Fall 10-16-2021

Most valued factors in rural tourism: An analysis of Portuguese customer comments on a booking platform

Rui Esteves
Polytechnic of Coimbra, Coimbra Business School Research Centre, ISCAC, ruiesteves23@gmail.com

Fernando Paulo Belfo
Instituto Politécnico de Coimbra, pbelfo@iscac.pt

Antonio Trigo
Instituto Politecnico de Coimbra, antonio.trigo@gmail.com

Fatores mais valorizados no turismo rural: Uma análise de comentários de clientes portugueses numa plataforma de reservas

Most valued factors in rural tourism: An analysis of Portuguese customer comments on a booking platform

Rui Esteves, Polytechnic of Coimbra, Coimbra Business School Research Centre| ISCAC, Coimbra, Portugal, ruiesteves23@gmail.com

Fernando Paulo Belfo, Polytechnic of Coimbra, Coimbra Business School Research Centre| ISCAC, Coimbra, Portugal, pbelfo@iscac.pt

António Trigo, Polytechnic of Coimbra, Coimbra Business School Research Centre| ISCAC, Coimbra, Portugal, antonio.trigo@gmail.com

Resumo

Este estudo pretende descobrir e analisar os fatores mais valorizados no turismo rural. Apresenta-se uma análise de uma amostra com 9.939 comentários obtidos da plataforma de reservas Booking.com sobre estadias em estabelecimentos de turismo rural localizados em Portugal. Usando técnicas de mineração de texto e modelação por tópicos com o objetivo de encontrar os fatores mais valorizados, obtiveram-se nuvens de palavras que permitem uma análise rápida e intuitiva aos dados. O algoritmo LDA permitiu extrair dez tópicos coerentes com as nuvens de palavras. Os resultados apontam para que o pequeno-almoço, a simpatia do pessoal, os anfitriões e o espaço exterior sejam os fatores mais relevantes para os hóspedes. Estes resultados têm implicação para a gestão das unidades de turismo rural pois identifica os fatores a ter em atenção pelos responsáveis destas unidades.

Palavras-chave: turismo rural; booking.com; *web scraping;* CRISP-DM; mineração de texto; algoritmo LDA.

Abstract

This study aims to discover and analyze the most valued factors in rural tourism. An analysis of a sample with 9,939 comments obtained from the booking platform Booking.com on stays in rural tourism establishments located in Portugal is presented. Using text mining techniques and topic modeling in order to find the most valued factors, word clouds were obtained that allow a quick and intuitive analysis of the data. The LDA algorithm allowed to extract ten topics coherent with the word clouds. The results point out that breakfast, the friendliness of the staff, the hosts and the outdoor space are the most relevant factors for guests. These results have implications for the management of rural tourism units as they identify the factors to be taken into account by those in charge of these units.

Keywords: *rural tourism; booking.com; web scraping; CRISP-DM; text mining; LDA algorithm.*

1. INTRODUÇÃO

Nos últimos anos temos assistido a uma grande inovação tecnológica, em particular, no que diz respeito ao desenvolvimento da Internet, cada vez mais suportada em dispositivos móveis. Esta inovação tem conduzido a um novo contexto e afirmação do ser humano, possibilitando uma nova

era nas relações sociais. O turismo não foge desta tendência. Segundo Yi et al. (2017) o turista procura muitas vezes experiências cada vez mais autênticas, no sentido de se encontrar a si próprio. A autenticidade inerente à experiência está relacionada com a sua identidade, e reforça o sentimento de desenvolvimento pessoal e autorrealização.

Essa valorização e validação pessoal tem muitas vezes como pano de fundo a partilha das mesmas nas redes sociais. A tecnologia tem facilitado cada vez mais este processo de partilha, sendo este, hoje em dia, quase imediato. É hoje frequente alguém comentar a utilização de um produto ou de um serviço numa rede social ou numa plataforma de comércio eletrónico e receber comentários muito rapidamente por parte de amigos e conhecidos.

Por outro lado, a atual capacidade computacional disponível, aliada a técnicas e algoritmos de aprendizado de máquina (*machine learning*), permitem processar enormes quantidades de informação disponível em variados conjuntos de dados de diversas áreas. O aprendizado de máquina tem sido cada vez mais utilizado nos mais diversos domínios, seja no setor público, como nas áreas fiscal (Seiça et al. 2019), educação (C. Pimenta et al. 2018) e medicina (Brandão et al. 2021) ou, no sector privado, como no marketing (Cui et al. 2006), na indústria de *media* e entretenimento (Sereday & Cui, 2017), na indústria de eventos (Loureiro et al. 2014), no turismo (P. Pimenta et al. 2009) e em muitas outras áreas, contribuindo para a criação de novos conhecimentos e ajudando as organizações a definir estratégias que lhes permitam aumentar seu desempenho.

Aliado a esse facto está a própria oferta comercial que nos últimos anos teve de se adaptar às novas exigências e escrutínio constante por parte dos utilizadores. Este trabalho foca-se no turismo rural, a nível da sua oferta de serviços de alojamento. A oferta de serviços de alojamento irá estar, em certa medida, relacionada com os dois conceitos referidos, respetivamente, a procura de uma nova experiência por parte do cliente e os comentários que se encontram disponíveis sobre essa experiência. Os comentários são cada vez mais tidos em conta antes da tomada de decisão sobre a contratação do serviço. Por essa razão, os comentários merecem uma análise adequada que possa permitir, eventualmente, reposicionar e redefinir a oferta dos ditos serviços de alojamento.

Segundo a direção-geral de agricultura e desenvolvimento rural do estado português, o turismo rural apresenta uma evolução do modelo de sociedade em que vivemos. Não se trata de um fenómeno passageiro, mas antes uma tendência com um crescimento regular nos últimos anos, especialmente por parte de uma clientela culta, com poder económico superior à média, exigente de qualidade, de genuinidade e em busca das diferenças que o tornam atraente face às restantes modalidades de turismo (DGADR, 2020b).

Segundo o Instituto Nacional de Estatística, em julho de 2017, o turismo rural e o turismo de habitação dispunham de uma oferta de 1 419 estabelecimentos e 23,2 mil camas. No final desse ano,

atingiram os 794,7 mil hóspedes, correspondendo a 1,7 milhões de dormidas, respetivamente crescimentos de 18,8% e de 17,0% face ao ano anterior (INE, 2018). Em julho de 2019, havia 1 687 estabelecimentos de turismo no espaço rural e de habitação em atividade disponibilizando um total de 26,6 mil camas. Em 2019, estes estabelecimentos registaram 948,4 mil hóspedes (+11,8% do que em 2018), que proporcionaram 2,0 milhões de dormidas (+9,7%). A estadia média no ano de 2019 correspondeu a 2,07 noites, em média, e a taxa de ocupação-cama foi de 24,1% (INE, 2020).

No futuro, a tendência deverá manter-se, existindo neste momento uma clara aposta na qualidade da oferta. As unidades de alojamento estão cada vez mais adequadas e oferecem um vasto leque de serviços. Os produtos regionais com muita qualidade e o cuidado de fazer bem as coisas simples dão um caráter especial a este sector, procurado muitas vezes pelo descanso e tranquilidade que pode transmitir.

Um dos fatores que distingue as unidades hoteleiras de turismo rural do resto da oferta turística é o facto de grande parte deste tipo de estabelecimentos serem projetos individuais e raramente geridos por grupos hoteleiros, existindo um tratamento mais personalizado e pessoal. O turismo rural tem usualmente como característica específica, o caráter dos proprietários, a sua experiência percecionada e o seu conhecimento do local onde está situado o empreendimento, o que torna cada alojamento um caso único. Além do contacto com a natureza e atividades rurais, a grande maioria destas unidades tem vindo a adaptar a oferta às novas exigências dos consumidores que procuram um refúgio não massificado e centrado na natureza, um contacto e atendimento mais personalizado e uma maior atenção ao detalhe. Os comentários recebidos dos hóspedes são cruciais para entender as preferências dos mesmos (Publituris, 2019).

O comportamento dos consumidores tem sido nos últimos anos um dos campos mais pesquisados na área do marketing e turismo. Em Cohen et al (2014) são enumeradas algumas dimensões essenciais a ter em conta ao analisarmos o comportamento destes consumidores. Estas são a tomada de decisão, os seus valores, as suas motivações, personalidade, espectativas, atitudes, perceções, satisfação, confiança e lealdade.

Diversas plataformas possibilitam hoje intensa interação social, recolhendo e partilhando comentários de hóspedes, e assim potenciam várias possibilidades de análise de dados, em particular, das emoções humanas registadas. Tal conhecimento é bastante pertinente a nível da criação de valor e tem um papel crucial no desenvolvimento da oferta do serviço e da antecipação do mesmo, no esforço de ir de encontro às expectativas do cliente. As redes sociais têm uma importância bastante grande nesta matéria pois contêm vastas quantidades de informação que podem ser analisadas no contexto turístico. Diversas redes sociais como o Twitter ou Facebook tornaram-se fontes de dados que podem ser analisadas de forma possibilitar uma maior perceção do comportamento do potencial turista (Vu et al. 2018).

2. REVISÃO DA LITERATURA

2.1. Setor Turístico

O sector do turismo tem tido, nas últimas 6 décadas, um crescimento constante e consistente, sendo um dos principais motores económicos a nível mundial. Para além de gerar rendimento, em grande parte proveniente do estrangeiro, é importante ainda na criação de emprego e estimulação da cultura a nível nacional e regional. Tem ainda um papel substancial no suporte às comunidades locais e representa uma importante fatia nas exportações nos países envolvidos (OCDE, 2020).

De acordo com os dados obtidos pela Organização para a Cooperação e Desenvolvimento Económico (OCDE), referentes ao ano de 2018 o turismo contribui diretamente, nos países membros desta organização, em média, 4.4% para o PIB (Produto Interno Bruto), 6.9% a nível de criação de emprego e 21.5% relativamente a exportações relacionadas com o sector (OCDE, 2020). É referido no mesmo relatório que, no que diz respeito a números de chegadas em 2018, o número superou 1.4 biliões, representando um crescimento de 5.6% relativamente a 2017.

O setor do turismo no nosso país representa um peso importante na atividade económica. Em 2019, as exportações de turismo representavam o quarto valor mais elevado na área do euro, com 8,6% do Produto Interno Bruto (PIB) português. Em 2020, as exportações contribuíram negativamente em -5,5% para a redução de 8,1% do PIB. A componente do PIB com a queda mais acentuada (-56,6%) foi a relativa à das exportações de turismo, explicando metade da redução das exportações totais. O nível pré-crise das exportações de bens e serviços será alcançado no início do ano de 2023, justificado pela recuperação mais gradual do turismo e dos serviços que lhe estão relacionados (Banco de Portugal, 2020).

O turismo em Portugal no seu geral, até 2019, cresceu nos últimos anos. O ano de 2020, foi uma exceção, principalmente por conta do estado de pandemia à escala mundial que se vive atualmente. No entanto tendo em conta resultados de 2019, um ano tido como "normal" a nível de mercado, o turismo revela-se um sector em crescimento e com bastante procura a nível internacional.

2.2. Turismo Rural

O turismo no espaço rural tem várias características que o tornam único neste sector de atividade. O turismo no Espaço rural tem várias características que o tornam único neste sector de atividade. De acordo com a Direcção-Geral de Agricultura e Desenvolvimento, este deve ser situado em espaços com ligação tradicional e significativa à agricultura ou ambiente e paisagem de caráter vincado, tanto a nível arquitetónico como dimensão e materiais de construção. É um tipo de turismo que assenta na sustentabilidade, na tradição e no acolhimento personalizado de acordo com os costumes da região (DGADR, 2020a). Este compreende vários grupos de empreendimentos. Inclui as casas de campo,

caracterizadas pela sua localização em aldeias e construção tradicional. Inclui igualmente o turismo de aldeia, caracterizado por cinco ou mais casas de campo situadas na mesma aldeia ou freguesia, ou em aldeias ou freguesias contíguas, quando exploradas de uma forma integrada por uma única entidade. Compreende também o agroturismo, cujos imóveis são situados perto de explorações agrícolas e permitem aos hóspedes participar em atividades nos trabalhos aí desenvolvidos, numa partilha de novas experiências e conhecimento. Por fim, os hotéis rurais podem também ser classificados como turismo rural pela sua localização, arquitetura e materiais de construção utilizados de acordo com os costumes da zona onde são construídos (DGADR, 2020a).

2.3. Algoritmos utilizados em trabalhos na área do Turismo

A Tabela 1 apresenta o resumo dos algoritmos encontrados na revisão de literatura efetuada na área do Turismo.

Algoritmos	Referências
Naive Bayes	Sánchez-Franco et al. (2019)
Latent dirichlet analysis (LDA)	Guo et al. (2017); Taecharungroj & Mathayomchan (2019); Yi Luo et al. (2020); Luo et al. (2021)
Modelo de regressão linear	Pokryshevskaya & Antipov (2017)
Técnicas de *text-mining* e análise de sentimento com recurso ao software *Leximancer*	Cheng & Jin (2019)
Construção de classificadores para os comentários e posterior análise de sentimento dos mesmos	Tsai et al. (2020)
Tratamento estatístico em clusters	Eusébio et al. (2017)

Tabela 1 – Algoritmos utilizados em trabalhos na área do Turismo

Em Sánchez-Franco et al. (2019) é estudada a possibilidade da identificação de termos relacionados com a experiência dos hóspedes no sentido de serem utilizados para melhorar o seu serviço a nível de hospitalidade, aplicando o algoritmo de classificação de *Naive Bayes*. O algoritmo *Latent Dirichlet Allocation* (LDA) prova ser algo popular neste contexto de estudos, representado num estudo efetuado por Guo et al. (2017) no sentido de identificar as dimensões chave referentes à satisfação com o serviço hoteleiro com base em 266.544 comentários de utilizadores em 25.670 hotéis de 16 países retirados da plataforma TripAdvisor.com. O mesmo algoritmo é utilizado também por Taecharungroj & Mathayomchan, (2019) e Yi Luo et al. (2020). Já em 2021 é também utilizado o mesmo modelo numa versão melhorada em Luo et al. (2021), neste caso em algoritmos de *Support Vector Machine* (SVM) e *Importance-Performance Analysis* (IPA), no sentido de analisar comentários turísticos obtidos nas visitas a 24 geoparques que fazem parte do património mundial da *United Nations Educational, Scientific and Cultural Organization* (UNESCO) de forma a fornecer sugestões aos órgãos de gestão no sentido de compreender melhor a perceção dos visitantes ao visitar os parques e avaliar as condições dos mesmos. Em Pokryshevskaya & Antipov (2017) foram obtidos de um hotel situado no Dubai 3.630 comentários da plataforma Booking.com

e respetivos perfis/características dos utilizadores para análise, no sentido de tentar prever a satisfação dos hóspedes em relação aos serviços disponibilizados a nível de hotelaria utilizando um modelo de regressão linear para detetar quais as características mais relevantes. Cheng & Jin (2019) procuraram objetivos semelhantes aplicando dados recolhidos da plataforma Airbnb. Neste caso foram detetados 4 tópicos essenciais da análise aos comentários sendo estes; a localização, as comodidades, o anfitrião e as recomendações.

Tendo em conta a importância dos comentários dos utilizadores e o enorme volume de informação dispersa, Tsai et al. (2020) propõe uma abordagem de sumarização dos comentários no sentido de classificar os mesmos de forma a salientar os mais importantes e relevantes para avaliação. A amostra utilizada conteve 23.430 comentários de 23.038 utilizadores na plataforma TripAdvisor.com. Foram construídos vários classificadores, relacionados com a estrutura frásica do comentário e características do comentador, alguns desenvolvidos em estudos anteriores. O estudo conclui que a classificação antecipada dos comentários ao tratamento efetuado na análise de sentimento é vantajosa e mais eficaz, traduzindo-se em melhores sumários para nomeadamente 6 aspetos distintos dos hotéis: localização, qualidade do sono, quarto, serviço, valor e limpeza.

No contexto do turismo em espaço rural português, não há muitos estudos sobre este tema. Eusébio et al. (2017) estudou a questão de quem é o principal consumidor de turismo rural em Portugal, e, em particular, fez uma análise ao turismo rural doméstico neste país.

3. METODOLOGIA

A metodologia utilizada suportou-se na metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM). O seu modelo de referência é composto por 6 fases principais, respetivamente, o entendimento do negócio, o entendimento dos dados, a preparação dos dados, a modelação, a avaliação do modelo e a implementação do modelo (Chapman et al. 2000). A pertinência e atualidade da metodologia CRISP-DM na área da mineração de dados e do *business intelligence* continuam a ser evidentes (Belfo & Andreica, 2018; Huber et al. 2019). As fases consideradas neste estudo correspondem às primeiras cinco fases da metodologia CRISP-DM.

A fase de entendimento do negócio corresponde essencialmente à revisão de literatura efetuada e já apresentada anteriormente. A revisão de literatura para este trabalho consistiu numa combinação entre uma revisão sistemática e revisão integrativa. Foi realizada com o auxílio da base de dados *ScienceDirect* utilizando as combinações de palavras-chave "*tourism sentiment analysis*", "*rural tourism*" "*text mining*", "*data mining*", "*online reviews*" e "*visitor satisfaction*". O resultado das pesquisas deu origem a um conjunto de 606 artigos ordenados pela sua relevância. Sendo que a nível tecnológico é importante existirem fontes atuais sob pena da revisão de literatura se tornar obsoleta, foram privilegiadas publicações entre 2015 e 2021, o que reduziu o resultado da pesquisa para 337

artigos. Foi também adicionada outra restrição relativamente ao tipo de artigo a considerar, que neste caso foi "artigo de pesquisa", reduzindo a amostra de artigos para 280. Destes 280, foram selecionados 17 artigos com base na leitura do resumo e relevância para o assunto a estudar. Por fim, por considerar que a questão de pesquisa é mais ampla do que aquela que pode ser gerada por uma revisão sistemática, consideraram-se mais algumas referências.

Com o objetivo de obter informação sobre os principais fatores de interesse e satisfação por parte dos hóspedes num estabelecimento de turismo rural foram obtidos dados sob a forma de comentários de 400 alojamentos classificados como "alojamento de turismo rural" em Portugal. Optou-se por realizar a recolha dos dados após o término da época de verão, pois seria interessante analisar o comportamento e os comentários dos hóspedes no contexto especial da pandemia COVID-19 que se viveu a nível mundial e em particular em Portugal, e que por sua vez teve um efeito de alavanca do mercado nacional, pelas restrições existentes na entrada de turistas de outros países em Portugal e da saída de turistas portugueses para o estrangeiro.

Optou-se pela obtenção dos dados na plataforma Booking.com, principalmente pelas seguintes razões (Pokryshevskaya & Antipov, 2017): apenas um utilizador que agendou um alojamento pela plataforma está autorizado a comentar, o que torna os comentários reais e fidedignos; para manter o rating atualizado, a plataforma arquiva comentários com mais de 24 meses; para cada comentário, os pontos positivos e negativos estão separados, facilitando bastante a análise de sentimento e obtenção dos dicionários de dados;

Para o efeito foi construído um algoritmo de *web scraping* utilizando a linguagem de programação *Python* (PSF, 2021), tendo obtido um total de 14.976 avaliações feitas por visitantes, das quais 9.939 continham texto com aspetos positivos, enquanto as restantes deixam esse campo em branco. Para além dos comentários positivos, foram também obtidos os seguintes dados: nota (de 1 a 10) dada ao alojamento; nome do utilizador; país de origem; número de comentários submetidos pelo utilizador na plataforma Booking.com; data do comentário; motivo da viagem; tipo de viajante; tipologia; número de noites em que o utilizador ficou no alojamento avaliado.

Em seguida, foram analisados os dados recolhidos de forma a extrair a informação relevante para o trabalho a efetuar. Para a realização desta análise foi utilizada a linguagem de programação R. A escolha desta ferramenta advém do facto de ser, a par do Python, uma linguagem bastante utilizada nos últimos anos que contém algumas bibliotecas dedicadas à análise de dados, manipulação de *arrays*, matrizes e visualização gráfica mais dedicadas à análise estatística (Chan, 2020). Normalmente esta fase é dividida em duas etapas. Em primeiro lugar é necessário proceder ao pré-processamento dos dados. Para o efeito são normalmente usados os seguintes passos: 1) limpeza dos dados, a qual consiste em identificar e eliminar palavras inúteis para a análise ou com erros ortográficos, *stop words* que são palavras que não acrescentam informação ao texto e que podem ser

eliminadas sem alterar o significado da frase (Wilbur & Sirotkin, 1992), palavras que não são o objetivo do estudo e também palavras que aparecem com pouca frequência; 2) tokenização, a qual tem o objetivo de separar as palavras importantes a analisar em grupos mais pequenos de palavras, ou mesmo pequenas frases que se denominam por *tokens*. Esta técnica é particularmente útil na identificação de localizações específicas associadas a determinada região turística ou mesmo os sentimentos associados gerados por parte dos utilizadores como é dito em (Li et al. 2018).

Para perceber a frequência das palavras mais usadas nos comentários foram formadas nuvens de palavras (*wordclouds*). Uma nuvem de palavras é uma forma de apresentação visual de dados textuais, tipicamente utilizada de forma a encontrar as palavras-chave. A informação é apresentada de forma que, palavras com maior frequência apareçam mais proeminentemente, sendo uma forma de sumarizar grandes quantidades de dados (Ahuja & Shakeel, 2017).

Na etapa seguinte, procedeu-se à modelação por tópicos de forma a extrair as dimensões gerais dos comentários obtidos. Optou-se por utilizar o algoritmo LDA, um dos modelos mais comuns para este tipo de situação. O algoritmo LDA consiste na aplicação de um modelo de identificação por tópicos, que captura eficientemente dimensões num determinado contexto sem realizar assunções a nível gramatical de um texto ou linguagem, o que permite uma análise com uma intervenção humana mínima (Guo et al. 2017). Na sua essência, o LDA assume que todas as palavras têm a sua origem num conjunto de tópicos que podem estar presentes em todos os comentários disponíveis, cada um com a sua proporção de cada tópico. É um método bastante utilizado quando se trata um conjunto não estruturado de dados. O LDA torna-se bastante útil por ser um método não supervisionado que permite de forma eficiente o tratamento de uma grande quantidade de dados. São extraídas dimensões referentes à satisfação dos clientes e as palavras que compõem as respetivas dimensões tendo como base documentos (neste caso cada comentário) pré-processados. O termo dimensão é definido como um conjunto das palavras que formam determinado tópico. Existem três parâmetros muito importantes neste modelo que devem ser tidos em conta. Em primeiro lugar, o parâmetro k, indicador do número de tópicos a formar, em segundo lugar o parâmetro β (*beta*) que basicamente calcula a probabilidade de determinado termo estar associado a um tópico. Este parâmetro indica a distribuição de palavras pelos tópicos criados. Um beta alto indica que cada tópico contém uma maior quantidade de palavras do conjunto de dados; e o parâmetro γ (*gamma*) que devolve a percentagem de cada tópico associada a determinado documento.

4. ANÁLISE E DISCUSSÃO DE RESULTADOS

Realizando uma pré-análise aos dados obtidos, conseguiu-se retirar algumas informações interessantes acerca do perfil do utilizador que realiza os comentários. Em média os comentários positivos têm 9,08 (numa escala de 0 a 10 da plataforma *Booking*). 95,8% dos comentários

analisados são realizados por utilizadores de nacionalidade portuguesa, o que faz sentido, tendo em conta o clima de pandemia e dificuldades acrescidas de circulação entre países. Dos restantes 4,2%, os países mais representados são o Brasil, França, Suíça Espanha e Reino Unido. 53,5% dos comentários obtidos dizem respeito ao ano de 2020 e grande parte dos utilizadores (68,9%) tem no máximo 10 comentários submetidos no Booking.com. No que diz respeito ao tipo de cliente, 48% dos comentários positivos são de casais e 37,1% de famílias com filhos que na maioria dos casos (95,7%), realizam a sua estadia em lazer e escolhem quartos ou casas (71,2%) como tipologia.

Os resultados obtidos nas duas abordagens, tanto a nível da nuvem de palavras, como da aplicação do modelo LDA revelam alguma coerência. A nuvem de palavras com os termos mais utilizados nos comentários positivos pode ser observada na Figura 1.

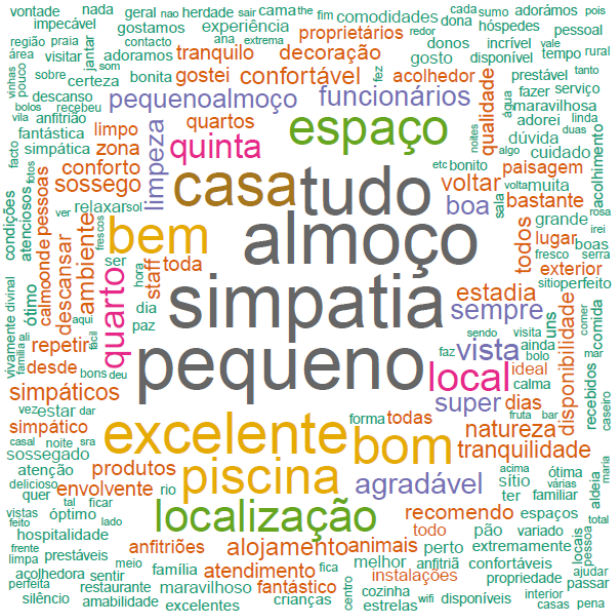

Figura 1 – Nuvem de palavras com os termos mais utilizados nos comentários positivos.

Na análise à nuvem de palavras, é dada uma grande relevância aos termos “pequeno-almoço” e “simpatia”, “tudo”, “localização, “espaço” e “piscina” entre outros. De uma forma bastante intuitiva obtemos evidência que esses termos são relevantes na avaliação geral do estabelecimento e que são referidos pelos hóspedes com maior frequência. É compreensível que estes termos sejam importantes pois são características particulares deste tipo de turismo, como já foi referenciado anteriormente. Existe também a referência aos vários tipos de estabelecimento nos termos “quinta” e “casa”, sendo casa dos dois o mais utilizado pela razão de que na maioria dos casos os estabelecimentos de turismo rural são casas de campo. Os termos “funcionários” e “disponibilidade” são também evidência de que o serviço do staff é um valor a ter em conta neste sector.

Com recurso ao modelo LDA foi possível identificar dez tópicos distintos que caracterizam os comentários positivos da amostra obtida, os tópicos podem ser visualizados na Tabela 2.

Número	Nome do tópico
1	Serviço Geral
2	Simpatia
3	Divisões da propriedade
4	Sentimento do cliente
5	Pequeno-almoço
6	Ambiente relaxante
7	Localização
8	Atividades
9	Anfitriões
10	Espaço exterior

Tabela 2 – Tópicos extraídos da amostra recolhida.

Os tópicos encontrados indicam de uma forma relativamente simples os tópicos mais importantes utilizando os termos em cada documento utilizado como amostra, como podemos verificar nos exemplos expostos na Figura 2.

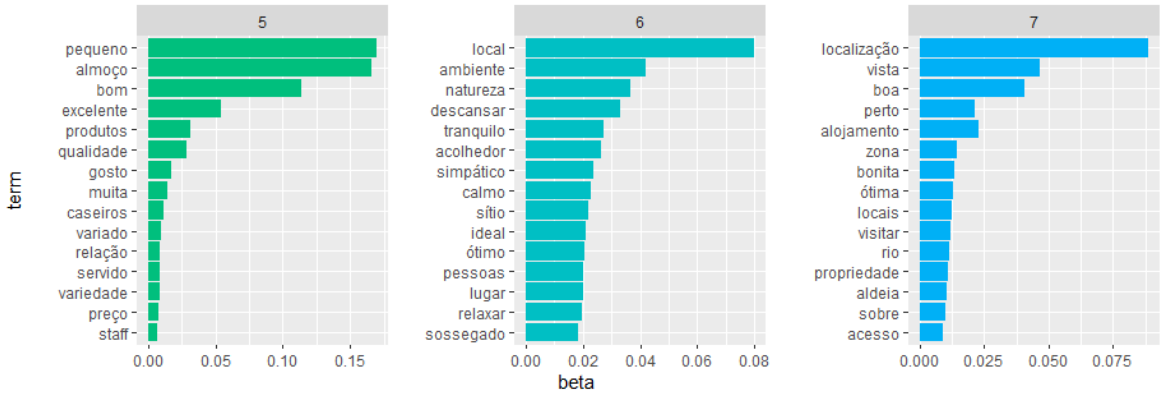

Figura 2 – Exemplo dos tópicos obtidos para pequeno-almoço, ambiente relaxante e localização.

Estes resultados esclarecem as áreas a investir ou melhorar neste tipo de estabelecimentos. Neste caso apenas foram tidas em conta as críticas positivas, mas o mesmo processo poderia ser aplicado para os comentários negativos se o objetivo fosse encontrar os pontos a melhorar. Uma particularidade interessante a ter em conta é o facto da grande maioria dos tópicos serem características controláveis pelo proprietário do estabelecimento. Características como o serviço, simpatia, divisões, pequeno-almoço, atividades, e o espaço exterior são algo que pode diretamente ser influenciado por uma decisão de administração e tendo e conta os resultados, é recomendável que o seja feito. O fator da localização e a promoção de um ambiente relaxante é parcialmente controlável, pois embora o proprietário tenha a liberdade para escolher o sítio onde construir um estabelecimento de turismo rural, estará sempre dependente da capacidade financeira e do espaço envolvente, eventuais construções em volta da propriedade que ponham em causa o impacto visual do estabelecimento e possibilidade de ruído que perturbe os hóspedes.

Com o volume de informação sob a forma de comentários a aumentar constantemente, torna-se útil não só identificar os tópicos mais relevantes contidos nos mesmos de um modo geral, mas também identificar qual o tópico ou tópicos mais relevantes em cada estabelecimento, ou mesmo, em um dos comentários. O LDA pressupõe que cada documento tem uma percentagem de um ou mais tópicos no seu conteúdo e o parâmetro γ (*gamma*) consegue obter essa informação. Na Tabela 3 podemos verificar os pesos relativos de cada um dos dez tópicos presentes para os primeiros dez documentos da amostra. Desta forma torna-se possível filtrar os comentários e dividir os mesmos pelos tópicos mais relevantes de forma a proceder a uma análise mais aprofundada e eventualmente mais útil e direcionada na tomada de decisão ou para avaliar o impacto decorrente de uma implementação de determinado ativo ou atividade num estabelecimento.

Comentários	Serviço Geral	Simpatia	Divisões da propriedade	Sentimento do cliente	Pequeno Almoço	Ambiente relaxante	Localização	Atividades	Anfitriões	Espaço Exterior
1	13.16%	10.53%	6.58%	6.58%	9.21%	10.52%	15.79%	6.58%	11.84%	9.21%
2	11.67%	11.67%	10.00%	10.00%	8.33%	8.33%	8.33%	11.67%	11.67%	8.33%
3	13.43%	8.96%	8.96%	11.94%	7.46%	10.45%	8.96%	10.45%	10.45%	8.96%
4	8.82%	14.71%	7.35%	16.18%	8.82%	7.35%	10.29%	11.77%	7.35%	7.35%
5	11.11%	9.26%	9.26%	11.11%	9.26%	12.96%	9.26%	9.26%	9.26%	9.26%
6	9.61%	11.54%	9.62%	11.54%	9.61%	9.62%	9.62%	9.62%	9.61%	9.62%
7	6.80%	17.48%	11.65%	8.74%	17.48%	8.73%	6.80%	8.74%	5.83%	7.77%
8	15.87%	11.11%	11.11%	12.70%	7.93%	7.94%	7.94%	7.94%	7.94%	9.52%
9	8.92%	14.29%	10.71%	8.93%	10.71%	8.93%	8.93%	10.71%	8.93%	8.93%
10	11.11%	12.96%	9.26%	11.11%	9.26%	9.26%	9.26%	9.26%	9.26%	9.26%

Tabela 3 – Percentagem de cada tópico em cada um dos primeiros dez comentários.

Vejamos em seguida três exemplos de comentários em particular e em que medida o algoritmo LDA quantifica a importância dos principais tópicos encontrados.

O comentário 1 consiste no seguinte texto: *"Localização muito agradável junto à nascente do rio Liz. As instalações estão bem arranjadas, recuperadas com gosto e mantendo todos os pormenores da traça antiga do edifício. O staff foi muito prestável e simpático. Ambiente acolhedor e comida muito boa. Aconselho vivamente"*. Neste comentário podemos verificar que o tópico localização é o mais relevante e presente tendo em conta os termos existentes (15,79%), sendo que os tópicos serviço geral (13,16%), ambiente relaxante (10,53%) e anfitriões (11,84%) também têm expressão.

O comentário 4 tem o texto: *"É a segunda vez que que ficámos hospedados e assim como da primeira vez adorámos. Desde a simpatia e acolhimento, a decoração, o conforto... ah, e o jantar,*

divinal. Iremos voltar de certeza". No comentário 4 é dado destaque à simpatia do pessoal (14,71%) e também o agrado do cliente face à experiência da estadia (16,18%).

Por fim, o comentário 8 consiste na seguinte redação: *"Adorei tudo: o alojamento, as refeições, espaço e a simpatia e amabilidade de todos os funcionários. As refeições excelentes. A repetir sem dúvida"*. Neste comentário é feita uma referência positiva ao serviço na sua generalidade (15,87%) e ao sentimento gerado ao usufruir da estadia (12,70%).

5. CONCLUSÕES E TRABALHOS FUTUROS

Os tópicos identificados são coerentes com a revisão de literatura realizada. Existe realmente uma reação positiva por parte do cliente quando existe qualidade no serviço oferecido pelo estabelecimento de um modo geral. A localização é um dos tópicos primordiais neste tipo de turismo, o que está alinhado com a procura da experiência mais relaxante e personalizada que este tipo de serviço pode oferecer, sendo este um dos fatores a ter em consideração aquando da definição do conceito, do planeamento e implementação de uma unidade de turismo rural.

O tópico com mais comentários positivos está relacionado com o pequeno-almoço, sendo que neste caso é valorizado o cuidado em colocar à disposição uma oferta variada de produtos típicos da região e de qualidade. A simpatia do pessoal é também um fator importante a ter em conta e um dos pontos particularmente importantes quando o estabelecimento em causa é uma casa de campo. É particularmente importante a simpatia do anfitrião, já que esta provoca no cliente um sentimento de familiaridade com o local e fá-lo sentir-se em casa. Estes fatores que se destacam nestes comentários por parte dos utilizadores estão em linha com outros estudos anteriores (Cheng & Jin, 2019).

São ainda considerados nos comentários positivos dos utilizadores o espaço exterior, pela envolvência rural, jardins, piscinas e ambiente convidativo para toda a família e agradável para crianças. As divisões do alojamento, consoante seja uma casa de campo ou um quarto num empreendimento são também pontos muito importantes e de alguma forma consensuais com este setor turístico particular. No caso do turismo rural, parece existir evidência de que, o cuidado em manter as divisões e a construção dos alojamentos num formato tradicional e com decorações típicas da região em que se encontram inseridas, num contexto de conforto e relaxamento, influencia positivamente o cliente e leva-o a manifestar essa satisfação.

Como trabalhos futuros, seria interessante comparar os resultados com outros estudos semelhantes a nível internacional e avaliar com maior pormenor se o perfil dos clientes vai ser alterado num futuro pós COVID-19. Sugere-se ainda o estudo da possível influência que temas atuais e universais, como as alterações climáticas e uma maior sensibilidade e preocupação ambiental, podem de alguma forma ter na escolha por este tipo de turismo.

REFERÊNCIAS

Ahuja, V., & Shakeel, M. (2017). Twitter Presence of Jet Airways-Deriving Customer Insights Using Netnography and Wordclouds. *Procedia Computer Science*, *122*, 17–24. https://doi.org/10.1016/j.procs.2017.11.336

Banco de Portugal. (2020). *Boletim Económico Dez. 2020*.

Belfo, F. P., & Andreica, A. B. (2018). A Comprehensive Methodology to Implement Business Intelligence and Analytics Through Knowledge Discovery in Databases. *Mining Intelligence and Knowledge Exploration. MIKE 2018. Lecture Notes in Computer Science*, *11308*, 102–111. https://doi.org/https://doi.org/10.1007/978-3-030-05918-7_10

Brandão, L., Belfo, F. P., & Silva, A. (2021). Wavelet-based cancer drug recommender system. *Procedia Computer Science, Communications in Computer and Information Science*, *181*, 487–494. https://doi.org/https://doi.org/10.1016/j.procs.2021.01.194

Chan, B. K. C. (2020). Data analysis using R programming. *Simultaneous Mass Transfer and Chemical Reactions in Engineering Science*, 39–60. https://doi.org/10.1016/b978-0-12-819192-7.00002-3

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-By-Step Data Mining Guide*. SPSS, CRISP-DM Consortium.

Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, *76*(May 2018), 58–70. https://doi.org/10.1016/j.ijhm.2018.04.004

Cohen, S. A., Prayag, G., & Moital, M. (2014). Consumer behaviour in tourism: Concepts, influences and opportunities. *Current Issues in Tourism*, *17*(10), 872–909. https://doi.org/10.1080/13683500.2013.850064

Cui, G., Wong, M. L., & Lui, H.-K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science*, *52*(4), 597–612.

DGADR. (2020a). *Características do Turismo no Espaço Rural*. Direção Geral de Agricultura e Desenvolvimento Rural. https://www.dgadr.gov.pt/diversificacao/turismo-rural/caracteristicas-do-turismo-no-espaco-rural

DGADR. (2020b). *O Interesse pelo Turismo no Espaço Rural*. Direção Geral de Agricultura e Desenvolvimento Rural. https://www.dgadr.gov.pt/diversificacao/turismo-rural/o-interesse-pelo-turismo-no-espaco-rural

Eusébio, C., Carneiro, M. J., Kastenholz, E., Figueiredo, E., & Soares da Silva, D. (2017). Who is consuming the countryside? An activity-based segmentation analysis of the domestic rural tourism market in Portugal. *Journal of Hospitality and Tourism Management*, *31*, 197–210. https://doi.org/10.1016/j.jhtm.2016.12.006

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467–483. https://doi.org/10.1016/j.tourman.2016.09.009

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. *Procedia Cirp*, *79*, 403–408. https://doi.org/10.1016/j.procir.2019.02.106

INE. (2018). *Estatísticas do Turismo 2017*. Instituto Nacional de Estatística, I.P.

INE. (2020). *Estatísticas do Turismo 2019*. Instituto Nacional de Estatística, I.P.

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, *68*, 301–323. https://doi.org/10.1016/j.tourman.2018.03.009

Loureiro, A., Lourenço, J., Costa, E., & Belfo, F. (2014). Indução de Árvores de Decisão na Descoberta de Conhecimento: Caso de Empresa de Organização de Eventos. In *VI Congresso Internacional de Casos Docentes em Marketing Público e Não Lucrativo*.

Luo, Yi, Tang, L., Kim, E., & Wang, X. (2020). Finding the reviews on yelp that actually matter to me: Innovative approach of improving recommender systems. *International Journal of Hospitality Management*, *91*(November 2019), 102697. https://doi.org/10.1016/j.ijhm.2020.102697

Luo, Yuyan, He, J., Mou, Y., Wang, J., & Liu, T. (2021). Exploring China's 5A global geoparks

through online tourism reviews: A mining model based on machine learning approach. *Tourism Management Perspectives*, *37*(December 2019), 100769. https://doi.org/10.1016/j.tmp.2020.100769

OCDE. (2020). *OECD Tourism Trends and Policies 2020*. 1–16.

Park, D. B., & Yoon, Y. S. (2011). Developing sustainable rural tourism evaluation indicators. *International Journal of Tourism Research*, *13*(5), 401–415. https://doi.org/10.1002/jtr.804

Pimenta, C., Ribeiro, R., Sá, V., & Belfo, F. P. (2018). Fatores que Influenciam o Sucesso Escolar das Licenciaturas numa Instituição de Ensino Superior Portuguesa. In *Atas da 18ª Conferência da Associação Portuguesa de Sistemas de Informação (CAPSI 2018) Associação Portuguesa de Sistemas de Informação*. Associação Portuguesa de Sistemas de Informação.

Pimenta, P., Belfo, F., & Trigo, A. (2009). Study the impact of Booking. com user scores and reviews in hotel management. *Book of Abstracts of the CENTERIS 2011–Conference on Enterprise Information Systems*, *30*, 8.

Pokryshevskaya, E. B., & Antipov, E. A. (2017). Profiling satisfied and dissatisfied hotel visitors using publicly available data from a booking platform. *International Journal of Hospitality Management*, *67*, 1–10. https://doi.org/10.1016/j.ijhm.2017.07.009

PSF. (2021). *Python*. Python Software Foundation.

Sánchez-Franco, M. J., Navarro-García, A., & Rondán-Cataluña, F. J. (2019). A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services. *Journal of Business Research*, *101*(June), 499–506. https://doi.org/10.1016/j.jbusres.2018.12.051

Seiça, A., Trigo, A., & Belfo, F. P. (2019). LexiNB - Uma Abordagem Bietápica de Classificação de Sentimentos em Tweets Relacionados com as Autoridades Fiscais Portuguesas. *Proceedings of the 19.ª Conferência Da Associação Portuguesa de Sistemas de Informação (CAPSI'2019) Held in Lisboa, Portugal, 11-12 October 2019. Paper 5.*, *October*, 11–12.

Sereday, S., & Cui, J. (2017). Using machine learning to predict future tv ratings. *Data Science, Nielsen*, *1*(3), 3–12.

Taecharungroj, V., & Mathayomchan, B. (2019). Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand. *Tourism Management*, *75*(July), 550–568. https://doi.org/10.1016/j.tourman.2019.06.020

Tsai, C. F., Chen, K., Hu, Y. H., & Chen, W. K. (2020). Improving text summarization of online hotel reviews with review helpfulness and sentiment. *Tourism Management*, *80*(February 2019), 104122. https://doi.org/10.1016/j.tourman.2020.104122

Turismo rural: segmento em mudança - Publituris - Publituris. (n.d.). Retrieved August 11, 2020, from https://www.publituris.pt/2019/08/08/turismo-rural-segmento-em-mudanca/

Vu, H. Q., Li, G., Law, R., & Zhang, Y. (2018). Tourist Activity Analysis by Leveraging Mobile Social Media Data. *Journal of Travel Research*, *57*(7), 883–898. https://doi.org/10.1177/0047287517722232

Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science*, *18*(1), 45–55. https://doi.org/10.1177/016555159201800106

Yi, X., Lin, V. S., Jin, W., & Luo, Q. (2017). The Authenticity of Heritage Sites, Tourists' Quest for Existential Authenticity, and Destination Loyalty. *Journal of Travel Research*, *56*(8), 1032–1048. https://doi.org/10.1177/0047287516675061