Instituto Politécnico de Coimbra
Instituto Superior de Contabilidade
e Administração de Coimbra

Rui Filipe Martins Esteves

Identificação de fatores que levam à escolha da realização de turismo rural em Portugal utilizando técnicas de Text Mining na plataforma Booking.com

Coimbra, outubro de 2021

Identificação de fatores que levam à escolha da realização de turismo rural em Portugal
Rui Filipe Martins Esteves
ISCAC | 2021

Instituto Politécnico de Coimbra
Instituto Superior de Contabilidade
e Administração de Coimbra

Rui Filipe Martins Esteves

Identificação de fatores que levam à escolha da realização de turismo rural em Portugal utilizando técnicas de Text Mining na plataforma Booking.com

Dissertação submetida ao Instituto Superior de Contabilidade e Administração de Coimbra para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Análise de Dados e Sistemas de Apoio à Decisão, realizada sob a orientação do Professor Fernando Paulo Belfo e do Professor António Trigo.

Coimbra, outubro de 2021

TERMO DE RESPONSABILIDADE

Declaro ser o autor desta dissertação, que constitui um trabalho original e inédito, que nunca foi submetido a outra Instituição de ensino superior para obtenção de um grau académico ou outra habilitação. Atesto ainda que todas as citações estão devidamente identificadas e que tenho consciência de que o plágio constitui uma grave falta de ética, que poderá resultar na anulação da presente dissertação.

PENSAMENTO

"A vida é a soma de todas as tuas escolhas."

Albert Camus

DEDICATÓRIA

À Joana, que me acompanha há quase metade destes 36 anos.

Ao Miguel, para que ele saiba que tudo é possível com algum trabalho e dedicação.

Nunca deixes de sonhar!

AGRADECIMENTOS

A realização desta dissertação de Mestrado contou com o apoio de várias pessoas, sem o qual não teria sido possível de concluir. Assim deixo aqui os meus agradecimentos.

Aos meus orientadores, Professor Fernando Paulo Belfo e Professor António Trigo, pela disponibilidade que manifestaram quando propus este tema.

Pelo seu acompanhamento constante, assim como as várias sugestões e críticas construtivas que tiveram durante todo o processo.

A todos os meus bons amigos que nunca me deixaram mal e pelos quais tenho uma enorme estima. Obrigado por todos os bons momentos que já tivemos e iremos seguramente continuar a ter.

Aos meus pais por todo o apoio incondicional e paciência que tiveram ao longo da minha vida.

Por último, quero agradecer à minha esposa Joana Lopes por todo o apoio que me deu desde que decidi melhorar o meu percurso académico e abraçar este desafio.

És acima de tudo a minha melhor amiga e a minha fonte de motivação. És a razão pela qual tento sempre ser melhor e dar o melhor de mim. Obrigado pela tua paciência e carinho durante todos estes anos.

Muito obrigado!

RESUMO

Este estudo pretende descobrir e analisar os fatores mais valorizados no turismo rural, tendo em conta os aspetos positivos e negativos inerentes à estadia em estabelecimentos deste sector turístico. Apresenta-se uma análise de uma amostra com 14.976 comentários obtidos da plataforma de reservas Booking.com sobre estadias em estabelecimentos de turismo rural localizados em Portugal entre 2018 e 2021. 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 *Latent Dirichlet Allocation* 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 quando falamos em aspetos positivos. Em termos de aspetos negativos, o pequeno-almoço, limpeza da casa de banho, dificuldades de acesso, rede e *WiFi* foram os mais relevantes para o estudo. Concluiu-se também que os tópicos encontrados dizem, em grande parte, respeito a fatores controláveis pelo proprietário do estabelecimento.

Foi também analisada a coerência, prevalência e relação entre tópicos para os fatores mais valorizados positivamente pelos utilizadores. Estes resultados têm implicação para a gestão das unidades de turismo rural pois identificam os fatores a ter em atenção pelos responsáveis destas unidades.

Palavras-chave: mineração de texto, análise de sentimento, modelação por tópicos, Latent Dirichlet Allocation, Booking.com, turismo rural

ABSTRACT

This study aims to discover and analyze the most valued factors in rural tourism, considering the positive and negative aspects inherent to staying in establishments in this tourism sector. An analysis of a sample with 14,976 comments obtained from the Booking.com reservation platform about stays in rural tourism establishments located in Portugal between 2018 and 2021 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 Latent Dirichlet Allocation algorithm allowed to extract ten topics coherent with the word clouds. The results indicate that breakfast, the friendliness of the staff, the hosts and the outdoor space are the most relevant factors for guests when we talk about positive aspects. In terms of negative aspects, breakfast, bathroom cleaning, access difficulties, network and WiFi were the most relevant for the study. It was also concluded that the topics found are largely related to factors that can be controlled by the owner of the establishment.

Coherence, Prevalence, and relationship between topics were also analyzed for the factors most positively valued by users. These results have implications for the management of rural tourism units as it identifies the factors to be considered by those responsible for these units.

Keywords: text mining, sentiment analysis, topic modelling, Latent Dirichlet Allocation, Booking.com, rural tourism

ÍNDICE GERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE ABREVIATURAS, ACRÓNIMOS E SIGLAS

API ..... Application Programming Interface

CRISP-DM.... Cross-Industry Standard Process for Data Mining

CSV..... Comma-Separated Values

DCBD ..... Descoberta de conhecimento em bases de dados

DGADR ..... Direção-Geral de Agricultura e Desenvolvimento Rural

DTM ..... Document-Term Matrix

GPS ..... Global Positioning System

HTML ..... HyperText Markup Language

INE..... Instituto Nacional de Estatística

IPA ..... Importance-Performance Analysis

KNN..... K-nearest neighbours

LARA..... Latent Aspect Rating Analysis

LDA ..... Latent Dirichlet Allocation

LSTMM ..... Long Short-Term Memory Model

OCDE..... Organização para a Cooperação e Desenvolvimento Económico

PIB ..... Produto Interno Bruto

PLN..... Processamento de Língua Natural

SVM..... Support Vector Machine

UNESCO ..... United Nations Educational, Scientific and Cultural Organization

XML..... Extensible Markup Language

1 INTRODUÇÃO

Nos últimos anos temos assistido a uma grande inovação tecnológica no que diz respeito ao desenvolvimento da Internet, aliada aos dispositivos móveis e próprias relações sociais, dando um novo contexto ao conceito de afirmação do ser humano. O turismo tem um lugar importante neste prisma. 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.

Cada vez mais, essa valorização e a validação pessoal tem muitas vezes como pano de fundo a partilha das mesmas nas redes sociais. Nesse aspeto a tecnologia facilitou bastante o processo, sendo este, neste momento imediato e, principalmente, suscetível a obter comentários muito rapidamente por parte de amigos e conhecidos.

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. Neste trabalho irei focar a minha atenção no Turismo Rural, a nível da sua oferta de serviços de alojamento. Esta área comercial junta um pouco os dois conceitos que já referidos: a procura de uma nova experiência e os comentários que se obtêm, que se podem e devem analisar de forma a extrair informação muitíssimo útil para a tomada de decisão.

Em Kastenholz et al. (2014) é referida a ideia de que um visitante ao viajar, procura em primeiro lugar uma experiência fora do seu ambiente habitual de residência, sendo esta ideia uma base para várias teorias na área da sociologia. No caso em particular do turismo rural é transmitida a ideia de um ambiente quase idílico a nível paisagístico que permite ao visitante uma experiência que não encontra num ambiente mais urbano e que, portanto, se traduz num escape ao seu quotidiano. O mesmo estudo vai mais longe e refere ainda a experiência turística numa procura de repouso e a descoberta da diferença, num ambiente rural, mas relativamente controlado, criando assim um paradoxo que indiretamente urbaniza os espaços rurais, ao integrar os mesmos cada vez mais a nível turístico.

Existem várias formas de entender e sintetizar os comentários dos utilizadores de forma a retirar conclusões válidas para a gestão. Neste trabalho serão utilizadas várias técnicas de *text mining* para o fazer, técnicas essas que serão enumeradas mais abaixo.

O turismo rural não é um fenómeno acidental ou temporário, mas principalmente uma evolução do modelo de sociedade em que vivemos. Os indicadores apontam para um crescimento regular da procura desta atividade 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).

Em 2017, segundo o Instituto Nacional de Estatística (INE), a oferta de turismo rural no Portugal cresceu 18,8% traduzindo-se em 1,7 milhões de dormidas (que revelam um crescimento de 17% comparativamente com o período homólogo) e que geraram 94,7 milhões de € de proveitos totais (TP, 2017). No futuro, a tendência deverá manter-se, existindo neste momento uma clara aposta na qualidade deste serviço. As unidades de alojamento estão cada vez mais adequadas e oferecem um vasto leque de serviços. Os produtos regionais têm muita qualidade e o cuidado de fazer as coisas simples bem, dão um caráter especial a este sector, procurado muitas vezes pelo descanso e a tranquilidade que pode transmitir.

É referido num artigo jornalístico em *Publituris* (2019) que o turismo Rural parece atravessar um período dourado, mas, ainda assim, existem vários desafios que se atravessam no caminho deste segmento, estando estes centrados no reforço da comunicação, facilitação da informação sobre a oferta disponível e a diferenciação da oferta, tanto em termos de serviços como de entretenimento.

Outro dos obstáculos encontrados é a falta de recursos humanos qualificados, pois o turismo é um setor que exige muito conhecimento de marketing, design e gestão, mas mais do que isso, ousadia e criatividade para se poder demarcar dos demais segmentos de mercado turístico. No entanto, o facto de grande parte dos alojamentos se situarem maioritariamente em zonas rurais, não torna atrativo a muitos jovens a perspetiva de enveredar por uma carreira ou plano de negócio nesta área.

O que distingue as unidades hoteleiras de turismo rural das restantes reside no facto da quase totalidade das unidades de turismo rural serem projetos individuais. Não existem muitas unidades nesta categoria que pertençam a grupos hoteleiros, e são, na sua grande maioria, projetos de pessoas para pessoas, mais marcados pela personalidade dos proprietários, pela sua experiência, o que torna cada unidade um lugar único e com personalidade própria conectado com a natureza e atividades rurais.

Partindo dos factos acima descritos, surge a minha motivação para a realização deste estudo, que passa por construir uma metodologia com o intuito de fornecer informação sobre os fatores e características mais importantes para o Turismo Rural e apresentar a mesma de uma forma fácil de interpretar e intuitiva a nível visual a órgãos de gestão responsáveis pela exploração de empreendimentos deste sector no nosso país.

Procura-se, nos últimos anos, 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 pelos hóspedes são cruciais para entender as preferências dos mesmos. O comportamento dos consumidores tem sido nos últimos anos um dos campos mais pesquisados na área do marketing e do turismo. Em Cohen et al. (2014) são enumerados algumas dimensões essenciais a ter em conta ao analisarmos o comportamento destes consumidores, sendo estas a tomada de decisão, os seus valores, as suas motivações, personalidade, espectativas, atitudes, perceções, satisfação, confiança e lealdade.

A análise de dados é um tema bastante atual e que oferece várias possibilidades para retirar informação pertinente dos comentários obtidos dos hóspedes, sendo estas possibilidades muito apoiadas nas emoções humanas. Tal informação é bastante pertinente a nível de criação de valor e tem um papel crucial no desenvolvimento da oferta do serviço e a antecipação do mesmo de modo a ir de encontro às expectativas do cliente (Vu et al., 2018). As redes sociais têm uma grande importância nesta matéria pois contêm vastas quantidades de informação que pode ser analisada no contexto do comportamento turístico. Diversas fontes como o *Twitter* ou *Facebook* tornaram-se fontes de dados que são analisados de forma possibilitar uma maior perceção do comportamento do potencial turista/cliente (Vu et al., 2018).

O objetivo deste estudo é descobrir e analisar os fatores mais valorizados no turismo rural em Portugal, tanto na perspetiva positiva como negativa. Para ser possível aferir sobre este assunto será realizada uma recolha de dados para posterior análise das avaliações obtidas da plataforma de reservas *Booking.com* sobre estadias em estabelecimentos de turismo rural localizados em Portugal.

Serão utilizadas técnicas de mineração de texto e modelação por tópicos que permitem uma análise rápida e intuitiva aos dados e que permitem, em conformidade com a motivação expressa acima atingir os objetivos propostos.

A estrutura deste trabalho está dividida em várias fases sendo que, em primeiro lugar, é realizada uma introdução ao tema, seguida da revisão de literatura que incide sobre o sector do Turismo, em particular no Turismo Rural em Portugal e também na pesquisa de trabalhos efetuados dentro da mesma temática, referindo a metodologia e algoritmos utilizados e conclusões do estudo.

Segue-se um capítulo sobre a metodologia a aplicar ao estudo a realizar, de modo a atingir os objetivos propostos, onde são destacados os processos de recolha de dados, preparação dos dados, análise volumétrica da amostra e os modelos a aplicar.

Existe um capítulo dedicado aos resultados obtidos e também um capítulo onde é efetuada uma discussão dos mesmos, seguido das conclusões a retirar do estudo.

2 REVISÃO DE LITERATURA

2.1 Sector do Turismo

O sector do turismo tem tido, nas últimas 6 décadas, um crescimento constante e consistente, sendo um dos principais motores económicos a nível mundial. Ao gerar rendimento, em grande parte vindo 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 é uma importante fatia na exportação nos países envolvidos (OCDE, 2020).

A OCDE (Organização para a Cooperação e Desenvolvimento Económico), da qual Portugal é membro desde 1961, consiste num fórum de contacto entre vários países membros que trata de temas ligados a políticas de desenvolvimento económico e bem-estar social das pessoas a nível mundial. A organização formou também um comité ligado ao turismo em 1948 e nos últimos anos tem vido a realizar relatórios de monitorização do sector.

De acordo com os dados obtidos da OCDE (2020), referentes ao ano de 2018 o turismo contribuiu diretamente para o PIB (Produto Interno Bruto) em média 4.4%, 6.9% a nível de criação de emprego, conforme podemos verificar na Figura 2.1. 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.

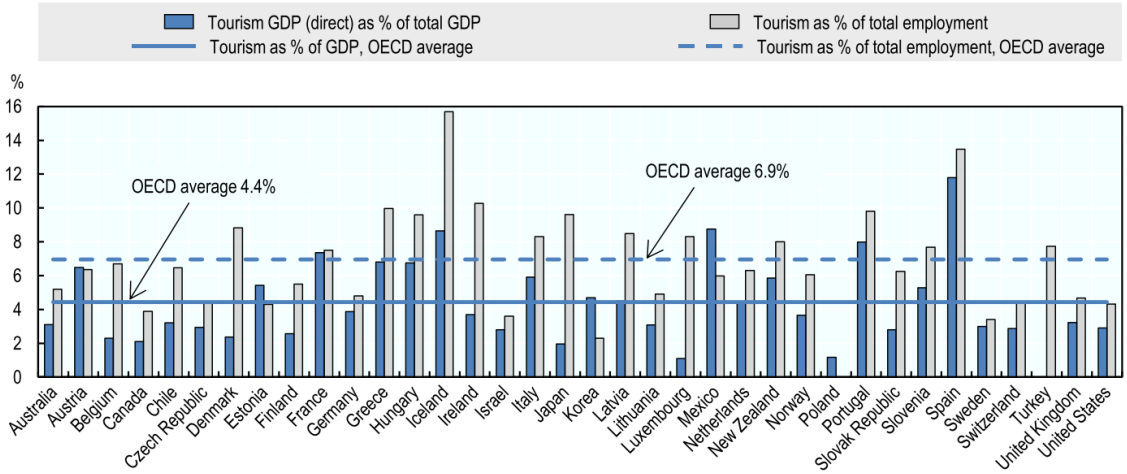


Figura 2.1 – Contributo do turismo no PIB em 2018 nos países membros da OCDE

Fonte: OCDE (2020)

Estes dados naturalmente não refletem ainda o clima de pandemia à escala mundial por COVID-19 que se viveu em 2020 e que ainda está presente em 2021. Este tema será abordado um pouco mais abaixo neste trabalho.

À medida que a tecnologia evolui, e o volume de dados em plataformas *online* aumenta, aumenta também a necessidade do próprio sector turístico evoluir e acompanhar as novas tendências tecnológicas. A OCDE refere a criação de medidas que preparam um futuro digital ao nível da criação de plataformas de turismo digital, apoiadas por ferramentas de *business intelligence*, com base na adoção de análise de dados de forma a arranjar soluções para os novos desafios que o sector seguramente terá no futuro, especialmente um futuro pós-pandémico como o que vamos ter.

2.2 Turismo em Portugal

O setor do turismo no nosso país representa um peso importante na atividade económica em Portugal. Em 2019, contou com números expressivos na ordem dos 52,3% do total a nível de exportações de serviços e 19,7% das exportações totais. Tem também um contributo importante ao nível do Produto Interno Bruto (PIB) nacional, com uma fatia de 8,7% no período (TP, 2020).

Figura 2.2 – Principais indicadores do turismo português em 2020

Fonte: TravelBI (2020)

O turismo em Portugal no seu geral, até 2019, teve um crescimento nos últimos anos (TravelBI, 2020). O ano de 2020 e talvez 2021 são uma exceção, principalmente por conta do estado de pandemia à escala mundial que se vive atualmente e que se traduziu numa diminuição em cerca de 61,3% (ver Figura 2.2) do número de hóspedes no país. Esta situação será referida em pormenor mais abaixo no presente trabalho. 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.

Com base no relatório do Instituto Nacional de Estatística (INE, 2020), referente ao ano de 2019, o número estimado de chegadas a Portugal de turistas não residentes atingiu os 24,6%, o que se traduziu num aumento de 7,9% face ao período homólogo. A maior parte das dormidas (90,2%) está situada em alojamentos de hotelaria, alojamento local e turismo de espaço rural, sendo que o Reino Unido foi o país com mais dormidas de turistas internacionais com uma substancial quota de 18,8% seguido da Alemanha com 12,3% e Espanha com 11,0%.

Em termos económicos o Turismo em Portugal traduziu-se em proveitos de 4,3 mil milhões de euros em 2019, revelando um crescimento de 7,8% face ao período homólogo.

Na sua grande maioria, as principais razões para a realização de viagens em Portugal foram "lazer, recreio ou férias", seguindo-se de "visitas a familiares" e "viagens de negócios".

Em termos comparativos com as restantes regiões no planeta, Portugal beneficia da tendência mundial para escolher a europa como destino de férias, conforme se pode visualizar na Figura 2.3.

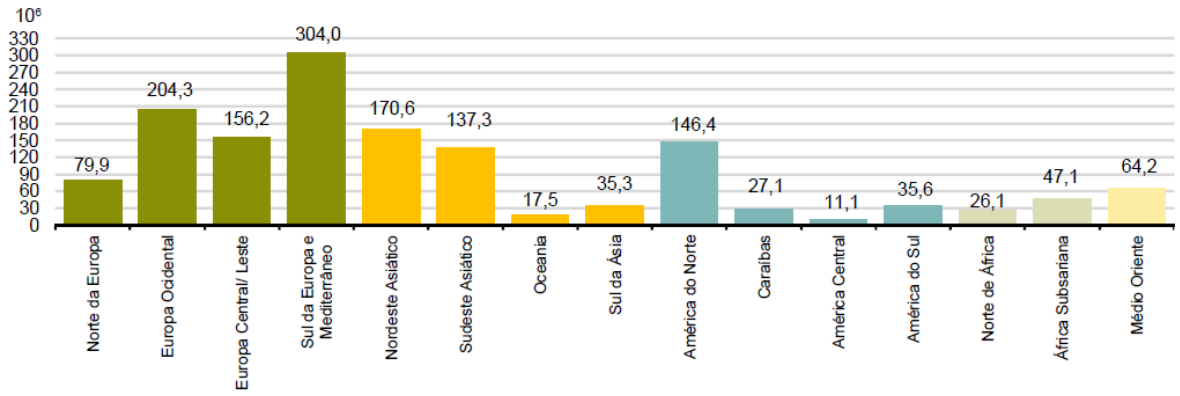


Figura 2.3 – Rácio entradas/saídas turísticas

Fonte: relatório INE (2020)

Já em 2016 o turismo em Portugal teve uma enorme expressão com influência direta e indireta em vários sectores de mercado. Conseguiu alargar a atividade nas chamadas "épocas baixas", e criou empregos na área, fomentando a economia nacional e dinamizando o próprio mercado interno. Ao estar exposto ao mercado internacional, promoveu uma diversificação do mesmo com crescimento expressivo no mercado americano, polaco e brasileiro. Este crescimento foi reconhecido internacionalmente sob a forma de vários prémios internacionais ganhos pelo sector (Economia, 2017).

Também em Economia (2017) são enumerados os várias ativos estratégicos que o turismo abraça para o futuro do sector. Pessoas, o clima e luz, vasta história e cultura nacional, natureza, água, a gastronomia e vinhos, os vários eventos artístico-culturais e desportivos que acontecem durante o ano, um bem-estar e uma qualidade de vida acima da média são fatores base para a projeção de um futuro para o turismo em Portugal, numa fase pós-pandémica. Vários dos aspetos enumerados enquadram-se perfeitamente no contexto do turismo rural, antevendo assim algum otimismo no desenvolvimento deste sector mais especializado.

2.3 Turismo Rural

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, conforme podemos verificar na Tabela 2.1 abaixo.

Tabela 2.1 – Tipos de Estabelecimento de Turismo Rural

Tipos de Estabelecimento	Características
Casa de Campo	Localização em aldeias e construção tradicional
Turismo de Aldeia	Cinco ou mais casas de campo situadas na mesma aldeia ou freguesia, ou em aldeias ou freguesias contíguas, sejam exploradas de uma forma integrada por uma única entidade
Agroturismo	Explorações agrícolas que permitem aos hóspedes participar em atividades nos trabalhos aí desenvolvidos, numa partilha de novas experiências e conhecimento
Hotel Rural	Localização, arquitetura e materiais de construção utilizados de acordo com os costumes da zona onde são construídos

Assim, temos casas de campo, caracterizadas pela sua localização em aldeias e construção tradicional, turismo de aldeia, quando cinco ou mais casas de campo situadas na mesma aldeia ou freguesia, ou em aldeias ou freguesias contíguas, sejam 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 também podem 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).

Para usufruir deste tipo de turismo, a plataforma *Booking.com* tem uma grande importância, trazendo uma maior facilidade no processo de reservas, assim como uma maior visibilidade a este sector, ficando assim, pelo menos teoricamente em pé de igualdade com os restantes sectores turísticos. Nos últimos anos, o uso desta plataforma cresceu exponencialmente desde a sua criação em 1996, reportando em 2018 mais de 27 milhões de estabelecimentos que cobrem mais de 130.000 destinos distribuídos por 227 países (Booking, 2020). Com estes recursos é fácil perceber a enorme influência que o *Booking.com* tem atualmente, e, nomeadamente no caso do turismo rural é bastante importante pois este tem a tendência para estar associado a pequenos estabelecimentos, muitas vezes com pouca expressão e experiência em marketing online, que se traduz em sites com design antiquado e limitado ao nível de experiência de utilizador, muitas vezes com fragilidades no que diz respeito a reservas e/ou pagamentos. Estas condicionantes sendo assumidas pela plataforma, libertam os funcionários deste tipo de responsabilidade (Gössling & Lane, 2015).

No entanto, existe também um efeito adverso que pode eventualmente tornar-se num obstáculo neste contexto. Por um lado, temos um mundo de facilidades provocadas pelo fácil acesso e experiência de utilizador intuitiva. Por outro lado, uma fraca sensibilidade face à maior proximidade necessária a ter com o cliente, os *updates* constantes de perfil, transparência de preços e condições de alojamento pode prejudicar de forma substancial a imagem do estabelecimento, obrigando por vezes à contratação de recursos humanos mais qualificados, que podem não estar ao alcance dos proprietários (Gössling & Lane, 2015).

Os últimos dois anos estão a ser considerados anos atípicos por conta do estado de pandemia de COVID-19 que se vive à escala mundial e em particular no nosso país. 2020 e 2021 têm sido um desafio em todos os contextos e é em particular no contexto social que se têm sentido um maior impacto. A pandemia foi provocada pelo novo coronavírus, identificado pela primeira vez em humanos na China, na cidade de Wuhan (DGS, 2020). O grande potencial de transmissão e efeitos graves a nível respiratório, o elevado número de mortes e o relativo desconhecimento a longo prazo dos efeitos provocados pelo vírus levaram a que muitos países tenham decidido implementar inúmeras restrições sociais para assim tentar mitigar os efeitos de transmissão do vírus.

As restrições impostas pelos governos de grande parte do mundo e a procura por sítios mais isolados, no sentido de evitar ajuntamentos levou a que o Turismo Rural ganhasse algum destaque em Portugal, sobretudo pelas características inerentes associadas ao isolamento dos vários alojamentos e o contacto com a natureza longe de multidões.

O sector turístico foi severamente afetado pela pandemia, no entanto o turismo de espaço rural foi o sector que registou uma menor quebra no contexto em que vivemos atualmente, tendo em maio de 2020 levado a um crescimento acentuado pelo termo nas pesquisas dos portugueses, tendência esta que se manteve durante os meses de verão (*IPDT*, 2020). Tal comportamento pode ser observado na Figura 2.4 que representa a procura do termo "Turismo Rural" no motor de busca Google utilizando a ferramenta *Google Trends.*

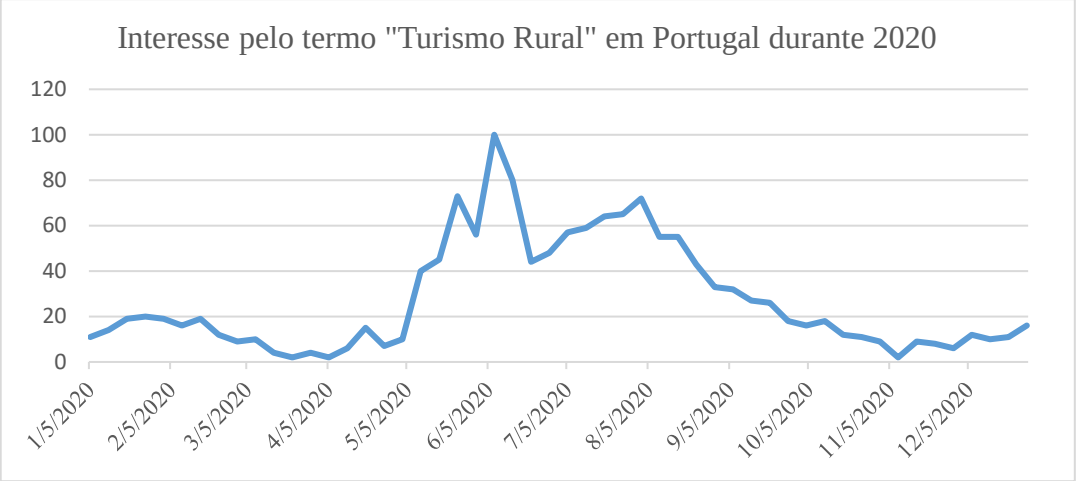


Figura 2.4 – Procura pelo termo "Turismo Rural" em portugal

Fonte: Google Trends (2020)

Esta ferramenta gratuita analisa uma amostra de pesquisas na página web para determinar quantas pesquisas foram realizadas sobre determinado tema num certo período temporal. Ao pesquisar na ferramenta um termo, esta devolve um gráfico que mostra a popularidade desse termo no motor de busca. Os números do gráfico não refletem o número real de procuras, mas sim, uma versão normalizada da mesma de 0 a 100, em que o número 100 é o ponto mais alto da procura do termo.

2.4 Descoberta de Conhecimento em Bases de Dados

A Descoberta de conhecimento em bases de dados (DCBD) tem como foco principal a exploração computorizada de grandes volumes de informação no sentido de descobrir padrões de interesse nos mesmos (Feldman et al., 1998). Nas últimas décadas tem existido um interesse crescente no sentido de desenvolver sistemas de extração de informação, tanto num contexto de utilidade como habilidade para o efeito, já que, existe uma enorme quantidade de informação apenas na sua forma natural e não estruturada, que necessita de ser tratada e estruturada para ser possível de interpretar e retirar informação válida para a criação de valor (Grishman, 1998).

Nos últimos anos, muito associado ao conceito de DCBD vem também o conceito de Big Data que se define em Baig et al. (2019) como um conjunto de dados em que se torna difícil o seu processamento à mão ou utilizando técnicas tidas como mais tradicionais ao nível do processamento e tratamento de dados. Devido a este facto, existem várias problemáticas associadas que incluem a recolha, tratamento, armazenamento, pesquisa, partilha, análise e visualização dos dados em questão.

Para tornar todo este processo de DCBD possível, são definidos em Fayyad et al. (1996) várias fases de todo um processo que levam a cabo esta tarefa.

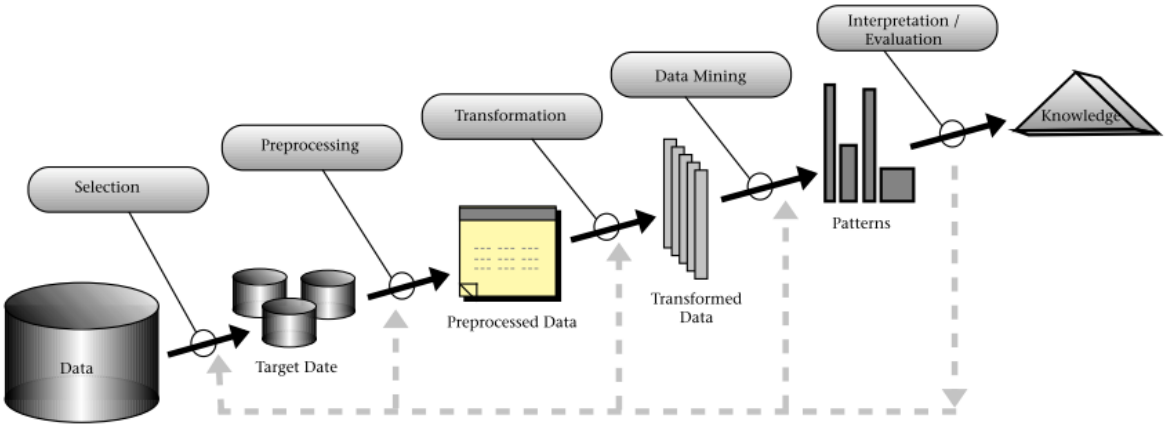


Figura 2.5 – Etapas incluídas do processo de descoberta de conhecimento em Base de Dados

Fonte: Hendrickx et al. (2015)

O processo envolve a seleção dos dados a observar, o seu pré-processamento e transformação, a etapa de *data mining* onde são implementados algoritmos que permitem descobrir padrões e retirar informação útil que depois é interpretada com o intuito de gerar conhecimento, conforme pode ser verificado na Figura 2.5.

O *Data Mining* tem sido cada vez mais utilizado nos mais diversos domínios, seja no setor público, como nas áreas fiscal (Seiça et al., 2019), educação (Pimenta, Ribeiro, Sá, & Belfo, 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 (Pimenta, Belfo, & Trigo, 2009) e em muitas outras áreas, contribuindo para a criação de novos conhecimentos e ajudando as organizações a definir estratégias que lhes permitam aumentar o seu desempenho.

Turistas, ao visitar partes diferentes do mundo, ou do próprio país de origem, pelos mais variados motivos, têm a tendência para procurar aconselhamento *online* antes de realizar uma reserva (cerca de 65% de acordo com Fang et al. (2016)). Por sua vez os estabelecimentos solicitam em regra que os seus hóspedes comentem a experiência durante a estadia que efetuaram criando assim uma base de dados que pode ser analisada

(Banerjee & Chua, 2016). Estando cada vez mais presente a tendência de partilhar e transmitir a experiência turística nas várias plataformas online, estas bases de dados crescem a todo o momento a um ritmo cada vez maior.

Em Fang et al. (2016) é igualmente referida a importância da avaliação e do seu conteúdo. Esta, para fornecer informação efetivamente válida, deve ser precisa e fácil de entender de forma a evitar conflitos de interpretação. É possível assim, por vezes, tirar elações sobre o próprio estilo da escrita sobre o status social e personalidade do autor do texto. O artigo sugere ainda que avaliações com maior clareza devem ser tidas como prioritárias como fonte de informação. Outro fator importante associado a uma avaliação, referido no mesmo artigo é a sua classificação que no fundo resume neste caso numericamente a avaliação que um consumidor dá a determinada experiência. Desta forma o leitor pode facilmente identificar a atitude e sentimento do autor. Na perspetiva de um potencial consumidor, estas avaliações são consideradas autênticas, influenciais e de confiança (X. Li & Hitt, 2008).

Os turistas têm as mais variadas razões para avaliar alojamentos, tanto para demonstrar o seu agrado como o seu desagrado, ou demonstrar a sua satisfação ou frustração face à experiência obtida, muitas vezes de forma altruísta para aconselhar futuros hóspedes. Este comentário é também muitas vezes dado de forma de crítica construtiva de forma a ajudar os próprios alojamentos a encontrar e resolver determinadas situações que possam colocar em causa a qualidade dos serviços oferecida (Banerjee & Chua, 2016).

Este trabalho irá incidir bastante na análise de texto (neste caso comentários de utilizadores) tendo como base os princípios do Processamento de Língua Natural (PLN).

O PLN é uma área de pesquisa e aplicação que explora a forma como computadores podem ser utilizados para entender e processar linguagem sob a forma de texto ou discurso, com o intuito de realizar tarefas úteis. Esta pesquisa, incide na procura de conhecimento sobre como o ser humano entende e utiliza a sua linguagem e tenta replicar o mesmo método utilizando algoritmos estatísticos ou matemáticos de forma a atingir determinados objetivos ou tarefas (Chowdhury, 2003).

2.5 Metodologias utilizadas em trabalhos na área do Turismo

A revisão de literatura para este trabalho 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 apenas escolhidas 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 16 artigos com base na leitura do *abstract* e relevância para o assunto a estudar. O esquema relativo à seleção de artigos para a revisão de literatura pode ser analisado na Figura 2.6 abaixo.

Figura 2.6 – Esquema utilizado para a Revisão de Literatura

Adicionalmente, e por pertinência para o contexto do estudo a realizar foi incluído um artigo publicado em 2011 por Park & Yoon (2011), pois aborda aspetos pertinentes relativos à criação de indicadores para gestão no sector do turismo rural e, dado que não existe uma grande quantidade de estudos nesta área, escolhi incluí-lo na revisão de literatura a realizar.

Em termos de trabalhos realizados nesta área, o turismo revela ser um sector de negócio que movimenta uma grande quantidade de recursos financeiros, sendo em alguns países, o principal motor financeiro e comercial. Para além disso, existe também o facto de a análise de dados ser um tema *"trend"* nos últimos anos, o que deu origem a vários artigos relacionados com o trabalho que aqui é proposto.

Sánchez-Franco et al. (2019) estudaram 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. Para o efeito foi obtida uma amostra de 47.172 avaliações de hotéis de Las Vegas classificados na plataforma *Yelp* no sentido de retirar indicadores que pudessem identificar e classificar a satisfação dos consumidores. Aplicando o algoritmo de classificação de Naive Bayes, foi possível concluir que o mesmo é bastante preciso e requer um baixo custo computacional e foi identificada a evidência de que hotéis com bons serviços ligados ao jogo de apostas, quartos com comodidades apreciadas pelos hóspedes e entrega dos funcionários ao serviço têm mais tendência para gerar melhores avaliações. O estudo reforça ainda a importância da aplicação deste tipo de metodologia para tirar melhor proveito da informação dispersa de forma a criar valor na tomada de decisão.

Outro estudo, neste caso efetuado por Khorsand et al. (2020) em Teerão e utilizando a plataforma *TripAdvisor* como fonte de dados, foi efetuado no sentido de prever uma avaliação de um hotel apenas com base na informação do hotel e do utilizador. Nesse sentido foram obtidas 4.718 avaliações referentes a estadias em 64 hotéis listados no *TripAdvisor* na cidade de Teerão e foram utilizados vários modelos de machine-learning de modo a escolher o mais indicado. Neste caso concluiu-se que o *K-nearest neighbors* (KNN) foi o mais indicado pois permite identificar as semelhanças e distâncias dos vários termos para a sua classificação e análise estatística. O estudo concluiu que 77% dos fatores obtidos são relacionados com caraterísticas associadas aos hóspedes (data de comentário e tipo de viagem por exemplo) e apenas os restantes são relacionados com os

estabelecimentos em si. Ainda assim são identificados várias áreas e comodidades chave para garantir a satisfação dos utilizadores de modo a obter melhores avaliações entre as quais Wi-Fi, bar, restaurante, estrelas do hotel, serviço de quartos, *staff* multilingue, limpeza a seco e sauna.

O algoritmo *Latent Dirichlet Allocation* (LDA) aparece como uma eficaz abordagem neste contexto, tendo sido utilizado 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 estudo em questão enumerou um conjunto de 30 tópicos associados à satisfação dos utilizadores em conjunto com as 20 palavras mais mencionadas em cada um dos tópicos. Foi também analisada a componente controlada ou não controlada do universo de tópicos existentes e ficou evidente que existem diferenças a nível demográfico na satisfação dos utilizadores.

Um estudo recente realizado em território chinês foi realizado por Luo et al. (2021) com vista a analisar comentários turísticos obtidos na visita a 24 Geoparques que fazem parte do património mundial da 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. Para o efeito foram obtidas 120.532 avaliações online e analisadas utilizando os algoritmos de *Support Vector Machine* (SVM), uma versão melhorada do algoritmo de LDA e o algoritmo por *importance-performance analysis* (IPA). Os resultados da aplicação dos vários modelos revelaram 10 atributos importantes e significativos para avaliação e talvez mais importante ainda revelaram alguns atributos avaliados negativamente pelos turistas sendo os seguintes o custo da viagem, serviços da viagem, conhecimento da matéria e transporte/alojamento. Destes comentários foram dadas algumas sugestões para melhorar os serviços e a experiência turística inerente.

Os algoritmos do modelo Naive Bayes e LDA são também utilizados por Taecharungroj & Mathayomchan (2019) com o objetivo de analisar avaliações obtidas online, com o intuito de fornecer informação para a gestão no sentido de melhorar as suas atrações num contexto de atrações turísticas em Phuket na Tailândia. Para este estudo foram obtidas 65.079 avaliações da plataforma TripAdvisor de atrações tão variadas como praias, ilhas, templos, ruas e mercados na região. Foi possível com este estudo, à semelhança do estudo referido anteriormente, definir vários atributos referentes aos diferentes cenários

turísticos e responder a questões como perceber quantas avaliações estão incluídas em cada atributo, o quão positivos são os atributos encontrados, quão frequentes e positivos são os termos encontrados em cada atributo e qual o grau de precisão que é possível obter na previsão dos comentários dos utilizadores aplicando o modelo de Naive Bayes nos comentários analisados.

Pokryshevskaya & Antipov (2017) realizaram um estudo bastante relevante para a temática que abordamos neste trabalho. Neste, no sentido de tentar prever a satisfação dos hóspedes em relação aos serviços disponibilizados a nível de hotelaria, foram obtidas de um hotel situado no Dubai 3.630 avaliações da plataforma *Booking.com* e respetivos perfis/características dos utilizadores para análise. O estudo conseguiu identificar os perfis com maior probabilidade de ficarem com uma opinião negativa, de forma a ser alvo de maior atenção por parte da gerência e também os perfis com maior probabilidade de terem uma opinião positiva acerca dos serviços disponibilizados e que, portanto, não requerem uma atenção extra. A análise foi realizada utilizando o modelo de regressão linear para detetar quais as características mais relevantes para o mesmo, baseando-se nos termos utilizados nos comentários para demonstrar a satisfação ou insatisfação durante a sua estadia. O que destaca este estudo dos demais é o foco maior nas caraterísticas do hóspede e não tanto nos comentários, ainda que estes sejam também importantes para análise,

Também a plataforma *Airbnb*, uma das plataformas *peer-to-peer* que, nos últimos anos conheceu mais sucesso, foi alvo de estudo por Cheng & Jin (2019). Para o efeito foram analisados 181.263 comentários de alojamentos da cidade de Sidney, Austrália no sentido de perceber quais os aspetos que os utilizadores desta plataforma mais valorizam. Foram utilizadas técnicas de *text-mining* e análise de sentimento com recurso ao *Leximancer*, software especializado para o efeito, de forma a atingir o objetivo proposto. Foram detetados 4 tópicos essenciais da análise aos comentários sendo os quais a localização, as comodidades, o anfitrião e as recomendações. Numa segunda fase foi realizada a análise de sentimento aos vários tópicos no sentido de perceber quais são os aspetos considerados mais positivos por parte dos utilizadores. Conclui-se que a experiência dos visitantes é maioritariamente positiva neste estudo.

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 justificação para este estudo parte do facto de existir uma quantidade enorme de comentários que não acrescentam valor, relevância ou sentimento válido para a tomada de decisão da gestão assim como da perspetiva do consumidor ao procurar recomendações para uma possível estadia ou experiência turística, ao ser necessário despender uma quantidade considerável de tempo à procura de comentários objetivos que lhe permitam formular uma opinião válida para a sua própria tomada de decisão. 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.

Yi Luo et al. (2020), também com a premissa da complexidade associada ao enorme volume de informação à disposição dos utilizadores, realizaram um estudo no sentido de entender a relação existente entre uma avaliação dada a um atributo específico, relacionado com a importância e o sentimento relativo a um estabelecimento, e a satisfação geral face ao mesmo. Neste caso a avaliação foi feita utilizando como base 4.704 restaurantes nas principais cidades dos Estados Unidos: Chicago, Las Vegas, Los Angeles, Nova Iorque e Orlando. foram obtidos 244.649 comentários da plataforma *Yelp.com* para a análise efetuada. Apesar do foco ser, neste caso, restaurantes e não estabelecimentos de hotelaria, a problemática é bastante relacionada com o foco deste trabalho e a importância do comentário do utilizador é comum nos dois temas. Para o estudo foi utilizado como abordagem o modelo *Latent Aspect rating analysis* (LARA). Este modelo tem como objetivo explicar e prever de forma dinâmica o valor numérico de uma entidade, que pode ser um negócio ou *rating* com base em atributos ou aspetos retirados de comentários escritos. A metodologia proposta conta com 3 passos essenciais: o pré-processamento dos dados, identificação dos aspetos e análise de sentimento dos mesmos. Para a identificação dos aspetos foi uma vez mais utilizado o algoritmo LDA que deu origem a 5 tópicos sendo os quais as comidas/bebidas, serviço, ambiente do restaurante, valor do restaurante e localização. Destes 5, o valor do restaurante, que neste

caso representa o rácio qualidade/preço para o utilizador, foi considerado o mais importante para a avaliação final do estabelecimento.

Tendo em conta que uma das principais preocupações no sector turístico é perceber o que provoca insatisfação na experiência dos turistas em determinado contexto, foi proposto um estudo por Kim et al. (2017). Para o efeito foi utilizada uma amostra de 19.835 comentários da plataforma *Virtualtourist.com*, de vários estabelecimentos de Paris organizados por 14 categorias: avaliação geral, restaurantes, vistas, hotéis, atividades, ambiente noturno, transporte, compras, desporto e ar livre, favoritos, atrações menos procuradas, o que levar, armadilhas para turistas, perigos e costumes locais. O estudo foi feito sobre a premissa de que mesmo existindo análise de sentimento aliado à contagem de palavras/termos, esta é muitas vezes insuficiente pois apesar de ser possível saber o alvo do descontentamento do utilizador, é difícil perceber a razão pela qual este está descontente. Empregando um método com o termo de *co-ocurrence* (português para coocorrente), em que se tenta encontrar padrões em palavras que ocorrem recorrentemente em pares, foi possível concluir que a categoria transporte foi alvo de descontentamento pela generalidade dos turistas e que as razões se prendem com as altas tarifas aplicadas no serviço de táxis, as fracas condições a nível de estrutura na rede de metro de Paris e a falta de limpeza no serviço rodoviário da cidade.

Hou et al. (2019) abordam a identificação dos temas e a comparação das diferenças existentes nos comentários dos utilizadores em 3 agências de viagens (*Ctrip, Tuniu and Tongcheng*) com plataforma online implementada na China. Para o efeito é utilizado um método de análise de associação semântica que consiste na extração de palavras temáticas de um conjunto de dados e na construção de uma rede visual de associação semântica dos termos e para uma maior compreensão das relações entre os mesmos. Para o efeito foi utilizada uma amostra de 165.429 comentários obtidas com base nos comentários retirados das 3 agências analisadas. O estudo encontrou diferenças consideráveis nas palavras temáticas entre as várias agências, e nas suas relações a nível estrutural. O que indica que existe uma relação próxima entre os comentários dos utilizadores e alvo do modelo de negócio aplicado por cada agência o que origina diferentes níveis de interesse e importância nos comentários dos utilizadores.

Devido ao facto de, por natureza, o turismo rural ser um nicho neste mercado e também o tamanho reduzido da maioria dos estabelecimentos não existem atualmente muitos

estudos dedicados às áreas da gestão, planeamento e monitorização desta atividade. No entanto, existem já alguns esforços para chegar a um consenso no que diz respeito a desenvolver de forma sustentável, indicadores válidos a utilizar como medida de controlo Park & Yoon (2011) definiram quatro dimensões fulcrais para esta matéria, e um total de 33 indicadores, sendo as dimensões as seguintes: Qualidade de Serviço, Instalações, Sistema de Gestão e Resultados. A qualidade do serviço está muito relacionada com a sua acessibilidade e conveniência para o cliente no sentido de facilitação de acessos, reservas e pagamento. Em termos de instalações foi dada especial atenção à qualidade dos quartos, divisões complementares do alojamento e preocupação ambiental. Para o Sistema de gestão a atenção é mais centrada na própria comunidade e na importância do capital humano na organização e na criação de uma experiência de criação de valor no espaço envolvente. Por fim, no que diz respeito aos Resultados é de especial valor a satisfação dos clientes, visitantes e moradores da comunidade envolvente assim como o aumento do número de visitantes a um determinado alojamento de turismo rural. Este referido estudo salientou a importância acrescida da satisfação do cliente e do sistema de reserva na disponível na Internet como indicadores de avaliação, o que de certa forma vai de encontro ao expectável e já referido anteriormente.

Foi desenvolvido um estudo por Pitchayadejanant & Nakpathom (2018) com o objetivo de perceber quais as atividades preferidas pelos turistas ao visitar terrenos de cultivo de frutas na Tailândia num contexto de agroturismo. Para o efeito foram utilizados métodos de *data-mining* no sentido de obter regras de associação e agrupamento, numa amostra obtida por questionário com 409 observações, de modo a encontrar padrões de comportamentos turístico. Da lista de atividades tipicamente disponíveis para os turistas realizarem, a colheita e prova de fruta foi a atividade que mais contribui para a satisfação dos turistas aliada também às caminhadas, compras de artigos típicos e alimentação de animais em segunda instância.

Aliado ao Turismo Rural existe uma tendência para um maior contacto com a natureza, muito aliado às características já anteriormente referidas quando foi abordado este sector em particular. O crescimento generalizado do desejo das pessoas pela observação e interação com a vida selvagem no seu habitat natural, associado a uma maior sensibilidade ambiental provocou um crescimento neste sector turístico que é estudado em Prakash et al. (2019). Neste caso foram estudadas as razões associadas ao descontentamento demonstrado pelos turistas em visita aos parques naturais no Sri Lanka.

Para a amostra deste estudo foram obtidas 206 avaliações da plataforma *TripAdvisor* com a pior classificação atribuída pelos utilizadores. Foram encontradas, por tratamento estatístico, 15 razões para o descontentamento dos turistas na sua experiência, sendo 75% dos quais relacionados com a gestão do parque e muito focados no elevado trânsito dentro da reserva, que de alguma forma afeta a imersão no ambiente mais puro e selvagem do parque.

Também ligado ao estudo da satisfação dos utentes em áreas verdes em zonas urbanas, foi efetuado um estudo por Sun & Shao (2020) para medir o grau de satisfação dos habitantes da zona de Shenzhen na China perante as áreas verdes ao seu dispor. Estes estudos, normalmente efetuados utilizando métodos de questionários e entrevistas, são ultrapassados quando existem atualmente comentários oriundos das redes sociais que podem fornecer informação em maior quantidade e atempadamente possível de analisar com poucos recursos financeiros. Como amostra foram utilizados 3.511 comentários da rede social *Weibo* que foram alvos da aplicação de 2 métodos de *machine learning*: Word2Vec e *Long Short-Term Memory Model* (LSTMM). O estudo expõe o grau de satisfação dos utentes quanto aos parques e conclui que a faixa etária que se demonstra mais satisfeita está situada entre os 21 e os 40 anos e de posses financeiras mais limitadas, dado o facto de muitos parques serem de entrada gratuita.

Ruhanen (2019) também realizou um estudo de *data-mining* utilizando como fonte o conteúdo de utilizadores online, investigando no sentido de entender o sentimento associado à experiência ecoturística, neste caso na Austrália, baseado em mais de 3022 comentários da plataforma *Tripadvisor.com*, O interesse deste estudo prendeu-se com a questão que é colocada verificando o crescimento deste tipo de turismo e se esse crescimento se deve ou não a uma maior preocupação e sensibilidade ambiental por parte dos consumidores. Da análise efetuada foram retirados os tópicos essenciais: Staff, Comodidades, Natureza, Comida, Aprendizagem e Experiência. Concluiu-se que embora exista uma maior procura para este tipo de experiência, esta é mais motivada pela própria experiência em si do que pela preocupação ambiental e a sua sustentabilidade. Grande parte dos comentários relativos ao fator ambiental e ao ecoturismo diziam respeito sobretudo à adaptação das infraestruturas e comodidades às áreas naturais envolventes e não a uma preocupação genuína com o tema.

Em Portugal e ligado ao turismo rural, Eusébio et al. (2017) estudaram a questão acerca de quem é o principal consumidor em áreas rurais e é portanto, efetuada uma análise do turismo rural doméstico em Portugal. O estudo foi realizado com base num questionário com 1.853 respostas em que 866 declararam ter realizado atividades relacionadas com turismo rural nos últimos 3 anos, relativamente à data do estudo efetuado. Das respostas obtidas foram obtidos 4 agrupamentos associados aos perfis dos turistas: visitantes ativos, visitantes passivos e observadores da natureza, visitantes inativos e visitantes em família durante as férias. O estudo concluiu que existe heterogeneidade nas características sociodemográficas assim como nos comportamentos durante a experiência turística. Concluiu-se também que os visitantes ativos são mais propensos a aderir a atividades relacionadas com turismo rural relacionadas com a fauna, flora e contemplação de paisagem rural. Foi dado também destaque às características ambientais em contexto rural, a tranquilidade envolvente e as atividades culturais e gastronómicas do local a visitar. Seguem abaixo as Tabela 2.2 e Tabela 2.3 com o quadro síntese referente à revisão de literatura efetuada para este trabalho e que pode ser verificado mais em pormenor no Apêndice 6.

Tabela 2.2 – Quadro síntese referente à revisão de literatura realizada

Autor(es)	Tema	Amostra	Métodos Utilizados
Sánchez-Franco et al. (2019)	Identificação de termos relacionados com a experiência dos hóspedes no sentido de serem utilizados para melhorar o seu serviço a nível de hospitalidade em hóteis de Las Vegas	47.172 comentários da plataforma *Yelp.com*	Naive Bayes
Khorsand et al. (2020)	Prever uma avaliação de um hotel na cidade de Teerão apenas com base na informação do hotel e do utilizador	4.718 comentários de 64 hotéis listados na plataforma *TripAdvisor.com*	K-nearest neighbors Naïve Bayes Árvore de decisão regressão logística Support Vector Machine Neural Network Random Forest Gradient Boosting Data Base sensitivity
Guo et al. (2017)	Identificação das dimensões chave referentes à satisfação com o serviço hoteleiro	266.544 comentários de utilizadores em 25.670 hotéis de 16 países na plataforma *TripAdvisor.com*	Latent dirichlet analysis (LDA)
Luo et al. (2021)	Analisar comentários turísticos obtidos na visita a 24 Geoparques que fazem parte do património mundial da 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	120.532 comentários acerca de 24 Geoparques obtidos das principais comunidades turisticas online na China	Support Vector Machine Versão modificada/melhorada do algoritmo Latent dirichlet analysis (LDA)
Taecharungroj & Mathayomchan (2019)	Analisar as opiniões com base em *reviews* obtidas online de forma a fornecer informação para a gestão no sentido de melhorar as suas atrações, sendo neste caso num contexto de atrações turísticas em Phuket na Tailândia	65.079 comentários da plataforma *TripAdvisor.com*	Naive Bayes e LDA
Pokryshevskaya & Antipov (2017)	Prever a satisfação dos hóspedes em relação aos serviços disponibilizados a nível de hotelaria	3.630 comentários da plataforma *Booking.com*	Modelo de regressão linear
Cheng & Jin (2019)	Perceber quais os aspetos que os utilizadores mais valorizam numa plataforma dedicada a aluguer de alojamento	181.263 comentários de alojamentos listados na plataforma *Airbnb* da cidade de Sidney	Foram utilizadas técnicas de *text-mining* e análise de sentimento com recurso ao software *Leximancer*
Tsai et al. (2020)	Propor uma abordagem de sumarização dos comentários no sentido de classificar os mesmos de forma a salientar os mais importantes e relevantes para posterior avaliação	23.430 comentários de 23.038 utilizadores na plataforma *TripAdvisor.com*	Construção de classificadores para os comentários e posterior análise de sentimento dos mesmos

Tabela 2.3 - Quadro síntese referente à revisão de literatura realizada (continuação)

Autor(es)	Tema	Amostra	Métodos Utilizados
Yi Luo et al. (2020)	Entender a relação existente entre uma avaliação dada a um atributo específico, relacionado com a importância e o sentimento relativo a um estabelecimento, e a satisfação geral face ao mesmo em restaurantes localizados em Chicago, Las Vegas, Los Angeles, Nova Iorque e Orlando, USA	244.649 comentários de 4.704 restaurantes na plataforma *Yelp*	LARA (*Latent Aspect rating analysis*)
Kim et al. (2017)	Perceber o que provoca insatisfação na experiencia dos turistas em determinado contexto	19.835 comentários da plataforma *Virtualtourist.com* de vários estabelecimentos de Paris organizados por 14 categorias	método *Co-ocurrence*
Hou et al. (2019)	Identificação dos temas e a comparação das diferenças existentes nos comentários dos utilizadores em 3 agências de viagens com plataforma online implementada.	165.429 comentários obtidas com base nos comentários retirados de 3 agências de viagens (Ctrip, Tuniu and Tongcheng)	Análise de associação semântica
Park & Yoon (2011)	Desenvolver de forma sustentável indicadores válidos a utilizar como medida de controlo na gestão no contexto de Turismo Rural	34 especialistas na área	Metodo Delphi
Pitchayadejanant & Nakpathom (2018)	Perceber quais as atividades preferidas pelos turistas ao visitar terrenos de cultivo de frutas na Tailândia num contexto de agro-turismo	409 observações obtidas por questionário	FP-Growth Algorithm e agrupamento
Prakash et al. (2019)	Estudar as razões associadas ao descontentamento demonstrado pelos turistas em visita aos parques naturais no Sri Lanka	206 *reviews* da plataforma *TripAdvisor* com a pior classificação atribuída pelos utilizadores	Tratamento estatístico
Sun & Shao (2020)	Medir o grau de satisfação dos habitantes da zona de Shenzhen na China perante as áreas verdes ao seu dispor	3.511 comentários da rede social *Weibo*	Word2Vec e *Long Short-Term Memory Model*
Ruhanen (2019)	Entender o sentimento associado à experiência ecoturística neste caso na Austrália e se o crescimento assinalado se deve ou não a uma maior preocupação e sensibilidade ambiental por parte dos consumidores.	3022 comentários da plataforma *Tripadvisor.com*	Tratamento manual aliado a análise semântica
Eusébio et al. (2017)	Qual o principal consumidor nas áreas rurais em Portugal	questionário com 1.853 respostas	Tratamento estatístico em agrupamento

3 METODOLOGIA

Por se tratar de um estudo na área da mineração de dados, a metodologia utilizada suportou-se na metodologia Cross-Industry Standard Process for Data Mining (CRISP-DM). Concebida em 1966, esta metodologia consiste num modelo de processos hierárquicos, com diversas tarefas, distribuídas por vários níveis de abstração. 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). Na Figura 3.1 podemos verificar as principais relações entre as mesmas.

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.

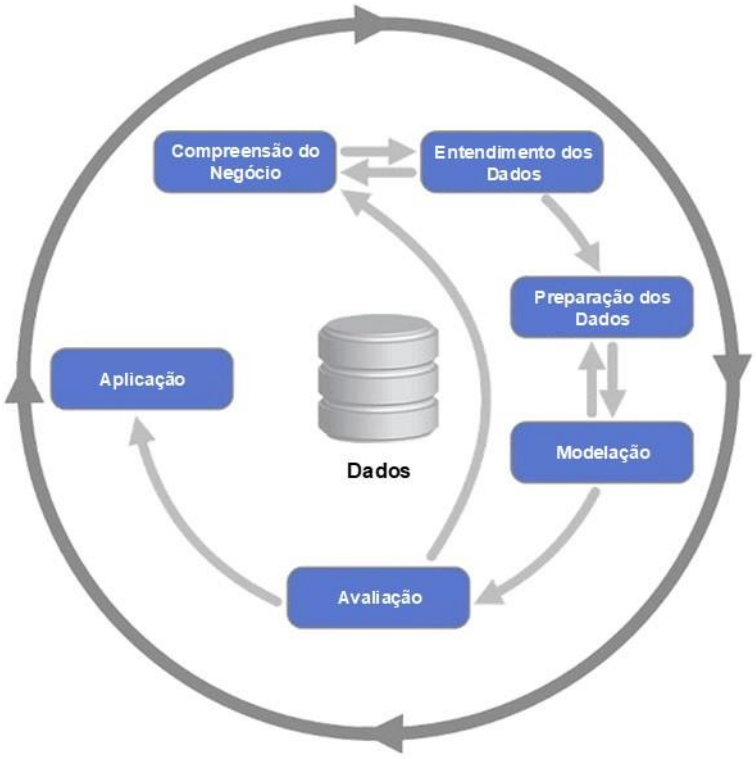


Figura 3.1 – Fases envolvidas no Modelo CRISP-DM

Fonte: Adaptado de Chapman et al. (2000)

Esta metodologia tem nos últimos anos sido utilizada como referência para projetos de *Data Mining.* As várias fases referidas acima são concebidas no contexto deste trabalho nos seguintes termos:

- Entendimento do negócio: A fase de entendimento do negócio corresponde essencialmente à revisão de literatura efetuada e já apresentada anteriormente.
- Entendimento dos Dados: Corresponde à recolha dos dados e, posteriormente, uma pré-análise volumétrica de forma a ficarmos com uma ideia geral do conjunto de dados disponível em termos descritivos e da qual é possível avançar com algumas suposições de forma imediata.
- Preparação dos dados: Envolve a alteração dos dados de origem e a eventual criação de novos dados, assim como as etapas e limpeza dos mesmos e eliminação de informação não pertinente. No caso do estudo a realizar, os dados referentes aos comentários serão alvo de especial atenção, sem, no entanto, descurar os restantes dados que podem oferecer informação relevante.
- Modelação dos dados: Criação de modelos, neste estudo, de natureza descritiva de forma a retirar a informação relevante para o estudo a realizar e a descoberta de padrões nos dados previamente tratados.
- Avaliação dos resultados: Discussão dos resultados alcançados, e consequente síntese com o intuito de fornecer conhecimento relevante para a tomada de decisão das partes interessadas, neste caso, proprietários de estabelecimentos de turismo rural em Portugal.

Li et al. (2018) realizaram uma revisão de literatura associada ao tratamento de *Big Data* aplicado no sector turístico em que basicamente foram definidas três grandes fontes de dados: dados gerados por utilizadores, que incluíram texto/fotos *online,* dados de dispositivos, como por exemplo os provenientes de sinal GPS, *roaming* e *bluetooth* e finalmente dados de transações que incluem pesquisas na internet, reservas feitas *online* e registos de visitas em páginas web. O foco neste trabalho esteve no primeiro grupo de fontes de dados, mais especificamente no texto presente nos comentários de utilizadores na plataforma *Booking.com.*

Para a recolha de dados, são normalmente utilizadas tecnologias de *web crawling,* que consistem na criação de algoritmos ou utilização de serviços dedicados para o efeito, no sentido de obter, de forma automática, informação diretamente de *websites* (Li et al., 2018). Tais técnicas são muitas vezes criadas em linguagens e programação como *Python,* R ou Java, normalmente acessíveis a título gratuito ou baixo custo e com uma comunidade online bastante ativa.

3.1 Entendimento dos dados

3.1.1 Recolha dos dados

Como já foi referido no capítulo anterior, a recolha de dados foi realizada tendo como fonte os 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 de pandemia de COVID-19 que se viveu a nível mundial e em particular no nosso país, e que por sua vez provocou um maior interesse neste segmento turístico, acentuado ainda pelas restrições existentes para entrada em outros países.

Optou-se pela obtenção dos dados na plataforma *Booking.com*, principalmente pelas seguintes razões, referidas por Pokryshevskaya & Antipov (2017):

1- Apenas um utilizador que agendou um alojamento pela plataforma está autorizado a comentar, o que torna os comentários reais e fidedignos;
2- Para manter o *rating* atualizado, a plataforma arquiva comentários com mais de 24 meses;
3- Para cada comentário, os pontos positivos e negativos estão separados, facilitando bastante a análise de sentimento e obtenção dos dicionários de dados;

Para a seleção da amostra de alojamentos a utilizar como fonte de dados, foi utilizado o software *Octoparse* (Octoparse, 2020), pois o *Booking.com* não dispõe de uma API que permita aceder aos dados gerados relativamente aos alojamentos e respetivos comentários. A escolha deste software deveu-se ao facto de existirem modelos já pré-definidos que permitem, com relativa facilidade, retirar informação de várias plataformas online e em particular do *Booking.com* sobre a forma de ficheiro CSV pronto a trabalhar.

Foi então necessário definir um critério de seleção temporal das reservas para termos à disposição uma listagem de estabelecimentos com tamanho suficiente para análise. O intervalo temporal definido foi de 18-10-2020 a 22-10-2020. A amostra escolhida continha 471 estabelecimentos. O site *Pordata.pt* lista 1.374 estabelecimentos classificados como Turismo Rural em Portugal em 2021, correspondendo assim a amostra obtida, a 34,3% do total de estabelecimentos deste sector no Pordata (2021). Para os estabelecimentos selecionados na amostra, foram obtidas as avaliações sob a forma de comentários apresentadas na página *web*.

Para o efeito, foi adaptado um algoritmo já utilizado por Wang (2020), utilizando a linguagem de programação *Python*. Esta linguagem é atualmente uma das mais utilizadas no mundo inteiro. Com uma vasta coleção de bibliotecas e uma enorme comunidade online, é um dos recursos mais versáteis nesta área de aplicação e é utilizada com sucesso em milhares de aplicações ligadas ao negócio em todo o mundo (*Python.Org*, 2021). A principal biblioteca que foi utilizada foi a *BeautifulSoup*, uma ferramenta criada especificamente para retirar dados de ficheiros XML ou HTML (*Beautiful Soup Documentation*, 2020).

O código utilizado tornou possível obter não apenas o texto dos comentários, mas também bastante informação relativa ao utilizador que se revelou útil e relevante para análise. A título de exemplo, na Figura 3.2 podemos verificar um exemplo da informação possível de obter com recurso ao algoritmo criado. Foram obtidos, quando disponíveis, os seguintes dados para cada comentário:

1- A nota (de 1 a 10) no comentário de cada estadia;
2- Nome do utilizador;
3- País de Origem;
4- Comentário geral ao alojamento;
5- Nº de comentários submetidos pelo utilizador no *Booking.com;*
6- Data do comentário;
7- Motivo da viagem;
8- Tipo de Viajante;
9- Tipologia;
10- Nº de noites que o utilizador ficou no alojamento;
11- Tipo de submissão;
12- Comentário negativo relativo ao alojamento/experiência;
13- Comentário positivo relativo ao alojamento/experiência;

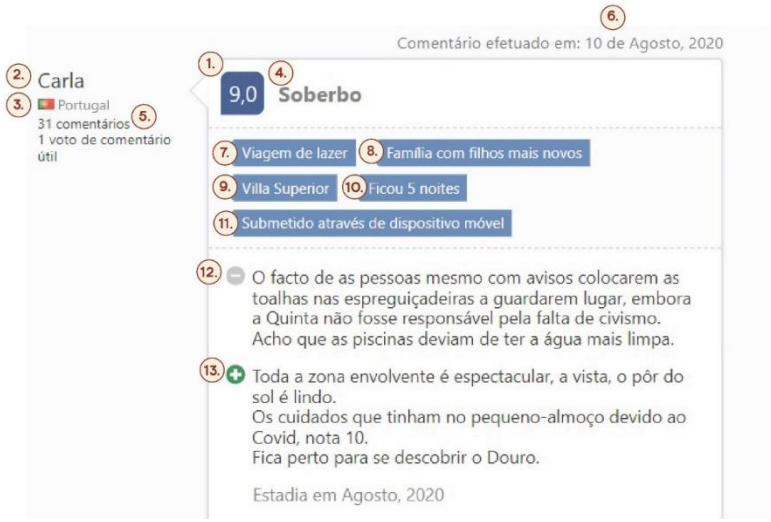


Figura 3.2 – Exemplo de avaliação de utilizador inserida na plataforma web

Fonte: Booking.com (2020)

No total, com recurso ao algoritmo foi obtida uma amostra com 14.976 comentários.

Para completar a base de dados obtida decorrente da utilização do software *Octoparse* e do algoritmo em *Python*, foram também obtidas as coordenadas de forma manual com recurso a pesquisa no *Google.pt*. Tal resultado da pesquisa é obtido de forma bastante objetiva, conforme podemos observar na Figura 3.3.

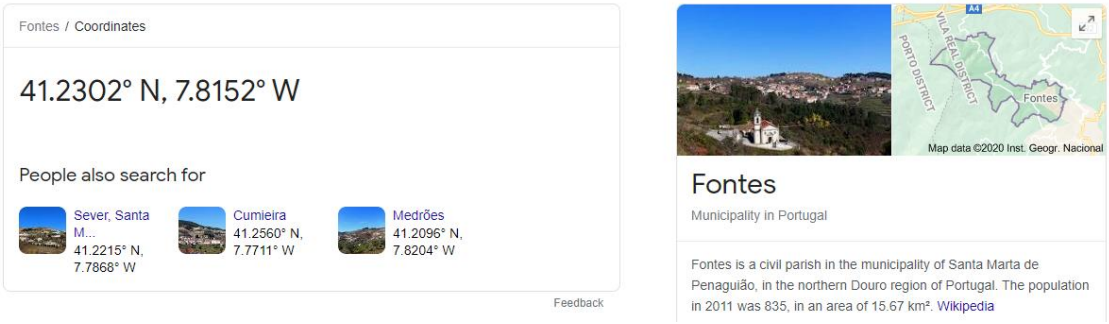


Figura 3.3 – Coordenadas obtidas com recurso a pesquisa no Google.pt

Fonte: Google.pt (2020)

Desta forma, com base nas coordenadas, é possível saber qual a localização exata da localidade extraída na amostra. Utilizando o software da Microsoft PowerBI, conseguimos uma representação geográfica que nos dá uma ideia mais objetiva da amostra recolhida. O PowerBI é um software com uma componente gratuita que permite reunir dados de várias fontes de modo a criar *Dashboards* com o objetivo de fornecer informação pertinente e que acrescente valor na tomada de decisão (Power BI, 2020).

Durante o ato de extração de dados para a amostra, verificou-se que alguns estabelecimentos não possuíam comentários e avaliações válidas para análise, tanto a nível de quantidade como qualidade dos dados, com ausência de informação para extrair. As razões para a exclusão desses comentários baseiam-se no facto de existirem avaliações em que o texto se resume a *emogis*, símbolos com expressões faciais, ou ocasiões em que o utilizador apenas atribui uma nota de 1 a 10 na sua avaliação, sem efetuar qualquer comentário. Excluídos da amostra estes dados, ficou o número de estabelecimentos avaliados reduzido a 400 unidades.

A localização geográfica no continente, na região autónoma dos Açores e da Madeira dos diversos estabelecimentos que fazem parte da amostra está representada, respetivamente, nas Figura 3.4, Figura 3.5 e Figura *3.6*.

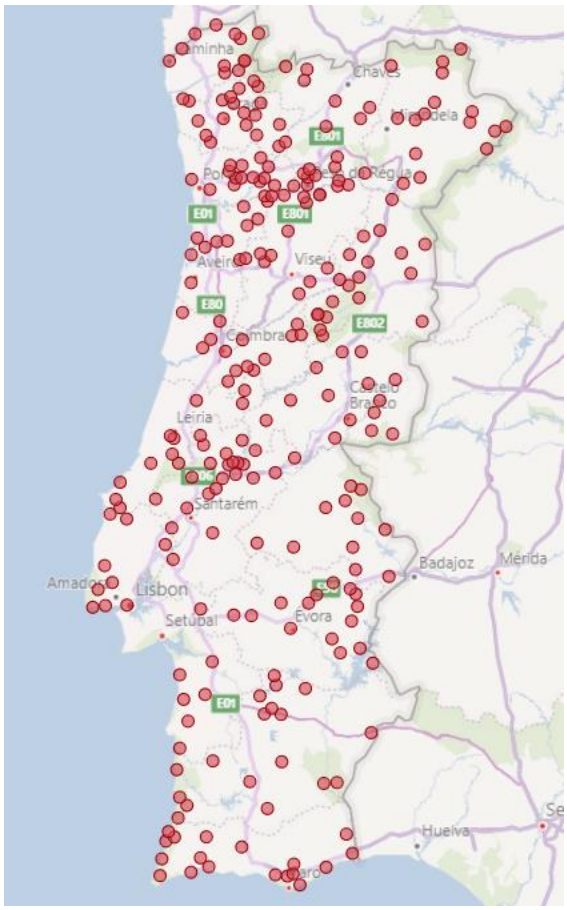


Figura 3.4 – Estabelecimentos em Portugal Continental

Figura 3.5 – Estabelecimentos nos Açores

Figura 3.6 – *Estabelecimentos na Ilha da Madeira*

A Figura *3.7* representa a distribuição dos estabelecimentos obtidos por distritos e regiões autónomas.

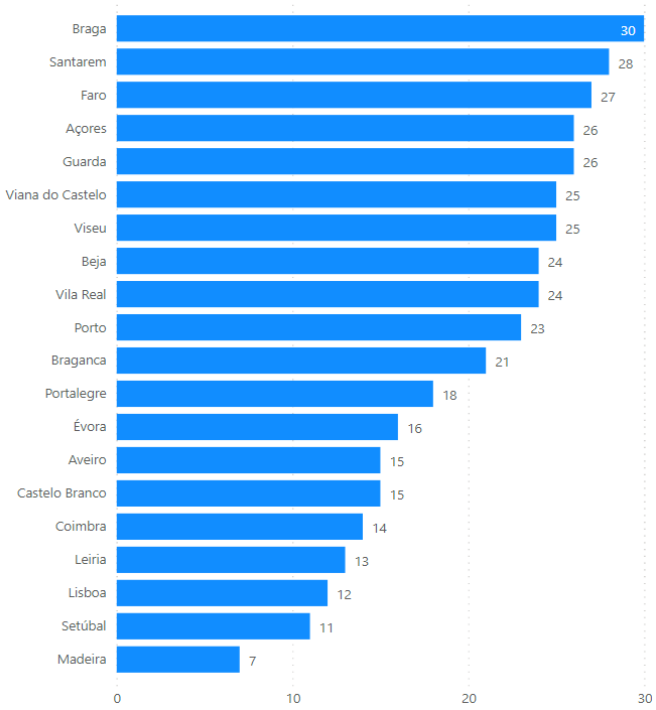


Figura 3.7 – *Estabelecimentos obtidos por distritos e regiões autónomas*

Para além do distrito ao qual pertence o estabelecimento, foi também acrescentado manualmente o subtipo de alojamento rural ao qual o estabelecimento pertence, seguindo os critérios de distinção referidos na Tabela 2.1 do capítulo 2.

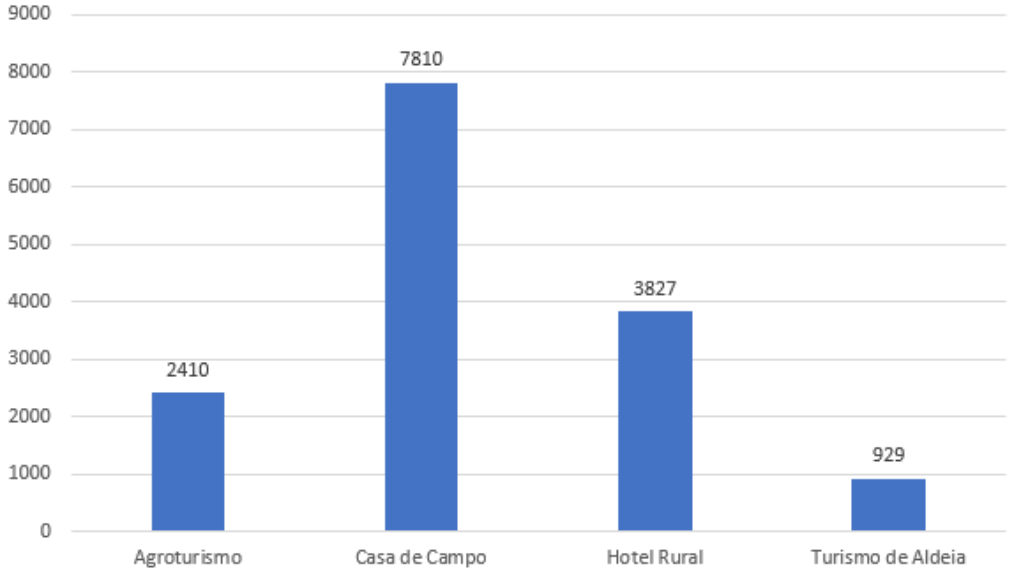


Figura 3.8 – Distribuição do número de comentários da amostra por tipo de estabelecimento

Podemos verificar a distribuição dos alojamentos por tipo na Figura 3.8. Do total dos 400 estabelecimentos, 52 (13%) correspondem a estabelecimentos onde os visitantes podem realizar atividades de agroturismo. A grande maioria (64,5%) dizem respeito a casas de campo com 258 alojamentos selecionados. Foram também obtidas informações sobre 71 hotéis rurais (17,8%) e sobre 19 estabelecimentos ligados ao turismo de aldeia (4,8%).

Na Figura 3.9 está esquematizado o fluxograma a implementar para os dados obtidos. Será realizada uma preparação dos dados, seguida de uma limpeza dos mesmos, etapas que serão detalhadas mais em pormenor nos capítulos seguintes.

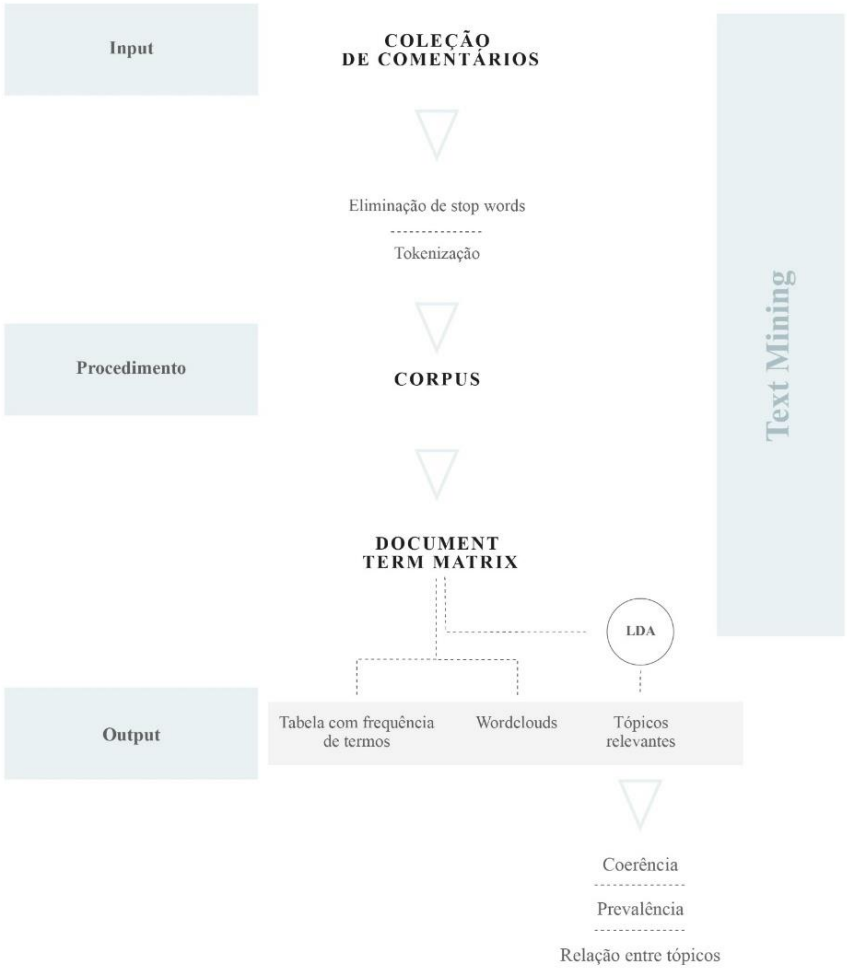


Figura 3.9 – Fluxograma geral a implementar no trabalho

Após a preparação dos dados teremos um conjunto de comentários já pronto para trabalhar implementando os modelos escolhidos para o estudo, que serão descritos também nos capítulos abaixo.

3.1.2 Análise volumétrica dos dados

Neste capítulo é efetuada uma análise volumétrica dos dados obtidos relativos às avaliações dos utilizadores obtidas. No total foram obtidas com recurso ao algoritmo desenvolvido em *Python* já referido acima 14.976 avaliações. De realçar que as avaliações foram obtidas numa data posterior à seleção dos alojamentos para amostra, pelo que já existem avaliações referentes ao ano de 2021. Este intervalo de tempo decorreu do tempo necessário para desenvolver o algoritmo e utilização do mesmo para a recolha dos dados.

Será realizada uma análise estatística de forma a resumir os valores obtidos de forma a ter uma ideia geral da composição dos dados. A estatística descritiva consegue, com recurso a tabelas ou gráficos, uma representação rápida e intuitiva de forma a atingir esse objetivo.

Para análise de cada variável foi utilizado o software Microsoft Excel, utilizando as ferramentas de análise de dados em conjunto com a criação de várias tabelas dinâmicas e respetivos gráficos de modo a extrair informação relevante para o estudo.

3.1.2.1 Nota dada na avaliação de cada estadia

A nota dada por cada utilizador ao alojamento numa determinada estadia tem um valor entre 1 e 10. Como a nota final resulta de uma ponderação de vários aspetos a avaliar, poderá resultar num valor com uma casa decimal. Neste caso, optou-se por truncar a nota final de forma a ficar apenas com a parte inteira da nota

Tabela 3.1 – Estatística descritiva da variável Nota Utilizador

Estatística descritiva Nota Utilizador	
Média	8.995272436
Erro Padrão	0.010059481
Mediana	9.2
Moda	10
Desvio Padrão	1.231043785
Variância da Amostra	1.515468801
Curtose	5.088226611
Assimetria	-1.862724124
Intervalo	9
Mínimo	1
Máximo	10
Soma	134713.2
Contagem	14976

Na Tabela 3.1 podemos verificar alguns dados referentes à estatística descritiva da variável, com especial atenção ao valor médio dos valores que está muito próximo da nota 9, denotando, portanto, que, a grande maioria das avaliações são bastante positivas. A realçar este resultado está o facto de a moda ser 10, ou seja, a maior parte das avaliações foi 10.

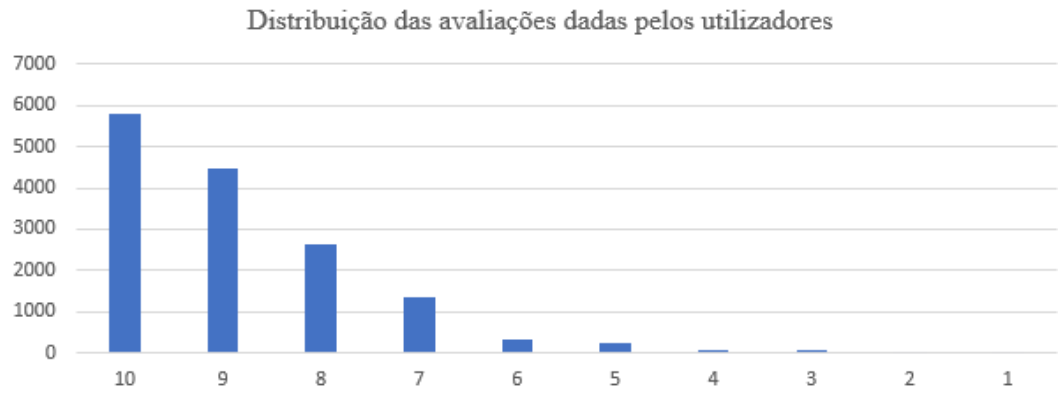


Figura 3.10 – Gráfico com a distribuição das notas dadas pelos utilizadores

Na Figura 3.10 podemos analisar a distribuição dos valores sob a forma de um gráfico.

3.1.2.2 País origem do utilizador

Esta variável identifica o país de origem do utilizador que faz a avaliação. É uma variável significativa já que pessoas de diferentes países podem ter patamares de exigência diferentes e podem dar valor a outros aspetos de cada alojamento. Na amostra recolhida 14.355 (95,9%) das avaliações são portuguesas, enquanto existem 621 (4,1%) que são de outros países, conforme a Tabela 3.2 e a Figura 3.11.

Tabela 3.2 – Distribuição da variável País origem do utilizador

País	Nr. avaliações	%
Portugal	14355	95.9%
Restantes países	621	4.1%
Total	**14976**	**100%**

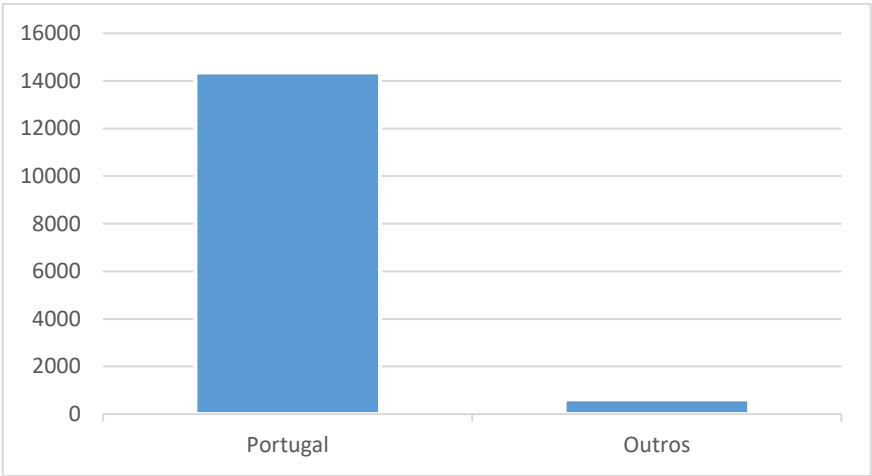


Figura 3.11 – Gráfico da distribuição da variável País origem do utilizador

Aprofundando as 621 avaliações de outros países, podemos ver quais são os restantes

países representados, verificando na Figura 3.12 que o Brasil é o país mais representado, seguido da França, Suíça, Espanha, Reino Unido e Luxemburgo, dentro dos mais representados.

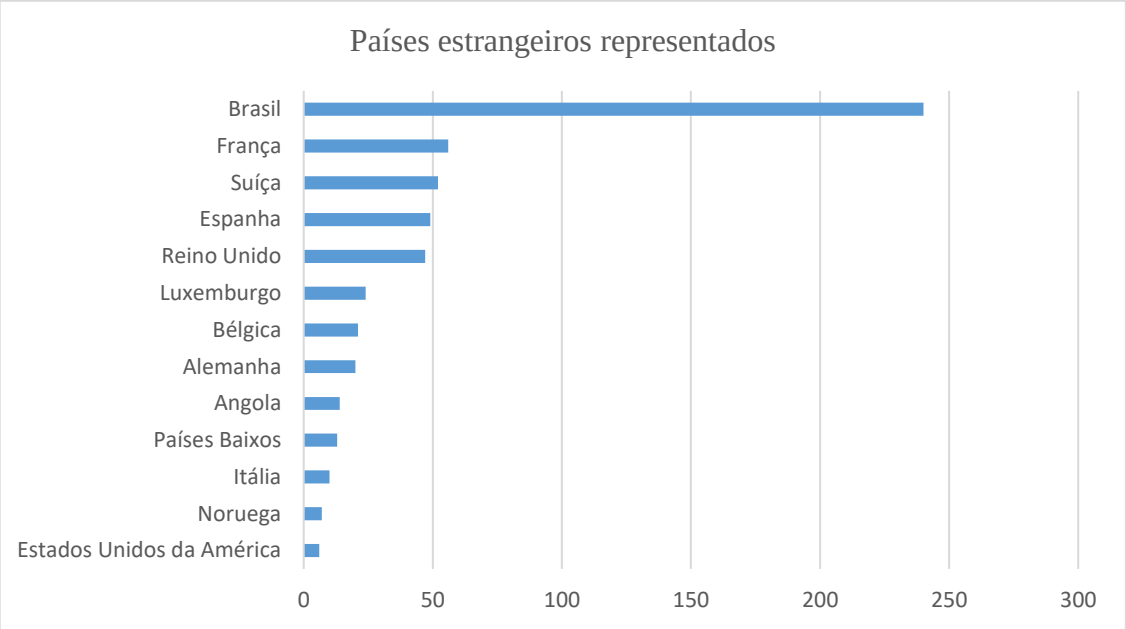


Figura 3.12 – Distribuição do número de avaliações por países mais representados para além de Portugal

Em termos de notas dadas à estadia nos estabelecimentos por hóspedes de outros países, podemos verificar pela estatística descritiva da Tabela 3.3 que em média as notas dadas são marginalmente superiores ao panorama geral.

Tabela 3.3 – Estatística descritiva para notas de utilizadores naturais de outros países que não Portugal

Nota Restantes Países	
Média	9.0719807
Erro Padrão	0.0559738
Mediana	9.6
Moda	10
Desvio Padrão	1.3948595
Variância da Amostra	1.945633
Curtose	6.601218
Assimetria	-2.3609317
Intervalo	9
Mínimo	1
Máximo	10
Soma	5633.7
Contagem	621

A informação pode também ser visualizada sob a forma de um gráfico na Figura 3.13.

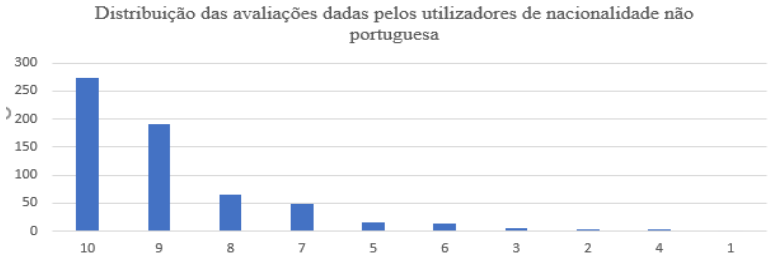


Figura 3.13 – Distribuição das Notas dadas pelos Utilizadores de nacionalidade não portuguesa

Da análise às avaliações, retiramos que 74,7% das vezes, os hóspedes dão uma nota superior a 9.

3.1.2.3 Avaliação geral

Esta variável não foi trabalhada pois inúmeras vezes não tem qualquer comentário disponível, e nos casos em que este existe é resumindo a uma palavra apenas, ou uma pequena frase, que normalmente também está presente no campo do comentário positivo, provocando alguma redundância nos dados.

3.1.2.4 Número de avaliações do utilizador

A variável em questão, tem um valor numérico, o qual representa o número de comentários que determinado utilizador já introduziu na plataforma. Desta forma conseguimos, de alguma forma, depreender que um utilizador com um maior número de avaliações é mais experiente e consequentemente, poderá ter um melhor termo de comparação ao avaliar um estabelecimento.

Tabela 3.4 – Estatística descritiva da variável Nº avaliações

Nº avaliações	
Média	9.377938034
Erro Padrão	0.105077189
Mediana	5
Moda	1
Desvio Padrão	12.85897534
Variância da Amostra	165.3532468
Curtose	33.91687412
Assimetria	4.178841512
Intervalo	236
Mínimo	1
Maximo	237
Soma	140444
Contagem	14976

Em termos estatísticos, visualizando a Tabela 3.4 pela análise descritiva, podemos verificar que em média cada utilizador tem 9 avaliações na plataforma *Booking.com.*

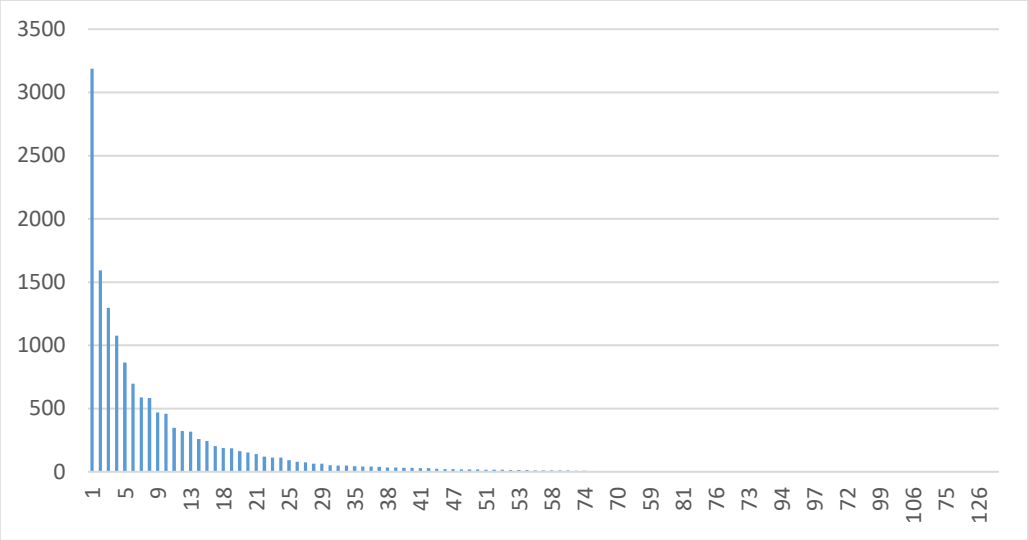


Figura 3.14 – Representação gráfica do número de avaliações de cada utilizador

No entanto, com a visualização gráfica da Figura 3.14 da distribuição de valores, percebemos que uma parte significativa dos utilizadores (8.020 ocorrências correspondendo a 53,6%) tem entre 1 e 5 avaliações.

3.1.2.5 Data da avaliação

Foi obtida a data da avaliação, no formato texto com o dia, mês e ano respeitante. Com esta informação conseguimos perceber a distribuição pelos vários anos. De notar que 50% das avaliações são respeitantes a 2020, o que é coerente com o aumento da procura por este tipo de sector específico de turismo/alojamento, em ano de pandemia. Na Tabela 3.5 podemos analisar a distribuição mais em pormenor.

Tabela 3.5 – Número de avaliações da amostra por ano

Ano	Nr. De avaliações	%
2018	2058	13.7%
2019	5198	34.7%
2020	7594	50.7%
2021	126	0.8%
Total	**14976**	**100.0%**

Podemos também representar a mesma sob a forma de um gráfico na Figura 3.15:

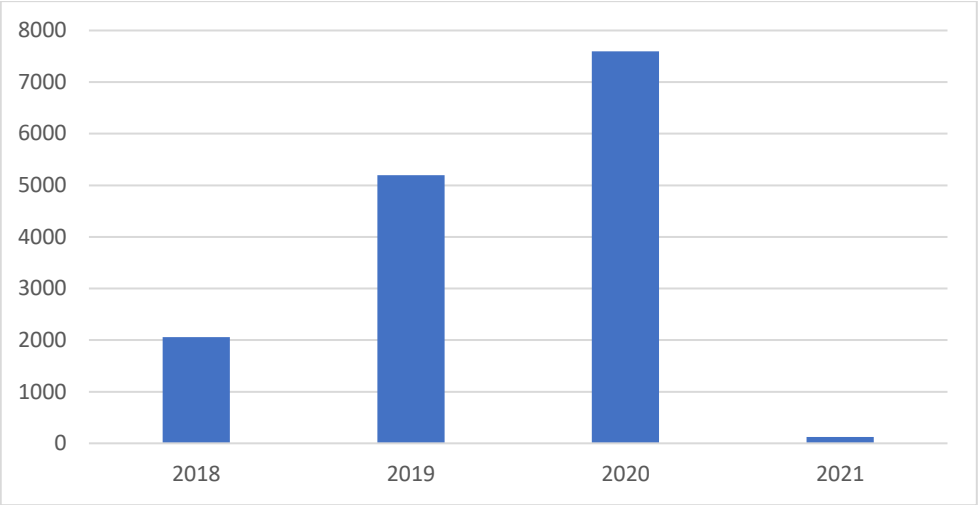


Figura 3.15 – Representação gráfica do número de avaliações da amostra por ano

3.1.2.6 Motivo da viagem

O motivo da viagem é uma variável que representa o contexto em que viagem aconteceu e originou a estadia em determinado alojamento. Na amostra recolhida existem 3 categorias: Viagem de lazer, Viagem de negócios e finalmente registos em que não é especificado o motivo da viagem. A respetiva distribuição dados pode ser visualizada na Tabela 3.6 e Figura 3.16.

Tabela 3.6 – Distribuição da variável Tipo de viagem

Tipo de Viagem	Nr. De viagens	%
Viagem de lazer	14247	95.1%
Viagem de negócios	554	3.7%
(em branco)	175	1.2%
Total	**14976**	**100.0%**

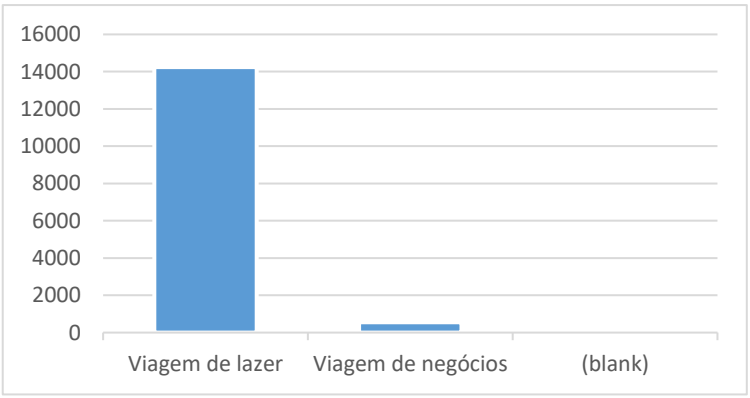


Figura 3.16 – Representação gráfica da distribuição da variável Tipo de viagem

Podemos verificar que a esmagadora das estadias (95,1%) foi realizada em lazer, por hóspedes possivelmente em período de férias ou fim de semana. No entanto, existem também ocorrências (3,7%) em que o hóspede escolhe um alojamento de turismo rural em situações profissionais.

3.1.2.7 Tipo de viajante

A variável determina o tipo de viajante que realizou a estadia num determinado alojamento. Na amostra recolhida foram obtidos 5 tipos de perfil cujas distribuições podem ser verificadas na Tabela 3.7 e Figura 3.17: Casal, Família com filhos mais novos, Grupo, Grupo de amigos e viajante individual.

Tabela 3.7 – Distribuição da variável Tipo de viajante

Tipo de Viajante	Nr. De ocorrências	%
Casal	7283	48.6%
Família com filhos mais novos	5383	35.9%
Grupo	936	6.3%
Grupo de amigos	633	4.2%
Viajante individual	741	4.9%
Total	**14976**	**100.0%**

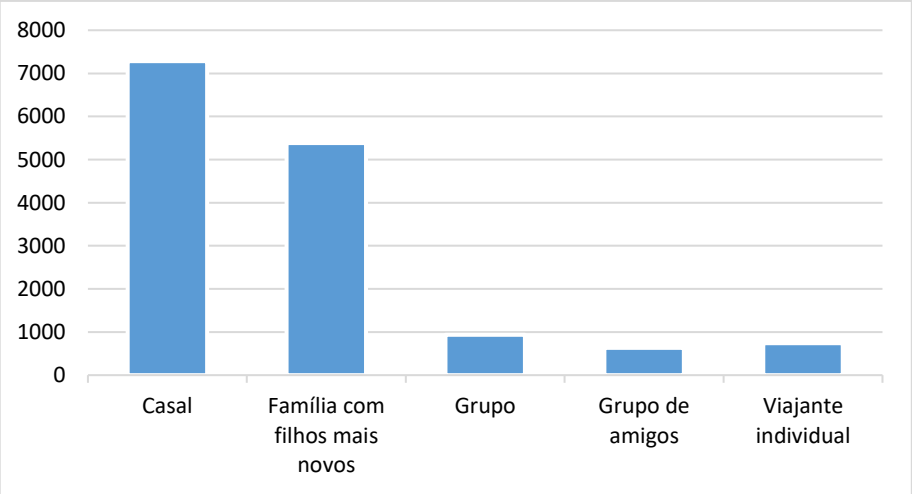


Figura 3.17 – Representação gráfica da distribuição da variável Tipo de viajante

Verificamos que neste caso, as viagens em casal e em família com filhos mais novos são as mais frequentes, com 84,5% das ocorrências.

3.1.2.8 Tipologia

A variável contém informação acerca da tipologia escolhida pelo utilizador durante o tempo de estadia. A distribuição pode ser analisada na Tabela 3.8 e Figura 3.18, onde

verificamos que os quartos, são normalmente a opção mais escolhida com 8.910 ocorrências (59,5%).

Tabela 3.8 – Distribuição da variável Tipologia

Tipologia	Nr. De viagens	%
Quartos	8910	59.5%
Casa	1841	12.3%
Apartamento	1307	8.7%
Suite	1237	8.3%
Estúdio	705	4.7%
Villa	486	3.2%
Bungalow	304	2.0%
Chalé	142	0.9%
Loft	16	0.1%
Cama em Dormitório	14	0.1%
Tendas	14	0.1%
Total	**14976**	**100.0%**

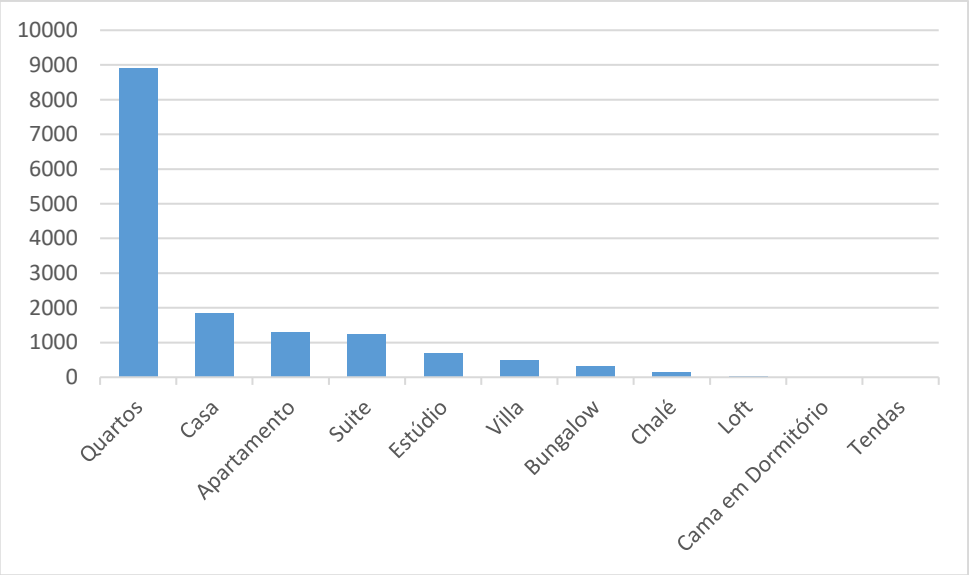


Figura 3.18 – Representação gráfica da distribuição da variável Tipologia

As restantes tipologias presentes na amostra são casas, apartamento, suite, estúdio, villa, bungalow, chalé, loft, cama em dormitório e tendas, que representam 40,5% das restantes ocorrências.

3.1.2.9 Número de noites

A variável mostra o número de noites que o utilizador tem na sua estadia. Nesta amostra podemos perceber pela análise estatística na Tabela 3.9 que, em média, os utilizadores

ficam cerca de 2 noites durante a sua estadia nos alojamentos.

Tabela 3.9 – Estatística descritiva da variável Nº noites

Nº Noites	
Média	2.130542201
Erro Padrão	0.011441431
Mediana	2
Moda	1
Desvio Padrão	1.400161927
Variância da Amostra	1.960453422
Curtose	5.521258668
Assimetria	1.958103634
Intervalo	13
Mínimo	1
Maximo	14
Soma	31907
Contagem	14976

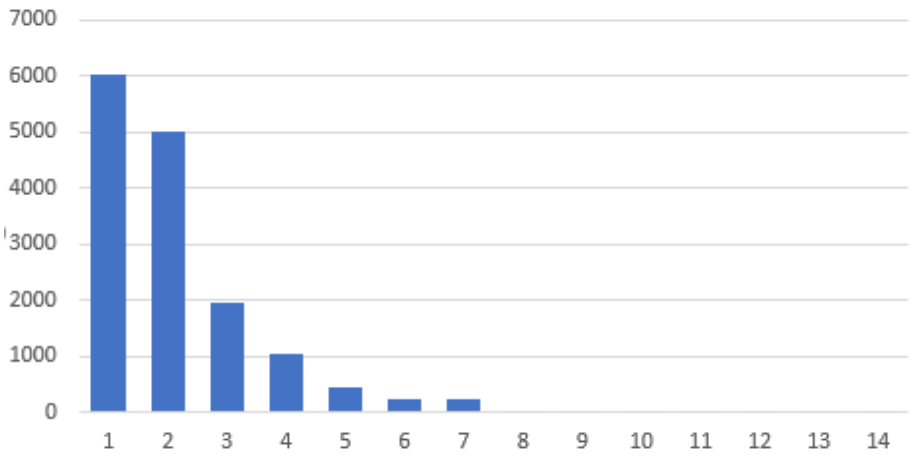


Figura 3.19 – Representação gráfica da distribuição do nº de noites

Podemos analisar a distribuição das noites em estadia também na Figura 3.19.

3.1.2.10 Comentário negativo relativo ao alojamento/experiência

O comentário negativo acerca da experiência vivida durante a estadia dispõe de informação bastante relevante para o trabalho proposto. Nesta amostra, constatou-se que, das 14.976 avaliações, 5.318 contêm texto com aspetos negativos, enquanto as restantes deixam esse campo em branco, conforme presente na Tabela 3.10 e Figura 3.20

Tabela 3.10 – Distribuição da variável comentário negativo

Existe comentário negativo	Contagem	%
Não	9658	64.5%
Sim	5318	35.5%
Total	**14976**	**100.0%**

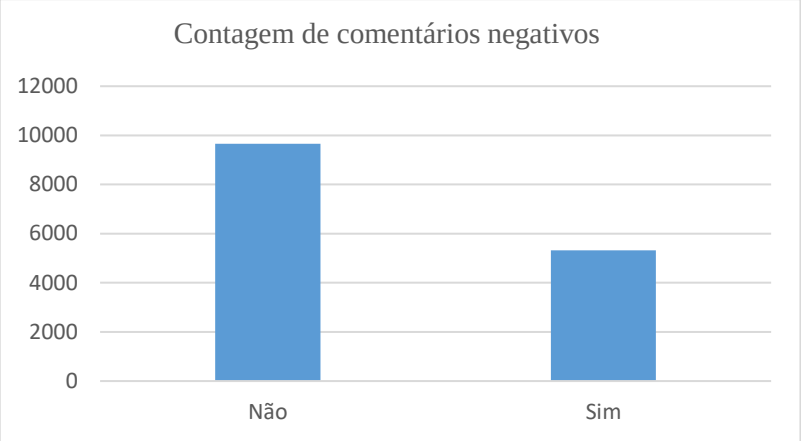


Figura 3.20 – Distribuição da variável comentário negativo quanto à existência de comentário

3.1.2.11 Comentário positivo relativo ao alojamento/experiência

O comentário positivo acerca da experiência vivida durante a estadia é igualmente um dos pontos chave de informação para o trabalho proposto. Nesta amostra, constatou-se que, das 14.976 avaliações, 9.939 contêm texto com aspetos positivos, enquanto as restantes deixam esse campo em branco, conforme podemos observar na Tabela 3.11 e Figura 3.21.

Tabela 3.11 – Distribuição da variável comentário positivo

Existe comentário positivo	Contagem	%
Não	5037	33.6%
Sim	9939	66.4%
Total	**14976**	**100.0%**

Constatamos que existe maís informação no que diz respeito a comentários positivos na amostra recolhida.

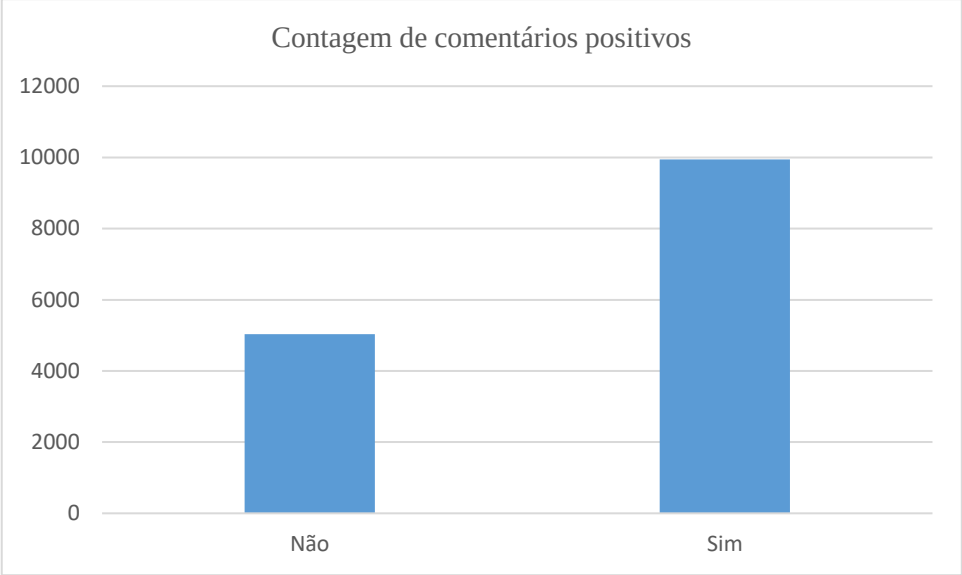


Figura 3.21 – Distribuição da variável comentário positivo quanto à existência de comentário

3.2 Preparação dos dados

Para algumas das colunas dos dados obtidos optou-se por efetuar algumas transformações no sentido de categorizar os dados lá contidos, assim como foram criadas novas colunas auxiliares. Assim foram feitas as seguintes alterações ao conjunto de dados da amostra:

- As notas dadas aos utilizadores no estabelecimento na plataforma Booking.com vão de 1 a 10 e a classificação final pode ter uma casa decimal. Para facilitar a análise deste indicador, foi truncada a parte decimal do número, ou seja, uma nota que tenha por exemplo 8.4 toma o valor de 8 após alteração;
- Criação de uma coluna com o nome "Restantes Países" que assume os termos "Portugal" ou "Outros" consoante o utilizador seja de nacionalidade portuguesa ou não;
- Criação de uma coluna com o nome "Ano" com o respetivo ano em que foi realizado o comentário com base na coluna com a data do mesmo;
- Dado o facto da existência de vários tipos de divisão que os hóspedes escolhem para a sua estadia, essa informação foi simplificada para uma mais fácil síntese e análise e agrupada em grandes grupos de "quartos", "casa", "apartamento", "estúdios", "suites", "villa", "bungalow", "chalé", "loft", "cama em dormitório" e "tendas".
- As variáveis "Tipo de submissão" e "Nome do utilizador" foram excluídas do estudo pois não foram consideradas relevantes.

Para o tratamento dos dados textuais foi utilizada a linguagem de programação R. O R é uma linguagem bastante utilizada na análise de dados, mais especificamente em análise estatística e para este efeito possui ferramentas de modelação linear e não linear, testes estatísticos, classificação e agrupamento de dados (*R-Project*, 2021).

Em primeiro lugar, após a extração dos dados foi necessário proceder ao pré-processamento dos mesmos. Para o efeito são normalmente utilizados os seguintes métodos conforme referido em Li et al. (2018):

- **Limpeza dos dados** – 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 (*Python Tutorialspoint*, 2021), palavras que não são o objetivo do estudo e também palavras que aparecem com pouca frequência.
- **Tokenização –** 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.

Para a limpeza dos dados foi utilizada a bibioteca *tm* existente para a linguagem R, também utilizada em Calheiros et al. (2017) pois possui várias funcionalidades específicas que permitem realizar tarefas de *Text-Mining* necessárias para analisar o texto e eliminar significativamente a dimensão dos dados, mantendo a informação relevante.

A limpeza dos dados foi realizada com as seguintes etapas:

- Transformação de caracteres em maiúsculas para minúsculas;
- Eliminação de caracteres numéricos;
- Eliminação de *Stopwords* de língua portuguesa;
- Eliminação de caracteres de pontuação gramática.
- Eliminação de espaços em branco em excesso.

Assim temos o seguinte exemplo de um dos comentários antes e depois da aplicação da limpeza de dados:

"Adorei tudo: o alojamento, as refeições espaço e a simpatia e amabilidade de todos os funcionários. As refeições excelentes . A repetir sem duvida."

Com a limpeza dos dados o comentário regista a seguinte informação para análise:

"adorei tudo alojamento refeições espaço simpatia amabilidade todos funcionários refeições excelentes repetir duvida"

3.3 Modelação

No que diz respeito à modelação e descoberta de padrões nos dados, foram aplicadas várias técnicas em várias fases distintas utilizando a linguagem de programação R de acordo com o fluxograma apresentado na Figura 3.9:

- Numa primeira fase foi apurada a frequência de palavras nos comentários de forma a conhecermos os termos mais mencionados de um modo geral. Esta informação foi apresentada sob a forma de tabela de frequência de termos e também sobre a forma de nuvem de palavras (*wordcloud*) para mais fácil interpretação. Um *wordcloud* é uma forma de apresentação visual de dados textuais, tipicamente utilizada de forma a encontrar as palavras-chave num conjunto de dados. A informação é apresentada de forma que, palavras com maior frequência apareçam mais proeminentemente, sendo uma forma bastante rápida de sumarizar grandes quantidades de dados (Ahuja & Shakeel, 2017). Para tal foi utilizada a biblioteca *Wordcloud*, especificamente criada para este efeito (ver código R no Apêndice 1);
- Foi criada uma *Document-term Matrix* que consiste numa matriz com duas dimensões em que as linhas são os documentos em análise (neste caso os comentários) e as colunas contêm os termos que aparecem em cada documento. Cada célula é preenchida com a frequência de cada termo em cada um dos documentos. Esta matriz é essencial em termos de *input* do modelo a testar (ver código R no Apêndice 2);
- Numa terceira fase procedemos à "modelação por tópicos" (*topic modelling*) de forma a extrair as dimensões gerais dos comentários obtidos. Optou-se por utilizar o modelo LDA (*Latent Dirichlet Allocation*) um dos modelos mais comuns para este tipo de situação. Para tal foi utilizada a biblioteca *TopicModels* e também a biblioteca *Tidytext* (ver código R no Apêndice 2);
- Numa quarta e última fase, foi utilizada a biblioteca *textmineR* para avaliar a coerência e prevalência dos tópicos positivos obtidos e a sua relação entre os

mesmos para melhor compreensão dos dados obtidos obtendo uma representação gráfica da coerência/prevalência e um dendrograma, respetivamente. Um dendrograma é um diagrama que representa a relação hierárquica em árvore entre objetos, neste caso os tópicos (ver código R no Apêndice 3).

3.3.1 Método *Latent Dirichlet Allocation*

Conforme referido por Guo et al. (2017), o algoritmo *Latent Dirichlet Allocation* (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. 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.

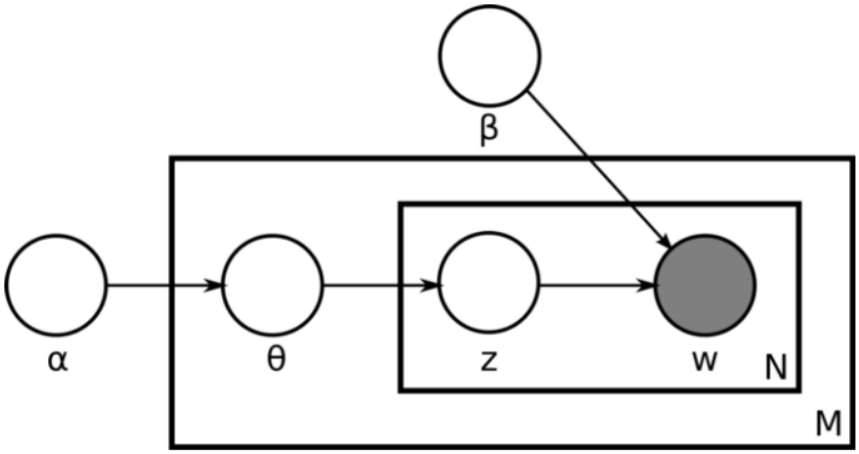


Figura 3.22 – Esquema do modelo LDA

Fonte: Christian, (2020)

O modelo apresentado na Figura 3.22 está na base do algoritmo LDA. Este assenta em vários parâmetros conforme descrito em Christian (2020):

- M representa o número de documentos, neste caso o número de comentários (a amostra tem M_i documentos);
- N representa o número de palavras em cada documento (cada documento tem N_i palavras);
- α (alfa) é um parâmetro específico do modelo que representa a distribuição de tópicos por documento;

- β (beta) indica a distribuição de palavras pelos tópicos criados. Um β alto indica que cada tópico contém uma maior quantidade de palavras do conjunto de dados;
- θ_i (theta), também por vezes representado por γ (gamma), representa a distribuição de cada tópico por cada documento;
- *z* representa o tópico para cada palavra num documento *m*;
- *w* representa cada uma das palavras.

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 o conjunto das palavras que formam determinado tópico. Existem três parâmetros muito importantes neste modelo aos quais deve ser dada especial atenção. Em primeiro lugar, o parâmetro k, indicador do número de tópicos a formar, em segundo lugar o parâmetro β (beta) que calcula a probabilidade de determinado termo estar associado a um tópico e o parâmetro γ (gamma), referido em Robinson et al. (2021), que devolve a percentagem de cada tópico associada a determinado documento. A este último parâmetro será dada uma especial atenção pois, 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 um dos comentários. Para a aplicação do modelo e cálculo dos parâmetros foi utilizado o método de Gibbs conforme efetuado por Guo et al. (2017).

Para a estimação do modelo são normalmente seguidos os seguintes passos referidos por Grün & Hornik (2011) :

1. Cálculo dos valores de distribuição de cada termo β para cada tópico;
2. Apuramento das proporções da distribuição de tópicos para cada documento *m*;
3. Para cada uma das palavras, escolha de um tópico *z* baseado na probabilidade condicionada $z_i: p(w_i|z_i, \beta)$ em que β é, como já foi referido acima, a representação da distribuição de termos pelos tópicos e contém a probabilidade de determinada palavra estar contida em determinado tópico.

3.3.2 Coerência, Prevalência e Relação entre os Tópicos

No sentido de avaliar a qualidade dos tópicos obtidos pela aplicação do Modelo LDA, é importante aferir sobre a coerência dos mesmos em relação aos termos obtidos, e se possível, a relação que os tópicos têm entre si. Tal é possível com aplicação da biblioteca de R *textmineR.*

A coerência entre os tópicos é definida em Rosner et al. (2014) como sendo o valor médio entre as similaridades dos termos mais frequentes em determinado tópico.

A prevalência, por sua vez, dá-nos, conforme dito em Christian (2020), a probabilidade da distribuição de determinado tópico por todos os documentos, ou seja, qual o tópico mais frequente no conjunto de dados obtidos.

Com base nestes dois valores, é possível ter uma base mais fidedigna no que diz respeito à verdadeira importância de cada tópico no conjunto de dados e é mais uma forma de validar os resultados obtidos da aplicação do método LDA.

A relação entre tópicos, que neste caso será visualizada sob a forma de um dendrograma, tem por base a métrica da distância de Hellinger se define em como sendo a distância entre dois vetores probabilísticos (EM, 2020). Este método oferece uma forma de agrupamento em que conseguimos de forma intuitiva analisar de que forma os tópicos se relacionam.

4 RESULTADOS

Os resultados obtidos destas duas abordagens, tanto a nível do *Wordcloud* como da aplicação do modelo LDA revelam alguma coerência. Foram formados *Wordclouds* para os comentários positivos e negativos, testando, com várias iterações, os vários valores que as variáveis obtidas apresentam, sob a forma de filtros para os comentários, com o intuito de detetar alterações expressivas nas palavras mais utilizadas.

4.1 Wordclouds

Em primeiro lugar, foram obtidas as tabelas de frequência e respetivos *Wordclouds* de todos os comentários. A Tabela 4.1 apresenta os 20 termos mais frequentes para os comentários positivos e negativos.

Tabela 4.1 – Os 20 termos mais frequentes para os comentários positivos e negativos

Negativos		**Positivos**	
Termos	**Frequência**	**Termos**	**Frequência**
nada	1314	pequeno	3297
pequeno	702	almoço	3211
casa	608	simpatia	3152
almoço	586	tudo	2628
ter	579	casa	2014
quarto	555	excelente	1912
piscina	364	piscina	1669
tudo	350	espaço	1639
apontar	267	localização	1371
pouco	258	vista	850
cama	254	agradável	834
limpeza	236	funcionários	773
poderia	221	limpeza	700
acesso	217	bem	692
espaço	211	alojamento	650
alojamento	196	quinta	650
gostei	191	estadia	650
melhor	190	ambiente	629
apenas	168	local	598
wifi	165	disponibilidade	580

A Figura 4.1 representa a distribuição na forma de nuvem de palavras para os comentários positivos.

Figura 4.1 – Wordcloud relativa aos comentários positivos

A Figura 4.2 representa a distribuição na forma de nuvem de palavras para os comentários negativos.

Figura 4.2 – Wordcloud relativa aos comentários negativos

4.2 Latent Dirichlet Allocation

Para a implementação do método LDA, como já foi referido acima, foi necessário apurar os valores de β (beta) que calcula a probabilidade de determinado termo estar associado a um tópico. Para tal foi utilizada a livraria *Tidytext*, como podemos verificar no Apêndice 2. Na Tabela 4.2 podemos visualizar uma amostra dos resultados obtidos utilizando esta técnica. A título de exemplo, temos que o termo "pequeno" e "almoço" tem cerca de 17,0% e 16,6% respetivamente, de probabilidade de estar associado ao tópico 5.

Tabela 4.2 – Amostra de valores Beta para vários termos presentes na amostra

Posição	Tópico	Termo	Beta
1	2	simpatia	0,173746834
2	5	pequeno	0,169974498
3	5	almoço	0,166061191
4	4	tudo	0,129548673
5	9	casa	0,118766202
6	5	bom	0,114079427
7	9	bem	0,104266444
8	10	espaço	0,103694860
9	10	piscina	0,099864005
10	7	localização	0,088811139

Com recurso ao modelo LDA foi possível identificar dez tópicos distintos que caracterizam os comentários positivos da amostra obtida e também 10 tópicos associados a comentários negativos submetidos pelos utilizadores. Os tópicos podem ser visualizados na Tabela 4.3 e podem ser analisados mais em pormenor nos Apêndices 4 e 5 deste trabalho. O processo de identificação dos tópicos foi, à semelhança do efetuado por Guo et al. (2017), baseado na conexão lógica existente entre os termos que compõem os tópicos produzidos.

O número de tópicos produzido para os comentários é um parâmetro que foi calculado correndo várias iterações do algoritmo, e ajustado analisando os resultados obtidos e se os termos mais relevantes se repetem de forma substancial nos vários tópicos produzidos, abordagem semelhante ao realizado por Calheiros et al. (2017).

Tabela 4.3 – Tópicos obtidos para os comentários positivos e negativos

Número	Nome do tópico para comentários positivos	Nome do tópico para comentários negativos
1	Serviço Geral	Pequeno-almoço
2	Simpatia	Casa de banho
3	Divisões da propriedade	Localização
4	Sentimento do cliente	Rede/Wi-Fi
5	Pequeno-almoço	Temperatura da Água
6	Ambiente relaxante	Espaço exterior
7	Localização	Conforto do Quarto
8	Atividades	Pouco tempo de estadia
9	Anfitriões	Cama
10	Espaço exterior	Sentimento do Cliente

Os tópicos encontrados indicam de uma forma relativamente simples os aspetos mais importantes utilizando os termos em cada documento utilizado como amostra, como podemos verificar nos exemplos expostos na Figura 4.3.

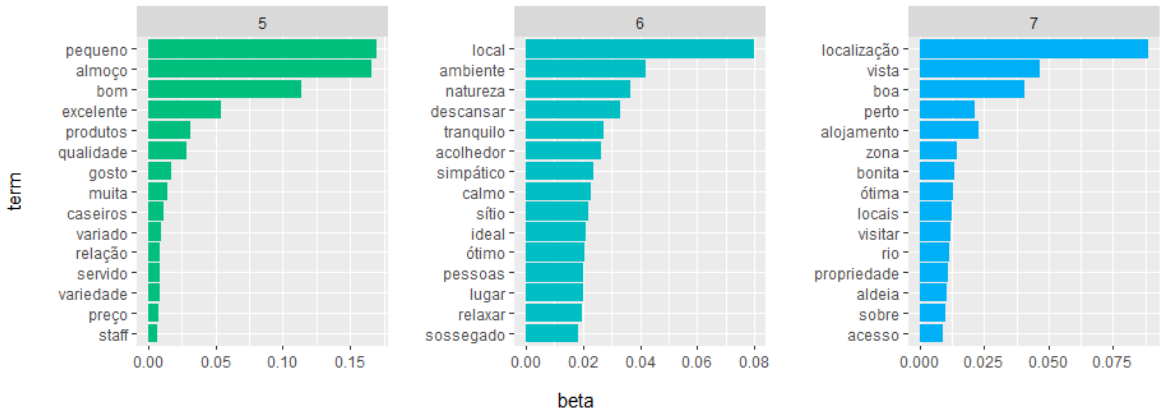


Figura 4.3 – Exemplo dos tópicos obtidos para pequeno-almoço, ambiente relaxante e localização

No primeiro exemplo temos o tópico obtido para os termos relacionados com o pequeno-almoço, com vários termos que caracterizam os produtos à disposição. De seguida, o termo obtido para o ambiente relaxante realça os aspetos sentidos pelos hóspedes durante a estadia e que deram origem aos comentários com características como a natureza, tranquilidade e ambiente calmo e sossegado. No último exemplo é dado destaque à localização do alojamento e a vista e zona onde a propriedade está inserida.

4.3 Coerência e Prevalência

No que diz respeito à aplicação da metodologia para cálculo e visualização destes indicadores foi necessário criar um novo conjunto de tópicos pois a livraria *textmineR,* cujo código base pode ser consultado no Apêndice 3. é independente da biblioteca *TopicModels* e assim necessita de uma nova iteração do modelo com a função *FitLdaModel()* em que é obrigatória a criação de uma nova DTM em que podem ser dados parâmetros pertencentes ao modelo anterior para assim conseguir resultados semelhantes aos dados com a biblioteca *TopicModels.* É importante referir o facto pois os tópicos criados via este método foram ligeiramente diferentes, mas ainda assim coerentes com os resultados anteriores.

Assim, foram obtidos, a título de exemplo, os seguintes dez tópicos apenas para os comentários positivos, que podemos visualizar na Figura 4.4.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_10
1	decoração	tudo	simpatia	espaço	localização	local	gostei	casa	pequeno	quinta
2	atendimento	simpatia	natureza	piscina	boa	funcionários	onde	bem	almoço	alojamento
3	local	estadia	sossego	agradável	perto	super	melhor	vista	excelente	animais
4	instalações	sempre	disponibilidade	quarto	simpática	simpáticos	estar	confortável	bom	pessoas
5	simpático	voltar	tranquilidade	staff	disponível	descansar	ser	piscina	produtos	paisagem
6	gosto	todos	acolhimento	envolvente	dona	tranquilo	ambiente	zona	qualidade	tranquilidade
7	bom	conforto	amabilidade	exterior	locais	ambiente	ter	dias	pão	excelentes
8	cama	recomendo	calma	limpeza	boas	proprietários	dia	limpo	muita	condições
9	sossegado	toda	paz	bastante	visitar	calmo	jantar	quarto	caseiros	fantástica
10	pessoal	repetir	silêncio	grande	rio	sítio	fazer	todas	variedade	donos
11	atencioso	desde	meio	quartos	prestável	ideal	ainda	uns	variado	sentir
12	extremamente	fantástico	proprietária	pequenoalmoço	sobre	ótimo	todo	bonita	delicioso	crianças
13	comodidade	lugar	contacto	espaços	bons	relaxar	familiar	comodidades	caseiro	propriedade
14	acolhedor	dúvida	quer	geral	proprietário	pequenoalmoço	quartos	recebidos	manhã	anfitriã
15	vila	maravilhoso	beleza	serviço	fica	acolhedor	tempo	descanso	café	óptima

Figura 4.4 – Tópicos obtidos para comentários positivos com recurso à biblioteca do R textmineR

De seguida, e adaptando a metodologia utilizada em Christian, (2020), foi testada a coerência e prevalência para os tópicos obtidos. O critério para a aplicação deste método foi o de aferir sobre a qualidade dos tópicos obtidos e da sua relação entre os mesmos e a sua importância no conjunto de dados como um todo.

A coerência, para este efeito, dá-nos o grau de semelhança semântica entre as palavras com maior relevância em cada tópico, permitindo assim perceber quais são os tópicos mais coesos e coerentes do conjunto de dados referente aos comentários positivos. Os resultados, implementando o algoritmo em R para o efeito, estão representados na Figura 4.5.

```
> mod_lda_5$coherence
       t_1        t_2        t_3        t_4        t_5        t_6        t_7        t_8        t_9       t_10
0.01828791 0.03938261 0.02879401 0.05344346 0.04313137 0.04854804 0.03854525 0.06575611 0.14371741 0.02595299
> |
```

Figura 4.5 – Resultados para a coerência dos tópicos obtidos com recurso à biblioteca textmineR

A prevalência por outro lado, devolve o tópico mais comum no conjunto de dados, ou seja, a probabilidade da distribuição de determinado tópico pelo conjunto dos dados. Os resultados, implementando também o algoritmo em R, estão representados na Figura 4.6

```
> mod_lda_5$prevalence
      t_1       t_2       t_3       t_4       t_5       t_6       t_7       t_8       t_9      t_10
 9.239901 10.925783  9.403791  9.944246  9.822643  9.710042 10.363240 10.361175 10.367112  9.862066
> |
```

Figura 4.6 – Resultados para a prevalência dos tópicos obtidos com recurso à biblioteca textmineR

Os resultados obtidos podem também ser visualizados de forma gráfica na Figura 4.7.

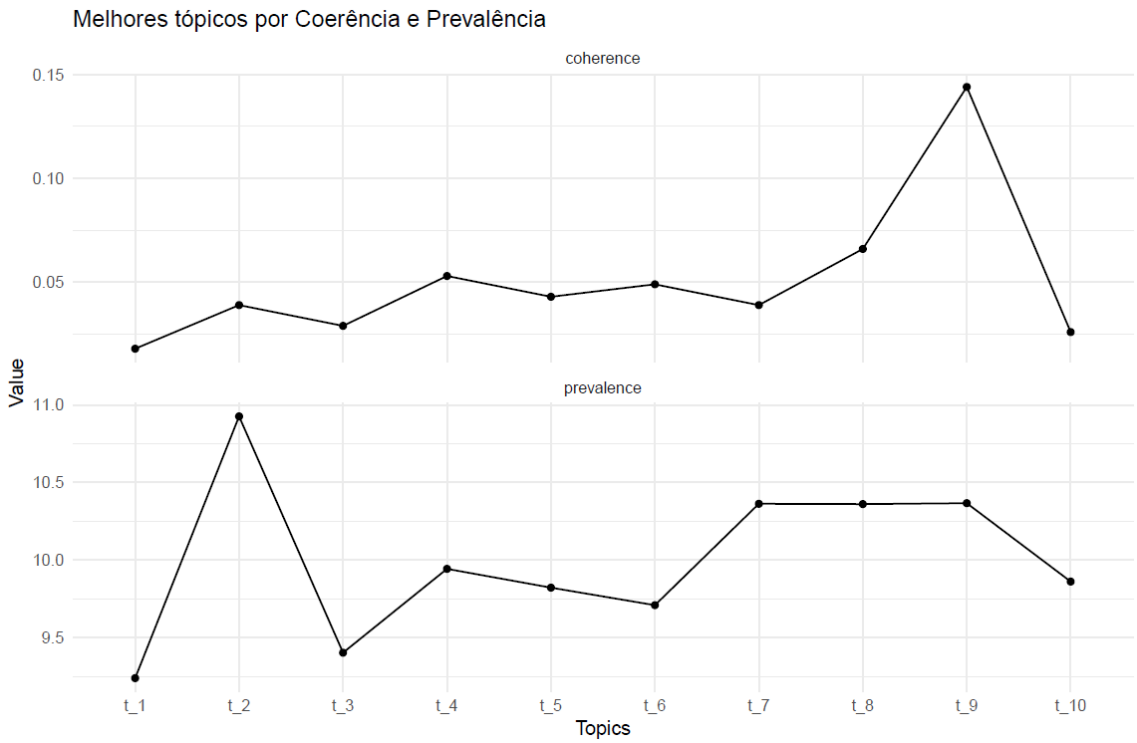


Figura 4.7 – Análise gráfica da coerência e prevalência dos tópicos criados

Podemos observar que o tópico 9 formado por termos relacionados com o pequeno-almoço apresenta a maior coerência, enquanto, em termos de prevalência, é o tópico 2 o que tem o valor mais elevado. Estes resultados serão discutidos mais em pormenor no capítulo de discussão de resultados.

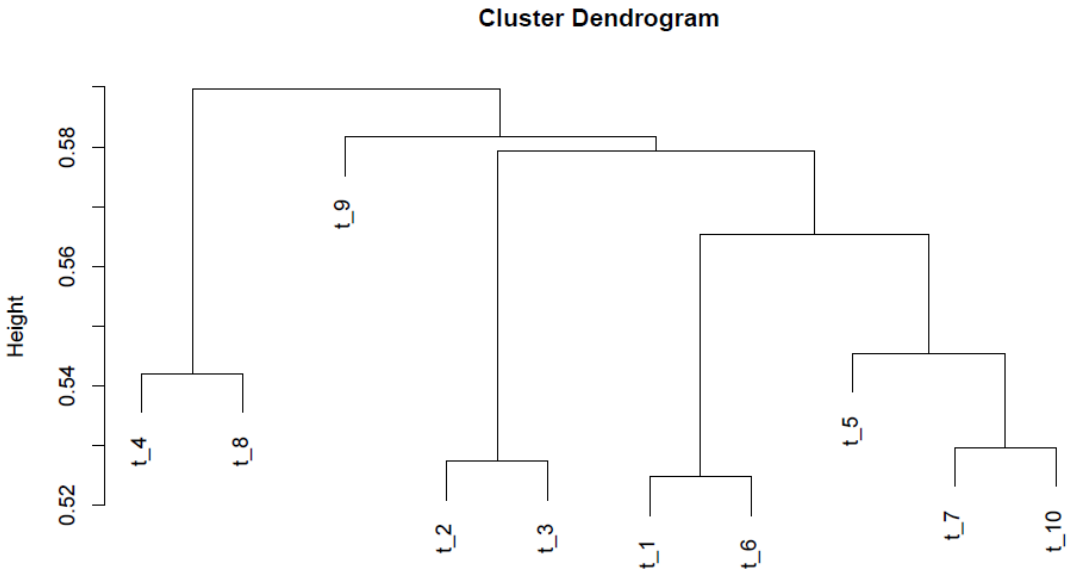


Figura 4.8 – Dendrograma referente aos tópicos criados.

Foi também obtido na Figura 4.8, adaptando a metodologia já descrita acima para o contexto deste trabalho, o dendrograma que estabelece as relações que o modelo encontra entre os tópicos obtidos, matéria que também será discutida no capítulo dedicado à discussão de resultados.

4.4 Aplicação de Wordclouds e LDA a subconjuntos de dados

Tendo em conta a qualidade dos dados obtidos e as variáveis obtidas, foram também realizados testes utilizando a metodologia aplicada com *wordclouds* e respetivo modelo LDA, assumindo diferentes valores para essas variáveis presentes na base de dados, com o intuito de tentar encontrar diferenças nas opiniões dos utilizadores e a eventual descoberta de outros padrões para além da visão geral obtida utilizando todos os comentários, positivos ou negativos.

Assim, foram consideradas vários critérios de seleção para os comentários positivos e negativos com as variáveis a assumir diferentes valores. Temos, portanto, os seguintes critérios para os quais foram selecionados os subconjuntos de dados:

- Nota atribuída pelo utilizador de 1 a 10;
- Tipo de alojamento: Casa de campo, Hotel Rural, Agroturismo e Turismo de Aldeia;
- Ano do comentário: 2018, 2019 e 2020;

- Tipo de viajante: Casal, Família com filhos mais novos, grupo ou grupo de amigos e viajante individual;
- Se o alojamento se situa em Portugal continental ou ilhas;
- Utilizador experiente (com maior número de comentários no *Booking.com*, neste caso assumiu-se um utilizador experiente como tendo 30 comentários ou superior);
- Viagem em contexto de lazer ou negócios;
- Tempo de estadia de uma noite ou mais do que uma noite;
- Estadia em quartos ou outras divisões que não quartos;
- Comentários de utilizadores de outros países que não Portugal.

4.4.1 Subamostra dos comentários positivos

Numa perspetiva mais específica ao nível do tipo de estabelecimento, as Casas de Campo e Hotéis Rurais revelam tópicos semelhantes ao panorama geral dos comentários positivos já analisado. No entanto, em alojamentos de Agroturismo existem ligeiras diferenças aliadas ao facto de existir uma ligação ao espaço exterior mais associado a atividades relacionadas com animais da quinta e entretenimento para as crianças, algo que pode ser analisado Figura *4.9* abaixo.

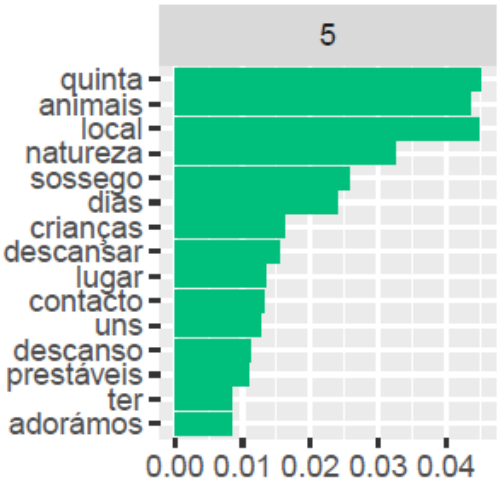


Figura 4.9 – *Tópico criado referente a estabelecimentos de Agroturismo*

Ao analisar os comentários tendo em conta a nota que receberam do utilizador (de 1 a 10) não foram encontradas diferenças tanto nos *wordclouds* como na aplicação do modelo LDA quando comparados com o quadro geral de todos os comentários. Este resultado pode estar relacionado com o facto de a amostra ser, na sua grande maioria, bastante

positiva enviesando assim os resultados, pois para notas menos boas existem menos comentários, tornando a amostra testada demasiado pequena para devolver resultados mais consistentes.

No que diz respeito a comentários positivos nos diferentes anos analisados, também não foram encontradas diferenças notórias nos termos utilizados pelos utilizadores. Termos como "simpatia", "pequeno-almoço", "localização", "piscina" e "casa" têm um enorme destaque, não se notando uma tendência diferente da visão geral que já existia com todos os comentários.

Em termos de diferenças notadas entre estabelecimentos localizados nas regiões autónomas dos Açores e Madeira e Portugal continental, há a registar a ocorrência de termos mais específicos e relacionados com as características geográficas da zona como "vista", "ilha", "mar" e "ananás", como podemos verificar na Figura 4.10.

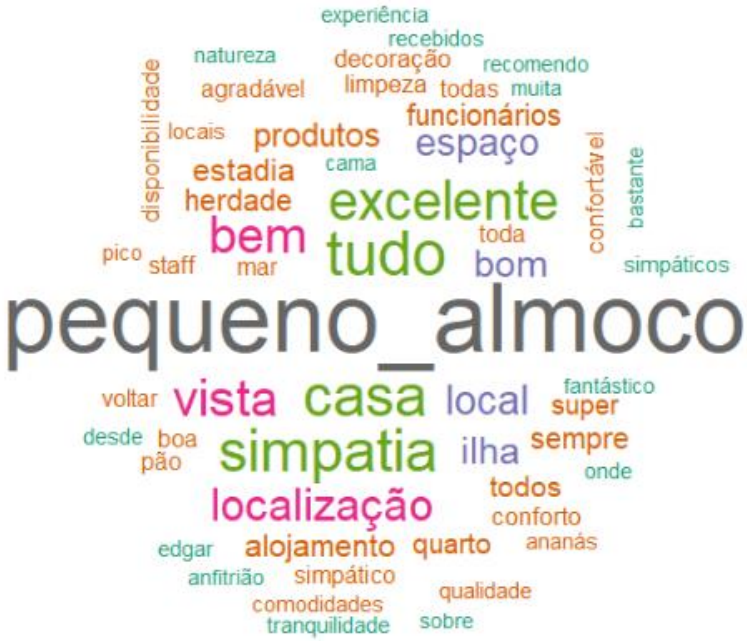


Figura 4.10 – Wordcloud para comentários positivos relativos a alojamentos em regiões autónomas

No modelo LDA extraído para estes estabelecimentos, os tópicos com mais peso e que são mais distintos do retrato geral, estão relacionados com a vistas nas várias ilhas, nomeadamente vistas marinhas e para a ilha do Pico nos açores e a localização do alojamento aliado à sua qualidade como podemos verificar na Figura 4.11.

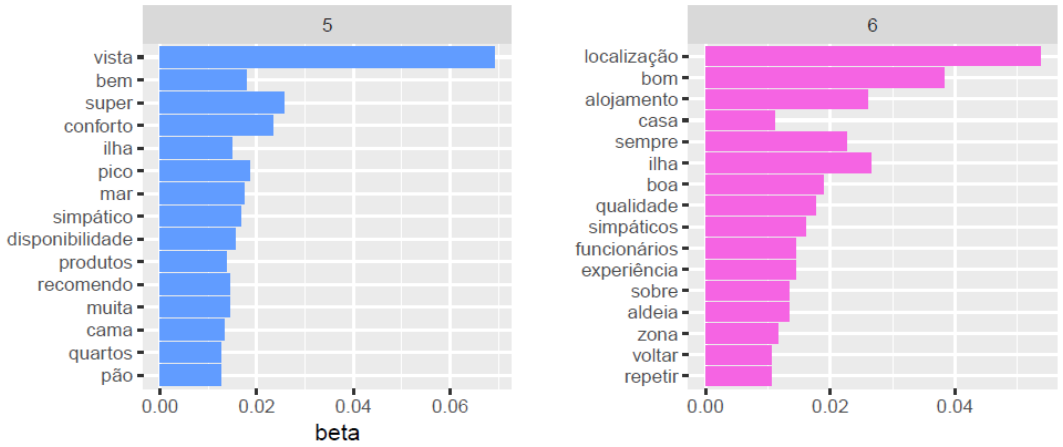


Figura 4.11 – Exemplos de Tópicos obtidos para comentários positivos relativos a alojamentos em regiões autónomas

Quando analisamos o tipo de utilizador no que diz respeito ao contexto social em que se insere a viagem, e constatando na análise volumétrica dos dados que grande parte do tipo de viajante aparece sobre a forma de casais e família com filhos mais velhos (84,5% dos casos), não existem, nestes dois casos específicos, grandes diferenças relativamente ao quadro geral de resultados no que diz respeito à frequência de palavras.

Já o modelo LDA revela tópicos ligeiramente diferentes e que no caso da família com filhos mais novos dão mais destaque ao espaço exterior e ao facto do local ser convidativo ao descanso e propício ao divertimento das crianças. Para hóspedes em contexto de viagem de negócios ou lazer os *wordclouds* foram coerentes com o panorama geral de resultados com todos os comentários.

No que diz respeito às diferentes divisões ou tipologias que os hóspedes escolhem, como já foi dito acima foram testados os dados relativamente a estadias em quartos e estadias em outras divisões que não quartos. Nos *wordclouds* não foram encontradas diferenças notórias. No modelo LDA, nomeadamente para hóspedes que ficaram em diferentes divisões que não quartos, existiu mais destaque para a localização num contexto de ligação a elementos geográficos como a vista, praia e rio, como podemos verificar na Figura 4.12.

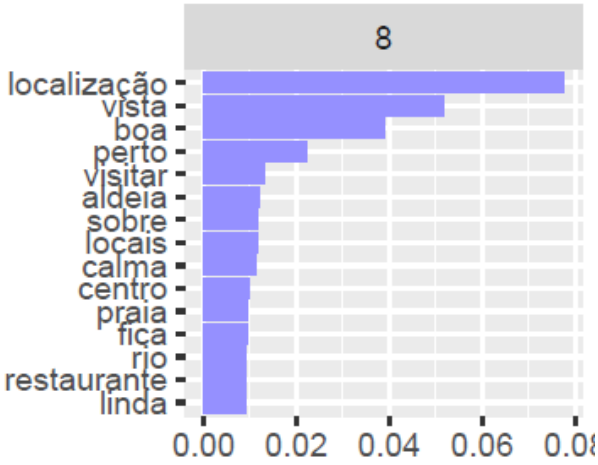


Figura 4.12 – Tópico originado de comentários de hóspedes com estadia em outras divisões que não quartos

Não foram encontradas diferenças relevantes quando foram comparados os subconjuntos com hóspedes que ficam mais do que uma noite em determinado alojamento e aqueles que ficaram apenas uma noite.

Finalmente, analisando os comentários de hóspedes considerados para este estudo como experientes (mais de 30 comentários submetidos) é de realçar o facto dos tópicos obtidos refletirem uma crítica objetiva à experiência obtida e não tanto o sentimento positivo que por vezes é expresso em restantes comentários.

4.4.2 Subamostra dos comentários negativos

Analisando os diferentes *wordclouds* obtidos tendo em conta os diferentes tipos de estabelecimento, o termo que se destaca mais proeminentemente é, uma vez mais, o termo "nada", sendo que os restantes termos são uniformes com a visão geral dos comentários negativos obtidos. Os modelos LDA para as casas de campo e hotéis rurais são semelhantes ao panorama geral da amostra, enquanto, os alojamentos de agroturismo e turismo de aldeia revelaram menos tópicos obtidos e de menor consistência. No que diz respeito às diferentes notas dadas, existe um maior foco nas críticas à limpeza nas notas abaixo de 7 conforme podemos visualizar na Figura 4.13 e foram também produzidos menos tópicos gerados pelo LDA, que focam os aspetos negativos no pequeno-almoço, quartos, limpeza e piscina do alojamento.

Figura 4.13 – Wordcloud para notas abaixo de 7

No que diz respeito a comentários negativos nos diferentes anos analisados, não foram encontradas diferenças notórias nos termos utilizados pelos utilizadores. O mesmo se aplica quando analisamos os resultados em utilizadores considerados experientes e não experientes ou entre alojamentos situados em Portugal continental ou nas regiões autónomas.

Existem algumas diferenças nos modelos LDA quanto ao contexto social em que a estadia se insere. Casais parecem dar mais importância ao desconforto da cama, o que não acontece quando falamos de viajantes em grupo de amigos. Viajantes individuais parecem dar mais importância nos comentários negativos a aspetos do quarto no geral, como podemos verificar na Figura 4.14.

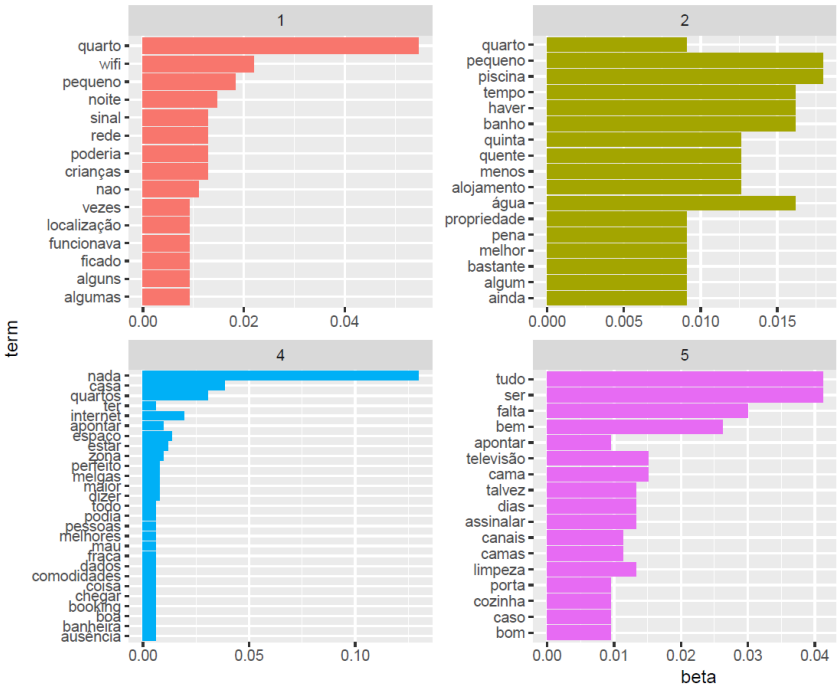


Figura 4.14 – Tópicos gerados relativamente a comentários negativos por parte de viajantes individuais

As opiniões negativas quando testadas para hóspedes em contexto de viagem de negócios ou lazer, hóspedes que ficam em quartos ou hóspedes que escolhem outro tipo de tipologias, assim como hóspedes oriundos de outros países e hóspedes que ficam uma noite ou mais do que uma noite foram coerentes com o panorama geral de resultados com todos os comentários tanto nos *wordclouds* como no modelo LDA extraído para o efeito.

5 DISCUSSÃO DOS RESULTADOS

5.1 Visão Geral

Na análise ao *Wordcloud*, é dado 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. Não foram notadas diferenças significativas nos termos face às notas dadas, já que os comentários são na sua grande maioria fortemente positivos, como pode ser verificado na média de notas dadas de 8,99 (de 1 a 10).

Por outro lado, em termos negativos, as palavras mais utilizados para expressar desagrado na experiência no alojamento foram centrados também no pequeno-almoço, limpeza, casa de banho e dificuldades de acesso, rede e wifi. De notar que em muitas situações, o utilizador refere no espaço para o comentário negativo que não tem nada a apontar ou que não encontrou nada que mereça comentário negativo durante a sua estadia, o que explica a razão pela qual a palavra "nada" é o termo mais utilizado. Esta situação vai de novo de encontro um pouco à amostra obtida que como já foi dito acima é maioritariamente positiva.

Neste contexto o facto de a fonte de dados ser da plataforma *Booking.com* parece ser acertada pois permite que seja possível isolar termos negativos e positivos independentemente da nota dada pelo utilizador. No entanto há que referir o facto de existir uma quantidade considerável de utilizadores que na sua avaliação refere apenas que adorou ou ficou extremamente satisfeito com a estadia, não dando motivos para tal, traduzindo apenas o sentimento do cliente. Esta é uma situação que acontece tanto no campo dos comentários positivos como negativos, daí existirem dois tópicos idênticos com denominação de sentimento de cliente para os dois tipos de comentários. O facto dos termos "nada", "tudo" e "apontar" terem destaque nos *wordclouds* é uma evidência desse

facto no dicionário de termos nos comentários negativos. É também de referir parece existir evidência de que este tipo de avaliações é mais comum em utilizadores menos experientes e com menos avaliações submetidas na plataforma *booking.com.*

Estes resultados esclarecem as áreas a investir ou melhorar neste tipo de estabelecimentos. 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.

5.2 Subamostras

É também interessante notar que analisando os subconjuntos de dados da amostra, existem maiores variações relativamente ao panorama geral de resultados nos comentários positivos quando comparado com os comentários negativos. Parece existir evidência de que os aspetos negativos são algo generalizados não existindo enormes diferenças, pelo menos no que diz respeito à amostra obtida. Existem, no entanto, alguns aspetos de realçar. Para as piores notas dadas pelos utilizadores, o pequeno-almoço, limpeza, condições do quarto e piscina têm especial destaque. Este facto é pertinente, já que, para os tópicos negativos gerais foram identificados aspetos negativos como a casa de banho e o espaço exterior. Tendo as duas informações como referência, parece existir a sugestão que, associado ao caso particular da falta de limpeza do alojamento, está mais associada à divisão da casa de banho, enquanto no caso do espaço exterior, as piores críticas são mais apontadas à piscina.

Ainda no prisma dos comentários negativos, quanto ao contexto social em que se insere a viagem, a preocupação que os casais e viajantes individuais têm em relação ao conforto da cama e quarto respetivamente, vem talvez com a intenção de passar uma estadia mais relaxante e recuperadora. Viajantes individuais em particular, revelam preocupações

como a falta de rede e localização, casa de banho e temperatura da água e conforto da cama. Grupos de amigos valorizam mais o espaço exterior não estando tão preocupados com o conforto do quarto em si.

No que diz respeito também aos subconjuntos analisados, é de referir que no caso dos comentários positivos existe alguma importância ao contexto geográfico em que o alojamento de turismo rural está situado, pois no caso das regiões autónimas existem referências tanto à proximidade marítima como aos produtos regionais como o ananás. Ainda relativamente aos comentários positivos, o pequeno-almoço revela ser um tópico bastante relevante em todos os estabelecimentos e comum nos aspetos positivos a referir nas avaliações dadas.

Ao analisar os comentários submetidos por hóspedes oriundos de outros países, é de referir a importância dada à experiência vivida no alojamento com especial atenção à decoração e à forma como são recebidos com realce para a zona norte do país e perto do rio Douro. A explicação para este facto reside na natureza dos dados da amostra, em que uma fatia considerável dos comentários surgirem de alojamentos próximos do rio Douro (32,4%).

Existem também ligeiras alterações no que diz respeito nos termos que formam o tópico no caso de avaliações dadas por grupos de amigos que podem estar relacionadas com o intuito da estadia em si não estar tão relacionada com a experiência rural e gastronómica, mas sim mais associada a divertimento e convívio em grupo.

5.3 Percentagem de cada tópico nos documentos

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 5.1* podemos verificar os pesos relativos de cada um dos dez tópicos presentes para os primeiros dez documentos correspondentes a comentários positivos da amostra, com recurso ao algoritmo criado em R com a biblioteca *Topicmodels*. 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.

Tabela 5.1 – Percentagem de cada tópico em cada um dos primeiros dez comentários

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%

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%).

5.4 Coerência, Prevalência e Relação entre os Tópicos

Analisando a informação obtida referente à coerência e prevalência dos dados obtidos referentes aos comentários positivos, com recurso à biblioteca *textmineR*, é possível retirar informação pertinente e que nos dá uma maior compreensão acerca dos dados e tópicos obtidos aplicando o Método LDA. Com base nos resultados obtidos na Figura 4.4, o Tópico 9, que apresenta termos relacionados com o pequeno-almoço é, de todos, o que tem maior coerência e portanto, maior qualidade, o que quer dizer que os termos incluídos estão mais associados entre si. Para este tópico em particular existe também uma prevalência substancial quando comparada com os restantes tópicos, pelo que, podemos concluir que é razoavelmente provável a existência de um comentário em que sejam empregues combinações entre os termos que constituem o tópico. Esta é uma conclusão algo esperada, dado que os termos do tópico 9 são muito ligados ao pequeno-almoço e também aos adjetivos utilizados para descrever o mesmo e os produtos que foram postos à disposição.

No que diz respeito à prevalência, o Tópico 2 é o que apresenta um maior valor. Este indica que é o tópico com maior probabilidade de existir no conjunto de comentários positivos obtidos. O tópico 2 é formado essencialmente com termos que traduzem o sentimento expresso pelo utilizador relativamente à experiência obtida, bastante presentes na amostra obtida, pelo que faz sentido o resultado obtido observando os resultados anteriores, nomeadamente nos *wordclouds* obtidos. Desta feita, é mais provável que um comentário analisado seja formado por termos constantes neste tópico.

Por último da análise que podemos tirar acerca do dendrograma obtido, parece existir a sugestão de que os tópicos 4 e 8 relacionados com aspetos como a casa e espaço exterior estão mais relacionados entre si. A conclusão faz sentido pois assenta numa das características do turismo rural que é a preocupação com o espaço em si mesmo, numa tentativa de oferecer ao hóspede uma experiência mais autêntica e a preocupação com o detalhe, decoração e arquitetura, muitas vezes de acordo com os costumes da região.

6 CONCLUSÃO

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 que revela maiores comentários positivos e negativos 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 e criticado quando a oferta de produtos é limitada ou não é dada atenção à qualidade dos mesmos. 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. Este facto parece estar em linha com uma das dimensões estudadas por Park & Yoon (2011) no sentido de obter indicadores de gestão válidos para este sector, nomeadamente no que diz respeito às instalações.

Os aspetos negativos como a Limpeza de casa de Banho, Privacidade, Temperatura e condições dos quartos e cama são também tópicos a ter em conta para garantir a satisfação dos hóspedes, já que estes são referidos com frequência. É importante que os órgãos de

gestão estejam sensíveis à informação produzida desta forma e com base num fluxo diário e quase imediato, já que, se os dados forem tratados é possível retirar informação atempada de pertinente para a tomada de decisão de forma eficiente e eficaz. É também importante ter presente, como já foi referido acima que embora o fator da envolvência rural e o contacto com a natureza de um determinado alojamento sejam aspetos positivos retirados da análise efetuada, deve existir um cuidado acrescentado ao acesso do alojamento e à qualidade de Wi-Fi/rede telemóvel do mesmo, pois demasiado isolamento, particularmente a nível tecnológico tem tendência a refletir-se de forma negativa nos comentários dos utilizadores.

6.1 Principais contributos

O tema tratado é atual e pertinente, o que justifica por si só a realização do estudo. O setor turístico é um setor de mudança, um setor adaptável e, na maioria dos casos, sempre em busca da satisfação do cliente, o que leva a questões constantes sobre como atingir os objetivos a que os órgãos de gestão se propõem.

O principal contributo deste estudo reside no facto de fornecer informação válida e de fácil interpretação para eventuais órgãos de gestão, independentemente da sua formação ou habilitações. Visualmente os resultados são na sua grande maioria intuitivos e a metodologia aplicada é de fácil aplicação, depois de desenvolvidos os algoritmos, o que permite um fornecimento dessa informação quase imediato, caso existam dados para analisar.

O facto de ser uma base de dados que fornece informação de alojamentos de todo o território português (continental e também regiões autónomas) dá confiança aos resultados numa perspetiva geral, mas permite também, uma análise mais aprofundada sobre determinado aspeto ou tipo de cliente, região do país ou tipo de alojamento.

A nível académico pode ser também significativo o facto de a revisão de literatura efetuada estar sintetizada num quadro resumo facilmente acessível para análise, já que existe uma grande quantidade de artigos nesta área e pode tornar-se por vezes complicado arranjar um ponto de partida para abordar uma temática semelhante.

6.2 Limitações

A metodologia utilizada e os métodos de *text-mining* estão sempre em constante evolução e atualização. Logo, existem várias abordagens também possíveis de ter com os dados à disposição que não foram realizadas, pelo facto de ser necessário criar um foco no trabalho a realizar e um caminho a escolher para analisar os dados. Pela revisão de literatura realizada, são referidas várias metodologias também possíveis de se implementar no estudo realizado.

A base de dados obtida foi apenas referente a estabelecimentos de turismo rural, não existindo um paralelismo com outros tipos de segmento turístico para podermos realizar comparações entre eles.

Para a análise à coerência, prevalência e relação entre tópicos, foi utilizada a biblioteca *textmineR* do R. Reforça-se a ideia de que para a utilização dessas ferramentas é necessário a criação de novos tópicos, que são ligeiramente diferentes dos dez tópicos iniciais criados a partir da biblioteca *TopicModels* e que serviram de base à análise geral dos comentários. No entanto as diferenças encontradas, não são críticas no que diz respeito à interpretação dos tópicos obtidos, funcionando até como uma forma de validação dos resultados iniciais.

O facto deste estudo ter sido realizado numa altura de pandemia, pode ter tido influência no interesse por este tipo de turismo, o que, num ano considerado "normal", pode não se verificar a 100%, pese embora a importância das questões ambientais e de alterações climáticas que podem dar força ao crescimento deste tipo de opções.

6.3 Trabalhos futuros

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.

Seria interessante comparar esta análise a outros segmentos de alojamento turístico com diferentes perfis de clientes analisados mais em pormenor.

A ferramenta Microsoft PowerBI, que foi utilizada exclusivamente para a localização geográfica dos estabelecimentos, tem um enorme potencial para a visualização de dados. Existindo recurso a uma atualização constante de dados turísticos, seria interessante a criação de uma plataforma exclusivamente criada para o setor do turismo rural, pela facilidade de criação de *dashboards* inerente à aplicação.

6.4 Considerações finais

Por fim, referir que embora a plataforma *Booking.com* tenha determinados aspetos com avaliações específicas para a generalidade dos estabelecimentos disponíveis para reserva, e que são na maior parte dos casos coerentes também com os resultados encontrados, no caso específico do turismo rural, parece existir evidência de aspetos específicos a considerar na avaliação de alojamentos que praticam este tipo de turismo. Esta opinião não é, no entanto, sem algumas reservas pois não foi efetuada a análise a outro tipo de estabelecimentos para tirar conclusões mais precisas.

Durante a execução deste estudo existiu a perceção de que existe uma vasta quantidade de metodologias aplicadas nesta área de estudos, em que, algumas assentam em abordagens preditivas enquanto outras assentam em abordagens mais descritivas. Neste trabalho foi utilizada uma abordagem descritiva.

O estudo efetuado foi bastante útil para aprofundar um pouco os conhecimentos em R e *Python*, linguagens de programação bastante utilizadas atualmente. A principal razão para esse facto foi a especificidade de algumas bibliotecas mais especializadas para certas áreas. A linguagem *Python* também dispõe de bibliotecas dedicadas à análise de dados textuais pelo que, teria sido possível realizar toda a análise nessa linguagem. A escolha pela linguagem R foi uma opção pessoal, pois achei a mesma mais intuitiva no que diz respeito à análise efetuada.

O tema escolhido é uma área que me interessa particularmente assim como a área da mineração de texto. Esta temática tem potencial para criar valor e informação válida para uma entidade com um custo financeiro associado bastante baixo para as entidades patronais, já que, grande parte das tecnologias normalmente aplicadas estão disponíveis no mercado a título gratuito.

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

Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2019). Big data adoption: State of the art and research challenges. *Information Processing and Management*, *56*(6). https://doi.org/10.1016/j.ipm.2019.102095

Banerjee, S., & Chua, A. Y. K. (2016). In search of patterns among travellers' hotel ratings in TripAdvisor. *Tourism Management*, *53*, 125–131. https://doi.org/10.1016/j.tourman.2015.09.020

Beautiful Soup 4.9.0 documentation. (2021, jul 24). Retrieved from https://www.crummy.com/software/BeautifulSoup/bs4/doc/

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

Booking.com. (2020). *Comentários Booking.com - comentários reais de hotéis, por hóspedes reais*. https://www.booking.com/reviews.pt-pt.html?label=gen173nr-1DEgdyZXZpZXdzKIICOOgHSDNYBGi7AYgBAZgBH7gBF8gBDNgBA-gBAYgCAagCA7gCkoGHiwbAAgHSAiQxOWExMWIzMC00MzY2LTQzNzktYTVkZi04ZTg3NGRiNmRlZTnYAgTgAgE;sid=60fc173f07df4371a8d77907ac42dec1;keep_landing=1&page=0&

Booking.com Announces Milestone of Five Million Reported Listings in Homes, Apartments and Other Unique Places to Stay. (2020, set 3). Retrieved from https://globalnews.booking.com/bookingcom-announces-milestone-of-five-million-reported-listings-in-homes-apartments-and-other-unique-places-to-stay/

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

Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment Classification of Consumer-

Generated Online Reviews Using Topic Modeling. *Journal of Hospitality Marketing and Management*, *26*(7), 675–693. https://doi.org/10.1080/19368623.2017.1310075

Campos, A. C., Mendes, J., do Valle, P. O., & Scott, N. (2018). Co-creation of tourist experiences: A literature review. *Current Issues in Tourism*, *21*(4), 369–400. https://doi.org/10.1080/13683500.2015.1081158

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

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*, 51–89. https://doi.org/10.1002/aris.1440370103

Christian, J. (2020, jul 29). *Topic Modeling in R With tidytext and textmineR Package*. Retrieved from https://medium.com/swlh/topic-modeling-in-r-with-tidytext-and-textminer-package-latent-dirichlet-allocation-764f4483be73

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.

Data Visualization | Microsoft Power BI. (2020, dec 12). Retrieved from https://powerbi.microsoft.com/en-us/

DGADR. (2020a, ago 17). *Características do Turismo no Espaço Rural*. Direção Geral de Agricultura e Desenvolvimento Rural. Retrieved from https://www.dgadr.gov.pt/diversificacao/turismo-rural/caracteristicas-do-turismo-no-espaco-rural

DGADR. (2020b, ago 17). *O Interesse pelo Turismo no Espaço Rural*. Direção Geral de Agricultura e Desenvolvimento Rural. Retrieved from

https://www.dgadr.gov.pt/diversificacao/turismo-rural/o-interesse-pelo-turismo-no-espaco-rural

Direção Geral da Saúde. (2020, dec 15). *Perguntas Frequentes Categoria - COVID-19*. Retrieved from https://covid19.min-saude.pt/category/perguntas-frequentes/

Economia, M. da. (2017). Estratégia 2027. *Estratégia 2027*, 66. http://estrategia.turismodeportugal.pt/sites/default/files/Estrategia_Turismo_Portugal_ET2027.pdf

EM. (2020, Oct 23). *Hellinger distance - Encyclopedia of Mathematics*. Retrieved from https://encyclopediaofmath.org/index.php?title=Hellinger_distance

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

Fang, B., Ye, Q., Kucukusta, D., & Law, R. (2016). Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management*, *52*, 498–506. https://doi.org/10.1016/j.tourman.2015.07.018

Feldman, R., Fresko, M., Hirsh, H., Aumann, Y., & Liphstat, O. (1998). *Knowledge Management : A Text Mining Approach . Knowledge Management : A Text Mining Approach*. *May 2014*.

Google.pt. (2020, dec 16). *fontes portugal coordinates - Google Search*. Retrieved from https://www.google.com/search?q=fontes+portugal+coordinates&oq=fontes&aqs=chrome.1.69i57j35i39j0i512l8.4786j0j4&sourceid=chrome&ie=UTF-8

Google Trends. (2020, dec 31). *turismo rural - Explore - Google Trends*. Retrieved from https://trends.google.pt/trends/explore?q=turismo rural&geo=PT

Gössling, S., & Lane, B. (2015). Rural tourism and the development of Internet-based accommodation booking platforms: a study in the advantages, dangers and implications of innovation. *Journal of Sustainable Tourism*, *23*(8–9), 1386–1403. https://doi.org/10.1080/09669582.2014.909448

Grün, B., & Hornik, K. (2011). Topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30. https://doi.org/10.18637/jss.v040.i13

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

Hendrickx, T., Cule, B., Meysman, P., Naulaerts, S., Laukens, K., & Goethals, B. (2015). Mining association rules in graphs based on frequent cohesive itemsets. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9078*(3), 637–648. https://doi.org/10.1007/978-3-319-18032-8_50

Hou, Z., Cui, F., Meng, Y., Lian, T., & Yu, C. (2019). Opinion mining from online travel reviews: A comparative analysis of Chinese major OTAs using semantic association analysis. *Tourism Management*, *74*(March), 276–289. https://doi.org/10.1016/j.tourman.2019.03.009

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

IPDT. (2020, dec 15). *PESQUISAS POR TURISMO RURAL EM PORTUGAL ATINGIRAM PICO DURANTE O MÊS DE MAIO - IPDT*. PESQUISAS POR TURISMO RURAL EM PORTUGAL ATINGIRAM PICO DURANTE O MÊS DE MAIO - IPDT. Retrieved from https://www.ipdt.pt/turismo-rural-pico-maio-covid/

Kastenholz, E., Eusébio, C., Figueiredo, E., Carneiro, M. J., & Lima, J. (2014). Cocriação de experiências turísticas sustentáveis. In *Reinventar o turismo rural em Portugal: Cocriação de experiencias túristicas sustentáveis* (Vol. 1).

Khorsand, R., Rafiee, M., & Kayvanfar, V. (2020). Insights into TripAdvisor's online reviews: The case of Tehran's hotels. *Tourism Management Perspectives*, *34*(August 2019), 100673. https://doi.org/10.1016/j.tmp.2020.100673

Kim, K., Park, O. joung, Yun, S., & Yun, H. (2017). What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management. *Technological Forecasting and Social Change*, *123*, 362–369. https://doi.org/10.1016/j.techfore.2017.01.001

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

Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, *19*(4), 456–474. https://doi.org/10.1287/isre.1070.0154

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.

Pitchayadejanant, K., & Nakpathom, P. (2018). Data mining approach for arranging and clustering the agro-tourism activities in orchard. *Kasetsart Journal of Social*

Sciences, *39*(3), 407–413. https://doi.org/10.1016/j.kjss.2017.07.004

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

PORDATA - Estabelecimentos de turismo de habitação e de turismo no espaço rural: total e por tipo de estabelecimento. (2021, jul 24). Retrieved from https://www.pordata.pt/Portugal/Estabelecimentos+de+turismo+de+habitação+e+de+turismo+no+espaço+rural+total+e+por+tipo+de+estabelecimento-2607

Prakash, S. L., Perera, P., Newsome, D., Kusuminda, T., & Walker, O. (2019). Reasons for visitor dissatisfaction with wildlife tourism experiences at highly visited national parks in Sri Lanka. *Journal of Outdoor Recreation and Tourism*, *25*(October 2017), 102–112. https://doi.org/10.1016/j.jort.2018.07.004

Python - Remove Stopwords - Tutorialspoint. (2021, mar 20). Retrieved from https://www.tutorialspoint.com/python_text_processing/python_remove_stopwords.htm

R-Project. (2021, abr 1). *R: What is R?*. Retrieved from https://www.r-project.org/about.html

Robinson, D., & Silge, J. (2021, mai 19). *6 Topic modeling | Text Mining with R*. Retrieved from https://www.tidytextmining.com/topicmodeling.html

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014). *Evaluating topic coherence measures*. 1–4. http://arxiv.org/abs/1403.6397

Ruhanen, L. (2019). The prominence of eco in ecotourism experiences: An analysis of post-purchase online reviews. *Journal of Hospitality and Tourism Management*, *39*(October 2018), 110–116. https://doi.org/10.1016/j.jhtm.2019.03.006

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.

Sun, Y., & Shao, Y. (2020). Measuring visitor satisfaction toward peri-urban green and open spaces based on social media data. *Urban Forestry and Urban Greening*, *53*(May), 126709. https://doi.org/10.1016/j.ufug.2020.126709

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

TP. (2017). *Turismo no Espaço Rural e Turismo de Habitação em Portugal | 2017*.

TP. (2020, jan 1). *Visão geral*. Retrieved from http://www.turismodeportugal.pt/pt/Turismo_Portugal/visao_geral/Paginas/default.aspx

TravelBI. (2020, dez 31). *Estatísticas Hóspedes*. Retrieved from https://travelbi.turismodeportugal.pt/pt-pt/Paginas/PowerBI/hospedes.aspx

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. (2019). Retrieved 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

Wang, J. (2020). *Sentiment Analysis & Topic Modeling for Hotel Reviews*. https://medium.com/swlh/sentiment-analysis-topic-modeling-for-hotel-reviews-6b83653f5b08

Web Scraping Tool & Free Web Crawlers | Octoparse. (2020, set 30). Retrieved from

https://www.octoparse.com/

Welcome to Python.org. (2021, fev 1). Retrieved from https://www.python.org/

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

APÊNDICES

APÊNDICE 1. Código em R para tratamento dos dados e Wordclouds

```
library(caret)
library(tm)
library(SnowballC)
library(data.table)
library(tidyr)
library(tidytext)
library(wordcloud)
library(reshape2)
library(textstem)
library(topicmodels)
library(Rmpfr)
library(LDAvis)
library(stringi)
library(e1071)
library(text2vec)
library(ggrepel)
library(Rtsne)
library(tibble)
library(scales)
library(ldatuning)

# leitura de ficheiro CSV
data <- fread("nome do ficheiro.csv")
df <- data.frame(data)
df <- df[complete.cases(df),]


#exemplos de variações para filtro de dados a analisar
#tipo_estabelecimento <- filter(data, Tipo_de_estabelecimento == "Hotel Rural")
#tipo_viajante <- filter(data, Viajante == "Grupo")
#nota_atribuida <- filter(data, Segmento_de_notas >= 8)
#nota_atribuida2 <- filter(data, Segmento_de_notas %in% c(8,9))
```

- **Limpeza de dados**

```
text_corpus <- VCorpus(VectorSource(data$review_text))

print(text_corpus)

text_corpus_clean <- tm_map(text_corpus,
                            content_transformer(tolower))
text_corpus_clean <- tm_map(text_corpus_clean, removeNumbers)
text_corpus_clean <- tm_map(text_corpus_clean,
                            removeWords, stopwords("portuguese"))
```

```
text_corpus_clean <- tm_map(text_corpus_clean, removePunctuation)
text_corpus_clean<-tm_map(text_corpus_clean, content_transformer(function(x)


text_corpus_clean <- tm_map(text_corpus_clean, stripWhitespace)
print(text_corpus_clean)

dtm <- DocumentTermMatrix(text_corpus_clean)
dtm <- as.matrix(dtm)
```

- **Wordcloud**

```
library(wordcloud)

wordcloud(text_corpus_clean, min.freq = 100, random.order = FALSE, width = 1600,
height = 1600,
        colors=brewer.pal(8, "Dark2"))
```

APÊNDICE 2. Código em R para modelação LDA com as bibliotecas *Topicmodels* e *Tidytext*

- **Modelo LDA**

```
text_dtm <- DocumentTermMatrix(text_corpus_clean)
text_dtm

findFreqTerms(text_dtm, lowfreq = 20) ## encontra termos que aparecem menos de 20 vezes

rowTotals <- apply(text_dtm , 1, sum) #soma as palavras em cada documento
text_dtm.new   <- text_dtm[rowTotals> 0, ]    #remove todos os documentos sem termos

text_lda <- LDA(text_dtm.new, k = 10, control = list(seed = 1234), method = "Gibbs")
text_lda


library(tidytext)

text_topics <- tidy(text_lda, matrix = "beta")


library(ggplot2)
library(dplyr)

text_top_terms <- text_topics %>%
  group_by(topic) %>%
  top_n(15, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

text_top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```

APÊNDICE 3. Código em R para modelação LDA com a biblioteca *textmineR*

Modelação LDA com a biblioteca TextmineR()

```
library(textminerR)


m <- Matrix::sparseMatrix(i=text_dtm.new$i,
                          j=text_dtm.new$j,
                          x=text_dtm.new$v,
                          dims=c(text_dtm.new$nrow, text_dtm.new$ncol),
                          dimnames = text_dtm.new$dimnames)

str(m)

set.seed(1234)
mod_lda <- FitLdaModel(dtm = m,
                       k = 10, # number of topic
                       iterations = 300,
                       #burnin = 180,
                       alpha = 6,beta = 0.05,
                       optimize_alpha = T,
                       calc_coherence = T)


mod_lda$top_terms <- GetTopTerms(phi = mod_lda$phi,M = 15)
data.frame(mod_lda$top_terms)
```

- **Coerência**

```
mod_lda$coherence
```

- **Prevalência**

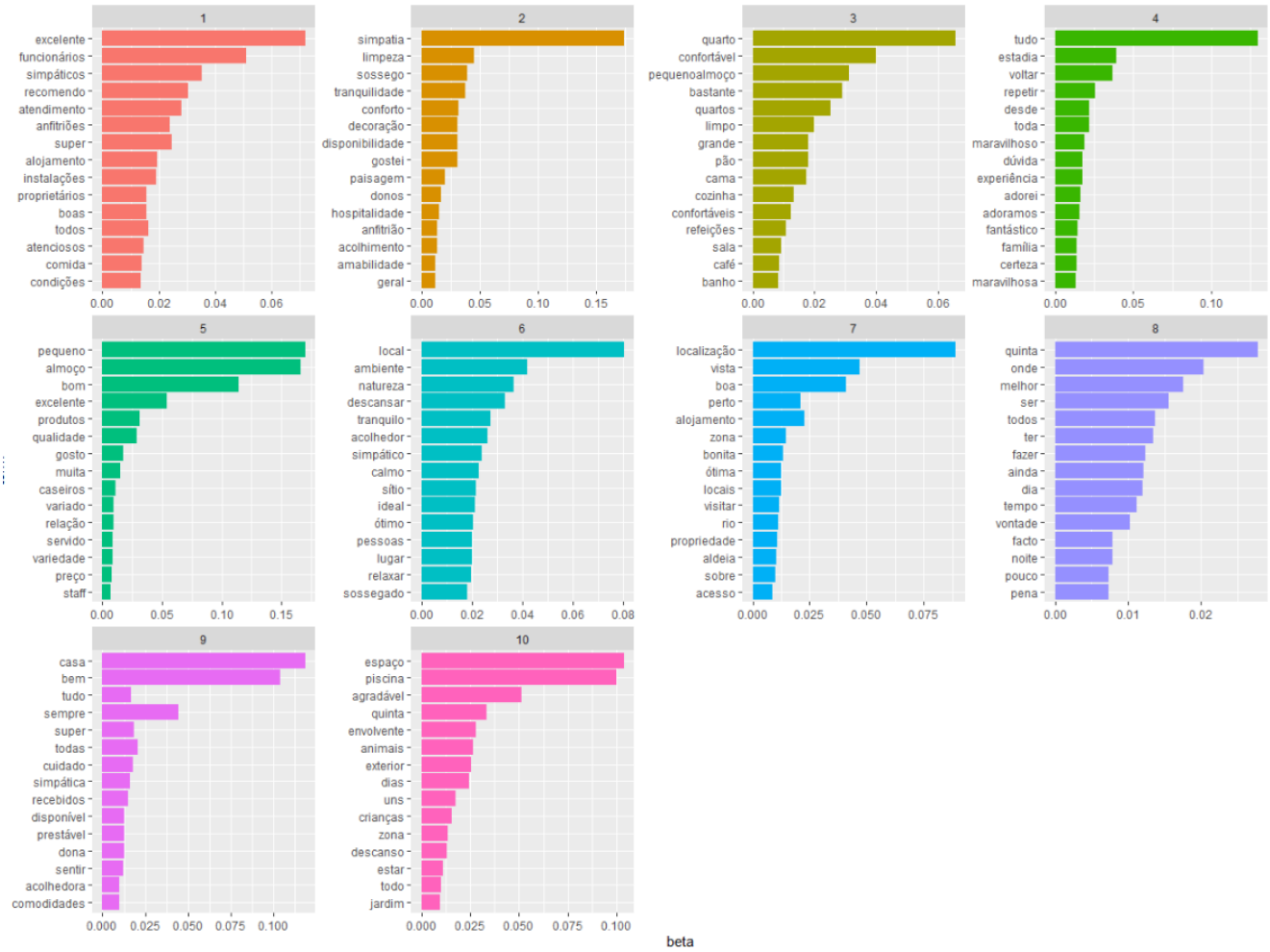
```
mod_lda$prevalence <- colSums(mod_lda$theta)/sum(mod_lda$theta)*100
mod_lda_$prevalence

mod_lda$summary <- data.frame(topic = rownames(mod_lda$phi),
                              coherence = round(mod_lda$coherence,3),
                              prevalence = round(mod_lda$prevalence,3),
                              top_terms =
apply(mod_lda$top_terms,2,function(x){paste(x,collapse = ", ")}))

modsum_5 <- mod_lda$summary %>%
  `rownames<-`(NULL)
```

```
modsum_5 %>% pivot_longer(cols = c(coherence,prevalence)) %>%
  ggplot(aes(x = factor(topic,levels = unique(topic)), y = value, group = 1)) +
  geom_point() + geom_line() +
  facet_wrap(~name,scales = "free_y",nrow = 2) +
  theme_minimal() +
  labs(title = "Melhores tópicos por Coerência e Prevalência",
       x = "Topics", y = "Value")
```

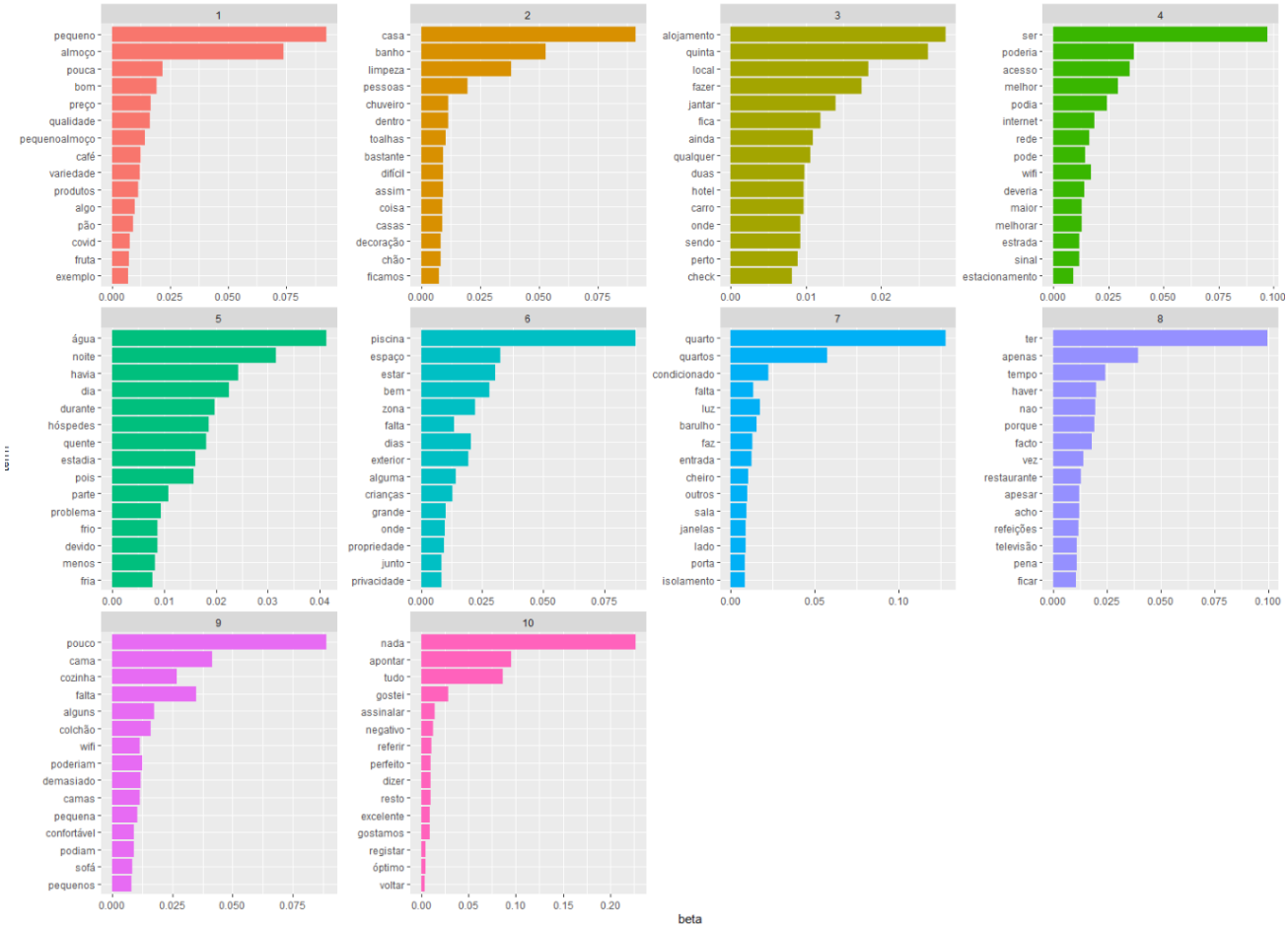
- **Dendrograma**

```
mod_lda$linguistic <- CalcHellingerDist(mod_lda$phi)
mod_lda$hclust <- hclust(as.dist(mod_lda$linguistic),"ward.D")
mod_lda$hclust$labels <- paste(mod_lda$hclust$labels, mod_lda$labels[,1])
plot(mod_lda$hclust)
```

APÊNDICE 4 - Tópicos extraídos para comentários positivos



APÊNDICE 5 - Tópicos extraídos para comentários negativos



APÊNDICE 6 – Quadro síntese da revisão de Literatura

Ano	Autor(es)	Tema	Amostra	Métodos Utilizados	Resultados/Conclusões
2019	Sánchez-Franco et al. (2019)	Identificação de temos relacionados com a experiência dos hóspedes no sentido de serem utilizados para melhorar o seu serviço a nível de hospitalidade em hóteis de Las Vegas	47.172 comentários da plataforma *Yelp.com*	Naive Bayes	Evidência de que hotéis com bons serviços ligados ao jogo de apostas, quartos com comodidades apreciadas pelos hóspedes e entrega dos funcionários ao serviço têm mais tendência para gerar melhores reviews.
2020	Khorsand et al. (2020)	Prever uma avaliação de um hotel na cidade de Teerão apenas com base na informação do hotel e do utilizador	4.718 comentários de 64 hotéis listados na plataforma *TripAdvisor.com*	K-nearest neighbors Naïve Bayes Árvore de decisão regressão logística Support Vector Machine Neural Network Random Forest Gradient Boosting Data Base sensitivity	O estudo concluiu que 77% dos fatores importantes analisados são relacionados com os hóspedes e apenas os restantes são relacionados com os estabelecimentos em si. São identificados várias áreas e comodidades chave para garantir a satisfação dos utilizadores de modo a obter melhores avaliações entre as quais Wi-Fi, bar, restaurante, estrelas do hotel, serviço de quartos, staff multilingue, limpeza a seco e sauna.
2017	Gao et al. (2017)	Identificação das dimensões chave referentes à satisfação com o serviço hoteleiro	266.544 comentários de utilizadores em 25.670 hotéis de 16 países na plataforma *TripAdvisor.com*	Latent dirichlet analysis (LDA)	O estudo em questão enumerou um conjunto de 30 tópicos associados à satisfação dos utilizadores em conjunto com as 20 palavras mais mencionadas em cada um dos tópicos. Foi também analisada a componente controlada ou não controlada do universo de tópicos existentes e ficou evidente que existem diferenças a nível demográfico na satisfação dos utilizadores.
2021	Luo et al. (2021)	Analisar comentários turísticos obtidos na visita a 24 Geoparques que fazem parte do património mundial da 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	120.532 comentários acerca de 24 Geoparques obtidos das principais comunidades turísticas online na China	Support Vector Machine Versão modificada/melhorada do algoritmo Latent dirichlet analysis (LDA)	Os resultados revelaram 10 atributos importantes e significativos para avaliação e talvez mais importante ainda revelaram alguns atributos avaliados negativamente pelos turistas sendo os seguintes o custo da viagem, serviços da viagem, conhecimento da matéria e transporte/alojamento.
2019	Taecharungroj & Mathayomchan (2019)	Analisar as opiniões com base em *reviews* obtidas online de forma a fornecer informação para a gestão no sentido de melhorar as suas atrações, sendo neste caso num contexto de atrações turísticas em Phuket na Tailândia	65.079 comentários da plataforma *TripAdvisor.com*	Naïve Bayes e LDA	Foi possível definir vários atributos referentes aos diferentes cenários turísticos e responder a questões como perceber quantas reviews estão incluídas em cada atributo, o quão positivos são os atributos encontrados, quão frequentes e positivos são os termos encontrados em cada atributo e qual o grau de precisão que é possível obter na previsão dos comentários dos utilizadores.
2017	Pokryshevskaya & Antipov (2017)	Prever a satisfação dos hóspedes em relação aos serviços disponibilizados a nível de hotelaria	3.630 comentários da plataforma *Booking.com*	Modelo de regressão linear	O estudo conseguiu identificar os perfis com maior probabilidade de ficarem com uma opinião negativa, de forma a ser alvo de maior atenção por parte da gerência e também os perfis com maior probabilidade de terem uma opinião positiva acerca dos serviços disponibilizados e que, portanto, não requerem uma atenção extra.

Ano	Autor(es)	Tema	Amostra	Métodos Utilizados	Resultados/Conclusões
2019	Cheng & Jin (2019)	Perceber quais os aspetos que os utilizadores mais valorizam numa plataforma dedicada a aluguer de alojamento	181.263 comentários de alojamentos listados na plataforma *Airbnb* da cidade de Sidney	Foram utilizadas técnicas de *text-mining* e análise de sentimento com recurso ao software *Leximancer*	Foram detetados 4 tópicos essenciais da análise aos comentários sendo os quais a localização, as comodidades, o anfitrião e as recomendações. Conclui-se que a experiência dos visitantes é maioritariamente positiva.
2020	Tsai et al. (2020)	Propor uma abordagem de sumarização dos comentários no sentido de classificar os mesmos de forma a salientar os mais importantes e relevantes para posterior avaliação	23.430 comentários de 23.038 utilizadores na plataforma *TripAdvisor.com*	Construção de classificadores para os comentários e posterior análise de sentimento dos mesmos	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.
2020	Yi Luo et al. (2020)	Entender a relação existente entre uma avaliação dada a um atributo específico, relacionado com a importância e o sentimento relativo a um estabelecimento, e a satisfação geral face ao mesmo em restaurantes localizados em Chicago, Las Vegas, Los Angeles, Nova Iorque e Orlando, USA	244.649 comentários de 4.704 restaurantes na plataforma *Yelp*	LARA (*Latent Aspect rating analysis*)	O estudo deu origem a 5 tópicos sendo os quais as comidas/bebidas, serviço, ambiente do restaurante, valor do restaurante e localização. Destes 5, o valor do restaurante, que neste caso representa o rácio qualidade/preço para o utilizador, foi considerado o mais importante para a avaliação final do estabelecimento.
2017	Kim et al. (2017)	Perceber o que provoca insatisfação na experiência dos turistas em determinado contexto	19.835 comentários da plataforma *Virtualtourist.com* de vários estabelecimentos de Paris organizados por 14 categorias	método *Co-ocurrence*	Conclui-se que a categoria transporte foi alvo de descontentamento pela generalidade dos turistas e que as razões se prendem com as altas tarifas aplicadas no serviço de táxis, as fracas condições a nível de estrutura na rede de metro de Paris e a falta de limpeza no serviço rodoviário da cidade.
2019	Hou et al. (2019)	Identificação dos temas e a comparação das diferenças existentes nos comentários dos utilizadores em 3 agências de viagens com plataforma online implementada.	165.429 comentários obtidas com base nos comentários retirados de 3 agências de viagens (Ctrip, Tuniu and Tongcheng)	Análise de associação semântica	O estudo encontrou diferenças consideráveis nas palavras temáticas entre as várias agências, e nas suas relações a nível estrutural. O que indica que existe uma relação próxima entre os comentários dos utilizadores e alvo do modelo de negócio aplicado por cada agência o que origina diferentes níveis de interesse e importância nos comentários dos utilizadores.
2011	Park & Yoon (2011)	Desenvolver de forma sustentável indicadores válidos a utilizar como medida de controlo na gestão no contexto de Turismo Rural	34 especialistas na área	Metodo Delphi	foram definidas quatro dimensões fulcrais para esta matéria, e um total de 33 indicadores, sendo as dimensões as seguintes: Qualidade de Serviço, Instalações, Sistema de Gestão e Resultados.

Ano	Autor(es)	Tema	Amostra	Métodos Utilizados	Resultados/Conclusões
2018	Pitchayadejanant & Nakpathom (2018)	Perceber quais as atividades preferidas pelos turistas ao visitar terrenos de cultivo de frutas na Tailândia num contexto de agro-turismo	409 observações obtidas por questionário	FP-Growth Algorithm e agrupamento	Da lista de atividades tipicamente disponíveis para os turistas realizarem, a colheita e prova de fruta foi a atividade que mais contribui para a satisfação dos turistas aliada também às caminhadas, compras de artigos típicos e alimentação de animais em segunda instância.
2019	Prakash et al. (2019)	Estudar as razões associadas ao descontentamento demonstrado pelos turistas em visita aos parques naturais no Sri Lanka	206 *reviews* da plataforma *TripAdvisor* com a pior classificação atribuída pelos utilizadores	Tratamento estatístico	Foram encontradas, 15 razões para o descontentamento dos turistas na sua experiência, sendo 75% dos quais relacionados com a gestão do parque e muito focados no elevado trânsito dentro da reserva
2020	Sun & Shao (2020)	Medir o grau de satisfação dos habitantes da zona de Shenzhen na China perante as áreas verdes ao seu dispor	3.511 comentários da rede social *Weibo*	Word2Vec e *Long Short-Term Memory Model*	O estudo expõe o grau de satisfação dos utentes quanto aos parques e conclui que a faixa etária que se demonstra mais satisfeita está situada entre os 21 e os 40 anos e de posses financeiras mais limitadas, dado o facto de muitos parques serem de entrada gratuita.
2019	Ruhanen (2019)	Entender o sentimento associado à experiência ecoturística neste caso na Austrália e se o crescimento assinalado se deve ou não a uma maior preocupação e sensibilidade ambiental por parte dos consumidores.	3022 comentários da plataforma *Tripadvisor.com*	Tratamento manual aliado a análise semântica	Conclui-se que embora exista uma maior procura para este tipo de experiência, esta é mais motivada pela própria experiência em si do que pela preocupação ambiental e a sua sustentabilidade. Grande parte dos comentários relativos ao fator ambiental e ao ecoturismo dizem respeito sobretudo à adaptação das infraestruturas e comodidades às áreas naturais envolventes e não a uma preocupação genuína com o tema.
2017	Eusébio et al. (2017)	Qual o principal consumidor nas áreas rurais em Portugal	questionário com 1.853 respostas	Tratamento estatístico em agrupamento	O estudo conclui que existe heterogeneidade nas características sociodemográficas assim como nos comportamentos durante a experiência turística. É dado também destaque às características ambientais em contexto rural, a tranquilidade envolvente e as atividades culturais e gastronómicas do local a visitar.