Proxectos INOU 2019.

Investigación aplicada na provincia de Ourense

Coordinadora:

de Blas Varela, Esther

Ourense, 2020

Universidade de Vigo • Campus de Ourense

Proxectos INOU 2019. Investigación aplicada na provincia de Ourense

Autores/as:
Cid Iglesias, Begoña
Gueimonde Canto, Ana Isabel
García Feal, Orlando
Pérez Losada, Fermín Emiliano
Rodríguez Toubes-Muñiz, Diego
Iglesias Sarmiento, Valentín
Ribadas Pena, Francisco José
Laza Fidalgo, Rosalía
Cotos Yáñez, Tomás Raimundo
Rivo López, Elena
Astray Dopazo, Gonzalo

Coordinadora:
de Blas Varela, Esther

Comisión de Avaliación:
Álvarez Díaz, Marcos
Garrote Velasco, Gil
López Periago, José Eugenio
Prada Rodríguez, Julio
Reboiro Jato, Miguel
Sampayo Fernández, José A.

Vicerreitoría do Campus de Ourense-Campus Auga
Universidade de Vigo
Ourense, 2020

Nº de páxinas: 232

ISBN: 978-84-8158-882-8

Edición
Vicerreitoría do Campus de Ourense - Campus Auga
www.uvigo.gal/campus/ourense-campus-auga

Maquetación
Rodi Artes Gráficas, S. L.

Índice

Desenvolvemento de ferramentas para apoiar a calidade e a mellora continua no sector hoteleiro ourensán

D. Frade-Amil,** A. Borrajo,* T. R. Cotos-Yáñez,** E. Lorenzo,* A. Pérez-González,** R. Pavón,* M. A. Mosquera-Rodríguez,** L. Borrajo* e R. Laza*

**Departamento de Informática, campus de Ourense, Universidade de Vigo. Subproxecto EPOCa*

***Departamento de Estatística e Investigación Operativa, campus de Ourense, Universidade de Vigo. Subproxecto AMOCa*

rlaza@uvigo.gal, pavon@uvigo.gal, lborrajo@uvigo.gal, eva@uvigo.gal, abvieitez2@esei.uvigo.gal, diego.frade.amil@gmail.com, cotos@uvigo.gal, mamrguez@uvigo.gal, anapg@uvigo.gal

O obxectivo principal deste proxecto coordinado é extraer de forma automática datos de comentarios sobre hoteis da provincia de Ourense das plataformas Booking.com e Tripadvisor.es (dous importantes portais nos que se recollen as opinións dos e das turistas) para analizar posteriormente aplicando técnicas estatísticas e/ou de minaría de datos.

Medición e análise da reputación en liña (*on-line)* son dous conceptos que non se poden entender de maneira independente. O emprego de boas ferramentas para captar os datos é moi importante e complexo, debido ao volume de información presente (moitas opinións), á súa dispersión e heteroxeneidade nos mecanismos de captación, á presentación dos datos (datos procedentes de distintas plataformas de reserva con distintos formatos) e á escasa (a miúdo nula) estruturación dos datos. Neste senso, logrouse un grande avance pola obtención dun sistema funcional de extracción automática das características dos hoteis da provincia, ademais das opinións vertidas sobre eles.

Técnicas básicas e avanzadas de modelización estatística e de exploración ou minaría de datos *(data mining)* permitiron explicar a valoración global do usuariado en

función da valoración dos diferentes servizos do hotel e dos comentarios de opinión sobre eles. Axustar un modelo que sirva para identificar tipoloxías de mensaxes ou temáticas de interese por parte dos e das hóspedes é unha ferramenta que achega información valiosa na toma de futuras decisións.

Palabras clave:

Big data, modelaxe de opinións, regresión xeneralizada, *data mining*.

1. Introdución

Na actualidade, o sector hoteleiro ourensán atópase nun excelente momento segundo os datos extraídos do Instituto Nacional de Estatística. Particularmente, segundo os datos difundidos recentemente, os hoteis da provincia aloxaron en febreiro 20413 persoas, e foi este o segundo mellor dato da última década a estas alturas do ano (soamente superado polos/as 22205 visitantes do ano 2018).

No ano 2017, o 70 % das reservas de aloxamento formalizáronse a través das plataformas en liña (*on-line)* segundo o estudo «Minerva Travel 2017» realizado por Google España [1]. Atendendo á mesma fonte, o 42 % busca información en liña do lugar visitado (que visitar, onde aloxarse, que actividades se poden realizar...) e, se se ten unicamente en conta a xente máis nova, esta porcentaxe increméntase ata o 74 %. Por esta razón, gozar dunha boa reputación en liña vai predispoñer futuros clientes e clientas para confiar nos produtos e nos servizos ofertados.

As plataformas en liña que permiten as reservas de aloxamento tamén comparten opinións de clientas e clientes que visitaron os establecementos con anterioridade. Estes datos poden ser explotados polos hoteis para detectar deficiencias e mellorar a calidade dos servizos prestados. Así tamén o identificaron D. H. Moya e J. Majó, os cales centraron a súa investigación en analizar as opinións recibidas por medios virtuais en 57 hoteis latinoamericanos. A partir destes comentarios, elaboraron un manual de boas prácticas con recomendacións sobre seis ámbitos hoteleiros (cuartos, alimentos e bebidas, recepción, centro de negocios, seguridade e xerencia) para optimizar a calidade nos hoteis e mellorar a reputación en liña [2]. Por tanto, a análise dos datos das interaccións producidas entre os clientes e clientas dos hoteis ofreceralles unha vantaxe competitiva sobre quen non o faga, xa que con elas e eles poden mellorar os servizos do hotel e espertar o interese de posibles clientas e clientes.

Medición e análise da reputación en liña son dous conceptos moi relacionados. O emprego de boas ferramentas para obter a información adecuada é moi importante e complexo debido ao gran volume de información manexada, ademais da heteroxeneidade dos datos. Neste senso, a explotación desta información implica a combinación de técnicas de procesamento de grandes volumes de datos *(big data)* e de análise estatística.

Segundo o estudo «Barómetro sobre la gestión de la satisfacción del huésped 2016» realizado por ReviewPro, o 67 % dos hoteis consideraron difícil xestionar a retroalimentación *(feedback)* proporcionada polas plataformas de reservas, e dous terzos destes establecementos consideraron eses datos difíciles de manexar, o que reforza a necesidade de ter ferramentas especializadas para facilitar a xestión da satisfacción da clientela.

Nos últimos anos, o uso de métodos estatísticos e de aprendizaxe automática orientados a resolver problemas emerxentes na análise de datos no sector turístico acadou un notable crecemento. Este feito reflectiuse no aumento do índice de impacto que teñen revistas científicas como *Tourism Management*, *Annals of Tourism Research* ou *Journal of Travel Research*. Por outra parte, a aparición e o aumento de plataformas web, onde o usuariado pode valorar a súa experiencia durante a súa estadía en establecementos hoteleiros, dotou o sector dun mecanismo efectivo para recoller información de especial importancia tanto para os xestores e xestoras dos establecementos coma para o futuro usuariado. De feito, o 86 % dos viaxeiros e viaxeiras non reservan aloxamento sen antes ler as opinións do usuariado sobre el (TripBarometer do ano 2017-2018).

As ferramentas de análise estatística, xunto coas de minaría de datos serven para atopar cales son os aspectos/características dun hotel do que falan as e os seus hóspedes. Técnicas básicas e avanzadas de análise clúster van permitir agrupar mensaxes ou características para identificar tipoloxías de mensaxe ou temáticas de interese por parte dos e das hóspedes. Finalmente, a minaría de datos vai permitir encontrar patróns que poidan achegar información valiosa na toma de futuras decisións [3].

Na seguinte sección preséntase o traballo levado a cabo no subproxecto EPOCa «Extracción e Preprocesamento de Opinións sobre a Calidade dos servizos prestados polo sector hoteleiro ourensán», na sección 3 descríbese o traballo desenvolvido no subproxecto AMOCa «Análise e Modelaxe de Opinións sobre a Calidade dos servizos

prestados polo sector hoteleiro ourensán» e, por último, a sección 4 estará dedicada ás conclusións extraídas dos resultados.

2. Subproxecto EPOCa: «Extracción e Preprocesamento de Opinións sobre a Calidade dos servizos prestados polo sector hoteleiro ourensán»

Big data é un termo que describe e engloba o gran volume de datos (tanto estruturados coma non estruturados), xunto coas técnicas, para tratar a súa explotación e entender a realidade. Isto fai que *big data* sexa tan útil e que, cunha cantidade tan grande de información, os datos se poidan moldear e analizar para poder identificar os problemas dunha forma máis comprensible.

Big data é unha gran mina para explotar debido á existencia dunha cantidade inmensa de datos almacenados que posúen un gran valor pero que necesitan ser tratados, refinados e preprocesados mediante ferramentas de analítica de datos. O preprocesamento de datos é fundamental para converter os datos almacenados en datos de calidade [4].

Google deseñou MapReduce [5], que se considera como a plataforma pioneira para procesar datos masivos, así como un paradigma para procesar datos mediante a división de ficheiros de datos. A pesar da súa popularidade, MapReduce e a súa versión de código aberto Hadoop [6] teñen limitacións en certos escenarios. Apache Spark [7] nace como unha alternativa que intenta darlles solución ás limitacións de MapReduce/Hadoop. Spark converteuse nunha das ferramentas máis potentes e populares no ecosistema do *big data*, grazas ás súas operacións de uso intensivo de memoria que lle permiten ser capaz de cargar datos en memoria e consultalos rapidamente.

Algúns métodos de preprocesamento que se poden aplicar no contexto *big data* son discretización e normalización, extracción de atributos, selección de atributos, conversión de atributos, técnicas para o preprocesamento de texto... [4]

As plataformas de reserva en liña, grazas ao *big data*, posúen enorme cantidade de información que os propios clientes e clientas foron deixando en internet sobre os seus gustos, hábitos e preferencias. Un rastro que pode axudar o sector a personalizar os seus servizos e a adaptarse á demanda.

Por todo o indicado anteriormente, o obxectivo principal do subproxecto EPOCa é extracción de datos dos establecementos hoteleiros ourensáns da plataforma de reservas Booking e do portal de promoción turística TripAdvisor, así como as tarefas de pre-

procesamento que permitan a explotación dos datos obtidos desde un punto de vista estatístico. Ambas as plataformas ofertan unha API (*Application Programming Interface*, interface para o desenvolvemento de aplicacións) RESTful (*Representational State Transfer*, transferencia de representación do Estado), aínda que no caso de Tripadvisor non permite consultar comentarios e, doutra parte, en Booking o acceso á API está limitado a negocios hoteleiros. Neste senso, foi necesario acudir ao emprego de técnicas de *web scrapping*, que permitiron procesar as páxinas web de forma automática para obter a información desexada [8].

Para desenvolver o proxecto, empregouse como linguaxe de programación Java porque ultimamente os sitios web están a desenvolverse en JavaScript, o que supón un problema para a recollida de datos polas cargas asíncronas, polo que unha biblioteca necesaria é Selenium + Jsoup.

Selenium foi desenvolvido por SeleniumHQ. O obxectivo principal desta tecnoloxía é acceder a calquera sitio web mediante o uso dun controlador que permite a manipulación dun navegador, pero a medida que a cantidade de información na rede creceu, Selenium comezou a usarse para recoller información.

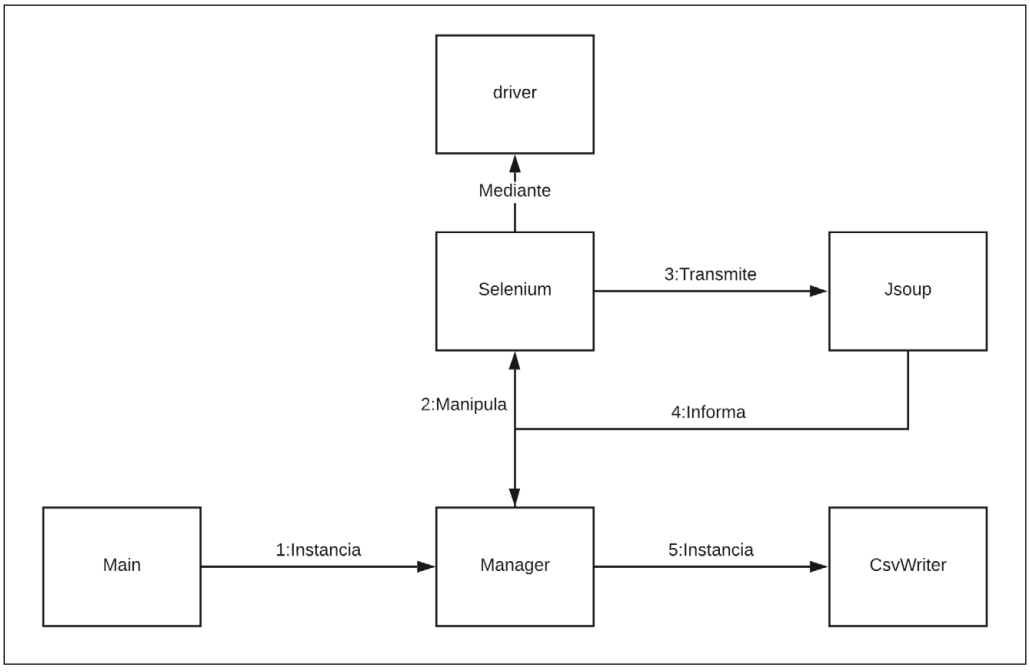

Figura 1. Arquitectura do sistema

Jsoup é unha biblioteca Java para traballar con HTML *(HyperText Markup Language)* do mundo real. Proporciona unha API moi cómoda para extraer e manipular datos, empregando o mellor de DOM *(Document Object Model)*, CSS *(Cascading Style Sheets)* e métodos similares ao *jquery*.

Na figura 1 preséntase a arquitectura do sistema: o *main* é o encargado de instanciar un *manager* construído a través dun URL (o URL da provincia). No construtor do Manager FirefoxDriver (Selenium) instánciase co URL que lle chega como parámetro; o resultado é un Firefox aberto nese URL. A información deste Firefox envíase á biblioteca Jsoup; finalmente, o Firefox péchase e instancíanse os CSV (véxase a figura 2).

```
final URL url = new URL( s: "https://www.tripadvisor.es/Hotels-g1768741-Province_of_Ourense_Galicia-Hotels.html");
Manager man = new Manager(url);
```

```
/**
 * Constructs a new instance of {@link Manager}
 *
 * @param url the URL of the province
 */
public Manager(URL url) {
    this.urlHotel = url;
    WebDriver driver = new FirefoxDriver();
    try {
        driver.get(url.toString());
        this.docHotel = Jsoup.parse(driver.getPageSource()).normalise();
    } finally {
        driver.quit();
    }
    this.csvDatasetWriter = new CSVDatasetWriter( csvDatasetPath: "prueba.csv");
    this.csvUbicationWriter = new CSVDatasetWriter( csvDatasetPath: "prueba-ubication.csv");
}
```

Figura 2. Código Java para instanciar un Manager FirefoxDriver

O comando *run ()* inicia a recollida de todos os datos da provincia. O primeiro que fai é saber cantas páxinas de hoteis ten esa provincia. Para isto, Jsoup permite pasar a través da árbore DOM empregando a clase *pagenum*. Desta forma, percorrerase cada

páxina e empregarase o *driver* de Firefox para conseguir os hoteis dispoñibles nesa páxina.

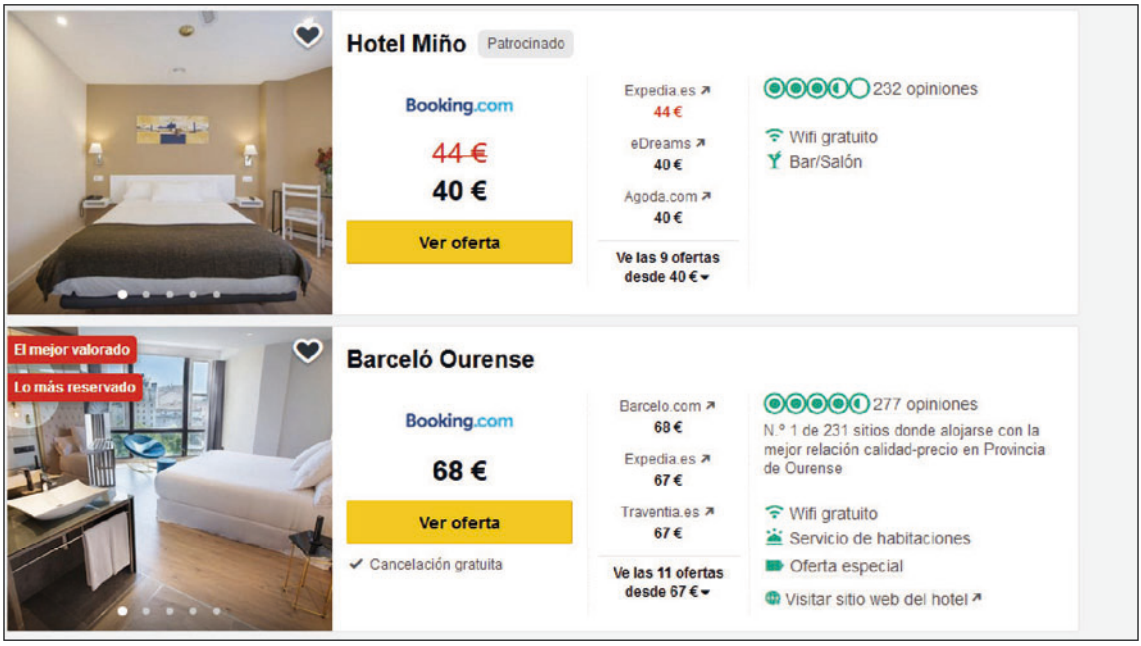

Figura 3. Información de dous hoteis dispoñibles en Booking

O *driver* cargará algo similar ao exemplo dos hoteis da figura 3, que para acceder á súa información ten que pasarlle os datos a Jsoup, e pasará por cada hotel para extraer os seus datos; para iso a árbore DOM percorrerase a través da clase *property_title*.

Unha vez almacenada toda a información, o seguinte que se obtén é a latitude e a lonxitude do hotel pola rúa onde está situado. Para iso emprégase a API de xeolocalización de Google, tal e como se pode ver na figura 4.

Outra información importante son os comentarios que se atopan ao final da páxina de cada hotel e que teñen, á súa vez, unha serie de páxinas de comentarios. Para cada páxina de comentarios abrirase un *driver* Firefox que amosará unha certa cantidade de comentarios que serán tratados por *Jsoup*.

```
String province = hotelDoc.getElementsByClass( className: "link").text().split( s: " ")[5];
defCsv.put( k: "nombreProvincia", province);


//Id of the location
int idLocation = Integer.parseInt(urlSplitted[1].substring(1));
defCsv.put( k: "idTripLocalizacion", idLocation);
//Name location
String nameLocation = hotelDoc.getElementsByClass( className: "ui_pill inverted").text();
defCsv.put( k: "nombreLocalizacion", nameLocation);

//Name hotel
String nameHotel = hotelDoc.getElementById("HEADING").text();
defCsv.put( k: "nombreHotel", nameHotel);
ubicationCsv.put( k: "nombreHotel", nameHotel);
```

```
if (!street.equals("")) {
    GeoApiContext context = new GeoApiContext.Builder()
            .apiKey("AIzaSyA4Nn-9ULHutLrGoANfUzHYAssOcmRrIu0")
            .build();
    FindPlaceFromText results = null;
    try {
        results = PlacesApi.findPlaceFromText(context, street, inputType).await();
    } catch (ApiException ex) {
        ex.printStackTrace();
    } catch (InterruptedException ex) {
        ex.printStackTrace();
    }
```

Figura 4. Código para obter a localización dun hotel

A información extraída das dúas plataformas indícase na táboa 1. En canto aos formatos de opinión, Booking recolle de forma separada os comentarios positivos e negativos de cada establecemento, xunto cunha puntuación nun rango de 0-10. Doutra parte, Tripadvisor non estrutura a opinión en aspectos a favor e en contra, e a puntuación atópase no rango de 0-5. Tendo en conta estas diferenzas, as tarefas de preprocesamento foron especialmente relevantes para os comentarios procesados. Á vista das particularidades detectadas no problema, o tratamento desta información fíxose empregando técnicas de manipulación de grandes volumes de datos (*big data*).

Tripadvisor	**Booking**
Localización	Localización
Nome da provincia (aínda que neste caso non sería necesario)	Nome da provincia (aínda que neste caso non sería necesario)
Identificador do hotel	Identificador do hotel
Nome do hotel	Nome do hotel
Categoría do hotel (estrelas)	Categoría do hotel (estrelas)
Servizos: aparcadoiro, piscina, almorzo, gardaría... [un total de 153 servizos clasificados con 1 (se posúe) e 0 (se non posúe)]	Categoría Booking (estrelas indicadas por Booking)
	Servizos: aparcadoiro, ceas temáticas, piscina, almorzo, chan de moqueta... [un total de 418 servizos clasificados con 1 (se posúe) 0 (se non posúe)]
Identificador de comentario (autoincremental)	Identificador de comentario (autoincremental)
Valoración total do hotel	Valoración total do hotel
Valoración total da localización	Valoración total da localización
Valoración total da limpeza	Valoración total da limpeza
Valoración total dos servizos	Valoración total das instalacións
Valoración total da relación calidade/prezo	Valoración total da relación calidade/prezo
Valoración individual do usuario/a	Valoración total da wifi
Identificador do comentario propio de Tripadvisor	Valoración total do persoal
Título do comentario	Valoración individual do usuario/a
Tipo de viaxe	Título do comentario
Comentario	Comentario positivo
Consellos	Comentario negativo
Votos útiles do comentario	Votos útiles do comentario
Data do comentario	Data do comentario
Usuario/a	
Número de contribucións do usuario/a	
Votos útiles do usuario/a	
Nivel de colaboración	

Táboa 1. Información extraída das plataformas Tripadvisor e Booking

A información almacenouse en arquivos CSV, os cales poderán ser analizados polo equipo do subproxecto AMOCa. Cómpre indicar que se xerou un arquivo CSV coa información de todos os comentarios de cada plataforma e, ademais, outros dous arquivos coa localización dos diferentes hoteis analizados.

3. Subproxecto AMOCa: «Análise e Modelaxe de Opinións sobre a Calidade dos servizos prestados polo sector hoteleiro ourensán»

Este proxecto cos datos obtidos dos comentarios sobre hoteis da provincia de Ourense das plataformas Booking e Tripadvisor (dous importantes portais nos que se recollen as opinións dos e turistas) obtén información aplicando técnicas estatísticas e/ou de minaría de datos.

Na seguinte sección repasarase o estado actual do sector hoteleiro na provincia de Ourense e describiremos brevemente os portais web TripAdvisor e Booking. Na subsección 3.1 describiranse os datos obtidos destas plataformas e empregaranse para aplicar as técnicas detalladas na subsección 3.2.

3.1. O sector hoteleiro ourensán

Segundo os datos do Instituto Galego de Estatística (IGE), no período comprendido entre agosto de 2018 e xullo de 2019 estímase que había abertos arredor de 160-170 establecementos hoteleiros en toda a provincia de Ourense. Estes déronlles emprego a máis de 600 traballadores/as e aloxamento a unha media mensual de 28800 usuarios/as, que realizaron unha estadía media inferior ás dúas noites. Se comparamos estes datos cos do ámbito autonómico e estatal, observamos que a estadía media na provincia de Ourense é similar ao dato galego, pero bastante inferior respecto ao estatal (figura 5 der.). No que respecta ao grao de ocupación, a situación é semellante,

Ourense atópase por debaixo dos datos galego e estatal (figura 5 esq.). Algo que si é semellante nos tres casos é a evolución da estadía media e da ocupación, que presenta uns altos índices nos meses do verán e valores máis baixos durante o inverno.

O IGE tamén ofrece datos do número de hoteis con estrelas localizados na provincia de Ourense no ano 2019 (véxase a táboa 2), xunto coa cantidade de cuartos e de prazas dispoñibles. Destaca que nesta provincia non se localiza ningún establecemento co máximo número de estrelas (5). Estes mesmos datos atópanse, ademais, disgregados por comarcas (véxase a táboa 3) e por concellos, onde se pode observar que só 38 dos 92 concellos teñen establecementos hoteleiros catalogados cunha ou máis estrelas.

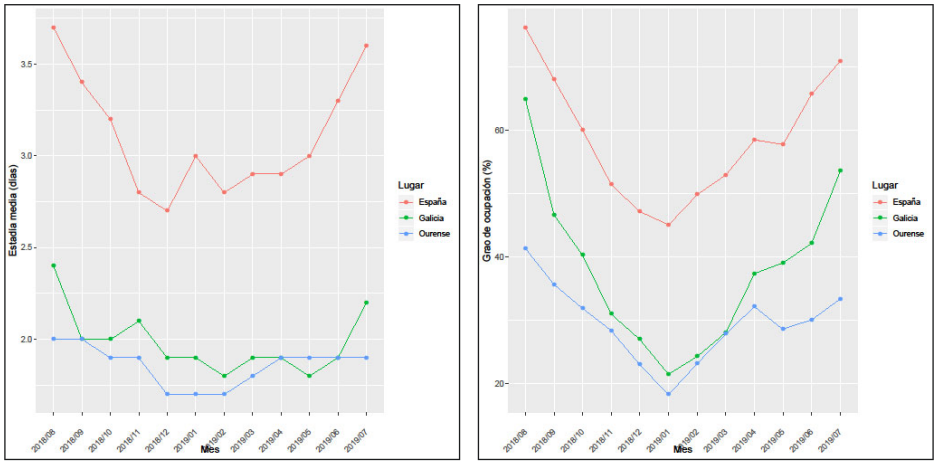

Figura 5. Comparativa da estadía media no ámbito provincial, galego e estatal (esq.), e comparativa do grao de ocupación no ámbito provincial, galego e estatal (der.)

Categoría	Establecementos	Cuartos	Prazas
5 estrelas	0	0	0
4 estrelas	15	837	1549
3 estrelas	8	219	500
2 estrelas	24	503	905
1 estrela	46	845	1528

Táboa 2. Datos dos hoteis con estrelas na provincia de Ourense

Lugar	Establecementos	Cuartos	Prazas
Allariz-Maceda	8	183	335
Baixa Limia	4	151	338
O Carballiño	10	256	563
A Limia	4	81	138
Ourense	32	1011	1793
O Ribeiro	5	188	370
Terra de Caldelas	4	35	44
Terra de Celanova	2	36	74
Terra de Trives	2	30	59
Valdeorras	10	192	329
Verín	4	83	162
Viana	8	158	277

Táboa 3. Datos dos hoteis con estrelas por comarca

3.2. Os portais TripAdvisor e Booking

3.2.1. Análise descritiva

Nesta sección resúmense as características dos establecementos hoteleiros da provincia de Ourense, así como as valoracións e os comentarios realizados polo usuariado, información extraída das plataformas TripAdvisor e Booking e proporcionada polo subproxecto EPOCa (véxase a táboa 4). A grandes trazos, a información proporcionada das mencionadas plataformas consta de dúas partes:

- Os datos identificativos de cada establecemento rexistrado (nome, localización, categoría –no seu caso–); servizos ofertados, tanto no propio establecemento coma na súa contorna; e unha ou varias valoracións globais.
- Cada comentario achega a opinión do usuario/a; a súa valoración xeral e sobre diversas características; e, en maior ou en menor medida, información sobre o propio usuario/a.

A estes datos engádense os relativos ao número de cuartos e de prazas de cada establecemento, información que proporciona o directorio de empresas e de establecementos turísticos (Turismo de Galicia, 2019).

	TripAdvisor	Booking
Comentarios	5806	22227
Establecementos	154	137
Concellos	57	38
Cuartos	2148	1355
Prazas	4360	2740
Servizos	153	418

Táboa 4: Resumo dos datos totais extraídos de TripAdvisor e Booking

3.2.2. Plataforma TripAdvisor

Desta plataforma disponse dun total de 5806 comentarios referidos a 154 establecementos hoteleiros, que se atopan distribuídos en 57 dos 92 concellos de Ourense e que teñen presenza en todas as comarcas da provincia (figura 6). O reconto do número de cuartos destes establecementos ascende a 2148 cuartos, que permiten aloxar un total de 4360 hóspedes (figura 7). TripAdvisor realiza unha clasificación dos hoteis en

estrelas tendo en conta a calidade dos seus servizos e das instalacións, unhas categorías que non teñen por que coincidir coas clases reais dos establecementos.

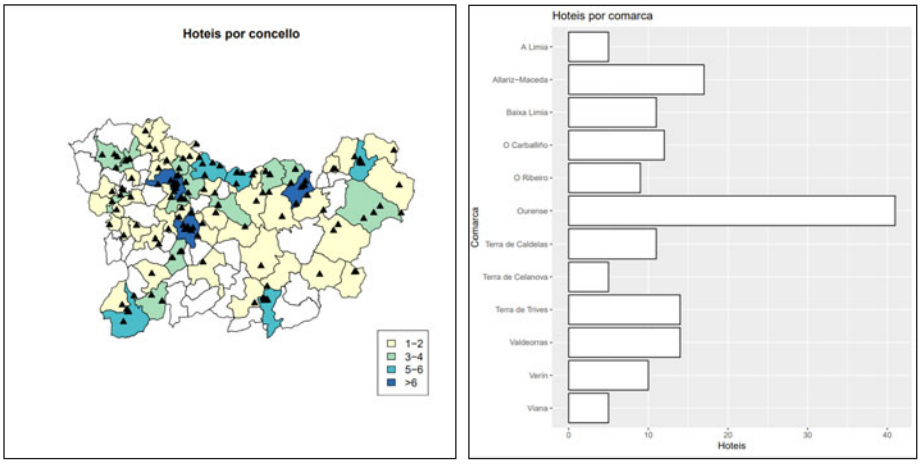

Figura 6. Comparativa da estadía media no ámbito provincial, galego e estatal (esq.), e comparativa do grao de ocupación no ámbito provincial, galego e estatal (der.)

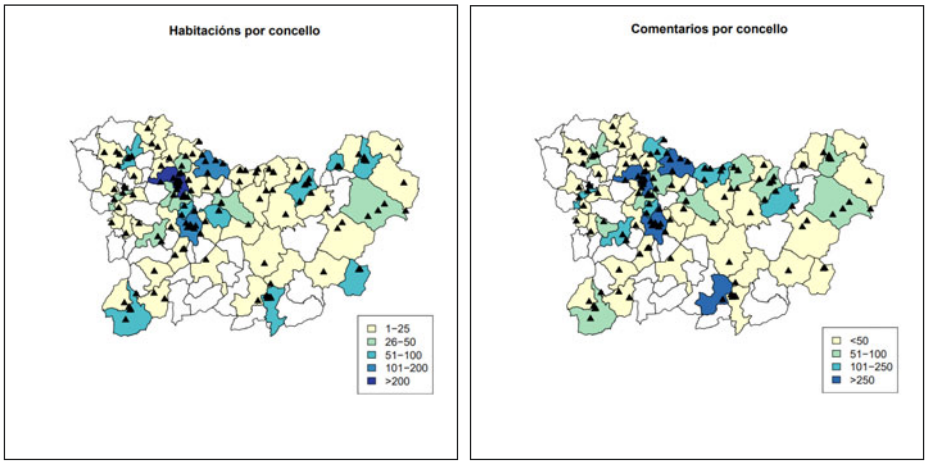

Figura 7. Comparativa da estadía media no ámbito provincial, galego e estatal (esq.), e comparativa do grao de ocupación no ámbito provincial, galego e estatal (der.)

As variables recollidas nesta plataforma poden dividirse en dous grupos:

- Información dos servizos. Trátase de 153 variables de tipo binario que indican os servizos ofertados polos establecementos como, por exemplo, se conta con restaurante, zona infantil, aparcadoiro de balde, internet de alta velocidade, piscina...

- Valoracións dos servizos. Puntúanse os seguintes aspectos:
  - Valoración total do establecemento. Puntuación calculada como a media das valoracións individuais que realiza cada usuario/a.
  - Valoracións da localización, limpeza, servizo e calidade-prezo. Valoracións xerais calculadas segundo as respostas dos usuarios/as.
  - Valoración individual. Puntuación xeral que cada usuario/a lle outorga ao establecemento.

A puntuación individual toma valores enteiros entre 0 e 5, mais as restantes poden tomar valores intermedios. Hai nunha alta porcentaxe de casos nos que o usuariado dá unha valoración de 3 puntos ou superior.

Cada comentario ou opinión permite tamén extraer certa información sobre o usuario/a que o realizou:

- Tipo de viaxe. Clasifícanse en cinco categorías (negocios, parella, familia, amizades e só).
- Data de estancia. Período no que realizou a súa estancia no establecemento.
- Contribucións realizadas. Número de achegas do usuario/a en TripAdvisor.
- Votos útiles. Número de persoas que consideran útil as achegas do usuario/a.
- Nivel de colaboración. Categoría á que pertence o usuario/a en función dunha puntuación outorgada polas diversas achegas que se poden realizar (opinións, fotos, publicacións, votos útiles...). Consta de seis niveis: 1 (máis de 300 puntos), 2 (máis de 500), 3 (máis de 1000), 4 (máis de 2500), 5 (máis de 5000) e 6 (máis de 10000).

Atendendo á categoría do establecemento, as peores valoracións medias recíbenas en xeral os de dúas estrelas. Para establecementos sen categoría ou cunha estrela, as valoracións dos distintos aspectos son moi similares. Por outra parte, os hoteis de tres estrelas reciben a menor puntuación na valoración individual (que dá lugar á xeral do establecemento) e os de catro estrelas, na relación calidade-prezo. Neste último caso, contraponse á alta puntuación que reciben en canto á súa localización. Ao considerar as valoracións dos usuarios e usuarias segundo o tipo de viaxe, obsérvase que as puntuacións máis altas en todas as clases de viaxe adoitan ser para a localización dos establecementos. Pola contra, a relación calidade-prezo é a menos valorada nas viaxes en familia, parella ou de amizades e a valoración individual rexistra os datos máis baixos no caso de usuarias e usuarios que se desprazan por negocios ou sós.

Na figura 8 pódese ver o número dos comentarios por concello (esquerda), así como a valoración media de cada municipio (dereita). Para calcular a citada valoración, tomouse a media das valoracións totais dos establecementos situados no correspondente concello, polo que tomará valores entre 0 e 5.

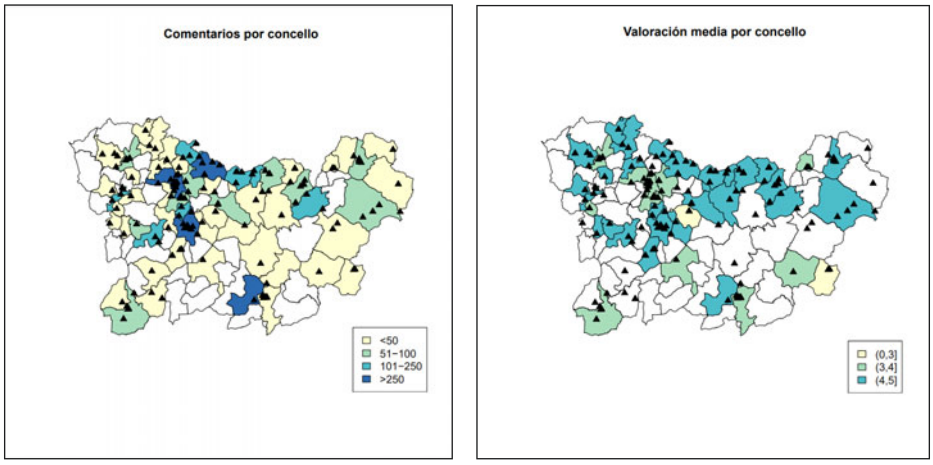

Figura 8. Comparativa da estadía media no ámbito provincial, galego e estatal (esq.), e comparativa do grao de ocupación no ámbito provincial, galego e estatal (der.)

3.2.3. Plataforma Booking

Desta plataforma extraéronse un total de 22227 comentarios relativos a 137 establecementos hoteleiros localizados en 38 dos 92 concellos cos que conta a provincia de Ourense (véxase a figura 9). A diferenza de TripAdvisor, non aparece rexistrado ningún establecemento en Booking localizado na comarca da Terra de Celanova.

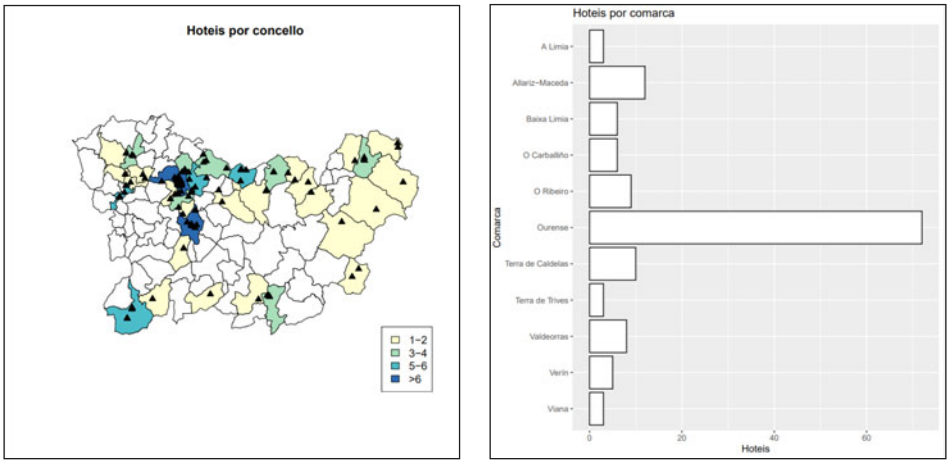

Figura 9. Número de establecementos por concello (esq.) e comarca (der.) rexistrados en Booking

O número de cuartos e de prazas dos establecementos ourensáns rexistrados en Booking ascende a 1355 e 2740, respectivamente, (véxase a figura 10). Esta plataforma clasifica os hoteis en función das súas estrelas e crea tamén categorías para os apartamentos ou vivendas de uso turístico rexistrados.

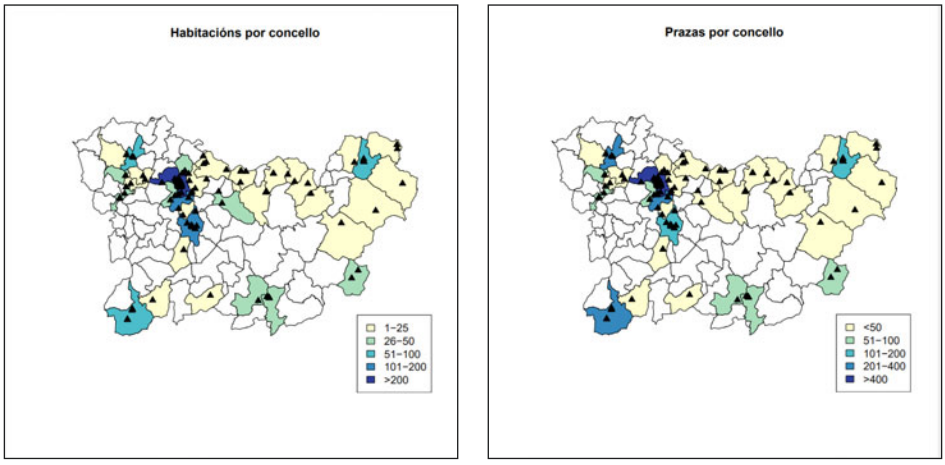

Figura 10. Número de establecementos por concello (esq.) e comarca (der.) rexistrados en Booking

Os dous grupos de variables recollidos en Booking son:

- Información dos servizos. Son 418 variables binarias que indican os servizos ofertados polos establecementos como, por exemplo, aparcadoiro público, conexión a internet, piscina, idiomas de atención á clientela...
- Valoracións dos servizos. Avalíanse oito aspectos (establecemento en xeral, persoal, instalacións/servizos, limpeza, confort, relación calidade-prezo, localización e wifi gratuíta), aos que se engade unha valoración que fai de xeito individual cada usuario ou usuaria.

Neste caso, as valoracións están puntuadas entre 0 e 10, e poden tomar calquera valor intermedio. Obsérvase que estas superan os 5 puntos na totalidade ou nunha porcentaxe moi elevada dos casos. Convén engadir, ademais, que a valoración xeral do establecemento se calcula como a media das valoracións individuais dos usuarios/as, motivo polo cal non aparece reflectida explicitamente na citada figura, como tamén ocorría en TripAdvisor.

Ao considerar as valoracións medias do usuariado por categoría de hotel e de apartamento, obsérvase que os establecementos de catro estrelas de ambos os tipos obteñen en xeral as puntuacións máis altas en todos os aspectos valorados.

Agora, a información do usuario/a que se pode obter limítase a un nome que non se corresponde cun usuario/a rexistrado ou único e a unha data de aloxamento. Ademais, para poder opinar en Booking, o usuario/a debe reservar ou hospedarse previamente no establecemento. Isto non ocorre en TripAdvisor, onde calquera persoa pode deixar a súa opinión sen necesidade de ter estado no establecemento. Do mesmo xeito ca antes, represéntanse o número de comentarios e a valoración media por concello na figura 11; esta última toma valores entre 0 e 10.

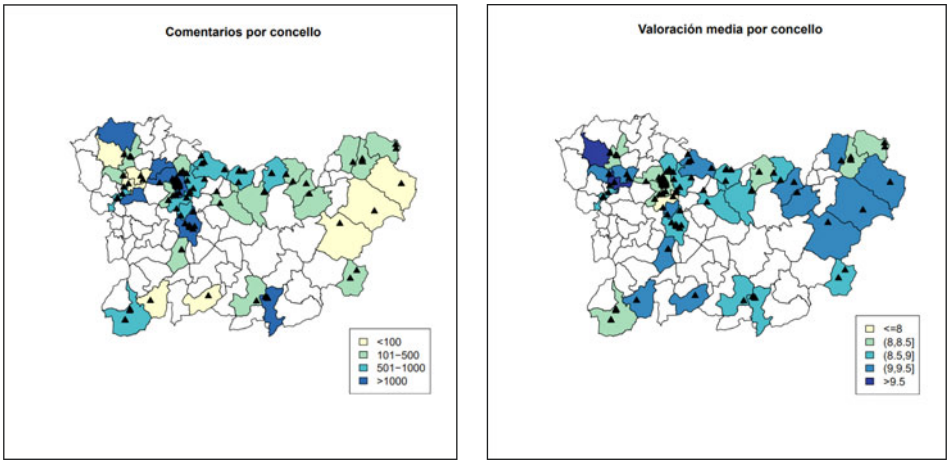

Figura 11. Número de establecementos por concello (esq.) e comarca (der.) rexistrados en Booking

3.3. Análise das opinións

As opinións do usuariado recollidas nas plataformas TripAdvisor e Booking contan cunha parte cuantitativa, correspondente ás puntuacións realizadas aos distintos aspectos do establecemento; e unha cualitativa, a relativa á experiencia da súa estancia no hotel que está reflectida no comentario realizado.

Para o tratamento dos comentarios, empregouse unha técnica de minaría de textos denominada análise de sentimentos, que permite converter textos en variables cuantitativas indicando o sentimento global do comentario.

Unha vez transformados os comentarios en variables numéricas, aplícanse modelos para estudar a valoración do sector hoteleiro.

3.3.1. Análise de sentimentos

A técnica de análise de sentimentos permitiu extraer unha valoración nunha escala continua dos comentarios do usuariado. A partir desta escala púidose modelar a valoración global do establecemento en función do comentario.

De forma xeral, unha vez dada unha opinión transfórmase en palabras recollidas nun dicionario onde previamente se lles asignou un valor ou unha clase de opinión.

Aínda que de xeito xeral se poida denominar análise de contido semántico, empregarase calquera das dúas denominacións para referirse a unha técnica de minaría de textos consistente en clasificar de xeito automático textos (ben sexan documentos enteiros, parágrafos, frases...) en función da información que conteñan. No noso caso, dado que as opinións adoitan ser positivas ou negativas, tratarase de extraer a polaridade de cada comentario. Para calcular esta polaridade, botarase man de dicionarios ou listaxes de palabras positivas e negativas, coas cales se comparará cada termo do comentario. As palabras que conforman os dicionarios teñen asociados valores -1 e 1 considerando termos negativos e positivos, respectivamente. O procedemento xeral (véxase a figura 12) consiste en atopar para cada comentario os termos que aparecen no dicionario, convertelos nos valores numéricos asociados e, posteriormente, transformalos nun valor *s*.

$$\text{sentence} \Rightarrow \begin{pmatrix} word_1 \\ word_2 \\ \vdots \\ word_k \end{pmatrix} \Rightarrow \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{pmatrix} \Rightarrow s$$

Figura 12. Esquema de funcionamiento da análise de sentimentos

Como resultado, obterase un valor numérico comprendido no intervalo [–1, 1]. O *software* estatístico R [9] dispón dunha serie de paquetes que permiten realizar esta análise como, por exemplo, *syuzhet*, *Rsentiment*, *sentimentr* ou *SentimentAnalysis* (véxase [10] e referencias dentro). Os dous últimos foron os empregados neste traballo, pois están baseados nos monogramas, é dicir, toman cada termo do comentario para comparalo co dicionario, a diferenza doutros que poden tomar combinacións de dous, tres ou máis termos (n-gramas).

3.3.2. SentimentAnalysis

Nun primeiro momento, utilizouse *SentimentAnalysis*, realizando unha serie de tarefas previas á súa execución, denominadas preprocesamento. Nas técnicas de minaría de datos os comentarios adoitan almacenarse nunha especie de base de datos para

xestionar posteriormente. Esta colección recibe o nome de corpus [11]. Neste inclúense unicamente os comentarios escritos en lingua castelá, que conforman a porcentaxe máis alta. Unha vez elaborado o corpus, aplícanselle distintos métodos de limpeza e de estruturación para un conxunto máis simplificado para posteriores análises. Entre as transformacións máis comúns están a conversión de maiúsculas en minúsculas, a eliminación das *stopwords* (as palabras máis frecuentes dunha lingua), signos de puntuación, números e espazos en branco innecesarios, e a lematización *(stemming)* ou transformación a lemas das palabras que compoñen os comentarios [11]. Ademais, tamén adoita ser habitual eliminar os termos cun número reducido ou relativamente grande de caracteres, xa que a información que achegan adoita ser ínfima. Unha vez depurado o corpus, acostúmase a crear a matriz de documentos-termos, cuxas filas se corresponden cos documentos (neste caso, comentarios) e as columnas cos termos ou palabras, xa que moitas aplicacións toman a devandita matriz como parámetro de entrada. Na táboa 5 recóllense as dimensións (é dicir, comentarios e termos) das matrices resultantes para os datos de TripAdvisor e de Booking. A modo ilustrativo, a figura 13 mostra, en forma de nubes de palabras, os termos máis frecuentes nos comentarios extraídos das mencionadas plataformas.

Plataforma	Comentarios (filas)	Termos (columnas)
TripAdvisor	5739	17637
Booking	20064	11255

Táboa 5. Dimensións das matrices de documentos-termos para cada plataforma

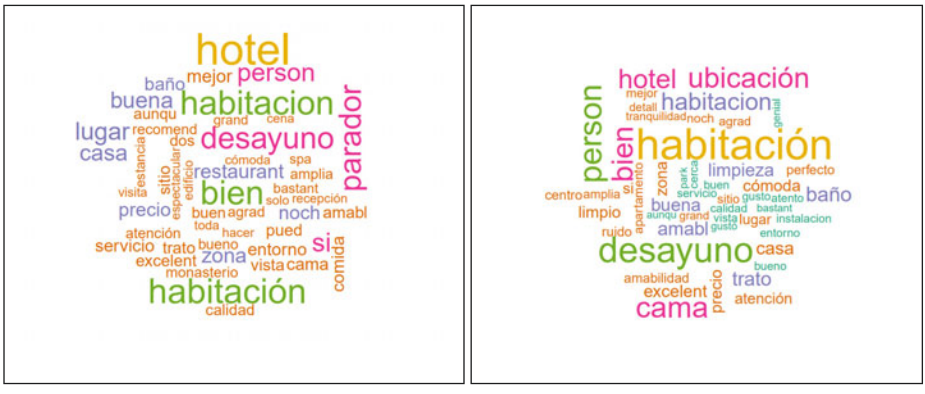

Figura 13. Termos máis frecuentes nos comentarios de TripAdvisor (esq.) e de Booking (der.)

De forma xeral, *SentimentAnalysis* devolve para cada texto e cada dicionario empregado os seguintes valores continuos:

• Puntuación de sentimento

$$\frac{\#positivo - \#negativo}{\#all}$$

onde *#positivo* e *#negativo* se refiren ao número de palabras positivas e negativas do dicionario empregado que aparecen no texto, respectivamente; e *#all* ao número de «palabras significativas» (substantivos, adxectivos, verbos, adverbios) do texto. Cómpre destacar que a puntuación está no intervalo [-1; 1] e que o valor 0 indica que hai tantas palabras positivas coma negativas.

• Puntuación de negatividade

$$\frac{\#negativo}{\#all}$$

tomando valores en [0; 1].

• Puntuación de positividade

$$\frac{\#positivo}{\#all}$$

tomando valores en [0; 1].

• Ratio

$$\frac{\#positivo + \#negativo}{\#all}$$

SentimentAnalysis conta con varios dicionarios, dos cales tomamos o *Harvard-IV* (referido como *DictionaryGI*), conformado por dúas coleccións de palabras positivas e negativas, respectivamente (táboa 5). Ademais, tomamos o léxico de opinión de Hu e Liu [12], que tamén está constituído por dúas coleccións de palabras negativas e positivas [12] (táboa 6). Xa que estes dicionarios recollen termos en inglés, tivemos que realizar a súa tradución para poder empregalos en castelán.

Palabras	DictionaryGI		Hu e Liu	
Positivas	1637	44,95 %	2005	29,55 %
Negativas	2005	55,05 %	4780	70,45 %
Total	3642		6785	

Táboa 6. Datos dos dicionarios

Na figura 14 obsérvase o resultado de estimar a densidade dos sentimentos ou do contido semántico (proporcionado pola rutina *analyzeSentiment()* de *SentimentAnalysis*) en función de distintos valores da valoración individual. Destácase que, empregando calquera dos dous dicionarios arriba mencionados, se obteñen polaridades moi próximas ao 0. As curvas de densidade están moi solapadas, polo que o carácter preditivo do sentimento é escaso. Isto é, dado un valor calquera do sentimento/polaridade, non se pode distinguir a que valor da puntuación individual se corresponde. Esta mesma situación ocorre cando empregamos os datos de Booking.

Estes malos resultados manifestan algúns problemas que apareceron:

- Eliminar as *stopwords* pode cambiar o sentido do comentario. Por exemplo, onde nun primeiro momento está a frase «No recomendaría este sitio» (comentario claramente negativo), descartando eses termos máis comúns do idioma queda a frase «Recomendaría sitio», que parece indicar algo positivo.
- A lematización en castelán non é perfecta, pois non reduce a lemas todas as palabras e, ás veces, faino de xeito «estraño». Para mostra, na figura 14 obsérvanse varios termos que comparten unha raíz ou un lema común, e que se deberían considerar como un único, e outros cuxa redución resulta «estraña» ou «absurda». Por exemplo, no caso de TripAdvisor *buena*, *bueno* e *buen* comparten a raíz *buen*, termo ao que se deberían reducir as dúas primeiras palabras. No caso de Booking, atopamos *aunqu**, lematización absurda de *aunque*.
- A polaridade e o dicionario empregado. A alta densidade de comentarios cuxo sentimento se sitúa arredor do valor 0 fai pensar que algo está a fallar no cálculo da polaridade, pois non parece normal que a maioría dos comentarios sexan neutros. Aquí inflúen dous factores:
    - O dicionario empregado. Xa que en castelán non se atopan dicionarios para a análise de sentimentos, tradúcense os arriba mencionados. A tradución automática do inglés ao castelán implica, en moitos casos, a perda de infor-

mación, posto que un termo inglés cunha soa forma pode ter máis dunha variante en castelán (formas masculina, feminina, singular e plural, por exemplo). Isto pode provocar que moitas palabras se pasen por alto ao aplicar a análise.

- A forma de calcular a polaridade. Dividir polo número total de palabras parece o razoable cando todas (ou case todas) estas palabras teñen asignado un valor ou un tipo de opinión positivo ou negativo. Pero se, tal e como se indicou antes, o dicionario pasa por alto moitas palabras, estas non serán detectadas como positivas ou como negativas, polo que o numerador tenderá a ser pequeno, encamiñando a polaridade a ser próxima a 0. O mesmo pasaría se o comentario é relativamente longo.

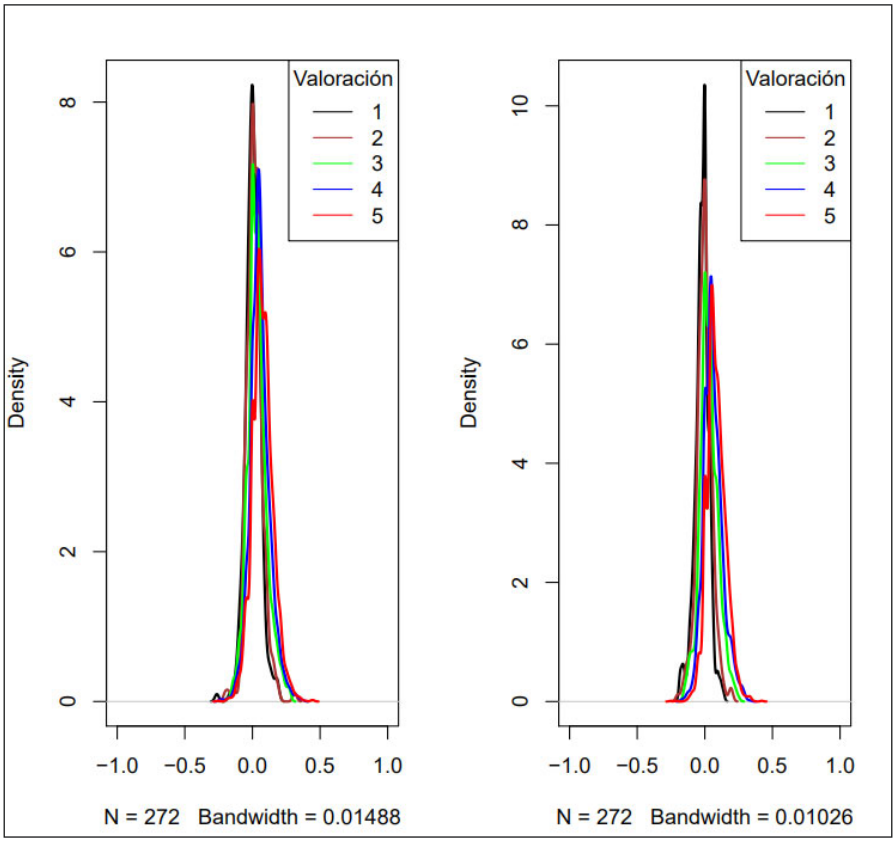

Figura 14. Densidades estimadas do sentimento empregando o dicionario *GI* (esq.) e o léxico de Hu e Liu (der.) en función da valoración individual para TripAdvisor

3.3.3. Sentimentr

Para tratar de paliar as eivas atopadas no uso de *SentimentAnalysis*, emprégase o paquete *Sentimentr*. Este conta cunha gran vantaxe que o fai preferible aos demais, pois ten en conta os potenciadores. Trátase de palabras que actúan sobre o monograma, alterando dun modo ou outro o seu valor: cambian o signo (negadores), aumentan ou diminúen o impacto do monograma sobre o total do comentario etc. Na táboa 7 vese un exemplo que ilustra esta situación: dadas dúas opinións aparentemente opostas, *SentimentAnalysis* valora ambas de xeito positivo, pois non ten en conta o adverbio de negación *no*, mentres que *Sentimentr* si que o considera, cambiando o signo do valor asociado ao adxectivo *bueno*.

	Comentarios	
Paquete	**«Este hotel es bueno»**	**«Este hotel no es bueno»**
SentimentAnalysis	0,5	0,5
Sentimentr	1	-1

Táboa 7. Comparativa entre *SentimentAnalysis* e *Sentimentr*

Convén salientar que no uso de *Sentimentr* prescindiuse de eliminar as *stopwords* e de realizar a lematización polos problemas vistos no apartado anterior. No caso do dicionario, empregouse unicamente o léxico de Hu e Liu (maior número de termos ca o dicionario *GI*) e tratouse de corrixir os posibles defectos de tradución, así como traducir tamén a listaxe de potenciadores desde o inglés ao castelán. Nun primeiro momento, aplicouse a rutina *sentiment()* para obter a polaridade de cada comentario, pero as densidades volvían acumularse arredor do 0, debido ao denominador (no que se teñen en conta todas as palabras e os potenciadores). Para tratar de mellorar o resultado, cambiouse o citado denominador por

#positivo + #negativo;

o que leva aparellado que nesa suma só se teñen en conta as palabras detectadas polo dicionario, número inferior ao total de termos que compoñen o comentario. A pesar destas consideracións, os resultados non foron moito mellores aos obtidos anteriormente, como mostran as densidades estimadas representadas na figura 15 para o caso de TripAdvisor.

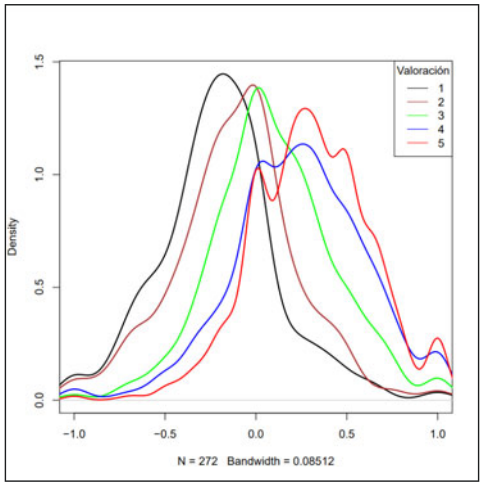

Figura 15. Densidades estimadas do sentimento en función da valoración individual para TripAdvisor

3.3.4 Análise da valoración individual

Xa que a valoración individual do hotel non é unha relación lineal perfecta coas valoracións das distintas características, estableceuse un marco teórico aditivo na súa relación entre a valoración individual do establecemento e as distintas características mais a valoración dos comentarios, i. e.:

$$Val.\ individual = f\ (val.\ caract.) + g\ (val.\ cont.\ semántico)$$

A función *g()* aplícase ao resultado dada pola análise de sentimentos da subsección anterior.

Desde o ámbito do turismo, considérase que na valoración individual inclúen as valoracións que o usuario ou usuaria fai das diferentes características do establecemento, así como a experiencia na súa estancia, que se ve traducida en forma de comentario. Baixo este marco teórico, proponse a construción de modelos explicativos que trate de explicar a mencionada puntuación individual a partir da información que se ten a disposición. A valoración individual corresponderase coa variable dependente Y, as demais valoracións son as variables explicativas X_1, ..., X_p e a explicativa relativa ao sentimento ou ao contido semántico denotarase por X_p. Para comprobar como inclúe o contido semántico na valoración individual, construiranse modelos de regresión sen e coa correspondente variable explicativa X_p.

A partir deste modelo teórico aplicáronse diferentes técnicas de regresión e de *data mining*, na procura dun modelo de predición e clasificación. Os modelos finais que deron mellor resultado foron os modelos de regresión xeneralizada, regresión ordinal e o de Random Forest.

Regresión xeneralizada. No caso de TripAdvisor, a valoración individual toma unicamente os valores enteiros 1, 2, 3, 4 e 5, que se poderían corresponder con outras tantas categorías ou clases de valoración, por exemplo, «moi mala», «mala», «regular», «bo» e «moi bo». Considerando entón a variable resposta *Y* como unha variable categórica, cabe pensar na posibilidade de construír modelos de regresión xeneralizada, no noso caso usaremos regresión loxística linear e aditiva.

Regresión loxística. A principal característica da regresión loxística é que a variable resposta sexa dicotómica ou binaria, isto é, que tome os valores 0 ou 1. Para iso, a partir de *Y* construiranse cinco variables binarias do seguinte xeito:

$$Y(j)_i = \begin{cases} 1, & \text{se } Y_i = j \\ 0, & \text{noutro caso} \end{cases} \quad 1 \leq i \leq n, \;\; j \in \{1,2,3,4,5\} \qquad (1)$$

A probabilidade de éxito será *P(Y(j) = 1)*. Nesta situación, asúmese unha relación lineal entre as variables explicativas e a *log-odds* do evento *Y(j) = 1*, polo que o correspondente modelo se formula como segue:

$$\log \frac{\mathbb{P}(Y(j)_i = 1)}{\mathbb{P}(Y(j)_i = 0)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{ip-1} + \beta_p x_{ip}.$$

Neste subproxecto, o axuste de cada modelo dá a probabilidade de que cada comentario se corresponda coa respectiva categoría que determinan as valoracións individuais. Tamén se pode calcular como:

$$\mathbb{P}(Y(j)_i = 1) = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_{p-1} x_{ip-1}+\beta_p x_{ip}}}{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_{p-1} x_{ip-1}+\beta_p x_{ip}}+1} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_{p-1} x_{ip-1}+\beta_p x}}$$

A rutina *glm()* permite calcular ambos os modelos loxísticos, un sen considerar a variable do sentimento e outro engadíndoa. Nas matrices de confusión elaboradas cos axustes obtidos vese a porcentaxe de clasificación correcta, que tende a mellorar considerando o contido semántico, pero que se mantén en valores similares aos obtidos cos modelos lineais e aditivos. Os estatísticos globais Accuracy e Kappa seguen na mesma liña ca no caso dos modelos citados (táboa 8).

	Accuracy	Kappa
Loxístico	0,5217	0,0914
Loxístico (contido)	0,5327	0,1291

Táboa 8. Estatísticos globais das matrices de confusión para os modelos loxísticos en TripAdvisor

Regresión loxística aditiva. Agora a modelización supón que a relación entre as *v.* explicativas e a función enlace é non paramétrica coa restrición de que sexa aditiva. É dicir, para *j = 1, ..., 5* defínese a *v.* binaria $Y_j = I\{Val.Ind=j\}$:

$$P(I_{\{\text{Val.Ind}=j\}} \,|\mathbf{x}_i) = link\,(f(\text{val. caract.}_i) + g(\text{val. cont. semántico}_i))$$

Definindo $p_{ij} = P(Y_j = 1\,|\mathbf{x}_i)$, obtéñense para cada j:

$$\hat{p}_{ij} = logit\left(\hat{f}_j(\text{val. caract.}_i) + \hat{g}_j(\text{val. cont. semántico}_i)\right),$$

e asígnase a clase correspondente como:

$$\hat{J}_i = \underset{j=1,\ldots,5}{\text{argmax}}\ \hat{p}_{ij}$$

Cómpre destacar que as probabilidades estimadas en j = 1, …, 5 non suman 1. Na táboa 9 amósase a matriz de confusión resultante da clasificación do modelo (columnas) e os valores reais (filas).

		Predición					
		1	**2**	**3**	**4**	**5**	**Total**
	1	**43,01**	0,00	8,82	21,32	26,84	271
	2	16,81	**0,00**	13,36	31,90	37,93	232
Referencia	3	4,64	0,00	**12,69**	35,14	47,52	645
	4	1,66	0,00	2,55	**23,39**	72,40	1553
	5	0,60	0,00	0,32	7,63	**91,45**	2754

Táboa 9. Matriz de confusión do modelo loxístico aditivo con TripAdvisor (%)

Como estatísticos globais, na táboa 10 dáse a Accuracy ou a proporción de datos correctamente clasificados e o Kappa que reflicte a mellora que se produce no modelo, se se considera toda a información dispoñible fronte a unha situación na que se carece de información.

		TripAdvisor	
N.º clases	Estat. / Modelo	Lineal	Aditivo
5	Accuracy	53,27	56,85
	Kappa	12,91	23,04

Táboa 10. Estatísticos resumo do modelo loxístico e do modelo de regresión xeneralizada con compoñentes aditivas

Idénticos resultados atópanse para Booking.

Regresión ordinal. Na regresión ordinal non é preciso modificar a variable resposta ou construír outras apropiadas, pois admite variables con máis de dúas categorías, que deben vir marcadas por unha orde. Así, para a variable resposta *Y* con cinco categorías defínese o correspondente modelo lineal como segue:

$$\log \frac{\mathbb{P}(Y_i \leq j)}{\mathbb{P}(Y_i > j)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{p-1} x_{ip-1} + \beta_p x_{ip}, \quad i \leq n, \;\; j \in \{1,2,3,4,5\}$$

A rutina *polr* do paquete MASS [13] permítenos construír ambos os modelos ordinais, un considerando a variable do contido semántico e o outro non. Os resultados (táboa 11) seguen na mesma liña dos modelos anteriores.

	Accuracy	Kappa
Ordinal	0,5275	0,1238
Ordinal (contido)	0,5453	0,1948

Táboa 11. Estatísticas resumo do modelo de regresión ordinal

Random Forest para clasificación. Esta técnica de clasificación, tamén denominada bosques aleatorios, está baseada nas árbores de decisión, método no que se realizan particións binarias dos datos de xeito recursivo e cuxo resultado se pode representar en forma de árbore [11]. No caso do Random Forest xéranse moitas árbores de decisión. Este método emprega o concepto de *bagging*, no que o conxunto de datos que servirán de adestramento se seleccionan ao azar con substitución, é dicir, que un mesmo dato pode saír varias veces. Con cada selección, adéstrase o modelo, mentres que o resto do conxunto de datos se empregará para probar o algoritmo.

A algorítmica do Random Forest vén dada por:

1. *Boostrap* con substitución: réplicas ←→ árbores.
2. Para cada réplica constrúese unha árbore de clasificación e, polo tanto, unha predición.
3. A predición final consiste en asignar a clase con maior frecuencia de aparición.

Neste subproxecto empregouse o paquete Random Forest [14] de R para levar a cabo esta técnica. Tomouse de xeito aleatorio unha mostra formada polo 80 % dos nosos datos, que se empregou para adestrar o algoritmo. Unha vez feito isto, aplicouse ao restante conxunto de datos. Os resultados (táboas 12 e 13) que devolve Random Forest seguen a liña dos que proporcionaron todos os métodos anteriores.

		Predición					
		1	**2**	**3**	**4**	**5**	**Total**
	1	**34,55**	3,64	18,18	16,36	27,27	271
	2	21,15	**3,85**	21,15	28,85	25,00	232
Referencia	3	5,30	3,79	**16,67**	36,36	37,88	645
	4	1,50	0,60	5,39	**25,75**	66,77	1553
	5	0,56	0,19	1,49	10,97	**86,80**	2754

Táboa 12. Matriz de confusión do modelo de Random Forest para TripAdvisor (%) con 80 % para *train* e 20 % de test

	Accuracy	**Kappa**
Ordinal	0,5356	0,2207

Táboa 13. Estatísticos resumo do modelo de Random Forest

4. Conclusións

Neste traballo obtivéronse de forma automática datos de comentarios sobre hoteis da provincia de Ourense extraídos dos portais TripAdvisor e Booking coa intención de obter información que posibilitase mellorar os servizos prestados por tales establecementos. Para tal fin, empregáronse métodos de regresión e de clasificación (análise de sentimentos, Random Forest), dado que todos os métodos devolveron resultados moi

similares e, á súa vez, con pouco poder preditivo. Se nos fixamos nos valores que toma a variable «valoración individual» (que desempeñou o papel de variable dependente ou resposta), hai unha categoría (valor 5) que sobresae entre as demais, pois alberga máis do 50 % dos datos, mentres que a porcentaxe restante se ten que repartir entre as outras catro categorías. Isto é sinal dunha mostra *desbalanceada* e provoca que a mellor clasificación se dea precisamente na categoría que conta con maior cantidade de datos. Estratexias como o *subsampling* ou similares permitirían *balancear* as categorías e mellorar o poder clasificador.

Como xa indicou, a ausencia dun dicionario de termos en castelán para a análise de sentimentos provocou algúns atrancos nesta tarefa, polo que a creación dunha listaxe de palabras para esta lingua sería un gran paso para mellorar os resultados desta análise. Tamén se podería considerar combinar o castelán co galego, dadas as similitudes e o uso indiscriminado de palabras de ambas as linguas. Tal dicionario non só debería recoller termos positivos e negativos, senón que sería preciso incluír:

- Erros ortográficos máis frecuentes na correspondente lingua.
- Neoloxismos, que son palabras ou expresións de recente creación nunha lingua ou collida prestada doutra desde hai pouco tempo.
- Emoticonos, dado o seu uso xeneralizado.
- Así mesmo, no caso de usar o paquete S*entimentr*, debería terse tamén unha boa listaxe de potenciadores.

5. Referencias

[1] "Minerva Travel 2017" https://www.thinkwithgoogle.com/intl/es-es/insights/comportamiento-de-los-viajeros-espa%C3%B1oles-en-2017-desde-la-inspiraci%C3%B3n-hasta-el-destino/

[2] Moya, D.H. e Majó, J. (2017). Análisis de comentarios en redes sociales para mejorar la reputación online hotelera. Turismo y Sociedad. 20, (jul. 2017), 169-190. doi:10.18601/01207555.n20.09.

[3] Olmeda, I. e Sheldon P.J. (2008) Data Mining Techniques and Applications for Tourism Internet Marketing. Pages 1-20 | Received 01 Oct 2000, Accepted 30 Jun 2001, Published online: 13 Oct 2008. doi: 10.1300/J073v11n02_01

[4] Dean, J. e Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. OSDI 2004.

[5] García S., Ramírez-Gallego S., Luengo J. e Herrera F. (2016). Big Data: Preprocesamiento y calidad de datos. Novática (n. 237)

[6] White, T. (2012). Hadoop, The Definitive Guide. O'Reilly Media.

[7] Karau, H., Konwinski, A., Wendell, P. e Zaharia, M. (2015). Learning Spark: Lightning-Fast Big Data Analytics. O'Reilly Media.

[8] Castrillo-Fernández O. (2015). Web Scraping: Applications and Tools. European Public Sector Information Platform (EPSI Platform). Topic Report Nº 2015/10.

[9] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

[10] Naldi M (2019, xaneiro). A review of sentiment computation methods with R packages. arXiv preprint arXiv:1901.08319.

[11] Calvo Torres M (2017). Text Analytics para Procesado Semántico (Traballo de fin de máster, Universidade de Vigo). Recuperado de \url{http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1475.pdf}

[12] Hu M, Liu B (2004, agosto). Mining and summarizing customer reviews. En: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168-177.

[13] Venables WN, Ripley BD (2002). Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, http://www.stats.ox.ac.uk/pub/MASS4

[14] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.