Universidad de Granada

Tesis Doctoral

Social Media Analytics para Smart-Tourism

Autor:
Marlon Santiago Viñán Ludeña

Director:
Dr. Luis M. de Campos Ibáñez

Programa de Doctorado en Tecnologías de la Información y la Comunicación

Departamento de Ciencias de la Computación e Inteligencia Artificial
Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

16 de septiembre de 2022

Editor: Universidad de Granada. Tesis Doctorales
Autor: Marlon Santiago Viñán Ludeña
ISBN: 978-84-1117-642-2
URI: https://hdl.handle.net/10481/79630

Declaración de Autoría

El doctorando / *The doctoral candidate*, **Marlon Santiago Viñán Ludeña**, y el director de la tesis / and the thesis supervisor: **Luis M. de Campos Ibáñez**.

Garantizamos, al firmar esta Tesis Doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección del director de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Guarantee, by signing this Thesis, that the work has been done by the doctoral candidate under the direction of the thesis supervisor and, as far as our knowledge reaches, in the performance of the work, the rights of other authors to be cited (when their results or publications have been used) have been respected.

Firma/Signed:

Director de la Tesis/ Thesis supervisor

Doctorando/Doctoral candidate

Granada, septiembre, 2022

Dedicatoria

Este camino ha estado lleno de alegrías y decepciones, que en ciertas ocasiones me llevaron a pensar a dejarlo de lado pero la constancia y perseverancia me ha servido para culminar con éxito esta tesis y finalmente conseguir mi sueño. Esta sólida formación académica la he conseguido gracias a mi director y tutor de tesis; gracias Luis por tu orientación y consejos que han sido necesarios para comenzar este fascinante camino en la academia e investigación.

Como no dedicar todo este esfuerzo a mi familia, a mis padres especialmente a mi madre, siempre ha sido mi soporte y lo mejor que la vida me ha regalado ya que sin su apoyo no hubiera llegado hasta aquí; gracias a mis hermanos que siempre me han ayudado en todo, especialmente a Mireya y Diego.

Marlon Santiago Viñán Ludeña

Índice general

Índice de figuras

Índice de cuadros

Resumen

Esta tesis tiene el propósito de brindar un análisis exhaustivo de datos turísticos provenientes de redes sociales, desde el campo de la ciencia de datos y la inteligencia artificial, con el fin de entender las percepciones de los usuarios al visitar un destino turístico para mejorar la administración de estos lugares. El análisis de datos de redes sociales se basa en tres pilares fundamentales: las plataformas; como Twitter, Instagram o TripAdvisor, los usuarios y la tecnología que transforma posts, imágenes o videos en datos, esta información tiene gran valor para especialistas en marketing, para la evaluación de la confianza de marca, satisfacción, etc. Por tanto, este trabajo tiene aportaciones importantes en el ámbito de Smart-Tourism.

Existen muchos términos que hacen referencia a «Smart» o «Inteligente», sin embargo; Smart-Tourism se refiere a la interacción y/o combinación de redes de comunicaciones, internet, sensores, internet de las cosas y el turismo; por tanto, el aporte realizado en esta tesis en el ámbito de Smart-Tourism es significativo, debido a que los datos que se analizaron y que en algunos casos sirvieron para entrenar un algoritmo, provienen de redes sociales que los usuarios usan mientras están visitando algún lugar y que para ello, es necesario las redes de comunicaciones, su dispositivo móvil e internet. Además, todas las herramientas que se han utilizado van en concordancia con la ciencia de datos utilizando sobre todo la minería de texto, el aprendizaje de máquina y el aprendizaje profundo.

En esta investigación se propone un framework para el análisis de datos de redes sociales, el mismo que puede ser utilizado en cualquier red social, para analizar y evaluar de forma general cuáles son los lugares más visitados de un destino turístico. Este enfoque puede ser usado para implementar una aplicación que pueda ayudar a un turista a saber los lugares más visitados de una forma resumida sin tener que ir a las redes sociales donde la información puede resultar abrumadora. Otro de los aportes de este trabajo, consiste en proponer una arquitectura basada en tecnologías de aprendizaje profundo denominada BERT (Bidirectional encoders representation from transformers) para mejorar la clasificación de sentimientos usando datos en español. Esto permite evaluar

la satisfacción de los visitantes. Además, la integración de esta tecnología con información del transporte, eventos especiales o condiciones climáticas puede ayudar a la construcción de aplicaciones que mejoren la calidad de las visitas de un turista. Por último, se propone un algoritmo basado en reglas para la detección de aspectos en textos o revisiones de redes sociales con el objetivo de evaluar la (in)satisfacción de los turistas de una determinada entidad (lugar, evento, etc.). Este enfoque permite saber de forma detallada qué es lo que los turistas piensan sobre una entidad. Creemos que toda esta tecnología desarrollada puede ayudar a los gestores turísticos a administrar de mejor manera sus negocios, entender las percepciones, conocer la (in)satisfacción, proponer planes de mejoras y reevaluar constantemente los servicios turísticos y estrategias implementadas.

Capítulo 1

Introducción

1.1. Motivación

La tecnología está presente prácticamente en todos dominios de investigación y en la vida cotidiana de gran parte de la población, que la utiliza a través de sus computadoras personales, teléfonos inteligentes, tabletas, etc. Cada usuario genera mucha información que puede ser analizada con el objetivo de mejorar su interacción con los servicios que utiliza. El sector turístico es uno de los sectores más importantes en todo el mundo, por la gran cantidad de recursos económicos que genera. Por tanto, es conveniente vincular la tecnología con la industria turística a través de los datos que generan tanto viajeros como las empresas del sector, con la finalidad de encontrar patrones que ayuden a la toma de decisiones por parte de los administradores y permitan mejorar la experiencia turística de un viajero que visita un determinado destino. A continuación se presenta un breve resumen del turismo, smart-tourism y el análisis de datos.

1.1.1. Turismo y su impacto

De acuerdo a la Organización Mundial del Turismo (OMT), el turismo se define como «un fenómeno social, cultural y económico que supone el desplazamiento de personas a países o lugares fuera de su entorno habitual por motivos personales, profesionales o de negocios. Esas personas se denominan viajeros (que pueden ser o bien turistas o excursionistas; residentes o no residentes) y el turismo abarca sus actividades, algunas de las cuales suponen un gasto turístico» (OMT, 2008).

Existen muchos tipos de turismo, en este estudio se destacan dos de ellos. (i) El **turismo doméstico** que se refiere al turismo interno, es decir, un viajero

disfruta de los atractivos turísticos y servicios sin salir de su país de origen. (ii) **Turismo internacional**[1], que consiste en las actividades realizadas fuera del país de residencia de los viajeros.

El turismo es una de las áreas que tiene mucho impacto debido a que dinamiza la economía tanto en países desarrollados como en los sub-desarrollados. Debido a la pandemia del COVID-19 el sector turístico ha sido uno de los más afectados a nivel mundial. Países asiáticos como Indonesia o China han disminuido sus ingresos por las restricciones de viaje impuestas para contener la pandemia. En el sur de Europa, que en verano tiene bastante afluencia de turistas, ha provocado que algunos negocios cierren, esto tiene un impacto en la sociedad, principalmente si una ciudad o región depende casi exclusivamente de los ingresos provenientes de la industria turística.

El turismo en el mejor de los casos trae inversión extranjera y desarrollo social debido a que se incrementan los puestos de trabajo y se mejora la calidad de vida de los habitantes de un determinado lugar. Para algunos investigadores este desarrollo se da en dos sentidos, (Inversini y Rega, 2020) debatiendo por un lado la construcción de infraestructura y servicios de un destino turístico y por el otro lado el desarrollo social de las comunidades que son parte de las iniciativas turísticas.

En países en desarrollo quienes lideran el desarrollo turístico son las empresas privadas, sin embargo existen instituciones gubernamentales conjuntamente con organizaciones sin fines de lucro no gubernamentales que permiten el desarrollo de comunidades rurales especialmente.

El turismo doméstico es uno de los que se ha impulsado y ha tenido un crecimiento importante en todas las regiones luego de que la mayoría de los países hayan impulsado un programa de vacunación a la población. A pesar de que las regiones de Asia y el Pacífico y Europa son las más afectadas por la pandemia, poco a poco se han ido reactivando con la finalidad de llegar a los niveles alcanzados a finales del 2019.

De acuerdo con la OMT[2] el turismo ha sido incluido en la Agenda 2030 para el desarrollo sostenible con la finalidad de acabar con la pobreza extrema, combatir la desigualdad y la injusticia y solucionar el cambio climático, siendo la industria turística un sector clave en algunos de esos objetivos. A continuación se mencionan los objetivos que están directamente relacionados con el turismo:

[1] En este texto se hará referencia al turismo en general

[2] https://www.unwto.org/es/turismo-agenda-2030

- *Trabajo decente y crecimiento económico*: De acuerdo a esta organización la industria turística es responsable de la creación de un 10 % aproximadamente de los puestos de trabajo.
- *Producción y consumo responsable*: Pese a que es muy importante en la economía de los países, es necesario que los administradores turísticos apliquen prácticas sostenibles y promuevan el consumo de los productos locales. Es necesario que los hoteles tengan un manejo responsable de residuos que no dañen el medio ambiente, pues en muchos de los casos son ecosistemas vulnerables.
- *Vida submarina*: Muchos de los destinos turísticos más importantes son pequeños estados insulares como es el caso de la Isla de Madeira en Portugal o las Islas Galápagos en Ecuador. Estos destinos son susceptibles al daño que puede producir la llegada excesiva de turistas. Por tanto, es importante que la industria desarrolle y establezca políticas de protección a los ecosistemas marinos.

Un turismo responsable y amigable con el ambiente es importante para que se pueda reactivar la economía. La importancia del turismo es vital para muchos países, esta importancia ha motivado la realización de este estudio, vinculando la tecnología con el sector del turismo. La forma de realizarlo es la siguiente: Cuando una persona decide visitar algún destino turístico muy a menudo busca referencias de amigos y familiares, sin embargo no siempre se encuentra información relevante de las personas que nos rodean. Por tanto, él o ella decide buscar información en las páginas web oficiales del sitio, blogs, microblogs (escritos por personas que han visitado el lugar), redes sociales que muestren las percepciones de los usuarios o sitios especializados de revisiones de los diferentes servicios turísticos de ese lugar o región. Esta información es muy importante, pero abrumadora y con mucho ruido; por tanto, es conveniente que esos datos sean analizados y procesados con la finalidad de tener información útil para los turistas y para los administradores, datos relevantes para la toma de decisiones.

1.1.2. Smart-Tourism

En las últimas décadas ha crecido significativamente la tecnología de la información y comunicación, la cual juega un papel primordial en el turismo (Inversini y Rega, 2020). El término «Smart Tourism» fue lanzado en China en el año 2013 como una política importante del turismo en China. Una forma de entender el término es la propuesta por (Li y col., 2017), donde la palabra *smart*

es sinónimo de sabiduría y la palabra Tourism o turismo se refiere a los viajes. Los autores proponen «smart tourism» o «turismo inteligente» en lugar de «turismo de sabiduría». Los términos «smart» e «intelligent» traducidos al español significan lo mismo, «inteligente». Sin embargo, la definición de «intelligent» en inglés significa que puede cambiar su estado o acción en respuesta a diferentes situaciones y tiene relación al ámbito de la tecnología, mientras que «smart» significa hacer lo correcto en situaciones complejas usando grandes cantidades de datos como entrada, poniendo énfasis en los resultados tecnológicos para las personas, por tanto en inglés estos dos términos son distintos. (Li y col., 2017)

Investigadores chinos han acuñado el término «Smart-Tourism» y la mayoría de las definiciones coincide en una combinación de redes de comunicaciones, internet, sensores, tecnología móvil, internet de las cosas y el turismo; a esto se añade tecnologías de posicionamiento y análisis de redes sociales.

En Europa, el término Smart-Tourism ha nacido con las iniciativas de ciudades inteligentes o «Smart-City», centrándose en la innovación, competitividad y en el desarrollo de aplicaciones inteligentes, con la finalidad de que los usuarios puedan compartir sus experiencias turísticas para posteriormente analizar los datos con nuevos algoritmos y presentar resultados útiles para los turistas (Gretzel y col., 2015).

Algunos autores clasifican a Smart-Tourism en tres niveles: (i) para los viajeros, brindar información turística y ajustar rápidamente los planes de viajes; (ii) para gerentes gubernamentales y de empresas turísticas, consiste en implementar un sistema completo que ofrezca precisión, conveniencia y la ubicuidad de las aplicaciones turísticas, a través de plataformas de servicios turísticos que ofrezcan a los visitantes servicios de hospedaje, comida, transporte, viajes, compras, etc.; por último, (iii) desde la parte técnica consiste en la interacción sistemática entre los recursos turísticos físicos (hoteles, restaurantes, lugares turísticos en general) y la información turística sobre esos recursos físicos (Wang y col., 2012).

Desde una perspectiva empresarial, «Smart-Tourism» se basa en una extensa estructura de información. Estos datos comúnmente son subidos a las redes sociales de forma activa o explícitamente por los usuarios o turistas, o implícitamente a través de sensores móviles o portátiles. Todos estos datos son compartidos voluntariamente por los usuarios. El poder económico y los beneficios que pueden tener todos los actores turísticos se deriva del control sobre las fuentes de la información (Bick y col., 2012). La clave está en la creación de valor a través del uso de datos/tecnología/infraestructura. (Gretzel y col., 2015).

Este estudio se basa en la creación de valor tanto para turistas como empresarios a través del uso gratuito de los datos que redes sociales nos proporcionan, que al ser analizados y procesados los resultados pueden convertirse en nuevas aplicaciones para los turistas y herramientas para la toma de decisiones para los empresarios.

1.1.3. Redes Sociales

Una red social es una estructura social compuesta por un conjunto de actores y uno o más lazos o relaciones definidos entre ellos. A principios de los años 2000, con los avances de la tecnología, este tipo de plataformas se volvieron muy populares entre la gente, siendo imprescindibles en cualquier dispositivo móvil.

En importante tener en cuenta que tanto «social media» como «social network», traducidos al español significan lo mismo. De acuerdo con el diccionario de Cambridge ambos términos: *social media* y *social network* se refieren a «sitios web o programas de computadora que permiten a las personas comunicarse y compartir información a través de la web o internet, usando una computadora o teléfono móvil». Otros autores mencionan que existe una diferencia entre estos dos términos, siendo las redes sociales la interacción entre las personas y el social media los datos que deja el usuario en estas plataformas. Sin embargo, en este estudio los tomaremos como sinónimos.

A pesar de que las redes sociales han sido estudiadas desde hace décadas por psicólogos antes de que aparecieran en la década de los 90 en forma de sitio web o aplicación, hoy en día se las conoce como un mundo virtual que puede funcionar en diferentes niveles de la vida cotidiana de las personas.

De acuerdo con Rdstation[3], el sitio web SixDegrees.com[4] se considera como la primera red social, permitiendo a los usuarios tener un perfil y contactar con otras personas. Luego surgieron sitios web como: Friendster, una Red social que permitía crear conexiones, mantener contactos y compartir contenido multimedia, fué cerrada en 2018 [5], MySpace[6], Orkut[7] o hi5[8]. Más adelante se consolidó

[3] https://www.rdstation.com/es/redes-sociales/
[4] http://www.sixdegrees.com
[5] https://en.wikipedia.org/wiki/Friendster
[6] https://myspace.com/
[7] http://www.orkut.com/index.html
[8] https://hi5.com/

el poderío de Facebook[9], Instagram[10](adquirida por Facebook en 2012), LinkedIn[11] y Twitter[12].

Las redes sociales, al ser tan variadas y estar presentes en cualquier nivel de la vida de una persona, se han convertido en parte de la rutina diaria, siendo beneficioso para las empresas de marketing, ya que pueden promocionar sus productos a una audiencia que ya está enganchada en este tipo de plataformas.

Las redes sociales se pueden dividir en 10 tipos de acuerdo con el portal *Psicología y mente*[13]:

- **Redes sociales horizontales**: Fueron creadas para el público en general sin distinción alguna, como por ejemplo, Facebook e Instagram[14].
- **Redes sociales verticales**: Dirigidas a un público especializado, por ejemplo TripAdvisor[15].
- **Redes sociales profesionales**: El objetivo de este tipo de redes es meramente profesional. La más popular entre ellas en LinkedIn
- **Redes sociales de ocio**: Son plataformas donde se comparten temas de música, ocio, videojuegos , etc.
- **Redes verticales mixtas**: Son plataformas que son una combinación de las dos últimas que se ha mencionado anteriomente, por ejemplo Unience combina temas profesionales con inversores.
- **Redes sociales universitarias**: Se enfocan en compartir experiencias de semestres anteriores con otros alumnos permitiendo entre otras cosas, descargar apuntes.
- **Noticias sociales**: En este caso las noticias son moderadas por los usuarios, es decir; las que tienen más votos son las que se publican, por ejemplo: Digg[16], Reddit[17] y Menéame[18].

[9] https://www.facebook.com
[10] https://www.instagram.com
[11] https://www.linkedin.com/
[12] http://www.twitter.com
[13] https://psicologiaymente.com/social/tipos-de-redes-sociales
[14] http://www.instagram.com
[15] http://www.tripadvisor.com
[16] https://digg.com/
[17] https://www.reddit.com/
[18] https://www.meneame.net/

- **Blogging**: En este tipo de plataformas se comparten historias, artículos, opiniones, enlaces o contenido multimedia, por ejemplo: Blogger[19], WordPress[20], o Medium[21].

- **Microblogging**: Parecido al anterior pero en este caso el texto y el contenido en general es muy corto. Las plataformas más populares son Twitter[22], Tumblr[23] y Sina Weibo[24].

- **Contenido compartido**: En esta clasificación sobresalen las plataformas de YouTube[25], Flickr[26] y Tiktok[27] .

En el ámbito del turismo existen muchas redes sociales especializadas que permiten a los turistas tener herramientas con las que puedan compartir sus experiencias en la web. Entre las más conocidas se tienen:

- **TripAdvisor**: Sitio web que tiene reseñas publicadas por usarios de sus experiencias en sus travesías turísticas.

- **Airbnb**[28]: Permite ofertar alojamiento a través de anfitriones locales.

- **Trivago**[29]: Esta plataforma permite comparar precios de hoteles alrededor del mundo.

- **Yelp**[30]: Es una plataforma que permite encontrar lugares de comida y bares, y valorar sus servicios a través de reseñas.

- **OpenTable**[31]: Plataforma que permite realizar reservaciones en restaurantes y revisar las reseñas de clientes.

[19] https://www.blogger.com/about/?bpli=1
[20] https://wordpress.com/
[21] https://medium.com
[22] http://www.twitter.com
[23] https://www.tumblr.com/
[24] https://weibo.com/
[25] https://www.youtube.com/
[26] https://www.flickr.com/
[27] https://www.tiktok.com/
[28] https://www.airbnb.com
[29] https://www.trivago.com
[30] https://www.yelp.com/
[31] https://www.opentable.com/

De acuerdo con DataReportal[32] las redes sociales que más se usan son: Facebook, con 2,936 millones de usuarios activos, seguido por YouTube con 2,562 millones de usuarios, WhatsApp con 2,000 millones de usuarios, Facebook Messenger con 1,300 millones de usuarios, Instagram con 1,221 millones de usuarios y también es importante mencionar a Twitter con 353 millones de usuarios.

El presente estudio usa datos de dos redes sociales, Twitter, Instagram y TripAdvisor. En estas plataformas, existe información importante sobre percepciones de destinos turísticos. Además el acceso a la información no es tan complejo y restrictivo como en otras redes sociales como Facebook.

1.1.4. Ciencia de datos

Los datos son muy importantes para la toma de decisiones de cualquier empresa, gobierno u organización. Según el diccionario de Cambridge[33] los datos se definen como información, especialmente hechos o números, recolectados con la finalidad de ser examinados y usados para la ayuda a la toma de decisiones. Por otro lado, se pueden considerar como un conjunto de valores de variables cualitativas o cuantitativas. Existen muchas fuentes de información como censos de población, registros médicos electrónicos, datos de sistemas de información geográfica, análisis y extrapolación de imágenes, tráfico de sitios web, datos personales, publicitarios, datos que circulan en redes sociales, etc.

Tener fuentes de datos implica de manera implícita que ellos pueden ayudarnos a entender cómo funciona el mundo (Blei y Smyth, 2017). Para que los datos se conviertan en información útil, es necesario que sean analizados, tomando en cuenta que la estadística es la disciplina más importante en cada etapa de la ciencia de datos (Weihs e Ickstadt, 2018). Existen seis categorías que los científicos de datos generalmente usan para responder a las preguntas de investigación[34].

- **Análisis descriptivo:** Es el primer análisis que se debe realizar en cualquier investigación, este consiste en describir o resumir los datos, es decir; obtener medidas de tendencia central (promedio, mediana, moda) o medidas de variabilidad (rango, desviación estándar o varianza).

- **Análisis exploratorio:** El principal objetivo de este análisis es examinar los datos y encontrar relaciones que anteriormente no fueron halladas, es

[32] Estadísticas acerca del uso de internet, disponible en: https://datareportal.com/
[33] https://dictionary.cambridge.org
[34] https://towardsdatascience.com/the-six-types-of-data-analysis-75517ba7ea61

decir; cómo las diferentes medidas pueden estar relacionadas entre ellas, pero sin verificar si dichas relaciones son causales. Este análisis permite formular hipótesis e impulsar el diseño de estudios futuros y mejorar la recopilación de datos.

- **Análisis inferencial:** El principal objetivo de este análisis es tomar una muestra relativamente pequeña de datos y formular conclusiones acerca de la población, este análisis es el objetivo del modelamiento estadístico. Por tanto, los datos tomados como muestra deben ser representativos de la población, en caso contrario las generalizaciones e inferencias que se hagan no serán precisas.

- **Análisis predictivo:** Este análisis consiste en tomar los datos actuales e históricos y realizar predicciones acerca del futuro. De la misma manera que en el análisis inferencial, la calidad de las predicciones depende de la elección correcta de las variables. Sin embargo, resulta mejor si se tiene una gran cantidad de datos y un modelo simple pero apropiado.

- **Análisis causal:** Este análisis consiste en saber qué pasa con una variable cuando se manipula otra variable, buscando la causa y el efecto de la relación entre esas variables. Generalmente este análisis se aplica a los resultados de estudios aleatorios que fueron diseñados para encontrar la causa.

- **Análisis mecanicista:** El objetivo de este análisis es comprender los cambios exactos en las variables que conducen a cambios en otras variables. Este tipo de análisis es de alguna manera predictivo, se usan metodologías de muy alta precisión y se aplica comúnmente en las ciencias físicas, ingeniería o ciencias biológicas.

En la actualidad un término que es muy utilizado es “Bigdata”. Un conjunto de datos puede tener tres características importantes, que dificultan su procesamiento: Volumen (necesidad de recolectar y almacenar grandes cantidades de datos), Velocidad (datos que tienen que procesarse con gran rapidez) y Variedad (los datos que podemos analizar vienen en muchos formatos). Más adelante se definieron tres características más: Veracidad (datos correctos y completos), Valor (que generan los datos para la toma de decisiones) y Visibilidad (herramientas que permitan mostrar aspectos, patrones, características de los datos). Cuando se trabaja con grandes cantidades de datos uno de los principales desafíos es pasar de datos no estructurados a datos estructurados.

Los datos estructurados son aquellos que conocemos tradicionalmente en bases de datos como MySQL u Oracle, tablas en hojas de cálculo, etc. Sin embargo, los datos que comúnmente encontramos ahora son mucho más desordenados. Actualmente, estos datos desordenados se generan diariamente en gran cantidad, a través del correo electrónico, posts de Facebook, Instagram, Twitter o cualquier interacción con redes sociales, mensajes de texto, hábitos de compra, datos que generan los teléfonos inteligentes, páginas que se visitan, datos sobre la cantidad de tiempo que se pasa en una determinada plataforma, aplicación o página web, fotos y datos de video; por tanto, el trabajo del científico de datos es extraer la información necesaria usando nuevas herramientas o técnicas, y transformarla en algo más ordenado y estructurado.

Existen algunos desafíos de tener grandes cantidades de datos. Primeramente, existen muchos datos crudos que es necesario almacenarlos y analizarlos. Los datos están en constante cambio, es decir que luego de haber finalizado nuestro análisis, hay más datos que se podrían incorporar al mismo análisis. La variedad de datos es abrumadora, hay muchas fuentes de información, algunas veces esto dificulta elegir cual fuente de información es la mejor para responder a la pregunta de ciencia de datos. Por último, raramente se encuentran datos que tengan un formato estructurado; por tanto, antes de buscar patrones o responder a la pregunta de investigación es necesario transformar los datos en un formato adecuado.

A pesar de los desafíos existentes existen muchos beneficios, uno de ellos es que las preguntas que antes no se podían responder debido a la falta de información ahora es posible; debido a que existen muchas más fuentes de información pudiendo realizar conexiones y nuevos descubrimientos. Otro de los beneficios es que se pueden identificar correlaciones ocultas; es decir, los investigadores pueden recolectar y analizar masivas cantidades de datos de un evento y cualquier cosa que esté relacionado con él, permitiendo encontrar patrones, generar conclusiones y predicciones.

La cantidad de información y las diversas fuentes que pueden registrar y transmitir datos se ha disparado, teniendo un volumen impresionante, estos datos tienen mucho valor en diferentes áreas, generando un nuevo mercado, el de la venta de datos. Muchos emprendedores y empresas como Google, Meta, Twitter o LinkedIn tienen un servicio que permite vender los datos que ellos poseen a quien los necesite, este nuevo mercado ha superado en valor al petróleo desde hace algunos años. Los datos generados siguen creciendo debido a que los usuarios de internet siguen incrementándose.

De acuerdo con el reporte publicado en DataReportal en Julio del 2022, hay 7.98 mil millones de personas en el planeta, de los cuales 4.70 mil millones son

usuarios activos de internet; los usuarios que acceden a internet vía teléfonos móviles son 4.38 mil millones y más de cinco millones de aplicaciones están disponibles en tiendas como GooglePlay o Apple Store. Además, el 66 % de las compañías se anuncian en línea. La mayor parte de inversión publicitaria se destina a televisión con un 29 %, búsqueda pagada 17 % y redes sociales con un 13 %. Las razones por lo que la gente usa las redes sociales son varias, la razón principal es que los usuarios buscan permanecer en contacto con sus familiares y amigos, con un 49.7 %, seguido por la diversión o distracción en su tiempo libre con un 36.9 %, y en tercer lugar se tiene el uso informativo con un 36.1 %.

Esta tendencia al alza de los usuarios de redes sociales ha provocado que las empresas analicen los datos publicados y mejoren sus estrategias de marketing para incrementar sus ganancias.

El proceso de análisis de datos es muy complejo, debido a que la mayoría de ellos son no estructurados y heterogéneos. Por esta razón, es necesario seguir un proceso que permita procesar toda la información captada para que pueda ser usada en la toma de decisiones. Este proceso generalmente consta de varias etapas: (i) La primera consiste en recopilar los datos, luego (ii) realizar una limpieza de los datos recolectados, posteriormente (iii) realizar un análisis exploratorio, luego (iv) utilizar el método adecuado para el procesamiento de estos datos de acuerdo con el objetivo planteado y por último (v) presentar de una forma amigable a través de visualizaciones y reportes los resultados obtenidos.

Debido a que mucha gente usa las redes sociales para comunicarse, encontrar información relevante, mantenerse informado o para el entretenimiento, estas plataformas albergan información muy relevante la cual se puede analizar para definir patrones que ayuden a las empresas a mejorar y focalizar sus estrategias de marketing. La evolución de los datos provenientes de redes sociales se puede dividir de acuerdo al contenido generado por una marca o firma de una compañia (FGC[35] por sus siglas en inglés), este contenido es generado a través de las páginas o cuentas de la empresa como un canal de marketing; y el contenido generado por el usuario (UGC[36] por sus siglas en inglés), que se refiere al contenido que clientes de una determinada marca o firma publican en redes sociales (Varsha y col., 2021).

El análisis de datos de redes sociales es un área importante de investigación que ha ganado gran interés en los últimos años. Generalmente consiste en (i) capturar todas las conversaciones y metadatos que ocurren naturalmente en las redes sociales, luego (ii) transformar estos datos en información útil y por último

[35] Firm generated content
[36] User generated content

encontrar formas de hablar sociológicamente sobre los datos transformados y explotar dicha información (Brooker, Dutton y Greiffenhagen, 2017).

Para transformar estos datos y encontrar patrones, en los últimos años se han desarrollado técnicas y tecnologías para analizar, captar, limpiar, procesar y visualizar esta gran cantidad de información. Estas herramientas incluyen muchas disciplinas entre ellas, la ciencia computacional, economía, matemáticas, estadística, entre otras (Ying y col., 2021). A continuación, se resumen las principales herramientas usadas en la analítica de redes sociales:

- **Minería de datos**: Según IBM[37] es un proceso que involucra la extracción, visualización y presentación de información de grandes cantidades de datos. Las técnicas más usadas son: (i) reglas de asociación, método que se basa en encontrar relaciones entre variables de un conjunto de datos. Generalmente se usa para entender los hábitos de consumo de clientes, permitiendo a los administradores desarrollar mejores estrategias de venta y la construcción de sistemas recomendadores. (ii) Las redes neuronales son otro método que consiste en procesar los datos de entrenamiento imitando la interconectividad del cerebro a través de capas y nodos. Otra de las técnicas usadas son los (iii) árboles de decisión, que es una técnica que usa métodos de regresión para la predicción basada en un conjunto de decisiones. Por último, el vecino más cercano es otro algoritmo que se suele utilizar para clasificar los datos en función de su proximidad y asociación con otros datos disponibles.

- **Minería de Texto:** Consiste en transformar los datos textuales no estructurados a un formato estructurado para su análisis. Usa procesamiento de lenguaje natural para entender y procesar automáticamente información tales como, sitios web, libros, correos electrónicos, revisiones, publicaciones de redes sociales, artículos, etc. Entre las técnicas más utilizadas están: asociación de palabras, agrupación de texto, categorización de texto, resumen de texto, análisis de temas, minería de opinión o análisis de sentimientos. Una de las principales aplicaciones de la minería de texto es en la inteligencia de negocios por ejemplo, donde un administrador puede estar interesado en conocer la percepción que tienen sus clientes acerca de un servicio específico, para conocer qué tan bien están posicionándose frente a la competencia; los datos pueden encontrarse en la web en forma de revisiones, tweets o posts (Zhai y Massung, 2016).

[37] IBM, https://www.ibm.com/cloud/learn/data-mining

- **Aprendizaje de máquina**: Según (Krohn, Beyleveld y Bassens, 2020), es un campo de la inteligencia artificial que se encarga de configurar el software de tal forma que pueda reconocer patrones sin la necesidad de que el programador especifique explícitamente cómo llevar a cabo esta tarea de reconocimiento. Existen muchos algoritmos desarrollados en este ámbito, los cuales se clasifican en: *Aprendizaje supervisado*, estos métodos necesitan datos etiquetados que sirvan de entrenamiento para este tipo de algoritmos. *Aprendizaje no supervisado* que sirve para descubrir patrones ocultos y agrupar datos sin la necesidad de tener algún conjunto de datos de entrenamiento, y el *aprendizaje semi-supervisado* que es una combinación de aprendizaje supervisado y no supervisado. Tanto el aprendizaje de máquina como la minería de datos usan los mismos algoritmos; sin embargo, un científico de datos utiliza datos extraídos de la información que ya existe para encontrar patrones y modelar procesos que ayuden a la toma de decisiones, mientras que el aprendizaje automático aprende de los datos existentes para realizar predicciones futuras.

- **Aprendizaje profundo:** Este tipo de algoritmos están compuestos por capas o redes neuronales más sofisticadas. De acuerdo con (Krohn, Beyleveld y Bassens, 2020), esencialmente están compuestas por una capa de entrada, tres o más capas ocultas que sirven para representar y aprender de los datos de entrada y una capa de salida que son los datos de predicción.

- **Análisis basado en grafos:** Consiste principalmente en el uso de teoría de grafos, cuyos nodos se pueden representar como personas o usuarios y los enlaces como interacciones y relaciones entre estos nodos.

Todos estos tipos de análisis han sido abordados por investigadores para poder entender lo que ocurre en redes sociales, buscar patrones, realizar inferencias y predicciones que permitan mejorar el entendimiento de los datos y ayuden a la toma de decisiones.

Los beneficios de transformar datos de redes sociales en información útil han sido aprovechados en áreas como administración, economía, política, turismo, etc. Por ejemplo, se puede mejorar el entendimiento de las preferencias de los usuarios en la elección de un producto determinado o usar tweets para realizar predicciones en una determinada elección.

El turismo uno de los sectores más importantes en la economía de un país, tanto así que la pandemia del COVID-19 hasta la fecha ha causado pérdidas en Latinoamérica de aproximadamente 2,4 billones de dólares, siendo Ecuador el

país más afectado debido a que las Islas Galápagos son visitadas todo el año. La Organización Mundial del Turismo[38] (UNWTO por sus siglas en inglés) indica que países como Francia, España o Estados Unidos son los más afectados por la pandemia debido a que en el 2019 encabezaron la lista al tener más ingresos por turismo internacional.

De acuerdo con UNWTO el turismo internacional podría volver a sus niveles del 2019 en dos a cuatro años. Tarde o temprano el turismo volverá a ser una fuente importante de ingresos que dinamiza la economía y contribuye al desarrollo de los países.

La hipótesis principal de esta Tesis Doctoral, es que dado el ámbito del turismo, el uso de datos de redes sociales y métodos de analítica de datos puede contribuir a entender y mejorar los destinos turísticos tanto para viajeros en su etapa pre-viaje como para administradores para mejorar sus servicios. Para esto, se plantea el estudio de viabilidad de nuevas técnicas de análisis que contribuyan al sector turístico enfocándonos en texto escrito tanto en inglés como en español publicado en estas plataformas.

1.2. Objetivos

El problema que se plantea, en general, en este estudio es el de analizar datos turísticos obtenidos de redes sociales. El problema es tratado desde el enfoque de la minería de texto. Uno de los inconvenientes es que no existen herramientas que sirvan de forma global para todos los idiomas existentes en el mundo. Debido a que el procesamiento de lenguaje natural se tiene que abordar teniendo en cuenta los lenguajes que se emplean, en este trabajo se toma en cuenta el lenguaje Español y el Inglés.

Por tanto, el objetivo primordial de esta Tesis Doctoral es el de explorar y proponer nuevos enfoques, métodos y herramientas en el contexto de minería de texto y turismo, basándose en el uso de técnicas de aprendizaje automático y análisis de sentimientos que permitan obtener patrones y entender el turismo de un destino. Para ello se pretenden abordar las siguientes tareas específicas:

- **O1:** Revisión exhaustiva de la literatura en lo referente a análisis de redes sociales y su influencia en el contexto Turístico y la comprensión de las técnicas existentes.
- **O2:** Propuesta de un enfoque o arquitectura que permita analizar el contenido publicado en redes sociales (Twitter, Facebook, Instagram), que

[38] https://www.e-unwto.org/doi/pdf/10.18111/9789284421237

incluya el proceso de extracción, limpieza y generación de información relevante para un destino turístico.

- **O3:** Análisis de un destino turístico usando técnicas de análisis de sentimientos de dos de las principales redes sociales de propósito general, Twitter e Instagram, que permitan encontrar los lugares más representativos de un destino turístico.
- **O4:** Análisis y uso de técnicas avanzadas de minería de texto para la identificación de las percepciones que los usuarios tienen sobre lugares, monumentos, manifestaciones culturales o servicios de un destino turístico.

Hay que indicar que todo el desarrollo experimental se ha llevado a cabo a través de un caso de estudio empleando datos turísticos obtenidos desde las redes sociales de Twitter, Instagram y TripAdvisor de la ciudad española de Granada y su provincia.

1.3. Estructura de la Tesis doctoral

Los siguientes capítulos de esta Tesis Doctoral están organizados en función de los avances que se han producido de acuerdo con la línea de investigación. En primer lugar, en el Capítulo 2 se realiza una revisión exhaustiva de literatura de todas las técnicas empleadas en análisis de redes sociales e influencia en turismo. A continuación, se plantea un enfoque en el Capítulo 3 para analizar los datos en un contexto estacional con una combinación de métodos de la ciencia de datos y minería de texto, permitiendo presentar una visión general de un destino turístico a las personas que piensen o vayan a visitar el lugar y dar una perspectiva clara a los administradores de las fortalezas y debilidades de sus servicios turísticos. En el Capítulo 4 se realiza un análisis comparativo de herramientas de análisis de sentimientos utilizadas tanto en español como en inglés. Además, se realiza el análisis de técnicas de aprendizaje profundo y se propone una mejora para el análisis de sentimiento de datos turísticos en español usando una fuente de datos de entrenamiento relativamente pequeña, mejorando la eficiencia, para luego usar las mejores técnicas en la detección de lugares o entidades y las percepciones o aspectos de los datos clasificados. En el Capítulo 5 se mejora la identificación de entidades y aspectos, lo que permite conocer de una forma más precisa la causa de (in)satisfacción que tienen los usuarios sobre un destino turístico. Por último, se presentan las conclusiones generales de este estudio en el Capítulo 6.

Capítulo 2

Revisión de la literatura sobre análisis de redes sociales, influencia social y turismo

Los pilares fundamentales de la analítica de redes sociales son: redes sociales, usuarios o personas e industria y tecnología que transforma conversaciones, comentarios, fotos, videos, me gusta, blogs, tweets, etc. en datos de mucho valor para analistas y especialistas en marketing. Su objetivo es analizar y monitorear el comportamiento de los usuarios, la lealtad a la marca y otros indicadores de desempeño, transformando estos datos en información útil (Misirlis y Vlachopoulou, 2018). En la sección 2.2 se identificaron publicaciones importantes en análisis de redes sociales y Smart-Tourism o SMAST (Social Media Analytics for Smart-Tourism por sus siglas en inglés).

El término «Smart» se ha convertido en una palabra de moda para describir los desarrollos tecnológicos, económicos y sociales que utilizan tecnologías dependientes de sensores y una gran cantidad de intercambio de datos e información (Gretzel y col., 2015). En primer lugar, es necesario tener una idea clara de qué es «Smart Tourism», este término se deriva del concepto de «Smart City» cuyo objetivo es mejorar la calidad de vida de todos los ciudadanos. El término «Smart Tourism» se refiere a la actividad donde el turista aplica nuevas tecnologías en sectores relacionados con servicios de experiencia turística, aplicaciones para reservas, alojamiento, transporte y restauración; además, se

relaciona como un fenómeno social donde la industria hotelera y turística existente se integran con el uso de tecnologías de la información y la comunicación (TIC) (Lee, 2017). En la actualidad, la actividad turística está estrechamente ligada a la tecnología.

Las aplicaciones y servicios relacionados con el turismo han sido influenciados por las redes sociales, que cada año aumentan el número de usuarios, y su impacto ha sido explotado por las empresas de marketing en general. La analítica de redes sociales centrada en el turismo se basa en el uso de tecnologías de la información y la comunicación para recopilar, limpiar, procesar, analizar y visualizar esos datos para transformarlos en información útil con el fin de mejorar tanto los servicios turísticos como la experiencia del turista.

Por lo tanto, se puede definir Social Media Analytics y Smart Tourism como un conjunto interdisciplinario de métodos y técnicas que permiten recopilar datos de las redes sociales (es decir, blogs, sitios de reseñas, uso compartido de medios, sitios de preguntas y respuestas, marcadores sociales, redes sociales, noticias, wikis, etc.) para procesar, analizar y visualizar información útil con el fin de mejorar los servicios y las aplicaciones turísticas.

Se ha realizado una revisión extensa de las publicaciones relacionadas con SMAST, para eso es fundamental crear un esquema de clasificación conceptual de la literatura encontrada, utilizando cuatro dimensiones o categorías, tales como: metodologías de investigación, tipo de análisis, temas de actualidad sobre turismo y tipo de plataforma de redes sociales (Misirlis y Vlachopoulou, 2018), que proporciona una descripción general de los problemas de investigación actuales en SMAST.

2.1. Social Media Analytics

El término «Social Media Analytics» SMA se refiere a «un campo emergente de investigación interdisciplinaria que tiene como objetivo combinar, ampliar y adaptar métodos para el análisis de redes sociales» (Stieglitz y col., 2014). Otra definición lo considera como un conjunto de herramientas para «recopilar, analizar, resumir y visualizar datos de redes sociales, generalmente impulsados por requisitos específicos de una aplicación de destino». (Zeng y col., 2010).

De acuerdo con (Zeng y col., 2010) la investigación en SMA sirve para facilitar la interacción y conversaciones entre comunidades virtuales y la extracción de patrones de interacción entre las personas. Social media contiene un conjunto enriquecido de datos o metadata (hashtags, opiniones, ratings, perfiles de usuario, etc.) que puede ser usado para muchos propósitos, uno de ellos es en

la política. En los últimos años se han desarrollados estrategias sofisticadas que permiten predecir el comportamiento de los votantes con la finalidad de modificar o corregir estrategias de los partidos políticos de acuerdo a la información que circula en social media (Zhang y col., 2010). Las redes sociales constituyen el vehículo ideal y la información base para evaluar la opinión pública sobre políticas o posiciones políticas. En algunos casos, este tipo de empresas usan datos privados de votantes y tratan de manipular la opinión publicando contenido falso de adversarios políticos con la finalidad de ganar elecciones que son moderadas por la voluntad popular a través del voto. Las elecciones para presidente de Estados Unidos en el 2016 son un ejemplo de ello con el escándalo de Cambridge Analytica.

Las redes sociales son una fuente de información importante en el contexto de los negocios y el marketing, además constituyen una plataforma de ejecución empresarial para el diseño de y la innovación de productos, la gestión de las relaciones con los consumidores usando plataformas de inteligencia de negocios que en la actualidad son muy populares (Gruhl y col., 2010).

En el contexto del turismo se han desarrollado varias herramientas para analizar datos de redes sociales, sin embargo para la información textual (posts, tweets, comentarios o respuestas a otros usuarios) que es la que tiene datos relevantes acerca de las percepciones de los usuarios, no hay herramientas suficientemente maduras que trabajen con múltiples lenguajes (inglés, español, francés, italiano, chino, etc.); por tanto, este estudio propone algunos enfoques que incluyen el lenguaje español e inglés para analizar el texto que los usuarios publican en redes sociales sobre algún destino turístico.

Los datos son lo más importante en cualquier empresa, industria, organización o investigación. La forma como se obtienen datos de redes sociales es a través de API (Application Programming Interface por sus siglas en inglés). Esta interfaz de programación de aplicaciones permite la comunicación entre aplicaciones a través de un conjunto de reglas. Twitter tiene su «Search API» y su «Streamming API», que son las más usadas mientras que Facebook e Instagram tienen su «Graph API» que funciona para obtener datos de publicaciones, comentarios, etc. (Stieglitz y Dang-Xuan, 2013). Si no se tiene una API, es posible obtener información minando directamente la página HTML usando técnicas y herramientas de web-crawling[1].

La información que se obtiene generalmente está en formato JSON[2] que es un formato sencillo con un conjunto de objetos Javascript.

[1] https://en.wikipedia.org/wiki/Web_crawler
[2] https://es.wikipedia.org/wiki/JSON

Para la búsqueda de información, en este trabajo se utilizó el enfoque de búsqueda por tema (topic-based approach), es decir se definió un conjunto de palabras clave que se usaron para buscar y obtener información desde las redes sociales.

El procesamiento de los datos es un tema amplio que se menciona ampliamente en el siguiente capítulo a través de un enfoque de análisis de datos turísticos, y por último el análisis de contenido, donde intervienen técnicas de procesamiento de lenguaje natural y análisis de sentimientos, los cuales se abordan en los capítulos 4 y 5 del presente trabajo.

2.2. Análisis de redes sociales y el turismo

Para analizar los trabajos relacionados con SMAST se identificó la metodología usada (cualitativa o cuantitativa), tipo de análisis (procesamiento de lenguaje natural, análisis estadístico, análisis de sentimientos, técnicas de modelado de ecuaciones estructurales y análisis predictivo), temas sobre el turismo (destinos y atracciones, toma de decisiones en la industria turística, satisfacción turística, movilidad y turismo, y búsqueda/estructuración de información sobre viajes) y plataformas de redes sociales en general (Facebook, Twitter, Instagram, MySpace, SinaWeibo, TripAdvisor, TravelBlogs, Ctrip, Yelp, Airbnb, etc.).

Los términos utilizados para la búsqueda fueron: («social media» AND (analytic OR analysis) AND (tourism OR «smart tourism»), estos términos se buscaron en el título, resumen y palabras clave de los artículos. Este proceso se realizó en dos bases de datos académicas, concretamente Scopus y Web of Science. Los artículos pertenecientes a libros, capítulos de libros, artículos en prensa y reseñas fueron excluidos. Para seleccionar los artículos se utilizaron los siguientes criterios: (i) Estudios que usen datos de por lo menos una red social, (ii) estudios que usen técnicas de análisis de datos; (iii) artículos que hayan sido publicados en idioma inglés; por último, (iv) artículos enfocados en el dominio del turismo, e-tourism o Smart-Tourism. En total, se encontraron 1581 (743 en Web of Science y 838 en Scopus) artículos entre 2015 y septiembre de 2021. Luego de descartar los que se repiten entre las bases de datos seleccionadas y los que no cumplan con los criterios antes mencionados, se seleccionaron 62 de ellos.

En la tabla 2.1, se muestra el título de cada trabajo, autor, año y revista donde han sido publicados cada uno de los artículos seleccionados.

ID	Título	Revista

P01	Adoption of travel information in user generated content on social media: the moderating effect of social presence (Chung, Han y Koo, 2015)	Behaviour and Information Technology
P02	SoCoMo marketing for travel and tourism: Empowering co-creation of value (Buhalis y Foerste, 2015)	Journal of Destination Marketing and Management
P03	The role of prior experience in the perception of a tourism destination in user-generated content (Marchiori y Cantoni, 2015)	Journal of Destination Marketing and Management
P04	Tourism analytics with massive user-generated content: A case study of Barcelona (Marine-Roig y Anton Clavé, 2015)	Journal of Destination Marketing and Management
P05	Identifying and ranking cultural heritage resources on geotagged social media for smart cultural tourism services (Nguyen, Camacho y Jung, 2017)	Personal and Ubiquitous Computing
P06	A novel popular tourist attraction discovering approach based on geo-tagged Social media big data (Peng y Huang, 2017)	ISPRS International Journal of Geo-Information
P07	Creating value from social big data: Implications for smart tourism destinations (Vecchio y col., 2018)	Information Processing and Management
P08	Generating travel-related contents through mobile social tourism: Does privacy paradox persist? (Hew y col., 2017)	Telematics and Informatics
P09	How can big data support smart scenic area management? An analysis of travel blogs on Huashan (Shao, Chang y Morrison, 2017)	Sustainability

P10	Smart tourism technologies in travel planning: The role of exploration and exploitation (Huang y col., 2017)	Information and Management
P11	Social media analytics and value creation in urban smart tourism ecosystems (Brandt, Bendler y Neumann, 2017)	Information and Management
P12	Social support and commitment within social networking site in tourism experience (Chung, Tyan y Chung, 2017)	Sustainability
P13	The relationship among tourists persuasion, attachment and behavioral changes in social media (Chung y Han, 2017)	Technological Forecasting and Social Change
P14	Digital technology in a smart tourist destination: The case of Porto (Costa Liberato, Alén-González y Azevedo Liberato, 2018)	Journal of Urban technology
P15	Do online information sources really make tourists visit more diverse places?: Based on the social networking analysis (Lee, Chung y Nam, 2019)	Information Processing and Management
P16	Development of social media strategies in tourism destination (Kiráľová y Pavlíčeka, 2015)	Procedia - Social and Behavioral Sciences
P17	Heritage tourism entrepreneurship and social media: Opportunities and challenges (Surugiu y Surugiu, 2015)	Procedia - Social and Behavioral Sciences
P18	How smart is your tourist attraction?: Measuring tourist preferences of smart tourism attractions via a FCEM-AHP and IPA approach (Wang y col., 2016)	Tourism Management
P19	A big data analytics method for tourist behaviour analysis (Miah y col., 2017)	Information and Management
P20	Content mining framework in social media: A FIFA world cup 2014 case analysis (Thomaz y col., 2017)	Information and Management

P21	Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges (Rashidi y col., 2017)	Transportation Research
P22	Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy (Chua y col., 2016)	Tourism Management
P23	Measuring tourism destinations using mobile tracking data (Raun, Ahas y Tiru, 2016)	Tourism Management
P24	Opinion mining from online hotel reviews — A text summarization approach (Hu, Chen y Chou, 2017)	Information Processing and Management
P25	Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings (Hu y Chen, 2016)	International Journal of Information Management
P26	Shared experience in pretrip and experience sharing in posttrip: A survey of Airbnb users (Bae y col., 2017)	Information and Management
P27	Effects of tourism information quality in social media on destination image formation: The case of Sina Weibo (Kim y col., 2017b)	Information & Management
P28	Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning (Ma y col., 2018)	International Journal of Hospitality Management
P29	Obtaining a better understanding about travel-related purchase intentions among senior users of mobile social network sites (Kim, Lee y Bonn, 2017)	International Journal of Information Management
P30	Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor (Chang, Ku y Chen, 2019)	International Journal of Information Management
P31	Social return and intent to travel (Boley y col., 2018)	Tourism Management

P32	What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management (Kim y col., 2017a)	Technological Forecasting and Social Change
P33	Will firm's marketing efforts on owned social media payoff? A Quasi experimental analysis of tourism products (Chang y col., 2018)	Decision Support Systems
P34	The effects of social media on emotions, brand relationship quality, and word of mouth: An empirical study of music festival attendees (Hudson y col., 2015)	Tourism Management
P35	The use of social media in travel information search (Chung y Koo, 2015)	Telematics and Informatics
P36	Using social network analysis to explain communication characteristics of travel-related electronic word-of-mouth on social networking sites (Luo y Zhong, 2015)	Tourism Management
P37	Likes—The key to my happiness: The moderating effect of social influence on travel experience (Sedera y col., 2017)	Information and Management
P38	Landmark reranking for smart travel guide systems by combining and analyzing diverse media (Shen y col., 2016)	IEEE Transactions on Systems, Man, and Cybernetics: Systems
P39	Progress on smart tourism research (Mehraliyev, Choi y Köseoglu, 2019)	Journal of Hospitality and Tourism Technology
P40	Exploring the Roles of DMO's Social Media Efforts and Information Richness on Customer Engagement: Empirical Analysis on Facebook Event Pages (Lee y col., 2021)	Journal of Travel Research

P41	Actualizing big data analytics for smart cities: A cascading affordance study (Zeng y col., 2020)	International Journal of Information Management
P42	Analysis of the performance and robustness of methods to detect base locations of individuals with geo-tagged social media data (Liu y col., 2021)	International Journal of Geographical Information Science
P43	Game on! A new integrated resort business model (Tham y Huang, 2019)	Tourism Review
P44	The influence of cultural origins of visitors when staying in the city that never sleeps (Moro y col., 2022)	Tourism Recreation Research
P45	Developing an artificial intelligence framework for online destination image photos identification (Wang, Luo y Huang, 2020)	Journal of Destination Marketing & Management
P46	Visualizing theme park visitors' emotions using social media analytics and geospatial analytics (Park y col., 2020)	Tourism Management
P47	Social media sentimentas an additional performance measure? Examples from iconic theme park destinations (Widmar y col., 2020)	Journal of Retailing and Consumer Services
P48	Application of social media analytics in tourism crisis communication (Park, Kim y Choi, 2019)	Current Issues in Tourism
P49	Exploring influential factors affecting guest satisfaction: Big data and business analytics in consumer-generated reviews (Lee y col., 2020)	Journal of Hospitality and Tourism Technology
P50	Thematic analysis of destination images for social media engagement marketing (Song, Park y Park, 2021)	Industrial Management & Data Systems

P51	Destination image through social media analytics and survey method (Lin y col., 2021)	International Journal of Contemporary Hospitality Management
P52	Improving the resident–tourist relationship in urban hotspots (Vu y col., 2021)	Journal of Sustainable Tourism
P53	Machine infelicity in a poignant visitor setting: comparing human and AI's ability to analyze discourse (MacCarthy y Shan, 2022)	Current Issues in Tourism
P54	Designing tourist experiences amidst air pollution: A spatial analytical approach using social media (Zhang y col., 2020)	Annals of Tourism Research
P55	Co-visitation network in tourism-driven peri-urban area based on social media analytics: A case study in Shenzhen, China (Sun, Shao y Chan, 2020)	Landscape and Urban Planning
P56	Mapping destination images and behavioral patterns from user-generated photos: a computer vision approach (Zhang, Chen y Lin, 2020)	Asia Pacific Journal of Tourism Research
P57	Chinese cultural theme parks: text mining and sentiment analysis (Zhang, Li y Hua, 2022)	Journal of Tourism and Cultural Change
P58	Branding luxury hotels: Evidence from the analysis of consumers' "big" visual data on TripAdvisor (Giglio y col., 2020)	Journal of Business Research
P59	#ILoveLondon: An exploration of the declaration of love towards a destination on Instagram (Filieri, Yen y Yu, 2021)	Tourism Management
P60	Smart tourism destinations: a critical reflection (Baggio, Micera y Del Chiappa, 2020)	Journal of Hospitality and Tourism Technology
P61	Data-focused managerial challenges within the hotel sector (Lamest y Brady, 2019)	Tourism Review

P62	Peeking inside the minds of tourists using a novel web analytics approach (Aggarwal y Gour, 2020)	Journal of Hospitality and Tourism Management

Cuadro 2.1: Artículos sobre SMAST

Los 62 trabajos seleccionados se distribuyen de la siguiente manera: 9 publicados en 2015, con un 14.52 %, 5 publicados en 2016 con un 8.1 %, 17 publicados en 2017 con un 27.42 %, 5 publicados en 2018 con un 8.1 %, 6 publicados en el 2019 con un 9.7 %, 12 publicados en el 2020 con un 19.35 % y 8 publicados hasta septiembre de 2021 con un 12.9 % . Así, el interés por SMAST es significativo, siendo el 2017 el año con mayor número de estudios publicados.

Además, se procede a clasificar cada artículo según cuatro dimensiones diferentes: (i) la metodología de investigación (MI), (ii) el tipo de análisis (TA), (iii) temas de actualidad en turismo (TAT) y (iv) tipo de redes sociales (TRS) usada. La búsqueda, selección y clasificación de cada trabajo permite resumir cada publicación, tener una visión clara sobre los temas en los que se enfoca cada trabajo y encontrar los temas de actualidad que generan más interés.

Por cada dimensión se seleccionaron algunos temas. Cada uno de ellos fue seleccionado de acuerdo con la lectura de los 62 artículos analizados. Para la dimensión MI se seleccionaron la Revisión de literatura / Enfoque teórico / Análisis exploratorio e Investigación basada en cuestionarios. Para la dimensión TA se seleccionó: Análisis predictivo, Análisis de contenido / Procesamiento de lenguaje natural, Análisis estadístico, Análisis de sentimientos, Análisis de la actividad en redes sociales / Análisis de redes sociales (grafos) y Técnicas de modelado de ecuaciones estructurales (SEM, por sus siglas en inglés). Para la dimensión TAT se seleccionaron los siguientes temas: Destinos y atracciones, Toma de decisiones / Marketing, Satisfacción con el viaje, Comportamiento de movilidad / Traslados hacia lugares turísticos, Información de viaje / Información de búsqueda / Boca a boca electrónica (eWOM) / Contenido generado por el usuario (UGC) y Privacidad de datos. Estos temas de actualidad están relacionados con el trabajo de (Shafiee y Ghatari, 2016), ellos mencionan temas como: calidad del servicio, reputación e imagen del destino, UGC como eWOM, experiencias, comportamientos y patrones de movimientos. Por último, la dimensión TRS tiene las siguientes clases: Horizontales (Facebook, Instagram, etc), Verticales y contenido compartido (TripAdvisor, Flickr, etc), Blogging y Microblogging.

Se utiliza el término tipología en lugar de taxonomía, clasificación utilizada en (Misirlis y Vlachopoulou, 2018); su clasificación se adapta a este estudio. Se descartan las categorías de marketing y campos de estudio y se agrega una categoría basada en el turismo. A partir del análisis de los trabajos seleccionados, se identificaron diferentes subcategorías específicamente relacionadas con el dominio del turismo, análisis de datos, metodología de la investigación y tipos de redes sociales donde investigadores podrán agregar, modificar o eliminar categorías. Cada trabajo está relacionado con Smart-Tourism, sin embargo; en la tipología se agregó la categoría «Temas de actualidad en el turismo».

En la tabla 2.2 se muestra la tipología usada para análisis de redes sociales y Smart-Tourism. La primera columna se refiere a las cuatro dimensiones MI, TA, TAT y TRS. La segunda columna corresponde a la micro clasificación, la misma que fue seleccionada de acuerdo con los artículos analizados. La última columna corresponde al identificador de cada subcategoría.

Macro clasificación	**Micro clasificación**	**ID**
MI	Revisión de literatura / Enfoque teórico / Análisis exploratorio	I01
MI	Investigación basada en cuestionarios	I02
TA	Análisis predictivo	I03
TA	Análisis de contenido / Procesamiento de lenguaje natural	I04
TA	Análisis estadístico	I05
TA	Análisis de sentimientos	I06
TA	Análisis de la actividad en redes sociales / Análisis de redes sociales (grafos)	I07
TA	Técnicas de modelado de ecuaciones estructurales (SEM)	I08
TAT	Destinos y atracciones	I09
TAT	Toma de decisiones / Marketing	I10
TAT	Satisfacción con el viaje	I11
TAT	Comportamiento de movilidad / Traslados hacia lugares turísticos	I12
TAT	Información de viaje / Información de búsqueda / eWOM/UGC	I13
TAT	Privacidad de datos	I14
TRS	Horizontales (Facebook, Instagram, etc)	I15

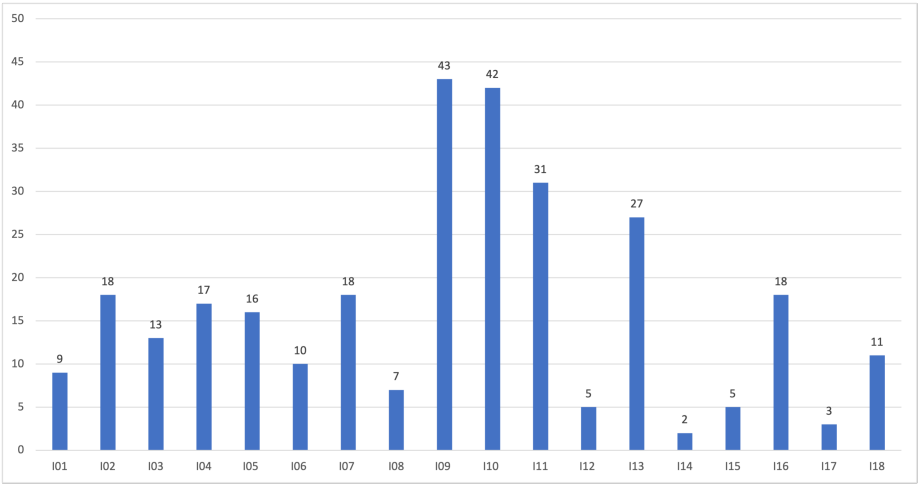

FIGURA 2.1: Mapeo de trabajos analizados usando S3M

TRS	Verticales y contenido compartido (TripAdvisor, Flickr, etc)	I16
TRS	Blogging	I17
TRS	Microblogging	I18

CUADRO 2.2: Tipología para los estudios basados en análisis de redes sociales y Smart-Tourism

En la figura 2.1 se muestra el mapeo de cada trabajo analizado utilizando la tipología de la tabla 2.2.

En la figura 2.1 se puede observar que los investigadores basan sus estudios principalmente en los destinos y atracciones (I09), toma de decisiones y marketing (I10) y satisfacción del usuario con el viaje (I11). Además, un número importante de investigadores se centra en la búsqueda de información/UGC/eWOM (I13). En cuanto a los estudios correspondientes al análisis de contenido y/o procesamiento de lenguaje natural (I04) y análisis de sentimientos (I06), no son tan numerosos, pudiendo combinar estas líneas de investigación con los destinos(I09), toma de decisiones (I10) y UGC/eWOM (I11).

De acuerdo con este análisis, los investigadores usan mayoritariamente datos de redes sociales como TripAdvisor o Flickr (I16) y Twitter o Sina Weibo (I18).

El objetivo es encontrar los estudios más importantes relacionados con métodos o técnicas de análisis de datos de redes sociales y el turismo. Algunos trabajos relevantes encontrados no usan precisamente datos de redes sociales, en vez de ello usan encuestas o entrevistas (I02) que contienen preguntas relacionadas con el uso de redes sociales en la industria turística.

Analizando cada uno de los trabajos seleccionados, hay mucho interés en la calidad de los datos generados a partir de las redes sociales, porque tener datos fiables influye en que la información sea de calidad, de tal forma que pueda ser presentada para que otros usuarios puedan elegir un destino para sus vacaciones y utilizada con fines de marketing por administradores y prestadores de servicios.

Otro tema encontrado que tiene una gran relevancia en el sector es la satisfacción en el viaje. Existen algunas investigaciones basadas en el uso de encuestas o cuestionarios y sus posibles sesgos asociados (por ejemplo, sesgos de deseabilidad social, sesgos de recuerdo a corto plazo, etc.). Los datos de redes sociales son relativamente fáciles de obtener; por tanto, es posible usarlos para evaluar el nivel de satisfacción a través del análisis de sentimientos; este tema se aborda ampliamente en los capítulos 4 y 5.

2.3. Influencia Social, redes sociales y Turismo

En esta sección se analizan los métodos y técnicas que se han desarrollado hasta el momento sobre la influencia que las redes sociales tienen en diferentes dominios y específicamente en el turismo, el ranking de influencia en las plataformas de redes sociales y los influencers.

2.3.1. Influencia en redes sociales y los Influencers

Uno de los primeros estudios analizados (Freberg y col., 2011) utiliza «California Q-sort» para cuantificar las percepciones subjetivas de cuatro personas influyentes en las redes sociales. Una vez que se han identificado los influencers para una marca u organización, proporciona un método para evaluar y comparar las impresiones subjetivas de las audiencias de cada influencer.

Otros investigadores (Francalanci y Hussain, 2015; Francalanci y Hussain, 2017) utilizan un enfoque visual para encontrar personas influyentes en Twitter; el concepto central de este enfoque es identificar personas influyentes navegando a través de la red de amigos de un usuario. La herramienta utiliza los recursos del usuario -usuarios a quien se sigue, favoritos, tweets, listas y seguidores-; recursos

de tweets -URL, hashtags, retweets, favoritos y menciones-; después de eso, utilizan el proceso de jerarquía analítica para ponderar diferentes parámetros, en función de su importancia relativa y encontrar los influyentes.

Otro estudio (Zhao y col., 2016) tiene en cuenta la experiencia de los usuarios; ellos proponen un método computacional para medir la correlación entre experiencia e influencia en las redes sociales, los autores recolectaron 13,684 celebridades -propagadores- de Sina Weibo, descubrieron que las celebridades con alta experiencia tienen una influencia más fuerte, además, la experiencia/estatus/reputación es más importante que la relevancia y la participación en las redes sociales.

En (Nargundkar y Rao, 2016), se diseñó un sistema llamado InfluenceRank que se basa en las características del perfil de Twitter y los tweets. Los autores utilizan el número de retweet, los favoritos y las respuestas como una característica del enfoque de aprendizaje automático basado en regresiones.

Otra investigación analiza la influencia que StockTwits.com -una red social donde los usuarios comparten publicaciones sobre acciones, índices y mercados financieros- tienen en las bolsas de valores a través de la combinación de características de los usuarios para comprender qué tipo de inversionista ejerce una mayor influencia a través de sus mensajes (Piñeiro-Chousa, Vizcaíno-González y Pérez-Pico, 2017).

Otro trabajo se centra en la validación de un instrumento llamado modelo de aceptación de tecnología, se aplica a usuarios de redes sociales (Facebook, Twitter, Instagram y Pinterest) para encontrar la influencia entre vendedores y compradores en la industria de la moda (Reiter, McHaney y Connell, 2017).

Un trabajo interesante se presenta en (Hassan y col., 2018), ellos desarrollaron un método que incluye varios criterios como: la cantidad de me gusta, seguidores/amigos/suscriptores, publicaciones, comentarios, enlaces, etc. para crear un modelo de evaluación de redes sociales que mida la influencia de los medios de comunicación de manera cuantificable mediante el reconocimiento, la generación de actividades y la credibilidad.

En otro estudio, los autores desarrollaron un sistema de minería de datos sociales utilizando una base de datos emocional construida a partir de datos no estructurados recopilados de Twitter y clasificaron el sentimiento del texto en ocho emociones con el objetivo de identificar y predecir la opinión sobre el tema que el usuario seleccionó (Kim y Hong, 2017).

En (Almeida-Santana y Moreno-Gil, 2017), los autores evaluaron las diferencias en la lealtad hacia los destinos cuando los turistas utilizan las redes sociales como fuente de información antes de su viaje, teniendo en cuenta la nacionalidad y las características sociodemográficas mediante la aplicación de una encuesta.

Descubrieron que las redes sociales tienen un mayor impacto en la lealtad actitudinal. En otro estudio, muestran que existe una relación positiva entre la relación con los compañeros, la expectativa académica, el interés por aprender y el autoconcepto académico es decir, modelaron la relación de los estudiantes con las redes sociales desde la perspectiva de su desempeño académico y la calidad de sus compañeros (Liu y Gu, 2017).

La influencia de las redes sociales también tiene sus desventajas, existe un estudio donde los investigadores determinan el impacto del uso de las redes sociales en la recuperación del estrés; sus resultados mostraron que después de que alguien usa su perfil de Facebook y es sometido a un factor de estrés social, lleva más tiempo recuperarse de esta situación (Rus y Tiemensma, 2017). Por otro lado, en (Chandawarkar, Gould y Grant Stevens, 2018) los autores identificaron a los principales influencers en cirugía plástica en Twitter y se relacionó su influencia en las redes sociales con la influencia académica; obtuvieron una red de influencia para proporcionar a otros cirujanos con quienes puedan interactuar y mejorar su propia influencia.

Las redes sociales han mejorado la comunicación al reducir los costos y aumentar la velocidad de difusión de la información. En (Zeitzoff, 2017) el autor hace un análisis de cómo saber quién promueve protestas, conflictos o políticos que apoyan una nueva política y estimar cuánto apoyo tienen. Analizan en términos generales los efectos e implicaciones de la tecnología de la comunicación en los conflictos dando los siguientes resultados: (i) reducción de barreras a la comunicación, (ii) mayor velocidad de información (iii) dinámica estratégica y adaptación y (iv) nuevos datos e información. Otro estudio relacionado con el anterior muestra cómo se expande la información a través de las redes sociales, los autores hacen un análisis a través de los datos de Twitter sobre el conflicto en Gaza durante 2012, proporcionando evidencia importante de que las redes sociales influyen activamente en la mediación internacional del conflicto (Zeitzoff, 2018).

A diferencia de otros estudios, los autores (Halpern, Valenzuela y Katz, 2017) analizaron y compararon dos plataformas de redes sociales. Se analizó cómo Facebook o Twitter influyen políticamente en la participación activa de la ciudadanía. Introdujeron un modelo teórico para comprender el mecanismo por el cual las plataformas de redes sociales podrían afectar la participación política y encontraron que las redes sociales aumentan esta participación y demostraron que Facebook tiene un efecto significativo en comparación a Twitter. Un artículo relacionado con el anterior, se basa en el análisis de las elecciones presidenciales de Estados Unidos en 2016 utilizando los datos que se generaron en Twitter. En otras palabras, los candidatos tuitearon para expresar su posición, políticas,

atacar a otros candidatos, expresar sus propias opiniones, animar a la gente a votar, etc. Utilizaron la plataforma de microblogging para expandir su mensaje. El objetivo de los autores fue comprender las estrategias de comunicación aplicadas a través de esta plataforma social (Buccoliero y col., 2020). Otro trabajo que va en la misma línea es (Senaweera y col., 2018), los autores analizan la influencia de las interacciones de la comunidad en la afinidad del usuario en Facebook; investigan la influencia de las redes sociales en las elecciones locales en Sri Lanka. Utilizaron una estructura gráfica de los usuarios de la red para representar las interacciones y observaron cambios en las estructuras temporales en las interacciones de los usuarios durante el período de elección. Sin embargo, no prueban si los cambios en las afinidades están relacionados con la elección.

En (Zhao, Zhan y Liu, 2018) los autores probaron un enfoque que integra diferentes perspectivas sobre la influencia de los influencers de Twitter con cuatro factores; salida (número de publicaciones y longevidad), salida reactiva (retweets, favoritos y cantidad de seguidores), salida proactiva (menciones, respuestas, referencias positivas) y posicionamiento en la red (grado de centralidad, centralidad de intermediación y PageRank), ellos hacen énfasis en la importancia de optimizar la influencia de las redes sociales con componentes emocionales positivos.

En (Dupré y col., 2018), los autores encontraron que los usuarios que están predispuestos a compartir sus emociones en redes sociales dependen de su personalidad. Esto se puede relacionar y aplicar para evaluar la influencia en las redes sociales en diferentes dominios. Otro estudio muestra el papel de las redes sociales y la emoción en las protestas de juicio político presidencial de Corea del Sur (Min y Yun, 2019). Los autores revelan la influencia de las redes sociales en la movilización de acciones colectivas y encontraron que los sentimientos de ira y el intercambio de información a través de las redes sociales y mensajes a través de los dispositivos móviles, aumentan la participación en las protestas.

Un estudio interesante menciona un mecanismo para medir el índice de influencia de influencers en plataformas de redes sociales como Facebook, Twitter e Instagram. Descubrieron que el compromiso, el alcance, el sentimiento y el crecimiento son factores clave para determinar los influencers en cada plataforma (Arora y col., 2019).

Por último, la literatura sobre la influencia de las redes sociales se ha mapeado en (Sun y Xie, 2019). Desde una perspectiva internacional, los países con fuerza central en este dominio son: Estados Unidos, China, Corea del Sur e Inglaterra; los centros de investigación se encuentran principalmente en China y Estados Unidos y desde la perspectiva de la academia, las universidades con

mayor fortaleza en esta línea se encuentran en China. Además, las plataformas más utilizadas en este mapeo son: Sina Weibo y Facebook.

2.3.2. Índice de influencia Social

En (Arora, Arora y Palvia, 2014) se presenta un estudio en el cual exploran las redes sociales en cuatro dimensiones: tecnológica, social, económica y ética, y desarrollaron un modelo matemático para medir las redes sociales a través de un número o índice. Esta medida se puede utilizar para comparar el desempeño de las redes sociales de las empresas en la industria de Tecnologías de la Información a través del compromiso, rentabilidad y crecimiento desde las redes sociales.

En (Thoma y col., 2015), evaluaron los recursos educativos para encontrar el impacto o la calidad de los sitios web de medicina de emergencia y cuidados intensivos; para realizar esta evaluación utilizaron PageRanks de Google, Alexa Ranks, seguidores de Twitter, la cantidad de «Me gusta» en Facebook y seguidores de Google+. Los autores muestran que el índice de redes sociales y la correlación con los factores de impacto sugieren que puede ser un indicador estable de impacto para los sitios web de educación médica. En un estudio que coincide con el anterior (Thoma y col., 2018) los autores desarrollaron un índice de redes sociales para cuantificar el impacto relativo de los recursos educativos (sitios web) utilizando el estudio METRIQ. Sus resultados señalan que el índice puede desempeñar un papel en la orientación de las personas hacia recursos de alta calidad que se pueden revisar con técnicas de evaluación crítica.

2.3.3. Influencia de redes sociales y turismo

En (Trunfio y Lucia, 2019) se proporciona un índice de redes sociales innovador para medir la participación de los turistas en las organizaciones de gestión de destinos (DMO por sus siglas en inglés) en las redes sociales; encontraron que los índices de redes sociales pueden usarse para evaluar y monitorear el desempeño de las redes sociales de DMO en la participación turística a través del tamaño de la audiencia, el contenido generado por el usuario y la interacción de los visitantes.

En (Varkaris y Neuhofer, 2017) se explora cómo las redes sociales influyen en la forma en que los usuarios buscan, evalúan y seleccionan un hotel, a través de contenidos generados por el usuario a través de un modelo teórico integrado denominado *«hotel consumer decision-journey through social media»* o proceso de elección de hoteles a través de redes sociales.

En (Stojanovic, Andreu y Curras-Perez, 2018), los autores analizan los efectos de la intensidad del uso de las redes sociales sobre el valor de una marca de destino a través de un estudio cuantitativo mediante el uso de encuestas; ellos muestran que la intensidad del uso de las redes sociales influye significativamente en el conocimiento de la marca.

En el turismo, específicamente en el alojamiento en hoteles, los investigadores analizan el sentimiento de las publicaciones o reseñas de hoteles escritas por viajeros y las respuestas de los seguidores en la plataforma china Sina Weibo, con respecto a quienes se hospedan en un hotel en Macao, utilizando influencers como líderes clave para la promoción de la marca. Sus resultados mostraron que el incremento de reservaciones en hoteles es consecuencia directa de las opiniones positivas en redes sociales (McCartney y Pek, 2018). En la misma línea, Huertas (Huertas, 2018) analiza, la característica de los videos en vivo de los usuarios y cómo influyen en las opiniones y comportamientos turísticos de otros usuarios, el autor mostró que la influencia de los videos en vivo en otros usuarios para generar una atracción positiva es limitada y depende del tipo de video, y concluye que las historias son más populares y utilizadas que los videos en vivo.

En (Wang, 2016), el autor explora las intenciones de los clientes de registrarse en hoteles a través de la página oficial que la empresa tiene en Facebook, los resultados ayudan a comprender las percepciones de los clientes potenciales y proporcionan información sobre la influencia de las redes sociales en la hostelería.

En (S P. Tussyadiah, Kausar y Soesilo, 2018), los autores exploran la relación entre la influencia de las redes sociales y el consumo de los usuarios en restaurantes y encontraron que los consumidores con mayor participación en las redes sociales son más susceptibles a la influencia del consumo global o lo que los autores denominan conformidad con la tendencia, el prestigio social, la percepción de calidad y la influencia de los amigos y/o seguidores en las redes sociales

Por último, en (Sedera y col., 2017) analizan que la influencia social de las redes sociales en los viajes y el turismo es ininterrumpida en todas las fases de los viajes -antes, durante y después-. Utilizando la teoría de la confirmación de expectativas descubrieron que un individuo está influenciado por su grupo social cercano (socios, amigos y familiares) y los gerentes de viajes y turismo deben aprovechar la estrategia de orientación de influencers.

Casi todos los trabajos analizados utilizan la *persuasión* como tipo de influencia social en redes sociales, es decir se basan en los usuarios más influyentes y sus características. En el ámbito turístico la mayor parte de los estudios usan

la investigación cualitativa exploratoria o modelo de ecuaciones estructurales y uno de ellos utilizó la teoría de confirmación de expectativas.

La identificación y explotación de usuarios influyentes es un tema clave en la influencia en redes sociales. Sin embargo, la influencia puede cambiar a lo largo del tiempo y no se ha analizado aún una estrategia efectiva para actualizar el nivel de influencia de los influencers.

2.4. Discusión

Analizando los trabajos tanto del análisis de redes sociales y Smart-Tourism como la influencia de redes sociales y el turismo, no existen repositorios (respetando la privacidad de los usuarios) que permitan a los investigadores realizar su trabajo y probarlos con sus pares académicos. Estos temas han servido para guiar la investigación y poder elegir el tema de investigación adecuado presentado en esta tesis. Luego de analizar cuál es el tema adecuado entre influencia social y análisis de datos en redes sociales enfocado en el turismo, el presente trabajo se basa en el análisis de datos enfocándonos en la minería de texto debido a que cuenta con una amplia perspectiva en investigación para el futuro. Una de las contribuciones es proponer un enfoque de análisis de datos turísticos provenientes de redes sociales y mejorar los métodos de clasificación de sentimientos en español. Otro enfoque, que también se presenta en esta tesis, se basa en encontrar las razones de (in)satisfacción de los usuarios en cuanto a servicios o lugares visitados de un destino turístico. En el capítulo siguiente se propone un enfoque para el análisis de redes sociales con un componente estacional.

Capítulo 3

Análisis de datos turísticos en redes sociales

El análisis de datos de redes sociales tiene enorme relevancia en muchos dominios y los investigadores continúan generando nuevas formas de obtener, procesar, estructurar, visualizar o presentar la información relevante o patrones que permitan mostrar las percepciones de los viajeros.

En el presente capítulo se desarrolla un framework para el procesamiento de toda la información que se ha capturado desde la red social Twitter, tomando en cuenta la perspectiva estacional, de tal forma que permita sacar conclusiones de acuerdo con la temporada del año. Con este enfoque, era necesario encontrar dos dimensiones principales. Primero elegir un destino turístico que tenga entrada de turistas todo el año independientemente de la estación climática. La provincia de Granada en España es uno de los sitios con un enorme potencial turístico y se convirtió en objeto de estudio en este capítulo debido a su patrimonio histórico y monumental, eventos, fiestas populares, música, gastronomía, entretenimiento y eventos deportivos. Y segundo, elegir las técnicas adecuadas de análisis de datos. En el capítulo 1 vimos algunas técnicas como el análisis descriptivo de los usuarios. Además, de acuerdo con la revisión de literatura realizada en el capítulo 2, se procedió a integrar el análisis de contenido y análisis de sentimientos de tal forma que se pueda resumir toda esa gran cantidad de información presente en redes sociales para que se útil tanto para empresarios turísticos como para viajeros.

Como se mencionó anteriormente en este texto, el turismo es uno de los campos más prometedores, ya que representa una parte importante del producto interno bruto de muchos países. Se ha publicado una gran cantidad de artículos sobre el uso de reseñas de viajes en línea o UGC. En esta época donde

las plataformas sociales son imprescindibles ya sea por ocio o por mantenerse informado, los turistas no dejan de compartir sus experiencias a través de estas plataformas por lo que es muy probable que los turistas y viajeros compartan sus experiencias antes, durante y después de sus viajes. Esta información está disponible en línea y en las redes sociales y ha sido analizada extensamente: Airbnb (Leoni, 2020; Martí, García-Mayor y Serrano-Estrada, 2020), Tripadvisor (Chang, Ku y Chen, 2020; Liu y col., 2020; Sudhakar y Gunasekar, 2020), Yelp (Srivastava y Kalro, 2019; Dai y col., 2018; DeAndrea y col., 2018), Expedia (Xu y col., 2019) y Ctrip (Wang y col., 2019). Este contenido permite a los viajeros dar sus opiniones sobre hoteles, restaurantes o lugares que han visitado, y los usuarios que planean visitar dichos lugares pueden usar UGC para leer recomendaciones y comentarios de antemano. Dado que los usuarios están influenciados por las opiniones de otras personas (Viñán-Ludeña y col., 2020; Ampountolas, Shaw y James, 2019) es importante poder analizar toda esta información para comprender cómo los viajeros utilizan e interactúan con las diferentes redes sociales (Viñan-Ludeña, 2019).

El creciente interés por los datos de redes sociales es inmenso. Esto se da en muchos ámbitos, como el político. En Alemania, por ejemplo, para las elecciones de septiembre de 2021, los candidatos adoptaron muchas estrategias en redes sociales para ganar la simpatía de la gente y ganar votos, incluso existen empresas que les ayudan a realizar esta tarea, usando datos de los usuarios de redes sociales, teniendo en cuenta que no se sabe a ciencia cierta si respetan nuestra privacidad. Por tanto, ya no es necesario realizar entrevistas o encuestas para obtener dicha información ya que los datos de las redes sociales se pueden utilizar para analizar y encontrar patrones en cualquier dominio, como es el caso de la imagen de un destino turístico (Marine-Roig, 2019). Sin embargo, cuando los usuarios buscan información sobre destinos turísticos en las redes sociales, pueden confundirse por la cantidad de información disponible y, en consecuencia, tomar decisiones incorrectas porque este tipo de búsqueda generalmente contiene mucho ruido y puede ser poco confiable. Por tanto, es fundamental que esta información sea procesada para que se muestren resultados útiles y fiables. En este enfoque no sólo se utilizaron las técnicas de análisis de datos mencionadas anteriormente, también fue necesario usar la tecnología de análisis de texto para identificar temas y analizar opiniones. Algunos modelos se han utilizado para realizar este tipo de tarea, como el algoritmo LDA (Latent Dirichlet Allocation), máquinas de vectores de soporte, ontología de dominio difuso o análisis de asociación semántica (Ovádek, 2020; Bohr, 2020; Claster, Cooper y Sallis, 2010; Tian y col., 2020). En resumen, este enfoque incorpora extracción de información, limpieza, procesamiento de datos, análisis descriptivo de

los datos de los usuarios y de contenido de las publicaciones, esto incluye la detección de sentimientos y emociones y la extracción de temas desde el texto; y se puede usar en diferentes plataformas de redes sociales como Twitter, Instagram, Facebook, etc. Permitiendo a los turistas antes de su viaje revisar las percepciones de los usuarios de tales redes sobre el destino y a los profesionales y administradores mejorar sus servicios al identificar los comentarios negativos que han recibido sobre un lugar o servicio turístico. Todo este proceso se realiza tomando en cuenta el contexto estacional, es decir, dividiéndolo en dos estaciones (primavera-verano y otoño-invierno) o las cuatro estaciones (primavera, verano, otoño e invierno). Se consideran dos temporadas porque al realizar el análisis de contenido con los datos de la temporada de primavera, los resultados son muy similares a los resultados con los datos de la temporada de verano; lo mismo sucedió con los datos de otoño e invierno.

Es importante destacar que este enfoque utiliza datos de Twitter debido a varias razones: (i) los usuarios publican las percepciones del lugar que visitan en redes sociales como Twitter, Instagram o Facebook, (ii) la facilidad de acceso a los datos de Twitter a través de su API o mediante diferentes herramientas de recolección de datos, (iii) los estudios revisados revelan que se pueden complementar los hallazgos de otras investigaciones realizadas, no solo utilizando plataformas específicas en el ámbito turístico como TripAdvisor o Yelp, finalmente, (iv) de acuerdo a los estudios analizados en el capítulo anterior hay bastante interés en las plataformas de microblogging (I18), sin embargo muchos de ellos se basan en la plataforma Sina Weibo que es la red social que funciona en China, por ello es importante explorar los datos de Twitter.

A continuación se presenta un análisis de los diferentes enfoques que han sido desarrollados por diferentes investigadores, arquitecturas y herramientas tecnológicas para el análisis de datos de redes sociales en el ámbito del turismo.

3.1. Trabajos relacionados

En el capítulo 2 ya se realizó una revisión de literatura, sin embargo, se enfocó en descubrir cuáles trabajos y cuántos de ellos se basan en la metodología, tipo de análisis, toma de decisiones, destinos y atracciones, satisfacción del viaje, comportamiento de movilidad o traslados hacia lugares turísticos, información de búsqueda (eWOM o UGC) y por último qué tipo de plataforma o red social han utilizado. La revisión que se presenta a continuación se enfoca en encontrar los trabajos publicados acerca del contenido generado por el turista, minería de texto en Twitter sobre datos turísticos y sobre todo los enfoques metodológicos

existentes, esto ayudó a establecer una línea base para el framework que se presenta en la siguiente sección.

3.1.1. Contenido generado por el turista

Las redes sociales tienen contenido generado por los viajeros en diferentes plataformas como Twitter, Facebook, Instagram, Yelp, etc. Con ello, los investigadores pueden identificar patrones para la gestión y planificación del turismo. Si bien algunos académicos utilizan el término reseñas de viajes en línea (OTR por sus siglas en inglés) o UGC, todo el contenido turístico generado en las redes sociales puede describirse en términos generales como contenido generado por el viajero (TGC por sus siglas en inglés), y esto abarca todos los contenidos sociales, plataformas (sitios de redes sociales, blogs/microblogs, comunidades de contenido, foros de discusión, etc.)

Las reseñas de viajes en línea son los comentarios publicados por los usuarios de las redes sociales en función de sus experiencias y sus comentarios y opiniones sobre un producto o servicio específico (Litvin, Goldsmith y Pan, 2008). Es importante identificar los motivos por los que alguien podría querer compartir sus experiencias en diferentes redes sociales ya que esto afecta la calidad del contenido publicado sobre turismo en un determinado lugar. A continuación se detallan algunos trabajos que usan OTR o TGC.

Existe una ontología que utiliza un sistema de recomendación para examinar el contenido OTR de las atracciones populares, de modo que se puedan tomar decisiones de viaje precisas e informadas (Pai y col., 2019). En el campo del turismo gastronómico, existe un enfoque para analizar OTR, extraer información útil de elementos textuales, construir una matriz de frecuencias de palabras y realizar un análisis de contenido cuantitativo y temático (Marine-Roig y col., 2019). En otro trabajo, la asociación semántica se usa para extraer palabras temáticas y construir una red (Hou y col., 2019). Otros autores, por su parte, realizan una clasificación jerárquica de los servicios que prestan los restaurantes, e identifican cuatro categorías: servicio, restaurante, hostelería y alimentación (Nascimento Filho, Flores y Limberger, 2019). En sus artículos (Assaker, 2020; Shin y col., 2019), los investigadores examinan los efectos de la confiabilidad, la experiencia, la utilidad percibida y la facilidad de uso percibida con respecto al uso previsto de las publicaciones en las redes sociales por parte de viajeros jóvenes y mayores. Muestran que la facilidad de uso percibida fue mayor entre las mujeres y los viajeros mayores; la utilidad percibida fue más fuerte entre los hombres; y la experiencia tuvo un impacto significativo en los viajeros más

jóvenes. Estos resultados permiten a los emprendedores gestionar mejor las campañas de marketing al centrarse en grupos específicos de usuarios según la edad, la utilidad percibida, la experiencia y la especialización.

El enfoque presentado en este capítulo hace uso de técnicas de minería de texto, en este caso usando Twitter y captando datos correspondientes al turismo. Por esta razón, a continuación se hace una revisión de trabajos relevantes en esta línea de investigación.

3.1.2. Minería de texto, Twitter y Turismo

Twitter junto con Sina Weibo son las plataformas más importantes de microblogging. En este capítulo nos enfocamos en Twitter, pues gracias a su interfaz de programación de aplicaciones (API por sus siglas en inglés), es posible acceder a los datos publicados y esto permite realizar una amplia variedad de investigaciones en muchos dominios diferentes. Uno de estos sectores es el turismo, convirtiéndose en una herramienta invaluable para los empresarios del sector turístico para el desarrollo de campañas de marketing y para la planificación y estrategias de toma de decisiones. Durante la última década, aunque la investigación se ha realizado utilizando todas las redes sociales, Twitter ha sido utilizado específicamente en turismo principalmente por investigadores de Estados Unidos y Japón (Curlin, Jaković y Miloloža, 2019). Las cuentas oficiales también se analizan para explorar segmentos de mercado de empresas para detectar comunidades de clientes en la red a través de minería de texto y análisis de clústeres. Esto no solo es beneficioso para fines de marketing y estrategia de cliente, sino que también permite a las empresas detectar categorías y enfocar mejor sus estrategias de marketing (Punel y Ermagun, 2018). Adicionalmente, se ha construido un sistema para obtener la información turística más actualizada de las redes sociales y extraer los temas más relevantes distinguiendo entre idiomas y utilizando un método para calcular el mejor número de temas usando similitud temática (Hoshino, Ishii y Yamada, 2018).

Contrastando estudios previos que realizaron análisis de datos en sitios de redes sociales de turismo específicos como TripAdvisor o Yelp para analizar el destino turístico en un contexto general, este capítulo amplía el conocimiento existente proponiendo un enfoque de análisis de datos con un contexto estacional a plataformas de redes sociales generales como Twitter o Instagram, que permite extraer toda la información necesaria que los viajeros necesitan para planificar su viaje y los gestores turísticos para mejorar sus servicios.

3.1.3. Enfoques metodológicos

En uno de los trabajos analizados, los autores realizan un estudio longitudinal utilizando datos de 10 Organizaciones de Gestión de Destino (DMOs por sus siglas en inglés) españolas para pronosticar la ocupación hotelera, su planteamiento se basa en 4 etapas: extracción de información (API de Twitter), análisis de contenido (definición de categorías de tweets y clasificación de tweets en categorías), minería de texto (extracción de información significativa de palabras clave en tweets) y red neuronal artificial para pronosticar la ocupación hotelera (Bigné, Oltra y Andreu, 2019). Otros investigadores exploran cómo las DMO regionales italianas emplean Facebook estratégicamente para promover y comercializar sus destinos, utilizan una herramienta cuya arquitectura se compone de cuatro módulos: (i) un módulo extractor, que extrae datos de las páginas de Facebook de las DMO a través de Graph API, (ii) un módulo analizador que procesa el resultado de cada consulta, (iii) un módulo analizador que calcula métricas de participación agregadas durante una ventana de tiempo definida por el usuario y (iv) un módulo de visualización de datos, que muestra los resultados a través de gráficos (Mariani, Di Felice y Mura, 2016). Mientras tanto, otro estudio explora cómo las DMOs europeas utilizan las redes sociales para promocionar y comercializar sus destinos y amplía la investigación existente sobre las redes sociales en el turismo utilizando un tamaño de muestra más grande con muchos idiomas. Se recopilaron datos de Twitter, Facebook, Instagram y YouTube (Uşaklı, Koç y Sönmez, 2017). Por otro lado, teniendo en cuenta las reseñas en línea, algunos investigadores compararon las principales plataformas de reseñas en línea como TripAdvisor, Expedia y Yelp, en términos de calidad de la información relacionada con las reseñas en línea sobre toda la población hotelera de Manhattan, Nueva York; utilizaron LDA y análisis de sentimiento en su análisis (Xiang y col., 2017). En la misma dirección, otros autores proponen un marco compuesto por LDA y análisis de sentimiento para extraer significado de los valiosos comentarios o reseñas aportados por los visitantes de hoteles de diferentes países (Guo, Barnes y Jia, 2017). Utilizando datos de Twitter, otros investigadores proponen un enfoque que explora las emociones de los visitantes del parque temático, que incluye análisis de redes sociales y análisis geoespacial, que se integran con la teoría del Modelo Circumplex de Afecto (Park y col., 2020). Finalmente, se propone un marco analítico espacial para comprender mejor las experiencias turísticas a partir de datos de redes sociales geoetiquetados en Beijing en 2013, los autores investigaron los efectos de la contaminación del aire en las experiencias de los turistas en términos de sus resultados conductuales, emocionales y de salud (Zhang y col., 2020).

Mediante el uso de algoritmos de aprendizaje automático para pronosticar las ventas y las tendencias turísticas, es posible predecir las tendencias sociales y comerciales en el turismo. Sin embargo, todavía existen algunos desafíos en la investigación en términos de identificar las necesidades potenciales de los consumidores en función de la granularidad de las palabras, la asociación semántica y la minería de textos combinada con el multilingüismo, que es un tema importante ya que los turistas provienen de todo el mundo.

En este capítulo se realiza un análisis de texto en las publicaciones de Twitter para explorar los principales temas discutidos, integrando un análisis descriptivo de los datos de los usuarios (es decir, usuarios, me gusta, retweets, comentarios) y análisis de minería de texto (frecuencia de palabras, análisis de sentimientos y modelado de temas)

3.2. Framework para el análisis de datos turísticos

Para identificar lo que las personas están hablando y publicando en las redes sociales sobre lugares, eventos, restaurantes, hoteles, etc., se presenta el siguiente framework para la recopilación, limpieza y análisis de datos, como se muestra en la Figura 3.1. Primero hay que identificar las palabras clave principales para el lugar de estudio; los datos almacenados se utilizan luego para construir un dataset. Posteriormente se realiza un análisis descriptivo, que incluye métricas de publicaciones con análisis geoetiquetado y métricas de usuarios (cantidad de retweets, likes, comentarios, videos, fotos y seguidores). Luego, el texto se limpia antes del proceso de tokenización eliminando menciones de usuario, URL, caracteres redundantes, emojis, caracteres especiales, números y signos de puntuación, se eliminan las stopwords y se realiza el proceso de lematización. Finalmente, se lleva a cabo un análisis de contenido que incluye el cálculo de la frecuencia de palabras, detección de sentimientos y emociones, nubes de palabras y la identificación de temas usando (LDA) (Blei, Ng y Jordan, 2003).

3.2.1. Recolección de datos

Twitter es una de las redes sociales más importantes del mundo y, según Kemp (Kemp, 2020), cuenta con 353.1 millones de usuarios activos en todo el mundo y es utilizada frecuentemente por personas de entre 25 y 49 años. En España, más de 37 millones de personas son usuarios activos de redes sociales, de los 47.75 millones de habitantes más del 80 % usa regularmente social media.

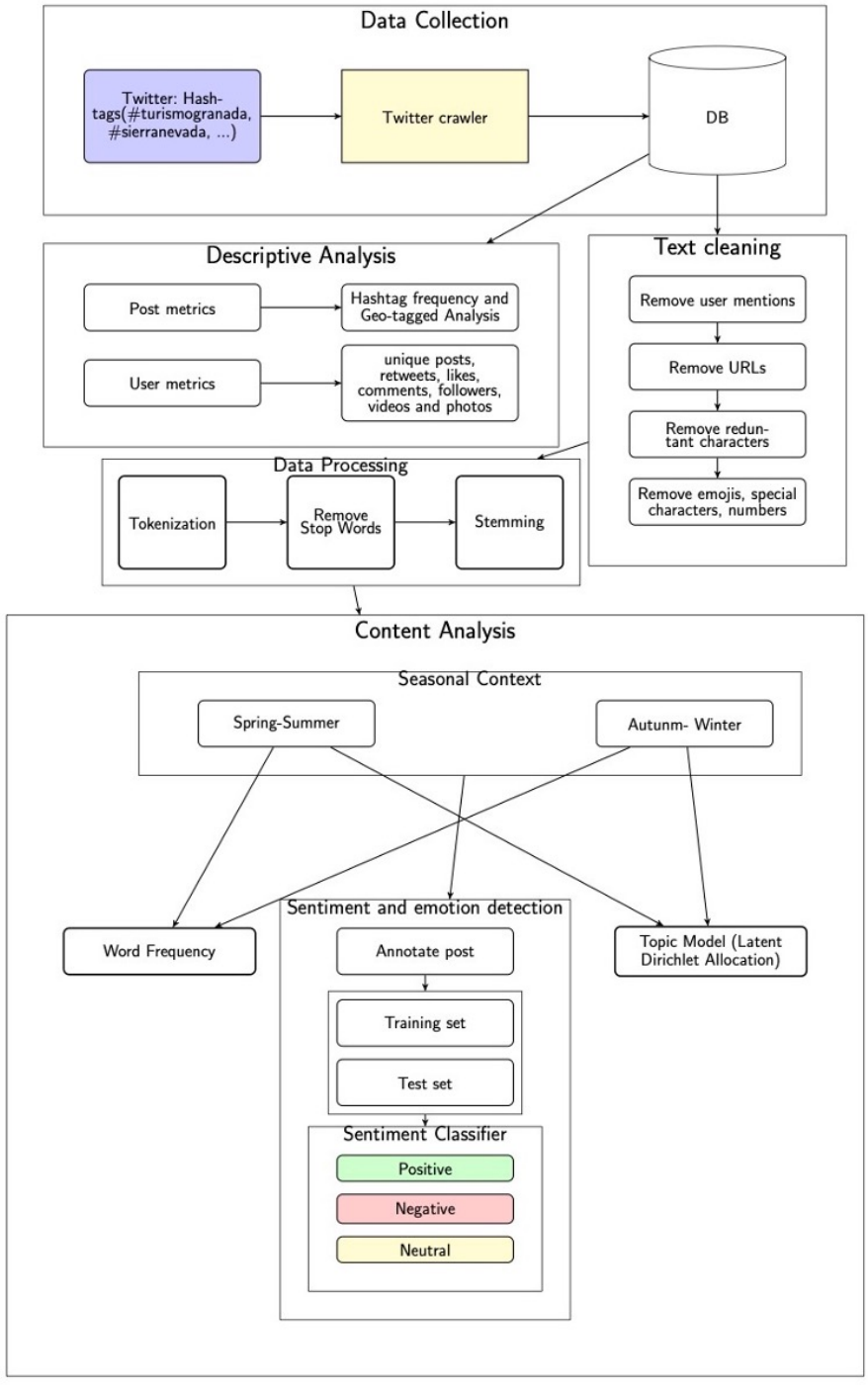

FIGURA 3.1: Framework para el análisis de los datos turísticos de redes sociales

Debido a la pandemia el uso de redes sociales se incrementó, en España esto representa un aumento del 27 %. Aproximadamente el 53 % de los usuarios de redes sociales usan Twitter. Los investigadores pueden utilizar los datos recopilados de Twitter para realizar estudios en muchos dominios científicos diferentes, como salud pública, derecho, informática, etc. (Kim, Nordgren y Emery, 2020; Colditz y col., 2018; Samoggia, Riedel y Ruggeri, 2020; Monachesi, 2020; Arthur y Williams, 2019; Ren, Jiang y Seipel, 2019). Dado que la API de Twitter evita que se acceda a los tweets publicados después de siete días, una herramienta de código abierto llamada Twint [1] se usó, ya que esta permite el acceso ilimitado a cualquier tweet publicado y recopila tweets de acuerdo con el nombre de usuario o la palabra clave.

Se recopilaron tweets de publicaciones en español e inglés, y se identificaron dos cuentas principales para los tweets en español: Granadaturismo, que es la cuenta oficial de Twitter del Departamento de Turismo de Granada, con el hashtag #teenseñomigranada en su biografía; y Alhambracultura, que es la cuenta oficial de Twitter del Patronato de la Alhambra y Generalife, con el hashtag #alhambra. Para obtener todos los tweets españoles sobre turismo en Granada, se consideraron los siguientes criterios de búsqueda: *granada turismo*, *gastronomia Granada*, *hoteles Granada* y *restaurantes Granada*. Otro criterio que también se tuvo en cuenta fue *¿Qué hacer en Granada?* identificado con el hashtag #planesgranada. Para obtener criterios de búsqueda adicionales, utilizamos la herramienta Twint para buscar tweets sobre *granadaturismo*. Usamos estos tweets para calcular las frecuencias de los hashtags y luego analizamos manualmente los cuarenta primeros. Entonces fue posible encontrar #albaicín[2], #sierranevada[3] y #welovegranada (que se utilizará para tweets escritos en inglés). Una vez identificados todos los criterios de búsqueda, se utiliza cada uno para extraer los tweets con la herramienta Twint.

En resumen, se utilizaron los siguientes criterios de búsqueda para encontrar tweets en español: *granadaturismo*, *teenseñomigranada*, *alhambracultura*, *#alhambra*, *granada turismo*, *gastronomia granada*, *hoteles granada*, *restaurantes granada*, *#planesgranada*, *#albaicín*, *#sierranevada*. Los términos usados para buscar información en inglés son: *#welovegranada*, *#granadatrip*, *#granadatravel*.

[1] Twint es una herramienta avanzada de crawling para Twitter escrita en Python que permite obtener tweets de Twitter sin su API, se puede obtener más información en la página de Github `https://github.com/twintproject/twint`

[2] una zona de Granada que es Patrimonio de la Humanidad

[3] una cadena montañosa cerca de Granada con una estación de esquí y las montañas más altas de la Península Ibérica.

Cada criterio de búsqueda se ejecuta por separado y los tweets capturados se almacenan en un archivo .csv. Luego se lee cada archivo y los datos del tweet se almacenan en una base de datos MySQL global. Los datos se centran en la ciudad de Granada en España, y dado que existen otras ciudades llamadas Granada, utilizamos exclusiones en las consultas de búsqueda de Twitter que funcionan con la API de Twitter. Por lo tanto, colocamos el caracter (-) antes de la palabra que queremos excluir: por ejemplo, podemos usar el criterio de búsqueda *granadaturismo -nicaragua -colombia -california*. Las palabras *Nicaragua*, *Colombia* y *California* están excluidas porque hay otras ciudades llamadas Granada en Colombia y Nicaragua, y *California* está excluida porque hay una cadena montañosa en el oeste de Estados Unidos entre el Valle Central de California y «Great Basin» llamada *Sierra Nevada*, que tiene el mismo nombre que la cordillera cercana a Granada. Estos criterios de exclusión se utilizaron con cada uno de los criterios de búsqueda. Aunque se tuvieron en cuenta criterios de búsqueda para encontrar tweets escritos en español e inglés utilizando criterios de búsqueda como granadaturismo, alhambracultura o #alhambra, también se capturaron tweets en muchos otros idiomas, como muestra la Tabla 3.1.

La herramienta Twint permite obtener la mayoría de los tweets que se han publicado desde 2008. Así, se recopilaron tweets desde 2008 hasta julio de 2020. La ventaja de esta herramienta es que no tiene restricciones de fecha para obtener los datos como tiene la API de Twitter.

3.2.2. Análisis descriptivo, limpieza de texto y procesamiento de datos

Una vez recolectadas las publicaciones, se procesaron con las siguientes operaciones: limpieza de texto, análisis descriptivo, tokenización, eliminación de stop-word y lematización.

1. **Limpieza de texto**: Debido a la heterogeneidad de la información, este proceso es un verdadero desafío por las siguientes razones: ciertos tweets comparten fotos o videos y solo contienen hashtags o emojis en el texto; muchos comentarios solo mencionan a otros usuarios, comparten fotos o videos y no contienen texto; algunas publicaciones contienen sorteos, promoción de productos o noticias sobre la ciudad; la mayoría de los usuarios no comparten su ubicación; algunos tweets usan emojis para expresar opiniones en los comentarios y publicaciones; y algunos tweets contienen preguntas para otros usuarios. Se excluyen los tweets que solo contienen

hashtags o emojis y no tienen texto o solo mencionan a los usuarios. Además, se analizó la ubicación geográfica para aquellos tweets en los que estaba disponible. Finalmente, se eliminaron los signos de puntuación, los usuarios y los enlaces, al igual que el término *RT* para los retweets, ya que cuando una persona retwetea, publica nuevamente el tweet original[4].

2. **Análisis descriptivo**: incluye métricas de publicación (frecuencia de hashtags) y se busca todos los tweets con una ubicación para identificar los lugares más visitados. En cuanto a métricas de usuario, se obtiene la frecuencia de publicaciones, retweets, comentarios, seguidores, videos y fotos únicos de cada usuario, y esto nos da una idea inicial de la influencia social del usuario.

3. **Tokenización**: Primero se identifican los idiomas presentes en el conjunto de datos y luego se divide cada tweet en palabras clave según el idioma. En esta tesis se toman tweets escritos en español e inglés porque estos idiomas tienen la mayor cantidad de información. Se incluye un idioma *und* porque el texto del tweet solo incluye hashtags, menciones de usuarios o enlaces, y no contiene ningún texto.

4. **Eliminación de palabras irrelevantes y lematización**: Se utiliza Natural Language Toolkit (NLTK) [5] para realizar este proceso ya que contiene stop-words tanto en español como en inglés. Para el proceso de «stemming» también se utiliza NLTK con un «stemmer» para español y otro para inglés, esto permite reducir las palabras a su base o raíz.

3.2.3. Análisis de contenido

El análisis de contenido se realizó en un contexto estacional y el año se dividió en dos períodos de primavera-verano y otoño-invierno. El análisis se centra principalmente en estructurar los datos a partir de textos e incluye los siguientes procesos:

1. **Frecuencia de palabras**: esto incluye un análisis de frecuencia de palabras por separado para español e inglés además de la nube de palabras para todos los tweets en un intento de identificar los lugares más comentados.

[4] Los retweets no se han eliminado porque muestran la aceptación del tweet original y por lo tanto, muestra aceptación de lo que se publicó en el tuit original

[5] NLTK es una plataforma líder para crear programas Python que funcionen con datos de lenguaje humano. Se puede encontrar más información en la página web `http://www.nltk.org`

2. **Análisis de sentimientos**: A pesar que en los siguientes capítulos se realiza un análisis más detallado de este tema, este proceso implica identificar y clasificar cada tweet según los sentimientos. Este análisis no considera el contexto estacional porque esta división no arrojó resultados relevantes, se realizó de manera global. Debido a la gran cantidad de tweets tanto en español como en inglés, se utilizan dos herramientas. La primera herramienta para clasificar tweets en español se llama Senti-py[6]. El paquete utiliza datos para entrenar al clasificador de varios sitios web como Tripadvisor y Twitter. La puntuación de polaridad es un valor flotante dentro del rango [0,1]: los valores menores a 0,3 son negativos, los valores mayores o iguales a 0,3 y menores o iguales a 0.6 son neutros y los valores mayores o iguales a 0,6 son positivos. La segunda herramienta para clasificar tweets en inglés se llama TextBlob[7]. La puntuación de polaridad es un valor flotante dentro del rango [-1.0, 1.0], donde los valores inferiores a 0 implican un sentimiento negativo, los valores superiores a cero representan un sentimiento positivo y un valor igual a 0 significa neutral. Esta información sobre sentimientos positivos o negativos es de suma importancia para los turistas a la hora de decidir si deben visitar algún lugar o no. Estos tweets también pueden ser utilizados por operadores turísticos, autoridades locales y residentes para mejorar la imagen del destino turístico. Este proceso se aborda ampliamente en los capítulos 4 y 5.

3. **Latent Dirichlet Allocation LDA**: para la extracción de temas. Para implementar estos algoritmos, se usó Python 3.7 con la biblioteca Gensim[8]. LDA intenta identificar los temas y lugares subyacentes más importantes para el destino turístico con el fin de descubrir las opciones disponibles para el turista según el período estacional definido, identificando lugares que pueden ser visitados tanto en primavera-verano como en otoño-invierno, o todo el año.

[6]Senti-py es un clasificador previamente entrenado en español que se basa en scikit-learn y NLTK. Se puede encontrar más información en la página de Github https://github.com/aylliote/senti-py

[7]TextBlob es una biblioteca de Python para procesar texto. Proporciona una API simple para tareas comunes de procesamiento de lenguaje natural (PNL por sus siglas en inglés), como «part-of-speech» o tipos de palabras que se pueden encontrar en una oración como sustantivos, verbos, pronombres, etc., extracción de frases nominales, clasificación de sentimientos, etc. Puede encontrar más información en la página web https://textblob.readthedocs.io/en/dev/

[8]Gensim es una biblioteca gratuita escrita en Python para el modelado de temas. Puede encontrar más información en la página web https://radimrehurek.com/gensim/index.html

3.3. Resultados del Análisis Descriptivo

El análisis descriptivo es un resumen que describe cuantitativamente las características de un conjunto de datos y nuestro trabajo considera a los usuarios que publican información sobre Granada. Ciertos atributos se consideran importantes como publicaciones únicas, número de retweets, me gusta, seguidores, fotografías, videos y comentarios. El análisis geográfico es posible si los usuarios han tuiteado con una ubicación geográfica activa.

Primeramente se obtienen la cantidad de tweets por lenguaje, número de publicaciones, comentarios y retweets como se puede observar en la tabla 3.1.

Language	**Unique posts**	**Comments**	**Retweets**
Spanish(es)	180215	22301	131
English (en)	24095	1316	4
Portuguese (pt)	10244	652	–
No text (und)	5336	1424	–
Catalan (ca)	4176	486	2
Dutch (nl)	3373	59	–
German (de)	1742	38	–
Italian (it)	1315	281	–
French (fr)	779	56	–
Indonesia (in)	655	27	–
Romanian (ro)	514	66	–
Albanian (sq)	475	29	–
Norwegian Nynorsk (nn)	464	43	–
Norwegian Bokmål (nb)	281	21	–
Pali (pl)	267	7	–
Japanese (ja)	221	3	–
Turkish (tr)	219	16	–
Arabic (ar)	180	16	–
Slovak (sk)	137	8	–
Shona (sn)	129	16	–
Finnish (fi)	117	14	–
Tagalog (tl)	113	8	–
Somali (so)	83	3	–
Slovenian (sl)	70	6	–
Haitian (ht)	68	7	–
Czech (cs)	63	1	–
Basque (eu)	57	10	–

Swedish (sv)	48	3	–
Russian (ru)	38	3	–
Korean (ko)	38	2	–
Danish (da)	34	2	–
Chinese (zh)	33	1	–
Lithuanian (lt)	25	4	–
Norwegian (no)	22	1	–
Thai (th)	21	–	–
Estonian (et)	19	–	–
Hungarian (hu)	19	3	–
Indonesian (id)	15	–	–
Bosnian (bs)	9	–	–
Latvian (lv)	9	–	–
Icelandic (is)	9	1	–
Greek, Modern (el)	8	–	–
Tsonga (ts)	8	–	–
Bulgarian (bg)	7	1	–
Hindi (hi)	6	–	–
Persian (fa)	6	–	–
Vietnamese (vi)	5	–	–
Malay (ms)	4	–	–
Urdu (ur)	4	–	–
Welsh (cy)	3	–	–
Hebrew (iw)	2	–	–
Galician (gl)	2	–	–
Swahili (sw)	1	–	–
Pashto (ps)	1	–	–
Belarusian (be)	1	–	–
Xhosa (xh)	1	–	–
Kazakh (kk)	1	–	–
Total	235787	26935	137

CUADRO 3.1: Dataset por Lenguaje

La tabla 3.1 revela que los usuarios de Twitter más frecuentes twittean en español e inglés, seguidos de tweets en portugués, catalán o sin texto (es decir, tweets que solo contienen hashtags, emojis o links). Sin embargo, la tabla también muestra que los tweets son publicados por usuarios en cualquier lugar de Europa y Asia, aunque en menor medida. El idioma *und* significa que un tweet

no tiene texto. Esta información proporciona un mayor contexto sobre las nacionalidades de los turistas que visitan Granada[9] Y esto permite a los gestores de viajes redefinir las estrategias de marketing.

La Tabla 3.2 muestra el número total de publicaciones, retweets y comentarios únicos para los diferentes hashtags o palabras clave.

Hashtag / Keyword	**Total**
alhambra	144036
alhambracultura	58372
granadaturismo	49014
sierranevada	26593
planesgranada	27116
“granada turismo”	42538
“gastronomia granada”	10052
“hoteles granada”	5283
welovegranada	3788
“restaurantes granada”	4459
Albaicín	733
granadatrip	167
Teenseñomigranada	142
granadatravel	153

Cuadro 3.2: Dataset de acuerdo a los Hashtags/Keyword

Los resultados presentados en la Tabla 3.2 muestran el número de tweets para cada criterio de búsqueda y, en consecuencia, el número de ocurrencias de cada criterio en el conjunto de datos completo. Esto no coincide con la cantidad de tweets existentes en nuestro conjunto de datos porque cada tweet puede hacer referencia a muchos hashtags/palabras clave/criterios en lugar de a un solo criterio de búsqueda. Un ejemplo de esto es el tweet: *“We invite everybody to visit #baza, its #history and the cultural and natural #heritage #welovegranada #greenwalkes #altiplanogranada @AytoBaza @granadaturismo @JAGGER0779 @turgranada @TurismoSpain ...”*, y dado que este tweet tiene el hashtag *welovegranada* y la palabra *granadaturismo*, se refiere a dos criterios de búsqueda.

[9]Sin embargo, existen ciertas limitaciones es decir, cuando diferentes países comparten el mismo idioma.

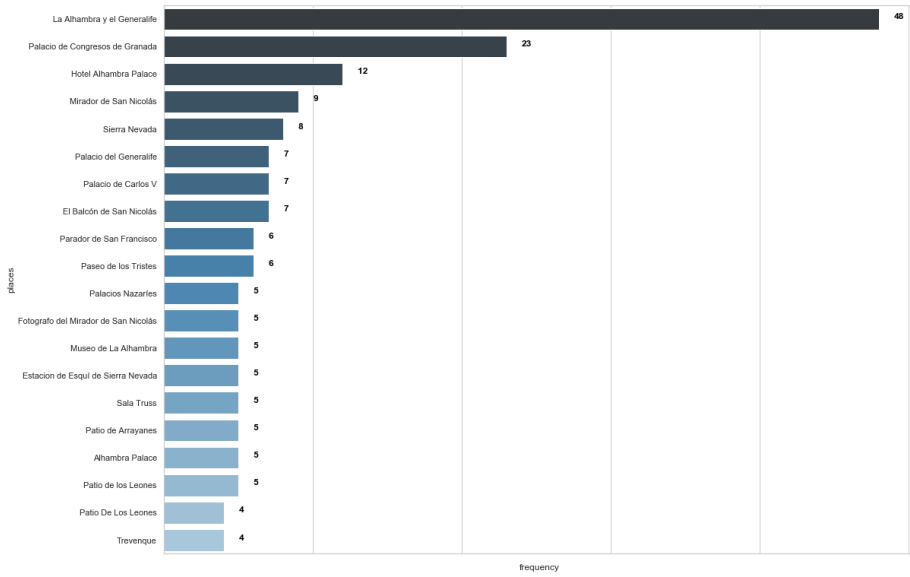

FIGURA 3.2: Tuits Geo-etiquetados (top 20)

Métricas de publicaciones

Se recopiló un total de 262,859 tweets que contenían 46,555 hashtags únicos. Los hashtags más comunes en el conjunto de datos son #Granada, #Alhambra, #SierraNevada, #Spain, #turismo, #España, #Andalucia, #travel, #planesgranada y #gastronomía. Está claro que los lugares más promocionados y también los más mencionados son Granada y Alhambra.

En el contexto del turismo, es importante conocer la ubicación de los tweets, estos se muestran en la Figura 3.2. De los 262.859 tweets, solo 432 tweets están asociados con una ubicación, es decir, tweets que han sido geoetiquetados.

La mayoría de los tuits geoetiquetados están en Granada. Según estos tuits, los lugares más visitados son *La Alhambra y el Generalife*, *Palacio de Congresos de Granada* (un centro de conferencias), *Sierra Nevada* y *Mirador de San Nicolás* (un famoso mirador situado en el Albaicín).

Métricas de usuario

Los 262,859 tweets fueron publicados por 66,409 usuarios únicos. Una vez que se identificó a los usuarios, se encontraron aquellos usuarios con publicaciones únicas, es decir, no eran retweets ni respuestas a otras publicaciones. Los usuarios más prolíficos son @alhambracultura, cuenta oficial del Patronato de la Alhambra y Generalife con 12,481 tuits; @granadaturismo, cuenta oficial de turismo de Granada con 12,135 tuits; @planesgranada, cuenta que ofrece información sobre ocio, cultura y gastronomía en Granada con 5,270 tuits; @granadaxmundo, cuenta que promociona la página web `granadaporelmundo.com` con blogs sobre Granada, videos y fotos con 2,744 tweets; y @makaralu, una bloguera local de Granada que promociona su ciudad en Twitter, con 1,652 publicaciones únicas. La figura 3.3 muestra los veinte usuarios principales que han publicado la mayor cantidad de tweets únicos. El análisis de los usuarios revela que la cuenta @alhambracultura es la más popular para el turismo en la región y la más activa en Twitter con mayor número de tuits en comparación con otras cuentas analizadas.

A partir de los datos de las publicaciones únicas, obtenemos la cantidad de retweets, me gusta y comentarios de cada publicación. Es posible, por tanto, encontrar aquellos usuarios que más retweets han recibido y estos son @alhambracultura con 107,362 retweets; @planesgranada con 97,985 retweets; @granadaturismo con 55,745 retweets; @spain, cuenta que promueve el turismo en España, con 16,596 retweets; y @ joselop44, un usuario residente en Granada, con 13,863 retweets. La tabla 3.3 muestra los veinte usuarios principales con más retweets, me gusta y comentarios en el conjunto de datos.

La popularidad en cualquier plataforma de redes sociales se mide por el número de seguidores por usuario, es decir, generalmente se cree que cuantos más seguidores tiene un usuario, más populares son. Como se muestra en la Figura 3.4, @ideal_granada, el periódico local con noticias sobre la ciudad y provincia de Granada, tiene el mayor número de seguidores, seguido de @viveandalucia, que es un canal oficial de promoción turística, y @alhambracultura.

Es muy fácil obtener el número de seguidores con la API de Twitter. Sin embargo, hay una serie de usuarios en el conjunto de datos que solo han comentado, publicado o retuiteado una vez sobre el turismo en Granada y, sin embargo, tienen millones de seguidores, y estos usuarios son de poca relevancia para los propósitos de este análisis. Para obtener resultados consistentes con este análisis, se consideran aquellos usuarios que han publicado tuits únicos sobre Granada tomando en cuenta un límite (es decir, usuarios con más de 500

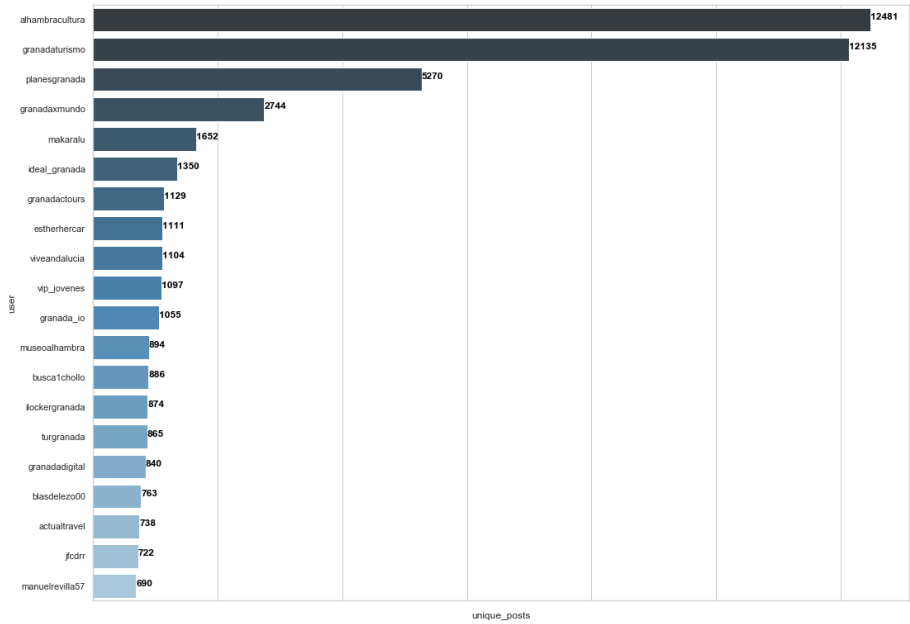

FIGURA 3.3: Total de publicaciones únicas de acueerdo a cada usuario (top 20)

user	**retweet_count**	**likes_count**	**comments_count**
alhambracultura	107362	194257	5445
planesgranada	97985	167283	4329
granadaturismo	55745	86821	2992
spain	16596	42839	918
joselop44	13863	23626	913
estherhercar	8080	27635	472
itsmalbert	8074	18856	123
yosoyesssa	12961	13886	793
manuelrevilla57	7050	16850	493
elhuffpost	5813	16369	591
ismaelquesada	9945	10917	430
ferminius	5490	14197	99
ideal_granada	6968	9434	542
aesthevic_	6592	8842	44
davidschez	4259	10998	78
tarandas61	2351	11393	286
viveandalucia	5344	7331	354
blasdelezo00	2389	9344	436
museoalhambra	3938	7403	228
rocio_diazj	3980	7125	327

Cuadro 3.3: Top 20 retuitts, likes y comentarios

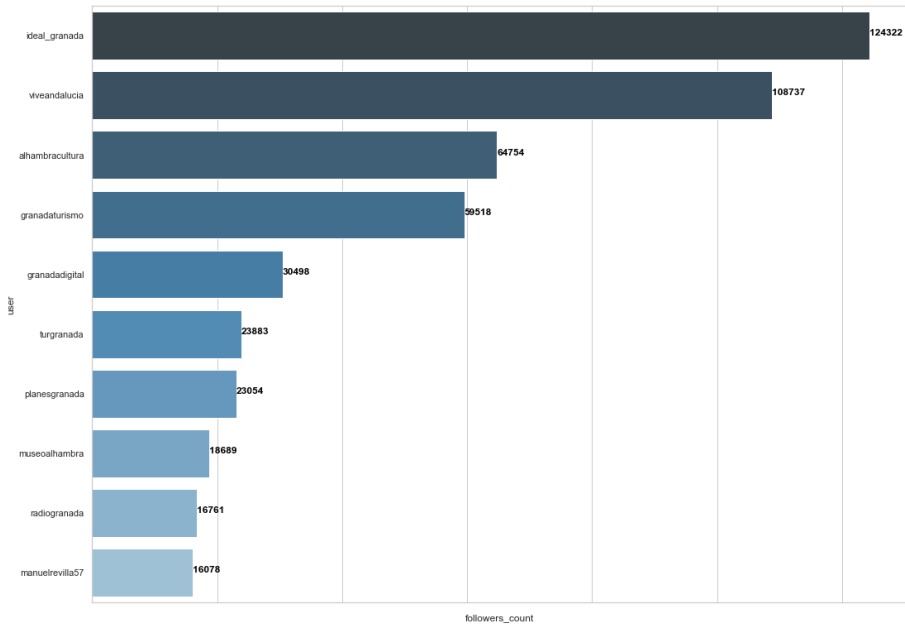

Figura 3.4: Seguidores por usuario (top 10)

publicaciones únicas). Este procedimiento se utilizó como filtro para asegurar que se obtengan los usuarios más representativos.

Las métricas multimedia también se calcularon examinando los tuits para detectar la presencia de fotografías o videos. En total de los 262,859 tweets, existen 123,809 fotos o videos, y de los 66,409 usuarios, solo 21,753 habían publicado contenido multimedia. Muchos usuarios solo publican imágenes de lugares que han visitado y, en general, los usuarios tienden a tomar fotos o videos de los lugares más importantes en los que han estado y proceden a publicarlos en las redes sociales. No hay duda de que las excelentes imágenes en las redes sociales atraerán, involucrarán y alentarán a sus amigos y seguidores a visitar estos destinos en el futuro. Un gran marketing visual también es extremadamente útil para los gerentes de turismo al establecer una conexión personal entre la marca y el cliente.

3.4. Resultados del Análisis de Contenido

El análisis de contenido o análisis de datos textuales consiste en determinar las variables de interés y la frecuencia de ocurrencia de las variables procesadas, y realizar una posterior reducción de la dimensión de los datos. En esta subsección, se calcula la frecuencia de las palabras, se realiza el análisis de sentimientos y la extracción de tópicos o temas (Joseph y col., 2017). La unidad para el análisis de texto usada es la «palabra».

3.4.1. Frecuencia de palabras

Para propósitos del análisis, la frecuencia de palabras se ha dividido en español (202,647 tweets) e inglés (25,415 tweets). Para cada grupo, el proceso de lematización se realiza por separado con la herramienta («stemmer») del idioma correspondiente. La tabla 3.4 muestra las palabras en inglés en el período primavera-verano, y la tabla 3.5 muestra las palabras en inglés en el período otoño-invierno. En ambos casos se puede ver que los lugares *La Alhambra* y *Sierra Nevada* se visitan durante todo el año (aunque hay más referencias a *Sierra Nevada* en otoño-invierno). En el caso de primavera-verano, sin embargo, los usuarios mencionan las vistas y los jardines, mientras que en el caso de otoño-invierno destacan la región de Andalucía, la nieve y el esquí. Como las palabras *hermoso* (beautiful) y *amor* (love) también aparecen en ambos períodos, podemos suponer que a los turistas les gustan mucho los miradores en diferentes puntos de la ciudad y diferentes lugares turísticos de la provincia.

Palabras	**Frecuencia**
granada (Granada)	11974
alhambra (Alhambra)	9817
spain (Spain)	4816
palac (palace)	2163
travel (travel)	1434
visit (visit)	1384
beauti (beautiful)	1226
sierranevada (SierraNevada)	1043
citi (city)	916
one (one)	788
day (day)	774
place (place)	713
view (view)	708
love (love)	688
andalucia (Andalucia)	624
garden (garden)	577
see (see)	561
andalusia (Andalusia)	543
go (go)	515
tour (tour)	501

CUADRO 3.4: Analisis de las frecuencias de palabras (Primavera-Verano) (inglés)

Words	**Frequency**
granada (Granada)	11132
alhambra (Alhambra)	8343
spain (Spain)	4507
palac (palace)	1709
sierranevada (SierraNevada)	1682
travel (travel)	1540
visit (visit)	1363
beauti (beautiful)	989
citi (city)	825
andalucia (Andalucia)	785
andalusia (Andalusia)	747
day (day)	738
one (one)	678
view (view)	655
love (love)	596
snow (snow)	595
place (place)	570
ski (ski)	524
tour (tour)	520
see (see)	511

CUADRO 3.5: Análisis de frecuencia de palabras (Otoño-Invierno) (inglés)

Words	**Frequency**
gran (Granada)	69752
alhambr (Alhambra)	38859
turism (turismo)	8212
sierranev (SierraNevada)	7250
visit (visita)	7179
espa (España)	6034
graci (gracias)	4612
ciud (ciudad)	4592
buen (bueno)	4317
hoy (hoy)	4206
hotel (hotel)	4071
albaicin (albaicin)	3992
fot (foto)	3905
viaj (viaje)	3683
gastronom (gastronomía)	3650
disfrut (disfruta)	3629
dia (dia)	3496
noch (noche)	3339
mejor (mejor)	3299
hac (hace)	3292

CUADRO 3.6: Análisis de frecuencia de palabras (Primavera-Verano) (español)

Words	**Frequency**
gran (Granada)	70431
alhambr (Alhambra)	35178
sierranev (SierraNevada)	11096
turism (turismo)	8424
visit (visita)	8386
espa (España)	6126
ciud (ciudad)	5626
buen (bueno)	4886
graci (gracias)	4536
gastronom (gastronomía)	4457
hoy (hoy)	4145
fot (foto)	4021
dia (dia)	3720
disfrut (disfruta)	3593
viaj (viaje)	3496
mejor (mejor)	3387
conoc (conoce)	3195
andaluc (Andalucia))	3184
albaicin (albaicin)	3141
restaur (restaurante)	3087

CUADRO 3.7: Análisis de frecuencia de palabras (Otoño-Invierno) (español)

FIGURA 3.5: Nube de palabras de los tweets

Las tablas 3.6 y 3.7 muestran las palabras en español para primavera-verano y otoño-invierno, respectivamente. Las palabras que destacan en los tuits en español también son *Granada*, *Alhambra* y *Sierra Nevada* tanto en los períodos primavera-verano como otoño-invierno, ya que son los lugares de la región más promocionados y visitados, y también se menciona el barrio *Albaicín*. El uso de palabras positivas como *bueno*, *gracias*, *disfruta*, *mejor* confirma aún más las opiniones positivas de los turistas. En las tablas que muestran las frecuencias de palabras de los tuits en español e inglés, la palabra completa se ha colocado entre paréntesis ya que el proceso de «stemming» se realiza antes del proceso de conteo.

La figura 3.5 muestra la nube de palabras sin ninguna división estacional y resalta las palabras que se repiten con más frecuencia en el conjunto de datos. Las palabras resaltadas son *Granada*, *Alhambra* y *sierranevada* y esto refuerza el hecho de que estos son los lugares turísticos más importantes de la región.

3.5. Resultados del Análisis de sentimientos y emociones

Numerosos estudios han realizado análisis de sentimiento en el campo del turismo. Uno de esos estudios es el de Ramanathan y Meyyappan 2019 que utiliza la ontología turística de Omán basada en ConceptNet. Los autores identifican las entidades (generalmente sustantivos) de cada tweet utilizando un etiquetador de «part-of-speech», y estos se comparan con conceptos en la ontología. El sentimiento se calcula para cada entidad utilizando su propio léxico basado en SentiStrength, SentiWordNet y «Opinion-Lexicon», que utilizan el análisis de sentimiento semántico conceptual para la extracción de entidades y AlchemyAPI para el mapeo de conceptos semánticos. Dado que los autores solo usan 4,432 tweets, uno de los problemas de este enfoque es la precisión en la identificación de entidades, y esto afecta significativamente la clasificación. En su artículo (Kirilenko y col., 2018), los autores comparan varios enfoques que se han utilizado en diferentes tipos de datos para determinar cuál obtiene los mejores resultados, y uno de estos conjuntos de datos incluye tweets. El enfoque que dio los mejores resultados fue el que utilizó SentiStrength. Sin embargo, recomiendan que los algoritmos de aprendizaje automático como SVM (Support Vector Machines) se utilicen con métodos basados en léxicos.

Se han analizado métodos de aprendizaje automático supervisados (Rao y col., 2017; Soong y col., 2019) y no supervisados (Aoudi y Malik, 2019) para clasificar las emociones. El enfoque o técnica de aprendizaje automático que ha obtenido los mejores resultados es el clasificador Multinomial Naive Bayes (Farisi, Sibaroni y Faraby, 2019). En esta sección, la herramienta TextBlob se utiliza para clasificar los tweets escritos en inglés y utiliza el modelo Multinomial Naive Bayes como método supervisado. Para los tweets en español, se usa la herramienta Senti-Py que está escrita en Python y que también usa un modelo Multinomial Naive Bayes como método. El modelo se alimenta con datos captados de varios sitios web que incluyen Tripadvisor y Twitter.

El principal objetivo del análisis de sentimientos en esta sección es clasificar el texto de acuerdo con tres categorías: positivo, negativo y neutro.

Para los tweets escritos en inglés, el análisis muestra (ver Figura 3.6) que del total de 25,415 tweets (publicaciones únicas, comentarios y retweets), 49.55 por ciento son neutrales, 44.01 por ciento son positivos y solo 6.43 % son negativos. También se revela que, si bien la mayoría de los tweets se refieren a promover o proporcionar información sobre Granada, otros tweets hablan favorablemente

de lugares de interés específicos de la ciudad y también de la ciudad misma. Los siguientes ejemplos de tweets expresan sentimientos positivos y neutrales:

1. Granada is an amazing city with jewels like this monastery http://bit.ly/jeron #ttot #travel @granadaturismo *[Granada es una ciudad increíble con joyas como este monasterio http://bit.ly/jeron #ttot #travel @granadaturismo]* (Positivo)

2. I love snow :D #snow #cool #great #mountain #cold #jmgt #granada #sierranevada http://instagram.com/p/jpFlLsKRaf/ *[Amo la nieve :D #snow #cool #great #mountain #cold #jmgt #granada #sierranevada http://instagram.com/p/jpFlLsKRaf/]* (Positivo)

3. Skiing in the #SierraNevada in #Granada !!! All you need to know before planning your trip http://ow.ly/8PEx7 *[Esquiar en la #SierraNevada en #Granada !!! Todo lo que necesita saber antes de planificar su viaje http://ow.ly/8PEx7]* (Neutro)

4. We remind you that the shops will be open tomorrow in #Granada! do your last-minute shopping in the city! *[¡Te recordamos que las tiendas estarán abiertas mañana en #Granada! ¡haz tus compras de última hora en la ciudad!]* (Neutro)

Sin embargo, un pequeño porcentaje de los tweets son negativos sobre la región y estos incluyen tweets negativos sobre temas de actualidad y visitas a la Alhambra:

1. First bad news: the Rosario and the Nazareno processions won't take place #SSantaGr

2. Alhambra visits are NOT well organized, long wait even with tickets bought in advance. @alhambracultura #Andalucia pic.twitter.com/iXv4ScY2GR

El análisis de los tuits en español revela que el 58.70 % del total de 202,647 tuits (publicaciones únicas, comentarios y retweets) son positivos, el 21.34 % son neutros y el 19.95 % son negativos, como se muestra en la Figura 3.7.

Los siguientes tweets son positivos y neutros:

1. Ahora a Granada lugar maravilloso por su historia, su gastronomía, su ambiente y por supuesto por el Casting de Somos Sur. (Positivo)

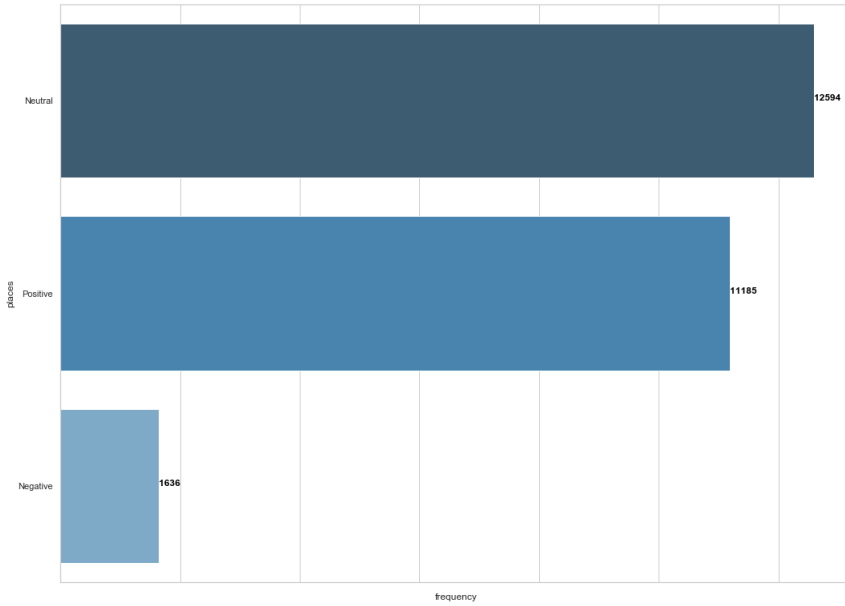

FIGURA 3.6: Distribución del sentimiento de los tuits (Inglés)

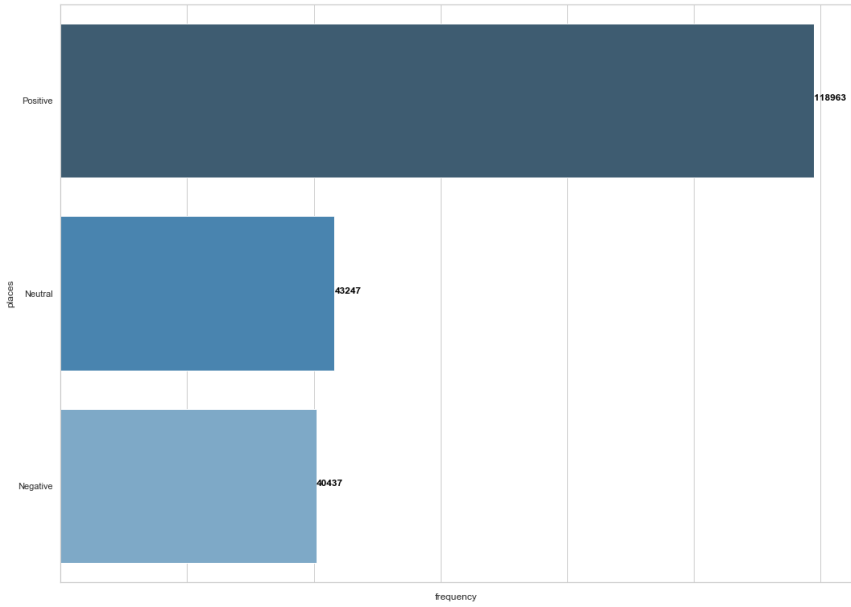

FIGURA 3.7: Distribución del sentimientos de los Tuits (Español)

2. Las sombras hacen que sierra nevada parezca estar sin nieve #Granada #SierraNevada (Neutro)

Los siguientes tweets son negativos:

1. En serio.....por que #granada esta tan oscura?? Señor Alcalde que no se ve ni la fachada de la catedral!!!!!!(Dicho por turistas) @granadaturismo

2. ÚLTIMA HORA: 2 fallecidos y 2 heridos, uno grave, al precipitarse un vehículo por un terraplén en la Carretera de #SierraNevada en #Granada

Cuando se dispone de suficientes datos, los lugares más populares (tweets positivos) y los lugares donde se han tenido experiencias negativas (tweets negativos) son mucho más evidentes. En este capítulo, no se considera las frecuencias de tweets geoetiquetados que han sido identificados como negativos o positivos ya que el número de estos es muy pequeño.

3.6. Resultados del Análisis temático

Muchos investigadores han utilizado «Latent Dirichlet Allocation (LDA)» para extraer temas de tweets en diferentes dominios (Yoosefi Nejad y col., 2020; Singh, Chauhan y Dhir, 2019; Abd-Alrazaq y col., 2020). Como explican los autores en su artículo (Blei, Ng y Jordan, 2003), LDA es un modelo de red bayesiana jerárquica de tres niveles, donde cada elemento de una colección se modela como una mezcla finita de un conjunto subyacente de temas/tópicos y el número de temas es un parámetro de entrada que debe fijarse. Cada tema define una distribución de probabilidad sobre las palabras clave, indicando la probabilidad de que cada palabra clave esté asociada a ese tema. En este caso, habiendo realizado una serie de experimentos previos, se eligieron cuatro temas para los tuits en español para el período primavera-verano, y tres temas para el período otoño-invierno. Se eligieron cuatro temas para los tuits en inglés para ambos períodos.

Las tablas 3.8 y 3.9 muestran las palabras clave más representativas y probables y sus probabilidades[10] asociadas a los diferentes temas de los tweets en español.

1. **Español (Primavera-Verano)**: En este caso, el hecho de que la Alhambra se mencione en casi todos los temas demuestra que la Alhambra es

[10]Por razones de claridad, se han incluido tanto palabras raíz como completas

Temas y Palabras clave
Tema 1 0.056*«alhambr(Alhambra)» + 0.047 «gran(Granada)» + 0.033 «visit(visita)» + 0.013 «monument(monumento)» + 0.010 «conoc(conoce)» + 0.007 «entrad(entrada)» + 0.007 «generalif(generalife)» + 0.007 «palaci(palacio)» + 0.007 «patrimoni(patrimonio)» + 0.006 «ciud(ciudad)»
Tema 2 0.070 «gran(granada)» + 0.025 «alhambr(Alhambra)» + 0.021 «viaj(viaje)» + 0.017 «sierranev(sierranevada)» + 0.009 «cuent(cuentame)» + 0.009*«pas(paseo)» + 0.009 «noch(noche)» + 0.008 «buen(bueno)» + 0.008 «sierr(sierra)» + 0.007 «nev(nevada)»
Tema 3 0.104 «gran(granada)» + 0.095 «alhambr(Alhambra)» + 0.027 «españ(España)» + 0.018 «ciud(ciudad)» + 0.014 «fot(foto)» + 0.013 «turism(turismo)» + 0.011 «albaicin(albaicin)» + 0.010 «maravill(maravilla)» + 0.008 «volv(volver)» + 0.008 «bonit(bonito)»
Tema 4: 0.083 «gran(granada)» + 0.041 «turism(turismo)» + 0.014 «restaur(restaurante)» + 0.011 «andaluci(Andalucía)» + 0.011 «sevill(sevilla)» + 0.010 «spain(spain)» + 0.010 «gastronom(gastronomia)» + 0.009 «andaluc(Andalucia)» + 0.009 «hotel(hotel)» + 0.009 «malag(Malaga)»

CUADRO 3.8: Temas y sus palabras clave en la estación Primavera-Verano (Español), k=4

Temas y Palabras clave
Tema 1 0.117 «gran(granada)» + 0.049 «alhambr(Alhambra)» + 0.034 «españ(España)» + 0.024 «sierranev(sierranevada)» + 0.012 «spain(spain)» + 0.011*«fot(foto)» + 0.011 «andaluci(andalucia)» + 0.011 «sierr(sierra)» + 0.011 «nev(nevada)» + 0.009 «oblig(obligado)»
Tema 2 0.076 «gran(granada)» + 0.066 «alhambr(Alhambra)» + 0.030 «visit(visita)» + 0.020*«ciud(ciudad)» + 0.013 «viaj(viaje)» + 0.013 «turism(turismo)» + 0.010 «monument(monumento)» + 0.010 «disfrut(disfruta)» + 0.009 «conoc(conoce)» + 0.008 «gui(guía)»
Tema 3 0.060 «gran(granada)» + 0.021 «alhambr(Alhambra)» + 0.017 «turism(turismo)» + 0.013 «gastronom(gastronomía)» + 0.010 «palaci(palacio)» + 0.009 «restaur(restaurante)» + 0.008 «sevill(Sevilla)» + 0.007 «tap(tapa)» + 0.007 «nuev(nuevo)» + 0.006 «granadin(granadino)»

CUADRO 3.9: Temas y sus palabras clave en la estación Otoño-Invierno (Español), k=3

sinónimo de la ciudad de Granada. El primer tema también considera los otros palacios colindantes como el Generalife, que fue la residencia real de verano en la época árabe; el segundo tema destaca las montañas de Sierra Nevada; el tercer tema trata sobre el Albaicín, el barrio árabe de Granada; y, finalmente, el cuarto tema es claramente gastronómico y visitas a otras ciudades famosas cercanas a Granada, como Sevilla y Málaga.

2. **Español(Otoño-Invierno)**: Una vez más, la Alhambra aparece en los tres temas (una visita a Granada sin visitar la Alhambra es casi inconcebible), con el primer tema centrado en Sierra Nevada, el segundo tema en visitas guiadas por la ciudad, y la tercera sobre gastronomía (las *tapas* son una tradición muy arraigada en Granada).

Los temas extraídos de los datos en inglés y las palabras clave se muestran en las tablas 3.10 y 3.11.

1. **Inglés (Primavera-Verano)**: En este caso, el primer tema se centra claramente en Sierra Nevada, mientras que los dos temas siguientes se centran en la Alhambra, el tercer tema se centra en la herencia árabe y el famoso Patio de los Leones. y el cuarto centrado tanto en la arquitectura árabe como en la gastronomía.

Temas y Palabras clave
Tema 1 0.036 «sierranevada(SierraNevada)» + 0.016 «sierra(sierra)» + 0.015 «nevada(nevada)» + 0.012 «mountain(mountain)» + 0.009 «snow(snow)» + 0.006 «ye(yes)» + 0.005 «ski(ski)» + 0.005 «hike(hike)» + 0.004 «alpujarra(Alpujarra)» + 0.004 «away(away)»
Tema 2 0.119 «granada(Granada)» + 0.104 «alhambra(Alhambra)» + 0.048 «spain(Spain)» + 0.019 «palac(palace)» + 0.015 «visit(visit)» + 0.015 «travel(travel)» + 0.012 «beauti(beautiful)» + 0.009 «citi(city)» + 0.008 «day(day)» + 0.008 «place(place)»
Tema 3 0.021 «palac(palace)» + 0.019 «spain(Spain)» + 0.016 «live(live)» + 0.009 «spectacular(spectacular)» + 0.009 «moorish(moorish)» + 0.008 «best(best)» + 0.008 «lion(lion)» + 0.008 «past(past)» + 0.008 «guid(guide)» + 0.007 «alhambra(Alhambra)»
Tema 4 0.017 «architectur(architecture)» + 0.015 «year(year)» + 0.014 «fortress(fortress)» + 0.009 «restaur(restaurant)» + 0.008 «locat(location)» + 0.008 «complex(complex)» + 0.008 «arab(arab)» + 0.008 «wonder(wonder)» + 0.007 «restaurant(restaurantes)» + 0.007 «granada(Granada)»

Cuadro 3.10: Temas y sus palabras clave de la estación Primavera-Verano (Inglés), k=4

Temas y Palabras clave
Tema 1 0.115 «granada(Granada)» + 0.100 «alhambra(Alhambra)» + 0.053 «spain(Spain)» + 0.022 «palac(palace)» + 0.016 «travel(travel)» + 0.015 «visit(visit)» + 0.012 «beauti(beautiful)» + 0.011 «citi(city)» + 0.010 «andalucia(Andalucia)» + 0.008 «andalusia(Andalusia)»
Tema 2 0.013 «wall(wall)» + 0.011 «arab(arab)» + 0.011 «great(great)» + 0.011 «find(find)» + 0.010 «spectacular(spectacular)» + 0.009 «guid(guide)» + 0.008 «moorish(moorish)» + 0.008 «live(live)» + 0.008 «citi(city)» + 0.007 «past(past)»
Tema 3 0.010 «restaur(restaurant)» + 0.009 «meet(meet)» + 0.007 «restaurant(restaurantes)» + 0.007 «design(design)» + 0.005 «food(food)» + 0.005 «entranc(entrance)» + 0.005 «learn(learn)» + 0.004 «calligraphi(calligraphic)» + 0.004 «cultur(culture)» + 0.004 «ok(ok)»
Tema 4 0.046 «sierranevada(SierraNevada)» + 0.021 «granada(Granada)» + 0.016 «sierra(sierra)» + 0.016 «nevada(nevada)» + 0.015 «ski(ski)» + 0.015 «snow(snow)» + 0.014 «mountain(mountain)» + 0.008 «winter(winter)» + 0.008 «snowboard(snowboard)» + 0.006 «locat(local)»

CUADRO 3.11: Temas y sus palabras clave de la estación Otoño-Invierno (Inglés), k = 4

2. **Inglés (Otoño-Invierno)**: En este caso, el primer tema se centra en las visitas a la Alhambra; el segundo tema se centra en la historia árabe y las visitas guiadas a la ciudad; el tercer tema se centra principalmente en la gastronomía; y el cuarto se basa en Sierra Nevada y su estación de esquí.

Una vez extraídos los temas relevantes se puede concluir que los aspectos más valorados de Granada tanto a nivel de turismo nacional e internacional como a lo largo del año son la Alhambra, Sierra Nevada (aunque para diferentes actividades en distintas épocas del año) y la gastronomía.

3.7. Discusión

En este capítulo se usó un framework para analizar el contenido generado por los viajeros mediante la identificación de palabras clave, la extracción de información a través de «crawlers» o «scrapers», limpieza de datos, análisis descriptivo y análisis de contenido. Para probar el framework, se recopilaron publicaciones de Twitter sobre la ciudad de Granada en España utilizando Twint para evitar las restricciones que Twitter tiene con su API. El uso de Twint permitió obtener un conjunto de datos completo de casi todos los tuits publicados sobre el turismo en Granada desde 2008. Los datos sin procesar se almacenaron en una base de datos relacional MySQL. Los datos se limpian posteriormente para eliminar cualquier caracter especial, URL, caracteres repetidos, etc. del texto, siendo este el proceso de análisis de datos más extenso e importante. Luego se identifican varios aspectos que son necesarios para el análisis descriptivo (por ejemplo, publicaciones únicas, comentarios, me gusta, retweets, contenido multimedia, etc.), los cuales brindan una idea clara de los usuarios más influyentes para los temas relevantes. Los datos se procesan posteriormente para realizar la tokenización y el proceso de «stemming», y para eliminar las palabras no relevantes o «stopwords». Durante la siguiente etapa, se calcula la frecuencia de palabras, se realiza un análisis de sentimientos y se clasifican las emociones. Finalmente, se lleva a cabo un análisis de los temas extraídos usando LDA para identificar los temas más importantes en el conjunto de datos.

Se recopilaron 262,859 tweets sobre Granada, que es un importante destino turístico en España. Los hashtags más importantes son #Alhambra y #SierraNevada, y el usuario más representativo es @AlhambraCultura. El análisis descriptivo proporciona una imagen amplia del turismo en Granada considerando a los usuarios que más publican sobre la ciudad. El enfoque propuesto utiliza un contexto estacional y los tweets publicados se dividen en dos períodos

(primavera-verano y otoño-invierno) para calcular la frecuencia de palabras y realizar la extracción de tópicos o temas, el análisis de sentimientos se excluye de este contexto estacional porque los resultados en este caso no son relevantes. Se calculó la frecuencia de palabras, y nuevamente Granada, Alhambra y España son las palabras más frecuentes en ambos períodos (primavera-verano y otoño-invierno) en inglés y en español, las palabras Granada, Alhambra y Turismo son las palabras más frecuentes en el período primavera-verano, y las palabras Granada, Alhambra y Sierra Nevada en el período otoño-invierno. En cuanto a los resultados de LDA para tweets en español, según la temporada primavera-verano los temas encontrados fueron: Alhambra y Generalife, Sierra Nevada, Albaicín y gastronomía; la temporada otoño-invierno muestra temas como: Alhambra, visitas guiadas a la ciudad y gastronomía. En cuanto a los resultados de los tuits en inglés, la temporada primavera-verano muestra temas como Sierra Nevada, Alhambra e historia árabe y arquitectura árabe y gastronomía; en la temporada otoño-invierno los temas que destacan son: Alhambra, historia árabe, gastronomía y Sierra Nevada y su estación de esquí. Los temas que muestra LDA con el conjunto de datos sobre Granada muestran los lugares más importantes que los turistas pueden visitar, como la Alhambra así como Sierra Nevada, que son lugares que se pueden visitar durante todo el año. Una de las ventajas de la provincia de Granada es la diversidad de lugares que posee y se puede disfrutar tanto en primavera-verano como en otoño-invierno, destacando la importancia de incluir el contexto estacional en el enfoque utilizado.

Este capítulo muestra como establecer un vínculo entre el turismo y la literatura de las redes sociales en el contexto del análisis de datos y contribuye a descubrir patrones y características del destino turístico a través de las redes sociales. Como se ha visto, existe mucha información relevante que se comparte en las redes sociales que refleja diferentes formas de interacción entre los usuarios antes de visitar un destino turístico. Esta información puede ser utilizada tanto por profesionales como por viajeros. Por otro lado, a través de este framework se puede ver la integración de entornos multidisciplinares que incluyen hoteles y turismo, tecnologías de la información, redes sociales, marketing y administración (Nusair, 2020) resaltando la integración de técnicas de análisis de datos y turismo digital para descubrir nuevas perspectivas para gestores turísticos y viajeros.

Un extenso análisis se ha realizado utilizando el framework. Este proceso tedioso puede ser automatizado a través de la construcción de una herramienta de análisis de datos turísticos de Twitter que muestre los resultados del análisis descriptivo (métricas de publicaciones y métricas de usuarios) y el análisis de contenido (análisis de sentimiento, frecuencia de palabras y análisis temático) en

este caso correspondiente a datos de Granada en España. Es posible realizar el análisis de cualquier destino turístico de forma sencilla, para que administradores y viajeros puedan analizar la información que existe sobre el destino turístico antes del viaje. Además, permite visualizar cuáles son los lugares más visitados y populares según la temporada del año, clasificar las experiencias positivas y negativas que han tenido los usuarios o viajeros al visitar la ciudad y verificar los usuarios más representativos que constantemente publican sobre el destino.

En este capítulo se muestra en detalle la forma en que se realiza el análisis de la información, por lo que puede resultar difícil para los viajeros comprender los procedimientos técnicos utilizados. Sin embargo, este enfoque permite resumir una gran cantidad de información para ser revisada por el turista antes de su viaje a través de una forma práctica y sencilla utilizando una herramienta web que implemente todas las técnicas utilizadas.

Es importante que los profesionales puedan realizar el análisis para un determinado destino turístico agregando las palabras que se utilizaron como criterios de búsqueda a la lista de palabras no relevantes o «stop-words». En el presente trabajo se ha probado en ambos sentidos y se ha decidido publicar los resultados en los que no se excluyen las palabras que son criterios de búsqueda debido a que, en el caso del análisis de frecuencias de palabras, basta con eliminar esas palabras de la lista; para LDA, si se eliminan las palabras que son los sitios más característicos y representativos de un destino turístico, los resultados serán poco claros y un poco confusos.

Para realizar el análisis de sentimiento existen muchas herramientas, entre ellas las pertenecientes a empresas como IBM o Microsoft que se puede realizar un análisis con mayor facilidad, sin embargo, las que se utilizaron en este capítulo son de código abierto; por tanto, toda la comunidad científica puede utilizarlas fácilmente. Las dos herramientas utilizadas tienen datos de entrenamiento diferentes, por lo que los resultados del análisis de sentimiento no se pueden comparar.

Los métodos utilizados para la detección del sentimiento en el texto de los tuits en español son deficientes. En este capítulo se ha utilizado Senti-py, debido a que se tiene un porcentaje de precisión superior a las demás herramientas que analizan texto en español, sin embargo, tiene una precisión de alrededor del 30 %. En el próximo capítulo se analizan las herramientas de clasificación de sentimientos tanto en español como en inglés y se mejora la clasificación en español usando métodos de aprendizaje profundo.

Capítulo 4

Análisis de sentimientos como herramienta para descubrir un destino turístico a través de datos de redes sociales

En el capítulo anterior se presentó un framework para el análisis de datos turísticos provenientes de redes sociales. Este enfoque utiliza el análisis de sentimientos como componente importante, para entender cómo percibe el turista un determinado destino turístico y realizar las mejoras necesarias por parte de los operadores y gestores turísticos. Sin embargo, las herramientas utilizadas no tienen una buena efectividad sobre todo con texto escrito en español. Por tanto, en el presente capítulo se analizan técnicas basadas en aprendizaje profundo aplicadas a la clasificación de sentimientos; posteriormente se utilizan las que tienen mejor efectividad con el objetivo de saber cuáles son los lugares, eventos o manifestaciones culturales mejor/peor valoradas y tratar de entender el porqué de las valoraciones negativas.

Según (Nakov y col., 2019), el objetivo del análisis de sentimientos es establecer si un texto expresa una opinión positiva, neutra o negativa sobre una entidad específica mencionada en el texto o en términos generales. Se han utilizado muchos métodos y modelos para detectar automáticamente el sentimiento en un texto, tanto en la industria como en la academia, ya que es difícil que un

modelo funcione con la misma precisión en diferentes dominios (política, marketing, marcas, turismo, etc.). En el capítulo anterior, las herramientas utilizadas no se enfocan precisamente al turismo, se las puede utilizar en cualquier dominio. En turismo, el análisis de sentimientos es muy importante para analizar las impresiones de los usuarios sobre los diferentes lugares de interés de un destino turístico específico, identificando cada lugar con publicaciones geolocalizadas (Paolanti y col., 2021). Esta información se puede recopilar de las redes sociales y permite a los profesionales descubrir qué lugares tienen más demanda, cuáles han sido evaluados negativamente y las razones de tales opiniones. Por tanto, el sentimiento representa una base importante para la toma de decisiones en materia de turismo.

La opinión pública sobre un producto o servicio es un recurso muy importante que utilizan los gestores para evaluarlo y descubrir qué falta, qué está mal o por qué los usuarios no están satisfechos al visitar un lugar o utilizar un servicio turístico. Dado que los métodos tradicionales como entrevistas y encuestas son muy costosos e ineficientes, los datos de redes sociales en forma de tweets (Twitter), publicaciones (Instagram, Facebook), foros, reseñas de viajes (TripAdvisor), etc. son, por tanto, muy utilizados para extraer la opinión pública sobre servicios o productos. Hay muchas técnicas basadas en el aprendizaje automático supervisadas y no supervisadas que se han desarrollado para identificar la polaridad del texto y, en los últimos años, se han desarrollado y utilizado métodos de aprendizaje profundo más precisos en la clasificación de texto (Liu y col., 2019b; Liu, 2015; Kirilenko y col., 2018).

En resumen, en el presente capítulo primeramente se analiza el desempeño de las herramientas de análisis de sentimiento para textos en inglés y español; segundo se propone un modelo de aprendizaje profundo para textos en español, que usa información de reseñas de los 30 destinos españoles más recomendados en TripAdvisor y tweets de un Taller de Análisis Semantico en la SEPLN (TASS)[1] (edición de 2019) para entrenar dicho modelo. Los datos de TripAdvisor utilizados se han compartido en una base de datos de acceso gratuito para que la comunidad pueda utilizarla para realizar más investigaciones; tercero, se selecciona la herramienta de mayor rendimiento para clasificar tweets y publicaciones de Instagram sobre la región de Granada relacionados con el turismo, con el fin de analizar cualquier información negativa, encontrar las deficiencias del destino turístico, identificar los hashtags que hacen referencia a un atractivo turístico, y las características importantes de los datos negativos que se puedan

[1]La página web del taller se puede encontrar en http://tass.sepln.org/

identificar mediante un análisis detallado del texto; y en cuarto lugar, comparar los resultados entre la información de Twitter e Instagram.

En el capítulo anterior se realizó una revisión de literatura de las técnicas utilizadas para analizar datos de redes sociales en el dominio del turismo en general. El análisis de investigaciones anteriores que se presenta a continuación se enfoca en los trabajos que se han publicado sobre análisis de sentimientos en general y en el dominio del turismo.

4.1. Trabajos relacionados

4.1.1. Análisis de Sentimientos

Según (Liu, 2012), «el análisis de sentimientos es el campo de estudio que analiza las opiniones, sentimientos, valoraciones, actitudes y emociones de las personas hacia entidades como productos, organizaciones de servicios, asuntos individuales, eventos, temas y sus atributos». En otras palabras, el análisis de sentimientos (AS), o más precisamente la minería de opiniones, consiste en identificar actitudes, estados de ánimo y emociones. Estos conceptos son muy importantes cuando se trata de comprender la psicología social de cómo un grupo o un individuo pueden modificar sus creencias, elecciones y percepciones del mundo (Liu, 2015). Un ejemplo de esto son los «influencers», que promocionan de forma persuasiva las marcas de la empresa en las redes sociales (Viñán-Ludeña y col., 2020).

De acuerdo con (Liu, 2012), el análisis de sentimientos se puede abordar desde tres niveles diferentes: en primer lugar, clasificar todo el documento para identificar si es positivo, negativo o neutro; en segundo lugar, clasificar las frases según sean positivas, negativas o neutras (sin opinión); y en tercer lugar, identificar no solo el sentimiento de la entidad sino también los aspectos relacionados con ella para facilitar una mejor comprensión del problema del análisis de sentimientos.

Los investigadores han desarrollado una serie de algoritmos para resolver el problema del análisis de sentimientos, y uno de ellos es el enfoque basado en el léxico que utiliza el contenido estadístico y semántico de diccionarios (Shi y col., 2019; Rosanensi y col., 2018). Este enfoque tiene, sin embargo, muchos problemas, y estos incluyen el hecho de que no considera el contexto de la oración, las palabras utilizadas como palabras de opinión pueden no expresar ningún sentimiento, resulta difícil lidiar con oraciones sarcásticas, oraciones sin palabras de opinión puede implicar un sentimiento positivo o negativo, etc. Otros

investigadores, por otro lado, utilizaron el enfoque de aprendizaje automático supervisado (clasificación de árbol de decisión, clasificación lineal con máquinas de vectores de soporte y redes neuronales, clasificación basada en reglas, clasificación probabilística con Naive Bayes, redes bayesianas y entropía máxima, etc.) (Gulnerman y Karaman, 2020; Mostafa, 2020; Chen y col., 2020).

En los últimos años, los investigadores se han centrado en modelos de aprendizaje profundo que mejoran la precisión de los modelos basados en aprendizaje automático. Se han desarrollado varias arquitecturas que superan a otros algoritmos (por ejemplo, redes neuronales convolucionales (CNN)[2], redes neuronales recurrentes (RNN)[3], memoria a largo o corto plazo (LSTM)[4], arquitecturas híbridas, etc.) (An, Kim y Moon, 2020; Li y col., 2019; Liu y col., 2019a; Yu y col., 2019; Gao y col., 2019; Martín y col., 2018). En el campo del turismo se analizarán algunos trabajos en los párrafos siguientes.

4.1.2. Análisis de sentimientos y turismo

Debido a la gran cantidad de datos que se generan a partir de las redes sociales (blogs, microblogs, reseñas, etc.), es necesario resumir estos datos para obtener información útil para los gestores turísticos y viajeros. En su artículo (Gu y col., 2018), los autores utilizaron el análisis de sentimientos para construir un sistema que consiste en analizar los datos de los blogs de turismo, resumir y visualizar su contenido, reduciendo así el tiempo dedicado a revisar blogs de viajes. Un análisis exploratorio se presenta en el artículo (Saura, Palos-Sanchez y Martin, 2018), en el que los autores utilizan datos de Twitter para identificar factores positivos, neutros y negativos que afectan la experiencia del usuario al visitar hoteles en España. El uso de la información de Twitter también permitió a los autores de su artículo (Cajachahua y Burga, 2017) realizar un análisis de correspondencia para comprender los sentimientos y opiniones sobre los tweets en inglés sobre el turismo en Perú. En tiempos de recesión, es importante analizar la imagen del destino, por lo tanto, en su trabajo (Gkritzali, Gritzalis y Stavrou, 2018), los autores utilizaron el análisis de sentimientos para analizar los mensajes del foro de TripAdvisor publicados sobre Atenas con el fin de comprender la imagen del destino.

La analítica de big data se ha convertido en un campo importante para encontrar patrones y analizar la calidad del servicio mediante el análisis de

[2]Convolutional neural network
[3]Recurrent neural network
[4]Long short-term memory

sentimientos (Serna y col., 2018; Fuentes-Medina, Hernández-Estárico y Morini-Marrero, 2018; Alcoba y col., 2017). Los autores (Martinez-Torres y Toral, 2019) se centraron en la industria hotelera para identificar reseñas positivas y negativas, engañosas y no engañosas a través de atributos únicos orientados a la polaridad, para evitar estafadores que intentan manipular a los consumidores publicando reseñas falsas. Sus métricas mostraron que la máquina de vectores de soporte (SVM)[5] es el mejor clasificador para las tres opciones para bag-of-words o bolsa de palabras (todas las palabras, atributos únicos y atributos únicos orientados a la polaridad).

En turismo, el análisis de sentimiento (AS) es muy importante en el apoyo a las decisiones (Liang, Liu y Wang, 2019; Nave, Rita y Guerreiro, 2018; Corallo y col., 2018; Micera y Crispino, 2017), por lo que en su artículo (Chen y col., 2020), los autores transforman AS en un problema de clasificación múltiple basado en métodos de aprendizaje automático y proponen un método híbrido que combine la representación de texto basada en Word2Vec y un mapa de conocimiento. Probaron tres conjuntos de datos de Tripadvisor y los resultados mejoraron con respecto a los de métodos similares. Su principal objetivo era crear un modelo de clasificación de sentimientos eficaz para las reseñas de viajes en línea. Chen y col. presentaron un modelo de detección de emociones con datos turísticos (Chen, Fan y Fu, 2019) y propusieron un clasificador de sentimientos que combine el léxico de los sentimientos y los métodos de aprendizaje automático basados en microblogs y datos de reseñas de viajes. Otro enfoque basado en el análisis de sentimientos y la agrupación semántica es presentado por (Abbasi-Moud, Vahdat-Nejad y Mansoor, 2019) para extraer las preferencias del usuario en un sistema basado en smart-tourism. En su artículo (Tao y col., 2019), los autores analizan las percepciones de los turistas con respecto a la calidad del aire para estudiar la satisfacción de la experiencia del turista, usaron el análisis de sentimientos con un enfoque de aprendizaje automático semi-supervisado con datos de Sina Weibo. Mientras que otros autores analizan las inconsistencias entre las calificaciones de los usuarios de TripAdvisor y el sentimiento de todas las revisiones (Valdivia y col., 2019), otro trabajo muestra que el uso de bi-gramas junto con la técnica SVM aumenta la precisión de la clasificación del sentimiento (Laoh, Surjandari y Prabaningtyas, 2019). De la misma manera, en su artículo (Athuraliya y Farook, 2018), los autores utilizan SVM para analizar las reseñas de los hoteles e identificar cualquier problema relacionado con el mantenimiento y la limpieza del hotel para que los gerentes del hotel puedan ser notificados y se puedan tomar las medidas correctivas

[5]Support Vector Machine

necesarias.

Muchos estudios se basan en la experiencia y satisfacción de los huéspedes del hotel (Imane y Abdelouahab, 2019; Prameswari, Surjandari y Laoh, 2017) para medir la calidad de los servicios del hotel. En su artículo (Rus y col., 2019), los autores midieron la calidad del hotel mediante el análisis de la percepción de otros servicios del hotel, como la habitación, las instalaciones, el entorno, el personal y la fiabilidad de la información publicada, utilizando Naive Bayes, SVM y algoritmos de los K vecinos más cercanos. Los autores (Fu y col., 2019) utilizan una perspectiva de meta-aprendizaje para realizar análisis de sentimientos. Mientras tanto, en su artículo (Ramanathan y Meyyappan, 2019), los autores utilizan una ontología basada en ConceptNet para extraer entidades y luego calcular el sentimiento de cada una utilizando el enfoque del léxico de sentimiento para obtener comentarios de la gente sobre el turismo en Omán. Otro enfoque innovador se presenta en el documento de (Uchinaka, Yoganathan y Osburg, 2019) donde los autores exploran el papel desempeñado por los residentes como embajadores del destino turístico, usando tweets sobre Onomichi (Japón), y sus hallazgos pueden ser la clave para sostener el turismo en vista de los crecientes sentimientos anti-turísticos. Otro estudio interesante aplicó AS para evaluar las reseñas en línea de los viajeros, mediante el cual los autores agruparon los datos en categorías y luego analizaron la calidad del servicio para revelar por qué los turistas albergan sentimientos negativos sobre cada categoría (Kim y col., 2017a).

Varios estudios utilizan el análisis de sentimientos para pronosticar las llegadas de turistas (Önder, Gunter y Scharl, 2019; Dragouni y col., 2016). El interesante estudio de Starosta, Budz y Krutwig, 2019 analizó los datos de las redes sociales sobre la inestabilidad política y económica y la agitación para mejorar el pronóstico de la llegada de turistas, utilizando datos de las redes sociales para realizar análisis de sentimientos a través de una red neuronal artificial (ANN)[6]. Las redes neuronales también fueron utilizadas por (Yadav y Bhojane, 2019), ellos emplearon tres enfoques de redes neuronales utilizando palabras preclasificadas, Hindi-SentiWordNet y oraciones preclasificadas, y descubrieron que los enfoques de redes neuronales tienen la mayor precisión en la clasificación de sentimientos. En el artículo de (Wang, Chiang y Sun, 2019) se propone un estudio útil para los viajeros que desean recuperar y resumir información útil de las reseñas de hoteles, donde los autores utilizaron léxicos múltiples para analizar características semánticas para realizar análisis de sentimientos. En su artículo (Moro y col., 2019), los autores muestran que existe una relación entre

[6] Artificial Neural Network

las características de gamificación y el comportamiento de los viajeros al escribir reseñas, los autores utilizaron análisis de sensibilidad para comprender cómo las características contribuyeron a explicar la puntuación de sentimiento. Otros investigadores utilizan la narración como una poderosa estrategia de promoción de destinos al analizar la estructura emocional de las historias en línea a través de AS (Zhang, Choe y Fesenmaier, 2019; Zhang y col., 2019; Zhang y Fesenmaier, 2018). Autores como (Rendalkar y Chandankhede, 2018) utilizaron la minería de opiniones para identificar el sarcasmo. (Yang y Chao, 2018) proponen una anotación de sentimiento a nivel de oración para reducir la sobrecarga de información al leer reseñas, y los autores utilizaron AS basado en palabras clave para realizar su estudio.

La calidad del sueño es otra característica que incide profundamente en la experiencia turística debido a que cuanto más descansado estás, más probabilidades tienes de disfrutar tu viaje. En su artículo (Mao, Yang y Wang, 2018), los autores proponen un framework para evaluar la calidad del sueño mediante el análisis de las reseñas de hoteles de TripAdvisor mediante AS. Mientras tanto, (Liu y col., 2018) define seis temas de turismo (comida, entretenimiento, compras, turismo, transporte y alojamiento) para definir el análisis de sentimiento temático, utilizando 53,140 elementos etiquetados manualmente, y no solo entrenaron el modelo de máxima entropía y la máquina de vectores de soporte, sino también mejoraron la precisión de los modelos adoptados mediante el uso de emoticones y palabras temáticas para complementar las características tradicionales de las palabras. El análisis espacial y espacio-temporal también es importante para identificar las percepciones de los viajeros, y en su artículo (Padilla y col., 2018), los autores aplicaron un análisis temporal y de múltiples ubicaciones utilizando datos de Twitter para analizar las emociones de los turistas en la ciudad de Chicago. Los investigadores (Yan, Zhou y Wu, 2018) analizaron la experiencia de viaje de los turistas a través de AS de las reseñas entre dos plataformas, y encontraron que las reseñas de viajes positivas son más frecuentes que las negativas.

Mientras que algunos enfoques usaron modelos de aprendizaje profundo para mejorar el análisis de sentimientos (Paolanti y col., 2021; Feizollah y col., 2019; Martín y col., 2018), otros usaron AS para construir sistemas de recomendación (Wang, Bao y Xu, 2019; Bueno y col., 2019; Zheng y col., 2018; Chaudhari, Kshirsagar y Nagori, 2018; Situmorang y col., 2018; Pronoza, Yagunova y Volskaya, 2016). Otros autores usan la ubicación con la minería de opiniones para comprender las percepciones de los usuarios y la experiencia de los viajeros a través de reseñas de hoteles (Annisa, Surjandari y Zulkarnain, 2019; Tariyal, Goyal y Tantububay, 2018; Putri y Kusumaningrum, 2017; Nakamura, Okada

y Hashimoto, 2015; Palakvangsa-Na-Ayudhya y col., 2011), mientras que otros investigadores combinan AS y el resumen textual (Prameswari y col., 2017), y otros crean diferentes aplicaciones utilizando AS en el campo del turismo (Guevara y col., 2018; Baralla., Ibba. y Zenoni., 2017; Gede Suardika, 2016; Shanshan Gao, Jinxing Hao y Yu Fu, 2015). Además, en su artículo (Alaei, Becken y Stantic, 2019), los autores examinan diferentes enfoques de análisis de sentimientos para identificar el desempeño de cada uno utilizando diferentes conjuntos de datos. Por último, algunos autores probaron un conjunto de algoritmos de aprendizaje profundo con datos de Twitter relacionados con el turismo; como resultado, la combinación de redes neuronales convolucionales con memoria a corto plazo logró la máxima precisión (Feizollah y col., 2019; Gao y col., 2019; Starosta y col., 2019; Li y col., 2018; Starosta, Budz y Krutwig, 2018; Martín y col., 2018; Kuhamanee y col., 2017).

Si bien la información disponible en Twitter ha sido ampliamente utilizada por los investigadores, Instagram es otra importante plataforma de redes sociales donde las personas publican sus experiencias turísticas. Sin embargo, muy pocos artículos se han centrado en Instagram en el dominio del turismo, uno de ellos es el artículo de (Palazzo y col., 2021), por ejemplo, donde los autores analizan el turismo sostenible mediante el uso de hashtags; (Falk y Hagsten, 2021), que estudian el número de visitantes a los sitios del Patrimonio Mundial, y (Gunter y Önder, 2021), que utilizan fotos con etiquetas geográficas para identificar las diferencias entre lugares populares de Viena.

A continuación se presenta la metodología utilizada incluyendo la recolección de datos, clasificación de sentimientos y el análisis de datos turísticos.

4.2. Metodología

En esta sección, se presenta el enfoque utilizado para la extracción de datos, la clasificación de datos y la gestión del turismo. Se usa un crawler para recopilar datos de Twitter e Instagram. En la Figura 4.1, se resume dicho enfoque.

En primer lugar, la fase de extracción de datos se realiza desde las plataformas de redes sociales usando un crawler. Luego, se evalúan varios métodos o herramientas de análisis de sentimientos para elegir las herramientas con la mejor precisión para cada idioma. Finalmente, se analiza la información de polaridad obtenida de los datos de Twitter e Instagram con las herramientas elegidas para identificar lugares a través de hashtags y aspectos (sustantivos), y proponer recomendaciones a administradores y viajeros sobre el destino turístico. En las

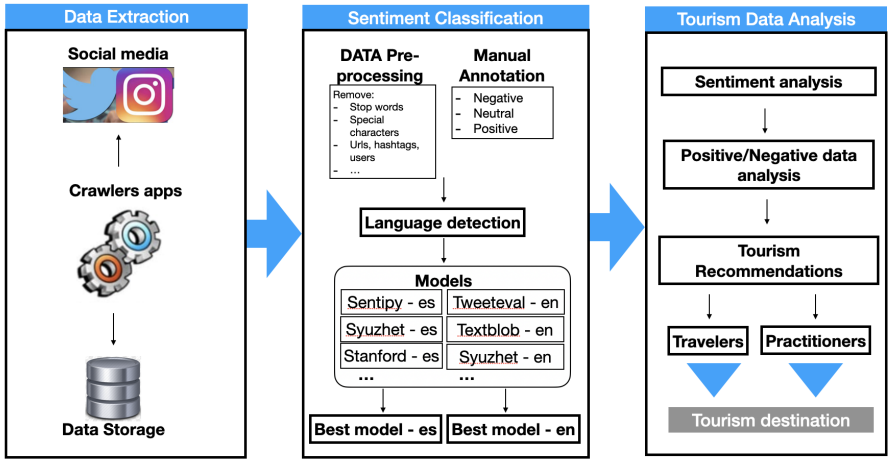

FIGURA 4.1: Framework para el análisis de sentimientos en turismo

siguientes subsecciones, cada etapa de este framework propuesto se describe en detalle.

4.2.1. Recolección de datos

Si bien se han desarrollado muchas aplicaciones y scripts para capturar datos de diferentes plataformas de redes sociales, también es posible utilizar la API de Twitter e Instagram. En el caso de Twitter, sin embargo, se recomienda el uso de una herramienta desarrollada en python llamada «Twint», que permite obtener tweets sin la API de Twitter (también se la utilizó en el capítulo anterior). Dado que Instagram tiene ciertas restricciones con respecto al uso de su API, se desarrolló un programa en Java que permite obtener datos utilizando la API y un proceso de programación de tareas para hacer frente a las restricciones diarias de acceso a los datos. Al igual que en el capítulo anterior se usarán datos de Granada, España y se consideran los idiomas español e inglés para el análisis. Las palabras clave elegidas para los datos en español tanto en Twitter como en Instagram son *granadaturismo*, *teenseñomigranada*, *alhambracultura*,

#alhambra, *granada turismo*, *turismogranada*, *gastronomia granada*, *gastronomiagranada*, *hoteles granada*, *hotelesgranada*, *granadahoteles*, *restaurantes granada*, *restaurantesgranada*, *granadarestaurantes*, *#planesgranada*, *#albaicín*, *#sierranevada*.

Las palabras clave elegidas para obtener datos en inglés en ambas plataformas son: *#welovegranada*, *#granadatrip*, *#granadatravel*, *granadatourism*, *granadatour*, *granadatours*, *granadatourisme*, *granadatourtravel*, *granadatravelcenter*, *granadatravels*, *travelgranada*, *granadatraveler*, *traveleringranada*, *sixsensestravelsgranada*, *granadatraveltips*, *triptogranada*, *thingstodoingranada*, *granadathingstodo*, *granadahotels*, *granadaluxuryhotels*, *cheaphotelsgranada*, *granadarestaurant*.

Hay que tener en cuenta que en el capítulo anterior se capturó datos de Twitter únicamente; sin embargo, no se tomó en cuenta algunos hashtags/palabras clave que se han puesto en esta sección. Esto sirvió para tratar de complementar el conjunto de datos de Twitter que se usará en este capítulo.

Además, se identificaron otros sitios turísticos que no estaban relacionados con Granada, España, y estos incluyen Granada Nicaragua, Granada Colombia o Sierra Nevada en California, Estados Unidos. Posteriormente se eliminaron todos los tweets o publicaciones relacionados con dichos sitios para eliminar el ruido, y las publicaciones o tweets que contengan las palabras «Nicaragua», «Caribe», «América del Sur», «Colombia», «Sudamérica», «California», «Estados Unidos» o «EEUU» también se eliminaron del conjunto de datos.

A pesar que se agregaron nuevas palabras clave o hastags para encontrar nuevos tweets, el dataset de Twitter que se utilizó en el capítulo anterior no varía significativamente con el dataset que se usa en este capítulo.

4.2.2. Clasificación de sentimientos

Los datos de Twitter se han elegido para etiquetar manualmente la polaridad de 1,000 tweets en inglés y 1,000 tweets en español, y así luego poder evaluar el comportamiento de las diferentes herramientas de AS con una muestra de datos de turismo sobre Granada. Los enfoques más utilizados en AS son el enfoque de aprendizaje supervisado y el enfoque basado en léxico. En este capítulo, se eligieron ambos enfoques para ambos idiomas, y se utilizaron las siguientes herramientas de AS para clasificar la polaridad en español e inglés:

1. **Senti-py**[7]: este es un clasificador de sentimientos para datos en español. El enfoque de aprendizaje supervisado, y más específicamente el modelo

[7]Esta herramienta se puede encontrar en https://github.com/aylliote/senti-py

multinomial Naive Bayes, se utiliza en esta herramienta y se entrenó con datos obtenidos de TripAdvisor, PedidosYa, Apestan, QuejasOnline, MercadoLibre, SensaCine, OpenCine, TASS y Twitter. Esta herramienta fue elegida debido a su popularidad entre la comunidad de Github.

2. **Stanford CoreNLP**[8]: En esta herramienta utiliza el enfoque basado en el léxico con el corpus Stanford Sentiment Treebank. El corpus se basa en el conjunto de datos recopilados de 11,855 oraciones individuales obtenidas de reseñas de películas (Socher y col., 2013), y también permite realizar análisis de sentimientos en diferentes idiomas, como inglés y español. Se usa esta herramienta para los datos en español.

3. **Syuzhet**[9]: este paquete R extrae el sentimiento del texto, y la herramienta utiliza un modelo basado en un léxico. El léxico Syuzhet utilizado en este paquete se desarrolló en el Laboratorio Literario de Nebraska, y los términos del diccionario se extrajeron de 165,000 oraciones codificadas por humanos de un corpus de novelas contemporáneas. Dado que esta herramienta es compatible con muchos idiomas (incluidos el español y el inglés), en este capítulo se la usa para ambos idiomas.

4. **Textblob**[10]: se utiliza un enfoque de aprendizaje supervisado en esta biblioteca de Python y un modelo Naive Bayes con datos de reseñas de películas[11] para el entrenamiento del modelo. Esta herramienta se utilizó para realizar análisis de sentimiento considerando los datos en inglés.

5. **TweetEval**[12]: este enfoque de aprendizaje supervisado se entrena con datos de Twitter. El conjunto de datos utilizado en este framework se recopiló con temas como *Donald Trump*, *iPhone*, *Aleppo*, *Palestina*, *refugiados sirios*, *Dakota Access Pipeline*, *medios occidentales*, *control de armas* y *vegetarianismo*. El modelo utilizado fue RoBERTa (un enfoque de preentrenamiento optimizado de Representaciones de codificador bidireccional

[8]Una herramienta para realizar el procesamiento del lenguaje natural: https://stanfordnlp.github.io/CoreNLP/sentiment.html

[9]Una herramienta basada en corpus para la detección de sentimiento: https://cran.r-project.org/web/packages/syuzhet/index.html

[10]Una biblioteca de Python para analizar datos textuales: https://textblob.readthedocs.io/en/dev/

[11]Página web de datos de reseñas de películas: http://www.cs.cornell.edu/people/pabo/movie-review-data/

[12]Repositorio de Tweetval: https://github.com/cardiffnlp/tweeteval

de Transformers (BERT)[13]) que admite la clasificación de tweets de varias clases. Es compatible con los datos en inglés (Barbieri y col., 2020).

Las tres categorías: positivo, negativo y neutro se eligieron para la clasificación de sentimientos. Las métricas que se analizarán son exactitud, precisión, recall y la puntuación F1 (Lecun, Bengio e Hinton, 2015):

1. **Exactitud**: la cantidad de documentos (tweets o publicaciones) que se clasifican correctamente (como positivos, negativos o neutrales) dividido por la cantidad total de documentos clasificados.

2. **Precisión**$_C$: para cada categoría C (positiva, negativa o neutral), este es el número de documentos correctamente clasificados como C dividido por el número de documentos clasificados en esta categoría.

3. **Recall**$_C$: para cada categoría C, este es el número de documentos correctamente clasificados como C dividido por el número de documentos que realmente pertenecen a esta categoría.

4. **F1**$_C$: para cada categoría C, esta métrica permite comparar el rendimiento combinado de precisión y recall a través de su media armónica:

$$F1_C = 2 * \frac{precision_C * recall_C}{precision_C + recall_C} \tag{4.1}$$

Cabe señalar que, para calcular un valor único para cada métrica de precisión, recall y F1, se promedian sus valores en cada categoría C. Una vez calculadas estas métricas, se procede a analizar y evaluar la mejor herramienta que se utilizará en la siguiente fase.

4.2.3. Análisis de datos turísticos

La parte final de este enfoque utiliza el mejor método obtenido en la sección anterior para obtener la polaridad de los tweets y publicaciones, y luego relacionar esta polaridad con los hashtags que aparecen en ellos referidos a diferentes entidades (por ejemplo, lugares específicos dentro del destino turístico, o conceptos como gastronomía). Una parte importante es identificar aquellos tweets y publicaciones que se consideran negativos, para así poder agrupar los hashtags característicos de un lugar e identificar los problemas que los usuarios

[13] Bidirectional Encoders Representations from Transformers

expresan en las redes sociales sobre este lugar. Esta información es importante para que los administradores de servicios puedan mejorarlos y los viajeros puedan recopilar información precisa sobre el destino turístico.

4.3. Resultados

En esta sección se presentan los resultados obtenidos después de utilizar los analizadores de sentimiento mencionados en la sección anterior. Los datos se recopilaron de Twitter e Instagram; dado que ambas plataformas de redes sociales son similares, es posible analizar los textos publicados de la misma manera. De las 90,725 publicaciones de Instagram recopiladas, 7,717 de ellas están escritas en inglés y 56,247 en español, y se descartaron las 26,761 publicaciones escritas en otros idiomas. También recopilamos 235,755 tweets, que incluyen 19,340 tweets en inglés y 144,947 tweets en español, y se descartaron los 71,468 tweets en otros idiomas. El idioma de los tweets y las publicaciones se identificó mediante gcld3 [14] (Google Compact Language Detector v3). Se usan datos de Twitter para elegir el mejor método de clasificación de sentimientos: 1,000 tweets en inglés y 1,000 tweets en español se clasificaron manualmente (como positivos, negativos o neutros), y luego cada clasificador se evaluó en este conjunto de datos. Las herramientas de análisis de sentimientos utilizadas en español fueron Sentipy, Syuzhet y CoreNLP. Los resultados se muestran en la tabla 4.1.

Modelo	**Exactitud**	**Precisión**	**Recall**	**F1**
CoreNLP (Stanford)	0.5720	0.3564	0.3958	0.3402
Syuzhet R	0.6050	0.3691	0.3906	0.3404
Sentipy	0.3590	0.4064	0.5521	0.3176

Cuadro 4.1: Resultados por cada modelo para el conjunto de datos de prueba de tweets en español

De los 1,000 tweets en español, 163 son positivos, 796 neutros y 41 negativos. Los resultados obtenidos son bastante modestos (con una exactitud inferior al 60 % y una puntuación de F1 en torno al 0,34). Aunque el clasificador Syuzhet tiene la mejor exactitud y puntaje F1, casi siempre tiende a predecir cada publicación como neutra con muy pocos aciertos en las clases negativas y positivas.

[14] una biblioteca de Python basada en redes neuronales para reconocer idiomas a partir del texto, https://pypi.org/project/gcld3/

La herramienta Sentipy, por otro lado, funciona mejor en cuanto a precisión y recall, ya que funciona mejor con clases positivas y negativas que son minoría. Sin embargo, como resultado de su mal desempeño con la clase neutra mayoritaria, este clasificador tiene una peor exactitud y puntuación F1.

Los clasificadores seleccionados para los datos en inglés fueron Tweeteval, Suyzhet y Textblob. Los resultados obtenidos se muestran en la Tabla 4.2

Modelo	Exactitud	Precisión	Recall	F1
Tweeteval	0.7560	0.6862	0.7660	0.7157
Textblob	0.5480	0.5296	0.6171	0.5128
Syuzhet R	0.4810	0.5325	0.6157	0.4807

Cuadro 4.2: Resultados por cada modelo para el conjunto de datos de prueba de tweets en inglés

De los 1,000 tweets en inglés, 109 son negativos, 658 neutros y 233 positivos. El clasificador que destaca con los tweets en inglés es Tweeteval, seguido de Textblob y Syuzhet. Estos resultados son significativamente mejores que los del español, lo que parece sugerir que existen mejores clasificadores de sentimientos en inglés que en español. Los resultados obtenidos por Tweeteval son buenos, con una exactitud superior al 75 % y una puntuación F1 superior a 0,7.

De este análisis preliminar, se puede concluir que Tweeteval es una herramienta útil para analizar los datos de tweets/publicaciones en inglés sobre el turismo en Granada, pero para obtener resultados más confiables para datos en español, es necesario mejorar los clasificadores en español. En la siguiente subsección, se estudian tres tipos de clasificadores que utilizan modelos de aprendizaje profundo.

4.3.1. Clasificadores de sentimientos basados en aprendizaje profundo para texto en español

Dataset de entrenamiento

Debido a los malos resultados obtenidos por los clasificadores anteriores para datos en español y debido a que no hay muchas alternativas disponibles (González y col., 2015; García, Gaines y Linaza, 2012), se propone utilizar clasificadores basados en tecnología de aprendizaje profundo con datos de turismo en español. Para ello, se identificaron los 30 lugares más visitados de España recomendados por TripAdvisor y se recopilaron reseñas sobre cada lugar mediante un script

de Python (30,805 opiniones en total). Cada reseña ha sido calificada por los usuarios de TripAdvisor en un rango entre 1 y 5 (siendo 5: Excelente, 4: Muy bueno, 3: Regular, 2: Malo y 1: Terrible), por lo tanto, cambiamos esta clasificación, es decir, las reseñas que tienen una calificación de 5 y 4 fueron etiquetadas como Positivas; las reseñas que han sido calificadas con 3 fueron etiquetadas como Neutras y las reseñas con calificaciones de 2 y 1 fueron etiquetadas como Negativas. Además, se etiquetaron manualmente 3,316 tweets sobre Granada (este conjunto de tweets no está incluido en el conjunto de datos propuesto en nuestro análisis); el proceso de etiquetado fue realizado por el autor de esta tesis y revisado por su director, y 4,800 tweets que fueron etiquetados en la edición 2019 del taller en español sobre Análisis de sentimientos[15]. Se obtuvo un total de 38,921 textos etiquetados, de los cuales 5,219 son negativos, 9,096 neutros y 24,606 son positivos, y estos fueron luego utilizados para entrenar los modelos de aprendizaje profundo.

Este conjunto de datos se ha puesto a disposición del público con fines de investigación y, a mi leal saber y entender, es el único conjunto de datos de turismo español (Tourism Spanish Dataset o TSD) disponible públicamente[16].

Arquitecturas de aprendizaje profundo

En el ámbito del turismo, Paolanti y col., 2021 realizaron una evaluación de aprendizaje profundo de diferentes modelos y descubrieron que se destaca una red basada en caracteres, los investigadores (Hao y col., 2021) analizaron las percepciones de los turistas chinos sobre Hong Kong a través del análisis de sentimientos basado en el aprendizaje profundo, (Chang, Ku y Chen, 2020) utilizó lingüística computacional, análisis visual y técnicas de aprendizaje profundo para analizar las reseñas de hoteles de TripAdvisor, y (Li y col., 2020a) propuso el modelo CNN-BiLSTM de dos canales integrado de léxico para realizar análisis de sentimientos.

Para comparar el rendimiento de la tecnología de aprendizaje profundo, hemos utilizado tres arquitecturas basadas en el libro de (Krohn, Beyleveld y Bassens, 2020) usando una biblioteca de Python llamada Keras[17]: memoria

[15]Página web del taller: `http://tass.sepln.org`

[16]Puede encontrar más información sobre el repositorio de Bitbucket en `https://bitbucket.org/msvl/tourismspanishdataset/src/master/`

[17]Esta API es la biblioteca de aprendizaje profundo líder y se puede encontrar más información en `https://keras.io`

bidireccional apilada a corto plazo y largo plazo[18] (BiLSTM), una red multi-convolucional y el modelo de representación de codificador bidireccional de transformers (BERT)[19].

Los clasificadores LSTM fueron propuestos por Hochreiter y Schmidhuber (Hochreiter y Schmidhuber, 1997) y se utilizan ampliamente en el procesamiento del lenguaje natural (NLP). Este tipo de modelo recibe entradas no solo de una secuencia de datos, sino también del punto de tiempo anterior en la secuencia de texto, y cada celda de una capa LSTM contiene estructuras más complejas. Si bien estos modelos se propagan hacia atrás a través de los pasos de tiempo, los LSTM bidireccionales se propagan hacia atrás y hacia adelante a través de los pasos de tiempo.

Se utiliza varias capas apiladas de la familia BiLSTM, y esta arquitectura ha producido resultados bastante aceptables al realizar la clasificación de texto. Hemos basado nuestra investigación en la arquitectura presentada en el libro de Khron y col. Luego implementamos varias modificaciones al modelo y sus hiperparámetros con la adición y eliminación de capas para mejorar los resultados. Sin embargo, la arquitectura final no es muy diferente de la presentada en el libro de Khron y col.

Uno de los modelos más comunes es el generado por convoluciones, y la red neuronal convolucional (Convnet o CNN) es una red neuronal artificial que presenta una o más capas convolucionales que permiten procesar eficientemente patrones espaciales (Lecun, Bengio e Hinton, 2015). Por tanto, una oración podría representarse como un vector multidimensional, con cada capa que tiene núcleos o filtros y cada uno se llama una ventana o «patch» que escanea el texto, y estas capas aprenden de la retro propagación. Hemos utilizado la arquitectura y los hiperparámetros presentados en el libro de Khron y col.

Uno de los últimos modelos de representación de lenguaje es (BERT), que utiliza un modelo de lenguaje que enmascara aleatoriamente varios tokens de la entrada y fusiona el contexto izquierdo y derecho, lo que permite un transformador bidireccional profundo (Devlin y col., 2019). Además, utiliza una arquitectura de aprendizaje profundo diseñada para pre-entrenar representaciones bidireccionales profundas a partir de texto sin etiquetar utilizando el contexto izquierdo y derecho en cada capa. Una vez que el modelo se ha pre-entrenado, se puede ajustar con una capa de salida adicional. En esta capa adicional, se usa el TSD para ajustar el modelo y el clasificador de sentimientos.

[18] Significa que este modelo contiene dos capas Bi-LSTM
[19] Bidirectional Encoder Representation from Transformers

Para aplicar este enfoque, se utilizan dos conjuntos de datos: uno para el entrenamiento previo y el otro para el entrenamiento en sí. Usamos BETO, un modelo BERT entrenado con un corpus español (Cañete y col., 2020) como el conjunto de datos previamente entrenado. El TSD se utilizó como el conjunto de datos de entrenamiento. El modelo previamente entrenado usa el mismo método de ajuste fino de clasificación utilizado por (Liu y col., 2019b).

La tabla 4.3 muestra las métricas de las tres arquitecturas analizadas en los 1,000 tweets en español utilizados anteriormente.

Modelo	**Exactitud**	**Precisión**	**Recall**	**F1**
Stacked Bi-LSTM	0.3400	0.3479	0.3500	0.2877
MultiConvnets	0.5450	0.3195	0.3274	0.3043
Spanish-BERT	0.7570	0.5635	0.6278	0.5875

CUADRO 4.3: Resultados de los modelos de aprendizaje profundo con el conjunto de datos de prueba de tweets en español

Con pocos datos de entrenamiento (aproximadamente 38,000), está claro que el modelo BERT para español funciona bien y se destaca no solo entre las otras arquitecturas de aprendizaje profundo analizadas sino también entre los modelos utilizados en la tabla 4.1.

4.3.2. Extracción de entidades y aspectos

Utilizando el modelo BERT para datos en español y Tweeteval para tweets o publicaciones en inglés, se ha obtenido el sentimiento de cada uno. Como los datos no cuentan con información de geolocalización, para observar la polaridad de cada entidad (lugar, evento, patrimonio, gastronomía o actividad turística) del destino turístico, se asoció la polaridad obtenida para cada tweet/post con los hashtags que aparecen en él. De esta forma, si un tweet que menciona, por ejemplo, tanto #alhambra como #generalife se clasifica como positivo, entonces esta polaridad también se asocia a los dos hashtags. Además, se agruparon diferentes hashtags que se refieren esencialmente a las mismas entidades. A modo de ejemplo, se utilizan una serie de hashtags diferentes para referirse a la Alhambra, el lugar turístico más importante de Granada: «#alhambra», «#laalhambra», «#alhambradegranada» y «#thealhambra». De esta manera,

se han identificado 38 entidades[20], que se muestran junto con las frecuencias de cada polaridad en la tabla 4.4 para los datos de Twitter.

[20]Solo consideramos aquellas entidades significativas con un número de menciones por encima de un umbral.

hashtag	**Neg**	**Neu**	**Pos**	**Total**	**%neg**	**%neu**	**%pos**
			Lugares				
salobreña	1	97	99	197	0.51	49.24	50.25
plazabibrambla	4	80	57	141	2.84	56.74	40.43
miradordesannicolás	4	131	88	223	1.79	58.74	39.46
catedral	0	109	68	177	0.00	61.58	38.42
almuñecar	9	463	274	746	1.21	62.06	36.73
valledelecrin	3	229	126	358	0.84	63.97	35.19
lanjaron	1	98	52	151	0.66	64.90	34.44
costatropical	2	236	101	339	0.59	69.62	29.79
alpujarra	11	682	285	978	1.12	69.73	29.14
alhambra	381	15292	6207	21880	1.74	69.89	28.37
generalife	11	565	225	801	1.37	70.54	28.09
granada	1209	43810	17261	62280	1.94	70.34	27.72
guadix	1	228	83	312	0.32	73.08	26.60
sierranevada	306	10807	3827	14940	2.05	72.34	25.62
patiodelosleones	3	199	66	268	1.12	74.25	24.63
sacromonte	9	490	162	661	1.36	74.13	24.51
monachil	20	224	76	746	6.25	70.00	23.75
albaicin	59	2661	838	3558	1.66	74.79	23.55
guejardelasierra	1	108	29	138	0.72	78.26	21.01
motril	3	200	53	256	1.17	78.13	20.70
realejo	3	187	43	233	1.29	80.26	18.45
palaciocarlosv	3	202	40	245	1.22	82.45	16.33
			Cultura				
aljibes	4	94	105	203	1.97	46.31	51.72
cultura	18	1204	413	801	1.10	73.64	25.26
lorca	11	257	87	355	3.09	72.39	24.51
flamenco	4	419	271	533	0.75	78.61	20.64
			Actividades - Eventos				
esquí	26	1147	459	1632	1.59	70.28	28.13
cruces-en-Granada	1	97	37	135	0.74	71.85	27.41
semanasanta	36	1103	256	1395	2.58	79.07	18.35
corpus	8	244	53	305	2.62	80	17.38
			Gastronomía				
restaurant	68	1601	836	2505	2.71	63.91	33.37
tapas	18	709	271	998	1.80	71.04	27.15
gastronomia	74	3708	1380	5162	1.43	71.83	26.73

CUADRO 4.4: Resultados del análisis de sentimentos para datos de Twitter (Inglés-Español)

Las entidades más conocidas o citadas de Granada (aunque no necesariamente las mejor valoradas) se pueden identificar como la «Alhambra», el «Albaicín» (dos patrimonios de la Humanidad), la «sierranevada» y «gastronomía». La tabla 4.4 muestra los lugares o entidades ordenados según el porcentaje positivo. En general, el número y porcentaje de tweets negativos es bastante reducido (especialmente en comparación con los tweets positivos), por lo que el grado de satisfacción con las entidades tiende a ser elevado.

Es importante señalar que aunque lugares como «salobreña», «aljibes», «plazabibrambla» o «mirador de san nicolas» no se mencionan con frecuencia en las redes sociales, los turistas que los visitan tienen una percepción bastante positiva de ellos. Por lo tanto, destinos como «salobreña», «plazabibrambla», «mirador de san nicolas», «catedral», «almuñecar», «valledelecrin», «lanjaron», «costatropical», «alpujarra» y «aljibes» pueden considerarse lugares relativamente infravalorados con un importante potencial turístico.

También identificamos las entidades para los datos de Instagram, y sus recuentos de frecuencia de polaridad se muestran en la tabla 4.5.

Hashtag	**Neg**	**Neu**	**Pos**	**Total**	**%neg**	**%neu**	**%pos**
			Lugares				
realejo	0	6	18	24	0	25	75
valledelecrin	0	28	60	88	0	31.82	68.18
almuñecar	0	10	20	30	0	33.33	66.67
miradordesannicolas	0	28	44	72	0	38.89	61.11
sacromonte	0	51	69	120	0	42.5	57.49
granada	28	8241	8334	16603	0.17	49.64	50.19
alpujarra	0	43	42	85	0	50.59	49.41
sierranevada	20	1263	1078	2361	0.85	53.49	45.66
monachil	0	11	9	30	0	55	45
albaicin	7	705	551	1263	0.55	55.82	43.63
cathedral	0	97	74	171	0	56.73	43.27
guadix	0	17	12	29	0	58.62	41.38
costatropical	0	12	8	20	0	60	40
alhambra	12	2577	1725	4314	0.28	59.74	39.98
generalife	0	146	85	231	0	63.20	36.79
palaciocarlosv	0	12	5	17	0	70.58	29.41
patiodelosleones	1	26	11	38	2.63	68.42	28.95
paseodelostristes	0	16	6	22	0	72.73	27.28
			Cultura				
flamenco	1	35	89	98	1.02	35.71	63.26
cultura	1	199	182	382	0.27	52.09	47.64
			Actividades - Eventos				
semanasanta	4	1049	1051	2104	0.19	49.86	49.95
ski	2	75	54	131	1.53	57.25	41.22
			Gastronomía				
restaurant	1	38	126	165	0.61	23.03	76.37
tapas	0	104	89	193	0	53.89	46.11
vino	0	11	8	19	0	57.89	42.12
gastronomía	0	24	13	37	0	64.86	35.13

Cuadro 4.5: Resultados del análisis de sentimentos para datos de Instagram (Inglés-Español)

Según estos datos, las entidades más populares vuelven a ser «Alhambra», «Albaicín», «sierranevada», con la inclusión de «semanasanta». Se puede ver que el porcentaje de publicaciones neutras de Instagram es considerablemente

más bajo que el de Twitter, y esto quizás refleje el mayor uso de Twitter para difundir información objetiva. El porcentaje de publicaciones negativas también es menor y el de publicaciones positivas es mayor que el número de publicaciones de Twitter. De la misma forma, los resultados del análisis de sentimiento de Instagram mostrados en la tabla 4.5 puntúan «realejo», «valledelecrin», «almuñecar», «miradordesannicolas», «sacromonte» y «flamenco» como entidades muy valoradas pero que no se mencionan excesivamente, por lo que nuevamente tienen un buen potencial turístico.

Para identificar los aspectos más importantes de los datos negativos, utilizamos el siguiente proceso. En primer lugar, se limpia cada publicación para eliminar los enlaces y los caracteres especiales. Luego, se usan las partes del discurso (POS)[21] en este proceso, mediante el cual las entidades fueron identificadas como sustantivos y los adjetivos nos brindan las percepciones positivas y negativas de estas entidades. Los datos ya están clasificados utilizando tanto Tweeteval para los textos en inglés como el modelo BERT para los datos en español, pero lo complementamos tratando de especificar los aspectos negativos identificando los sustantivos y adjetivos. Se determina la relación o frecuencia entre entidad y adjetivos detectando si tanto la entidad como el aspecto están presentes en un determinado tweet/publicación. El tamaño del resultado que devuelve la consulta anterior constituye la frecuencia del par sustantivo-adjetivo. Por ejemplo, si buscáramos tweets que contengan tanto “Granada” (entidad) como “incorrecto” (aspecto), el resultado mostraría que estas palabras aparecen en 7 tweets. NLTK[22], la biblioteca de Python, se usó para reconocer las partes del habla (POS) (sustantivos y adjetivos) de cada tweet/publicación escrita en inglés, y el framework de Python llamado Spacy[23] para identificar sustantivos y adjetivos escritos en español. El proceso anterior se realizó solo para adjetivos negativos, para ello, se calculó la puntuación de sentimiento para cada adjetivo encontrado usando Texblob para tweets/publicaciones en inglés y CoreNLP Stanford para español, seleccionando así solo adjetivos negativos.

La tabla 4.6 resume los pares de [entidad-adjetivo] más importantes extraídos de los 224 tweets negativos en inglés. El número que se muestra después de cada adjetivo es la frecuencia con la que aparece cada par en este subconjunto de datos. Por ejemplo, algunos usuarios relacionan Granada con adjetivos como equivocado, imposible, decepcionado, malo, terrible, estúpido, horrible, etc.

[21] Part of Speech

[22] Se puede encontrar más información sobre Python Natural Language Toolkit en http://www.nltk.org

[23] Se puede encontrar más información sobre el procesamiento de lenguaje natural para la identificación de partes del habla en https://spacy.io

Entidad	Adjetivos
	Lugares
granada	wrong 7 impossible 6 disappointed 9 due 8 bad 8 terrible 6 other 8 stupid 3 horrible 3 long 2 disappointing 2 miserable 2 few 1 little 3 bloody 1 confusing 1 sad 7 pale 1 narrow 1 unnecessary 1 anxious 1 lazy 1 sorry 4 angry 1 annoyed 1 fake 2 weird 2 awful 2 sharp 1 violent 1 due 8 green 1 sick 1 past 2 dead 2 useless 2 confused 1 dreadful 1 sloppy 1 dangerous 1 vicious 1 annoying 1 guilty 1
alhambra	wrong 13 impossible 7 disappointed 15 due 10 bad 10 terrible 8 other 7 stupid 4 horrible 3 long 4 disappointing 2 miserable 3 few 3 little 3 bloody 3 confusing 2 sad 8 pale 1 narrow 1 unnecessary 1 anxious 1 lazy 1 sorry 7 angry 1 annoyed 1 fake 3 weird 2 awful 4 sharp 1 violent 1 due 8 green 1 sick 1 past 4 dead 2 poor 4 useless 3 expensive 2 confused 1 dreadful 1 sloppy 1 dangerous 1 vicious 1 annoying 1 guilty 1
sacromonte	due 1 pale 1 narrow 1 anxious 1
albaicin	confusing 1
sierranevada	impossible 1 bad 1
	Actividades - Eventos
holy-week	disappointed 1 stupid 1
	Otros
ticket	wrong 2 impossible 2 disappointed 6 bad 1 terrible 1 other 1 stupid 1 long 1 disappointing 1 few 1 confusing 2 angry 1 fake 1 awful 2 poor 1 dreadful 1
booked	wrong 1 disappointed 1 due 3 bad 1 stupid 1 angry 1
museum	disappointed 1 due 1 pale 1 narrow 1 anxious 1
narrow streets	due 1 pale 1 narrow 1 anxious 1 due 1

Cuadro 4.6: Las características negativas más importantes (Inglés-Twitter)

En la otra sección de la Tabla 4.6, podemos ver que una serie de problemas se relacionan con la compra de boletos. Por ejemplo, el tweet *«alhambracultura incredibly unprofessional ticket sale for #alhambra in #granada web doesn't work and they hang up, phoning from australia!! (alhambracultura increíblemente poco profesional venta de entradas para #alhambra en #granada web no funciona y cuelgan, llamando desde australia !!)»* se refiere a la compra de entradas para la Alhambra, algo que algunos viajeros encuentran difícil. Otros usuarios mencionan el proceso de reserva *«I can't believe how much time I had to spend on booking two tickets to visit alhambracultura. Very poor online booking system Ticketmaster (No puedo creer cuánto tiempo tuve que gastar en reservar dos entradas para visitar alhambracultura. Sistema de reserva online muy deficiente Ticketmaster)»*. También hay algunos comentarios negativos sobre las calles estrechas de la ciudad. También encontramos frases nominales ilustrativas (negativas) como:

1. people can't order tickets (la gente no puede pedir boletos)

2. unprofessional ticket sale (venta de entradas no profesional)

3. terrible customer service (terrible servicio al cliente)

4. white elephant construction (construcción de elefante blanco)

5. 2hr queue (cola de 2 horas)

La tabla 4.7 resume los pares de [entidad-adjetivo] más importantes obtenidos de los más de 4,000 tweets negativos en español.

Entidad	Adjetivos
Lugares	
granada	terrible 15 artificial 2 cruel 4 grave 33 impersonal 3 mediocre 5 manual 1 error 29 intolerable 8
alhambra	terrible 13 cruel 4 grave 9 impersonal 2 mediocre 4 manual 1 error 30 intolerable 2
sierranevada	terrible 2 artificial 1 grave 3 mediocre 1 intolerable 1
albaicin	terrible 3 grave 1
catedral	terrible 1
patiodelosleones	terrible 1
monachil	grave 1
palaciocarlosv	grave 1
Cultura	
culture	terrible 8 artificial 1 grave 7 impersonal 1 error 12 terrible 8
lorca	mediocre 1
Actividades - Eventos	
semanasanta	mediocre 1 error 1
Gastronomía	
restaurant	mediocre 1

Cuadro 4.7: Características negativas más importantes (Español-Twitter)

Muchos de los tweets en español hacen referencia a la gestión del destino turístico. Un ejemplo es el tweet *"En Granada: El dinero de la Alhambra y la sierra se lo lleva Sevilla, intolerable. Veis como los catalanes no pueden ser más españoles ??"*. O el tweet *Otra cosilla @Granada_Limpia! A ver si Uds pueden eliminar de la acera los restos de los líquidos que se derraman del contenedor verde, que a veces huele fatal y da una sensación terrible!.*

El mismo análisis se realizó con Instagram. Las entidades más importantes que se encuentran en las publicaciones escritas en inglés son *Granada*, con algunos adjetivos relevantes como «inútil», «feo», «incorrecto», «malo», «terrible», «extraño» u «oscuro»; *Alhambra*, con adjetivos como «incorrecto», «malo»,

«terrible»; *SierraNevada* con adjetivos como «horrible», «malo», «terrible», «oscuro», «pequeño»; *Albaicin* con adjetivos como «feo», «incorrecto», «pequeño» o «terrible».

Es mucho más difícil identificar la polaridad en las publicaciones de Instagram porque el texto es más extenso que los tweets. Un factor adicional es que el mismo texto puede contener el mismo contenido en diferentes idiomas, y que a menudo la entidad se describe pero no se nombra específicamente, debido a que hace referencia a las imágenes subidas a Instagram.

De forma similar, se realizó el proceso de identificación de las entidades y adjetivos más importantes con las publicaciones escritas en español. Las entidades más importantes son *Granada* con adjetivos como «grave», «terrible», «cruel» o «rival», y *Alhambra* con adjetivos como «cruel», «rival» o «victima».

4.4. Discusión

En el presente capítulo se probaron una serie de herramientas que detectan automáticamente la polaridad en diferentes textos. Se recopilaron 90,725 publicaciones de Instagram (7,717 publicaciones en inglés y 56,247 en español) y 235,755 tweets (19,340 tweets en inglés y 144,947 tweets en español) relacionados con el turismo en Granada, España. Luego se aplicaron estas herramientas para evaluar los tweets y las publicaciones. Se analizaron Instagram y Twitter ya que estas plataformas son utilizadas por los viajeros para comentar sobre un destino turístico. Existe una gran diferencia entre los resultados obtenidos por las diversas herramientas en un conjunto de prueba de 1000 tweets en ambos idiomas. Aunque la herramienta Tweeteval obtuvo los mejores resultados para los textos en inglés con una exactitud de 75.60 % y una puntuación de F1 de 71.57 %, ninguna de las herramientas probadas en textos en español fue sobresaliente, aunque Syuzhet obtuvo un mejor rendimiento con una exactitud de 60.50 % y una puntuación F1 de 34,04 %. Luego se experimentó con modelos de aprendizaje profundo alimentados con datos de turismo en un intento por mejorar la exactitud y encontramos que un modelo BERT para español obtenía resultados mucho mejores con una exactitud del 75.70 % y una puntuación F1 de 58.71 %. Luego usamos Tweeteval para textos en inglés y el modelo BERT en español que se desarrolló para textos en español para clasificar los datos como negativos, neutrales o positivos. Se revelaron las entidades turísticas más populares del destino y estas tendían a ser las mismas en Twitter e Instagram, y también las mejor (y peor) valoradas. También se encontraron una serie de entidades menos mencionadas que habían sido evaluadas de manera bastante

positiva, y se considera que estos lugares infravalorados tienen un potencial turístico importante.

También se observó que el porcentaje de publicaciones neutrales en Instagram es considerablemente menor que el de Twitter, lo que quizás refleje un mayor uso de Twitter para difundir información objetiva. El porcentaje de publicaciones negativas en Instagram es menor y el porcentaje de publicaciones positivas en Instagram es mayor que en Twitter.

Para intentar explicar las razones por las que algunos usuarios tienen sentimientos negativos sobre su viaje a Granada, también se realizó un análisis más detallado de los tweets/posts negativos, para esto, se identificó los sustantivos(entidades) y adjetivos(características) de tal manera que se pueda tener más información sobre el porqué se tiene una valoración negativa del lugar, evento o servicio turístico.

Al procesar todas y cada una de las publicaciones y tweets, se puede no solo identificar entidades y aspectos importantes y resaltar los lugares más visitados y los lugares importantes infravalorados, sino también determinar por qué los viajeros pueden tener percepciones negativas sobre el destino turístico.

El modelo BERT en español para datos turísticos que se desarrolló, es fácil de implementar y usar porque se proporciona el código y el TSD. El modelo en español también puede ser adoptado en aplicaciones prácticas con herramientas de gestión de destinos por administradores y organizaciones turísticas y mejorado con investigaciones futuras. El análisis de sentimientos para la detección de entidades y aspectos puede proporcionar a los gerentes, organizaciones o agencias de viajes información sobre qué cosas deben cambiarse o mejorarse en términos de un lugar específico o proveedor de servicios en un destino turístico determinado. Esto, a su vez, permite diseñar nuevas estrategias de marketing y aplicar mejores políticas para que estos lugares se conviertan en destinos turísticos inteligentes.

El siguiente capítulo trata de mejorar la detección de entidades, aspectos y las palabras que expresen el sentimiento de los dos anteriores. Este proceso es mucho más complejo, debido a que se intenta descubrir exactamente qué es lo que los usuarios están publicando y cuál es su percepción sobre un lugar, evento o servicio determinado. Por tanto, el último capítulo de esta tesis se trata del análisis de sentimientos basado en aspectos con un enfoque en la insatisfacción de los usuarios luego de haber visitado un destino turístico.

Capítulo 5

Análisis de Sentimientos basado en Aspectos en Turismo

5.1. Introducción

En el capítulo anterior se analizaron algunas herramientas que permiten el análisis de sentimientos a nivel de oración. Además, se entrenó un modelo basado en aprendizaje profundo para tener mejores resultados en cuanto a la clasificación de sentimientos con texto en español. En este capítulo se analiza la insatisfacción de los usuarios luego de haber visitado un destino turístico. La identificación de las quejas de los usuarios por un determinado lugar o servicio se realizó mediante el análisis de sentimientos basado en aspectos. Esta técnica es mucho más compleja debido a que se trata de identificar a qué aspecto o característica se refieren los usuarios cuando emiten una opinión acerca de un lugar o servicio.

La satisfacción tiene una importancia fundamental en el ámbito del turismo, ya que los gestores pueden obtener información sobre la experiencia del cliente, su satisfacción y proveer una retroalimentación a través de comentarios publicados en plataformas sociales. De acuerdo a los autores en su articulo (Kim y Kim, 2022), el principal objetivo de las actividades de marketing de las empresas es centrarse en la gestión de la satisfacción del cliente, ya que puede aumentar la fidelidad de los clientes y la intención de recompra y contribuir a una mayor rentabilidad. El análisis de sentimientos es una herramienta bastante útil para saber la satisfacción o insatisfacción de un usuario.

En este capítulo se analizan todas las etapas primordiales en el análisis de sentimientos basado en aspectos y se propone un enfoque para evaluar la insatisfacción de los turistas luego de haber viajado a un destino turístico. Por tanto, en la sección 5.2 se aborda el análisis de sentimientos basado en aspectos en todas sus etapas comenzando por la identificación de la entidad (lugar o servicio del destino turístico); la extracción de aspectos y la identificación de su polaridad tomando en cuenta los enfoques publicados hasta la fecha; posteriormente, se analiza la representación vectorial de cada aspecto y la clusterización; por último, la generación de resúmenes es una parte importante para que los resultados sean entendibles para el usuario final, en este capítulo se escogieron dos herramientas basadas en BERT que ayudan a generar resúmenes automáticos usando los aspectos identificados anteriormente. En la sección 5.3 se analizan los trabajos relacionados que aplican análisis de sentimientos basados en aspectos en turismo; además, se revisaron enfoques que utilizan técnicas computacionales para abordar la insatisfacción turística. En la sección 5.4 se propone un enfoque que consta del pre-procesamiento de datos, extracción de entidades, identificación de aspectos y opiniones, representación vectorial de aspectos, agrupación de aspectos, visualización y sumarización. Por último, en la sección 5.5 se aplica este enfoque para identificar la insatisfacción de los turistas con datos de Twitter, Instagram y TripAdvisor de los lugares turísticos más importantes de Granada, tales como: Alhambra, Albaicin, Generalife y Sacromonte.

5.2. Antecendentes

Para entender correctamente el enfoque utilizado en este capítulo, es necesario introducir algunos conceptos teóricos sobre el análisis de sentimientos basado en aspectos.

5.2.1. Análisis de Sentimientos basado en Aspectos (ASBA)

ASBA es una subtarea del análisis de sentimientos. En ASBA, una opinión se define como una quíntupla (e, a, s, h, t), e es una entidad, lugar o servicio (por ejemplo, un monumento, un barrio, un restaurante, etc.), a corresponde a uno de sus atributos o aspectos (por ejemplo, el precio, la limpieza, etc.), s es el sentimiento sobre la entidad-aspecto (positivo/negativo/neutro), h es el titular de la opinión (en los medios sociales es fácil saber qué reseña/post pertenece a un individuo o usuario) y t es el momento en que se emite la opinión (las

plataformas de medios sociales almacenan y muestran la fecha de cada post o comentario) (Liu, 2015).

A continuación, identificamos las siguientes subtareas para aplicar ASBA y también proporcionamos una breve introducción teórica de cada una:

- *Extracción de aspectos*: esta tarea identifica los términos o atributos de una entidad en los textos y su respectiva polaridad (por ejemplo, el tuit: «Alhambra es el monumento más hermoso del mundo»; en este caso la entidad es la «Alhambra»; el aspecto, una característica o rasgo de la Alhambra, es decir, «monumento»; y la opinión puede reconocerse a través de una palabra o frase de opinión, en este caso «hermoso» o «más hermoso»). Hay muchos enfoques para abordar esta tarea, en la siguiente sección analizamos los más importantes.

- *Categorización de aspectos o clustering*: Se trata de una tarea de clasificación, que permite agrupar las palabras relacionadas con un mismo aspecto, ya que un aspecto puede estar asociado a diferentes palabras. Presentamos diferentes métodos de agrupación de aspectos en la sub-sección 5.2.3.

- *Sumarización*: El objetivo de esta tarea es proporcionar o generar un resumen estructurado a partir de todos los recursos encontrados en las tareas anteriores. En la sub-sección 5.2.4 analizamos algunos enfoques sobre el resumen en ASBA.

5.2.2. Extracción de aspectos

En su artículo (Luo, Huang y Zhu, 2019), los autores utilizan diferentes enfoques para identificar aspectos

- *Métodos basados en reglas:* Las reglas se suelen escribir a mano para extraer aspectos del texto. En (Luo, Huang y Zhu, 2019), los autores emplearon seis reglas para extraer los aspectos. También construyeron un gráfico de aspectos para reducir el espacio de aspectos, realizaron la agrupación y finalmente identificaron los aspectos más destacados. En (Rana y Cheah, 2017) se presenta un modelo basado en dos reglas, en el que los autores extraen los aspectos relacionados con las opiniones independientes del dominio (primer pliegue) y luego extraen los aspectos asociados con las opiniones dependientes del dominio utilizando enfoques de frecuencia y similitud (segundo pliegue). En (Li y col., 2020b) se presenta un método híbrido basado en reglas, en el que los autores integran la representación

del significado abstracto de los aspectos con la estructura sintáctica del texto y demuestran la utilidad de la representación semántica profunda. En (Dragoni, Federici y Rexha, 2019), se propone un método supervisado de extracción de aspectos que combina reglas sintácticas, léxicos (por ejemplo, SenticNet (Cambria y col., 2016)) y análisis textual (por ejemplo, Stanford CoreNLP[1]) para construir una herramienta de seguimiento de las reseñas en tiempo real. En (Marcacini y col., 2018), los autores proponen un enfoque llamado «Cross-domain Aspect Label Propagation through Heterogeneous Networks», que utiliza aspectos etiquetados y no etiquetados y reglas lingüísticas para realizar la extracción de aspectos mediante un algoritmo de propagación.

- *Métodos basados en el modelado de temas:* Estos métodos extraen temas del texto y aspectos de los temas. En (Wang y col., 2014) se presenta un enfoque supervisado, en el que los autores utilizan en primer lugar aspectos semilla obtenidos de las descripciones de los productos; en segundo lugar, las reseñas de productos se clasifican de acuerdo con estos aspectos semilla; y finalmente, proponen el LDA de grano fino y el LDA etiquetado de grano fino unificado para descubrir aspectos relacionados con los aspectos semilla. En (He y col., 2021), se propone un modelo temático basado en características jerárquicas (Hierarchical Features-based Topic Model o HFTM por sus siglas en inglés) para extraer aspectos de las reseñas en línea y luego capturar características específicas. En (Ekinci y Omurca, 2020) se presenta el modelo Concept-LDA, en el que los autores utilizan LDA para extraer aspectos latentes construyendo un espacio de características antes de enriquecerlo con conceptos y entidades extraídos de Babelfy[2]. En (Liao y col., 2017), se propone un marco de Aprendizaje de Representación de Fusión de Relaciones (Fusion Relation Embedded Representation Learning o FREERL), donde los autores combinan la co-ocurrencia estadística en incrustaciones de entidades de objetos y atributos. En (Shams y Baraani-Dastjerdi, 2017), los autores proponen un método iterativo que consta de tres etapas: en primer lugar, se obtienen aspectos preliminares mediante LDA; en segundo lugar, se identifican los temas relevantes a través de las relaciones de co-ocurrencia; y en tercer lugar, se incorpora el conocimiento extraído en el modelo LDA, y un número específico de iteraciones mejora la calidad del proceso de extracción de aspectos. Por último,

[1] Una herramienta para realizar el procesamiento del lenguaje natural: https://stanfordnlp.github.io/CoreNLP/sentiment.html

[2] http://babelfy.org/about

en (Ozyurt y Akcayol, 2021) se presenta un método bastante competitivo, en el que los autores crean un LDA de segmentos de frases para realizar una adaptación del algoritmo LDA para la extracción de aspectos.

- *Métodos basados en redes neuronales:* Estos métodos aplican arquitecturas de aprendizaje profundo. En (Santos, Marcacini y Rezende, 2021) se propone una extracción de aspectos multidominio utilizando BERT, combinando 15 conjuntos de datos de diferentes dominios para entrenar el modelo. Los resultados mostraron una alternativa competitiva en comparación con los modelos de un solo dominio. En (Poria, Cambria y Gelbukh, 2016), los autores utilizan una red neuronal convolucional profunda y una serie de patrones lingüísticos para realizar la detección de aspectos. En (Zhang y col., 2021) se presenta un modelo de red neuronal convolucional con filtros dinámicos para extraer los aspectos en un documento, donde los aspectos se categorizan con un modelo neural de temas. En (Singh Chauhan y col., 2020) se presenta un enfoque que aplica patrones lingüísticos (aspectos de una y varias palabras) para etiquetar aspectos y construye un conjunto de datos que se utiliza para entrenar el modelo de aprendizaje profundo. En (Alekseev y col., 2020), los autores proponen un enfoque de red neuronal no supervisada que utiliza un clasificador probabilístico de extracción de aspectos que es útil tanto para los textos fuera del dominio como para los del dominio. Posteriormente, filtra las frases con una baja probabilidad de estar dentro del dominio y entrena el modelo con las frases restantes. Por último, en (Ansar y col., 2021) se propuso una versión modificada del enfoque de frecuencia de términos (es decir, el enfoque de frecuencia de términos inversa), mediante el cual los autores identifican los aspectos significativos y capturan las dependencias a largo plazo en el análisis del sentimiento.

5.2.3. Clusterización de aspectos

Antes de agrupar los aspectos, es importante convertir cada aspecto extraído en una representación vectorial para poder entender las relaciones léxicas entre ellos. Las incrustaciones de palabras son representaciones vectoriales de las palabras que tienen en cuenta las palabras circundantes. Estos vectores pueden generarse con métodos como las redes neuronales, los modelos probabilísticos de matriz de co-ocurrencia, etc. Las herramientas más utilizadas en esta tarea son Word2vec (Mikolov y col., 2013) y Glove (Pennington, Socher y Manning, 2014).

Una vez representados los vectores, es necesario agrupar o clusterizar aspectos similares. Según (Ansar y col., 2021), los aspectos pueden agruparse según sus puntuaciones de similitud (por ejemplo, la similitud del coseno). Los autores aplicaron dos algoritmos de clustering (es decir, el algoritmo de clustering de enlace simple y el algoritmo de clustering de promedio de grupo).

En (Kumar, Saini y Sharan, 2020) se presenta un enfoque basado en reglas de asociación para la detección de clústeres de aspectos, en el que los autores encuentran palabras representativas de la categoría de aspecto utilizando la asociación estadística entre las palabras de revisión y la categoría de aspecto a través de reglas de asociación basadas en la clase. A continuación, se entrenan las incrustaciones de palabras en un conjunto de datos de dominio específico y las incrustaciones de palabras se utilizan para encontrar la asociación semántica entre las palabras de revisión y las categorías de aspecto. Por último, se generan reglas de asociación basadas en la clase.

5.2.4. Sumarización

Según Hu y Liu, 2004, la sumarización tiene dos características importantes: en primer lugar, identifica los objetivos de opinión (es decir, los aspectos) y sus sentimientos, y en segundo lugar, es necesario cuantificar cuántas opiniones positivas y negativas hay sobre los objetivos de opinión. A partir de estas características, esta tarea puede presentarse en un gráfico de barras, en el que cada barra por encima del eje X muestra el número de opiniones positivas y las que están por debajo del eje X corresponden a las opiniones negativas sobre cada aspecto. Otros enfoques, por su parte, aplican resúmenes de texto.

En general, los resúmenes no ordenan los aspectos y no pueden mostrar el aspecto más importante sobre una entidad ni cómo se relacionan los aspectos entre sí. Estas limitaciones se abordan en el artículo (Carenini, Cheung y Pauls, 2013) en el que los autores proponen una herramienta «extractiva» y «abstractiva». Por su parte, el artículo (Di Fabbrizio, Stent y Gaizauskas, 2014) propone un método híbrido que combina técnicas de generación de lenguaje natural y de selección de frases destacadas. Estos dos enfoques aprovechan la generación de lenguaje natural (Natural Language Generation NLG) para generar nuevas frases a partir de los datos extraídos de su corpus para generar resúmenes más coherentes. Para identificar cómo se relacionan los aspectos y las opiniones, en su artículo (Carenini, Cheung y Pauls, 2013), los autores proponen la aplicación de una taxonomía de características definida por el usuario para los aspectos y para ello se utiliza una gran cantidad de datos de entrenamiento (Di Fabbrizio, Stent y Gaizauskas, 2014). Por último, en el artículo (Gerani, Carenini y Ng,

2019) se presenta un enfoque que genera un resumen basado en aspectos a partir de oraciones/revisiones de una entidad sin una taxonomía de características ni datos de entrenamiento, donde los autores toman como entrada un conjunto de revisiones sobre la entidad (objetivo), identifican los aspectos, su polaridad y la fuerza de las opiniones sobre cada aspecto en cada oración, antes de generar los resúmenes con herramientas de generación de lenguaje natural utilizando el grado de relevancia de los aspectos además de la asociación entre ellos.

A pesar de que en el capítulo anterior se revisaron los trabajos publicados en análisis de sentimientos a nivel de oración, en este capítulo nos enfocamos únicamente en aquellos que se basan en ASBA y el turismo.

5.3. Trabajos relacionados

5.3.1. ASBA y turismo

Uno de los objetivos de ASBA es ofrecer información a partir de textos (es decir, posts, tweets, reseñas, blogs, etc.) sobre entidades (lugares, servicios, etc.), sus atributos o aspectos (precio, limpieza, etc.) y opiniones sobre ellos (positivas/neutrales/negativas). Los investigadores han propuesto diferentes enfoques de ASBA para el ámbito del turismo. En (Moreno-Ortiz, Salles-Bernal y Orrequia-Barea, 2019), los autores validan un esquema de anotación para el análisis de sentimiento basado en aspectos utilizando reseñas sobre alojamiento, restauración y alquiler de coches. Sin embargo, sólo se centran en el proceso de construcción del corpus, que es una subtarea de ASBA. El artículo (Afzaal, Usman y Fong, 2019) presenta una aplicación móvil de turismo en la que los autores aplican un método de extracción de aspectos basado en árboles y algoritmos de aprendizaje automático para identificar aspectos y realizar una tarea de clasificación. Esta aplicación proporciona información útil y permite a los visitantes tomar mejores decisiones en su viaje. En (Maity y col., 2020) se presenta un diseño prospectivo, en el que se utiliza un léxico para identificar características (aspectos) a partir de reseñas de viajes sobre hoteles o centros turísticos. En (Stepaniuk y Sturgulewska, 2021), los autores crearon una metodología para analizar y visualizar las respuestas emocionales de los usuarios de las redes sociales de un grupo cerrado de Facebook. Se utilizó ASBA para descomponer semánticamente 300 fotos seleccionadas y los resultados mostraron la comprensión y visualización de las fotos (memes), así como las respuestas emocionales del receptor del contenido visual.

En (Valdivia y col., 2020) se presenta una metodología basada en las reseñas negativas de TripAdvisor que aplica el enfoque de aprendizaje profundo presentado en (Poria, Cambria y Gelbukh, 2016). Los autores utilizan el algoritmo k-means para la agrupación de aspectos y el proceso de sumarización se realiza mediante el descubrimiento de subgrupos a través del uso de reglas de descripción proporcionando información sobre los aspectos negativos de las reseñas.

Una aplicación interesante de ASBA es la identificación de fallos de servicio en el sector hotelero, la satisfacción de los huéspedes del hotel y las experiencias de los usuarios. (Sann y Lai, 2020): los autores identificaron los elementos de fallo de servicio (aspectos) y los agruparon según el ciclo de los huéspedes del hotel y sus correspondientes operaciones. También compararon los patrones de expresión utilizados por asiáticos y no asiáticos para comprender la homofilia de los fallos del servicio, así como sus experiencias en el hotel.

Por último, el ASBA también puede utilizarse para evaluar la reputación de un destino turístico, por lo que en su artículo (Ali y col., 2021), los autores emplearon una técnica para combinar el modelado de temas (LDA) y los algoritmos basados en el léxico para recopilar información sobre la reputación de un destino turístico utilizando las reseñas de TripAdvisor sobre diferentes lugares y monumentos de la ciudad de Marrakech.

5.3.2. (In)Satisfacción turística

Hay muchos estudios sobre la satisfacción del turismo a través de experiencias pasadas utilizando datos de las redes sociales. En su artículo (Wei, Zhang y Ming, 2022), los autores propusieron el análisis de sentimiento a nivel de atributos, recogieron reseñas en línea de TripAdvisor para encontrar experiencias positivas/negativas a nivel de atributos de los usuarios; los autores encontraron que este tipo de atributos tienen una gran influencia en la satisfacción general de los usuarios. Mientras tanto, en el artículo (Sutherland y Kiatkawsin, 2020), los autores examinan los temas de interés de la experiencia del cliente y la satisfacción del sector del alojamiento utilizando datos de Airbnb a través de Latent Dirichlet Allocation (LDA). Otro estudio combina métodos de aprendizaje automático (es decir, LDA para el análisis de datos textuales, k-means para la segmentación de datos y reglas difusas para la predicción del nivel de satisfacción) y enfoques basados en encuestas para el análisis de la satisfacción de los clientes durante el COVID-19 (Nilashi y col., 2022). En su artículo (Oh

y Lee, 2021), los autores proponen el uso de técnicas de inteligencia artificial para analizar los efectos de los componentes de las fiestas locales en la satisfacción de los turistas.

Es importante identificar los factores para la selección de hoteles para la satisfacción de un turista y por ello, en su trabajo (Ahani y col., 2019), los autores examinaron las reseñas anteriores de los viajeros de TripAdvisor a través de enfoques de toma de decisiones multicriterio y de soft computing, para entender la satisfacción de los viajeros y las preferencias de los hoteles. Un enfoque de aprendizaje automático supervisado se presenta en (Sánchez-Franco, Navarro-García y Rondán-Cataluña, 2019) para identificar las características relevantes y clasificar la satisfacción de los clientes de Yelp. Por último, en (Tao y col., 2019), los autores utilizaron herramientas de procesamiento del lenguaje natural, análisis de sentimientos y redes neuronales artificiales utilizando datos de Sina Weibo para evaluar las percepciones de los turistas sobre la calidad del aire y su satisfacción.

Investigadores han publicado varios estudios que abarcan o evalúan la insatisfacción de los turistas. En su artículo (Prakash y col., 2019), los autores analizan la causas en la insatisfacción de los turistas al visitar la vida silvestre en Sri Lanka, ellos utilizan 206 reseñas obtenidas de TripAdvisor encontrando que la queja predominante es la mala gestión del parque. En su trabajo (Hu y col., 2019), los autores investigan las causas de las quejas en el sector hotelero, ellos utilizan un modelo temático estructural para analizar 27,864 revisiones de hoteles en TripAdvisor, encontrando que la mayor parte de quejas se refieren a problemas con el servicio del hotel. Por otro lado, en su artículo (Fernandes y Fernandes, 2018), los autores caracterizan el perfil del usuario que se queja del servicio para identificar los factores que influyen en la publicación de su comentario negativo; 1,191 opiniones de huéspedes obtenidas desde TripAdvisor fueron utilizadas y analizadas a través del análisis de contenido. En (Mate, Trupp y Pratt, 2019), los autores examinan las estrategias utilizadas por administradores de hoteles en respuesta a las reseñas negativas publicadas en TripAdvisor y propusieron un framework basado en el análisis de contenido y entrevistas realizadas a administradores de tal forma que se pueda tener a disposición un amplio rango de estrategias que se podrían aplicar en respuesta a las quejas presentadas por los usuarios. La insatisfacción también se ha investigado en los aeropuertos por (Taheri y col., 2020); los autores obtuvieron los datos de cuestionarios y aplicaron los mínimos cuadrados parciales y análisis multigrupo para probar su modelo en dos aeropuertos de Irán. La insatisfacción en el turismo médico en el que los pacientes viajan a otro sitio para buscar

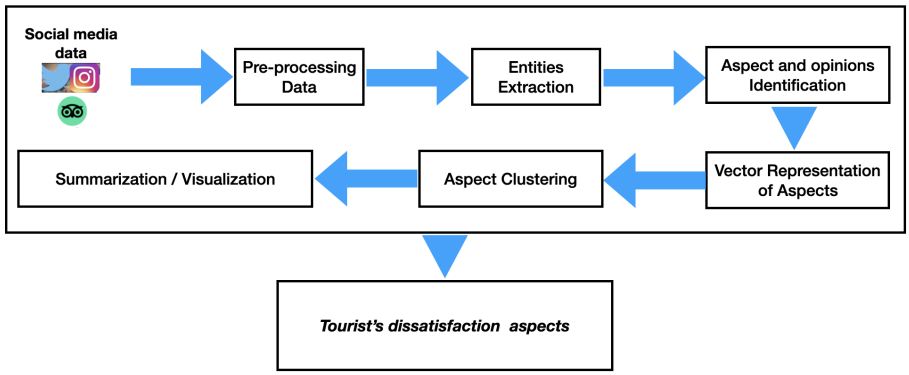

FIGURA 5.1: Enfoque propuesto (ASBA)

tratamiento ha sido investigado por Lam-González y col., 2021; los autores aplicaron encuestas en línea a 354 participantes, el análisis se realizó a través del modelo de ecuaciones estructurales basado en la covarianza; ellos encontraron que la equidad en el servicio (entre un residente y un visitante) es un factor primordial para disminuir la insatisfacción. Por último, en (Rodrigues, Brochado y Troilo, 2020), los autores analizan los atributos de satisfacción e insatisfacción en spas; ellos utilizaron el análisis de contenido con 1,254 revisiones obtenidas desde Booking.com y encontraron cinco atributos esenciales: el spa, el personal, la habitación, la ubicación y la piscina.

A continuación se describe el enfoque utilizado para saber la insatisfacción de los usuarios al visitar un destino turístico.

5.4. Metodología

Para evaluar la insatisfacción de los turistas, se propone un enfoque que aplica el ASBA como una subtarea del análisis de sentimientos, analizando las opiniones negativas sobre lugares, servicios, eventos, etc. de un destino turístico. En la figura 5.1, resumimos el enfoque utilizado.

Cada etapa del enfoque propuesto se amplía en las siguientes subsecciones.

5.4.1. Pre-procesamiento de datos

En este capítulo, utilizamos los datos de Twitter e Instagram recolectados en el capítulo anterior. Este enfoque también se ha probado con datos de TripAdvisor (se utilizó una aplicación construida con Python para recoger las reseñas según el lugar donde se haya publicado).

Una vez cargados los datos de entrada, cada tuit, post o reseña se divide en frases y se realizan las siguientes operaciones para cada frase:

- Eliminación de enlaces mediante patrones de URL
- Eliminación de las menciones de los usuarios
- Eliminación de hashtags
- Eliminación de caracteres no ASCII (no se tienen en cuenta los emoticonos porque el objetivo es encontrar los atributos o aspectos de la entidad)
- Eliminación de los signos de puntuación
- Eliminación de «stop words», es decir, de las palabras irrelevantes incluidas en el texto

5.4.2. Extracción de entidades

Hasta ahora, los investigadores han aplicado el ASBA con un conocimiento previo del objetivo o entidad (por ejemplo, restaurantes, hoteles, ordenadores portátiles, etc.). Sin embargo, en este estudio necesitamos identificar las entidades que los usuarios mencionan en las redes sociales. También utilizamos las reseñas de TripAdvisor y, en esta red social, es necesario especificar la entidad o el servicio para extraer las reseñas sobre una entidad concreta, a diferencia de Twitter o Instagram, donde no se especifica la entidad o el servicio. Por lo tanto, utilizamos Twitter para identificar las entidades más importantes en un destino turístico y luego aplicamos el enfoque propuesto con los tres conjuntos de datos (Twitter, Instagram y TripAdvisor). Dado que es necesario identificar los lugares del destino turístico (por ejemplo, Alhambra[3], Albaicín[4], etc.), el servicio o el evento (por ejemplo, las procesiones de Semana Santa), proponemos el siguiente proceso semiautomático para la identificación de entidades:

[3]La Alhambra es un palacio y fortaleza situado en Granada, Andalucía, España.
[4]El Albaicín es un barrio de la ciudad de Granada.

1. Las entidades candidatas se identificaron utilizando todos los tuits mediante el enfoque[5] propuesto en Schweter y Akbik, 2020, que es un modelo BERT de reconocimiento de entidades con nombre (NER) de 4 clases para el español. Las cuatro clases de nombre identificadas son la persona (PER), la localización (LOC), la organización (ORG) y los misceláneos (MISC).

2. A continuación se calculó la frecuencia para cada entidad identificada.

3. Por último, un residente del destino turístico agrupó las entidades según el lugar, el servicio o el evento. Es importante señalar que utilizamos un umbral de frecuencia para descartar las entidades irrelevantes.

5.4.3. Identificación de aspectos y opiniones

Este enfoque aplica las reglas gramaticales para la identificación de aspectos. Esto es importante, ya que los sustantivos suelen ser aspectos y los adjetivos están relacionados con las palabras de opinión. En este capítulo, utilizamos datos en inglés, pero estas reglas pueden aplicarse en cualquier otro idioma, como el español, el francés, etc.

Estas reglas gramaticales pueden ser convertidas a representaciones de tal forma que se puedan entender por un ordenador. Las «Stanford typed dependencies» proporcionan una representación sencilla de las reglas y relaciones gramaticales que son accesibles a personas sin conocimientos lingüísticos. Estas dependencias son las relaciones binarias entre dos palabras de la oración, ordenadas en una jerarquía y contienen 56 relaciones gramaticales. Dado que una relación gramatical se da entre un *head* y un *dependent*, debemos identificar la relación gramatical, head y dependent. Por ejemplo, la frase «La Alhambra es el lugar más hermoso» tiene las siguientes relaciones bajo esta representación: (i) **nsubj**(lugar, Alhambra), (ii) **amod**(lugar, hermoso), (iii) **cop**(lugar, es), (iv) **det**(lugar, el), (v) **advmod**(más, hermoso). Las relaciones gramaticales (por ejemplo, nsubj, amod, etc.) fueron necesarias para construir un algoritmo que permita identificar los aspectos, las mismas se describen a continuación.

Utilizamos la librería CoreNLP[6] desarrollada por el Stanford NLP Group para realizar la identificación de aspectos y opiniones a través de dependencias gramaticales expresadas por un tipo de relación, head y dependent. Se seleccionaron algunas reglas de (Dragoni, Federici y Rexha, 2019) y se utilizaron dos

[5] El sitio web de Flair Hugging Face se encuentra en https://huggingface.co/flair/ner-spanish-large.

[6] https://stanfordnlp.github.io/CoreNLP/

léxicos para identificar la polaridad de las palabras: SenticNet[7] y el léxico de opinión propuesto por Liu, 2010:

- **Compuesto:** Si el head y el dependent son sustantivos, los unimos utilizando el caracter «_» para obtener un aspecto compuesto. Por ejemplo, en la frase «Grandiosa vista del atardecer desde el antiguo barrio árabe», una de las tripletas es (compuesto, atardecer, vista) y en esta tripleta, tanto «atardecer» como «vista» son sustantivos, por lo que el aspecto resultante es vista_atardecer).

- **Modificador de adjeto «amod»**: Esta regla se aplica si el head es un aspecto (sustantivo) y el dependent es un adjetivo con valor de polaridad. Por ejemplo, en la frase «Muy pobre sistema de reservas online», una de las tripletas es (amod, sistema_reservas, pobre), y en esta tripleta, tenemos un aspecto compuesto «sistema_reservas» como head, y un adjetivo negativo «poor» como dependent.

- **Sujeto nominal «nsubj»**: El head debe tener un valor de polaridad para aplicar esta regla: por ejemplo, en el caso de la triple (nsubj, fantástico, vista), tenemos una opinión positiva (fantástico) sobre el aspecto (vista).

- **Conjunción «conj»**: Si el head y el dependent son aspectos, si uno de ellos está presente en la regla «amod», al otro aspecto se le debe asignar el mismo adjetivo. Por ejemplo, en la frase «Este lugar tiene una gran música y decoración», tenemos las tripletas (amod, música, gran) y (conj, música, decoración), y así podemos generar la tripleta (amod, decoración, gran). Por el contrario, si el head y el dependent son adjetivos o palabras de polaridad, si uno de ellos está presente en una regla «nsubj», al otro adjetivo se le debe asignar el mismo aspecto. Por ejemplo, en la frase «Este servicio es malo y caro», tenemos las tripletas (nsubj, malo, servicio) y (conj, malo, caro) y así podemos generar la tripleta (nsubj, caro, servicio).

- **Negación**: Si las palabras como no, nunca, ni, no puede, etc. están presentes en el texto, la polaridad de las palabras de opinión se modifica.

En resumen, primero identificamos las entidades a partir del texto para cada tuit, post o reseña (con la excepción de las reseñas de TripAdvisor, ya que debemos conocer la entidad). En segundo lugar, dividimos el texto en frases, y

[7]Se trata de una sencilla API para utilizar SenticNet (`http://sentic.net/`) y se puede obtener más información en `https://pypi.org/project/senticnet/`.

para cada frase, identificamos la entidad más importante según la frecuencia y asociamos esa entidad con la frase. Por ejemplo, si un texto tiene las dos entidades de Alhambra (2,952) y Albaicín (197), seleccionamos Alhambra como la entidad importante en esta frase para el tweet o el post (el proceso de identificar la entidad más importante se excluye para las reseñas de TripAdvisor). En tercer lugar, identificamos las reglas y partes de la oración para cada palabra. Por último, aplicamos las reglas descritas anteriormente a la identificación de aspectos y opiniones. Este procedimiento se resume en el Algoritmo 1.

Algorithm 1 Identificación de aspectos y opiniones

$entities \leftarrow entities\_identification\_process()$
for t in tweet/post/review **do**
  if $t == review$ **then**
    $important\_entity \leftarrow Null$
  else
    $pub\_entities \leftarrow get\_entities(t, entities)$
    $important\_entity \leftarrow get\_entity(pub\_entities, entities)$
  end if
  $aspect\_opinion\_rules \leftarrow [amod, nsub, dobj, compound, conjunction, negation]$
  $sentences \leftarrow split\_text(t)$
  for s in sentences **do**
    $rules \leftarrow get\_rules(s)$
    $result \leftarrow get\_aspects\_opinions(rules, aspect\_opinion\_rules)$
  end for
end for

5.4.4. Representación vectorial de los aspectos

El proceso de transformación de las palabras en representaciones vectoriales es un paso importante en el procesamiento del lenguaje natural. Estas representaciones pueden utilizarse para la categorización de aspectos o la agrupación. Tras completar la extracción de aspectos, transformamos cada aspecto en una representación vectorial de dimensión finita utilizando «ConceptNet (CN) NumberBatch» (Speer, Chin y Havasi, 2017). Este conjunto de vectores semánticos fue seleccionado porque se construye utilizando un conjunto que combina datos

de ConceptNet[8], word2vec, GloVe y OpenSubtitles 2016[9]. La incrustación de palabras o la representación de vectores numéricos del texto utilizada en este estudio a través de CN Numberbatch pre-entrenado nos permite mantener las relaciones semánticas y contextuales de los aspectos de nuestro conjunto de datos para poder agruparlos.

5.4.5. Agrupación de aspectos

Dado que la representación vectorial de palabras es necesaria para la agrupación de aspectos, aplicamos esta representación a cada aspecto. Esto permite resumir la información mediante visualizaciones o resúmenes para comprender la satisfacción del viajero con el destino turístico.

Una vez representado cada aspecto como un vector, se realiza la agrupación de aspectos. El clustering es el proceso por el cual los conjuntos de objetos se agrupan en clases teniendo en cuenta que los objetos de un mismo grupo son más similares entre sí que los de otro grupo. Por lo tanto, tenemos que identificar qué aspectos son similares entre sí: por ejemplo, los aspectos «customer_service», «reservation_system» y «service» son similares y podrían agruparse en un clúster, mientras que el aspecto «tour_experience» debería agruparse en otro clúster. Utilizamos k-means para realizar la agrupación de aspectos; además, utilizamos la métrica Silhouette para determinar el número óptimo de clusters (Rousseeuw, 1987).

5.4.6. Visualización / Sumarización

Una vez obtenido el conjunto de datos final, utilizamos árboles de palabras para examinar las opiniones en las que aparecen diferentes aspectos. La estructura del árbol muestra las combinaciones entidad-clúster-aspecto-palabra(s)_de_opinión, con el nodo raíz correspondiente a la entidad (nivel 0), el nivel 1 perteneciente al clúster de aspectos, el nivel 2 correspondiente a los aspectos y, por último, el nivel 3 que muestra las palabras de sentimiento u opiniones sobre el aspecto. El gráfico nos da una idea general sobre la satisfacción de la entidad.

Por último, realizamos el proceso de sumarización, que nos proporciona toda la información relevante y más importante de los tweets, posts y reseñas sin tener que leer cada post de las redes sociales. Hay dos categorías de resumen: *resumen textual extractivo*, que extrae las frases significativas del texto y *resumen textual abstractivo*, que es un método avanzado para identificar las secciones

[8] https://conceptnet.io
[9] https://www.opensubtitles.org/en/search/subs

importantes del texto, interpretar el contexto y compilar un resumen con la información principal de una manera diferente. En este capítulo, utilizamos la primera categoría porque tenemos una serie de posts, tweets y reseñas en lugar de un documento grande y esta es la mejor opción para resumir el contenido de los medios sociales.

5.5. Resultados

En esta sección se presentan los resultados obtenidos mediante el enfoque propuesto. El objetivo es analizar las percepciones de los turistas sobre su insatisfacción con un destino turístico. Utilizamos datos de Twitter e Instagram del capítulo anterior. El conjunto de datos corresponde a Granada, un importante destino turístico español. Se utilizaron 19,340 tweets y 7,717 posts de Instagram, todos ellos escritos en inglés. Dado que ASBA exige que se descarten los saludos, las preguntas, los cumplidos y las despedidas, filtramos los tuits y las publicaciones en consecuencia y obtuvimos 2,613 tweets en inglés, 7,712 publicaciones en inglés (sólo se eliminaron 5 publicaciones de Instagram, ya que generalmente no contienen saludos, a diferencia de Twitter) y 25,483 reseñas de TripAdvisor sobre Granada en España. A continuación, eliminamos las URL, las menciones de usuarios, los hashtags, los caracteres no ASCII, los signos de puntuación y las palabras irrelevantes.

5.5.1. Extracción de entidades

En esta etapa, utilizamos los datos de Twitter para identificar las entidades mediante la herramienta BERT descrita en la sección 5.4. A continuación, calculamos la frecuencia de cada entidad y, finalmente, agrupamos cada una de ellas. Este proceso se realizó con 2,613 tuits en inglés e incluimos 21,143 tuits en español para que hubiera un número representativo de tuits. Nos dimos cuenta de que las palabras clave que pertenecen a entidades específicas podían estar en tuits en español y en inglés: por ejemplo, en el tuit en inglés «The procession won't continue its parade #SSantaGr» y el tuit en español «La Hermandad del Huerto tampoco sale a procesionar #SSantaGr #SemanaSanta #Granada», ambos tuits contienen el mismo hashtag/palabra clave #SSantaGr, que hace referencia a la Semana Santa. Al incluir los datos en español, obtuvimos las siguientes entidades: Alhambra, Albaicín, Generalife, Semana Santa, Palacio de Carlos V, Patio de los Leones, Alpujarra, Guadix, Mirador de San Nicolás,

Sacromonte, Motril, Realejo, Parque de las Ciencias, Federico García Lorca, Almuñécar, Valle de Lecrin, Paseo de los Tristes, Sierra Nevada. Estas entidades identificadas corresponden principalmente a lugares de Granada, aunque también hay eventos culturales (por ejemplo, la Semana Santa o el poeta Federico García Lorca, que nació en Granada). Como hay menos tuits, posts y opiniones negativas que positivas, para realizar el análisis de insatisfacción sólo hemos utilizado las siguientes entidades:

- Alhambra[10]
- Albaicin [11]
- Generalife[12]
- Sacromonte[13]

Una vez identificados los lugares y eventos de un destino turístico, continuamos con las siguientes etapas del enfoque. La visualización o árbol de palabras consta de cuatro niveles: el nivel uno corresponde a la entidad (por ejemplo, Alhambra); el nivel dos contiene los números que corresponden a los clústeres encontrados con el algoritmo k-means; el nivel tres corresponde a los aspectos, que suelen ser sustantivos (por ejemplo, sistema de audio(audio_system), belleza(beautiful), lugar y entrada(place and entrance), etc.); y el nivel cuatro corresponde a las palabras de sentimiento, que suelen ser adjetivos (por ejemplo, horrible(horrible), aburrido(boring), triste(sad), etc.). Cada árbol de palabras fue creado utilizando la librería D3.js[14].

Para generar los resúmenes automáticamente, seleccionamos dos herramientas construidas con tecnología BERT. La primera de ellas es BART y fue propuesta por el equipo de Facebook (Lewis y col., 2019), mediante la cual los autores pre-entrenaron su modelo utilizando el idioma inglés y lo afinaron y mejoraron con el conjunto de datos CNN_Dailymail[15] (que contiene más de 300,000 artículos de noticias únicos escritos por periodistas de CNN y Daily

[10]La Alhambra es un palacio y fortaleza islámicos situados en Granada, Andalucía, España.
[11]El Albaicín es un barrio de Granada con las calles estrechas y sinuosas de su pasado medieval
[12]El Generalife es un palacio de verano y una finca de los gobernantes nazaríes del Emirato de Granada.
[13]El Sacromonte es un barrio gitano y es la capital del flamenco en Granada.
[14]Esta es una librería de Javascript para producir visualizaciones interactivas y se puede encontrar más información en https://d3js.org/.
[15]más información sobre este conjunto de datos se puede encontrar en https://huggingface.co/datasets/cnn_dailymail.

Mail). Esta herramienta obtiene buenos resultados en las tareas de diálogo abstracto, respuesta a preguntas y sumarización.

La segunda herramienta se construyó utilizando un transformador creado por el equipo de Google (Raffel y col., 2019). Este transformador se llama «T5», que significa Transformador de Transferencia de Texto a Texto[16] y se mejoró con 4,515 ejemplos de artículos de noticias de «The Hindu», un diario de noticias hindú y «The Guardian»[17].

Es posible generar resúmenes sobre entidades teniendo en cuenta todos los aspectos pertenecientes a una entidad específica, o seleccionando un clúster de una entidad que incluya todas las reseñas, tuits o posts relacionados con aspectos pertenecientes a ese clúster. Los resúmenes también pueden generarse utilizando frases pertenecientes a dos o más clústeres.

A continuación, procederemos a analizar las percepciones de insatisfacción de los usuarios para cada entidad examinando las tres plataformas de medios sociales.

5.5.2. La Alhambra

TripAdvisor

Las opiniones de TripAdvisor son subjetivas, por lo que proporcionan percepciones de opinión sobre un lugar o servicio. Aplicamos las mismas herramientas tanto en Twitter como en Instagram para identificar aspectos y opiniones. Aunque los usuarios suelen calificar sus reseñas de TripAdvisor como positivas o negativas, el objetivo es encontrar los aspectos y las opiniones positivas o negativas de estas reseñas. Por lo tanto, utilizamos todas las reseñas independientemente de su calificación, ya que las reseñas positivas también pueden contener aspectos negativos.

Dado que la Alhambra es el lugar más visitado de Granada, había muchas referencias a ella (16,116 reseñas de TripAdvisor). Dividimos el árbol de palabras de la Alhambra en dos para que sea más claro de ver. Las figuras 5.2 y 5.3 muestran los aspectos y las respectivas opiniones negativas en las reseñas de TripAdvisor.

[16] Más información sobre esto se puede encontrar en https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html.

[17] Más información sobre esta herramienta se puede encontrar en https://huggingface.co/mrm8488/t5-base-finetuned-summarize-news.

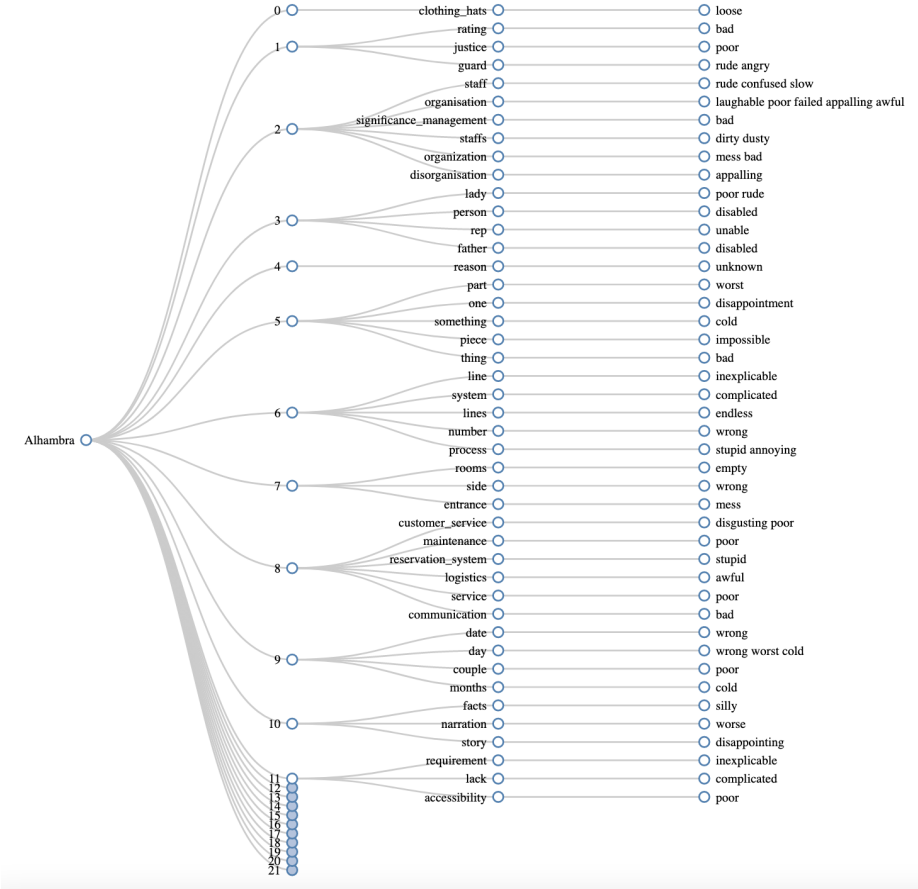

FIGURA 5.2: Percepciones negativas sobre la Alhambra-parte 1 (TripAdvisor)

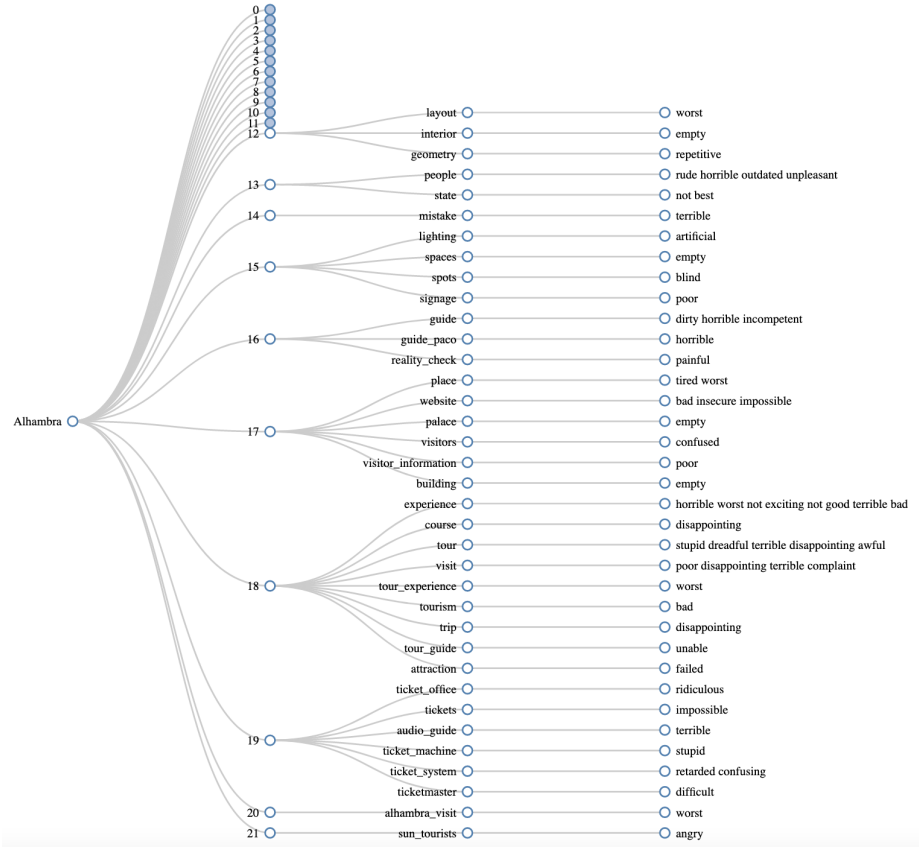

Figura 5.3: Percepciones negativas sobre la Alhambra-parte 2 (TripAdvisor)

Aunque hemos detectado muchos clústeres sobre la Alhambra, sólo hemos seleccionado el clúster 19 para mostrar los resúmenes automáticos:

- **Resumen de la herramienta BART**: «worst part is the ridiculous ticket office. Once inside there is little or no information on signboards. Unless you purchase their terrible audio guide or have prebooked a tour you are walking blind. If you can't get to the palace try to get tickets online. If it's too late you can walk to a viewing point for free.». Traducido al

español: «Lo peor es la ridícula taquilla. Una vez dentro hay poca o ninguna información en los carteles. A menos que compres su terrible audioguía o hayas reservado una visita guiada, vas a ciegas. Si no puedes llegar al palacio, intenta conseguir las entradas por Internet. Si es demasiado tarde, puedes ir andando a un mirador de forma gratuita.»

- **Resumen de la herramienta T5**: «Stayed up like a fool thinking I would get ticket to the palace. Tried to get tickets impossible, turn up at stayed up like a fool thinking i would get ticket must book them also they did not tell you that you could walk to a viewing point for free good luck. once you are inside there is little or no information on signboards so unless you purchase an audio guide or have prebooked a tour you are walking blind.». Traducido al español: «Me quedé despierto como un tonto pensando que conseguiría entradas para el palacio. Intenté conseguir entradas, imposible, me quedé despierto como un tonto pensando que iba a conseguir entradas y que tenía que reservarlas. Una vez que estás dentro hay poca o ninguna información en los carteles, así que a menos que compres una audioguía o hayas reservado un tour, estás caminando a ciegas.»

Twitter

En la Figura 5.4, mostramos los tuits negativos acerca de la Alhambra.

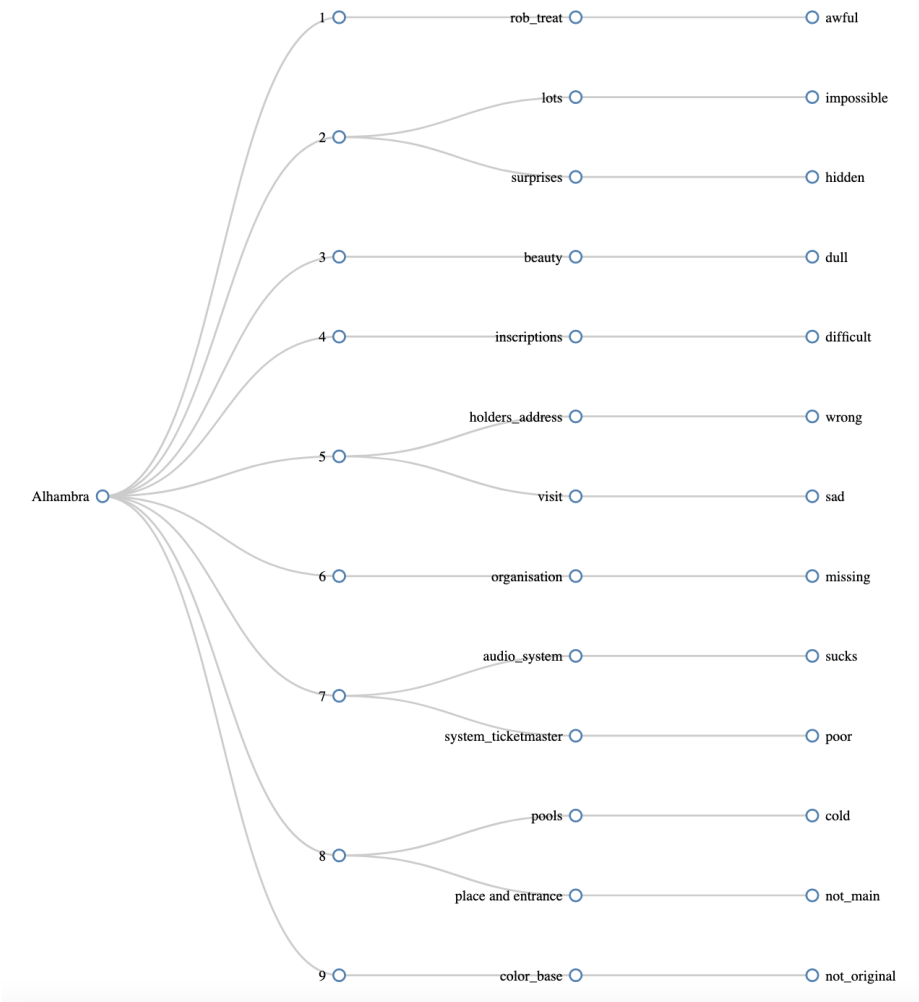

Figura 5.4: Percepciones negativas sobre la Alhambra (Twitter)

Los siguientes resúmenes se obtuvieron de las herramientas anteriormente utilizadas:

- **Resumen de la herramienta BART**: «Icy rain not enough to dull

beauty of Alhambra but how did the nasrids keep warm. The classic arabic inscriptions are difficult to read as they re decorated with ornaments and flowers. I saw people with tickets who couldnt enter. That's a rob awful treat.» Traducido al español: «La lluvia helada no es suficiente para empañar la belleza de la Alhambra, pero ¿cómo se calentaban los nazaríes? Las inscripciones árabes clásicas son difíciles de leer porque están decoradas con adornos y flores. Vi gente con entradas que no podía entrar. Eso es un robo horrible.»

- **Resumen de la herramienta T5**: «I saw lots of people with tickets who couldn't enter. That's a rob awful treat. Alhambracultura is chaotic entry organisation families and seniors queuing for entry then missing slots. Audiosystem with guide it sucks. Alhambracultura its not the original color base on what I saw.». Traducido al español: «He visto a mucha gente con entradas que no ha podido entrar. Eso es un robo horrible. Alhambracultura es un caos en la organización de la entrada, familias y personas mayores haciendo cola para entrar y luego perdiendo los huecos. El sistema de audio con guía es una mierda. Alhambracultura no es el color original por lo que he visto.»

Instagram

La Figura 5.5 muestra las percepciones negativas acerca de la Alhambra.

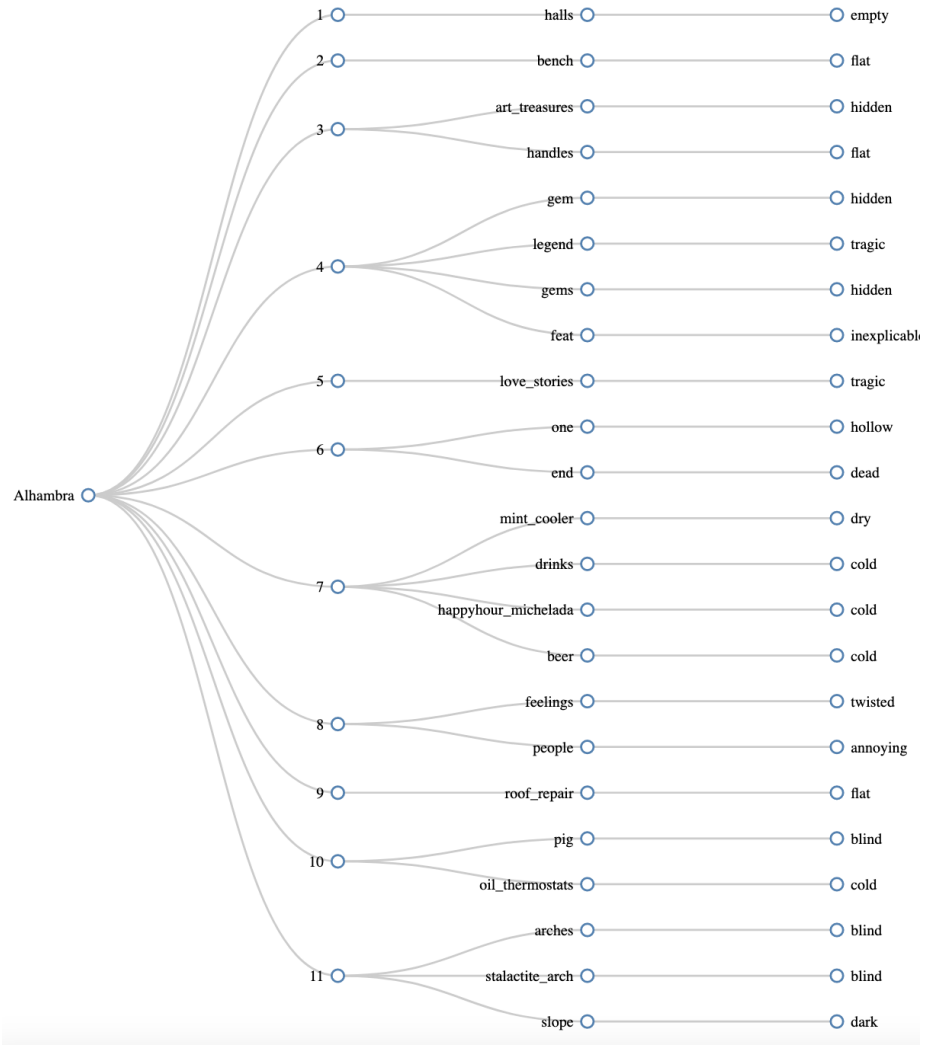

FIGURA 5.5: Percepciones negativas sobre la Alhambra (Instagram)

Los aspectos y sus opiniones encontradas podrían estar relacionados con la

Alhambra y sus alrededores porque los empresarios, dueños de bares y restaurantes suelen utilizar el hashtag/palabra clave #Alhambra en sus posts, por ejemplo: «¿Sediento? ven a probar nuestra michelada fría... happyhour michelada #alhambra ...»

Este árbol de palabras proporciona una visión general de las percepciones negativas sobre la Alhambra en Instagram. Sin embargo, como estas percepciones están relacionadas con la historia o las historias de amor tristes, los restaurantes, la comida, etc., no hay quejas sobre el servicio turístico. Utilizamos las herramientas del resumidor para analizar los clústeres 1, 8 y 9.

- **Resumen de la herramienta BART**: «These stormy times are gone but i still feel the marks of the history in the empty halls of Alhambra. Not sure how to solve my twisted feelings so i resort to mindlessly sorting through pictures. Before and afters of a flat roof repair in Alhambra in Spain.» Traducido al español: «Estos tiempos tormentosos se han ido, pero todavía siento las marcas de la historia en las salas vacías de la Alhambra. No estoy seguro de cómo resolver mis retorcidos sentimientos, así que recurro a ordenar sin sentido las fotos. Antes y después de la reparación de un tejado plano en la Alhambra en España.»
- **Resumen de la herramienta T5**: «Not sure how to solve my twisted feelings so i resort to mindlessly sorting through pictures. Before and afters of a flat roof repair in Alhambra.» Traducido al español: «No estoy seguro de cómo resolver mis sentimientos retorcidos así que recurro a ordenar sin sentido las fotos. Antes y después de la reparación de un tejado plano en Alhambra.»

En sus posts de Instagram, los viajeros y usuarios suelen compartir sus experiencias en puntos concretos de su viaje para mostrar vistas y paisajes del destino turístico. También suelen mostrar sus comidas y bebidas (por ejemplo, cerveza fría, happyhour_michelada fría). También hay una tendencia a que estos posts se escriban de forma poética o metafórica. En consecuencia, ni el enfoque propuesto ni los resumidores funcionan bien con los datos de Instagram.

5.5.3. Albaicín

TripAdvisor

El Albaicín es otro lugar importante de Granada. Obtuvimos 2,456 reseñas y las frases expresan opiniones positivas o negativas sobre los aspectos. La figura 5.6 muestra el árbol de palabras sobre los aspectos y opiniones negativas.

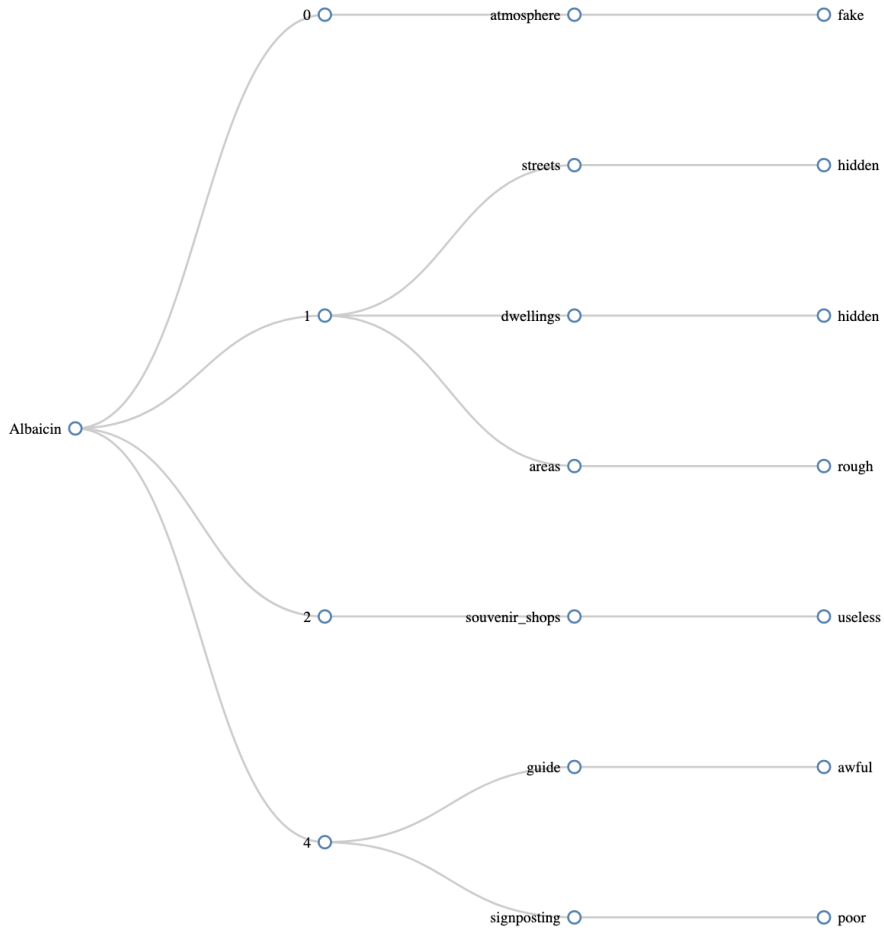

FIGURA 5.6: Percepciones negativas sobre el Albaicín (TripAdvisor)

Los siguientes resúmenes se obtuvieron de las herramientas utilizadas anteriormente:

- **Resumen de la herramienta BART**: «Would be charming but for amount of dog faeces and dodgy loiterers. Poor signposting and poorer

tourist maps mean it is easy to get lost and end up in rough looking areas full of heavily tattooed men who look anything but welcoming. Small winding streets but dwellings mainly hidden behind high walls.» Traducido al español: «Sería encantador si no fuera por la cantidad de heces de perro y los vagabundos de dudosa reputación. La escasa señalización y los mapas turísticos más pobres hacen que sea fácil perderse y acabar en zonas de aspecto rudo llenas de hombres muy tatuados que parecen cualquier cosa menos acogedores. Las calles son pequeñas y sinuosas, pero las viviendas están escondidas detrás de altos muros.»

- **Resumen de la herramienta T5**: «full of heavily tattooed men who look anything but welcoming.. would be charming but for amount of dog faeces and dodgy loiterers.. we had a fixed walk with an awful guide called diego.. too much pain for my legs and nothing worth seeing.. you turn a corner and there is a street full of people and traditional shops.. diego's guide said it was "too much pain for my legs and nothing worth seeing".. » Traducido al español: «lleno de hombres muy tatuados que parecen de todo menos acogedores.. sería encantador si no fuera por la cantidad de heces de perro y vagabundos de mala muerte.. tuvimos un paseo fijo con un guía horrible llamado diego.. demasiado dolor para mis piernas y nada que merezca la pena ver.. giras una esquina y hay una calle llena de gente y tiendas tradicionales.. el guía de diego dijo que era "demasiado dolor para mis piernas y nada que merezca la pena ver"..»

No realizamos clustering ni resumen con los datos de Twitter porque no hay suficiente información negativa sobre esa entidad. Sólo obtuvimos 5 frases con opiniones positivas y 1 con una opinión negativa («Este es el tema del Albaicín, encontrar algo irreal y nunca más porque es un laberinto»). Del mismo modo, sólo obtuvimos 16 publicaciones en Instagram y de éstas, 4 fueron negativas (un ejemplo es «puedes perderte vagando por las estrechas calles del Albaicín así que debes tener siempre tu google maps»).

5.5.4. Generalife

TripAdvisor

El Generalife es un lugar importante para visitar en Granada. Obtuvimos 3,895 opiniones y empleamos el enfoque propuesto para obtener sólo unas pocas opiniones y aspectos negativos sobre esta entidad. La figura 5.7 muestra los aspectos y sus opiniones.

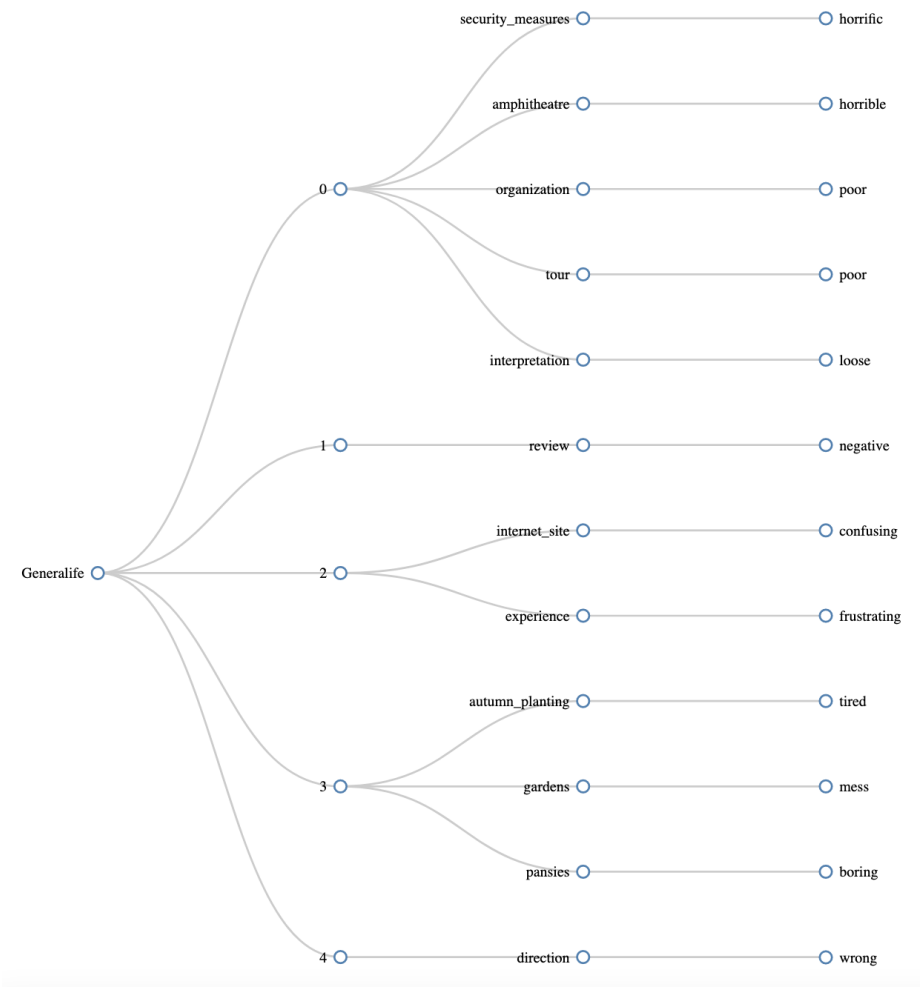

FIGURA 5.7: Percepciones negativas sobre el Generalife (TripAdvisor)

Los siguientes resúmenes se obtuvieron de las herramientas utilizadas en secciones anteriores:

- **Resumen de la herramienta BART**: «The gardens were in a mess

and the plants on display were not the type of plants that would have been used so the character of the gardens was nothing like the original. The gardens were very tired with little or no consideration for autumn planting or colour. The internet site was very confusing you don't get any idea what you are buying.» Traducido al español: «Los jardines estaban desordenados y las plantas expuestas no eran el tipo de plantas que se habrían utilizado, por lo que el carácter de los jardines no se parecía en nada al original. Los jardines estaban muy cansados con poca o ninguna consideración para la plantación de otoño o el color. El sitio de Internet era muy confuso, no se sabe lo que se está comprando.»

- **Resumen de la herramienta T5**: «Alhambra itself was nice however the tour was poor. The gardens were in a mess and the plants on display were not the type that would have been used. The staff at the venue need to do more to fulfill the hype and plant natives or give the effect of what the plantings were like originally.» Traducido al español: «La Alhambra en sí misma era bonita, pero la visita guiada era pobre. Los jardines estaban desordenados y las plantas expuestas no eran del tipo que se hubiera utilizado. El personal del lugar debe hacer más para cumplir con la publicidad y plantar plantas nativas o dar el efecto de cómo eran las plantaciones originalmente.»

Al no disponer de suficientes datos de percepción negativa en Twitter o Instagram, no hemos podido realizar el enfoque propuesto.

5.5.5. Sacromonte

TripAdvisor

El Sacromonte es otro lugar importante para visitar en Granada y obtuvimos 570 reseñas sobre él. Sin embargo, una vez empleado el enfoque propuesto, se obtuvieron muy pocos aspectos. La figura 5.8 muestra las opiniones y aspectos negativos sobre esta entidad.

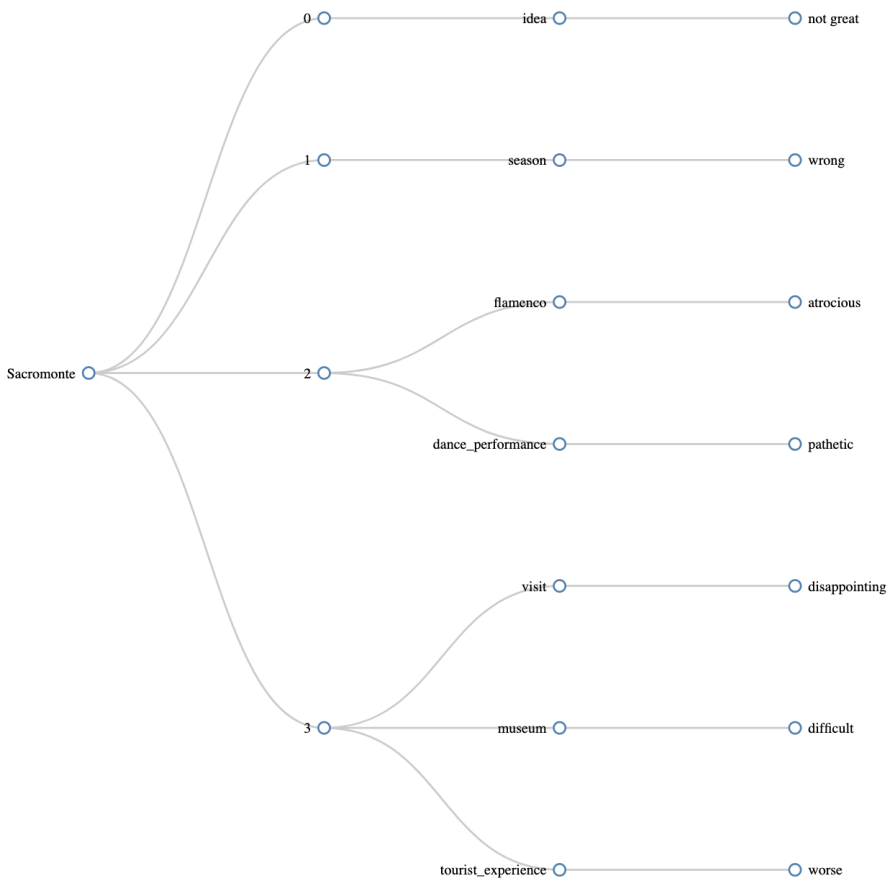

FIGURA 5.8: Percepciones negativas sobre Sacromonte (TripAdvisor)

Los siguientes resúmenes se obtuvieron de las herramientas utilizadas anteriormente:

- **Resumen de la herramienta BART**: «A long uphill walk to visit caves but the museum was closed even though it said opening hours were from 10:00 am. The museum of rural life at sacramonte is difficult to reach without a car and involves climbing several flights of steps from the road.

The intention was to see a touristytype flamenco ie focus on the woman dancer with all her polka dotted finery.. the show itself was pretty much a disaster.» Traducido al español: «Una larga caminata cuesta arriba para visitar las cuevas pero el museo estaba cerrado aunque decía que el horario de apertura era a partir de las 10:00 am. El museo de la vida rural en sacramonte es de difícil acceso sin coche e implica subir varios tramos de escaleras desde la carretera. La intención era ver un espectáculo de flamenco de tipo turístico, es decir, centrado en la mujer bailaora con todas sus galas de lunares... el espectáculo en sí fue un desastre.»

- **Resumen de la herramienta T5**: «I was accosted by a very aggressive young man who threatened me with bodily injury while walking through the old part of sacromonte. The museum of rural life at sacromonte is difficult to reach without a car and involves climbing several flights of steps from the road. It is the worse tourist experience i have ever had.» Traducido al español: «Fui abordado por un joven muy agresivo que me amenazó con lesiones corporales mientras caminaba por la parte antigua de sacromonte. El museo de la vida rural de sacromonte es de difícil acceso sin coche y supone subir varios tramos de escaleras desde la carretera. Es la peor experiencia turística que he tenido.»

Al no existir datos de Twitter o Instagram para esta entidad, no fue necesario realizar el proceso propuesto.

5.6. Discusión

El contenido de las redes sociales generado por los usuarios desempeña un papel fundamental en el seguimiento de la satisfacción o insatisfacción de los turistas. En el presente capítulo hemos analizado los datos de Twitter, TripAdvisor e Instagram utilizando un enfoque de análisis de sentimiento basado en aspectos. Este enfoque incluye un algoritmo semi-supervisado para identificar las entidades más importantes (lugares o eventos) en un destino turístico; un algoritmo basado en reglas para la identificación de aspectos y el cálculo de opiniones (y aunque esto puede aplicarse utilizando datos en otros idiomas, en este capítulo sólo utilizamos tuits, posts y reseñas en inglés); un proceso de visualización, que incluye la representación vectorial de los aspectos a través de ConceptNet Numberbatch para mantener las relaciones semánticas y contextuales dentro de los aspectos y un proceso de clustering utilizando k-means y word-tree para comprender mejor los factores de insatisfacción. La última etapa

de este enfoque aplica dos herramientas de sumarización diferentes que generan resúmenes cortos. Aunque éstos pueden generarse con el árbol de palabras completo o por clúster, si hay muchos aspectos u opiniones, es aconsejable aplicar estas herramientas para cada clúster. La información proporcionada permite a los gestores y otros operadores mejorar sus servicios.

Examinamos 19,340 tweets, 7,712 publicaciones en Instagram y 25,483 reseñas en TripAdvisor sobre Granada. Las entidades analizadas fueron la Alhambra, el Albaicín, el Generalife y el Sacromonte. Según los datos de Twitter, la entidad más importante es la «Alhambra», ya que es el monumento más visitado de Granada. Sin embargo, los aspectos negativos más importantes y las opiniones o insatisfacciones de los turistas se refieren a la taquilla, el sistema de reserva de entradas online y las audioguías que narran la historia de los palacios, etc. Estas pistas son esenciales para mejorar los servicios de este complejo palacio fortificado. El conjunto de datos de Twitter, sin embargo, no contiene aspectos u opiniones negativas para las demás entidades. Los datos de TripAdvisor también muestran que la entidad más importante es la «Alhambra» y los turistas se mostraron más insatisfechos con el personal y la organización, las colas, el sistema de reserva de entradas, la atención al cliente, la logística, la inseguridad de la página web, la información al visitante, la audioguía, el sistema de entradas, etc. Otra entidad es el Albaicín y en este caso, los turistas expresaron su insatisfacción con las tiendas de souvenirs, los guías, la mala señalización, etc. El Generalife es otra entidad y los aspectos y opiniones negativas mencionan el horrible anfiteatro, la mala organización, el mal recorrido, la confusa página web, etc. El Sacromonte es otra entidad con turistas que expresan su descontento mencionando los terribles espectáculos de flamenco, pésimas actuaciones de baile, etc. En la misma línea, los datos de Instagram reflejan que la entidad más importante es la «Alhambra» y algunos de los aspectos y opiniones más importantes incluyen malas impresiones, salas vacías, gente molesta, etc.

Como demostraron los resultados, TripAdvisor y Twitter proporcionan más información subjetiva que Instagram. En cuanto al número de aspectos mostrados, TripAdvisor podría reflejar un mayor uso para difundir quejas sobre un determinado recurso, Twitter tiene un número moderado de factores de insatisfacción, mientras que los usuarios de Instagram comparten noticias y actividades de ocio (es decir, historias tristes, historias de amor trágicas, malas películas, empresas terribles, etc.).

Este análisis puede utilizarse tanto para evaluar la situación actual de los lugares, eventos y servicios relacionados con un destino o evento turístico concreto como para realizar una auditoría independiente que permita identificar los problemas para poder tomar las medidas necesarias e introducir mejoras.

A continuación, se presenta el capítulo final de esta tesis donde se mencionan algunas observaciones finales; además, se presenta el trabajo futuro en esta línea y se listan las publicaciones realizadas en este proceso de investigación.

Capítulo 6

Conclusiones Generales

6.1. Observaciones finales

En esta tesis hemos realizado un análisis exhaustivo del uso de redes sociales en el dominio turístico a través de técnicas computacionales, siendo el análisis de texto y en concreto el análisis de sentimientos un factor clave para entender las percepciones que un usuario tiene sobre los destinos turísticos.

En el primer capítulo se analizó el impacto del turismo en las economías, siendo de vital importancia en economías emergentes debido a que ha sido incluido en la Agenda 2030 para el desarrollo sostenible, un tema muy importante teniendo en cuenta el calentamiento global que nos afecta a todos. Por otro lado, si recordamos el término Smart Tourism, se refiere a la interacción y/o combinación de redes de comunicaciones, internet, sensores, internet de las cosas y el turismo; por tanto, el aporte realizado en esta tesis en el ámbito de Smart-Tourism es significativo, debido a que los datos que se analizaron y que en algunos casos sirvieron para entrenar un algoritmo, provienen de redes sociales que los usuarios usan mientras están visitando algún lugar y que para ello, es necesario las redes de comunicaciones, su dispositivo móvil e internet. Además, todas las herramientas que se han utilizado van en concordancia con la ciencia de datos utilizando sobre todo la minería de texto, el aprendizaje de máquina y el aprendizaje profundo.

Como en todo trabajo de investigación es necesario realizar una revisión sistemática de literatura para saber y entender qué es lo que existe en esta línea de investigación; por tanto, en el capítulo dos se analizan los trabajos más relevantes. Luego de definir los criterios de búsqueda se encontraron 62 trabajos publicados desde el año 2015 hasta septiembre de 2021, siendo TripAdvisor, Flickr, Twitter y Sina Weibo las plataformas que más utilizan los investigadores

en este ámbito y uno de los temas que tiene mucha relevancia es la satisfacción turística, que se abordó ampliamente en los capítulos 4 y 5. Además, se abordó la influencia de redes sociales y el turismo encontrando muchos estudios que se basan en como las reseñas de los influencers tiene impacto sobre los visitantes en el ámbito del alojamiento. Otros autores exploran como los turistas interactúan con las redes sociales previo, durante y después de su viaje. Además, en un artículo se propone un índice de redes sociales para medir la participación turística en la gestión. Sin embargo, muchos de ellos utilizan los cuestionarios como medio para la obtención de datos.

En el capítulo tres se propuso un framework para el análisis de datos turísticos captados desde Twitter, que consta de un proceso de recolección de datos, análisis descriptivo, limpieza del texto de cada tweet, un proceso de tokenización, eliminación de palabras irrelevantes y lematización. Posteriormente se realiza un análisis de contenido que consta de un análisis de frecuencia de palabras identificando los hashtags mas importantes y lugares más comentados, se hace análisis de sentimientos con herramientas existentes y el análisis con Latent Dirichlet Allocation (LDA) para la extracción de temas. Este enfoque se probó utilizando datos de Twitter del turismo en Granada, España. Los resultados mostraron la gran variedad de tweets escritos en muchos idiomas predominando el español e inglés; además, se evidenció que la «Alhambra» es la palabra más mencionada, la cuenta «@alhambracultura» es la cuenta que tiene la mayor cantidad retweets, likes y comentarios. El análisis de sentimientos mostró que la mayor cantidad de posts son neutros, seguido de tweets positivos y una menor cantidad de negativos. Por último, el análisis de contenido se hizo tomando en cuenta las estaciones del año (otoño-invierno y primavera-verano). Este extenso análisis se realizó tomando en cuenta datos en español e inglés. A pesar de que existen herramientas de pago que pueden ayudar a realizar todas las tareas propuestas en este enfoque, se trata de mostrar cómo realizarlo con herramientas de código abierto para que pueda ser fácilmente automatizado en un software como herramienta para la gestión turística. Finalmente, al analizar las técnicas, se pudo evidenciar que una de las más importantes es el análisis de sentimientos; sin embargo, los clasificadores analizados no fueron tan precisos. Por tanto, en el capítulo 4 se trató de mejorar este rendimiento.

En el capítulo cuatro nos enfocamos en el análisis de sentimientos a nivel de oración para el descubrimiento de un destino turístico. Primeramente, se analizó el desempeño de las herramientas de análisis de sentimientos para textos en inglés y en español. Se pudo evidenciar que para textos en inglés hay muy buenas herramientas como es el caso de TweetEval; sin embargo, para textos en español no pudimos encontrar una herramienta que tenga resultados parecidos a

TweetEval por tanto, se propuso un modelo de aprendizaje profundo para textos en español, este modelo utiliza reseñas de 30 destinos turísticos recolectados de TripAdvisor y tweets de TASS edición 2019 para su entrenamiento. Este modelo se basa en BERT y fue comparado con otros modelos de aprendizaje profundo mejorando notablemente los resultados de clasificación en comparación con otros modelos y herramientas existentes. Luego de elegir las herramientas para textos en español e inglés se procedió a extraer las entidades y características de los lugares y servicios encontrados, con el objetivo de explicar las razones por las que algunos usuarios tienen sentimientos positivos/negativos sobre un destino turístico. Además, pudimos descubrir los lugares y aspectos más importantes y también los lugares infravalorados. Por último, se pudo encontrar también las razones por las que algunos usuarios tienen opiniones negativas sobre un determinado lugar o servicio. Tanto el modelo «Spanish-BERT» como los datos de entrenamiento utilizados se han compartido para que cualquier persona pueda utilizarlos o mejorarlos. A diferencia del capítulo tres, en este capítulo se utilizaron datos de Twitter y de Instagram para analizar datos turísticos de Granada.

Finalmente, abordamos el análisis de sentimientos basado en aspectos para entender la insatisfacción de los usuarios al visitar un determinado lugar. En este capítulo se utilizan datos de Twitter, Instagram y TripAdvisor. Se propuso un framework para realizar el análisis con ASBA que incluye una fase muy importante que es el pre procesamiento de datos, la extracción de entidades, la identificación de aspectos o características del lugar o servicio con su respectiva opinión, la representación de aspectos a través de vectores (word-embedding), la clusterización de aspectos, la elaboración de resúmenes automáticos y visualización de los resultados para que puedan ser entendidos por administradores y gestores turísticos. Además, se propuso un algoritmo semi-supervisado para identificar las entidades más importantes de un destino turístico, un algoritmo basado en reglas para la identificación de aspectos y cálculo de opiniones que puede ser aplicado con cualquier idioma, un proceso de visualización de aspectos y las palabras relacionadas con las opiniones de los usuarios y se utilizaron dos herramientas que usan BERT para la elaboración de resúmenes automáticos muy cortos que dan una perspectiva general de forma rápida de cuáles son las quejas de los turistas.

A continuación mencionamos los trabajos futuros en esta línea de investigación.

6.2. Trabajo futuro

En cuanto a los datos de redes sociales, es necesario analizar las reseñas, comentarios, posts, tweets, etc. para filtrar la información falsa de la real; modelos de aprendizaje profundo conjuntamente con la identificación de características lingüísticas pueden resultar útiles para detectar la información falsa. Esto es importante debido a que dueños de hoteles, restaurantes, lugares o servicios con baja reputación pueden contratar a personas que se dediquen a escribir reseñas o posts positivos para mejorar su reputación; por el contrario una entidad turística puede utilizar «fake reviews/posts/tweets» para desprestigiar a la competencia.

Se ha demostrado que el análisis de sentimientos es una herramienta poderosa; sin embargo, los resultados pueden mejorase con conjuntos de entrenamiento mucho mayores que el que se utilizó en el capítulo 4 y combinarse con una taxonomía estandarizada para diferenciar las frases positivas, negativas o neutras, de tal forma que pueda ser utilizada en cualquier idioma, añadiendo mayor consistencia y validez a los resultados obtenidos. El análisis de sentimientos se puede utilizar para construir un modelo de predicción de la demanda turística, utilizando los datos de sitios web de turismo y de las publicaciones obtenidas desde las redes sociales con sus respectivas puntuaciones de opinión; de modo que dichos datos puedan transformarse en series temporales y transformarse en variables explicativas en modelos predictivos.

Además, se podría entrenar un modelo BERT para mejorar la identificación de la tupla aspecto-opinión, que es una etapa importante en el análisis del sentimiento basado en aspectos. La construcción de un sistema recomendador que incorpore los factores de insatisfacción y las estrategias de gestión y supervise su eficacia puede ser útil para los profesionales y gestores del sector turístico.

En la presente tesis se han analizado datos textuales; sin embargo, sería muy interesante complementarlo con el análisis de imágenes y video mediante técnicas y modelos de aprendizaje profundo, esto permitiría mejorar la identificación de los lugares más populares a través del contenido multimedia.

6.3. Lista de publicaciones

Se han publicado dos capítulos de libro que fueron presentados en dos conferencias, se publicaron dos artículos en revistas de alto impacto y un tercer artículo que está en revisión. A continuación se presentan los detalles: :

- Viñan-Ludeña, MS. (2019). A Systematic Literature Review on Social Media Analytics and Smart Tourism. In: Katsoni, V., Segarra-Oña, M. (eds)

Smart Tourism as a Driver for Culture and Sustainability. Springer Proceedings in Business and Economics. Springer, Cham. https://doi.org/10.1007/978-3-030-03910-3_25. La editorial Springer aparece en SPI (Scholarly Publishers Indicators), en la posición 4 de 259 en la categoría general (con un ICEE, Indicador de Calidad de Editoriales según los Expertos, de 33,061) y en la posición 1 de 40 en la categoría de economía (con ICEE de 8,021)

- Viñán-Ludeña, M.S., de Campos, L.M., Jacome-Galarza, LR., Sinche-Freire, J. (2020). Social Media Influence: A Comprehensive Review in General and in Tourism Domain. In: Rocha, Á., Abreu, A., de Carvalho, J., Liberato, D., González, E., Liberato, P. (eds) Advances in Tourism, Technology and Smart Systems. Smart Innovation, Systems and Technologies, vol 171. Springer, Singapore. https://doi.org/10.1007/978-981-15-2024-2_3. La editorial Springer aparece en el SPI (Scholarly Publishers Indicators), en la posición 4 de 259 en la categoría general (con un ICEE, Indicador de Calidad de Editoriales según los Expertos, de 33,061) y en la posición 1 de 40 en la categoría de economía (con ICEE de 8,021)

- Viñán-Ludeña, M.S. and de Campos, L.M. (2022), Analyzing Tourist Data on Twitter: a Case Study in the Province of Granada at Spain, Journal of Hospitality and Tourism Insights, Vol. 5 No. 2, pp. 435-464. https://doi.org/10.1108/JHTI-11-2020-0209. Esta revista aparece en el Journal Citation Reports (JCR) y en Emerging Sources Citation Index (ESCI), y pertenece a la categoría *Hospitality, Leisure, Sport and Tourism* con un factor de impacto de *JCI(0.65)*, rango 59/131 en cuartil Q2 (Artículo publicado).

- Viñán-Ludeña, M.S. and de Campos, L.M. (2022), Discovering a tourism destination with social media data: BERT-based sentiment analysis, Journal of Hospitality and Tourism Technology, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JHTT-09-2021-0259. La revista aparece en el Journal Citation Reports (JCR) y Social Sciences Citation Index (SSCI) y pertenece a la categoría *Hospitality, Leisure, Sport and Tourism* con un factor de impacto de *5.576*, rango *20/57* en cuartil *Q2* (Artículo publicado).

- Viñán-Ludeña, M.S. and de Campos, L.M. (2022), Evaluating Tourist Dissatisfaction with Aspect-based Sentiment Analysis using Social Media Data. La revista a la que se envió este trabajo es *Information Technology &*

Tourism, aparece en el Journal Citation Reports (JCR) y Social Sciences Citation Index (SSCI), y tiene un factor de impacto de *6.093* con rango *18/57* en cuartil *Q2* (En revisión).

Bibliografía

Abbasi-Moud, Zahra, Hamed Vahdat-Nejad y Wathiq Mansoor (2019). «Detecting Tourist's Preferences by Sentiment Analysis in Smart Cities». En: *2019 IEEE Global Conference on Internet of Things, GCIoT 2019*. Institute of Electrical y Electronics Engineers Inc.

Abd-Alrazaq, Ala, Dari Alhuwail, Mowafa Househ, Mounir Hai y Zubair Shah (2020). «Top concerns of tweeters during the COVID-19 pandemic: A surveillance study». En: *Journal of Medical Internet Research* 22.4.

Afzaal, Muhammad, Muhammad Usman y Alvis Fong (2019). «Tourism Mobile App With Aspect-Based Sentiment Classification Framework for Tourist Reviews». En: *IEEE Transactions on Consumer Electronics* 65.2, págs. 233-242.

Aggarwal, Shikha y Alekh Gour (2020). «Peeking inside the minds of tourists using a novel web analytics approach». En: *Journal of Hospitality and Tourism Management* 45, págs. 580-591.

Ahani, Ali, Mehrbakhsh Nilashi, Elaheh Yadegaridehkordi, Louis Sanzogni, A. Rashid Tarik, Kathy Knox, Sarminah Samad y Othman Ibrahim (2019). «Revealing customers' satisfaction and preferences through online review analysis: The case of Canary Islands hotels». En: *Journal of Retailing and Consumer Services* 51, págs. 331-343.

Alaei, Ali Reza, Susanne Becken y Bela Stantic (2019). «Sentiment Analysis in Tourism: Capitalizing on Big Data». En: *Journal of Travel Research* 58.2, págs. 175-191.

Alcoba, J., S. Mostajo, R. Paras y R.A. Ebron (2017). «Beyond quality of service: exploring what tourists really value». En: *Lecture Notes in Business Information Processing* 279, págs. 261-271.

Alekseev, Anton, Elena Tutubalina, Valentin Malykh y Sergey Nikolenko (2020). «Improving unsupervised neural aspect extraction for online discussions using out-of-domain classification». En: *Journal of Intelligent & Fuzzy Systems* 39.2, págs. 2487-2496.

Ali, Twil, Bidan Marc, Bencharef Omar, Kaloun Soulaimane y Safaa Larbi (2021). «Exploring destination's negative e-reputation using aspect based

sentiment analysis approach: Case of Marrakech destination on TripAdvisor». En: *Tourism Management Perspectives* 40, pág. 100892.

Almeida-Santana, Arminda y Sergio Moreno-Gil (2017). «New trends in information search and their influence on destination loyalty: Digital destinations and relationship marketing». En: *Journal of Destination Marketing & Management* 6.2. Special edition on Digital Destinations, págs. 150-161.

Ampountolas, Apostolos, Gareth Shaw y Simon James (2019). «The role of social media as a distribution channel for promoting pricing strategies». En: *Journal of Hospitality and Tourism Insights* 2.1, págs. 75-91.

An, H.-W., K. Kim y N. Moon (2020). «Design of Establishment System of Satisfaction Index for Tourist Sites According to the Weather Using Deep Neural Network». En: *Lecture Notes in Electrical Engineering* 536 LNEE, págs. 310-315.

Annisa, R., I. Surjandari y Zulkarnain (2019). «Opinion mining on mandalika hotel reviews using latent dirichlet allocation». En: vol. 161, págs. 739-746.

Ansar, Wazib, Saptarsi Goswami, Amlan Chakrabarti y Basabi Chakraborty (2021). «An efficient methodology for aspect-based sentiment analysis using BERT through refined aspect extraction». En: *Journal of Intelligent & Fuzzy Systems* 40.5, págs. 9627-9644.

Aoudi, Samer y Asif Malik (2019). «Lexicon Based Sentiment Comparison of iPhone and Android Tweets during the Iran-Iraq Earthquake». En: *ITT 2018 - Information Technology Trends: Emerging Technologies for Artificial Intelligence*. Institute of Electrical y Electronics Engineers Inc., págs. 233-238.

Arora, Amit, Anshu Saxena Arora y Shailendra Palvia (2014). «Social Media Index Valuation: Impact of Technological, Social, Economic, and Ethical Dimensions». En: *Journal of Promotion Management* 20.3, págs. 328-344.

Arora, Anuja, Shivam Bansal, Chandrashekhar Kandpal, Reema Aswani y Yogesh Dwivedi (2019). «Measuring social media influencer index- insights from facebook, Twitter and Instagram». En: *Journal of Retailing and Consumer Services* 49, págs. 86-101.

Arthur, Rudy e Hywel T.P. Williams (2019). «Scaling laws in geo-located Twitter data». En: *PLoS ONE* 14.7.

Assaker, Guy (2020). «Age and gender differences in online travel reviews and user-generated-content (UGC) adoption: extending the technology acceptance model (TAM) with credibility theory». En: *Journal of Hospitality Marketing and Management* 29.4, págs. 428-449.

Athuraliya, B. y C. Farook (2018). «"Revyew" Hotel Maintenance Issue Classifier and Analyzer using Machine Learning and Natural Language Processing». En: *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, págs. 274-280.

Bae, Sung Joo, Hyeonsuh Lee, Eung-Kyo Suh y Kil-Soo Suh (2017). «Shared experience in pretrip and experience sharing in posttrip: A survey of Airbnb users». En: *Information & Management* 54.6. Smart Tourism: Traveler, Business, and Organizational Perspectives, págs. 714-727.

Baggio, Rodolfo, Roberto Micera y Giacomo Del Chiappa (2020). «Smart tourism destinations: a critical reflection». En: 11.3, págs. 407-423.

Baralla., Gavina, Simona Ibba. y Riccardo Zenoni. (2017). «Aposentu: A Social Semantic Platform for Hotels». En: *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR,* INSTICC. SciTePress, págs. 269-274.

Barbieri, Francesco, Jose Camacho-Collados, Luis Espinosa Anke y Leonardo Neves (nov. de 2020). «TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification». En: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, págs. 1644-1650.

Bick, Markus, Katherina Bruns, Jens Sievert y Frank Jacob (2012). «Value-in-use of mobile technologies». En: *MMS 2012: Mobile und Ubiquitäre Informations Systeme*. Ed. por Andrea Back, Markus Bick, Martin Breunig, Key Pousttchi y Frédéric Thiesse. Bonn: Gesellschaft für Informatik e.V., págs. 56-67.

Bigné, Enrique, Enrique Oltra y Luisa Andreu (2019). «Harnessing stakeholder input on Twitter: A case study of short breaks in Spanish tourist cities». En: *Tourism Management* 71, págs. 490-503.

Blei, D M, A Y Ng y M I Jordan (2003). *Latent Dirichlet Allocation Michael I. Jordan*. Inf. téc., págs. 993-1022. URL: http://www.jmlr.org/papers/v3/blei03a.

Blei, David M. y Padhraic Smyth (2017). «Science and data science». En: *Proceedings of the National Academy of Sciences* 114.33, págs. 8689-8692.

Bohr, Jeremiah (2020). «Reporting on climate change: A computational analysis of U.S. newspapers and sources of bias, 1997–2017». En: *Global Environmental Change* 61.

Boley, B. Bynum, Evan J. Jordan, Carol Kline y Whitney Knollenberg (2018). «Social return and intent to travel». En: *Tourism Management* 64, págs. 119-128.

Brandt, Tobias, Johannes Bendler y Dirk Neumann (2017). «Social media analytics and value creation in urban smart tourism ecosystems». En: *Information*

& Management 54.6. Smart Tourism: Traveler, Business, and Organizational Perspectives, págs. 703-713.

Brooker, Phillip, William Dutton y Christian Greiffenhagen (2017). «What would Wittgenstein say about social media?» En: *Qualitative Research* 17.6, págs. 610-626.

Buccoliero, Luca, Elena Bellio, Giulia Crestini y Alessandra Arkoudas (2020). «Twitter and politics: Evidence from the US presidential elections 2016». En: *Journal of Marketing Communications* 26.1, págs. 88-114.

Bueno, I., R.A. Carrasco, R. Ureña y E. Herrera-Viedma (2019). «Application of an opinion consensus aggregation model based on OWA operators to the recommendation of tourist sites». En: *Procedia Computer Science* 162, págs. 539-546.

Buhalis, Dimitrios y Marie Foerste (2015). «SoCoMo marketing for travel and tourism: Empowering co-creation of value». En: *Journal of Destination Marketing & Management* 4.3. Smart Destinations, págs. 151-161.

Cajachahua, Luis e Indira Burga (2017). «Sentiments and opinions from Twitter about Peruvian touristic places using correspondence analysis». En: *CEUR Workshop Proceedings*. Vol. 2029, págs. 178-189.

Cambria, Erik, Soujanya Poria, Rajiv Bajpai y Bjoern Schuller (dic. de 2016). «SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives». En: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, págs. 2666-2677.

Carenini, Giuseppe, Jackie Chi Kit Cheung y Adam Pauls (2013). «Multidocument summarization of evaluative text». En: *Computational Intelligence* 29.4, págs. 545-576.

Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang y Jorge Pérez (2020). «Spanish Pre-Trained BERT Model and Evaluation Data». En: *PML4DC at ICLR 2020*.

Chandawarkar, Akash A, Daniel J Gould y W Grant Stevens (mar. de 2018). «The Top 100 Social Media Influencers in Plastic Surgery on Twitter: Who Should You Be Following?» En: *Aesthetic Surgery Journal* 38.8, págs. 913-917.

Chang, Hsin-Lu, Yen-Chun Chou, Dai-Yu Wu y Sou-Chein Wu (2018). «Will firm's marketing efforts on owned social media payoff? A quasi-experimental analysis of tourism products». En: *Decision Support Systems* 107, págs. 13-25.

Chang, Yung Chun, Chih Hao Ku y Chien Hung Chen (2020). «Using deep learning and visual analytics to explore hotel reviews and responses». En: *Tourism Management* 80.

Chang, Yung-Chun, Chih-Hao Ku y Chun-Hung Chen (2019). «Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor». En: *International Journal of Information Management* 48, págs. 263-279.

Chaudhari, V. A., V. Kshirsagar y M. Nagori (2018). «Integrating Sentiment Analysis and User Descriptors with Ratings in Sightseer Recommender System». En: *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, págs. 1-7.

Chen, Bingyang, Lulu Fan y Xiaobao Fu (2019). «Sentiment classification of tourism based on rules and LDA topic model». En: *Proceedings - 2019 International Conference on Electronic Engineering and Informatics, EEI 2019*. Institute of Electrical y Electronics Engineers Inc., págs. 471-475.

Chen, Wen, Zhiyun Xu, Xiaoyao Zheng, Qingying Yu y Yonglong Luo (2020). «Research on Sentiment Classification of Online Travel Review Text». En: *Applied Sciences* 10.15, pág. 5275.

Chua, Alvin, Loris Servillo, Ernesto Marcheggiani y Andrew Vande Moere (2016). «Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy». En: *Tourism Management* 57, págs. 295-310.

Chung, Namho y Heejeong Han (2017). «The relationship among tourists' persuasion, attachment and behavioral changes in social media». En: *Technological Forecasting and Social Change* 123, págs. 370-380.

Chung, Namho, Heejeong Han y Chulmo Koo (2015). «Adoption of travel information in user-generated content on social media: the moderating effect of social presence». En: *Behaviour & Information Technology* 34.9, págs. 902-919.

Chung, Namho y Chulmo Koo (2015). «The use of social media in travel information search». En: *Telematics and Informatics* 32.2, págs. 215-229.

Chung, Namho, Inessa Tyan y Hee Chung Chung (2017). «Social Support and Commitment within Social Networking Site in Tourism Experience». En: *Sustainability* 9.11.

Claster, William B., Malcolm Cooper y Philip Sallis (2010). «Thailand - Tourism and conflict. Modeling sentiment from twitter tweets using naïve bayes and unsupervised artificial neural nets». En: *Proceedings - 2nd International Conference on Computational Intelligence, Modelling and Simulation, CIMSim 2010*, págs. 89-94.

Colditz, Jason B., Kar Hai Chu, Sherry L. Emery, Chandler R. Larkin, A. Everette James, Joel Welling y Brian A. Primack (2018). *Toward real-Time infoveillance of twitter health messages.*

Corallo, A., A. Trono, L. Fortunato, F. Pettinato y L. Schina (2018). «Cultural event management and urban e-planning through bottom-up user participation». En: *International Journal of E-Planning Research* 7.1, págs. 15-33.

Costa Liberato, Pedro Manuel da, Elisa Alén-González y Dália Filipa Veloso de Azevedo Liberato (2018). «Digital Technology in a Smart Tourist Destination: The Case of Porto». En: *Journal of Urban Technology* 25.1, págs. 75-97.

Curlin, Tamara, Božidar Jaković e Ivan Miloloža (2019). «Twitter usage in Tourism: Literature Review». En: *Business Systems Research Journal* 10.1, págs. 102-119.

Dai, Weijia (Daisy), Ginger Jin, Jungmin Lee y Michael Luca (2018). «Aggregation of consumer ratings: an application to Yelp.com». En: *Quantitative Marketing and Economics* 16.3, págs. 289-339.

DeAndrea, David C., Brandon Van Der Heide, Megan A. Vendemia y Mao H. Vang (2018). «How People Evaluate Online Reviews». En: *Communication Research* 45.5, págs. 719-736.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee y Kristina Toutanova (jun. de 2019). «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». En: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, págs. 4171-4186.

Di Fabbrizio, Giuseppe, Amanda Stent y Robert Gaizauskas (2014). «A hybrid approach to multi-document summarization of opinions in reviews». En: *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, págs. 54-63.

Dragoni, Mauro, Marco Federici y Andi Rexha (2019). «An unsupervised aspect extraction strategy for monitoring real-time reviews stream». En: *Information Processing & Management* 56.3, págs. 1103-1118.

Dragouni, Mina, George Filis, Konstantinos Gavriilidis y Daniel Santamaria (2016). «Sentiment, mood and outbound tourism demand». En: *Annals of Tourism Research* 60, págs. 80-96.

Dupré, Damien, Gary McKeown, Nicole Andelic y Gawain Morrison (2018). «Willingness to Share Emotion Information on Social Media: Influence of Personality and Social Context». En: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, págs. 665-672.

Ekinci, Ekin y Sevinç İlhan Omurca (2020). «Concept-LDA: Incorporating Babelfy into LDA for aspect extraction». En: *Journal of Information Science* 46.3, págs. 406-418.

Falk, Martin Thomas y Eva Hagsten (2021). «Visitor flows to World Heritage Sites in the era of Instagram». En: *Journal of Sustainable Tourism* 29.10, págs. 1547-1564.

Farisi, Arif Abdurrahman, Yuliant Sibaroni y Said Al Faraby (2019). «Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier». En: *Journal of Physics: Conference Series*. Vol. 1192. 1. Institute of Physics Publishing, pág. 012024.

Feizollah, A., S. Ainin, N. B. Anuar, N. A. B. Abdullah y M. Hazim (2019). «Halal Products on Twitter: Data Extraction and Sentiment Analysis Using Stack of Deep Learning Algorithms». En: *IEEE Access* 7, págs. 83354-83362.

Feizollah, Ali, Sulaiman Ainin, Nor Badrul Anuar, Nor Aniza Binti Abdullah y Mohamad Hazim (2019). «Halal Products on Twitter: Data Extraction and Sentiment Analysis Using Stack of Deep Learning Algorithms». En: *IEEE Access* 7, págs. 83354-83362.

Fernandes, Teresa y Filipa Fernandes (2018). «Sharing Dissatisfaction Online: Analyzing the Nature and Predictors of Hotel Guests Negative Reviews». En: *Journal of Hospitality Marketing & Management* 27.2, págs. 127-150.

Filieri, Raffaele, Dorothy A. Yen y Qionglei Yu (2021). «#ILoveLondon: An exploration of the declaration of love towards a destination on Instagram». En: *Tourism Management* 85, pág. 104291.

Francalanci, Chiara y Ajaz Hussain (2015). «NavigTweet: A Visual Tool for Influence-Based Twitter Browsing». En: Springer International Publishing, págs. 183-198.

— (2017). «Influence-based Twitter browsing with NavigTweet». En: *Information Systems* 64, págs. 119-131.

Freberg, Karen, Kristin Graham, Karen McGaughey y Laura A. Freberg (2011). «Who are the social media influencers? A study of public perceptions of personality». En: *Public Relations Review* 37.1, págs. 90-92.

Fu, Yu, Jin-Xing Hao, Xiang (Robert) Li y Cathy H.C. Hsu (2019). «Predictive Accuracy of Sentiment Analytics for Tourism: A Metalearning Perspective on Chinese Travel News». En: *Journal of Travel Research* 58.4, págs. 666-679.

Fuentes-Medina, M.L., E. Hernández-Estárico y S. Morini-Marrero (2018). «Study of the critical success factors of emblematic hotels through the analysis of content of online opinions: The case of the Spanish Tourist Paradors». En: *European Journal of Management and Business Economics* 27.1, págs. 42-65.

Gao, Jinfeng, Ruxian Yao, Han Lai y Haitao Wu (2019). «Sentiment Analysis of Tourism Reviews: An exploratory study based on CNNs built on LSTM model». En: *ICEB 2019 Proceedings (Newcastle Upon Tyne, UK). 55.*

García, A., S. Gaines y M.T. Linaza (2012). «A Lexicon based sentiment analysis retrieval system for tourism domain». En: *e-Review of Tourism Research* 10.2, págs. 35-38.

Gede Suardika, I. (2016). «Sentiment analysis system and correlation analysis on hospitality in Bali». En: *Journal of Theoretical and Applied Information Technology* 84.1, págs. 88-95.

Gerani, Shima, Giuseppe Carenini y Raymond T. Ng (2019). «Modeling content and structure for abstractive review summarization». En: *Computer Speech & Language* 53, págs. 302-331.

Giglio, Simona, Eleonora Pantano, Eleonora Bilotta y T.C. Melewar (2020). «Branding luxury hotels: Evidence from the analysis of consumers' "big" visual data on TripAdvisor». En: *Journal of Business Research* 119, págs. 495-501.

Gkritzali, A., D. Gritzalis y V. Stavrou (2018). «Is Xenios Zeus Still Alive? Destination Image of Athens in the Years of Recession». En: *Journal of Travel Research* 57.4, págs. 540-554.

González, M.D.M., E.M. Cámara, M.T.M. Valdivia y S.M.J. Zafra (2015). «ESOLHotel: Building an Spanish opinion lexicon adapted to the tourism domain [eSOLHotel: Generación de un lexicón de opinión en español adaptado al dominio turístico]». En: *Procesamiento de Lenguaje Natural* 54, págs. 21-28.

Gretzel, Ulrike, Marianna Sigala, Zheng Xiang y Chulmo Koo (2015). «Smart tourism: foundations and developments». En: *Electronic Markets* 25.3, págs. 179-188.

Gruhl, Daniel, Meena Nagarajan, Jan Pieper, Christine Robson y Amit Sheth (2010). «Multimodal social intelligence in a real-time dashboard system». En: *The VLDB Journal* 6, págs. 825-848.

Gu, Y. H., S. J. Yoo, Z. Jiang, Y. J. Lee, Z. Piao, H. Yin y S. Jeon (2018). «Sentiment analysis and visualization of Chinese tourism blogs and reviews». En: *2018 International Conference on Electronics, Information, and Communication (ICEIC)*, págs. 1-4.

Guevara, J., J. Costa, J. Arroba y C. Silva (2018). «Harvesting opinions in Twitter for sentiment analysis». En: *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*, págs. 1-7.

Gulnerman, A.G. y H. Karaman (2020). «Spatial reliability assessment of social media mining techniques with regard to disaster domain-based filtering». En: *ISPRS International Journal of Geo-Information* 9.4.

Gunter, Ulrich e Irem Önder (2021). «An Exploratory Analysis of Geotagged Photos From Instagram for Residents of and Visitors to Vienna». En: *Journal of Hospitality & Tourism Research* 45.2, págs. 373-398.

Guo, Yue, Stuart J. Barnes y Qiong Jia (2017). «Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation». En: *Tourism Management* 59, págs. 467-483.

Halpern, Daniel, Sebastián Valenzuela y James E. Katz (nov. de 2017). «We Face, I Tweet: How Different Social Media Influence Political Participation through Collective and Internal Efficacy». En: *Journal of Computer-Mediated Communication* 22.6, págs. 320-336.

Hao, Jin Xing, Rui Wang, Rob Law y Yan Yu (2021). «How do Mainland Chinese tourists perceive Hong Kong in turbulence? A deep learning approach to sentiment analytics». English. En: *International Journal of Tourism Research* 23.4, págs. 478-490.

Hassan, Shahizan, Norshuhada Shiratuddin, Nor Laily Hashim y Feng Li (2018). «Evaluating Social Media: Towards a Practical Model for Measuring Social Media Influence». En: *Media Influence*. IGI Global, págs. 293-309.

He, Jin, Lei Li, Yan Wang y Xindong Wu (2021). «Hierarchical features-based targeted aspect extraction from online reviews». En: *Intelligent Data Analysis* 25.1, 205-223.

Hew, Jun-Jie, Garry Wei-Han Tan, Binshan Lin y Keng-Boon Ooi (2017). «Generating travel-related contents through mobile social tourism: Does privacy paradox persist?» En: *Telematics and Informatics* 34.7, págs. 914-935.

Hochreiter, Sepp y Jürgen Schmidhuber (nov. de 1997). «Long Short-Term Memory». En: *Neural Computation* 9.8, págs. 1735-1780.

Hoshino, Yuko, Eriko Ishii y Mitsuho Yamada (2018). «A Study of Recommended Tourist Spot Information Extraction Using SNS». English. En: *International Conference on Tourism Research*, págs. 259-262,IX.

Hou, Zhiping, Fasheng Cui, Yongheng Meng, Tonghui Lian y Caihua Yu (2019). «Opinion mining from online travel reviews: A comparative analysis of Chinese major OTAs using semantic association analysis». En: *Tourism Management* 74, págs. 276-289.

Hu, Minqing y Bing Liu (2004). «Mining Opinion Features in Customer Reviews». En: *Proceedings of the 19th National Conference on Artifical Intelligence*. AAAI'04. San Jose, California: AAAI Press, 755–760.

Hu, Nan, Ting Zhang, Baojun Gao e Indranil Bose (2019). «What do hotel customers complain about? Text analysis using structural topic model». En: *Tourism Management* 72, págs. 417-426.

Hu, Ya-Han y Kuanchin Chen (2016). «Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings». En: *International Journal of Information Management* 36.6, Part A, págs. 929-944.

Hu, Ya-Han, Yen-Liang Chen y Hui-Ling Chou (2017). «Opinion mining from online hotel reviews – A text summarization approach». En: *Information Processing & Management* 53.2, págs. 436-449.

Huang, C. Derrick, Jahyun Goo, Kichan Nam y Chul Woo Yoo (2017). «Smart tourism technologies in travel planning: The role of exploration and exploitation». En: *Information & Management* 54.6. Smart Tourism: Traveler, Business, and Organizational Perspectives, págs. 757-770.

Hudson, Simon, Martin S. Roth, Thomas J. Madden y Rupert Hudson (2015). «The effects of social media on emotions, brand relationship quality, and word of mouth: An empirical study of music festival attendees». En: *Tourism Management* 47, págs. 68-76.

Huertas, Assumpció (2018). «How live videos and stories in social media influence tourist opinions and behaviour». En: *Information Technology & Tourism* 19.1, págs. 1-28.

Imane, E.H. e I. Abdelouahab (2019). «Social big data analysis of Five Star hotels: A case study of hotel guest experience and satisfaction in Marrakech». En: *African Journal of Hospitality, Tourism and Leisure* 8.3.

Inversini, Alessandro e Isabella Rega (2020). «Digital Communication and Tourism for Development». En: *Handbook of Communication for Development and Social Change*. Singapore: Springer Singapore, págs. 667-677.

Joseph, Nimish, Arpan Kumar Kar, P. Vigneswara Ilavarasan y Shankar Ganesh (2017). «Review of Discussions on Internet of Things (IoT): Insights from Twitter Analytics». En: *Journal of Global Information Management* 25.2, págs. 38-51.

Kemp, Simon (2020). *Digital 2021: 60 percent of the world's population is now online - We Are Social*. URL: https://wearesocial.com/blog/2021/04/60-percent-of-the-worlds-population-is-now-online (visitado 26-07-2021).

Kim, Kun, Oun joung Park, Seunghyun Yun y Haejung Yun (2017a). «What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management». En: *Technological Forecasting and Social Change* 123, págs. 362-369.

Kim, M y C Hong (2017). «Unstructured Social Media Data Mining System Based on Emotional Database and Unstructured Information Management Architecture Framework». En: *Advanced Science Letters* 23.3, págs. 1668-1672.

Kim, Myung Ja, Choong-Ki Lee y Mark Bonn (2017). «Obtaining a better understanding about travel-related purchase intentions among senior users of mobile social network sites». En: *International Journal of Information Management* 37.5, págs. 484-496.

Kim, Sung-Eun, Kyung Young Lee, Soo Il Shin y Sung-Byung Yang (2017b). «Effects of tourism information quality in social media on destination image formation: The case of Sina Weibo». En: *Information & Management* 54.6. Smart Tourism: Traveler, Business, and Organizational Perspectives, págs. 687-702.

Kim, Yae-Ji y Hak-Seon Kim (2022). «The Impact of Hotel Customer Experience on Customer Satisfaction through Online Reviews». En: *Sustainability* 14.2.

Kim, Yoonsang, Rachel Nordgren y Sherry Emery (2020). «The story of goldilocks and three twitter's APIs: A pilot study on twitter data sources and disclosure». En: *International Journal of Environmental Research and Public Health* 17.3.

Kirilenko, Andrei P., Svetlana O. Stepchenkova, Hany Kim y Xiang (Robert) Li (2018). «Automated Sentiment Analysis in Tourism: Comparison of Approaches». En: *Journal of Travel Research* 57.8, págs. 1012-1025.

Kiráľová, Alžbeta y Antonín Pavlíčeka (2015). «Development of Social Media Strategies in Tourism Destination». En: *Procedia - Social and Behavioral Sciences* 175. Proceedings of the 3rd International Conference on Strategic Innovative Marketing (IC-SIM 2014), págs. 358-366.

Krohn, Jon, Grant Beyleveld y Aglae Bassens (2020). *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence*. Pearsonś Addison-Wesley.

Kuhamanee, T., N. Talmongkol, K. Chaisuriyakul, W. San-Um, N. Pongpisuttinun y S. Pongyupinpanich (2017). «Sentiment analysis of foreign tourists to Bangkok using data mining through online social network». En: *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, págs. 1068-1073.

Kumar, Ashish, Mayank Saini y Aditi Sharan (2020). «Aspect category detection using statistical and semantic association». En: *Computational Intelligence* 36.3, 1161-1182.

Lam-González, Yen E., Richard Clouet, Niurka Cruz Sosa y Javier de León (2021). «Dissatisfaction Responses of Tourists in the Havana World Heritage Site». En: *Sustainability* 13.19.

Lamest, Markus y Mairead Brady (2019). «Data-focused managerial challenges within the hotel sector». En: 74.1, págs. 104-115.

Laoh, Enrico, Isti Surjandari y Nadhila Idzni Prabaningtyas (2019). «Enhancing hospitality sentiment reviews analysis performance using SVM N-grams method». En: *2019 16th International Conference on Service Systems and*

Service Management, ICSSSM 2019. Institute of Electrical y Electronics Engineers Inc.

Lecun, Yann, Yoshua Bengio y Geoffrey Hinton (2015). «Deep learning». En: *Nature* 521.7553, págs. 436-444.

Lee, Hyunae, Namho Chung y Yoonjae Nam (2019). «Do online information sources really make tourists visit more diverse places?: Based on the social networking analysis». En: *Information Processing & Management* 56.4, págs. 1376-1390.

Lee, Minwoo, Yanjun (Maggie) Cai, Agnes DeFranco y Jongseo Lee (2020). «Exploring influential factors affecting guest satisfaction». En: *Journal of Hospitality and Tourism Technology* 11.1, págs. 137-153.

Lee, Minwoo, Jung Hwa Hong, Sunghun Chung y Ki-Joon Back (2021). «Exploring the Roles of DMO's Social Media Efforts and Information Richness on Customer Engagement: Empirical Analysis on Facebook Event Pages». En: *Journal of Travel Research* 60.3, págs. 670-686.

Lee, Seung Jae (2017). «A review of audio guides in the era of smart tourism». En: *Information Systems Frontiers* 19.4, págs. 705-715.

Leoni, Veronica (2020). «Stars vs lemons. Survival analysis of peer-to peer marketplaces: the case of Airbnb». En: *Tourism Management* 79, pág. 104091.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov y Luke Zettlemoyer (2019). «BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension». En: *CoRR* abs/1910.13461. URL: http://arxiv.org/abs/1910.13461.

Li, Q., S. Li, S. Zhang, J. Hu y J. Hu (2019). «A review of text corpus-based tourism big data mining». En: *Applied Sciences (Switzerland)* 9.16.

Li, Qin, Shaobo Li, Jie Hu, Sen Zhang y Jianjun Hu (2018). «Tourism Review Sentiment Classification Using a Bidirectional Recurrent Neural Network with an Attention Mechanism and Topic-Enriched Word Vectors». En: *Sustainability* 10.9.

Li, Wei, Luyao Zhu, Yong Shi, Kun Guo y Erik Cambria (2020a). «User reviews: Sentiment analysis using lexicon integrated two-channel CNN–LSTM family models». En: *Applied Soft Computing* 94, pág. 106435.

Li, Xuelian, Bi Wang, Lixin Li, Zhiqiang Gao, Qian Liu, Hancheng Xu y Lanting Fang (2020b). «Deep2s: Improving Aspect Extraction in Opinion Mining With Deep Semantic Representation». En: *IEEE Access* 8, págs. 104026-104038.

Li, Yunpeng, Clark Hu, Chao Huang y Liqiong Duan (2017). «The concept of smart tourism in the context of tourism information services». En: *Tourism Management* 58, págs. 293-300.

Liang, X., P. Liu y Z. Wang (2019). «Hotel selection utilizing online reviews: A novel decision support model based on sentiment analysis and DL-VIKOR method». En: *Technological and Economic Development of Economy* 25.6, págs. 1139-1161.

Liao, Jian, Suge Wang, Deyu Li y Xiaoli Li (2017). «FREERL: Fusion relation embedded representation learning framework for aspect extraction». En: *Knowledge-Based Systems* 135, págs. 9-17.

Lin, Michael S., Yun Liang, Joanne X. Xue, Bing Pan y Ashley Schroeder (2021). «Destination image through social media analytics and survey method». En: *International Journal of Contemporary Hospitality Management* 33.6, págs. 2219-2238.

Litvin, Stephen W., Ronald E. Goldsmith y Bing Pan (2008). «Electronic word-of-mouth in hospitality and tourism management». En: *Tourism Management* 29.3, págs. 458-468.

Liu, Bing (2010). «Sentiment analysis and subjectivity». En: *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca.*

— (2012). «Sentiment analysis and opinion mining». En: *Synthesis Lectures on Human Language Technologies* 5.1, págs. 1-184.

— (2015). «Preface». En: *Sentiment Analysis*. Cambridge: Cambridge University Press, págs. xi-xiv.

Liu, Matthew Tingchi, Yongdan Liu, Ziying Mo y Kai Lam Ng (2020). «Using text mining to track changes in travel destination image: the case of Macau». En: *Asia Pacific Journal of Marketing and Logistics*.

Liu, P., D. Nie, X. He, W. Zhang, Z. Huang y K. He (2019a). «Sentiment analysis of Chinese tourism review based on boosting and LSTM». En: págs. 664-668.

Liu, S., Y. Tian, Y. Feng e Y. Zhuang (2018). «Comparison of Tourist Thematic Sentiment Analysis Methods Based on Weibo Data». En: *Beijing Daxue Xuebao (Ziran Kexue Ban)/Acta Scientiarum Naturalium Universitatis Pekinensis* 54.4, págs. 687-692.

Liu, Yan y Xiaoqing Gu (2017). «What Contributes to Chinese Adolescents' Academic Self-Concept? —An Analysis of Social Media Influence of Peers». En: *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*, págs. 373-376.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer y Veselin Stoyanov (2019b). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*.

Liu, Zhewei, Anshu Zhang, Yepeng Yao, Wenzhong Shi, Xiao Huang y Xiaoqi Shen (2021). «Analysis of the performance and robustness of methods to detect base locations of individuals with geo-tagged social media data». En: *International Journal of Geographical Information Science* 35.3, págs. 609-627.

Luo, Qiuju y Dixi Zhong (2015). «Using social network analysis to explain communication characteristics of travel-related electronic word-of-mouth on social networking sites». En: *Tourism Management* 46, págs. 274-282.

Luo, Zhiyi, Shanshan Huang y Kenny Q. Zhu (2019). «Knowledge empowered prominent aspect extraction from product reviews». En: *Information Processing & Management* 56.3, págs. 408-423.

Ma, Yufeng, Zheng Xiang, Qianzhou Du y Weiguo Fan (2018). «Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning». En: *International Journal of Hospitality Management* 71, págs. 120-131.

MacCarthy, Martin y Hairong Shan (2022). «Machine infelicity in a poignant visitor setting: comparing human and AI's ability to analyze discourse». En: *Current Issues in Tourism* 25.8, págs. 1289-1306.

Maity, Aranyak, Sritama Ghosh, Saikat Karfa, Moutan Mukhopadhyay, Saurabh Pal y Pijush Kanti Dutta Pramanik (2020). «Sentiment analysis from travellers' reviews using enhanced conjunction rule based approach for feature-specific evaluation of hotels». En: *Journal of Statistics and Management Systems* 23.6, págs. 983-997.

Mao, Zhenxing, Yang Yang y Mingshu Wang (2018). «Sleepless nights in hotels? Understanding factors that influence hotel sleep quality». En: *International Journal of Hospitality Management* 74, págs. 189-201.

Marcacini, Ricardo Marcondes, Rafael Geraldeli Rossi, Ivone Penque Matsuno y Solange Oliveira Rezende (2018). «Cross-domain aspect extraction for sentiment analysis: A transductive learning approach». En: *Decision Support Systems* 114, págs. 70-80.

Marchiori, Elena y Lorenzo Cantoni (2015). «The role of prior experience in the perception of a tourism destination in user-generated content». En: *Journal of Destination Marketing & Management* 4.3. Smart Destinations, págs. 194-201.

Mariani, Marcello M., Marco Di Felice y Matteo Mura (2016). «Facebook as a destination marketing tool: Evidence from Italian regional Destination Management Organizations». En: *Tourism Management* 54, págs. 321-343.

Marine-Roig (2019). «Destination Image Analytics through Traveller-Generated Content». En: *Sustainability* 11.12, pág. 3392.

Marine-Roig, Estela y Salvador Anton Clavé (2015). «Tourism analytics with massive user-generated content: A case study of Barcelona». En: *Journal of Destination Marketing & Management* 4.3. Smart Destinations, págs. 162-172.

Marine-Roig, Estela, Berta Ferrer-Rosell, Natalia Daries y Eduard Cristobal-Fransi (2019). «Measuring gastronomic image online». En: *International Journal of Environmental Research and Public Health* 16.23.

Martí, Pablo, Clara García-Mayor y Leticia Serrano-Estrada (2020). «Taking the urban tourist activity pulse through digital footprints». En: *Current Issues in Tourism*, págs. 1-20.

Martinez-Torres, M. R. y S. L. Toral (2019). «A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation». En: *Tourism Management* 75, págs. 393-403.

Martín, C.A., J.M. Torres, R.M. Aguilar y S. Diaz (2018). «Using deep learning to predict sentiments: Case study in tourism». En: *Complexity* 2018.

Mate, Mani Jeremiah, Alexander Trupp y Stephen Pratt (2019). «Managing negative online accommodation reviews: evidence from the Cook Islands». En: *Journal of Travel & Tourism Marketing* 36.5, págs. 627-644.

McCartney, Glenn y Rowena Pao Cheng Pek (2018). «An Examination of Sina Weibo Travel Blogs' Influence on Sentiment towards Hotel Accommodation in Macao». En: *Journal of China Tourism Research* 14.2, págs. 146-157.

Mehraliyev, Fuad, Youngjoon Choi y Mehmet Ali Köseoglu (2019). «Progress on smart tourism research». En: *Journal of Hospitality and Tourism Technology* 10.4, págs. 522-538.

Miah, Shah Jahan, Huy Quan Vu, John Gammack y Michael McGrath (2017). «A Big Data Analytics Method for Tourist Behaviour Analysis». En: *Information & Management* 54.6. Smart Tourism: Traveler, Business, and Organizational Perspectives, págs. 771-785.

Micera, R. y R. Crispino (2017). «Destination web reputation as "smart tool" for image building: the case analysis of Naples city-destination». En: *International Journal of Tourism Cities* 3.4, págs. 406-423.

Mikolov, Tomas, Kai Chen, Greg Corrado y Jeffrey Dean (2013). «Efficient Estimation of Word Representations in Vector Space». En: *CoRR* abs/1301.3781.

Min, Hee y Seongyi Yun (2019). «Role of Social Media and Emotion in South Korea's Presidential Impeachment Protests». En: *Issues & Studies* 55.01, pág. 1950002.

Misirlis, Nikolaos y Maro Vlachopoulou (2018). «Social media metrics and analytics in marketing – S3M: A mapping literature review». En: *International Journal of Information Management* 38.1, págs. 270-276.

Monachesi, Paola (2020). «Shaping an alternative smart city discourse through Twitter: Amsterdam and the role of creative migrants». En: *Cities* 100.

Moreno-Ortiz, Antonio, Soluna Salles-Bernal y Aroa Orrequia-Barea (2019). «Design and validation of annotation schemas for aspect-based sentiment analysis in the tourism sector». En: *Information Technology & Tourism* 4, págs. 535-557.

Moro, Sérgio, Pedro Ramos, Joaquim Esmerado y Seyed Mohammad Jafar Jalali (2019). «Can we trace back hotel online reviews' characteristics using gamification features?» En: *International Journal of Information Management* 44, págs. 88-95.

Moro, Sérgio, Paulo Rita, Pedro Ramos y Joaquim Esmerado (2022). «The influence of cultural origins of visitors when staying in the city that never sleeps». En: *Tourism Recreation Research* 47.1, págs. 78-90.

Mostafa, Lamiaa (2020). «Machine Learning-Based Sentiment Analysis for Analyzing the Travelers Reviews on Egyptian Hotels». En: *Advances in Intelligent Systems and Computing*. Vol. 1153 AISC. Springer, págs. 405-413.

Nakamura, S., M. Okada y K. Hashimoto (2015). «An Investigation of Effectiveness Using Topic Information Order to Classify Tourists Reviews». En: *2015 International Conference on Computer Application Technologies*, págs. 94-97.

Nakov, Preslav, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani y Veselin Stoyanov (2019). *SemEval-2016 Task 4: Sentiment Analysis in Twitter*.

Nargundkar, Ashish e Y. S. Rao (2016). «InfluenceRank: A machine learning approach to measure influence of Twitter users». En: *2016 International Conference on Recent Trends in Information Technology (ICRTIT)*, págs. 1-6.

Nascimento Filho, Francisco Barbosa do, Luiz Carlos Da Silva Flores y Pablo Flôres Limberger (2019). «Análise do posicionamento dos restaurantes de São Paulo estrelados pelo guia Michelin com base nas On-line Travel Reviews (OTRS)». En: *Revista Brasileira de Pesquisa em Turismo* 13.2, págs. 1-15.

Nave, M., P. Rita y J. Guerreiro (2018). «A decision support system framework to track consumer sentiments in social media». En: *Journal of Hospitality Marketing and Management* 27.6, págs. 693-710.

Nguyen, Tuong Tri, David Camacho y Jai E. Jung (2017). «Identifying and ranking cultural heritage resources on geotagged social media for smart cultural tourism services». En: *Personal and Ubiquitous Computing* 21.2, págs. 267-279.

Nilashi, Mehrbakhsh, Rabab Ali Abumalloh, Behrouz Minaei-Bidgoli, Waleed Abdu Zogaan, Ashwaq Alhargan, Saidatulakmal Mohd, Sharifah Nurlaili Farhana Syed Azhar, Shahla Asadi y Sarminah Samad (2022). «Revealing

travellers' satisfaction during COVID-19 outbreak: Moderating role of service quality». En: *Journal of Retailing and Consumer Services* 64, pág. 102783.

Nusair, Khaldoon (2020). «Developing a comprehensive life cycle framework for social media research in hospitality and tourism: A bibliometric method 2002-2018». En: *International Journal of Contemporary Hospitality Management* 32.3, págs. 1041-1066.

Oh, Hoonseong y Sangmin Lee (2021). «Evaluation and Interpretation of Tourist Satisfaction for Local Korean Festivals Using Explainable AI». En: *Sustainability* 13.19.

OMT (2008). *Glosario de términos de turismo*. URL: https://www.unwto.org/es/glosario-terminos-turisticos (visitado 17-09-2021).

Ovádek, Michal (2020). «"Popular tribunes" and their agendas: topic modelling Slovak presidents' speeches 1993–2020». En: *East European Politics*.

Ozyurt, Baris y M. Ali Akcayol (2021). «A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA». En: *Expert Systems with Applications* 168, pág. 114231.

Padilla, J.J., H. Kavak, C.J. Lynch, R.J. Gore y S.Y. Diallo (2018). «Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter». En: *PLoS ONE* 13.6.

Pai, Mao Yuan, Ding Chau Wang, Tz Heng Hsu, Guan Yu Lin y Chao Chun Chen (2019). «On ontology-based tourist knowledge representation and recommendation». En: *Applied Sciences (Switzerland)* 9.23.

Palakvangsa-Na-Ayudhya, S., V. Sriarunrungreung, P. Thongprasan y S. Porcharoen (2011). «Nebular: A sentiment classification system for the tourism business». En: *2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE)*, págs. 293-298.

Palazzo, Maria, Agostino Vollero, Pierluigi Vitale y Alfonso Siano (2021). «Urban and rural destinations on Instagram: Exploring the influencers' role in #sustainabletourism». En: *Land Use Policy* 100, pág. 104915.

Paolanti, Marina, Adriano Mancini, Emanuele Frontoni, Andrea Felicetti, Luca Marinelli, Ernesto Marcheggiani y Roberto Pierdicca (2021). «Tourism destination management using sentiment analysis and geo-location information: a deep learning approach». En: *Information Technology and Tourism*, págs. 1-24.

Park, Deukhee, Woo Gon Kim y Soojin Choi (2019). «Application of social media analytics in tourism crisis communication». En: *Current Issues in Tourism* 22.15, págs. 1810-1824.

Park, Seunghyun Brian, Jinwon Kim, Yong Kyu Lee y Chihyung Michael Ok (2020). «Visualizing theme park visitors' emotions using social media analytics and geospatial analytics». En: *Tourism Management* 80, pág. 104127.

Peng, Xia y Zhou Huang (2017). «A Novel Popular Tourist Attraction Discovering Approach Based on Geo-Tagged Social Media Big Data». En: *ISPRS International Journal of Geo-Information* 6.7.

Pennington, Jeffrey, Richard Socher y Christopher D. Manning (2014). «GloVe: Global Vectors for Word Representation». En: *Empirical Methods in Natural Language Processing (EMNLP)*, págs. 1532-1543. URL: http://www.aclweb.org/anthology/D14-1162.

Piñeiro-Chousa, Juan, Marcos Vizcaíno-González y Ada María Pérez-Pico (2017). «Influence of Social Media over the Stock Market». En: *Psychology & Marketing* 34.1, págs. 101-108.

Poria, Soujanya, Erik Cambria y Alexander Gelbukh (2016). «Aspect extraction for opinion mining with a deep convolutional neural network». En: *Knowledge-Based Systems* 108. New Avenues in Knowledge Bases for Natural Language Processing, págs. 42-49.

Prakash, Supun Lahiru, Priyan Perera, David Newsome, Tharaka Kusuminda y Obelia Walker (2019). «Reasons for visitor dissatisfaction with wildlife tourism experiences at highly visited national parks in Sri Lanka». En: *Journal of Outdoor Recreation and Tourism* 25, págs. 102-112.

Prameswari, P., I. Surjandari y E. Laoh (2017). «Opinion mining from online reviews in Bali tourist area». En: *2017 3rd International Conference on Science in Information Technology (ICSITech)*, págs. 226-230.

Prameswari, P., Zulkarnain, I. Surjandari y E. Laoh (2017). «Mining online reviews in Indonesia's priority tourist destinations using sentiment analysis and text summarization approach». En: *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, págs. 121-126.

Pronoza, E., E. Yagunova y S. Volskaya (2016). «Aspect-based restaurant information extraction for the recommendation system». En: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9561 LNCS, págs. 371-385.

Punel, Aymeric y Alireza Ermagun (2018). «Using Twitter network to detect market segments in the airline industry». En: *Journal of Air Transport Management* 73, págs. 67-76.

Putri, IR y R Kusumaningrum (2017). «Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia». En: *Journal of Physics: Conference Series* 801, pág. 012073.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li y Peter J. Liu (2019). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*.

Ramanathan, Vallikannu y T. Meyyappan (2019). «Twitter text mining for sentiment analysis on people's feedback about Oman tourism». En: *2019 4th MEC International Conference on Big Data and Smart City, ICBDSC 2019*. Institute of Electrical y Electronics Engineers Inc.

Rana, Toqir A. y Yu-N Cheah (2017). «A two-fold rule-based model for aspect extraction». En: *Expert Systems with Applications* 89, págs. 273-285.

Rao, Yanghui, Jianhui Pang, Haoran Xie, An Liu, Tak Lam Wong, Qing Li y Fu Lee Wang (2017). «Supervised intensive topic models for emotion detection over short text». En: *Lecture Notes in Computer Science*. Vol. 10177 LNCS. Springer Verlag, págs. 408-422.

Rashidi, Taha H., Alireza Abbasi, Mojtaba Maghrebi, Samiul Hasan y Travis S. Waller (2017). «Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges». En: *Transportation Research Part C: Emerging Technologies* 75, págs. 197-211.

Raun, Janika, Rein Ahas y Margus Tiru (2016). «Measuring tourism destinations using mobile tracking data». En: *Tourism Management* 57, págs. 202-212.

Reiter, Lauren, Roger McHaney y Kim Y. Hiller Connell (2017). «Social media influence on purchase intentions: instrument validation». En: *International Journal of Web Based Communities* 13.1, pág. 54.

Ren, Zheng, Bin Jiang y Stefan Seipel (2019). «Capturing and characterizing human activities using building locations in America». En: *ISPRS International Journal of Geo-Information* 8.5.

Rendalkar, Shubham y Chaitali Chandankhede (2018). «Sarcasm Detection of Online Comments Using Emotion Detection». En: *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, págs. 1244-1249.

Rodrigues, Helena, Ana Brochado y Michael Troilo (2020). «Listening to the murmur of water: essential satisfaction and dissatisfaction attributes of thermal and mineral spas». En: *Journal of Travel & Tourism Marketing* 37.5, págs. 649-661.

Rosanensi, Melati, Miftahul Madani, Rizki Tri Puji Wanggono, Arief Setyanto, Andhika Agus Selameto y Sri Ngudi Wahyuni (2018). «Analysis Sentiment And Tourist Response To Rinjani Mountain Tour Based On Comments From Photo Upload In Instagram». En: *2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE)*, págs. 184-188.

Rousseeuw, Peter J. (1987). «Silhouettes: A graphical aid to the interpretation and validation of cluster analysis». En: *Journal of Computational and Applied Mathematics* 20, págs. 53-65.

Rus, Annisa Marlin Masbar, Rossi Annisa, Isti Surjandari y Zulkarnain (2019). «Measuring hotel service quality in borobudur temple using opinion mining». En: *2019 16th International Conference on Service Systems and Service Management, ICSSSM 2019*. Institute of Electrical y Electronics Engineers Inc.

Rus, Holly M. y Jitske Tiemensma (2017). «Social Media under the Skin: Facebook Use after Acute Stress Impairs Cortisol Recovery». En: *Frontiers in Psychology* 8, pág. 1609.

S P. Tussyadiah, II, Devi Roza Kausar y Primidya K. M. Soesilo (2018). «The Effect of Engagement in Online Social Network on Susceptibility to Influence». En: *Journal of Hospitality & Tourism Research* 42.2, págs. 201-223.

Samoggia, Antonella, Bettina Riedel y Arianna Ruggeri (2020). «Social media exploration for understanding food product attributes perception: the case of coffee and health with Twitter data». En: *British Food Journal*.

Sann, Raksmey y Pei-Chun Lai (2020). «Understanding homophily of service failure within the hotel guest cycle: Applying NLP-aspect-based sentiment analysis to the hospitality industry». En: *International Journal of Hospitality Management* 91, pág. 102678.

Santos, Brucce Neves Dos, Ricardo Marcondes Marcacini y Solange Oliveira Rezende (2021). «Multi-Domain Aspect Extraction Using Bidirectional Encoder Representations From Transformers». En: *IEEE Access* 9, págs. 91604-91613.

Saura, J.R., P. Palos-Sanchez y M.A.R. Martin (2018). «Attitudes expressed in online comments about environmental factors in the tourism sector: An exploratory study». En: *International Journal of Environmental Research and Public Health* 15.3.

Schweter, Stefan y Alan Akbik (2020). *FLERT: Document-Level Features for Named Entity Recognition*.

Sedera, Darshana, Sachithra Lokuge, Maura Atapattu y Ulrike Gretzel (2017). «Likes—The key to my happiness: The moderating effect of social influence on travel experience». En: *Information & Management* 54.6. Smart Tourism: Traveler, Business, and Organizational Perspectives, págs. 825-836.

Senaweera, Malith, Ruwanmalee Dissanayake, Nuwini Chamindi, Anupa Shyamlal, Charith Elvitigala, Sameera Horawalavithana, Primal Wijesekera, Kasun Gunawardana, Manjusri Wickramasinghe y Chamath Keppitiyagama (2018). «The Influence of Community Interactions on User Affinity in Social Networks: A Facebook Case Study». En: *2018 National Information Technology Conference (NITC)*, págs. 1-6.

Serna, A., A. Casellas, G. Saff y J.K. Gerrikagoitia (2018). «Big data and service quality: Barcelona's hospitality and tourism industry». En: *Bridging Tourism Theory and Practice* 9, págs. 213-227.

Shafiee, Sanaz y Ali Rajabzadeh Ghatari (2016). «Big data in tourism industry». En: *2016 10th International Conference on e-Commerce in Developing Countries: with focus on e-Tourism (ECDC)*, págs. 1-7.

Shams, Mohammadreza y Ahmad Baraani-Dastjerdi (2017). «Enriched LDA (ELDA): Combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction». En: *Expert Systems with Applications* 80, págs. 136-146.

Shanshan Gao, Jinxing Hao y Yu Fu (2015). «The application and comparison of web services for sentiment analysis in tourism». En: *2015 12th International Conference on Service Systems and Service Management (ICSSSM)*, págs. 1-6.

Shao, Jun, Xuesong Chang y Alastair M. Morrison (2017). «How Can Big Data Support Smart Scenic Area Management? An Analysis of Travel Blogs on Huashan». En: *Sustainability* 9.12.

Shen, Junge, Jialie Shen, Tao Mei y Xinbo Gao (2016). «Landmark Reranking for Smart Travel Guide Systems by Combining and Analyzing Diverse Media». En: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 46.11, págs. 1492-1504.

Shi, Y., L. Zhu, W. Li, K. Guo e Y. Zheng (2019). «Survey on Classic and Latest Textual Sentiment Analysis Articles and Techniques». En: *International Journal of Information Technology and Decision Making* 18.4, págs. 1243-1287.

Shin, Seunghun, Namho Chung, Zheng Xiang y Chulmo Koo (2019). «Assessing the Impact of Textual Content Concreteness on Helpfulness in Online Travel Reviews». En: *Journal of Travel Research* 58.4, págs. 579-593.

Singh, Shiwangi, Akshay Chauhan y Sanjay Dhir (2019). «Analyzing the startup ecosystem of India: a Twitter analytics perspective». En: *Journal of Advances in Management Research* 17.2, págs. 262-281.

Singh Chauhan, Ganpat, Yogesh Kumar Meena, Dinesh Gopalani y Ravi Nahta (2020). «A two-step hybrid unsupervised model with attention mechanism for aspect extraction». En: *Expert Systems with Applications* 161, pág. 113673.

Situmorang, K. M., A. N. Hidayanto, A. F. Wicaksono y A. Yuliawati (2018). «Analysis on Customer Satisfaction Dimensions in Peer-to-Peer Accommodation using Latent Dirichlet Allocation: A Case Study of Airbnb». En: *2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, págs. 542-547.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng y Christopher Potts (2013). «Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank». En: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA, págs. 1631-1642.

Song, Seobgyu, Seunghyun "Brian" Park y Kwangsoo Park (2021). «Thematic analysis of destination images for social media engagement marketing». En: *Industrial Management & Data Systems* 121.6, págs. 1375-1397.

Soong, Hoong Cheng, Norazira Binti A. Jalil, Ramesh Kumar Ayyasamy y Rehan Akbar (2019). «The essential of sentiment analysis and opinion mining in social media : Introduction and survey of the recent approaches and techniques». En: *ISCAIE 2019 - 2019 IEEE Symposium on Computer Applications and Industrial Electronics*. Institute of Electrical y Electronics Engineers Inc., págs. 272-277.

Speer, Robyn, Joshua Chin y Catherine Havasi (2017). «ConceptNet 5.5: An Open Multilingual Graph of General Knowledge». En: *AAAI Conference on Artificial Intelligence*.

Srivastava, Vartika y Arti D. Kalro (2019). «Enhancing the Helpfulness of Online Consumer Reviews: The Role of Latent (Content) Factors». En: *Journal of Interactive Marketing* 48, págs. 33-50.

Starosta, K., S. Budz y M. Krutwig (2018). «Artificial neural-network-based emotion classification in the online media for tourism businesses». En: *Proceedings of the 5th European Conference on Social Media, ECSM 2018*, págs. 412-423.

Starosta, K., C.B. Onete, S. Budz y M. Krutwig (2019). «Differences in travelers' perceptions of popular tourist destinations estimated by a LSTM neural network: A comparison between the UK and Germany». En: *Tourism* 67.4, págs. 405-422.

Starosta, Kejo, Sonia Budz y Michael Krutwig (2019). «The impact of German-speaking online media on tourist arrivals in popular tourist destinations for Europeans». En: *Applied Economics* 51.14, págs. 1558-1573.

Stepaniuk, Krzysztof y Anna Sturgulewska (2021). «Hitchhiking Experiences and Perception of Affective Label Polarity in Social Networking Sites—Potential Memetic Implications for Digital Visual Content Management». En: *Sustainability* 13.1.

Stieglitz, Stefan y Linh Dang-Xuan (2013). «Social media and political communication: a social media analytics framework». En: *Social Network Analysis and Mining* 3.4, págs. 1277-1291.

Stieglitz, Stefan, Linh Dang-Xuan, Axel Bruns y Christoph Neuberger (2014). «Social Media Analytics». En: *Business & Information Systems Engineering* 6.2, págs. 89-96.

Stojanovic, Igor, Luisa Andreu y Rafael Curras-Perez (2018). «Effects of the intensity of use of social media on brand equity». En: *European Journal of Management and Business Economics* 27.1, págs. 83-100.

Sudhakar, Sooriya y Sangeetha Gunasekar (2020). «Examining online ratings and customer satisfaction in airlines». En: *Anatolia* 31.2, págs. 260-273.

Sun, Xueying y Fu Xie (2019). «Visual Analysis of Social Network Influence Based on Knowledge Mapping». En: *IOP Conference Series: Materials Science and Engineering* 490.4, pág. 042039.

Sun, Yao, Yiwen Shao y Edwin H.W. Chan (2020). «Co-visitation network in tourism-driven peri-urban area based on social media analytics: A case study in Shenzhen, China». En: *Landscape and Urban Planning* 204, pág. 103934.

Surugiu, Marius-Răzvan y Camelia Surugiu (2015). «Heritage Tourism Entrepreneurship and Social Media: Opportunities and Challenges». En: *Procedia - Social and Behavioral Sciences* 188. Heritage as an alternative driver for sustainable development and economic recovery in South East Europe -Project SEE/B/0016/4.3/X SAGITTARIUS, págs. 74-81.

Sutherland, Ian y Kiattipoom Kiatkawsin (2020). «Determinants of Guest Experience in Airbnb: A Topic Modeling Approach Using LDA». En: *Sustainability* 12.8.

Sánchez-Franco, Manuel J., Antonio Navarro-García y Francisco Javier Rondán-Cataluña (2019). «A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services». En: *Journal of Business Research* 101, págs. 499-506.

Taheri, Babak, Hossein Olya, Faizan Ali y Martin Joseph Gannon (2020). «Understanding the Influence of Airport Servicescape on Traveler Dissatisfaction and Misbehavior». En: *Journal of Travel Research* 59.6, págs. 1008-1028.

Tao, Yuguo, Feng Zhang, Chunyun Shi y Yun Chen (2019). «Social Media Data-Based Sentiment Analysis of Tourists' Air Quality Perceptions». En: *Sustainability* 11.18, pág. 5070.

Tariyal, Ankit, Sachin Goyal y Neeraj Tantububay (2018). «Sentiment Analysis of Tweets Using Various Machine Learning Techniques». En: *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)*, págs. 1-5.

Tham, A y D Huang (2019). «Game on! A new integrated resort business model». En: *Tourism Review* 74.6, págs. 1153-1166.

Thoma, Brent, Teresa M. Chan, Puneet Kapur, Derek Sifford, Marshall Siemens, Michael Paddock, Felix Ankel, Andy Grock, Michelle Lin, Charlotte Alexander, Mohammed Alkhalifah, Abdulaziz S. Almehlisi y Saeed Alqahtani ... (2018). «The Social Media Index as an Indicator of Quality for Emergency Medicine Blogs: A METRIQ Study». En: *Annals of Emergency Medicine* 72.6, págs. 696-702.

Thoma, Brent, Jason L Sanders, Michelle Lin, Quinten S Paterson, Jordon Steeg y Teresa M Chan (2015). «The Social Media Index: Measuring the Impact of Emergency Medicine and Critical Care Websites». En: *Western Journal of Emergency Medicine* XVI.2.

Thomaz, Guilherme M., Alexandre A. Biz, Eduardo M. Bettoni, Luiz Mendes-Filho y Dimitrios Buhalis (2017). «Content mining framework in social media: A FIFA world cup 2014 case analysis». En: *Information & Management* 54.6. Smart Tourism: Traveler, Business, and Organizational Perspectives, págs. 786-801.

Tian, Ye, Chenru Chen, Xinyi Chen, Qianqian Zhang y Ruizhi Sun (2020). «Research on real-time analysis technology of urban land use based on support vector machine». En: *Pattern Recognition Letters* 133, págs. 320-326.

Trunfio, Mariapina y Maria Della Lucia (2019). «Engaging Destination Stakeholders in the Digital Era: The Best Practice of Italian Regional DMOs». En: *Journal of Hospitality & Tourism Research* 43.3, págs. 349-373.

Uchinaka, Sanae, Vignesh Yoganathan y Victoria-Sophie Osburg (2019). «Classifying residents' roles as online place-ambassadors». En: *Tourism Management* 71, págs. 137-150.

Uşaklı, Ahmet, Burcu Koç y Sevil Sönmez (2017). «How 'social' are destinations? Examining European DMO social media usage». En: *Journal of Destination Marketing and Management* 6.2, págs. 136-149.

Valdivia, Ana, Emiliya Hrabova, Iti Chaturvedi, M. Victoria Luzón, Luigi Troiano, Erik Cambria y Francisco Herrera (2019). «Inconsistencies on TripAdvisor reviews: A unified index between users and Sentiment Analysis Methods». En: *Neurocomputing* 353, págs. 3-16.

Valdivia, Ana, Eugenio Martínez-Cámara, Iti Chaturvedi, M. Victoria Luzón, Erik Cambria, Yew-Soon Ong y Francisco Herrera (2020). «What do people think about this monument? Understanding negative reviews via deep learning, clustering and descriptive rules». En: *Journal of Ambient Intelligence and Humanized Computing* 11.1, págs. 39-52.

Varkaris, Eleftherios y Barbara Neuhofer (2017). «The influence of social media on the consumers' hotel decision journey». En: *Journal of Hospitality and Tourism Technology* 8.1, págs. 101-118.

Varsha, P. S., Shahriar Akter, Amit Kumar, Saikat Gochhait y Basanna Patagundi (2021). «The Impact of Artificial Intelligence on Branding: A Bibliometric Analysis (1982-2019)». En: *Journal of Global Information Management* 29.4, págs. 221-246.

Vecchio, Pasquale Del, Gioconda Mele, Valentina Ndou y Giustina Secundo (2018). «Creating value from Social Big Data: Implications for Smart Tourism Destinations». En: *Information Processing & Management* 54.5, págs. 847-860.

Viñan-Ludeña, Marlon-Santiago (2019). «A Systematic Literature Review on Social Media Analytics and Smart Tourism». En: *Smart Tourism as a Driver for Culture and Sustainability*. Springer, Cham, págs. 357-374.

Viñán-Ludeña, Marlon Santiago, Luis M. De Campos, Luis Roberto Jacome-Galarza y Javier Sinche-Freire (2020). «Social media influence: a comprehensive review in general and in tourism domain». En: *Smart Innovation, Systems and Technologies*. Vol. 171. Springer, págs. 25-35.

Vu, Huy Quan, Birgit Muskat, Gang Li y Rob Law (2021). «Improving the resident–tourist relationship in urban hotspots». En: *Journal of Sustainable Tourism* 29.4, págs. 595-615.

Wang, H., T. Jin, B. Zhou, K.R. Shui y M. Zhou (2012). «Smart tourism». En: *Smart Tourism*, págs. 10-12.

Wang, H.C., Y.H. Chiang e Y.F. Sun (2019). «Use of multi-lexicons to analyse semantic features for summarization of touring reviews». En: *Electronic Library* 37.1, págs. 185-206.

Wang, Hsiu-Yuan (2016). «Predicting customers'intentions to check in on Facebook while patronizing hospitality firms». En: *Service Business* 10.1, págs. 201-222.

Wang, J., B.-K. Bao y C. Xu (2019). «Sentiment-Aware Multi-modal Recommendation on Tourist Attractions». En: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11295 LNCS, págs. 3-16.

Wang, Renwu, Jiaqi Luo y Songshan (Sam) Huang (2020). «Developing an artificial intelligence framework for online destination image photos identification». En: *Journal of Destination Marketing & Management* 18, pág. 100512.

Wang, Tao, Yi Cai, Ho fung Leung, Raymond Y.K. Lau, Qing Li y Huaqing Min (2014). «Product aspect extraction supervised with online domain knowledge». En: *Knowledge-Based Systems* 71, págs. 86-100.

Wang, Wanfei, Shun Ying, Jiaying Lyu y Xiaoguang Qi (2019). «Perceived image study with online data from social media: the case of boutique hotels in China». En: *Industrial Management and Data Systems* 119.5, págs. 950-967.

Wang, Xia, Xiang (Robert) Li, Feng Zhen y JinHe Zhang (2016). «How smart is your tourist attraction?: Measuring tourist preferences of smart tourism

attractions via a FCEM-AHP and IPA approach». En: *Tourism Management* 54, págs. 309-320.

Wei, Zihan, Mingli Zhang y Yaxin Ming (2022). «Understanding the effect of tourists' attribute-level experiences on satisfaction – a cross-cultural study leveraging deep learning». En: *Current Issues in Tourism* 0.0, págs. 1-17.

Weihs, Claus y Katja Ickstadt (2018). «Data Science: the impact of statistics». En: *International Journal of Data Science and Analytics* 6.3, págs. 189-194.

Widmar, Nicole Olynk, Courtney Bir, McKenna Clifford y Natalya Slipchenko (2020). «Social media sentimentas an additional performance measure? Examples from iconic theme park destinations». En: *Journal of Retailing and Consumer Services* 56, pág. 102157.

Xiang, Zheng, Qianzhou Du, Yufeng Ma y Weiguo Fan (2017). «A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism». En: *Tourism Management* 58, págs. 51-65.

Xu, Yukuan, Zili Zhang, Rob Law y Ziqiong Zhang (2019). «Effects of online reviews and managerial responses from a review manipulation perspective». En: *Current Issues in Tourism* 23.17, págs. 2207-2222.

Yadav, M. y V. Bhojane (2019). «Semi-Supervised Mix-Hindi Sentiment Analysis using Neural Network». En: *2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, págs. 309-314.

Yan, Qiang, Simin Zhou y Sipeng Wu (2018). «The influences of tourists' emotions on the selection of electronic word of mouth platforms». En: *Tourism Management* 66, págs. 348-363.

Yang, H.-L. y A.F.Y. Chao (2018). «Sentiment annotations for reviews: an information quality perspective». En: *Online Information Review* 42.5, págs. 579-594.

Ying, Song, Stavros Sindakis, Sakshi Aggarwal, Charles Chen y Jiafu Su (2021). «Managing big data in the retail industry of Singapore: Examining the impact on customer satisfaction and organizational performance». En: *European Management Journal* 39.3, págs. 390-400.

Yoosefi Nejad, Mohsen, Maryam Sadat Delghandi, Ahmed Omar Bali y Mehdi Hosseinzadeh (2020). «Using Twitter to raise the profile of childhood cancer awareness month». En: *Network Modeling Analysis in Health Informatics and Bioinformatics* 9.3.

Yu, C., X. Zhu, B. Feng, L. Cai y L. An (2019). «Sentiment analysis of Japanese tourism online reviews». En: *Journal of Data and Information Science* 4.1, págs. 89-113.

Zeitzoff, Thomas (2017). «How Social Media Is Changing Conflict». En: *Journal of Conflict Resolution* 61.9, págs. 1970-1991.

— (2018). «Does Social Media Influence Conflict? Evidence from the 2012 Gaza Conflict». En: *Journal of Conflict Resolution* 62.1, págs. 29-63.

Zeng, Daniel, Hsinchun Chen, Robert Lusch y Shu-Hsing Li (2010). «Social Media Analytics and Intelligence». En: *IEEE Intelligent Systems* 25.6, págs. 13-16.

Zeng, Delin, Yenni Tim, Jiaxin Yu y Wenyuan Liu (2020). «Actualizing big data analytics for smart cities: A cascading affordance study». En: *International Journal of Information Management* 54, pág. 102156.

Zhai, ChengXiang y Sean Massung (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Association for Computing Machinery y Morgan Claypool.

Zhang, Kun, Ye Chen y Zhibin Lin (2020). «Mapping destination images and behavioral patterns from user-generated photos: a computer vision approach». En: *Asia Pacific Journal of Tourism Research* 25.11, págs. 1199-1214.

Zhang, Tingting, Bin Li y Nan Hua (2022). «Chinese cultural theme parks: text mining and sentiment analysis». En: *Journal of Tourism and Cultural Change* 20.1-2, págs. 37-57.

Zhang, W., Y. Choe y D.R. Fesenmaier (2019). «The defining features of emotions in online stories». En: *e-Review of Tourism Research* 16.2-3, págs. 136-145.

Zhang, W. y D.R. Fesenmaier (2018). «Assessing emotions in online stories: comparing self-report and text-based approaches». En: *Information Technology and Tourism* 20.1-4, págs. 83-95.

Zhang, W., J.J. Kim, H. Kim y D.R. Fesenmaier (2019). «The tourism story project: Developing the behavioral foundations for an ai supporting destination story design». En: *e-Review of Tourism Research* 17.2, págs. 169-187.

Zhang, Weiwu, Thomas J. Johnson, Trent Seltzer y Shannon L. Bichard (2010). «The Revolution Will be Networked: The Influence of Social Networking Sites on Political Attitudes and Behavior». En: *Social Science Computer Review* 28.1, págs. 75-92.

Zhang, Xiaowei, Yang Yang, Yi Zhang y Zili Zhang (2020). «Designing tourist experiences amidst air pollution: A spatial analytical approach using social media». En: *Annals of Tourism Research* 84, pág. 102999.

Zhang, Zusheng, Yanghui Rao, Hanjiang Lai, Jiahai Wang y Jian Yin (2021). «TADC: A Topic-Aware Dynamic Convolutional Neural Network for Aspect Extraction». En: *IEEE Transactions on Neural Networks and Learning Systems*, págs. 1-13.

Zhao, Wayne Xin, Jing Liu, Yulan He, Chin-Yew Lin y Ji-Rong Wen (2016). «A computational approach to measuring the correlation between expertise and social media influence for celebrities on microblogs». En: *World Wide Web* 19.5, págs. 865-886.

Zhao, Xinyan, Mengqi Zhan y Brooke F. Liu (2018). «Disentangling social media influence in crises: Testing a four-factor model of social media influence with large data». En: *Public Relations Review* 44.4, págs. 549-561.

Zheng, X., Y. Luo, L. Sun, J. Zhang y F. Chen (2018). «A tourism destination recommender system using users' sentiment and temporal dynamics». En: *Journal of Intelligent Information Systems* 51.3, págs. 557-578.

Önder, I., U. Gunter y A. Scharl (2019). «Forecasting tourist arrivals with the help of web sentiment: A mixed-frequency modeling approach for big data». En: *Tourism Analysis* 24.4, págs. 437-452.