Facultad de Ciencias Económicas y Empresariales

Análisis automático de revisiones online, a través de estrategias basadas en aprendizaje automático y procesamiento de lenguaje natural

Autor: Alejandra Tabasco Ruiz
Director: Jenny Alexandra Cifuentes Quintero

MADRID| Abril 2023

Resumen

En los últimos años, el uso de la tecnología para evaluar y expresar opiniones sobre productos y servicios ha experimentado un notable aumento por parte de los usuarios. La información generada por este fenómeno es de gran utilidad para las empresas, ya que les permite comprender con mayor precisión el comportamiento de sus clientes y establecer estrategias en relación a sus productos, basándose en información real. Esto se traduce en una oportunidad única para mejorar la calidad y satisfacción del cliente, así como en una herramienta clave para el desarrollo de negocios competitivos en un mercado cada vez más dinámico. No obstante, la gestión de grandes volúmenes de datos puede dificultar la tarea de categorización y análisis manual, debido a que exige un gran esfuerzo y un tiempo considerable para su procesamiento. Por ello, se hace indispensable el uso de técnicas de minería de textos y procesamiento del lenguaje natural que permitan analizar la información extraída de los usuarios y, de esta manera, estudiar sus opiniones y necesidades e identificar posibles oportunidades de mejora en los productos o servicios ofrecidos.

Es en dicho contexto donde se enmarca el objetivo de este proyecto en el cual, mediante un proceso de modelado de tópicos y de análisis de sentimientos, se analizarán las reseñas de usuarios con el fin de obtener información de utilidad que permita entender el sentimiento y los factores de interés que afectan al consumidor. Mediante el uso de las técnicas mencionadas, este trabajo presenta el análisis de tres casos de uso correspondientes a los siguientes sectores: turismo, con 37.343 reseñas de la Sagrada Familia; restauración, con 139.763 opiniones obtenidas de TripAdvisor sobre restaurantes de Malasia y comercio electrónico, con 59.815 reseñas de teléfonos móviles de Amazon. Estas áreas de aplicación se han seleccionado debido a que representan un volumen significativo de investigaciones en el campo del análisis de reseñas en línea.

En cuanto al sector turístico, los resultados permiten identificar la existencia de diez tópicos que concentran un gran volumen de la discusión, destacando entre ellos los temas relacionados con la compra de entradas y las vidrieras de la basílica. En términos generales, los usuarios expresan un sentimiento mayoritariamente positivo, aunque se han identificado puntos de mejora en la espera para comprar las entradas y la falta de variedad en los medios de accesibilidad. Por otro lado, en el sector de comercio electrónico se han encontrado cinco tópicos relevantes, siendo los temas más abordados aquellos relacionados con la batería y

la cámara de los dispositivos, así como las tarjetas SIM. Los usuarios expresan nuevamente un sentimiento positivo en su mayoría, aunque se han identificado sentimientos negativos en temáticas asociadas a la usabilidad y la vida útil de los dispositivos. Por último, en el sector de la restauración se han modelado trece tópicos, destacando los temas relacionados con las esperas para acceder a los restaurantes, además de opiniones asociadas al servicio y el personal. En términos generales, los usuarios expresan una percepción positiva, aunque se han identificado aspectos a mejorar en la lentitud del servicio y los tiempos de espera.

Abstract

In the last couple of years, the use of technology by users to evaluate and express opinions about products and services has experienced a notable increase. The information generated by this phenomenon is very useful for companies, since it allows them to understand more precisely the behavior of their customers and establish strategies in relation to their products, based on real information. This translates into a unique opportunity to improve quality and customer satisfaction, as well as a key tool for developing a competitive businesses in an increasingly dynamic market. However, managing large volumes of data can make manual categorization and analysis difficult, as it requires considerable effort and time to process. For this reason, the use of text mining and natural language processing techniques is essential to analyze the information extracted from users and, in this way, study their opinions and needs and identify possible opportunities for improvement in the products or services offered.

It is in this context where the objective of this project is framed, in which, through a process of topic modeling and sentiment analysis, user reviews will be analyzed in order to obtain useful information that allows us to understand the sentiment and the factors of interest that affect the consumer. Through the use of the techniques mentioned previously, this project presents the analysis of three use cases corresponding to the following sectors: tourism, with 37,343 reviews of the Sagrada Familia; restaurant, sector with 139,763 opinions obtained from TripAdvisor on Malaysian restaurants and e-commerce, with 59,815 reviews of Amazon mobile phones. These application areas have been selected because they are present in a significant volume of the research done in the field of online review analysis.

Regarding the tourism sector, the results allow us to identify the existence of ten topics that concentrate a large volume of discussion, highlighting among them the issues related to the purchase of tickets and the stained glass windows of the basilica. In general terms, users express a mostly positive sentiment, although points for improvement have been identified in the waiting time when buying tickets and the lack of variety in the means of accessibility. On the other hand, in the e-commerce sector, five relevant topics have been found, the most addressed issues being those related to the device's battery and camera, as well as SIM cards. Users once again express a mostly positive sentiment, although negative sentiments have been identified in issues associated with usability and the use life of the devices. Finally, in the restaurant sector, thirteen topics have been modeled, highlighting issues related to waiting

to access restaurants, as well as opinions associated with service and staff. In general terms, users express a positive perception, although aspects to be improved in the slowness of the service and waiting times have been identified.

Agradecimientos

Quisiera agradecer en primer lugar a mi directora, Alexandra, por su paciencia y dedicación a la hora de ayudarme durante todo el proyecto.

Agradecer también a mis padres y hermanos por todo el cariño y el apoyo que me dan cada día.

Índice general

Índice de figuras

Índice de tablas

Acrónimos

BiLSTM	Bidirectional Long-Short Term Memory
CNN	Convolutional Neural Network
ICAI	Insitituto Católico de Artes e Industrias
LDA	Latent Dirichlet Analysis
ML	Machine Learning
SentiWV	Sentiment Word Vector
STM	Structural Topic Model
SVM	Support Vector Machines
TF-IDF	Term Frequency-Inverse Document Frequency
VADER	Valence Aware Dictionary for sEntiment Reasoning
WDE	Weakly-supervised Deep Embedding

Capítulo 1

Introducción

Estudios preliminares en el campo del marketing y de análisis del comportamiento del consumidor han definido la satisfacción de un cliente como la evaluación subjetiva que este hace de un producto o servicio basado en la comparación de un conjunto inicial de expectativas, con su desempeño en la vida real (Anderson, Fornell, y Lehmann, 1994). Diversas investigaciones han propuesto que este sentimiento de satisfacción del cliente desempeña un papel importante en la motivación de su comportamiento de fidelización, como dar comentarios positivos, o dejar algún tipo de recomendación (Kim, Ng, y Kim, 2009). De hecho, gracias al desarrollo de nuevas tecnologías asociadas a nuevas estrategias de comercialización, se ha facilitado que los consumidores basen sus decisiones de compra en las opiniones que otros clientes han dejado sobre el producto. Un ejemplo de ello, se describe en el trabajo realizado por (Smith y Anderson, 2016) en el cual se indica que en un estudio realizado a 2.000 norteamericanos adultos, el 82 % de los consumidores habría investigado sobre el producto antes de comprarlo. En la Figura 1.1 se muestra un desglose de ese porcentaje en función de las edades de los clientes, siendo los más propensos a mirar reseñas de otras personas los que pertenecen al grupo de edad de 18 a 29 años. Así mismo, los resultados del estudio también indican que en el momento de la compra, los clientes estarían dispuestos a pagar más por un producto que tuviese muy buenas reseñas de otros consumidores. De hecho, el estudio afirma también que estas revisiones podrían afectar a la compra del 80–87 % de compañías, como restaurantes, hoteles y otros servicios (Riaz, Fatima, Kamran, y Nisar, 2019). En esta misma línea de análisis, un estudio realizado por el Centro de Información *China Internet Network* en 2015 evidenció que el 82.1 % de los compradores revisaban las opiniones *online* de otros usuarios y el 41.1 % afirmaban tener en cuenta estas reseñas siempre que iban a realizar una compra *online* (Sun, Niu, Yao, y Yan, 2019).

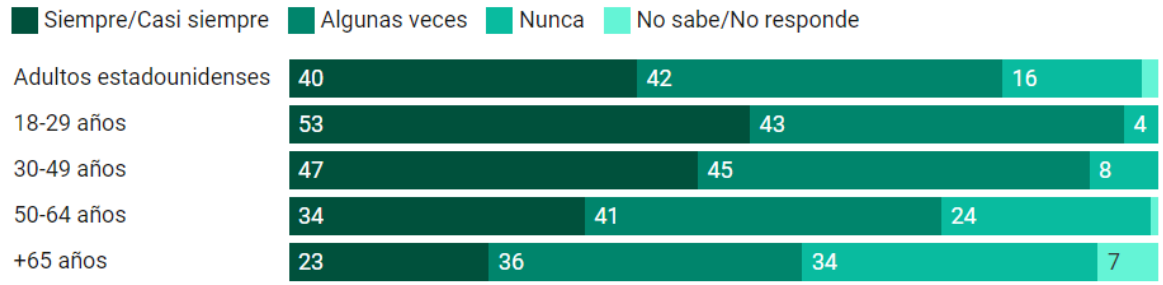


Figura 1.1: Porcentaje de adultos estadounidenses que dicen leer reseñas que otras personas han publicado en línea al comprar algo por primera vez
Fuente: Elaboración propia. Datos: (Smith y Anderson, 2016).

El análisis de estas revisiones *online* facilita la estrategia comercial para productores y vendedores y las compras comparativas de los consumidores individuales, quienes pueden analizar los factores que favorecen y desfavorecen la adquisición de determinados productos. Sin embargo, la evaluación de esta información puede convertirse en una tarea que involucra una gran cantidad de recursos, en términos de tiempo y esfuerzo, si se tienen en cuenta los grandes volúmenes de datos que pueden llegar a generarse. Por esta razón, los métodos de análisis automático de texto han atraído la atención de investigadores y comercializadoras alrededor del mundo. Una primera aproximación para extraer la información han sido los métodos basados en la frecuencia de aparición de ciertas palabras. Sin embargo, aunque estas estrategias son muy intuitivas y son fáciles de implementar, la frecuencia de aparición de las palabras suele resultar en conclusiones ambiguas (Hu, Zhang, Gao, y Bose, 2019). La principal razón de estos resultados se debe a que los usuarios no siempre escriben una opinión completamente negativa, sino que suelen incluir también ciertos aspectos positivos de la experiencia en la misma reseña, o viceversa. Es por ello que diversas investigaciones han enfatizado sobre la importancia de identificar los temas de discusión que aparecen en cada tipo de reseña, así como el análisis del sentimiento involucrado en cada una de las categorías identificadas, con el fin de hacer un análisis estadístico que permita establecer los factores de importancia, en conjunto con el nivel de satisfacción. Estas técnicas son usadas, por ejemplo, por (Mirai, Kannan, y Nobuhiko, 2020) en su artículo donde, mediante el modelado de tópicos sobre una base de reseñas de Sephora, obtuvo diez temas de discusión de los consumidores. De estos temas se identificó cuales estaban más presentes en reseñas positivas y cuales estaban más ligados a reseñas negativas, de esta forma se pueden identificar rápidamente que aspectos del producto hay que mejorar para aumentar la satisfacción del cliente. Estas técnicas no solo son útiles para analizar productos físicos sino que también resultan de gran ayuda para medir la satisfacción del cliente respecto a servicios ofrecidos. Por ejemplo, (Putranto, Sartono, y Djuraidah, 2021) en su artículo analizaron 50.000 reseñas de hoteles en Indonesia, obteniendo cinco tópicos principales. De esos temas obtenidos, el más relevante para el cliente dependía de la categoría del hotel. Con dicho análisis combinado con técnicas

de regresión se pudo obtener también una clasificación numérica para los hoteles, basada en la experiencia del cliente.

Es en este contexto de gran abundancia de datos y de altas tasas de crecimiento electrónico donde se enmarca el desarrollo del presente trabajo, en el cual, con la ayuda de estrategias de modelado de tópicos y análisis de sentimientos, se pretende ofrecer una metodología de análisis para la revisiones *online* de los consumidores de una marca.

1.1. Objetivos

1.1.1. Objetivo General

El objetivo general de este trabajo de fin de grado es analizar las revisiones *online* de los clientes y extraer información relacionada con los principales temas de discusión, así como del nivel de satisfacción de los consumidores de una compañía. Para ello, se llevará a cabo un proceso de modelado de tópicos y de análisis de sentimientos con el fin de ofrecer una metodología de análisis automático, junto con un conjunto de resultados de interés para consumidores, minoristas y potenciales usuarios. La información extraída permitirá comprender mejor el sentimiento y los factores de interés del cliente respecto a la marca, y así ayudará a las partes interesadas a tomar mejores decisiones en relación a su producto o servicio.

1.1.2. Objetivos Específicos

- Analizar la importancia de las reseñas *online* de contenidos generados por los usuarios, en el proceso de toma de decisiones de una compañía.
- Analizar los resultados, previamente documentados en la literatura, sobre estrategias de minería de textos y procesamiento de lenguaje natural, empleadas para el análisis automático de revisiones *online* de los usuarios de una compañía.
- Analizar las categorías de discusión y el nivel de satisfacción de las revisiones *online* de los usuarios de una compañía, en áreas de amplia aplicación, mediante la implementación de estrategias de modelado de tópicos y de análisis de sentimientos.

1.2. Organización de la Memoria

Esta memoria está organizada en cinco diferentes capítulos. En este primer capítulo se ha contextualizado la motivación del desarrollo de este trabajo, detallando el objetivo general y los objetivos específicos del mismo. En el siguiente capítulo se presenta una revisión de la bibliografía sobre estrategias de minería de textos y procesamiento de lenguaje natural, en

relación al análisis automático de revisiones *online* que permitan extraer información sobre la satisfacción de los usuarios.

En el capítulo 3 se lleva a cabo una explicación de la metodología que se ha seguido para elaborar este proyecto. En él se explican los procedimientos desarrollados para la recolección de los datos y su limpieza y pre-procesamiento. También se detalla el análisis descriptivo preliminar realizado, así como el proceso de modelado de tópicos y de análisis de sentimientos. En el capítulo 4 se comentan los resultados obtenidos.

Por último, en el capítulo 5 se exponen las conclusiones del proyecto y futuros desarrollos en esta línea de investigación.

Capítulo 2

Técnicas de minería de textos en el análisis automático de reseñas *online*

Actualmente los clientes escriben reseñas *online* sobre su experiencia durante el proceso de compra y sobre el uso de un producto o servicio (Hou, Yannou, Leroy, y Poirson, 2019). De esta manera, los consumidores evalúan en detalle todas las partes de su compra, incluyendo desde la calidad del producto/servicio, hasta sus servicios de seguimiento de entrega y atención al cliente (Moghaddam y Ester, 2012). En este contexto y teniendo en cuenta que el número de plataformas en Internet y de sitios web de reseñas, propios de las empresas y externos, han aumentado considerablemente, el volumen asociado a la cantidad de revisiones *online* en diferentes ámbitos se ha incrementado también. De ahí que sea más fácil identificar y analizar las evaluaciones y la satisfacción del cliente, mediante el análisis de las reseñas, a través de métodos de minería de textos. Para ello, se suelen abordar dos tipos de estrategias de procesamiento de texto: el análisis de sentimientos y el modelado de tópicos.

El análisis de sentimientos permite la clasificación de las palabras en positivas, negativas o neutras, enfocándose principalmente en las emociones subyacentes del texto. Por su parte, el modelado de tópicos se encarga de agrupar automáticamente las reseñas de los clientes con base en las categorías principales, mediante la implementación de modelos probabilísticos. Teniendo en cuenta el desarrollo de diversas estrategias de procesamiento automático de textos, estos análisis han sido implementados en diferentes aplicaciones en el área del marketing y la analítica empresarial. En términos del análisis de sentimientos, en (Gräbner, Zanker, Fliedl, Fuchs, et al., 2012) se realizó un sistema de clasificación de las reseñas positivas/negativas de clientes de TripAdvisor, basado en *lexicon*, el cual obtuvo un 90 % de precisión. Con este enfoque se destacan también las palabras con mayor frecuencia de las revisiones de los clientes incluyendo: ‘Hotel’, ‘Habitación’, ‘Personal’, ‘Ubicación’, ‘Estancia’ y ‘Desayuno’. Como solución alternativa, (Soni y Sharaff, 2015) utilizan un modelo estocástico, el modelo de Hidden Markov, para identificar el sentimiento (positivo/negativo) en las opiniones de los consumidores sobre diversas categorías de productos (cámara digital, móvil,

ordenador portátil, equipo de música, etc.). Los resultados muestran que el modelo entrenado es prometedor, con métricas de desempeño (exactitud, precisión, *recall*) considerablemente altas. En la misma línea de investigación, (Kumar, Desai, y Majumdar, 2016) realizaron también un análisis de sentimiento de las reseñas dejadas en tres productos (Apple Iphone 5S, Samsung J7 y Redmi Note 3) en Amazon utilizando tres algoritmos distintos: Naïve Bayes, Regresión y SentiWord-Net. Los autores concluyeron que Naïve Bayes presentaba un mejor desempeño en el proceso de clasificación (positivo/negativo) que los otros dos algoritmos. De forma similar, (Guan et al., 2016) realizaron la clasificación de reseñas de 3 categorías de productos (cámaras digitales, teléfonos móviles y ordenadores portátiles), con base en el sentimiento, utilizando *Weakly-supervised Deep Embedding* (WDE). Los resultados obtenidos con el método propuesto se compararon con los de otros enfoques, incluyendo Lexicon, *Support Vector Machines* (SVM), *Sentiment Word Vector* (SentiWV), y *Convolutional Neural Network* (CNN-rand). Las métricas de desempeño mostraron que el método propuesto (WDE) era más efectivo que los demás enfoques implementados.

De forma complementaria, (Sari, Alamsyah, y Wibowo, 2018) realizaron en conjunto un proceso de clasificación manual por temáticas y de análisis de sentimiento, usando un cuestionario de calidad del servicio llamado e-Servqual y el algoritmo Naïve Bayes en 609 reseñas de Tokopedia. El resultado del cuestionario permitía la clasificación de las revisiones en diversas categorías: 'Fiabilidad', 'Personalización', 'Sensibilidad', 'Confianza' y 'Diseño web', y el algoritmo de clasificación permitió identificar que la categoría con mayor sentimiento negativo fue la de 'Personalización', mientras que las dimensiones de 'Confianza' y 'Diseño web' tuvieron la mayor cantidad de sentimientos positivos. En la misma línea, (Vanaja y Belwal, 2018) realizaron un análisis de sentimientos de las revisiones de clientes de Amazon, utilizando dos algoritmos: SVM y Naïve Bayes. Los resultados mostraron un mejor desempeño para el enfoque Naïve Bayes, en la clasificación de revisiones positivas y negativas. Este mismo método fue implementado por (Sharif, Hoque, y Hossain, 2019) y por (Laksono, Sungkono, Sarno, y Wahyuni, 2019) para analizar el sentimiento de un conjunto de 1.000 reseñas de restaurantes bengalíes y de 337 reseñas de 10 restaurantes presentes en TripAdvisor, respectivamente. Los resultados de la clasificación con Naïve Bayes fueron comparados con otros métodos como árboles de decisión, *random forest* o textblob, mostrando en todos los casos un mejor desempeño para el enfoque propuesto. Sin embargo, en un estudio más reciente, (Rajeswari, Mahalakshmi, Nithyashree, y Nalini, 2020) propusieron una estrategia híbrida de análisis de sentimientos de revisiones *online*, combinando técnicas basadas en léxico (SentiWordNet) con técnicas tradicionales de *machine learning* (ML) como Naïve Bayes, árboles de decisión, SVM, y regresión. Esta evaluación se llevó a cabo en tres conjuntos de datos: 2.000 reseñas de Amazon, 1.6 millones de *tweets* obtenidos de Kaggle y 50.000 reseñas de películas de IMDB, y los resultados obtenidos mostraron que la combinación de *lexicon* con SVM y con la regresión logística superaba a los demás enfoques.

Así mismo, (Bachtiar, Paulina, y Rusydi, 2020) han explorado el desempeño de SVM

y Naïve Bayes en el análisis de sentimientos, aumentando el conjunto de datos incluyendo fuentes diferentes como TripAdvisor, Booking, Expedia, Agoda y Pegi-Pegi. Así, un total de 1.561 reseñas fueron analizadas respecto a las siguientes categorías: 'Ubicación', 'Habitación', 'Comida', 'Precio' y 'Servicio'. Los resultados indicaron que los clientes estaban satisfechos con el servicio en la mayoría de categorías, excepto para la asociada a 'Comida', donde se pedía incluir mejoras urgentes. El modelo de clasificación de SVM presentó mejores resultados en términos de precisión, *recall* y *F1-Score*. Otras fuentes de datos incluidas en este tipo de análisis ha sido Twitter. En este sentido, recientemente, (Prananda y Thalib, 2020) analizaron 3.111 *tweets* asociados a la marca de servicios GO-JEK. Los resultados obtenidos mostraron que los consumidores estaban satisfechos con los servicios, obteniéndose 666 reseñas positivas, 2055 neutras y 127 negativas. En este caso de aplicación, se realizó la comparativa entre los modelos: Naïve Bayes, redes neuronales, SVM y árboles de decisión, con este último presentando los mejores resultados en términos de *F1-Score* y precisión. De forma alternativa a los modelos tradicionales de ML, (Sun et al., 2019) realizaron la propuesta de una metodología basada en la minería automática de ontologías de productos difusos (*fuzzy product ontology mining*) para extraer el conocimiento semántico de los comentarios de los clientes en línea con etiquetas positivas o negativas. De esta manera, se analizaron más de 500.000 reseñas *online* de la web Zol y los resultados mostraron que el método propuesto presenta una notable mejora en el rendimiento respecto a otros métodos de referencia, como las reglas de asociación y los campos aleatorios condicionales.

Finalmente, con el auge que han tenido las redes neuronales en los últimos años, principalmente las estrategias de aprendizaje profundo, este enfoque ha sido también utilizado para el análisis de sentimiento en reseñas *online*. Es así como (Hossain, Sharif, Hoque, y Sarker, 2020) realizaron un análisis de sentimiento de 8.435 reseñas de restaurantes bengalíes mediante el uso de *Bidirectional Long-Short Term Memory* (BiLSTM). De esta manera, con el fin de evaluar la efectividad de este modelo, se compararon sus resultados con otras técnicas de ML. Los resultados mostraron que el modelo propuesto superaba a las demás técnicas con una exactitud del 91.35 %, frente a la de las otras técnicas (85 % regresión logística, 81.9 % árbol de decisión, 84.7 % *random forest*, 89.5 % Naïve Bayes y 88.3 % SVM). Por su parte, las técnicas de *clustering* también han sido utilizadas para modelar el sentimiento de una reseña *online*. Un ejemplo de ello es el trabajo realizado por (Riaz et al., 2019), en el cual se realizó el análisis de 1.2 millones de reseñas de seis categorías de productos diferentes de Amazon. Para ello, mediante la técnica de extracción de *keygraph keyword*, se identificaron las palabras clave de las reseñas, y posteriormente mediante el uso del algoritmo de *k-means* se agruparon los datos en base al sentimiento. Los resultados indicaron que la categoría 'Portátil' tenía un mayor sentimiento positivo mientras que 'Cámara' era la categoría con mayor sentimiento negativo.

En términos de la identificación de categorías de opinión, en un trabajo preliminar realizado por (Levy, Duan, y Boo, 2013), se analizaron 1.946 comentarios de una estrella proce-

dentes de diez sitios web de reseñas de usuarios, así como las 225 respuestas de la gerencia de 86 hoteles en Washington, D.C. a diferentes reclamaciones. Con base en este conjunto de datos, se analizaron las características del hotel, el propósito del viaje del usuario, la localización del hotel, entre otros elementos. Así, tras una clasificación manual inicial, se realizó una prueba de chi-cuadrado y una regresión logística para examinar más a fondo las variables que influían en las respuestas específicas de los directivos a las correspondientes reseñas. Los resultados mostraron que las áreas problemáticas abordadas con mayor frecuencia son: 'Baños', 'Limpieza', 'Ruido', '*Check-in*' y 'Aparcamiento', entre otras. Por su parte, uno de los primeros trabajos en implementar modelos probabilísticos para el proceso de modelado de tópicos de revisiones *online* es el propuesto en (Guo, Barnes, y Jia, 2017), dónde los autores describen el uso de *Latent Dirichlet Analysis* (LDA) con el fin de extraer los temas de discusión relacionados con la satisfacción del consumidor. Los resultados indicaron que las categorías más relevantes estaban asociadas a la experiencia en las habitaciones y la calidad del servicio. Además, según el análisis de las reseñas en línea, los clientes de los hoteles de dos y tres estrellas identificaron varias dimensiones importantes no relacionadas con el precio, como el 'Baño' y el *'Check-in'* y *'Check-out'*. Por su parte, se extrajo información relevante para los gestores y propietarios de hoteles de 5 estrellas, indicándoles que el área de mayor relevancia para sus clientes es la 'Sensación de estar en su hogar'.

Posteriormente, (Boo y Busser, 2018) aplicaron Leximancer para extraer los conceptos clave y mostrar la estructura conceptual de 696 reseñas de 173 hoteles. En primer lugar, los autores realizaron un análisis de las características de los hoteles, como el número de estrellas, la pertenencia a una cadena hotelera o su fecha de apertura, identificando 8 categorías diferentes y 811 términos de sentimiento favorables y 19 negativos. Posteriormente, las relaciones entre categorías fueron modeladas, destacando que la conexión entre 'Personal' y 'Recomendación' fue más fuerte que las demás conexiones entre conceptos, lo que podría implicar que el personal puede ser el atributo más influyente al momento de dejar una revisión de recomendación del servicio. Un método alternativo fue propuesto por (Hu et al., 2019), quienes utilizaron *Structural Topic Model* (STM) para analizar 27.864 reseñas de hoteles en Nueva York. Los resultados permitieron identificar 10 categorías, cuyas apariciones en las reseñas negativas eran sustancialmente mayores que las de las reseñas positivas. Además, los resultados del análisis indicaron que las reclamaciones de los clientes de los hoteles de gama alta estaban relacionadas principalmente con problemas de servicio, mientras que en las de los clientes de los hoteles de gama baja estaban asociadas a problemas con las instalaciones. El estudio presentado en (E. Park, Chae, Kwon, y Kim, 2020) también se utilizó este método para analizar 85.505 reseñas de 225 restaurantes obtenidas de la web de Yelp. El objetivo principal estaba orientado a extraer información sobre la opinión de los clientes sobre las prácticas ecológicas de los restaurantes. Los resultados obtenidos mostraron que los clientes reconocían las prácticas ecológicas adoptadas por los restaurantes pero se limitaban únicamente al ámbito de la comida, como puede ser el origen orgánico de las mismas o la

presencia de alternativas veganas. Asimismo, se encontró que aunque había algunas reseñas que mencionaban otras prácticas como el reciclaje, no eran suficientes para llegar a modelar un tópico completo. Finalmente, el estudio concluyó que las categorías que más afectan a la satisfacción del cliente son las relacionadas con el servicio y la calidad de la comida.

En el área del comercio electrónico, (Yang, 2020) implementaron también LDA para realizar el análisis de 430.000 reseñas de aperitivos de distintas compañías. Los resultados permitieron identificar diez categorías sobre las que se realizó un agrupamiento para obtener los aspectos que se deberían mejorar en cada una de las empresas, y de esta manera, aumentar la satisfacción de los consumidores. Este análisis mostró que las marcas debían trabajar en mejorar aspectos como: 'Reputación de la marca', 'Calidad del producto', 'Nivel de servicio', 'Estrategia de marketing' y 'Revisión de la información'. En esta línea de trabajo, (Mirai et al., 2020) propusieron el uso de un modelo híbrido de LDA junto a un método de incrustación de palabras (word2vec) con el fin de analizar 8.551 reseñas de máscara de pestañas de la web de Sephora. El objetivo de este estudio consistió en analizar la relación entre los atributos del producto y la satisfacción del cliente. Los resultados permitieron identificar 9 categorías: 'Desprendimiento', 'Complementario', 'Contenido', 'Cejas', 'Pestañas', 'Prueba', 'Cortesía', 'Too faced', 'Cepillo'. Del análisis, también se encontró que la categoría 'Pestañas' estaba más relacionado con altos niveles de satisfacción, mientras que el tópico 'Cepillo' con niveles más bajos.

En el área de hostelería y turismo, (Adiguzel, Elsherbiny, Quintana, y González-Martel, 2021) y (Putranto et al., 2021) implementaron LDA para analizar 8.376 reseñas de tres hoteles obtenidas de TripAdvisor y 50.000 reseñas de 510 hoteles en Indonesia, respectivamente. (Adiguzel et al., 2021) identificaron 20 tópicos de discusión, siendo 'Amantes de los hoteles', 'Personal', 'Clientes que regresan', 'Calidad del servicio' y 'Conveniencia de la ubicación' los más importantes. Por su parte, las categorías 'Vida Nocturna' y 'Animación Hotelera' contaban con menor relevancia. Por otro lado, (Putranto et al., 2021) identificaron cinco tópicos principales: 'Precio comida', 'Servicios', 'Ubicación', 'Instalaciones' y 'Comodidad', siendo 'Servicios' el tema más comentado en todos los hoteles menos en los de tres estrellas, en los cuales el tema más comentado fue 'Precio comida'. Finalmente, (S. Park, Cho, Park, y Shin, 2021) propusieron en su artículo un análisis de las opiniones de los clientes mediante propagación de sentimiento, para incluir también el contexto de las palabras en el análisis. Para ello analizaron 311.550 reseñas de coches de una empresa Coreana, provenientes de diez páginas web distintas. Los resultados mostraron que los temas negativos más comentados por los clientes fueron: 'Ruido', 'Mantenimiento y reparación', 'Rendimiento de conducción' y 'Tren motriz y transmisión'. Particularmente, este análisis es de gran interés pues incluyó en el procesamiento términos que no son gramaticalmente correctos pero que los clientes usan tradicionalmente como jerga en internet.

En la Tabla 2.1 se muestra un resumen de este estudio de la literatura en relación al análisis automático de las reseñas *online* de los usuarios. En ella se indican la referencia, el

tamaño del conjunto de datos y la fuente, el algoritmo utilizado, el sector de aplicación y los resultados obtenidos. Del resumen mostrado, es importante destacar que cada una de las investigaciones incluidas han analizado conjuntos de datos provenientes de fuentes diferentes y con un tamaño de observaciones que varía en cada uno de los casos. Debido a esta gran diversidad en los datos de entrada, no es posible concluir sobre el desempeño de los algoritmos teniendo en cuenta solamente las métricas de desempeño indicadas por las diversas investigaciones. Sin embargo, a partir de la revisión si que es posible identificar las técnicas más utilizadas tanto para tareas de modelado de tópicos como de análisis de sentimientos. En el caso particular del modelado de tópicos, es posible observar que la técnica más utilizada es LDA. El uso extendido de está técnica se debe principalmente a ventajas como la posibilidad de tratar documentos de longitudes variables, y la simplicidad en su implementación, pues es una técnica que no necesita de datos de entrenamiento y produce información que es más semánticamente interpretable que otros métodos (Albalawi, Yeap, y Benyoucef, 2020). Asimismo, esta estrategia permite encontrar tópicos coherentes cuando se ajustan correctamente sus parámetros, y el número de tópicos identificados suele ser menor que el obtenido con técnicas de incrustación de palabras, lo que permite que sean más interpretables (Egger y Yu, 2022). Por su parte, para obtener el análisis de sentimientos, las técnicas más usadas son aquellas que incluyen reglas de decisión. Estas pueden estar directamente asociadas al uso de diccionarios como Lexicon o técnicas tradicionales de ML como puede ser Naïve Bayes o SVM. En términos de las estrategias basadas en diccionarios, su uso difundido en diversas líneas de investigación se debe a la simplicidad del modelado, lo que facilita su implementación al no necesitar datos etiquetados o de un proceso de entrenamiento, para obtener los resultados del análisis (Devika, Sunitha, y Ganesh, 2016).

Teniendo en cuenta que el objetivo de este proyecto consiste en el análisis de las reseñas *online* de los clientes, con el fin de extraer información de los principales temas tratados, así como del nivel de satisfacción de los mismos, se considerará LDA para el modelado de tópicos y técnicas basadas en diccionarios, como VADER, para el análisis de sentimientos. Esta decisión se basa en los resultados del estudio de la literatura realizado, donde se destaca el uso extendido de estas técnicas en esta área de investigación.

Referencia	Tamaño del Data Set (Fuente)	Tarea	Algoritmo	Sector	Resultados
(Gräbner et al., 2012)	10.969 (TripAdvisor)	Análisis de sentimiento, clasificación	Lexicon	Turismo	Palabras con mayor frecuencia: 'Hotel', 'Habitación', 'Personal', 'Ubicación', 'Estancia' y 'Desayuno'. Precisión Clasificación: 90 %
(Soni y Sharaff, 2015)	No especificado (Amazon)	Análisis de sentimiento, clasificación	Modelo de Hidden Markov	Comercio electrónico	Precisión del 95.7 % y exactitud del 93.5 % para una proporción de los datos de entrenamiento del 0.8
(Kumar et al., 2016)	No especificado (Amazon)	Análisis de sentimiento, clasificación	Naïve Bayes, Regresión y SentiWordNet	Comercio electrónico	Naïve Bayes: (Precisión, *Recall*) Apple Iphone 5S (67.5 %, 87 %), Samsung J7 (55.7 %, 77 %), Redmi Note 3 (50.4 %, 79.6 %).
(Guan et al., 2016)	1.1 millones (Amazon)	Análisis de sentimiento, clasificación	WDE	Comercio electrónico	Exactitud WDE: 87.7 %
(Sari et al., 2018)	609 (Tokopedia)	Análisis de sentimiento, clasificación	Naïve Bayes	Comercio electrónico	Dimensiones modeladas: 'Fiabilidad', 'Personalización', 'Sensibilidad', 'Confianza' y 'Diseño web'. Exactitud clasificación: 90 %
(Vanaja y Belwal, 2018)	No especificado (Amazon)	Análisis de sentimiento, clasificación	Naïve Bayes, SVM	Comercio electrónico	Exactitud Naïve Bayes: 90.423 %, Exactitud SVM: 83.423 %
(Sharif et al., 2019)	1.000 (Internet)	Análisis de sentimiento, clasificación	Multinomial Naïve Bayes	Restauración	Multinomial Naïve Bayes Exactitud: 80.48 % con validación cruzada K=6
(Laksono et al., 2019)	337 (TripAdvisor)	Análisis de sentimiento, clasificación	Naïve Bayes	Restauración	Naïve Bayes Precisión: 72.06 %, Textblob Precisión: 69.12 %
(Rajeswari et al., 2020)	2.000 (Amazon), 1.6 millones (Kaggle) y 50.000 (IMDB)	Análisis de sentimiento, clasificación	Naïve Bayes, Árboles de decisión, SVM, Regresión	Varios	Exactitud Lexicon + SVM: 69 %, Exactitud Lexicon + Regresión Logística: 72 %
(Bachtiar et al., 2020)	1.561 (TripAdvisor, Booking, Expedia, Agoda y Pegi-Pegi)	Análisis de sentimiento, clasificación	SVM, Naïve Bayes	Turismo	Categorías analizadas (Exactitud SVM, Naïve Bayes): 'Ubicación' (93 %, 92 %), 'Habitación'(80 %, 78 %), 'Comida'(68 %, 64 %), 'Precio'(90 %, 85 %) y 'Servicio'(92 %, 88 %).
(Prananda y Thalib, 2020)	3.111 (Twitter)	Análisis de sentimiento, clasificación	Redes neuronales , SVM, árbol de decisión y Naïve Bayes	Servicios electrónicos	Precisión árboles de decisión: 55 %, *F1-score*: 55 %
(Sun et al., 2019)	500.000 (Zol)	Análisis de sentimiento, clasificación	Minería de ontologías de productos difusos	Comercio electrónico	Características más comentadas: 'Apariencia', 'Batería', 'Pantalla', 'Sistema' y 'Red'.
(Hossain et al., 2020)	8.435 (Facebook, grupos de internet y Yelp)	Análisis de sentimiento, clasificación	BiLSTM	Restauración	Exactitud BiLSTM: 91.35 % .
(Riaz et al., 2019)	1.2 millones (Amazon, ebay, alibaba)	Análisis de sentimiento, *clustering*	Keygraph keyword y k-means	Comercio electrónico	Pureza de la agrupación: 0.97, Exactitud: 70 %
(Levy et al., 2013)	1.946 reseñas y 225 respuestas (Internet)	Análisis de contenido, regresión	Test de chi-cuadrado, Regresión Logística	Turismo	Áreas problemáticas más frecuentes: 'Baños', 'Limpieza', 'Ruido', *'Check-in'*, 'Aparcamiento', 'Restaurantes', 'Facturación', 'Tamaño de la habitación' y 'Personal de limpieza'.

Referencia	Tamaño del Data Set (Fuente)	Tarea	Algoritmo	Sector	Resultados
(Guo et al., 2017)	266.544 (Trip Advisor)	Modelado de tópicos	LDA	Turismo	Categorías más relevantes: 'Problemas *check-in y check-out*',' Malas revisiones', 'Instalaciones', 'Satisfacción', 'Mala comunicación', 'Hogareño'
(Boo y Busser, 2018)	696 (Internet)	Análisis de contenidos	Leximancer	Turismo	Categorías clave: 'Personal', 'Reunión', 'Hotel', 'Propiedad', 'Trabajo', 'Ubicación', 'Recomendación' y 'Servicios'.
(Hu et al., 2019)	27.864 (TripAdvisor)	Modelado de tópicos	STM	Turismo	Categorías negativas modeladas: 'Fallos severos del servicio', 'Suciedad', 'Reserva y cancelación', 'Tipo de habitación', 'Cargos extra', 'Instalaciones de la habitación', 'Ruido', 'Errores', 'Comparación de experiencias' e 'Instalaciones públicas'.
(E. Park et al., 2020)	85.505 (Yelp)	Modelado de tópicos	STM	Restauración	40 categorías identificadas: 'Ingredientes locales/orgánicos', 'Menú Vegano', 'Precio de la comida', 'al servicio', 'Clasificación por estrellas de los restaurantes', 'Buena comida y servicio', entre otras
(Yang, 2020)	435.350 (BESTORE, BE&CHEERY, Three Squirrels, Qiaqia, y Wolong)	Modelado de tópicos	LDA	Comercio electrónico	Categorías modeladas: 'Reputación de la marca', 'Calidad del producto', 'Nivel de servicio', 'Estrategia de marketing' y 'Revisión de la información'.
(Mirai et al., 2020)	8.551 (Sephora)	Modelado de tópicos	Word2vec y LDA	Comercio electrónico	Tópicos modelados: 'Desprendimiento', 'Complementario', 'Contenido', 'Cejas', 'Pestañas', 'Prueba', 'Cortesía', 'Too faced', 'Cepillo'.
(Adiguzel et al., 2021)	8.376 (TripAdvisor)	Modelado de tópicos	LDA	Turismo	Categorías más relevantes: 'Amantes de los hoteles', 'Personal', 'Clientes que regresan', 'Calidad del servicio' y 'Conveniencia de la ubicación'.
(Putranto et al., 2021)	50.000 (Internet)	Modelado de tópicos, regresión	LDA	Turismo	Tópicos principales: 'Precio|Comida', 'Servicios', 'Ubicación', 'Instalaciones' y 'Comodidad'.
(S. Park et al., 2021)	311.550 (Internet)	Modelado de tópicos y análisis de sentimiento	LDA	Comercio	Quejas más comunes: 'Ruido', 'Mantenimiento', 'Reparación', 'Rendimiento de conducción', 'Tren motriz' y 'Transmisión'.

Tabla 2.1: Artículos que abordan el análisis de sentimiento y el modelado de tópicos en reseñas *online* de consumidores de productos y servicios.

Capítulo 3

Metodología de Análisis

En este capítulo se describe el proceso metodológico abordado en la realización del análisis cuantitativo presentado en este trabajo. De esta manera, el desarrollo de este estudio considera cinco etapas de análisis, descritas a continuación (ver Figura 3.1):

1. **Selección y recolección de datos**. En esta etapa se realiza un breve análisis de los conjuntos de datos abiertos disponibles y se realiza la selección de los datos a considerar en este trabajo. Es importante destacar que la elección se basa en los sectores de aplicación más estudiados en el estado del arte del análisis automático de reseñas *online* (ver Tabla 2.1).

2. **Pre-procesamiento de los datos**. Teniendo los conjuntos de datos seleccionados, esta etapa pretende asegurar la mayor calidad en la información recolectada. Así, esta fase involucra la limpieza y transformación de los datos requeridos para el modelado de tópicos y el análisis de sentimientos

3. **Análisis descriptivo de N-gramas**. Esta etapa consiste en obtener un análisis descriptivo de las palabras, bigramas (conjuntos de dos palabras) y trigramas (conjuntos de tres palabras) más importantes en cada conjunto de datos analizados. Los resultados permitirán obtener un análisis descriptivo preliminar de los conceptos más relevantes discutidos en cada conjunto de datos.

4. **Modelado de tópicos**. Esta etapa se encarga de obtener un modelo que permita categorizar el conjunto de datos en un conjunto de temáticas relevantes al sector de aplicación. Para ello, se realizará la implementación del algoritmo *Latent Dirichlet Analysis* (LDA) debido a sus amplias ventajas computacionales y a su gran aplicabilidad en trabajos de investigación relacionados al tema de estudio.

5. **Análisis de sentimientos**. Finalmente, esta etapa pretende analizar el carácter positivo o negativo de las emociones subyacentes en los comentarios online de los clientes.

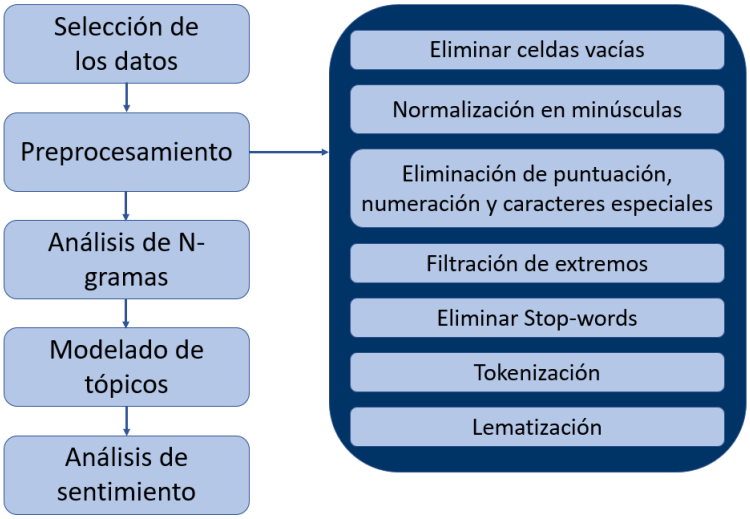


Figura 3.1: Etapas de la metodología seguida en este estudio
Fuente: Elaboración propia

A continuación se presentan los detalles del desarrollo de cada una de las etapas resumidas anteriormente.

3.1. Selección de datos

Para la realización del análisis de las reseñas *online* de los clientes se han elegido varios conjuntos de datos, pertenecientes a distintos sectores de aplicación. Así, la selección de los datos a analizar se orienta principalmente a los sectores con mayor análisis mostrados en el estado del arte. De esta manera, con base en los resultados del capítulo 2, las categorías más analizadas incluyen: Restauración, Turismo y Comercio electrónico. Teniendo en cuenta esta consideración, se ha realizado la búsqueda de conjuntos de datos abiertos en repositorios de datos ampliamente usados en el sector de la analítica de datos, incluyendo Kaggle, data.world, Github, y Google Dataset Search. La razón de incluir esta búsqueda se basa en el hecho de que la mayoría de los artículos considerados en el estado del arte usan un conjunto de datos específico, lo que dificulta la comparación entre metodologías y su respectiva evaluación. De esta manera, se busca encontrar un conjunto de datos que tenga el mayor número de observaciones posible, así como una amplia diversificación del texto recolectado que permita que los modelos implementados puedan ser usados posteriormente con mayor confianza en el área de aplicación. Así mismo, se han seleccionado solo conjuntos de datos en inglés, debido a la gran difusión de este idioma, y a la alta calidad de análisis que tienen las librerías de análisis automático de texto en inglés. Así, en la Tabla 3.1 se resumen las características de los conjuntos de datos preseleccionados. Para cada categoría se ha escogido un conjunto de datos teniendo en cuenta las consideraciones descritas anteriormente.

Nombre del Data Set	Tamaño del Data Set	Ciudad	Elemento	Sector
Monuments Datasets	43.540	Barcelona	Sagrada Familia	Turismo
Hotel Reviews	38.932	Varias ciudades de E.E.U.U.	Varios hoteles	Turismo
Reviews of London-based hotels	27.330	Londres	Diez hoteles	Turismo
Trip Advisor Hotel Reviews	20.491	Varias ciudades	Varios hoteles	Turismo
Marina Bay Sands Hotel Reviews on Tripadvisor	10.232	Singapur	Marina Bay Sands Hotel	Turismo
Amazon Cell Phones Reviews	67.965	No especificado	Móviles de distintas marcas (Apple, Google, HUAWEI..)	Comercio electrónico
Amazon Earphones Reviews	14.337	No especificado	Diez tipos de auriculares inalámbricos	Comercio electrónico
Restaurant Review Sentiment Analysis	10.000	No especificado	Restaurante	Restauración
Opinion Mining in Costumer Reviews for McDonalds Restaurants	1.525	Varias ciudades de E.E.U.U.	Reseñas McDonalds	Restauración
TripAdvisor Restaurants Info for 31 Euro-Cities	1.666	Ciudades europeas	Varios restaurantes	Restauración
Restaurant Reviews in Dhaka, Bangladesh	16.151	Dhaka	Varios restaurantes	Restauración
Malaysia Restaurant Review Datasets	139.763	Varios	Varios restaurantes	Restauración

Tabla 3.1: Conjuntos de datos preseleccionados para el análisis.

3.2. Pre-procesamiento de los datos

Como se explicó anteriormente, en la etapa de pre-procesamiento se llevarán a cabo tareas de limpieza y tratamiento de datos, necesarias para obtener una calidad aceptable de los mismos y de esta manera, realizar un análisis satisfactorio en etapas posteriores. Para ello, en primer lugar, se eliminan los valores que faltan dentro del conjunto de datos, equivalentes a todas aquellas observaciones que contengan valores nulos o estén vacías. Teniendo en cuenta el resultado del análisis anterior, y considerando el modelo elegido para la caracterización de los tópicos, se realiza la normalización en minúsculas del texto, se elimina puntuación, numeración y caracteres especiales. Una vez realizada esta limpieza, se procede a la filtración de los extremos, es decir, se eliminan palabras que se repiten mucho (están presentes en más del 95 % de las reseñas) o se repiten muy poco (menos del 2 % en todo el corpus), ya que en ninguno de esos casos, corresponden a términos que puedan aportar información al análisis.

Posteriormente, se procede a la eliminación de las palabras vacías (*stopwords*), definidas como palabras de uso muy común en el lenguaje como pueden ser 'yo', 'como', 'esos', etc. La razón por la que esta tarea es fundamental en tareas de análisis automático de texto es que, si se eliminan las palabras que se utilizan con mucha frecuencia en un idioma determinado, es posible centrar el objetivo de análisis en las palabras importantes que definen el contenido de los datos.

Finalmente, se realizan las tareas asociadas a la tokenización y lematización del conjunto de datos. La primera consiste en dividir el texto en unidades más pequeñas llamadas *tokens*. De esta forma, una reseña que incluya el texto: 'Este producto me encanta' quedaría dividido después de esta etapa en: 'Este', 'producto', 'me', 'encanta'. Por otro lado, la lematización consiste en hallar el *lema* de cada una de las palabras. Se define el lema como la forma más básica de una palabra, mediante la cual, la encontraríamos como entrada en un diccionario. De esta manera, por ejemplo, el lema de la palabra 'viajando' correspondería a 'viajar'.

3.3. Análisis descriptivo de N-gramas

El análisis incluido en esta etapa consiste en analizar la frecuencia y la relevancia de ciertas palabras o grupos de palabras presentes en el texto. Así, un unigrama sería el resultado de estudiar una sola palabra y su relevancia en el conjunto de observaciones. Un bigrama sería el análisis de dos palabras consecutivas y un trigrama, el de tres palabras consecutivas. Con el fin de analizar la respectiva relevancia de los n-gramas, no se analizará únicamente la frecuencia de aparición de estos términos, si no que se cuantificará además la importancia de los mismos, respecto al resto del documento. Para ello se va a utilizar la frecuencia de término – frecuencia inversa de documento (*Term Frequency-Inverse Document Frequency* - TF-IDF). (Huang et al., 2021) definen este método como una forma de conocer la importancia

de ciertos términos, mediante el modelado y el cálculo de su peso respecto al resto. Esta herramienta es útil ya que al incluir el término IDF cobran más importancia aquellos términos que son más relevantes, mientras que las palabras más comunes y que pueden comportarse como palabras vacías, pierden importancia al tener un menor IDF. En ese mismo artículo, los autores definen la forma de calculo en dos partes:

- *Term frequency* (TF). Se define como el número de veces que una palabra (w) aparece en un documento (d), entendiendo cada documento como cada reseña analizada.

$$TF(w, d) = \text{count}(w, d) \tag{3.1}$$

- *Inverse Document Frequency* (IDF): Este concepto tiene en cuenta el número de veces que se usa una palabra (w) en un conjunto de reseñas, asignándole un valor más bajo a los términos más habituales. Se calcula como:

$$IDF(w) = log\left(\frac{M+1}{df(w)}\right), \tag{3.2}$$

  donde M sería el número de documentos totales y df(w) sería el número de documentos que contienen la palabra w.

Considerando los términos previamente expuestos, el TF-IDF se obtendría multiplicando ambos términos:

$$TF\text{-}IDF = TF(w, d) \times IDF(w) \tag{3.3}$$

3.4. Modelado de tópicos

Con el fin de obtener un modelo de categorización automática de las reseñas analizadas, se utilizará el modelo *Latent Dirichlet Allocation* (LDA), una técnica de aprendizaje automático no supervisado que identifica los temas latentes en documentos. Se ha escogido este modelo ya que, como se ha comprobado en el capítulo 2, se trata de una técnica ampliamente utilizada en la literatura para realizar tareas asociadas al modelado de tópicos. Asimismo, presenta una serie de ventajas como la posibilidad de tratar documentos de longitudes variables, y de obtener resultados sin necesidad de datos de entrenamiento. Además sus resultados están asociados a conclusiones semánticamente más interpretables, lo que generalmente se relaciona con un menor número de tópicos en comparación con otros métodos (Albalawi et al., 2020; Egger y Yu, 2022). Esta metodología fue propuesta por (Blei, Ng, y Jordan, 2003), quienes basan el modelo en la idea de representar cada documento (cada reseña en nuestro caso) como mezclas aleatorias de temas o tópicos latentes. Estos temas a su vez son caracterizados mediante un modelo de distribución de diversas palabras. Ambas distribuciones de probabilidad son representadas mediante un modelo de Dirichlet.

En la Figura 3.2 se muestra una representación gráfica del modelo, donde se pueden observar las dependencias entre los distintos parámetros. En función de si estos parámetros se encuentran dentro de uno de los rectángulos o no indica se indica si estos se aplican a nivel de corpus (conjunto de todas las observaciones), de documento o de palabra. A continuación se define cada uno de ellos:

- M: denota el numero total de documentos presentes en el corpus, en nuestro caso el número total de reseñas de clientes.
- N: denota el número de palabras dentro de un documento.
- α y β: son los parámetros de la distribución de Dirichlet. El primero indica la distribución de temas por documento. Así, un α alto indica que es probable que cada documento contenga una combinación de una mayor cantidad de temas. β, por su parte, está relacionado con la distribución de palabras por tema. Un valor alto de este parámetro indica que cada tema contendrá una combinación de una mayor cantidad de palabras.
- θ_m: representa la distribución de tópicos para un documento m.
- z_{mn}: representa el tópico para la n-sima palabra en el documento m.
- w_{mn}: caracteriza la n-sima palabra en el documento m.

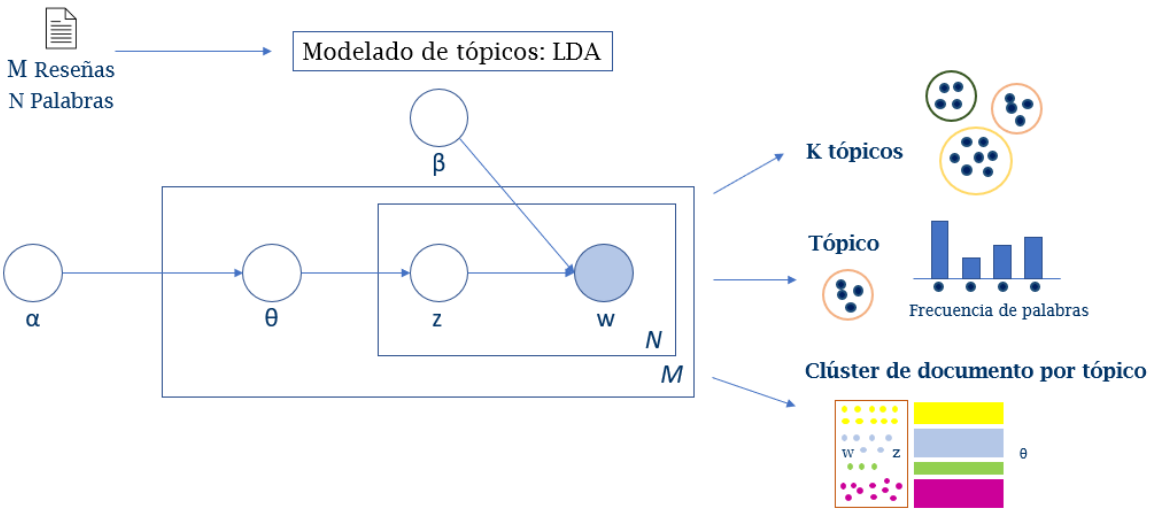


Figura 3.2: Representación gráfica del modelo de LDA
Fuente: Elaboración propia

El número de tópicos *K* y los parámetros α y β son parámetros de entrada del modelado, mientras que el resto son un resultado del mismo. Teniendo en cuenta estos parámetros, el modelo se va ajustando mediante un proceso iterativo por documento y por palabra. Para ello, por un lado, se calcula la proporción de palabras de todo el documento que están asignadas a cada tópico *k* y la proporción de otros documentos donde esa misma palabra también ha sido asignada al tópico *k*. En función del resultado del producto entre ambas proporciones, se ajusta la probabilidad de asignación de cada palabra a cada tópico.

Por último, para medir la bondad del ajuste del modelo al conjunto de datos usado se va a emplear el índice de coherencia, que equivale a una puntuación media de la similitud de palabras por pares de las M palabras más relevantes del tema. La implementación de esta métrica crea vectores de contenido de palabras utilizando sus co-ocurrencias y, después, calcula el índice utilizando la información mutua normalizada por puntos o NPMI y la similitud del coseno (Tolegen, Toleu, Mussabayev, y Krassovitskiy, 2022). Cuanto mayor sea el número, mejor será la puntuación de coherencia. Para obtener el valor global se calcula la media de las puntuaciones de coherencia por pares de las N palabras principales que describen cada tópico. Con ello, si se obtiene un coeficiente de coherencia alto significa que las palabras de un mismo tema están estrechamente relacionadas. En este trabajo, se utilizará esta métrica para elegir el número de tópicos que mejor representa cada uno de los corpus analizados. Para ello, se obtendrá una gráfica evidenciando el cambio en el coeficiente de coherencia en función de diferentes valores para el número de tópicos. A partir de este gráfico, se escogerá el valor *k* para el cual el coeficiente de coherencia muestre el final de un crecimiento rápido, ya que suelen ser los que producen tópicos más fácilmente interpretables. Es necesario tener en cuenta que un número de tópicos muy bajo puede traducirse en tópicos distintos que ha sido agrupados juntos. Mientras que un valor de *k* muy elevado tampoco produce resultados idóneos ya que causa la aparición de subgrupos iguales o muy similares.

Una vez se realiza la selección del número de tópicos *k*, se obtiene el mapa de distancia intertópica. En dicho gráfico el área de cada circulo representa la importancia de dicho tópico con respecto al resto. Los centros de dichos círculos se establecen calculando la distancia entre tópicos, para posteriormente proyectarla en dos dimensiones mediante una escala multidimensional (Sievert y Shirley, 2014). Por ello, una buena elección de *k* conllevaría a un mapa de distancia, dónde los círculos no estén superpuestos y estén distribuidos a lo largo del mismo. De esta manera, esta estrategia será utilizada para evaluar el modelo de tópicos obtenido para cada conjunto de datos. En la Figura 3.3 se muestra un ejemplo gráfico de un mapa de distancia intertópica. En él se puede observar como los tópicos 2 y 4 son similares entre sí, están cerca uno del otro y hay un solapamiento parcial entre ellos. Sin embargo, los tópicos 1 y 3 son muy diferentes y serían fácilmente separables entre sí, razón por la cual se encuentran alejados uno del otro en el mapa.

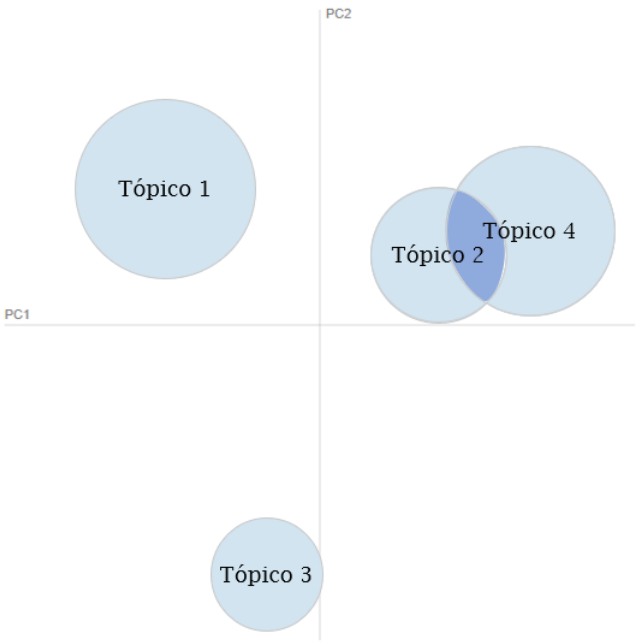


Figura 3.3: Ejemplo de mapa de distancia intertópica
Fuente: Elaboración propia

3.5. Análisis de sentimientos

Finalmente, el proceso de análisis de sentimientos consiste en cuantificar las emoción subyacente en las reseñas dejadas por los clientes, mediante la clasificación de las palabras en positivas, negativas o neutras. Para llevar a cabo este análisis se va a emplear *Valence Aware Dictionary for sEntiment Reasoning* (VADER). Según explican (Hutto y Gilbert, 2014) en su artículo, se trata de un modelo para realizar análisis de sentimiento en el que se combinan características léxicas con cinco reglas que engloban las convenciones sintácticas y gramaticales que se suelen utilizar para poner mayor énfasis en la expresión de sentimiento. En esta metodología, la técnica de *Wisdom of the crowd* fue usada para establecer una estimación de la intensidad de cada una de las características escogidas. Estas características contenían numerosas expresiones comunes para expresar sentimiento como los emoticonos, acrónimos o jergas específicas. Estas características fueron evaluadas en una escala de -4 (extremadamente negativo) a 4 (extremadamente positivo). Además, el estudio permitió identificar diversas características del texto que afectan a la percepción de su correspondiente sentimiento. De esta manera, se encontraron cinco heurísticas generalizables que cambiaban el sentimiento percibido en el texto:

- *Puntuación*. En concreto el símbolo de exclamación permite modificar la intensidad del sentimiento expresado.

- *Mayúsculas.* El uso de una palabra entera en mayúscula permite cambiar la intensidad del sentimiento que se quiere expresar sin cambiar la semántica del texto.

- *Adverbios de grado.* Afectan a la intensidad del sentimiento que se quiere expresar, intensificándola o disminuyéndola. Por ejemplo, 'Esta comida esta buena' frente a 'Esta comida está realmente buena'.

- *La conjunción 'pero'*. Esta indica un cambio en la polaridad del sentimiento, siendo el predominante el que va después de la conjunción. Por ejemplo, 'La comida esta buena, pero el servicio es horrible' tiene una mezcla de dos polaridades pero predomina la mostrada en la segunda mitad de la frase.

- *Negación de polaridad.* Es posible captar hasta el 90 % de los casos en los que la negación cambia la polaridad del texto fijándose en las tres palabras que la preceden. Por ejemplo, 'El servicio aquí no es realmente bueno'.

En este estudio, se ha decidido emplear el modelo VADER, ya que de acuerdo a los resultados presentados en el capítulo 2, se trata de una técnica muy utilizada en la literatura para llevar a cabo el análisis de sentimiento en el corpus. Su creciente uso en investigación se debe a un conjunto de ventajas, entre las que se encuentran sus altas métricas de desempeño en el análisis de datos de redes sociales, el soporte ofrecido para el análisis de emoticonos, así como la posibilidad de obtener resultados confiables sin requerir datos de entrenamiento (Bonta y Janardhan, 2019). En este proyecto se va a utilizar esta estrategia para calcular la puntuación compuesta o *compound score*. Esta se calcula añadiendo la puntuación de cada palabra de acuerdo a las reglas mencionadas anteriormente y normalizando el resultado para obtener una puntuación entre 1 (muy positivo) y -1 (muy negativo).

En la sección 3.2 se han explicado las tareas asociadas al pre- procesamiento de los datos. Entre ellas, se encuentran la normalización de palabras, y la eliminación de caracteres especiales y *stopwords*. Sin embargo, como se ha podido comprobar en la descripción del modelo VADER, esta estrategia es sensible a características del texto como pueden ser las mayúsculas, los símbolos de puntuación o la presencia de ciertas conjunciones. Por esta razón, a la hora de realizar el análisis de sentimientos no se van a incluir estas etapas en el pre-procesado de los datos.

Capítulo 4

Resultados Experimentales

En este capítulo se presentan los resultados de la metodología, expuesta en el capítulo 3, sobre cada uno de los conjuntos de datos seleccionados y analizados en este trabajo. Como se mencionó previamente, se han seleccionado tres sectores de aplicación, teniendo en cuenta su amplio estudio en la literatura: Turismo, Restauración y Comercio Electrónico. El acceso al repositorio con todo el código utilizado para los análisis descritos en el presente trabajo se encuentra disponible en (Tabasco, 2023).

4.1. Turismo

4.1.1. Selección y pre-procesamiento de los datos

El conjunto de datos seleccionado contiene reseñas de visitas a la Basílica de la Sagrada Familia, ubicada en la ciudad de Barcelona, España (Valdivia et al., 2018). Estas reseñas van desde mayo de 2011 a octubre de 2016 y cuentan con un total de 43.533 observaciones. Se ha elegido este *dataset* por dos motivos principalmente: en primer lugar, debido al gran volumen de observaciones disponibles, lo que contribuye a la obtención de mejores resultados y de conclusiones más fiables; en segundo lugar, el enfoque y su aplicación directa al área del turismo.

Con el fin de realizar un análisis descriptivo preliminar de la información, en primer lugar, se elaboró un estudio de la distribución del número de palabras por reseña. De esta manera, en la Figura 4.1 se muestra información relevante de la distribución, eliminando valores atípicos encontrados en las observaciones del conjunto de datos. El resultado evidencia que la longitud de las reseñas es en su mayoría corta, con una mediana de 50 palabras y con un 75 % de las reseñas conteniendo menos de 82 palabras. También se puede observar que la parte superior a la mediana es más amplia a la inferior, mostrando una asimetría a la derecha de los datos lo que indica una mayor concentración en valores bajos de la distribución (opiniones cortas) que en valores altos.

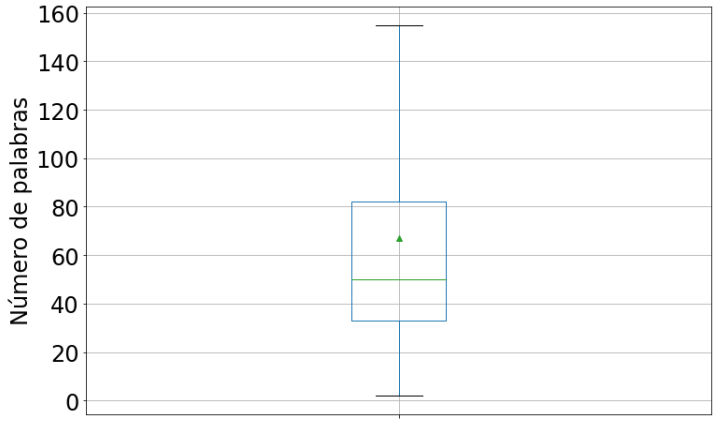


Figura 4.1: Diagrama de cajas del número de palabras por reseña - Sector turismo
Fuente: Elaboración propia

Asimismo, se ha construido un gráfico que muestra la evolución del número de reseñas a lo largo del tiempo (ver Figura 4.2). Se puede observar una tendencia creciente en el número de reseñas, lo cual puede deberse principalmente al aumento del uso de herramientas tecnológicas para compartir opiniones sobre productos o servicios, como se ha comentado en el capítulo 1. Por otro lado, en la serie de tiempo se observa también un componente estacional, produciéndose un aumento a partir del mes de abril y un descenso a partir del mes de octubre, aproximadamente. Esto se debe a que durante los meses de verano, el turismo en Barcelona aumenta (Observatori del Turisme a Barcelona: ciutat i regió, 2021), por lo que también se incrementa el número de reseñas realizadas por los usuarios.

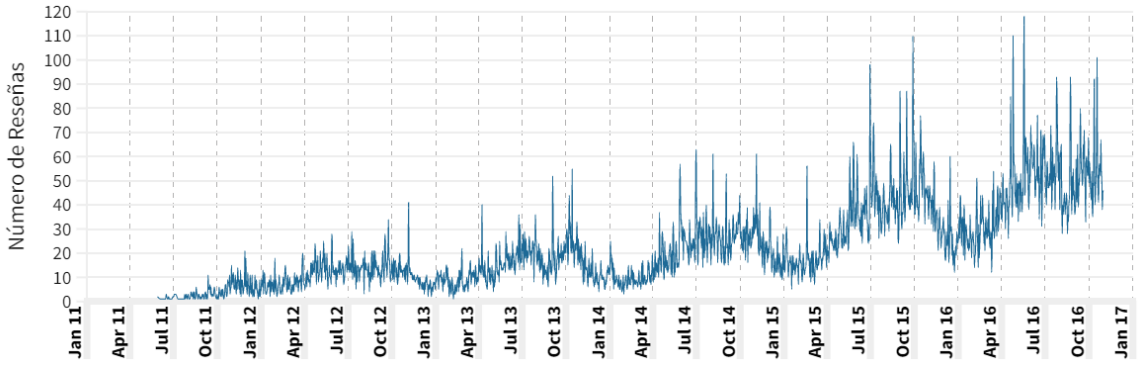


Figura 4.2: Serie temporal del número de reseñas - Sector turismo
Fuente: Elaboración propia

Una vez se obtienen los datos de forma estructurada, es necesario llevar a cabo las tareas de limpieza y pre-procesamiento descritas en la sección 3.2. Esta etapa incluye la eliminación de observaciones vacías, de signos de puntuación, numeración y de caracteres especiales. Además, el texto se normaliza de forma que todas las palabras se encuentren en minúsculas y se procede a la filtración de los extremos. Este procedimiento implica la eliminación de palabras que aparecen muy poco (menos de 2 veces en todo el corpus) y palabras que se repiten mucho (están presentes en más del 95 % de las reseñas), ya que en ninguno de los

casos son palabras que aporten información al análisis. El conjunto de datos inicial contenía 43.533 reseñas, de las cuales una vez finalizadas las tareas de limpieza y pre-procesamiento se reducen a 37.343, por lo que se ha eliminado aproximadamente el 15 % de los datos respecto a los que se adquirieron inicialmente.

4.1.2. Análisis descriptivo de N-gramas

Teniendo como base el conjunto de datos pre-procesado, se realiza posteriormente un análisis descriptivo por N-gramas sobre el conjunto de datos. Para ello, y de forma exploratoria, en primer lugar se muestra en la Figura 4.3 una nube de palabras elaborada a partir de los datos procesados.

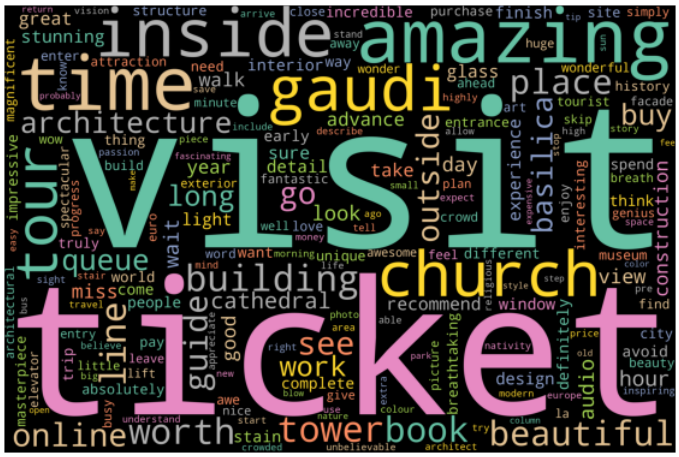


Figura 4.3: Nube de palabras - Sector turismo
Fuente: Elaboración propia

Esta figura consiste en un conjunto de palabras representadas en distintos tamaños que permiten visualizar la frecuencia de las palabras contenidas en el conjunto de datos. Un mayor tamaño de la palabra representa una mayor frecuencia de aparición de la misma en los datos analizados. En el caso del conjunto estudiado en esta sección se observa que las palabras más relevantes son: ‘visita’, ‘entrada’, ‘tiempo’, ‘interior’, ‘increíble’,‘tour’, ‘gaudi’ e ‘iglesia’. Otras palabras que también aparecían con mucha frecuencia eran: ‘Barcelona’, ‘Sagrada’ y ‘Familia’, pero se eliminaron ya que estaban presentes en casi todas las reseñas y no aportaban información relevante al análisis. Inicialmente, es posible observar una diversidad de conceptos en las opiniones de los usuarios. Por un lado, se identifican palabras como ‘entrada’, ‘tour’, ‘visita’, etc. En este caso, la Sagrada Familia al ser un monumento tan importante de la ciudad de Barcelona, es a su vez muy visitado, recibiendo constantemente una gran afluencia de turistas, lo que lleva a muchos a recomendar la compra de entradas por internet para evitar esperas. Por otro lado, los usuarios también hacen referencia al interior de la basílica, su arquitectura y Gaudí. No es extraño que los turistas hagan referencia a estos términos ya que el interior de la basílica, de arquitectura modernista y diseñado por Antoni

Gaudí, es uno de los signos de identidad de la ciudad de Barcelona y un atractivo turístico a nivel mundial.

Teniendo como base este descriptivo inicial, se ha realizado un análisis de los N-gramas más importantes del corpus mediante el cálculo del TF-IDF de cada palabra, técnica explicada en la sección 3.3. Para ello, inicialmente se obtiene la distribución de la métrica TF-IDF para cada uno de los unigramas, o lo que es lo mismo, para las palabras individuales que forman parte del conjunto de datos analizados. Los resultados se muestran en la Figura 4.4, evidenciando que las diez palabras con mayor relevancia en el corpus son: 'visita', 'increíble', 'entrada', 'interior', 'tiempo', 'iglesia', 'valor', 'bello', 'tour', y 'edificio'. Estos resultados son compatibles con la nube de palabras mostrada anteriormente, donde se observa que los turistas hacen referencia a la importancia de la compra de entradas así como a las impresiones sobre el interior de la iglesia, con términos como 'bello' o 'increíble'. Además se incluyen conceptos asociados a otros servicios que se ofrecen en la basílica como pueden ser los tours guiados.

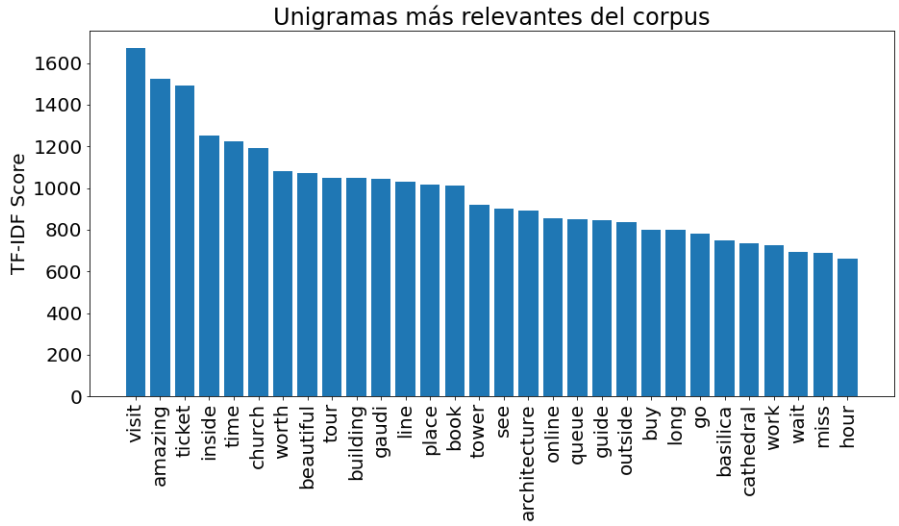


Figura 4.4: TF-IDF de los 30 unigramas más relevantes - Sector turismo
Fuente: Elaboración propia

Posteriormente, se ha calculado el TF-IDF de cada bigrama del corpus o, en otras palabras, de cada conjunto de dos palabras. Como se puede observar en la Figura 4.5, las diez parejas de palabras con mayor relevancia en el conjunto de datos son: 'comprar entrada', 'entrada *online*, 'reservar entrada', 'merece la pena visitar', ´vidriera', 'audio guía', 'entrada anticipada', 'tour guiado', 'cristalera' y 'reserva *online*'. Se observa un patrón de gran relevancia en la mayoría de estas parejas resaltando las largas filas para adquirir las entradas y la importancia de reservar las mismas con antelación (*online*). Por otro lado, también se hace referencia a otros servicios que se ofrecen como pueden ser las visitas guiadas o las audio-guías. Estos son servicios de gran interés para los turistas, ya que ayudan a conocer mejor las características de los monumentos visitados, así como su historia y algunos datos

de interés. Es por ello que es importante conocer que clase de opinión tienen los usuarios de estos servicios, con el fin de identificar cuales son los puntos fuertes del servicio y cuales son posibles puntos de mejora.

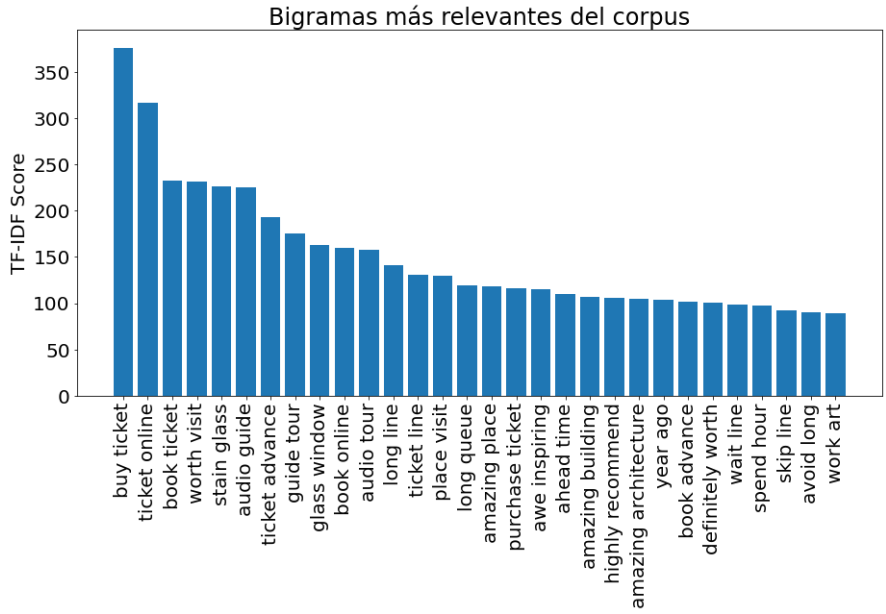


Figura 4.5: TF-IDF de los 30 bigramas más relevantes - Sector turismo
Fuente: Elaboración propia

Finalmente, este análisis también es realizado sobre los trigramas o grupos de tres palabras que tienen mayor relevancia. Los resultados se muestran en la Figura 4.6, siendo los trigramas más importantes: 'comprar entrada *online*', 'vidriera', 'reservar entrada *online*', 'comprar entrada por adelantado', 'reservar entrada por adelantado' y 'dejar sin aliento'. De nuevo, se observa que los visitantes consideran de gran importancia comprar las entradas con antelación. Se comprueba que, de los cincos primeros grupos de palabras más relevantes, tres de ellos tratan el tema de la compra de entradas anticipadamente, lo que indica que es un tema que afecta en gran medida a la opinión de los usuarios. Otra característica que es importante resaltar es la relevancia que dan las reseñas a las vidrieras de la basílica. Esto se debe probablemente a que estas cristaleras forman una parte clave del atractivo turístico de la basílica, creando un ambiente de misticismo y conteniendo iconografía y simbolismo cristiano, como nombres de santos y santuarios. Además, en función del momento del día y de la época del año, la luz que entra a través de ellas crea diferentes juegos de colores, lo que contribuye a que sea un tema relevante en las opiniones de los visitantes.

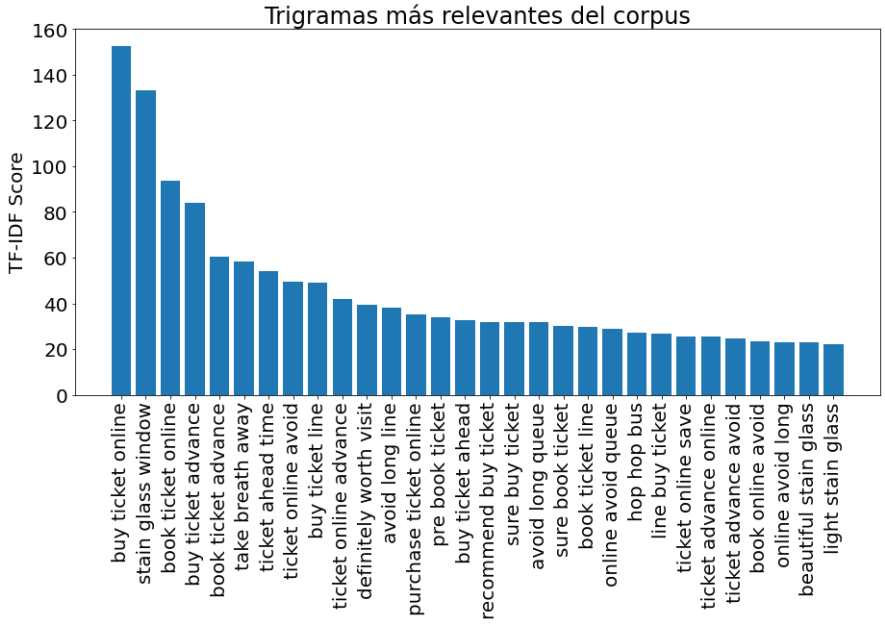


Figura 4.6: TF-IDF de los 30 trigramas más relevantes - Sector turismo
Fuente: Elaboración propia

4.1.3. Modelado de tópicos

Una vez realizado el análisis descriptivo preliminar de las opiniones de los usuarios, se realiza el proceso de modelado de tópicos explicado en la sección 3.4. Para ello, en primer lugar, es necesario ajustar el número de tópicos con el fin de obtener un modelo óptimo e interpretable del corpus. Para ello, se ha calculado el índice de coherencia para diferentes valores de número de tópicos variando de 2 hasta 15 tópicos (ver Figura 4.7). Estos resultados permiten evidenciar el número de tópicos que proporciona un mayor índice de coherencia.

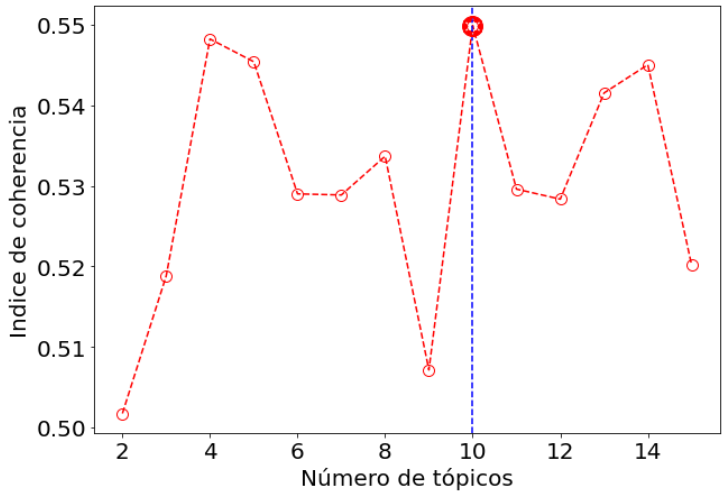


Figura 4.7: Índice de coherencia para cada número de tópicos - Sector turismo
Fuente: Elaboración propia

Así, en la Figura 4.7 se observa que el valor de *k* para el que se obtiene un mayor índice de coherencia es 10. Teniendo en cuenta este valor, se analiza entonces la distancia inter-tópica con el fin de evaluar como de interpretables son los tópicos modelados. Como se ha explicado en la sección 3.4, se busca un modelo donde los círculos no estén muy superpuestos, si no que se distribuyan a lo largo del gráfico, permitiendo obtener un conjunto de tópicos altamente interpretables. De esta manera, en la Figura 4.8 se muestra el mapa de distancia entre tópicos para un valor de *k* igual a 10. Asimismo, es posible observar que pese a que los círculos presentan cierta superposición, ninguno de ellos está contenido, de forma completa o casi completa, dentro de otro, lo cual indica que hay elementos diferenciadores entre cada uno de los tópicos modelados.

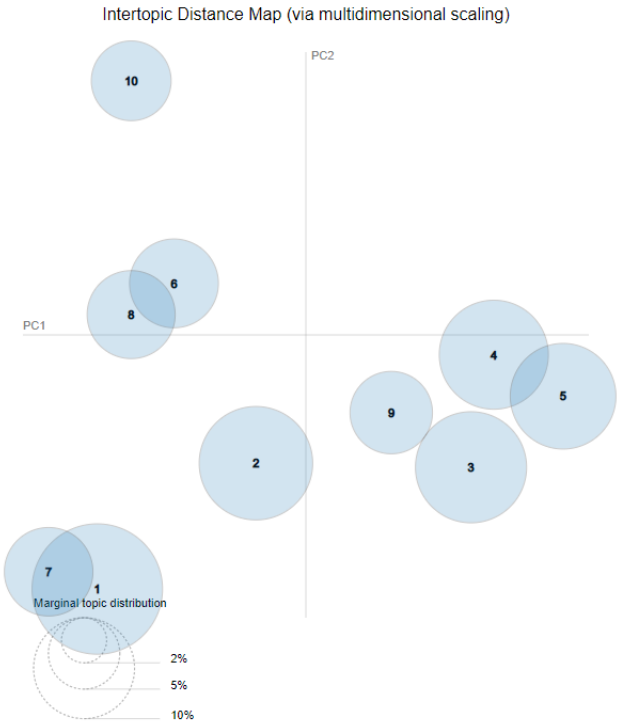


Figura 4.8: Distancia entre tópicos para *k*=10 - Sector turismo
Fuente: Elaboración propia

Como resultado, en la Tabla 4.1 se muestran los tópicos con las diez palabras clave asociadas a cada uno de ellos. Se puede observar que ciertos tópicos tienen palabras en común, como 'visita' o 'increíble' ya que son términos que los usuarios pueden utilizar para hablar de distintos temas, como puede ser la 'Experiencia', en general, o particularidades como las 'Vidrieras'. Además, con el fin de mejorar el carácter interpretable de los tópicos, se han incluido los cinco bigramas y trigramas más relevantes de cada una de las categorías. En el Anexo A se ha incluido una tabla con todos los bigramas y trigramas obtenidos en el análisis, en el idioma original (inglés).

Tópico	Categoría	Palabras clave	Bigramas más relevantes	Trigramas más relevantes
1	Compra de entradas	entrada, fila, *online*, comprar, reservar, tiempo, cola, visitar, largo, por adelantado	comprar entrada, entrada *online*, reservar entrada, entrada por adelantado, reserva *online*	comprar entrada *online*, reservar entrada *online*, comprar entrada por adelantado, reservar entrada por adelantado, entrada antes de tiempo
2	Esperas para entrar	visita, linea, temprano, espera, largo, cola, entrada, día, tiempo, ir, mañana	pronto por la mañana, larga cola, larga fila, esperar fila, esperar una hora	pronto evitar multitudes, esperar larga fila, evitar larga fila, esperar hora fila, llegar pronto mañana
3	Experiencia	dentro, valor, fuera, increíble, visita, edificio, mirar, ir, bello, cola	merece la pena visitar, sin aliento, definitivamente vale la pena, edificio increíble, quitar el aliento	dejar sin aliento, definitivamente vale la pena visitar, merece la pena pagar la entrada, merece la pena el interior, increíble lugar visita
4	Museo	edificio, gaudi, detalle, museo, gastar, hora, arquitectura, interesante, increíble, visita	estar hora, merece la pena visita, visitar museo, museo sótano, atención al detalle	estar hora mirando, perder museo sótano, museo nivel bajo, visitar museo sótano, museo sótano interesante
5	Medios de accesibilidad al monumento	iglesia, tiempo, aparcar, metro, área, tienda, andar, turista, entrada, personas	estación de metro, metro fácil, fácil acceso, bus turístico, fácil encontrar	estación metro derecha, estación de metro cerca, *hop hop bus*, fácil acceder metro, fácil encontrar metro
6	Experiencia en la visita a la torre	torre, vista, ascensor, elevador, andar, escaleras, natividad, pasión, ciudad, altura	torre del nacimiento, torre de la pasión, escalera de caracol, grandes vistas, ascensor torre	escalera de caracol estrecha, ir torre nacimiento, ir torre pasión, pagar extra torre, recomendar visitar torre
7	Servicios de la basílica	tour, guía, audio, recomendar, visita, altamente, bueno, coger, autobús, basílica	audio tour, reservar tour, tour guiado, audio tour, altamente recomendado	reservar tour guiado, recomendar tour guiado, julia travel tour, tour guiado merecer la pena, recomendar audio guía
8	Impresiones del monumento y su arquitecto	gaudi, trabajo, obra maestra, genio, admiración, completo, palabra, construcción, diseño, mundo	imponente, antoni gaudi, obre maestra gaudi, gaudi genio, maravilla mundial	arquitecto antoni gaudi, aniversario muerte gaudi, maravilla arquitectónica mundial, patrimonio mundial unesco, realmente impresionante
9	Vidrieras	luz, cristal, vidriera, dentro, iglesia, ventana, catedral, interior, bello, gaudi	vidriera, día soleado, luz natural, final de la tarde, luz interior	vidriera, cristalera bella, vidrieras increíbles, luz fluye interior, vidriera interior
10	Construcción	visita, iglesia, alucinante, año, lugar, arquitectura, ver, bello, trabajo, terminado	trabajo en progreso, 100 años, obra de arte, increíble lugar, pieza arquitectura	increíble pieza de arquitectura, hace 100 años, construcción 100 años, terminar 10 años, visitar 10 años

Tabla 4.1: Tópicos, categoría, palabras clave, bigramas y trigramas más relevantes.

Por otro lado, se ha analizado la distribución de los tópicos en el corpus como se muestra en la Figura 4.9. De esta manera, se comprueba que los tópicos 1 (‘Compra de Entradas’), 3 (‘Experiencia’) y 10 (‘Construcción’) son los que tienen más revisiones asociadas. Estos resultados son consistentes con el descriptivo inicial de los datos, mostrando una mayor relevancia a la compra anticipada de entradas y al valor de la experiencia. Por otro lado, se observa también la presencia de otros tópicos que no se encontraban en el análisis preliminar de los datos, como pueden ser el tópico 5 (‘Medios de accesibilidad al monumento’) o el 4 (‘Museo’). El tópico 6 (‘Experiencia en la visita a la torre’) es el que cuenta con menor número de opiniones dentro del corpus, esto se debe probablemente a que es un servicio adicional, que no se incluye con la compra de una entrada estándar a la basílica, por ello muchos turistas deciden no pagar un extra para acceder a ellas.

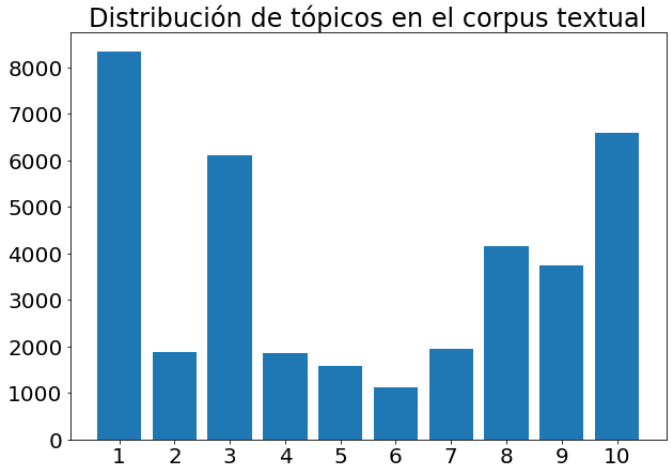


Figura 4.9: Distribución de tópicos en el corpus textual - Sector turismo
Fuente: Elaboración propia

Por último, como se ha mencionado anteriormente, se han construido los bigramas y trigramas por tópico para profundizar en el análisis de los mismos. Respecto al tópico 1, ‘Compra de entradas’, se observa la importancia que tiene para los usuarios la compra de las mismas *online*. Los visitantes utilizan distintas expresiones pero todas llevan a recomendaciones asociadas a la reserva de las entradas de forma anticipada, con el fin de evitar las colas. En la misma línea, el tópico 2 ‘Esperas para entrar’, está relacionado con las entradas, sin embargo, en este caso, los usuarios hacen especial énfasis en las esperas asociadas a la compra presencial de las entradas. Del análisis de los trigramas, es posible extraer recomendaciones asociadas como por ejemplo el ir pronto por la mañana para evitar las multitudes (*‘early avoid crowds’* o *‘visit early morning’*).

Por su parte, respecto al tópico 3: ‘Experiencia’, los visitantes expresan su opinión, en general, sobre la visita, mencionando su gran valor (*‘definitely worth visit’*) y como el edificio, la arquitectura y las vidrieras dejan sin aliento al visitante (*‘take breath away’*). Esto, de forma preliminar, evidencia una opinión positiva de los turistas respecto a la experiencia

de visitar la basílica. En la misma línea, el tópico 4, 'Museo', asocia principalmente opiniones dirigidas a la visita del museo. En ellas, los usuarios recomiendan su visita, comentando que les ha parecido interesante. También recalcan que se encuentra ubicado en el sótano del edificio, por lo que es fácil perderse la oportunidad de visitarlo.

En referencia al tópico 5 'Medios de accesibilidad al monumento', en él se indican, en términos generales, los medios de transporte que los usuarios recomiendan para llegar hasta la basílica. De las opiniones, en general, se prefiere el metro, por la cercanía y facilidad de acceso a este medio particular de transporte. También se mencionan otros transportes como los autobuses turísticos ('*hop hop bus*'). El tópico 6 'Experiencia de visita a la torre', por su parte, agrupa las opiniones de los turistas relacionadas con la visita a las distintas torres de la basílica, siendo las más recurrentes la torre de la fachada de la Pasión y la de la Natividad o Nacimiento. Los comentarios incluyen también opiniones sobre las vistas que se pueden disfrutar desde lo alto, así como de los accesos a dichas torres (escaleras de caracol y ascensores). Además, se recalca que merece la pena pagar extra por acceder a ellas, recomendando fuertemente su visita.

En relación al tópico 7 'Servicios de la basílica', se agrupan las opiniones de los visitantes sobre los distintos servicios ofrecidos en la basílica, incluyendo tours guiados y audio guías. Se observa sobretodo en los trigramas que son servicios recomendados por los usuarios ('*recommend guide tour o guide tour worth*'). De nuevo, al igual que con las entradas, en este servicio también se recomienda que se realice la reserva por adelantado. El tópico 8 'Impresiones del monumento y su arquitecto' está principalmente asociado a sentimientos y connotaciones positivas sobre la basílica, se incluyen principalmente adjetivos asociados al diseño y al arquitecto como 'obra maestra', 'genio', 'obra de arte' , 'maravilla arquitectónica mundial' o 'imponente', etc. Además, también hablan de su pertenencia al Patrimonio de la Humanidad declarado por la UNESCO.

En la misma línea, el tópico 9 'Vidrieras' habla específicamente de la característica de la basílica que más llama la atención de los turistas, siendo esta la presencia de múltiples vidrieras y cristaleras. Anteriormente ya se ha resaltado su importancia y su simbología, por lo que no es de extrañar que estén presente en las reseñas de los usuarios. Sin embargo, en este tópico se da especial importancia a conceptos asociados al cristal, la luz, o las ventanas, utilizando adjetivos para describirlos como 'bellas'. Además también se recomienda ir al final de la tarde, momento en el que los juegos de luces y colores son más impresionantes. Finalmente, en el tópico 10 'Construcción', los visitantes hablan principalmente de la arquitectura y las características propias de la construcción. Una gran cantidad de opiniones están relacionadas al trabajo en proceso ('*work in progress*') y a la duración de más de 100 años desde el inicio de su construcción.

4.1.4. Análisis de sentimiento

Finalmente, con el objetivo de analizar las emociones subyacentes a las revisiones online de los turistas, se ha seguido la metodología explicada en la sección 3.5. De esta manera, al utilizar la estrategia VADER, se ha calculado la puntuación asociado al sentimiento de cada una de las opiniones. Como resultado, en la Figura 4.10 se muestra un gráfico de cajas con la distribución de la puntuación compuesta para todas las reseñas. Así, es posible observar que las opiniones de los usuarios son muy positivas, presentando un valor de mediana de 0.8747 (muy cercano a 1, valor asociado a un sentimiento positivo). Además, el primer cuartil se sitúa casi en 0.70, lo que quiere decir que solo el 25 % de los datos tienen una puntuación menor que este valor. Finalmente, podemos ver en la figura que el bigote inferior tiene un valor menor a 0.35 (previa eliminación de *outliers*), lo que indica que a excepción de algunos valores atípicos la menor puntuación obtenida se sitúa cerca de dicho valor.

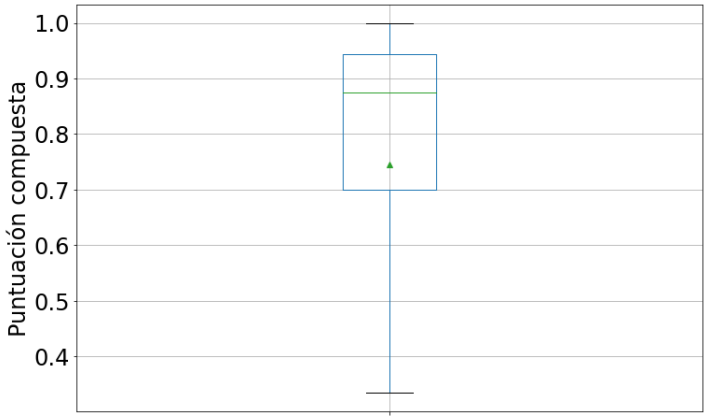


Figura 4.10: Gráfico de cajas de la puntuación compuesta - Sector Turismo
Fuente: Elaboración propia

Por otro lado, en la Figura 4.11 se muestra la evolución temporal del sentimiento acumulado diariamente en las reseñas. Se observa una tendencia creciente hacia sentimientos positivos, lo que concuerda con el sentimiento general de satisfacción que los usuarios parecen tener en relación a esta experiencia. Se evidencia también una componente estacional, más pronunciada en los años 2015 y 2016, en la que el sentimiento aumenta en los meses de verano y disminuye en la temporada invernal. Esto, como se ha comentado anteriormente, coincide con los meses en los que crece el número de turistas en la ciudad de Barcelona (Observatori del Turisme a Barcelona: ciutat i regió, 2021).

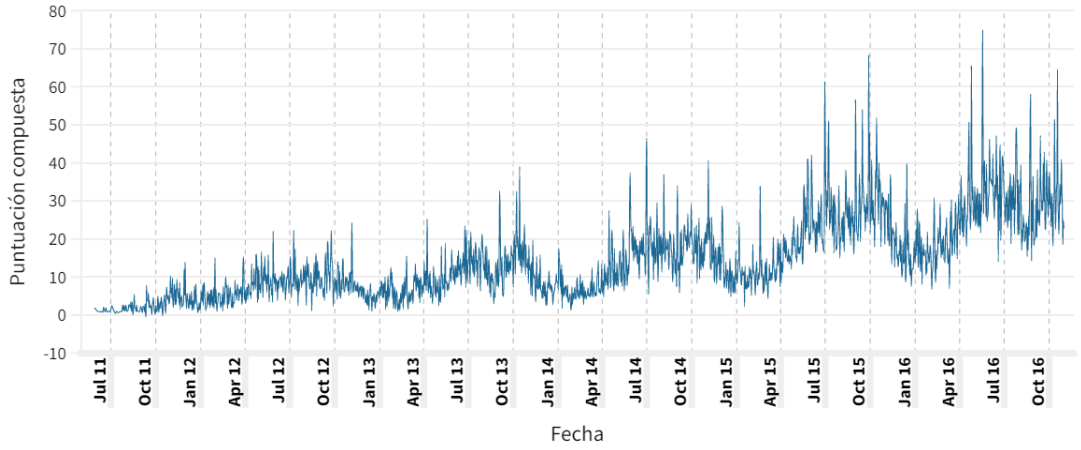


Figura 4.11: Evolución del sentimiento a lo largo de los años - Sector turismo
Fuente: Elaboración propia

A continuación, se va a analizar la evolución temporal del sentimiento para cada uno de los tópicos obtenidos en el modelado. Comenzando por el tópico 1 'Compra de entradas' (ver Figura 4.12a), se observa que en su mayor parte los sentimientos obtenidos para este tópico son positivos, pese a que podría pensarse que hablar de comprar entradas por adelantado para evitar esperas tendría una connotación negativa. En este sentido, es importante considerar que el cálculo de la puntuación se basa en la polaridad expresada. Con ello, si el usuario únicamente recomienda comprar las entradas por adelantado para evitar esperas esa reseña no tendrá una puntuación negativa, aunque trate un tema que podría considerarse molesto para el visitante. Además en el gráfico también se puede destacar cierta estacionalidad, aumentando su sentimiento durante los meses de verano. A diferencia del caso anterior, en el tópico 2 'Esperas para entrar' se observa que los sentimientos expresados tienen un componente negativo muy relevante (ver Figura 4.12b). En términos generales, la puntuación compuesta de este tópico presenta valores más bajos que en el caso anterior, incluso llegando a tener una gran cantidad de puntuaciones negativas. Estos resultados son coherentes ya que tener que esperar largas filas para poder entrar se trata de un tema que, en general, ocasiona molestias a los turistas.

Con respecto al tópico 3 'Experiencia' (ver Figura 4.12c) se ve con menos claridad la presencia de una componente estacional, pese a que los mayores picos se siguen produciendo a finales de verano, no se ve una diferencia tan clara con los meses de invierno como en el tópico 1. Por otro lado, se observa que existe una mayor dispersión del sentimiento diario, sobretodo en los meses de verano del 2015 y 2016, pero el sentimiento es en su mayor parte positivo. El tópico 4 'Museo' (ver Figura 4.12d), por su parte, se comprueba que es un tema que genera una diversidad de sentimientos, en su mayor parte positivos pero con un componente negativo relevante. Esto puede ser debido, como se ha visto en los trigramas de la Tabla 4.1, a que el museo se encuentra localizado en el sótano y por ello muchos turistas se lo perdían y no lo visitaban, lo cual podría generar cierto sentimiento negativo a la hora de

comentar la experiencia.

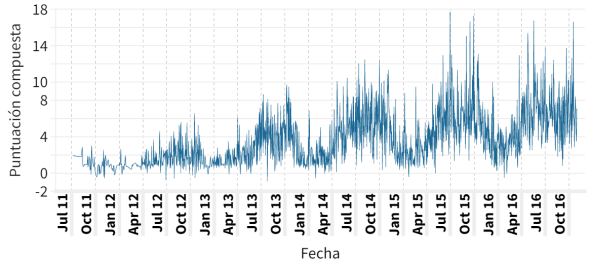
El tópico 5 ‘Medios de accesibilidad al monumento’ (ver Figura 4.12e) y el tópico 6 ‘Experiencia en la visita a la torre’ (ver Figura 4.12f) son dos tópicos que presentan mayor desviación en la puntuación compuesta en cuanto a la evolución del sentimiento de valores positivos a negativos, sin apreciarse un componente claro de tendencia o estacionalidad. Cabría destacar con respecto al tópico 5, el hecho de que las componentes negativas del sentimiento podrían estar relacionadas con algunas inconformidades con los medios de acceso al monumento. Pese a que en los trigramas de la Tabla 4.1 hacen referencia a la facilidad de acceso en metro, también destacan el tiempo dedicado en el acceso si no se cuenta con una entrada de forma anticipada, lo que podría contribuir al sentimiento negativo. Con respecto al tópico 7 ‘Servicios de la basílica’ se puede apreciar un sentimiento mayoritariamente positivo, resultado que es coherente con los análisis de la Tabla 4.1 donde los usuarios recomendaban los servicios de tours guiados ofrecidos en la basílica (ver Figura 4.12g).

En el tópico 8 ‘Impresiones del monumento y su arquitecto’, se observa de nuevo cierta estacionalidad, siendo los meses de invierno los que tienen menos variación en cuanto al sentimiento (ver Figura 4.12h). En este gráfico sí que se observan ciertas fechas con sentimientos más negativos, siendo el pico más bajo en agosto de 2014. Sin embargo, el sentimiento es en su mayor parte positivo, pese a existir ciertos picos puntuales donde la puntuación compuesta es negativa. De nuevo son resultados coherentes con los mencionados anteriormente, ya que los usuarios utilizaban términos como ‘imponente’, ‘obra maestra’, ‘maravilla’, etc.

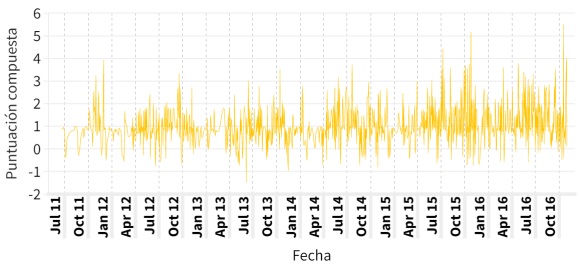
En relación al tópico 9 ‘Vidrieras’ mostrado en la Figura 4.12i, el sentimiento expresado es muy positivo donde la puntuación compuesta alcanza valores considerablemente altos, en comparación a la mayoría de los demás tópicos. De nuevo, estos resultados son razonables ya que las vidrieras son uno de los atractivos de la basílica que más impresión crean en los visitantes. Por último, se analiza la variación del sentimiento para el tópico 10 ‘Construcción’ (ver Figura 4.12j). En este tópico, el sentimiento de los usuarios es en su mayor parte positivo, existiendo pocos picos negativos con valores muy cercanos a 0 (neutro). Estos resultados no son de extrañar pues los visitantes utilizan términos como ‘obra de arte’ o ‘increíble pieza de arquitectura’ para referirse a la basílica en este tópico. De nuevo se aprecia una estacionalidad en el sentimiento, siendo los picos más altos alcanzados en abril y mayo de 2016.

Tras analizar la evolución temporal de los tópicos, así como la evolución general se puede concluir que el sentimiento que predomina en las reseñas es positivo, con algunos picos puntuales que aparecen especialmente en los meses de verano. Esto coincide con la época del año donde Barcelona recibe más turistas, siendo un periodo de vacaciones y donde el clima es un factor que potencia la actividad turística. Por lo tanto, se puede concluir que los visitantes de la Sagrada Familia tienen un sentimiento mayoritariamente positivo, siendo los temas predominantes de discusión, la compra de entradas por adelantado, las vidrieras y la arquitectura de la basílica. Por otro lado, para mejorar la experiencia en la visita a los turista se podrían tener en cuenta mejoras relacionadas con los temas que generan un sentimiento

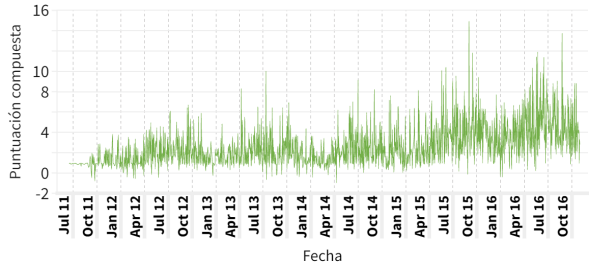
más negativo en los usuarios. Entre dichos temas se encuentra señalizar o indicar mejor la localización del museo, además de mejorar la experiencia durante la compra de las entradas.



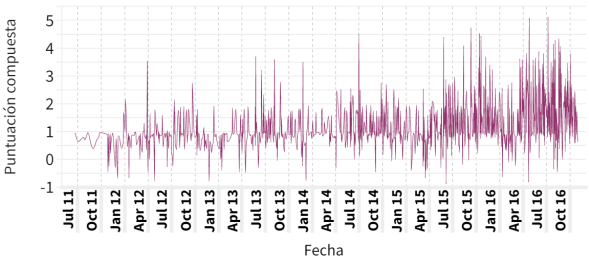
(a) Tópico 1: Compra de entradas



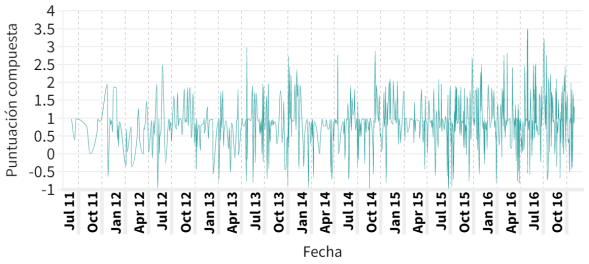
(b) Tópico 2: Esperas para entrar



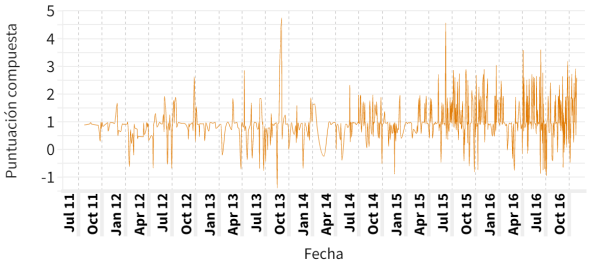
(c) Tópico 3: Experiencia



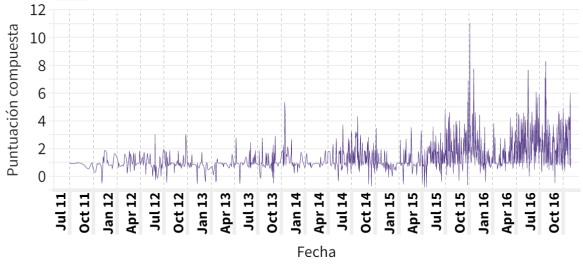
(d) Tópico 4: Museo



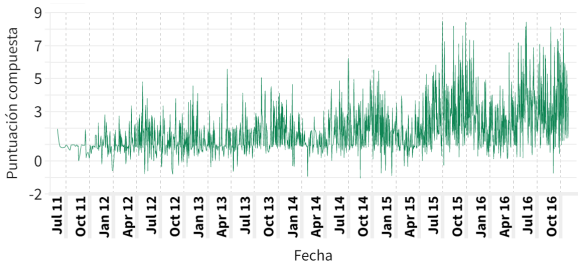
(e) Tópico 5: Medios de accesibilidad



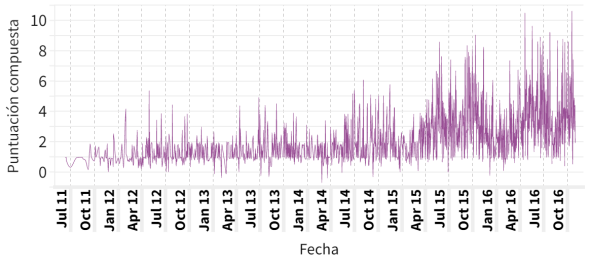
(f) Tópico 6: Experiencia en la visita a la torre



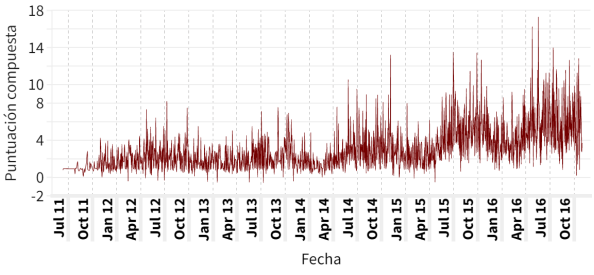
(g) Tópico 7: Servicios de la basílica



(h) Tópico 8: Impresiones de monumento y arquitecto



(i) Tópico 9: Vidrieras



(j) Tópico 10: Construcción

Figura 4.12: Evolución del sentimiento a lo largo del tiempo para cada tópico
Fuente: Elaboración propia

4.2. Comercio electrónico

4.2.1. Selección y pre-procesamiento de los datos

El conjunto de datos seleccionado en esta área de aplicación corresponde a reseñas de teléfonos móviles provenientes de Amazon, desde 2003 hasta 2019, y cuenta con un total de 67.965 reseñas (Nibras, 2019). Estos datos cuentan con información de teléfonos de las siguientes marcas: ASUS, Apple, Google, HUAWEI, Motorola, Nokia, OnePlus, Samsung, Sony y Xiaomi. De nuevo se ha elegido este conjunto de datos ya que presentan una gran cantidad de reseñas, lo que como se ha comentado con anterioridad, contribuye a la obtención de conclusiones más fiables.

Con el fin de proporcionar un descriptivo inicial de los datos, en la Figura 4.13 se muestra un diagrama de cajas asociado a la distribución del número de palabras por reseña. Eliminando datos atípicos, el diagrama muestra que el 75 % de los datos contienen menos de 63 palabras por reseña, siendo el máximo 140 palabras aproximadamente. Además, se observa que la parte superior del diagrama, por encima de la mediana, es más amplia que la inferior, indicando una concentración mayoritaria en valores bajos de la distribución (asimetría positiva).

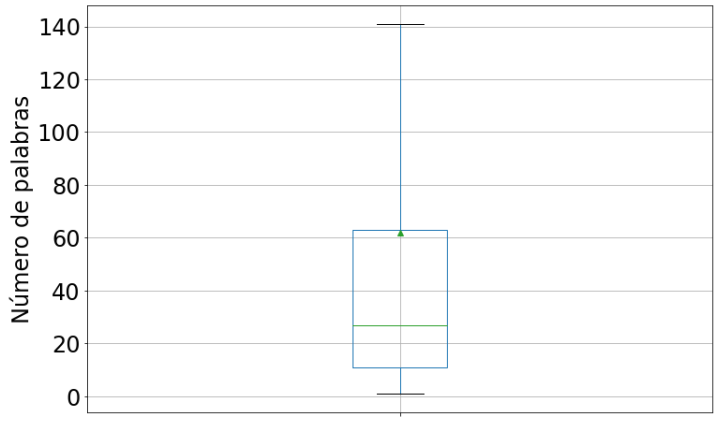


Figura 4.13: Diagrama de cajas del número de palabras por reseña - Sector comercio electrónico

Fuente: Elaboración propia

Por otro lado, con el fin de analizar la evolución del número de reseñas a lo largo de los años, la Figura 4.14 muestra la dinámica de la distribución temporal de las revisiones. Pese a que existen datos desde el 2003 no es hasta el 2012 cuando empieza a haber un número elevado de reseñas. Los resultados muestran una tendencia creciente, alcanzando un valor máximo en diciembre de 2019. Una posible causa de este aumento en los datos temporales, como se ha explicado con anterioridad, puede deberse al incremento en el uso de la tecnología para compartir opiniones sobre diversos productos o servicios.

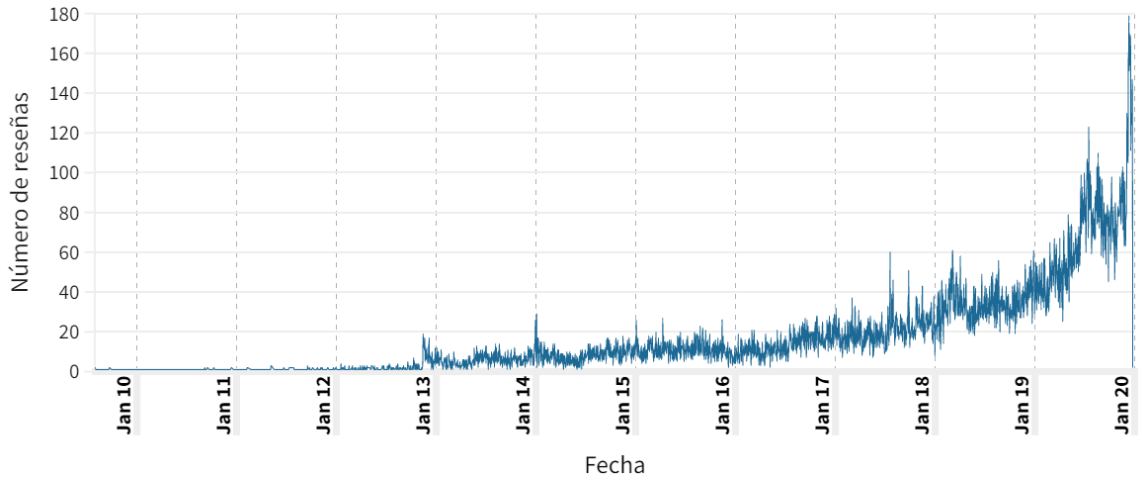


Figura 4.14: Serie temporal del número de reseñas - Sector comercio electrónico
Fuente: Elaboración propia

Tras este análisis inicial, se lleva a cabo la limpieza y pre-procesado de las reseñas, de acuerdo a lo expuesto en el capítulo 3. Además, en este caso ha sido necesario eliminar algunas reseñas que no estaban en el lenguaje de análisis (inglés). Inicialmente el corpus contaba con un total de 67.985 entradas y, al culminar las tareas comentadas, las reseñas se redujeron en un 12 % llegando a un valor de 59.815 reseñas.

4.2.2. Análisis descriptivo de N-gramas

Una vez obtenido el conjunto de datos pre-procesado, se ha desarrollado el análisis descriptivo de los N-gramas. En primer lugar se muestra en la Figura 4.15 una nube de palabras. Como ya se ha explicado en anteriores apartados, esta herramienta permite analizar de forma más visual cuales son los términos que aparecen más en el corpus. Particularmente en este campo, los términos más usados en las revisiones son: ‘funcionar’, ‘bueno’, ‘estupendo’, ‘batería’, ‘pantalla’, ‘comprar’ y ‘cámara’. Es importante destacar las palabras referidas a elementos de los teléfonos como la batería o la pantalla, ya que son características esenciales que permiten a los clientes tomar la decisión al elegir una marca determinada. Por otro lado, se encuentran adjetivos positivos como ‘bueno’ o ‘estupendo’ que son usados para describir algunas características del producto.

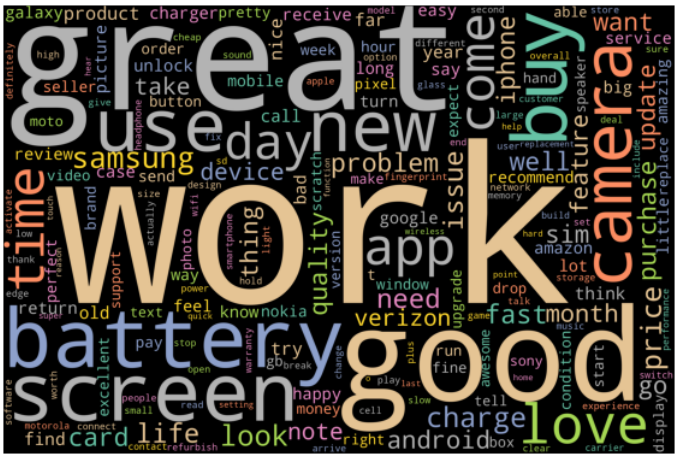


Figura 4.15: Nube de palabras - Sector comercio electrónico
Fuente: Elaboración propia

Partiendo de estos resultados, se ha llevado a cabo un análisis descriptivo de los N-gramas más relevantes del corpus a través del cálculo del TF-IDF de cada término, como se ha detallado en la sección 3.3. Con esta técnica se obtiene, en primer lugar, el gráfico de los unigramas (ver Figura 4.16), siendo los diez más relevantes ‘genial’, ‘bueno’, ‘funcionar’, ‘encantar’, ‘batería’, ‘comprar’, ‘nuevo’, ‘precio’, ‘pantalla’, ‘usar’. Se puede comprobar que son resultados acorde con lo obtenido en la nube de palabras. Los usuarios mencionan sobretodo las características relacionadas con los teléfonos, así como adjetivos positivos sobre sus características. Además, es importante tener en cuenta que aparecen otras propiedades que pueden influir en la opinión del usuario en cuanto a la compra de un dispositivo, tales como el precio del mismo o el hecho de que se encuentre en un estado casi nuevo al momento de la venta.

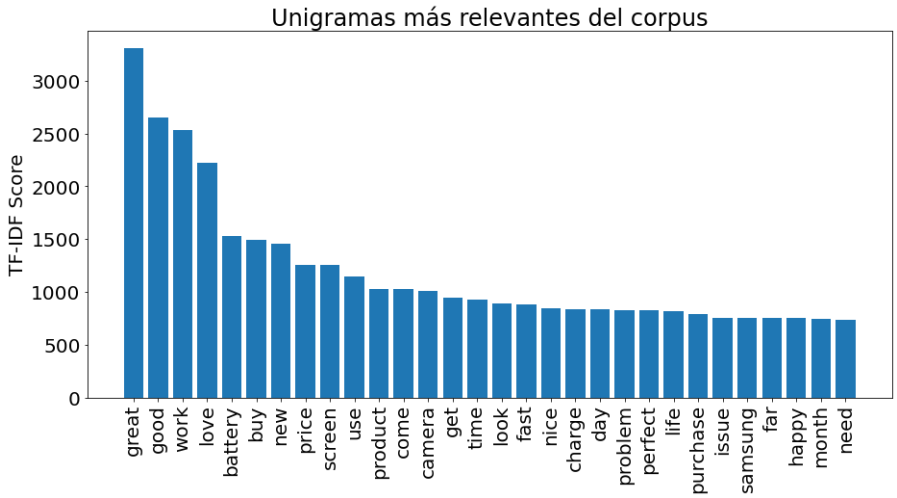


Figura 4.16: Unigrama con las 30 palabras más relevantes - Sector comercio electrónico
Fuente: Elaboración propia

A continuación, se ha calculado el TF-IDF de cada conjunto de dos palabras o bigrama, mostrado en la Figura 4.17. Los resultados muestran que las diez parejas de palabras más

relevantes son: ‘funciona estupendamente’, ‘vida de la batería’, ‘tarjeta sim’, ‘como nuevo’, ‘precio estupendo’, ‘precio bueno’, ‘funciona perfectamente’, ‘buen producto’, ‘fácil de usar’ y ‘producto estupendo’. En este caso se comprueba la relevancia que tiene el precio en las opiniones de los usuarios, que viene acompañado de adjetivos positivos para referirse a dicha característica del producto. Además, se pueden destacar otras características como puede ser la vida de la batería o la tarjeta SIM. Esta última se debe a que las reseñas pertenecen tanto a teléfonos de venta libre como a aquellos proporcionados por compañías telefónicas. Finalmente, otra característica que los usuarios resaltan es la facilidad de uso del dispositivo.

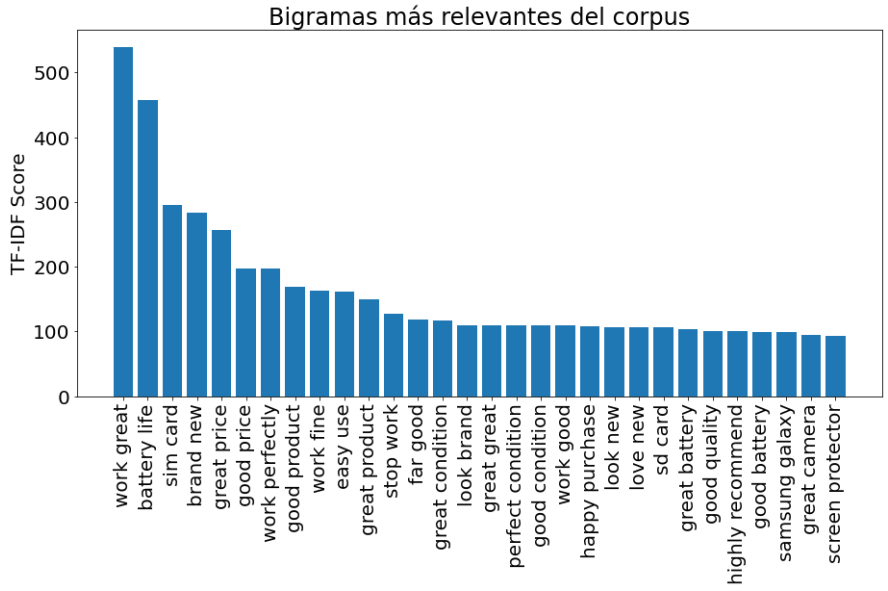


Figura 4.17: Bigrama con las 30 parejas de palabras más relevantes - Sector comercio electrónico

Fuente: Elaboración propia

Por último, se ha calculado el TF-IDF de cada conjunto de tres palabras, o trigrama (ver Figura 4.18). Los diez más relevantes son: ‘se ve como nuevo’, ‘estupenda vida de batería’, ‘gran gran precio’, ‘buena vida de batería’, ’vida de batería buena’, ‘vida de batería estupenda’, ‘vida de batería larga’, ‘tarjeta sim funciona’, ‘batería dura un día’ y ‘encantar funciona estupendamente’. Los resultados del análisis indican que la batería es una característica altamente satisfactoria para los usuarios de los teléfonos, quienes la describen con adjetivos muy positivos. Otras características que aparecen de nuevo en el análisis son las relacionadas a la tarjeta SIM y al estado del teléfono.

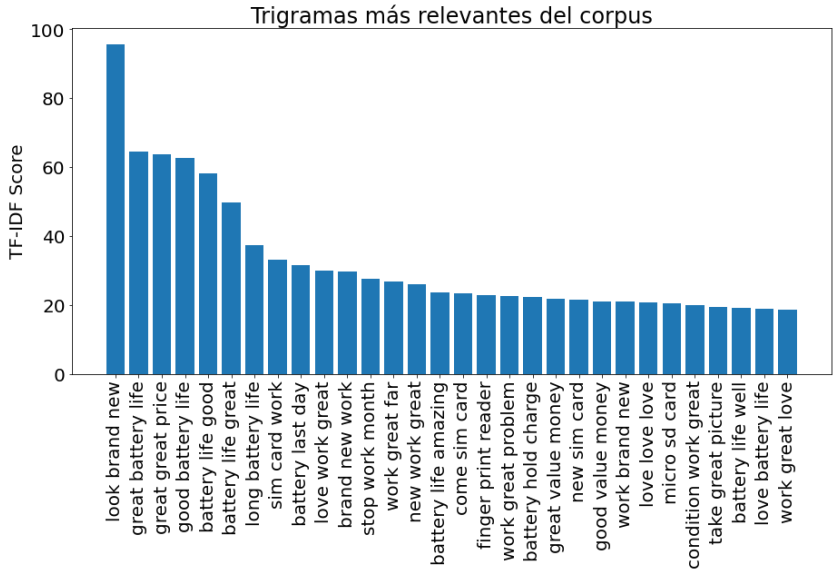


Figura 4.18: Trigrama con las 30 palabras más relevantes - Sector comercio electrónico
Fuente: Elaboración propia

4.2.3. Modelado de tópicos

Tras realizar el análisis descriptivo preliminar, en esta sección se presentan los resultados obtenidos en el modelado de tópicos, siguiendo el proceso descrito en la sección 3.4. En primer lugar, se requiere identificar el número óptimo de tópicos necesario para ajustar el modelo al conjunto de datos. Con este fin, se ha realizado el cálculo del índice de coherencia en función del número de tópicos. Estos resultados se muestran para un rango de 2 a 15 tópicos, en la Figura 4.19. En esta gráfica puede observarse que los valores más altos de índice coherencia están entre 0.47 y 0.48, con un máximo en 12 tópicos. En este caso, se ha decidido explorar los valores de número de tópicos iguales a k=5, y a k=10, los cuales son los primeros valores para los que se evidencia un valor alto del índice de coherencia. En particular, k=10 es el que registra el mayor valor de los dos.

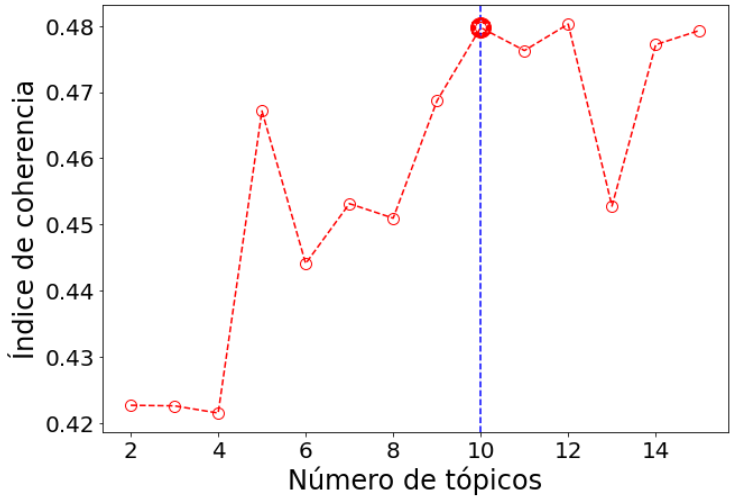


Figura 4.19: Índice de coherencia en función del número de tópicos seleccionado - Sector comercio electrónico

Fuente: Elaboración propia

Con dichos valores de *k* se procede al cálculo del mapa de distancia intertópica, para analizar la distancia entre los tópicos. Como se explicó en la sección 3.4, se busca aquel modelo donde los círculos no se encuentren demasiado superpuestos para obtener tópicos interpretables y que aporten valor al análisis con elementos diferenciadores. En la Figura 4.20, y en la Figura 4.21 se muestran dichos gráficos para k=5, y k=10, respectivamente. Se puede observar que el modelo de diez tópicos queda descartado ya que hay dos tópicos que prácticamente están contenidos al completo dentro del otro. Por lo que aunque tenga un índice de coherencia elevado no contribuye a obtener resultados relevantes. Por estas razón, se ha decidido seleccionar el modelo de cinco tópicos, cuyo mapa de distancia intertópica no muestra ninguna intersección significativa que pueda dificultar el proceso de interpretación de los tópicos.

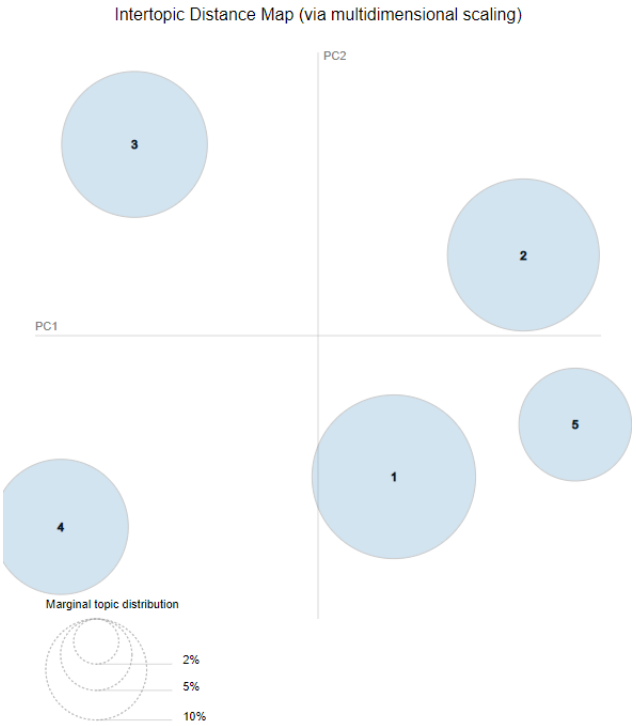


Figura 4.20: Distancia entre tópicos para k=5 - Sector comercio electrónico
Fuente: Elaboración propia

Figura 4.21: Distancia entre tópicos para k=10 - Sector comercio electrónico
Fuente: Elaboración propia

Al ajustar el modelo LDA para un valor de $k = 5$, se han obtenido los resultados que se describen en la Tabla 4.2. En dicha tabla se destacan las diez palabras clave más relevantes asociadas a cada uno de los cinco tópicos identificados por el modelo ajustado. Además se han incluido los cinco bigramas y trigramas más relevantes para cada uno de los tópico. En el Anexo B se muestra una tabla con el resto de bigramas y trigramas del análisis, en su idioma original (inglés).

Teniendo en cuenta estos resultados, se ha obtenido la distribución de opiniones en cada uno de los tópicos como se observa en la Figura 4.22. Esta gráfica evidencia que los tópicos más relevantes son el 2 (‘Cámara y batería’), 4 (‘Tarjeta SIM’) y 5 (‘Opinión general’). Estos resultados son coherentes, ya que los temas identificados se corresponden con aquellos que los usuarios abordan con frecuencia al realizar una reseña de un dispositivo móvil, y están estrechamente ligados a sus características físicas más relevantes.

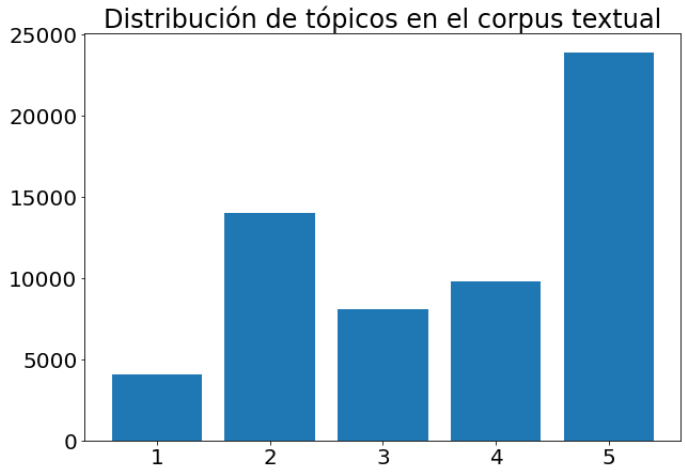


Figura 4.22: Distribución de tópicos en el corpus textual - Sector comercio electrónico
Fuente: Elaboración propia

Para completar el análisis, respecto al tópico 1 ‘Desbloqueo y almacenamiento’, es importante destacar que los usuarios expresan su opinión sobre los distintos métodos de desbloqueo del teléfono, siendo los más destacados el reconocimiento facial y la huella dactilar. Es razonable que estos temas sean tratados por los usuarios en sus reseñas ya que añaden una comodidad extra como es el desbloqueo rápido, sin necesidad de códigos, y añade mayor seguridad al dispositivo. Así mismo, otro tema de gran interés abordado en las reseñas es el almacenamiento de los teléfonos, el cual resulta un factor determinante en la elección del modelo por parte de los consumidores.

En relación al tópico 2, denominado ‘Cámara y batería’, cabe destacar que se trata de aspectos que influyen significativamente en la satisfacción del usuario. De hecho, al analizar los bigramas y trigramas presentes en las reseñas, se puede observar que los usuarios parecen estar bastante satisfechos con ambas características, utilizando términos como ‘bueno’ o ‘estupendo’ para referirse a ellas. En relación al tópico 3, ‘Usabilidad’, los clientes parecen

satisfechos de nuevo con esta característica de los productos ya que utilizan términos como ‘estupendo’ o ‘fácil’ para referirse al uso del dispositivo. El tópico 4, ‘Tarjeta SIM’, se encuentra presente en las reseñas debido a que se trata de una preocupación recurrente entre los usuarios de teléfonos móviles, tanto de aquellos que son de venta libre como de aquellos vinculados a compañías telefónicas. En este sentido, los consumidores expresan su inquietud por asegurarse de que la tarjeta SIM funcione correctamente en el dispositivo adquirido, lo cual es un aspecto crucial para poder utilizar el teléfono de manera efectiva. Así, se puede apreciar la importancia que tiene la compatibilidad del teléfono con la tarjeta SIM en la satisfacción del cliente y en la experiencia de uso del dispositivo.

Por último, en el tópico 5 ‘Opinión general’, los usuarios expresan su opinión global sobre el producto, lo que da lugar a una gran variedad de casos. Al analizar los trigramas presentes en las reseñas, se puede observar que algunos clientes se muestran satisfechos con el producto, destacando su buen funcionamiento o el hecho de que llegó en óptimas condiciones. Sin embargo, también existen trigramas que indican lo contrario, es decir, que el dispositivo presenta problemas o fallos técnicas, incluso poco después de la compra. En definitiva, este tópico refleja la percepción general que tienen los consumidores sobre el producto, incluyendo tanto aspectos positivos como negativos, lo cual resulta de gran utilidad para conocer la calidad y fiabilidad del producto analizado.

Tópico	Categoría	Palabras clave	Bigramas más relevantes	Trigramas más relevantes
1	Desbloqueo y almacenamiento	bueno, batería, gb, precio, huella , píxel, cámara, dispositivo, calidad, android	lector huella, huella dactilar, sensor de huellas, reconocimiento facial, almacenamiento interno	lector de huellas dactilares, sensor huellas dactilares, reconocimiento facial funciona, lector huellas funciona, gb almacenamiento interno
2	Cámara y batería	estupendo, cámara, bueno, pantalla, batería, precio, rápido, bien, vida, encantar	vida batería, buena cámara, buena batería, batería dura, cámara encantar	buena vida batería, vida batería larga, vida batería increíble, encantar vida batería, encantar buen precio
3	Usabilidad	app, usar, tiempo, pantalla, funcionar, querer, necesitar, nokia, cosa, bueno	fácil uso, fácil usuario, aprender usar, sistema operativo, estupendo fácil	estupendo fácil usar, encantar fácil uso, agradable fácil uso, funcionar estupendamente fácil, simple fácil uso
4	Tarjeta SIM	funcionar, sim, tarjeta, verizon, desbloquear, comprar, móvil, samsung, servicio, usar	tarjeta sim, verizon funciona, sim dual, compatible verizon, traer sim	tarjeta sim funcionar, trae tarjeta sim, verizon tarjeta sim, insertar tarjeta sim, pérdida tiempo dinero
5	Opinión general	funcionar, estupendo, nuevo, comprar, batería, cambiar, venir, bueno, encantar, pantalla	funciona genial, buen producto, dejar funcionar, compra feliz, envío rápido	dejar funcionar mes, funcionar perfectamente, funcionar genial encantar, llegar estupenda condición

Tabla 4.2: Tópicos, categoría, palabras clave, bigramas y trigramas más relevantes - Sector comercio electrónico

4.2.4. Análisis de sentimiento

Para concluir con el análisis de este conjunto de datos, se ha seguido la metodología descrita en la sección 3.5, con el fin de examinar las emociones subyacentes en las reseñas de los usuarios. De esta manera, mediante el uso de la estrategia VADER, se ha calculado la puntuación asociada al sentimiento expresado en cada una de las reseñas. Este enfoque permite cuantificar y clasificar el tono emocional de las opiniones, lo que resulta de gran utilidad para comprender la actitud general de los consumidores hacia el producto analizado. Así, la distribución de la puntuación compuesta se visualiza en la Figura 4.23, mediante un gráfico de cajas. Se puede apreciar que, a diferencia del conjunto de datos anterior, en este caso hay un volumen relevante de puntuaciones negativas. Así mismo, tenemos una distribución asimétrica a la izquierda con una mayor concentración en valores positivos de sentimiento. Específicamente, la mediana tiene un valor de 0.62, lo que indica que el 50 % de los datos están por encima de ese valor. Además, el primer cuartil se sitúa en 0 (sentimiento neutro), por lo que solo el 25 % de los datos tiene asociado un sentimientos negativo.

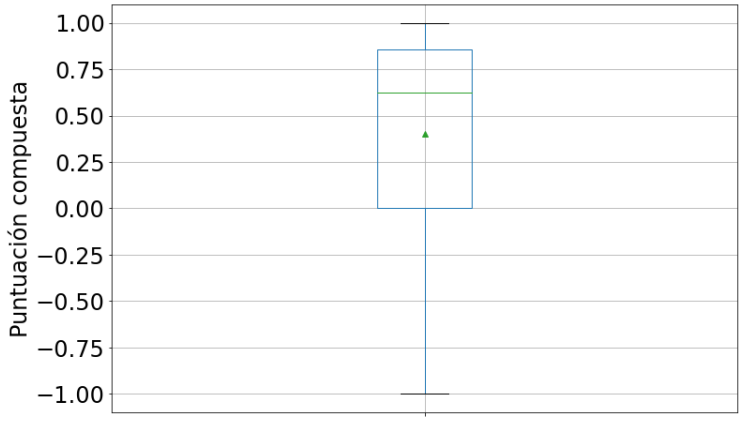


Figura 4.23: Gráfico de cajas de la puntuación compuesta - Sector comercio electrónico
Fuente: Elaboración propia

Por otro lado, en la Figura 4.24 se muestra la evolución temporal del sentimiento expresado por los clientes en sus reseñas, acumulando su valor diariamente. Se observa una tendencia al alza, pero no es hasta 2019 cuando se produce un aumento más pronunciado en cuanto al sentimiento expresado.

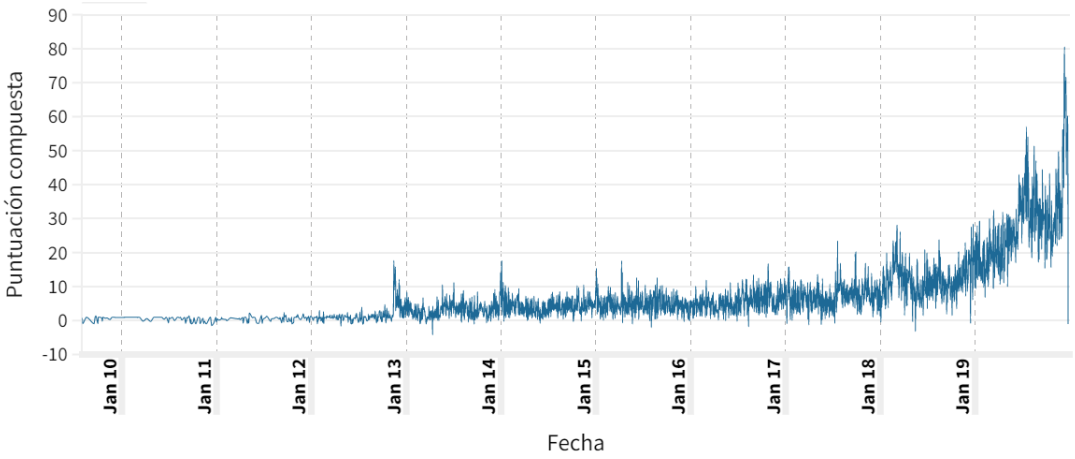
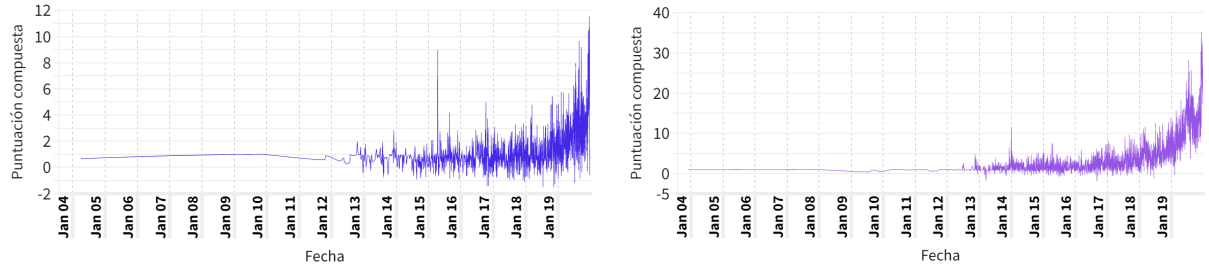


Figura 4.24: Evolución del sentimiento a lo largo de los años - Sector comercio electrónico
Fuente: Elaboración propia

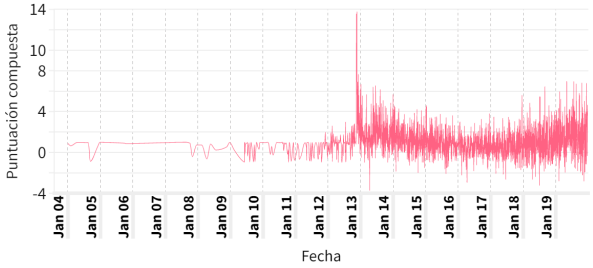
A continuación, se va a estudiar la evolución del sentimiento para cada uno de los tópicos modelados. En todos ellos se puede observar que, aunque existan datos desde el 2003, no es hasta el 2012 cuando empieza a haber un número elevado de reseñas, por lo que antes de esta fecha el análisis del sentimiento no tiene una dinámica relevante a analizar. Empezando por el tópico 1 'Desbloqueo y almacenamiento' (ver Figura 4.25a) se comprueba que los usuarios tienen un sentimiento mayoritariamente positivo con una tendencia creciente. Estos resultados son coherentes con la información de la Tabla 4.2, donde los usuarios expresan su conformidad con el funcionamiento de los distintos medios de desbloqueo de los dispositivos y la capacidad de almacenamiento. En el tópico 2 'Cámara y batería' (ver Figura 4.25b) se observa que el sentimiento es en su mayor parte positivo, con una tendencia creciente. De nuevo son resultados consistentes con los trigramas de la Tabla 4.2, donde los usuarios utilizan términos como 'bueno' o 'encantar' para referirse a estas características de los dispositivos. Además, los clientes consideran la vida de la batería como algo muy positivo calificándola como 'larga' y 'buena'. Por su parte, en el tópico 3 'Usabilidad' (ver Figura 4.25c) no se encuentra una tendencia tan clara, si no que el volumen de datos positivos y negativos es significativo en ambos casos. Las causas de estos sentimientos negativos no se pueden analizar con claridad con la información recogida en la Tabla 4.2, por ello como futuro desarrollo se podría estudiar con más detenimiento esta categoría para identificar posibles puntos de mejora.

Con respecto al tópico 4 'Tarjeta SIM' (ver Figura 4.25d), se trata de nuevo de un tópico donde el sentimiento no presenta una clara tendencia, alcanzando valores de sentimiento positivo elevados pero con un componente negativo importante. Son resultados coherentes con la información recogida en la Tabla 4.2, donde los usuarios se muestran conformes con el funcionamiento de las tarjetas SIM y el servicio de red ofrecido por compañías como Verizon. Sin embargo, también se encuentran trigramas como 'pérdida tiempo dinero', que estarían relacionados con sentimientos negativos de inconformidad.

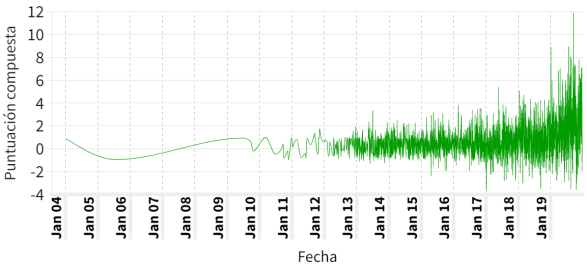


(a) Tópico 1: Desbloqueo y almacenamiento

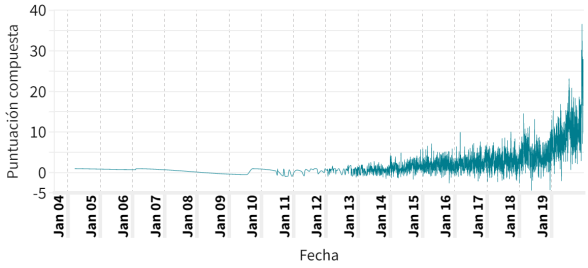
(b) Tópico 2: Cámara y batería



(c) Tópico 3: Usabilidad



(d) Tópico 4: Tarjeta SIM



(e) Tópico 5: Opinión general

Figura 4.25: Evolución del sentimiento a lo largo del tiempo para cada tópico - Sector comercio electrónico

Fuente: Elaboración propia

Por último, en el tópico 5 ‘Opinión general’ (ver Figura 4.25e) se aprecia un sentimiento mayoritariamente positivo, con una tendencia creciente. Resultados consistentes con los trigramas de la Tabla 4.2 donde los clientes expresan su satisfacción con el funcionamiento y el envío de los productos (‘funcionar perfectamente’ o ‘llegar estupenda condición´). Sin embargo, también se observan zonas donde la puntuación compuesta alcanza valores negativos. Este resultado puede deberse principalmente al descontento que sienten los usuarios con la duración del producto, ya que indican, por ejemplo, que deja de funcionar poco tiempo después de la compra.

Tras analizar la evolución temporal de los tópicos, así como el sentimiento general, se puede concluir que el sentimiento que predomina es positivo, siendo las características que más aprecian los usuarios la calidad de la cámara y la larga duración de la vida de la batería. Aunque en menor medida, el sentimiento asociado a los medios de desbloqueo y almacenamiento de los dispositivos es otro tema que genera satisfacción en los clientes. Como puntos a mejorar estarían los esfuerzos asociados a aumentar el tiempo de funcionamiento de los

dispositivos.

4.3. Restauración

4.3.1. Selección y pre-procesamiento de los datos

El conjunto de datos seleccionado (Ng, 2022) corresponde a un dataset de reseñas recogidas de TripAdvisor, sobre restaurantes de varios estados y ciudades de Malasia. El conjunto de datos inicial cuenta con un total de 139.763 estradas desde octubre de 2008 hasta abril de 2022. Para el tema de restauración, se ha elegido este conjunto de datos debido al volumen considerable de observaciones, lo que contribuye a la obtención de resultados y conclusiones más fiables. Por otro lado, al ser reseñas de diferentes restaurantes y zonas de Malasia, permite identificar, de forma global en este país, las características que afectan la opinión de los usuarios cuando se trata de restaurantes, además de identificar los puntos más importantes y críticos a la hora de establecer una valoración.

Para comenzar, se ha llevado a cabo un análisis preliminar de estos datos. Así, en la Figura 4.26 se muestra un gráfico descriptivo sobre la distribución del número de palabras por reseña. De esta manera, se observa que el 75 % de los datos tiene menos de 82 palabras. Además, se puede comprobar que los datos se encuentran más dispersos en la parte superior, evidenciando una asimetría positiva, siendo el máximo número de palabras por comentario de 160 (previa eliminación de *outliers*). Esta visualización nos indica una menor concentración en valores altos del número de palabras (opiniones largas) que en valores cortos.

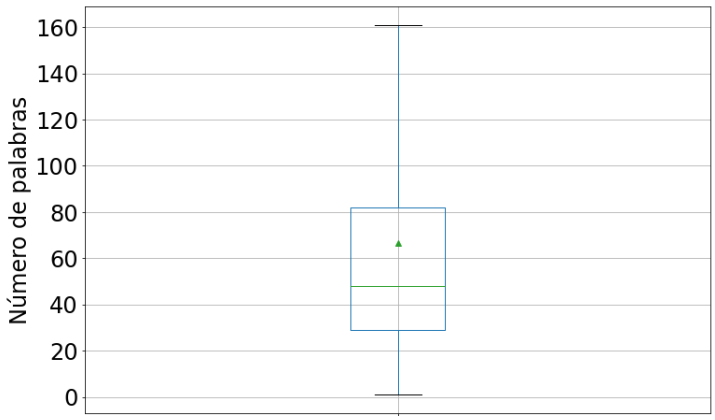


Figura 4.26: Diagrama de cajas del número de palabras por reseña - Sector restauración
Fuente: Elaboración propia

Por otro lado, en la Figura 4.27 se visualiza la evolución temporal del número de reseñas. Se puede observar como los datos presentan una tendencia de aumento a lo largo de los años, con algunos picos en fechas puntuales. Esto se debe probablemente, como ya se ha comentado con anterioridad, al aumento del número de usuarios que escriben sus opiniones *online* sobre

sus experiencias, en las últimas décadas. Una característica a destacar de la gráfica es el descenso pronunciado que se produce en febrero de 2020, no recuperando un crecimiento hasta julio del 2020. Este periodo particular está asociado al inicio de la pandemia del Covid-19, la cual afectó a muchos sectores, entre ellos el de la restauración, por lo que es normal que no existan entradas en ese periodo debido a los confinamientos y la preocupación de la sociedad por evitar potenciales lugares de contagio, como podían ser los restaurantes.

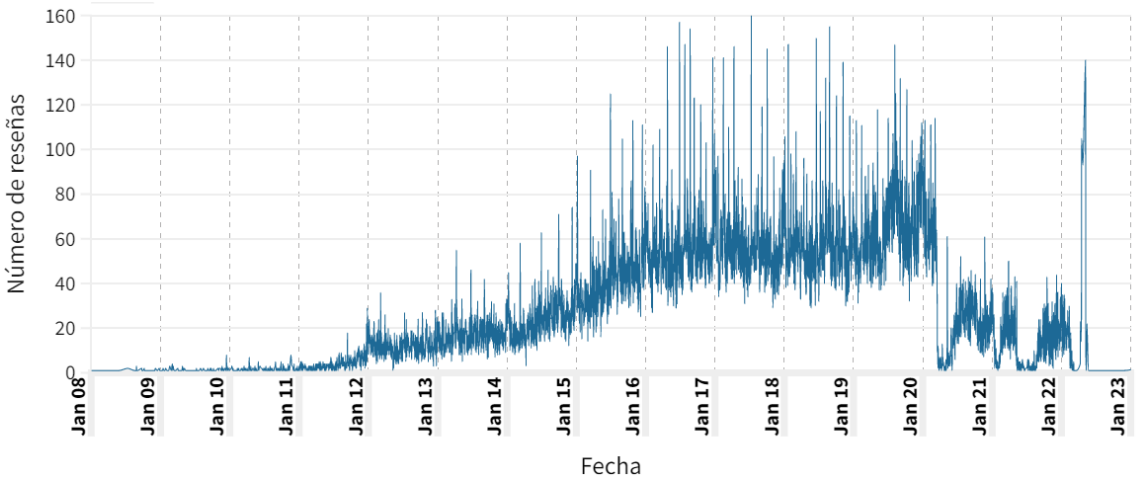


Figura 4.27: Evolución temporal del número de reseñas - Sector restauración
Fuente: Elaboración propia

Una vez recolectados los datos, es necesario realizar las tareas de limpieza y pre-procesamiento descritas en detalle en la sección 3.3. Estas tareas incluyen la eliminación de celdas vacías, de signos de puntuación, así como de numeración y caracteres especiales. Posteriormente, se filtran los extremos y se normalizan los datos para que todas las palabras estén en minúscula. A diferencia de los casos anteriores, el número de reseñas tras llevar a cabo estas tareas sigue siendo el mismo, lo que sugiere que la recolección de los datos de la plataforma incluyó una estrategia de pre-procesado que garantizaba una buena calidad de los datos adquiridos.

4.3.2. Análisis descriptivo de N-gramas

En esta sección se muestran los resultados obtenidos del análisis descriptivo de N-gramas realizado sobre el conjunto de datos. Para comenzar se muestra en la Figura 4.28 una nube de palabras. En este caso se observa que las palabras que aparecen con mayor frecuencia son: ‘comida’, ‘bueno’, ‘restaurante’, ‘servicio’, ‘estupendo’ y ‘personal’. Los clientes hacen referencia a características de los restaurantes que pueden afectar a la satisfacción del cliente como pueden ser la comida o el trato en el servicio. También usan adjetivos positivos como ‘bueno’, ‘estupendo’ ‘delicioso’ o ‘amigable’ en sus descripciones sobre ellos.

Figura 4.28: Nube de palabras - Sector restauración
Fuente: Elaboración propia

Para profundizar en este análisis se ha realizado un estudio de los N-gramas más importantes mediante el cálculo de la puntuación TF-IDF de cada palabra, según se ha explicado en el capítulo 3. Como resultado, se muestra en la Figura 4.29 la distribución de esta métrica para los unigramas, o palabras individuales. De ellas las diez más relevantes son: ‘comida’, ‘bueno’, ‘servicio’, ‘estupendo’, ‘lugar’, ‘restaurante’, ‘agradable’, ‘personal’, ‘amistoso’ y ‘venir’. Estos resultados están en concordancia con los obtenido en la nube de palabras, destacando que los usuarios hablan sobretodo del servicio y la comida y utilizan adjetivos positivos para referirse a ello en sus reseñas.

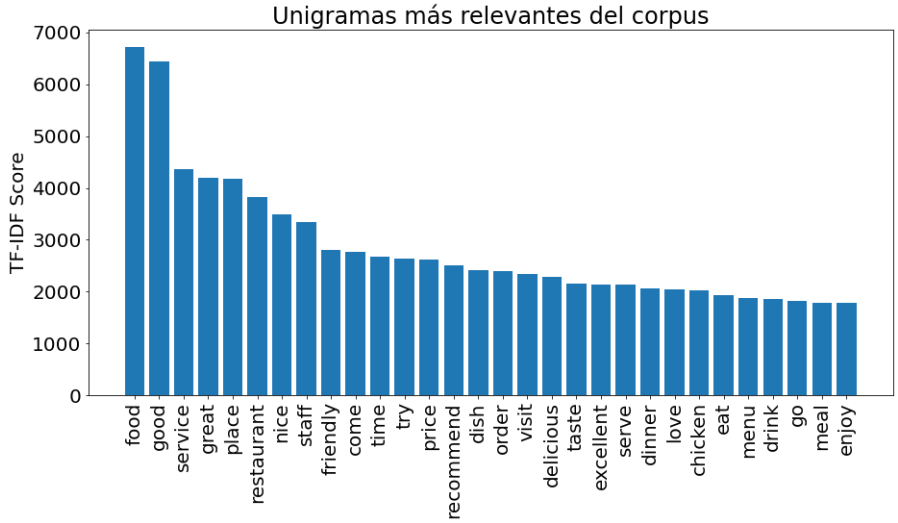


Figura 4.29: TF-IDF de los 30 unigramas más relevantes - Sector restauración
Fuente: Elaboración propia

Para continuar, se ha calculado el TF-IDF de cada bigrama o conjunto de dos palabras del corpus. En este caso los diez más relevantes están asociados a la ‘buena comida’, ‘buen servicio’, ‘altamente recomendado’, y ‘personal amistoso’. De nuevo se comprueba que los

clientes parecen estar satisfechos con el servicio y el personal de los restaurantes, destacado la comida y el servicio con diferentes calificativos positivos.

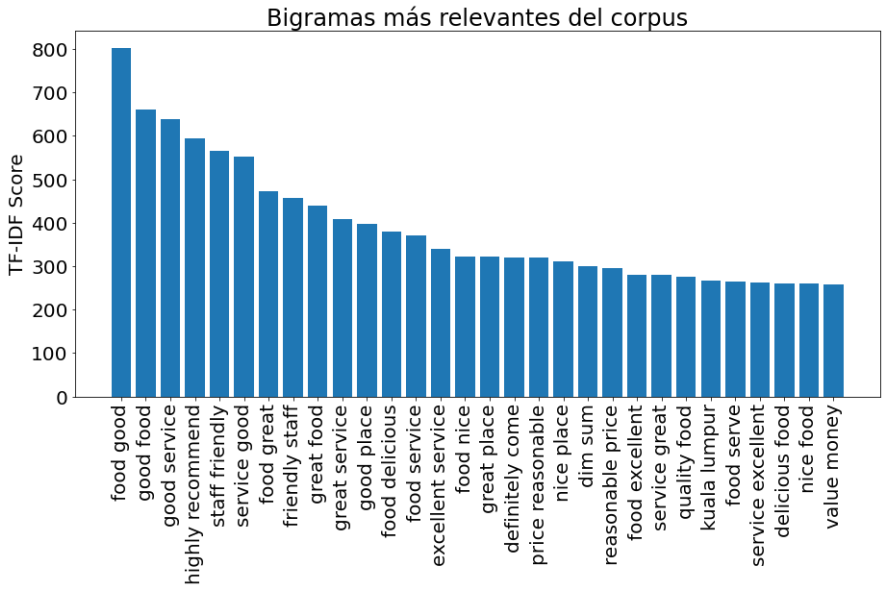


Figura 4.30: TF-IDF de los 30 bigramas más relevantes - Sector restauración
Fuente: Elaboración propia

Por último, se ha calculado el TF-IDF de cada trigrama o conjunto de tres palabras. En este caso los diez más relevantes están asociados con temas como: ‘comida buena servicio’, ‘personal amistoso útil’, y ’comida precio razonable’. De nuevos se comprueba que los clientes están muy satisfechos con el personal, el servicio y la comida de los restaurante pero además se introduce una nueva característica en las reseñas, el precio. Los clientes parece que encuentran razonable el precio que están pagando y consideran que el valor está bien en relación a lo que pagan por las consumiciones.

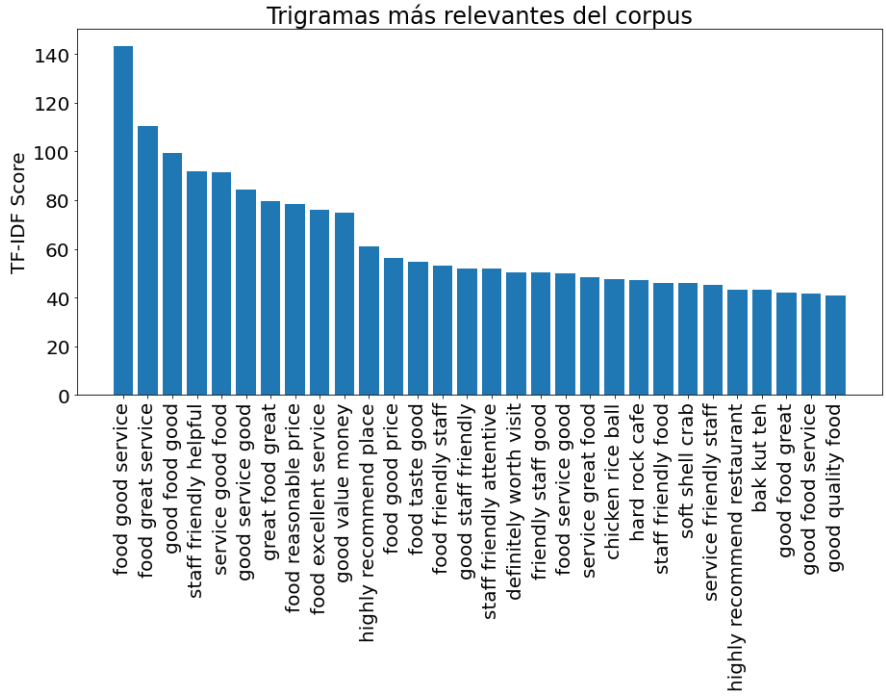


Figura 4.31: TF-IDF de los 30 trigramas más relevantes - Sector restauración
Fuente: Elaboración propia

4.3.3. Modelado de tópicos

En este apartado, se detallan los resultados obtenidos en el proceso del modelado de tópicos, descrito en la sección 3.4. Para comenzar, es necesario ajustar el número de tópicos idóneo para estos datos. Para ello, se ha calculado el índice de coherencia para diferentes valores de número de tópicos, como se visualiza en la gráfica 4.32. Como se ha explicado en la sección 3.4, se busca aquel número de tópicos que coincida con un pico de crecimiento en el índice de coherencia. En este caso, es posible observar que para k=4, k=8 y k=13, el índice de coherencia presenta un pico de crecimiento. Con estos valores de k, se realiza el análisis de la distancia entre los tópicos, mediante la visualización del mapa de distancia intertópica. Se busca aquel modelo que tenga el mayor índice de coherencia pero que también genere tópicos interpretables, lo cual puede evidenciarse con un mapa en el cual los círculos no presenten mucha superposición.

Del resultado del análisis, puede evidenciarse que con $k = 13$ (mayor índice de coherencia) se obtiene la distancia intertópica mostrada en la Figura 4.33. Como es posible observar en el mapa, aunque existe una pequeña superposición, su naturaleza es parcial y en ningún caso hay tópicos contenidos dentro de otros, lo que significa que hay elementos diferenciadores para cada una de las categorías modeladas.

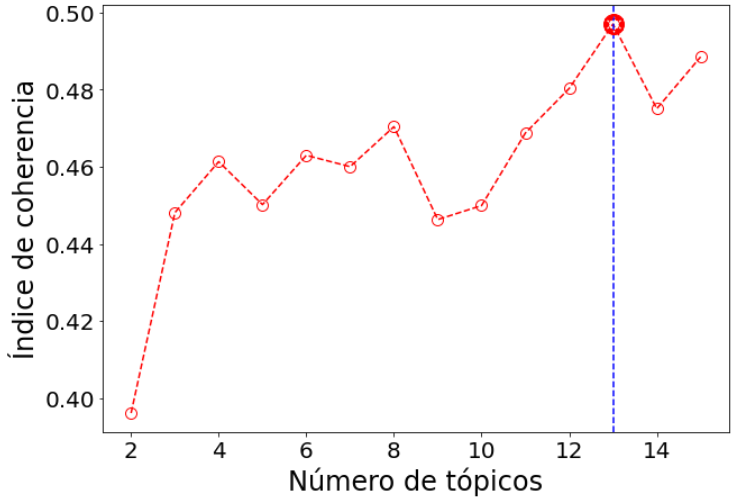


Figura 4.32: Índice de coherencia para cada número de tópicos - Sector restauración
Fuente: Elaboración propia

Figura 4.33: Distancia entre tópicos para *k*=13 - Sector restauración
Fuente: Elaboración propia

Por otro lado, se ha calculado la distribución de los tópicos en el corpus (ver Figura 4.34). En ella se puede comprobar que los tópicos con más reseñas asociadas son el 3 (‘Opinión del servicio’), el 6 (‘Esperas’) y el 8 (‘Comida india’). Es de esperar que los dos primeros aparezcan en muchas reseñas ya que son temas que afectan en gran medida a la opinión de los usuarios y que normalmente tienen mucho peso a lo hora de emitir una valoración. Por otro lado, los tópicos con menor presencia en el corpus, incluyen el 5 (‘Opinión de la experiencia’),

el 10 (‘Carnes’) y el 13 (‘Comida japonesa’). No es de extrañar el caso de los dos últimos ya que hablan de temas que son muy particulares a las especialidades que ofrece cada restaurante y no pueden ser generalizables a las experiencias, en contraste con otras categorías como el precio o el servicio. Con respecto al tópico 5 (‘Opinión de la experiencia’), la baja pertenencia de reseñas a este tópico puede estar causada por que trata generalidades de temas que están presentes en otras categorías.

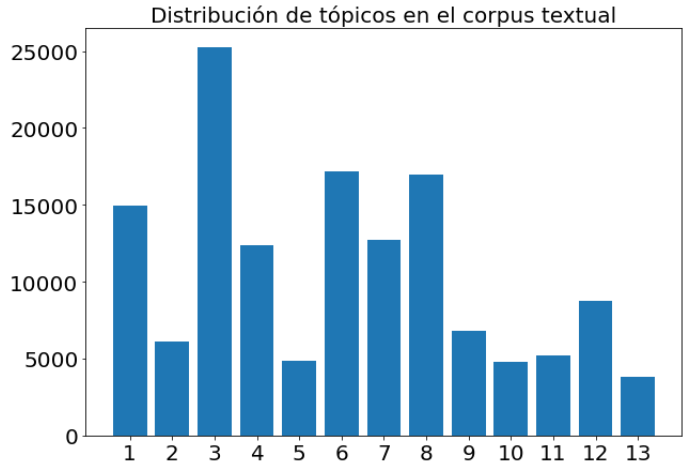


Figura 4.34: Distribución de tópicos en el corpus textual - Sector restauración
Fuente: Elaboración propia

En la Tabla 4.3 se detallan los tópicos modelados junto con las diez palabras clave de cada uno de ellos, además de los cinco bigramas y trigramas más relevantes de cada categoría, el resto de ellos, en su idioma original, se han incluido en una tabla en el Anexo C.

Como resultado, en primer lugar es posible encontrar dos grandes categorías de tópicos. Por un lado, se encuentran los tópicos 1 (‘Opciones de menú’), 2 (‘Comida italiana’), 8 (‘Comida india’), 10 (‘Carnes’), 12 (‘Comida china’) y 13 (‘Comida japonesa’), donde los usuarios hacen referencia a distintos platos y preparaciones, característicos de la cocina en este país. Estos tópicos suelen estar acompañados de adjetivos positivos como ‘bueno’, ‘buen servicio’, ‘altamente recomendado’, ‘buena comida’, etc, lo cual indica que los usuarios están satisfechos con la oferta proporcionada por estos restaurantes. Por otro lado, se encuentran los tópicos 3 (‘Opinión del servicio’), 4 (‘Opinión del lugar’), 5 (‘Oferta para todos los gustos’), 6 (‘Esperas’), 7 (‘Opinión general’), 9 (‘Precios y otros servicios’) y 11 (‘Tipos de celebración’), los cuales relacionan temáticas fuera de la oferta gastronómica, pero igualmente importantes a la hora de formar una opinión en el consumidor.

Analizando estos tópicos con más detalle, es posible identificar que en el tópico 3 (‘Opinión del servicio’), los usuarios hablan del personal y el trato recibido en los distintos restaurantes, frecuentemente calificado de forma positiva, con conceptos recurrentes como ‘buen servicio’ y personal ‘amistoso’ o ‘útil’. Por su parte, en el tópico 4 (‘Opinión del lugar’), se observa que los clientes parecen estar satisfechos con el ambiente del lugar, calificándolo

como ‘estupendo’ o ‘agradable’. Además, se podría destacar que los usuarios manifiestan una actitud positiva hacia los lugares tranquilos donde es posible relajarse y pasar tiempo con amigos. Para complementar esta línea, el tópico 5 (‘Oferta para todos los gustos’) incluye descripciones de los usuarios sobre temas generales de la experiencia, como puede ser la oferta de platos sin gluten o la carta de vinos.

Por otra parte, en el tópico 6 (‘Esperas’) los clientes hablan de un tema que generalmente asocia sentimientos negativos, como son las esperas. Los usuarios expresan su inconformidad con la lentitud del servicio, quejándose de esperas largas y resultando en una mala opinión del servicio del restaurante. A su vez, el tópico 7 (‘Opinión general’) involucra opiniones generales sobre la visita, manifestando principalmente su opinión positiva de esta experiencia. Así, se encuentran conceptos recomendando la visita y calificando de estupenda tanto la comida como el servicio. En el tópico 9 (‘Precios y otros servicios’) los usuarios tratan un tema que es de importancia a la hora de generar una buena sensación en el consumidor, el precio, el cual parece ser razonable para la mayoría de las opiniones. Se incluyen conceptos adicionales asociados a la facilidad para encontrar aparcamiento o el horario de apertura. Por último, en el tópico 11 (‘Tipos de celebración’), los clientes describen las razones (eventos) por las que han ido al restaurante, como pueden ser cumpleaños o año nuevo. Además también se menciona el nombre de un chef, otra característica que podría contribuir en gran medida a la reputación de los restaurantes.

Estas experiencias parecen ser positivas en la mayoría de los restaurantes, los cuales cuentan con revisiones asociadas a un personal amistoso y a espacios agradables y tranquilos, características que contribuyen a mejorar la experiencia del cliente. Además, comentarios asociados al precio razonable ayudan a generar una sensación de satisfacción en los usuarios. Sin embargo, el tópico 6 (‘Esperas’) involucra temas que pueden generar descontento en lo clientes, como por ejemplo largos tiempos de espera, por lo que los esfuerzos de mejora en esta linea de mercado podrían centrarse en este tema.

Tópico	Categoría	Palabras clave	Bigramas más relevantes	Trigramas más relevantes
1	Opciones de menú	pollo, arroz, bueno, probar, comida, curry, plato, pedir, pescado, gusto	pollo con arroz, nasi lemak, bola de arroz, arroz frito, buena comida	bola de arroz y pollo, bak kut teh (guiso de chuleta de cerdo), curry de cabeza de pescado, arroz en hoja de plátano, cabeza de pescado curry
2	Comida italiana	pizza, pasta, bueno, postre, crema, helado, probar, cordero, tarta, ensalada	helado, altamente recomendado, sopa de setas, comida italiana, masa fina	helado de vainilla, tarta de chocolate lava, pasta de calamar en su tinta, pizza de masa fina, comida italiana autentica
3	Opinión del servicio	comida, bueno, servicio, agradable, personal, amistoso, lugar, estupendo, precio, recomendar	buen servicio, personal amistoso, servicio bueno, excelente servicio, amistoso útil	buen servicio buen, personal útil amistoso, servicio bueno comida, personal amistoso bueno, buen servicio amistoso
4	Opinión del lugar	lugar, bueno, estupendo, bebida, agradable, comida, cerveza, musica, café, bar	lugar estupendo, lugar bueno, lugar agradable, lugar tranquilo, buena atmósfera	lugar agradable tranquilo, buen lugar tranquilo, buen lugar relajarse, lugar agradable estar, lugar estar amigos
5	Oferta para todos los gustos	personal, mesa, restaurante, vino, cena, menú, comida, noche, reservar, plato	experiencia de comida, altamente recomendado, buena cena, lista vinos, sin gluten	buena experiencia de cena, restaurante altamente recomendado, reservar mesa antes, buena lista vinos, opción sin gluten
6	Esperas	comida, pedir, tiempo, mesa, venir, esperar, restaurante, servicio, servir, preguntar	largo tiempo, servicio lento, espera larga, mal servicio, 10 minutos	tardar mucho tiempo, esperar mucho tiempo, esperar media hora, 10 cambios de servicio, esperar 20 minutos
7	Opinión general	estupendo, comida, servicio, increíble, personal, gracias, experiencia, recomendar, amistoso, maravilloso	altamente recomendado, gran servicio, comida estupenda, estupenda comida, excelente servicio	comida estupenda servicio, estupenda comida estupenda, comida excelente servicio, lugar altamente recomendado, comida buena servicio
8	Comida india	comida, restaurante, bueno, lugar, visitar, local, precio, recomendar, comida, servicio	comida india, comida local, comida calidad, valor dinero, precio razonable	buena comida india, comida sur india, buena calidad comida, comida norte india, comida precio razonable
9	Precios y otros servicios	comida, restaurante, precio, lugar, kuching (ciudad), bueno, aparcamiento, localizar, melaka (ciudad), área	precio razonable, espacio aparcamiento, aire acondicionado, centro comercial, hora punta	comida precio razonable, amplio espacio aparcamiento, comida precio medio, abierto 24 horas, difícil encontrar aparcamiento
10	Carnes	filete, bueno, pedir, hamburguesa, carne, pollo, vaca, costilla, cerdo, cocinar	medio hecho, costilla de cerdo, chuleta de pollo, cocinado perfecto, buen filete	filete de costilla, salsa de pimienta negra, filete cocinado perfecto, cocinado medio hecho, pedir medio hecho
11	Tipos de celebración	desayuno, bueno, cena, comida, cumpleaños, familia, bufete, servicio, comida, amigo	celebración cumpleaños, canción cumpleaños, te tarde, chef nathan, venir definitivamente	cantar el cumpleaños feliz, celebrar cumpleaños amigo, celebrar cumpleaños mujer, año nuevo chino, chef nathan steven
12	Comida china	plato, sopa, cerdo, sabor, pollo, freír, fideo, bueno, pato, salsa	dim sum, cerdo barbacoa, arroz frito, rollo primavera, comida china	xiao long bao (plato tradicional chino), rollo de fideo de arroz, restaurante de dim sum, sopa tom yam, cangrejo de cascara blanda
13	Comida japonesa	menú, fresco, experiencia, comida, sushi, set, chef, plato, japones, pescado	restaurante japones, comida japonesa, shashimi fresco, marisco fresco, cangrejo blando	cangrejo de cascara blanda, buen restaurante japonés, experiencia cena única, sopa bacalao negro, helado de te

Tabla 4.3: Tópicos, categoría, palabras clave, bigramas y trigramas más relevantes - Sector restauración.

4.3.4. Análisis de sentimientos

Finalmente, se se procede a analizar el sentimiento de cada revisión mediante la implementación del algoritmo VADER, explicado en la sección 3.5. Como resultado, en la Figura 4.35 se observa la distribución de las puntuaciones obtenidas para las revisiones, eliminando los datos atípicos de la distribución. Se comprueba que las opiniones de los usuarios son en su mayor parte muy positivas, contando con un valor de mediana de 0.9 (siendo 1 el valor máximo asociado a un sentimiento positivo). Analizando el primer cuartil se observa que solo el 25 % de los datos se encuentra con una puntuación inferior a 0.75. Además, a excepción de los valores atípicos que se han eliminado del análisis (*outliers*) la puntuación más baja se encuentra por encima de 0.4. Por último, se puede comprobar que la distribución tiene una asimetría negativa, evidenciando una mayor concentración de puntuaciones en valores altos de sentimiento. A su vez, en la Figura 4.36 se muestra la evolución temporal del sentimiento expresado en las reseñas. Como se puede observar, el sentimiento que expresan los clientes es muy positivo, destacándose su tendencia al alza. En enero de 2020 se observa un descenso en las opiniones que no comienza a crecer de nuevo hasta mayo del 2020, recuperando valores elevados en mayo de 2022. Este resultado coincide con la pandemia del Covid-19, que afectó a nivel mundial a diversos sectores, incluyendo el de la restauración.

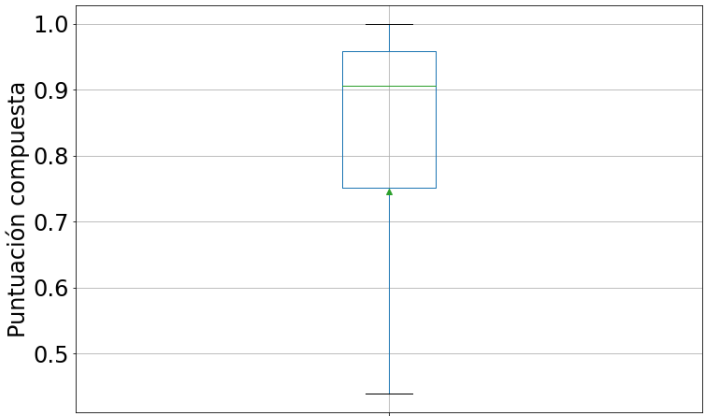


Figura 4.35: Gráfico de caja de la puntuación compuesta - Sector restauración
Fuente: Elaboración propia

A continuación, se analiza la evolución del sentimiento a lo largo de los años para cada uno de los tópicos. Si se analiza el grupo de tópicos que hacen referencia a las distintas opciones del menú y platos que ofrecen los restaurantes (tópicos 1 ‘Opciones de menú’ en la Figura 4.37a, 2 ‘Comida italiana’ en la Figura 4.37b, 8 ‘Comida india’ en la Figura 4.37c, 10 ‘Carnes’ en la Figura 4.37d, 12 ‘Comida china’ en la Figura 4.37e y 13 ‘Comida japonesa’ en la Figura 4.37f), se observa que todos ellos presentan una puntuación mayoritariamente positiva. En el tópico 10, por su parte, se puede apreciar una puntuación inferior a la del resto de tópicos de este grupo, además de contar con mayor cantidad de picos negativos (ver Figura 4.37d).

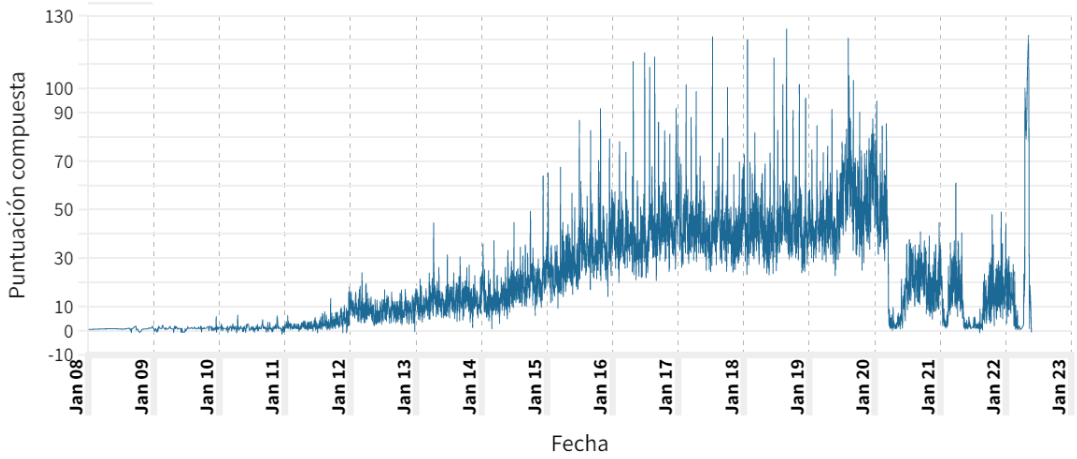


Figura 4.36: Evolución del sentimiento a lo largo de los años - Sector restauración
Fuente: Elaboración propia

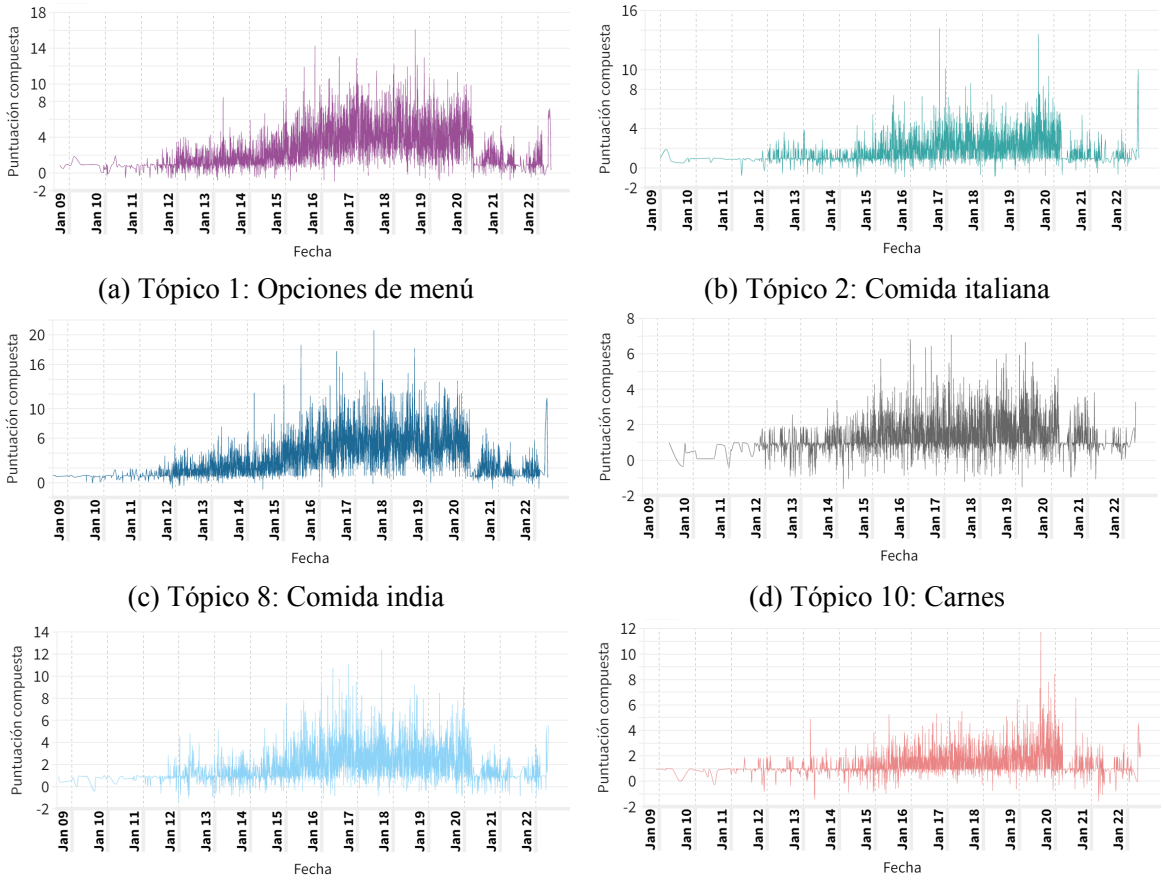
Analizando ahora el grupo de tópicos que abordan otros temas que no están relacionados con la comida ofrecida, se encuentra en primer lugar el tópico 3 ‘Opinión del servicio’ (ver Figura 4.38a). Se observa que es uno de los tópicos con mayores puntuaciones, lo que indica un alto grado de satisfacción por parte de los clientes. Esto no es de extrañar ya que en la Tabla 4.3, los clientes utilizaban términos como ‘bueno’, ‘útil’ o ‘amistoso’ para referirse al servicio. En el tópico 4 ‘Opinión del lugar’, por su parte, se observa de nuevo un alto grado de satisfacción por parte de los usuarios, sin estacionalidad y presentando picos puntuales hacia puntuaciones positivas (ver Figura 4.38b). De nuevo, este resultado está en concordancia con los obtenidos en la Tabla 4.3, donde los usuarios se referían a los lugares como ‘estupendo’ o ‘agradable’.

Con respecto al tópico 5 ‘Oferta para todos los gustos’, se observa mayoritariamente un sentimiento positivo (ver Figura 4.38c). Estos resultados son razonables ya que en los bigramas y trigramas obtenidos para este tópico los usuarios recomiendan los restaurantes y se refieren a las experiencias como ‘buenas’. Sin embargo, se pueden apreciar picos puntuales hacia valores negativos de sentimiento, lo cual podría estar relacionado con algunos temas que se mencionan en los trigramas de la Tabla 4.3, como puede ser la reserva anticipada de la mesa (‘reservar mesa antes’). En este contexto, es probable que si los clientes recomiendan reservar con antelación es porque suele ser difícil ser atendido sin reserva o que se requiera esperar, lo que puede repercutir en una opinión negativa. En esta misma línea, el tópico 6 ‘Esperas’ , se trata de un tópico que alcanza los valores negativos mas grandes en comparación con el resto de tópicos (ver Figura 4.38d). Este resultado no es de extrañar, puesto que en los trigramas de la Tabla 4.3, los clientes hablan de largas esperas y servicio lento, características negativas que afectan en gran medida a la experiencia de los clientes de un restaurante.

En el tópico 7 ‘Opinión general’ se aprecia un sentimiento muy positivo, alcanzando picos con grandes puntuaciones compuestas, y prácticamente sin ningún valor negativo (ver Figura 4.38e). De nuevo son resultados esperados ya que en los bigramas y trigramas ob-

tenidos para este tópico los usuarios utilizan en sus términos más relevantes ‘excelente’ o ‘estupendo’ para hablar de su experiencia. Por su parte, para el tópico 9 ‘Precios y otros servicios’ se observa que predomina un sentimiento mayoritariamente positivo, que concuerda con los trigramas de la Tabla 4.3 donde los usuarios hacen referencia a unos precios razonables (ver Figura 4.38f). Sin embargo, también se pueden apreciar zonas de la gráfica donde la puntuación compuesta tiene valores negativos. Este resultado puede estar relacionado con servicios como el aparcamiento, ya que los usuarios incluían en los conceptos de mayor relevancia, trigramas asociados a ‘difícil encontrar aparcamiento’. Por último, en el tópico 11 ‘Tipos de celebración’, se trata de un tópico que no presenta una tendencia significativa, si no que mantiene valores mayoritariamente positivos con picos puntuales hacia puntuaciones positivas (ver Figura 4.38g).

Tras analizar la evolución temporal de estos tópicos, se puede concluir que los clientes de estos restaurantes tienen un sentimiento mayoritariamente positivo, lo que se refleja en sus reseñas. Los temas más tratados en las reseñas son los referentes a las distintas opciones del menú, donde los clientes hablan de los platos que han probado y utilizan adjetivos positivos para referirse a ellos. Otros de los temas que más contribuyen a esta satisfacción es el servicio y la atmósfera y el ambiente de los restaurantes. Entre los temas que se podrían mejorar, para aumentar el grado de satisfacción del cliente, estarían la reducción de los tiempos de espera y la mejora de servicios adicionales como puede ser el aparcamiento. Además, también podría analizarse en más detalle los resultados del tópico 10 ‘Carnes’, con el fin de explorar posibles puntos de mejora asociados directamente a los clientes de este tipo de restaurantes.



(a) Tópico 1: Opciones de menú

(b) Tópico 2: Comida italiana

(c) Tópico 8: Comida india

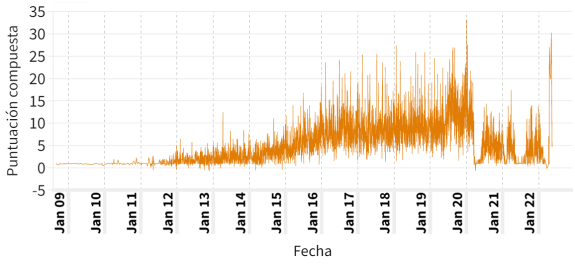
(d) Tópico 10: Carnes

(e) Tópico 12: Comida china

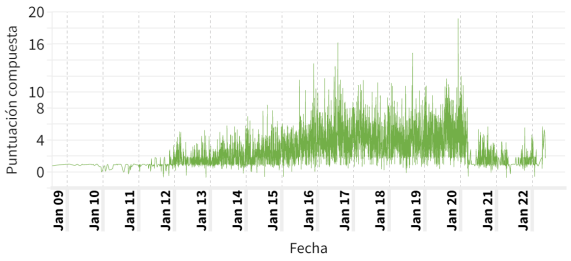
(f) Tópico 13: Comida japonesa

Figura 4.37: Evolución del sentimiento a lo largo del tiempo para los tópicos relacionados con el tipo de comida - Sector restauración

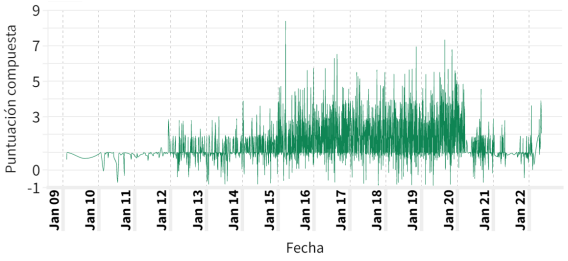
Fuente: Elaboración propia



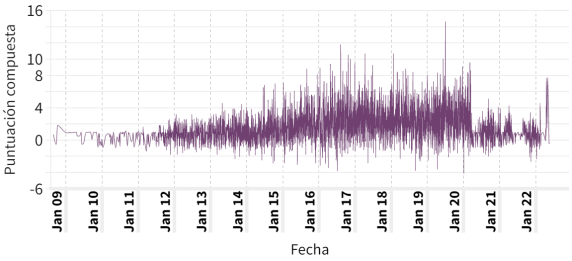
(a) Tópico 3: Opinión del servicio



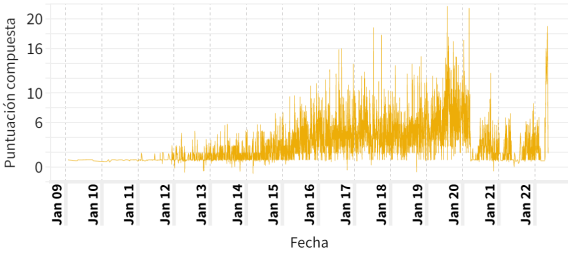
(b) Tópico 4: Opinión del lugar



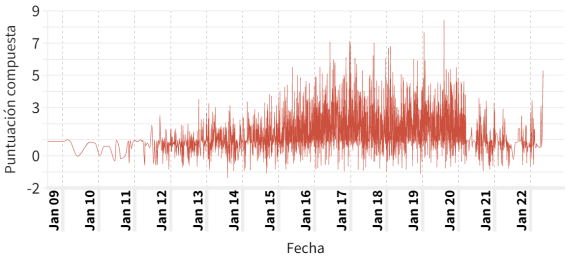
(c) Tópico 5: Oferta para todos los gustos



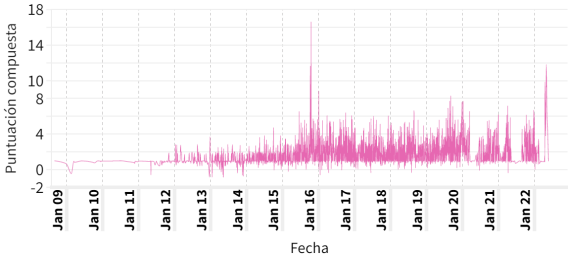
(d) Tópico 6: Esperas



(e) Tópico 7: Opinión general



(f) Tópico 9: Precios y otros servicios



(g) Tópico 11: Tipos de celebración

Figura 4.38: Evolución del sentimiento a lo largo del tiempo para el resto de tópicos - Sector restauración

Fuente: Elaboración propia

Capítulo 5

Conclusiones y trabajos futuros

El presente trabajo de grado se desarrolla en un contexto donde los clientes han incrementado el uso de nuevas tecnologías para expresar sus opiniones sobre productos y servicios, generando una gran abundancia de datos que son valiosos para productores y comercializadoras. Analizar los sentimientos de los consumidores y encontrar áreas de mejora que aumenten su satisfacción y fidelización se ha convertido en una prioridad para estas empresas. Sin embargo, la gran cantidad de información que se produce requiere métodos de análisis automático que permitan identificar los principales temas de discusión, así como el sentimiento involucrado en cada una de las categorías.

Con el fin de identificar las técnicas más adecuadas para llevar a cabo este estudio, se realizó un análisis exhaustivo del estado del arte de las estrategias de minería de textos y procesamiento de lenguaje natural, aplicadas al análisis de reseñas *online* para la extracción de información sobre la satisfacción de los usuarios. Tras el análisis, se destaca que la técnica LDA es ampliamente utilizada en tareas de modelado de tópicos debido a sus grandes ventajas, como la capacidad de procesar documentos de longitud variable y su fácil implementación, ya que no requiere de datos de entrenamiento. Por otra parte, respecto al análisis de sentimientos, se ha observado que las técnicas más empleadas incorporan reglas de decisión en diccionarios o algoritmos de aprendizaje automático convencionales. En este proyecto, se ha optado por la primera alternativa, mediante el uso del modelo VADER.

La metodología de análisis empleada en este proyecto ha involucrado una selección cuidadosa de los datos. En particular, se han escogido tres conjuntos de datos representativos de las categorías de turismo, comercio electrónico y restauración, las cuales son las más comúnmente analizadas en el estado del arte. Una vez que se han limpiado y procesado los datos, se ha realizado un análisis descriptivo preliminar para evaluar la longitud media de las opiniones y su evolución temporal. A continuación, se ha llevado a cabo un análisis de N-gramas, en el que se ha calculado el TF-IDF para cada unigrama, bigrama y trigrama, permitiendo obtener los términos más relevantes del conjunto de reseñas analizado. Una vez terminado este paso, se realizó el modelado de tópicos mediante el cálculo del índice de coherencia para diferen-

tes números de tópicos. Para concluir, se obtuvieron los mapas de distancia inter-tópica para aquellos valores de número de tópicos con mayor índice de coherencia, lo que permitió identificar el número óptimo de tópicos que caracterizaba el conjunto de datos. De esta forma, se logró obtener resultados interpretables y con elementos diferenciadores en el análisis de las reseñas. Finalmente, se ha llevado a cabo un análisis adicional para examinar el sentimiento presente en las reseñas. En este sentido, se ha calculado la puntuación compuesta tanto para el conjunto de datos en su totalidad como para cada uno de los tópicos.

Siguiendo la metodología descrita, se ha llevado a cabo el análisis de las reseñas correspondientes a las categorías mencionadas anteriormente. A continuación, se presentan las principales conclusiones obtenidas.

- **Turismo**:

  Se han analizado 37.343 reseñas de la Sagrada Familia para identificar diez tópicos principales. Los temas de mayor relevancia en el corpus están asociados a la "Compra de entradas", "Experiencia" y "Construcción". De esta manera, los usuarios destacan la compra anticipada de entradas, las vidrieras y la arquitectura de la basílica. El sentimiento general es positivo, especialmente en relación a las vidrieras. No obstante, se han detectado algunas áreas de mejora que generan sentimientos negativos, como la señalización del museo y la experiencia de compra de entradas que puede resultar larga y tediosa para los clientes.

- **Comercio electrónico**:

  En esta categoría, se han analizado 59.815 reseñas de distintas marcas de teléfonos móviles adquiridos en Amazon. Tras aplicar el modelado de tópicos, se identificaron cinco categorías, siendo los temas más comunes la batería, la cámara y la tarjeta SIM. En cuanto al sentimiento expresado por los usuarios, se observó que la mayoría de los comentarios son positivos, destacando la larga duración de la batería y la calidad de la cámara como las características más satisfactorias. No obstante, también se identificaron algunos aspectos que generan insatisfacción en los clientes, como la corta vida útil de los dispositivos y su usabilidad.

- **Restauración**:

  Para esta categoría, se ha realizado un análisis de 139.763 reseñas de restaurantes ubicados en diferentes estados y ciudades de Malasia, recogidas de TripAdvisor. Tras el modelado de tópicos, se han identificado trece categorías, siendo las más relevantes la "Opinión del servicio" y las "Esperas". Respecto al sentimiento expresado por los clientes, predominan las opiniones positivas, destacando la satisfacción con el personal de servicio y la atmósfera del lugar. Sin embargo, se han identificado aspectos que generan insatisfacción entre los clientes, como las esperas y la velocidad del servicio, así como la falta de servicios adicionales, como aparcamiento.

Finalmente, como trabajo futuro en esta área de análisis e investigación, las empresas podrían beneficiarse significativamente al implementar sistemas de recomendación alimentados con los resultados obtenidos del modelado de tópicos y análisis de sentimientos. De esta manera, estos algoritmos de decisión podrían ayudar a las empresas a identificar patrones en los datos que les permitan anticiparse a las necesidades de los clientes y ofrecerles soluciones personalizadas. Esto no solo mejoraría la satisfacción del cliente, sino que también podría aumentar la fidelidad y la retención de los mismos.

Apéndice A

Análisis sector turismo

Tópico	Categoría	Bigramas	Trigramas
1	Compra de entradas	buy ticket, ticket online, book ticket, ticket advance, book online, ticket line, audio guide, long line, worth visit, purchase ticket, ahead time, avoid long, online avoid, long queue, book advance, avoid queue, wait line, ticket ahead, pre book, sure book, audio tour, online advance, skip line, stain glass, line long, guide tour, line ticket, highly recommend, advance avoid, time slot	buy ticket online, book ticket online, buy ticket advance, book ticket advance, ticket ahead time, ticket online avoid, buy ticket line, ticket online advance, avoid long line, stain glass window, pre book ticket, sure book ticket, buy ticket ahead, sure buy ticket, purchase ticket online, avoid long queue, book ticket line, recommend buy ticket, line buy ticket, online avoid queue, ticket advance online, ticket online save, ticket advance avoid, online avoid long, book online avoid, visit buy ticket, online ahead time, book online advance, ticket online ahead, ticket online skip
2	Esperas para entrar	buy ticket, early morning, long line, ticket online, long queue, wait line, worth wait, audio guide, ticket line, line long, stain glass, book ticket, wait hour, worth visit, book online, late afternoon, arrive early, avoid crowd, stand line, queue long, purchase ticket, audio tour, line ticket, early avoid, guide tour, hour wait, long wait, visit early, ticket advance, time visit	buy ticket online, stain glass window, buy ticket line, hop hop bus, buy ticket advance, book ticket online, early avoid crowd, visit earl morning, purchase ticket online, early morning late, ticket ahead time, wait long line, take breath away, morning late afternoon, early morning avoid, line buy ticket, avoid long line, buy ticket ahead, long line ticket, arrive early morning, go early morning, early beat crowd, wait line hour, early queue long, ticket online avoid, book ticket line, early long queue, buy ticket early, go late afternoon, purchase ticket line
3	Experiencia	worth visit, breath away, definitely worth, amazing building, take breath, outside inside, inside outside, buy ticket, audio guide, stain glass, worth see, breath take, long queue, go inside, beautiful building, amazing place, worth wait, audio tour, place visit, ticket online, see outside, amazing architecture, book ticket, blow away, entrance fee, queue long, absolutely stunning, look outside, book online, spend hour	take breath away, stain glass window, definitely worth visit, buy ticket online,book ticket online, worth entrance fee, pay extra tower, hop hop bus, inside take breath, outside amazing inside, book ticket advance, buy ticket advance, worth go inside, building worth visit, definitely worth see, beautiful building see, amazing place visit, pre book ticket, amazing building see, worth visit inside, inside breath take, beautiful stain glass, inside definitely worth, worth pay inside, amazing amazing building, audio guide worth, building work go, stunning worth visit, audio tour worth, great view city
4	Museo	spend hour, audio guide, guide tour, worth visit, spend time, mind blow, amazing building, audio tour, attention detail, buy ticket, stain glass, gaudi work, gaudi masterpiece, ticket online, book ticket, visit museum, tour guide, amazing architecture, amazing structure, ticket advance, mind blowing, piece architecture, book online, spend day, museum underneath, simply amazing, definitely worth, hour look, work art, blow away	buy ticket online, spend hour look, stain glass window, plan spend hour, definitely worth visit, buy ticket advance, audio guide tour, book ticket online, spend couple hour, book ticket advance, spend hour inside, miss museum basement, sure visit museum, take breath away, museum low level, easily spend hour, mind blow architecture, visit museum basement, spend hour walk, amazing spend hour, purchase ticket online, hour look detail, easily spend day, miss museum underneath, pre book ticket, building worth visit, take guide tour, museum basement interesting, book guide tour, museum underneath church

Tópico	Categoría	Bigramas	Trigramas
5	Medios de accesibilidad al monumento	buy ticket, audio guide, metro station, ticket online, worth visit, tourist attraction, easy metro, book ticker, guide tour, audio tour, easy access, book online, ticket advance, entrance fee, spend hour, book advance, ticket line, stain glass, park guell, long line, purchase ticket, selfie stick, spend time, las ramblas, building site, amazing architecture, place visit, la caixa, tourist bus, easy find	buy ticket online, book ticket online, stain glass window, buy ticket advance, metro station right, near metro station, gaudy las vegas, hop hop bus, easily accessible metro, book entrance ticket, easy metro station, time view outside, church sacred family, las vegas describe, walk entire church, gaudis original beautiful, gorgeous gaudis original, ticket ahead time, easy find metro, book ticket line, order ticket online, la caixa bank, book ticket advance, city tour bus, buy ticket line, local metro station, need book advance, get metro station, guide audio guide, attraction worth visit
6	Experiencia en la visita a la torre	nativity tower, audio guide, passion tower, spiral staircase, buy ticker, view city, ticket online, great view, stain glass, audio tour, visit tower, tower view, book ticket, tower elevator, afraid height, trip tower, tower tour, worth visit, go tower, glass window, lift tower, elevator walk, book online, narrow spiral, tower walk, pay extra, elevator tower, ticket advance, nativity facade, tower lift	stain glass window, buy ticket online, narrow spiral staircase, go nativity tower, book ticket online, go passion tower, walk spiral staircase, great view city, ticket ahead time, pay extra tower, buy ticket advance, book ticket advance, amazing view city, walk narrow spiral, tower great view, tower elevator walk, purchase ticket online, choose nativity tower, ticket online avoid, narrow spiral stair, tower amazing view, recommend go tower, passion facade tower, tower view city, beautiful stain glass, nativity tower elevator, audio guide tower, tower passion tower, passion tower lift, narrow wind staircase
7	Servicios de la basílica	guide tour, audio guide, tour guide, audio tour, highly recommend, sip line, buy ticket, worth visit, ticket online, book tour, hop hop, book guide, julia travel, stain glass, hop bus, line tour, take guide, tour worth, book ticket, self guide, book online, english speak, long line, tour book, guide audio, recommend guide, tour tour, tour bus, awe inspiring, bus tour	book guide tour, hop hop bus, take guide tour, skip line tour, recommend guide tour, tour julia travel, stain glass window, guide tour worth, guide audio tour, buy ticket online, english speak guide, audio guide tour, tour audio guide, guide tour book, recommend audio tour, self guide tour, tour highly recommend, book ticket online, self guide audio, guide tour julia, guide tour guide, guide tour audio, buy ticket advance, tour tour guide, recommend audio guide, get audio guide, english speak tour, audio guide informative, book skip line, skip line ticket
8	Impresiones del monumento y su arquitecto	awe inspiring, word describe, gaudi masterpiece, antoni gaudi, work art, gaudi genius, gaudi work, work progress, wonder world, buy ticket, ticket online, audio guide, guide tour, stain glass, 100 year, gaudi vision, genius gaudi, breath away, antonio gaudi, gaudi design, audio tour, worth visit, awe inspiring, complete 2026, architectural wonder, year ago, architectural masterpiece, ahead time, amazing building	take breath away, buy ticket online, stain glass window, architect antoni gaudi, anniversary gaudi death, buy ticket advance, world heritage site, unesco world heritage, book ticket online, construction 100 year, truly awe inspiring, world describe beauty, world describe amazing, 100 year ago, 100th anniversary gaudi, 2026 100th anniversary, architectural wonder world, bring tear eye, avoid long line, ticket ahead time, 2026 100 year, avoid long queue, ticket online avoid, architect antoni gaudi, 100 year gaudi, year gaudi death, way ahead time, antoni gaudi genius, wonder modern world
9	Vidrieras	stain glass, glass window, boy ticket, awe inspiring, ticket online, audio guide, sunny day, light come, worth visit, light stain, ticket advance, sun shine, book ticket, beautiful church, stained glass, guide tour, come stain, audio tour, beautiful stain, breath away, inside church, book online, natural light, late afternoon, year ago, inside outside, shine stain, light inside, outside inside, highly recommend	stain glass window, light stain glass, come stain glass, beautiful stain glass, shine stain glass, buy ticket online, light come stain, colour stain glass, glass window beautiful, take breath away, buy ticket advance, inside stain glass, colour stain glass, streaming stain glass, sun shine stain, glass window amazing, book ticket online, light streaming stain, glass window inside, book ticket advance, buy ticket line, sunny day light, amazing stain glass, glass window light, light shine stain, stream stain glass. architecture stain glass, ticket ahead time, stunning stain glass, stain glass amazing
10	Construcción	worth visit, year ago, place visit, amazing place, amazing architecture, work progress, 100 year, beautiful church, work art, piece architecture, 10 year, buy ticket, church see, amazing church, audio guide, amazing building, visit church, stain glass, visit place, piece art, time visit, ticket online, guide tour, gaudi work, truly amazing, visit visit, church visit, visit time, beautiful place, visit amazing	stain glass window, amazing piece architecture, buy ticket online, construction 100 year, visit year ago, 10 year ago, definitely worth visit, amazing place visit, amazing place visit,100 year ago, great place visit, visit 10 year, 15 year ago, 20 year ago, beautiful church see, amazing work art, book ticket online, buy ticket advance, build 100 year, good place visit, ticker ahead time, amazing church see, church sacred family, book ticket advance, beautiful church world, finish 10 year, ticket online avoid, hop hop bus, beautiful building see, piece architecture see, beautiful place visit

Tabla A.1: Bigramas y trigramas más relevantes del sector turismo

Apéndice B

Análisis sector comercio electrónico

Tópico	Categoría	Bigramas	Trigramas
1	Desbloqueo y almacenamiento	battery life, sd card, fingerprint reader, meet expectation, finger print, 64 gb, good price, good quality, good device, headphone jack, exceed expectation, pixel xl, print reader, 128 gb, micro sd, good battery,, gb ram, good smartphone, sim card, good cell, 32 gb, excellent price, card slot, dual sim, memory card, fingerprint sensor, face recognition, 16 gb, fingerprint scanner, internal storage	finger print reader, sd card slot, micro sd card, good battery life, battery life good, gb sd card, great battery life, finger print sensor, 64 gb sd, long battery life, gb micro sd, google pixel xl, gb internal storage, gb internal memory, finger print scanner, fingerprint reader work, good fingerprint reader, use fingerprint reader, reader face recognition, app sd card, battery life great, 128 gb storage, facial recognition work, product meet expectation, micro sd slot, 16 gb internal, fingerprint reader face, sd card storage, decent battery life, external sd card
2	Cámara y batería	battery life, great price, good price, great camera, good quality, great great, work great, good battery, great battery, good camera, easy use, samsung galaxy, camera good, highly recommend, far good, camera great, life good, good value, screen protector, big screen, battery last, picture quality, great good, pretty good, love camera, great quality, life great, super fast, well expect, camera amazing	great great price, good battery life, great battery life, battery life good, battery life great, long battery life, battery last day, battery life amazing, good value money, battery life well, good good price, love battery life, take great picture, amazing battery life, love big screen, camera battery life, samsung galaxy note, battery last long, galaxy s7 edge, great good price, long last battery, battery life long, mate 10 pro, excellent battery life, samsung galaxy s7, battery life excellent, love great price, battery life camera, case screen protector, battery life day
3	Usabilidad	easy use, battery life, work great windows phone, text message, lumia 920, sd card, touch screen, sim card, easy set, personal use, nokia lumia, work fine, user friendly, great easy, operating system, learn use, call text, year old, stop work, long time, download app, samsung galaxy, sound quality, listen music, home button, power button, battery last, bell whistle	great easy use, battery life good, great battery life, love easy use, long battery life, good battery life, good battery life, easy use work, battery life great, nokia lumia 920, battery last day, micro sd card, take great picture, sorry video unsupported, video unsupported browser, nice easy use, nokia lumia 928, send text message, work great easy, use social medium, need bell whistle, simple easy use, easy use great, watch youtube video, battery life day, samsung galaxy note, battery life excellent, nokia lumia 900, android operating system, battery life poor, battery last long
4	Tarjeta SIM	sim card, work verizon, work great, straight talk, customer service, work fine, dual sim, sd card, work perfectly, card work, brand new, compatible verizon, samsung galaxy, say unlock, work mobile, new sim, verizon store, international version, verizon network, waste time, wi fi, unlock work, factory unlock, work carrier, work at, verizon sim, come sim, battery life, waste money, able use	sim card work, come sim card, new sim card, verizon sim card, work straight talk, use sim card, read sim card, work verizon network, mobile sim card, insert sim card, sim card tray, use straight talk, work sim card, nano sim card, look brand new, sim card old, work metro pc, waste time money, sim card verizon, sin card activate, micro sd card, sim card slot, switch sim card, wi fi calling, samsung galaxy s7, at sim card, sim card use, report lose steal, micro sim card, say sim card

Tópico	Categoría	Bigramas	Trigramas
5	Opinión general	work great, brand new, work perfectly, good product, battery life, great product, work fine, great condition, stop work, perfect condition, look brand, good condition, look new, love new, happy purchase, great price, work good, work perfectly, great buy, fast shipping, far good, condition work, new work, great love, excellent condition, sim card, arrive time, exactly describe, hold charge, buy new	look brand new, brand new work, love work great, stop work month, work great far, new work great, battery hold charge, work brand new, work great problem, condition work great, great battery life, great condition work, work great love, come great condition, come perfect condition, great great price, come brand new, work perfectly fine, battery life great, great value money, condition work perfectly, look work new, work great issue, brand new scratch, screen stop work, time work great, excellent condition work, love love love

Tabla B.1: Bigramas y trigramas más relevantes del sector comercio electrónico

Apéndice C

Análisis sector restauración

Tópico	Categoría	Bigramas	Trigramas
1	Opciones de menú	chicken rice, nasi lemak, rice ball, fry rice, food good, banana leaf, indian food, highly recommend, taste good, price reasonably, fry chicken, fish head, good food, service good, kut teh, chicken curry, roti canai, tandori chicken, bak kut, butter chicken, bean sprout, staff friendly, beef rendang, cheese naan, good place, reasonably price, curry fish, asam pedas, food delicious, good service	chicken rice ball, bak kut teh, curry fish head, banana leaf rice, fish head curry, char kway teow, char kuey teow, ayam buah keluak, north indian food, soft shell crab, char koay teow, bean sprout chicken, good indian food, bah kut-teh, nasi lemak nasi, order nasi lemak, sweet sour chicken, food reasonable price, banana leaf meal, try nasi lemak, chicken bean sprout, fresh fruit juice, south indian food, good chicken rice, good nasi lemak, food taste good, veg non veg, good value money, ice lemon tea, try chicken rice
2	Comida italiana	ice cream, main course, highly recommend, mushroom soup, italian food, aglio olio, italian restaurant, food good, service good, lamb shank, foie gras, staff friendly, good service, definitely come, good food, squid ink, lava cake, good pizza, service excellent, food delicious, taste good, chocolate cake, friendly staff, great service, thin crust, food great, melt mouth pizza good, great food, definitely recommend	vanilla ice cream, chocolate lava cake, squid ink pasta, thin crust pizza, main course dessert, french onion soup, french onion soup, coconut ice cream, starter main course, cake ice cream, italian restaurant kl, good italian restaurant, seafood aglio olio, scoop ice cream, ice cream dessert, staff friendly attentive, middle eastern food, spaghetti aglio olio, chocolate ice cream, good italian food, authentic italian food, staff friendly helpful, ice cream delicious, ice cream good, good value money, soft shell crab, wood fire oven squid ink spaghetti, service excellent food, highly recommend restaurant, tea ice cream
3	Opinión del servicio	food good, good service, good food, staff friendly, service good, friendly staff, highly recommend, food nice, food great, food delicious, great food, definitely come, nice food, good place, great service, excellent service, nice place, price reasonable, reasonable price, food service, delicious food, nice environment, friendly helpful, food excellent, value money, service staff, service great, good environment, quality food, food service	food good service, good food good, staff friendly helpful, service good food., food great service, food reasonable price, food excellent service, good value money, friendly staff good, nice food nice, food friendly staff, service friendly staff, good staff friendly, good service nice, food taste good, great food great, good food nice, staff friendly food, good service food, nice food good, food good price, good service friendly, good service staff, food service good, highly recommend place, food nice environment, staff friendly attentive, food delicious service, food good staff

Tópico	Categoría	Bigramas	Trigramas
4	Opinión del lugar	great place, good place, food good, nice place, staff friendly good food, live music, happy hour, great food, food drink, food great, place chill, good service, service good, live band, great service, jonker street, place hang, great atmosphere, highly recommend, cold beer, reasonably price, good selection, drink good, service great, good music, food nice, hard rock, place good	nice place chill, good place chill, hard rock cafe, staff friendly helpful, good food great, great food great, great place chill, good place hang, place hang fried, great place hang, friendly staff good, friendly staff great, nice place hang, food great service, service good food, good service good, food good service, great place relax, staff friendly attentive, friendly staff nice, food friendly staff, changkat bukit bintang, good staff friendly, good food drink, food reasonable price, place chill friend, food drink good, good food great, walk jonker street, great place drink
5	Oferta para todos los gustos	food good, staff friendly, dining experience, main course, fine dining, highly recommend, good food, food great, food excellent, wine list, food delicious, food service, friendly staff, staff attentive, good service, great food, gluten free, service good, book table, friendly helpful, excellent service, bla bla, dining dark, ice cream, great service, service staff, service excellent, kuala lumpur, staff helpful, great experience	bla bla bla, staff friendly helpful, food good service, fine dining experience, fine dinning restaurant, casa del mar, food excellent service, food great service, highly recommend restaurant, staff friendly attentive, starter main course, friendly helpful staff, good food good, staff attentive friendly, food excellent staff, book table advance, pre dinner drink, bottle red wine, wine list extensive, main course dessert, staff extremely friendly, good wine list, staff helpful attentive, gluten free option, candle light dinner, service excellent waiter, experience dining dark, wine list good, soft shell crab, good dinning experience
6	Esperas en la visita a la torre	food good, food serve, good food, order food, dim sum, quality food, food average, service good, good service, food ok, long time, food service, food come, service slow, food quality, wait long, take order, food arrive, customer service, food great, bad service, staff friendly, 10 minute, main course, 20 minute, food taste, service food, 15 minute, food order, wait staff	food good service, take long time, wait long time, wait half hour, 10 service change, wait 20 minute, wait 15 minute, hard rock cafe, wait 30 minute, wait 10 minute, wait 45 minute, wait hour food, chinese new year, food taste good, food great service, fine dinning restaurant, read good review, quality food service, service food good, food good price, chicken rice ball, wait 30 minute, wait long food, food ok service, waste time money, food take long, food bad service, good value money, soft shell crab, staff friendly helpful
7	Opinión general	highly recommend, great service food great, great food, excellent service, staff friendly, food amazing, food service, service great, good service, food delicious,definitely come, friendly staff, dining experience food good, good food, food excellent, amazing food, great experience, delicious food, service excellent, definitely recommend, great place, special thank, amazing service, service staff, great atmosphere, recommend place, service food service good	food great service, great food great, food excellent service, highly recommend place, food good service, great service great, food amazing service, great dining experience, highly recommend visit, highly recommend restaurant, great food service, food delicious service, definitely worth visit, staff super friendly, food friendly staff, service highly recommend, look forward visit, staff friendly helpful, good food great, food great atmosphere, food amazing staff, food service excellent, food great staff, make feel home, staff friendly attentive, friendly staff great, take good care, great service food, great food drink
8	Comida india	food good, kuala lumpur, good food, highly recommend, indian food, service good, good service, quality food, worth visit, food service, great food, value money, food great, good place, local food, food excellent, food delicious, good restaurant, good value, service excellent, reasonable price, staff friendly, visit restaurant, excellent service, price reasonable, reasonably price, western food, good quality, fine dining, malaysian food	food good service, good value money, restaurant kuala lumpur, food reasonable price, good indian food, definitely worth visit, south indian food, service good food, food good price, good food good, food excellent service, visit kuala lumpur, food great service, good quality food, north indian food, good place eat, highly recommend place, fine dining restaurant, highly recommend restaurant, great food great, quality food service, service excellent food, good service good, food reasonably price, great value money, food kuala lumpur, review trip advisor, food service good, service great food, kuala lumpur food

Tópico	Categoría	Bigramas	Trigramas
9	Precios y otros servicios	food good, hard rock, price reasonable, good food, service good, reasonable price, dim sum, food average, good service, food ok, parking space, air condition, food service, staff friendly, rock cafe, shopping mall, good place, peak hour, reasonably price, car park, restaurant locate, food quality, easy find, taste good, food nice, food great, open air, service fast, food serve, good price	hard rock cafe, food reasonable price, food good service, food good price, weekend public holiday, service good food, ample parking space food taste good, food reasonably price, good food good, dim sum restaurant, food average price, good price reasonable, opens 24 hour, smoking non smoking, good service good, hard find parking, visit hard rock, food average service, hard rock cafe, price slightly high, locate ground floor, food pretty good, easy find parking, food service good, food nice price, non smoking area, great food service, food great service, staff friendly helpful
10	Carnes	medium rare, pork rib, chicken chop, service good, good steak, rib eye, cook perfection, good service, fish chip, steak cook, food good, highly recommend, mash potato, staff friendly, good place, good food, beet burger, lamb chop, chicken wing, tender juicy, steak good, pork chop, grill chicken, definitely come, taste good, black pepper, pork belly, pork knuckle, mashed potato, melt mouth	rib eye steak, black pepper sauce, steak cook perfection, cook medium rare, order medium rare, steak medium rare, pork belly burger, steak cook perfectly, grill chicken chop, good value money, order rib eye, good service good, hard rock cafe, medium rare steak, sweet potato fry, meat fall bone, staff friendly helpful, steak perfectly cook, tender fall bone, order beef burger, soft shell crab, pork shoulder steak, bbq pork rib, meat tender juicy, good steak kl, truffle mash potato order fish chip, service good food, medium medium rare, food service good
11	Tipos de celebración	good service, excellent service, food good, great service, birthday celebration, good food, service good, chef steven, definitely come, food great, food service, celebrate birthday, birthday song, adli azu, chef nathan, afternoon tea, service excellent, birthday dinner, staff friendly, dim sum, great food, food excellent, food delicious, highly recommend, special thank, lunch dinner, good place, service food, paya serai, good work	sing birthday song, food good service, food great service, chef steven nathan, steven nathan haziq, chef nathan chef, food excellent service, celebrate friend birthday, adli azu chef, chef steven haziq, breakfast lunch dinner, chef nathan steven, adi azu edward, azu chef steven, service good food, celebrate wife birthday, service adli azu, great food great, chef steven chef, nathan steven haziq, good food great, nathan chef haziq, good service good, good food good, thank adli azu, chinese new year, steven haziq ehsan, edward chef steven, chef nathan haziq, staff friendly helpful
12	Comida china	dim sum, char siew, fry rice, highly recommend, service good, salt egg, roast pork, taste good, food good, roast duck, spring roll, chicken rice, stir fry, sweet sour, deep fry, price reasonable, good food, pork belly, tom yam, chinese restaurant, pork rib, chinese food, good service, xiao long, signature dish, cheong fun, roasted pork, bean sprout, staff friendly, long bao	xiao long bao, chee cheong fun, dim sum restaurant, tom yam soup, soft shell crab, good dim sum, halal dim sum, dim sum place, sweet sour pork, dim sum good, salt egg yolk, chinese new year, variety dim sum, dim sum dish, tom yum soup, pork char siew, yong tau foo, sweet sour chicken, bak kut teh, chicken rice ball, dim sum serve, mango sticky rice, hot sour soup, dim sum lunch, char siew roast, pineapple fry rice, double roasted pork, char siew pau, dim sum taste, service good food
13	Comida japonesa	dining experience, fine dining, japanese restaurant, highly recommend, japanese food, soft shell, food good, shell crab, main course, la carta, set menu, service good, ice cream, fresh seafood, sashimi fresh, chef yau, good food, melt mouth, staff friendly, kuala lumpur, le petit, sushi sashimi, food delicious, good japanese, food service, good service, food serve, food fresh, fresh sashimi, grand yatt	soft shell crab, le petit chef, casa del mar, fine dining experience, fine dining restaurant, la carta menu, good japanese restaurant, ala carte menu, unique dining experience, good japanese food, stay casa del, food service good, garlic fry rice, dining experience food, staff friendly helpful, seven terraces hotel, course set menu, wonderful dining experience, green tea ice, tea ice cream, good dining experience, great dining experience, experience dining dark, michelin star restaurant, shangri la hotel, salmon belly sashimi, grand hyatt kl, food great service, good fine dining, black cod miso

Tabla C.1: Bigramas y trigramas más relevantes del sector restauración

Referencias

Adiguzel, F., Elsherbiny, M., Quintana, J. T. A., y González-Martel, C. (2021). Topic modelling application for luxury hotel reviews.

Albalawi, R., Yeap, T. H., y Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, *3*, 42.

Anderson, E. W., Fornell, C., y Lehmann, D. R. (1994). Customer satisfaction, market share, and profitability: Findings from sweden. *Journal of marketing*, *58*(3), 53–66.

Bachtiar, F. A., Paulina, W., y Rusydi, A. N. (2020). Text mining for aspect based sentiment analysis on customer review: A case study in the hotel industry. En *Iicst* (pp. 105–112).

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

Bonta, V., y Janardhan, N. K. N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, *8*(S2), 1–6.

Boo, S., y Busser, J. A. (2018). Meeting planners' online reviews of destination hotels: A twofold content analysis approach. *Tourism Management*, *66*, 287–301.

Devika, M., Sunitha, C., y Ganesh, A. (2016). Sentiment analysis: a comparative study on different approaches. *Procedia Computer Science*, *87*, 44–49.

Egger, R., y Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, *7*.

Gräbner, D., Zanker, M., Fliedl, G., Fuchs, M., y cols. (2012). Classification of customer reviews based on sentiment analysis. En *Enter* (pp. 460–470).

Guan, Z., Chen, L., Zhao, W., Zheng, Y., Tan, S., y Cai, D. (2016). Weakly-supervised deep learning for customer review sentiment classification. En *Ijcai* (pp. 3719–3725).

Guo, Y., Barnes, S. J., y Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism management*, *59*, 467–483.

Hossain, E., Sharif, O., Hoque, M. M., y Sarker, I. H. (2020). Sentilstm: a deep learning approach for sentiment analysis of restaurant reviews. En *International conference on hybrid intelligent systems* (pp. 193–203).

Hou, T., Yannou, B., Leroy, Y., y Poirson, E. (2019). Mining customer product reviews for

product development: A summarization process. *Expert Systems with Applications*, *132*, 141–150.

Hu, N., Zhang, T., Gao, B., y Bose, I. (2019). What do hotel customers complain about? text analysis using structural topic model. *Tourism Management*, *72*, 417–426.

Huang, Y., Wang, R., Huang, B., Wei, B., Zheng, S. L., y Chen, M. (2021). Sentiment classification of crowdsourcing participants' reviews text based on lda topic model. *IEEE Access*, *9*, 108131–108143.

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).

Kim, W. G., Ng, C. Y. N., y Kim, Y.-s. (2009). Influence of institutional dineserv on customer satisfaction, return intention, and word-of-mouth. *international journal of hospitality management*, *28*(1), 10–17.

Kumar, K. S., Desai, J., y Majumdar, J. (2016). Opinion mining and sentiment analysis on online customer review. En *2016 ieee international conference on computational intelligence and computing research (iccic)* (pp. 1–4).

Laksono, R. A., Sungkono, K. R., Sarno, R., y Wahyuni, C. S. (2019). Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes. En *2019 12th international conference on information & communication technology and system (icts)* (pp. 49–54).

Levy, S. E., Duan, W., y Boo, S. (2013). An analysis of one-star online reviews and responses in the washington, dc, lodging market. *Cornell Hospitality Quarterly*, *54*(1), 49–63.

Mirai, I., Kannan, P., y Nobuhiko, T. (2020). Customer review analysis using word embedding model considering text topics.

Moghaddam, S., y Ester, M. (2012). Aspect-based opinion mining from product reviews. En *Proceedings of the 35th international acm sigir conference on research and development in information retrieval* (pp. 1184–1184).

Ng, C. K. (2022, Jul). *Malaysia restaurant review datasets.* Descargado de `https://www.kaggle.com/datasets/choonkhonng/malaysia-restaurant-review-datasets`

Nibras, G. (2019, Dec). *Amazon cell phones reviews.* Descargado de `https://www.kaggle.com/datasets/grikomsn/amazon-cell-phones-reviews?select=20191226-reviews.csv`

Observatori del Turisme a Barcelona: ciutat i regió. (2021). Informe de la actividad turística.. Descargado de `https://www.observatoriturisme.barcelona/en`

Park, E., Chae, B., Kwon, J., y Kim, W.-H. (2020). The effects of green restaurant attributes on customer satisfaction using the structural topic model on online customer reviews. *Sustainability*, *12*(7), 2843.

Park, S., Cho, J., Park, K., y Shin, H. (2021). Customer sentiment analysis with more sensi-

bility. *Engineering Applications of Artificial Intelligence*, *104*, 104356.

Prananda, A. R., y Thalib, I. (2020). Sentiment analysis for customer review: Case study of go-jek expansion. *Journal of Information Systems Engineering and Business Intelligence*, *6*(1), 1–8.

Putranto, Y., Sartono, B., y Djuraidah, A. (2021). Topic modelling and hotel rating prediction based on customer review in indonesia. *International Journal of Management and Decision Making*, *20*(3), 282–307.

Rajeswari, A., Mahalakshmi, M., Nithyashree, R., y Nalini, G. (2020). Sentiment analysis for predicting customer reviews using a hybrid approach. En *2020 advanced computing and communication technologies for high performance applications (accthpa)* (pp. 200–205).

Riaz, S., Fatima, M., Kamran, M., y Nisar, M. W. (2019). Opinion mining on large scale data using sentiment analysis and k-means clustering. *Cluster Computing*, *22*(3), 7149–7164.

Sari, P. K., Alamsyah, A., y Wibowo, S. (2018). Measuring e-commerce service quality from online customer review using sentiment analysis. En *Journal of physics: Conference series* (Vol. 971, p. 012053).

Sharif, O., Hoque, M. M., y Hossain, E. (2019). Sentiment analysis of bengali texts on online restaurant reviews using multinomial naïve bayes. En *2019 1st international conference on advances in science, engineering and robotics technology (icasert)* (pp. 1–6).

Sievert, C., y Shirley, K. (2014, junio). LDAvis: A method for visualizing and interpreting topics. En *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70). Baltimore, Maryland, USA: Association for Computational Linguistics. Descargado de `https://aclanthology.org/W14-3110` doi: 10.3115/v1/W14-3110

Smith, A., y Anderson, M. (2016). Online shopping and e-commerce.

Soni, S., y Sharaff, A. (2015). Sentiment analysis of customer reviews based on hidden markov model. En *Proceedings of the 2015 international conference on advanced research in computer science engineering & technology (icarcset 2015)* (pp. 1–5).

Sun, Q., Niu, J., Yao, Z., y Yan, H. (2019). Exploring ewom in online customer reviews: Sentiment analysis at a fine-grained level. *Engineering Applications of Artificial Intelligence*, *81*, 68–78.

Tabasco, A. (2023). *Tfg-customer-satisfaction.* `https://github.com/atabasco/TFG-Customer-Satisfaction`. GitHub.

Tolegen, G., Toleu, A., Mussabayev, R., y Krassovitskiy, A. (2022). A clustering-based approach for topic modeling via word network analysis. En *2022 7th international conference on computer science and engineering (ubmk)* (pp. 192–197).

Valdivia, A., Martínez-Cámara, E., Chaturvedi, I., Luzon, M. V., Cambria, E., Ong, Y., y

Herrera, F. (2018, 12). What do people think about this monument? understanding negative reviews via deep learning, clustering and descriptive rules. *Journal of Ambient Intelligence and Humanized Computing*, 39-52. doi: 10.1007/s12652-018-1150-3

Vanaja, S., y Belwal, M. (2018). Aspect-level sentiment analysis on e-commerce data. En *2018 international conference on inventive research in computing applications (icirca)* (pp. 1275–1279).

Yang, Q. (2020). Data mining of new snack e-commerce reviews based on text sentiment analysis and latent dirichlet allocation topic model. En *The 3rd international conference on economy, management and entrepreneurship (icoeme 2020)* (pp. 372–378).