a equivocarse con la clasificación de servicios minoritarios, ya que hay pocos ejemplos en el conjunto de datos.

Por su parte, "RoBERTa-BNE-large" eleva mínimamente el rendimiento de los modelos desarrollados por la "Biblioteca Nacional de España" gracias a su mayor capacidad y sus casi 4 horas de entrenamiento. Las métricas mejoran solo en un 1 % de sensibilidad; sin embargo, esta pequeña mejora muestra que el modelo se ajusta mejor a etiquetas menos frecuentes.

Con estos resultados, se llega a la conclusión de que el modelo "RoBERTa-BNE-large" satisface los requisitos de calidad y estabilidad con un mayor tiempo de entrenamiento, el cual se ha aceptado ya que solo se realiza una vez, de modo que se adopta como solución definitiva para el módulo de comentarios, sin necesidad de explorar arquitecturas adicionales.

4.8. Despliegue

Como parte final del proyecto, se ha implementado una aplicación web interactiva, desplegada en local, desarrollada con el "microframework" "Flask", que actúa como interfaz principal

para la exploración visual y analítica de los resultados generados por los modelos elegidos en la fase anterior. El objetivo fundamental de esta aplicación es facilitar la evaluación de la coherencia entre los servicios promocionados por los hoteles y la percepción real de los clientes.

4.8.1. Arquitectura y diseño

La aplicación sigue el patrón de arquitectura "Modelo-Vista-Controlador", lo que permite separar de manera clara las responsabilidades de cada parte del sistema, facilitando así el mantenimiento y la escalabilidad del proyecto:

- **Modelo:** Representa la lógica de negocio y está compuesto por módulos de análisis basados en modelos avanzados de inteligencia artificial. Concretamente, los módulos son:
  - **comments_services.py:** Encargado de la detección automática de los servicios mencionados en las reseñas de los usuarios mediante el modelo "RoBERTa-BNE-large", reforzado por reglas semánticas especiales para asignaciones más precisas.
  - **bert_description.py:** Utiliza el modelo "RoBERTa-BNE-large", entrenado con los datos provenientes de las descripciones, para detectar servicios en las descripciones.
  - **blip2_flow.py:**: Aplica un modelo de tipo "Visual Question Answering" (BLIP-2 + Flan-T5) para interpretar automáticamente las imágenes subidas por los clientes y asignarles etiquetas relacionadas con los servicios ofrecidos por el hotel.
  - **sentimientos.py**: Utiliza el paquete "pysentimiento" para evaluar la polaridad emocional de los comentaros asociados a cada servicio.
  - **graficado.py**: Recopila los resultados de los ficheros anteriores y calcula métricas de coherencia además de realizar visualizaciones claras mediante gráficos de puntos y circulares.
- **Vista:** Está compuesta por dos plantillas HTML:
  - **index.html**: La página inicial del sistema, que presenta un formulario intuitivo donde el usuario selecciona un lugar, un hotel en específico y los módulos (comentarios, descripciones e imágenes) que desea analizar.
  - **result.html**: Página dinámica generada tras la ejecución de los módulos seleccionados, que muestra tablas organizadas, gráficas claras y resultados numéricos interpretables, así como información detallada sobre el sentimiento global y específico por servicio.
- **Controlador:** Está representado por el archivo principal ``app.py''. Este archivo gestiona las peticiones "HTTP" del usuario, recopila información proporcionada por los formularios, invoca los distintos módulos de análisis según corresponda, coordina

el flujo de datos entre ellos y entrega los resultados organizados a la vista para su representación final.

4.8.2. Funcionamiento general

Al ingresar a la aplicación, el usuario visualiza una página principal con una interfaz intuitiva y sencilla mostrada en la imagen 4.8.

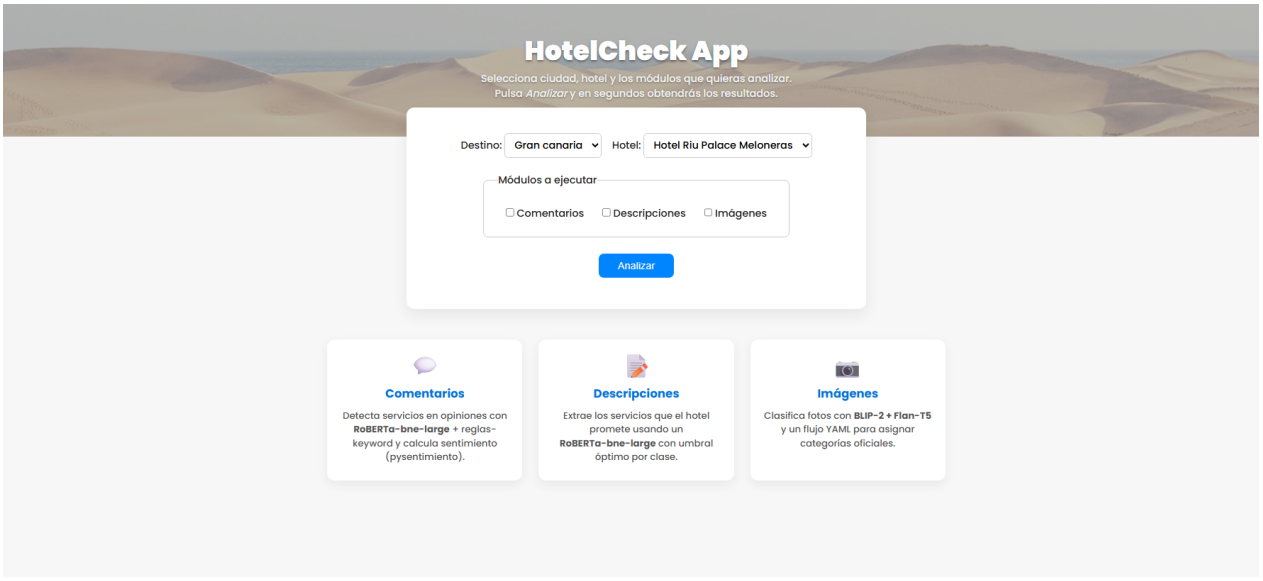


Figura 4.8: Interfaz principal de la aplicación

Aquí se puede elegir el destino y el hotel específico sobre el que se quiere realizar el análisis. Además, el usuario puede seleccionar los módulos que desea evaluar (comentarios, descripciones o imágenes). Esta selección se realiza cómodamente mediante menús desplegables y casillas de verificación etiquetadas. Esto se muestra en las imágenes 4.9, 4.10 y 4.11.

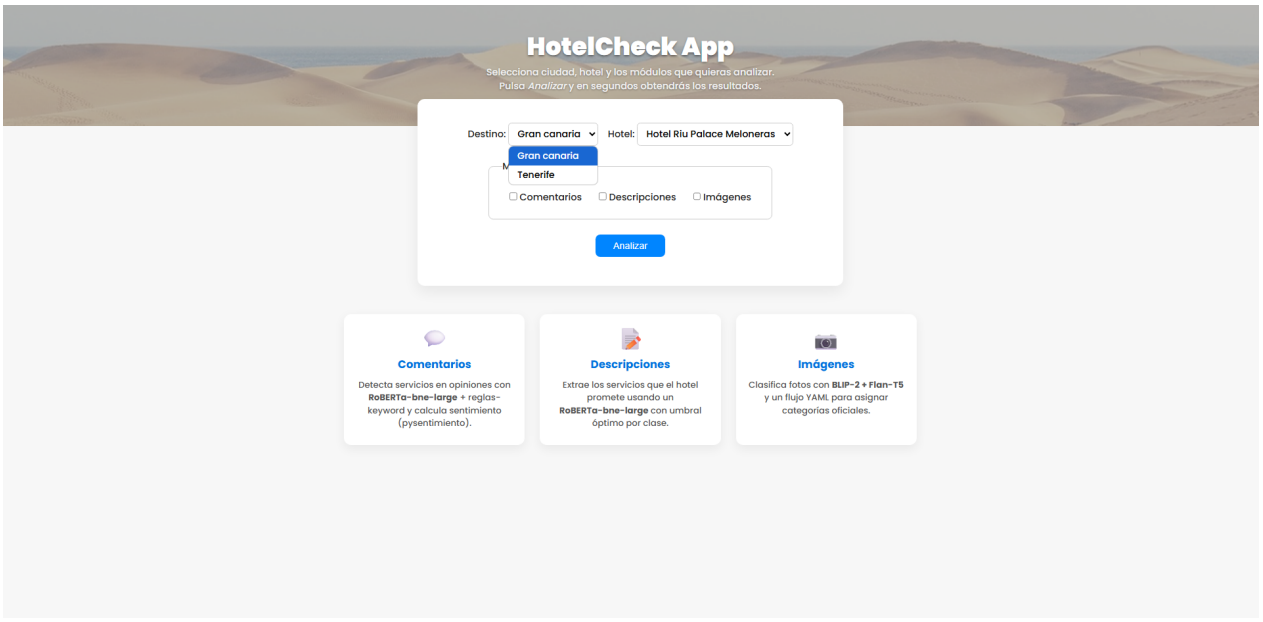


Figura 4.9: Desplegable para elegir destino

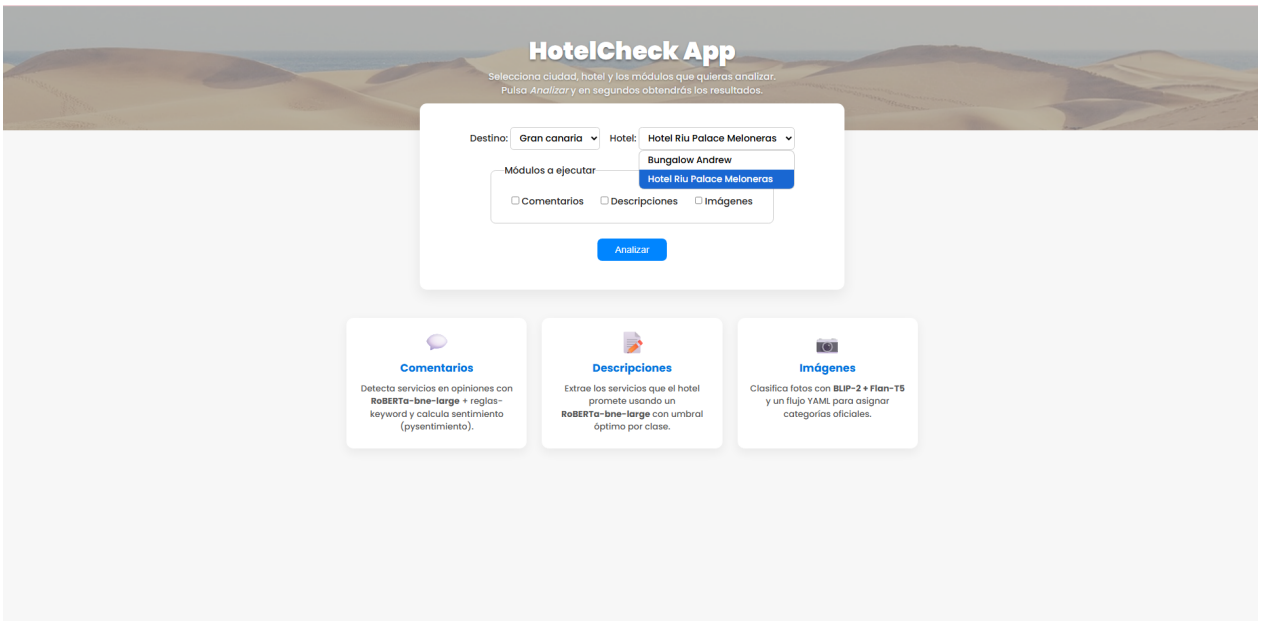


Figura 4.10: Desplegable para elegir hotel

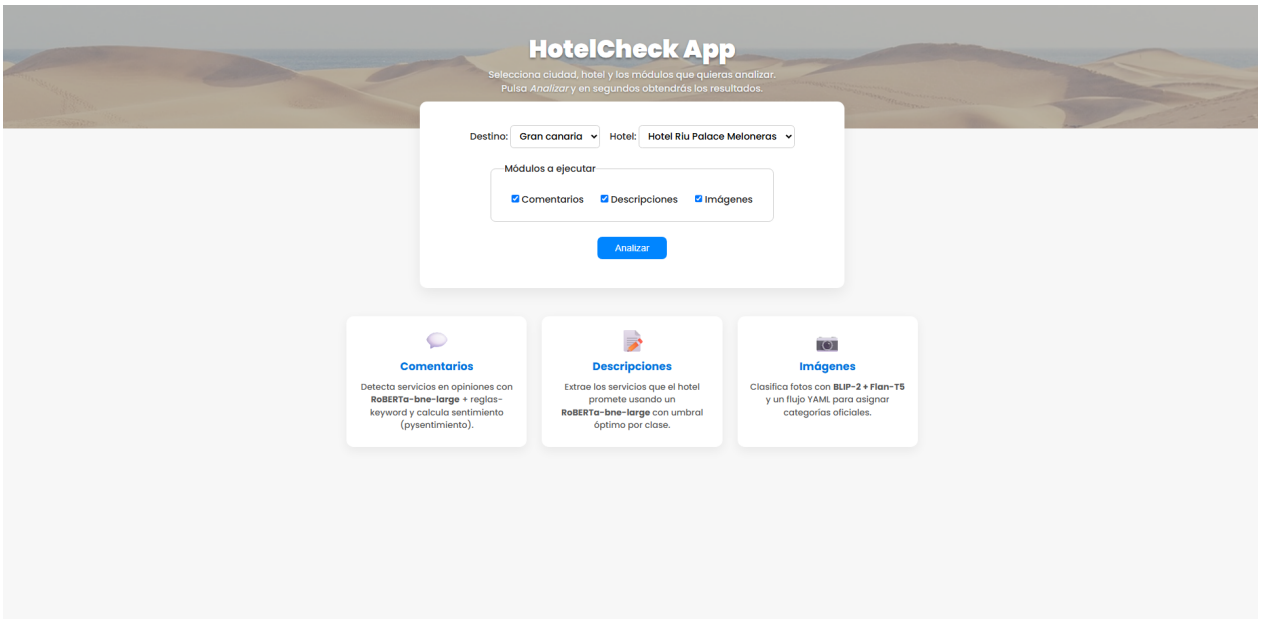


Figura 4.11: Selección de utilidades

Una vez pulsado el botón “Analizar”, la aplicación realiza el análisis solicitado y presenta los resultados.

4.8.3. Módulo de imágenes

Al activar el módulo de imágenes, los resultados del modelo “BLIP-2” se agrupan y visualizan de tres maneras principales. En primer lugar, se presenta una tabla que recoge los servicios detectados y su frecuencia, es decir, cuántas veces aparece cada servicio en el conjunto de imágenes del hotel analizado, como se puede visualizar en la figura 4.12. En segundo lugar, se muestra una gráfica de dispersión (gráfico de puntos) que compara, para cada servicio, si está detectado en las imágenes, si está ofertado oficialmente en la ficha del hotel, o si ambas condiciones se cumplen. Las coincidencias se agrupan en la parte superior del eje vertical, seguidas de los servicios que únicamente aparecen en las imágenes y, por último, aquellos que figuran solo en el listado oficial. Por último, se incluye un gráfico circular que representa la proporción global de coincidencias, sólo detectados y sólo ofertados, como se observa en la figura 4.13.

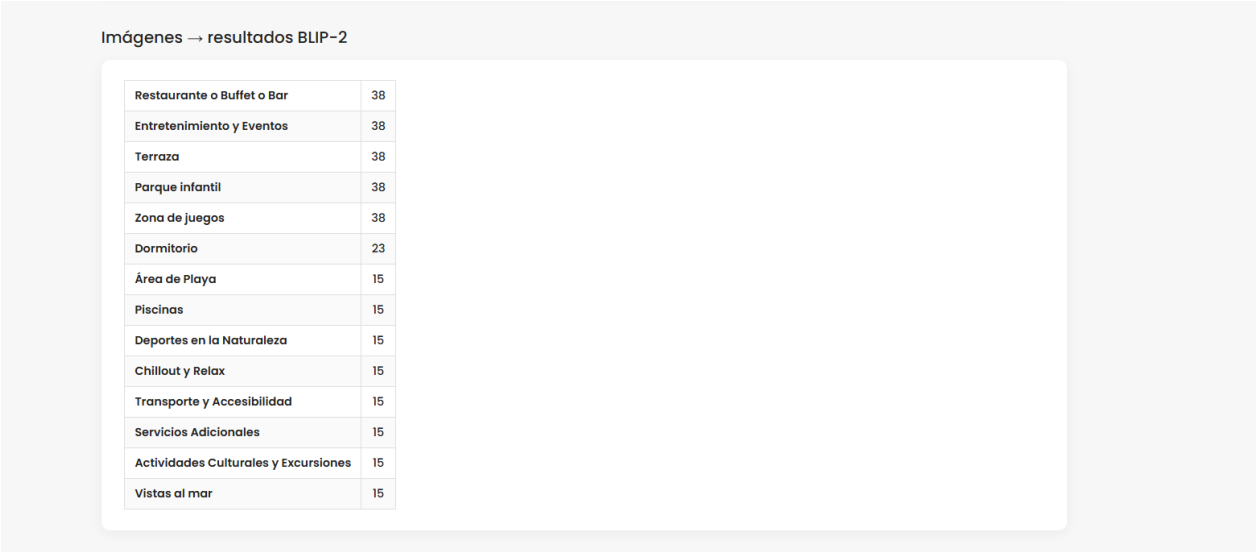


Figura 4.12: Número de apariciones de cada servicio en las imágenes

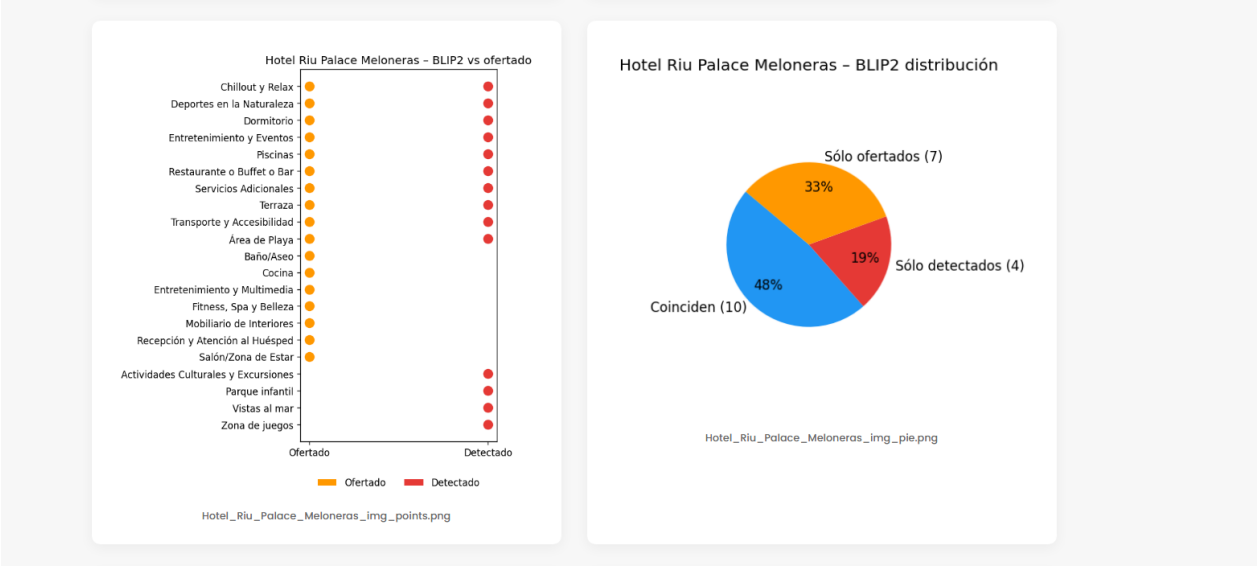


Figura 4.13: Gráficos relacionados con las imágenes

4.8.4. Módulo de descripciones

Al seleccionar este módulo, la aplicación analiza las descripciones de los hoteles y extrae los servicios presentes en ella. Los resultados se presentan en tablas y gráficas, al igual que en el módulo anterior, tal y como se puede ver en la figura 4.14 que muestra todos los servicios detectados, y la figura 4.15 en la que se pueden visualizar las diferentes gráficas generadas en la aplicación.

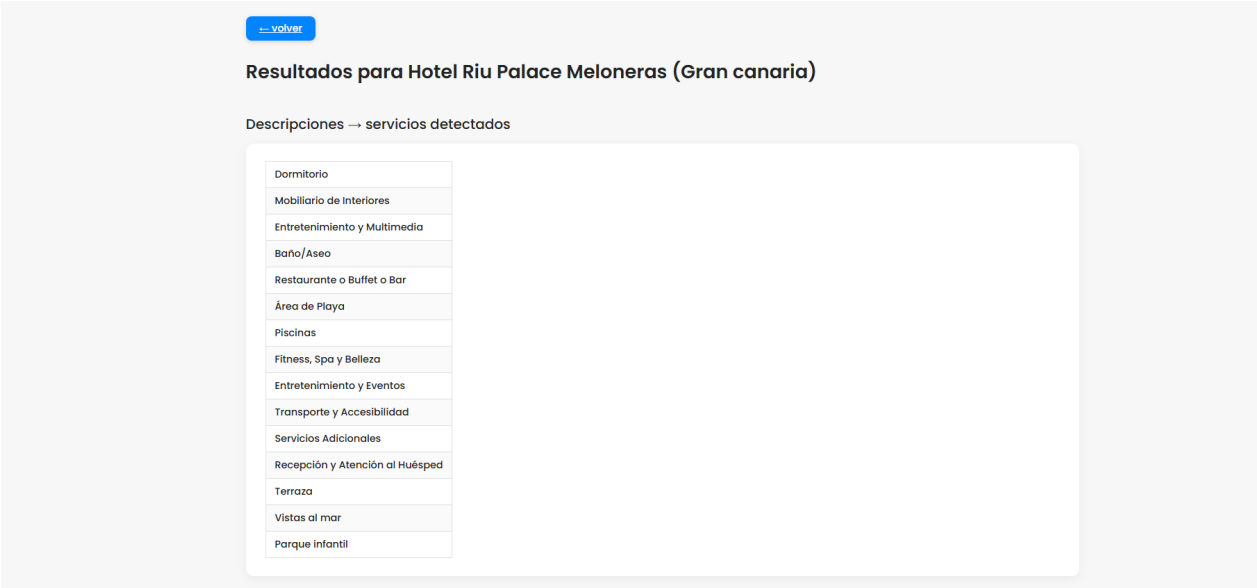


Figura 4.14: Servicios detectados en las descripciones

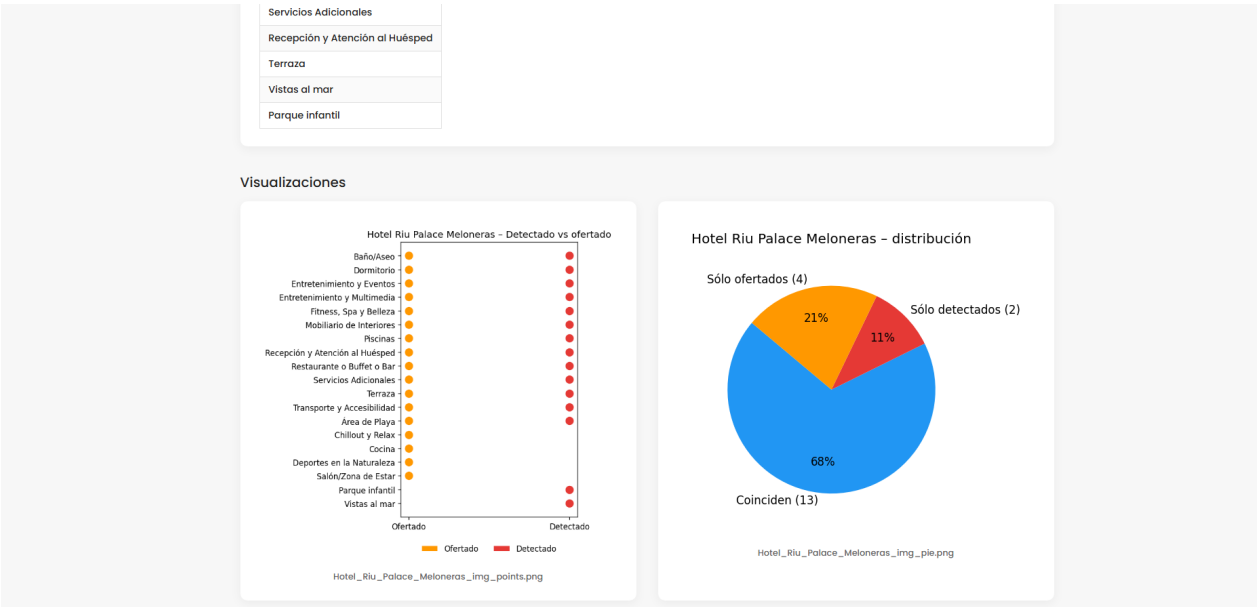
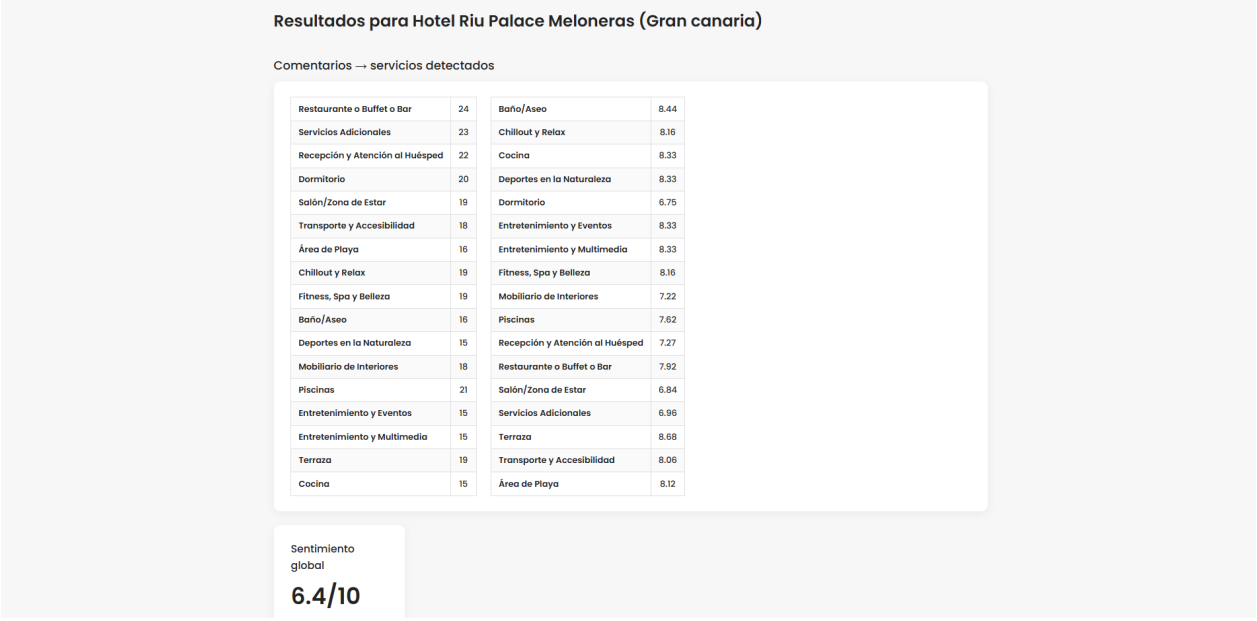


Figura 4.15: Gráficas relacionadas con las descripciones

4.8.5. Módulo de comentarios

Al activar el módulo de comentarios, la aplicación identifica automáticamente los servicios mencionados en los comentarios de los clientes y genera una tabla que muestra los servicios detectados con su frecuencia de aparición. Asimismo, la aplicación también muestra una evaluación del sentimiento global asociado a dichos comentarios en una escala de 0 a 10,

además de calcular y presentar la nota de satisfacción de los clientes por cada servicio en una tabla, tal y como se puede observar en la figura 4.16



Resultados para Hotel Riu Palace Meloneras (Gran canaria)

Comentarios → servicios detectados

Restaurante o Buffet o Bar	24	Baño/Aseo	8.44
Servicios Adicionales	23	Chillout y Relax	8.16
Recepción y Atención al Huésped	22	Cocina	8.33
Dormitorio	20	Deportes en la Naturaleza	8.33
Salón/Zona de Estar	19	Dormitorio	6.75
Transporte y Accesibilidad	18	Entretenimiento y Eventos	8.33
Área de Playa	16	Entretenimiento y Multimedia	8.33
Chillout y Relax	19	Fitness, Spa y Belleza	8.16
Fitness, Spa y Belleza	19	Mobiliario de Interiores	7.22
Baño/Aseo	16	Piscinas	7.62
Deportes en la Naturaleza	15	Recepción y Atención al Huésped	7.27
Mobiliario de Interiores	18	Restaurante o Buffet o Bar	7.92
Piscinas	21	Salón/Zona de Estar	6.84
Entretenimiento y Eventos	15	Servicios Adicionales	6.96
Entretenimiento y Multimedia	15	Terraza	8.68
Terraza	19	Transporte y Accesibilidad	8.06
Cocina	15	Área de Playa	8.12

Sentimiento global

6.4/10

Figura 4.16: Tablas de resultados relacionadas con los comentarios

Estos resultados también se visualizan mediante gráficos adicionales, que comparan los servicios detectados en los comentarios con los servicios ofrecidos en la ficha del hotel. Esto se puede observar en la siguiente figura 4.17

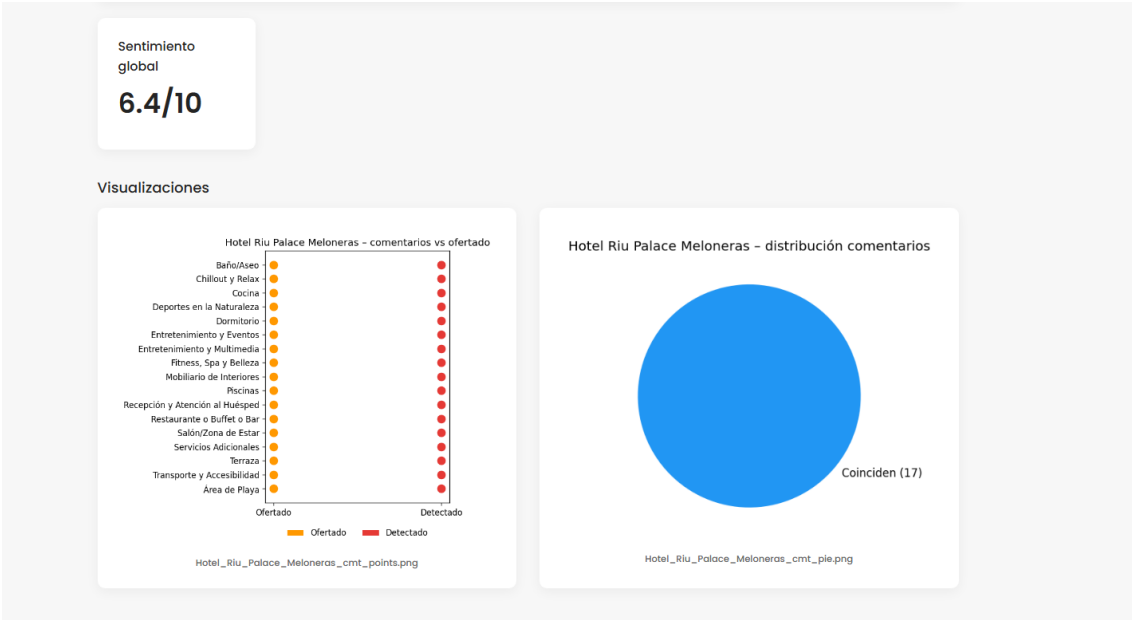


Figura 4.17: Gráficas relacionadas con los comentarios

Capítulo 5

Resultados

A semejanza de la estructura metodológica utilizada en capítulos anteriores, los resultados se presentan en 3 bloques complementarios.

- El primero se centra en la **evidencia visua**l, comparando los servicios que el hotel publica en **"Booking.com"** con lo que realmente aparece en las fotografías procesadas por **"BLIP-2"**, ofreciendo una nota de coherencia y un desglose servicio a servicio.
- El segundo bloque, evaluará la **consistencia textual** de las descripciones de los alojamientos, mediante el modelo **"RoBERTa-BNE-large"**, asignando también una nota de coherencia y varias gráficas.
- Por último, se presenta un bloque, dedicado a los **comentarios de los clientes**, que contrastará la percepción de los huéspedes con un análisis de sentimiento, con detección de servicios. Otrorgando una nota de satisfacción de los clientes, desglosada por cada servicio.

Para este proceso, se ha elegido de manera aleatoria un hotel de Gran Canaria; la única premisa utilizada era que debía poseer más de 40 comentarios. Con ello, se ha elegido el hotel **"Hotel Riu Palace Meloneras"**.

5.1. Coherencia basada en imágenes

El "Hotel Riu Palace Meloneras" sirve de ejemplo ilustrativo porque concentra la casuística más habitual.

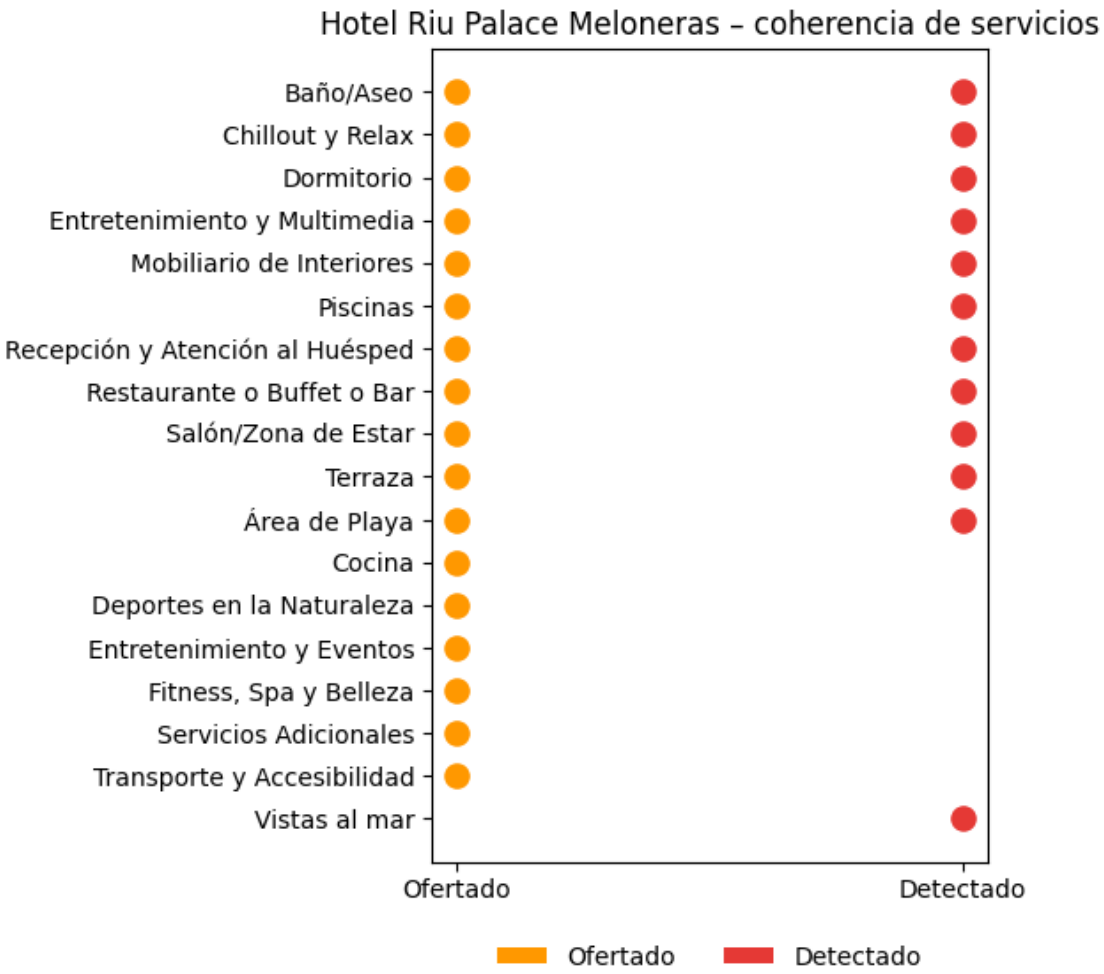


Figura 5.1: Ofertado-Detectado (imágenes)

En la figura 5.1, se muestra la matriz "Ofertado-Detectado", con dos columnas fijas, naranja para los servicios que la cadena anuncia y rojo para los servicios detectados por el modelo, y los 23 servicios listados en el eje vertical. En primer lugar, se observan 11 filas que poseen ambos valores, indicando que el hotel los ofrece y el modelo los detecta. Sin embargo, hay otras 5 filas que exhiben únicamente el punto naranja, ya que son servicios presentes en la ficha comercial del establecimiento, que no han quedado reflejados en la galería del hotel. También existe el caso inverso, al poder visualizar que hay un servicio "Vistas al mar" que aparece sólo marcado en rojo; por tanto, el hotel ofrece este servicio, pero no lo refleja en su ficha comercial.

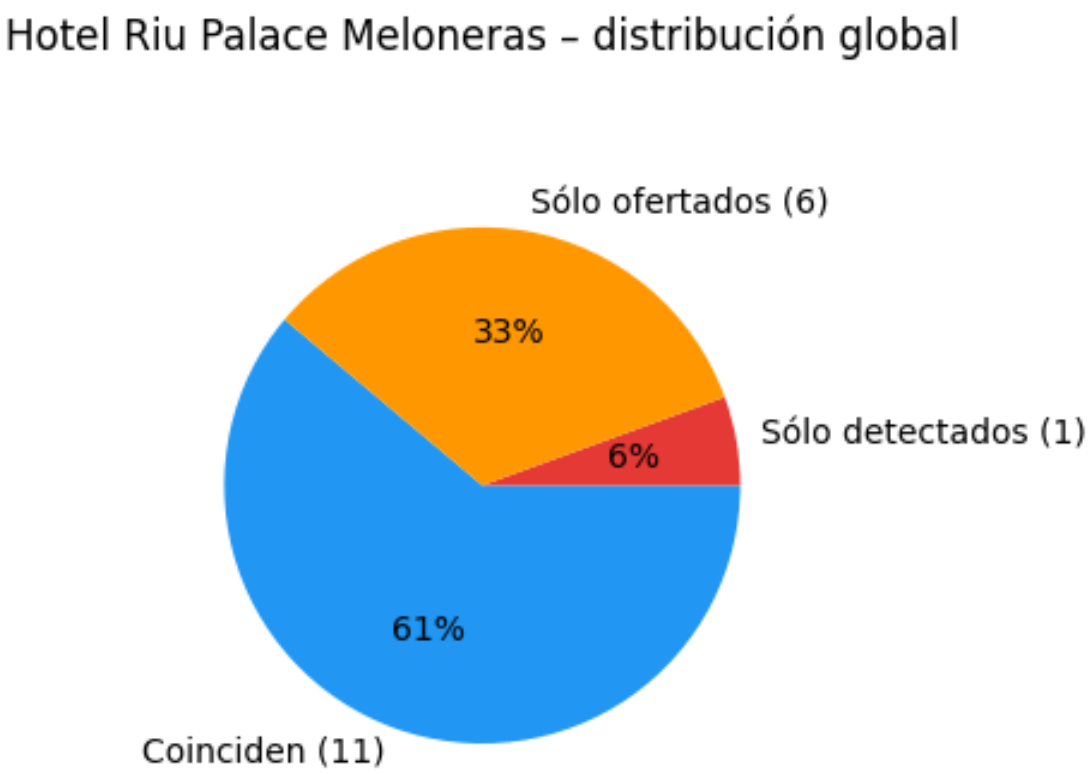


Figura 5.2: Diagrama de sectores (imágenes)

Esta información se ve mejor reflejada en la figura 5.2, donde el diagrama de sectores cuantifica las tres situaciones anteriores: donde existe un 61 % de coincidencias, un 33 % de servicios ofertados pero que no se reflejan en las imágenes y un 6 % de servicios detectados en las imágenes pero no aparecen ofertados por el establecimiento.

Métrica	**Valor (%)**
Precisión	91.67
Cobertura	64.71
F1	75.86
Jaccard	61.11
Nota (0–10)	**7.59**

Tabla 5.1: Resumen de métricas para el módulo de imágenes del *Hotel Riu Palace Meloneras*.

Por último, en la tabla 5.1 generada, se cuantifican con cuatro métricas clásicas:

- **Precisión:** indica la fiabilidad del modelo, en este caso el 91,62 % de los servicios que "BLIP-2" marca aparecen efectivamente en la oferta del hotel.
- **Cobertura o "recall":** esta métrica muestra que el 64,71 % de los servicios anunciados se ven reflejados en la fotografías.
- **F1:** F1 es la *media armónica* entre precisión y cobertura (75,86 %), por lo que penaliza en cuanto uno de los dos extremos se queda corto.
- **Jaccard:** Mide la fracción de elementos compartidos entre dos conjuntos A y B

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

de modo que $J = 0$ indica ningún solapamiento y $J = 1$ conjuntos idénticos.

Finalmente, el sistema transforma el valor de F1 a una escala de 0–10 multiplicándolo por diez y redondeando a dos decimales. Bajo este criterio, el *Hotel Riu Palace Meloneras* obtiene una calificación visual de 7,59, lo que indica que el modelo apenas incurre en falsos positivos; sin embargo, todavía queda en torno a un tercio de los servicios anunciados sin respaldo fotográfico.

5.2. Coherencia en las descripciones hoteleras

Para ilustrar el comportamiento del modelo seleccionado en este bloque, se ha llevado a cabo las mismas gráficas que en el apartado anterior, haciendo uso del mismo hotel.

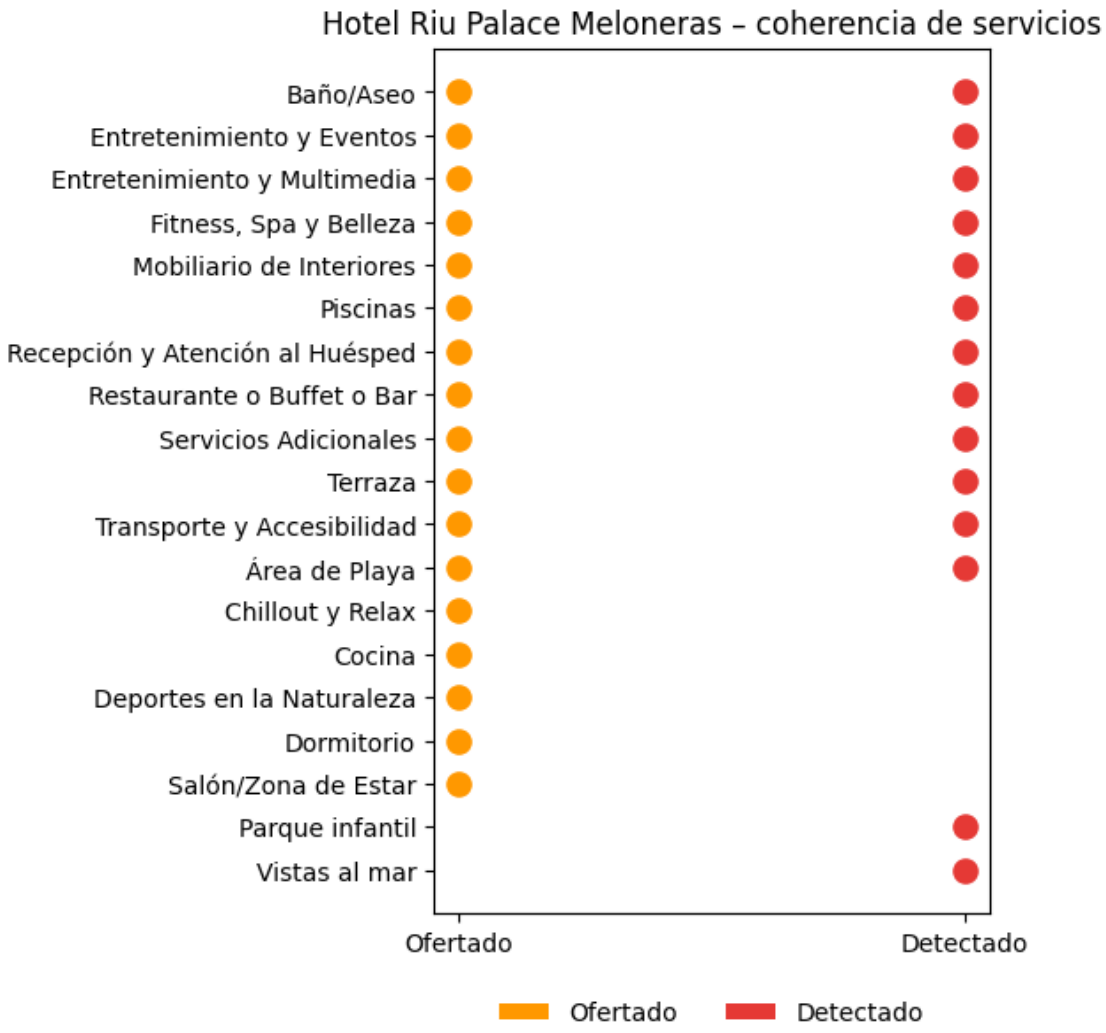


Figura 5.3: Ofertado-Detectado (descripciones)

Al igual que en el bloque visual, en la figura 5.3 cada servicio se representa con dos columnas, "ofertado" y "detectado". En este caso, existen 12 servicios coincidentes y solo 5 que muestran únicamente el punto referente a las ofertas del hotel. No obstante, lo más interesante son los dos servicios que solo se detectan pero no aparecen en la ficha del hotel; se confirma que el servicio "Vistas al mar", está presente en las instalaciones, pero el establecimiento lo ignora por completo.

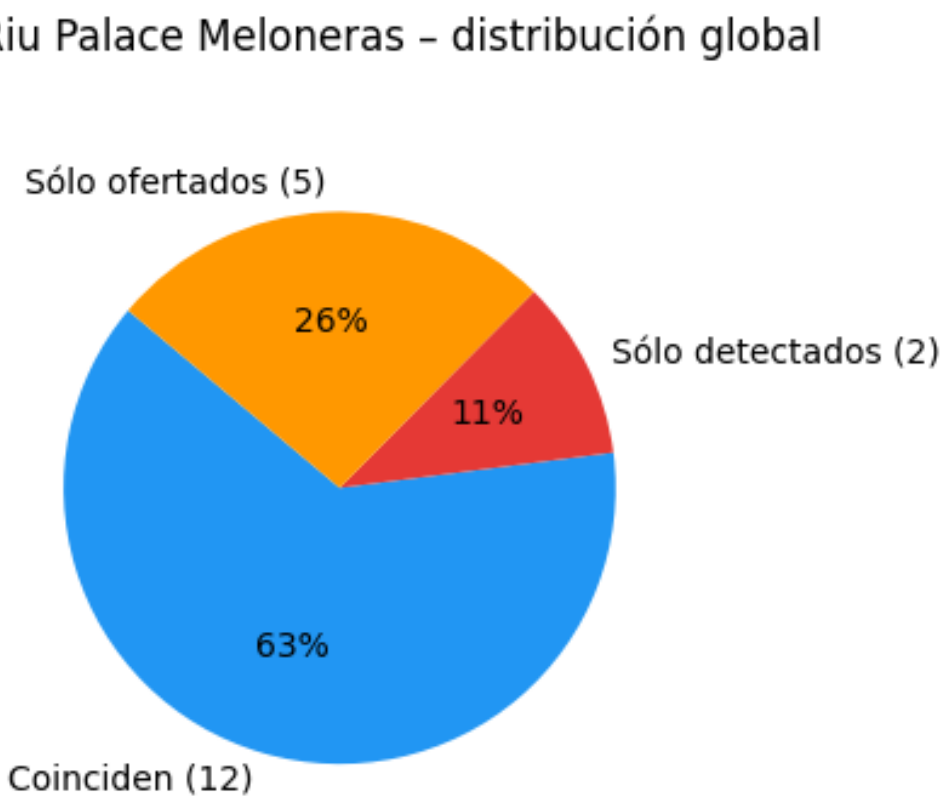


Figura 5.4: Ofertado-Detectado (descripciones)

El diagrama de sectores de este bloque representado en la figura 5.4, confirma los datos anteriores, cuantificando estos resultados en porcentajes.

- 63 % de coincidencias.
- 26 % de sólo ofertados.
- 11 % de sólo detectados.

Por último, la tabla de valores resultantes es la siguiente:

Métrica	**Valor (%)**
Precisión	85.71
Cobertura	70.59
F1	77.42
Jaccard	63.16
Nota (0–10)	**7.74**

Tabla 5.2: Métricas del módulo de descripciones para el *Hotel Riu Palace Meloneras*.

La precisión elevada mostrada en la tabla 5.2 confirma que el modelo rara vez muestra servicios inexistentes; sin embargo, la cobertura revela que casi un 30 % de lo anunciado queda descrito con tal ambigüedad que el modelo no lo reconoce, o simplemente no aparece en la descripción. El F1, se coloca en 0,774 y, en la escala definida para el proyecto, se traduce en una nota de 7,74 sobre 10.

Por consecuencia, el hotel ofrece un discurso textual bastante alineado con su catálogo, pero convendría revisar la redacción de los cinco servicios que sólo aparecen en la columna "ofertado" para asegurar que el vocabulario comercial coincide con el que los huéspedes y los modelos visualizan en las descripciones.

5.3. Opinión de los clientes vía reseñas

En este último bloque, se han analizado las reseñas del hotel tratado en este apartado. Cada texto se procesó con el modelo del lenguaje natural, comentado y elegido para esta tarea anteriormente, otorgando una lista de servicios presente, además de también pasar por un análisis de sentimiento para otorgarle una nota.

Servicio	**Nota 0–10**
Terraza	8.68
Baño/Aseo	8.44
Cocina	8.33
Deportes en la Naturaleza	8.33
Entretenimiento y Eventos	8.33
Entretenimiento y Multimedia	8.33
Fitness, Spa y Belleza	8.16
Chillout y Relax	8.16
Área de Playa	8.12
Transporte y Accesibilidad	8.06
Restaurante o Buffet o Bar	7.92
Piscinas	7.62
Recepción y Atención al Huésped	7.27
Mobiliario de Interiores	7.22
Servicios Adicionales	6.96
Salón/Zona de Estar	6.84
Dormitorio	6.75
Nota global (comentarios)	**6.40**

Tabla 5.3: Valoración media (escala 0–10) por servicio en los comentarios del *Hotel Riu Palace Meloneras*. La nota global se calcula directamente sobre la puntuación de cada reseña, por lo que no equivale al promedio simple de las notas de servicio.

En primer lugar, conviene aclarar que la nota global (6,4) mostrada en la tabla 5.3 no es la media aritmética de las notas por servicio. El cálculo de esta nota se realiza directamente sobre los comentarios. Esta nota asignada globalmente puede distar de manera notable de la evaluación de los servicios individualmente, debido a que un comentario muy positivo puede hacer alusión a una gran cantidad de servicios. Y como las notas por servicio se ponderan por el número de menciones que reciben, las categorías pueden absorber más comentarios positivos y, por tanto, mejorar esa nota respecto a la global del hotel.

Una vez aclarada esta cuestión, en esta tabla se puede observar que las notas de los servicios no varían en más de 2 puntos, otorgando una buena media, aunque existe una gran capacidad de mejora.

Capítulo 6

Conclusiones y Trabajo Futuro

6.1. Conclusiones

Este proyecto, ha mostrado que la Inteligencia Artificial, especialmente los modelos del lenguaje y los modelos multimodales, pueden ayudar y contribuir de manera notable a mejorar la imagen proyectada que los hoteles muestran en internet.

Tras analizar el estado del arte y seleccionar la metodología "CRISP-DM", se construyó un conjunto de datos propio con 550 hoteles españoles, con más de 19.000 imágenes y reseñas. Esta fase ha sido esencial en el proyecto, pues ha permitido, entrenar y comparar, numerosos modelos para las diferentes necesidades del proyecto.

En el módulo visual, **"BLIP-2 FLAN-T5-XL"** permitió alcanzar una gran precisión, permitiendo a los hoteles visualizar de manera rápida y sencilla, cuáles de sus servicios quedaban excluidos de las imágenes que los clientes veían para realizar la reserva en su alojamiento, así como, permitía a la dirección del hotel ver qué servicios estaban presentes en las imágenes pero ellos no tenían incluidos en la ficha de servicios. Para los clientes, esto se traduce en mayor transparencia, ya que insta a los hoteles a mejorar los servicios mostrados en las imágenes que inspiran su decisión de comprar, al reflejar con fidelidad la oferta real del establecimiento, aumentando la confianza en el momento de formalizar la reserva.

Por otro lado, en el módulo referente a las descripciones, el uso del modelo "RoBERTa-BNE-large" revela a la dirección hotelera qué servicios están mencionados en las descripciones y cuáles se diluyen en formulaciones ambiguas o directamente se omiten. Para la gestión del hotel, esta lectura funciona como un corrector semántico, ya que señala qué servicios, se deben incluir en el texto para que todas las características del hotel queden reflejadas con la terminología adecuada. Desde la perspectiva del futuro huésped, el beneficio es inmediato, al encontrar descripciones claras, completas y coherentes con el catálogo ofrecido por el alojamiento, ya que esto permite que el usuario pueda evaluar la propuesta del establecimiento sin malentendidos, fortaleciendo la percepción de profesionalidad antes de confirmar la reserva.

Por último, el análisis de comentarios aporta una capa de valor decisiva: convierte cada reseña

en una puntuación numérica que refleja tanto la satisfacción global del huésped como su valoración específica de cada servicio. En lugar de un promedio, el hotel recibe un panel dónde se muestra que por ejemplo, "Piscinas" roza la excelencia, mientras que "Zona de juegos" tiene una gran capacidad de mejora. Este enfoque, permite a los alojamientos orientar la inversión y ver qué servicios poseen la mayor capacidad de mejora posible. Al mismo tiempo, los huéspedes ganan voz en la mejora continua de los establecimientos, porque sus opiniones en vez de diluirse en internet, dejan un rating automático que guía de manera concreta al hotel, para realizar mejoras.

6.2. Trabajo futuro

La herramienta desarrollada demuestra que es posible auditar, de forma automática, la coherencia entre la información que el hotel publica y la experiencia que describen los huéspedes. Aun así, existen muchos puntos de mejora:

- **Etiquetado manual de imágenes:** Construir un conjunto de datos etiquetado para afinar al máximo "BLIP-2" mediante *"fine-tuning"* supervisado y reducir el número de falsos negativos en servicios poco visibles
- **Vinculación con indicadores de negocio:** Relacionar la nota de coherencia con KPIs reales (reservas, cancelaciones, reclamaciones) para cuantificar el retorno económico de corregir inconsistencias en la imagen digital.
- **Detección temporal de cambios:** Programar re-escaneos mensuales y destacar variaciones significativas en la coherencia, alertando al hotel cuando una reforma, un servicio nuevo o la rotación de fotografías altere la percepción digital.
- **Explicabilidad y confianza:** Integrar mapas de calor en las fotos y resaltar *tokens* clave en los textos para que el gestor vea qué parte de la evidencia apoya cada predicción.
- **Agregación multicanal:** Incluir fuentes adicionales para comparar la coherencia entre plataformas y detactar inconsistencias específicas de cada escaparate digital.
- **Recomendador de mejoras fotográficas:** Aplicar visión por computador para analizar la composición de la imagen y proponer directrices que aumenten la tasa de conversión de las imágenes.
- **Generación automática de textos:** Integrar un módulo que, tras comparar la descripción con los servicios ofertados, genere nuevas descripciones de forma automática para compensar aquellos servicios no detectados.

Bibliografía

[1] World Travel & Tourism Council, "Travel & tourism economic impact research 2024: Global trends," 2024.

[2] UN Tourism, "International tourism recovers pre-pandemic levels in 2024," 2024.

[3] World Travel & Tourism Council, "Travel & tourism sector to create 14m additional jobs in 2025," 2025.

[4] World Economic Forum, "Travel tourism development index 2024: Advancing sustainable and digital transformation," 2024.

[5] World Travel & Tourism Council, "New travel & tourism social impact report reveals significant increase in female employment," 2023.

[6] Instituto Nacional de Estadística, "Estadística de movimientos turísticos en fronteras (frontur). diciembre 2024 y año 2024. datos provisionales," 2025.

[7] Ministerio de Industria y Turismo – Turespaña, "El empleo turístico continúa al alza y alcanza 2,72 millones de afiliados en noviembre de 2024," 2024.

[8] Instituto Nacional de Estadística, "Encuesta de gasto turístico (egatur). diciembre 2024 y año 2024. datos provisionales," 2025.

[9] Aena, S.M.E., "Los aeropuertos del grupo aena registran más de 369,4 millones de pasajeros en 2024," 2025.

[10] Renfe Operadora, "Renfe cierra 2024 con récord histórico al contabilizar más de 537 millones de viajeros," 2025.

[11] Turespaña, "España, hub estratégico para las principales compañías de cruceros," 2025.

[12] Instituto Nacional de Estadística, "Encuesta de ocupación en alojamientos turísticos extrahoteleros. abril 2025," 2025.

[13] Nexotur, "El 90 % de los españoles reserva hoteles de forma online," 2024.

[14] Ministerio de Industria y Turismo, "El ministerio de industria y turismo ha invertido ya más de 225m€ en digitalización de destinos y empresas turísticas," 2025.

[15] Comisión Nacional de los Mercados y la Competencia, "Resolución del expediente s/0031/21 booking," 2024.

[16] Hotel Tech Report, "The 2025 state of hotel guest tech report unveils key trends shaping the future of hospitality," 2024.

[17] P. Cuesta-Valiño, S. Kazakov, and P. Gutiérrez-Rodríguez, "The effects of the aesthetics and composition of hotels' digital photo images on online booking decisions," *Humanities and Social Sciences Communications*, vol. 10, p. 59, 2023.

[18] SHR Group, "Hotel content mistakes that cost you bookings and how ai fixes them," 2025.

[19] SiteMinder, "Hotel reviews: How to manage online guest reviews at your property," 2024.

[20] L. Tang, X. Wang, and E. Kim, "Predicting conversion rates in online hotel bookings with customer reviews," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 4, 2022.

[21] B. Bramwell and L. Rawding, "Tourism marketing images of industrial cities," *Annals of Tourism Research*, vol. 23, no. 1, 1996.

[22] A. D. A. Tasci and W. C. Gartner, "Destination image and its functional relationships," *Journal of Travel Research*, vol. 45, no. 4, 2007.

[23] Z. Li and K. Zhang, "Seeing is believing? visual discrepancy between managerial and user-generated images and its impact on hotel ewom," *International Journal of Hospitality Management*, vol. 113, p. 103529, 2023.

[24] Escuela de Ingeniería Informática, Universidad de Las Palmas de Gran Canaria, "Memoria del plan de estudios del grado en ciencia e ingeniería de datos," 2020.

[25] Tripadvisor Business Insights, "Photos on tripadvisor: Best practices for accommodations," 2024.

[26] Expedia Group Media Solutions, "Destination marketing guide 2025 – image video best practices," 2025.

[27] R. Zhang, Q. Guo, Q. Li, N. Hu, and D. D. Zeng, "Expectation vs. reality: Predicting satisfaction with real-world experiences using online review images," *arXiv preprint arXiv:2011.01434*, 2020.

[28] R. Filieri and F. McLeay, "What makes online reviews helpful? a diagnosticity-adoption framework to explain informational and normative influences in e-wom," *Journal of Business Research*, vol. 68, no. 6, pp. 1261–1270, 2015.

[29] Z. Xiang, Z. Schwartz, J. Gerdes, and M. Uysal, "Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews," *International Journal of Hospitality Management*, vol. 66, pp. 57–67, 2017.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, 2019.

[32] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[34] H. Temiz, "Automatic and accurate classification of hotel bathrooms from images with deep learning," *International Journal of Engineering Research and Development*, vol. 14, no. 3, pp. 211–218, 2022.

[35] T. B. e. a. Brown, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[36] J. Li, M. Shou, N. Aafaq, and G. Chechik, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[37] J.-B. e. a. Alayrac, "Flamingo: a visual language model for few-shot learning," *arXiv preprint arXiv:2204.14198*, 2022.

[38] H. Liu, C. Zhang, H. Hu, and Z. Wang, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[39] Marriott International, "Homes villas by marriott bonvoy makes finding a vacation home easy with generative ai search tool," 2024.

[40] IHG Hotels & Resorts, "Ihg hotels resorts builds a new travel planner powered by google cloud ai," 2024.

[41] Skift, "Inside accor's tech strategy: Loyalty, ai and agility at scale," 2025.

[42] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018.

[43] C. Guerra, "Procesamiento del lenguaje natural con transformers: Tema 5 - modelos del lenguaje," 2024. Accedido: 20-Junio-2025.

[44] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[45] G. d. E. Plan de Tecnologías del Lenguaje (PlanTL), Secretaría de Estado de Digitalización e Inteligencia Artificial, "Plantl-gob-es roberta-base-bne y roberta-large-bne," 2022.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[48] J. Li, D. Li, S. Savarese, and S. Chuang, "Bootstrapping language-image pre-training for unified vision-language understanding and generation," 2022.

[49] P. Chapman, J. Clinton, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, *CRISP-DM 1.0: Step-by-Step Data Mining Guide*. CRISP-DM Consortium, 2000.

[50] Kenneth Jensen, "Crisp-dm process diagram," 2012.