Diseño de una arquitectura para el análisis de sentimiento de información turística en redes sociales como apoyo a un sistema de recomendación

por

Giosvany Miranda Valdés

Tesis presentada en conformidad con los requisitos
para el MSc en Economía, Finanzas y Computación

Universidad de Huelva & Universidad Internacional de Andalucía

Septiembre 2023

uhu.es

un Universidad Internacional de Andalucía

An Architecture design for the sentiment analysis of tourism information in social networks as support for a recommendation system

Giosvany Miranda Valdés

Máster en Economía, Finanzas y Computación

Prof. Dr. Antonio J. Tallón Ballesteros
Universidad de Huelva y Universidad Internacional de Andalucía

2023

Abstract

With the explosion of content on social media platforms and tourist attraction review websites, big data analysis is imperative to extract relevant information to support decision-making processes. Sentiment analysis of tourist reviews and opinions enriches the preference databases used in recommender systems increasing the accuracy of their recommendations. The present work shows a distributed and balanced Big Data architecture, where text mining techniques, sentiment analysis and classification are integrated, which allow the analysis of the Andalusian tourist information that appears on the TripAdvisor website. The results in the polarity classification obtained by the proposed BERT-BiLSTM model allow reaching results higher than 80% accuracy, which validates the efficiency of the tool in the recommendation process.

JEL classification: C, O, Y, Z.

Keywords: sentiment analysis, big data, deep learning, recommendation systems, social media.

Resumen

Con la explosión de contenido en las plataformas de redes sociales y sitios web de reseñas de atracciones turísticas, el análisis de grandes volúmenes de datos es imprescindible para extraer información relevante como soporte a los procesos de toma de decisiones. El análisis de sentimiento de las reseñas y opiniones de turistas, enriquece las bases de datos de preferencias utilizadas en los sistemas de recomendación con vistas a elevar la precisión de sus recomendaciones. El presente trabajo, muestra una arquitectura Big Data distribuida y balanceada, donde se integran técnicas de minería de textos, análisis de sentimiento y clasificación, que permiten analizar la información turística de Andalucía que aparece en el sitio web TripAdvisor. Los resultados en la clasificación de la polaridad obtenidos por el modelo BERT-BiLSTM propuesto permiten alcanzar resultados superiores al 80% de precisión, lo cual valida la eficiencia de la herramienta en el proceso de recomendación.

Clasificación JEL: C, O, Y, Z

Palabras clave: Big Data, análisis de sentimiento, redes neuronales profundas, sistemas de recomendación, redes sociales

Índice de contenidos

Índice de Tablas

Índice de Figuras

1 Introducción

"Inteligencia artificial, aprendizaje profundo, aprendizaje automático… te dediques a lo que te dediques, si no lo comprendes tienes que ponerte con ello y aprender qué es. Porque de lo contrario serás un dinosaurio dentro de 3 años."

Mark Cuban.

Con la expansión de las tecnologías móviles, la internet y las redes sociales, en la actualidad existe una explosión de información relevante sobre diversas temáticas, en particular el turismo, donde los usuarios expresan sus perspectivas, opiniones o experiencias sobre cualquier aspecto relacionado a un producto o servicio en particular a través de diversas plataformas.

A partir de la crisis generada por la pandemia de COVI-19, a partir de 2021 el turismo ha comenzado su camino creciente hacia la normalidad, alcanzando niveles similares a los alcanzados antes de la epidemia. En el caso de Andalucía se cierra el 2022 con un ascenso del 53,6% con más de 30,8 millones de turistas, que, si bien se encuentra por debajo del nivel alcanzado en 2019, muestra la recuperación actual del sector (SmartData Andalucía, 2023).

Es una tendencia que cada vez más turistas deciden viajar de manera independiente, buscando su elección en la web, favorecido por el auge en el uso de las tecnologías de la información, donde los turistas pueden encontrar más información, seleccionar y reservar sus destinos sin necesidad de una agencia de viaje intermediaria. La consulta de información oficial que podemos encontrar en las webs de los proveedores de productos y servicios turísticos se une a la extraída de las redes sociales, las cuales comienzan a ganar protagonismo.

Las reseñas en línea son claros detonantes de la elección del destino, ya que, expresan la opinión de los revisores, sus emociones, experiencias reales y en muchos casos ofrecen recomendaciones a otros consumidores. Unido a este aumento en el número de opiniones que otorgan los usuarios en línea, encontramos la necesidad cada vez más creciente de utilizar técnicas de procesamiento automatizado de estas, ya que resulta imposible que podamos leer y analizar todas estas reseñas, por lo que debemos auxiliarnos en métodos que nos brinden una información más "resumida". Los

sistemas permiten ayudar a las personas a encontrar lugares que pueden ser de interés, siendo esta situación de vital importancia en los problemas de la vida real (van Setten et al. 2004).

A través de la utilización de técnicas para la captura y almacenamiento de la información obtenida desde las redes sociales caracterizada por las 4V's de Big Data (Variedad, Velocidad Volumen y Veracidad), minería y clasificación de textos y el análisis de sentimiento, apoyado con el creciente aumento en el uso de la inteligencia artificial y el aprendizaje profundo, podemos determinar criterios para la evaluación de lugares de interés.

Las arquitecturas basadas en sistemas distribuidos permiten obtener en tiempo real la información deseada basada en la computación en la nube para el tratamiento de grandes volúmenes de datos, siendo este un campo de constante evolución en el campo de la informática y las comunicaciones.

Por ello, nos planteamos la siguiente pregunta de investigación: ¿Es posible con el uso de las tecnologías de procesamiento de datos masivos, el análisis de sentimiento y otras técnicas de minería de texto obtener información de las redes sociales que permita enriquecer la efectividad de las recomendaciones para futuros turistas?

El objetivo principal de este trabajo es la integración de estas tecnologías en una arquitectura robusta y flexible que permita el análisis de información contenida en redes sociales para actualizar en tiempo real la base de datos de preferencia de un sistema de recomendación sobre información turística. Al mismo tiempo, nos planteamos los siguientes objetivos específicos:

1. Llevar a cabo una revisión bibliográfica exhaustiva sobre el impacto del concepto de Big Data y el análisis de sentimiento en el perfeccionamiento de la calidad de las recomendaciones para el turismo.

2. Diseñar una arquitectura orientada al uso de grandes volúmenes de datos que permita el manejo eficiente en tiempo real y posterior análisis de los datos extraídos de redes sociales desde un sistema de recomendación.

3. Implementar una herramienta que permita la extracción, clasificación y almacenamiento de información turística presente en redes sociales.

4. Implementar una herramienta que permita la interpretación y análisis de polaridad de la información contenida en la base de datos con el uso de las redes profundas.

En el Capítulo 2 del presente trabajo se realizará una detallada revisión de la literatura asociada a las temáticas de estudio, con énfasis en su novedad y actualidad. Como parte del mismo, se describen temáticas como la evolución de la información turística en redes sociales, las arquitecturas Big Data, el análisis de sentimiento y los conceptos asociados al aprendizaje profundo.

Por su parte, en el Capítulo 3 se define la arquitectura propuesta, resaltando las particularidades presentes en cada capa de la misma y las principales herramientas a utilizar para el desarrollo y despliegue en un entorno de aplicación.

El Capítulo 4 muestra los resultados obtenidos en cada fase del proyecto. En el mismo, se describen los recursos utilizados para el despliegue de la arquitectura, la información derivada del proceso de captura de información y se evalúan los resultados finales al aplicar los modelos de aprendizaje profundo en los datos extraídos previamente tratados.

Como parte de la memoria se expresan las Conclusiones del trabajo, así como Proyecciones de Trabajo Futuro, trazando estrategias que garanticen la continuidad de la investigación en diversos escenarios. Las Referencias Bibliográficas organizadas en formato APA cierran la organización de la memoria.

2 Revisión de la literatura

"Investigar es ver lo que todo el mundo ha visto, y pensar lo que nadie más ha pensado."

Albert Szent-Györgyi

2.1 Impacto de las redes sociales en la elección de destinos turísticos

Con el rápido desarrollo de las tecnologías de la información, la Internet ha ido penetrando poco a poco en muchos ámbitos de nuestras vidas. La industria del turismo se ha extendido gradualmente desde un enfoque más presencial a un estado de turismo conectado o en línea, apoyado en su carácter de industria de servicios, donde su carácter intangible, hace que el turista apoye su elección en las valoraciones previas que posee, las cuales le permiten inferir la calidad del servicio a percibir.

Aunque es cierto que la tecnología se utiliza en el sector turístico desde hace varias décadas, no fue hasta pasado la primera década del siglo XXI que realmente se pudo apreciar un impacto en la evolución y automatización generalizada de tareas, desde la reserva y venta en línea, los sistemas de recomendación y otras oportunidades definidas por Buhalis y Law con el concepto de turismo electrónico, describiéndola como "la digitalización de trámites en el sector turístico" (Essien y Chukwukelu, 2022).

Con la aparición de las compañías de viajes en línea y diversas comunidades que potencian el turismo desde las redes, un número cada vez mayor de turistas busca su próximo destino en la web, y analiza los comentarios sobre experiencias de viaje de otros viajeros antes de tomar una decisión (Chen et al., 2023).

Sin lugar a dudas, internet contiene un gran cúmulo de información que puede utilizarse para el análisis y la predicción de múltiples aspectos vinculados al sector turístico, lo que ha conllevado a un acelerado aumento de las investigaciones con el fin de obtener el liderazgo en el sector. El análisis de información en plataformas como Google Trends, basada en la recolección de datos

con un índice diario, semanal y mensual en tiempo real de consultas que los usuarios envían a los motores de búsqueda, constituye una de las vías más utilizadas (Laaroussi et al., 2023).

Por otro lado, los datos provenientes de redes sociales en forma de tweets (Twitter), publicaciones (Instagram, Facebook), foros, reseñas de viajes (TripAdvisor, Expedia), entre otros, han jugado un papel clave en la industria del turismo y su evolución a nuevas formas de gestión en los últimos años, convirtiéndose en importantes fuentes de datos, de las cuales es posible extraer opiniones sobre servicios o productos.

Una de las redes más utilizadas por la comunidad turística es TripAdvisor, la cual facilita la revisión de hoteles, restaurantes, y destinos alrededor del mundo. Esta web social facilita a los usuarios con revisiones de diferentes viajeros y expertos en la temática. La cantidad total de reseñas y calificaciones que podemos encontrar en un sitio como Tripadvisor a nivel mundial ha aumentado significativamente desde el año 2014, alcanzando la marca de mil millones en 2021.

Al cierre de 2022, la empresa resalta que su plataforma en línea reportó más de mil millones de reseñas, duplicando lo que existía hasta el año anterior y logrando en solo un año números similares a los que se tenían en los 7 años precedentes (Statista, 2023). Sólo de la región de Andalucía, España, existen un total de 5.813.793 revisiones y opiniones en esta plataforma.

Viera y colaboradores realizaron un estudio de la influencia de las redes sociales en la selección de destinos, concluyendo que el 66,5% de los viajeros, habían utilizado las redes sociales para informarse o decidir sobre su destino turístico. Los resultados proporcionan información a las diversas organizaciones vinculadas al sector y demuestran que Facebook e Instagram poseen el mayor impacto en la actualidad (Viera et al. 2023).

Recientemente, el uso de las tecnologías de Big Data ha experimentado un aumento en el turismo (Mariani et al., 2022), donde los investigadores han apostado por contar con sus beneficios para un mejor proceso de recomendación.

Pero el impacto de este proceso de digitalización no solo lo podemos encontrar en el mercado del turismo, su evolución y predicciones del mercado, sino que existe un interés creciente en la integración de tecnologías como la robótica (Akdim et al., 2021), blockchain (Demirel et al., 2022)

y la inteligencia artificial (Ivanov y Webster, 2021), demostrando que vamos dando los primeros pasos hacia una evolución sin límites en la industria del turismo.

En cuanto a los métodos analíticos, los de clasificación de textos turísticos han sufrido cambios en su concepción en los últimos años, pasando de utilizar el análisis de redes sociales básico a esquemas que utilizan aprendizaje automático y aprendizaje profundo, evolucionando desde simples estadísticas de frecuencia de palabras de textos generados por usuarios hasta la extracción de la semántica profunda y las asociaciones dentro de los textos logrando alcanzar una mayor precisión en obtener las percepciones y emociones de los turistas en sus comentarios. En la Figura 1, Liu y colaboradores muestran cómo han ido evolucionando estos métodos desde los inicios de esta rama hasta la actualidad.

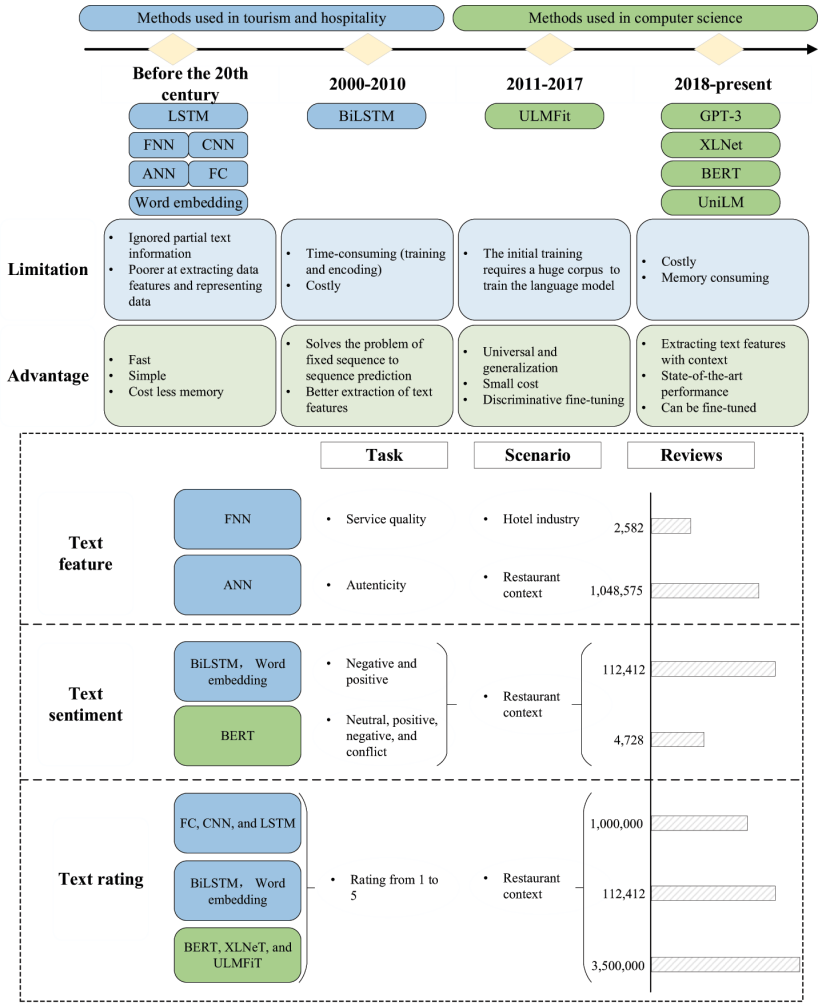

Figura 1: Comparación de métodos utilizados en el tratamiento de información turística. Fuente: Liu et al., 2023.

2.2 Big Data y su arquitectura

El aumento creciente de las redes móviles, la computación en la nube, las redes sociales y otras tecnologías emergentes, traen consigo un considerable volumen de información, denominada "Big Data". Aunque existen innumerables definiciones sobre este término, una de las más utilizadas es la descrita por el grupo consultor Gartner donde plantea "Big Data es un conjunto de activos de información de gran volumen, velocidad y variedad que exigen formas rentables e innovadoras de procesamiento de datos para mejorar la comprensión y la toma de decisiones" (Beyer et al., 2012).

Las arquitecturas basadas en Big Data están compuestas por infraestructuras y bases de datos heterogéneas, así como diversas herramientas para su análisis y visualización. Seleccionar la arquitectura que más se ajuste al problema planteado es la clave para poder explotar todo el poderío de las tecnologías asociadas al Big Data (Miranda y Delgado, 2015).

Con el desarrollo de la investigación definimos un ciclo de vida típico de Big Data generalmente como la combinación de los procesos de recopilación de datos, extracción, limpieza, pre-procesamiento, almacenamiento, análisis y visualización, contando en cada fase del proceso con herramientas y tecnologías Big Data como Apache Hadoop o Apache Spark.

En su trabajo, Macak y colaboradores hacen un estudio de diferentes arquitecturas de Big Data, en diferentes dominios de actuación, donde analizan las similitudes y diferencias entre los diferentes dominios, proponiendo un conjunto de guías prácticas sobre cómo construir y mejorar estas arquitecturas. En la Figura 2, podemos observar una propuesta de arquitectura de dominio general, que sirve de referencia a la propuesta a realizar en el presente trabajo.

Además, Macak y colaboradores plantean que desde la perspectiva de la herramienta Big Data, es generalizado el uso de Apache Hadoop para computación distribuida y como sistema de archivos, expandiéndose en los últimos tiempos la utilización de Apache Spark. Además, muchos dominios indican el uso de bases de datos NoSQL, como Apache HBase, MongoDB, o Redis, como los ejemplos más típicos (Macak et al., 2020).

Otra de las conclusiones importantes plantea que las aplicaciones de Big Data requieren amplios recursos y entornos para almacenar, procesar y analizar estas enormes colecciones de datos de forma distribuida. Singh y colaboradores, ponen su confianza en el uso de contenedores con

computación en la nube como medio para proporcionar una solución al problema, sin embargo, proponen de manera adicional un mecanismo de equilibrio de carga preciso y apropiado, con el uso de Docker Swarm (Singh et al., 2023).

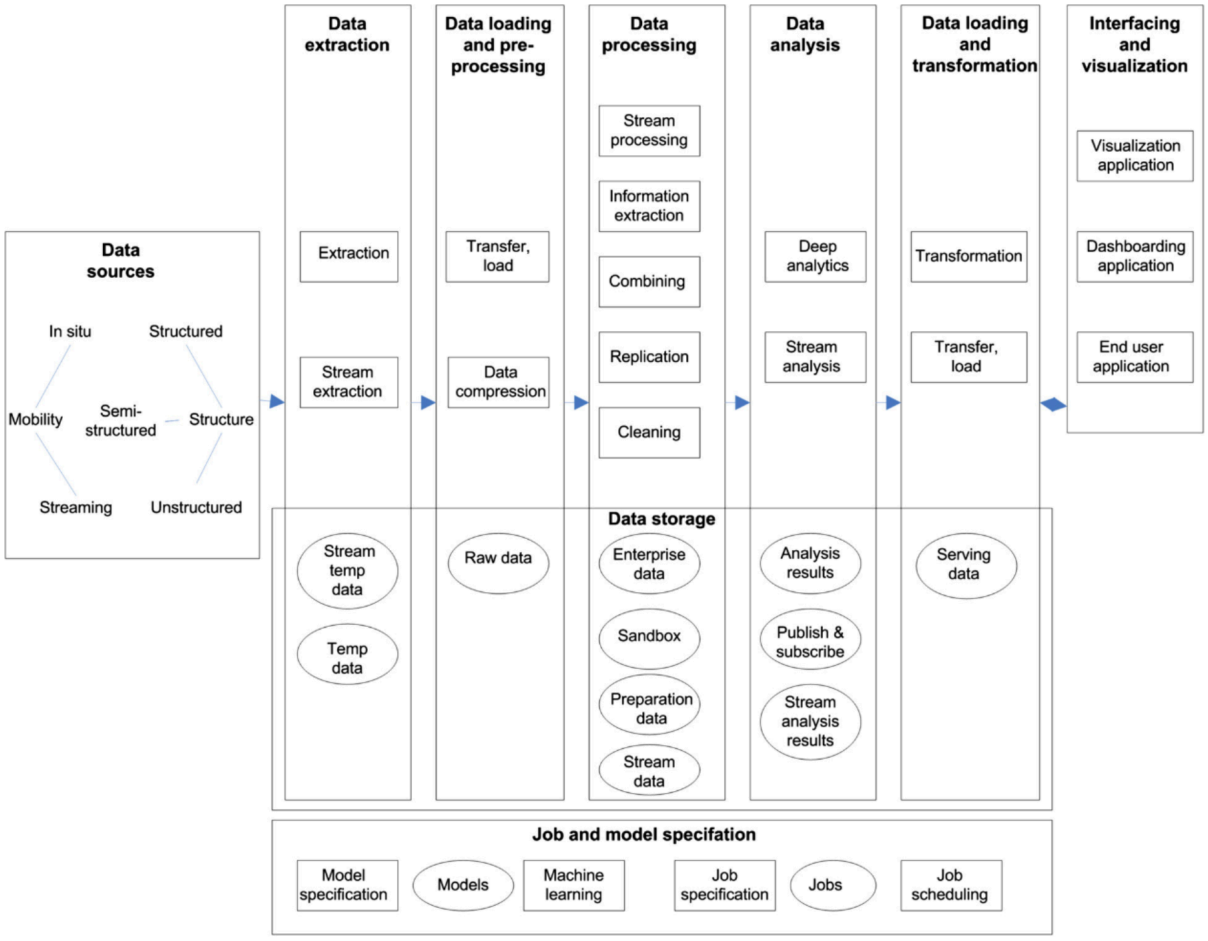

Figura 2. Ejemplo de arquitectura de Big Data de dominio general. Fuente: Macak et al., 2020

2.3 Tecnologías de minería y procesamiento de datos

2.3.1 Tecnologías de procesamiento de contenidos

Con el paso de los años internet ha mantenido un crecimiento acelerado, donde las oportunidades de consultar información relevante por parte de los usuarios muchas veces se complejizan, dado su volumen. Cualquier información mostrada en cualquier sitio web de viajes representa una fuente potencial de datos para investigadores de la hostelería o futuros clientes del servicio (Han y Anderson, 2021).

Sin embargo, recolectar esta información de manera manual, con la búsqueda de contenidos en internet se hace imposible por los altos costes en tiempo que posee. Es por esto, que la necesidad de automatizar la búsqueda de información en sitios de viajes y redes sociales adquiere una vital importancia.

Las teorías de búsqueda de información sugieren que los mecanismos que proporcionan una "señal" en cuanto a lo que puede contener una pieza de información puede ser una ayuda importante para los consumidores que intentan decidir qué información usar (Al-Natour y Turetken, 2020).

En la revisión de la literatura podemos observar diferentes formas de automatizar el acceso de la información, entre estas se encuentra el uso de las Interfaces de Programación de Aplicaciones (API por sus siglas en inglés) provistas por los diferentes sitios, las cuales, si bien facilitan el proceso de desarrollo de aplicaciones, no se encuentran disponibles para todos los sitios o poseen un costo elevado en sus versiones con mayores prestaciones. Tal es el caso del API de Twitter (https://developer.twitter.com/en) que en su versión gratuita solo permite acceder a 1.500 tweets mensuales y posee mayores prestaciones a las variantes de pago con subscripciones o costos que pueden rondar los $5.000,00 USD por mes.

Por otro lado, aparecen aplicaciones comerciales que nos permiten acceder a datos de diferentes plataformas, las cuales, aunque poseen períodos de prueba o versiones gratuitas, poseen limitaciones en cuanto a las funcionalidades o volumen de datos a consultar, siendo necesario el pago de cuotas para ello. Entre estas podemos encontrar algunas como Bright Data (https://brightdata.com) u Octoparse (https://www.octoparse.com/). Dados los precios de estas

variantes, en innumerables trabajos los autores han preferido desarrollar sus propios “web scraper”.

Una vez extraídos los contenidos, nos encontramos que los comentarios obtenidos necesitan ser pre-procesados antes de su análisis, utilizando herramientas y técnicas para el procesamiento de lenguaje natural, el cual posee como principales retos el tamaño del vocabulario, su naturaleza y ambigüedad (Nadkarni, 2011).

El procesamiento de lenguaje natural presenta diversas fases, comenzando por la tokenización, donde se realiza el análisis léxico del texto de entrada para reconocer las diferentes palabras o frases que lo componen y configurar una salida.

Como segundo paso, es necesaria la eliminación de palabras vacías, Este paso es muy importante pues con ello podemos reducir el tamaño del vocabulario a analizar y, de igual manera, se reducen los índices a generar.

Con vistas a continuar agrupando conceptos similares y optimizando el procesamiento de los índices, se propone como siguiente paso la extracción de lemas en el contenido sustituyendo aquellas palabras con forma flexionada (número, género, conjugación, etc.) por su “lema”. Este proceso, siempre produce una palabra válida en el idioma, de ahí que se proponga este paso.

A su vez, como parte complementaria de nuestro proceso de clasificación, utilizamos la evaluación media de los lugares turísticos, basado en una calificación de estrellas, siendo esta la valoración que hace el consumidor de su experiencia con un producto/servicio y que se traduce en una escala numérica ordinal (1 al 5).

2.3.2 Análisis de sentimiento

En los últimos años existe un aumento en las investigaciones relacionadas con el campo del análisis de sentimiento, motivado con el aumento de las informaciones generadas por los usuarios en las redes sociales, teniendo un amplio campo de aplicación en el marketing, inteligencia de negocios y los sistemas de recomendación, entre otros.

El análisis de sentimiento puede ser definido como “conjunto de métodos, técnicas y herramientas para extraer y detectar información subjetiva desde el lenguaje” (Mantyla et al., 2018), aunque por

lo general se utiliza para determinar la polaridad de un documento de texto, una revisión u opinión, clasificando en positiva, negativa o neutral (Alaei et al., 2019).

Este es un dominio que se encuentra todavía muy lejos de su madurez, los alcances cada vez mayores de la Internet empujan los límites de las aplicaciones de análisis de sentimientos, abriendo oportunidades para la investigación y el estudio en este campo (Bansal y Kumal, 2022).

El análisis de sentimiento puede ser metodológicamente dividido en fases como la extracción de información, pre-procesamiento de información, clasificación y la evaluación de la calidad en la clasificación (Mehraliyev et al., 2021).

Tal cual es su objetivo, el análisis de sentimiento permite examinar la relación entre las emociones y su polaridad, apoyando diferentes aplicaciones como el análisis de destinos, los sistemas de recomendación, así como soporte a las empresas de turismo para diseñar mejor los sistemas de retroalimentación para mejorar la calidad de la información recibida y así, potenciar sus atractivos basado en las reseñas y calificaciones textuales en línea de los clientes.

La aplicación de aprendizaje profundo en la industria turística es relativamente nueva, obteniendo aplicaciones diseñadas para brindar mejores servicios a los clientes a través de sistemas de recomendación. Preethi, propone un sistema de recomendación basado en análisis de sentimientos cuya base es el aprendizaje profundo (Preethi et al., 2017). También se pueden utilizar para rastrear y analizar el comportamiento del cliente (Generosi et al., 2018).

Desde 2018, los modelos de lenguaje pre-entrenados representados por un codificador bidireccional o transformadores (BERT) han abierto una nueva era en el campo del procesamiento del lenguaje natural. La aplicación de modelos de representación de lenguaje profundo con el paradigma de “ajuste fino previo al entrenamiento” se ha convertido en el enfoque dominante (Chen et al., 2023).

Las redes neuronales profundas y transformadores representan el presente y futuro ya que puede simular la estructura jerárquica del cerebro humano y hacer que tenga una semántica emocional más profunda, elevando la capacidad de representación, por lo que está siendo ampliamente utilizado en el análisis de sentimiento (Caiet al. 2020).

En este trabajo, proponemos utilizar un modelo combinado, utilizando BERT-BiLSTM para construir los modelos y predecir la orientación del sentimiento. BERT-BiLSTM es un algoritmo mejorado basado en representaciones de codificador opcionales de transformador (BERT) y memoria bidireccional a largo plazo (BiLSTM).

El proceso de pre-entrenamiento del modelo BERT consiste en realizar graduales a los parámetros del modelo, de modo que se puede describir e ir ajustando la esencia del lenguaje. Para lograr este objetivo, BERT propone dos pre-entrenamientos tareas: Modelado de lenguaje enmascarado (LM enmascarado) y Siguiente Oración.

Vale destacar que BERT-BiLSTM adopta BERT como su parte aguas arriba y BiLSTM como su parte aguas abajo (Cai et al., 2020). BERT posee la capacidad de aprender a partir de la "cercanía" entre las palabras que componen en lenguaje, mientras que BiLSTM es competente en el aprendizaje del contexto general, por lo que la utilización de ambos se acerca a la lógica del lenguaje humano. Como modelo de aprendizaje profundo, las redes LSTM tienen muchas ventajas: tienen fuertes capacidad de ajuste (Bi et al., 2023).

2.4 Sistemas de recomendación

Los sistemas de recomendación emergen como una herramienta cuya misión es personalizar la información que recibe el usuario acorde a sus necesidades, preferencias o gustos (González et al., 2013). Los mismos, Permiten obtener información sobre los usuarios con vistas a sugerir información de interés potencial para ellos, apoyando con esto el proceso de búsqueda de información, proveyendo solo aquella que pudiera ser relevante.

Utilizar un sistema de recomendaciones es un sitio web de viajes u otras aplicaciones vinculadas al sector turístico, nos permite promover las atracciones relevantes basadas en la predicción de las preferencias de los turistas. Aunque los motores de búsqueda pueden aliviar el problema de la sobrecarga de información en cierta medida, los turistas deben elegir las palabras clave apropiadas lo cual tiende a ser inexacto y engorroso, de ahí la importancia de estos sistemas (Wang, 2022).

Un gran número de sistemas de recomendación basan sus algoritmos en identificar similitudes en las preferencias de usuarios que se encuentran en un mismo grupo que otros (Esmaeili et al., 2020,

mientras otros se basan en el análisis de los comentarios realizados (Xian et al., 2017), pero no en todos los casos se aplica análisis de sentimiento como es el desarrollado por Abbasi-Moud que confecciona un sistema en el lenguaje de programación Python para el análisis de los comentarios en TripAdvisor (Abbasi-Moud et al., 2021) o el realizado en Andalucía para evaluar el impacto de la epidemia de COVID-19 en el desarrollo del turismo en la región según el análisis de sentimiento en la red social TripAdvisor (Flores Ruiz et al., 2021).

3 Propuesta de arquitectura

"Lo que importa no es aquello que miras, sino lo que ves."

Henry David Thoreau

Nuevas fuentes de datos incluyendo redes sociales, información obtenida de sitios web, dispositivos móviles, sensores y otra generada de manera automática debe ser manejada de manera eficiente con vistas a obtener por las organizaciones y usuarios en general, información consolidada, integrada y veraz que tribute a una mejor inferencia en el proceso de recomendación.

El manejo de información, almacenamiento y el análisis de los mismos otorgan valor agregado a este gran volumen de datos que encontramos en la web. Si no existiera alguno de estos procesos no es posible obtener una arquitectura adecuada.

El uso de la máquina y el aprendizaje profundo ha sido eficaz en la segmentación del mercado y la predicción de las preferencias de los clientes a través del análisis de big data social.

En el presente trabajo desarrollamos una arquitectura para analizar un gran conjunto de datos abiertos en sitios de redes sociales para la segmentación, análisis de sentimientos y predicción de los gustos de los viajeros utilizando técnicas de reducción de dimensionalidad y aprendizaje profundo.

En la Figura 3, podemos observar la arquitectura propuesta, la cual será ampliada a detalle durante el presente capítulo.

Esta arquitectura se encuentra compuesta por 6 capas, las cuales van desde la captura de la información primaria desde las diversas fuentes de datos hasta la visualización de los resultados de la recomendación, pasando por fases como el almacenamiento, el análisis y procesamiento de los datos, en una infraestructura adecuada para el manejo de grandes volúmenes de datos de manera eficiente y con un correcto balance de la carga de trabajo. La interacción constante de los usuarios en la web y la generación acelerada de contenidos nos obliga a habilitar una capa de ingestión de datos (basada en el mapeo de la web y APIs), la cual se encuentra activa en todas las fases del proceso identificando y moviendo datos desde las fuentes en la web hasta la base de datos de información.

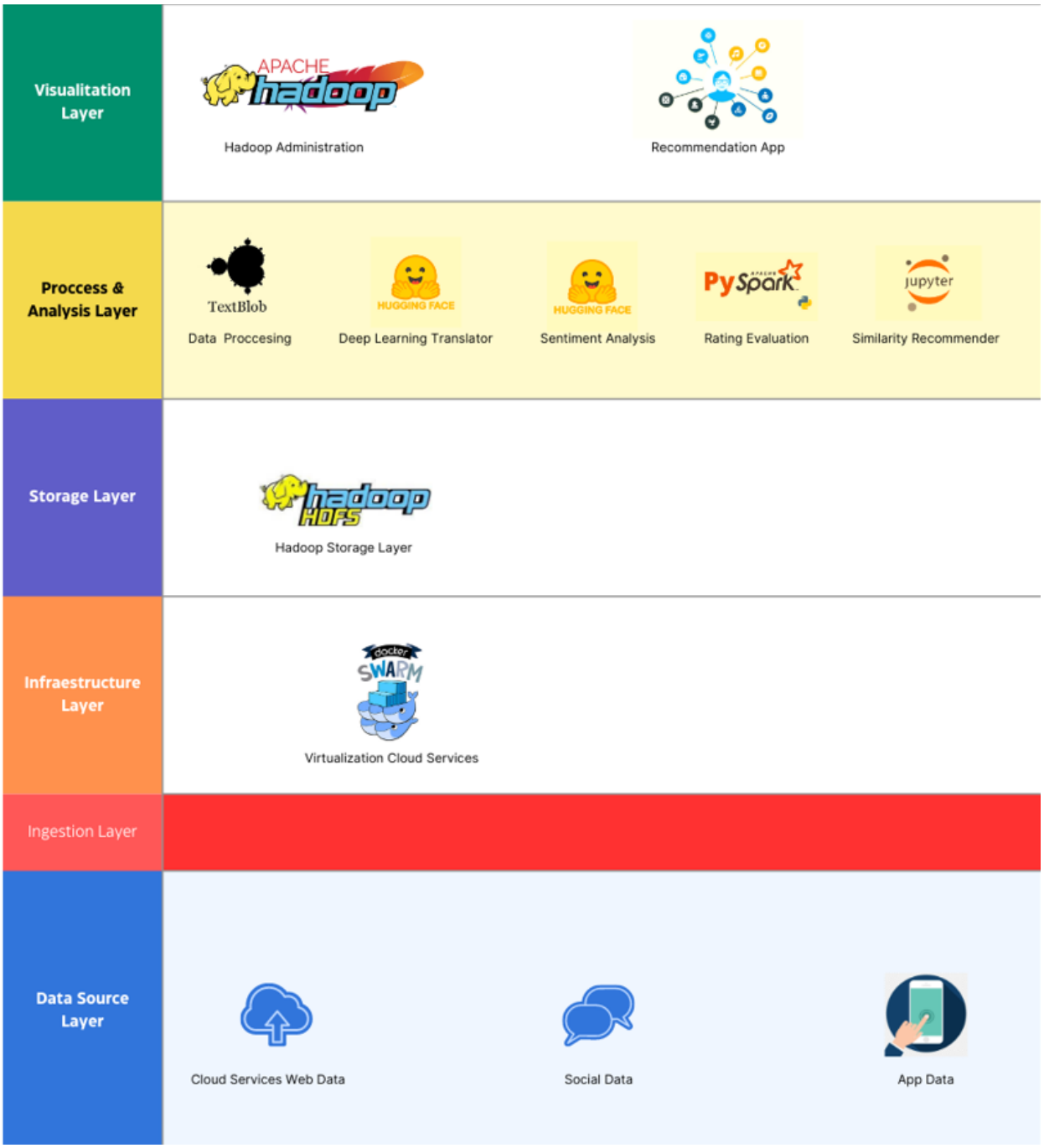

Figura 3: Arquitectura de Big Data propuesta.

En los siguientes epígrafes iremos introduciendo las diferentes capas, su concepción y los detalles para su implementación como parte del presente trabajo.

3.1 Data Source Layer

Un primer paso de gran importancia es definir las fuentes de datos desde las cuales se va a extraer la información, identificando las formas de acceso o captura de información en las mismas.

El concepto de desarrollo de esta capa se basa en una interfaz genérica, sobre la cual se pueden incorporar diferentes clases que heredan sus características y hacen que la incorporación de nuevas fuentes de datos sea un proceso sencillo que no afecte el funcionamiento de la plataforma.

Con el avance de la tecnología y los medios digitales, compartir experiencias sobre sitios de interés se hace cada día más fácil. Las redes sociales y las plataformas de reseñas se convierten en las herramientas ideales para compartir experiencias, tal es el caso de TripAdvisor y Yelp o redes sociales como Twitter o Instagram. La Tabla 1 nos muestra las principales fuentes de información disponibles para obtener reseñas desde la web.

Tabla 1: Principales sitios con información turística y sus vías de acceso.

Main Tourism social media websites / Access Way	API	Web Scraper	Manual	Obs
TripAdvisor	X	X	X	Content API
Yelp	X	X	X	Yelp Fusion
Google Review	X	X	X	Google My Business
Expedia	X	X	X	Rapid Guest Reviews
Twitter (X)	X	X	X	Twitter API v 2
Facebook/Instagram	X	X	X	API Graph

Tal y como podemos observar en la Tabla 1, las principales redes sociales y sitios de reseñas han habilitado su propia API para que los desarrolladores puedan interactuar con la información que se comparte en estos sitios. Aunque existen muchas que brindan acceso para el análisis de la totalidad de las reseñas (Expedia - https://developers.expediagroup.com), encontramos otras que como principal inconveniente presentan limitaciones en el acceso a la información.

En esta variante podemos encontrar que existen casos en donde el uso de la API viene vinculada a funcionalidades restringidas donde solo se puede acceder solamente a la información de una cuenta

donde se posee el id y las credenciales (TripAdvisor - https://tripadvisor-content-api.readme.io/reference/overview) u otros casos donde el volumen de los datos es limitado en opción gratuita (Twitter - https://developer.twitter.com/en/docs/twitter-api).

Por este motivo, surge muchas veces la necesidad de desarrollar o utilizar herramientas que permitan recopilar información de forma automática de la Web. A merced de los cuestionamientos que pudiera tener este método en cuanto a acceso e información a datos personales establecidos por el Reglamento General de Protección de Datos (RGPD) y otras regulaciones afines, en nuestro trabajo solo nos centraremos en la obtención de información de las revisiones y los sitios, sin recabar información personal de los usuarios que pudiera estar limitada por esta reglamentación.

Para el futuro uso de nuestra biblioteca por parte de otros desarrolladores se necesita cumplir con ciertos requerimientos que estarán establecidos en su acuerdo de Licencia Pública General de GNU/GPL.

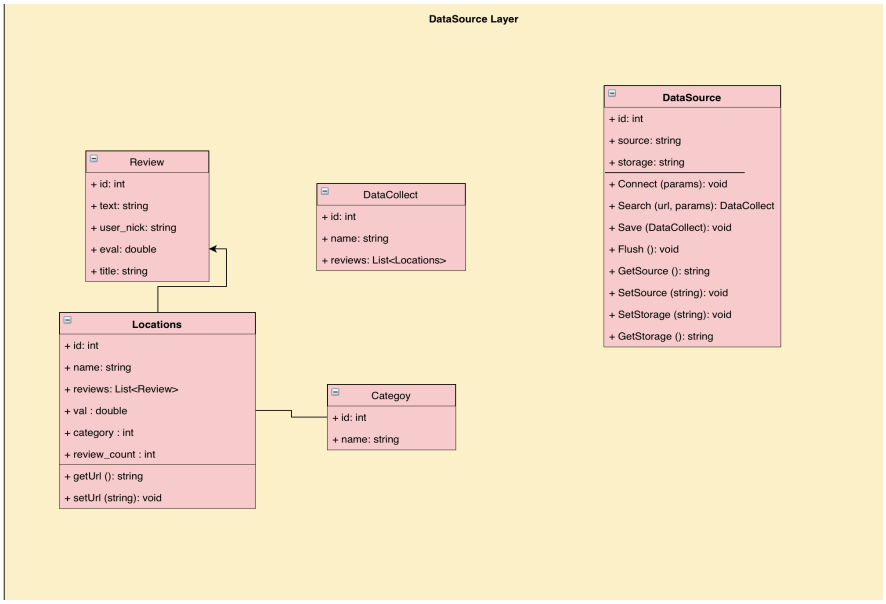

Figura 4: Diseño de clases para DataSource Layer.

3.2 Ingestion Layer

La capa de ingestión se encuentra presente en todas las fases de nuestra arquitectura, pues la misma se retroalimenta de las solicitudes que le llegan desde la capa de visualización y va enriqueciendo nuestra base de datos distribuida desde las diferentes fuentes de datos.

En su diseño implementa los diferentes métodos de búsqueda y permite interactuar con las diversas fuentes de datos a las cuales podemos acceder. Como concepto, esta ingestión de datos se puede realizar tanto mediante tareas programadas como de manera puntual dado la necesidad de información por el sistema de recomendación en un momento determinado.

Esta ingestión está presente en la clase *DataSource,* donde para cada una de las diferentes fuentes de datos se pueden realizar acciones como la búsqueda (Search) y la salva de datos en la base de datos distribuida (Save).

El método *Search* permite utilizar diferentes técnicas que permitan obtener los datos de manera ágil y flexible. Este, es sin duda, un paso de vital importancia a la hora de diseñar una arquitectura Big Data, para lo cual, hay que tener muy claro, el tipo y fuente de datos, así como la estructura de nuestros datos.

Dado que Python (https://www.python.org/downloads/) es un lenguaje de programación muy sencillo para aprender e intuitivo, es muy utilizado por muchos desarrolladores para esta tarea, con bibliotecas que permiten tanto conectarse e interactuar con diferentes APIs, como de profundizar en la web.

Entre las principales bibliotecas utilizadas en el proceso de interacción con la web, se encuentran Beautiful Soup, que es una biblioteca que facilita extraer información de páginas web. Se asienta sobre un analizador HTML o XML, proporcionando modismos Pythonic para iterar, buscar y modificar el árbol de análisis (https://pypi.org/project/beautifulsoup4/). Selenium por su parte (https://www.selenium.dev/) nos va a permitir contar con un entorno de pruebas en la web donde podemos grabar/reproducir diferentes comportamientos de los sitios web.

En nuestro caso se ha desarrollado un *TripAdvisorScraper*, el cual va a permitir la interacción con esta web de reseñas, donde dado una url que identifique el término de búsqueda, profundice en los

resultados de la misma, obteniendo información básica de los lugares de interés y las reseñas asociadas a los mismos.

Esta tarea puede realizarse en tres momentos de nuestro proceso de desarrollo:

1. **Población inicial de información:** De acuerdo a los términos de búsqueda se realiza una primera corrida de datos en nuestra infraestructura distribuida balanceada, la cual realiza una primera carga de datos exhaustiva en nuestra base de datos.

   En nuestro caso se utilizan como palabras claves, aquellas que fueron extraídas del estudio de los principales recursos turísticos de Andalucía, utilizando términos como *Andalucía, Cádiz, Guadalquivir, Doñana, Huelva, Aracena, Odiel, Río Tinto, Córdoba, Cazorla, Granada, Sevilla, Almería, Málaga*, entre otros.

2. **Actualización sincronizada de información:** Mediante una tarea distribuida se almacena la fecha y términos de la última búsqueda y se define la frecuencia, actualizando los elementos desde la última fecha de búsqueda hasta la actualidad.

   A los efectos de prueba, en nuestro caso se diseñó una tarea programada que se ejecutaba cada 7 días a las 0.00 h utilizando los términos de la búsqueda generalizada anterior y otros que se pudieran ir agregando que aumentaran el entorno de la búsqueda.

3. **Actualización puntual:** Se concibe como parte del uso de la aplicación de recomendación que se solicite información de un sitio de interés que no se encuentra contenida en la base de datos, para esto se realiza una solicitud de búsqueda en el momento, donde de encontrarse información se devuelve al usuario un extracto de la misma, para evitar demoras e ineficiencia en el procesamiento de datos.

   Este término se agrega a los términos de búsqueda para la próxima actualización sincronizada.

3.3 Infraestructure Layer

Uno de los principales retos que conlleva el manejo de grandes volúmenes de datos es encontrar una arquitectura que permita la administración e integración eficiente de los diferentes componentes que forman parte de la misma.

Una arquitectura para utilizar en plataformas de Big Data requiere procesar y atender lotes de cientos de millones de transacciones simultáneamente, con lo cual una aplicación alojada en un entorno con recursos limitados puede colapsar y quedar inactiva, de ahí que el balanceo de las cargas y manejo eficiente más que una característica deseada se convierte en una obligación.

Utilizar un contenedor nos otorga más flexibilidad a la que nos puede otorgar una máquina virtual, debido a que no se encuentran acoplados al sistema operativo ni restringidos por la infraestructura del host anfitrión, por lo que es posible la portabilidad entre diversas distribuciones de sistemas operativos y/o en la nube. A su vez nos facilita en las primeras fases del proyecto compartir características entre el entorno de desarrollo/pruebas con el entorno de despliegue de la aplicación.

Docker es un sistema basado en software libre que facilita la automatización del despliegue de las aplicaciones en contenedores. En un contenedor, donde las aplicaciones son virtualizadas y ejecutadas, la forma en la que está diseñado permite un entorno liviano con una ejecución más eficiente de los códigos, pudiendo testear el mismo antes de su ejecución en entornos reales.

Por su parte, Docker Swarm permite ejecutar estos contenedores en una granja de nodos, permitiendo gestionar varios clústeres de manera descentralizada en un entorno maestro-esclavo, teniendo como elemento distintivo el balanceo de cargas desde los nodos maestros. Entre sus características más relevantes podemos encontrar la redistribución de las cargas de trabajo si algún nodo falla asegurando una alta disponibilidad, la administración de los grupos de contenedores, ofreciendo la posibilidad de agregar, eliminar, balancear la carga entre ellos, etc.

Para su monitoreo se utiliza Swarmpit (https://swarmpit.io) el cual mediante una interfaz web permite administrar y monitorizar un clúster de Docker Swarm.

3.4 Storage Layer

En la capa de infraestructura desplegada se instala Hadoop HDFS como capa de almacenamiento de la información. Como base a los trabajos a realizar en nuestro entorno de pruebas se propone desplegar contenedores para la representación de dos nodos esclavos y un nodo maestro. En la Figura 5, podemos observar los nodos de Hadoop desplegados en nuestra arquitectura.

Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
hadoop-slave2:50010 (172.28.0.3:50010)	1	In Service	393.37 GB	24 KB	55.84 GB	332.74 GB	0	24 KB (0%)	0	3.3.7
hadoop-slave1:50010 (172.28.0.2:50010)	1	In Service	393.37 GB	24 KB	55.84 GB	332.74 GB	0	24 KB (0%)	0	3.3.7
hadoop-master:50010 (172.28.0.4:50010)	1	In Service	393.37 GB	24 KB	55.84 GB	332.74 GB	0	24 KB (0%)	0	3.3.7

Figura 5: Despliegue de nodos Hadoop en infraestructura diseñada con Docker Swarm.

El Sistema de Ficheros Distribuidos de Hadoop (HDFS, por sus siglas en inglés) es el componente principal del ecosistema Hadoop. Esta pieza hace posible almacenar grandes volúmenes de datos con tipos de datos estructurados, semi-estructurados y no estructurados como imágenes, vídeo, datos de sensores, etc.

HDFS se encuentra optimizado para almacenar grandes cantidades de datos y mantener varias copias para garantizar una alta disponibilidad y la tolerancia a fallos.

Nuestra capa de almacenamiento va a almacenar la información derivada de los datos que se obtienen desde las diferentes fuentes, poblados desde la capa de ingestión. Aunque en la actualidad a los efectos de prueba para este trabajo se tratan básicamente datos estructurados la infraestructura se encuentra preparada para el trabajo con tipos de datos no estructurados como la posición, videos, imágenes y otras que pueden ser recursos complementarios para nuestro sistema de recomendación.

3.5 Proccess and Analysis Layer

Esta capa de procesamiento y análisis presenta en su interior varias etapas que iremos describiendo en el presente trabajo, pasando por el pre-procesamiento de las revisiones, así como varias capas del modelo BERT-BiLSTM que incluyen la traducción de los textos y el análisis de la polaridad.

3.5.1 Pre-procesamiento de textos

Para el procesamiento de los textos comenzaremos con técnicas de procesamiento de lenguaje natural básicas. Luego, en un primer momento realizamos la extracción de las palabras mediante un análisis léxico del contenido ("tokenización"), tratando de en este primer paso de alcanzar mayor cobertura a costo de la precisión, por este motivo se mantienen los números, no se discrimina entre mayúsculas y minúsculas y se eliminan guiones, tildes y signos de puntuación, construyendo así, un conjunto de términos que formen el índice con una mayor amplitud. Los emoticones y emojis no se eliminan y son convertidos a su significado textual usando la biblioteca emoji de Python y su método emoji.demojize.

Para realizar estos procesos y todo el tratamiento del texto utilizamos la biblioteca TextBlob. TextBlob es una biblioteca de Python para el procesamiento del lenguaje natural (NLP). TextBlob utiliza Natural Language ToolKit (NLTK) para realizar sus tareas. NLTK es una biblioteca que brinda fácil acceso a una gran cantidad de recursos léxicos y permite a los usuarios trabajar con categorización, clasificación y muchas otras tareas. TextBlob es una biblioteca simple que admite análisis y operaciones complejos sobre datos textuales.

Posteriormente, se eliminan las palabras vacías que no aportan significado al texto, para esto, elaboramos una lista con todas las palabras presentes en la biblioteca nltk, la cual presenta 17 idiomas, entre los que se incluyen los que se encuentran en las revisiones.

```
from nltk.corpus import stopwords
```

Este paso es muy importante pues con ello podemos reducir el tamaño del vocabulario a analizar y, de igual manera, se reducen los índices a generar.

Con vistas a continuar agrupando conceptos similares y optimizando el procesamiento de los índices, se propone como siguiente paso la lematización del contenido sustituyendo aquellas palabras con forma flexionada (número, género, conjugación, etc.) por su "lema". A diferencia del steamming siempre produce una palabra válida en el idioma, de ahí que se proponga este paso.

La opción *lemmatize* de TexBlob devuelve el lema de una palabra usando la función morphy de WordNet, la cual procesa la morfología de la palabra de acuerdo al vocabulario de esta base de datos léxica. Para otros idiomas, se utiliza el stemming de SnowBall, versión más avanzada que Porter (Snowball, 2023)

```
>>> from textblob import Word

>>> w = Word("octopi")

>>> w.lemmatize()

'octopus'

>>> w = Word("went")

>>> w.lemmatize("v")  # Pass in WordNet part of speech (verb)

'go'
```

Para culminar esta fase inicial se vectorizan las palabras utilizando modelos Word2Vec, los cuales se encuentran implementados en Huggingface en los diferentes idiomas (https://huggingface.co/models?sort=trending&search=word2vec) .

Esta se trata de una comunidad de desarrollo que permite el acceso a innumerables herramientas para el procesamiento del lenguaje natural, donde podemos construir, entrenar y desplegar los modelos de machine learning. Es un espacio en el que usar modelos pre-entrenados con muchísimos parámetros para aplicarlos a las tareas que necesitamos, evitando en muchos casos el entrenamiento de modelos y su consiguiente costo en tiempo.

3.5.2 Modelos Transformers

Para llevar a cabo las tareas de traducción de los textos y análisis de sentimiento proponemos la utilización de una arquitectura transformer, la cual es una arquitectura codificador-decodificador que posee entre sus principales características la auto atención múltiple, codificación de la posición y las conexiones residuales (Vaswani et al., 2017).

Como primer paso del proceso se deben traducir los diferentes textos de acuerdos a los idiomas de los comentarios, utilizando para esto los modelos pre-entrenados derivados del proyecto OPUS-MT disponibles en Huggingface (Tiedemann y Thottingal, 2020).

Una vez unificados en esta primera fase del proyecto, se identifica la polaridad de los textos con un modelo combinado BERT-BiSLTM.

Aunque se ha hecho muy popular, las siglas del modelo BERT significan representación de codificador bidireccional de transformadores (Bidirectional Encoder Representations from Transformers).

Se trata de una técnica basada en redes neuronales para el pre-entrenamiento del procesamiento del lenguaje natural (NLP). Los modelos BERT pueden interpretar el contexto completo de una palabra analizando las palabras que vienen antes y después, lo que resulta muy útil para comprender la intención de búsqueda que tiene el usuario al realizar la consulta. En la propuesta se procesan las palabras de la búsqueda de manera lineal, por lo que proponemos complementar con un modelo BiLSTM, el cual permite analizar el contexto con la utilización de la “memoria” a más largo plazo (Wu et al., 2022).

Los LSTM bidireccionales son una extensión de los LSTM tradicionales, que entrenan dos LSTM en lugar de uno en la secuencia de entrada, con el segundo LSTM entrenado en la copia invertida de la secuencia de entrada, que proporciona contexto adicional a la red y da como resultado un aprendizaje aún más completo sobre el problema.

El modelo LSTM bidireccional utilizado es simple, que consta de solo 7 capas, con una capa de entrada de incrustación y una capa de incrustación, seguida de una capa convolucional, una capa de agrupación máxima, una capa LSTM bidireccional, una capa de abandono y finalmente una capa densa para obtener 3 clases para clasificación (Kang et al. 2023).

BERT se utiliza como capa codificadora para nuestro modelo y utilizar la autoatención del transformador para una codificación más rica que los modelos de integración anteriores, como Word2Vec. Luego, se decide utilizar BiLSTM después de recuperar las codificaciones BERT para obtener una representación más contextual de las entradas.

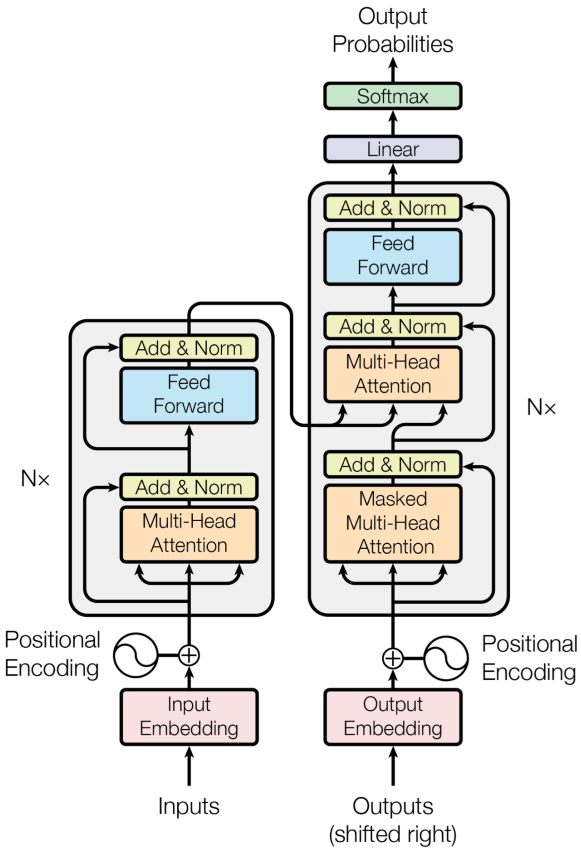

Figura 6: Arquitectura transformer. Fuente: Vaswani et al., 2017.

La capa LSTM utiliza de manera eficiente características pasadas y futuras una vez que se han codificado las entradas. Mientras que BERT brinda codificaciones contextuales con atención entre la secuencia de entrada, BiLSTM brinda contexto sin atención. Elegimos utilizar BiLSTM para conservar la coherencia de los modelos bidireccionales. La capa calcula dos representaciones ocultas diferentes para la secuencia de entrada: una para el contexto de la izquierda y otra para el contexto de la derecha. Este mecanismo es similar a la autoatención en BERT, pero no utiliza la atención en sí misma (Pearce et al. 2021).

Al realizar el ajuste fino, la mayoría de los hiper parámetros del modelo permanecen en el valor predeterminado, con la excepción del tamaño del lote, la tasa de aprendizaje y el número de épocas de entrenamiento.

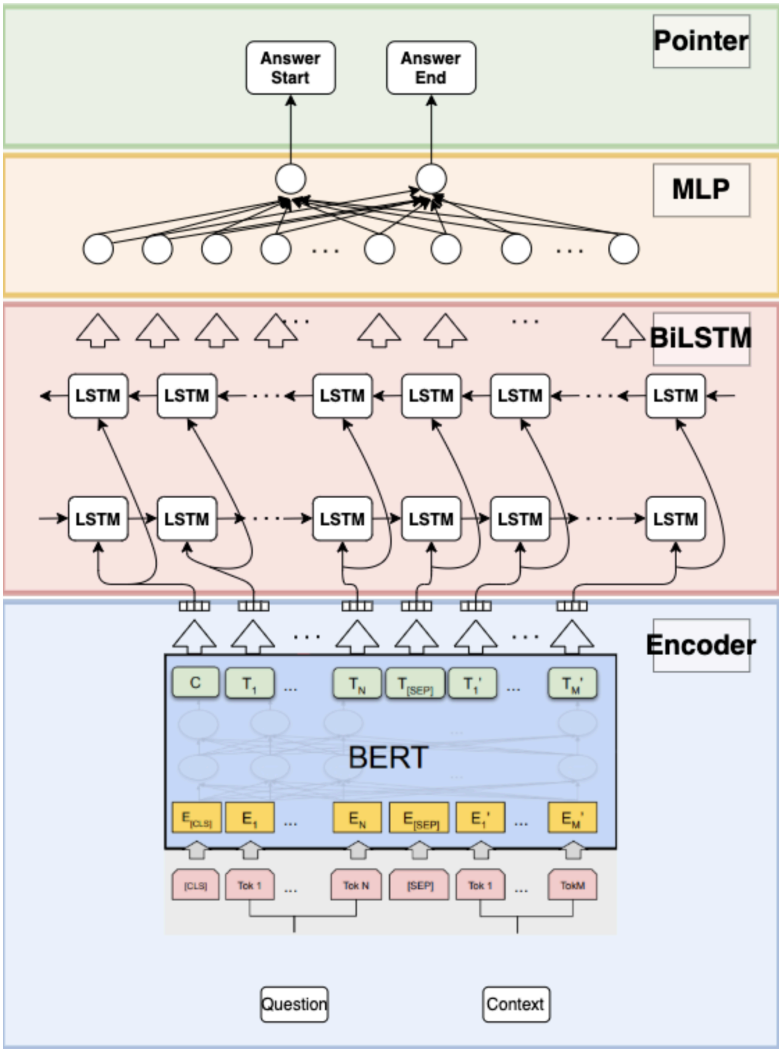

Figura 7: Modelo BERT-BiLSTM. Fuente: Pearce et al. 2021.

El resultado final es evaluado y se implementa un resultado que combina los resultados del análisis de sentimiento del texto, con la valoración numérica del sitio o actividad turística, obteniendo un valor ponderado de acuerdo a la fórmula (1).

$$R = W_1 E + W_2 \sum_{i=1}^{n} {}^{P_i}/_{n} \qquad (1)$$

donde:

R: Valor de la recomendación

W_1: Peso de la evaluación numérica del sitio

E: Evaluación del sitio

W_2: Peso de la polaridad del análisis de sentimiento

P_i: Polaridad de la revisión i

N: Cantidad de revisiones.

Para esto se utiliza en este entorno big data Spark (https://spark.apache.org), el cual nos provee de un marco rápido y potente que proporciona una API para realizar un procesamiento distribuido masivo sobre conjuntos de datos resistentes. A su vez, utilizamos Jupyter Notebook (https://jupyter.org) como aplicación para editar, ejecutar y compartir código Python en una vista web.

3.6 Visualization Layer

La capa de visualización es la que posee menor nivel de abstracción y nos permite interactuar con la administración de los datos y de cara al cliente como aplicación final.

Existen numerosas salidas para esta capa que pueden estar representadas en una herramienta de recomendación en una página de proveedor de servicios turísticos, agencia de viajes o simplemente desde una aplicación móvil, donde variables como la ubicación y el contexto geográfico puede ser incluido en nuestra arquitectura.

En todos los casos estas recomendaciones se basan en filtrado colaborativo asumiendo que existen relaciones entre los productos o servicios y los intereses de los usuarios. Muchos sistemas de recomendaciones utilizan filtrado colaborativo para encontrar estas relaciones y para dar una

recomendación precisa de un producto o servicio que el usuario puede estar interesado. En el caso de nuestra aplicación utiliza el filtrado colaborativo basado en los artículos/servicios.

La interacción de esta capa con las restantes se puede observar en la Figura 8.

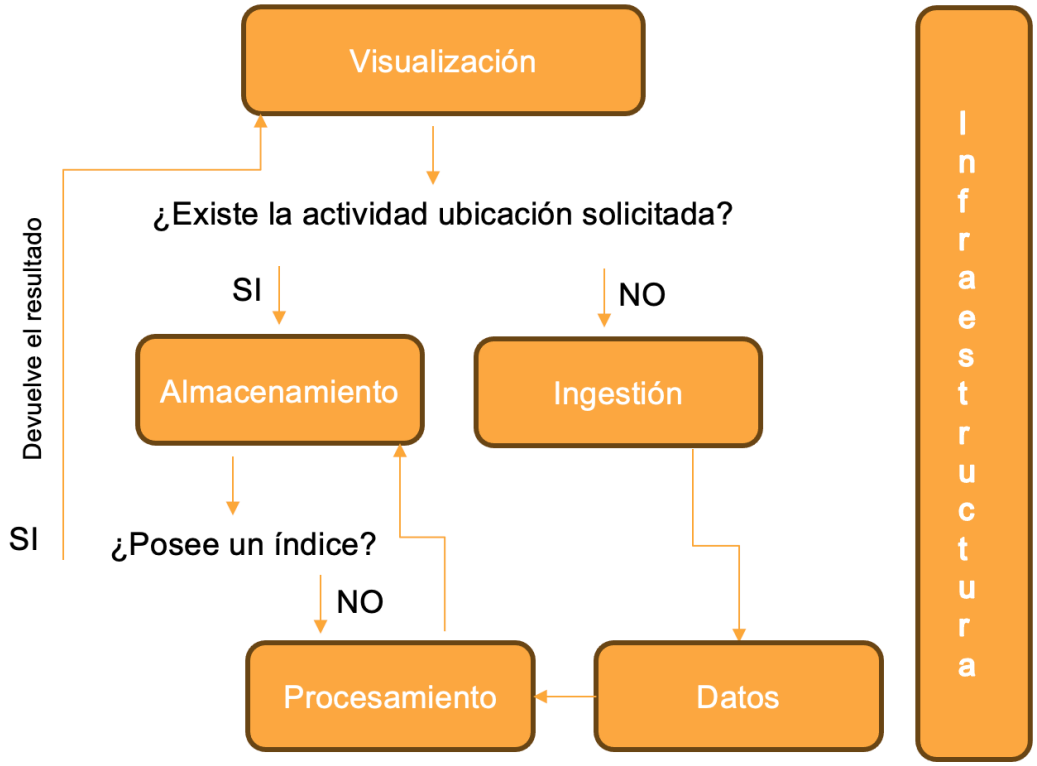

Figura 8: Relaciones entre las capas de la arquitectura.

Aunque se proponen diferentes aplicaciones destino que pueden formar parte de esta capa de visualización, el desarrollo de las mismas no se encuentra en el alcance del presente trabajo.

4 Resultados

¨Los números hablan por sí mismos en términos de todo lo que hemos hecho. ¨

Paul Rand

4.1 Preparación de arquitectura

Es importante resaltar que los resultados para la arquitectura se obtuvieron con máquinas virtuales instaladas en dos computadoras de escritorio con las siguientes características:

1) Nombre del dispositivo: DESKTOP-UL7OR9F

   Procesador: Intel(R) Core(TM) i7-8850H CPU @ 2.60GHz 2.59 GHz

   RAM instalada: 16,0 GB (15,8 GB usable)

   Tipo de sistema Sistema operativo Windows de 64 bits, procesador basado en x64.

La configuración de la máquina virtual utiliza el CPU al 100% de límite de ejecución y 8 Gb de memoria RAM.

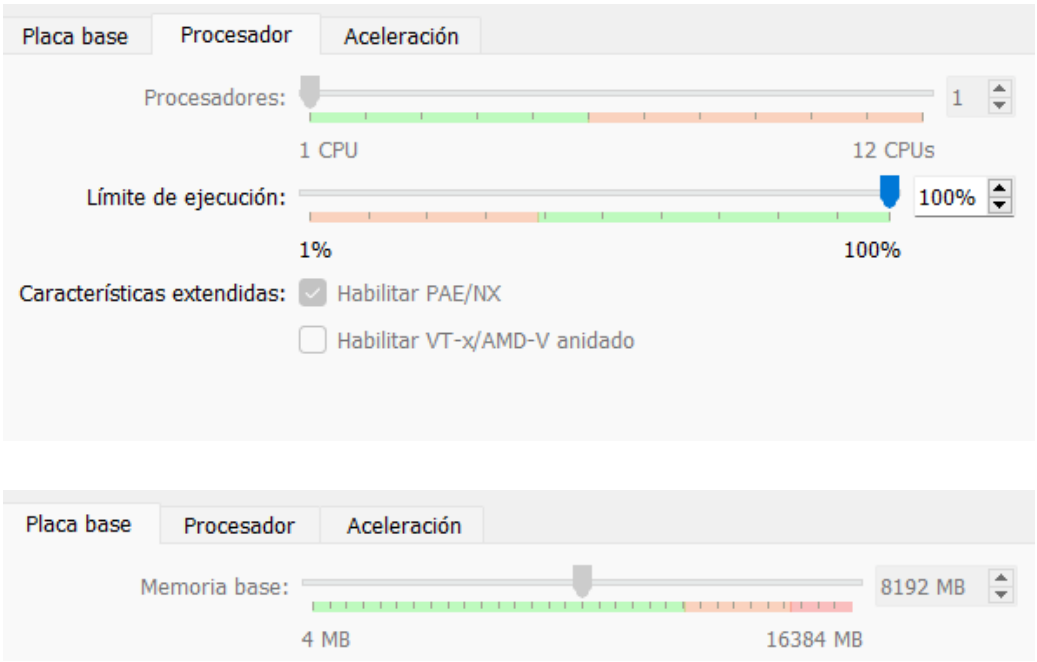

Figura 9: Recursos utilizados.

2) Nombre del dispositivo: MacBook Pro C02C983BMD6T

   Procesador: 2,3 GHz Intel Core i9 de 8 núcleos

   RAM instalada: 32 GB 2667 MHz DDR4

   Tipo de sistema MacOS 13.5.1 (22G90)

En este caso la configuración de la máquina virtual utiliza el CPU al 100% de límite de ejecución y 16 Gb de memoria RAM.

Para la ejecución de los modelos transformer se utiliza la versión gratuita de la herramienta en la nube Google Colab (https://colab.research.google.com/?hl=es) para ejecutar los modelos y funcionalidades de Huggingface, obteniendo acceso a tiempos de ejecución de GPU y TPU por hasta 12 horas, lo que ha obligado a seccionar las pruebas. Su tiempo de ejecución de GPU viene con CPU Intel Xeon a 2.20 GHz, 13 GB de RAM, acelerador Tesla K80 y 12 GB de VRAM GDDR5.

Por su parte, el tiempo de ejecución de TPU consta de una CPU Intel Xeon a 2.30 GHz, 13 GB de RAM y una TPU en la nube con 180 teraflops de potencia computacional.

4.2 Extracción y tratamiento de contenidos desde redes sociales

Para la población inicial de nuestra base de datos se extrae a manera de ejemplo información turística de TripAdvisor, como uno de los sitios turísticos más utilizados para las reseñas de visitas de internet. Se tienen en cuenta un grupo de palabras claves para realizar la búsqueda, extraídas de la Guía Turística de Andalucía, entre las que podemos encontrar Andalucía, Cádiz, Guadalquivir, Doñana, Huelva, Aracena, Odiel, Río Tinto, Córdoba, Cazorla, Granada, Sevilla, Almería, Málaga, entre otros. De la guía ¨Principales cifras mercados turísticos extranjeros en Andalucía en 2022¨, extraemos los principales mercados emisores de turismo en la provincia donde destacan además del mercado interno español otros mercados como Reino Unido, Francia, Países Nórdicos, Alemania, Portugal e Italia, por lo que tenemos en cuenta los comentarios en español, inglés, francés, alemán, portugués e italiano.

Los principales datos obtenidos acerca de las atracciones turísticas son el nombre, categoría, dirección, coordenadas, y evaluación. Tomando en cuenta las coordenadas de cada punto geográfico se geo codifican con GeoPy y se verifican que las mismas se encuentran realmente en la provincia de Andalucía. Con esto descartamos un 12% de los resultados obtenidos quedando para el análisis 2.423.300 opiniones de 25.654 sitios o atracciones ubicados en la provincia de Andalucía, distribuidos en idiomas como se expresa en la Tabla 2. Se etiquetan estos comentarios de acuerdo al idioma con vistas a tener caracterizados los mismos en el proceso de traducción.

Existen opiniones de 1.239.732 usuarios, con frecuencias entre 1 y 32 opiniones por cada uno. Donde los textos poseen una longitud promedio de 143 palabras, encontrando opiniones de una sola palabra y hasta 1.721. Posee emoticonos, acrónimos, URLs y otros caracteres que pueden introducir “ruido” en el análisis del texto y que se necesitan tener en cuenta en la etapa de pre-procesamiento.

Antes de la vectorización, tomamos una muestra de revisiones de 100 sitios, los cuales presentan un total de 3.589 revisiones en idioma español los cuales fueron etiquetados de acuerdo a la interpretación del autor sobre la polaridad de las mismas en tres categorías (Positivo, Negativo y Neutro). Esta muestra será utilizada para el entrenamiento del modelo propuesto y su comparación con otros modelos presentes en Huggingface para el análisis de sentimiento.

Para la conversión de nuestro vocabulario a vectores utilizamos Word2Vec, donde podemos realizar el proceso mediante el algoritmo CBOW o mediante Skip-gram. Para ambos modelos representamos las palabras mediante one-hot encoding. Aunque CBOW es un entrenamiento varias veces más rápido, y sus resultados son superiores para aquellas palabras más frecuentes Skip-gram posee la capacidad de predecir el contexto y no las palabras exactas, por lo que se considera que para el caso de estudio sus resultados son superiores.

Por este motivo entre los modelos presentes en Huggingface se selecciona el relacionado al proyecto Wikipedia2Vec (Yamada et al., 2020), el cual se encuentra entrenado para 12 idiomas, entre los que se encuentran los del objeto de estudio de este trabajo.

Tabla 2: Distribución de las revisiones por idioma

	Sitios	%**	Revisiones	%
Español	25.140	98,0	2.136.900	88,0
Inglés	8.209	32,0	213.434	8,8
Francés	2.052	8,0	32.832	1,4
Alemán	1.539	6,0	20.007	0,9
Portugués	1.283	5,0	11.547	0,5
Italiano	780	3,0	8.580	0,4
Total*	-	-	2.423.300	100%

**El número de sitios no se corresponde con los visitados pues existen revisiones en diferentes idiomas para un mismo sitio.*

***El porcentaje se encuentra en función de los 25.654 sitios extraídos.*

Una vez concluidas las diferentes etapas del proceso de procesamiento del lenguaje (tokenización, eliminación de palabras vacías, extracción de lemas) y conversión de nuestro vocabulario obtenemos un total de 1.306 vectores.

4.3 Análisis de sentimiento con BERT-BiLSTM

Para realizar nuestra tarea de reconocimiento utilizamos el diseño que aparece en la Figura 10, donde en los subsiguientes epígrafes mostraremos los resultados en cada una de las etapas (modelos de traducción, modelo BERT-BiLSTM). Como parte del análisis del capítulo se realizará una comparativa entre el modelo propuesto y otros presentes en Huggingface para el análisis de sentimiento.

Para la evaluación de los resultados finales utilizaremos las métricas de Precisión (precision), Exhaustividad (recall), Valor de F (F-score) y Exactitud (accuracy).

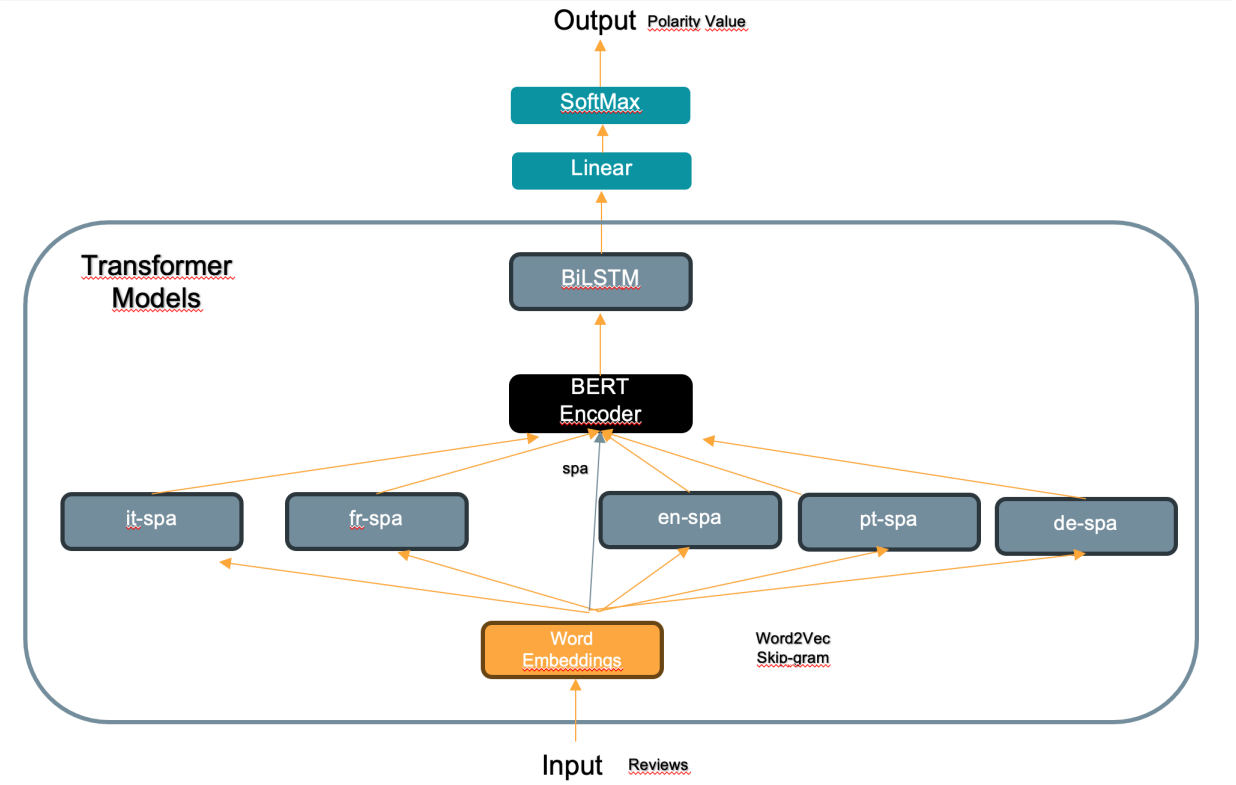

Figura 10: Diseño de solución de análisis de sentimiento basado en modelos transformers.

4.3.1 Traducción de contenidos

Para la selección del modelo de traducción entre los existentes en Huggingface vamos a utilizar la métrica BLEU (Reiter, 2018) de acuerdo a los diferentes idiomas. Como para las traducciones no es necesario un nivel de exactitud elevado, podemos utilizar esta métrica que mide la precisión de los n-gramas con respecto a un conjunto de traducciones de referencia.

El modelo seleccionado es el que se presenta en el proyecto Opus-MT (Tiedemann y Thottingal, 2020), el cual presenta para los diferentes idiomas utilizados los resultados promedio expresados a continuación en la Tabla 3.

Tabla 3: Evaluación de la traducción automática a español Opus-MT. Fuente: Tiedemann y Thottingal, 2020.

Idioma	BLEU
en-spa	40,3
fr-spa	37,3
de-spa	34,2
it-spa	39,8
pt-spa	44,8

4.3.2 Análisis de sentimiento

Para el desarrollo de nuestro modelo, comenzamos por definir los datos para el entrenamiento y prueba del modelo. Antes de realizar la vectorización de la información, tomamos una muestra de revisiones de 100 sitios, los cuales presentan un total de 3.589 revisiones en idioma español los cuales fueron etiquetados de acuerdo a la interpretación del autor sobre la polaridad de las mismas en tres categorías (Positivo, Negativo y Neutro).

Para la selección de los conjuntos de entrenamiento-prueba en un esquema 80% - 20% se utiliza la validación cruzada K-fold, donde seleccionamos K=5. La idea de la validación cruzada K-fold es dividir el conjunto de datos en K grupos y tratar cada grupo como un conjunto de datos de entrenamiento y validación para evaluar el modelo (Zhang y Liu, 2023).

El modelo se entrena con el vector de palabras características del conjunto de entrenamiento. Tomando referencias de trabajos anteriores se adopta como optimizador un optimizador Adam con función de pérdida de entropía cruzada binaria y se añade la activación 'sigmoidea', utilizada para clasificar el sentimiento de los datos de reseñas en tres categorías (Wu et al., 2023). El umbral del valor del sentimiento se establece el rango entre 0,4 y 0,6 para definir polaridad neutra, por lo cual las puntuaciones superiores a 0,6 se perciben como positivas y las puntuaciones inferiores a 0,4 se consideran negativas.

Para la selección de los parámetros óptimos se utiliza Optuna, biblioteca que permite a los usuarios construir el espacio de búsqueda de parámetros dinámicamente (Akiba et al., 2019).

Tabla 4: Parámetros de configuración del modelo.

Parámetros	Descripción	Valor
Épocas	Tiempo de entrenamiento	16
Capas Ocultas	Cantidad de capas ocultas	536
Lotes	Tamaño del lote	128
Tasa	Tasa de aprendizaje	10^{-5}
Neuronas	Neuronas por capa	300
NFinal	Neuronas en capa final	3
Capa Final	Algoritmo en capa final	SoftMax
Activación	Función de Activación	Relu

4.3.3 Evaluación de resultados

Los resultados obtenidos por el modelo propuesto con el conjunto de entrenamiento son satisfactorios, en la Tabla 5 se realiza un análisis del comportamiento con relación a los modelos de análisis de sentimiento en español pysentimiento (Pérez et al., 2021), basado en el modelo pre-entrenado BETO (Cañete et al., 2020) y el propuesto por Lampert multilenguaje (Lampert y Lampert, 2021) presentes en Huggingface. Aunque alguno de estos modelos solo posee un umbral para dos clases (positivo y negativo) se utilizan los umbrales de este trabajo, incluyendo la etiqueta para Neutral en valores entre 0,4 y 0,6.

En general, el modelo BERT-BiLSTM propuesto consigue un mejor rendimiento en comparación con el modelo pysentimiento y el multilenguaje de Lampert para el idioma español en todas las mediciones.

En el caso de la comparativa con el modelo pysentimiento, posee un Valor de F 16 % mejor en la evaluación de los positivos, un 13 % en los neutros y un 14,8 % en los negativos, resultados favorables, aunque pueden estar relacionados con el conjunto de entrenamiento de este modelo, orientado a Twitter, el cual presenta un vocabulario diferente al entrenado para nuestro trabajo. De igual manera, aunque el entrenamiento de pysentimiento se realiza con un conjunto de entrenamiento superior al modelo BERT BiLSTM propuesto, tampoco se trata de un conjunto grande de datos. Los resultados para las restantes métricas son igualmente superiores, siendo el modelo propuesto superior.

Tabla 5: Evaluación de los resultados de los modelos.

	Sentimiento	Precision	Recall	F1 Score	Accuracy
BERT BiLSTM	Positivo	0,834	0,812	0,823	0,805
	Neutro	0,801	0,796	0,799	
	Negativo	0,836	0,813	0,824	
pySentiment (BETO based)	Positivo	0,672	0,654	0,663	0,671
	Neutro	0,665	0,672	0,669	
	Negativo	0,679	0,673	0,676	
Multilingual Lampert Model (BERT based)	Positivo	0,707	0,693	0,701	0,713
	Neutro	0,689	0,693	0,691	
	Negativo	0,713	0,714	0,713	

En el caso del modelo multilenguaje propuesto por Lampert, el conjunto de entrenamiento es significativo (18.437.530 tweets) y aunque es obtenido también de la plataforma Twitter, sus resultados son superiores a los de pysentimiento. Sin embargo, en comparación con la propuesta del presente trabajo los resultados son un 12,2 % inferiores en Valor-F en comentarios positivos, 10,8 % en los neutros y 11,1 % en los negativos. Se comporta de igual manera con valores inferiores en alrededor del 10,0 % para las restantes métricas de estudio.

5 Conclusiones

Las redes sociales y sitios de reseñas turísticas manifiestan en la actualidad un crecimiento acelerado de la información, la cual es consultada por los diferentes usuarios para planificar sus viajes, tarea nada fácil dado el volumen, variedad y velocidad con que se mueve.

A partir de un amplio estudio de las tendencias en arquitecturas, técnicas para la obtención de los datos, tratamiento de los textos y análisis de sentimiento, el presente trabajo nos muestra una propuesta de arquitectura distribuida y balanceada, la cual, mediante la integración de tecnologías para el procesamiento masivo de datos, la minería de textos y los modelos de redes profundas, sirve como soporte a un sistema de recomendación de atracciones turísticas.

Python, como lenguaje de programación simple, pero a la vez potente, posee en la actualidad todas las bibliotecas necesarias para el desarrollo de aplicaciones, de ahí que su utilización en todas las fases de nuestro proyecto no sea casual, sino la muestra de su consolidación tanto en ambientes académicos como en soluciones de la vida real.

Los modelos basados en redes profundas han ganado en los últimos años un gran protagonismo en un sinnúmero de aplicaciones por sus excelentes resultados. La propuesta de un modelo basado en BERT BiLSTM para el análisis de la polaridad de las reseñas obtenidas de la web TripAdvisor a partir de los comentarios de sitios turísticos andaluces en 6 idiomas, nos permite alcanzar valores superiores al 80 % de precisión utilizando la medida Valor de F, la que combina los valores de la precisión y exhaustividad.

El modelo propuesto, aunque sujeto a mejoras, se complementa perfectamente con modelos pre-entrenados existentes para la vectorización del vocabulario y la traducción automática, presentando un rendimiento superior a estudios anteriores publicados en la plataforma Huggingface.

Combinando los resultados obtenidos en el análisis de la polaridad de las reseñas de cada atracción turística y la evaluación media de las mismas, podemos brindar una calificación más acertada a utilizar en un sistema de recomendación.

6 Trabajos Futuros

El desarrollo del presente trabajo ha permitido un acercamiento inicial a un amplio universo en crecimiento donde existen un grupo de tecnologías emergentes con resultados crecientes y con un elevado nivel de aplicación.

Las limitaciones en recursos para el procesamiento y almacenamiento de los datos, nos obligan a escalar la arquitectura propuesta a un entorno de servidores dedicados que nos permitan alcanzar el máximo partido al almacenamiento y procesamiento de estos grandes volúmenes de datos procedentes de diferentes fuentes.

Por otro lado, el desarrollo de bibliotecas que nos permitan ampliar nuestras fuentes de datos a redes sociales y otros sitios de reseñas con el uso de técnicas de web scrapping y APIs para desarrolladores, permitirán la creación de una base de datos con mayor número de datos y por ende un mayor conjunto de entrenamiento-prueba para nuestro modelo, alcanzando mejores resultados para la recomendación. Ampliar el alcance de los contenidos a sitios o atracciones fuera del ámbito de Andalucía permitirá de igual manera el crecimiento de los datos con que cuenta el sistema.

Se recomienda de igual manera, realizar entrenamientos del modelo con un mayor número de elementos en los conjuntos de entrenamiento-prueba, ajustando los parámetros para alcanzar una mayor precisión en los resultados. Extender la propuesta a un modelo multilenguaje basado en BERT-BiLSTM hace que podamos prescindir de la capa de traducción utilizada, pudiendo alcanzar con un correcto entrenamiento mejores resultados.

Finalmente, desarrollar un sistema de recomendación para probar los resultados obtenidos y evaluar su impacto es el siguiente paso en la hoja de ruta trazada, otorgando un enfoque práctico en un entorno real a la investigación. Como parte del desarrollo del sistema se debe evaluar la incorporación de otros contenidos que permitan la recomendación, como el contexto geográfico, el reconocimiento de patrones en imágenes, entre otros.

Referencias Bibliográficas

Abbasi-Moud, Z., Vahdat-Nejad, H. & Sadri, J. (2021). Tourism recommendation system based on semantic clustering and sentiment analysis. *Expert Systems with Applications*, 167, 114324.

Alaei, A., Becken, S. & Stantic, B. (2019). Sentiment Analysis in tourism: capitalizing on big data. *Journal of travel research,* 58 (2). 175-191.

Al-Natour, S., & Turetken, O. (2020). A comparative assessment of sentiment analysis and star ratings for consumer reviews. International Journal of Information Management, 54, 102132. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2020.10213

Akdim, K., Belanche, D. and Flavi, M. (2021). "Attitudes toward service robots: analyses of explicit and implicit attitudes based on anthropomorphism and construal level theory", International Journal of Contemporary Hospitality Management, available at: https://doi.org/10.1108/IJCHM-12-2020-1406.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 2623-2631.

Bansal, A., & Kumar, N. (2022). Aspect-Based Sentiment Analysis Using Attribute Extraction of Hospital Reviews. New Generation Computing, 40(4). https://doi.org/10.1007/s00354-021-00141-3.

Beyer, M. A., & Laney, D. (2012). The importance of 'big data': A definition. Gartner Report, June version, ID G00235055.

Bi, J.-W., Han, T.-Y., & Yao, Y. (2023). Collaborative forecasting of tourism demand for multiple tourist attractions with spatial dependence: A combined deep learning model. Tourism Economics, 135481662311539. https://doi.org/10.1177/13548166231153908.

Bigne, E., Ruiz, C., Perez-Cabañero, C., & Cuenca, A. (2023). Are customer star ratings and sentiments aligned? A deep learning study of the customer service experience in tourism destinations. Service Business, 17(1), 281–314. https://doi.org/10.1007/s11628-023-00524-0.

Bortoluzzi, D.A., Lunkes, R.J., Santos, E.A.D. and Mendes, A.C.A. (2020). "Effect of online hotel reviews on the relationship between defender and prospector strategies and management controls", International Journal of Contemporary Hospitality Management, 32(12), 3721-3745.

Cai, R., Qin, B., Chen, Y., Zhang, L., Yang, R., Chen, S., & Wang, W. (2020). Sentiment Analysis About Investors and Consumers in Energy Market Based on BERT-BiLSTM. IEEE Access, 8, 171408–171415. https://doi.org/10.1109/ACCESS.2020.3024750

Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., & Pérez, J. (2023). Spanish pre-trained bert model and evaluation data. PML4DC at ICLR, 2020.

Chen, T., Duan, Y., Ahmad, F., & Liu, Y. (2023). Detecting Novelty Seeking From Online Travel Reviews: A Deep Learning Approach. IEEE Access, 11, 43869–43881. https://doi.org/10.1109/ACCESS.2023.3253040.

Demirel, E., Zeren, S.K. and Hakan, K. (2022). "Smart contracts in tourism industry: a model with blockchain integration for post pandemic economy", Current Issues in Tourism, 25(12), 1895-1909.

Esmaeili, L., Mardani, S., Golpayegani, S.A.H. & Madar, Z.Z (2020). A novel tourism recommender system in the context of social commerce. Expert Systems with Applications, 149, 1133.

Essien, A., & Chukwukelu, G. (2022). Deep learning in hospitality and tourism: a research framework agenda for future research. International Journal of Contemporary Hospitality Management, 34(12), 4480–4515. https://doi.org/10.1108/IJCHM-09-2021-1176.

Flores-Ruiz, D., Elizondo-Salto, A. & Barroso-Gonzalez, M.D.L.O. (2021). Using social media in tourism sentiment analysis: A case study of Andalusia during the COVID-19 pandemic. Sustainability, 13(7), 45-67.

Generosi, A., Ceccacci, S. and Mengoni, M. (2018). "A deep learning-based system to track and analyze customer behavior in retail store", In 2018 IEEE 8th International Conference on Consumer Electronics-Berlin (ICCE-Berlin), IEEE, pp. 1-6.

González, G., Delgado, T., Capote, J. & Cruz, R. (2013). Context-Aware Recommender Systems based on ontologies. Spatially Enablement in Support of Economic Development and Poverty Reduction. GSDI Association Press, 227-243.

Han, S., & Anderson, C. K. (2021). Web scraping for hospitality research: Overview, opportunities, and implications. *Cornell Hospitality Quarterly*, *62*(1), 89-104.

Ivanov, S. and Webster, C. (2021). "Willingness-to-pay for robot-delivered tourism and hospitality services–an exploratory study", International Journal of Contemporary Hospitality Management, 33(11).

Kang, H. W., Chye, K. K., Ong, Z. Y., & Tan, C. W. (2022). Sentiment Analysis on Malaysian Airlines with BERT. The Journal of The Institution of Engineers, Malaysia, 82(3). 47-52. https://doi.org/10.54552/v82i3.98.

Laaroussi, H., Guerouate, F., & Sbihi, M. (2023). A novel hybrid deep learning approachfor tourism demand forecasting. International Journal of Electrical and Computer Engineering (IJECE), 13(2), 1989-1996. https://doi.org/10.11591/ijece.v13i2.

Lampert, J., & Lampert, C. H. (2021). Overcoming rare-language discrimination in multi-lingual sentiment analysis. IEEE Big Data 2021, pp. 5185-5192.

Liu, J., Hu, S., Mehraliyev, F., & Liu, H. (2023). Text classification in tourism and hospitality – a deep learning perspective. International Journal of Contemporary Hospitality Management. 35(7), 68-81, https://doi.org/10.1108/IJCHM-07-2022-0913.

Macak, M., Ge, M., & Buhnova, B. (2020). A Cross-Domain Comparative Study of Big Data Architectures. International Journal of Cooperative Information Systems, 29(4), 2030001. https://doi.org/10.1142/S0218843020300016.

Mantyla, M.V., Graziotin, D. & Kuutila, M. (2018). The Evolution of sentiment analysis – a review of research topics, venues and top cited papers. *Computer Science Review,* 27, 16-32.

Mariani, M. & Baggio, R. (2022). Big data and analytics in hospitality and tourism: A systematic literature review. *International Journal of Contemporary Hospitality Management,* 34 (1), 231-278.

Mehraliyev, F., Chan, I. C. C., & Kirilenko, A. P. (2022). Sentiment analysis in hospitality and tourism: a thematic and methodological review. International Journal of Contemporary Hospitality Management, 34(1), 46–77. https://doi.org/10.1108/IJCHM-02-2021-0132.

Miranda, G., & Delgado, T. (2015). Big data architecture for social media sentiment analysis supporting context aware recommendation systems. In 5th International Workshop on Knowledge Discovery, Knowledge Management and Decision Support, EUREKA 2015.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544-551.

Reiter, E. (2018). A structured review of the validity of BLEU. Computational Linguistics, 44(3), 393-401.

Pearce, K., Zhan, T., Komanduri, A., & Zhan, J. (2021). A comparative study of transformer-based language models on extractive question answering. arXiv:2110.03142.

Pérez, J. M., Giudici, J. C., & Luque, F. (2021). pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. arXiv e-prints, arXiv-2106.

Preethi, G., Krishna, P.V., Obaidat, M.S., Saritha, V. and Yenduri, S. (2017), "Application of deep learning to sentiment analysis for recommender system on cloud", In 2017 International Conference on Computer, Information and Telecommunication Systems (CITS), IEEE, pp. 93-97.

Singh, N., Hamid, Y., Juneja, S., Srivastava, G., Dhiman, G., Gadekallu, T. R., & Shah, M. A. (2023). Load balancing and service discovery using Docker Swarm for microservice based big data applications. Journal of Cloud Computing, 12(1), 4. https://doi.org/10.1186/s13677-022-00358-7.

SmartData Andalucía (2023). Últimos datos turísticos, disponible en: https://smartdata.andalucia.org.

Snowball. (2023). "Snowball: A language for stemming algorithms" disponible en: http://snowball.tartarus.org/texts/introduction.html.

Statista, Number of user reviews and opinions on Tripadvisor worldwide 2014-2022, disponible en: https://www.statista.com/statistics/684862/tripadvisor-number-of-reviews/.

Tiedemann, J., & Thottingal, S. (2020) OPUS-MT--Building open translation services for the World. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. European Association for Machine Translation, EAMT 2020.

Xiang, Z., Du, Q., Ma, Y. & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, 58, 51-67.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., & Matsumoto, Y. (2020). Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. EMNLP 2020, 23.

van Setten, M., Pokraev, S. & Koolwaaij, J. (2004). Context-Aware Recommendation in the Mobile Tourism Application COMPASS. *Springer, pp. 235-24.*

Vieira, B. M., Borges, A. P., & Vieira, E. P. (2023). The role of social networks for decision-making about tourism destinations. International Journal of Internet Marketing and Advertising, 18(1), 12-28. https://doi.org/10.1504/IJIMA.2023.128148.

Wang, X. (2022). Implementation of Personalized Information Recommendation Platform System Based on Deep Learning Tourism. Journal of Sensors, Especial Issue Image Analysis of Vision Sensors 2022, 6221413. https://doi.org/10.1155/2022/6221413.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, NIPS 2017, 30.

Wu, D. C., Zhong, S., Qiu, R. T. R., & Wu, J. (2022). Are customer reviews just reviews? Hotel forecasting using sentiment analysis. Tourism Economics, 28(3), 795–816. https://doi.org/10.1177/13548166211049865.

Zhang, X., & Liu, C. A. (2023). Model averaging prediction by K-fold cross-validation. Journal of Econometrics, 235(1), 280-301.