INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Data Mining aplicado a *reviews online* e estudo da insatisfação em hotéis da cidade de Lisboa

João Pedro Esteves de Oliveira Gonçalves

Mestrado em Business Analytics

Orientadora:
Professora Doutora Diana Aldea Mendes, Prof. Associada, Iscte Business School, Departamento de Métodos Quantitativos para Gestão e Economia

Setembro, 2023

BUSINESS
SCHOOL

Departamento de Métodos Quantitativos para Gestão e Economia

Data Mining aplicado a *reviews online* e estudo da insatisfação em hotéis da cidade de Lisboa

João Pedro Esteves de Oliveira Gonçalves

Mestrado em Business Analytics

Orientador:
Professora Doutora Diana Aldea Mendes, Prof. Associada, Iscte Business School, Departamento de Métodos Quantitativos para Gestão e Economia

Setembro, 2023

Agradecimentos

Os meus agradecimentos são dirigidos a todos aqueles que não só me ajudaram na realização deste trabalho, como também àqueles que estiveram presente ao longo do meu mestrado e percurso académico.

À minha orientadora, Professora Diana Aldea Mendes, pela atenção e acompanhamento no desenvolvimento da dissertação.

A todos os professores com quem me cruzei no ISCTE e que me proporcionaram as fundações de conhecimento que possibilitaram a concretização deste estudo.

Um agradecimento especial aos meus pais, Maria de Fátima Bombaça Gonçalves e Mário André Salgueiro que, através dos seus esforços, permitiram-me seguir o meu caminho.

À minha restante família, avós, tios e primos pelo incentivo, compreensão e apoio incondicional.

Aos meus amigos Raquel Cravidão e Pedro Jesus pelas suas amizades e companheirismo que, por um lado, marcaram presença nas minhas provas de superação e, por outro, permitiram um ciclo de estudos no ISCTE bem mais feliz e alegre.

Aos meus amigos de Setúbal e em especial ao meu melhor amigo João Cruz, que apesar de pouco conseguirem ajudar tecnicamente, sempre estiveram presentes e dispostos a motivar-me.

Resumo

Para a evolução da hotelaria, os hotéis devem aproveitar a revolução em torno do *Analytics*, do *Big Data* e do *Machine Learning* no intuito de compreender melhor os seus clientes-tipo para perceber onde devem investir, de modo a satisfazer as expetativas dos hóspedes consoante diferentes classes de hotel. O objetivo deste estudo é analisar comentários negativos de *reviews online* a fim de perceber em que aspetos divergem reclamações de hotéis de classe baixa, das de hotéis de luxo. Desta forma, extraíram-se comentários negativos de cerca de 150 mil *reviews* do portal Booking.com de 216 hotéis lisboetas e adotou-se a metodologia CRISP-DM. Os dados tipo texto foram estruturados através de um processo de *Text Mining* que envolveu pré-processamento, a análise de frequência, além da revisão de literatura para a construção de 18 variáveis independentes. A variável dependente correspondeu à classe do hotel criada a partir das estrelas dos hotéis. Seguidamente, os dados foram modelados segundo um algoritmo de árvores de decisão, acabando por ter uma capacidade preditiva baixa. A análise do modelo revelou que em hotéis de classe baixa os hóspedes queixaram-se das atividades de recreação ou da falta delas, e também da relação qualidade-preço. Quanto a hotéis de luxo, o estudo mostrou que não abordam tanto o valor, as instalações, ou interações com o *staff*, mas sim o *catering* e o reporte de cheiros desagradáveis. Concluindo, a produção deste estudo permite expandir o conhecimento em torno da insatisfação em hotéis, como pode fornecer *insights* para os gerentes aplicarem nos hotéis.

Palavras-chave: Hotelaria; *Machine Learning*; *Reviews online*; *Text Mining*; Árvores de decisão; Insatisfação.

Sistema de Classificação JEL: M15; Z31.

Abstract

For the evolution of hospitality, hotels should take advantage of the revolution around Analytics, Big Data and Machine Learning to better understand their typical customers, know where they should invest, and satisfy the expectations of guests according to different hotel classes. This study aims to analyze negative comments from online reviews to understand how low-class hotels complaints differ from those of luxury hotels. Therefore, negative comments were extracted from around 150 thousand reviews from Booking.com and 216 Lisbon hotels, and the CRISP-DM methodology was adopted. The text-type data were structured through a Text Mining process that involved pre-processing, frequency analysis, and a literature review to construct 18 independent variables. The dependent variable corresponded to the hotel class that was created from the number of hotel stars. Then, the data were modeled according to a decision tree algorithm, which ended up having a low predictive capacity. Model analysis revealed that in lower-class hotels, guests complained about recreational activities or the lack of them, and also about the quality-price ratio. As for luxury hotels, the study showed that negative comments don't address value, facilities, or interactions with staff so much as catering and the report of unpleasant smells. In conclusion, the production of this study allows extending the knowledge around dissatisfaction in hotels, as it can provide insights managers can apply in hotels.

Keywords: Hospitality; Machine Learning; Online reviews; Text Mining; Decision trees; Dissatisfaction.

JEL Classification System: M15; Z31.

Índice

Índice de Quadros

Índice de Figuras

Lista de abreviaturas, acrónimos e siglas

AI – *Artificial Intelligence*

BD – *Big Data*

CONCOR – *Convergence of Literated Correlation*

CRISP-DM – *Cross Industry Standard Process for Data Mining*

DM – *Data Mining*

Et al. – *et alii* ("e outros"), *et aliae* ("e outras"); *et alia* ("e outros")

eWOM – *Electronic Word-of-Mouth*

IoT – *Internet of Things*

JEL – *Journal of Economic Literature*

LDA – *Latent Dirichlet allocation*

ML – *Machine Learning*

SNA – *Semantic Network Analysis*

STM – *Structural Topic Model*

TIC – Tecnologias da Informação e Comunicação

TM – *Text Mining*

WOM – *Word-of-Mouth*

Introdução

A satisfação dos hóspedes demonstrada em *reviews online* tem sido cada vez mais alvo de atenção, tanto de investigadores que estudam os principais motivos de satisfação dos hóspedes, como também de gestores de hotéis que procuram traduzir o conhecimento extraído numa melhoria de desempenho.

No entanto, a insatisfação em hotéis é um tema ainda pouco investigado dentro da comunidade científica, especialmente quando comparado com o da satisfação. Pelo que, é objetivo deste estudo explorar esta temática, de modo a expandir a literatura em volta da insatisfação de hóspedes.

Assim, esta introdução tem como finalidade contextualizar e definir o problema de investigação, formular as questões de investigação às quais é pretendido responder, bem como apresentar os objetivos, além dos possíveis contributos da investigação. Para além disso, também será abordada a estrutura e, mesmo que de forma breve, a metodologia do presente trabalho.

Contextualização do problema

Hoje em dia, o turismo carateriza-se por ser um setor fundamental para o desenvolvimento económico dos países (Furtado *et al*., 2022). Trata-se de um setor que abrange a indústria hoteleira, pelo que a relação entre os dois é estreita, sendo o turismo diretamente influenciado pela hotelaria. Assim, pode-se considerar que os hotéis são ativos essenciais no setor do turismo (Mucharreira *et al*., 2019).

Em sequência, para o desenvolvimento do setor do turismo e para a competitividade da hotelaria, os hotéis devem aproveitar a revolução tecnológica de modo a estudar melhor os seus clientes e perceber em que devem investir para melhorar o seu desempenho.

Atualmente, há a expetativa de que a rede 5G crie valor ao combinar as Tecnologias da Informação e Comunicação (TIC) - que abrange a *Artificial Intelligence* (AI), o *Big Data* (BD) e a *Internet of Things* (IoT) - com os negócios. Desta forma, integrar o físico com o digital pode alavancar a indústria do turismo e tornar hotéis em melhores lares para os seus hóspedes (Gaur *et al*., 2021).

Ademais, as TIC começaram a impactar o setor hoteleiro devido ao surgimento do *social media* e das plataformas onde os hóspedes compartilham as suas experiências (Bizirgianni & Dionysopoulou, 2013). Esta nova realidade veio alterar a forma como os hotéis gerem a sua comunicação *online* e revolucionou a forma como clientes reservam e planeiam as suas viagens (Casado-Díaz *et al*., 2020). Além disso, passou a ser comum, para os clientes,

avaliarem as suas estadias em *reviews online*. Nestas, o cliente deixa comentários em texto e classificações numéricas para indicar a sua avaliação do serviço (Siering *et al*., 2018).

Do ponto de vista dos hotéis, as *reviews* são consideradas relevantes na medida em que permitem estudar o comportamento do hóspede. Lockyer (2005) referiu que, no setor do turismo, havia uma lacuna entre as perceções dos gerentes de hotéis em relação aos atributos de satisfação do cliente e o que os hóspedes acreditam ser essencial quando reservam e avaliam a sua estadia num hotel.

Dito isto, as *reviews* são consideradas como um fator de elevada importância na explicação da satisfação do cliente e do comportamento de compra. Deste modo, tornou-se imprescindível, para os hotéis, estudar esta nova fonte de dados diretamente ligada aos seus clientes, cuja análise permite obter *insights* para a melhoria do desempenho do serviço (Fernandes & Fernandes, 2018). Como as *reviews online* incluem *feedback* negativo e positivo da sua estadia, responder com rapidez e eficiência a avaliações negativas de clientes é uma maneira de tentar melhorar o serviço (Alsayat, 2022).

De outro ponto de vista, as *reviews online* contribuem ativamente para formar perceções positivas ou negativas de hotéis, influenciando a angariação de clientes e lealdade ou retenção de antigos clientes (Han & Ryu, 2007; Schuckert *et al*., 2015). Ou seja, as avaliações *online* também têm a função de promover o hotel e de potenciar tanto a sua notoriedade como a sua reputação (Tang & Kim, 2022).

Por conseguinte, potenciais clientes que estejam interessados em efetuar um *booking online* também podem mudar de ideias após ler comentários negativos deixados por outros hóspedes que tenham tido más experiências (Sharifi, 2019). Como as opiniões de clientes podem melhorar ou prejudicar a reputação de um hotel, comentários negativos têm o potencial de manchar a imagem dos hotéis e persuadir potenciais clientes a procurar serviços concorrentes. Além disso, Schuckert *et al*. (2016) referem mesmo que potenciais clientes consideram que as *reviews* com classificações baixas são mais úteis do que aquelas com classificações altas e, portanto, são mais propensas a confiar nas *reviews* com *rating* mais baixo no contexto da indústria hoteleira.

Assim, clientes de hotéis utilizam cada vez mais a *Internet* para compartilhar as suas experiências em relação aos serviços prestados e as suas opiniões, especialmente as negativas, transmitem grande confiabilidade a potenciais clientes e podem mesmo influenciar as suas decisões e escolhas (Perinotto *et al*., 2023). Em suma, com o rápido crescimento das plataformas de *booking online*, as *reviews*, opiniões e recomendações passaram a ser uma fonte de oportunidades e desafios na indústria hoteleira.

Definição do problema de investigação

Atualmente, tornou-se imperativo explorar dados relativos a hotéis e a *reviews* dos respetivos hóspedes de modo a esclarecer quais são os principais motivos de reclamação.

Além disso, um estudo conduzido por Hu *et al*. (2019) concluiu que as reclamações dos clientes variam em diferentes categorias de hotel. Ou seja, em hotéis bem classificados por número de estrelas os hóspedes tendem a queixar-se de aspetos diferentes de hotéis pior classificados por número de estrelas.

Desta forma, o principal objetivo desta investigação passa por estudar o comportamento de insatisfação de hóspedes em hotéis, através da análise de *reviews online,* com recurso ao *Machine Learning* (ML) - ferramenta usada para construir algoritmos que aprendem a detetar padrões em dados e fazer previsões com base nesses padrões (Hillier, 2023) -, de modo a compreender em que aspetos diferem reclamações de hotéis bem classificados por número de estrelas de reclamações de hotéis mal classificados por número de estrelas.

Questão de investigação e objetivos

Assim, a questão de investigação principal que foi formulada para este estudo é a seguinte: Quais são os motivos de reclamação em hotéis na cidade de Lisboa? E em que aspetos diferem comentários negativos associados a hotéis de classe baixa de comentários negativos associados a hotéis de classe elevada?

Além de ser objetivo da investigação responder à questão, também é importante entender a relação entre as reclamações *online* por parte dos hóspedes em Lisboa e a classe do hotel em que se hospedaram, bem como perceber quais são os fatores influenciadores da insatisfação dos hóspedes.

Adicionalmente, é objetivo explorar a literatura em torno do ML aplicado ao estudo da insatisfação de hóspedes. Outra meta deste projeto é formar um modelo preditivo capaz de associar um comentário negativo de um hóspede à classe de hotel ao qual este se hospedou.

Pertinência da investigação

Este estudo visa, de um modo geral, enriquecer a literatura sobre a indústria hoteleira e turística através da utilização de ferramentas de análise de BD, de *Text Mining* (TM) e de *Data Mining* (DM) para prever o comportamento de insatisfação de hóspedes perante diferentes classes de hotéis.

Para investigadores e estudos posteriores ficará uma metodologia que contempla os vários aspetos do ML referenciados anteriormente, um modelo cuja finalidade é a previsão

da classe do hotel ao qual uma reclamação de uma *review online* é dirigida. Também é importante a referenciação das variáveis com mais importância no estudo do comportamento de insatisfação do hóspede face a diferentes classes de hotéis, além de uma breve exploração dos comentários negativos de hotéis da cidade de Lisboa.

Além disso, é esperado que se atinjam conclusões dirigidas tanto a investigadores como a gestores de hotéis na cidade de Lisboa, que ficarão a conhecer os principais motivos de reclamação de acordo com a sua classe de hotel. Assim, os gerentes dos hotéis podem mitigar os pontos fracos e reforçar os pontos fortes das operações dos seus hotéis, em função das expectativas e necessidades dos seus hóspedes-alvo.

Estrutura e organização da investigação

De modo a entender o comportamento de insatisfação dos hóspedes em hotéis da cidade de Lisboa e a concretizar os objetivos supramencionados, foi desenvolvida a presente dissertação, organizada em capítulos que refletem as diferentes fases da investigação.

Primeiramente, de forma introdutória, foi contextualizado o problema em volta do comportamento de insatisfação de hóspedes demonstrada via *reviews online* e como o seu estudo pode trazer diversos contributos tanto a investigadores como a gestores de hotéis à procura de melhorias de desempenho do serviço.

Segundamente, será elaborada a revisão da literatura que auxiliará na escolha das melhores técnicas e métodos que permitirão responder à questão de investigação e atingir os objetivos propostos. Desta forma, este capítulo irá abordar o estudo em foco da satisfação e da insatisfação de hóspedes através de estudos anteriores. Além disso, irão ser discutidas as técnicas de ML que poderão ser executadas no decorrer da investigação.

Já no capítulo seguinte será apresentada a metodologia a utilizar no estudo das *reviews* e na previsão do comportamento de reclamação do hóspede. Serão dados a conhecer todo o processo e técnicas que permitirão atingir os objetivos propostos.

Depois, serão apresentados e discutidos os resultados da investigação. Isto abrange as descobertas da investigação, além das contribuições tanto teóricas como práticas da mesma.

Por fim, será realizada a conclusão da dissertação e as recomendações. Além disso, também serão referidas as limitações do estudo, bem como as sugestões de pesquisa futura.

CAPÍTULO 1

Revisão da literatura

O objetivo da revisão da literatura será reunir, através de fontes credíveis e fidedignas, informação relevante e imparcial que permita esclarecer o problema e formular o melhor método de investigação com a finalidade de estudar o comportamento de insatisfação de hóspedes demonstrada em *reviews online*.

Deste modo, primeiramente será abordada a temática das *reviews online* inserida no contexto da indústria hoteleira. Seguidamente, será exposto o envolvimento do BD no problema e o seu relacionamento com o TM. Além disso, estudar-se-ão as técnicas de TM capazes de transformar dados não estruturados, como é o caso dos comentários das *reviews*, em dados estruturados (organizados em tabela) que possibilitem o seu estudo.

Também serão investigadas as técnicas e algoritmos de DM eficazes na modelação dos dados e a partir dos quais poder-se-á retirar o conhecimento que responda às questões de investigação formuladas.

Por fim, ainda serão apresentadas as conclusões acerca do comportamento de insatisfação já presentes na literatura.

1.1. Indústria hoteleira e *reviews online*

No século XXI, com o progresso das TIC, o comportamento do cliente e os canais de *marketing* têm-se alterado consideravelmente. A ampla adoção da *Internet* permitiu aos clientes compararem cada vez mais os produtos, serviços e até as alternativas de pagamento com base em informações disponíveis, de forma gratuita, nos diversos motores de busca e nos *sites* das empresas (Key, 2017).

Tal como em vários setores, a indústria hoteleira não é exceção e desde a consolidação da *Internet* e da utilização do *e-commerce* para efetuar reservas de hotéis, os clientes também começaram a partilhar as suas experiências e avaliações na *Internet* na forma de *reviews online*. Assim, as *reviews online* referem-se à avaliação dos clientes sobre a satisfação com a sua estadia (Hernandez-Ortega, 2020).

Deste modo, os fenómenos da globalização e da disseminação das redes sociais proporcionam uma fonte de informação sem precedentes a potenciais clientes que ficaram cada vez mais informados acerca dos serviços que vão adquirir (Shiau *et al*., 2018). No setor hoteleiro, os potenciais clientes podem agora explorar os locais para onde vão viajar e ler as opiniões dos hóspedes que ficaram em cada estabelecimento anteriormente (Gonçalves *et al*., 2018). Por conseguinte, potenciais clientes ficam cientes do *background*

cultural do destino e podem tomar medidas para prevenir quaisquer constrangimentos ou situações conflituosas aquando das estadias (Zhou *et al*., 2018).

À comunicação entre clientes acerca de produtos, serviços ou empresas, independente de influência comercial, deu-se o nome de *Word-of-Mouth* (WOM) (Westbrook, 1987). A comunicação WOM é um processo em que clientes conversam entre si e compartilham informações ou opiniões, além de orientar os compradores a ficarem mais interessados ou até mesmo a desistirem de produtos ou de serviços específicos (Banerjee & Fudenberg, 2004). Com o desenvolvimento das TIC, Litvin *et al*. (2008) assumem que o *Eletronic Word-of-Mouth* (eWOM) pode ser definido como toda a comunicação informal, de cliente para cliente, dirigida por meio de tecnologias baseadas na *Internet*, relacionada com o desempenho de bens ou serviços específicos.

Já foi estudado que potenciais clientes confiam nas avaliações *online* e que estas servem como uma referência importante para as decisões de compra (Lee *et al*., 2011). Um estudo revela mesmo, depois de inquirir mil pessoas, que 67,7% das pessoas inquiridas são impactadas pelas avaliações *online*. Adicionalmente, mais de metade dos entrevistados (54,7%) admitiram que as avaliações *online* são relevantes no processo de tomada de decisão (Hinckley, 2015). Segundo Kim *et al*. (2018), para potenciais clientes, o eWOM é considerado ainda mais confiável do que anúncios comerciais. Desta forma, pessoas interessadas em efetuar um *booking* podem ficar convencidos ou até a mudar de ideias após ler os comentários *online* de outros viajantes (Sharifi, 2019).

De outro ponto de vista, as *reviews online* podem não só ser usadas por gerentes como uma ferramenta de *marketing* - as avaliações dos clientes podem ser importantes para a promoção do serviço -, bem como a sua análise pode permitir alocar recursos de acordo com o tipo de hóspede que recebe e, assim, melhorar o desempenho e a qualidade do serviço (Tang & Kim, 2022). Segundo Kim e Park (2017), as avaliações *online* desempenham um papel ainda mais significativo na explicação das métricas de desempenho do hotel do que o *feedback* tradicional, pois as *reviews online* de clientes podem ter influência na fidelidade do cliente - se clientes avaliarem a sua experiência como positiva, os seus níveis de satisfação, prontidão ou vontade de recompra aumentarão (Jani & Han, 2013) -, na promoção do hotel e, consequentemente, na performance da empresa (Anastasiei & Dospinescu, 2019).

Por conseguinte, é cada vez mais necessário à indústria hoteleira tirar partido do conteúdo gerado pelos clientes, sob forma de *review online*, na tentativa melhorar o desempenho dos hotéis na prestação de serviços, aumentar a competitividade e tornar este setor do turismo mais *smart*. No fundo, usar a tecnologia para oferecer serviços mais personalizados e com foco no turista (Moro *et al*., 2022).

Concluindo, para melhorar o desempenho do hotel, além de procurar a maximização da satisfação, também é preciso mitigar a insatisfação, que decorre de uma falha no serviço no qual o hóspede depositava uma certa expetativa incumprida (Sann *et al*., 2022). Assim, os gerentes de hotéis devem melhorar os seus serviços e entender as preocupações dos clientes com base em *reviews* negativas. Além disso, os gerentes de hotéis podem tomar melhores decisões estratégicas se tiverem uma melhor compreensão dos perfis, necessidades e atitudes dos seus clientes *target* (Kitsios *et al*., 2021), isto após o estudo do seu comportamento de insatisfação.

1.2. BD e TM aplicados a *reviews online*

O *Analytics* - processo metódico usado para extrair, organizar, interpretar, visualizar e tirar conclusões dos dados (Hillier, 2023) - e o BD foram reconhecidos como alguns dos impulsionadores tecnológicos da revolução e transformação digital dos negócios. Além de representar um novo paradigma tecnológico, o conceito de BD foi definido como uma vasta quantidade de dados, estruturados ou não estruturados (em variadíssimos formatos), produzidos a grande velocidade devido aos avanços tecnológicos e ao crescimento e difusão da automação, da *Internet* e de dispositivos conectados - associado ao surgimento da IoT - (Mariani & Nambisan, 2021).

Já Zhang e Kim (2021) consideram que a análise de BD é um conceito amplo definido como um conjunto de técnicas que permitem descobrir padrões ocultos, *insights* e relações interessantes na compreensão dos mais variados contextos, através de grande quantidade e variedade de dados. Além disso, Gaur *et al*. (2021) referem que, na atualidade, a análise de BD supera outros métodos mais convencionais de análise.

Antigamente, os investigadores usavam principalmente metodologias de pesquisa, como inquéritos a clientes, entrevistas ou *brainstorming* para estudar o comportamento do cliente (Kitsios *et al*., 2021). E, só na última década, um número crescente de investigadores começaram a realizar análises BD para descobrir padrões em dados que podem traduzir-se em conhecimento e que podem tornar os negócios mais competitivos (Erevelles *et al*., 2016). Desta forma, o conteúdo gerado por utilizadores da *Internet* surgiu como uma fonte de dados oportuna e rica para detetar padrões de comportamento de clientes através de análises BD, especialmente no que concerne à indústria hoteleira (Nilashi *et al*., 2022).

Ou seja, o BD surge como uma oportunidade para entender o mercado hoteleiro e para apoiar a tomada de decisão dos gerentes. Desta forma, o conteúdo gerado pelos clientes na *Internet*, as *reviews online*, permitem aos gerentes de hotéis obter *feedback* dos hóspedes e aprimorar características específicas do serviço, a fim de melhorar o serviço e a sua

proposta de valor ou apoiar as atividades de *marketing* e aumentar a procura (Kitsios *et al*., 2021).

Assim, os investigadores têm estudado cada vez mais o comportamento do cliente através de análises BD ao conteúdo gerado pelo mesmo, no contexto do turismo e da hospitalidade (Nilashi *et al*., 2019; Shen *et al*., 2021).

Por um lado, existem estudos acerca do comportamento do cliente que adotam abordagens de pesquisa quantitativa, através de *ratings* numéricos. Por outro lado, revelar a satisfação do cliente utilizando técnicas BD aplicadas a dados não estruturados, como é o caso do eWOM, pode ser uma maneira eficaz de entender melhor o cliente (Alsayat, 2022). Segundo Nie *et al*. (2020), os comentários de *reviews online* têm até um melhor desempenho ao descrever perceções e sentimentos que classificações quantitativas, como os *ratings*.

No entanto, para complementar a análise BD e poder analisar as *reviews* e os comentários *online* é necessário transformar os dados não estruturados - o texto - em dados estruturados. É aqui que surge o TM que refere-se à extração de informações através de técnicas de processamento de linguagem natural para descobrir padrões úteis desconhecidos e extrair conhecimento de texto (Vijay Gaikwad, 2014). Já na Figura 1.1 é possível visualizar um processo composto por tarefas próprias de uma metodologia de TM (Ban *et al*., 2019).

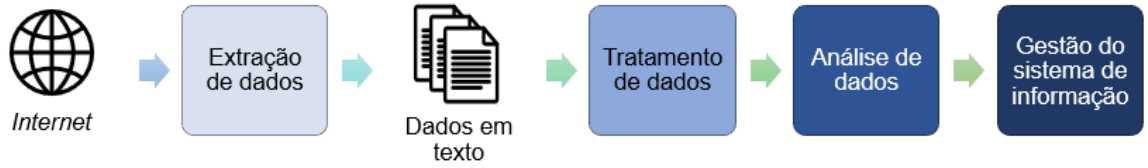

Figura 1.1 - Processo de *Text Mining* (Ban *et al.*, 2019)

Embora abordagens de TM ainda apresentassem algumas limitações - um computador não conseguia discernir diferenças subtis no significado de palavras e apenas tinha em conta a frequência das palavras-chave - em 2008, segundo Matthes e Kohring, a verdade é que o TM foi aprimorado e, atualmente, este tipo de técnicas já é bastante eficaz na transformação de texto em dados estruturados (Wei *et al*., 2015).

Desta forma, começaram-se a utilizar técnicas de TM em comentários de *reviews online* para auxiliar no estudo do cliente e, mais concretamente, entender os principais determinantes da satisfação/insatisfação em hotéis (Handani *et al*., 2022; Tang & Kim, 2022).

Técnicas de TM muito utilizadas para a pesquisa da (in)satisfação do cliente são a análise de frequência de palavras, a *Semantic Network Analysis* (SNA) – análise de centralidade e análise *Convergence of Literated Correlation* (CONCOR) – e a análise fatorial (Wei & Kim, 2022).

A análise de frequência das palavras examina a importância dos termos num texto ou conjunto de textos, medindo a frequência com que estes aparecem (Chaar-Perez, 2023).

Já a SNA examina a relação entre as palavras e exibe a sua ligação através da estrutura da rede. Ou seja, analisa a semântica padrão das palavras nos textos através da relação entre a frequência de palavras e o seu contexto. Para o estudo da centralidade dos termos há duas formas de análise: uma é através do *Freeman's Degree Centrality* - refere-se ao número de ligações diretas que uma palavra possui -; e a outra é através do *Eigenvector Centrality* - quanto maior o grau de centralidade de uma palavra, mais conexão ela tem com outros nós da rede - (Zhang & Kim, 2021). Após a análise da centralidade das palavras, para as agrupar em *clusters*, muitos investigadores recorrem à análise CONCOR, um método eficaz de agrupamento por correlação de palavras em texto e que determina um nível apropriado de grupos (Fu *et al*., 2022; Handani *et al*., 2022; Tang & Kim, 2022).

Adicionalmente, muitos investigadores ainda aplicam a análise fatorial, um processo de redução e de busca entre as semelhanças das variáveis. Isto é, um processo cujo objetivo é reduzir um grande número de variáveis e que seleciona apenas aquelas que são essenciais para que representem o total de fatores (Wei & Kim, 2022).

Outras técnicas que têm ganho relevo na determinação dos principais tópicos de satisfação ou de queixa em hotéis são *a Latent Dirichlet Allocation* (LDA) e, mais recentemente, a *Structural Topic Model* (STM). A LDA e a STM são técnicas ML não supervisionadas que podem capturar dimensões específicas de um contexto e processar vastas revisões não estruturadas e com pouca intervenção manual (Song *et al*., 2022; Srivastava & Kumar, 2021). Isto é, analisam as palavras de um corpo de texto e identificam os tópicos ou temas latentes. Ambas as técnicas tratam-se de *Bayesian generative topic models* e, a partir delas, é possível absorver os principais tópicos de um texto automaticamente. Já muitos investigadores decidem aplicar estas técnicas de criação automática de tópicos de modo a descobrir os temas abordados em conjuntos de texto de *reviews online* (Srivastava & Kumar, 2021).

1.3. DM aplicado a *reviews online*

Segundo Santos e Ramos (2017, p. 168) o DM " […] é uma das fases do processo de descoberta de conhecimento em bases de dados […]" e "[…] consiste na procura de

relacionamentos e padrões ou modelos que estão implícitos nos dados armazenados em grandes bases de dados […]".

Deste modo, o processo de DM gera modelos a partir de dados passados que são usados posteriormente para previsões, deteção de padrões ou identificação de regras que caraterizam dados históricos. Ou seja, as técnicas de DM oferecem métodos viáveis para a deteção de relações causais, para a especificação de quais variáveis têm dependência significativa do problema de interesse e para a construção de modelos que ajudam a prever o futuro (Nourani & Molajou, 2017).

Por conseguinte, nos últimos anos, investigadores têm vindo a adotar procedimentos DM na condução dos seus estudos nas indústrias do turismo e da hospitalidade e, mais concretamente, no entendimento da (in)satisfação do cliente (Hu *et al*., 2019; Zhao *et al*., 2019). Com a ajuda de uma abordagem DM, os gerentes de hotéis podem receber informações valiosas que lhes permitem obter uma melhor perceção sobre o comportamento do cliente e desenvolver estratégias eficazes de *marketing* e de retenção de clientes (Golmohammadi *et al*., 2012).

Embora os dados recolhidos do *feedback* do cliente, os comentários que deixam *online*, sejam dados textuais não estruturados, a maioria das abordagens de DM lida apenas com dados estruturados. Além disso, os dados recolhidos para o intuito de estudar o cliente desta forma são geralmente volumosos e de pouca utilidade na sua forma mais *raw* (Khade, 2016). É a informação escondida nos textos que tem valor para os modelos de DM e é por isso que é tão importante combinar técnicas de BD e de TM antes de se proceder ao DM neste tipo de problemas.

Noutro sentido, a tarefa associada ao DM perante este tipo de problemas é, geralmente, a descrição - que permite identificar regras que caraterizam os dados analisados - ao invés da previsão, uma vez que o objetivo é, de forma prioritária, adquirir um conhecimento mais alargado dos dados analisados e, secundariamente, obter o modelo com resultados mais precisos (Santos & Ramos, 2017, p. 168).

Existe uma grande variedade de técnicas de DM, as quais são consumáveis através de diferentes algoritmos. No entanto, em problemas a envolver a determinação dos influenciadores da (in)satisfação (segundo a classe de hotel), as técnicas que geralmente são concretizadas baseiam-se em regressões lineares (Tang & Kim, 2022; Zhang & Kim, 2021) ou em árvores de decisão (Nilashi *et al*., 2021; Sann *et al*., 2022; Singh & Alhamad, 2022).

Em relação à regressão linear, trata-se de um modelo estatístico que modela a relação entre uma variável dependente e uma ou mais variáveis dependentes. No entanto, só pode ser utilizada quando se pretende prever uma variável quantitativa e com valores contínuos, pelo que a variável dependente, nestes casos a satisfação, é quantificada. Já os

coeficientes que explicam a relação entre os determinantes e a satisfação podem ser estipulados a partir do método dos mínimos quadrados, que procura minimizar o erro existente entre os valores reais dos dados e os valores estimados (Santos & Ramos, 2017, p. 176).

Contudo, o objetivo deste estudo é entender as diferenças entre as queixas de clientes de hotéis de classe alta e queixas de clientes de hotéis de classe baixa. Pelo que seria necessário realizar dois modelos, um para cada tipo de hotel e, posteriormente, comparar os resultados.

Quanto às árvores de decisão, "como o nome indica, são constituídas por estruturas em árvore que representam um conjunto de decisões. Possuem uma representação simples, sendo facilmente interpretadas pelos utilizadores". Além disso, este tipo de algoritmos de indução de árvores de decisão, de aprendizagem supervisionada, permite responder a problemas de classificação - "enquadramento de um conjunto de dados dentro de classes predefinidas, identificando a classe a que cada elemento pertence" - (Santos & Ramos, 2017, p. 178). Neste tipo de problemas, o conjunto de dados disponível é dividido em dois subconjuntos: o primeiro, o conjunto de treino é utilizado para formar o modelo; e o segundo, o conjunto de teste, tem como objetivo avaliar a qualidade do modelo que é formado.

As árvores de decisão são compostas por nós, ramos e folhas. Cada nó interno representa um atributo a classificar. Cada ramo iguala os valores possíveis para esses atributos. E cada folha é um nó terminal que indica a classe em que cada registo pode ser classificado (Lan *et al*., 2018; Lee *et al*., 2018; Taamneh, 2018). Uma árvore de decisão pode ser também representada por regras, e cada folha da árvore traduz-se numa regra que integra uma conjunção de valores associados aos atributos (que existem nos ramos) e que ligam a folha à raiz da árvore (Santos & Ramos, 2017, p. 173).

Por um lado, as árvores de decisão têm bastantes vantagens: são fáceis de entender e interpretar (trata-se de uma técnica simbólica), lida tanto com dados numéricos como categóricos, funciona muito bem com um grande conjunto de dados num curto espaço de tempo (suporta BD) e proporciona excelentes resultados com boas visualizações (Sann *et al*., 2022).

Por outro lado, as árvores de decisão podem ter problemas de sobreajustamento: o modelo pode ajustar-se em demasia ao conjunto de treino e, assim, os *outliers* deste conjunto podem influenciar o desemprenho na classificação do conjunto de teste. Para evitar isto, pode-se efetuar um *pruning* da árvore o que pode permitir melhorar o desempenho do modelo (Santos & Ramos, 2017, p. 175).

Existem muitos algoritmos de árvores de decisão que podem ser aplicados e até dependem de ferramenta para ferramenta; no entanto, os algoritmos CART, CHAID, C5.0 e QUEST são os mais usados e mais reconhecidos (Sann *et al*., 2022).

1.4. Estudo da insatisfação em *reviews online*

À medida que os mercados mudam de *company-driven* para *customer-driven*, as empresas têm aumentado cada vez mais o interesse na experiência e na satisfação dos clientes. Além disso, o estudo do cliente tem sido reconhecido como um fator de relevo e com impacto nas atividades de uma empresa (Fu *et al*., 2022).

Definir os fatores de satisfação e insatisfação do cliente pode ajudar a indústria hoteleira a usar o eWOM de forma mais eficaz, uma vez que as avaliações *online* descrevem o sentimento do cliente face ao serviço de forma mais coerente e abrangente, porque o texto não é estruturado (Xu *et al*., 2017). A sua determinação pode ajudar a entender melhor o cliente (Li *et al*., 2013) e permite aos hotéis melhorar tanto na prestação dos seus serviços, como também a praticar *marketing* da forma mais correta e direcionada para o *target* dos hotéis (Kitsios *et al*., 2021).

Na literatura de *marketing*, a satisfação é definida como uma reação favorável emergente de uma avaliação positiva das experiências de consumo (Oliver, 2015). No entanto, por um lado, a satisfação é um tópico já abordado por muitos investigadores que, aqui e ali, têm referido os fatores de insatisfação, embora esta última seja uma temática ainda não muito investigada e com escassez na literatura, em específico.

Por outro lado, para os consumidores, críticas negativas podem ser especialmente influentes e valiosas, pois podem ajudar a prevenir experiências negativas (Fernandes & Fernandes, 2018), uma vez que potenciais clientes usam *sites* de classificações *online* para pesquisar informações e ler críticas negativas e positivas antes de decidir onde reservar.

Além disso, está estudado que as pessoas concentram-se, especialmente, em comentários e avaliações negativas para escolher um hotel (Schuckert *et al*., 2016), como já referido.

Na maioria das vezes, os fatores de insatisfação até diferem dos fatores de satisfação. Xu e Li (2016) verificaram que os fatores de satisfação não eram semelhantes aos fatores de insatisfação do cliente e que estes ainda dependiam do tipo de hotel. Também aperceberam-se de que os fatores de insatisfação do cliente eram muito mais precisos que os de satisfação. Especificamente, chegaram à conclusão (através de uma técnica de TM semelhante à STM e à LDA, a *Latent Semantic Analysis*) de que os fatores de insatisfação na generalidade dos hotéis eram os seguintes: “Wi-Fi”, “Facilities”, “Parking”, “Bathroom”, “Noise”, “Swimming pool” e “Room cleanliness”.

Contudo, há uma pesquisa (Singh & Alhamad, 2022) que divide os fatores não só em satisfatórios ou insatisfatórios, mas em quatro tipos de classificação: satisfatórios, insatisfatórios, críticos ou neutros. Segundo o estudo, instalações e conforto foram os

fatores críticos do serviço, enquanto que a limpeza, *staff* e localização foram os fatores que traduzem satisfação; já o valor de dinheiro gasto no serviço é a principal causa de insatisfação; a disponibilidade de pequeno-almoço e de refeições são fatores neutros.

Noutro estudo que divide os fatores em satisfatórios ou em insatisfatórios há a referência de que clientes satisfeitos concentram-se, geralmente, em atributos mais intangíveis da sua acomodação, como comportamento do *staff* e serviço. No entanto, os consumidores insatisfeitos concentram-se, pelo contrário, nas dimensões tangíveis do hotel (como por exemplo, mobiliário, preço, entre outros) (Zhou *et al*., 2014). Thu (2020) também refere que o fator relativo ao preço, quando está presente em *reviews*, é quase sempre mencionado apenas em críticas negativas.

Zhang e Kim (2021) também atestaram, através de uma análise de regressão linear, que fatores se correlacionavam negativamente com a satisfação e identificaram três fatores entre os quais a empatia familiar, o valor e a qualidade alimentar.

Noutro sentido, Hu *et al*. (2019) aferiram que as reclamações dos clientes variam em diferentes categorias de hotel. Isto é, as *reviews* negativas em hotéis de classe alta não abordam os mesmos tópicos de *reviews* negativas de hotéis de classe baixa. Através de uma análise STM a cerca de 28 mil *reviews* a hotéis de New York City, este estudo determinou que, para hotéis de classe alta, as reclamações dos clientes estão principalmente relacionadas a problemas na prestação de serviços e a questões de preço - isto pode dever-se a altas expetativas de hóspedes aquando da sua estadia em hotéis de luxo -, enquanto que os clientes de hotéis de classe baixa são frequentemente incomodados por aspetos mais tangíveis, como problemas relacionados às instalações ou limpeza, aspetos mais *core* dos estabelecimentos hoteleiros.

Sann *et al*. (2022) também investigaram em profundidade a temática da insatisfação segundo diferentes classes de hotéis. A partir de 1992 queixas *online* de 350 hotéis localizados no Reino Unido, estruturaram os comentários em texto via codificação manual, utilizaram ainda dados extra *review* (o tamanho do hotel e o seu número de quartos) e aplicaram vários algoritmos de árvores de decisão – CART, CHAID, C5.0 e QUEST – para atingir o modelo que melhor descrevesse a relação entre os atributos de insatisfação e a classe dos hotéis. O modelo que apresentou melhor desempenho foi aquele que aplicou o algoritmo CHAID, o qual apresentou as conclusões identificadas na Figura 1.2.

Figura 1.2 - Principais conclusões do estudo conduzido por Sann *et al*. (2022)

CAPÍTULO 2

Metodologia

Para a concretização desta investigação, foi decidido implementar-se a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM). O CRISP-DM é uma metodologia muito utilizada na resolução de problemas de DM e descreve um modelo de referência que define as tarefas a executar e os resultados esperados para a realização de cada uma das fases. É ainda um processo caraterizado pelo possível retrocesso a fases anteriores para mudar decisões, incluir novos dados ou acrescentar novas técnicas ou passos, sendo que o processo pode ser retomado em seguida (Santos & Ramos, 2017, p. 149).

Na Figura 2.1 podem-se observar as seis fases constituintes da metodologia CRISP-DM (Chapman *et al*., 1999):

- Compreensão do problema - pretende-se compreender os objetivos da investigação tendo em conta o problema;
- Compreensão dos dados - execução de tarefas de recolha, de estudo e descrição dos dados, além da identificação dos seus problemas;
- Preparação dos dados - obtenção do conjunto de dados para a fase seguinte e a sua exploração;
- Modelação - seleção das técnicas de modelação, ajustadas nos seus parâmetros;
- Avaliação - avaliação dos modelos obtidos na fase anterior; e
- Desenvolvimento - produção da presente dissertação.

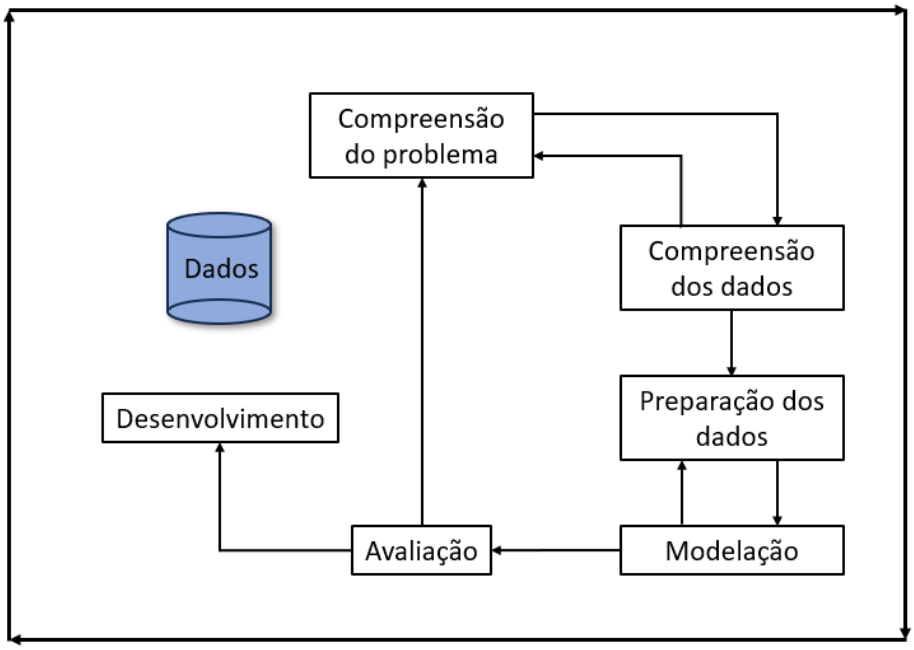

Figura 2.1 - Metodologia CRISP-DM (adaptado de CRISP-DM 1.0, 1999)

2.1. Compreensão do problema

No sentido daquilo que foi abordado na Introdução desta dissertação, a análise de *reviews online* pode constituir uma oportunidade para entender o hóspede quanto às suas necessidades e expetativas. Tendo isto em atenção, os gerentes dos hotéis podem minorar os pontos fracos e melhorar os pontos fortes das operações dos seus hotéis e, assim, tentar maximizar a satisfação dos seus clientes.

Deste modo, como também já foi referido, este estudo dirige-se não só a investigadores como também aos gerentes de hotéis, especialmente aos da cidade de Lisboa, e pretende expandir o conhecimento em volta do tema da insatisfação de hóspedes demonstrada em *reviews online*.

Mais especificamente, o objetivo do presente projeto é analisar os comentários negativos presentes nas *reviews* com recurso ao ML e entender quais são os determinantes da insatisfação segundo diferentes classes de hotéis. Ou seja, é objetivo de DM descrever reclamações dirigidas a hotéis de classe alta e descrever reclamações dirigidas a hotéis de classe baixa (problema de classificação).

Além disso, também é necessário averiguar que variáveis são mais importantes na construção do modelo obtido.

Assim, a tarefa associada ao DM deste problema é a descrição uma vez que permite identificar regras que caraterizam os dados analisados e aumentar o conhecimento acerca dos dados, ao invés da previsão que procura atingir o modelo com maior percentagem de acerto (Santos & Ramos, 2017, p. 168).

Para isso, é necessária a recolha massiva de todos os comentários negativos de *reviews online* de um *site* de *booking*, dirigidos a hotéis da cidade de Lisboa, bem como do número de estrelas do respetivo hotel associado. Desta forma, as técnicas a utilizar para análise têm de estar preparadas para BD uma vez que é perspetivado que se trabalhe com uma grande quantidade de dados.

Já para a preparação dos dados serão utilizadas técnicas de TM com o objetivo de tornar os comentários negativos em dados estruturados.

E, visto que este é um problema de classificação, de aprendizagem supervisionada - as classes que vão dirigir o processo de classificação dos dados são conhecidas à partida (Santos & Ramos, 2017, p. 169) e neste caso é a classe do hotel que é alta ou baixa -, é necessário recorrer a algoritmos que criam modelos simbólicos que permitam a extração de regras e de conhecimento.

Para concluir, é de referir que definiu-se que as fases posteriores serão feitas com recurso ao *Colaboratory* (mais conhecido por "*Colab*"), um produto da *Google*, que permite escrever e executar código na linguagem de código *Python* através do navegador (*Google*

Colaboratory, 2023). Trata-se, assim, de uma ferramenta adequada à resolução deste tipo de problemas a envolver ML e BD. Note-se que para o cumprimento da metodologia foram utilizadas, principalmente, as seguintes bibliotecas de *Python*: *pandas*; *nltk*; *matplotlib*; e *scikit-learn*.

2.2. Compreensão dos dados

A fase de compreensão dos dados integra a execução de três tarefas que foram realizadas no desenvolvimento do projeto: a recolha de dados; a descrição de dados; e a verificação da qualidade dos dados.

2.2.1. Recolha de dados

O conjunto de dados para trabalhar nesta investigação foram recolhidos em formato *Excel* a partir do *Apify*, uma página especialista em soluções de *web scraping* e de automação. Ou seja, uma plataforma com soluções integradas capazes de realizar o *scraping* de informações de vários *sites*, incluindo *sites* de *booking*, e organizá-las em questão de minutos.

Por conseguinte, a extração dos dados foi feita a 14 de maio de 2023, por um robô *scraper* da plataforma, nomeado de *Booking Reviews Scraper*, e cujo *link* da solução é, então, o seguinte: https://apify.com/voyager/booking-reviews-scraper.

Através do *Booking Reviews Scraper* é possível extrair as *reviews* de hotéis listados no portal Booking.com e os dados extraídos podem ser, além dos comentários e avaliações dos hóspedes: informações dos hotéis, como número de estrelas, avaliação geral e outras caraterísticas; informações básicas dos autores das *reviews*; a duração da estadia; informações do quarto; data da estadia.

Do ponto de vista do problema descrito é necessário recolher, por um lado, o Id e o número de estrelas dos hotéis de Lisboa - para medir a classe do hotel - e, por outro lado, extrair todos os comentários negativos das *reviews* de cada hotel.

Para correr o *Booking Reviews Scraper*, do *Apify*, além de se selecionar os dados a extrair, foi necessário colocar, como *input*, uma lista que continha os *links* dos 216 hotéis da cidade de Lisboa presentes no portal *Booking.com* dos quais é possível e pretendido que se extraiam os dados. É ainda de referir que foi possível obter essa lista de *links* através de um *scraper script* elaborado pelo próprio através da biblioteca *BeautifulSoup* do *Python* aplicado na página de hotéis da cidade de Lisboa de *Booking.com*.

Na Figura 2.2 é possível observar, como exemplo, de onde são retirados o Id - identificador no *link* do hotel - e o número de estrelas - que da imagem são convertidas em algarismos - de cada página de hotel (assinalados a vermelho).

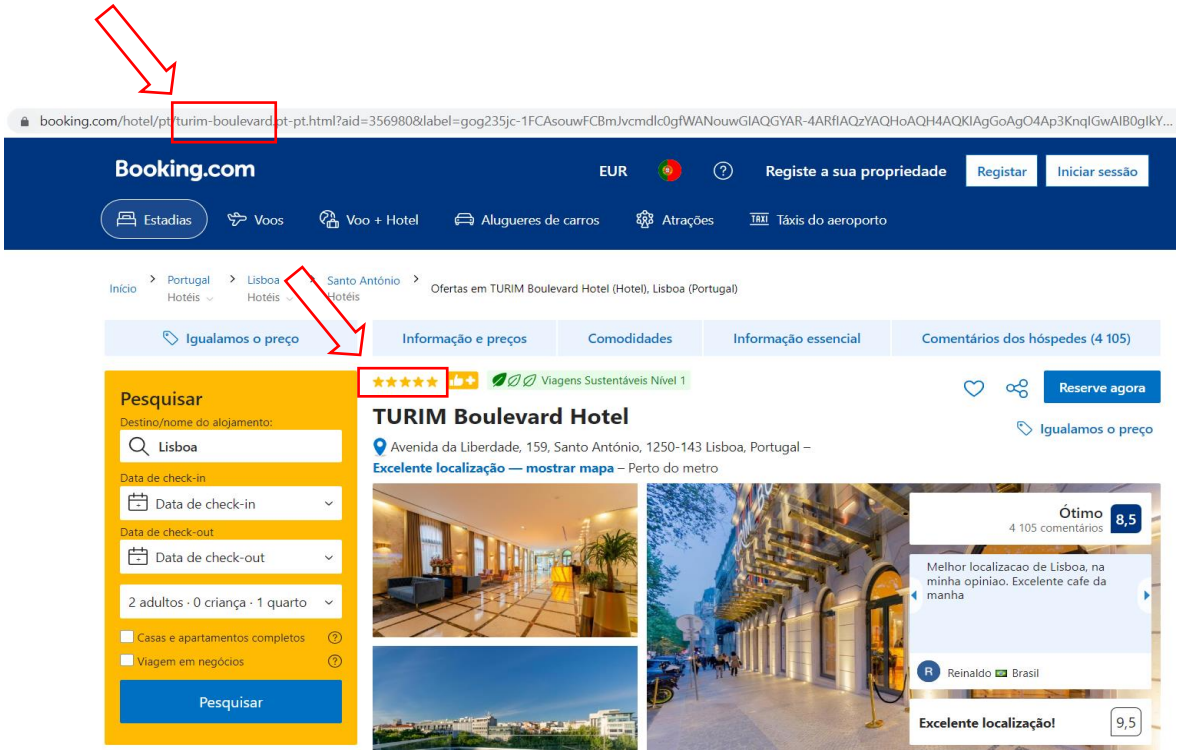

Figura 2.2 - Fonte das informações retiradas de um hotel em *Booking.com*

Já os comentários negativos presentes nas *reviews online* têm um espaço específico nas páginas de cada hotel, como se pode observar na Figura 2.3 (assinalados a vermelho). É a partir destes elementos que o *Booking Reviews Scraper* extrai os comentários meramente negativos dos hóspedes.

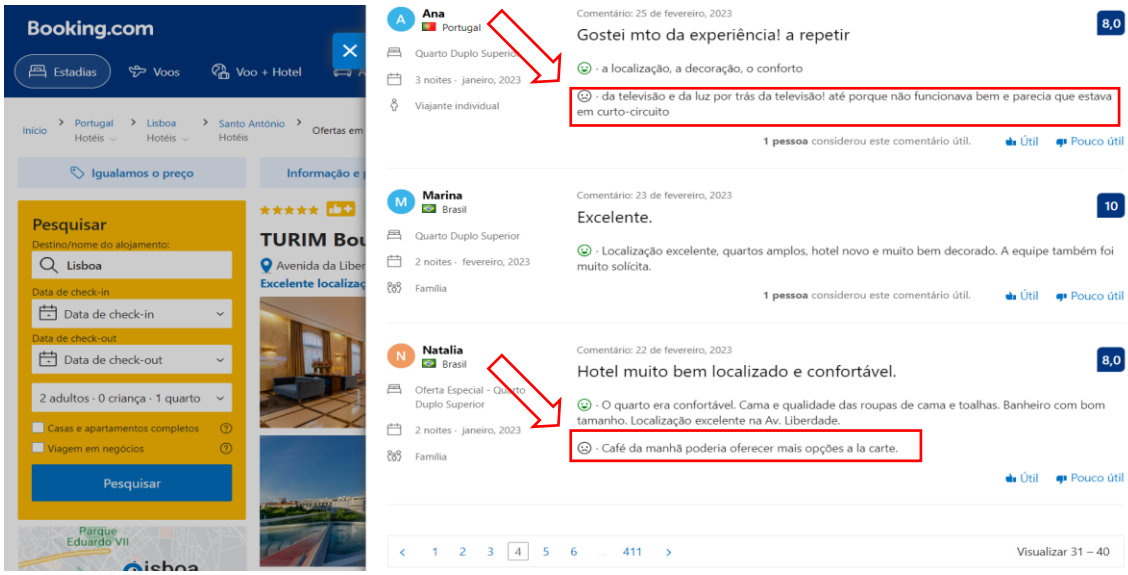

Figura 2.3 - Fonte dos comentários negativos retirados da página de um hotel em *Booking.com*

2.2.2. Descrição dos dados

Assim, para cada um dos 216 hotéis da cidade de Lisboa presentes no portal *Booking.com* foram extraídos o Id, o número de estrelas e todos os comentários negativos presentes nas *reviews* de cada página.

No total, foram extraídos do portal cerca de 150 mil comentários negativos (150 111 comentários, em concreto) relativos aos tais 216 hotéis de Lisboa listados em *Booking.com*. Como hotéis de classe baixa consideraram-se hotéis de 1, de 2 ou de 3 estrelas, já os hotéis de classe alta consideraram-se hotéis de 4 ou de 5 estrelas. E como tal, no Quadro 2.1 pode-se observar a distribuição dos hotéis por número de estrelas e por classe. Dos 216 hotéis de onde se extraíram os comentários, pode-se verificar que 90 hotéis são de classe baixa e que 126 são hotéis de classe alta.

Quadro 2.1 - Distribuição de hotéis por número de estrelas e por classe

	Estrelas	Número de Hotéis	Totais
Classe baixa	1	3	90
	2	17	
	3	70	
Classe alta	4	90	126
	5	36	
			216

2.2.3. Qualidade dos dados

Após uma breve análise e verificação da base de dados, não se detetaram erros decorrentes da sua extração ou registos com campos omissos.

No entanto, é necessário reformular os dados a respeito do número de estrelas dos hotéis para contruir a variável *target* do modelo, a classe do hotel. Isto é, aos hotéis de 1, 2 ou 3 estrelas será atribuída a classe baixa de hotel e aos hotéis de 4 ou 5 estrelas será atribuída a classe alta de hotéis.

Outro dos problemas com os dados, agora a envolver os comentários, é que estes são escritos nos mais diversos idiomas. E, por serem do tipo texto, será fundamental estruturá-los tendo em conta técnicas de TM de modo a atingir o conjunto de dados para a fase de modelação.

É ainda importante notar (ainda que com alguma subjetividade) que o problema a resolver envolve BD, uma vez que os registos recolhidos ultrapassam os 150 mil (grande quantidade de dados) e são compostos tanto por dados estruturados – número de estrelas dos hotéis – como por dados não-estruturados – comentários dos hóspedes em texto –

(grande variedade de dados). Pelo que as técnicas a ser utilizadas têm de estar preparadas para trabalhar com esta base de dados.

2.3. Preparação dos dados

A seguir à fase de compreensão dos dados vem a fase de preparação dos mesmos para atingir o conjunto de dados para modelação e análise. Deste modo, nesta fase de preparação dos dados, realizaram-se, essencialmente, tarefas de seleção, limpeza, construção e formatação dos dados com recurso a técnicas de TM. Adicionalmente, ainda produziu-se a exploração do conjunto de dados que transitou para a fase seguinte (será exposta em capítulo posterior, nos Resultados e discussão).

Primeiramente, foi formulada a nova variável "Hotel_Class" que representa a classe do hotel correspondente a cada comentário. Esta nova variável foi construída a partir dos dados extraídos relativos aos números de estrelas. Deste modo, a hotéis de 1, 2 ou 3 estrelas foi atribuído o valor "0" que representa hotéis de classe baixa, enquanto que a hotéis de 4 ou 5 estrelas foi atribuído o valor "1" que representa hotéis de classe alta. Em seguida, por já não serem necessárias, foram removidas do *dataframe* as variáveis que dizem respeito aos identificadores e aos números de estrelas dos hotéis.

Já para a estruturação dos comentários negativos que estão no tipo texto e transcritos nos mais diversos idiomas, foi seguida a metodologia de TM presente na Figura 2.4.

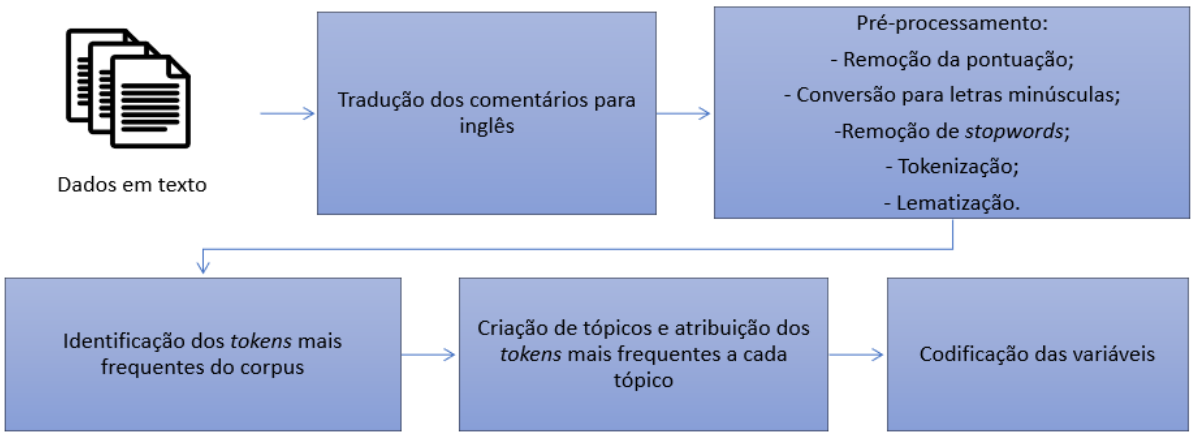

Figura 2.4 - Processo de TM adotado

Num primeiro momento, os comentários como estavam escritos nos mais variados idiomas foram traduzidos através da extensão do *Google Translator* para *Google Sheets*. Assim, o idioma original foi automaticamente detetado e traduzido para inglês, o idioma escolhido para uniformizar os comentários.

Depois, foram realizadas atividades típicas de pré-processamento: a pontuação foi removida; todas as letras foram convertidas em minúsculas; foram removidas *stopwords*; depois foi realizada a tokenização dos textos para dividir cada comentário em *tokens* individuais e possibilitar a sua estruturação; a seguir foi concretizada a lematização, para reduzir as palavras às suas formas base.

Deste modo, foi criada uma variável em que a cada registo está associada uma lista de *tokens* dos comentários negativos. No Quadro 2.2 pode-se observar, como exemplo, comentários tal como foram extraídos e o resultado do seu pós-processamento.

Quadro 2.2 - Exemplo de comentários negativos e pós-processamento correspondente

negative_comment	negative_comment_treated
- a bit expensive	['expensive']
- a bit monotonous breakfast - dark room (insufficient light from outside)	['monotonous', 'breakfast', 'dark', 'room', 'insufficient', 'light', 'outside']
- A bit of faded glory in the rooms, but that also has something. - No facilities in the room (kettle, coffee, tea, etc.)	['faded', 'glory', 'room', 'also', 'something', 'facility', 'room', 'kettle', 'coffee', 'tea', 'etc']
- a bit old bathroom (but still acceptable) - no fresh vegetables at breakfasts - it would be great to have more shelves in wardrobe	['old', 'bathroom', 'still', 'acceptable', 'fresh', 'vegetable', 'breakfast', 'shelf', 'wardrobe']
- A constant muffled noise; - Strange smell whenever I entered the room.	['constant', 'muffled', 'noise', 'strange', 'smell', 'whenever', 'entered', 'room']

Após a transformação dos comentários negativos em listas de *tokens*, é necessária a criação de tópicos a envolver os determinantes da insatisfação, além da sua identificação em cada comentário.

Desta forma, procedeu-se à tentativa de criação de tópicos automática via LDA ou STM. No entanto, os resultados obtidos não tinham a qualidade desejada e não identificavam os tópicos criados de forma correta nos comentários negativos. Assim, foi decidido efetuar a criação de tópicos com recurso ao agrupamento dos *tokens* mais frequentes do corpus, de forma manual. Ou seja, a um tópico será associada uma lista de *tokens* e, se um comentário negativo ("negative_comment_treated") contiver algum desses tokens, o comentário será identificado com esse tópico.

Por conseguinte, foram identificados os 500 *tokens* mais frequentes do corpus. No Quadro 2.3 podemos observar, dentro dos 500 *tokens* mais frequentes, os 10 *tokens* mais frequentes e os 10 menos frequentes.

Quadro 2.3 - Os 500 *tokens* mais frequentes do corpus

Rank	Token	Frequência
1º	"room"	67617
2º	"breakfast"	28599
3º	"hotel"	28420
4º	"small"	18196
5º	"bathroom"	15289
6º	"little"	15046
7º	"bed"	12586
8º	"noise"	9945
9º	"shower"	9929
10º	"night"	9461
...	...	...

Rank	Token	Frequência
...	...	...
491º	"comment"	579
492º	"quiet"	578
493º	"shame"	578
494º	"7"	577
495º	"taking"	574
496º	"complicated"	572
497º	"received"	572
498º	"mirror"	569
499º	"pressure"	568
500º	"gel"	566

Segundamente, os 500 termos mais frequentes foram organizados nos 18 tópicos seguintes: “Value”, “Facilities”, “Parking”, “Recreation”, “Room_Issue”, “Bathroom_Issue”, “Service_Encounter”, “Booking_Issue”, “Cleanliness_Maintenance”, “Wifi”, “TV”, “Catering”, “Location_Access”, “Lighting_Issues”, “Smelling_Issues”, “Air_Conditioning”, “Noise_Issues” e “Safety_Privacy”. Estes tópicos foram definidos tendo em conta uma breve exploração dos 500 *tokens* mais frequentes e os determinantes da satisfação/insatisfação percecionados da literatura analisada. No Quadro 2.4 pode-se observar os tópicos a criar e os autores que os abordaram nos seus estudos, bem como o ano de publicação e o método utilizado.

Quadro 2.4 - Tópicos a criar e investigação subjacente

Tópicos	Autores que consideraram o Tópico como determinante da satisfação/insatisfação, Ano, Método
'Value'	(Srivastava & Kumar, 2021, STM), (L. Zhou et al., 2014, Codificação manual), (Song et al., 2022, LDA)
'Facilities'	(Hu et al., 2019, STM), (Fu et al., 2022, Análise de frequência/CONCOR/fatorial)
'Parking'	(Srivastava & Kumar, 2021, STM), (Li et al., 2013, Revisão de literatura)
'Recreation'	(Srivastava & Kumar, 2021, STM), (Hu et al., 2019, STM)
'Room_Issue'	(L. Zhou et al., 2014, Codificação manual), (Song et al., 2022, LDA)
'Bathroom_Issue'	(L. Zhou et al., 2014, Codificação manual), (Hu et al., 2019, STM)
'Service_Encounter'	(Srivastava & Kumar, 2021, STM), (Fu et al., 2022, Análise de frequência/CONCOR/fatorial)
'Booking_Issue'	(Srivastava & Kumar, 2021, STM), (Hu et al., 2019, STM)
'Cleanliness_Maintenance'	(Song et al., 2022, LDA), (Nilashi et al., 2021, LDA)
'Wifi'	(Hu et al., 2019, STM), (Li et al., 2013, Revisão de literatura)
'TV'	(Li et al., 2013, Revisão de literatura)
'Catering'	(Srivastava & Kumar, 2021, STM), (Hu et al., 2019, STM)
'Location_Access'	(Hu et al., 2019, STM), (L. Zhou et al., 2014, Codificação manual), (Song et al., 2022, LDA)
'Lighting_Issues'	(Hu et al., 2019, STM)
'Smelling_Issues'	(Srivastava & Kumar, 2021, STM)
'Air_Conditioning'	(Li et al., 2013, Revisão de literatura)
'Noise_Issues'	(Srivastava & Kumar, 2021, STM), (Hu et al., 2019, STM)
'Safety_Privacy'	(Sann et al., 2022, Codificação manual)

Porém, 260 *tokens* não correspondiam a nenhum dos tópicos definidos ou teriam mais que uma interpretação em frase, pelo que foram removidos de análise.

Assim, para a construção dos tópicos foram utilizados 240 *tokens* dos 500 mais frequentes do corpus. Na Figura 2.5, pode-se observar um *wordcloud* que reflete esses 240 *tokens* cujo tamanho e intensidade refletem a sua frequência no corpus. Deste modo, neste *wordcloud* estão apenas incluídos os 240 *tokens* que foram distribuídos por tópicos, sendo que aqueles que têm maior destaque, em termos de tamanho e intensidade da letra, foram os mais frequentes do corpus: “room”; “small”; “breakfast”; “bathroom”; “time”; ”bed”; “staff”; “noise”; …

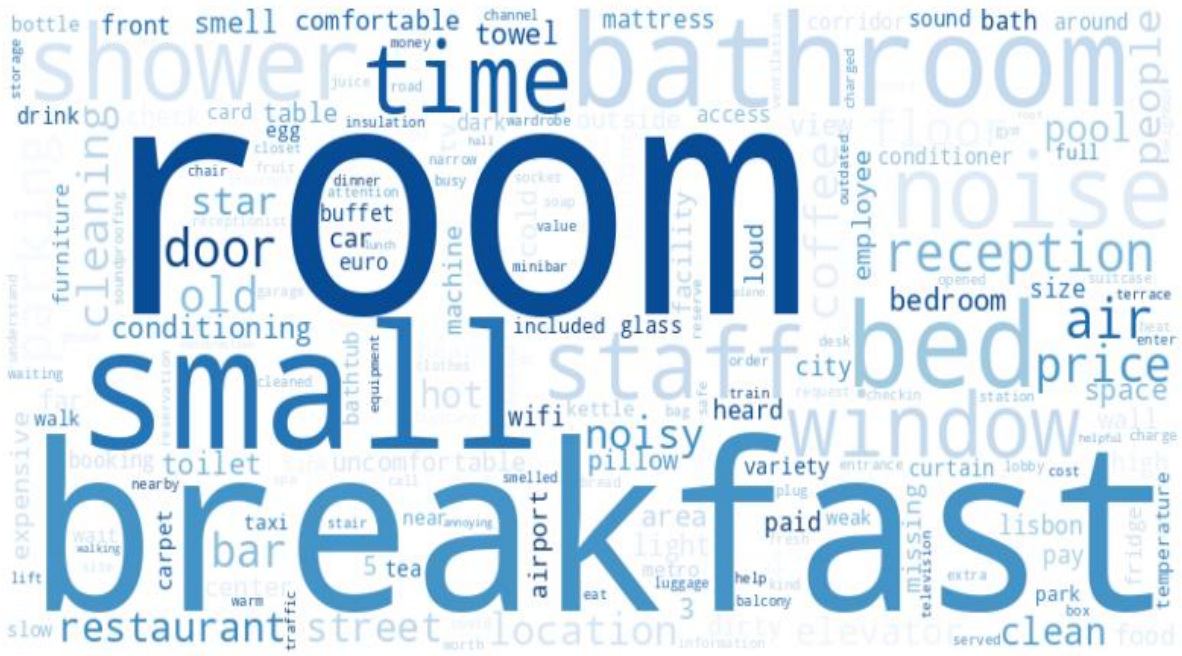

Figura 2.5 – *Wordcloud* dos *tokens* mais frequentes

Já no Quadro 2.5 podemos observar a identificação dos tópicos, e a distribuição dos *tokens* por esses mesmos tópicos.

Por conseguinte, as variáveis foram codificadas manualmente através dos tópicos referidos por meio de variáveis booleanas. Isto é, por exemplo, se um comentário referir o *token* ‘money’, contido no tópico “Value”, a variável com o mesmo nome retornará o valor ’True’ para esse registo. Caso o comentário não refira esse *token*, nem nenhum daqueles contidos no tópico “Value”, a variável retornará o valor ‘False’.

Assim, foi possível obter 18 variáveis booleanas (além da variável *target* “Hotel_Class”), que referem se cada comentário atribuído a uma classe de hotel (seja de hotel de classe alta ou hotel de classe baixa) aborda tal tópico, ou não.

Quadro 2.5 - Os tópicos e os *tokens* subjacentes

Value	Facilities	Parking
'average', 'worth', 'charged', 'money', 'extra', 'weak', 'euro', 'pay', 'paid', 'high', '5', '3', '4', 'star', 'price', 'expensive', 'cost', 'charge', 'value', 'worst'	'ground','rooftop', 'hall', 'enter', 'stair', 'decoration', 'dated', 'outdated', 'construction', 'terrace', 'lift', 'entrance', 'machine', 'building', 'old', 'floor', 'elevator', 'facilities', 'facility', 'corridor', 'lobby', 'roof', 'courtyard', 'hallway'	'parking', 'car', 'garage'
Recreation	**Room_Issue**	**Bathroom_Issue**
'pool','spa', 'gym'	'pillow', 'sheet', 'sofa', 'key', 'wardrobe', 'storage', 'chair', 'minibar', 'plug', 'clothes', 'accommodation', 'closet', 'desk', 'carpet', 'door', 'size', 'narrow', 'mattress', 'table', 'wall', 'window', 'small', 'bed', 'room', 'comfortable', 'furniture', 'curtain', 'socket', 'fridge', 'uncomfortable', 'equipment', 'space', 'bedroom', 'area', 'refrigerator', 'bedding', 'mirror'	'towel', 'sink', 'bathtub', 'dryer', 'bath', 'toilet', 'bathroom', 'shower', 'gel', 'shampoo', 'soap'
Service_Encounter	**Booking_Issue**	**Cleanliness_Maintenance**
wait', 'helpful', 'speak', 'provided', 'lady', 'informed', 'rude', 'help', 'receptionist', 'employee', 'reception', 'staff', 'time', 'check', 'slow', 'attention', 'portuguese', 'ordered', 'english', 'requested', 'welcome', 'busy', 'kind', 'understand', 'opened', 'information', 'luggage', 'call', 'checkin', 'suitcase', 'request', 'covid', 'bag', 'phone', 'waiting', 'balcony'	'card', 'booking', 'missing', 'reserved', 'site', 'reserve', 'included', 'reservation'	'cleaning', 'clean', 'cleanliness', 'glass', 'maintenance', 'cleaned', 'dirty'
Wifi	**TV**	**Catering**
'internet', 'wifi'	'channel', 'box', 'tv', 'television'	'coffee', 'taste', 'milk', 'lunch', 'served', 'fresh', 'eat', 'egg', 'kettle', 'variety', 'coffe', 'breakfast', 'restaurant', 'food', 'bar', 'buffet', 'dinner', 'drink', 'juice', 'fruit', 'tea', 'bottle', 'order', 'bread', 'menu', 'dish', 'cup', 'catering'
Location_Access	**Lighting_Issues**	**Smelling_Issues**
'taxi', 'subway', 'central', 'train', 'nearby', 'access', 'located', 'walk', 'park', 'around', 'front', 'lisbon', 'airport', 'far', 'view', 'location', 'street', 'center', 'city', 'near', 'metro', 'road', 'station', 'walking'	'lighting', 'light', 'dark'	'smelled', 'sink', 'smelled', 'smell', 'mold'
Air_Conditioning	**Noise_Issues**	**Safety_Privacy**
'cool', 'heating', 'heat', 'warm', 'ac', 'temperature', 'conditioner', 'cold', 'conditioning', 'hot', 'air', 'ventilation'	'music', 'noise', 'annoying', 'plane','insulation', 'soundproofing', 'heard', 'outside', 'hear', 'noisy', 'loud','sound', 'traffic'	'privacy', 'safe','full', 'people', 'neighbor'

Por fim, a terminar a fase de preparação dos dados para modelação, foram removidos os registos que retornavam o valor 'False' em todas as variáveis independentes, pois esses registos apresentam queixas de fatores que não têm relevância para análise ou porque representam comentários que referem que não têm razões ou motivos de queixa do hotel (muito frequente no corpus), pelo que estes registos foram eliminados graças a este último passo.

Depois deste processo seguiram para entrada na fase de modelação 132 533 entradas e cujos dados já se apresentam completamente estruturados em 18 variáveis independentes booleanas e a variável dependente *target*: "Hotel_Class" (tornada igualmente booleana). Note-se que foram retiradas, no fim da preparação dos dados, as variáveis não estruturadas, os comentários negativos, e as listas de *tokens* que correspondiam aos comentários pós-processamento.

No Quadro 2.6 podem-se observar, a título de exemplo, cinco comentários negativos de *reviews* e a respetiva estruturação pós processo de TM. Apenas de notar que para a concretização deste quadro mudou-se o tipo de variáveis de booleano para binário por questões de apresentação estética.

Quadro 2.6 - Exemplo de 5 comentários pós estruturação

negative_comment	Hotel_Class	Value	Facilities	Parking	Recreation	Room_Issue	Bathroom_Issue	Service_Encounter	Booking_Issue	Cleanliness_Maintenance	Wifi	TV	Catering	Location_Access	Lighting_Issues	Smelling_Issues	Air_Conditioning	Noise_Issues	Safety_Privacy
- a bit expensive	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
- A noise was heard throughout the night. It looked like an extractor or something.	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
- bed was not that confortable - bathroom a bit small	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
the only thing to mention is that it was very hot inside the rooms	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
Room was too small as compare to the price we have paid	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

2.4. Modelação

Na fase de Modelação pretende-se a satisfação do objetivo de DM referido na fase de Compreensão do problema - descrever a distinção entre reclamações dirigidas a hotéis de classe alta de reclamações dirigidas a hotéis de classe baixa - através da execução das seguintes tarefas: seleção das técnicas de modelação, definição dos mecanismos de teste e a construção dos modelos.

É de relembrar que o conjunto de dados inclui as 18 variáveis independentes booleanas, que se referem aos tópicos criados (ver o significado de cada variável no Quadro 2.7), mais a variável dependente, a “Hotel_Class”.

Quadro 2.7 - Significado das variáveis independentes

Variáveis	Significado
Value	Verdadeiro: Se o comentário negativo refere valores dispendidos no serviço ou se abordam a relação qualidade-preço | Falso: Caso
Facilities	Verdadeiro: Se o comentário negativo refere as condições das instalações do hotel | Falso: Caso contrário
Parking	Verdadeiro: Se o comentário refere problemas com estacionamento, carro ou garagem | Falso: Caso contrário
Recreation	Verdadeiro: Se o comentário refere problemas com atividades de recriação como piscina, spa, ginásio, ou até mesmo a falta delas | Falso:
Room_Issue	Verdadeiro: Se o comentário apresenta desagrado com o quarto do hotel, seja tamanho, comforto, mobília, … | Falso: Caso contrário
Bathroom_Issue	Verdadeiro: Se o comentário apresenta desagrado com a casa de banho, seja do duche, gel disponível, secador, toalhas, … | Falso: Caso contrário
Service_Encounter	Verdadeiro: Se o comentário refere dificuldades de comunicação com o staff ou desagrado com o serviço de receção | Falso: Caso contrário
Booking_Issue	Verdadeiro: Se o comentário refere erros nas reservas ou falhas nos pagamentos | Falso: Caso contrário
Cleanliness_Maintenance	Verdadeiro: Se o comentário menciona falta de limpeza, manutenção ou sujidade | Falso: Caso contrário
Wifi	Verdadeiro: Se o comentário relata problemas de conexão à Internet|
TV	Verdadeiro: Se o comentário refere a falta de TV ou demonstra desagrado com os canais disponibilizados | Falso: Caso contrário
Catering	Verdadeiro: Se o comentário menciona o serviço de comidas e bebidas ou de pequeno almoço | Falso: Caso contrário
Location_Access	Verdadeiro: Se o comentário relata desagrado com a localização do hotel ou com os seus acessos | Falso: Caso contrário
Lighting_Issues	Verdadeiro: Se o comentário menciona falta de iluminação, seja nas instalações dos hotéis ou arredores | Falso: Caso contrário
Smelling_Issues	Verdadeiro: Se o comentário reporta cheiro desagradáveis, seja nas instalações do hotel ou nas suas intermediações | Falso: Caso contrário
Air_Conditioning	Verdadeiro: Se o comentário menciona o ambiente ou a temperatura do hotel ou problemas de ar condicionado| Falso: Caso contrário
Noise_Issues	Verdadeiro: Se o comentário reporta barulhos seja de outros hóspedes ou mesmo das intermediações do hotel | Falso: Caso contrário
Safety_Privacy	Verdadeiro: Se o comentário aborda temas como a privacidade, segurança ou até refere se o hotel está cheio de hóspedes | Falso: Caso

De modo a aferir as técnicas de modelação a utilizar e, uma vez que o problema é de classificação e que são pretendidas técnicas simbólicas com vista à descrição e análise dos modelos, os algoritmos a selecionar devem ser baseados em árvores de decisão. Para isso será utilizada a biblioteca *scikit-learn*, integrada em *Python*, e que foi desenvolvida especificamente para aplicação prática de ML (Pedregosa *et al*., 2011). O algoritmo

DecisionTreeLearner, presente na biblioteca *scikit-learn* é um algoritmo baseado em árvores de decisão que se equivale ao CART (já referido na Revisão da literatura).

Noutro sentido, de modo a atestar a potencialidade preditiva do modelo construído pelo algoritmo *DecisionTreeLearner* também foram realizados testes nesta fase com outros tipos de algoritmos, nomeadamente o *Random Forest* (*RandomForestClassifier*), o *XGBoost* (*XGBClassifier*) e Redes Neuronais (*MLPClassifier*), que são mais conhecidos pelo seu desempenho preditivo que árvores de decisão. Apesar dos algoritmos mencionados por último não se tratarem de técnicas simbólicas ou que permitam o cumprimento do objetivo no que toca à descrição de um modelo, é sempre importante efetuar testes com diferentes algoritmos para o mesmo conjunto de dados até mesmo para aferir o potencial preditivo possível para o *dataset* em questão.

Em termos de mecanismos de teste, ficou definida a divisão *holdout* (C3.ai Inc., 2023) do conjunto dos dados (a que permitia a melhor obtenção de resultados): 70% dos dados são destinados ao treino do modelo, enquanto que 30% dos dados são testados de modo a averiguar a performance do modelo. É de notar que o método *holdout* foi escolhido sobretudo pela grande quantidade de dados e que as percentagens atribuídas aos conjuntos foram selecionadas, após várias tentativas, a fim de obter os melhores resultados em termos de acurácia do modelo com o conjunto de teste.

Já o processo de construção do modelo que melhor servisse à investigação foi bastante iterativo no sentido de atingir, entre várias tentativas, um modelo simplificado, com conclusões coerentes e que não desprimorasse o desempenho na previsão (principalmente em termos de acurácia). Para ajudar a definir os hiperparâmetros ótimos da árvore de decisão, foi utilizado o método *GridSearchCV* que efetua uma pesquisa sobre valores de parâmetros especificados para um estimador de modo a maximizar a acurácia do modelo (Pedregosa *et al.*, 2011).

Os parâmetros sujeitos a esta busca foram: o 'max_depth' - controla a profundidade máxima da árvore de decisão -; o 'min_samples_split' - define o número mínimo de amostras necessárias para que um nó interno da árvore possa ser dividido -; o 'min_samples_leaf' - determina o número mínimo de amostras que um nó folha da árvore deve ter -; e o 'criterion' - especifica a medida usada para avaliar a qualidade das divisões nos nós da árvore - (Pedregosa *et al.*, 2011).

A visualização e a análise do modelo selecionado serão expostas em capítulo posterior, nos Resultados e discussão.

2.5. Avaliação

A capacidade de classificação dos algoritmos pode ser avaliada usando vários indicadores de desempenho e representa um papel crucial em projetos de DM. A seleção de métricas apropriadas é considerada tão crucial quanto selecionar o algoritmo de aprendizagem para um determinado problema de classificação (Mathew *et al*., 2023).

Assim sendo, a avaliação da qualidade do modelo de classificação terá em consideração diversas métricas baseadas na matriz de confusão apresentada, ideologicamente, no Quadro 2.8 (Visa *et al.*, 2011).

Quadro 2.8 - Matriz de confusão ideológica (adaptado de Visa *et al.*, 2011)

Classe observada	**Classe prevista (pelo modelo)**	
	Hotel classe baixa	Hotel classe alta
Hotel classe baixa	a	b
Hotel classe alta	c	d

Em que “a” representa os casos corretamente previstos de comentários associados a hotéis de classe baixa; “b” representa os comentários erradamente associados a hotéis de classe alta; “c” representa os comentários erradamente associados a hotéis de classe baixa; e “d” que representa os casos corretamente previstos de comentários associados a hotéis de classe alta.

Em concreto, recorreu-se às métricas Acurácia, Sensibilidade e Especificidade demonstradas no Quadro 2.9.

Quadro 2.9 - Métricas de avaliação da qualidade do modelo

Métrica	**Fórmula**	**Significado**
Acurácia	$A = \frac{a+d}{a+b+c+d}$	A acurácia é a proporção de instâncias corretamente classificadas para ambas as classes de hotéis em relação ao total de instâncias.
Sensibilidade	$S = \frac{d}{c+d}$	A sensibilidade mede a capacidade do modelo em identificar corretamente os comentários relativos a hotéis de classe alta.
Especificidade	$E = \frac{a}{a+b}$	A especificidade mede a capacidade do modelo em identificar corretamente os comentários relativos a hotéis de classe baixa.

O motivo de se escolher a Acurácia como uma das métricas a utilizar tem a ver com a questão de ser de fácil interpretação e por indicar a precisão global do modelo. Já a Sensibilidade e a Especificidade foram as outras métricas escolhidas para avaliar o modelo visto que é através destas que é possível entender se o modelo tem mais dificuldade em classificar os comentários referentes a hotéis de classe alta ou a hotéis de classe baixa.

Sendo que é possível, assim, observar a confiabilidade da previsão associada a cada uma destas classes de hotel.

Os resultados desta fase serão posteriormente apresentados no capítulo Resultados e discussão e, além disso, será também apresentada uma breve avaliação dos modelos construídos pelos algoritmos não simbólicos mencionados na fase anterior.

2.6. Desenvolvimento

Esta fase final corresponde ao desenvolvimento da presente dissertação que sistematiza todo o processo, desde a averiguação do problema, à investigação e à procura de resultados que descrevam o problema definido.

A análise e discussão dos resultados, incluindo a exploração do modelo selecionado serão efetuadas no capítulo seguinte, nos Resultados e discussão. Já as contribuições da investigação tanto para a teoria como para a prática e o elencar das principais conclusões, limitações e pesquisa futura serão apresentados no último capítulo, a Conclusão.

CAPÍTULO 3

Resultados e discussão

Este capítulo visa dar sequência à metodologia abordada no capítulo anterior evidenciando os seus resultados, interpretá-los e efetuar as respetivas análises tendo em conta o problema. Primeiramente, neste capítulo, serão explorados os dados que seguiram para a fase de Modelação (já depois da fase de Preparação). Em seguida, será apresentada a Avaliação do modelo final como também a sua análise e comparação com a literatura estudada no capítulo da Revisão da literatura. Por fim, serão dadas sugestões aos gerentes de hotéis da cidade de Lisboa com base nos resultados da presente investigação.

3.1. Exploração dos dados

Os dados que serão explorados correspondem àqueles que entraram na fase de Modelação. Isto é, já depois dos comentários negativos das *reviews* terem sido estruturados nas 18 variáveis booleanas que indicam se tais aspetos foram referidos no comentário subjacente. Além disso, é de relembrar que foram retirados do *dataset* os registos em que não foram acusados quaisquer dos aspetos em busca. Assim, serão explorados 132 533 registos de comentários negativos de *reviews* que se referem a 216 hotéis da cidade de Lisboa (126 hotéis de classe alta e 90 de classe baixa).

No gráfico presente na Figura 3.1 até se pode observar que apesar de a amostra estar subjacente a mais hotéis de classe alta que de hotéis de classe baixa, a verdade é que até se encontram no *dataset* mais *reviews* de comentários negativos dirigidos a hotéis de classe baixa do que dirigidos a hotéis de classe alta. Neste sentido, pode-se concluir, como é natural, que foram dirigidas mais reclamações *online* a hotéis de classe baixa que a hotéis de classe alta.

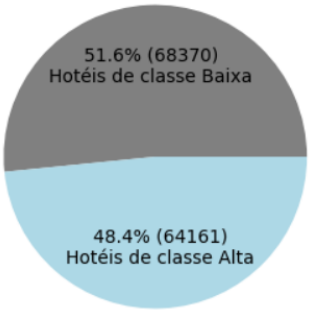

Figura 3.1 - Distribuição dos comentários negativos das *reviews* por classe de hotel

Passando à exploração das 18 variáveis que caraterizam os comentários negativos, podem-se observar as distribuições de cada uma na Figura 3.2.

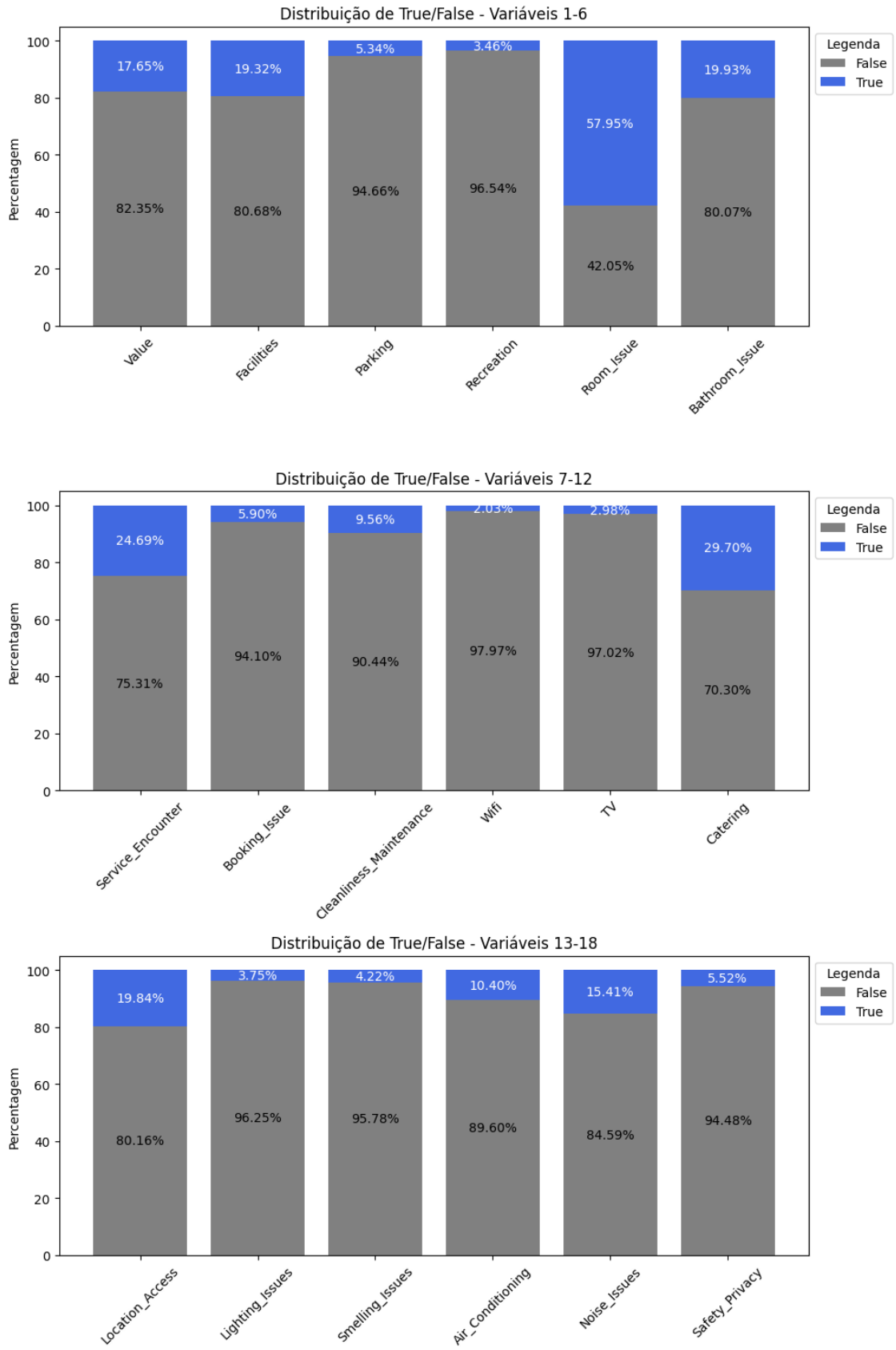

Figura 3.2 - Distribuição das variáveis independentes

O aspeto que teve maior frequência, o que surgiu em mais *reviews* negativas, foi o 'Room_Issue' (foi identificado em cerca de 58% dos comentários). Ou seja, mais de metade dos comentários negativos da amostra referiram que tiveram problemas com o quarto do hotel, tenha tido a ver com a cama, a área do quarto, arrumação ou outro problema a envolver o quarto.

Outros aspetos que foram bastante mencionados nas *reviews* foram:

- O 'Catering' (presente em cerca de 30% dos comentários) – desaprovação dos serviços de refeição dos hotéis;
- O 'Service_Encounter' (25%) – críticas às interações com o *staff* e serviços de receção;
- O 'Bathroom_Issue' (20%) – problemas com a casa de banho, tenha tido a ver com a sanita, o duche ou outro problema a envolver a casa de banho;
- A 'Location_Access' (20%) – reprovação da localização e dos acessos dos hotéis; e
- As 'Facilities' (20%) – críticas às instalações do hotel .

Em sentido contrário, os aspetos de crítica menos mencionados (aqueles que foram referidos em menos de 5% das *reviews*) foram:

- os 'Smelling_Issues' (apenas cerca de 4% dos comentários referiram este aspeto) – reporte de cheiros desagradáveis;
- os 'Lighting_Issues' (4%) – reporte de falta de iluminação; a 'Recreation' (3%) – problemas em recreações hoteleiras (como a piscina, spa ou ginásio) ou a falta ou encerramento delas;
- a 'TV' (3%) – críticas à falta de TV ou desaprovação da lista de canais disponível;
- e o 'Wifi' (2%) – reporte de falhas na ligação à *Internet*.

Também se pode observar a medida de associação entre as variáveis através da matriz de associações presente no *heatmap* da Figura 3.3. Uma vez que era intuito medir a força de associação entre variáveis booleanas, foi utilizado o coeficiente V de Cramér. Quanto maior for este coeficiente, mais forte é a associação entre as variáveis em teste.

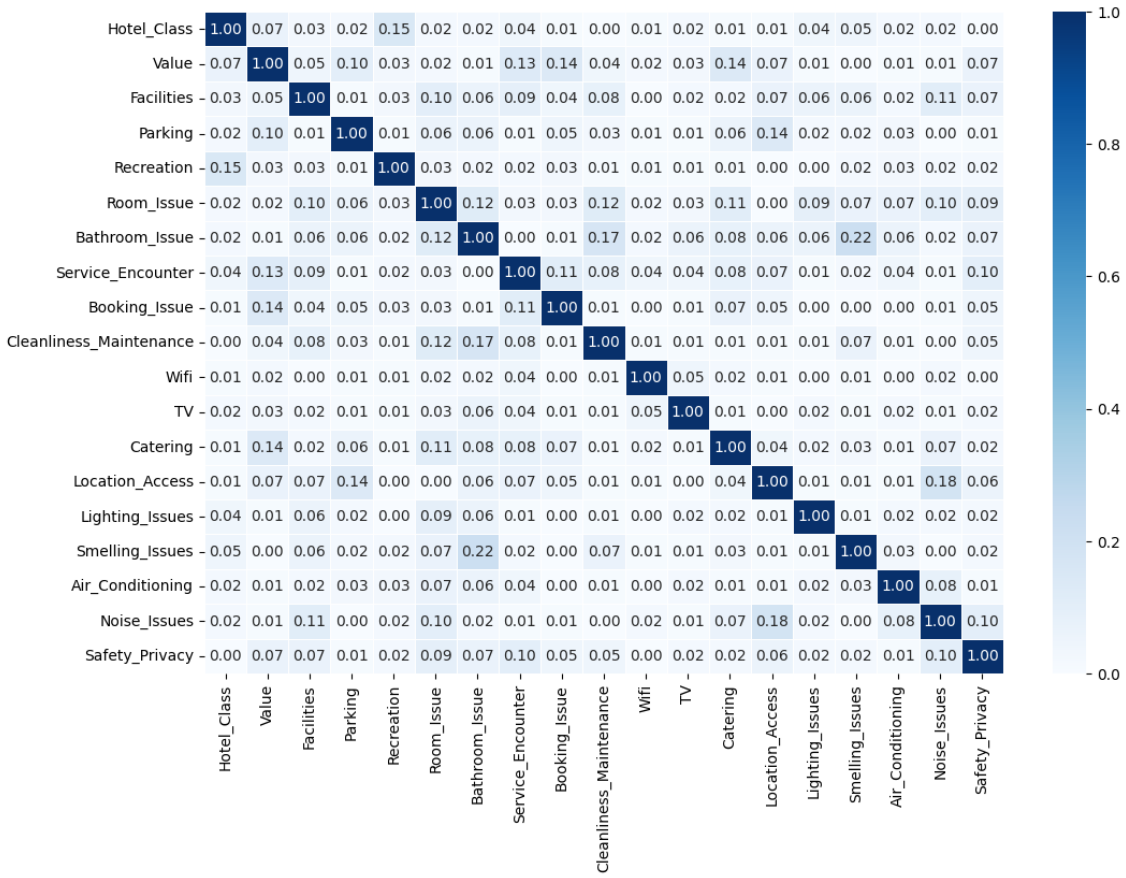

Figura 3.3 - Matriz de associações V de Cramér

É possível concluir que as associações entre todas as variáveis são fracas ou quase nulas. Em relação à variável *target*, a 'Hotel_Class', é de notar que a associação menos fraca foi de 0.15 em relação à variável 'Recreation'. No entanto, houve três pares de variáveis que obtiveram maiores associações (mesmo assim fracas): 'Bathroom_Issues' com 'Smelling_Issues' (0.22) ou com 'Cleanliness_Maintenance' (0.17) - pode significar que problemas com cheiros ou falta de limpeza estão muitas vezes conotados com problemas com a casa de banho e canalizações -; e 'Location_Access' com 'Noise_Issues' (0.18) - os problemas com barulho podem estar muitas vezes associados à má localização do hotel, ou por ter muito tráfego à volta, ou por se situar perto de alguma estação de transportes ou de outros locais com poluição sonora.

3.2. Apresentação do modelo

Após várias tentativas, buscas e análises de modelos, o modelo que melhor cumpriu para com os requisitos descritos anteriormente, na fase de Modelação, teve os seguintes hiperparâmetros: 'max_depth' = 7; 'min_samples_split' = 35000; 'min_samples_leaf' = 3000; 'criterion' = 'gini'. A árvore de decisão ficou, assim, limitada a 7 níveis de profundidade, cada nó interior com um mínimo de 35 mil *samples* e cada folha com um mínimo de 3 mil *samples* (a árvore de decisão tem no total 7 folhas).

Pode-se visualizar na Figura 3.4 a representação em árvore do modelo selecionado.

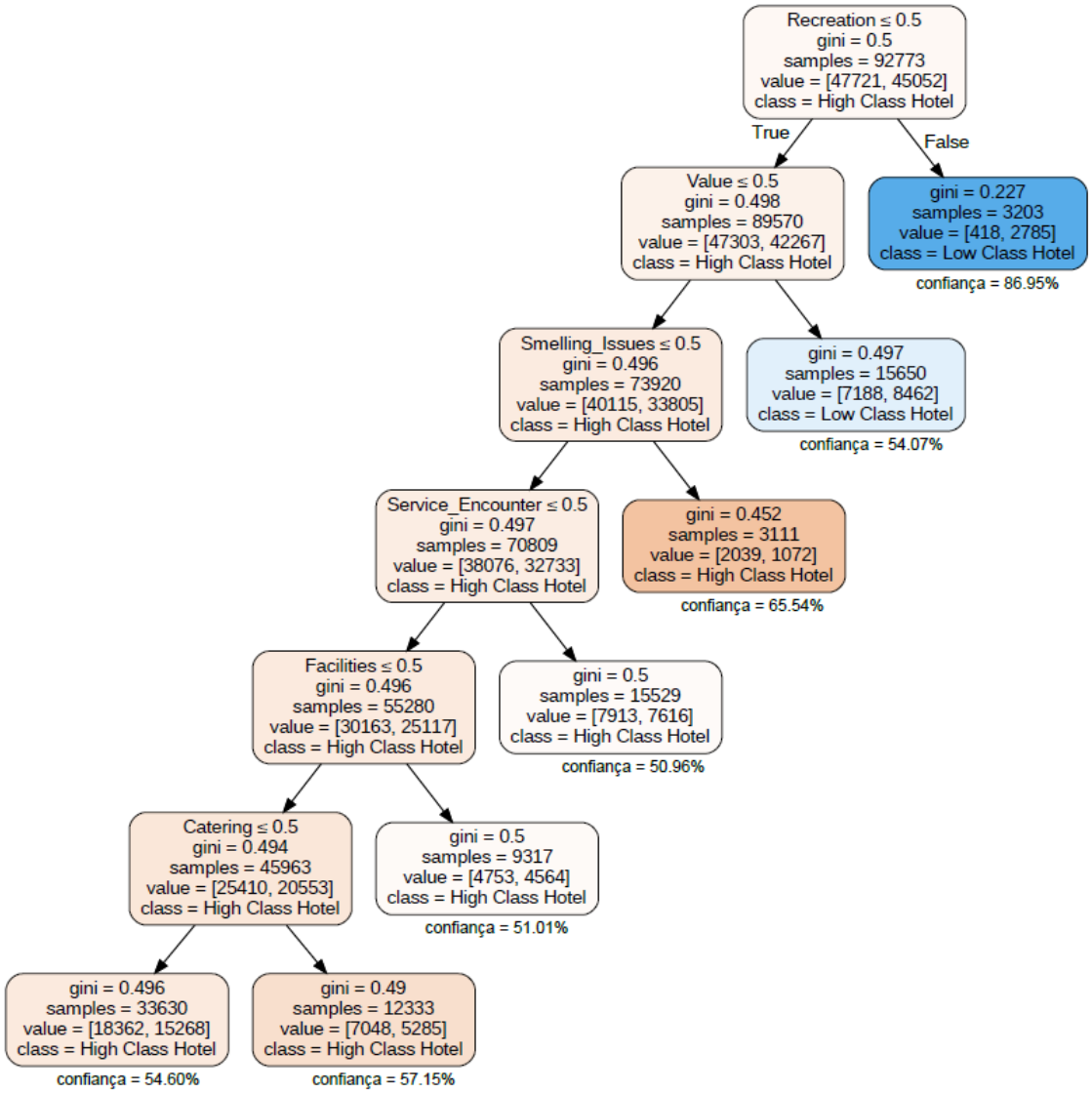

Figura 3.4 - Árvore de decisão

Na construção do modelo, das 18 variáveis independentes do *dataset*, aquelas que tiveram influência na construção do modelo e que se tornaram preditores, foram aquelas indicadas no Quadro 3.1, junto do seu nível de importância no modelo (as restantes variáveis não tiveram qualquer relevância no modelo). É de referir, assim, que os preditores mais importantes na construção do modelo foram, sobretudo, a 'Recreation' (que também foi a variável com maior correlação com o *target*), o 'Value' e o 'Smelling_Issues'. De forma mais secundária, tiveram ainda alguma importância na construção do modelo as variáveis 'Service_Encounter', 'Facilities' e 'Catering'.

Quadro 3.1 - Preditores e respetiva importância no modelo

Preditores	Importância
'Recreation'	0.745575
'Value'	0.136950
'Smelling_Issues'	0.062946
'Service_Encounter'	0.024065
'Facilities'	0.021534
'Catering'	0.008930

3.3. Avaliação do modelo

Em seguida, é apresentada a matriz de confusão (Quadro 3.2) que permite visualizar a performance do modelo de maneira mais detalhada, mostrando como as previsões do modelo se comparam com os valores reais dos dados.

Quadro 3.2 - Matriz de confusão

Classe observada	Classe prevista (pelo modelo)	
	Hotel classe baixa	Hotel classe alta
Hotel classe baixa	17435	3215
Hotel classe alta	14275	4835

É possível desde logo notar que foram previstos muitos comentários associados a hotéis de classe baixa, aspeto confirmado quando temos em consideração os valores de sensibilidade, 25,30%, e de especificidade, 84,43%. Por conseguinte, o modelo tem muito mais dificuldade em identificar os comentários relativos a hotéis de classe alta, que os comentários relativos a hotéis de classe baixa.

Já o valor da acurácia de 56,01% é indicativo de um modelo com um desempenho moderadamente fraco no contexto de classificação em duas categorias, uma vez que está a classificar corretamente pouco mais de metade das previsões e que as classes até são muito equilibradas. Pode verificar o desempenho do modelo no Quadro 3.3.

Quadro 3.3 - Indicadores de desempenho do modelo simbólico

Indicador de desempenho	Valor (%)
Acurácia	56,01%
Sensibilidade	25,30%
Especificidade	84,43%

O problema do modelo, como já referido, reside no baixo valor da sensibilidade - na capacidade de prever corretamente os comentários relativos a hotéis de classe alta. No entanto, ao tentar alterar a profundidade da árvore, ou criar mais variáveis para a análise, ou fazer *tuning* dos hiperparâmetros ou até mesmo alterando as percentagens de *holdout*, os modelos ficavam cada vez mais difíceis de analisar sem ocorrerem melhorias significativas em termos de indicadores de desempenho. Além de que, à medida que a sensibilidade pouco subia, a especificidade descia a pique em certas tentativas de aumentar a complexidade da árvore.

Deste modo, é possível afirmar por estes indicadores que o algoritmo de árvores de decisão tem dificuldade em conseguir classificar comentários negativos de *reviews online* e que, apesar de se tratar de uma técnica simbólica cujo ponto forte é a facilidade de interpretar, e cujo ponto fraco costuma ser a capacidade preditiva, a verdade é que para o conjunto de dados em questão nenhum dos outros algoritmos testados na fase de modelação (*Random Forest*, *XGBoost* e *Redes Neuronais*) obtiveram bons desempenhos preditivos (ou pelo menos desempenhos substancialmente melhores que o *CART*), como se pode observar no Quadro 3.4.

Quadro 3.4 - Indicadores de desempenho dos modelos não simbólicos

Modelo	Acurácia (%)	Sensibilidade (%)	Especificidade (%)
Random Forest	55,55%	40,92%	69,24%
XGBoost	56,27%	38,37%	73,02%
Redes Neuronais	56,29%	41,95%	69,71%

Dos algoritmos não simbólicos utilizados para atestar o potencial do *dataset,* aquele que teve melhor desempenho na previsão foi o *XGBoost*, que não teve um desempenho muito superior ao modelo baseado em árvores de decisão, especialmente quando em comparação estão os valores das acurácias dos modelos (*CART*: 56,01% e *XGBoost*: 56,27%).

Por um lado, o algoritmo baseado em árvores de decisão até é mesmo o mais capaz em prever os comentários negativos relativos a hotéis de classe baixa, sendo o modelo com maior valor de especificidade. Por outro lado, é o modelo com menor capacidade de prever comentários negativos relativos a hotéis de classe alta, pelo menor valor de sensibilidade de entre os modelos.

3.4. Interpretabilidade do modelo

Apesar do fraco desempenho na capacidade de predição, como o modelo selecionado para interpretar foi criado a partir de uma técnica simbólica - a árvore de decisão (cuja representação pode-se aferir na Figura 3.4) -, é possível interpretar o modelo selecionado e extrair regras que responderão ao problema.

Deste modo, pode-se testemunhar o efeito da interação dos preditores na classificação e, assim, averiguar qual é a diferença entre as *reviews online* negativas que um hotel de classe alta recebe, das de um hotel de classe baixa.

No Quadro 3.5, é possível visualizar as regras extraídas da árvore de decisão (Figura 3.4) que servirão para análise (foram selecionadas aquelas que teriam sentido de analisar ou que teriam um nível de confiança aceitável). Do lado esquerdo estão as regras que perfilam os tipos de reclamações que hotéis de classe baixa recebem nas *reviews online*, e do lado direito perfilam os tipos de reclamações que hotéis de classe alta recebem nas *reviews online*.

Quadro 3.5 - Regras extraídas da árvore de decisão

Regras extraídas da árvore de decisão	
Hotéis de classe baixa	**Hotéis de classe alta**
Regra 1: 'Recration' = 'True' (confiança = 86.95%; suporte = 3203)	Regra3: 'Recreation' = 'False' & 'Value' = 'False' & 'Smelling_Issues' = 'True' (confiança = 65.54%; suporte =3111)
Regra2: 'Recration' = 'False' & 'Value' = 'True' (confiança = 54.07%; suporte = 15650)	Regra 4: 'Recreation' = 'False' & 'Value' = 'False' & 'Smelling_Issues' = 'False' & 'Service_Encouter' = 'False' & 'Facilities' = 'False' & 'Catering' = 'True' (confiança = 57.15%; suporte =12333)
-	Os restantes ramos não têm interesse pela sua interpretabilidade ou pelo seu fraco nível de confiança

De acordo com a Regra 1, caso o comentário negativo da *review online* aborde o aspeto ‘Recreation’, o modelo garante, com 86.95% de confiança e com um suporte de 3203 instâncias, que esse comentário é dirigido a um hotel de classe baixa.

Segundo a Regra 2, caso o comentário negativo da *review online* não aborde o aspeto ‘Recreation’, mas aborde o aspeto ‘Value’, o modelo garante, com 54.07% de confiança e

com um suporte de 15650 instâncias, que esse comentário é igualmente dirigido a um hotel de classe baixa.

Conforme a Regra 3, caso o comentário negativo da *review online* não aborde os aspetos 'Recreation' e 'Value', mas aborde o aspeto 'Smelling_Issues' o modelo garante, com 65.54% de confiança e com um suporte de 3111 instâncias, que esse comentário é dirigido a um hotel de classe alta.

Consoante a Regra 4, caso o comentário negativo da *review online* não aborde os aspetos 'Recreation', 'Value', 'Smelling_Issues', 'Service_Encouter' e 'Facilities', mas aborde o aspeto 'Catering' o modelo garante, com 57.15% de confiança e com um suporte de 12333 instâncias, que esse comentário é dirigido a um hotel de classe alta.

Deste modo, os hotéis de classe baixa (hotéis de 1,2 ou 3 estrelas) recebem comentários negativos acerca de problemas com atividades de recreação ou até mesmo a falta delas, isto é, reclamações a ter a ver com a piscina, spa, ginásio, ou a falta deste tipo de serviços. Pode entender-se que há uma certa expetativa em relação aos hóspedes de hotéis de classe baixa em relação às atividades recreativas deste tipo de hotéis que não estão a ser cumpridas.

Adicionalmente, associadas a este tipo de hotéis estão críticas relacionadas aos valores despendidos nas estadias dos hotéis. Nos comentários negativos das *reviews online* são abordados temas correferidos da relação qualidade-preço dos serviços prestados nos hotéis. Apesar da opção de *booking* por hotéis de classe baixa, a verdade é que os hóspedes deste tipo de hotéis ainda não ficam satisfeitos em relação aos preços praticados por estes hotéis.

Já os hotéis de classe alta (hotéis de 4 ou 5 estrelas) não costumam receber críticas em relação à falta de atividades recreativas ou em relação a problemas neste aspeto, nem no que diz respeito aos valores despendidos, nem tão pouco referem problemas com as interações com o *staff* ou com as instalações de luxo. Pode-se verificar, assim, que as expetativas dos hóspedes, no que concerne a estes aspetos - atividades recreativas, relação qualidade-preço, interações com o *staff* e instalações -, são cumpridas em hotéis de luxo. Isto pode dever-se ao facto de hotéis de classe alta da cidade de Lisboa terem várias alternativas em termos de atividades recreativas, como piscina, spa ou ginásio (e em bom funcionamento) e instalações de luxo para uma boa relação qualidade-preço em relação a outras cidades da Europa. Além disso, estes resultados provam que as equipas de *staff* dos hotéis de luxo lisboetas exercem polidez, cortesia, e lidam positivamente com os problemas dos hóspedes e promovem boas interações sociais com os mesmos.

No entanto, há um foco de *reviews online* de hotéis de classe alta que reportaram cheiros desagradáveis ou problemas no serviço de *catering*, seja acerca da qualidade da comida, serviço inadequado ou a falta de opções de comidas ou bebidas no menu. O facto

de ocorrerem várias críticas ao serviço alimentar pode dever-se ao choque cultural entre hóspedes estrangeiros, como referido por Mariani e Matarazzo (2021), ou pela alta expetativa (não correspondida) que se deposita na comida e bebida de hotéis de luxo.

No que compreende à comparação com estudos anteriores que procuravam responder ao mesmo problema (já referidos no capítulo Revisão da literatura), as suas conclusões são um pouco divergentes em relação à presente investigação.

Para hotéis de classe baixa os estudos de Hu *et al.* (2019) como de Sann *et al.* (2022) referiam que os clientes incomodavam-se com problemas mais tangíveis, nomeadamente limpeza ou falta de condições nas instalações dos hotéis, embora os seus estudos não contemplassem nenhuma variável que retratasse a variável 'Recreation' como acontece neste estudo. Adicionalmente, no estudo de Sann *et al.* (2022) é mesmo referido que em hotéis de classe baixa as reclamações não abordam o preço ou valor, conclusão completamente contrária à do presente estudo. Praticamente, isto pode dever-se à diferença de contextos, ou seja, de um lado está um estudo realizado a *reviews* de hotéis do Reino Unido, já a presente investigação analisa *reviews* de hotéis da cidade de Lisboa.

Já para hotéis de classe alta, os estudos descritos anteriormente referem que questões de preço e prestação de serviços são os principais temas de reclamação nos hotéis de luxo, enquanto que na presente investigação, como já referido, os tópicos são o *catering* e o reporte de cheiros desagradáveis, e a não referência de preços ou de interações com o *staff,* resultado contrário à da literatura.

3.5. Sugestões para gerentes de hotéis

Como já referido, compreender as causas das reclamações dos clientes é fundamental para que os hotéis melhorem a qualidade dos seus serviços, a satisfação do cliente e o seu desempenho financeiro (Hu *et al*., 2019).

Assim, a perceção destes *insights* pode capacitar os gerentes dos hotéis a estabelecer respostas ou estratégias apropriadas para lidar com as reclamações dos clientes. Consequentemente, a gestão das reclamações dos hóspedes deverá tornar-se mais eficaz ao serem propostos os valores e prioridades para cada tipo de cliente de modo a também gerir as suas expetativas.

Assim, devem ser traçadas estratégias de gestão das reclamações dos clientes. Segundo os resultados deste estudo, os hotéis de classe baixa devem ter atenção aos processos e instalações que envolvem o serviço das suas atividades recreativas. Por um lado, dos gerentes dos hotéis devem ter atenção ao bom funcionamento das piscinas ou dos ginásios e, por outro lado, os gerentes destes hotéis devem fazer um esforço para organizar ou proporcionar atividades recreativas, de qualidade, para os seus hóspedes. Uma sugestão

pode decorrer da formalização de uma parceria com agências de turismo de forma a aprontar passeios turísticos ou a vender ingressos de atividades recreativas fora do hotel.

Além disso, os hotéis de classe baixa da cidade de Lisboa devem ter em atenção aos seus valores praticados e não cobrar valores desproporcionais à sua valia enquanto serviços. Por outras palavras, hotéis de 1, 2 ou 3 estrelas não se devem fazer cobrar como hotéis de luxo na consequência de ocorrer uma falha para com a expetativa do hóspede.

Já quanto aos hotéis de classe alta, hotéis de luxo, estes devem evitar ao máximo a questão que envolve cheiros desagradáveis. Houve um foco de reclamações em relação aos cheiros sentidos pelos hóspedes de hotéis de luxo, pelo que os gerentes destes hotéis devem ter em atenção as condições de canalizações, a limpeza e manutenção dos quartos e casas de banho e até mesmo cheiros que possam vir do exterior dos hotéis. Além disso, podem contactar a Câmara Municipal de Lisboa na tentativa de discutir a manutenção e limpeza das intermediações do hotel.

Noutro sentido, os hotéis de 4 ou 5 estrelas devem expandir o seu serviço de *catering* com melhores opções e mais variedade de modo a agradar hóspedes de todos os cantos do mundo, anular o fator da distância cultural em relação às refeições e, assim, cumprir para com as expetativas dos seus hóspedes.

Conclusão

A realização deste estudo visou enriquecer e expandir a literatura sobre a indústria hoteleira e turística, utilizando o ML de modo a entender o comportamento de insatisfação dos hóspedes. Em concreto, esta investigação tratou de responder ao problema definido que consistia na compreensão dos aspetos em que divergem reclamações dirigidas a hotéis de classe baixa de reclamações dirigidas a hotéis de classe alta, através de uma metodologia própria de problemas de DM - a CRISP-DM.

O conjunto de dados a modelar foi obtido por meio de técnicas de TM e da análise de frequência de *tokens* a um corpus que continha mais de 150 mil comentários negativos de *reviews online* dirigidas a hotéis da cidade de Lisboa listados no portal Booking.com e, pelo facto de se ter trabalhado com um número tão elevado de instâncias, considerou-se que o problema também envolveu BD. Durante este processo averiguou-se que em hotéis de classe baixa são realizados mais comentários negativos nas *reviews* que em hotéis de classe alta.

Aditivamente, através da revisão da literatura, este estudo definiu 18 aspetos de reclamação frequente e, por conseguinte, 18 variáveis independentes correspondentes. Já a modelação deste conjunto de dados (a partir de um algoritmo de árvore de decisão) à variável *target* deste estudo, a classe do hotel ao qual a *review* é dirigida, e a posterior análise do modelo simbólico formado fornecem aos hoteleiros uma compreensão mais precisa da insatisfação do cliente segundo diferentes classes de hotéis. Isto, apesar da performance preditiva do modelo selecionado ser considerada moderadamente baixa.

É ainda de notar que foram testados outros tipos de algoritmos não simbólicos na tentativa de aferir a potencialidade preditiva do *dataset*. No entanto, as métricas de desempenho desses modelos foram em linha com o modelo baseado em árvores de decisão, permitindo até concluir que o conjunto de dados em questão trata-se de uma limitação no que toca a obter melhores resultados em termos de desempenho preditivo.

Passando às conclusões retiradas a partir do modelo simbólico, para hotéis de classe baixa (hotéis de 1, 2 ou 3 estrelas), os gerentes de hotéis devem ter preocupação em proporcionar atividades recreativas de qualidade aos seus hóspedes. Se os hotéis de baixa classe conseguirem ter espaços como piscina, spa ou ginásio, ou conseguirem organizar passeios turísticos ou eventos, podem tornar-se vantagens competitivas de hotéis desta classe que vão de encontro com as expetativas dos seus hóspedes.

Além disso, os hotéis de classe baixa devem ter noção dos preços a que colocam os seus serviços e não fazerem-se cobrar a preços de hotéis de luxo, uma vez que é expectável para o hóspede padrão deste tipo de hotéis querer pagar pouco pela sua estadia.

Já para hotéis de classe alta ou hotéis de luxo (hotéis de 4 ou 5 estrelas) houve um foco de reclamações em relação a cheiros desagradáveis nas hospedagens. Isto pôde dever-se às infraestruturas dos hotéis de Lisboa serem velhas tais como as suas canalizações que podem, assim, causar sensações desagradáveis de olfato aos seus hóspedes. Fica a sugestão aos gerentes dos hotéis de luxo da cidade de Lisboa de ter atenção a este aspeto, efetuarem manutenções às suas infraestruturas, principalmente as canalizações de forma frequente, e sensibilizar e instruir o *staff* de formas de suster e mitigar este tipo de problemas.

Além disso, é necessário aos hotéis de luxo cumprir melhor para com as expetativas dos seus hóspedes em relação aos seus serviços de *catering*, seja melhorar a qualidade da comida ou a sua variedade também de modo a ter um menu mais inclusivo para outras culturas ou estilos de vida.

De um outro modo, ficou patente de acordo com os resultados do estudo que em hotéis de luxo da cidade de Lisboa o aspeto valor e as interações com o *staff*, não são fatores de relevo, resultados contrários àqueles concluídos por literatura anterior.

Limitações

Uma das limitações deste trabalho está relacionada com os próprios dados. Apesar dos comentários serem extraídos de campos das *reviews online* onde supostamente os avaliadores devem descrever somente os aspetos negativos das suas estadias nos hotéis, a verdade é que foi notado que vários avaliadores descreveram aspetos positivos da sua estadia no mesmo campo, o que pode tornar os resultados ligeiramente enviesados já que não é possível codificar manualmente cada comentário devido à grande quantidade de dados. De outra forma, os avaliadores quando referem um aspeto neste campo da *review*, podem não estar a mencioná-lo necessariamente de forma negativa, o que não foi tido em consideração no processo de estruturação dos comentários negativos. Ou seja, foi assumido que qualquer dos aspetos reconhecidos nos comentários negativos foi mencionado pelo avaliador de forma exclusivamente negativa.

Outro dos aspetos que limitou o estudo tratou-se do processo de criação das variáveis. Inicialmente, estaria prevista a utilização de técnicas de criação automática de tópicos, como a LDA ou a STM. No entanto, a fraca disposição de resultados destas técnicas (as variáveis programadas segundo os tópicos criados automaticamente, quando confrontadas com a análise manual dos comentários, davam muitos erros) fez com que se optasse por uma abordagem alternativa. Desta forma, já depois de realizado o pré processamento dos comentários, optou-se por agrupar os 500 *tokens* mais frequentes em tópicos que na literatura estariam relacionados com o tema da (in)satisfação na hospitalidade. Por

conseguinte, 260 *tokens* formaram 18 aspetos de reclamação de hotéis (sendo que a atribuição de cada *token* a um tópico pode ser um pouco subjetiva) e, por fim, foram criadas as 18 variáveis independentes correspondentes.

Já na fase de avaliação do modelo verificou-se que, apesar de o algoritmo de DM escolhido ser simbólico e, a partir do modelo gerado, ser possível retirar conclusões, na verdade, é que o algoritmo de árvore de decisão aplicado não se trata de um algoritmo que permita uma grande capacidade preditiva para o conjunto de dados em questão. No entanto, verificou-se até pelo teste de outros algoritmos conhecidos por maiores capacidades preditivas que o problema não estaria relacionado com o algoritmo em si, mas sim com os próprios dados. Apesar da grande quantidade de dados fornecidos aos algoritmos e da disponibilidade de 18 variáveis independentes no *dataset*, como todas as variáveis construídas são do tipo booleano (em vez de variáveis quantitativas), que indicam se tal comentário negativo refere ou não tais aspetos, limitam desde logo a capacidade de aprendizagem de um algoritmo e a construção de melhores modelos. A decisão de ter um elevado número de variáveis independentes (18 variáveis) também adveio da necessidade de alimentar o algoritmo com mais informação e tentar obter um modelo com maior desempenho preditivo, como já referido, sem grande sucesso.

Assim, não foi possível a construção de um modelo com um bom desempenho preditivo (acurácia do modelo baseado em árvores de decisão igual a 56,01%) e, apesar de o modelo até prever bem os comentários negativos associados a hotéis de classe baixa (especificidade = 84,43%), o modelo tem uma grande dificuldade em classificar os comentários associados a hotéis de luxo (sensibilidade = 25,30%), muito embora as classes até estejam equilibradas em termos de número de comentários negativos associados.

Sugestões de pesquisa futura

As limitações identificadas podem vir a ser, desde logo, pontos de partida para investigações futuras.

Desta forma, por um lado, futuros investigadores que queiram formular uma investigação semelhante poderão aplicar técnicas de criação automática de tópicos, como a LDA ou a STM, como estava inicialmente projetado para o processo de estruturação dos comentários das *reviews online*. Por outro lado, devem ser aplicados novos algoritmos de DM que possam ter maior capacidade de previsão com este tipo de dados, na tentativa de obter modelos classificativos com melhor desempenho, muito embora os algoritmos do DM caraterizados por terem as melhores prestações de aprendizagem não se tratem de técnicas simbólicas, o que limita, assim, a extração de conhecimento através do modelo.

De um outro prisma, apesar de o objetivo do presente trabalho ser trabalhar unicamente com os comentários negativos das *reviews*, de modo a melhorar a capacidade preditiva dos modelos, também poderão ser utilizadas variáveis que não sejam somente relativas ao conteúdo das *reviews*. Por exemplo, poderão ser englobados(as): os preços das estadias; caraterísticas dos avaliadores, como nacionalidade, objetivo da estadia (se é de trabalho, familiar ou com amigos); datas e duração da estadia; caraterísticas dos hotéis, como o tamanho do hotel, o número de quartos, se tem ou não piscina, ginásio, spa, elevador. Assim, até pela vantagem de muitas destas variáveis não serem booleanas, a capacidade preditiva de modelos futuros poderia ser melhor.

De modo a expandir o conhecimento e a literatura acerca da insatisfação dos hóspedes no setor do turismo, fica a sugestão de futuros estudos aplicarem metodologias semelhantes em conjuntos de dados referentes a hotéis de outras cidades, países ou regiões. Adicionalmente, este tipo de investigação pode ser igualmente realizada em relação à indústria do alojamento local, em vez da hospitalidade.

Por fim, resta recomendar que os responsáveis dos hotéis façam o mesmo tipo de estudo para os seus respetivos hotéis com a finalidade de entender melhor o comportamento do seu hóspede-tipo e, assim, ajustar as conclusões aos seus hotéis, pois aquelas que são retiradas deste estudo são generalizadas para os hotéis da cidade de Lisboa.

Referências bibliográficas

Alsayat, A. (2022). Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-022-07992-x

Anastasiei, B., & Dospinescu, N. (2019). Electronic word-of-mouth for online retailers: Predictors of volume and valence. *Sustainability (Switzerland)*, *11*(3). https://doi.org/10.3390/su11030814

Ban, H. J., Choi, H., Choi, E. K., Lee, S., & Kim, H. S. (2019). Investigating key attributes in experience and satisfaction of hotel customer using online review data. *Sustainability (Switzerland)*, *11*(23). https://doi.org/10.3390/su11236570

Banerjee, A., & Fudenberg, D. (2004). Word-of-mouth learning. *Games and Economic Behavior*, *46*(1), 1–22. https://doi.org/https://doi.org/10.1016/S0899-8256(03)00048-4

Bizirgianni, I., & Dionysopoulou, P. (2013). The Influence of Tourist Trends of Youth Tourism through Social Media (SM) & Information and Communication Technologies (ICTs). *Procedia - Social and Behavioral Sciences*, *73*, 652–660. https://doi.org/10.1016/j.sbspro.2013.02.102

C3.ai Inc. (2023). *Holdout Data*. https://c3.ai/glossary/data-science/holdout-data/

Casado-Díaz, A. B., Andreu, L., Beckmann, S. C., & Miller, C. (2020). Negative online reviews and webcare strategies in social media: effects on hotel attitude and booking intentions. In *Current Issues in Tourism* (Vol. 23, Issue 4, pp. 418–422). Routledge. https://doi.org/10.1080/13683500.2018.1546675

Chaar-Perez, K. (2023, February 17). *Text Mining & Analysis*. https://pitt.libguides.com/textmining

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). *CRISP-DM 1.0 Step-by-step data mining guide*. DaimlerChrysler.

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research*, *69*(2), 897–904. https://doi.org/10.1016/j.jbusres.2015.07.001

Fernandes, T., & Fernandes, F. (2018). Sharing Dissatisfaction Online: Analyzing the Nature and Predictors of Hotel Guests Negative Reviews. *Journal of Hospitality Marketing and Management*, *27*(2), 127–150. https://doi.org/10.1080/19368623.2017.1337540

Fu, W., Wei, S., Wang, J., & Kim, H. S. (2022). Understanding the Customer Experience and Satisfaction of Casino Hotels in Busan through Online User-Generated Content. *Sustainability (Switzerland)*, *14*(10). https://doi.org/10.3390/su14105846

Furtado, A., Ramos, R. F., Maia, B., & Costa, J. M. (2022). Predictors of Hotel Clients' Satisfaction in the Cape Verde Islands. *Sustainability (Switzerland)*, *14*(5). https://doi.org/10.3390/su14052677

Gaur, L., Afaq, A., Solanki, A., Singh, G., Sharma, S., Jhanjhi, N. Z., My, H. T., & Le, D. N. (2021). Capitalizing on big data and revolutionary 5G technology: Extracting and visualizing ratings and reviews of global chain hotels. *Computers and Electrical Engineering*, *95*. https://doi.org/10.1016/j.compeleceng.2021.107374

Golmohammadi, A. R., Jahandideh, B., & O'Gorman, K. D. (2012). Booking on-line or not: A decision rule approach. *Tourism Management Perspectives*, *2–3*, 85–88. https://doi.org/10.1016/j.tmp.2012.03.004

Gonçalves, H. M., Silva, G. M., & Martins, T. G. (2018). Motivations for posting online reviews in the hotel industry. *Psychology and Marketing*, *35*(11), 807–817. https://doi.org/10.1002/mar.21136

Google Colaboratory. (2023). https://research.google.com/colaboratory/intl/pt-PT/faq.html

Han, H., & Ryu, K. (2007). Moderating role of personal characteristics in forming restaurant customers' behavioral intentions: An upscale restaurant setting. *Journal of Hospitality and Leisure Marketing*, *15*(4), 25–54. https://doi.org/10.1300/J150v15n04_03

Handani, N. D., Riswanto, A. L., & Kim, H. S. (2022). A Study of Inbound Travelers Experience and Satisfaction at Quarantine Hotels in Indonesia during the COVID-19 Pandemic. *Information (Switzerland)*, *13*(5). https://doi.org/10.3390/info13050254

Hernandez-Ortega, B. (2020). What about "U"? The influence of positive online consumer reviews on the individual's perception of consumption benefits. *Online Information Review*, *44*(4), 863–885. https://doi.org/10.1108/OIR-10-2018-0304

Hillier, W. (2023, May 11). *What's the Difference Between Data Science, Data Analytics, and Machine Learning?* https://careerfoundry.com/en/blog/data-analytics/data-science-vs-data-analytics-vs-machine-learning/#in-depth-what-is-machine-learning

Hinckley, D. (2015). *New Study: Data Reveals 67% of Consumers are Influenced by Online Reviews*. https://moz.com/blog/new-data-reveals-67-of-consumers-are-influenced-by-online-reviews

Hu, N., Zhang, T., Gao, B., & Bose, I. (2019). What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management*, *72*, 417–426. https://doi.org/10.1016/j.tourman.2019.01.002

Jani, D., & Han, H. (2013). Personality, social comparison, consumption emotions, satisfaction, and behavioral intentions. *International Journal of Contemporary Hospitality Management*, *25*(7), 970–993. https://doi.org/10.1108/IJCHM-10-2012-0183

Key, T. M. (2017). Domains of Digital Marketing Channels in the Sharing Economy. *Journal of Marketing Channels*, *24*(1–2), 27–38. https://doi.org/10.1080/1046669X.2017.1346977

Khade, A. A. (2016). Performing Customer Behavior Analysis using Big Data Analytics. *Procedia Computer Science*, *79*, 986–992. https://doi.org/10.1016/j.procs.2016.03.125

Kim, S., Kandampully, J., & Bilgihan, A. (2018). The influence of eWOM communications: An application of online social network framework. *Computers in Human Behavior*, *80*, 243–254. https://doi.org/10.1016/j.chb.2017.11.015

Kim, W. G., & Park, S. A. (2017). Social media review rating versus traditional customer satisfaction. *International Journal of Contemporary Hospitality Management*, *29*(2), 784–802. https://doi.org/10.1108/IJCHM-11-2015-0627

Kitsios, F., Kamariotou, M., Karanikolas, P., & Grigoroudis, E. (2021). Digital marketing platforms and customer satisfaction: Identifying ewom using big data and text mining. *Applied Sciences (Switzerland)*, *11*(17). https://doi.org/10.3390/app11178032

Lan, T., Zhang, Y., Jiang, C., Yang, G., & Zhao, Z. (2018). Automatic identification of Spread F using decision trees. *Journal of Atmospheric and Solar-Terrestrial Physics*, *179*, 389–395. https://doi.org/10.1016/j.jastp.2018.09.007

Lee, J., Park, D., & Han, I. (2011). The different effects of online consumer reviews on consumers' purchase intentions depending on trust in online shopping malls. *Internet Research*, *21*(2), 187–206. https://doi.org/10.1108/10662241111123766

Lee, P. J., Hu, Y. H., & Lu, K. T. (2018). Assessing the helpfulness of online hotel reviews: A classification-based approach. *Telematics and Informatics*, *35*(2), 436–445. https://doi.org/10.1016/j.tele.2018.01.001

Li, H., Ye, Q., & Law, R. (2013). Determinants of Customer Satisfaction in the Hotel Industry: An Application of Online Review Analysis. In *Asia Pacific Journal of Tourism Research* (Vol. 18, Issue 7, pp. 784–802). https://doi.org/10.1080/10941665.2012.708351

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management*, *29*(3), 458–468. https://doi.org/10.1016/j.tourman.2007.05.011

Lockyer, T. (2005). The perceived importance of price as one hotel selection dimension. *Tourism Management*, *26*(4), 529–537. https://doi.org/10.1016/j.tourman.2004.03.009

Mariani, M. M., & Matarazzo, M. (2021). Does cultural distance affect online review ratings? Measuring international customers' satisfaction with services leveraging digital platforms and big data. *Journal of Management and Governance*, *25*(4), 1057–1078. https://doi.org/10.1007/s10997-020-09531-z

Mariani, M. M., & Nambisan, S. (2021). Innovation Analytics and Digital Innovation Experimentation: The Rise of Research-driven Online Review Platforms. *Technological Forecasting and Social Change*, *172*. https://doi.org/10.1016/j.techfore.2021.121009

Mathew, J., Kshirsagar, R., Abidin, D. Z., Griffin, J., Kanarachos, S., James, J., Alamaniotis, M., & Fitzpatrick, M. E. (2023). A comparison of machine learning methods to classify

radioactive elements using prompt-gamma-ray neutron activation data. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-36832-8

Matthes, J., & Kohring, M. (2008). The Content Analysis of Media Frames: Toward Improving Reliability and Validity. *Journal of Communication*, *58*(2), 258–279. https://doi.org/10.1111/j.1460-2466.2008.00384.x

Moro, S., Rita, P., Ramos, P., & Esmerado, J. (2022). The influence of cultural origins of visitors when staying in the city that never sleeps. *Tourism Recreation Research*, *47*(1), 78–90. https://doi.org/10.1080/02508281.2020.1821328

Mucharreira, P. R., Antunes, M. G., Abranja, N., Justino, M. R. T., & Quirós, J. T. (2019). The relevance of tourism in financial sustainability of hotels. *European Research on Management and Business Economics*, *25*(3), 165–174. https://doi.org/10.1016/j.iedeen.2019.07.002

Nie, R. xin, Tian, Z. peng, Wang, J. qiang, & Chin, K. S. (2020). Hotel selection driven by online textual reviews: Applying a semantic partitioned sentiment dictionary and evidence theory. *International Journal of Hospitality Management*, *88*. https://doi.org/10.1016/j.ijhm.2020.102495

Nilashi, M., Abumalloh, R. A., Almulihi, A., Alrizq, M., Alghamdi, A., Ismail, M. Y., Bashar, A., Zogaan, W. A., & Asadi, S. (2021). Big social data analysis for impact of food quality on travelers' satisfaction in eco-friendly hotels. *ICT Express*. https://doi.org/10.1016/j.icte.2021.11.006

Nilashi, M., Abumalloh, R. A., Minaei-Bidgoli, B., Abdu Zogaan, W., Alhargan, A., Mohd, S., Syed Azhar, S. N. F., Asadi, S., & Samad, S. (2022). Revealing travellers' satisfaction during COVID-19 outbreak: Moderating role of service quality. *Journal of Retailing and Consumer Services*, *64*. https://doi.org/10.1016/j.jretconser.2021.102783

Nilashi, M., Mardani, A., Liao, H., Ahmadi, H., Manaf, A. A., & Almukadi, W. (2019). A hybrid method with TOPSIS and machine learning techniques for sustainable development of green hotels considering online reviews. *Sustainability (Switzerland)*, *11*(21). https://doi.org/10.3390/su11216013

Nourani, V., & Molajou, A. (2017). Application of a hybrid association rules/decision tree model for drought monitoring. *Global and Planetary Change*, *159*, 37–45. https://doi.org/10.1016/j.gloplacha.2017.10.008

Oliver, R. L. (2015). *Satisfaction: a behavioral perpective on the consumer* (2nd ed.). Routledge.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perinotto, A. R. C., Camarço, J. C. F., Braga, S. D. S., & Gonçalves, M. F. (2023). Perceptions on services in Ceará-Brazil luxury hotels registered on TripAdvisor . *Journal of Global Scholars of Marketing Science*, *33*(1), 11–30. https://doi.org/10.1080/21639159.2020.1808841

Sann, R., Lai, P. C., Liaw, S. Y., & Chen, C. T. (2022). Predicting Online Complaining Behavior in the Hospitality Industry: Application of Big Data Analytics to Online Reviews. *Sustainability (Switzerland)*, *14*(3). https://doi.org/10.3390/su14031800

Santos, M., & Ramos, I. (2017). *Business Intelligence: Da informação ao conhecimento* (3rd ed.). FCA.

Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and Tourism Online Reviews: Recent Trends and Future Directions. *Journal of Travel and Tourism Marketing*, *32*(5), 608–621. https://doi.org/10.1080/10548408.2014.933154

Schuckert, M., Liu, X., & Law, R. (2016). Stars, Votes, and Badges: How Online Badges Affect Hotel Reviewers. *Journal of Travel and Tourism Marketing*, *33*(4), 440–452. https://doi.org/10.1080/10548408.2015.1064056

Sharifi, S. (2019). Examining the impacts of positive and negative online consumer reviews on behavioral intentions: Role of need for cognitive closure and satisfaction guarantees. *Journal of Hospitality Marketing and Management*, *28*(4), 397–426. https://doi.org/10.1080/19368623.2019.1531804

Shen, Z., Yang, X., Liu, C., & Li, J. (2021). Assessment of indoor environmental quality in budget hotels using text-mining method: Case study of top five brands in China. *Sustainability (Switzerland)*, *13*(8). https://doi.org/10.3390/su13084490

Shiau, W. L., Dwivedi, Y. K., & Lai, H. H. (2018). Examining the core knowledge on facebook. *International Journal of Information Management*, *43*, 52–63. https://doi.org/10.1016/j.ijinfomgt.2018.06.006

Siering, M., Deokar, A. V., & Janze, C. (2018). Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews. *Decision Support Systems*, *107*, 52–63. https://doi.org/10.1016/j.dss.2018.01.002

Singh, H. P., & Alhamad, I. A. (2022). A Novel Categorization of Key Predictive Factors Impacting Hotels' Online Ratings: A Case of Makkah. *Sustainability (Switzerland)*, *14*(24). https://doi.org/10.3390/su142416588

Song, Y., Liu, K., Guo, L., Yang, Z., & Jin, M. (2022). Does hotel customer satisfaction change during the COVID-19? A perspective from online reviews. *Journal of Hospitality and Tourism Management*, *51*, 132–138. https://doi.org/10.1016/j.jhtm.2022.02.027

Srivastava, A., & Kumar, V. (2021). Hotel attributes and overall customer satisfaction: What did COVID-19 change? *Tourism Management Perspectives*, *40*. https://doi.org/10.1016/j.tmp.2021.100867

Taamneh, M. (2018). Investigating the role of socio-economic factors in comprehension of traffic signs using decision tree algorithm. *Journal of Safety Research*, *66*, 121–129. https://doi.org/10.1016/j.jsr.2018.06.002

Tang, M., & Kim, H. S. (2022). An Exploratory Study of Electronic Word-of-Mouth Focused on Casino Hotels in Las Vegas and Macao. *Information (Switzerland)*, *13*(3). https://doi.org/10.3390/info13030135

Thu, H. N. T. (2020). Measuring guest satisfaction from online reviews: Envidence in Vietnam. *Cogent Social Sciences*, *6*(1). https://doi.org/10.1080/23311886.2020.1801117

Vijay Gaikwad, S. (2014). Text Mining Methods and Techniques. In *International Journal of Computer Applications* (Vol. 85, Issue 17).

Visa, S., Ramsay, B., Ralescu, A., & van der Knaap, E. (2011). Confusion Matrix-based Feature Selection. *Proceedings of the Twentysecond Midwest Artificial Intelligence and Cognitive Science Conference*, 120–127.

Wei, S., & Kim, H. S. (2022). Online Customer Reviews and Satisfaction with an Upscale Hotel: A Case Study of Atlantis, The Palm in Dubai. *Information (Switzerland)*, *13*(3). https://doi.org/10.3390/info13030150

Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications*, *42*(4), 2264–2275. https://doi.org/10.1016/j.eswa.2014.10.023

Westbrook, R. A. (1987). Product/Consumption-Based Affective Responses and Postpurchase Processes. *Journal of Marketing Research*, *24*(3), 258–270. https://doi.org/10.1177/002224378702400302

Xu, X., & Li, Y. (2016). The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. *International Journal of Hospitality Management*, *55*, 57–69. https://doi.org/10.1016/j.ijhm.2016.03.003

Xu, X., Wang, X., Li, Y., & Haghighi, M. (2017). Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors. *International Journal of Information Management*, *37*(6), 673–683. https://doi.org/10.1016/j.ijinfomgt.2017.06.004

Zhang, X., & Kim, H. S. (2021). Customer experience and satisfaction of disneyland hotel through big data analysis of online customer reviews. *Sustainability (Switzerland)*, *13*(22). https://doi.org/10.3390/su132212699

Zhao, Y., Xu, X., & Wang, M. (2019). Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International Journal of Hospitality Management*, *76*, 111–121. https://doi.org/10.1016/j.ijhm.2018.03.017

Zhou, L., Ye, S., Pearce, P. L., & Wu, M. Y. (2014). Refreshing hotel satisfaction studies by reconfiguring customer review data. *International Journal of Hospitality Management*, *38*, 1–10. https://doi.org/10.1016/j.ijhm.2013.12.004

Zhou, Q. (Bill), Zhang, J., Zhang, H., & Li, X. (Robert). (2018). Is all authenticity accepted by tourists and residents? The concept, dimensions and formation mechanism of negative authenticity. *Tourism Management*, *67*, 59–70. https://doi.org/10.1016/j.tourman.2017.12.024