Segmentación de clientes del sector turístico de Santander mediante el uso de análisis de sentimientos

Diana Carolina Gómez Vargas

Trabajo de Grado para optar por el título de Ingeniera Industrial

Director: Henry Lamos Díaz

Phd. en Física, Matemática

Universidad Industrial de Santander

Facultad de Ingenierías Físico-mecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2023

Agradezco a mi familia y a mis amigos por su constante apoyo, agradezco a la Universidad Industrial de Santander y a sus profesores por aportar tanto en mi crecimiento profesional como en mi crecimiento interpersonal. Así mismo, un agradecimiento especial al profesor Henry Lamos, por su ayuda y acompañamiento. Finalmente, honrar la memoria del profesor David Puentes, quien siempre mostró un profundo compromiso con la enseñanza y sus estudiantes. Su legado seguirá inspirándonos.

Tabla de contenido

Lista de Figuras

Lista de Tablas

Lista de Apéndices

(Los apéndices se encuentran en la carpeta adjunta)

Apéndice A. Base de Datos

Apéndice B. Código de Extracción de Datos

Apéndice C. Código de Polaridad

Apéndice D. Código de Análisis de sentimientos

Apéndice E. Código de Clasificación por categorías

Resumen

Título: Segmentación de clientes en el sector turístico de Santander mediante el uso de análisis de sentimientos[1]

Autor: Diana Carolina Gómez Vargas[2]

Palabras clave: Turismo, Procesamiento de lenguaje natural (PLN), minería de datos, análisis de sentimientos, Transformer, Aprendizaje profundo

Descripción: Este proyecto se centra en el diseño de un modelo destinado a comprender de manera más profunda a los clientes en el sector turístico de Santander. Se logra esto mediante el análisis de las opiniones de los usuarios que se encuentran en las reseñas de hoteles en TripAdvisor. El incremento del turismo en la región, junto con las tendencias emergentes después de la pandemia, ha generado un cambio en las preferencias de los viajeros.

Para afrontar este desafío, se aplicaron técnicas de Procesamiento de Lenguaje Natural (PLN) y Minería de Datos para evaluar dichas opiniones. Esta iniciativa permite establecer una conexión más sólida con cada tipo de viajero. A través de este proyecto, se identificaron patrones y tendencias ocultas en estas opiniones, lo cual revela preferencias esenciales relacionadas con la gastronomía, la ubicación, el servicio y otros aspectos fundamentales. Estos hallazgos pueden servir como una guía valiosa para las estrategias de marketing en el sector hospitalario de Santander.

---

[1] Proyecto de grado

[2] Facultad de Ingenierías Físico-mecánicas. Escuela de Estudios Industriales y Empresariales. Programa de Ingeniería Industrial. Director: Henry Lamos Díaz.

Abstract

Title: Customer segmentation in Santander's tourism sector using sentiment analysis[3]

Authors: Diana Carolina Gómez Vargas[4]

Keywords: Tourism, Natural language processing (NLP), Data mining, sentiment analysis, Transformer, Deep learning.

Description: This project focuses on designing a model to gain a deeper understanding of customers in the Santander tourism sector. This is achieved through the analysis of user opinions found in hotel reviews on TripAdvisor. The increase in tourism in the region, coupled with emerging trends post-pandemic, has led to shifts in traveler preferences.

To address this challenge, Natural Language Processing (NLP) and Data Mining techniques were employed to evaluate these opinions. This initiative enables a stronger connection with each type of traveler. Through this project, hidden patterns and trends in these opinions were identified, revealing crucial preferences related to cuisine, location, service, and other fundamental aspects. These findings can serve as a valuable guide for marketing strategies in the Santander hospitality sector.

[3] Bachelor thesis
[4] Mechanical Physicist Engineering Faculty. Industrial and entrepreneurial School. Industrial Engineering. Director: Henry Lamos Díaz.

Introducción

Un aspecto fundamental en el desarrollo de la civilización humana ha sido la recopilación y procesamiento de datos. Por ejemplo, en el antiguo Imperio Romano, se registraban los impuestos de los ciudadanos mediante inscripciones en tablillas. En el siglo XIX, se utilizaban tarjetas perforadas para almacenar datos en la máquina de Babbage. En el siglo XX, surgieron sistemas electrónicos para el almacenamiento y gestión de bases de datos, allanando el camino para los avances tecnológicos actuales. Estos avances permiten la creación de modelos o herramientas de análisis de datos que pueden adaptarse al crecimiento constante de información, tras la invención y desarrollo de internet. Estos nuevos datos, provenientes de diversas fuentes, son analizados por empresas, comunidades científicas y entidades gubernamentales y no gubernamentales en busca de valor, con el objetivo de fomentar la innovación mediante la creación de nuevos productos y servicios. Un gran volumen de estos datos corresponde a la categoría de "datos no estructurados", que incluye correos electrónicos, tweets, mensajes instantáneos (SMS, WhatsApp, Viber, Line, etc.), chats en tiempo real (Gmail, etc.), vídeos, fotos, registros de conexiones, registros y datos generados por sensores. Estos datos no estructurados representan el 95% de toda la información disponible para organizaciones y empresas (IBM, 2021).

Debido a la continua necesidad de recopilar y analizar información, surge la minería de datos, cuyo objetivo es descubrir patrones ocultos en los datos. Actualmente, existen técnicas eficientes y efectivas para el análisis tanto de datos estructurados como no estructurados. Una rama de la minería de datos es la minería de texto, que busca extraer información útil y relevante de diversos formatos de documentos, como páginas web, correos electrónicos, redes sociales, artículos de revistas, entre otros.

Además, este campo de la lingüística computacional ha adquirido relevancia en los últimos años en el sector turístico. Esto se debe en gran parte al volumen de comentarios y reseñas proporcionados por los usuarios de estos servicios. Cuando se analizan adecuadamente, estos comentarios ofrecen información valiosa a las empresas del sector, lo que les permite diseñar y ejecutar estrategias de marketing y mejorar la calidad del servicio, lo que a su vez aumenta la satisfacción del cliente.

El sector turístico en la región oriental de Colombia, como Santander, ha experimentado un impresionante aumento del 353.3% en el PIB entre 2020 y 2021(Alguero, 2021), Sin embargo, esta recuperación económica puede potenciarse aún más mediante el uso de herramientas contemporáneas que permitan orientarse y analizar los aspectos más críticos para promover procesos de innovación. Estos procesos involucran la estructuración de procedimientos, la recolección de datos y la formulación de estrategias enfocadas en alcanzar los objetivos establecidos por las tendencias actuales, con el fin de mantenerse a la vanguardia en un entorno cada vez más competitivo en América Latina y el Caribe.

En esta investigación, el objetivo es realizar una segmentación de mercados utilizando un enfoque de análisis de sentimientos, aprovechando las reseñas y datos sobre hoteles en Santander recopilados a través de la plataforma TripAdvisor. Este esfuerzo tiene como finalidad capacitar al sector para analizar tendencias y, en consecuencia, mejorar la calidad del servicio y la satisfacción del cliente.

Tabla 1

Cumplimiento de objetivos

Objetivos	Cumplimiento
Identificar las necesidades de información para realizar una segmentación de clientes	2. Planteamiento del problema 3. Marco de referencia
Realizar una revisión de literatura que permita obtener un diagnóstico situacional acerca del análisis de sentimientos y la minería de datos en el sector turismo	5. Análisis de literatura
Recolectar y procesar los datos de opinión de los clientes	6. Datos
Identificar las características relevantes de los clientes	7. Análisis de sentimientos 8. Clasificación de clientes
Validar el modelo segmentación de clientes	9. Resultados

1. Objetivos

1.1. Objetivo General

Diseñar un modelo de segmentación de clientes mediante análisis de sentimientos para el sector turístico de Santander

1.2. Objetivos Específicos

- Identificar las necesidades de información para realizar una segmentación de clientes
- Realizar una revisión de literatura que permita obtener un diagnóstico situacional acerca del análisis de sentimientos y la minería de datos en el sector turismo
- Recolectar y procesar los datos de opinión de los clientes
- Identificar las características relevantes de los clientes
- Validar el modelo segmentación de clientes

2. Planteamiento del Problema

En los años recientes el turismo ha tenido un crecimiento económico superior al que presenta la economía mundial; razón por la cual, el Gobierno Nacional de Colombia ha promovido e invertido en este sector de tal forma que permitió posicionar al país en el puesto 55 entre 140 países según la edición de Travel & Tourism Competitiveness Report 2019.

En la actualidad, el turismo se ha diversificado en términos de la exigencia de los viajeros, las tendencias que surgieron a partir del 2021 a raíz de todas las situaciones de contingencia que se presentaron por la pandemia del covid-19; se basan en lo que describe (FORBES, 2022) como las nuevas experiencias de viajeros de lujo, ya que buscan las siguientes aspectos: La sustentabilidad, como balance de relación económica, ambiental y social con las comunidades locales; además de esto, un agente experimentado, con el fin de evitar las molestias y en pocas palabras pagar por ahorrar tiempo y energía del viajero, claro que no aplica para todos, ya que algunos otros empezaron a ver la planeación anticipada como una gran opción para ahorrar dinero por lo que se observan reservaciones nacionales con 58 días de anticipación e internacionales con 80 días de anticipación, lo cual conduce a buscar las mejores épocas de viaje, es decir, la mejor época donde puedas tomar las fotos tranquilo, sin la molestia de tener que lidiar con grandes grupos de personas, altos precios, por eso la temporada baja se ha vuelto el nuevo atractivo a tomar en cuenta, además de buscar destino diferentes, la nueva moda es lo que no está de moda, así que los viajeros buscan por lugares nuevos que no sean conocidos para tener la libertad y tranquilidad que buscan los viajeros luego de un largo tiempo de confinamiento, ya que la seguridad, versatilidad y adaptabilidad a cada viajero se convierte en lo primordial como se explica en (Procolombia, 2021).

La reactivación que se dio a finales del año 2021 (Portafolio, 2021), es una oportunidad de crecimiento y cambio para Santander, ya que está dentro de los 10 primeros departamentos con mayor flujo turístico en Colombia (MINCIT, 2021), territorio que se encuentra en el top 7 de países con mayores visitantes en Latinoamérica[5] por lo que refleja el interés en la investigación y desarrollo del sector turístico en la actualidad.

Así mismo, se ha observado que ante el cambio de perspectiva que ofreció la pandemia, surgieron enfoques alternativos en la comprensión del comportamiento de las personas, debido a la exploración de investigaciones asociadas al estudio del procesamiento del lenguaje natural (PLN), en donde describe opiniones, gustos y disgustos, cambiando no solo la forma de analizar los comentarios sino la forma en que se agrupan personas según sus intereses en común, como sucede en el estudio de redes sociales (SN por sus siglas en inglés, *Social Networks*). Debido a las herramienta de tratamiento de texto asociadas al PNL y al acceso a internet, ha dejado de ser estrictamente necesaria la formulación y recolección de datos por medio de cuestionarios físicos o virtuales para conocer la opinión y perspectiva de una persona sobre un tema en específico, como lo es un atractivo turístico, ya que ellos mismos ofrecen esta información dentro de los comentarios que se encuentran en plataformas de internet, haciendo el trabajo de recolección de datos más sencillo y menos costoso para los analistas, además de que se pueden obtener datos más diversos en personas, que en el caso de los cuestionarios ya que se limitan a la capacidad financiera y logística del proyecto.

En la presente investigación se quiere abordar el reto que presentan las nuevas tendencias de los viajeros con el uso de algoritmos de Análisis de Sentimientos (AS) aplicado a la segmentación de clientes (Herrera, 2020). Se implementará el AS como herramienta de análisis

[5] Organización Mundial del Turismo (2021), Panorama del Turismo internacional, Edición 2020. DOI: https://doi.org/10.18111/9789284422746.

de texto, ya que permite identificar opiniones y sentimientos de los integrantes que hacen parte de una red social (Medhat et al., 2014) (Baviera, 2017), (Cardoso et al., 2019), (Matuschka, 2021). Una vez identificados los sentimientos de las personas respecto a temas o palabras relativas a las tendencias, se realizará la clasificación de comunidades según la temporada en que se hospedaron en los hoteles (REYES & OVIEDO, 2013) (Juan Guillermo Martínez Cano, 2018). Lo cual resulta en una segmentación de clientes que ofrece información e identificación de posibles grupos de interés, apoyando así, el proceso de toma de decisiones.

3. Marco de Referencia

En este capítulo, se expone el marco de antecedentes y el marco teórico relacionado con la segmentación de clientes mediante análisis de sentimientos en el mercado turístico. En el marco de antecedentes, se presentan estudios previos relevantes en esta área, mientras que en el marco teórico se abordan los conceptos fundamentales necesarios para el desarrollo del trabajo. El objetivo es establecer una base sólida de conocimientos que respalde la implementación efectiva de la segmentación de clientes basada en análisis de sentimientos en el sector turístico de interés.

3.1. Marco de Antecedentes

(Mohalem, 2018) propone la creación de un sistema de recomendación basado en los datos recolectado en la web, datos open Access (o de libre acceso) que proveen información de vital importancia si se procesa y estructura con una orientación definida como es el caso de un sistema de recomendaciones que se adecue a la interacción que el usuario tenga con el sistema con el fin de que estas recomendaciones sean más finas al detalle de interés o momento del usuario, realizando esto por medio de la herramienta MapReduce que está diseñada para el

procesamiento de grandes cantidades de datos con el fin de clasificarlas y asignar tareas alineadas con la resolución de problemas del mundo real tal como sucede en el presente proyecto en donde se debe procesar una gran cantidad de datos con el fin de identificar y agrupar individuos con características o patrones similares.

(Alaei et al., 2019) Este artículo es la base de investigación para este documento ya que muestra una revisión de literatura y evaluación de los métodos disponibles para el análisis de sentimientos aplicado al turismo, caracterizando cada algoritmo y realizando un cuadro comparativo según autor, clasificador, base de datos, número de reseñas, número de clases, entre otros. Con el fin de aportar aplicaciones según el algoritmo o método de análisis de sentimientos en el sector turismo.

(Jiménez & Jiménez, 2019) En su tesis evalúan el desempeño que tienen dos modelos de aprendizaje no supervisado con respecto al procesamiento de datos, más específicamente de tweets relacionados con los incendios forestales presentados en California, Estados Unidos en el año 2018, en donde se obtiene que el algoritmo LDA frente al algoritmo K-means, ya que es la opción óptima al momento de decidir sobre la mayor eficiencia que tienen en obtener los resultados deseados a partir de una gran base de datos, por lo cual brinda herramientas y metodologías para el desarrollo de minería de texto acorde a los objetivos y recursos determinados en el presente proyecto.

(Ayala & Rudas, 2019) y (García & Suárez, 2021) exponen en sus tesis la aplicación de técnicas de agrupamiento (clustering) para diferentes fines tales como la detección de comunidades y el análisis de tendencias; fines y metodologías altamente relacionados con el proyecto que se lleva a cabo, por lo que generan contraste en cuanto a la diversidad de técnicas

asociadas a este tema como evidencia literaria de la importancia que tiene el clustering dentro del ámbito de análisis de datos y minería de texto.

(Pineda, 2021) aporta en su tesis una revisión literaria concisa y comparativa de diferentes clasificadores y algoritmos capacitados para el desarrollo del análisis de sentimientos, especificando sus fortalezas y debilidades, además de mencionar los ámbitos más adecuados para cada uno de estos clasificadores que sustentan la ejecución del análisis de sentimientos ejecutado con el fin de realizar una predicción basado en un sistema multivariable en donde intervienen el análisis del lenguaje natural, el de indicadores técnicos y el de precios.

Finalmente, en su tesis de posgrado, (Gauch et al., 2022) realizan una contribución significativa al implementar el algoritmo BERT en el análisis de sentimientos en reseñas de películas. Su enfoque se destaca por su objetivo de mejorar la precisión en comparación con los métodos existentes en el campo. Para lograrlo, combinan técnicas de clasificación de polaridad y distribución de sentimientos a múltiples escalas, mientras incorporan una capa adicional del BiLSTM para adaptar BERT al contexto específico del proyecto.

3.2. Marco Teórico

3.2.1. Inteligencia Artificial (IA)

IA es construida bajo la hipótesis de que el pensamiento mecanizado es posible.

El nacimiento de la IA tal y como es conocida hoy en día, comenzó con la publicación de Alan Turing de "Computing Machinery and Intelligence" en 1950. En dicho escrito, Turing examinó la idea de cómo saber si las máquinas pueden pensar. De esta manera, se lleva a conocer el Juego de la Imitación que requiere de tres jugadores. El jugador A es un ordenador y el El jugador B es un humano. Cada uno debe convencer al jugador C (un humano que no puede ver ni

al Jugador A o Jugador B) de que son humanos. Si el jugador C no puede determinar quién es humano y quién no lo es de forma consistente, el ordenador gana. (TURING, 1950)

(TURING, 1950) también parte de refutar una objeción por no decir paradigma expresado por Lady Lovelace's, la cual se refiere a que "una máquina solo puede hacer lo que le indiquemos hacer", ya que a través de los años no solo se ha determinado que las máquinas no solo pueden aprender, sino que pueden realizar un sin número de cosas que se creían reservadas únicamente para el ser humano como las siguientes aplicaciones (Rouhiainen, 2018):

- Reconocimiento de imágenes estáticas, clasificación y etiquetado.
- Mejoras del desempeño de la estrategia algorítmica comercial.
- Procesamiento eficiente y escalable de datos de pacientes
- Mantenimiento predictivo
- Detección y clasificación de objetos
- Distribución de contenido en las redes sociales
- Protección contra amenazas de seguridad cibernética

3.2.2. Aprendizaje Automático

El machine learning (en su traducción en Inglés), es un tipo de IA que le brinda la facultad a un sistema de aprender a partir de datos externos que se modelan de acuerdo a la programación interna, que en el caso del ser humanos es que aprende de acuerdo a estímulos externos, teniendo en cuenta los aprendizajes previos, para llevarlo a un ejemplo sencillo: sería decir que para aprender a resolver ecuaciones se necesita saber sumar, restar, multiplicar y dividir, en el machine learning es similar, se programa a una máquina para aprender qué hacer con una entrada de datos externa, con el fin de convertir datos en programas sin requerir hacerlos (Russell, 1994). Pero esto, es un proceso complejo en el que a medida que se incrementa la

adquisición de datos de entrenamiento el modelo que se quiera llevar a cabo será cada vez más preciso al propósito por el cual se creó (IBM, 2022), y dependiendo de la presión, el objetivo y los recursos se elige entre la variedad de tecnologías, tales como, Aprendizaje profundo, minería de texto, minería de datos, entre otros que pueden ser utilizados en el aprendizaje supervisado o no supervisado (Gartner, 2019)

El aprendizaje automático utiliza algoritmos para aprender y determinar los patrones de datos para tomar decisiones, un ejemplo de ello es los filtros de spam en el correo electrónico, en donde identifica con base a palabras y oraciones, entre otros para decidir si son correos basura y separarlos de los demás. Y para mayor detalle del funcionamiento del aprendizaje automático se muestran los tres subconjuntos, a continuación:

Figura 1.

Tipos de Aprendizaje Automático

Nota: Información tomada de (Rouhiainen, 2018)

Una buena forma de explicar las diferencias entre estas tres clases es mediante un ejemplo y observar cómo interviene o resulta en cada caso. Si se tienen 10.000 fotografías y los algoritmos deben identificar las fotos en las que este un gato.

*"En el **aprendizaje supervisado**, los algoritmos usan datos que ya han sido etiquetados u organizados previamente para indicar cómo tendría que ser categorizada la nueva información. Con este método, se requiere la intervención humana para proporcionar retroalimentación. Volviendo a nuestro ejemplo, enseñaremos previamente al algoritmo fotos donde apareciera un gato para que luego pudiera identificar imágenes similares.*

*En el **aprendizaje no supervisado**, los algoritmos no usan ningún dato etiquetado u organizado previamente para indicar cómo tendría que ser categorizada la nueva información, sino que tienen que encontrar la manera de clasificarlas ellos mismos. Por tanto, este método no requiere la intervención humana. En el ejemplo, los algoritmos tendrían que clasificar ellos mismos todas las fotos en las que apareciera un gato en una categoría.*

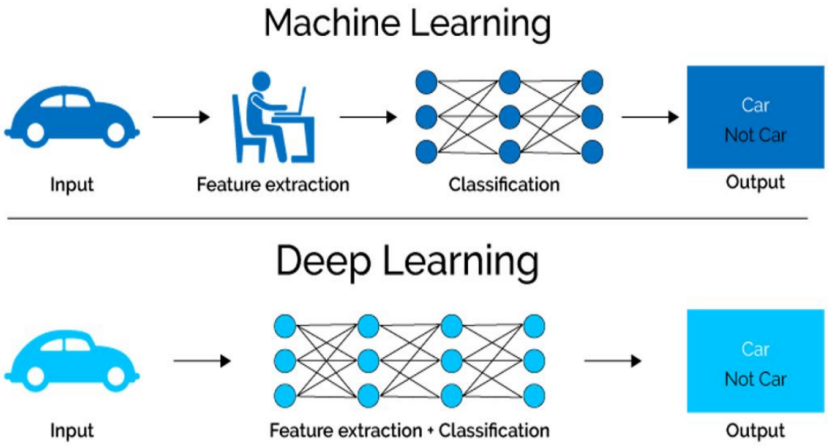
*Por último, con el **aprendizaje por refuerzo**, los algoritmos aprenden de la experiencia. En otras palabras, tenemos que darles «un refuerzo positivo» cada vez que aciertan. La forma en que estos algoritmos aprenden se puede comparar con la de los perros cuando les damos «recompensas» al aprender a sentarse, por ejemplo."* (Rouhiainen, 2018)

3.2.3. Aprendizaje profundo

El aprendizaje profundo (Deep learning en su traducción en inglés) se estructura a partir de la configuración de parámetros básicos acerca de los datos y entrena a la computadora para que aprenda a través del reconocimiento de patrones y mediante el uso de múltiples capas de procesamiento (Gartner, 2019), y más allá de ser un tipo de aprendizaje automático, no son lo mismo, su principal diferencia se deriva en el hecho de que el aprendizaje profundo se basa en la extracción de características y clasificación simultáneamente luego de recibir una entrada, algo que en machine learning ocurre por separado, como se observa en la Figura 8 (Puente Ríos, 2021).

Figura 2.

Diferencia entre Aprendizaje Automático y Aprendizaje Profundo.



Fuente: House of Bots (2018). Most Popular 20 Free Online Courses to Learn Deep Learning.

3.2.4. Procesamiento de lenguaje natural

Se comprende como la capacidad de un software o máquina de entender la información que se transfiere por medio del lenguaje (como español o inglés), es decir, procesar los datos administrados como letras, palabras o sonidos que se interpretan como oraciones con un sentido, con una idea a trasmitir al receptor que en este caso es una máquina que debe realizar un procesamiento de lenguaje natural (PLN, o NLP por sus siglas en inglés).

La ciencia encargada del estudio del PLN es la lingüística computacional, los lingüistas son descritos por (Gelbukh, 2010) como aquellos que cumplen con la función de analizar, interpretar y describir el lenguaje en estudio y como la humanidad el lenguaje cambia y evoluciona, así que esos cambios se documentan y se reportan las reglas en los diccionarios que permitirán a las computadoras comprender el lenguaje humano. Pero esto resultaba muy difícil,

lento y poco eficiente hasta la llegada del internet, que facilitó la adquisición de grandes volúmenes de datos a procesar por la máquina.

El sistema PNL según (Augusto Cortez Vásquez et al., 2009) caracteriza por niveles: fonológico, morfológico, sintáctico, semántico y pragmático; que tienen como aplicaciones la traducción automática, extracción de la información y resúmenes, recuperación de la información, resolución cooperativa de problemas, tutores inteligentes y reconocimiento de voz.

3.2.5. Transformer

La arquitectura Transformer fue presentada por primera vez en el artículo de (Vaswani et al., 2017), se basa en el aprendizaje automático y el procesamiento del lenguaje natural (NLP). Destaca por su capacidad para capturar relaciones de largo alcance en el texto, superando las limitaciones de las arquitecturas recurrentes tradicionales. Esto se logra mediante el mecanismo de atención y las capas de codificador y decodificador. El mecanismo de atención asigna pesos a las palabras del texto, permitiendo capturar relaciones contextuales importantes. Las capas de codificador y decodificador procesan la información secuencialmente, generando representaciones de alta calidad. La arquitectura Transformer ha demostrado un rendimiento sobresaliente en tareas de procesamiento del lenguaje, como traducción automática, generación de texto y clasificación de sentimientos.

3.2.5.1. BERT

BERT (Bidirectional Encoder Representations from Transformers) es una arquitectura revolucionaria en el campo del procesamiento del lenguaje natural (NLP). Fue presentada por (Devlin et al., 2019) y ha demostrado un rendimiento excepcional en una amplia gama de tareas de NPL. BERT se basa en la idea de pre-entrenamiento del lenguaje, donde un modelo de

lenguaje se entrena en grandes cantidades de texto no etiquetado para aprender representaciones de palabras contextualizadas.

Lo que hace a BERT especial es su capacidad para capturar relaciones contextuales bidireccionales en el texto. En lugar de depender de arquitecturas recurrentes tradicionales, BERT utiliza la arquitectura Transformer, que permite el procesamiento paralelo y la captura de dependencias a largo plazo en el texto. Esta arquitectura se basa en el mecanismo de atención, que asigna pesos a las palabras del contexto y captura las relaciones más relevantes.

El enfoque de pre-entrenamiento y ajuste detallado de BERT ha revolucionado el campo del NPL, ya que permite a los modelos aprovechar grandes cantidades de datos no etiquetados antes de ajustarse a tareas específicas con datos etiquetados.

Sin embargo, es importante tener en cuenta que cada versión de BERT tiene sus propias características y desempeño comparativo. Por ejemplo, algunas versiones populares de BERT son “DistilBERT”, “BERTbase”, “RoBERTa” y “ALBERT”. Cada una de ellas presenta ciertas ventajas y desventajas en términos de tamaño del modelo, velocidad de procesamiento, precisión y requisitos computacionales. Es esencial evaluar cuidadosamente las necesidades del proyecto y los recursos disponibles al elegir la versión adecuada de BERT, teniendo en cuenta los siguientes detalles por cada versión:

- **DistilBERT:** Esta versión de BERT, se enfoca en la comprensión del modelo base para reducir su tamaño y acelerar su tiempo de procesamiento. Aunque DistilBERT tiene un tamaño más pequeño y más rápido, su rendimiento puede ser ligeramente inferior al modelo base original.

- **BERTbase:** Es la versión base original de BERT y se utiliza como referencia para comparar otras variantes. Proporciona un buen equilibrio entre tamaño del modelo y rendimiento. BERTbase es ampliamente utilizado y ha demostrado ser efectivo en diversas tareas de procesamiento de lenguaje natural.
- **RoBERTa:** se basa en el enfoque de optimización de entrenamiento. Utiliza una técnica de entrenamiento. Utiliza una técnica de entrenamiento más prolongada y más datos, lo que resulta en un modelo más preciso y robusto. RoBERTa ha logrado mejorar el rendimiento en muchas tareas de NLP en comparación con la versión base de BERT.
- **ALBERT:** se diferencia por su estructura. ALBERT utiliza un enfoque de factorización de atención y comparte parámetros entre las capas para reducir la cantidad total de parámetros, lo que resulta en un modelo más eficiente en términos de memoria y cómputo. Aunque puede haber una ligera disminución en el rendimiento en comparación con BERTbase, ALBERT se destaca en entornos con recursos computacionales limitados.

3.2.5.2. Zero-Shot Text Classification via Self-Supervised Tuning

La Zero-Shot Text Classification es una técnica que busca clasificar textos sin etiquetas específicas, aprovechando el aprendizaje de categorías desconocidas. Por otro lado, el Self-Supervised Tuning es un enfoque de aprendizaje automático que utiliza el pre-entrenamiento de modelos de lenguaje para mejorar el rendimiento en tareas específicas.

En el contexto de la investigación (Liu et al., 2023), se propone el método SSTuning, el cual utiliza el objetivo de predicción de la primera oración para entrenar modelos de lenguaje. Este enfoque busca asociar el texto con su etiqueta correspondiente, mejorando así la capacidad

de realizar clasificación de texto sin etiquetas. Los resultados experiméntales demuestran que SSTuning supera a otros enfoques y muestra estabilidad con diferentes diseños de verbalizadores.

3.2.6. Análisis de Sentimientos (or Opinion mining)

(Feldman, 2013) tal como se expresa en el artículo, es el proceso que tiene como objetivo encontrar las emociones que están implícitas dentro de las opiniones de los autores sobre entidades específicas; también se define como “el estudio computacional de opiniones, sentimientos y emociones expresadas en textos” (Dubiau & Ale, 2013)

La información que ofrecen los sistemas de AS son la polarización de sentimientos con respecto a cualquier clase de entidad, producto o idea, que pueda implementarse en como recurso de fidelización de clientes, o de cualquier estrategia que pueda predecir y modelar el comportamiento de las personas. Los algoritmos más mencionados para la clasificación de sentimientos supervisada son: Naive bayes, el cual se basa en el teorema de bayes, es decir, las opiniones se clasifican según la probabilidad de que sea positivo o negativo; modelo de máxima entropía, es un método de clasificación que se enfoca en determinar la distribución de probabilidad que cumpla con todos los parámetros del modelo y que maximice la entropía; Support vector machines (SVM), es un método de clasificación binaria que se basa en encontrar el hiperplano que separe los vectores del conjunto de datos en dos grupos y Árbol de decisión (Decision Trees), la clasificación se basa en las reglas inferidas por el conjunto de decisiones tomados en el árbol de decisiones. (Dubiau & Ale, 2013) Estos son solo algunos de los algoritmos que hacen posible la clasificación de sentimientos en grupos, más comúnmente en 2 grupos.

3.2.6.1.TextBlob

(Loria, 2020) Es una biblioteca de procesamiento de lenguaje natural en Python que proporciona una interfaz sencilla para realizar diversas tareas de análisis de texto, incluido el análisis de sentimientos. El modelo de análisis de sentimientos de TextBlob se basa en técnicas de aprendizaje automático supervisado.

El proceso de TextBlob implica varios pasos, en primer lugar, el texto se divide en oraciones y palabras individuales, lo que se conoce como tokenización. A continuación, se asigna a cada palabra una puntuación de polaridad y subjetividad basada en un modelo previamente entrenado. La polaridad se refiere a la positividad o negatividad de una palabra, mientras que la subjetividad indica en que medida la palabra es objetiva o subjetiva.

Luego, se calcula la polaridad y subjetividad del texto completo utilizando las puntuaciones de las palabras individuales. La polaridad se calcula promediando las puntuaciones de polaridad de todas las palabras, mientras que la subjetividad se calcula promediando las puntuaciones de subjetividad.

El modelo de análisis de sentimientos de TextBlob es conocido por su facilidad de uso y su capacidad para proporcionar resultados rápidos y precisos en varios dominios de texto. Sin embargo, también tiene algunas limitaciones, como la dependencia de modelos previamente entrenados, lo que puede afectar su rendimiento en textos fuera del dominio de entrenamiento. Además, la precisión del modelo puede variar según el idioma y la calidad de los datos utilizados para el entrenamiento.

3.2.7. Detección de comunidades

La investigación de estructuras comunitarias en redes es un problema relevante para el análisis de interacciones sociales, investigaciones biológicas o investigaciones tecnológicas. A su vez estas estructuras de sistemas diversos y redes complejas pueden definirse como:

"Una comunidad es una colección de nodos que son homogéneos dentro del grupo y heterogéneos con otros grupos en la red, y este tipo de red se conoce como estructura comunitaria." (Cui et al., 2018)

Aunque el concepto de comunidad está comúnmente ligado a la clasificación de, por ejemplo, sistemas sociales, redes metabólicas o sistemas computacionales (Radicchi et al., 2004). Depende del contexto en el que se encuentre puede ser sinónimo de grupo, cluster, módulo, etc. Contextos tan relevantes como la definición e identificación de comunidades en redes (Radicchi et al., 2004), y los algoritmos de detección de comunidades ha demostrado ser efectivo tanto en redes virtuales como las de la vida real (Cui et al., 2018)

3.2.8. Clusterización

Es una clasificación de la información de individuos que tiene como fin la segmentación de personas en grupos, personas con características similares que puede ser comúnmente utilizados para formular estrategias de mercadeo, así mismo, esta clasificación puede ser usada en muchos campos del análisis de datos, como lo son el reconocimiento de patrones, el análisis de imágenes, aprendizaje automático, significado de los datos y minería de datos.

Tipos de cluster

Los algoritmos se dividen según el tipo de agrupación de datos, estos pueden ser jerárquicos o particionales, los jerárquicos se basan en grupos predefinidos para realizar la

clasificación, los particionales definen todos los grupos al mismo tiempo, cada tipo de clasificación tiene una subclasificación según (Soni Madhulatha, 2012) depende en lo que se basa el algoritmo, como se observa a continuación:

Agrupación Jerárquica

- Agrupación jerárquica aglomerativa: Algoritmo BIRCH, Algoritmo de desplazamiento medio
- Agrupación divisiva

Clúster Particional

- Basado en centroides: Algoritmo k-means, algoritmo k-medoids.
- Basado en la densidad: algoritmo DBSCAN, algoritmo SNN, Algoritmo OPTICS.
- Basado en modelos/distribución: Algoritmo de mezcla gaussiana.

3.2.9. K-means

Es un algoritmo de agrupamiento o método de cuantificación vectorial, tiene como objetivo dividir las observaciones o datos en grupos, en donde cada grupo cuenta con un centroide que se encarga de caracterizar al grupo, de tal manera que cada observación se asigna al centroide con menor distancia o diferencia a la observación, es decir, se utilizan centroides (centros de conglomerados) para modelar los datos, tal como lo indica el nombre se realiza el agrupamiento de k-medias.

Además, surge del problema de agrupamiento en donde se analiza la identificación de grupos homogéneos, o también llamados clusters, en donde se conocen diferentes métodos de agrupamiento, pero con mayor popularidad según (Likas et al., 2003), debido a que es un

procedimiento de búsqueda local que minimiza el error de agrupamiento, pero su rendimiento suele depender de las condiciones iniciales

Figura 3.

Algoritmo de agrupamiento k-means

$$E\left(m_1, \ldots, m_M\right) = \sum_{i=1}^{N} \sum_{k=1}^{M} I(x_i \in C_k) \|x_i - m_k\|^2$$

Nota: Información tomada de (Likas et al., 2003)

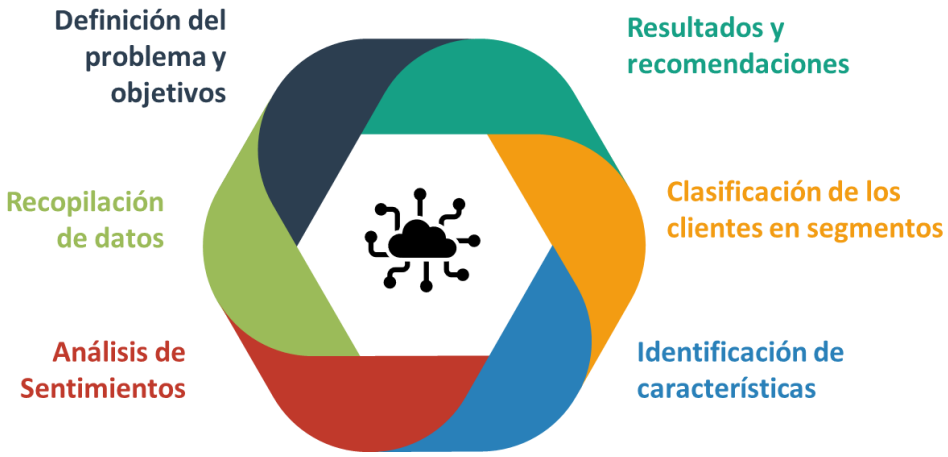
4. Metodología

4.1. Metodología de investigación

las fases de metodología para diseñar un modelo de segmentación de clientes mediante análisis de sentimientos para el sector turístico de Santander incluyen la definición del problema y objetivos, la recopilación y preprocesamiento de los datos, el análisis de sentimientos, la identificación de características relevantes de los clientes, la agrupación de los clientes en segmentos, la validación del modelo y la interpretación de los resultados y recomendaciones.

Figura 4.

Metodología



4.1.1. Fase 1: Definición del problema y objetivos

En esta fase se identifica el problema y se definen los objetivos generales y específicos que se quieren alcanzar con el modelo de segmentación de clientes, con base en la investigación "Sentiment analysis in tourism capitalizing on big data" que podemos abreviar como SAT, se identifican las palabras clave, terminologías y operadores requeridos para la formulación de la ecuación de búsqueda en las bases de datos Web of Science y Scopus donde luego de una serie de intentos en donde se elimina el ruido documental con el fin de obtener los documentos que aporten a la presente investigación.

4.1.2. Fase 2: recopilación y procesamiento de los datos

En esta fase, se recopilan datos de los principales municipios turísticos de Santander de la plataforma digital Tripadvisor, incluyendo el nombre del hotel, la reseña y la fecha de estadía. Estos datos son procesados de manera adecuada para su posterior análisis. Se requiere aplicar

técnicas de limpieza y procesamiento de datos, como eliminación de valores atípicos, corrección de errores y estandarización, con el fin de asegurar la calidad y utilidad de los datos.

4.1.3. Fase 3: Análisis de sentimientos

En esta fase, se aplicará la técnica de clasificación de texto basada en aprendizaje automático a los datos procesados para determinar la polaridad de las opiniones de los clientes. Esta técnica permitirá clasificar las opiniones de los clientes como positivas, negativas o neutras en relación con el sector turístico de Santander. Se utilizará un conjunto de datos etiquetados para entrenar el modelo y una vez que se haya completado el entrenamiento, se aplicará a las nuevas opiniones.

4.1.4. Fase 4: Identificación y Evaluación de las Características Clave

En esta fase, se lleva a cabo la identificación de las características que resultan relevantes para la asignación de categorías en el proceso de segmentación, que comprende la calificación de los aspectos relevantes de la experiencia en un hotel. Estas características abarcan aspectos como limpieza de las habitaciones y áreas comunes, la calidad del servicio brindado por el personal, la excelencia en la oferta gastronómica, la variedad y calidad de las actividades disponibles para los huéspedes, conveniencia de la ubicación del hotel, la estética del diseño y decoración, así como el mantenimiento y conservación de las instalaciones.

Mediante las consideraciones de estos elementos, se busca obtener una comprensión integral de los diferentes aspectos que influyen en la clasificación de los hoteles dentro de la segmentación. Estas características desempeñan un papel fundamental en la percepción y satisfacción de los clientes, ya que influyen directamente en la calidad de su experiencia durante su estancia en el hotel. Por tanto, su identificación y evaluación adecuada permiten establecer

criterios claros y consistentes para la asignación de categorías, facilitando la toma de decisiones informadas por parte de los clientes al seleccionar un hotel según sus preferencias y necesidades.

4.1.5. Fase 5: Clasificación de los clientes en segmentos

En este proceso se implementa una metodología basada en Zero-Shot Text Classification via Self-Supervised Tuning para categorizar las reseñas de los clientes según su afinidad a las categorías establecidas, utilizando técnicas de segmentación. Mediante este enfoque, se busca asignar de manera automática y precisa las reseñas a las categorías correspondientes, permitiendo obtener una visión estructurada de la experiencia del cliente en relación con diferentes aspectos del servicio, como limpieza, servicio, comida, actividades, ubicación, diseño, decoración y mantenimiento.

Para lograrlo, se emplea el proceso de self-supervised tuning, optimizando un modelo del lenguaje pre-entrenado para adaptarlo a las características únicas de las reseñas de los clientes. Esto implica entrenar el modelo en un conjunto de datos generado a partir de las reseñas existentes, lo cual le permite aprender a asociar patrones y características lingüísticas con cada categoría.

4.1.6. Fase 6: Resultados y recomendaciones

En esta fase, se procede a la interpretación exhaustiva de los resultados obtenidos, con el objetivo de proporcionar recomendaciones estratégicas para mejorar tanto la segmentación de clientes como las estrategias de marketing y promoción.

4.2. Validación del modelo

Con el fin de validar el procedimiento de segmentación de clientes utilizando el modelo Zero-Shot Text Classification via Self-Supervised Tuning, se llevan a cabo pruebas en conjuntos

de datos previamente etiquetados. El objetivo principal de esta etapa es confirmar la capacidad del modelo para ser reproducido y utilizado exitosamente en una variedad de contextos y situaciones. Durante este proceso de validación, se evalúa la precisión y el rendimiento general del modelo al clasificar los textos en las categorías establecidas.

Se realiza una comparación entre las predicciones del modelo y las etiquetas reales de los datos de prueba, utilizando métricas como precisión, recuperación y puntuación F1, para determinar la efectividad y confiabilidad del modelo en la segmentación de clientes. Es esencial llevar a cabo múltiples iteraciones de validación, realizando ajustes y mejoras según sea necesario, con el objetivo de garantizar la robustez y la capacidad de generalización del modelo en diversos escenarios del sector turístico.

La validación del modelo Zero-Shot Text Classification via Self-Supervised Tuning en la segmentación de clientes es un paso crucial para asegurar su adecuado funcionamiento y aplicabilidad práctica. Los resultados obtenidos en esta etapa proporcionan la confianza necesaria para utilizar el modelo en la toma de decisiones informadas en estrategias de marketing y promoción, así como en la mejora de la experiencia del cliente en el sector turístico.

5. Revisión de Literatura

En la primera fase de investigación se realiza una revisión de literatura para identificar y analizar información relacionada a la segmentación de clientes, teniendo como herramienta algoritmos de análisis de sentimientos enfocados en el sector turístico.

A partir de lo anterior, se realiza la búsqueda de literatura gris en plataformas digitales institucionales, como: la Organización Mundial del Turismo (OMT), el Ministerio de Comercio, Industria y Turismo (MINCIT), Procolombia, Prosantander y de la Gobernación de Santander; con el fin de recolectar y analizar información relativa al sector turismo tanto a nivel nacional

como internacional y, de este modo, poder obtener un diagnóstico de la situación actual del sector.

Adicionalmente, se hizo una búsqueda en las bases de datos bibliográficas tales como Web of Science y Scopus que permiten recopilar información sobre la literatura científica con el fin de hallar diversas metodologías desarrolladas en el análisis de sentimientos, minería de datos y detección de comunidades dentro del turismo por los diferentes autores a lo largo del tiempo.

5.1. Análisis Bibliométrico

Inicialmente, se realizó una búsqueda exhaustiva utilizando el motor de búsqueda de Google, seguido de un acceso a la versión especializada en contenido académico, con el objetivo de recopilar información detallada sobre el entorno turístico a nivel nacional e internacional y su relación con las nuevas tecnologías. Esta estrategia permitió identificar palabras clave relevantes en los documentos de interés y reveló que, a nivel internacional, existen más de 100 investigaciones y avances en la integración de nuevas tecnologías en la industria turística, mientras que a nivel nacional se evidenció una baja producción científica en este ámbito, especialmente en relación con la segmentación de clientes basada en minería de texto y análisis de sentimientos.

Además, se consideraron importantes organizaciones de turismo, tanto a nivel nacional como internacional, como la OMT, MINCIT y Procolombia, cuyos objetivos principales incluyen la implementación de tecnologías relacionadas con la industria 4.0 en el sector turístico. Este enfoque demuestra una alineación entre la agenda nacional e internacional en términos de la adopción de nuevas tecnologías en el turismo. Para la búsqueda de documentos académicos, se exploraron bases de datos bibliográficas reconocidas, como Web of Science y Scopus, seleccionando finalmente Web of Science debido a la abundancia de documentos relevantes. Esto

condujo a la identificación de 159 artículos para el análisis bibliométrico, donde se destacaron palabras clave mencionadas al menos 5 veces según se muestra en la figura 1.

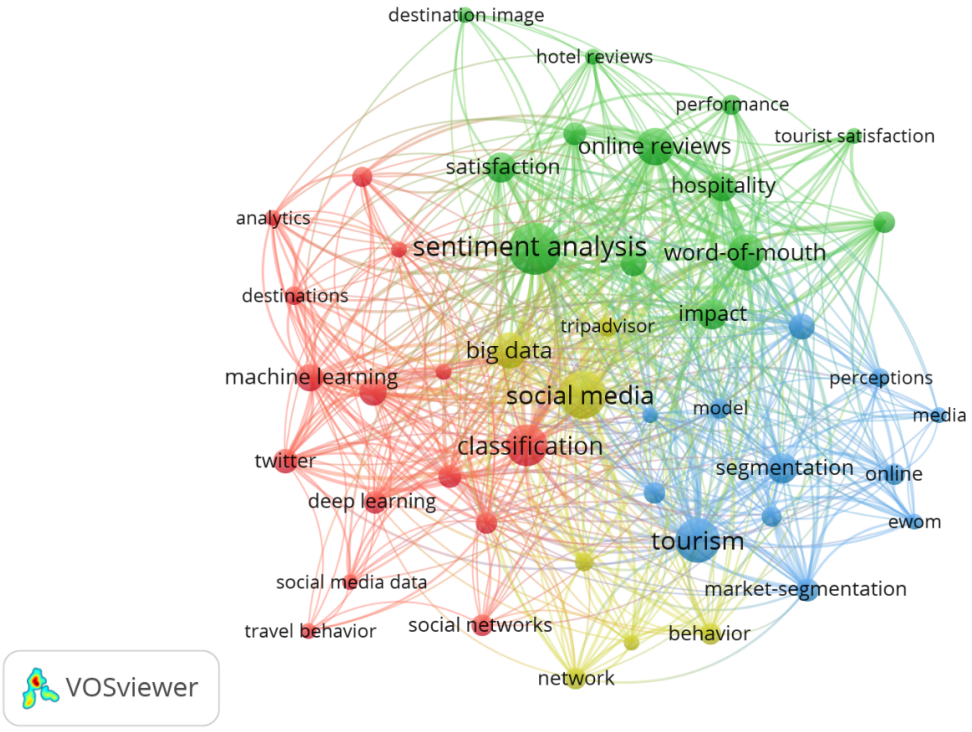
Figura 5.

Ecuación de búsqueda

TS= ((segmentation OR classification OR specialized AND market OR "market segmentation") AND (touris* OR "touris* industry" OR "travel industry" OR "travel sector" OR "international touris*" OR "touris* behavior" OR "travel behavior" OR "touris* segments") AND (sa OR "sentiment analysis" OR social AND networks OR community AND detection))

Figura 6.

Mapa de correlación de palabras clave.



*Nota: figura adaptada del software VosViewer*En la figura presentada, se observa cómo las palabras clave utilizadas en la ecuación de búsqueda se relacionan con otros términos clave

encontrados en los documentos, como machine learning, Deep learning, social media, Big data, TripAdvisor y online reviews. Estas conexiones son resultado de la investigación previa de varios autores, quienes han desarrollado ideas y soluciones en diferentes contextos utilizando estos términos. Además, se establece un criterio de inclusión que requiere un mínimo de 2 documentos por autor y 5 citaciones para realizar un análisis de correlación entre los autores. Se destacan 6 autores en este análisis, quienes se enfocan en el desarrollo de algoritmos, incluyendo redes neuronales, k-means y otros algoritmos de aprendizaje supervisado y no supervisado, con el objetivo de abordar problemáticas específicas del sector turismo. Esto se puede apreciar en la figura 2.

Figura 7.

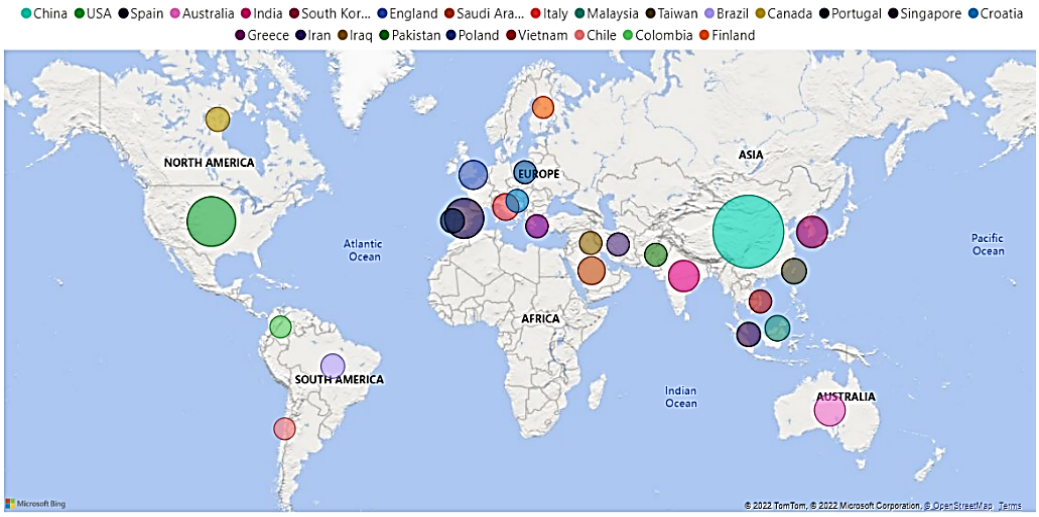
Mapa de correlación de autores

Nota: figura adaptada del software VosViewer

Posteriormente se analizaron los países con mayor aporte científico sobre el tema de análisis de sentimientos al sector turismo, evidenciando la participación de 25 países, donde se observa como lidera con mayor número de publicaciones China con 44, seguido de los Estados Unidos y los demás países tal como se observa en la figura 3, cabe destacar que Colombia se encuentra en el puesto 24 en el mundo y en el puesto tres con respecto a Latinoamérica con dos publicaciones en los años 2019 y 2020.

Figura 8.

Mapa de distribución de las publicaciones en el mundo

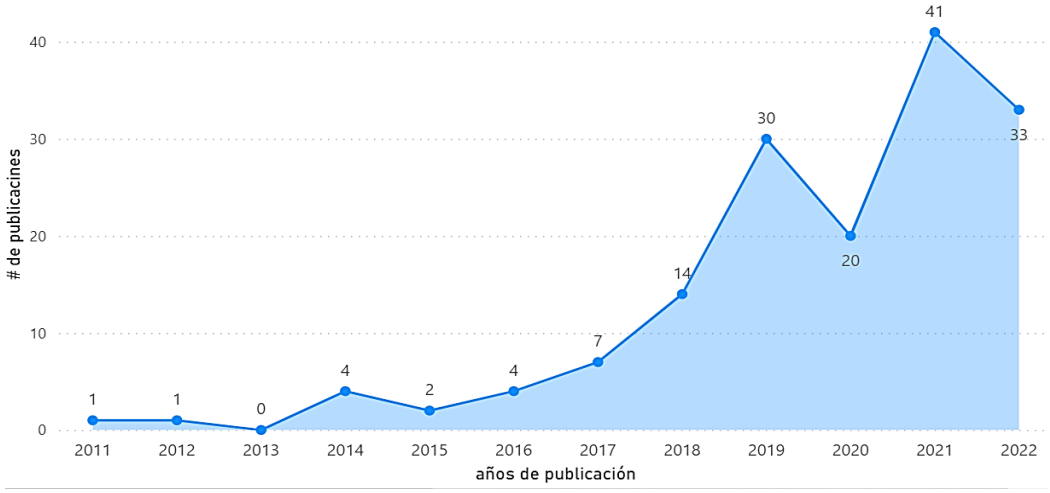


Nota: figura adaptada del software Power BI

Al encontrar que la actividad de Colombia en este tipo de investigaciones es reciente, surge la interrogante de si se ha presentado este comportamiento de manera generalizada a través del tiempo, por tanto, se realizó el análisis del recorrido histórico en el mundo de estas temáticas en donde se evidencia información desde el siglo XXI, sin embargo, se observa que desde el año 2011 la temática ha tomado relevancia como se presenta en la figura 3, además de esto se observa el comportamiento en la cantidad de publicación de documentos, evidenciando que para el año 2021 se presenta un incremento significativo con un total de 41, sin embargo, en el 2020 se dio un cambio en el comportamiento de la curva, este pico decreciente se dio debido a la pandemia del año 2020 que afecto a nivel mundial a todos los sectores pero en especial a turístico (World turism organization, 2021), que directamente afecto la producción científica en este tipo de documentos.

Figura 9.

Histórico de publicaciones



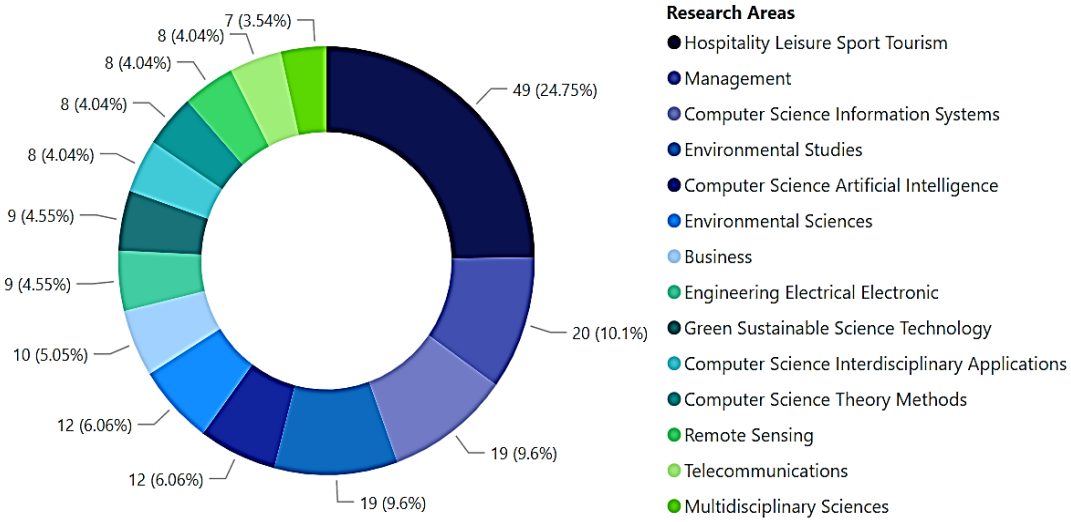
Nota: figura adaptada del software Power BI

Finalmente, en la figura 5 se observa que la distribución de la participación según sus áreas de investigación sobre el análisis de redes sociales y turismo la lideran las ciencias sociales y otros temas con un 24,56%, seguido de la economía empresarial y la ciencia computacional en un 18,71% y 14,04% respectivamente.

Así mismo, se encuentran áreas como ciencia ecológica, ciencia de la tecnología, ingeniería transporte, sensores remotos, la ciencia de la información y la geografía que se encuentra como la de menor participación en estos temas de análisis.

Figura 10.

Participación según su área de investigación



Nota: figura adaptada del software Power BI

5.2. Análisis Preliminar de literatura

El enfoque de investigación y desarrollo dado por el sector turístico se ha caracterizado en los últimos años por ser innovador, tecnológico y sostenible, debido al crecimiento económico y a la llegada de la "industria 4.0", lo que ha permitido conseguir una conexión más profunda con los deseos de los consumidores. Gracias a la producción académica generada desde la década de los noventa, el sector económico del turismo se ha beneficiado del uso de nuevos algoritmos de inteligencia artificial, aprendizaje automático, minería de datos, etc. Dando paso a herramientas que posibilitan modelar las respuestas o comportamientos de los consumidores ante los estímulos de avisos publicitarios (Curry & Moutinho, 1993), evaluar de forma integral las reseñas y cuantificar del impacto de la reputación en línea de los negocios turísticos analizados por (Phillips et al., 2015), predecir la decisión del consumidor (West et al., 1997), desarrollar un

nuevo producto (R. Jeffrey Thieme, 2000), predecir el precio de mercado de las acciones (Haider Khan et al., 2011) y segmentar el mercado (Boone & Roehm, 2002).

Si bien el aprendizaje automático surgió en la década de los cincuenta, continua vigente. Constituyendo el reto de conseguir que una maquina ejecute acciones a partir de información sistemáticamente adquirida, similar al aprendizaje humano, ha tomado relevancia en los últimos años, como es posible observar en el trabajo de (Kirilenko et al., 2018) en el que se comparan los enfoques y metodologías de aprendizaje automático para el análisis de sentimientos en el turismo. Sin embargo, es frecuente encontrar metodologías de aprendizaje profundo, debido a que se determinan por su alto rendimiento en la identificación de patrones, como el reconocimiento facial para la caracterización del perfil de turismo a raíz de redes sociales (Vilar González, 2022), el reconocimiento visual y de voz para mejorar la experiencia de atención al cliente(Tony Garry Assoc Sergio Biggemann, 2021), la clasificación de viajeros según su comportamiento (Cui et al., 2018), la clasificación de reseñas turísticas según los sentimientos (Li et al., 2018a), predicción de precios del mercado de acciones según indicadores financieros (Haider Khan et al., 2011) y predicción del estado de financiamiento de proyectos de tecnología (Puente Ríos, 2021).

Una de las cosas que tienen en común el aprendizaje automático y el aprendizaje profundo es que ambos se pueden utilizar para hacer minería de datos, el conjunto de técnicas más utilizadas en la actualidad debido a que se enfoca en el análisis de datos no estructurados, que representan más del 80% de los datos empresariales. Además, el 95% de las empresas da prioridad al manejo de los datos no estructurados (IBM, 2021). La minería de datos en los años setenta era conocida como pesca de datos o arqueología de datos (Amisaday Huerta Zamora, 2016), y actualmente se conocen ramas de la lingüística computacional como la minería de textos y la minería de sentimientos.

La minería de textos tiene como objetivo la búsqueda de conocimiento en grandes cantidades de texto y tiene aplicaciones tales como el análisis del valor turístico sostenible del grafiti tour en Bogotá a través de las reseñas de TripAdvisor (Seok et al., 2020) o el análisis de redes sociales mediante análisis de reseñas en línea de hoteles en China (He et al., 2017). En estos casos, intervienen tanto la minería de texto como la minería de sentimientos, ya que la minería de sentimientos o minería de opiniones se utiliza para conocer la información subjetiva de las reseñas, como en la clasificación de sentimientos de reseñas de turismo en hoteles chinos usando una red neuronal recurrente bidireccional con un mecanismo de atención y enriquecida por temas (Li et al., 2018a), el desarrollo de una arquitectura de Big Data para evaluar la satisfacción de los consumidores en el sector turismo de Boyacá, Colombia (Algecira Arbelaez, 2020) y la aproximación a las preferencias de los turistas en Guadua, Colombia, mediante el análisis de sentimientos (Morales Beltrán, 2020).

En cuanto al análisis de sentimientos (SA), este se remonta a los 70’s pero en los años recientes se le ha dado mayor relevancia en todos los ámbitos según (Brob, 2013)(Pang et al., 2002), los cuales exponen que se debe al aumento de la información en internet, al avance de la nuevas tecnologías y el desarrollo de nuevas aplicaciones que utilicen esta información para identificar patrones, y comprender como se lleva a cabo la toma de decisiones en grupos de población o de mercado como en el trabajo de (Ribeiro et al., 2016), para esto se dispone de la lingüística computacional, el análisis de texto y el procesamiento de lenguaje natural (PNL), herramientas que contribuyen al desarrollo del análisis de sentimientos, que se basa en dos ideas: clasificación de texto léxico y no léxico que se pueden desarrollar para la detección de la intensidad del sentimiento de textos (Thelwall et al., 2010), ya que el texto léxico parte de un conjunto de palabras a las cuales se les asigna un sentimiento típico (positivo, negativo o neutral)

mientras que, el enfoque no léxico se basa en el aprendizaje automático en donde un algoritmo de clasificación se entrena en una estructura de texto temáticamente cercano como sucede en la investigación de (Kirilenko et al., 2018). así mismo, el SA dispone de tres técnicas generales: el análisis de contenido se usa para identificar y cuantificar las características de contenido del texto no estructurado como es el caso del análisis de quejas de un hotel revisando reseñas de una estrella realizado por (Levy et al., 2013), pero el análisis de sentimiento basado en documentos se utiliza para definir la polaridad (positiva o negativa) a nivel de documento, utilizando algoritmos de clasificación como máquina de vectores (siglas en inglés SVM) y Naive bayes para la clasificación se sentimientos como la que hicieron (Zhang et al., 2011) con base a las reseñas de un restaurante cantonés. Al igual que los análisis anteriores, el análisis basado en temas se refiere a la clasificación de sentimientos, pero esta se realiza una vez que se han extraído los temas, como lo hace (Ren & Hong, 2017) en su análisis de imágenes de destinos en línea.

Cuando se habla de un análisis de sentimientos, se refiere a que se realizará una clasificación de polaridad. La clasificación de polaridad se basa en una clasificación binaria, es decir, de dos clases (p. ej., positivo o negativo). Según los resultados de la Figura 3, la mayoría de las investigaciones sobre análisis de sentimientos y su aplicación en el turismo utilizaron la clasificación de polaridad binaria (Bjørkelund et al., 2012ª) (Gindl et al., 2010) (Kang et al., 2012) (Shimada et al., 2011).

El uso de análisis de documentos no se debería descartar del todo, ya que hay herramientas complementarias como el mecanismo de atención que, alineado con el modelo Bigrula, como en el caso de (Li et al., 2018b), puede resolver las falencias de un análisis de documento común. De esta manera, se eligen palabras relevantes en el documento por medio del modelo Bigrula y, luego, con el mecanismo de atención se le da un peso de importancia a cada

palabra según el contexto y la aplicación. De esta manera, se construye el vector del documento a partir de una secuencia de vectores de palabras para, finalmente, realizar la clasificación de documentos, calculando la probabilidad de que el documento pertenezca a una de las categorías que se plantean.

Por otro lado, (Aggarwal & Gour, 2020) en su análisis web para conocer la mente de los turistas, utilizan tres enfoques matemáticos. El primero es un enfoque geométrico que se puede ver representado en el algoritmo Support Vector Machine (SVM), que busca encontrar la mejor distancia lineal entre los datos que contienen sentimientos positivos y negativos. Para el caso de Naïve Bayes, manifiesta un enfoque probabilístico mediante la estimación de la probabilidad del sentimiento dado el texto. Por último, está el enfoque neuronal de las redes neuronales artificiales (ANN) y los algoritmos derivados de Deep Learning. Al imitar el funcionamiento neuronal de un cerebro, se cree que arroja mejores resultados, pero eso no limita la popularidad de SVM y Naïve Bayes que son conocidos debido a su rápido rendimiento.

5.2.1. Redes sociales y su impacto en la toma de decisiones

Las redes sociales se han convertido en una de las mayores bases de datos debido a que permiten conectar y crear lazos sociales con gran rapidez y facilidad, sin importar las habilidades sociales o la distancia entre el receptor y el emisor del mensaje. Según lo describe (Larkin & Fink, 2016) en su análisis del comportamiento de los aficionados a los deportes de fantasía, la relación con la tecnología puede transformarse en una motivación persistente para generar, consumir y compartir contenido en todo momento a través de redes como Twitter, Instagram y Facebook, o mediante plataformas profesionales analizadas por (Bjørkelund et al.,

2012b)(Rabanser & Ricci, 2005) que se relacionan con los viajes y el turismo, como TripAdvisor, Expedia, VirtualTourist y LonelyPlanet. Estas plataformas cuentan con una gran cantidad de datos; por ejemplo, TripAdvisor cuenta con 350 millones de visitantes únicos por mes en su sitio web y genera más de 320 millones de reseñas que cubren alojamiento, restaurantes y atracciones (TripAdvisor, 2016), lo cual indica que son una gran base de datos disponible para el entrenamiento preciso de algoritmos de aprendizaje automático.

Adicionalmente, (Alaei et al., 2019)(Xiang & Gretzel, 2010) indican que la información adquirida a través de redes sociales o plataformas que permiten y generan comentarios, opiniones y experiencias de los individuos es más confiable que la de los sitios web de empresas o instituciones. Sin embargo, no solo las personas comparten opiniones en las redes sociales, sino que también lo hacen los comerciantes, las agencias (Wang, 2014)(Zheng et al., 2016), las empresas (Culnan et al., 2010) y hasta los políticos (Broersma & Graham, 2012), creando y compartiendo contenido que influye en las opiniones de todos aquellos que conforman las comunidades.

5.2.2. Detección de comunidades

El problema de identificar comunidades, redes sociales o clasificar grupos de personas a partir de ciertos parámetros es abordado por la detección de comunidades, que identifica nodos y patrones de asociación para crear grupos, categorías y clasificaciones que puedan utilizarse en aplicaciones tales como el análisis del flujo de estudiantes entre grados dentro del sistema de acceso a la universidad pública en España (Carmona et al., 2022), el análisis de la estructura actual del sistema de transporte aéreo estadounidense (Carmona et al., 2022), el análisis de la movilidad de los turistas urbanos basado en la actividad de los usuarios en redes sociales (CLAVICILLAS CABALTERA, 2021) y un estudio sobre las comunidades digitales de los

ayuntamientos españoles (Criado et al., 2018), entre otros. En cada una de estas investigaciones, se utilizan algoritmos basados en un cluster jerárquico, como el algoritmo Girvan-Newman utilizado en (Carmona et al., 2022), o cluster particional, como los algoritmos Walktrap y convolución (Carmona et al., 2022). El algoritmo k-means es otro de los más utilizados debido a su facilidad y adaptabilidad en la implementación, y se ha utilizado en investigaciones como las de (Aggarwal & Gour, 2020; Alaei et al., 2019; Ince et al., 2013; Victoriano et al., 2020) Por lo tanto, se determina que este algoritmo será utilizado en la presente investigación.

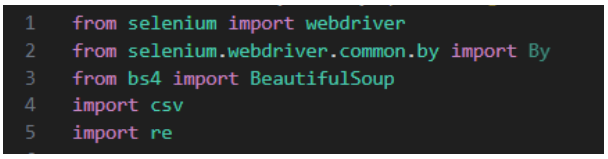
6. Datos

En esta sección, se detalla el proceso llevado a cabo para la recolección y tratamiento de los datos necesarios para la presente investigación. El objetivo fue recopilar los nombres de los hoteles, las reseñas y las fechas de estadía correspondientes al departamento de Santander, Colombia, encontrados en la plataforma TripAdvisor hasta mayo de 2023, en idioma español.

Dado que no se disponía de un registro significativo de datos en más del 85% de los municipios de Santander, teniendo en cuenta que existen 87 municipios en total (Invest in Santander, 2023), se tomó la decisión de seleccionar los 12 principales municipios turísticos del departamento. Estos municipios fueron elegidos siguiendo criterios específicos, como contar con una cantidad considerable de datos, como más de 10 hoteles o más de 100 reseñas en TripAdvisor. Esta selección permitió garantizar la calidad y cantidad de datos necesarios para un manejo eficiente de un modelo Transformer. Es importante destacar que los modelos Transformer se benefician de conjuntos de datos grandes, diversos y representativos, lo cual favorece su capacidad de procesamiento y generalización.

Figura 11.

Extracción de datos: importación de librerías



Nota: figura adaptada de Visual Studio Code

La elección de los municipios se llevó a cabo mediante una consulta previa en la plataforma de TripAdvisor, donde se verificó la disponibilidad y cantidad de hoteles y reseñas. Posteriormente, se procedió a la extracción de los datos utilizando herramientas de programación en Python. Se emplearon bibliotecas como BeautifulSoup y Selenium, junto con el controlador web de Selenium, para automatizar la extracción de información de TripAdvisor, simulando la interacción estándar del usuario en la plataforma. Es importante tener en cuenta que cada página en TripAdvisor puede mostrar hasta 30 hoteles. Por lo tanto, cuando se recolecta información de más de 30 hoteles, se debe realizar un cambio en el URL correspondiente a la línea 7 en el código que se muestra en la figura 10 que por efectos ilustrativos se tiene el caso de la extracción de datos del municipio de Bucaramanga para pasar a la siguiente página o para cambiar el municipio de análisis.

Figura 12.

Apertura y visualización del navegador

```
7   url = "https://www.tripadvisor.com/
    Hotels-g297474-oa-Bucaramanga_Santander_Department-Hotels.html"
8
9   # Configurar Selenium para utilizar el controlador adecuado (por ejemplo,
    ChromeDriver)
10  driver = webdriver.Chrome()  # Asegúrate de tener instalado el controlador
    de Chrome adecuado y en tu PATH
11
12  # Abrir la página web con Selenium
13  driver.get(url)
14
15  # Esperar a que la página se cargue completamente (puedes ajustar el
    tiempo de espera según sea necesario)
16  driver.implicitly_wait(10)
17
18  # Obtener el contenido HTML de la página actual
19  page_source = driver.page_source
20
21  # Cerrar el navegador controlado por Selenium
22  driver.quit()
23
24  # Analizar el contenido HTML de la página
25  soup = BeautifulSoup(page_source, 'html.parser')
26
27  # Extraer información de la página
28  hotel_names = soup.find_all('a', class_='property_title')
29
30  # Crear una lista para almacenar las reseñas
31  reviews = []
```

Nota: figura adaptada de Visual Studio Code

Además, en el proceso de extracción de datos se emplearon diferentes estructuras de control, como ciclos ´if´, ´for´ y ´while´, con el fin de iterar según las condiciones y requerimientos necesarios para obtener todos los datos deseados de manera completa.

El ciclo ´for´ se utilizó en dos instancias. En primer lugar, se utilizó para iterar sobre los nombres de los hoteles y extraer las reseñas correspondientes. Para cada hotel, se obtuvo su nombre y se construyó la URL de la página de reseñas correspondiente. Esto se puede apreciar en las líneas 34 a 36 en la figura 11 del código. En segunda instancia, se puede observar en la línea 65 del código en la figura 13, donde se itera sobre las reseñas de los hoteles con el propósito de agregar cada reseña extraída a la lista de reseñas.

Figura 13.

Estructura Condicional: for

```
33  # Iterar sobre los nombres de los hoteles
34  for hotel in hotel_names:
35      hotel_name = hotel.text.strip()
36      hotel_url = "https://www.tripadvisor.co" + re.search(r'/Hotel_Review-(.
        *?)-Reviews', hotel['href']).group(0) +
        "-Bucaramanga_Santander_Department.html#REVIEWS"
37
38      # Configurar Selenium nuevamente para abrir la página del hotel
39      driver = webdriver.Chrome()  # Utiliza el mismo controlador que antes
```

Nota: figura adaptada de Visual Studio Code

El siguiente ciclo utilizado fue el ´while´, el cual se utilizó para recorrer las páginas de reseñas de cada hotel. Se establecieron variables para el número de página y el límite de páginas a extraer antes del ciclo ´while´, como se muestra en la figura 12 del código. Además, se tuvo en cuenta la verificación del estado del botón "siguiente" para controlar la ejecución del bucle ´while´.

Figura 14.

Estructura Condicional: while

```
41      # Variables para el bucle de paginación
42      page_num = 0
43      page_limit = 272  # Límite de páginas de reseñas a extraer
44      next_button_disabled = False
45
46      while page_num < page_limit and not next_button_disabled:
47          # Calcular el valor "orX" para la paginación
48          page_offset = page_num * 5
49          page_url = hotel_url.replace('-Reviews', '-or{}-Reviews'.format(page_offset))
```

Nota: figura adaptada de Visual Studio Code

Por último, el ciclo if se utilizó para encontrar, extraer y guardar la fecha de estadía si estaba disponible en cada reseña. En caso de no encontrar la fecha, se omitió el paso de agregar esos datos a la lista correspondiente. Esto se puede observar en las líneas 73-77 en la figura 13 del código.

El uso de estas estructuras de control permitió llevar a cabo el proceso de extracción de datos de manera eficiente y precisa, garantizando la obtención de la información necesaria para su posterior análisis y almacenamiento.

Figura 15.

Estructura Condicional: for e if

```
61        # Extraer las reseñas del hotel
62        hotel_reviews = hotel_soup.find_all('div', class_='WAllg')
63
64        # Agregar las reseñas a la lista
65        for review in hotel_reviews:
66            review_title = review.find('div', class_='KgQgP').text.strip()
67            review_text = review.find('div', class_='fIrGe').text.strip()
68
69            # Buscar el elemento de fecha de estadía
70            date_element = review.find('span', class_='teHYY _R Me S4 H3')
71            date_text = ''
72
73            if date_element:
74                # Extraer el texto de la fecha de estadía
75                date_text = date_element.text.strip().replace('Fecha de la estadía:', '')
76
77            reviews.append([hotel_name, review_title, review_text, date_text])
78
79        # Verificar si el botón "Siguiente" está deshabilitado
80        next_button = hotel_soup.find('span', class_='ui_button nav next primary disabled')
81        next_button_disabled = next_button is not None
82
83        page_num += 1
```

Nota: figura adaptada de Visual Studio Code

Una vez finalizada la recolección de datos para cada municipio seleccionado, se procedió a guardar la información en archivos CSV con el nombre correspondiente a cada municipio como se registra en la figura 14. Además, se tuvo especial cuidado en cumplir con las políticas de protección de datos establecidas por TripAdvisor y otras plataformas similares. Se revisaron las reglas y condiciones de uso para evitar exceder los límites de solicitudes por servidor o dirección IP.

Figura 16.

Guardar en archivo csv

```
    # Cerrar el navegador controlado por Selenium
    driver.quit()

# Ruta del archivo CSV
csv_file_path = "Bucaramanga3.csv"

# Guardar las reseñas en un archivo CSV
with open(csv_file_path, "w", newline="", encoding="utf-8") as file:
    writer = csv.writer(file)
    writer.writerow(["Hotel Name", "Review Title", "Review Text", "Stay Date"])
    writer.writerows(reviews)

print("Las reseñas se han guardado correctamente en 'Bucaramanga.csv'.")
```

Nota: figura adaptada de Visual Studio Code

Durante el desarrollo del programa, se enfrentaron desafíos relacionados con la automatización del cambio del enlace de la página y el máximo número de páginas. Tras múltiples intentos, se determinó que la forma más efectiva era realizar este cambio manualmente en lugar de automatizarlo, debido a los problemas que surgían al intentar una automatización completa.

A pesar de estos desafíos, se perseveró en el proceso de extracción de datos y finalmente se logró obtener un conjunto de datos. Este conjunto abarca un total de 11 municipios, 406 hoteles y 11,680 reseñas en español. Estos datos fueron recolectados a lo largo de un período de 17 años, abarcando desde el año 2006 hasta mayo de 2023.

Este conjunto de datos proporciona una amplia perspectiva temporal y geográfica, lo que permite realizar análisis detallados y comparativos en el ámbito turístico de los municipios seleccionados. La información recopilada incluye opiniones y experiencias de los usuarios, reflejando la evolución y la calidad de los servicios turísticos a lo largo del tiempo.

La disponibilidad de esta cantidad significativa de datos facilita el desarrollo de modelos y algoritmos de procesamiento de lenguaje natural, como los modelos de Transformer, con el objetivo de analizar y extraer información relevante de las reseñas de los hoteles. Estos datos representan una valiosa fuente de información para comprender las preferencias de los turistas, identificar tendencias y patrones, así como para mejorar la toma de decisiones en el sector turístico.

Cabe destacar que el rango de tiempo de 17 años proporciona una visión amplia y permite analizar cambios a lo largo del tiempo, así como identificar posibles variaciones estacionales o tendencias de largo plazo en la industria turística de la región. La cantidad de hoteles y reseñas recopiladas garantiza una muestra representativa y confiable para llevar a cabo análisis estadísticos y modelado predictivo.

7. Análisis de sentimientos

Se realizó el análisis de sentimientos utilizando dos metodologías distintas. En la primera metodología, se empleó la lógica de polaridad de la biblioteca TextBlob de Python. Esta biblioteca permite determinar si un texto tiene una polaridad positiva, negativa o neutra, lo que brinda una visión general del sentimiento expresado en el texto.

Por otro lado, en la segunda metodología se utilizó BERT, una poderosa arquitectura de modelos de lenguaje basada en transformers. Con BERT, se logró una identificación y etiquetado más preciso de la emoción que predomina en el texto. El modelo BERT asigna probabilidades a diferentes emociones y categorías, permitiendo una mayor granularidad en la comprensión de los sentimientos expresados en el texto.

Ambas metodologías aportan enfoques complementarios para el análisis de sentimientos. Mientras que TextBlob proporciona una clasificación general de polaridad, BERT ofrece una clasificación más detallada y precisa de las emociones predominantes en el texto. Esto permite una comprensión más profunda de los sentimientos y emociones presentes en el contenido analizado.

Es importante tener en cuenta que la elección de la metodología depende del contexto y los objetivos del análisis de sentimientos. TextBlob puede ser adecuado para análisis rápidos y sencillos, mientras que BERT se destaca en casos donde se requiere una mayor precisión y una comprensión más fina de las emociones expresadas en el texto.

En esta investigación, es interesante contrastar los resultados de ambas metodologías y validarlos con datos pre-etiquetados de TripAdvisor. Esto enriquecerá el análisis, corroborando la polaridad y las emociones identificadas en el texto analizado. La combinación de enfoques y fuentes de información proporcionará una comprensión más completa del sentimiento expresado.

7.1. Análisis de sentimiento basado en polaridad

Una vez se tienen los datos almacenados en un solo documento CSV se toma la dirección de almacenamiento del documento y se coloca en el código de manera que se conozca de donde se extrae la información a analizar por la Liberia textblob de Python. A continuación, se detalla el proceso paso a paso:

1. **Importación de bibliotecas:** Se importan las bibliotecas necesarias, como csv para trabajar con archivos CSV, y la clase TextBlob de la biblioteca TextBlob para realizar el análisis de sentimientos.

Figura 17.

Código de análisis basado en polaridad: importación de librerías

```
import csv
from textblob import TextBlob
```

Nota: figura adaptada de Visual Studio Code

2. **Ruta del archivo CSV y creación de una lista para almacenar los resultados:** Se especifica la ubicación del archivo CSV que contiene los datos de entrada y se crea una lista vacía llamada sentiment_results, que se utilizará para almacenar los resultados del análisis de sentimientos.

Figura 18.

Código de análisis basado en polaridad: acceso a la base de datos

```
# Ruta del archivo CSV
csv_file_path = "C:\\Users\\PC\\Documents\\Python\\Bases de datos\\Santander.csv"

# Crear una lista para almacenar los resultados del análisis de sentimientos
sentiment_results = []
```

Nota: figura adaptada de Visual Studio Code

3. **Definición de etiquetas de polaridad:** se define un diccionario llamado sentiment_labels que contiene etiquetas asociadas a diferentes rangos de polaridad. Estas etiquetas representan los diferentes sentimientos o emociones expresados en el texto.

Figura 19.

Código de análisis basado en polaridad: etiquetas de polaridad

```
# Etiquetas para los diferentes rangos de polaridad
sentiment_labels = {
    'muy positivo': (0.6, 1.0),
    'positivo': (0.2, 0.6),
    'neutral': (-0.2, 0.2),
    'negativo': (-0.6, -0.2),
    'muy negativo': (-1.0, -0.6)
}
```

Nota: figura adaptada de Visual Studio Code

4. **Función para asignar etiquetas:** Se define una función llamada assign_label que asigna una etiqueta a la polaridad de acuerdo con los rangos definidos en el diccionario sentiment_labels.

Figura 20.

Código de análisis basado en polaridad: función para asignar etiquetas

```
# Función para asignar la etiqueta correspondiente a la polaridad
def assign_label(polarity):
    for label, (lower, upper) in sentiment_labels.items():
        if lower <= polarity < upper:
            return label
    return 'unknown'
```

Nota: figura adaptada de Visual Studio Code

5. **Análisis de sentimientos utilizando TextBlob:** Durante el proceso de análisis de sentimientos, se lleva a cabo la apertura del archivo CSV especificado en la figura 19, donde se itera sobre cada fila utilizando un lector de CSV. En cada iteración, se extrae el texto de la reseña del campo "Review Text". A continuación, se procede al análisis de sentimientos utilizando la librería TextBlob. Se crea un objeto TextBlob con el texto de la reseña y se utiliza el método sentiment.polarity para obtener la polaridad

del sentimiento, representada por un valor numérico que indica la positividad o negatividad. Posteriormente, se realiza la asignación de etiquetas llamando a la función assign_label, que asigna la etiqueta correspondiente a la polaridad obtenida del análisis de sentimientos.

Figura 21.

Código de análisis basado en polaridad: análisis de sentimientos

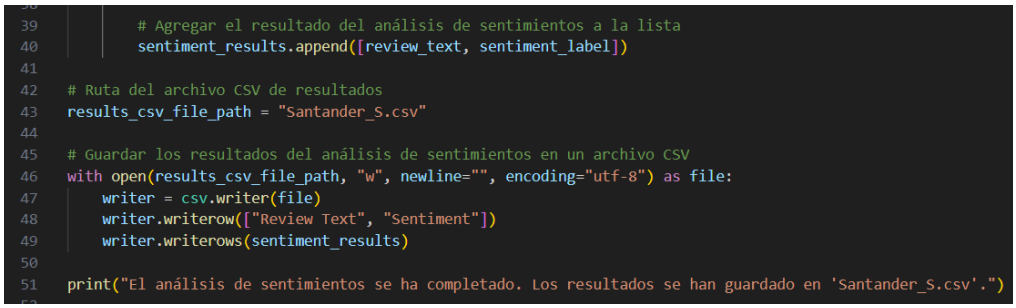
```
# Abrir el archivo CSV con el conjunto de caracteres "latin-1"
with open(csv_file_path, "r", encoding="latin-1") as file:
    reader = csv.DictReader(file)
    for row in reader:
        review_text = row['Review Text']

        # Realizar el análisis de sentimientos con TextBlob
        blob = TextBlob(review_text)
        sentiment = blob.sentiment.polarity

        # Asignar la etiqueta correspondiente a la polaridad
        sentiment_label = assign_label(sentiment)
```

Nota: figura adaptada de Visual Studio Code

6. **Almacenamiento y guardado de resultados:** El texto de la reseña y la etiqueta del sentimiento se agregan como una nueva fila a la lista sentiment_results. A continuación, se especifica la ruta del archivo CSV de resultados y se guarda la lista en dicho archivo. Cada fila del archivo contiene el texto de la reseña y la etiqueta del sentimiento correspondiente. Finalización del análisis: Se imprime un mensaje informando que el análisis de sentimientos ha sido completado exitosamente y que los resultados se han guardado en el archivo "Santander_S.csv".

Figura 22.

Código de análisis basado en polaridad: almacenamiento de resultados



```
        # Agregar el resultado del análisis de sentimientos a la lista
        sentiment_results.append([review_text, sentiment_label])

# Ruta del archivo CSV de resultados
results_csv_file_path = "Santander_S.csv"

# Guardar los resultados del análisis de sentimientos en un archivo CSV
with open(results_csv_file_path, "w", newline="", encoding="utf-8") as file:
    writer = csv.writer(file)
    writer.writerow(["Review Text", "Sentiment"])
    writer.writerows(sentiment_results)

print("El análisis de sentimientos se ha completado. Los resultados se han guardado en 'Santander_S.csv'.")
```

Nota: figura adaptada de Visual Studio Code

7.2. Análisis de sentimiento basado en BERT

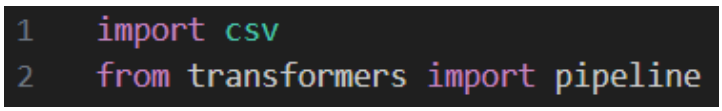
BERT ha demostrado un rendimiento sobresaliente en una amplia gama de tareas de procesamiento del lenguaje natural, como el análisis de sentimientos, la clasificación de texto y la respuesta a preguntas. Su capacidad para entender el contexto y las relaciones entre las palabras lo convierte en una herramienta poderosa para la comprensión y generación de texto de alta calidad.

A continuación, se describen los pasos clave del código que muestra cómo utilizar BERT para el análisis de sentimientos en un conjunto de datos contenido en un archivo CSV:

1. **Importación de las bibliotecas necesarias:** Se importa la biblioteca csv para trabajar con archivos CSV y la clase pipeline de la biblioteca Transformers para cargar el clasificador de sentimientos basado en BERT.

   Figura 23.

   Código de análisis basado en BERT: Importación de librerías



```
import csv
from transformers import pipeline
```

Nota: figura adaptada de Visual Studio Code

2. **Especificación de las rutas de los archivos CSV y creación de una lista:** Se definen las rutas del archivo CSV de entrada que contiene los datos a analizar y el archivo CSV de salida donde se guardarán los resultados del análisis de sentimientos, y se crea una lista vacía llamada sentiment_results, que se utilizará para almacenar los resultados del análisis de sentimientos.

Figura 24.

Código de análisis basado en BERT: acceso a la base de datos

```
4   # Ruta del archivo CSV de entrada
5   input_csv_file_path = "C:\\Users\\PC\\Documents\\Python\\Bases de datos\\Santander.csv"
6
7   # Ruta del archivo CSV de salida para los resultados
8   output_csv_file_path = "Santander_E.csv"
9
10  # Crear una lista para almacenar los resultados del análisis de sentimientos
11  sentiment_results = []
```

Nota: figura adaptada de Visual Studio Code

3. **Carga del clasificador de sentimientos basado en BERT:** Se utiliza el método pipeline de la biblioteca Transformers para cargar el clasificador de sentimientos basado en BERT. Este modelo ha sido entrenado previamente en una gran cantidad de datos para comprender el contexto y las emociones en el texto.

Figura 25.

Código de análisis basado en BERT: cargar el modelo BERT

```
13  # Cargar el clasificador de sentimientos
14  classifier = pipeline("text-classification", model='bhadresh-savani/
    albert-base-v2-emotion', return_all_scores=True)
```

Nota: figura adaptada de Visual Studio Code

4. **Análisis de Sentimientos:** Se abre el archivo CSV de entrada y se itera sobre cada fila utilizando un lector de CSV. Para cada fila, se extrae el texto que se desea analizar y

se realiza la predicción de sentimientos utilizando el clasificador de sentimientos basado en BERT. El resultado de la predicción, que representa la emoción detectada en el texto, se agrega a la lista sentiment_results para cada texto analizado.

Figura 26.

Código de análisis basado en BERT: análisis de sentimientos

```
# Leer el archivo CSV de entrada y realizar el análisis de sentimientos para cada texto en la columna 'Review_text'
with open(input_csv_file_path, "r", encoding="latin-1") as file:
    reader = csv.DictReader(file)
    for row in reader:
        if 'Review Text' in row:
            review_text = row['Review Text']

            # Realizar la predicción de sentimientos utilizando el clasificador de sentimientos
            prediction = classifier(review_text)
            sentiment_results.append(prediction[0])
```

Nota: figura adaptada de Visual Studio Code

5. **Guardado de los resultados en un archivo CSV de salida:** Se abre el archivo CSV de salida y se utiliza un escritor de CSV para escribir los resultados del análisis de sentimientos. Cada fila del archivo contiene el texto analizado y la etiqueta de la emoción detectada. Finalmente, se imprime un mensaje indicando que el análisis de sentimientos ha sido completado y que los resultados se han guardado en el archivo especificado.

Figura 27.

Código de análisis basado en BERT: almacenamiento de resultados

```
# Guardar los resultados del análisis de sentimientos en un archivo CSV
with open(output_csv_file_path, "w", newline="", encoding="utf-8") as file:
    writer = csv.DictWriter(file, fieldnames=["label", "score"])
    writer.writeheader()
    writer.writerows(sentiment_results)

print("El análisis de sentimientos se ha completado. Los resultados se han guardado en 'Santander_E.csv'.")
```

Nota: figura adaptada de Visual Studio Code

8. Clasificación de clientes

En esta sección, se presenta una descripción detallada del proceso llevado a cabo en el código, cuyo objetivo es realizar una segmentación de clientes con base a las reseñas de hoteles en Santander obtenidas de la plataforma TripAdvisor. El código utiliza el modelo DAMO-NLP-SG/zero-shot-classify-SSTuning-base, el cual ha sido entrenado en tareas de clasificación de sentimientos y posee una capacidad de comprensión contextual avanzada.

En este código, se utilizan etiquetas específicas para segmentar las reseñas y comprender mejor las preferencias y experiencias de los clientes. La elección de estas etiquetas, como "limpieza", "servicio", "comida", "actividades", "ubicación", "diseño y decoración" y "mantenimiento", se basa en aspectos clave que influyen en la satisfacción y la calidad de la estancia en un hotel. Al analizar las reseñas en función de estas etiquetas, es posible identificar áreas de mejora y fortalezas específicas de cada hotel, lo que facilita la toma de decisiones estratégicas para brindar un servicio excepcional y adaptado a las necesidades de los clientes.

A través de un proceso completo que abarca desde la lectura del archivo CSV de entrada hasta el almacenamiento de los resultados segmentados en un nuevo archivo CSV, este enfoque permite obtener una visión detallada y estructurada de las opiniones de los clientes en relación con los aspectos clave de su experiencia en los hoteles de Santander. Al comprender mejor las preferencias y necesidades de los clientes, los hoteles pueden adaptar sus servicios y estrategias de marketing para ofrecer experiencias más personalizadas y satisfactorias. A lo largo de la siguiente sección, se explicarán en detalle cada paso del proceso, brindando una comprensión más completa de cómo se realiza la segmentación de clientes con base a las reseñas de hoteles en Santander:

1. **Importación de librerías y carga del modelo:** El código importa las librerías necesarias, como pandas, transformers y torch. A continuación, se carga el modelo preentrenado DAMO-NLP-SG/zero-shot-classify-SSTuning-base utilizando AutoModelForSequenceClassification y se instala el tokenizer correspondiente utilizando AutoTokenizer.

   Figura 28.

   Código de clasificación: Importación de librerías

```
1  import pandas as pd
2  from transformers import AutoTokenizer, AutoModelForSequenceClassification
3  import torch, string, random
4
5  tokenizer = AutoTokenizer.from_pretrained("DAMO-NLP-SG/
   zero-shot-classify-SSTuning-base")
6  model = AutoModelForSequenceClassification.from_pretrained("DAMO-NLP-SG/
   zero-shot-classify-SSTuning-base")
```

   Nota: figura adaptada de Visual Studio Code

2. **Configuración inicial:** Se definen las etiquetas de categorías de sentimientos en la lista list_label, y se crea una lista alfabética list_ABC utilizando la librería string. Se verifica la disponibilidad de un dispositivo GPU y se asigna a la variable device.

   Figura 29.

   Código de clasificación: configuración inicial

```
8  list_label = ["limpieza", "servicio", "comida", "actividades",
   "ubicación", "diseño y decoración", "mantenimiento"]
9  list_ABC = [x for x in string.ascii_uppercase]
10
11 device = torch.device('cuda') if torch.cuda.is_available() else torch.
   device('cpu')
```

   Nota: figura adaptada de Visual Studio Code

3. **Función check_text:** Se define la función check_text que recibe el modelo, el texto a analizar, la lista de etiquetas y un parámetro opcional shuffle. La función realiza los siguientes pasos:

a. **Preparación del texto:** Se formatea el texto para incluir las opciones de etiquetas y el texto de la reseña, separados por el token especial [SEP].

Figura 30.

Código de clasificación: preparación de texto

```
13  def check_text(model, text, list_label, shuffle=False):
14      list_label = [x+'.' if x[-1] != '.' else x for x in list_label]
15      list_label_new = list_label + [tokenizer.pad_token]* (20 - len
        (list_label))
16      if shuffle:
17          random.shuffle(list_label_new)
18      s_option = ' '.join(['('+list_ABC[i]+') '+list_label_new[i] for i in
        range(len(list_label_new))])
19      text = f'{s_option} {tokenizer.sep_token} {text}'
```

Nota: figura adaptada de Visual Studio Code

b. **Codificación y clasificación:** Se realiza la codificación del texto utilizando el tokenizer y se envía al modelo para obtener los logits, que son las salidas antes de la función de activación softmax. Luego, se aplica la función softmax y se obtienen las probabilidades para cada etiqueta de categoría. Se determina la etiqueta con la mayor probabilidad como la predicción.

Figura 31.

Código de clasificación: codificación y clasificación

```
21      model.to(device).eval()
22      encoding = tokenizer([text], truncation=True, max_length=512,
        return_tensors='pt')
23      item = {key: val.to(device) for key, val in encoding.items()}
24      logits = model(**item).logits
25
26      logits = logits if shuffle else logits[:,0:len(list_label)]
27      probs = torch.nn.functional.softmax(logits, dim=-1).tolist()
28      predictions = torch.argmax(logits, dim=-1).item()
29      probabilities = [round(x, 5) for x in probs[0]]
```

Nota: figura adaptada de Visual Studio Code

c. **Retorno de resultados:** La función devuelve la predicción y las probabilidades asociadas a cada etiqueta de categoría.

Figura 32.

Código de clasificación: resultados

```
31      return predictions, probabilities
```

Nota: figura adaptada de Visual Studio Code

4. **Lectura y análisis de las reseñas:** El código lee el archivo CSV especificado utilizando la librería panda y lo almacena en el DataFrame 'data'. Luego, se itera sobre cada fila del DataFrame utilizando el método iterrows(). Para cada fila, se extrae el texto de la columna 'Review Text'. Se realiza la llamada a la función check_text para obtener la predicción de sentimientos y las probabilidades para el texto de la reseña. La predicción se asigna a la variable 'label' y se agregan el texto de la reseña, la etiqueta y las probabilidades a las listas correspondientes.

Figura 33.

Código de clasificación: lectura y análisis de reseñas

```
33  # Leer el archivo CSV
34  data = pd.read_csv(r'C:\Users\PC\Documents\Python\Bases de datos\Santander.
    csv', encoding='ISO-8859-1')
35
36  # Crear listas para almacenar la información de las reviews
37  reviews = []
38  labels = []
39  probabilities_list = []
40
41  # Analizar la columna 'Review Text' para cada fila del CSV
42  for index, row in data.iterrows():
43      review_text = row['Review Text']
44      print(f'Analizando review {index + 1}...')
45      prediction, probabilities = check_text(model, review_text, list_label)
46      label = list_label[prediction]
47
48      reviews.append(review_text)
49      labels.append(label)
50      probabilities_list.append(probabilities)
```

Nota: figura adaptada de Visual Studio Code

5. **Creación y guardado del DataFrame de resultados:** Se crea un nuevo DataFrame utilizando pandas con las listas de reseñas, etiquetas y probabilidades. Se crean columnas adicionales en el DataFrame para cada etiqueta de categoría, almacenando las probabilidades correspondientes. Finalmente, se guarda el DataFrame en un nuevo archivo CSV utilizando el método to_csv() de pandas.

Figura 34.

Código de clasificación: crear y guardar DataFrame

```
# Crear un DataFrame con la información recolectada
df = pd.DataFrame({'Review Text': reviews, 'Label': labels})
for i, label in enumerate(list_label):
    df[label + ' Probability'] = [probs[i] for probs in probabilities_list]

# Guardar el DataFrame en un nuevo archivo CSV
df.to_csv('Santander_L.csv', index=False)
print('¡Análisis completado y resultados guardados en el archivo CSV!')
```

Nota: figura adaptada de Visual Studio Code

9. Resultados

Una vez finalizados los análisis de sentimientos y la clasificación de las reseñas según sus categorías principales, se procede a consolidar toda la información en un único archivo CSV. A continuación, se realiza una exhaustiva limpieza de los datos para prepararlos de manera óptima para su importación al sistema Power Query. Esta etapa de limpieza implica corregir detalles y realizar ajustes necesarios en la base de datos, garantizando así la calidad y coherencia de los datos para un análisis y visualización posterior más efectivos.

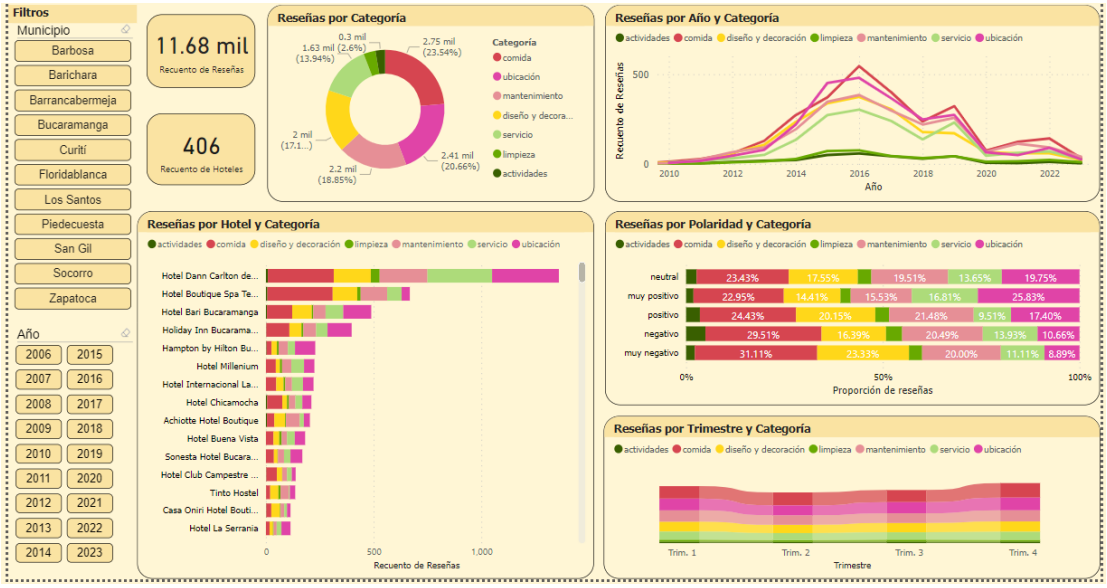
Una vez que los datos han sido debidamente preparados, son importados al sistema Power Query, una herramienta que ofrece la posibilidad de realizar transformaciones y ajustes adicionales de acuerdo con las necesidades específicas del análisis. Esta valiosa herramienta

facilita la manipulación y organización de los datos, otorgando un mayor control sobre su estructura y formato.

Posteriormente, se aprovechan las capacidades de visualización de datos que proporciona Power BI Desktop, para crear un atractivo y eficiente tablero interactivo. Este tablero presenta los resultados obtenidos de manera clara y comprensible, permitiendo a los usuarios explorar y analizar dinámicamente los datos. Esto simplifica la identificación de patrones, tendencias y conclusiones relevantes.

Figura 35.

Análisis categórico de Santander



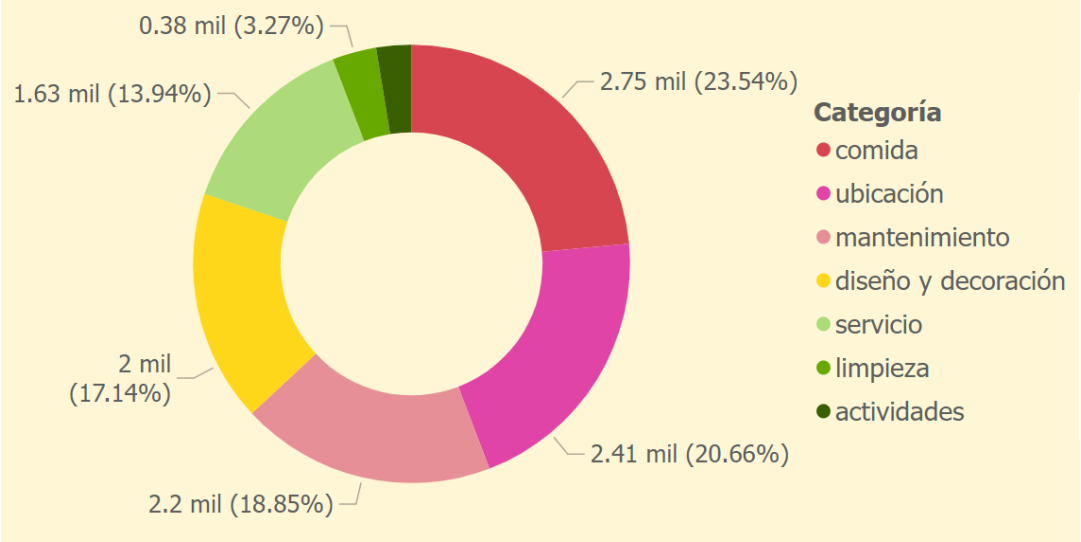
Nota: figura adaptada de Power BI

En la figura 33 se muestra un tablero que describe el análisis categórico basado en la evaluación de variables como el tiempo (anual y trimestral) y la polaridad, con la opción de aplicar filtros por municipio y/o año. El objetivo de este análisis es observar cómo se comportan

las reseñas según sus categorías en diferentes contextos. De manera general, se destaca que las proporciones de reseñas en cada categoría se mantienen consistentes, a pesar de las variaciones en las variables o filtros utilizados. El orden jerárquico de las categorías, de mayor a menor proporción de reseñas, se muestra en la figura 34 de la siguiente manera: comida, ubicación, mantenimiento, diseño y decoración, servicio, limpieza y actividades. Este análisis ofrece una visión clara y estructurada del impacto que las variables evaluadas tienen sobre las categorías de reseñas, permitiendo así comprender mejor las preferencias y percepciones de los usuarios en diferentes períodos y localidades.

Figura 36.

Proporción categórica de las reseñas



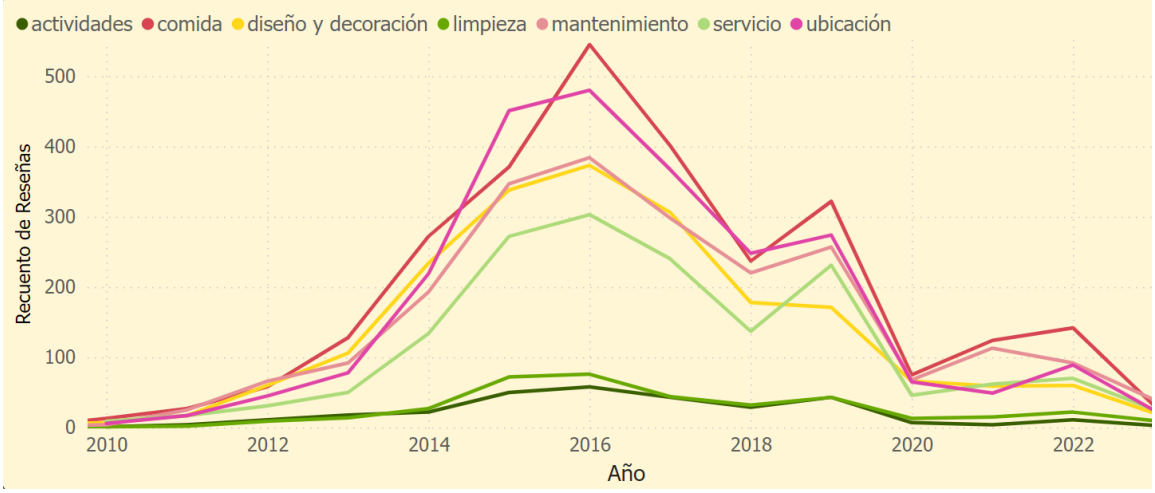
Nota: figura adaptada de Power BI

El análisis de las reseñas de hoteles en el departamento de Santander, publicadas en TripAdvisor desde el primero de noviembre de 2006 hasta el primero de mayo de 2023, proporciona información valiosa para identificar segmentos de clientes y evaluar sus experiencias en diferentes categorías. En la figura 35 se pueden observar picos máximos en el número de

reseñas durante los años 2015, 2016, 2019 y 2022. Estos picos podrían estar relacionados con factores como los períodos de crecimiento del dólar frente al peso colombiano y la percepción de seguridad y estabilidad que surgieron a raíz del proceso de paz firmado en 2016. Además, es importante destacar que el año 2022 presentó un pico máximo en las reseñas, lo cual podría explicarse por la reapertura económica después de la pandemia de COVID-19. La crisis económica mundial y las restricciones de movilidad y contacto físico impuestas durante la pandemia afectaron significativamente al sector turístico. Sin embargo, a medida que las medidas de restricción se relajaron y se implementaron protocolos de seguridad, es posible que más personas hayan optado por viajar y disfrutar de experiencias turísticas en Santander, lo que se reflejó en un aumento en las reseñas durante el año 2022.

Figura 37.

Histórico anual de las reseñas según su categoría



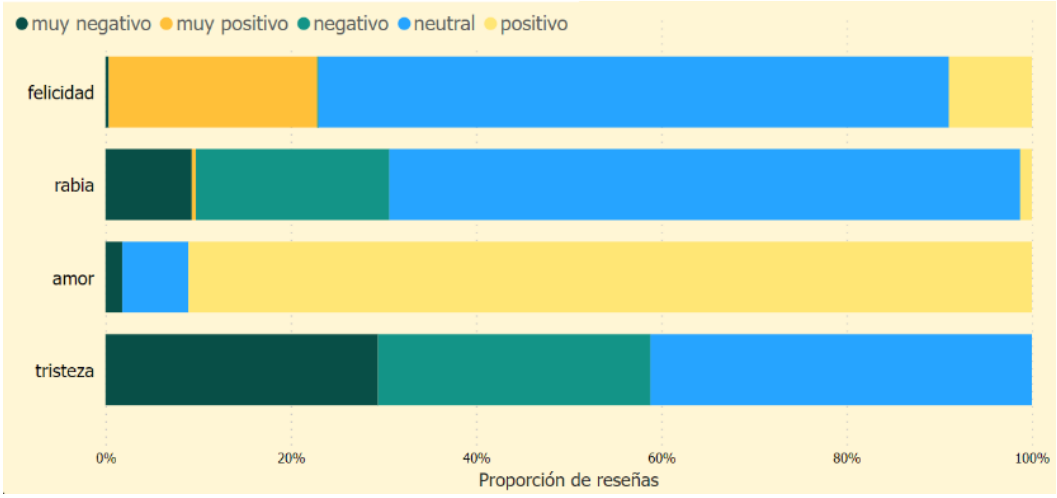
Nota: figura adaptada de Power BI

El análisis de polaridad fue realizado por un único analista, complementando el análisis categórico para obtener una visión más detallada del comportamiento de los clientes según las

reseñas de TripAdvisor. Además, en la figura 36 se incorporó el análisis de sentimientos y polaridad con el propósito de comprobar la hipótesis de que la polaridad positiva o neutra en los comentarios generalmente está asociada con emociones positivas, como felicidad y amor, mientras que la polaridad negativa está relacionada con emociones negativas, como tristeza y rabia. Estas emociones de menor intensidad en los datos fueron respaldadas por las estadísticas de puntuación por estrellas en la plataforma de TripAdvisor.

Figura 38.

Proporción de reseñas según su polaridad y sentimiento



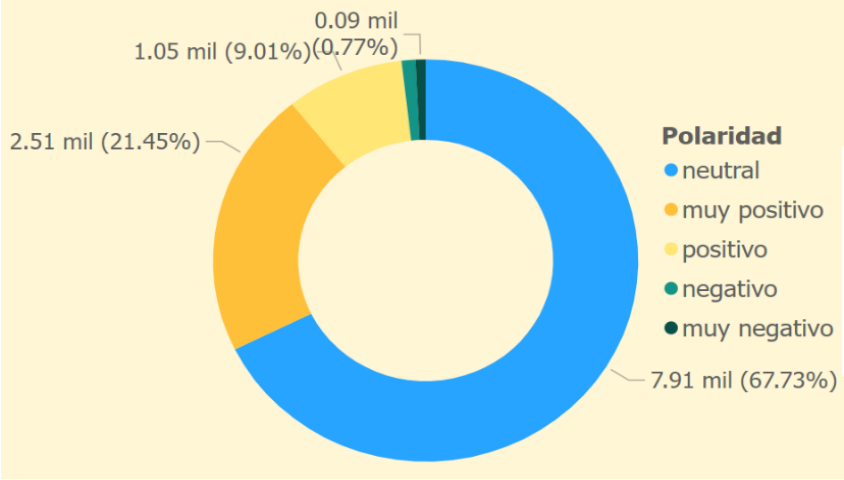
Nota: figura adaptada de Power BI

La combinación del análisis de sentimientos y polaridad permitió obtener una comprensión más profunda del contenido emocional expresado en las reseñas, brindando una visión completa de la actitud y satisfacción de los clientes hacia los servicios turísticos. Al presentar ambos análisis en una figura, se facilitó la identificación de patrones y tendencias relevantes relacionados con las experiencias de los clientes y sus percepciones generales hacia los hoteles y municipios de Santander.

Es importante destacar que la proporción de reseñas con relación a su polaridad se mantiene constante en todas las categorías, siendo la polaridad neutral la más predominante con un 67.73%, seguida de muy positivo con un 21.46% (2505 reseñas), positivo con un 9.01% (1052 reseñas), negativo con un 1.04% (122 reseñas) y muy negativo con un 0.77% (90 reseñas).

Figura 39.

Proporción de polaridad de las reseñas



Nota: figura adaptada de Power BI

A continuación, se presenta una descripción detallada de cada categoría y su segmento de clientes correspondiente:

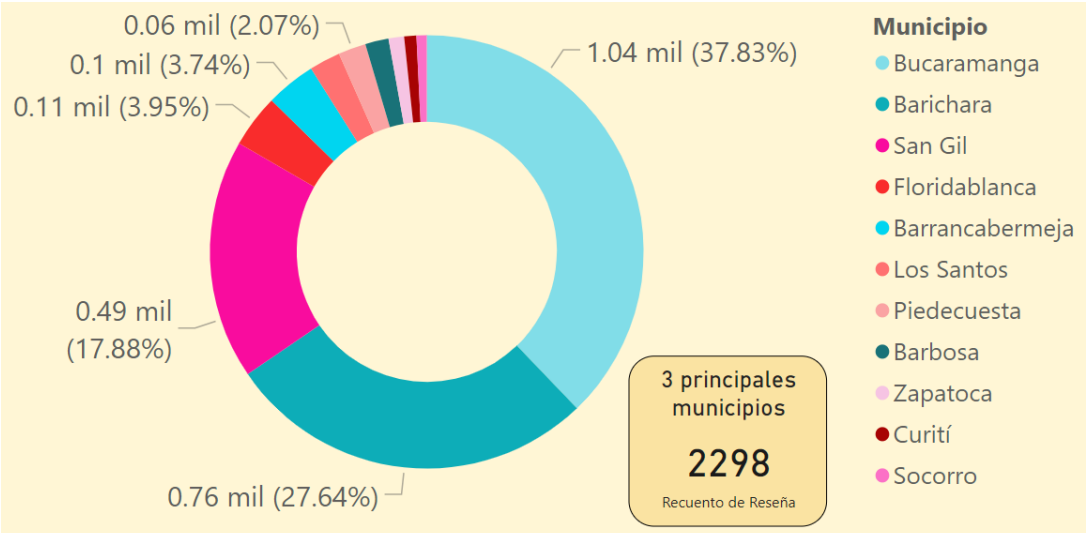
Figura 40.

Segmentación de clientes según la categoría de interés

1. **Comida-Turista gastronómico:** Este tipo de viajero busca experiencias culinarias auténticas y de alta calidad en sus destinos. Están dispuestos a explorar la gastronomía local, probar platos tradicionales y experimentar con sabores únicos.

Figura 41.

Proporción de reseñas por municipio: Comida



Nota: figura adaptada de Power BI

La dimensión culinaria emerge como un aspecto fundamental en las reseñas de este destino, abarcando un destacado 23.54% del total de opiniones en el departamento. Un total de 2.75 mil reseñas y 274 hoteles resaltan la importancia que la gastronomía tiene en la experiencia de los viajeros que buscan deleitar sus paladares. Entre los municipios que deslumbran en esta categoría figuran Bucaramanga, Barichara y San Gil, que atraen la atención con un impresionante 2.3 mil (83.5%) de todas las reseñas relacionadas con la comida.

El auténtico encanto culinario de esta región colombiana se despliega en cada plato y bocado. Bucaramanga, en particular, es un epicentro gastronómico donde las calles se llenan de aromas tentadores y sabores que narran la rica historia de la región. Desde puestos callejeros que ofrecen antojitos tradicionales hasta restaurantes innovadores que fusionan ingredientes locales con técnicas modernas, la escena culinaria de Bucaramanga es una invitación a una experiencia multisensorial.

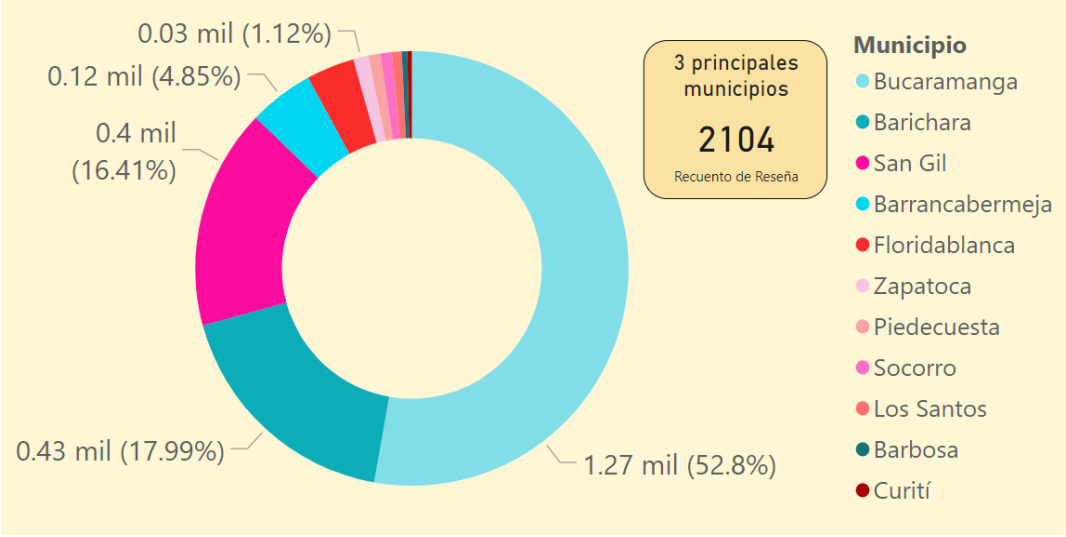
Barichara y San Gil, por su parte, no se quedan atrás en la exploración de sabores. Estos municipios brindan a los amantes de la comida la oportunidad de descubrir platos auténticos y exquisitos que han sido transmitidos a lo largo de generaciones. La conexión con los ingredientes locales y las recetas tradicionales crea un puente entre el pasado y el presente, permitiendo a los visitantes saborear la autenticidad de la cultura santandereana.

La plenitud de comentarios positivos en esta categoría refleja el placer y la satisfacción que experimentan los viajeros gastronómicos al explorar las delicias culinarias del departamento. Para el turista amante de la gastronomía, cada bocado es una oportunidad de conectarse con la cultura y la tradición, y de descubrir una nueva dimensión de la riqueza de Santander.

2. **Ubicación-Turista Explorador:** Estos turistas buscan destinos que les brinden acceso fácil a lugares importantes y populares, así como oportunidades para la aventura y la exploración de nuevos entornos. Valorarán la ubicación estratégica de los hoteles y la conveniencia de estar cerca de lugares de interés y actividades emocionantes.

Figura 42.

Proporción de reseñas por municipio: Ubicación



Nota: figura adaptada de Power BI

El aspecto de la ubicación emerge como un factor esencial en las reseñas de este destino, representando un 20.66% del total de opiniones en el departamento. Un total de 2413 reseñas y 267 hoteles son testigos del impacto que tiene la ubicación en la experiencia de los viajeros. Entre los municipios más destacados en esta categoría resaltan Bucaramanga, Barichara y San Gil, los cuales concentran asombrosamente el 87% de todas las reseñas relacionadas con la ubicación.

En este escenario, Bucaramanga lidera de manera indiscutible en la categoría de ubicación. La capital del departamento se distingue por su estratégica posición geográfica,

que se traduce en una ventaja para los visitantes. Los hoteles de Bucaramanga son reconocidos por su proximidad a lugares de interés clave y atractivos populares para los turistas, gracias a un sistema de vías de acceso sobresaliente y una variada oferta de medios de transporte. Esto otorga a los turistas la oportunidad de sumergirse rápidamente en la vibrante energía de la ciudad y de acceder cómodamente a experiencias únicas.

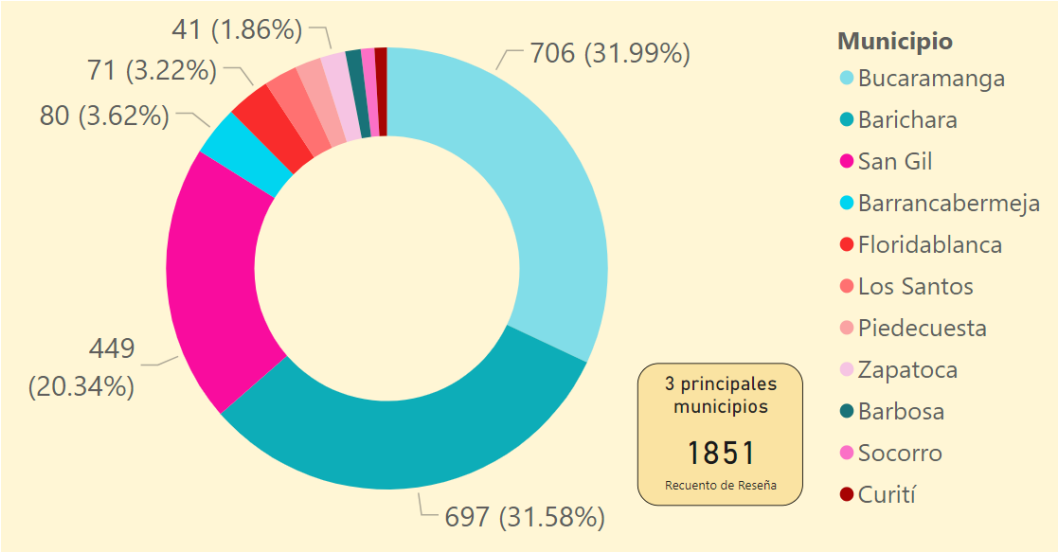
No obstante, la joya de esta región radica en que no se limita únicamente a Bucaramanga. Los municipios vecinos, como Barichara y San Gil, también hacen gala de su ubicación privilegiada en medio de entornos naturales y culturales cautivadores. Estos destinos brindan a los amantes de la aventura la posibilidad de explorar paisajes impresionantes y de embarcarse en emocionantes actividades al aire libre.

La abundante presencia de comentarios positivos en esta categoría resalta el sentimiento general de satisfacción y alegría entre los viajeros. Para el turista explorador o de aventura, la ubicación estratégica de los hoteles no solo brinda comodidad, sino también la puerta de entrada a un mundo de descubrimiento y emoción que espera ser explorado.

3. **Mantenimiento- Turista Confort:** Este tipo de viajero valora altamente su comodidad y la calidad de su estancia en un hotel. Busca instalaciones bien cuidadas, habitaciones limpias y un entorno que le permita relajarse y disfrutar de su tiempo en el destino.

Figura 43.

Proporción de reseñas por municipio: Mantenimiento



Nota: figura adaptada de Power BI

La categoría de mantenimiento emerge como un factor crucial en las reseñas de este destino, abarcando un significativo 18.85% del total de opiniones en el departamento. Un total de 2201 reseñas y 283 hoteles reflejan la importancia que el mantenimiento de las instalaciones tiene en la experiencia de los viajeros que buscan una estancia confortable y agradable. Entre los municipios que capturan la atención en esta categoría se encuentran Bucaramanga, Barichara y San Gil, que atraen la mirada con un sólido 1851 (84%) de todas las reseñas vinculadas con el mantenimiento.

Bucaramanga se alza nuevamente en el primer lugar, consolidando su posición como líder en la atención a los detalles que marcan la diferencia. La ciudad no solo recibe un flujo constante de turistas, sino que también demuestra su compromiso con el bienestar y la satisfacción de los visitantes. La presencia de grandes cadenas hoteleras en Bucaramanga no es solo un reflejo de su popularidad, sino también de la solvencia económica que

permite realizar un mantenimiento óptimo de las instalaciones. Esto garantiza que los turistas en busca de comodidad y relajación encuentren en la ciudad un oasis de bienestar.

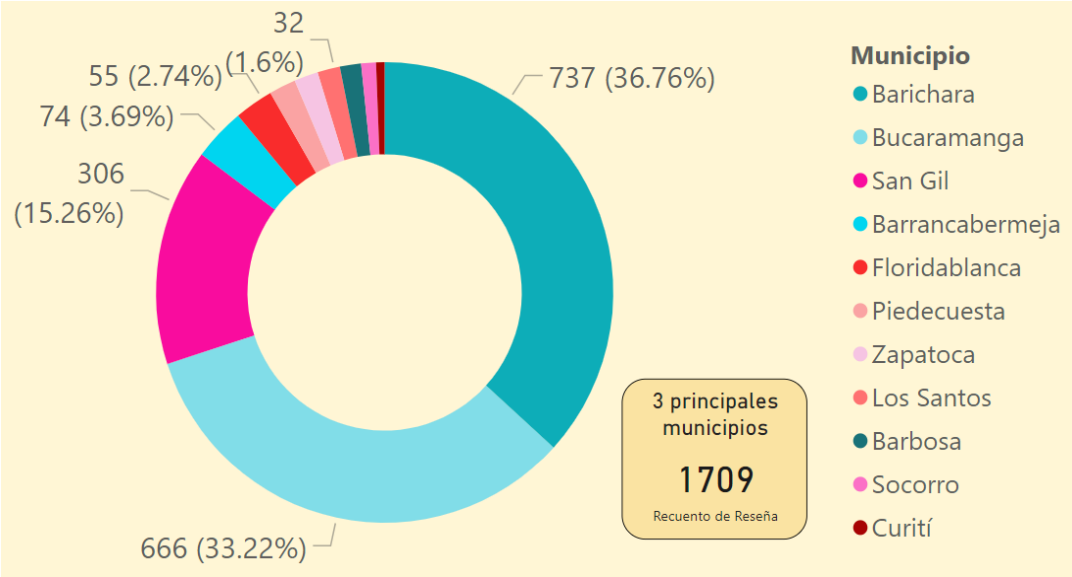
Barichara y San Gil, a su vez, también destacan en el frente del mantenimiento. Estos municipios comprenden que la experiencia del turista va más allá de los paisajes y las atracciones, extendiéndose al entorno donde descansan. Por lo tanto, se esfuerzan en mantener sus alojamientos en condiciones impecables, asegurando que cada rincón esté cuidado con atención y cariño.

La preponderancia de comentarios positivos en esta categoría subraya la satisfacción y la gratitud que los viajeros confortables sienten al encontrar instalaciones bien mantenidas. Para el turista que valora su comodidad, cada detalle bien cuidado es un testimonio del compromiso de la región con una experiencia de alojamiento excepcional.

4. **Diseño y Decoración-Turista cultural:** Estos tipos de viajeros valoran profundamente la estética visual y la experiencia cultural en sus destinos. Se sienten atraídos por la arquitectura, el diseño y la decoración que reflejan la identidad cultural y la belleza única de un lugar.

Figura 44.

Proporción de reseñas por municipio: Diseño y Decoración



Nota: figura adaptada de Power BI

En el vibrante mosaico de categorías, Diseño y Decoración se destaca, ocupando un sólido 17.14% de las reseñas que fluyen en el corazón de este departamento. Un total de 2 mil reseñas, cuidadosamente esparcidas por 258 hoteles, revelan la importancia que la estética y el diseño tienen en la experiencia de los viajeros que buscan la belleza y la autenticidad en cada rincón. En este fascinante paisaje, los municipios de Barichara, Bucaramanga y San Gil emergen como faros de creatividad y estilo, concentrando un impresionante 1709 (85%) del conjunto de reseñas en esta categoría.

Sin lugar a duda, Bucaramanga, el orgulloso portador del título de "ciudad bonita", exhibe una combinación de armonía arquitectónica y detalles cautivadores. Cada rincón de la ciudad es un tributo a la estética que va más allá de lo superficial, extendiéndose a la esencia misma de la cultura local. Sin embargo, es Barichara quien toma el centro de atención en este apartado. El famoso "pueblito más lindo de Colombia" seduce a los

amantes del diseño con su arquitectura colonial y sus callejones adoquinados que respiran encanto. La piedra Barichara, auténtico tesoro cultural, moldea su paisaje urbano, creando una experiencia estética que trasciende el tiempo.

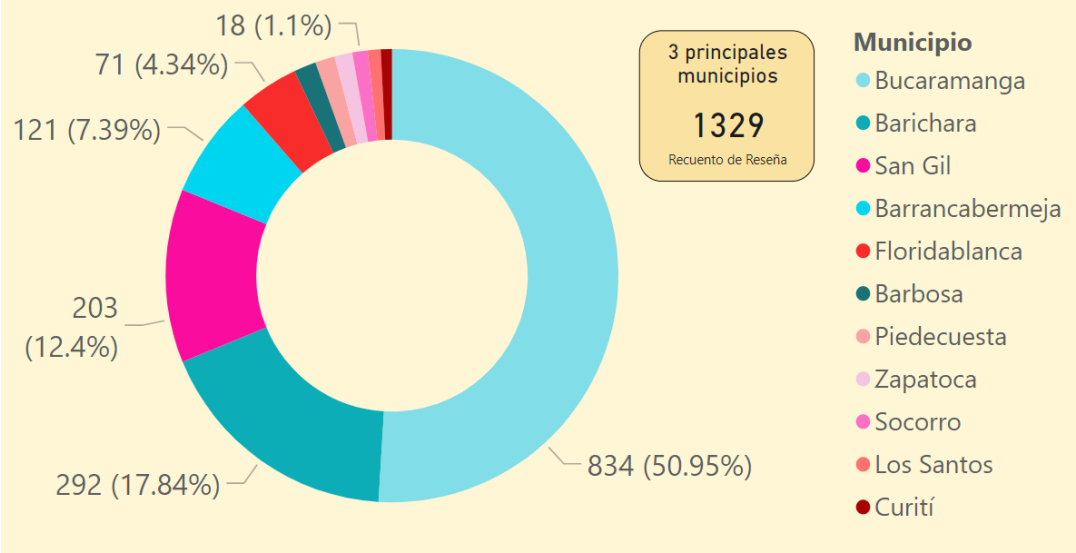
Tan impactante es su valor cultural y estético que Barichara ha sido reconocido incluso en las pantallas de cine. La película de Disney "Encanto" inmortalizó su encanto, presentando ante el mundo los tesoros arquitectónicos, culturales y folclóricos de Colombia. Esta proyección internacional es un testimonio del poder del diseño y la decoración para contar historias y emocionar a los viajeros.

En las reseñas, la emoción y el cariño nuevamente dominan en esta categoría, señalando que el turista estético y cultural encuentra en cada rincón una fuente de inspiración visual. Cada fachada, cada detalle decorativo, es una ventana a la rica herencia cultural y a la creatividad del destino.

5. **Servicio-Turista de lujo:** Estos tipos de viajeros buscan una experiencia de alto nivel en su estancia. Valoran un servicio al cliente impecable, personalizado y atención a los detalles para garantizar una experiencia de viaje memorable.

Figura 43

Proporción de reseñas por municipio: Servicio



Nota: figura adaptada de Power BI

En el dinámico panorama del servicio, surge una transformación marcada por las nuevas experiencias que buscan los viajeros de lujo. Esta categoría, que representa un sólido 13.94% de las reseñas en el departamento, con un total de 1628 reseñas y 222 hoteles, se redefine para satisfacer las demandas actuales de los huéspedes más exigentes. Los municipios que brillan en este contexto son Bucaramanga, Barichara y San Gil, acaparando un impresionante 1329 (81.6%) de todas las reseñas relacionadas con el servicio.

En este nuevo paradigma, el turista de lujo busca mucho más que comodidad y atención personalizada. Ahora, la sustentabilidad adquiere un papel protagónico, actuando como un equilibrio crucial entre la relación económica, ambiental y social con las comunidades locales. Además, la excelencia en el servicio toma la forma de un agente experimentado,

eliminando molestias y permitiendo al viajero gastar con sabiduría su tiempo y energía. Si bien esta tendencia no abarca a todos, algunos buscan la planificación anticipada como una estrategia para ahorrar dinero. Esto se traduce en reservaciones nacionales con 58 días de anticipación e internacionales con 80 días de anticipación, lo que guía la búsqueda de las mejores épocas para viajar, momentos en los que se pueden capturar fotografías sin la aglomeración de grandes grupos, y evitar los altos precios asociados a las temporadas más concurridas. En consecuencia, la temporada baja se ha convertido en un nuevo atractivo digno de considerar.

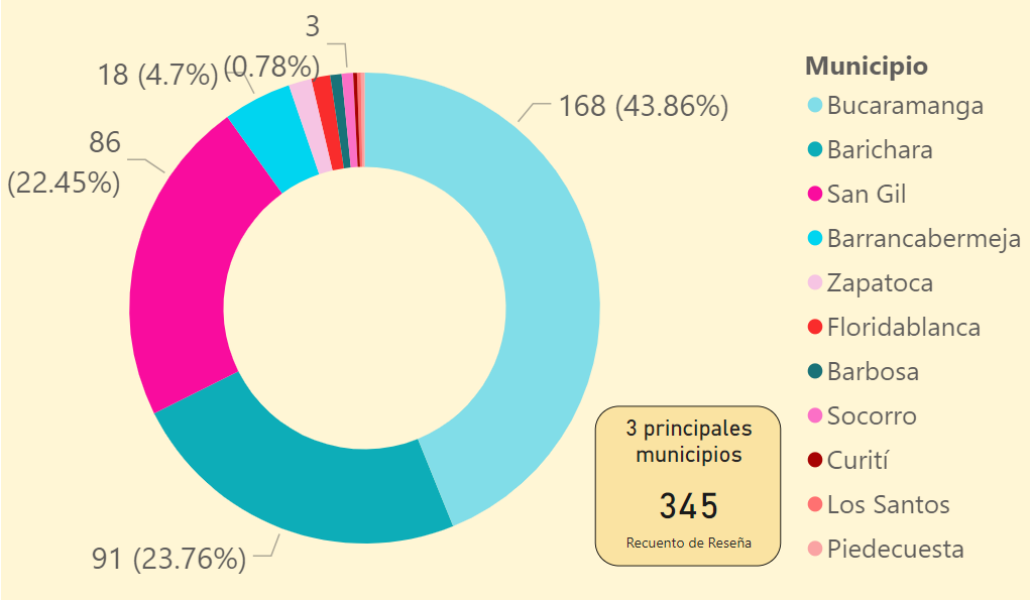
Además de las transformaciones en los patrones de reserva, los viajeros de lujo buscan destinos poco explorados. La nueva tendencia es lo que no está en la corriente principal, así que buscan lugares que aún no sean conocidos para obtener la libertad y tranquilidad anheladas después de un período de confinamiento prolongado. En este contexto, la seguridad, versatilidad y adaptabilidad para cada viajero son primordiales, valores que se alinean con las necesidades actuales. Esta evolución de la categoría de servicio coincide con las directrices establecidas por ProColombia (2021), que subrayan la importancia de estas transformaciones en la experiencia del viajero.

En las reseñas, la alegría y el amor siguen siendo recurrentes, resaltando la capacidad de la región para abrazar estas nuevas tendencias y proporcionar experiencias que satisfacen a la perfección las demandas de los viajeros de lujo. Cada sonrisa, cada detalle cuidadosamente planificado, refleja la dedicación de la industria turística hacia una experiencia que va más allá de las expectativas convencionales.

6. **Limpieza- Turista interesado por la Higiene:** Este tipo de viajero valora enormemente la limpieza y la higiene en su experiencia de alojamiento. Buscan ambientes impecables y bien cuidados para garantizar su comodidad y tranquilidad durante su estancia.

Figura 44

Proporción de reseñas por municipio: Limpieza



Nota: figura adaptada de Power BI

En el tapiz de categorías, emerge con vital importancia la dimensión de la limpieza, que representa el 3.27% de las reseñas en el departamento. Un total de 382 reseñas y 143 hoteles subrayan la relevancia que la higiene tiene en la experiencia de los viajeros más concienzudos. Entre los municipios que destacan en esta esfera se encuentran Bucaramanga, Barichara y San Gil, concentrando una notable cifra del 345 (90%) de todas las reseñas relacionadas con la limpieza.

Para el turista obsesivo por la higiene, la limpieza trasciende ser un mero detalle. Es un factor que impacta directamente en su bienestar y en la capacidad de disfrutar de su

estancia sin preocupaciones. Barichara, en particular, se convierte en un faro de excelencia en esta categoría. Las reseñas muestran que la percepción de los turistas hacia la limpieza es neutra, positiva y muy positiva, lo que demuestra el compromiso del municipio en brindar un ambiente limpio y pulcro que cumple y supera las expectativas.

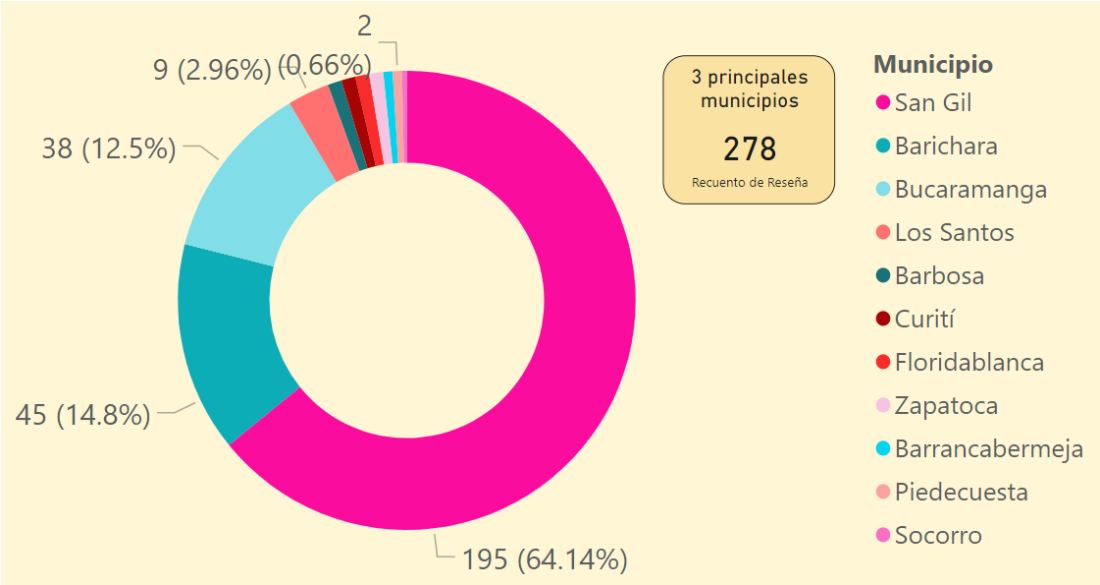
Es digno de mencionar que el municipio de Piedecuesta, con solo 1 reseña en esta categoría, resalta la singularidad de la atención que algunos destinos ponen en la limpieza como una prioridad.

En estas reseñas, los sentimientos de felicidad y amor nuevamente se hacen presentes, indicando que el turista obsesivo por la higiene encuentra en cada rincón una tranquilidad que supera sus expectativas. Cada detalle limpio, cada espacio reluciente, refleja la dedicación de la región hacia una experiencia de alojamiento que cuida de la comodidad y la paz mental de estos viajeros meticulosos.

7. **Actividades-Turista Aventurero:** Estos tipos de viajeros buscan emociones, actividades al aire libre y experiencias únicas que les permitan sumergirse en la cultura y el entorno natural del destino.

Figura 46

Proporción de reseñas por municipio: Actividades



Nota: figura adaptada de Power BI

En el escenario de categorías, las Actividades representan un 2.6% del conjunto de reseñas, con un total de 304 reseñas repartidas en 98 hoteles. San Gil, Barichara y Bucaramanga toman el centro de atención en esta categoría, concentrando un sorprendente 278 (91%) de todas las reseñas. No es sorprendente que San Gil, reconocida como la capital de los deportes extremos en Colombia, lidere esta categoría. Con su oferta diversa de emocionantes experiencias al aire libre, el municipio brinda una experiencia enriquecedora para el turista aventurero.

Los corazones intrépidos encuentran su hogar en San Gil, donde la adrenalina fluye junto a la belleza natural del entorno. Desde deportes extremos hasta actividades al aire libre, este destino promete una gama de experiencias que despiertan la pasión por la aventura.

Barichara y Bucaramanga también se unen al coro, ofreciendo opciones cautivadoras que atraen a los viajeros en busca de actividades enriquecedoras y llenas de emociones.

Las reseñas en esta área continúan mostrando una tendencia de percepción neutral, positiva y muy positiva, que refleja la satisfacción del turista aventurero al encontrar actividades que satisfacen su deseo de emociones y desafíos. Cada nueva experiencia al aire libre, cada oportunidad para sumergirse en la cultura local y el entorno natural agrega un capítulo emocionante a su viaje.

Socorro, con solo 1 reseña en esta categoría, destaca la exclusividad de las ofertas de actividades en algunos destinos. Sin embargo, para aquellos que buscan vivir al máximo, los municipios líderes ofrecen una ventana a un mundo de posibilidades emocionantes.

10. Conclusiones

- **Segmentación Significativa:** A través del análisis de sentimientos en reseñas de hoteles en TripAdvisor, este proyecto logró una segmentación de clientes sólida y significativa para el sector turístico de Santander. Cada categoría de viajero reveló preferencias y actitudes únicas, lo que ofrece una base esencial para la toma de decisiones estratégicas.

- **Diversidad de Preferencias:** Los resultados destacaron la diversidad de preferencias entre los diferentes tipos de viajeros. Desde los amantes de la gastronomía hasta los buscadores de aventuras, cada segmento valora aspectos específicos como la comida, la ubicación, el servicio y la limpieza, lo que subraya la necesidad de enfoques personalizados en la industria.

- **Adaptación al Cambio:** La evolución de las preferencias de los viajeros en respuesta a eventos como la pandemia y factores económicos fue evidente en las tendencias de las

reseñas a lo largo del tiempo. El análisis demuestra que el sector turístico de Santander pudo adaptarse a estas fluctuaciones y satisfacer las necesidades cambiantes de los viajeros.

- **Rol de la Tecnología:** La combinación de herramientas como el procesamiento de lenguaje natural y Power BI demostró ser crucial para extraer insights valiosos de las reseñas. La visualización efectiva y el análisis de sentimientos en conjunto permitieron comprender las percepciones de los clientes de manera más profunda y detallada.

- **Impacto en la Experiencia:** Las emociones positivas prevalecientes en las reseñas subrayan la importancia de ofrecer experiencias enriquecedoras y satisfactorias. Los resultados reafirman que el enfoque en detalles como la estética, la limpieza, el servicio personalizado y las actividades al aire libre puede generar emociones positivas y construir una imagen positiva de Santander como destino turístico.

Recomendaciones

- La extracción y análisis de datos se hace en un solo idioma en este caso español por temas de practicidad y recursos limitados, pero si se desea hacer un análisis más profundo de las reseñas, el modelo de chat GPT permite tomar la totalidad de las reseñas, identificar y etiquetar el idioma, para posteriormente realizar el análisis de sentimientos y clasificación por categoría sin importar el idioma, puede ser dando la instrucción de que traduzca las categorías al idioma de la reseñas o traduciendo todas las reseñas a un idioma específico.

- Para la extracción lo ideal sería correr un solo código para la extracción de todos los municipios, para esto se recomendaría encontrar una lógica en donde pregunte al inicio de cada ciclo por el link de la página web de cada municipio a analizar o en los posible solo por la ubicación y se dirija al link correspondiente. Además, optimizar el proceso de tal manera que el programa no realice más ciclos de los necesarios para la extracción de los datos.

- Al realizar el análisis de los datos se recomienda consultar a personas con experiencia dentro del sector que puedan dar opiniones y aspectos de utilidad para la investigación, ya que el un análisis es más enriquecedor entre más contexto se tenga del entorno.

- La realización de proyectos relacionados al uso de lenguajes de programación suele requerir manejar altos niveles de frustración porque involucran el ciclo de depuración, es decir, prueba, error, corregir y repetir el ciclo hasta que el código arroje el resultado deseado, además de que se recomienda estar en actitud de aprendizaje constante con cada código y cada intento, ya que cada uno muestra un nuevo desafío.

Referencias Bibliográficas

Aggarwal, S., & Gour, A, 2020. Peeking inside the minds of tourists using a novel web analytics approach. Journal of Hospitality and Tourism Management, 45(March), 580–591. https://doi.org/10.1016/j.jhtm.2020.10.009

Alaei, A. R., Becken, S., & Stantic, B, 2019. Sentiment Analysis in Tourism: Capitalizing on Big Data. Journal of Travel Research, 58(2), 175–191. https://doi.org/10.1177/0047287517747753

Amazon Machine Learning: Guía para desarrolladores, 2022

Amazon Machine Learning: Guía para desarrolladores. (s/f). Amazon.com. de https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/machinelearning-dg.pdf#cross-validation

Augusto Cortez Vásquez, M., Hugo Vega Huerta, M., Jaime, L., & Quispe, P, 2009. Procesamiento de lenguaje natural.

Ayala, P., & Rudas, J, 2019. Técnicas de Clustering Fuzzy para el análisis de tendencias en redes sociales. En Universidad Industrial de Santander (Vol. 3).

Bjørkelund, E., Burnett, T. H., & Nørvåg, K, 2012. A study of opinion mining and visualization of hotel reviews. Proceedings of the 14th International Conference on Information Integration and Web-Based Applications & Services - IIWAS '12, 229. https://doi.org/10.1145/2428736.2428773

Brob, J, 2013. Analysis of Customer Reviews Using Distant Supervision Techniques. University of Berlin.

Broersma, M., & Graham, T, 2012. SOCIAL MEDIA AS BEAT. Journalism Practice, 6(3), 403–419. https://doi.org/10.1080/17512786.2012.663626

Buhalis, D., & Law, R, 2008. Progress in information technology and tourism management: 20 years on and 10 years after the Internet-The state of eTourism research. Tourism Management, 29(4), 609–623. https://doi.org/10.1016/j.tourman.2008.01.005

Broersma, M., & Graham, T, 2012. SOCIAL MEDIA AS BEAT. Journalism Practice, 6(3), 403–419. https://doi.org/10.1080/17512786.2012.663626

Cui, Y., He, Q., & Khani, A, 2018. Travel Behavior Classification: An Approach with Social Network and Deep Learning. Transportation Research Record, 2672(47), 68–80. https://doi.org/10.1177/0361198118772723

Cui, Y., Meng, C., He, Q., & Gao, J, 2018. Forecasting current and next trip purpose with social media data and Google Places. Transportation Research Part C: Emerging Technologies, 97(September), 159–174. https://doi.org/10.1016/j.trc.2018.10.017

Culnan, M. J., McHugh, P. J., & Zubillaga, J. I., 2010). How large U.S. companies can use Twitter and other social media to gain business value. INDIANA UNIV, OPER & DECISION TECHNOL DEPTKELLEY SCH BUS, E 10 ST, BLOOMINGTON, IN 47405-1701, 9(4), 243–259.

Curry, B., & Moutinho, L, 1993. Neural Networks in Marketing: Modelling Consumer Responses to Advertising Stimuli. European Journal of Marketing, 27(7), 5–20. https://doi.org/10.1108/03090569310040325

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://github.com/tensorflow/tensor2tensor

Dubiau, L., & Ale, J. M., 2013. Análisis de Sentimientos sobre un Corpus en Español: Experimentación con un Caso de Estudio.

Feldman, R., 2013. Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82–89. https://doi.org/10.1145/2436256.2436274

García, D. N. C., & Suárez, P. C. S., 2021. Un algoritmo genético para la detección de comunidades en redes sociales mejorado mediante una técnica de clustering. En Universidad Industrial de Santander.

Gauch, S., Director Justin Zhan, T., member Ukash Nakarmi, C., & member Yanjun Pan, C., 2022. Movie Reviews Sentiment Analysis Using BERT.

Gelbukh, A., 2010. Procesamiento de lenguaje natural y sus aplicaciones Related papers. En Artículo invitado. Komputer Sapiens.

Gindl, S., Weichselbraun, A., & Scharl, A., 2010. Cross-Domain Contextualization of Sentiment Lexicons. www.google.com/language

Girvan, M., & Newman, M. E. J., 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12), 7821–7826. https://doi.org/10.1073/pnas.122653799

Hansen, D. L., Shneiderman, B., Smith, M. A., & Himelboim, I., 2011. Social network analysis: Measuring, mapping, and modeling collections of connections. Analyzing Social Media

Networks with NodeXL: Insights from a Connected World, 31–51. https://doi.org/10.1016/B978-0-12-817756-3.00003-0

Hernández-Méndez, J., Muñoz-Leiva, F., & Sánchez-Fernández, J., 2015. The influence of e-word-of-mouth on travel decision-making: consumer profiles. Current Issues in Tourism, 18(11), 1001–1021. https://doi.org/10.1080/13683500.2013.802764

Hernandez, O.,2013. Turismo Deportivo: Promoción Para La Diversificación de la Oferta Turistica En Manzanillo, Colima. Turismo y Desarrollo Local, 15.

Hunziker, W., & Krapf, K., 1942. Fundamentos de la Teoría General del Turismo.

Huyan, W., & Li, J., 2021. Research on rural tourism service intellectualization based on neural network algorithm and optimal classification decision function. Journal of Ambient Intelligence and Humanized Computing, 0123456789. https://doi.org/10.1007/s12652-021-03039-6

IBM, 2022. ¿Qué es Machine Learning?

Jiménez, N., & Jiménez, J., 2019. Evaluación de modelos de aprendizaje no supervisado para el análisis de contenido de tweets generados ante un desastre. En Universidad Industrial de Santander.

Kang, H., Yoo, S. J., & Han, D., 2012. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications, 39(5), 6000–6010. https://doi.org/10.1016/j.eswa.2011.11.107

Kirilenko, A. P., Stepchenkova, S. O., Kim, H., & Li, X. (Robert), 2018. Automated Sentiment Analysis in Tourism: Comparison of Approaches. Journal of Travel Research, 57(8), 1012–1025. https://doi.org/10.1177/0047287517729757

Kulkarni, S., & Rodd, S. F., 2020. Context Aware Recommendation Systems: A review of the state of the art techniques. Computer Science Review, 37, 100255. https://doi.org/10.1016/j.cosrev.2020.100255

Larkin, B. A., & Fink, J. S., 2016. Fantasy Sport, FoMO, and Traditional Fandom: How Second-Screen Use of Social Media Allows Fans to Accommodate Multiple Identities. Journal of Sport Management, 30(6), 643–655. https://doi.org/10.1123/jsm.2015-0344

Li, Q., Li, S., Zhang, S., Hu, J., & Hu, J., 2019. A Review of Text Corpus-Based Tourism Big Data Mining. Applied Sciences, 9(16), 3300. https://doi.org/10.3390/app9163300

Likas, A., Vlassis, N., & J. Verbeek, J., 200. The global k-means clustering algorithm. Pattern Recognition, 36(2), 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2

Liu, C., Zhang, W., Chen, G., Wu, X., Luu, A. T., Chang, C. H., & Bing, L., 2023. Zero-Shot Text Classification via Self-Supervised Tuning. http://arxiv.org/abs/2305.11442

López González, J., Martínez Niño, D. K., Herrera, L., & Abarca, R. M., 2021. Comparación de métodos para clasificar comentarios de lugares turísticos por medio de análisis de sentimiento. Nuevos sistemas de comunicación e información, 2013–2015. https://repositorio.ulima.edu.pe/handle/20.500.12724/12195

Loria, S., 2020. Textblob.

Martínez Quintana, V., 2017. El turismo de naturaleza: Un producto turístico sostenible. Arbor, 193(785). https://doi.org/10.3989/arbor.2017.785n3002

MINCIT., 2021. Informes de Turismo. MINCIT. https://www.mincit.gov.co/estudios-economicos/estadisticas-e-informes/informes-de-turismo

Mohalem, A. J. V., 2018. Herramienta para el análisis de big data aplicado a un sistema de recomendación utilizando MapReduce. En Universidad Industrial de Santander.

Moreno, M., & Coromoto, M., 2011. Turismo y producto turístico. Evolución, conceptos, componentes y clasificación. Visión Gerencial, 0(1), 135-158–158.

Nilashi, M., Samad, S., Ahani, A., Ahmadi, H., Alsolami, E., Mahmoud, M., Majeed, H. D., & Abdulsalam Alarood, A., 2021). Travellers decision making through preferences learning: A case on Malaysian spa hotels in TripAdvisor. Computers and Industrial Engineering, 158(September 2020), 107348. https://doi.org/10.1016/j.cie.2021.107348

Osgood, C. E., Suci, G. J., & Tannonbaum, P. H., 1976. La medida del significado. Gredos.

P. de la Hoz, A., 2021. Entendiendo el turismo de salud: un análisis sociodemográfico. Escenarios: Empresa Y Territorio, 2(2). Recuperado a partir de http://revistas.esumer.edu.co/index.php/escenarios/article/view/177

Pang, B., Lee, L., & Vaithyanathan, S., 2002. Thumbs up? Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP ’02, 79–86. https://doi.org/10.3115/1118693.1118704

Pantano, E., Priporas, C. V., & Stylos, N., 2016. 'You will like it!' using open data to predict tourists' response to a tourist attraction. Tourism Management, 60, 430–438. https://doi.org/10.1016/j.tourman.2016.12.020

Phillips, P., Zigan, K., Santos Silva, M. M., & Schegg, R., 2015. The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis. Tourism Management, 50, 130–141. https://doi.org/10.1016/j.tourman.2015.01.028

Pineda, J. D. M., 2021. Un modelo para la predicción del movimiento del precio de las acciones del mercado bursátil basado en un análisis de sentimiento y datos históricos de la BVC. En Universidad Industrial de Santander.

Priscila Falcón Serra, 2020. Clasificación y tipos de turismo. Entorno Turistico. https://www.entornoturistico.com/clasificacion-y-tipos-de-turismo/

Portafolio. 2021). Inversión en turismo en Colombia en el 2021. Portafolio. https://www.portafolio.co/negocios/inversion/inversion-en-turismo-en-colombia-en-el-2021-557273

Procolombia, 2021. Cinco tendencias para viajar en 2021. https://procolombia.co/noticias/cinco-tendencias-para-viajar-en-2021#:~:text=Viajes cortos y seguros%2C turistas,las nuevas formas de viajar.&text=Estas son las cinco tendencias,a partir de diversos estudios.

Rabanser, U., & Ricci, F., 2005. Recommender Systems: Do They Have a Viable Business Model in e-Tourism? En Information and Communication Technologies in Tourism 2005 (pp. 160–171). Springer-Verlag. https://doi.org/10.1007/3-211-27283-6_15

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. Proceedings of the National Academy of Sciences, 101(9), 2658–2663. https://doi.org/10.1073/pnas.0400054101

Ren, G., & Hong, T., 2017. Investigating online destination images using a topic-based sentiment analysis approach. Sustainability (Switzerland), 9(10). https://doi.org/10.3390/su9101765

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F., 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science, 5(1). https://doi.org/10.1140/epjds/s13688-016-0085-1

Riquelme, J. C., Ruiz, R., & Gilbert, K., 2006. Minería de Datos: Conceptos y Tendencias. Revista Iberoamericana de Inteligencia Artificial. No, 29, 11–18. http://www.aepia.org

Rouhiainen, L., 2018. Inteligencia artificial : 101 cosas que debes saber hoy sobre nuestro futuro. Alienta.

Shimada, K., Inoue, S., Maeda, H., & Endo, T., 2011. Analyzing Tourism Information on Twitter for a Local City. 2011 First ACIS International Symposium on Software and Network Engineering, 61–66. https://doi.org/10.1109/SSNE.2011.27

Suthaharan, S., 2000. Support Vector Machine. Springer, PRMU2000-1, 63–68. https://doi.org/10.1007/978-1-4899-7641-3

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A., 2010. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558. https://doi.org/10.1002/asi.21416

TURING, A. M., 1950. I.—COMPUTING MACHINERY AND INTELLIGENCE. Mind, LIX(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I., 2017. Attention Is All You Need. http://arxiv.org/abs/1706.03762

Victoriano, R., Paez, A., & Carrasco, J. A., 2020. Time, space, money, and social interaction: Using machine learning to classify people's mobility strategies through four key dimensions. Travel Behaviour and Society, 20(February), 1–11. https://doi.org/10.1016/j.tbs.2020.02.004

Wang, F. Y., 2014. Scanning the issue and beyond: Real-time social transportation with online social signals. En IEEE Transactions on Intelligent Transportation Systems (Vol. 15, Issue 3, pp. 909–914). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/TITS.2014.2323531

World turism organization., 2021. 2020: EL PEOR AÑO DE LA HISTORIA DEL TURISMO, CON MIL MILLONES MENOS DE LLEGADAS INTERNACIONALES. https://www.unwto.org/es/news/2020-el-peor-ano-de-la-historia-del-turismo-con-mil-millones-menos-de-llegadas-internacionales

World turism organization., 2022. Glosario de términos de turismo | OMT. https://www.unwto.org/es/glosario-terminos-turisticos

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

Zapata, G., Murga, J., Raymundo, C., Dominguez, F., Moguerza, J. M., & Alvarez, J. M., 2019. Business information architecture for successful project implementation based on sentiment analysis in the tourist sector. Journal of Intelligent Information Systems, 53(3), 563–585. https://doi.org/10.1007/s10844-019-00564-x

Zheng, X., Chen, W., Wang, P., Shen, D., Chen, S., Wang, X., Zhang, Q., & Yang, L. ,2016. Big Data for Social Transportation. IEEE Transactions on Intelligent Transportation Systems, 17(3), 620–630. https://doi.org/10.1109/TITS.2015.2480157