COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

Análisis de Percepciones en la Hostelería Digital: Minería de Texto en las Revisiones de Airbnb en Redes Sociales

Autor: Alejandra Bandeira Eguiraun
Director: Jenny Alexandra Cifuentes Quintero

MADRID| Abril 2024

Resumen

La economía colaborativa ha transformado la forma en que intercambiamos bienes y servicios, destacando la importancia de la cooperación y el acceso compartido sobre la posesión exclusiva. Plataformas como Airbnb han sido pioneras en este sector, permitiendo a individuos de todo el mundo rentar sus espacios de vivienda a viajeros, lo cual refleja un cambio significativo en la industria del hospedaje. Para una marca como Airbnb, es relevante comprender las percepciones y opiniones de las personas para afinar sus servicios y estrategias de mercado, mejorando así la experiencia del usuario y reforzando su posicionamiento en el mercado. En este contexto, las redes sociales han surgido como herramientas valiosas para captar estas percepciones. X, en particular, se destaca por su capacidad de facilitar discusiones en tiempo real y a gran escala.

En este contexto investigativo, el presente Trabajo de Fin de Grado se centra en analizar la percepción de los usuarios de X sobre Airbnb, aplicando técnicas de modelado de tópicos y análisis de sentimientos a las publicaciones en la plataforma. Se ha optado por emplear la metodología de Latent Dirichlet Allocation (LDA) para el modelado de tópicos y el diccionario VADER para el análisis de sentimientos. La selección de estas herramientas está respaldada por estudios previos que han demostrado la eficacia de LDA en la identificación de temas latentes dentro de grandes conjuntos de datos y la capacidad de VADER para reconocer el lenguaje informal propio de las redes sociales.

En relación al modelado de tópicos, se identificaron 5 categorías principales en el corpus analizado, etiquetadas como “experiencia de la estancia”, “modelo de negocio”, “peligros de Airbnb”, “anfitriones” y “monetización”. Por otro lado, el análisis de sentimientos mostró una predominante connotación positiva en las publicaciones examinadas, con más del setenta y cinco por ciento de ellas recibiendo una calificación de sentimiento igual o superior a 0. Durante el estudio, se evaluó la tendencia temporal de los sentimientos, identificando que los usuarios generalmente tienen una actitud neutral o ligeramente positiva. Además, se constató que ciertos eventos tienen un impacto significativo en la expresión de los sentimientos de los usuarios, como la controversia suscitada por el anuncio de Airbnb de no eliminar los anuncios de asentamientos israelíes ilegales en Cisjordania y la oferta pública inicial de la empresa.

Palabras clave: Economía colaborativa, Análisis de sentimientos, Modelado de tópicos, Redes sociales

Abstract

The shared economy has transformed the way we exchange goods and services, highlighting the importance of cooperation and shared access over exclusive possession. Platforms like Airbnb have been pioneers in this sector, allowing individuals from all around the world to rent their houses to travelers, reflecting a significant change in the hospitality industry. For a brand like Airbnb, it is important to understand people's perceptions and opinions in order to target their services and market strategies, thus improving the user experience and strengthening their market positioning. In this context, social media has emerged as a valuable tool to capture these perceptions. X, in particular, stands out for its ability to facilitate discussions in real time and on a large scale.

In this research context, this study focus on analyzing the perception of X users on Airbnb, by applying topic modeling techniques and sentiment analysis to the publications on the platform. The Latent Dirichlet Allocation (LDA) methodology has been chosen for topic modelling and the VADER dictionary for the sentiment analysis. The selection of these tools is supported by previous studies that have demonstrated the effectiveness of LDA in identifying latent topics within large data sets and VADER's ability to recognize the informal language in social networks.

In relation to topic modelling, five main categories were identified in the corpus analyzed, labeled as "stay experience", "business model", "Airbnb's dangers", "hosts" and "monetization". On the other hand, the sentiment analysis showed a predominant positive connotation in the publications examined, with more than seventy-five percent of them receiving a sentiment score equal or greater than 0. During the study, the sentiments temporal tendency was evaluated, identifying that users generally have a neutral or a slightly positive attitude. In addition, it was found that certain events have a significant impact on the expression of users' feelings, such as the controversy surrounding Airbnb's announcement not to eliminate advertisements for illegal Israeli settlements in the West Bank and the company's initial public offering.

Keywords: Shared economy, Sentiment analysis, Topic modelling, Social networks

Índice general

Índice de figuras

Índice de tablas

Acrónimos

API	Application Programming Interface
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
B2P	Business-to-peer
CNMC	Comisión Nacional de los Mercados y la Competencia
EEUU	Estados Unidos
MLG	Modelo Lineal Generalizado
IDF	Inverse Document Frequency
IEQ	Indoor Environment Quality
IPO	Initial Public Offering
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
ML	Machine Learning
P2P	Peer-to-peer
PLN	Procesamiento de Lenguaje Natural
RAE	Real Academia Española
TF	Inverse Document Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
URL	Uniform Resource Locator
VADER	Valence Aware Dictionary and Sentiment Reasoner

Capítulo 1

Introducción

1.1. Motivación

En un mundo cada vez más interconectado, el desarrollo de las tecnologías de la información y el crecimiento de la web han sido fundamentales en la creación de plataformas en línea que fomentan la creación de contenido, el intercambio y la colaboración entre usuarios (Hamari, Sjöklint, y Ukkonen, 2016). Ejemplos notables de estas plataformas incluyen Wikipedia, YouTube, BlaBlaCar y GitHub, donde la interacción y la cooperación entre usuarios son elementos clave. Estas plataformas no solo han facilitado el intercambio de información y conocimientos, sino que también han permitido el intercambio de bienes y servicios, dando lugar a la emergía de la economía colaborativa o *shared economy*.

La Real Academia Española (RAE) define la "economía" como la ciencia que estudia los métodos más eficaces para satisfacer las necesidades humanas materiales con bienes escasos, y el término "colaborativo" como el acto de trabajar conjuntamente para realizar una obra (RAE, 2022). En este sentido, la Comisión Nacional de los Mercados y la Competencia (CNMC) describe la **economía colaborativa** como un modelo emergente impulsado por avances tecnológicos, que facilita el acceso a bienes y servicios mediante la compartición, el reciclaje y la reutilización (CNMC, 2016). Esta definición subraya la relevancia de la tecnología en la creación de nuevos sistemas de producción y consumo.

Aunque la idea de economía colaborativa no es reciente, tal como lo demuestran prácticas ancestrales como el trueque y el trabajo comunal, ha ganado un nuevo significado en la era digital (Carlemany, 2022). Esta reinterpretación contemporánea, estimulada por contribuciones académicas relevantes (Botsman y Rogers, 2010; Gansky, 2010), se enfoca en cómo los modelos de negocio se han adaptado para capitalizar el crecimiento de Internet. A pesar de la falta de un consenso sobre una definición exacta de economía colaborativa, y su continua discusión (Hamari et al., 2016), es fundamental reconocer su naturaleza colectiva y comunitaria, así como su dependencia de las tecnologías peer-to-peer (P2P) y business-to-peer (B2P) (Díaz Foncea, Marcuello, y Montreal-Garrido, 2016). En este contexto eco-

nómico actual, los consumidores han modificado su enfoque hacia el consumo, optando por acceder a activos o servicios infrautilizados a través de plataformas digitales, a menudo a cambio de compensaciones monetarias o de otro tipo (Durán-Sánchez, Álvarez-García, del Río, Maldonado-Erazo, et al., 2016). Es así como la economía colaborativa ha extendido su influencia a ámbitos tan variados como el *crowdfunding* (Duque, 2021) y la inversión de capital de riesgo en empresas emergentes con gran potencial (Espinosa Fernández, 2018)

Airbnb, un pilar de la economía colaborativa en la industria hotelera, nació en 2007 como resultado de una idea innovadora entre amigos en San Francisco. Brian Chesky, Nathan Blecharczyk y Joe Gebbia transformaron una solución temporal para hospedar a tres invitados en su casa en una plataforma global que hoy acumula más de 4 millones de anfitriones en más de 220 países y regiones, con más de 100.000 anuncios activos (Statista, Enero 2023). Su crecimiento ha sido exponencial, con ingresos reportados en 2022 de unos 8.400 millones de dólares (ver Figura 1.1) , y su exitosa salida a bolsa en 2020 valoró la compañía en 101.600 millones de dólares, equivalente a 165 dólares por acción (Expansión, Diciembre 2020).

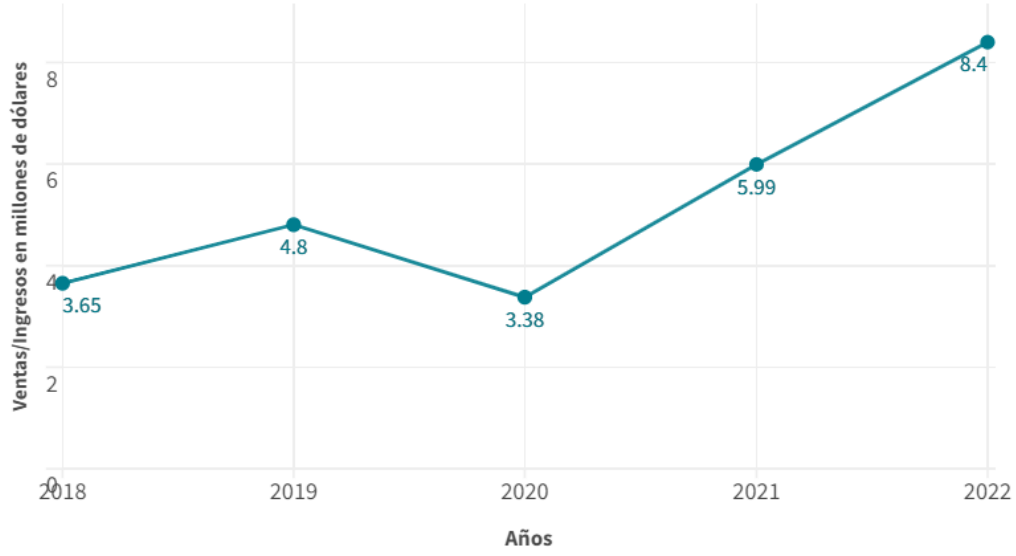

Figura 1.1: Evolución de ventas/ingresos de Airbnb de 2018-2022. Fuente de datos: (Journal, 2023). Elaboración propia

Airbnb se distingue de las opciones tradicionales de hospedaje al permitir que cualquier persona ofrezca una habitación en su hogar para hospedar a viajeros, reflejando así el espíritu de ayuda y colaboración frente al capitalismo puro (Edelman y Luca, 2014). El funcionamiento de la plataforma es sencillo: los usuarios buscan alojamiento introduciendo su destino y fechas deseadas, escogiendo entre diversas opciones como habitaciones individuales o apartamentos completos. La reserva se confirma tras la aprobación del anfitrión, basada en las opiniones y fotos del huésped, y el pago se realiza una vez aceptada la estancia.

Sin embargo, este modelo de negocio también presenta desafíos, especialmente en cuanto

a la calidad del servicio ofrecido. En este enfoque, las valoraciones y comentarios de los usuarios son esenciales para establecer la confianza necesaria en estas transacciones (Hamari et al., 2016). Así, las redes sociales como X (antes Twitter) juegan un rol crucial en la difusión y análisis de estas opiniones, proporcionando una perspectiva valiosa sobre las experiencias de los usuarios con Airbnb.

La influencia de las redes sociales en este contexto es innegable. Se han convertido en una ventana abierta a las opiniones y experiencias de una vasta audiencia, brindando un espacio donde los usuarios expresan sus impresiones, críticas, y recomendaciones de manera espontánea y directa. Este flujo constante de información resulta invaluable para entender cómo los consumidores interactúan con la economía colaborativa y Airbnb en particular, permitiendo identificar tendencias, preferencias y preocupaciones de forma casi instantánea (Tsimonis y Dimitriadis, 2014). Además, el análisis de estas interacciones en redes sociales ofrece una metodología escalable y rápida, crucial para manejar grandes volúmenes de datos y extraer información valiosa en tiempo real. La capacidad de analizar detalladamente estos datos convierte a la minería de datos en redes sociales en una herramienta esencial para comprender las dinámicas de un mercado en constante evolución y tomar decisiones informadas. En este sentido, la minería de texto se revela como un enfoque particularmente potente para explorar las complejidades de las percepciones y opiniones compartidas en plataformas como X (antes Twitter), proporcionando información clave para el análisis de la economía colaborativa y su impacto en la industria hotelera (Elgendy y Elragal, 2014). En este contexto, X se destaca como una plataforma especialmente ventajosa para este tipo de análisis. Su naturaleza abierta hace que sea un recurso rico para capturar las opiniones en tiempo real. Sus publicaciones, debido a su brevedad, ofrecen declaraciones concisas y directas de diversas experiencias, lo que facilita un análisis más ágil. Además, la estructura de X, con su sistema de hashtags y la posibilidad de seguimiento de temas tendencia, permite rastrear discusiones relacionadas a temas específicos (Izquierdo Expósito, Álvarez Rodríguez, y Nuño Barrau, 2017).

En el marco de la economía colaborativa, este estudio se centra en el análisis de las percepciones de los usuarios sobre Airbnb a través de las redes sociales, con un énfasis particular en publicaciones que reflejan la calidad del servicio ofrecido. La relevancia de este análisis radica en su capacidad para influir en la reputación online de la plataforma, afectar las decisiones de los consumidores y guiar la formulación de estrategias de mejora y marketing. Al explorar tanto las experiencias positivas como los desafíos señalados por los usuarios, el objetivo es obtener una comprensión integral de las expectativas y comportamientos de los consumidores. Esto no solo proporciona información valiosa para Airbnb y otras plataformas de economía colaborativa, sino que también contribuye al entendimiento general de cómo las interacciones en redes sociales pueden impactar la confianza de los usuarios en este dinámico sector del mercado.

1.2. Objetivos

El objetivo principal de este proyecto es realizar un análisis de las revisiones en línea de los usuarios de Airbnb en X (antes Twitter) utilizando técnicas avanzadas de minería de texto, como el modelado de tópicos y el análisis de sentimientos. El propósito es evaluar la percepción de la ciudadanía hacia esta línea de la industria hotelera, identificando las diversas características que pueden tener un impacto significativo en la percepción general de la calidad dentro de este sector.

1.2.1. Objetivos Específicos

- Contextualizar la relevancia y el impacto de la minería de texto en el análisis de revisiones en línea de usuarios de Airbnb en X, destacando cómo esta tecnología puede proporcionar una comprensión más profunda de las opiniones de los usuarios y su relación con la percepción de calidad en el sector de la hostelería.

- Desarrollar una revisión detallada de la literatura para identificar y comprender las diversas técnicas de minería de texto que se pueden aplicar específicamente en el análisis de tweets sobre Airbnb. Se busca conocer en detalle las metodologías y enfoques utilizados en investigaciones anteriores para analizar y extraer información relevante de las revisiones de los usuarios realizadas en X (antes Twitter).

- Identificar categorías de discusión clave que reflejen las preferencias y experiencias de los usuarios, así como evaluar la polaridad y la intensidad de sentimientos expresados en las revisiones, contribuyendo así a una comprensión más completa de la percepción de calidad en el sector de la hostelería.

1.3. Estructura del documento

Este Trabajo de Fin de Grado, con el objetivo de alcanzar las metas planteadas, se estructurará en cinco capítulos bien definidos. El primer capítulo introducirá la motivación detrás de este estudio y detallará los objetivos específicos de análisis, estableciendo así el escenario para la investigación. En el capítulo 2, dedicado al estado del arte, se realizará una revisión de la literatura existente. Aquí, se explorarán estudios similares para comprender las estrategias previamente adoptadas en el análisis de la percepción ciudadana en la industria hotelera, proporcionando un marco teórico para el estudio.

El capítulo 3, por su parte, se centrará en describir la metodología empleada en este trabajo. Se explicarán en detalle los pasos seguidos, desde la limpieza y procesamiento de los datos hasta la implementación de técnicas específicas como un análisis descriptivo detallado,

modelado de categorías de discusión y análisis de sentimientos. Este capítulo también abordará las herramientas y tecnologías utilizadas, así como los criterios de selección de datos. En el capítulo 4, se presentarán y discutirán los resultados obtenidos mediante la metodología propuesta, ofreciendo un análisis de su relevancia en el contexto del estudio. Finalmente, el capítulo 5 concluirá el trabajo resumiendo los principales resultados, discutiendo sus implicaciones y sugiriendo posibles direcciones para futuras investigaciones.

Capítulo 2

Analizando las percepciones sociales sobre la hostelería: Revisión de la Literatura

En la era digital actual, las redes sociales han evolucionado significativamente más allá de su función original de conectar personas, transformándose en extensos repositorios de información. Estas plataformas capturan y presentan una diversidad de opiniones, percepciones y experiencias compartidas por millones de usuarios alrededor del mundo, convirtiéndose en un espejo fiel de la sociedad y sus tendencias. Esta metamorfosis es de especial relevancia en industrias orientadas al servicio, como la hostelería, donde las redes sociales actúan como un canal inmediato y transparente para acceder a las opiniones y a la realimentación de los clientes. En este entorno digital, las redes sociales no solo reflejan las expectativas y experiencias de los usuarios, sino que también revelan sus niveles de satisfacción o insatisfacción con los servicios recibidos, ofreciendo así un panorama en tiempo real de las reacciones del consumidor. Esta característica las convierte en herramientas invaluables para las empresas del sector, permitiéndoles sintonizar con las necesidades y preferencias del cliente de manera rápida y eficaz.

En este contexto, la analítica de datos se destaca como una herramienta vital para manejar y extraer significado de los enormes volúmenes de información generada en estas plataformas. Técnicas avanzadas en áreas como la minería de texto y el procesamiento de lenguaje natural (PLN) permiten analizar de forma automática y eficiente estas grandes cantidades de información. Estos métodos no solo identifican patrones y tendencias en las opiniones y comentarios de los usuarios, sino que también facilitan una comprensión más profunda de los factores que influyen en la percepción del público sobre los servicios de hostelería. De esta manera, el análisis de las redes sociales a través de la analítica de datos se convierte en un enfoque de gran relevancia para extraer información valiosa sobre la percepción de los consumidores en la industria hotelera. Al analizar automáticamente las discusiones, reseñas y

comentarios provenientes de estas plataformas, se puede obtener una comprensión detallada de lo que los clientes valoran en sus experiencias de hospedaje e identificar áreas de mejora para los proveedores de servicios. Este enfoque no solo es beneficioso para las empresas del sector en su toma de decisiones estratégicas, sino que también contribuye al conocimiento académico sobre la interacción entre consumidores y servicios en el entorno digital.

Dada la relevancia de las redes sociales como fuentes de información sobre las experiencias y opiniones de los consumidores, varios investigadores han enfocado sus esfuerzos en realizar estudios en este campo, especialmente en la industria de la hostelería. Estas investigaciones buscan comprender la forma en la que las percepciones y experiencias compartidas en plataformas digitales pueden influir en la reputación y el éxito de servicios como los ofrecidos por Airbnb y otros participantes de la economía colaborativa. En esta línea de estudio, (Chang y Wang, 2018) realizaron una investigación focalizada en las reseñas de dos de las principales plataformas de alojamiento nocturno en la economía colaborativa en el mercado de Estados Unidos: Airbnb y HomeAway. Para su análisis, seleccionaron un conjunto de 200 reseñas del período de enero de 2017, priorizando los comentarios más recientes para capturar las tendencias actuales de los usuarios. El estudio se centró en varios factores clave, como el grado de satisfacción del anfitrión (*host*), la limpieza del alojamiento, su localización y la satisfacción general con la experiencia. Para llevar a cabo el análisis de sentimientos, utilizaron herramientas como el Linguistic Inquiry and Word Count (LIWC) y un modelo basado en la evaluación del ratio de sentimiento descrito en (Koh, 2011).

Los resultados del estudio revelaron que en Airbnb predominaban los sentimientos positivos sobre los negativos, con un ratio de sentimiento de 0,8, lo que indica una tendencia general hacia experiencias positivas en la plataforma. Sumado a estos análisis, el estudio también exploró cómo el ratio de sentimiento influía en las decisiones de los usuarios de diferentes generaciones, mostrando variaciones en la valoración de atributos específicos. Los resultados indicaron que la percepción y el proceso de toma de decisiones variaban significativamente entre las generaciones Z, Y y X, cada una asignando diferentes niveles de importancia a los atributos de los alojamientos. Para la Generación Z, aquellos nacidos entre 1990 y 2010, la decisión se basó principalmente en la primera impresión, sin verse afectada significativamente por el ratio de sentimiento. Por otro lado, la Generación Y, es decir, los nacidos entre 1982 y 1994, priorizó factores como la limpieza, el valor, la precisión y la calificación global al tomar sus decisiones de alojamiento. Finalmente, para la Generación X, nacidos entre 1965 y 1980, la limpieza y la calificación global fueron determinantes en su elección, mientras que las imágenes disponibles en las reseñas tuvieron un impacto menor en su proceso de decisión.

Además de la investigación en plataformas de alojamiento como Airbnb y HomeAway, otros estudios han explorado distintos aspectos de la economía colaborativa, incluyendo aplicaciones como Wallapop, que facilita el intercambio de objetos de segunda mano. En este contexto, (Prada y Iglesias, 2020) realizaron un estudio sobre la huella digital de los usuarios

de Wallapop en Twitter, ahora conocida como X. Su investigación se centró en determinar si la huella digital de los usuarios podía predecir comportamientos inadecuados en el futuro. Este estudio resaltó, como elemento clave del éxito en la economía colaborativa, la capacidad de generar confianza entre desconocidos, utilizando sistemas de retroalimentación. Además, como parte de la investigación, los autores desarrollaron un modelo predictivo que utilizaba las huellas digitales dejadas por los usuarios en plataformas sociales como X para predecir la reputación de los mismos en Wallapop. Con este fin, los investigadores emparejaron las cuentas de los usuarios en X y Wallapop, y mediante las APIs de ambas aplicaciones recopilaron datos de 19.325 usuarios. El modelo de predicción seleccionado fue el Modelo Lineal Generalizado (MLG), un modelo paramétrico de la familia Distribución Binomial Negativa. Este modelo utilizó el dataset de X como predictores y la variable Y como el ratio de malas revisiones por usuarios en Wallapop. Los resultados confirmaron que es posible predecir la probabilidad de que un usuario reciba una mala review en Wallapop basándose en su actividad en X. Además, se identificaron como predictores estadísticamente significativos el recuento de publicaciones, la hora punta para una publicación o el recuento de seguidores y amigos. Finalmente, tras entrenar el modelo con un método de validación cruzada con 10-folds, se obtuvo una AUC (area under the curve) de 9,83 % sobre la curva de ganacias, demostrando una capacidad predictiva suficiente.

Continuando con la exploración de diferentes métodos de análisis en el ámbito de la economía colaborativa, otro enfoque notable es el estudio realizado por (Kiatkawsin, Sutherland, y Kim, 2020), que se centra en las reseñas de Airbnb en dos de las ciudades más grandes del mundo: Hong Kong y Singapur. Utilizando un modelo no supervisado de aprendizaje conocido como Latent Dirichlet Allocation (LDA), el objetivo de su estudio era identificar los temas de conversación comunes en las reseñas de los alojamientos de Airbnb para así descubrir si existen asuntos únicos o muy representativos de cada destino. La principal característica del LDA es su capacidad para extraer temas de discusión latentes a partir de los textos, basándose en la co-ocurrencia de palabras. Este algoritmo se apoya en una matriz de documentos-términos para comprender la manera en la que las palabras se combinan y transmiten significados, resultando particularmente útil para analizar grandes conjuntos de datos textuales. Particularmente, en este estudio, se analizaron 185.695 reseñas de Hong Kong y 93.571 de Singapur. Los resultados del estudio muestran agrupaciones de las 30 palabras clave más frecuentes, organizadas en distintas categorías que reflejan aspectos clave de la experiencia de alojamiento. En Hong Kong, se identificaron 12 categorías, abarcando desde la accesibilidad, las condiciones del alojamiento y los servicios, hasta la localización, el vecindario, la gestión, la comunicación, la evaluación, el valor percibido y la satisfacción general de los usuarios. En contraste, Singapur, al ser una ciudad más compacta que Hong Kong y contar con un conjunto de datos más reducido, presentó un patrón simplificado con solo 5 categorías principales identificadas: apartamento, ubicación, gestión, anfitrión y evaluación. Este esquema más reducido sugiere diferencias en la forma en que los huéspedes

experimentan y valoran sus estancias en Singapur, posiblemente reflejando características únicas de la oferta de alojamiento y las expectativas culturales en este entorno urbano.

De manera similar, el estudio de (Zhang y Fu, 2020) profundiza en la experiencia de los usuarios de Airbnb en Beijing, China, diferenciando entre dos grupos distintos de usuarios: los locales y los extranjeros. Mediante el uso de estrategias de modelado de tópicos como LDA, los investigadores analizaron 180.452 reseñas recopiladas a través del sitio web Inside Airbnb. Las reseñas en chino se atribuyeron a usuarios domésticos, mientras que las escritas en inglés se consideraron de usuarios internacionales. Tras un proceso de tokenización, lematización y eliminación de palabras irrelevantes, se aplicó el algoritmo LDA, que reveló 10 categorías principales por idioma, basadas en las 15 palabras más recurrentes. Notablemente, 8 de estas categorías resultaron ser comunes para ambos grupos, destacando temas universales que impactan a todos los consumidores, tales como la ubicación, los servicios adicionales, la sensación de estar en casa, los procedimientos de *check-in* y *check-out*, la experiencia en general, el transporte, la decoración y la atención por parte del anfitrión. No obstante, las dos categorías restantes difirieron significativamente entre los dos perfiles de usuarios: para los extranjeros, lo más importante resultó ser la flexibilidad de la política de cancelación y las recomendaciones locales, elementos cruciales para quienes exploran una ciudad por primera vez. Por otro lado, para los usuarios chinos, las categorías adicionales se centraron en la limpieza y la posibilidad de regresar al mismo alojamiento en futuras visitas, indicando una valoración especial por el estilo y la decoración del alojamiento, así como la facilidad para realizar el *check-in* y *check-out*.

En el mismo campo de investigación, otro estudio realizado por (Sutherland y Kiatkawsin, 2020) se enfoca en analizar los temas de interés que influyen en la experiencia de los clientes dentro de la economía colaborativa de alojamientos, en particular en Airbnb. Este estudio se basa en un extenso dataset de reseñas de Airbnb de Nueva York, que comprende 1.086.000 reseñas, reflejando así la amplitud y diversidad del mercado de alojamiento en esta ciudad. La elección de Nueva York como objeto de estudio se debe no solo a la abundancia de datos disponibles, sino también a la diversa gama de invitados y anfitriones que participan en la plataforma. Mediante el uso del algoritmo LDA, se analizaron los datos de la plataforma de código abierto Inside Airbnb, identificando las categorías de discusión con base en las 50 palabras clave más relevantes. Los resultados del análisis revelaron un total de 43 temas distintos. De estos, 8 estaban relacionados con la evaluación global del alojamiento, 12 con la localización, 15 con aspectos tangibles e intangibles del apartamento y los 8 restantes se centraban en temas como la gestión y administración de los servicios. Finalmente, el empleo de un análisis jerárquico de clusters, específicamente el método de Ward, permitió deducir que existe una relación directa entre las experiencias emotivas y sociales de los usuarios. Por ejemplo, se destacó una conexión significativa entre el deseo de "volver a visitar la ciudad" y la "ubicación en el centro de la ciudad", una zona de alto interés turístico.

Un estudio paralelo se realizó en Boston, otra ciudad de gran relevancia en los Estados

Unidos, centrándose en su industria hotelera local (Lawani, Reed, Mark, y Zheng, 2019). Mediante la utilización de la plataforma Inside Airbnb para la recogida de datos, se buscó investigar de manera empírica si las reseñas generadas por los usuarios podrían actuar como una fuente de valor, donde las reseñas sirvieran como métricas de calidad para influir en las decisiones de los usuarios y, por lo tanto, impactar en los precios de los alojamientos. El estudio analizó 22.651 reseñas recolectadas a lo largo de 2016, aplicando técnicas de análisis de sentimientos al comparar el contenido con el diccionario AFINN, y empleando un modelo espacial de precios hedónicos para estimar el impacto de diversos factores en la fijación de precios. Los resultados indicaron que el precio de un alojamiento en Airbnb no solo se ve determinado por sus atributos intrínsecos y su localización, sino también por el precio de alojamientos competidores cercanos y otros factores de calidad. De manera específica, se encontró que un incremento en el número de habitaciones de un apartamento podía elevar su precio en un 24 %, mientras que la adición de un baño extra podría aumentar el precio en un 12 %. Asimismo, la proximidad al distrito financiero de Boston, una ciudad con un marcado perfil financiero y estudiantil, mostró tener un efecto significativo: una reducción del 1 % en la distancia al centro financiero implicaba un aumento del 0,09 % en el precio.

Siguiendo con el análisis de la utilidad de las reseñas en línea para la evaluación de servicios en la economía colaborativa, (Villeneuve y O'Brien, 2020) desarrollaron un enfoque interesante en su estudio. El objetivo era encontrar un método eficaz para evaluar la calidad del ambiente interior de las viviendas temporales de alojamiento, como las ofrecidas por Airbnb. Reconociendo las dificultades y costos asociados con la realización de encuestas post-ocupación, optaron por un análisis basado en el análisis automático de las revisiones dejadas por los arrendatarios. Para llevar a cabo este análisis, utilizaron un conjunto de datos compuesto por 1,35 millones de reseñas de Airbnb provenientes de varias ciudades canadienses, como Victoria, Vancouver, Ottawa, Toronto y Montreal, recogidas entre junio de 2009 y marzo de 2019. La diversidad de estas ciudades era crucial para la investigación, ya que permitía una evaluación más amplia y representativa de la calidad ambiental a través de la realimentación de los usuarios. Al procesar los datos, se eliminaron las palabras que no eran en inglés, simplificando así el análisis en un país bilingüe donde eran comunes las palabras en francés, así como palabras con acentos y caracteres especiales. En la fase de análisis de sentimientos, los investigadores desarrollaron un diccionario específico compuesto por términos asociados a condiciones ambientales deficientes y expresiones de insatisfacción, tales como olor ("smell"), polvo ("dusty") , ruido ("noise"), helado ("freezing") y cálido ("warm") entre otros. Se optó por excluir las quejas irrelevantes o aquellas de connotación positiva, ya que no aportaban información relevante para el estudio. El análisis reveló que las reseñas que reflejaban indicadores de baja calidad del ambiente interior presentaban, de forma estadísticamente significativa, puntuaciones de sentimiento más bajas en comparación con aquellas que no mencionaban estos aspectos. Este resultado indica que las percepciones negativas sobre el ambiente interior de un alojamiento impactan notablemente en la impresión general de

los huéspedes. Este resultado enfatiza la relevancia de la calidad ambiental en la percepción del cliente y en la imagen que proyecta el alojamiento, resaltando su importancia en la gestión y reputación de estas propiedades.

Por su parte, el estudio realizado por (Rensing, 2022) aborda una premisa innovadora: la posibilidad de identificar las necesidades y generar nuevas ideas a partir del análisis de publicaciones en X, utilizando técnicas avanzadas de machine learning (ML). Este enfoque se centra no solo en discernir la relevancia del contenido de los mensajes, sino también en determinar su capacidad para ser detectados automáticamente mediante estas algoritmos automáticos. Tras un cuidadoso proceso de preprocesamiento, (Rensing, 2022) seleccionó una muestra representativa de 10.000 publicaciones para aplicar el método BERT (Bidirectional Encoder Representations from Transformers). Este algoritmo, conocido por su habilidad para entender el contexto de una palabra basándose en las palabras circundantes, clasifica los textos según la probabilidad de que contengan información relevante sobre las necesidades de los usuarios. Aquellos con una probabilidad superior al 50 % se marcan con un 1, indicando relevancia, mientras que los marcados con un 0 se consideran no relevantes. Del análisis resultaron 831 publicaciones clasificadas, a partir de las cuales se extrajeron diversas conclusiones. Mediante el proceso de clasificación, se logró segmentar estas publicaciones en diferentes categorías de interés, como apoyo al usuario, problemas técnicos, seguridad, transparencia e ideas innovadoras. La precisión del algoritmo BERT, estimada en un 0,95, demuestra su eficacia para identificar necesidades e ideas útiles no solo para usuarios de Airbnb, sino también aplicables a otros campos de estudio y plataformas que interactúan con X .Sin embargo, los autores señalan una consideración importante: los datos de la red social X no constituyen una muestra aleatoria y, para obtener conclusiones más amplias y robustas, sería necesario incrementar la interacción y diversificar las fuentes de datos.

Alineado al desarrollo de la economía colaborativa, surge el concepto de economía circular, que promueve la reutilización y el reciclaje de recursos en los sistemas de producción y consumo. Esta filosofía, centrada en la sostenibilidad y la minimización de residuos, encuentra similitudes en prácticas como el uso de viviendas pertenecientes a terceros y el consumo de productos de segunda mano, aspectos también fundamentales en la economía colaborativa (Curtis y Lehner, 2019). En este contexto, (De Lima, 2022) investigan el sentimiento hacia la economía circular en X y analizan las interacciones y discusiones entre usuarios relacionadas con este tema. Para este propósito, los autores implementan una metodología de análisis triple sobre un dataset de 63.590 publicaciones de la red social X recopilados a través de su API. Tras un proceso inicial de preparación y procesamiento de los datos, el estudio se enfoca en tres ejes de análisis: la exploración de hashtags y palabras clave, la búsqueda de términos específicos relacionados con la economía circular, y un análisis de sentimientos para determinar la polaridad emocional de los textos. Los resultados revelaron una prevalencia de publicaciones con un sentimiento ligeramente positivo, indicando una tendencia general en la sociedad hacia la adopción de prácticas más sostenibles y alineadas con los principios

de la economía circular.

Al concluir esta revisión de la literatura, es importante destacar la profundidad de las investigaciones realizadas para entender la percepción de los usuarios en la industria hotelera, en especial dentro del marco de la economía colaborativa y plataformas como Airbnb. La minería de texto, mediante técnicas como el modelado de tópicos y el análisis de sentimientos, ha surgido como un enfoque valioso para identificar las categorías de discusión y los sentimientos subyacentes expresados por los usuarios. Entre las metodologías más utilizadas para el modelado de tópicos destacan LDA y BERT, mientras que para el análisis de sentimientos se han aplicado técnicas basadas en diccionarios como AFINN, así como diversas estrategias de ML tradicionales. LDA es especialmente reconocido por su capacidad para identificar temas latentes dentro de grandes conjuntos de datos textuales, permitiendo una comprensión estructurada de las conversaciones y opiniones compartidas. Esta técnica se beneficia de su eficiencia en la agrupación de palabras co-ocurrentes en temas significativos, facilitando el análisis temático de las reseñas. Por otro lado, las técnicas basadas en diccionarios, como el diccionario AFINN, sobresalen en el análisis de sentimientos por su simplicidad y precisión al calificar la polaridad de los textos, aprovechando listas predefinidas de palabras con valores sentimentales asignados.

De acuerdo con los resultados presentados en esta revisión, se planea emplear LDA como la técnica principal de modelado de tópicos, debido a su probada eficacia y versatilidad en la extracción de temas relevantes de los datos. Asimismo, se contempla la utilización de diccionarios con reglas particulares adaptadas a los datos de redes sociales, como VADER (*Valence Aware Dictionary and Sentiment Reasoner*), para el análisis de sentimientos. VADER, en particular, destaca por su capacidad para capturar el matiz de los sentimientos expresados en el lenguaje de las redes sociales, incluyendo emoticonos y abreviaturas, lo que lo hace particularmente adecuado para el análisis de texto en plataformas como X.

Para concluir, resulta importante destacar que, en los estudios revisados, se han aplicado diversas estrategias de filtrado con el objetivo de identificar categorías temáticas que aporten claridad al objeto de estudio. De particular relevancia son las categorías identificadas en las investigaciones, como “limpieza”, “ubicación”, “seguridad”, “interacción con el anfitrión” y “precio”, que, de forma recurrente, aparecen en todos los análisis. Esto sugiere que, a través de distintos contextos de la economía colaborativa, los usuarios comparten preocupaciones y sugerencias similares, mostrando preocupaciones similares en la experiencia de alojamiento. La recurrencia de estas categorías en múltiples estudios destaca el consenso entre los usuarios sobre los aspectos más valorados y críticos de los servicios de alojamiento colaborativo, lo que sugiere la importancia de abordar estas áreas para mejorar la satisfacción del cliente y potenciar la calidad del servicio ofrecido. Basándose en las investigaciones analizadas previamente, la Tabla 2.1 proporciona un resumen de los estudios revisados. Esta tabla incluye detalles sobre el método de recopilación de datos empleado en cada estudio, los modelos de análisis implementados y las principales conclusiones derivadas de cada investigación.

Tabla 2.1: Resumen de los estudios sobre percepción pública llevados a cabo mediante el análisis de Tweets.

Referencia	Tamaño del Dataset	Método de Adquisición	Objetivo	Algoritmo	Resultados
(Chang y Wang, 2018)	200 revisiones	Filtrado de revisiones desde enero de 2017, 2 páginas web de Estados Unidos: Airbnb y HomeAway. 20 reseñas más recientes de los 10 alojamientos más populares.	Analizar el grado de satisfacción del anfitrión, la limpieza del alojamiento, la localización y el grado de satisfacción con la experiencia.	Análisis de sentimientos: LIWC y ratio de sentimiento.	La Generación Z se guía principalmente por la primera impresión, sin verse influenciada por el ratio de sentimiento. La Generación Y prioriza la limpieza, el valor, la precisión y la calificación global al tomar decisiones. Por otro lado, la Generación X enfatiza la importancia de la limpieza y la calificación global, prestando poca atención a las imágenes en las reseñas.
(Lawani et al., 2019)	22.651 revisiones	Datos obtenidos de Inside Airbnb de 2016. Cada reseña incluía: características del alojamiento, coordenadas geográficas, precio por noche y reseñas previas de inquilinos.	Investigar empíricamente el impacto del contenido generado por los usuarios como una fuente de valor significativa, evaluando el impacto de las reseñas en las decisiones de los usuarios.	Análisis de sentimientos mediante el diccionario AFFIN y Modelo espacial de precios hedónicos.	El precio de una habitación en la plataforma no solo depende de sus características intrínsecas y su ubicación, sino también del precio de sus competidores en el vecindario y otros atributos de calidad. Por ejemplo, el número de habitaciones de un apartamento aumenta el precio en un 24 %, en contraste con el 12 % que incrementa el precio si el piso cuenta con un baño adicional.
(Prada y Iglesias, 2020)	19.325 revisiones	A través de la API de X. Filtrado por publicaciones desde Diciembre de 2015 que estuviesen relacionadas con Wallapoop.	Analizar el comportamiento de los usuarios de Wallapop en X y su influencia para predecir futuros comportamientos de nuevos usuarios de la plataforma.	MLG.	El estudio confirmó que es posible predecir la probabilidad de que un usuario reciba una mala review en Wallapop basándose en su actividad en X. Predictores significativos estadisticamente: recuento de publicaciones, la hora punta para una publicación o el recuento de seguidores y amigos.
(Kiatkawsin et al., 2020)	185.695 *revisiones* y 93.571 *revisiones*	Inside Airbnb. Datos recopilados desde abril de 2020.	Analizar las reseñas de alojamientos en la plataforma en las ciudades de Hong Kong y Singapur, y modelar las categorías de discusión con el fin de identificar aspectos únicos asociados a cada destino.	Modelado de tópicos: LDA.	Categorías Identificadas: 12 para la ciudad de Hong Kong, incluyendo accesibilidad, comunicación y evaluación, y 5 para Singapur, destacando apartamento y ubicación.
(Sutherland y Kiatkawsin, 2020)	1.086.000 revisiones	Inside Airbnb, filtrando por localización y por usuario.	Analizar los temas de interés que impulsan la experiencia de un cliente dentro de la economía colaborativa de los alojamientos.	Modelado de tópicos: LDA y Cluster Jerárquico de Ward.	43 categorías identificadas: 8 relacionadass con la evaluación global del alojamiento, 12 con la localización, 15 con aspectos tangibles e intangibles del apartamento y los 8 restantes se centraban en temas como la gestión y administración de los servicios.

Referencia	Tamaño del Dataset	Método de Adquisición	Objetivo	Algoritmo	Resultados
(Villeneuve y O'Brien, 2020)	1,35 millones de revisiones	Entre junio de 2009 y marzo de 2019 a través Inside Airbnb, Ciudades canadienses: Victoria, Vancouver, Toronto, Ottawa, Montreal y Quebec. Reseñas en francés.	Evaluar la calidad del ambiente interior (IEQ) de las viviendas temporales de alojamiento.	Análisis de sentimientos: Diccionarios.	Aquellas revisiones que contienen índices de calidad del ambiente interior (IEQ) negativos tienen una puntuación de sentimiento estadísticamente más baja que aquellas que no los tienen, indicando que un mal comentario sobre la calidad del ambiente interior de un alojamiento reduce significativamente la impresión general de los inquilinos.
(Zhang y Fu, 2020)	180.452 reseñas	Inside Airbnb desde abril de 2019.	Analizar las reseñas de usuarios de Airbnb en Beijing, China, con el fin de descubrir y comparar los temas clave que resaltan tanto de los usuarios locales como de los extranjeros.	Modelado de Tópicos: LDA.	Entre los dos grupos hay 8 categorías comunes y 2 diferentes. Los usuarios extranjeros priorizan las políticas de flexibilidad y valoran las recomendaciones locales, mientras que los usuarios nacionales se concentran más en la comodidad, la limpieza y la posibilidad de volver a alquilar el apartamento.
(Rensing, 2022)	10.000 publicaciones	API de X: desde febrero de 2021 hasta enero de 2022. Publicaciones en inglés, mayores de 50 caracteres y temas relacionados con Airbnb.	Analizar las interacciones de los usuarios en la cuenta de soporte de Airbnb en X con el fin de identificar las principales necesidades de los clientes mediante el análisis de tweets.	Técnicas de machine learning: BERT.	Con una precisión de 0,95, el modelo BERT aplicado a X es de gran utilidad a la hora de identificar necesidades e ideas de los usuarios de Airbnb.
(De Lima, 2022)	63.590 publicaciones	API de X: desde julio de 2021 hasta octubre de 2021. Palabras clave como economía circular. Publicaciones en inglés.	Evaluar el sentimiento expresado en las publicaciones de X relacionadas con la economía circular e identificar las principales temáticas que impulsan la interacción y el debate entre los usuarios sobre este tema.	Análisis de *hashtags* y de palabras clave; análisis de las publicaciones con *Lexical Search* ; análisis de sentimientos con el software MAXQDA.	Prevalencia de publicaciones con un sentimiento ligeramente positivo, indicando una tendencia general en la sociedad hacia la adopción de prácticas más sostenibles y alineadas con los principios de la economía circular.

Capítulo 3

Enfoque Metodológico en el Análisis de Textos

Este capítulo se dedica a detallar la metodología adoptada para la realización de este Trabajo de Fin de Grado, cuyo propósito es investigar las percepciones de los usuarios respecto a la industria de la hostelería a través de las redes sociales. Tal como se visualiza en la Figura 3.1, el enfoque metodológico se organiza en varias etapas, cada una compuesta por distintas actividades clave.

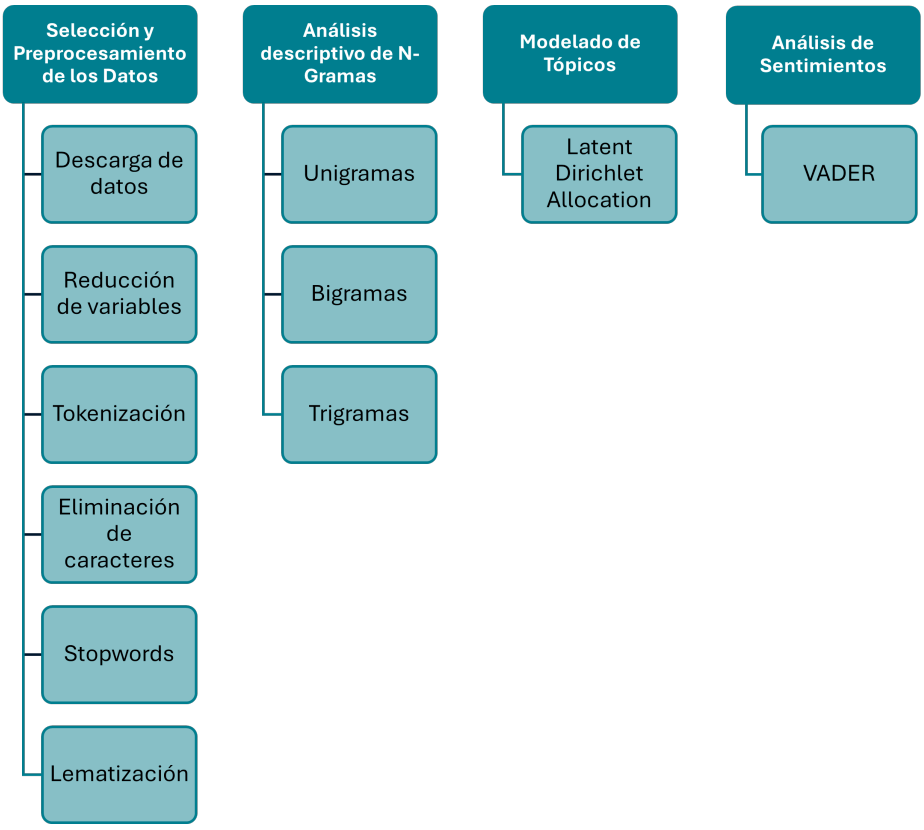

Figura 3.1: Metodología de análisis implementada. Elaboración propia

En la primera etapa, **Selección y Preprocesamiento de datos**, se llevará a cabo la recopilación de los datos necesarios para el análisis, seguida de su adecuada preparación. Este procesamiento incluye la limpieza y organización de los datos para garantizar su calidad y fiabilidad para los análisis subsiguientes. La segunda fase, denominada **Análisis descriptivo de N-gramas**, se centra en la evaluación de las palabras y secuencias de palabras (N-gramas) más significativas dentro del conjunto de datos. Para ello, se empleará la técnica de *Term Frequency - Inverse Document Frequency* (TF-IDF), una métrica que permite identificar la importancia de cada palabra en el contexto del corpus analizado. Esta aproximación permitirá identificar, en una primera instancia, las temáticas y términos claves que predominan en general en las conversaciones de los usuarios sobre la hostelería en las redes sociales.

Posteriormente, en la sección dedicada al **Modelado de Tópicos**, se abordará la implementación del método *Latent Dirichlet Allocation* (LDA) usado para descubrir los temas principales que emergen de las discusiones de los usuarios. Este enfoque permite identificar de forma automática los tópicos de discusión predominantes en grandes volúmenes de texto, revelando de formá más detallada las principales áreas de interés y preocupación entre los participantes. Finalmente, en la última etapa, **Análisis de Sentimientos**, se explicará el procedimiento adoptado para la detección y evaluación del sentimiento subyacente en las publicaciones de los usuarios. Este análisis se centrará en discernir las emociones positivas, negativas o neutrales expresadas en los comentarios, lo que facilitará una comprensión integral de las actitudes de los usuarios hacia la industria de la hostelería en el entorno de las redes sociales.

3.1. Selección y Pre-procesamiento de los Datos

La primera etapa de la metodología abarca tanto la recopilación como la preparación de los datos para su análisis. En cuanto a la recolección de datos de la red social X, anteriormente conocida como Twitter, es importante tener en cuenta ciertas restricciones impuestas tras su transformación en julio de 2023. Este cambio no solo incluyó una modificación en su logo y nombre, sino que también afectó diversas funcionalidades disponibles para los usuarios (Blanco, 2023). Un ejemplo notable es la alteración en las políticas de descarga de publicaciones mediante la API, que anteriormente estaba disponible gratuitamente para investigadores con cuentas de desarrollador. Dado que la obtención directa de datos de X puede resultar significativamente costosa bajo las nuevas condiciones, en este trabajo se optó por acceder al conjunto de datos a través de Mendeley Data. Este repositorio de acceso abierto proporciona una alternativa para adquirir los datos necesarios para el estudio, garantizando así la viabilidad de la investigación en el contexto de las recientes modificaciones en la plataforma. Para satisfacer los requisitos específicos establecidos en los objetivos de este trabajo, se llevó a cabo una búsqueda de diversos conjuntos de datos. Esta búsqueda estuvo guiada por criterios

como el volumen de datos y las palabras clave asociadas a su recolección. Finalmente, se seleccionó un conjunto de datos compuesto por 21.097 publicaciones, recopiladas desde marzo de 2019 hasta octubre de 2019, empleando el hashtag Airbnb como filtro principal. No obstante, es importante destacar que la selección se limitó a aquellas publicaciones redactadas en inglés, lo que resultó en una muestra depurada de 13.902 publicaciones (Teh, 2021). Esta delimitación es necesaria para mejorar la precisión del análisis, dado que el idioma uniforma la base para un procesamiento y posterior comparación efectiva de los datos.

Con el fin de mejorar la calidad de los datos recolectados para este estudio, se implementó un proceso de limpieza y pre-procesamiento (Uysal y Gunal, 2014). Como paso inicial en este proceso, se creó una tabla que detalla las variables recopiladas, su tipología y un breve texto informativo, facilitando así la realización de un análisis descriptivo preliminar de los datos. Este conjunto de datos incluye tres variables relevantes: el texto de la publicación, la fecha de la misma y la ubicación del usuario, tal como se detalla en la Tabla 3.1.

Nombre de la Variable	Descripción	Tipo de Variable
Date	Fecha de la publicación	Fecha
Tweet Text	Contenido de la publicación original	Carácter
Location (Continent)	Ubicación del usuario de X	Carácter

Tabla 3.1: Variables seleccionadas. Elaboración propia.

Partiendo de este punto, se procede al pre-procesamiento de datos,una etapa fundamental que implica la limpieza de la información recopilada para eliminar elementos irrelevantes que no aportan significado al análisis. El primer paso en este proceso es la tokenización del texto. Esta técnica descompone el texto en unidades menores denominadas tokens, que pueden ser palabras, símbolos, números o caracteres. La tokenización es necesaria, ya que transforma los datos en una serie de palabras individuales, facilitando así su análisis posterior. Por ejemplo, al aplicar la tokenización a la frase “El Airbnb era muy acogedor”, se obtendría una lista de tokens como “El”, “Airbnb”, “era”, “muy” y “acogedor”. Este enfoque no solo simplifica el manejo de los datos, sino que también los estructura para etapas analíticas más avanzadas, permitiendo un tratamiento más eficiente del conjunto de datos.

Tras completar la etapa inicial de tokenización, se procede a la segunda fase del preprocesamiento, que consiste en la normalización y limpieza del texto. Esta etapa implica convertir todas las letras a minúsculas y eliminar caracteres especiales que no contribuyen información valiosa al análisis. Elementos como signos de puntuación, caracteres especiales (@, &, %, etc.), números, enlaces, menciones, URLs, hashtags y emoticonos son excluidos para depurar el texto. La siguiente fase del proceso se enfoca en la eliminación de *stopwords*, es decir, palabras comunes en el idioma que, pese a su frecuencia, no aportan significado relevante al contexto del análisis. Esto incluye preposiciones, conjunciones y artículos como “de”, “la”, “y” en español, o “the”,“and”,“of” en inglés. Para realizar esta limpieza se recurre a

diccionarios específicos del idioma que contienen un listado completo de *stopwords*.

La penúltima fase del procesamiento de datos es la lematización de las palabras. Este proceso implica reducir las palabras a su forma base o lema, eliminando variaciones morfológicas como conjugaciones verbales, morfemas de género y número. De este modo, palabras como “casita”, “casa”, “casucha”, “casero” y “casona” se simplifican a su lema común “casa”, facilitando así un análisis más uniforme. La lematización permite agrupar variantes de una misma palabra, mejorando la precisión del análisis al tratar variaciones léxicas como una única entidad. Para finalizar la fase de preprocesamiento, es importante verificar y eliminar cualquier publicación duplicada. La presencia de entradas repetidas en el conjunto de datos podría afectar adversamente el análisis, potencialmente provocando sobreajuste (*overfitting*) en los modelos o introduciendo sesgos en los resultados.

Es importante destacar que el proceso previamente descrito de preparación y preprocesamiento de datos textuales es necesario para el análisis de N-Gramas y el modelado de tópicos, asegurando la focalización del análisis en términos con un peso semántico significativo. Este enfoque no solo mejora la calidad de los resultados, sino que también promueve una interpretación más precisa del conjunto de datos, eliminando información irrelevante y redundante. Sin embargo, cuando se aborda el análisis de sentimientos, el número de etapas de preprocesamiento se reduce. Se conservan mayúsculas, minúsculas, signos de puntuación y *stopwords*, ya que estos elementos proporcionan los matices necesarios para una evaluación precisa de las emociones expresadas en los textos.

3.2. Análisis descriptivo de N-Gramas

Una vez las publicaciones han sido pre-procesadas, se lleva a cabo el análisis descriptivo de N-Gramas, cuyo propósito es identificar los tokens y las secuencias que destacan por su relevancia en el conjunto de datos. Los N-Gramas se caracterizan por ser agrupaciones de N tokens consecutivos dentro de una secuencia textual, siendo N cualquier número natural. Por ejemplo, el análisis de uni-gramas se centra en identificar palabras individuales y determinar su importancia en el corpus. En el caso de los bi-gramas, se examinan pares de palabras consecutivas junto con su valor de relevancia. Para los tri-gramas, el foco está en las secuencias de tres palabras y su importancia, y así sucesivamente.

Por estas razones, el primer análisis descriptivo de N-Gramas es comúnmente utilizado en el procesamiento de texto. Este enfoque se adopta como un paso inicial, que precede tanto al modelado de tópicos como al análisis de sentimientos. Su propósito es descubrir las ideas centrales y detectar patrones y tendencias predominantes en un texto o conjunto de publicaciones. Sin embargo, la observación de la frecuencia con que aparece un N-Grama en una publicación individual no brinda información completa de su significado o relevancia. La verdadera utilidad surge al analizar la importancia de estos elementos dentro del contex-

to global del corpus. En consecuencia, este estudio implementa el cálculo del TF-IDF una métrica que proporciona una evaluación más precisa de la relevancia de los N-Gramas en todo el conjunto de datos (Yun-tao, Ling, y Yong-cheng, 2005). Este método no solo realza los términos significativos en relación con el documento individual, sino que también pondera su importancia en el contexto más amplio del corpus, facilitando así un análisis más significativo.

La métrica TF-IDF se determina mediante la multiplicación de dos componentes clave, tal y como se ilustra en la siguiente expresión:

$$\text{TF-IDF}(i) = \text{TF}(i) \times \text{IDF}(i) \tag{3.1}$$

donde TF (*Term Frequency)* representa la relevancia de un término específico dentro de un documento, refiriéndose a la cantidad de veces que dicho término se manifiesta en el documento. Como una medida de frecuencia relativa, se calcula de la siguiente manera:

$$\text{TF}(i) = \frac{\text{Frecuencia absoluta de la palabra } i \text{ en el corpus}}{\text{Número total de palabras en el corpus}}. \tag{3.2}$$

Por otro lado, la métrica IDF (*Inverse Document Frequency*) refleja la importancia de un término a lo largo de todo el corpus, es decir, su grado de unicidad o rareza. Se determina mediante el logaritmo natural con base en el número *e* de la proporción entre el número total de documentos en el corpus y el número de documentos que incluyen el término *i*, tal como se muestra en la siguiente ecuación :

$$\text{IDF}(i) = log_e \frac{\text{Número total de documentos en el corpus}}{\text{Número de documentos que incluyen la palabra } i}. \tag{3.3}$$

La métrica TF-IDF es de gran utilidad en el análisis de tokens y palabras dentro de un corpus, proporcionando una puntuación que destaca la relevancia de cada palabra en un documento específico. Esta puntuación se deriva de la multiplicación ponderada, reflejando la importancia relativa de una palabra: cuanto más frecuente es una palabra en un documento en particular, y a la vez menos frecuente en el corpus en general, mayor será su puntuación TF-IDF. Esto significa que los términos que son comunes en un documento pero raros en otros recibirán una valoración alta, señalando su potencial relevancia temática única para ese documento. En contraste, las palabras que son comunes a muchos documentos reciben una puntuación baja en TF-IDF, ya que su presencia generalizada las hace menos significativas para identificar temas o contenidos específicos. En pocas palabras, la métrica TF-IDF favorece los términos que ofrecen una contribución distintiva a la especificidad de un documento dentro del conjunto de datos analizado.

3.3. Modelado de Tópicos

El proceso de modelado de tópicos busca identificar y comprender los principales temas de discusión presentes en un texto, lo cual facilita una interpretación más detallada del significado contextual de las palabras dentro de categorías determinadas. Para este propósito, en el presente trabajo se ha seleccionado la técnica de *Latent Dirichlet Allocation* (LDA) por su reconocida eficacia y las ventajas significativas que brinda en el ámbito de la investigación. Tal como se discutió en el Capítulo 2, una amplia gama de estudios académicos previamente revisados han adoptado LDA como su metodología de análisis. Esta preferencia se debe a la notable capacidad de LDA para procesar y analizar grandes cantidades de datos no estructurados, además de su caracter escalable y flexible.

LDA es una técnica de aprendizaje no supervisado de ML que permite identificar tópicos o temas "latentes" basándose en modelos bayesianos (D. M. Blei, Ng, y Jordan, 2003). Al ser una técnica de aprendizaje no supervisado, no es necesario entrenar previamente el modelo con documentos previamente etiquetados. LDA es particularmente eficaz para descubrir patrones y tópicos ocultos dentro de grandes volúmenes de texto, ya que modela cada tópico como un agrupamiento de palabras con características similares. Además, esta técnica asume que los tópicos se distribuyen de manera probabilística a lo largo de los documentos, permitiendo que un único documento pueda ser clasificado bajo múltiples categorías. Para lograr esto, LDA asigna a cada palabra dentro de un documento una probabilidad que indica la categoría a la cual es más probable que pertenezca, en base a la distribución de tópicos observada en el conjunto de datos.

La Figura 3.2 ofrece una representación visual del proceso de modelado de tópicos utilizando LDA, en la cual se representan los siguientes términos:

- M: Número total de documentos dentro del corpus.
- N: Número de palabras en un documento.
- W: Cada palabra observada del documento M.
- α: Primer parámetro de Dirichlet para establecer la distribución de tópicos por documento.
- η: Segundo parámetro de Dirichlet para establecer la distribución de palabras por tópico.
- θ: Se extrae de una muestra a priori de Dirichlet, parametrizada por α.
- β: Se extrae de una muestra a priori de Dirichlet, parametrizada por η.
- z: Tópico asignado por palabra del documento M.

- k: Número óptimo de tópicos.

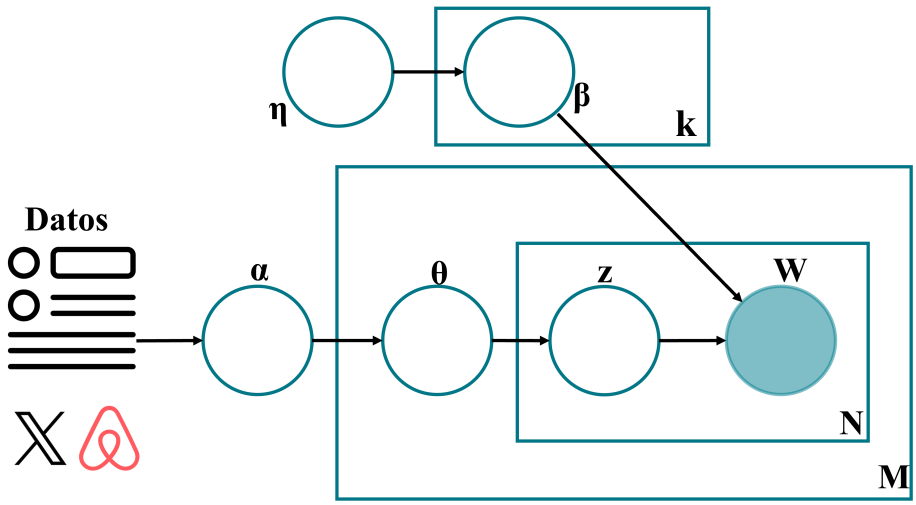

Figura 3.2: Descripción del modelo LDA. Elaboración propia

En resumen, como se describe en la figura, el modelo LDA se estructura en tres niveles principales de representación. M simboliza el conjunto de datos sujeto a análisis, que se compone de N palabras, indicadas por la variable W. Los parámetros α y η permiten modelar las distribuciones de probabilidad relacionadas con el número de tópicos por documento y el número de palabras por tópico, respectivamente. El valor de k se refiere al número óptimo de tópicos a identificar, mientras que β describe cómo se distribuyen las palabras dentro de cada tópico específico. La variable Z se encarga de determinar la asignación de cada palabra a un tópico particular, y θ representa la distribución de los tópicos identificados en cada documento del conjunto de datos. Este esquema destaca el funcionamiento interno del LDA, ofreciendo un marco conceptual para entender la forma en la que el modelo caracteriza las relaciones entre las palabras, los tópicos y los documentos. En el proceso de construcción del modelo LDA, se parte del supuesto de que todos los documentos están compuestos por una combinación de diversos temas, y que cada tema, a su vez, se define por una distribución específica de palabras. El objetivo de LDA es descubrir precisamente estas distribuciones tanto de temas como de palabras. Se considera que cada documento en el corpus está vinculado a una distribución única de temas, mientras que cada tema se asocia a su propia distribución de palabras. Para lograr identificar estas distribuciones, el modelo LDA se apoya en técnicas de inferencia estadística, permitiendo estimar de manera efectiva la distribución de temas presentes en cada documento y la distribución de palabras característica de cada tema. (D. Blei, Ng, y Jordan, 2001).

Un aspecto fundamental en la implementación del modelo LDA es determinar la cantidad de temas *k*, en los que se clasificará el conjunto de documentos, teniendo en cuenta que LDA asume por defecto la existencia de múltiples categorías. Esta elección definirá la eficacia del

modelo, pues un valor de *k* excesivamente alto puede conducir a un sobreajuste (*overfitting*), donde se crean categorías innecesarias que fragmentan en exceso la información. En contraste, un valor de *k* demasiado bajo puede resultar en un infraajuste (*underfitting*), agrupando de manera incorrecta las palabras en categorías demasiado amplias y vagas. Por lo tanto, es necesario identificar el número óptimo de temas para mantener un equilibrio que prevenga tanto el sobreajuste como el infraajuste, garantizando así que las categorías generadas sean significativas y reflejen de manera precisa la estructura temática del corpus.

Para determinar el número óptimo de tópicos (*k*), se requiere calcular el valor de coherencia, que proporciona una medida de la similitud semántica entre las palabras más frecuentes dentro de cada tópico (Röder, Both, y Hinneburg, 2015). Un valor de coherencia bajo sugiere una escasa relación semántica entre las palabras agrupadas bajo un mismo tópico, mientras que un valor alto indica una fuerte conexión semántica entre ellas. La selección adecuada del número de tópicos es determinante para alcanzar resultados significativos. Con este objetivo, se elaboran múltiples modelos variando el número de tópicos (*k*) y se mide el valor de coherencia de cada uno. El modelo con el valor de coherencia más elevado se selecciona como indicativo del número óptimo de tópicos. Esta metodología asegura que los tópicos identificados sean tanto coherentes como representativos de conceptos claramente diferenciados, optimizando la calidad y la relevancia de los resultados obtenidos del análisis. Además, como un enfoque complementario, se calcula la distancia intertópica, una métrica que evalúa la disimilitud entre los tópicos identificados en el texto. Esta métrica facilita la representación de los tópicos en un espacio bidimensional de forma circular, donde el tamaño de cada área tópica es proporcional al volumen de palabras que lo componen (Sievert y Shirley, 2014). El objetivo de este análisis es conseguir una distribución espacial de tópicos que evite superposiciones, lo cual permite una visualización más clara de la distinción e interpretabilidad de los tópicos identificados.

En conclusión, el propósito del LDA es modelar la distribución de probabilidad que determina la pertenencia de una palabra a un tópico específico, así como la asociación de ese tópico con un documento particular. Los resultados de este análisis permiten identificar un conjunto de tópicos distintivos del corpus, los cuales se caracterizan por agrupar las palabras con la mayor probabilidad de asociación a cada categoría. Para complementar este análisis, se recurre a herramientas como el índice de coherencia y la distancia intertópica. Estas métricas no solo contribuyen a la selección del número óptimo de tópicos y a la visualización de sus relaciones espaciales, sino que también son necesarias para validar y asegurar la coherencia de los tópicos identificados.

3.4. Análisis de Sentimientos

El análisis de sentimientos es una técnica utilizada en el PLN para determinar la actitud emocional subyacente en un texto. Su objetivo es identificar y categorizar las emociones expresadas en un conjunto de datos y clasificarlo como positivo, negativo o neutro. En este Trabajo de Fin de Grado, el análisis de sentimientos facilita la comprensión de cómo los usuarios de X perciben la industria hotelera y el fenómeno de la economía colaborativa. El propósito es examinar las reacciones y actitudes de la ciudadanía hacia plataformas como Airbnb, explorando sus sentimientos respecto al uso de estas aplicaciones en el contexto de la hostelería moderna.

Como se discutió previamente en el Capítulo 2, el análisis de sentimientos se aplica ampliamente en diversas áreas y disciplinas. La estrategia predominante involucra el uso de diccionarios de sentimientos, los cuales consisten en conjuntos de palabras y frases previamente evaluadas y clasificadas según el sentimiento o emoción que representan. Existe una variedad de diccionarios especializados diseñados para esta finalidad, y la elección de uno sobre otro suele depender del ámbito específico del estudio. En este contexto, para el presente Trabajo de Fin de Grado, se ha seleccionado el diccionario VADER (*Valence Aware Dictionary and Sentiment Reasoner*) debido a su adecuación y eficacia para analizar las percepciones y emociones de los usuarios de X. VADER se destaca por su popularidad en el análisis de texto proveniente de redes sociales y comentarios en línea, atribuible a su habilidad para interpretar el lenguaje coloquial y las expresiones idiomáticas (Hutto y Gilbert, 2014). Esta capacidad de VADER se debe a su diseño orientado a analizar no solo el contenido textual, sino también elementos adicionales que son frecuentes en la comunicación digital. Estos incluyen el uso de caracteres especiales, signos de puntuación y emoticonos, entre otros, que en el contexto de las redes sociales, determinan la expresión de emociones y sentimientos. La incorporación de estos elementos en el análisis permite que VADER proporcione una comprensión matizada de las percepciones expresadas por los usuarios en sus interacciones online, especialmente en plataformas como Airbnb, donde el tono y la intención emocional son indicativos clave de la satisfacción o insatisfacción del usuario con su experiencia.

VADER permite analizar la polaridad de un texto, así como su respectiva intensidad, asignando a cada palabra una puntuación que varía entre -4 y +4. Los valores extremos, +4 y -4, representan las emociones más positivas y negativas, respectivamente, mientras que 0 indica neutralidad. Una característica distintiva de VADER, en comparación con otros diccionarios de sentimientos, es su habilidad para interpretar aspectos textuales específicos, tales como la puntuación y el uso de mayúsculas. De manera significativa, VADER ajusta la polaridad de un mensaje si este se encuentra completamente en mayúsculas, interpretando dicha expresión como un intensificador de la emoción expresada. Además, VADER permite identificar la negatividad en las palabras y reconocer cambios de polaridad que ocurren por el uso de conjunciones como “pero” o “sin embargo”, así como el efecto de adverbios intensificado-

res como “extremadamente” o “muy”. Esta capacidad se extiende al análisis de negaciones, sarcasmo y el empleo de emoticonos, lo cual lo convierte en una herramienta particularmente adaptada al dinámico lenguaje de las redes sociales y los comentarios en línea (Hutto y Gilbert, 2014).

Para finalizar, se calcula la puntuación global de cada publicación con el fin de facilitar una comparación entre ellas. Este proceso involucra sumar todas las puntuaciones individuales asignadas a las palabras presentes en las publicaciones, y luego normalizar el resultado para que oscile dentro de un rango de -1 a 1. Utilizando el promedio de estas puntuaciones como indicador, se logra obtener una medida numérica que captura la valoración emocional global inherente a cada texto. Esta metodología ha probado ser eficaz en diversos ámbitos, abarcando desde el análisis de contenido en redes sociales hasta la evaluación de diversas reseñas de productos (Pano y Kashef, 2020).

En este contexto, es importante tener en cuenta que para obtener resultados significativos y realizar un análisis profundo con VADER, no se requiere un preprocesamiento de datos como el mencionado en el apartado 3.1 de este capítulo. De hecho, no es necesario normalizar el texto, eliminar puntuación o descartar *stopwords* para preservar todos los matices y sutilezas del lenguaje que son relevantes para interpretar correctamente las emociones. VADER está diseñado para analizar el texto tal y como se presenta, aprovechando elementos como la puntuación y el uso de mayúsculas para enriquecer el análisis de sentimientos. Solo se excluirán aquellos elementos que carecen de valor emocional, tales como URLs, menciones o cifras numéricas, asegurando así que se mantenga la integridad del mensaje y su carga emocional.

Capítulo 4

Resultados

En este capítulo, se presentan los resultados que fueron obtenidos tras la implementación de la metodología descrita en el Capítulo 3. El código utilizado para el análisis de las publicaciones en X fue subido a un repositorio en GitHub, donde se ha hecho accesible públicamente para su revisión (Bandeira, 2024). Para la etapa inicial de procesamiento de los datos, se empleó al software R, seleccionado por su eficiencia y capacidad para manejar análisis estadístico de datos complejos. En etapas posteriores, relacionadas con el desarrollo y ajuste de modelos, se recurrió al uso de Python debido a su flexibilidad y al extenso conjunto de librerías disponibles para ML y PLN. Esta metodología permitió que se abordara de manera comprensiva el análisis, desde la preparación de los datos hasta la aplicación de modelos de análisis de N-Gramas, modelado de tópicos y análisis de sentimientos.

4.1. Preparación de los Datos

Como se indicó en el Capítulo 3, debido a las limitaciones impuestas por la reciente actualización de la plataforma X, la recopilación de datos se efectuó utilizando un conjunto disponible en un repositorio de acceso abierto. Para obtener una evaluación inicial de los datos y establecer una base de análisis, se analizó el volumen de publicaciones a través del tiempo. Las Figuras 4.1 y 4.2 ilustran la distribución de la frecuencia diaria de publicaciones durante un intervalo aproximado de dos meses, específicamente del 18 de marzo al 25 de abril y del 21 de septiembre al 14 de octubre, periodo en el que se realizó la recolección de los datos.

Se observan fluctuaciones significativas en el volumen de publicaciones, con un aumento pronunciado hacia finales de septiembre de 2019. Específicamente, el 23 de septiembre se registró el mayor número de publicaciones. Para una interpretación adecuada de estos datos, es relevante considerar los eventos significativos que podrían coincidir con las fechas de los picos observados. En septiembre de 2019, Airbnb anunció oficialmente sus planes de convertirse en una empresa pública a través de una oferta pública inicial, prevista para el

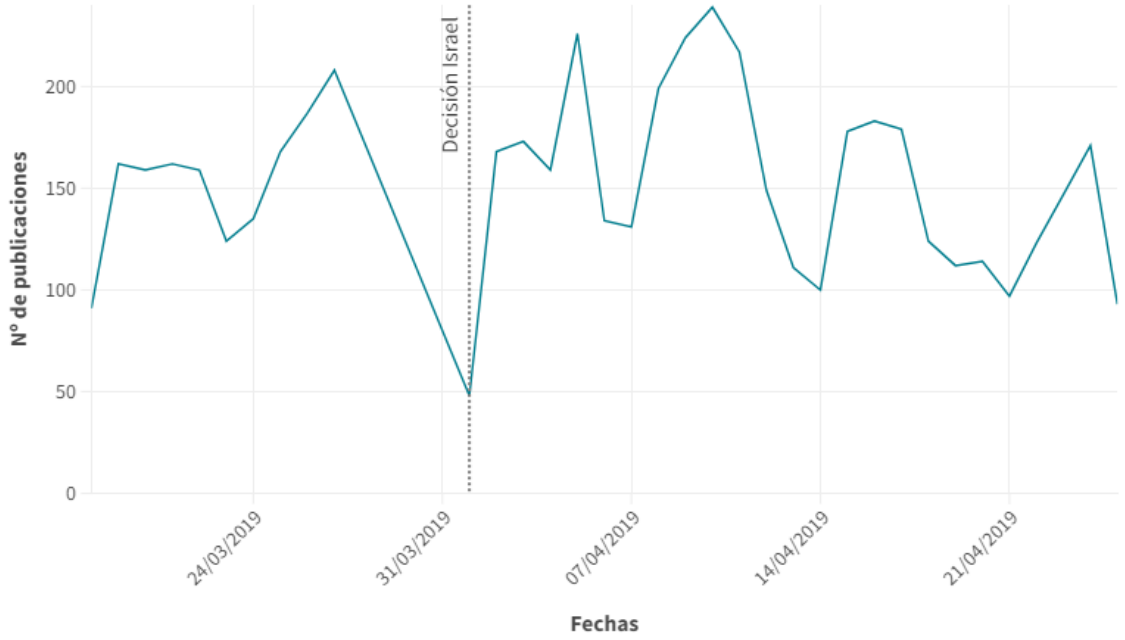

Figura 4.1: Publicaciones diarias. Elaboración propia

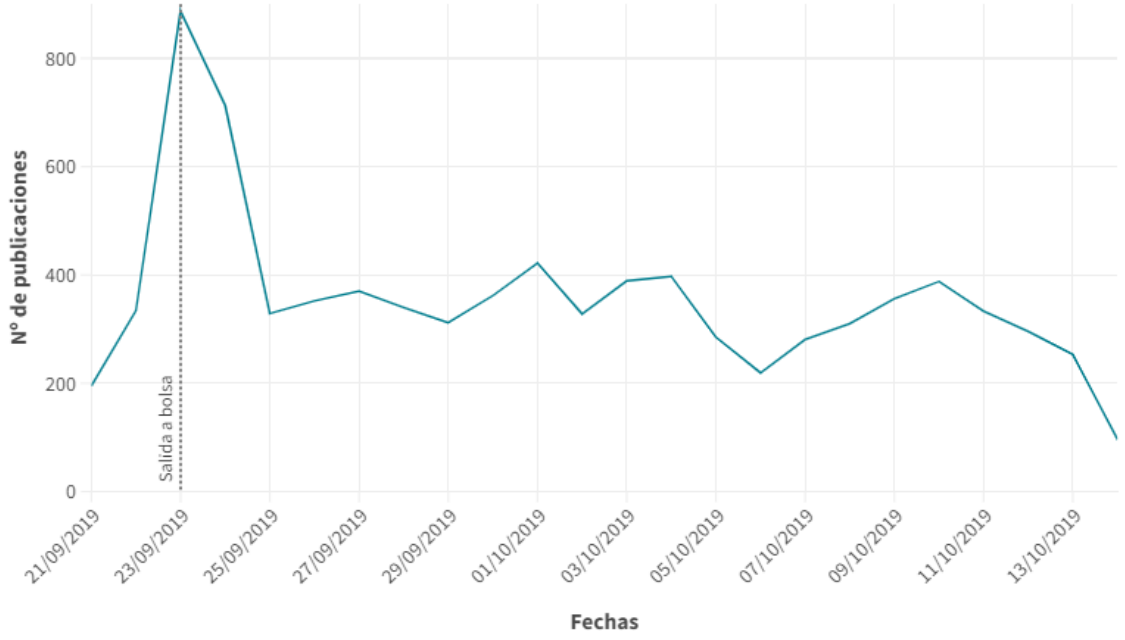

Figura 4.2: Publicaciones diarias. Elaboración propia

año 2020. Este anuncio se alinea temporalmente con el aumento observado en el número de publicaciones, reflejando el considerable interés de la comunidad de usuarios respecto a la transición de Airbnb al mercado bursátil (Airbnb, 2019). También se puede observar como el día con menor número de publicaciones es entorno a principios de abril, época en

la que Airbnb anunció su decisión de no eliminar los anuncios de asentamientos ilegales isrealíes en Cirsjordania, decisión previamente adoptada en 2018. (Middle East Eye, 2019) Este vínculo temporal sugiere una correlación entre las actividades corporativas significativas de la empresa y las discusiones en línea, destacando la relevancia de tales eventos en las interacciones de los usuarios con la plataforma.

Con el fin de alcanzar una mayor comprensión del contenido de las publicaciones dentro del conjunto de datos analizado, se realizaron varios análisis sobre la distribución de hashtags y los emoticonos disponibles en el corpus. Así, la Figura 4.3 presenta un análisis sobre la frecuencia de uso de diversos hashtags. El gráfico destaca los 15 hashtags más utilizados, con uno relacionado directamente con Airbnb dominando la lista, evidenciando que es el tema predominante entre las publicaciones. El análisis de los demás hashtags muestra una fuerte asociación con temas de turismo y ocio (tales como 'travel" , "vacation", "vacationrental", "homeaway", "tourism", "holiday"), lo que sugiere una preferencia marcada por Airbnb como alternativa de alojamiento para vacaciones, en contraposición a las alternativas hoteleras tradicionales. Además, la presencia de hashtags como "Vrbo" entre los hashtags más frecuentes ilustra la existencia de comparaciones entre Airbnb y otras plataformas que ofrecen servicios similares. Es notable, igualmente, la presencia de un hashtag dedicado a Japón, reflejando su popularidad como destino turístico entre los usuarios de Airbnb. Este interés puede estar influenciado por la inclusión de Japón en la lista de los 19 destinos recomendados por Airbnb para visitar en el año siguiente, según se anunció en una comunicación corporativa a finales de 2018 (Airbnb, 2018).

Finalmente, para adquirir una comprensión inicial sobre las opiniones de los usuarios, se llevó a cabo un análisis de la frecuencia con la que se utilizaron los emoticonos en el conjunto de publicaciones recopiladas, tal como se ilustra en la Figura 4.4. Se observa que el emoticono más frecuente es el que denota risa, lo cual podría sugerir una actitud generalmente positiva hacia Airbnb. En contraste, el segundo emoticono más común muestra una cara con expresión de llanto y tristeza, lo que podría interpretarse como una señal de experiencias o percepciones negativas. Además, el análisis revela la presencia de tres diferentes emoticonos relacionados con el tema del precio, lo que destaca su importancia en las discusiones de los usuarios y sugiere que este podría ser un tema de interés para posteriores análisis. De igual manera, es notable la abundancia de publicaciones con una connotación positiva, evidenciada en el uso frecuente de corazones y caras sonrientes. Este conjunto de emoticonos refleja una predominancia de experiencias satisfactorias compartidas por los usuarios en relación con Airbnb, ofreciendo una impresión inicial del sentimiento general que prevalece en el corpus analizado.

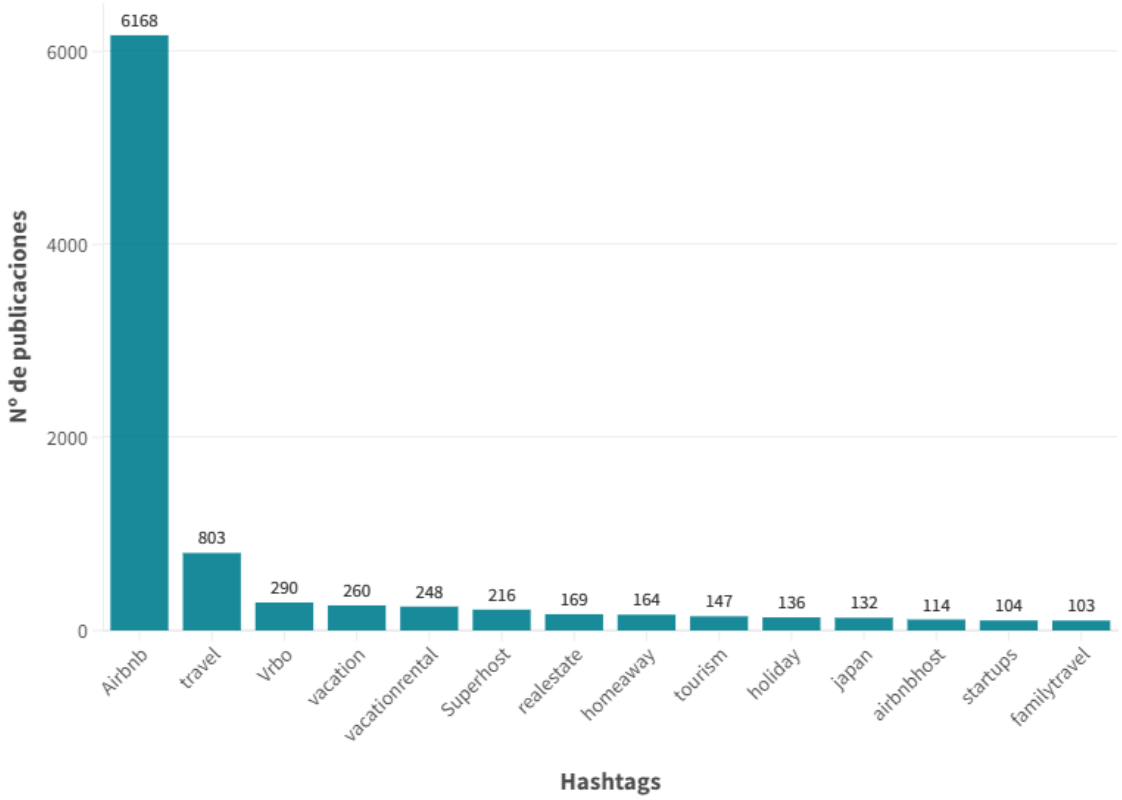

Figura 4.3: Distribución de publicaciones por Hashtag. Elaboración propia

4.2. Pre-procesamiento de los Datos

Para avanzar con el análisis, se requiere un preprocesamiento de datos conforme a lo expuesto en el Capítulo 3. Es importante notar que las estrategias de preprocesamiento difieren entre el modelado de tópicos y el análisis de sentimientos, como se explicó previamente. Por tanto, se describirán y explicarán los procedimientos específicos adoptados para cada método en las secciones correspondientes. En esta etapa, se procedió a eliminar URLs, menciones, hashtags y otros elementos no textuales que no contribuyen con información significativa en el análisis. A continuación, se realizó la eliminación de las *stopwords* más frecuentes en inglés, simplificando el contenido para un análisis más focalizado. Posteriormente, las palabras fueron tokenizadas y lematizadas, es decir, reducidas a su forma base. Este paso mejora la precisión de los análisis posteriores. Durante este proceso, se identificaron y ajustaron diversas excepciones vinculadas a las *stopwords* y a la lematización, adaptando el procedimiento para abordar situaciones particulares no cubiertas por las bibliotecas estándar utilizadas, garantizando así la calidad del preprocesamiento de los datos.

Para proporcionar una visión preliminar y detallada de los lemas obtenidos del corpus analizado, se ha desarrollado una nube de palabras. Como se puede observar en la figura 4.5, se destacan las 100 palabras más recurrentes del corpus ya lematizadas. El tamaño de

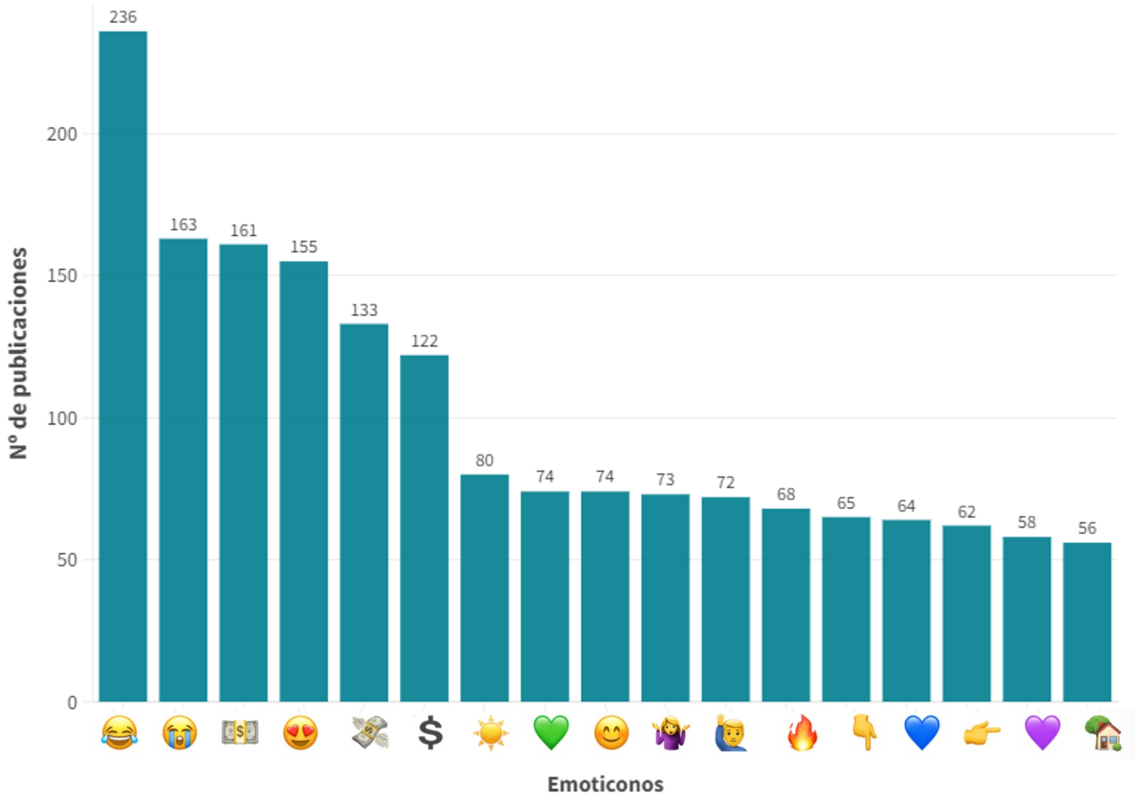

Figura 4.4: Distribución de publicaciones por emoticonos. Elaboración propia

las palabras en la figura es directamente proporcional a su frecuencia, indicando que las palabras de mayor tamaño son las que más aparecen en el corpus. Se destaca la presencia de términos como “book”, “stay”, y “new”, seguidos por una segunda categoría que incluye “host”, “home”, “rental” y “guest”. La presencia de estos términos en la nube de palabras indica una concentración del discurso en términos prácticos asociados a la hospitalidad, tales como la reserva, la estadía y la interacción con anfitriones y alojamientos todos vinculados temáticamente con el foco de investigación de este trabajo de Fin de Grado. Además, es notable que los verbos aparezcan en su forma infinitiva y los sustantivos se presenten en género masculino y número singular, evidenciando así la eficacia del preprocesamiento realizado.

4.3. Análisis descriptivo de N-Gramas

Con el fin de profundizar el análisis, se llevó a cabo una evaluación detallada de los N-Gramas más destacados a través del uso de la métrica TF-IDF, tal como se describe en la sección 3.2. La Figura 4.6 muestra la distribución de esta métrica específicamente para los unigramas, caracterizando así la relevancia de términos individuales dentro del corpus. En este análisis, destaca como palabra con mayor puntuación “book” (reservar), lo cual destaca el foco de las acciones de reserva en las discusiones sobre Airbnb. Posteriormente, se encuen-

Figura 4.5: Nube de Palabras. Elaboración propia

tran palabras como “stay” (estancia),“host” (anfitrión) y “rental” (alquiler) con puntuaciones significativas, reflejando las dimensiones esenciales de la experiencia Airbnb. Esto refleja el modelo de negocio de Airbnb, que gira en torno a las estancias cortas y la interacción con los anfitriones que alquilan sus propiedades. Otros términos como “guest” (huésped), “house” (casa), “home” (hogar), “place” (lugar) y “room” (habitación) resaltan los aspectos del alojamiento, sugiriendo descripciones de las propiedades de Airbnb. Adjetivos como “good” (bueno), “great” (excelente) y “beautiful” (bonito) probablemente se relacionan con las opiniones positivas de los huéspedes sobre su experiencia, destacando que la calidad y la estética son factores significativos en las reseñas. Términos temporales y relacionados con la logística de la estancia, como“time” (tiempo), “day” (día), “check” (registro), “night” (noche) y “city” (ciudad) sugieren una preocupación por la planificación durante las estancias. Finalmente, palabras como “hotel”, “property” (propiedad) y “experience” (experiencia) apuntan hacia una comparación con la industria hotelera tradicional, así como a la importancia de la calidad de la experiencia global.

Con el objetivo de ampliar el análisis de N-Gramas, se extendió el estudio a los bigramas, que son secuencias de dos palabras adyacentes, mediante la aplicación de la métrica TF-IDF. Este enfoque permitió generar la Figura 4.7, que muestra la distribución de esta métrica entre pares de términos conectados. De manera similar a los unigramas, una puntuación TF-IDF elevada en los bigramas señala que estos pares de palabras poseen una importancia particular dentro de los documentos y no se utilizan de manera generalizada en todo el corpus, destacando su relevancia específica en relación con los temas abordados. El bigrama con la relevancia más alta es “hide camara” (ocultar cámara), haciendo referencia a la preocupa-

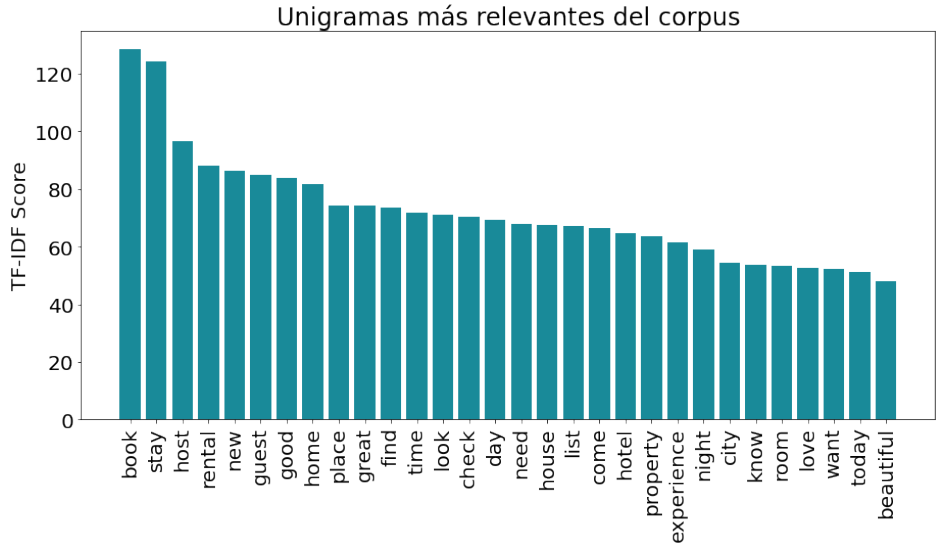

Figura 4.6: Métrica TF-IDF para Unigramas. Elaboración propia

ción sobre la presencia de cámaras ocultas en los alojamientos, lo cual se ha convertido en un tema de debate recurrente (EL PAÍS, 2024). Esta notable puntuación refleja una profunda inquietud entre los usuarios respecto a su privacidad y seguridad, evidenciando aspectos específicos relacionados con la confianza y la protección personal como puntos críticos en las conversaciones sobre experiencias de alojamiento. Otros bigramas como "vacation rental" (alquiler vacacional), "book stay" (reservar estancia) o "place stay"(lugar de estar) se centran en aspectos asociados a las acciones y preferencias más comunes de los usuarios en Airbnb, como la búsqueda y reserva de alojamientos para estancias vacacionales.

Otros conjuntos de bigramas como "web client" (cliente de la web) y "family find" (encuentro familiar) destacan la diversidad del público objetivo de Airbnb, evidenciando su uso tanto para viajes familiares como para reservas realizadas a través de internet, abarcando así una amplia gama de usuarios más allá de cualquier nicho específico de viajeros. El bigrama "short term" (corto plazo) resalta una de las características distintivas de Airbnb, asociada a la naturaleza temporal de los alojamientos ofrecidos, lo que indica que la plataforma se especializa en alquileres de corta duración. Finalmente, es relevante mencionar "link bio" (enlace en biografía) y "freelance ambassador" (embajador freelance), términos que sugieren estrategias de *marketing* digital y programas de referidos utilizados en la plataforma. Aquí se incluye la promoción a través de enlaces en biografías de redes sociales y la colaboración con embajadores independientes o *influencers*, quienes recomiendan Airbnb, a menudo compartiendo códigos de descuento para atraer nuevos usuarios.

Para finalizar con este análisis de N-Gramas, se extendió el estudio a los trigramas más destacados, presentando en la Figura 4.8 aquellas secuencias de tres palabras que obtuvieron las mayores puntuaciones según la métrica TF-IDF. Al explorar trigramas, se captura con mayor detalle el contexto de las conversaciones y se brinda información adicional de

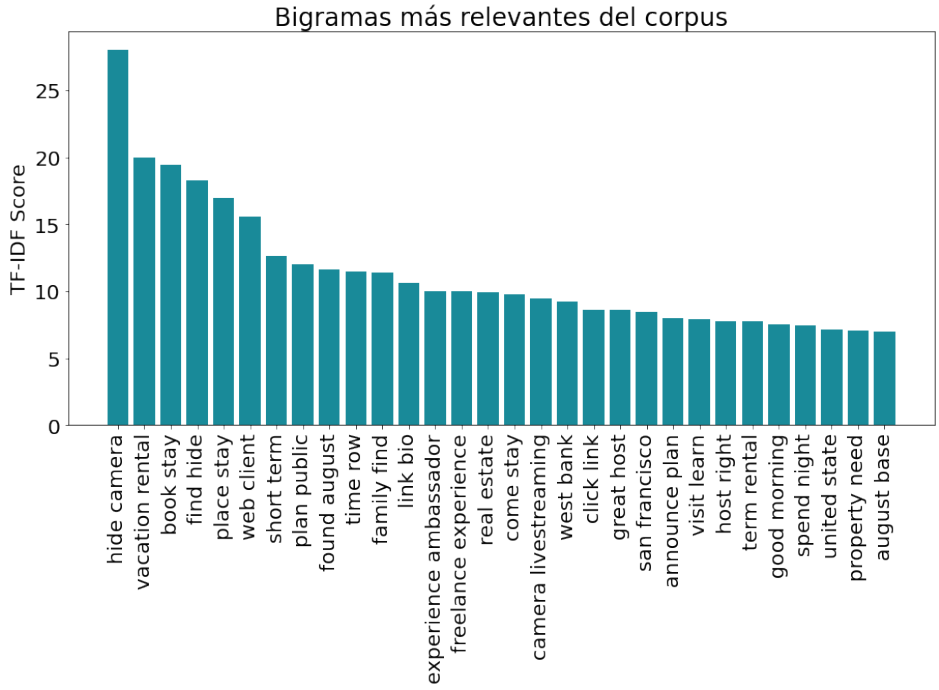

Figura 4.7: Métrica TF-IDF para Bigramas. Elaboración propia

las interacciones y preferencias de los usuarios. El trigrama más destacado, “find hide camera” (encontrar cámara oculta), resalta, al igual que ciertos bigramas, la preocupación de los huéspedes respecto a su privacidad y seguridad. Este temor también se evidencia en otros trigramas como “family find hide” y “hide camera livestreaming”, destacando la inquietud por la presencia de cámaras no declaradas. Adicionalmente, expresiones como “investment property need”, “learn profit investment” y “profit investment property” señalan la importancia de la inversión y rentabilidad de las propiedades en Airbnb, destacando el interés de los anfitriones en el retorno de inversión. Asimismo, trigramas como “short term rental” y “found august base” vinculan el uso de Airbnb con alquileres vacacionales, particularmente en el mes de agosto o para estancias de corta duración. En contraste, secuencias como “united state found”, “francisco united state” o “san francisco united” sugieren que Estados Unidos, y en especial la ciudad de San Francisco, figuran como destinos primordiales en las publicaciones analizadas.

En conclusión, el análisis de unigramas, bigramas y trigramas ha proporcionado una comprensión inicial descriptiva de los temas predominantes en el corpus de datos examinado. Los unigramas muestran la prevalencia de términos que reflejan acciones vinculadas a Airbnb, así como atributos y cualidades de los alojamientos. Por otro lado, los bigramas y trigramas refuerzan y amplían los resultados previos, aportando contexto adicional y profundizando en la comprensión sobre la naturaleza de los alojamientos y las inquietudes de los usuarios. Estos incluyen desde la seguridad y privacidad de los huéspedes hasta aspectos de inversión y rentabilidad para los anfitriones, incluyendo también algunas preferencias turísticas de la

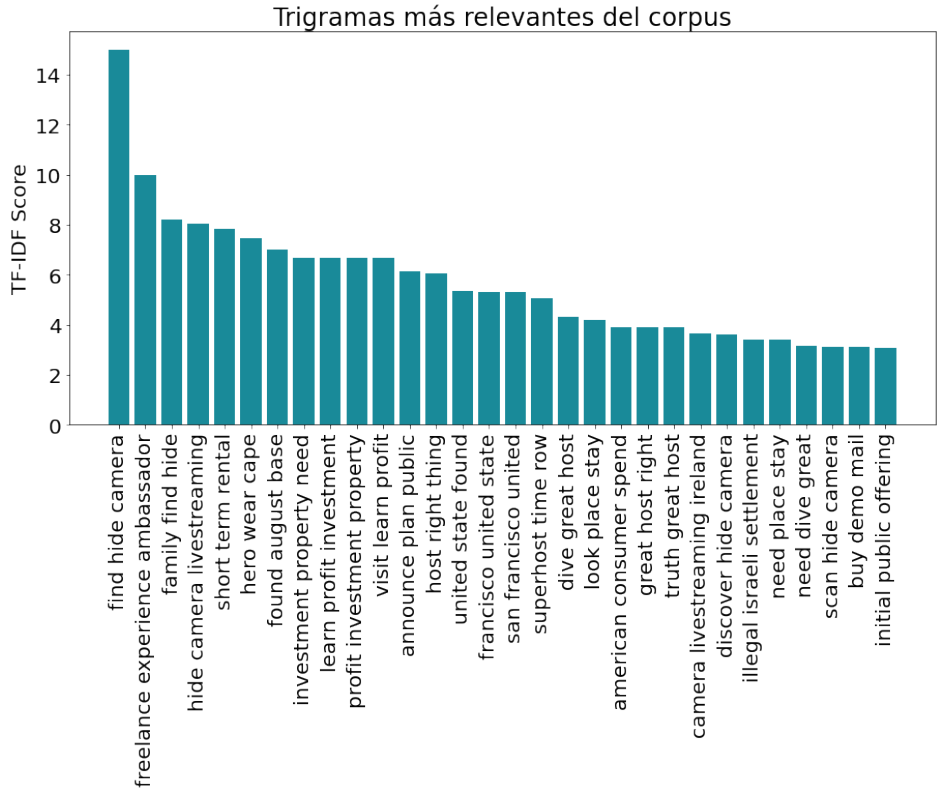

Figura 4.8: Métrica TF-IDF para Trigramas. Elaboración propia

comunidad de Airbnb.

4.4. Modelado de Tópicos

Tras finalizar el preprocesamiento de datos y realizar un análisis descriptivo inicial centrado en los N-gramas más destacados del corpus, se procedió a la fase de modelado de tópicos. Esta técnica de procesamiento automático es fundamental para identificar las principales categorías de discusión dentro de un corpus, tal como se explicó en la sección 3.3. Para determinar el modelo que mejor representa el corpus analizado, se requiere establecer el número óptimo de tópicos. Este proceso implica calcular el índice de coherencia para varios posibles números de tópicos, seleccionando aquel que ofrezca el mejor equilibrio entre interpretabilidad (un número menor de tópicos) y precisión (un índice de coherencia más alto). Como se muestra en la Figura 4.9, se evaluaron los índices de coherencia para configuraciones que van desde dos hasta catorce tópicos, lo que permite optimizar la estructura del modelo en función de la coherencia temática y la claridad interpretativa.

Por esta razón, aunque el valor de coherencia más alto se alcanza para k igual a 12, es notable que entre $k = 4$ y $k = 5$ se produce un incremento significativo en el índice de coherencia. Esta mejora es considerablemente mayor que la observada entre $k = 5$ y $k = 12$. Para asegurar que el modelo mantenga una alta interpretabilidad y un equilibrio adecuado

con el índice de coherencia, se ha optado por seleccionar $k = 5$ como el número óptimo de tópicos. Esta decisión se fundamenta en que la inclusión de los ocho tópicos adicionales no proporciona un incremento proporcional en valor informativo que justifique su complejidad adicional. Se prioriza, así, un modelo más sencillo y manejable que facilita la comprensión y la aplicación práctica de los resultados obtenidos.

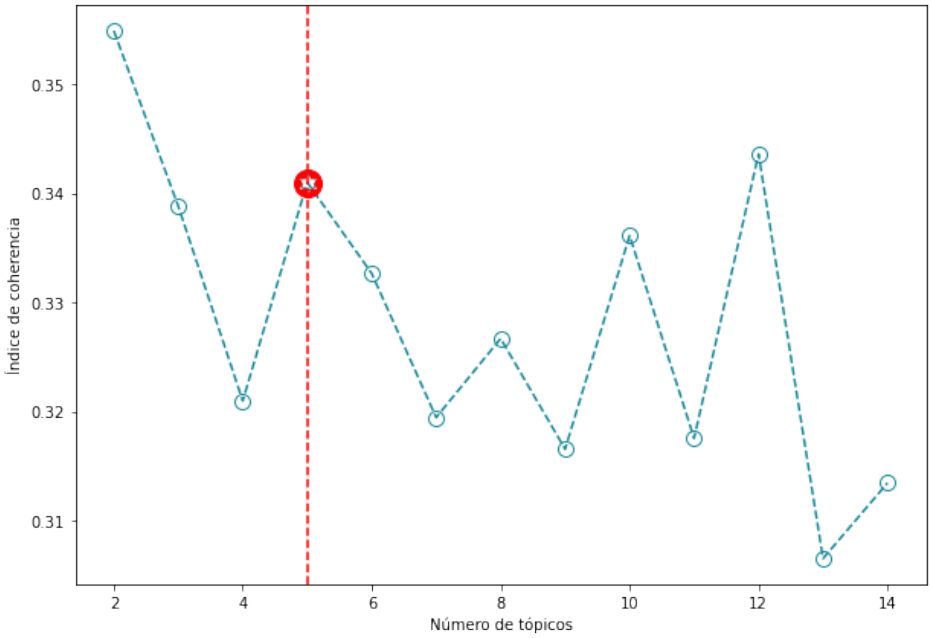

Figura 4.9: Número óptimo de tópicos según índice de coherencia. Elaboración propia

Para confirmar la elección óptima de k, se llevó a cabo un análisis adicional de la distancia intertópica entre los diversos temas identificados, utilizando para ello la herramienta PyLDAvis. Como se muestra en la Figura 4.11, cinco tópicos fueron identificados, cada uno caracterizado de manera distintiva. La revisión de la distancia entre estos tópicos revela que existe una considerable separación entre ellos, lo que indica una significativa heterogeneidad. No obstante, se observa un ligero solapamiento entre los tópicos 1 y 2. A pesar de esta superposición, las diferencias entre ellos son lo suficientemente marcadas como para justificar un análisis independiente de cada tópico. En conjunto, esta visualización apoya la decisión de establecer cinco tópicos bien delimitados y diferenciados, lo que contribuye a mejorar el carácter interpretativo de los tópicos identificados.

Una vez establecido el número óptimo de tópicos, se procedió al análisis y etiquetado de cada uno de ellos con el fin de identificar la categoría principal. En la Tabla 4.1, se presentan para cada tópico sus palabras clave más frecuentes, así como los bigramas y trigramas asociados. Los tópicos principales han sido identificados y etiquetados en el conjunto de datos de este Trabajo de Fin de Grado, mostrando una clara diferenciación de las temáticas abordadas por cada uno:

- **Tópico 1:** Experiencia de la estancia

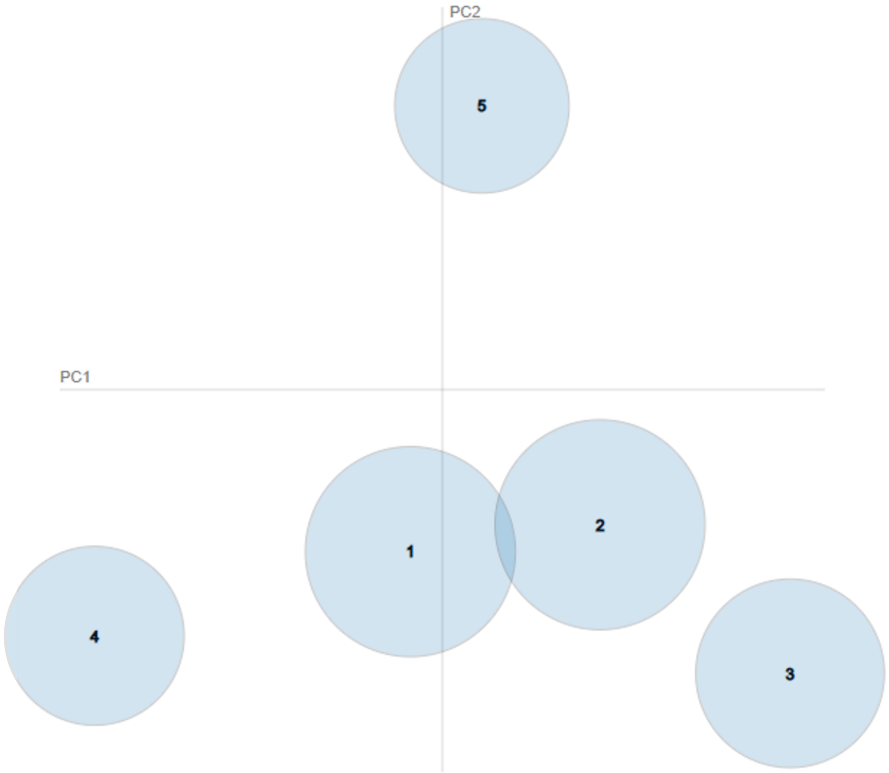

Figura 4.10: Visualización de la distancia intertópica. Elaboración propia

- **Tópico 2:** Modelo de negocio
- **Tópico 3:** Peligros de Airbnb
- **Tópico 4:** Anfitriones
- **Tópico 5:** Monetización

El primer tópico, el tópico 1 se enfoca con las publicaciones relacionadas con la **experiencia de la estancia**. Esta categoría abarca las publicaciones que tratan sobre la experiencia general de los usuarios con respecto a su hospedaje, incluyendo aspectos como la reserva de la estancia, la calidad y las características del alojamiento. Entre las principales palabras clave, se encuentran "stay" (estancia),"book" (reserva), "night" (noche), "beautiful" (bonito), "guest" (invitado), "room" (habitación). Al analizar los principales bigramas y trigramas como "san francisco" o "good morning tokio" (buenos días Tokyo), se puede evidenciar que muchos de estas publicaciones hacen referencia al destino de los viajeros, siendo San Francisco en Estados Unidos o Tokio en Japón unos de los principales destinos.

Por su parte, el tópico 2 incluye las publicaciones vinculadas con el **modelo de negocio** de Airbnb. Este tópico cubre tanto la gestión operacional de las reservas de alojamientos de corta duración como aspectos financieros relacionados con la especulación sobre la salida

de Airbnb al mercado bursátil. Está representado por palabras clave como “rental” (alquiler), “list” (listar) o “book” (reservar), bigramas como “vacation rental” (alquiler vacación), “plan list” (planear listar) , “rental market” (mercado alquiler); y trigramas como “announce plan public” (anunciar plan público), “short term rental” (alquiler a corto plazo) y “stock market splash” (estallido mercado de valores).

El tercer tópico refleja los **peligros de Airbnb**. Se identifican numerosas publicaciones que expresan los temores de los usuarios, utilizando palabras clave como “check” (comprobar) o “room” (habitación); o con bigramas o trigramas como “family find ” (familia encontar), “find hide camera” (encontrar camara escondida), o “hide camera livestreaming” (camara escondida en directo). Tal como se muestra en el gráfico 4.11, este tópico acumula el mayor número de publicaciones, alcanzando un total de 1795. Aunque las palabras clave individualmente no parecen alarmantes, el análisis de los bigramas y trigramas muestra una creciente preocupación por la privacidad, sugiriendo reservas y desconfianza hacia algunos alojamientos. Esta temática, altamente representada en el corpus, destaca la importancia de la seguridad y la privacidad para los usuarios de la plataforma.

El cuarto tópico se centra en los **anfitriones**, quienes son figuras clave en el modelo de negocio de Airbnb. Este tópico agrupa publicaciones que utilizan palabras clave como “host” (anfitrión), “experience” (experiencia) o “good” (bueno). Además, los bigramas y trigramas nos ayudan a definir de manera más concreta el tópico gracias a “host right” (anfitrión correcto), “superhost time” (superanfitrión tiempo) o “truth great host” (verdad genial anfitrión). Estas expresiones reflejan las experiencias positivas que los usuarios han tenido con sus anfitriones, resaltando evaluaciones favorables y el impacto positivo de la hospitalidad en la satisfacción del huésped.

Finalmente, el último tópico recoge aquellas publicaciones relacionadas con la **monetización** en el entorno de Airbnb, enfocándose específicamente en temas vinculados al **desarollo profesional y la gestión de inversiones**. Palabras clave como “learn” (aprender), “property” (propiedad), “company” (compañía), bigramas y trigramas como “investment property” (propiedad de inversión), “experience ambassador” (experiencia embajador), “freelance experience” (experiencia por cuenta propia), “learn profit investment” (aprender beneficio inversión), “visit learn profit” (visitar aprender beneficio) caracterizan las publicaciones que tratan sobre el uso de Airbnb para obtener beneficios, ya sea como una forma de inversión o de manera independiente, como embajador o *influencer*.

Por último, se analizó la distribución de los tópicos dentro del corpus, como se muestra en la Figura 4.11. Se observa que las publicaciones relacionadas con el modelo de negocio y los peligros asociados a Airbnb predominan en frecuencia. Por otro lado, las publicaciones que abordan temas relacionados con los anfitriones son menos comunes, reflejando así los intereses predominantes entre los usuarios de la plataforma X. Estos resultados muestran las preocupaciones y prioridades de los usuarios, destacando la relevancia de las cuestiones de seguridad y las estrategias empresariales dentro de la comunidad.

Tópico	Categoría	Palabras clave	Bigramas	Trigramas
1	Experiencia de la estancia	“stay” (estancia), “book” (reserva), “night” (noche), “beautiful” (bonito), “guest” (invitado), “room” (habitación)	“place stay” (lugar estancia), “san francisco”, “book stay” (reservar estancia)	“san francisco united ”, “look place stay” (buscar sitio quedar), “good morning tokio” (buenos dias Tokio)
2	Modelo de negocio	“rental” (alquiler), “list” (listar), “book” (reservar), “property” (propiedad)	“vacation rental” (alquiler vacación), “plan list” (planear listar) , “rental market” (mercado alquiler), “plan public” (plan público)	“announce plan public” (anunciar plan público), “short term rental” (alquiler a corto plazo),“stock market splash” (estallido mercado de valores)
3	Peligros de la aplicación de Airbnb	“check” (comprobar), “room” (habitación), “place”(lugar)	“family find ” (familia encontar), “hide camera”(camara escondida), o “camera livestreaming” (camara en directo)	“find hide camera”(encontrar camara escondida), “hide camera livestreaming” (camara escondida en directo)
4	Anfitriones	“host” (anfitrión), “experience” (experiencia), “good” (bueno)	“host right” (anfitrión correcto), “superhost time” (superanfitrión tiempo)	“truth great host” (verdad genial anfitrión), “great host right” (anfitrión genial verdad)
5	Monetización	“learn” (aprender), “property” (propiedad), “company” (compañía)	“investment property” (propiedad de inversión), “experience ambassador” (experiencia embajador), “freelance experience” (experiencia por cuenta propia)	“learn profit investment” (aprender beneficio inversión), “visit learn profit” (visitar aprender beneficio), “freelance experience ambassador” (embajador experiencia por cuenta propia)

Tabla 4.1: Palabras clave, bigramas y trigramas por tópico. Elaboración propia.

4.5. Análisis de sentimientos

Este último capítulo está enfocado en analizar el sentimiento subyacente de los usuarios de X hacia la aplicación Airbnb. Tal y como se explico en el capítulo 3.4, se empleará el diccionario VADER, gracias a su adapatabilidad y eficacia a la hora de analizar las percepciones y emociones de las publicaciones de X. VADER tiene en cuenta para su análisis tanto las mayúsculas, como los signos de puntuación como las *stopwords* ya que aportan información relevante a la hora de calificar el sentimiento de un texto. Es por ello, que para este capítulo se ha realizado un nuevo pre-procesamiento de datos, eliminando únicamente los hashtags, las URLs, las menciones y los números.

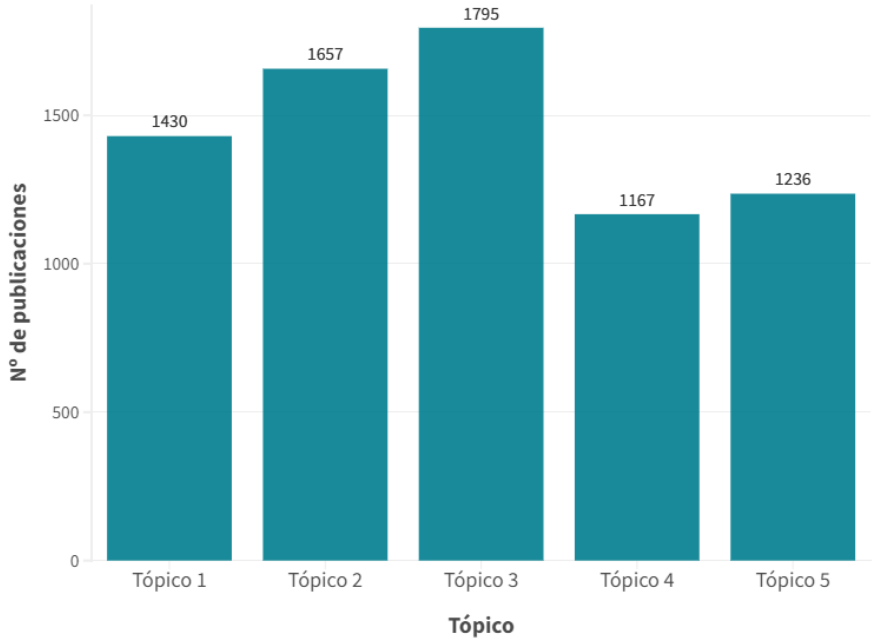

Figura 4.11: Número de publicaciones por tópico. Elaboración propia.

Teniendo en cuenta las anteriores consideraciones, se realizó un análisis de sentimientos sobre el conjunto de datos, asignando una puntuación normalizada a cada publicación variando entre -1 y 1. Para presentar de forma descriptiva una visión general de los resultados, se elaboró un diagrama de caja. Como se ilustra en la Figura 4.12, predominan los valores positivos, siendo solamente el primer percentil el que recoge valores negativos. Este resultado sugiere que el 75 % de las publicaciones restantes presenta una connotación neutra o positiva. Adicionalmente, la media y la mediana, ubicadas alrededor de 0.5, apuntan a emociones positivas predominantes en las publicaciones.

Para complementar los resultados del análisis de sentimientos, se realizó un histograma que muestra la distribución de las puntuaciones de sentimiento de las publicaciones, tal como se muestra en la Figura 4.13. La revisión del histograma muestra que la frecuencia más alta corresponde a una puntuación de 0, sugiriendo que un volumen significativo de las publicaciones posee un carácter neutro. A la derecha del gráfico, se observa una acumulación de publicaciones con puntuaciones positivas, lo que indica un sentimiento favorable. En contraste, las publicaciones con sentimientos negativos representan una proporción menor. Por lo tanto, se puede inferir que en el conjunto de datos analizado, la tendencia general refleja un sentimiento positivo de los usuarios de la plataforma X hacia Airbnb.

Para lograr una comprensión más detallada de las emociones reflejadas, se efectuó un análisis de la evolución temporal de la puntuación de sentimiento acumulada de las publicaciones por día. Tal como se visualiza en las Figuras 4.14 y 4.15, el pico de puntuación positiva más alto se registra en septiembre de 2019. Este periodo no solo coincide con el día

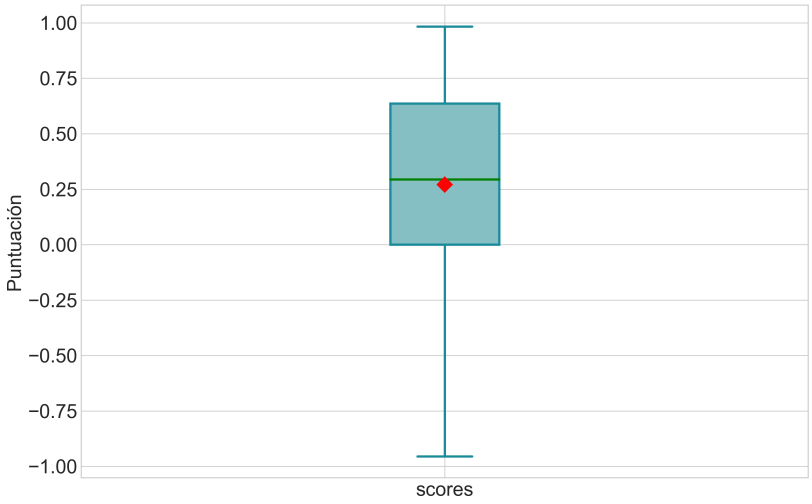

Figura 4.12: Diagrama de caja de la puntuación. Elaboración propia.

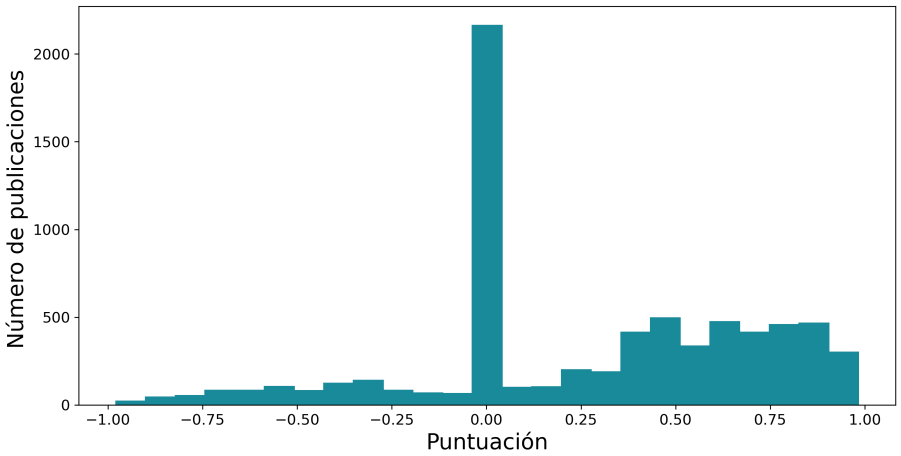

Figura 4.13: Histograma de la puntuación. Elaboración propia.

de mayor volumen de publicaciones, tal como se identificó en la Figura 4.2, sino que también corresponde al momento en que Airbnb anunció su intención de cotizar en bolsa. Asimismo, se observa que durante el periodo analizado, en ninguna fecha la puntuación acumulada de sentimiento es completamente negativa, sugiriendo que, en su conjunto, la percepción de los usuarios hacia Airbnb es favorable. Aunque hay días con puntuaciones más elevadas que otros, en términos generales, la tendencia emocional expresada en este corpus de datos puede clasificarse como positiva. Sin embargo, es importante resaltar el día en el que se registró la puntuación más baja, en abril de 2019, coincidiendo con el momento en que Airbnb anunció su decisión de no proceder con la eliminación de los anuncios de asentamientos israelíes ile-

gales en Cisjordania, una decisión adoptada en noviembre de 2018 (Middle East Eye, 2019).

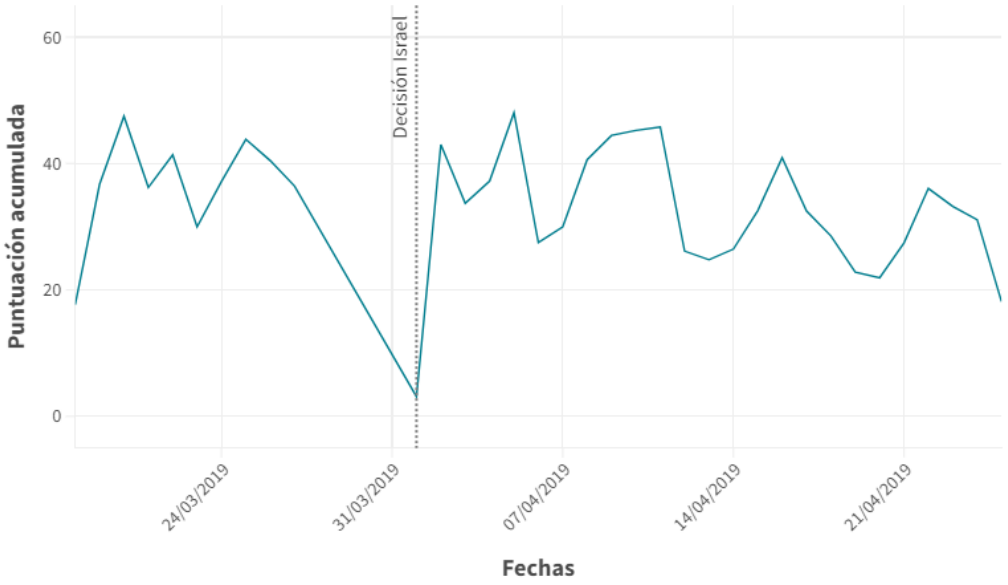

Figura 4.14: Puntuación de sentimiento de las publicaciones agregada por día. Elaboración propia.

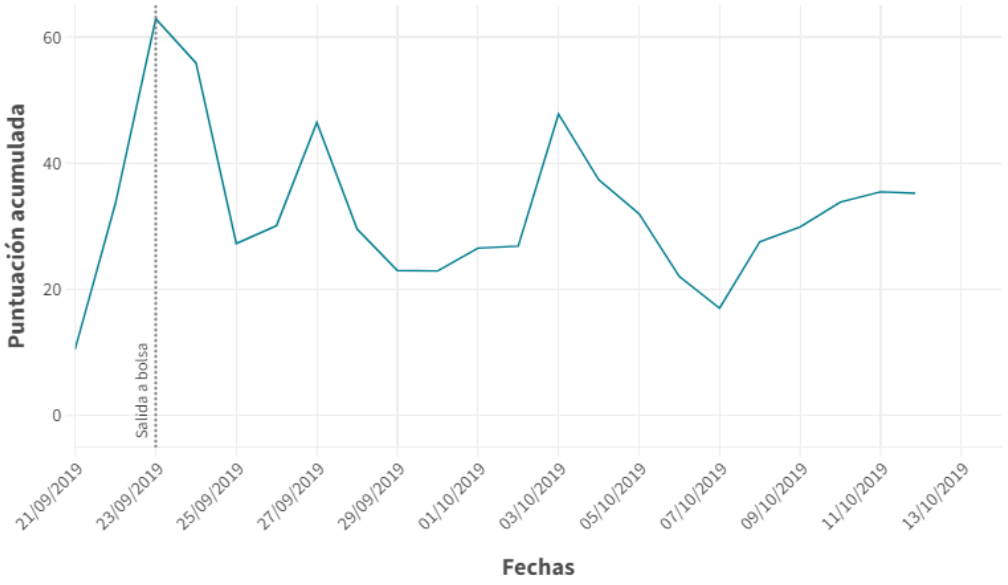

Figura 4.15: Puntuación de sentimiento de las publicaciones agregada por día. Elaboración propia.

Además del análisis temporal del sentimiento asociado a las publicaciones en su conjunto, se ha llevado a cabo también un análisis temporal del sentimiento para cada uno de los tópicos identificados en la sección 4.4. Este enfoque proporciona información sobre la

dinámica temporal del sentimiento por tópico, identificando aquellos que se relacionan más estrechamente con reacciones positivas o negativas. La Figura 4.16 muestra esta evolución, diferenciando la trayectoria del sentimiento para cada tópico identificado.

Respecto al primer tópico, **la experiencia de la estancia**, se observa que el sentimiento asociado mantiene una positividad estable a lo largo del tiempo, con fluctuaciones mínimas, lo que lo categoriza como uno de los tópicos con emociones más equilibradas. Este patrón sugiere que la percepción general de los usuarios sobre la experiencia de hospedarse en un Airbnb tiende hacia una valoración ligeramente positiva. En la evaluación del sentimiento correspondiente al tópico 2, **modelo de negocio**, este se destaca por registrar la mayor puntuación positiva dentro de todo el conjunto de datos. La notable tendencia positiva, observada durante el mes de septiembre, corrobora las hipótesis previamente planteadas sobre la inminente oferta pública inicial de la empresa. En este contexto, se identificó que el anuncio de este evento se correlacionó con un marcado incremento en la percepción positiva dentro de la plataforma X, lo cual refleja una recepción favorable por parte de los potenciales inversores y accionistas.

La gráfica del tópico 3, dedicada a **los peligros de Airbnb**, evidencia que este tópico posee la mayor variabilidad en términos de sentimiento, con oscilaciones notables hacia valoraciones menos positivas. Esto evidencia la preocupación existente entre los usuarios acerca de aspectos críticos como la privacidad. Se observan puntuaciones que descienden hasta alcanzar niveles neutros, lo que podría ser indicativo de la diversidad de opiniones respecto a los riesgos percibidos, reflejando el amplio rango de sentimientos que las reservas en la plataforma pueden generar en la comunidad de usuarios. En relación con el cuarto tópico, dedicado a los **anfitriones**, se observa que este tópico exhibe un perfil claramente positivo y con baja variabilidad, lo cual indica que los inquilinos tienen una percepción muy favorable hacia los anfitriones. Este análisis corrobora las suposiciones hechas en el estudio de bigramas y valida que palabras clave como “good” (bueno) y “great” (genial), mencionadas en la Tabla 4.1, reflejan genuinamente las opiniones positivas de los usuarios hacia sus anfitriones.

Finalmente, analizando el último tópico que agrupa las publicaciones relacionadas con la **monetización**, se puede observar que es el tópico más neutro de todos los identificados. Sus publicaciones no exceden una puntuación acumulada de 10 puntos, indicando que el sentimiento de los usuarios hacia Airbnb como forma de inversión o como forma de desarrollo personal es neutral. Sin embargo, es el único tópico que tiene asociada una puntuación negativa, relacionada con las publicaciones sobre los asentamientos en Israel y la decisión económica de Airbnb de donar dichos ingresos.

A partir del análisis de sentimientos realizado sobre el conjunto de datos, se ha identificado una tendencia predominante durante el periodo de recolección de datos. En general, se observa una mayoría de publicaciones de carácter positivo y neutro, con una notable escasez de publicaciones que expresen emociones negativas. Este patrón muestra que, en las publicaciones de la plataforma X, los usuarios tienden a mostrar sentimientos neutros y mo-

deradamente positivos hacia Airbnb. Asimismo, se ha evidenciado que ciertos eventos han generado respuestas positivas entre los usuarios, como la anunciada salida a bolsa de Airbnb, destacando el interés inversor en esta iniciativa. De la misma manera, otros eventos han propiciado respuestas negativas por parte de los usuarios, como la decisión en torno a los asentamientos en Israel. Finalmente, tópicos relacionados con la experiencia de los usuarios y la interacción con los anfitriones han mostrado niveles principalmente neutros, lo que señala oportunidades para mejoras en estos aspectos dentro del mercado.

Figura 4.16: Puntuación de sentimiento por tópico. Elaboración propia.

Capítulo 5

Conclusiones

La economía colaborativa, en la que se enmarca Airbnb, ha revolucionado la manera en que concebimos el intercambio de bienes y servicios. Desde su creación en 2007, Airbnb no solo ha crecido hasta contar con más de cuatro millones de anfitriones, sino que también ha encarnado el espíritu de altruismo y cooperación entre individuos. La plataforma ofrece una alternativa dinámica a las opciones de alojamiento tradicionales, permitiendo a personas de todo el mundo alquilar habitaciones o viviendas enteras para hospedar a viajeros. Este modelo no solo facilita el acceso a alojamientos únicos y personalizados, sino que también fomenta un sentido de comunidad y conexión global.

En el contexto de crecimiento y expansión global de Airbnb, es de gran relevancia explorar las opiniones y percepciones de los ciudadanos sobre la marca. Las redes sociales, especialmente X (antes Twitter), desempeñan un papel clave en este análisis por su capacidad para proporcionar una plataforma accesible donde los usuarios pueden compartir sus experiencias y valoraciones en tiempo real. X se destaca por su naturaleza inmediata y alcance global, facilitando un flujo constante de datos actualizados. La plataforma es particularmente eficaz para seguir tendencias y detectar patrones de discusión mediante el uso de hashtags, lo que permite agrupar y analizar datos sobre temas específicos como Airbnb. Bajo estas consideraciones, el presente Trabajo de Fin de Grado propone realizar un análisis de la reputación de la marca Airbnb. El objetivo es identificar los principales tópicos de discusión y los sentimientos subyacentes en las publicaciones asociadas a la marca en X.

Con este propósito, se llevó a cabo inicialmente una revisión de la literatura sobre las estrategias de minería de texto empleadas para extraer información relevante de publicaciones en redes sociales, especialmente en temas relacionados con la economía colaborativa. Esta revisión permitió identificar las técnicas más eficaces para el análisis propuesto. Se concluyó que la estrategia LDA era la más adecuada para el modelado de tópicos, dada su capacidad para descubrir temas latentes en grandes volúmenes de datos. Para el análisis de sentimientos, se determinó que el diccionario VADER era el más apropiado, gracias a su habilidad para captar las sutilezas de los sentimientos expresados en el lenguaje de las redes sociales, inclu-

yendo la interpretación de emoticonos. Los resultados de esta fase de revisión respaldaron la selección de las metodologías aplicadas en el análisis de las publicaciones relacionadas con Airbnb.

El proceso metodológico adoptado en este trabajo abarca varias etapas para el análisis exhaustivo de las publicaciones. Comienza con una etapa de pre-procesamiento de datos, necesaria para preparar y limpiar el conjunto de datos para los análisis subsiguientes. A continuación, se realiza un análisis descriptivo de n-gramas, utilizando la métrica de relevancia TF-IDF para identificar las palabras y frases más significativas dentro del corpus. Posteriormente, se emplea LDA para el modelado de tópicos, con el objetivo de descubrir y categorizar los temas latentes que predominan en las publicaciones. Finalmente, se lleva a cabo el análisis de sentimientos utilizando el diccionario VADER, que permite evaluar la carga emocional de las publicaciones y capturar con precisión el sentimiento expresado por los usuarios en relación con Airbnb.

En el marco de este estudio, se analizaron 13.902 publicaciones recogidas entre marzo, abril, septiembre y octubre de 2019. El análisis descriptivo preliminar reveló fluctuaciones significativas en el volumen de publicaciones. Se observó un descenso notable en abril, coincidiendo con el anuncio de Airbnb de no eliminar los anuncios de asentamientos israelíes ilegales en Cisjordania, y un aumento considerable a finales de septiembre, alrededor del anuncio oficial de su transición a una entidad pública. Además, se identificó una amplia presencia de hashtags relacionados con turismo y ocio, destacando a Airbnb como una opción preferida para alojamiento vacacional. A través del análisis de n-gramas, los unigramas revelaron una predominancia de términos asociados a acciones de Airbnb, mientras que los bigramas y trigramas proporcionaron un contexto adicional y profundizaron en las preocupaciones de los usuarios sobre temas como la seguridad y la rentabilidad.

En cuanto al modelado de tópicos, se han identificado cinco tópicos en el conjunto de datos analizado. El primero de ellos hace referencia a la experiencia de la estancia de los usuarios, incluyendo aspectos como la reserva, la calidad y las caracteristicas del alojamiento. El segundo tópico agrupa aquellas publicaciones relacionadas con el modelo de negocio de Airbnb, tanto la parte operacional como aspectos financieros como la oferta inicial pública. El tópico 3 refleja los peligros de Airbnb, reflejando la preocupación de los usuarios y la desconfianza hacia algunos apartamentos. El cuarto tópico se centra en los anfitriones, figuras clave en la aplicación de Airbnb y reflejando la buena relación entre anfitriones y huéspedes. Y finalmente, el útimo tópico recoge las publicaciones relacionadas con la monetización, es decir, con temas relacionados con el desarollo profesional y la gestión de inversiones en busca de un retorno.

Finalmente, el análisis de sentimientos realizado sobre el corpus reveló una predominante connotación positiva en las publicaciones examinadas, con más del setenta y cinco por ciento de ellas obteniendo una calificación de sentimiento igual o superior a 0. Durante el estudio, se evaluó la tendencia temporal de los sentimientos, identificando que los usuarios generalmente

mantienen una actitud neutral o ligeramente positiva hacia la plataforma. El análisis destacó que ciertos eventos tienen un impacto significativo en los sentimientos expresados en las publicaciones, como la controversia sobre los asentamientos en Israel y el anuncio de la oferta pública inicial de Airbnb. Además, se determinó que el tópico que generó el mayor sentimiento positivo estaba relacionado con el modelo de negocio de Airbnb, evidenciando la aceptación y el entusiasmo de los usuarios por las oportunidades de inversión y crecimiento que ofrece la empresa.

Para futuros trabajos en este campo, sería interesante considerar la expansión de la ventana temporal utilizada. Esta expansión no solo debe limitarse a un incremento en la cantidad de años analizados, sino también en la variedad de meses seleccionados para la recolección de datos. Este enfoque más amplio permitiría capturar variaciones en las tendencias y comportamientos a lo largo del tiempo, proporcionando una visión más comprensiva y menos sesgada que la que se podría obtener al centrarse únicamente en un año específico. Además, la incorporación de información demográfica detallada es crucial para profundizar en el análisis. Datos como el género, la edad, y el lugar de residencia de los usuarios, podrían revelar patrones distintivos en las preferencias y comportamientos de distintos segmentos de la población. También facilitaría la segmentación de los usuarios en grupos más específicos, pudiendo ser muy útil a la hora de comprender las percepciones de los usuarios hacia Airbnb.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que *ChatGPT* u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Alejandra Bandeira Eguiraun, estudiante de Doble Grado en ADE y Business Analytics (E-2 + Analytics) de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Análisis de Percepciones en la Hostelería Digital: Minería de Texto en las Revisiones de Airbnb en Redes Sociales", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa *ChatGPT* u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Brainstorming de ideas de investigación: Utilizado para idear y esbozar posibles áreas de investigación.

2. Referencias: Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.

3. Interpretador de código: Para realizar análisis de datos preliminares.

4. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.

5. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado *ChatGPT* u otras herramientas similares). Soy consciente de

las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 21 Abril 2024

Firma: Alejandra Bandeira Eguiraun

Referencias

Airbnb. (2018). *Airbnb reveals the 19 destinations to visit in 2019*. https://news.airbnb.com/airbnb-reveals-the-19-destinations-to-visit-in-2019/. (2024-03-27)

Airbnb. (2019). *Airbnb announces intention to become a publicly-traded company during 2020*. Descargado de https://news.airbnb.com/airbnb-announces-intention-to-become-a-publicly-traded-company-during-2020/ (Accedido el 10 de abril de 2024)

Bandeira, A. (2024). *Tfg airbnb*. https://github.com/alejandrabandeira/TFG_Airbnb. GitHub.

Blanco, U. (2023). *¿qué significa el cambio de twitter a x y cómo afecta a los usuarios?* https://cnnespanol.cnn.com/2023/07/25/que-significa-cambio-twitter-x-como-afecta-usuarios-orix//.

Blei, D., Ng, A., y Jordan, M. (2001). Latent dirichlet allocation. *Advances in neural information processing systems*, *14*.

Blei, D. M., Ng, A. Y., y Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Botsman, R., y Rogers, R. (2010). What's mine is yours. *The rise of collaborative consumption*, *1*.

Carlemany, U. (2022). *7 ejemplos de economía colaborativa que te sorprenderán.* https:https://www.universitatcarlemany.com/actualidad/blog/economia-colaborativa/.

Chang, W.-L., y Wang, J.-Y. (2018). Mine is yours? using sentiment analysis to explore the degree of risk in the sharing economy. *Electronic Commerce Research and Applications*, *28*, 141–158.

CNMC. (2016). *"conclusiones preliminares sobre los nuevos modelos de prestación de servicios y la economía colaborativa".* (acceso Noviembre 20, 2023) https://www.cnmc.es/CNMC/Prensa/TabId/254/ArtMID/6629/ArticleID/1684/La-CNMC-somete-a-consulta-p250blica-las-Conclusiones-preliminares-del-estudiosobre-los-nuevos-modelos-de-prestaci243n-de-servicios-y-la-econom237a-colaborativa.aspx.

Curtis, S. K., y Lehner, M. (2019). Defining the sharing economy for sustainability. *Sustainability*, *11*(3), 567.

De Lima, F. A. (2022). # circular economy–a twitter analytics framework analyzing twitter data, drivers, practices, and sustainability outcomes. *Journal of Cleaner Production*, *372*, 133734.

Díaz Foncea, M., Marcuello, C., y Montreal-Garrido, M. (2016). *Economía social y economía colaborativa: encaje y potencialidades* (Inf. Téc.).

Duque, S. (2021). *Crowdinvestment o inversiÓn colectiva como futuro de la economÍa colaborativa.* https://www.linkedin.com/pulse/crowdinvestment-o-inversi%C3%B3n-colectiva-como-futuro-de-la-duque-e-/?originalSubdomain=es/.

Durán-Sánchez, A., Álvarez-García, J., del Río, M. d. l. C., Maldonado-Erazo, C. P., y cols. (2016). Economía colaborativa: análisis de la producción científica en revistas académicas. *Revista de Gestão e Secretariado*, *7*(3), 1–20.

Edelman, B. G., y Luca, M. (2014). Digital discrimination: The case of airbnb. com. *Harvard Business School NOM Unit Working Paper*(14-054).

EL PAÍS. (2024, 13 de Mar). Airbnb prohíbe las cámaras de seguridad en el interior de las casas. *El País*. Descargado de https://elpais.com/tecnologia/2024-03-13/airbnb-prohibe-las-camaras-de-seguridad-en-el-interior-de-las-casas.html (Consultado el [Fecha de acceso])

Elgendy, N., y Elragal, A. (2014). Big data analytics: a literature review paper. En *Advances in data mining. applications and theoretical aspects: 14th industrial conference, icdm 2014, st. petersburg, russia, july 16-20, 2014. proceedings 14* (pp. 214–227).

Espinosa Fernández, M. T. (2018). La economía colaborativa. orígenes, evolución y retos futuros.¿ en qué consiste realmente este nuevo fenómeno?

Expansión. (Diciembre 2020). *Airbnb logra la mayor salida a bolsa de 2020 en wall street.* https://expansion.mx/mercados/2020/12/10/airbnb-logra-mayor-salida-bolsa-2020/.

Gansky, L. (2010). *The mesh: Why the future of business is sharing*. Penguin.

Hamari, J., Sjöklint, M., y Ukkonen, A. (2016). The sharing economy: Why people participate in collaborative consumption. *Journal of the association for information science and technology*, *67*(9), 2047–2059.

Hutto, C., y Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. En *Proceedings of the international aaai conference on web and social media* (Vol. 8, pp. 216–225).

Izquierdo Expósito, V., Álvarez Rodríguez, P., y Nuño Barrau, A. (2017). Comunicación y divulgación de contenidos artísticos a través de las redes sociales: Facebook y twitter.

Journal, T. W. S. (2023). https://www.wsj.com/market-data/quotes/ABNB/financials/annual/income-statement.

Kiatkawsin, K., Sutherland, I., y Kim, J.-Y. (2020). A comparative automated text analysis of airbnb reviews in hong kong and singapore using latent dirichlet allocation. *Sustainability*, *12*(16), 6673.

Koh, N. S. (2011). *The valuation of user-generated content: a structural, stylistic and semantic analysis of online reviews*. Singapore Management University (Singapore).

Lawani, A., Reed, M. R., Mark, T., y Zheng, Y. (2019). Reviews and price on online platforms: Evidence from sentiment analysis of airbnb reviews in boston. *Regional Science and Urban Economics*, *75*, 22–34.

Middle East Eye. (2019, 9 de Abril). Airbnb reverses decision delisting illegal israeli settlements. *Middle East Eye*. Descargado de `https://www.middleeasteye.net/news/airbnb-reverses-decision-delisting-illegal-israeli-settlements`

Pano, T., y Kashef, R. (2020). A complete vader-based sentiment analysis of bitcoin (btc) tweets during the era of covid-19. *Big Data and Cognitive Computing*, *4*(4), 33.

Prada, A., y Iglesias, C. A. (2020). Predicting reputation in the sharing economy with twitter social data. *Applied Sciences*, *10*(8), 2881.

RAE. (2022). *Diccionario de la lengua española, 23.ª ed.* (acceso Noviembre 20, 2023) `https://dle.rae.es/econom%C3%ADa?m=form/`.

Rensing, L. (2022). *Twitter as a data mine: Can user needs be derived from twitter data?-an example of airbnb* (B.S. thesis). University of Twente.

Röder, M., Both, A., y Hinneburg, A. (2015). Exploring the space of topic coherence measures. En *Proceedings of the eighth acm international conference on web search and data mining* (pp. 399–408).

Sievert, C., y Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topics. En *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).

Statista. (Enero 2023). *Principales datos de airbnb a nivel mundial a fecha de diciembre de 2022.* (acceso Noviembre 16, 2022) `https://es.statista.com/estadisticas/1218479/principales-indicadores-de-actividad-de-airbnb-en-el-mundo/`.

Sutherland, I., y Kiatkawsin, K. (2020). *Determinants of guest experience in airbnb: a topic modeling approach using lda. sustainability 12 (8): 3402.*

Teh, P. L. (2021). *Viewing airbnb from twitter: Factors associated with users' utilization.* Mendeley Data, V1. (Dataset) doi: 10.17632/g6sk4z4fbw.1

Tsimonis, G., y Dimitriadis, S. (2014). Brand strategies in social media. *Marketing Intelligence & Planning*, *32*(3), 328–344.

Uysal, A. K., y Gunal, S. (2014). The impact of preprocessing on text classification. *Information processing & management*, *50*(1), 104–112.

Villeneuve, H., y O'Brien, W. (2020). Listen to the guests: Text-mining airbnb reviews to explore indoor environmental quality. *Building and Environment*, *169*, 106555.

Yun-tao, Z., Ling, G., y Yong-cheng, W. (2005). An improved tf-idf approach for text classification. *Journal of Zhejiang University-Science A*, *6*(1), 49–55.

Zhang, Z., y Fu, R. J. (2020). Accommodation experience in the sharing economy: a comparative study of airbnb online reviews. *Sustainability*, *12*(24), 10500.