ISCTE IUL
Instituto Universitário de Lisboa

Departamento de Ciências e Tecnologias da Informação

Modelação de comentários em plataformas online sobre alojamento local: O caso Airbnb

Sónia Alexandra Miranda Morais

Dissertação submetida como requisito parcial para obtenção do grau de

Mestre em Sistemas Integrados de Apoio à Decisão

Orientador(es):
Doutor Ricardo Daniel Santos Faro Marques Ribeiro,
Prof. Auxiliar do Departamento de Ciências e Tecnologias de Informação do ISCTE-IUL

Doutor Fernando Manuel Marques Batista,
Prof. Auxiliar do Departamento de Ciências e Tecnologias de Informação do ISCTE-IUL

30 de Dezembro de 2019

Resumo

O crescimento do setor do turismo e mais propriamente do alojamento local, proporcionou nos últimos anos um aumento significativo de comentários *online* que se refletem nas decisões de cada turista na hora de reservar um alojamento. Um dos problemas associados à plataforma Airbnb é a infinidade de comentários existentes para cada alojamento, que expressam muitas das vezes as experiências realizadas pelos hóspedes e que não estão a ser considerados com a devida importância na tomada de decisão do proprietário. Este estudo pretende analisar os aspetos discutidos nos comentários, e sentimentos que advêm deste tipo de experiências, bem como as falhas e necessidades que são importantes de colmatar para se poder usufruir/fornecer de uma melhor qualidade do serviço neste tipo de alojamentos.

A obtenção e identificação destes aspetos, que vão desde a propriedade até ao próprio proprietário, irá auxiliar os proprietários a tomar decisões quanto a melhorias nos respetivos alojamentos locais, conseguindo ter uma rápida indicação do que pode ser melhorado bem como a alcançar o bom *feedback* na plataforma Airbnb e também a tomarem conhecimento dos aspetos que estão a ser discutidos na atualidade. Por forma a analisar estas opiniões e sentimentos dos hóspedes relativamente a toda a experiência turística, foram utilizadas técnicas de *text mining*, como a análise de sentimentos e modelação por tópicos, tais como o *Latent Semantic Analysis* (LSA) e *Latent Dirichlet Allocation* (LDA) para alojamentos localizados em Lisboa, Portugal, na plataforma Airbnb.

Palavras chave*:* *Text Mining*, Análise de sentimentos, Modelação por tópicos, Processamento de linguagem natural, Airbnb, Comentários *online*

Abstract

The growth of the tourism sector and more specifically of local accommodation has in recent years provided a significant increase of online comments that are reflected in the decisions of each tourist when booking a accommodation. One of the problems associated with the Airbnb platform is the plethora of existing reviews for each accommodation, which often express the experiences of guests and are not being considered with due importance in the owner's decision making. This study aims to analyze the aspects discussed in the comments, and feelings that come from this type of experiences, as well as the flaws and needs that are important to address in order to enjoy/provide a better quality of service in this type of accommodation.

Obtaining and identifying these aspects, ranging from property to owner, will help homeowners make decisions on how to improve their local accommodation, giving them a quick indication of what can be improved as well as getting good feedback on the property on the Airbnb platform and also becoming aware of what is currently being discussed on actual days. In order to analyze these guests' opinions and feelings about the whole tourist experience, text mining techniques, such as sentiment analysis, and topic modeling, such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) were used, for accommodation located in Lisbon, Portugal on the Airbnb platform.

Keywords: Text Mining, Sentiment analysis, Topic modelling, Natural language processing, Airbnb, Online reviews

Agradecimentos

Os meus agradecimentos são dirigidos a todos aqueles que me ajudaram na realização deste trabalho e ao longo de todo o percurso deste mestrado, nomeadamente:

• ao meu orientador e coorientador, Professor Ricardo Ribeiro e Professor Fernando Batista pelo apoio e atenção;

• aos Professores de mestrado pelo conhecimento transmitido;

• aos meus pais e amigos por me terem apoiado incondicionalmente;

• aos meus amigos de mestrado, em especial, não só por todo o apoio, como que juntos conseguimos ultrapassar várias “batalhas” nesta fase das nossas vidas.

Lisboa, Outubro de 2019

Sónia Morais

Índice

Lista de Figuras

Lista de Tabelas

1 Introdução

No decorrer dos últimos anos, a constante inovação tecnológica trouxe consigo a vinda de plataformas *online* de P2P (*peer-to-peer*, designado de ponto-a-ponto) e consequentemente, o aparecimento de novos conceitos como é o exemplo da economia partilhada ou *sharing economy,* que podem ou não estar diretamente interligados com estas plataformas. A economia partilhada possibilita ao empreendedor partilhar um bem ou serviço através de plataformas *online* existentes e por conseguinte, esta partilha ser visualizada por milhões de utilizadores ativos espalhados pelo mundo, como é o caso das plataformas já conhecidas Uber, Airbnb, entre outras. Esta última, pertencendo ao setor do alojamento, tornou-se num dos modelos colaborativos mais bem-sucedidos na economia partilhada (Cheng e Jin, 2019). Sendo uma plataforma *online* P2P de alojamento turístico local, permite a qualquer indivíduo, partilhar o seu bem exclusivo com outras pessoas, resultando em rendimentos para o próprio. Por sua vez, os hóspedes que vivenciam esta experiência têm a possibilidade de deixar um comentário ou avaliação sobre a sua estadia em determinado alojamento, visível a todos os utilizadores de forma global.

Estes comentários cada vez mais em larga escala refletem, muitas das vezes, opiniões e sentimentos dos hóspedes relativamente a toda a experiência turística. Posto isto, a extração de conhecimento ou informação a partir das diferentes opiniões torna-se de extrema importância, ajudando os proprietários de determinado alojamento a reagir mais rapidamente aos pedidos dos hóspedes, assim como a tomarem conhecimento dos aspetos que estão a ser discutidos na atualidade. Este estudo emprega técnicas de *text mining,* tais como a análise de sentimentos, modelação por tópicos e outras técnicas ligadas ao processamento de linguagem natural (PLN), por forma a analisar dados não estruturados de um *dataset* de grandes dimensões, correspondente a alojamentos localizados em Lisboa, capital de Portugal extraídos da plataforma Airbnb.

A evolução ao longo dos anos do turismo em Portugal, mais propriamente no setor do alojamento, tem sido cada vez mais crescente[1] e isso é percetível se considerarmos o surgimento das plataformas *online* que estão cada vez mais disponíveis. Estas plataformas cada vez mais dispõem de uma maior oferta de alojamento, bem como uma crescente aceitação por parte dos turistas. Devido a esta grande aceitação, torna-se cada vez mais importante

[1] `https://www.pordata.pt/Municipios/Alojamentos+tur{%}c3{%}adsticos+total+e+por+tipo+de+alojamento-746`, acedido a 08-12-2018

efetuar análises a estes dados não estruturados que até então não tinham sido tão considerados.

Neste estudo, foi aplicada a modelação por tópicos (*topic modelling*) que permite descobrir os principais temas abordados numa coleção de documentos não estruturados, sem recorrer a qualquer tipo de anotação nos documentos analisados constituindo assim, uma abordagem de aprendizagem não supervisionada (Blei et al., 2003). Os métodos de análise de tópicos utilizados neste estudo são o *Latent Semantic Analysis* (LSA) e o *Latent Dirichlet Allocation* (LDA).

Por sua vez, a análise de sentimentos está entre as técnicas mais populares, podendo ser facilmente usada em dados com opiniões sobre um determinado aspeto de um produto ou serviço, sendo muito útil para este tipo de análises de dados não estruturados (Tsytsarau e Palpanas, 2016). De maneira a compreender-se este tipo de informação, para este estudo são analisadas as polaridades das opiniões expressas no geral, bem como os sentimentos gerados em cada categoria de um determinado tópico. Deste modo, para a aplicação desta análise, foi abordada uma técnica de aprendizagem não supervisionada, o *Vader Sentiment* (*Valence Aware Dictionary and sEntiment Reasone*).

Por forma a manter a boa reputação do proprietário e a satisfação do hóspede, foram analisadas as inconformidades existentes entre a descrição dos anúncios da propriedade e aquilo que o hóspede presencia aquando da sua estadia. Estas inconformidades surgem quando o proprietário refere que o alojamento detém determinada caraterística, mas na realidade o hóspede constata que não está conforme.

Por fim, outro dos aspetos a salientar neste estudo, é a análise do *score* que está associada a cada alojamento, que pode não estar de acordo com os comentários escritos pelos hóspedes. Este *score* é indicado por cada hóspede na plataforma, depois de cada estadia neste tipo de alojamentos.

1.1 Definição do problema e objetivos

O Airbnb (plataforma *online* integrada totalmente num modelo colaborativo) permite a partilha ou aluguer de alojamentos locais. O Airbnb disponibiliza a visualização de todos os comentários fornecidos por antigos hóspedes em cada alojamento, permitindo assim, aos turistas observar se determinado alojamento dispõe de comentários considerados satisfatórios para si. Com este mecanismo fornecido por parte do Airbnb é possível, não só visualizar os comentários relativos à experiência do hóspede, mas também relativos ao próprio proprietário de determinado alojamento, bem como comentários realizados aos hóspedes, de forma a que os proprietários futuros possam considerar a vinda destes (Bridges e Vásquez, 2018). Estudos recentes realizados nos Estados Unidos da América aos consumidores das plataformas *online*, Yelp, TripAdvisor, entre outras, indicaram que 91% dos consumidores com idades compreendidas entre os 18 e os 34 anos confiam tanto nos comentários *online* como nas recomendações pessoais, tal com se pode observar na Figura 1.1. Isto porque

estas idades estão mais recetivas a este tipo de informação que surgiu mais recentemente, no contexto da era digital, ao invés das idades com mais de 55 anos, que se tornam mais céticos correspondendo apenas a 61%. No entanto, a maioria dos consumidores exige vários comentários para confiar nos mesmos, o que se deve ao fato de o consumidor se sentir mais confiante em usar um produto ou serviço ao qual vários consumidores anteriores puderam experimentar. Sendo que, para o Airbnb, os turistas tendem a decidir sobre os seus próprios planos de viagem, consoante a informação que é fornecida pelos hóspedes anteriores (Black e Kelley, 2009; Bridges e Vásquez, 2018; Camilla, 2011; Filieri et al., 2015; Ye et al., 2009).

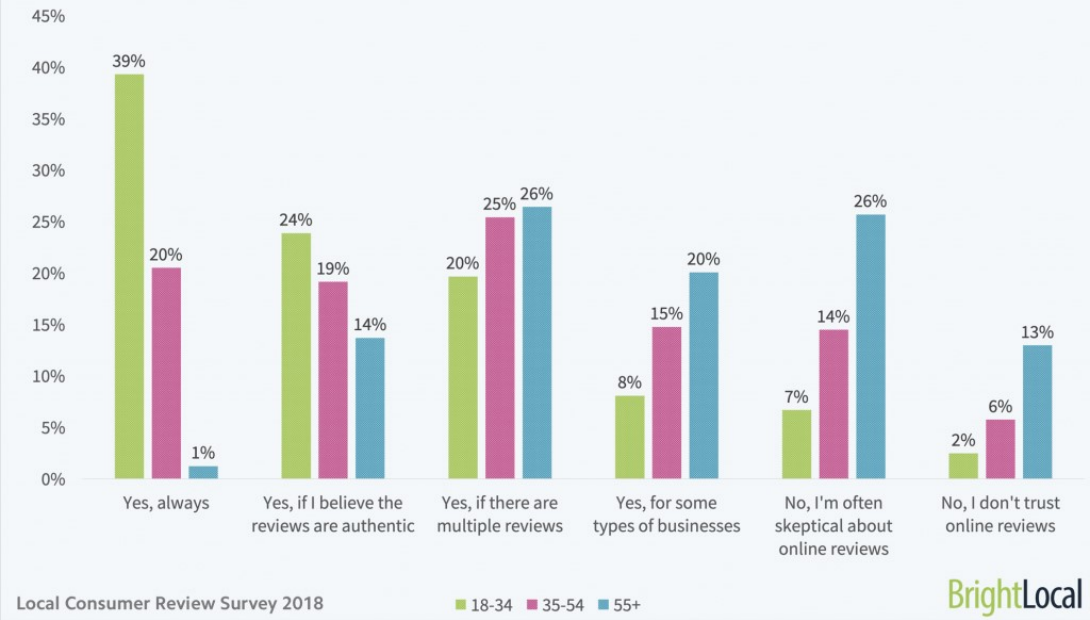


Figura 1.1: Opinião dos consumidores relativamente ao tipo de confiança que têm nos comentários *online*. Fonte: *BrightLocal*, acedido a 09-12-2018

Para o Airbnb, os sistemas de comentários advêm sempre de uma partilha ou anúncio que o proprietário de um determinado alojamento colocou na plataforma. Esta partilha é composta normalmente pela descrição do alojamento, a sua disponibilidade, alguns aspetos relevantes alusivos à propriedade e, por fim, todos os comentários realizados pelos hóspedes. Desta forma, alguns dos proprietários que partilham os seus alojamentos nesta plataforma podem querer perceber se o(s) seu(s) alojamento(s) vão ao encontro dos desejos e necessidades dos respetivos hóspedes. Neste momento, se um proprietário de um ou mais alojamentos tiver uma infinidade de comentários, sejam eles negativos ou positivos, não consegue compreender se a proporção de comentários foi mais positiva do que negativa, ou vice-versa, ou até mesmo conseguir identificar quais os aspetos da experiência mais relevantes para cada hóspede. O proprietário tem que estar preparado para todo o tipo de comentários que possam surgir e a reputação positiva no Airbnb é um dos pontos favoráveis para o proprietário, sendo que, esta reputação é ganha pelo bom *feedback* que o proprietário terá nos comentários do seu alojamento. No entanto, o proprietário tem de ter

também em atenção quando um comentário é escrito por uma pessoa famosa ou influente, neste caso, o hóspede é altamente influenciado na sua tomada de decisão, ou até mesmo comentários escritos por utilizadores anteriores, refletindo muitas das vezes experiências reais. Outro ponto bastante importante, é a "idade" de um comentário, o que implica que comentários mais recentes geralmente fornecem aos hóspedes informações mais atualizadas.

Deste modo, este estudo tem como objetivo principal proceder à identificação das falhas e necessidades que os hóspedes mencionam nos comentários, por forma a evitar falhas futuras no serviço de cada alojamento, melhorando a qualidade de serviço da plataforma Airbnb. Utilizando as técnicas de *text mining*, este estudo passa pela extração e análise dos comentários dos hóspedes, detetando o sentimento associado a cada comentário e os tópicos mais relevantes da experiência dos hóspedes, por forma a entender o que as pessoas discutem sobre um determinado alojamento turístico. Assim, este estudo é uma grande ajuda para a tomada de decisão do proprietário dando *insights* de melhoria para o seu alojamento local, em busca de comentários positivos e alcance dos seus objetivos estratégicos, financeiros e operacionais.

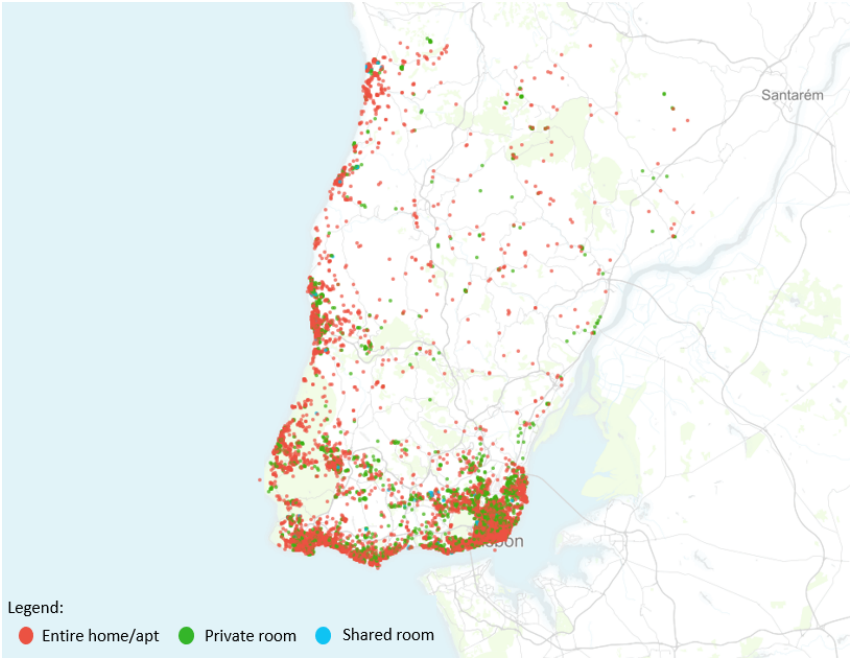


Figura 1.2: Geolocalização dos alojamentos em Lisboa. Fonte: *Inside Airbnb*, acedido a 09-12-2018

Para o desenvolvimento desta análise foi retirado o *dataset* de *Inside Airbnb*, obtendo-se uma amostra (com dados de alojamentos e comentários até abril de 2019) que contém 769.636 comentários relativos aos alojamentos turísticos locais da cidade de Lisboa, nos mais variados idiomas. Para efeitos de análise apenas foi selecionado o idioma em inglês, com uma amostra de 466.296 comentários, ou seja, mais de metade da amostra original. Quanto ao número de alojamentos, a plataforma Airbnb disponibiliza para a capital, cerca

de 22.242 propriedades disponíveis para serem reservadas. Na Figura 1.2 é possível observar os tipos de espaço (apartamento/casa inteira, quarto privado e quarto partilhado) e a sua distribuição.

1.2 Motivação e relevância do tema

Perante o cenário atual da plataforma Airbnb, os comentários *online* surgem cada vez mais em larga escala. Estes comentários refletem muitas vezes opiniões e sentimentos de experiências anteriores e é de extrema importância realizar a análise destes dados não estruturados, utilizando técnicas de *text mining*, por forma a ajudar os proprietários de determinado alojamento a reagirem mais rapidamente aos pedidos dos hóspedes para uma eficaz tomada de decisão, assim como, a tomarem conhecimento das suas falhas e necessidades que estão a ser discutidas na atualidade. Posto isto, a motivação passa por melhorar os níveis de serviço de cada alojamento disponível, levando também a um aumento da qualidade de serviço da plataforma Airbnb, o que em termos de relevância acaba por ajudar tanto o proprietário e o hóspede, como o próprio Airbnb. O proprietário pelo fato de que irá poder reagir mais rapidamente às falhas e necessidades dos hóspedes. No caso do hóspede terá sempre alojamentos que respeitam os seus desejos e necessidades, aumentando a sua satisfação aquando da reserva deste tipo de alojamentos. E por fim, o Airbnb, que com o descrito anteriormente para cada interveniente consegue aumentar o nível de serviço dos alojamentos de Lisboa.

Com o rápido crescimento do turismo, os proprietários pretendem que os seus alojamentos sejam cada vez melhores, mantendo uma reputação positiva no Airbnb pelo bom *feedback* que o proprietário terá nos comentários dos próprios alojamentos. Desta forma, importa perceber quais os aspetos que possam fornecer comentários positivos e negativos e aspetos abordados, de modo a atuar de imediato, por forma a que determinado alojamento esteja sempre a ir ao encontro dos desejos e necessidades dos respetivos hóspedes.

1.3 Questões e objetivos de investigação

As questões de investigação servem para o estudo se focar em distintas análises, para que posteriormente possam ser respondidas. Esta dissertação pretende dar resposta às seguintes questões:

1. Através dos comentários dos hóspedes, quais são os aspetos mais relevantes da sua experiência? Tais como por exemplo a localização, a comunicação, a limpeza etc.
   (a) Quais deles os hóspedes consideram mais positivos e mais negativos?
2. Será que os comentários negativos dos hóspedes chocam com as informações dos proprietários sobre os alojamentos?

(a) Consoante o tipo de espaço, o tipo de constatações é diferente? Há evidências claras que isto muda consoante o tipo de espaço?

3. O *score* associado aos indicadores espelha o que é descrito nos comentários?

4. Relativamente aos alojamentos na cidade de Lisboa, os hóspedes escolhem tendencialmente os alojamentos cujos proprietários detêm o estatuto de *superhost*[2]?

   (a) Existe variação pelo tipo de espaço?

   (b) Em comparação com os *hosts* regulares[3], quais as categorias mais abordadas nos comentários?

De acordo com as questões de investigação estabelecidas, as mesmas vão de encontro aos objetivos da investigação. Os objetivos passam por analisar os comentários dos hóspedes, por forma a identificar os aspetos mais relevantes da sua experiência, quer sejam falhas ou necessidades que acabam por ajudar o proprietário na sua tomada de decisão ou, também, o fato de contribuírem para a melhoria da plataforma Airbnb ao nível do serviço da cidade de Lisboa. Por forma a serem identificadas as falhas e necessidades que os vários comentários transmitem, importa não só analisar os aspetos padrão mais relevantes do Airbnb (precisão, limpeza, valor, *check-in*, conforto e localização), bem como outros aspetos igualmente importantes para o hóspede, de modo a proceder à identificação destes aspetos como negativos ou positivos. Outro ponto bastante importante e que também está diretamente relacionado com o objetivo deste estudo é a identificação dos comentários negativos dos hóspedes que chocam com as informações dos proprietários sobre os alojamentos, conseguindo desta maneira verificar pelos comentários *online*, se este tipo de constatações varia consoante o tipo de espaço. Para mais um processo para a tomada de decisão do proprietário, é efetuada uma análise à classificação em estrelas através dos comentários, de modo a verificar se a classificação espelha corretamente o que é descrito nos comentários dos hóspedes. Por fim, é realizada a análise aos *hosts* regulares e também aos *superhosts*, onde para este, é verificado se tendencialmente os hóspedes escolhem as propriedades cujo proprietário detém o estatuto de *superhost*, sendo que, quanto às categorias mais abordadas é realizada uma análise aos tópicos relativos a cada proprietário.

1.4 Abordagem metodológica

A metodologia a ser abordada na concretização dos objetivos do presente estudo é o CRISP-DM (CRoss Industry Standard Process for Data Mining). Apesar de ser uma metodologia preparada para projetos de *data mining*, William Vorhies, um dos criadores do

[2] *Superhost*: Estatuto dado a um proprietário Airbnb, em que consiste num programa desenvolvido pela plataforma em que é uma forma de celebrar e premiar os proprietários mais experientes com melhores avaliações do Airbnb.

[3] *Host* regular: Nome dado a um proprietário Airbnb, que ainda não está abrangido pelo programa da plataforma.

CRISP-DM argumenta que todos os projetos de *data science* se iniciam na compreensão do negócio e que os dados precisam de ser processados, aplicando algoritmos de *data science* (Vorhies, 2016). Esta abordagem, ilustrada na Figura 1.3, é constituída pelos seguintes passos: compreensão do negócio (*business understanding),* compreensão dos dados (*data understanding),* preparação dos dados (*data preparation),* modelação (*modeling),* avaliação (*evaluation)* e desenvolvimento (*deployment*), sendo descrita em termos de um modelo de processo hierárquico, compreendendo quatro níveis de abstração (do geral ao específico): atividades, tarefas genéricas, tarefas especializadas e instâncias do processo (Wirth, 2000).

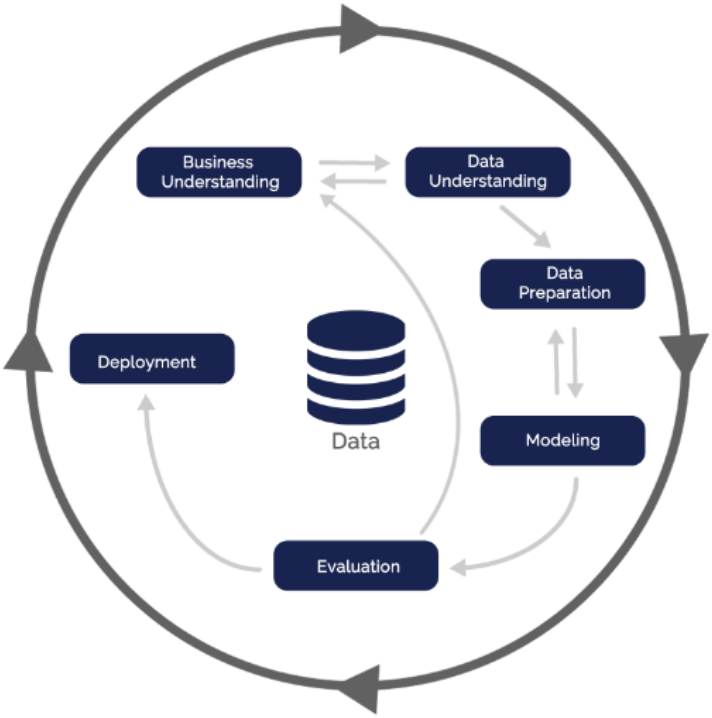


Figura 1.3: Metodologia CRISP-DM. Fonte: *Otaris*, acedido a 10-12-2018

Segundo Azevedo (2014) e Braz et al. (2009), descrevem as seguintes seis fases desta metodologia:

1. Compreensão do Negócio - foca-se na compreensão do problema, objetivos e requisitos do projeto a partir de uma perspetiva de negócio.

2. Compreensão dos Dados - consiste em interpretar a recolha de dados e explorar a informação, mas também é responsável pela avaliação da qualidade dos dados. Esta fase é bastante relevante, quando se trabalha com dados de texto não estruturados, em que é necessário perceber as caraterísticas destes dados cometendo assim o menor número de erros possíveis.

3. Preparação dos Dados - é uma das fases essenciais que consiste em extrair, limpar e formatar os dados brutos iniciais, por forma a garantir que os dados se apresentam no formato adequado para a criação do modelo e assim obter os melhores resultados.

4. Modelação - são aplicadas diversas técnicas, algoritmos para obter a informação importante. Durante esta fase, é frequente ter de voltar a efetuar atividades de pré-processamento, isto porque podem ser identificadas variáveis com pouco significado.

5. Avaliação - consiste em avaliar os resultados obtidos das técnicas definidas no ponto anterior. É assim, determinado o modelo que clarifica as expectativas para garantir que atinge adequadamente os objetivos do negócio.

6. Desenvolvimento – Implementação estratégica dos resultados obtidos. Depois de avaliado, é aprovado para o início do seu desenvolvimento.

1.5 Estrutura e organização da dissertação

Nesta dissertação, optou-se por uma estrutura que está coerente com a abordagem metodológica abordada na Secção 1.4. O que torna todo o estudo mais coerente com os próprios objetivos do projeto, no entanto, como este estudo se trata de um trabalho de investigação científica, a fase de desenvolvimento desta abordagem não foi contemplada, desta forma, a estrutura da investigação foi dividida em seis capítulos essenciais: introdução, enquadramento, revisão da literatura, modelação de comentários no Airbnb, resultados, conclusões e trabalho futuro. O Capítulo 2 pretende realizar um enquadramento da plataforma Airbnb e também do *text mining*. O Capítulo 3 apresenta o trabalho que tem vindo a ser desenvolvido na área de *text mining* para se proceder à realização da tarefa seguinte de modelação de comentários no Airbnb. O Capítulo 4 é apresentado o conjunto de dados, bem como todo o processamento realizado, demonstração das análises e modelos utilizados para responder às questões de investigação. O Capítulo 5 apresenta os resultados obtidos de toda a análise. Por fim, o Capítulo 6 são retratadas as conclusões do estudo e trabalho futuro.

2 Enquadramento

Tendo em conta que o Airbnb vai ser o foco do nosso estudo, este capítulo aborda a origem do Airbnb, o Airbnb como economia partilhada, o proprietário no Airbnb e a confiança que existe neste tipo de alojamentos P2P. Descreve também as técnicas de *text mining* mais comuns utilizadas em tarefas relacionadas.

2.1 O Airbnb

Tal como referido pelo Mundo das Marcas[1], em 2007 dois estudantes Brian Chesky e Joe Gebbia, colegas de curso na Escola de Design de Rhode Island resolveram tornar-se empreendedores. Para a realização de um sonho, deixaram os seus empregos e mudaram-se para a cidade de São Francisco com o intuito de iniciarem o seu próprio negócio como tantos outros jovens estudantes.

Com o decorrer do tempo e cada um usufruindo da experiência de viver numa cidade nova, o dinheiro começa a esgotar-se. Preocupados com este contratempo, foram avisados que o aluguer do apartamento onde estavam a morar iria sofrer um aumento e tinham apenas 14 dias para encontrar uma forma de pagar todas as despesas associadas, o que acabou por piorar ainda mais a situação. Deste modo, os dois jovens tiveram que proceder a um *brainstorming* sobre o tema e surgiu dos dois um pensamento empreendedor. Nesse momento observaram rapidamente que o espaço vazio na sala parecia ser muito mais útil e uma boa oportunidade para rentabilizarem o espaço.

Com uma conferência de *design* na cidade a acontecer no mês de outubro e a disponibilidade do segmento hoteleiro a ficar cada vez mais limitada, Chesky e Gebbia decidem desde logo colocar a sua ideia em prática, desenvolvendo um pequeno *site* com o objetivo de oferecerem um serviço de alojamento na sala do apartamento em São Francisco aos visitantes, composto por um colchão de ar para dormir e pequeno almoço. Este conceito de *Air Bed & Breakfast* veio dar origem mais tarde ao nome da empresa Airbnb.

A ideia empreendedora destes dois jovens estudantes teve bastante adesão e as perguntas dos clientes acerca deste serviço inovador sobre equacionar a possibilidade de levarem este serviço para outros locais dos Estados Unidos da América, fizeram com que Chesky e

[1] `http://mundodasmarcas.blogspot.com/2014/11/airbnb.html`, acedido a 20-12-2018

Gebbia quisessem elevar o seu pequeno negócio a outro nível. Sem conhecimentos ao nível informático decidiram falar com Nathan Blecharczy, licenciado em ciências informáticas e programador de profissão. Com os seus conhecimentos tecnológicos, responsabilizou-se pelo desenvolvimento da plataforma *online* por forma a ajudar estes dois jovens na partilha de alojamentos e permitindo deste modo às pessoas alugarem o todo ou parte da sua própria casa como forma de um rendimento extra.

Em 2008, com o empreendedorismo de Brian Chesky, Joe Gebbia e pela grande ajuda de Nathan Blecharczy foi lançada nos Estados Unidos da América a plataforma Airbnb. Começou por ser um *site* com a designação *airbedandbreakfast.com* e descreve-se como *"um mercado comunitário de confiança para as pessoas arrendarem, descobrirem e reservarem alojamentos únicos por todo o Mundo"*[2].

Em março de 2009, a designação do *site* foi abreviada apenas para *airbnb.com* e o seu alcance foi expandido para mais tipos de alojamento. As pessoas começaram a utilizar o serviço para partilharem alojamentos invulgares, como é o caso de casas na árvore, barcos, cabanas, cavernas, castelos medievais, ilhas e até iglos.

O Airbnb é cada vez mais reconhecido pelo seu alojamento turístico alternativo e a sua expansão por todo o mundo é cada vez mais notória, atualmente com cerca de 81 mil cidades, em mais de 191 países e com cerca de mais de 5 milhões de proprietários por todo o mundo [3]. Sendo uma plataforma *online* P2P de alojamento turístico, possibilita a qualquer pessoa partilhar o seu bem exclusivo, resultando em rendimentos para o próprio. As propriedades partilhadas nesta plataforma *online* podem ir desde as mais modestas às mais luxuosas, abrangendo desta forma os vários nichos de mercado. Estas estão divididas em três tipos de espaço: apartamento/casa inteira, quarto privado e quarto partilhado. Todo o processo de reserva exige que haja confiança de parte a parte e isto deve-se à questão de que as reservas no Airbnb são efetuadas a completos estranhos. Os hóspedes que têm a oportunidade de vivenciar esta experiência, têm a possibilidade de deixar um comentário ou avaliação sobre a sua estadia em determinado alojamento (num prazo de 14 dias, após o *check-out*) e ainda proceder a uma avaliação numa escala de um a cinco estrelas, por forma a classificar determinada estadia em função de seis aspetos ou variáveis, nomeadamente: limpeza, valor, comunicação, *check-in*, localização e precisão.

Alojamento local em Lisboa, Portugal

A entidade Turismo de Portugal define a noção de alojamento local da seguinte forma: *"Os estabelecimentos de alojamento local (AL) são aqueles que prestam serviços de alojamento temporário, nomeadamente, a turistas, mediante remuneração desde que não reúnam os requisitos para serem considerados empreendimentos turísticos".*[4]

[2] `https://press.atairbnb.com/fast-facts/`, acedido a 20-12-2018

[3] `https://press.atairbnb.com/fast-facts/`, acedido a 20-12-2018

[4] `http://business.turismodeportugal.pt/pt/Planear_Iniciar/Como_comecar/Alojamento_Local/Paginas/default.aspx`, acedido a 20-12-2018

A evolução ao longo dos anos no turismo em Portugal, mais propriamente no setor do alojamento tem sido cada vez mais crescente relativamente ao número de alojamentos [5]. É evidente este crescimento quando se verifica uma união com as plataformas tecnológicas que se têm à disposição. Ricardo Macieira, diretor (*country lead*) do Airbnb em Portugal, refere ao Jornal de Negócios em novembro de 2016, que *"Portugal e Lisboa têm sido destinos muito importantes, e mostrado um crescimento bastante interessante (...) Nos dados que foram efetuados no verão, mostraram cerca de 70% nas vindas. Lisboa, foi assim o quarto destino mais procurado da Europa"*.[6]

Observando a análise para a cidade de Lisboa o Jornal Expresso[7], em 2016, refere que no ano anterior à publicação ficaram na cidade 450 mil pessoas através da plataforma Airbnb. Segundo Àngel Mesado, diretor (*public policy manager*) do Airbnb em Portugal e Espanha, referiu que em 2016 Lisboa estava no *top* 10, considerando que esta cidade é cada vez mais importante para o Airbnb. No entanto, senão fosse o Airbnb, os turistas não iriam para a cidade de Lisboa. De acordo com o responsável, Àngel Mesado, reforça que 30% dos hóspedes que usaram a aplicação para ficarem hospedados afirmam que, de outra forma, não ficariam na cidade por muito tempo, ou simplesmente não a visitariam.

Os proprietários que recebem os hóspedes em Lisboa tendencialmente proporcionam uma boa experiência, o que faz com que o hóspede queira sempre voltar a viajar à cidade que visitou. Ricardo Macieira indica que os proprietários das casas em Lisboa tendem a receber de forma calorosa os hóspedes, oferecendo algo de boas-vindas ou até mesmo mostrar os melhores pontos da cidade. Contudo, estas experiências positivas têm dado bons resultados na cidade de Lisboa ao Airbnb, onde numa escala de um a cinco estrelas, com alguns parâmetros que vão desde a limpeza até à localização, os hóspedes têm avaliado as suas estadias em média de 4,7. Na Figura 2.1 é possível verificar que na capital portuguesa os hóspedes tendem a reservar mais o apartamento/casa inteira e/ou quarto privado do que propriamente o quarto partilhado. É ainda observável que as zonas com maior aglomeração de propriedades, provêm especialmente de sítios mais turísticos, como por exemplo Belém e a zona da Baixa (Vale, 2018).

Apesar da plataforma Airbnb estar mais focada para a comunidade, está também pensada para o turismo de negócios. Ricardo Macieira explicou à Event Point que: *"Cada vez mais as pessoas encaram as viagens de negócios como parte integrante da sua vida social normal. A primeira prioridade, claro, é chegar a tempo à reunião de negócios, mas isso não impede que possam explorar a cidade de destino no seu tempo livre. Com o Airbnb é possível juntar o melhor dos dois mundos, viver a cidade como uma pessoa local ao mesmo tempo que se trabalha"*. Com este objetivo definido, existem muitas empresas que estão

[5] https://www.pordata.pt/Municipios/Alojamentos+tur{%}c3{%}adsticos+total+e+por+tipo+de+alojamento-746, acedido a 08-12-2018

[6] https://www.jornaldenegocios.pt/negocios-iniciativas/observatorio-sectores/observatorio-turismo/detalhe/airbnb-objectivo-e-focalizar-na-qualidade-da-experiencia, acedido a 20-12-2018

[7] https://expresso.pt/economia/2016-07-03-Muitos-turistas-nao-iriam-a-Lisboa-sem-a-Airbnb, acedido a 20-12-2018

registadas para utilizar o Airbnb em viagens de negócios.[8]

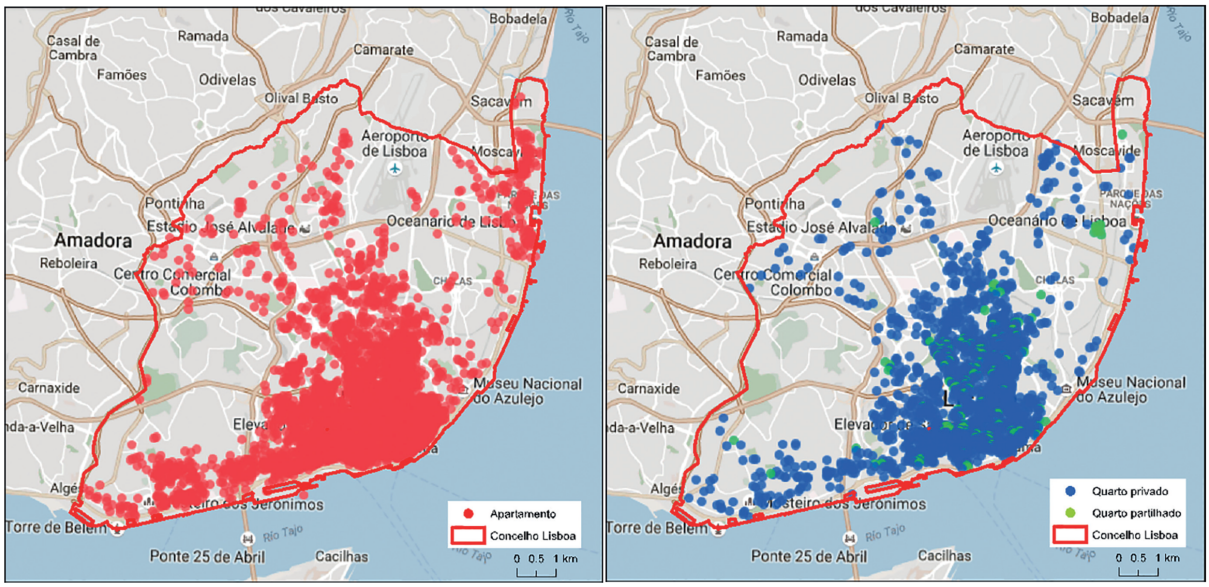


Figura 2.1: Gráficos relativos às estadias na cidade de Lisboa em 2016 (à esquerda alugueres de casa inteira e à direita alugueres de quarto privado ou partilhado) (Vale, 2018)

2.1.1 Airbnb e a economia partilhada

A economia partilhada é um conceito que surgiu com o intuito de satisfazer as necessidades dos consumidores, desde a partilha de automóveis, refeições e alojamentos, que até anteriormente eram serviços fornecidos principalmente por empresas e não por indivíduos empreendedores (Zervas et al., 2013). Este conceito, sendo um modelo completamente focado nas plataformas *online* P2P, consegue ser um modelo altamente utilizado e bem aceite no mercado, devido à celeridade de troca de informação que a *Internet* de hoje em dia disponibiliza. Os acionistas através destas plataformas conseguem proporcionar emprego para os seus membros com a partilha dos seus recursos e por conseguinte, o consumidor poder usufruir dos mesmos. Desta forma, possibilita a estes membros tornarem-se também eles empreendedores, elevando assim, as receitas dos acionistas.

No entanto, os mais entendidos neste tema realçam que os fatores mais decisivos que estimulam o desenvolvimento da economia partilhada são os económicos. Estes são determinados pelos problemas existentes e pela necessidade de uso efetivo dos recursos disponíveis, nomeadamente, no meio *online*. Os fatores tecnológicos (aplicações móveis, redes sociais e *Internet*), bem como os fatores sociais e culturais, têm também alguns efeitos significativos. Atualmente, os fatores ambientais que estimulam os processos de consumo colaborativo não são tão relevantes como os referidos anteriormente (Ivanova, 2017). De

[8] `http://www.eventpointinternational.com/pt/item/10-reports/2171-airbnb-assume-se-/como-solucao-para-quem-viaja-em-negocios`, acedido a 20-12-2018

acordo com Albinsson e Yasanthi Perera (2012); Ozanne e Ballantine (2010); Lutz e Newlands (2018), para além de referirem estes fatores como igualmente importantes, é ainda referenciado que tanto os prestadores de serviço, como os consumidores podem, de fato, procurar interação através da economia partilhada, ou seja, existindo vários participantes de ambos os lados do mercado, esperam assim conhecer pessoas e até criar ligações entre eles. Em contraste, e no contexto da plataforma de alojamento Airbnb, os autores Cheng e Jin (2019) referem que o papel do proprietário talvez seja o de um facilitador e não o de um construtor de relações sociais entre o hóspede e o proprietário.

Como os fatores de cariz tecnológico estão cada vez mais em fase crescente, existe uma maior confiança no comércio e facilidade de pagamento *online*, promovendo em certa parte o desenvolvimento da economia partilhada em alternativa aos mercados considerados mais tradicionais. Apesar destes mercados existirem há mais tempo, em termos tecnológicos estão mais restritos.

Do ponto de vista da indústria, o sucesso quase instantâneo e o crescimento acelerado da plataforma *online* Airbnb faz com que haja uma competição direta representando uma quebra nas empresas do segmento hoteleiro mais convencional, segundo Lehr (2015); Bridges e Vásquez (2018). De acordo com Moufahim (2013); Bridges e Vásquez (2018), em alguns casos o Airbnb também oferece aos turistas a possibilidade de usufruírem de uma alternativa mais acessível do que pagar preços altos por quartos em grandes cadeias de hotéis, ao mesmo tempo em que lhes permite desfrutar mais de uma experiência de "sentir-se em casa". Esta rivalidade e preocupação entre o segmento hoteleiro mais tradicional com a plataforma *online* P2P origina uma teoria da inovação disruptiva, que foi proposta e popularizada por Clayton Christensen (Bower e Christensen, 1995). Esta teoria esboça um processo através do qual um produto/serviço disruptivo transforma um mercado, derrubando por vezes empresas anteriormente dominantes. Um produto/serviço disruptivo geralmente tem um desempenho inferior em relação aos atributos-chave de desempenho dos produtos/serviços predominantes, no entanto possibilita a oferta de um conjunto distinto de benefícios, normalmente focados no custo (mais barato), em serem mais convenientes ou mais simples. O custo é previsivelmente um fator importante nas decisões sobre os produtos/serviços predominantes, segundo Chu e Choi (2000); Dolnicar e Otter (2003); Guttentag (2015). Este produto/serviço, normalmente cria um mercado completamente novo, distinto dos mercados predominantes, que inicialmente são limitados em dimensão e margens de lucro, pelo que é pouco atrativo para as empresas líderes que se contentam em concentrar-se nos seus mercados mais rentáveis e continuar a melhorar marginalmente os seus produtos através de "inovações sustentáveis". No entanto, com o tempo o produto disruptivo melhora, tornando-o atrativo para um maior número de clientes e atraindo níveis crescentes do mercado principal. De acordo com Bower e Christensen (1995); Schmidt e Druehl (2008); Guttentag (2015), a teoria da inovação disruptiva descreve como as empresas podem desmoronar-se, não por ficarem atrás do seu rápido avanço ou ignorarem os seus principais consumidores, mas por excluírem a invasão ascendente de um produto

disruptivo que não possui atributos tradicionalmente favorecidos, mas consegue oferecer benefícios alternativos. Ou seja, no caso da plataforma Airbnb, este tipo de alojamento é normalmente mais barato do que o alojamento tradicional e o alojamento Airbnb para além disso introduz benefícios adicionais associados à permanência numa residência. A procura de um serviço como o Airbnb não é um dado adquirido, uma vez que o Airbnb está consideravelmente ausente em muitos dos aspetos que são mais importantes para os turistas quando selecionam o alojamento do hotel, como é o caso da qualidade do serviço, a simpatia do pessoal, a reputação da marca e a segurança (Guttentag, 2015; Dolnicar e Otter, 2003; Chu e Choi, 2000). Quanto ao custo, os proprietários dos anúncios do Airbnb conseguem fixar preços muito competitivos para os seus espaços, isto porque os custos fixos primários, como é o caso do aluguer e da eletricidade já estão abrangidos, o que pode existir são custos de mão-de-obra mínimos. Geralmente, são cobradas ao hóspede, taxas de limpeza e taxas de serviço, que corresponde ao serviço prestado 24 horas por dia pelo Airbnb ao hóspede e à respetiva manutenção da plataforma.

Relativamente à comparação de custos entre a plataforma Airbnb e as empresas do segmento hoteleiro, são bastante distintas, as tarifas médias do Airbnb para uma casa inteira são geralmente inferiores às dos hotéis de quatro e cinco estrelas e as tarifas médias do Airbnb para um quarto privado são aproximadamente comparáveis às dos hotéis de uma e duas estrelas (Guttentag, 2015).

Tanto os proprietários, como o comércio conseguem adquirir um rendimento extra com esta plataforma. Mais em particular o comércio local do que propriamente o proprietário. Segundo um estudo[9] do impacto económico realizado em Lisboa, pelo Airbnb, os visitantes que usaram esta plataforma geraram um negócio na cidade que ascendeu no ano de 2015 a €268 milhões, dos quais €42,8 milhões foram para os donos das casas onde ficaram os turistas e o maior lucro de €228 milhões devem-se a gastos dos hóspedes no comércio local, como por exemplo os restaurantes, lojas que estão localizadas no bairro onde o hóspede está hospedado. No entanto, passados dois anos a cidade de Lisboa recebeu 1,03 milhões de visitantes e um proprietário típico ganhou 7.685 euros, alugando o alojamento por 103 noites pela plataforma Airbnb.[10]

A economia partilhada é cada vez mais abordada na literatura (Bridges e Vásquez, 2018; Cheng e Jin, 2019; Lutz e Newlands, 2018) e tendo em conta toda esta evolução, surge assim um aumento na cultura da partilha por parte das pessoas, que tendem a ser cada vez mais liberais no que remete ao que é seu, a sua propriedade, o seu bem pessoal, deixando de existir cada vez mais o sentimento de pertença. Quanto ao valor, no caso da plataforma *online* Airbnb o aluguer dos apartamentos/casas inteiras têm o preço médio noturno mais alto, enquanto que os quartos partilhados têm o preço mais baixo, segundo Cansoy e Schor (2016). No entanto e de acordo com Lutz e Newlands (2018) a casa inteira,

[9] https://expresso.pt/economia/2016-07-03-Muitos-turistas-nao-iriam-a-Lisboa-sem-a-Airbnb, acedido a 22-12-2018

[10] https://www.jornaldenegocios.pt/empresas/turismo---lazer/detalhe/airbnb-em-portugal-acolheu-mais-de-26-milhoes-de-hospedes-em-2017, acedido a 22-12-2018

que permite ao hóspede usufruir de todo o alojamento, é a opção mais comum global do Airbnb, seguido por quarto privado e por último quarto partilhado. Para além disto, este aluguer é realizado por completos estranhos que usufruem deste tipo de plataformas *online* (Karlsson et al., 2017).

De acordo com a notícia reportada pelo Jornal de Negócios[11] relativa ao número de casas Airbnb ter triplicado em 2014, constatou-se que quanto à legislação as regras do alojamento local apenas indicam que não existem restrições de dias por ano para arrendar os imóveis, em Portugal. Já noutros países da Europa, como Londres, existe um limite de 90 noites por cada ano, no entanto, para Amesterdão existe uma maior restrição de 60 noites por ano.

Devido à forte expansão da economia partilhada têm surgido críticas à volta dos apartamentos na cidade de Lisboa, uma vez que o alojamento local tem limitado a variação populacional da própria cidade, isto porque, e segundo o Jornal de Negócios, os apartamentos disponíveis para arrendar, não turísticos, diminuiu significativamente e os preços tendencialmente aumentaram, como se pode observar na Figura 2.2. Pode-se constatar que existiu uma evolução dos hóspedes em Lisboa, relativamente aos anos anteriores. O gráfico refere ainda, que 5.25 milhões de hóspedes passaram a noite em Lisboa no último ano, dez vezes mais do que a população da cidade.

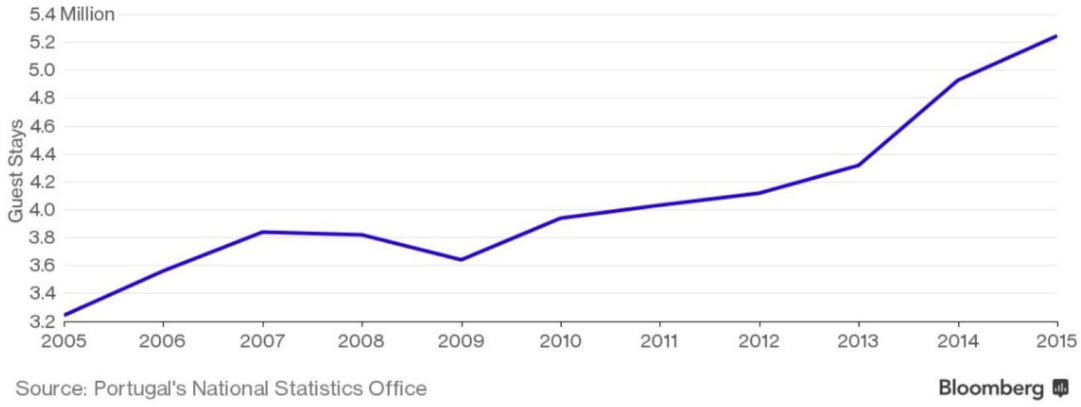


Figura 2.2: Gráfico demonstrativo da evolução dos turistas em Lisboa. Fonte: *Bloomberg*

2.1.2 O proprietário do alojamento na plataforma Airbnb

Segundo a definição da plataforma Airbnb, um anfitrião ou responsável do alojamento, doravante proprietário, designa-se por: "*O principal ponto de contacto para os hóspedes antes, durante e após uma reserva, é a pessoa que aparece como anfitrião na reserva. Se o anfitrião principal também for o administrador do anúncio, as avaliações e os comentários*

[11] `https://www.jornaldenegocios.pt/empresas/turismo---lazer/detalhe/numero_de_casas_em_lisboa_no_airbnb_triplicou_desde_2014`, acedido a 23-12-2018

dos hóspedes irão aparecer no seu perfil e afetarão o seu estatuto de superhost".[12]

No entanto, não foi possível descobrir muita informação sobre a perspetiva do proprietário no Airbnb. Ainda que, existem algumas análises importantes de serem referidas que estão diretamente relacionadas também com o hóspede.

Relativamente à definição de proprietário dada pelo próprio Airbnb, *"(. . .) as avaliações e os comentários dos hóspedes irão aparecer no seu perfil e afetarão o seu estatuto de superhost"*. Este estatuto de *superhost* é um programa desenvolvido pela plataforma Airbnb e é uma forma de celebrar e premiar os proprietários mais experientes com melhores avaliações do Airbnb. Com este programa, o *superhost* terá mais visibilidade, mais rendimento, e também mais recompensas. Um *superhost* é assim definido também pelo Airbnb como: *"anfitriões experientes e um exemplo perfeito para outros anfitriões que oferecem experiências extraordinárias aos seus hóspedes"*.[13] Um proprietário só se torna um *superhost*, se conseguir atingir os seguintes objetivos:[14]

1. Hospedarem pelo menos 10 estadias no último ano ou, se hospedarem reservas de longa duração, 100 noites em pelo menos 3 estadias;
2. Classificação geral média de 4,8 ou superior com base em comentários de pelo menos 50% dos seus hóspedes do Airbnb no ano passado;
3. Não podem ter cancelamentos no último ano, a não ser que tenham existido circunstâncias atenuantes;
4. Resposta a 90% das novas mensagens dentro de 24 horas.

Do ponto de vista do hóspede, é especialmente importante saber se um proprietário cancelou reservas que tenham sido confirmadas, o que representa uma informação valiosa para os hóspedes, uma vez que torna a reserva mais confiável. Além disso, uma taxa de resposta elevada indica um proprietário bem organizado. No geral, um *superhost* é um sinal de qualidade excecional e pode, portanto, ajudar na construção de uma boa reputação (Teubner, Timm and Saade, Norman and Kawlitschek, Florian and Weinhardt, Christof, 2016).

2.1.3 Confiança no alojamento Airbnb

A confiança é referida na literatura como um ponto fulcral neste tipo de alojamentos locais pela plataforma P2P, tanto do lado dos proprietários como dos hóspedes (Karlsson et al., 2017; Moon et al., 2019; Tussyadiah e Park, 2018). Neste tipo de situações, a confiança é importante quando as expectativas de confiança fazem a diferença numa decisão.

[12] `https://www.airbnb.pt/help/article/1536/what-s-the-difference-between-a-primary-/host-and-a-co-host`, acedido a 27-12-2018

[13] `https://www.airbnb.pt/help/article/828/what-is-a-superhost`, acedido a 27-12-2018

[14] `https://www.airbnb.pt/superhost`, acedido a 27-12-2018

Observando-se do ponto de vista social, a confiança está centrada nos deveres morais. Os potenciais hóspedes fazem uma reserva através da plataforma Airbnb sobre as expectativas confiantes de que todas as partes envolvidas no sistema de serviços, incluindo os proprietários e a empresa por trás da plataforma *online*, irão agir de forma competente e obediente. De uma perspetiva racional, a confiança centra-se no interesse próprio; um aumento na confiança diminuirá o custo de transação associado à proteção de si mesmo das possibilidades do comportamento oportunista dos outros (Tussyadiah e Park, 2018; Lauer e Deng, 2007). No entanto, todos estes utilizadores tem a opção de criarem perfis virtuais, com foto e informações pessoais descritivas e a possibilidade de efetuarem uma troca de comunicação direta por mensagens (Guttentag, 2015), permitindo assim a possibilidade de se conhecerem melhor no mundo virtual, fomentando a confiança entre si. Tussyadiah e Park (2018); W. Lehman e Sztompka (2001), sugeriram dois tipos de informação sobre o tipo de pessoa que toma uma determinada decisão de confiança: os traços inerentes das pessoas que levam à confiabilidade primária e o contexto no qual estas pessoas operam, que leva à confiabilidade derivada. Ao estimar a confiabilidade primária, as pessoas aplicam três critérios: reputação, desempenho e aparência. A reputação está associada ao registo de atos passados e à consistência do registo. Desempenho refere-se a ações reais, conduta atual e resultados obtidos atualmente. A aparência está associada ao olhar e à autoapresentação, que em contextos *offline* envolve a forma como as pessoas se vestem, a disciplina corporal e a civilidade, bem como o *status* atribuído. Com base nestes conceitos, as informações que as pessoas podem usar para estimar a confiabilidade primária dos proprietários de alojamento P2P incluem o sistema de comentários antigos e mais recentes (para estimar a reputação e o desempenho), bem como os perfis dos proprietários (para estimar a aparência). No entanto, o sistema de comentários do Airbnb é bastante distinto das outras plataformas. Em 2018, Bridges e Vásquez (2018) referem que o Airbnb é a única plataforma que permite a reciprocidade entre proprietários e hóspedes, ao invés da maioria das empresas que emprega uma comunicação unidirecional. Ou seja, quando um proprietário faz um comentário a um hóspede, o hóspede só consegue visualizar esse comentário se o mesmo escrever um comentário a este proprietário. Desta forma, permite a outros proprietários poderem considerar ou não esses hóspedes e os hóspedes considerarem reservar as suas estadias nos alojamentos desses proprietários. Toda esta informação disponível na plataforma acaba também por ajudar o hóspede a tomar decisões sobre determinado alojamento, confiando assim nas palavras de hóspedes antigos que descrevem de uma forma detalhada toda a sua estadia. Logo, pode ser sugerido que a forma como os proprietários expressam-se no contexto *online* possa influenciar a estimativa dos potenciais hóspedes quanto à confiabilidade dos proprietários.

De acordo com o estudo de Festila e Dueholm Müller (2017), os hóspedes que participaram nesta pesquisa referiram que ofereceram pequenos presentes apesar da taxa da reserva ser cobrada. Isto demonstra que o hóspede sente a necessidade de agradecer por ter ficado no alojamento, mesmo pagando pela estadia e em certa parte surge um senti-

mento de confiança tanto do lado do hóspede como do proprietário. Simplesmente o proprietário reserva os seus alojamentos a estranhos *online*, o que desde logo tem de haver empatia. Os hóspedes por meio de agradecimento tendem assim a retribuir pelo fato de terem sido tão bem-recebidos. No entanto, e neste mesmo estudo um testemunho de um hóspede referiu curiosamente que apesar de não ter interagido com os seus proprietários, sentiu a necessidade pelo ato de hospedagem nas suas propriedades de uma pessoa estranha. A importância deste atributo ressalta uma distinção fundamental entre os alojamentos do Airbnb e os hotéis tradicionais, destacando assim parte da proposta de valor única que o Airbnb introduziu. A interação social no segmento hoteleiro é quase nula, não havendo o tipo de proximidade que se pode observar na plataforma Airbnb. A relação que é mantida com o proprietário pode sempre levar a uma experiência mais pessoal, ou talvez um relacionamento de proximidade, fazendo sentir o hóspede mais involucrado na experiência. No entanto, esta experiência é mais significativa quando o hóspede reserva a sua estadia num quarto privado ou partilhado, pois fica alojado na casa do proprietário havendo sempre um tipo de relacionamento diferente aquando da reserva da casa inteira, onde não existe tanta proximidade. No entanto, os autores Festila e Dueholm Müller (2017) indicam que, para alguns utilizadores o Airbnb é apenas uma experiência semelhante à de um hotel, mas a um custo relativamente mais baixo.

2.2 *Text Mining*

Text mining é o processo de explorar e analisar grandes quantidades de dados de texto, dados não estruturados ou semiestruturados, por forma a extrair informações úteis de diferentes recursos escritos para propósitos específicos (Tan et al., 1999). É uma tecnologia que auxilia por *software* a possibilidade de identificação de conceitos, padrões, tópicos, palavras-chave e outros atributos das bases de dados disponíveis para análise. O *text mining* é diferente daquilo com que o utilizador está familiarizado na pesquisa na *web*. Nesta pesquisa, o utilizador está tipicamente a procurar por algo que já é conhecido e que foi escrito por outro utilizador. O problema destas pesquisas é deixar de lado todo o material que atualmente não é relevante para as suas necessidades, de modo a encontrar a informação relevante. No *text mining*, o objetivo é descobrir até agora informações desconhecidas, algo que ninguém ainda sabe e portanto, não poderia ter ainda escrito[15] (Gaikwad, 2014). O *text mining* é muitas vezes comparado com o *data mining* (Gaikwad, 2014; Pande e Khandelwal, 2014). No entanto, as ferramentas tradicionais de *data mining* são incapazes de lidar com dados no formato textual, uma vez que exigem tempo e esforço para a extração de informações (Talib et al., 2016). Como a forma mais natural de armazenar informações é o texto, acredita-se que o *text mining* tenha um potencial comercial superior ao *data mining* (Sumathy, 2013; Tan et al., 1999).

[15] `http://people.ischool.berkeley.edu/~hearst/text-mining.html`, acedido a 08-12-2018

De acordo com Tan et al. (1999), 80% da informação de uma empresa tem como conteúdo documentos textuais. Este conteúdo apresenta uma quantidade de informação de grande utilidade para as organizações, como opiniões dos clientes ou saber como a empresa está no mercado. As ferramentas de *data mining* são projetadas para lidar com dados estruturados de bases de dados, mas o *text mining* tem como principal atividade analisar grandes conjuntos de dados não estruturados ou semiestruturados, como e-mails, documentos de texto integral e arquivos *HTML,* etc. (Pande e Khandelwal, 2014) O *text mining* é um campo multidisciplinar, que é constituído pela extração de texto, transformação de texto, processamento de texto, métodos de *clustering* e classificação (Sumathy, 2013; Tan et al., 1999). O *text mining* tem como principal objetivo resolver problemas do mundo real e para este estudo em questão, será analisada a informação dos comentários dos hóspedes, por forma a encontrar padrões e tendências, visto que os utilizadores (proprietários) são sobrecarregados com um grande volume de informação na *web* e precisam cada vez mais de métodos para os ajudar na sua tomada de decisão.

2.2.1 Desafios e questões do *text mining*

A complexidade da linguagem natural é a principal questão desafiadora no *text mining*, muitos problemas ocorrem durante o processo de *text mining*, que causa efeito na eficiência e eficácia aquando da tomada de decisões. Estas complexidades normalmente advêm da etapa intermédia do *text mining* (Talib et al., 2016). A ambiguidade é um dos problemas que está bastante associado à linguagem natural, pois tem a capacidade de ser compreendido de duas ou mais maneiras possíveis, isto porque uma palavra pode ter vários significados e várias palavras podem ter o mesmo significado. Esta ambiguidade leva ao ruído na informação extraída, no entanto, não pode ser totalmente eliminada da linguagem natural, pois dá flexibilidade e usabilidade (Gaikwad, 2014; Sumathy, 2013). Existem muitas possibilidades de interpretação de uma frase ou frases de um determinado documento, que podem levar a muitos significados. Embora uma série de pesquisas tenham sido realizadas para resolver o problema da ambiguidade, o trabalho ainda é imaturo (Gaikwad, 2014). De acordo com os autores Gaikwad (2014); Sumathy (2013), com técnicas de *text mining* conseguiram encontrar facilmente os nomes das diferentes entidades e as relações que possam existir entre elas podendo ser facilmente encontrados no *corpus* de documentos. Contudo, algumas análises semânticas são especialmente caras e muitas vezes operam na ordem de poucas palavras por segundo.

Para além destes desafios, a integração do conhecimento de domínio é uma área importante, pois realiza operações específicas num *corpus* específico e atinge os resultados desejados. De acordo com os requisitos da área, especialistas são necessários para trabalhar de forma colaborativa nos diversos domínios para extrair resultados mais eficazes e precisos (Talib et al., 2016).

2.2.2 Aplicabilidade do *text mining*

Em termos de aplicabilidade da técnica de *text mining*, a mesma pode ser empregue nas distintas áreas (Sumathy, 2013; Talib et al., 2016):

- Banca, seguradoras, mercados financeiros;
- No âmbito académico e de investigação;
- Indústrias de cuidados de saúde e farmacêutica;
- *Social Media* e tecnologias de informação;
- *Business Inteligence*, bioinformática e segurança nacional;
- Telecomunicações, energia e outros serviços.

Áreas de aplicação como motores de pesquisa, sistema de gestão de relacionamento com clientes, filtros de e-mails, análise de sugestões de produtos e deteção de fraudes utilizam *text mining* para *opinion mining*, extração de recursos, sentimento, previsão e análise de tendências (Talib et al., 2016). Para além da aplicabilidade, é também importante perceber a interação que existe com outros domínios (Talib et al., 2016), que se encontra ilustrada na Figura 2.3.

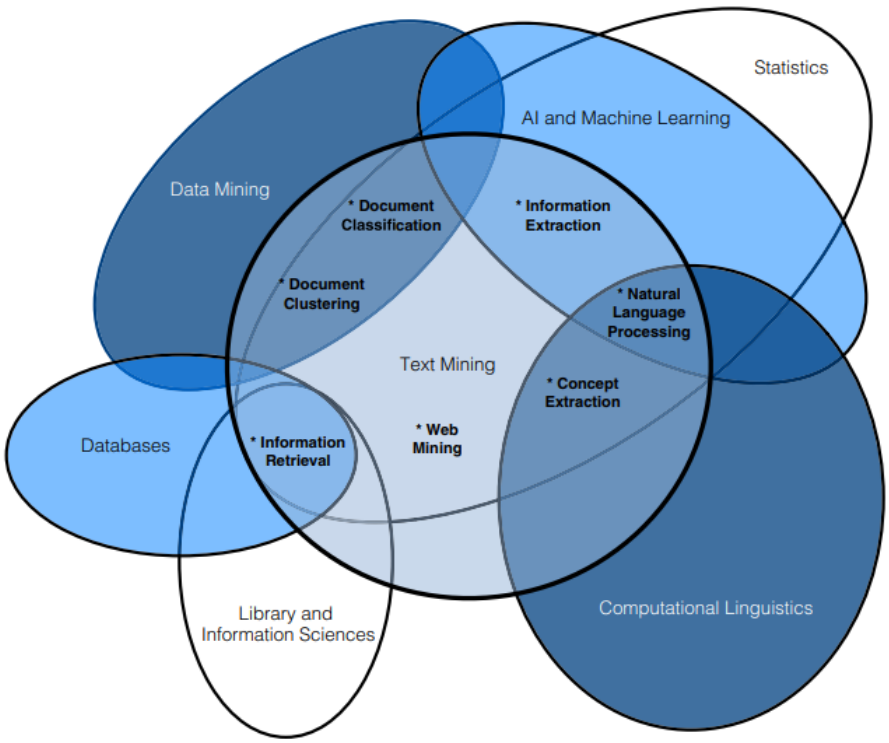


Figura 2.3: Diagrama de *Venn* da interação entre o *text mining* e outros domínios Talib et al. (2016)

2.2.3 Framework do *text mining*

A *framework* do *text mining* é constituída por duas fases genéricas:

1. O *Text Refining*, que consiste na transformação de documentos de texto de forma livre, para um formato intermédio que tende a ser um formato estruturado ou semi-estrututrado.

2. O *Knowledge Destillation*, que efetua a extração de padrões ou conhecimento dos documentos que estão no formato intermédio, logo informações relevantes de acordo com os objetos de interesse num domínio específico. O *clustering*, visualização e categorização de documentos, são exemplos de *mining* a partir de um formato intermédio com base em documentos (Sumathy, 2013; Tan et al., 1999).

A Figura 2.4 apresenta esta *framework*.

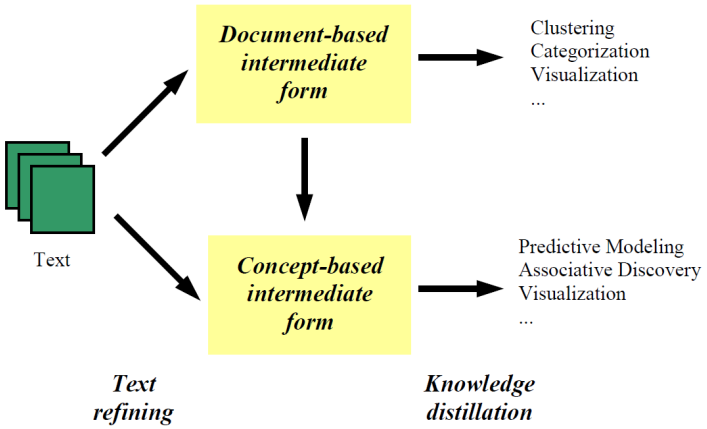


Figura 2.4: *Framework* genérico do *text mining*. Fonte: (Sumathy, 2013; Tan et al., 1999)

A *framework* anterior instancia as seguintes etapas:

- Extração da Informação;
- Transformação da Informação;
- Processamento da Informação;
- *Data Mining*;
- Avaliação e análise dos resultados.

Extração da Informação

A extração da informação é uma tarefa que permite extrair automaticamente informações estruturadas de documentos não estruturados e/ou semiestruturados (Kumar e Bhatia, 2013). O método de extração da informação tem por objetivo, localizar itens específicos dentro de um documento textual não estruturado, para posteriormente estruturá-los em formatos processáveis por máquinas, por exemplo base de dados relacional ou arquivos *XML*[16].

Transformação da Informação

Para a transformação da informação são contempladas algumas abordagens que permitem a representação do documento.

Bag of Words Esta abordagem permite a representação de cada documento como um vetor ponderado de termos. O seu objetivo é a identificação do peso associado a cada termo que corresponde ao número de ocorrências desse termo no documento, ignorando tanto a gramática, como a ordem em que estes aparecem no texto (Blake, 2011).

Document-by-term matrix (DTM) Esta matriz é constituída pelas colunas que representam os termos que aparecem em qualquer parte do *corpus*, pelas linhas que representam os documentos e na intersecção entre estes é possível identificar-se o valor da frequência absoluta do termo no documento. Cada entrada da matriz é uma contagem do número de vezes que o termo correspondente aparece em cada documento. A matriz documento-termo é uma tabela estruturada de números que pode, em princípio, ser analisada utilizando técnicas padrão. No entanto, na prática, a dimensão da matriz muitas vezes fica muito grande, criando desafios computacionais e de memória. Assim, as análises baseadas no *text mining* requerem um pré-processamento especial do *corpus*, onde o objetivo geral é enfatizar as palavras significativas removendo as não-informativas e manter o número de termos únicos que aparecem no *corpus* (Mankad et al., 2016).

Term frequency-inverse Document frequency (TF-IDF) Na abordagem TF-IDF é escolhido um vocabulário básico de termos e para cada documento do *corpus*, é feita uma contagem do número de ocorrências de cada palavra. Após uma normalização adequada, a contagem de frequência do termo é comparada a uma contagem inversa de frequência de documentos, que mede o número de ocorrências de uma palavra em todo o *corpus*. O resultado final é um *document-by-term matrix,* cujas colunas contêm

[16] `https://document.onl/documents/text-mining-infufscbr-alvaresine5644g2textopdf-augusto-fredigo-hack.html`, acedido a 15-03-2019

os valores TF-IDF para cada um dos documentos do *corpus*. Assim, o TF-IDF reduz documentos de comprimento arbitrário a listas de números de comprimento fixo (Blei et al., 2003). A frequência do termo $(tf_{i,j})$ poderá ser calculada como:

$$tf_{i,j} = \frac{n_{i,j}}{n_j} \tag{2.1}$$

onde $n_{i,j}$ é o número de ocorrências do *token* i no documento j e n_j é o número total de *tokens* no documento j.

O IDF é usualmente calculado como:

$$idf = log\left(\frac{N}{df_i}\right) \tag{2.2}$$

onde N é o número de documentos da base de dados e df_i é a frequência dos documentos que têm o *token* i. Finalmente, o peso de cada *token* de cada análise é calculado como (Xu et al., 2017):

$$w_{i,j} = tf_{i,j} \times idf_i \tag{2.3}$$

Processamento da Informação

É nesta fase que se processa toda a estruturação do documento, através da etapa de pré-processamento do texto.

Processamento de Linguagem Natural (PLN) O processamento da linguagem natural (PLN) diz respeito ao auto processamento e análise de informação textual não estruturada. É um campo da ciência da computação, que inclui a manipulação da linguagem humana e a compreensão do sistema informático. Pode-se dizer que é a forma de o sistema informático compreender e analisar a linguagem humana e os seus significados, e transformar e alcançar alguma informação útil e relevante (Sumathy, 2013; Talib et al., 2016). O PLN divide-se em duas tarefas que são a geração automática de texto (GAT) e a compreensão da linguagem natural (CLN). A GAT garante que o texto gerado seja gramaticalmente correto e fluente. A maioria dos sistemas GAT incluem um realizador sintático para assegurar que as regras gramaticais são estabelecidas (nomes, verbos, adjetivos, entre outros). A CLN consiste em pelo menos um dos seguintes componentes: *tokenizer*, análise do léxico, análise de sintaxe e análise do semântico (Sumathy, 2013). Para a extração de sinónimos e abreviaturas de dados textuais, a técnica de co-referenciamento é frequentemente utilizada para PLN. A linguagem natural é muito complexa, isto porque um texto extraído de diferentes fontes não tem palavras idênticas ou abreviaturas. Existe sempre a necessidade de detetar tais questões e estabelecer regras para a sua identificação uniforme (Talib et al.,

2016). Um dos exemplos da sua aplicabilidade, e de acordo com (Joshi et al., 2018) indica que esta técnica tem sido utilizada para analisar o comportamento humano com base nas suas atividades no *Social Media*. Algumas das técnicas utilizadas para o processamento da informação, são o *tokenization, part-of-speech,* remoção *stopwords*, *stemming,* entre outras.

Data Mining

A abordagem de *data mining* refere-se a encontrar informação relevante ou a descobrir conhecimento a partir de grandes volumes de dados. O *data mining* tenta descobrir regras e padrões estatísticos de forma automática a partir de dados. As ferramentas de *data mining* podem prever comportamentos e tendências futuras, permitindo que as empresas tomem decisões baseadas em conhecimento positivo. O objetivo geral do processo de *data mining* é extrair informações de um conjunto de dados e transformá-las numa estrutura compreensível para posterior análise (Kumar e Bhatia, 2013; Sumathy, 2013). Como exemplos nesta fase, utilizam-se os métodos de classificação, *clustering*, *mapping*, etc.

2.2.4 Análise de sentimentos

A análise de sentimentos é um tipo de análise de texto, sob a ampla área de processamento de linguagem natural, linguística computacional e *text mining*, que analisa o sentimento numa determinada unidade textual com o objetivo de compreender as polaridades das opiniões expressas e os tipos de emoções em relação aos vários aspetos de um assunto. Para Redhu et al. (2018), a análise de sentimentos é também conhecida como *Opinion Mining* e é o processo de quantificar o valor emocional de uma série de palavras ou texto, para obter uma compreensão das atitudes, opiniões e emoções expressas. A polaridade dos sentimentos é uma característica particular do texto, normalmente dicotomizado em positivo e negativo, ou por um intervalo de valores, onde por exemplo, valores entre]0, 1] o texto é positivo, e com valores entre [-1 , 0[o texto é negativo. Além de que, o sentimento pode mudar quando são utilizados *emoticons*, capitalização, pontuação, etc.

Segundo Thet et al. (2010), é cada vez mais importante analisar as opiniões expressas em várias plataformas *web* para a tomada de decisões eficazes das organizações. Os sentimentos tais como opiniões, atitudes, pensamentos, julgamentos e emoções, são estados privados dos indivíduos que não estão abertos à observação ou verificação objetivas. São expressos num tipo de linguagem que usa expressões subjetivas (Thet et al., 2010). Utilizando a análise de sentimentos e o *text mining*, as organizações podem obter *insights* do consumidor a partir da resposta sobre os seus produtos e serviços. Poderá ter ainda mais utilidade no estudo da satisfação dos clientes com os serviços e em caso de reclamações e problemas, encontrar as possíveis causas para isso (Redhu et al., 2018). Os avanços na tecnologia da *Internet* e o desenvolvimento de aplicações P2P têm causado mudanças subs-

tanciais na indústria do turismo. Neste momento, a *Internet* é o principal meio da pesquisa de informações sobre viagens. Os turistas comunicam cada vez mais uns com os outros e partilham as suas perspetivas e experiências em *sites* de redes sociais, gerando um número infinito de comentários e opiniões diárias.

De seguida, é descrita uma das técnicas utilizadas para este tipo de análise. Existindo tantas outras, que permitem igualmente classificar o sentimento de um determinado texto.

Valence Aware Dictionary and sEntiment Reasone (Vader) Disponibilizado na biblioteca NLTK[17] da linguagem de programação *Python*, o léxico *vader*, baseia-se em léxicos de palavras relacionadas com o sentimento, sendo que cada palavra do léxico é analisada quanto ao seu sentimento positivo ou negativo. Através do *vader*, são produzidas quatro métricas: positiva, negativa, neutra e *compound* ou agregada. Estas métricas, como o próprio nome indica, identificam as polaridades positivas, negativas e neutras de cada texto, sendo que, a métrica *compound* permite calcular a soma de todas as classificações do léxico, sendo estas normalizadas com os valores entre -1 e +1. O texto é considerado positivo se o *score* do *compound* ou *score* agregado for superior ou igual a 0,5, neutra se o *score* for superior a -0,5 e inferior a 0,5 e negativo se o *score* for inferior ou igual a -0,5 (Karim e Das, 2018).

2.2.5 Modelação por tópicos

A modelação por tópicos tem como objetivo analisar as palavras de um conjunto de documentos, descobrindo os temas que estão associados. Desta forma, e através de algoritmos de aprendizagem não supervisionada, são identificados os tópicos através dos textos dos documentos. Estes tópicos são conjuntos de palavras que ocorrem em documentos e que estão semanticamente relacionados entre si (fazendo sentido dentro do mesmo contexto).

De seguida, são abordadas as várias metodologias usadas para a descoberta de tópicos subjacentes a partir de um conjunto de documentos de texto.

Latent Semantic Analysis (LSA) O *Latent Semantic Analysis (LSA)* ou *Latent Semantic Indexing (LSI)*, é uma técnica estatística que, utiliza a matriz *document-by-term* (mencionada na Subsecção 2.2.3) que provavelmente é a forma mais comum de estruturar os dados contidos no *corpus*. Esta matriz contém uma coluna para cada termo que aparece em qualquer parte do *corpus* e uma linha para cada documento (Mankad et al., 2016). Também chamada de análise de componentes principais (PCA), o produto deste processo é uma série de vetores ortogonais, chamados vetores próprios e uma série de valores próprios que refletem a variância que é capturada em cada um destes vetores (Pearson, 1901). O benefício de usar esta abordagem, segundo Blake

[17] `https://www.nltk.org/`, acedido a 04-01-2019

(2011) é que estes vetores capturam palavras com um significado semelhante, mesmo que as palavras possam não aparecer num único documento. Diferente do modelo de espaço vetorial, no qual os documentos de texto são mapeados num espaço vetorial multidimensional literalmente construído a partir do vocabulário dos documentos de texto, o LSA constrói o espaço vetorial usando técnicas de redução de dimensão como a *Singular Value Decomposition* (SVD)[18] e mapeia os documentos de texto no espaço vetorial de ordem superior. Esta abordagem supera parcialmente o problema da variabilidade das escolhas de palavras humanas no modelo espacial vetorial (Xu et al., 2017). O LSA tem sido aplicado em áreas enormes como sistemas de recomendação (Resnick e Varian, 1997; Xu et al., 2017), recuperação de imagens (Cascia et al., 1998; Xu et al., 2017), reconhecimento de voz (Bellegarda, 2000; Xu et al., 2017), e recuperação de vídeo (Xu et al., 2017).

Latent Dirichlet Allocation (LDA) A modelação por tópicos refere-se a uma classe de algoritmos que resume automaticamente grandes arquivos de texto descobrindo "tópicos" ocultos ou temas que são discutidos dentro de um conjunto de documentos. O LDA é um algoritmo poderoso e amplamente utilizado na modelação por tópicos onde a estrutura de tópicos ocultos é inferida a partir de textos originais usando uma estrutura probabilística (Blei e Lafferty, 2009). A ideia central por detrás deste método é que todos os documentos partilham o mesmo conjunto de tópicos, mas cada documento apresenta uma mistura probabilística diferente desses tópicos. Além disso, as palavras são distribuídas de forma diferente para diferentes tópicos. Por outras palavras, é mais provável que certas palavras sejam usadas com determinados tópicos. O LDA emprega uma estrutura de estimação bayesiana (probabilidade de um evento, dado que outro evento já ocorreu) não supervisionada para os dados de texto fornecidos para inferir os tópicos e decompõe cada documento numa mistura de tópicos (Mankad et al., 2016). No âmbito do turismo, o LDA é o mais adequado para lidar com grandes comentários *online* não estruturados, criando assim significados que são mais realistas (Guo et al., 2017). Na Figura 2.5 é possível visualizar o processo do modelo LDA.

Através da Figura 2.5 pode-se descrever o processo do LDA como, dado o número M de documentos, o número N de palavras e o número K anterior de tópicos, o modelo é treinado para produzir:

- *psi*, a distribuição de palavras para cada tópico K*;*
- *phi*, a distribuição de tópicos para cada documento i.

[18] `http://web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm`, acedido a 10-01-2019

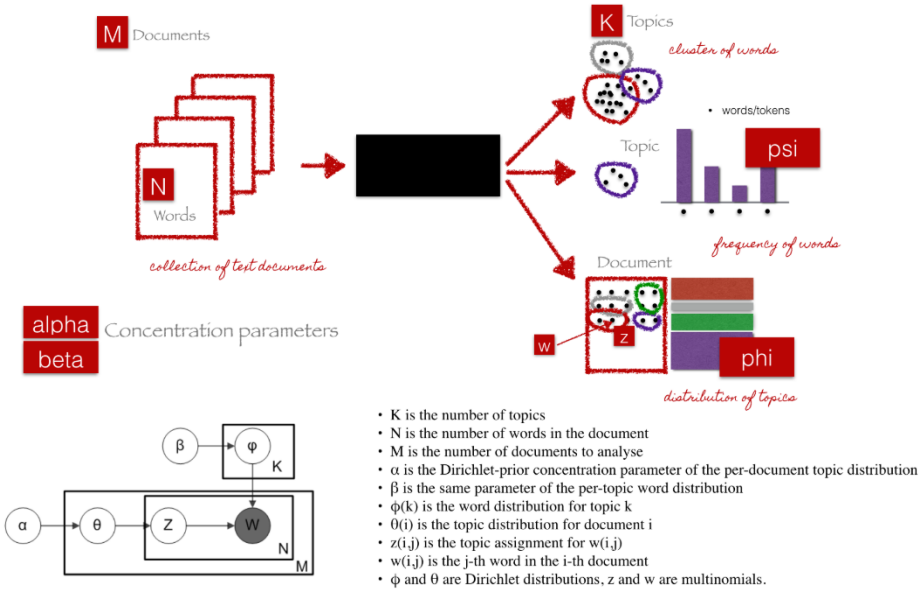


Figura 2.5: Processo do Modelo de Latent Dirichlet Allocation. Fonte: *PyTexas 2015*, acedido a 10-01-2019

Hierarchical Dirichlet Process (HDP) O *Hierarchical Dirichlet Process* (HDP), surgiu como uma metodologia para encontrar padrões e estruturas em grandes quantidades de informação de dados. Sendo uma metodologia aplicada ao texto, o HDP é um modelo probabilístico que permite que cada documento exiba vários tópicos (Williamson et al., 2009). O número de tópicos não precisa de ser especificado com antecedência e pode crescer à medida que a coleção de dados cresce. Além disso, o HDP permite que documentos não analisados anteriormente despoletem tópicos até nunca antes vistos. Esta é uma propriedade particularmente atrativa para analisar coleções em crescimento e mudança. O HDP permite que novos tópicos surjam naturalmente como consequência do modelo de probabilidade (Blei et al., 2010).

2.2.5.1 Avaliação dos modelos

Para a avaliação dos modelos aqui presentes (LDA e HDP), é utilizada a métrica da perplexidade, em que dado um modelo treinado a perplexidade tenta medir, através da probabilidade logarítmica normalizada qual o comportamento do modelo quando recebe um novo conjunto de dados. Quanto menor a perplexidade, melhor é o modelo (Blei et al., 2003). De seguida, é mostrada a fórmula da perplexidade (2.4) (Whye Teh et al., 2006):

$$exp\left(-\frac{1}{I}logp\left(w_1,...,w_I|DadosTreino\right)\right) \tag{2.4}$$

onde p (·) é a função de probabilidade para um determinado modelo.

2.2.5.2 Medidas de coerência de tópicos

As medidas de coerência de tópicos classificam um único tópico medindo o grau de similaridade semântica entre as palavras com pontuação mais alta no tópico. Essas medidas ajudam a distinguir entre tópicos que são semanticamente interpretáveis e tópicos que são artefactos de inferência estatística (Stevens et al., 2012). De seguida, é mostrada a fórmula da coerência (2.5):

$$CoherenceScore = \sum_{i<j} score(w_i, w_j) \tag{2.5}$$

onde w_i, w_j são as principais palavras do tópico.

Existem as mais variadas medidas de coerência, mas nesta secção apenas serão apresentadas três delas (Pradhan et al., 2016):

Métrica *C_v* Esta medida baseia-se numa *sliding window*[19] ou janela deslizante de tamanho 110, em que mede a coerência de um tópico utilizando uma variação da NPMI[20] e a similaridade de *cosseno* (entre o vetor de cada palavra e a soma de todos os vetores de palavras);

Métrica *C_uci* É uma medida que se baseia numa janela deslizante de tamanho 10 e na abordagem PMI (abordagem para medir a associatividade entre duas palavras) de todos os pares de palavras das palavras principais fornecidas;

Métrica *C_umass* Para esta medida, a coocorrência de uma palavra do tópico é suportada por todas as palavras anteriores que tenham uma classificação de ordem superior no tópico. Logo, para cada palavra, o logaritmo da sua probabilidade condicional utilizando todas as outras palavras anteriores é calculado e a média aritmética dessa soma fornece a medida da coerência.

[19] Subconjunto de palavras sucessivas de tamanho N. Pode ser deslocado palavra por palavra para qualquer um dos lados da janela.

[20] *Normalized Pointwise Mutual Information* (NPMI) - Abordagem utilizada para medir a associatividade entre duas palavras, normalizando o valor obtido para o intervalo [-1,1]. Em que o limite inferior -1 significa nenhuma coocorrência, o valor 0 significa independência entre as duas palavras e 1 significa uma coocorrência completa.

3 Revisão da Literatura

Neste capítulo, irá se proceder à análise da literatura de forma a entendermos como é modelada a experiência do hóspede em relação às ofertas de alojamento dos proprietários. Deste modo, através da literatura relacionada, analisou-se os aspetos mais relevantes da experiência do hóspede no Airbnb, bem como abordagens utilizadas para a classificação dos sentimentos associados aos comentários escritos pelos hóspedes. Outro dos pontos que também será abordado, é a relação entre os comentários negativos dos hóspedes, com as informações dos proprietários sobre os alojamentos, ou seja, aquilo que o hóspede observou na página de determinado alojamento do Airbnb e ao presenciar na realidade constatou que o alojamento não estava conforme. Relativamente à classificação em estrelas, para o Airbnb não é possível verificar se a classificação de uma a cinco estrelas para cada alojamento, coincide corretamente com o que é descrito nos comentários, ou seja, comentários negativos que podem corresponder a alojamentos bem classificados e vice-versa, deste modo, foram identificados alguns estudos que abordam esta análise. Por fim, foi necessário analisar estudos referentes à identificação das categorias relevantes de um *superhost* no Airbnb, de modo a compreender melhor se este aspeto tem impacto na relação do hóspede com o alojamento, como é o caso de um hóspede preferir ou não os alojamentos cujos proprietários são *superhosts*. Todos os temas descritos anteriormente relativos a este capítulo estão diretamente relacionados com as questões de investigação abordadas na Secção 1.3, do Capítulo 1, no entanto, constatou-se que estes estudos focam-se mais na descrição dos resultados e não tanto na descrição da modelação utilizada para estes estudos.

3.1 Aspetos mais relevantes da experiência do hóspede no Airbnb

As experiências são formas de compreender as interações entre pessoas e lugares, essencialmente produzidas de uma forma pessoal na medida em que cada indivíduo as entende e interpreta de forma diferente (Paulauskaite et al., 2017; Jennings e Weiler, 2006; Pine II e Gilmore, 1998). No entanto, Pine II e Gilmore (1998) definem a noção de experiências como *"aquelas que o cliente encontra únicas, memoráveis e sustentáveis ao longo do tempo, que gostariam de repetir e construir, e que entusiasticamente são promovidas por meio do “passa a palavra”"*.

Nesta secção, analisam-se os estudos que estão relacionados com as experiências dos hóspedes, nomeadamente, quais os aspetos que os hóspedes observam como mais importantes e que estão diretamente relacionados com o próprio alojamento ou até mesmo com o proprietário.

Tussyadiah e Zach (2017) focam-se na identificação de atributos proeminentes dos alojamentos P2P através dos comentários *online* escritos pelos hóspedes. A análise dos dados para este estudo seguiu várias etapas, que incluem pré-processamento, análise lexical e de *clusters*, utilizando o *software* de *text mining KH Coder*[1]. O pré-processamento do conjunto de dados foi realizado pelos autores, segundo lista abaixo e conforme o exemplo da seguinte Figura 3.1. Para a técnica de *Part-Of-Speech Tagging*, os autores usaram o *Stanford POS Tagger*[2].

1. *Tokenization*: O texto é dividido em palavras e frases;
2. Remoção *stopwords*: remoção de palavras comuns, que não transmitem informação, tais como artigos definidos ou indefinidos e verbos auxiliares, incluindo *a*, *an*, *and*, *the*, etc.;
3. *Part-Of-Speech Tagging (POS)*: é realizada a atribuição de uma etiqueta a cada palavra baseada na sua morfologia e contexto, tais como substantivo, verbo, adjetivo, etc.;
4. Lematização: processa as palavras, por forma a ignorar o tempo verbal ou o género, ou até mesmo o número dos substantivos (singular e plural), de modo a identificar a raiz da palavra.

Original text:	A great place to stay! The space is clean and comfortable. The hosts are warm and helpful.
Tokenization:	A /great /place /to /stay /! /The /space /is /clean /and /comfortable /. /The /hosts /are /warm /and /helpful /.
Elimination of stop words:	/great /place //stay /! //space //clean //comfortable /. //hosts //warm //helpful /.
Part-of-speech (POS) tagging and lemmatization:	great/Adj place/Noun stay/Verb ! space/Noun clean/Adj comfortable/Adj . host/Noun warm/Adj helpful/Adj .
Preprocessed text:	Great place stay ! space clean comfortable. host warm helpful.

Figura 3.1: Exemplo de pré-processamento (Tussyadiah e Zach, 2017)

Os dados usados neste estudo foram obtidos no *site Inside Airbnb*, que disponibiliza a informação que está visível publicamente na plataforma Airbnb. O conjunto de dados contém

[1] `https://khcoder.net/en/`, acedido a 01-03-2019
[2] `https://nlp.stanford.edu/software/tagger.shtml`, acedido a 01-03-2019

informações sobre os alojamentos da cidade de Portland em Oregon, Estados Unidos da América. Com a eliminação de dados omissos e com a seleção dos comentários apenas no idioma inglês, o *dataset* ficou com 41.560 comentários de 1.617 propriedades e em média cada comentário contém cinco frases (ou seja, um total de 215.497 frases no conjunto de dados).

O *dataset* em questão contém 3.530.597 *tokens* e 33.059 palavras diferentes. Após a eliminação de *stopwords*, 1.473.197 *tokens* e 21.561 palavras diferentes (ou seja, termos representativos). Os autores procederam à análise de *Term Frequency* (TF), que permite observar o número de ocorrências de palavras no *dataset*. O TF médio é de 68,33 (ou seja, as palavras aparecem 68 vezes em média). A análise revela os atributos mais frequentemente mencionados nos comentários dos hóspedes por serem indicativos dos seus fatores de satisfação. De seguida, os autores analisaram os termos mais relevantes sobre a experiência de alojamento P2P, para obter palavras compostas importantes (N-gramas), que podem ser uma combinação de duas palavras (bigrama), uma combinação de três palavras (trigrama) e assim por diante. Estes termos foram identificados utilizando uma abordagem de reconhecimento automático de termos, através do módulo *TermExtract* incluído no programa *KH Coder*.

Através da análise de *clusters,* os autores identificaram grupos de comentários que discutem tópicos semelhantes, representando atributos de alojamento P2P. Para se considerar se os comentários são semelhantes ou diferentes, o estudo utilizou a distância de *Jaccard*, que compara o peso da soma dos termos partilhados com o peso da soma dos termos que estão presentes em qualquer um dos dois documentos. Estes documentos contendo conjuntos de palavras semelhantes geralmente discutem o mesmo tópico. Portanto, neste estudo, a análise de *clusters* auxilia na identificação de grupos de comentários que discutem tópicos semelhantes, representando atributos dos alojamentos de P2P. A análise hierárquica de *clusters* produziu deste modo cinco *clusters* de comentários. A respetiva identificação dos temas pode ser observada na Tabela 3.1.

Temas	**Número de comentários**
Serviço	19.463
Instalações	4.353
Localização	6.890
Acolhimento	3.451
Conforto de um lar	7.263

Tabela 3.1: Temas identificados e respetivo número de comentários

Relativamente ao identificado anteriormente, a maioria dos comentários identificados por serviço e localização (65% e 60%, respetivamente) foram escritos sobre experiências de alojamentos numa casa inteira, enquanto que os comentários relativos a instalações, acolhimento e conforto de um lar foram distribuídos uniformemente pelas experiências de alojamento numa casa inteira e num quarto privado. Os comentários sobre experiências

num quarto partilhado foram extremamente reduzidos (1 a 2% de todas as opiniões).

Os comentários no *cluster* serviço contêm palavras que descrevem as características de serviço dos proprietários. Estes comentários, incluem a comunicação (ou seja, comunicação acessível com os proprietários), capacidade de resposta (os proprietários respondem rapidamente a perguntas, comunicação imediata), processo de reserva e tempo de *check-in* e *check-out*. O *cluster* instalações centra-se nos aspetos físicos do bem, incluindo o espaço (por exemplo, quarto, casa de banho, cozinha), bens ou utensílios (por exemplo, cama, toalhas) e outras comodidades. Para a localização, foram abordados os tópicos referentes às descrições das localizações, bem como vantagens da localização em termos de proximidade com outros pontos de interesse, como é o caso de restaurantes e lojas e também o acesso ao transporte público facilitado, e por fim, as características das zonas onde as propriedades estão localizadas. Quanto ao acolhimento, os autores referem-se à sensação de estar "em casa" enquanto se encontram num alojamento P2P. Os comentários abordados neste *cluster* estão diretamente relacionados com as interações sociais com os proprietários e também com a dedicação que os proprietários têm em acolher e alojar os hóspedes. Por fim, no *cluster* conforto de um lar, o grupo de comentários cingiu-se ao conforto do alojamento, à sua envolvente e à hospitalidade dos proprietários.

Comparativamente com o estudo anterior, o estudo de Guo et al. (2019) foca-se na análise de uma plataforma distinta do Airbnb. A plataforma *online* de partilha de alojamentos da China, Xiaozhu, permitiu aos autores explorarem os aspetos mais importantes para os hóspedes chineses. Esta é uma plataforma que existe somente na China, permitindo a partilha de casas inteiras e quartos privados. Os autores escolheram a cidade de Pequim para a extração dos dados desta plataforma, pois é a capital da China que atrai um grande número de visitantes de todas as partes do mundo que compram, viajam e fazem reservas em alojamentos todos os anos. Para a extração dos dados, os autores incluíram os alojamentos com data de 17 a 25 de dezembro de 2018. O resultado final foi de 20.571 comentários de hóspedes de casas inteiras e 6.020 comentários de hóspedes de quartos privados, com aproximadamente 20.000 casas inteiras e 7.000 quartos privados.

Ao contrário dos idiomas ocidentais, não existem espaços entre as palavras de uma frase em chinês. Assim, este estudo adotou o *software ROST CM 6.0*[3] para identificar a frequência das palavras nos comentários. Este *software* suporta dicionários personalizados e foi usado para distinguir e extrair palavras de alta frequência e palavras emocionais relevantes para este estudo.

Como primeira análise, os autores analisaram os comentários dos hóspedes na plataforma Xiaozhu e estabeleceram um dicionário personalizado, que incluía as palavras «host», «facilities», «location», «decoration», «traffic», «price», entre outras. De seguida, aplicaram um filtro de palavras não relacionadas ao dicionário, tais como «I», «and» e «in». Com base nas palavras de alta frequência resultantes, os autores realizaram uma análise

[3] `https://www.cnblogs.com/ROST123/p/6547861.html`, acedido a 05-03-2019

de *cluster* para identificar os principais aspetos que influenciam a experiência dos hóspedes. Foram ainda realizadas análises de redes semânticas para obter o relacionamento e a conexão entre as palavras, por forma a verificar o significado de e entre as palavras.

Relativamente aos resultados de *text mining* e de acordo com o *software ROST,* as palavras mais frequentes para o alojamento casa inteira são «host», «clean», «convenient», «room», «location», «subway», entre outras. Após este processo, foi realizada uma matriz de comentários vs. termos mais frequentes, onde a linha refere-se a um comentário específico e a coluna representa um termo específico. Sendo que, a célula da matriz é codificada como 1 quando um comentário específico menciona esse termo e 0 quando um comentário específico não menciona determinado termo. De seguida, os autores realizaram uma análise hierárquica de *cluster* usando o *SPSS* 16.0 para classificar as palavras de alta frequência. As palavras como «next time», «super», «satisfied», «house» ou «experience» foram consideradas como irrelevantes e não consideradas na análise de *cluster*, incluindo assim 21 palavras de frequência alta na análise de *clusters*, conforme Figura 3.2.

Number of clusters	Case																																							
	Problem		Host		Communication		Hospitality		Room		Tidy		Cleanliness		Sanitation		Location		Convenient		Transportation		Supermarket		Subway		Environment		District		Warm		Quietness		Decoration		Facility		Kitchen		Cooking
1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
2	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
3	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
4	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x		x	x	x	x	x
5	x	x	x	x	x	x	x		x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x		x	x	x	x	x
6	x	x	x	x	x		x		x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x		x	x	x	x	x
7	x	x	x	x	x		x		x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x		x		x	x	x
8	x	x	x	x	x		x		x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x		x	x	x	x	x		x	x	x		x		x	x	x
9	x	x	x	x	x		x		x	x	x	x	x	x	x		x	x	x	x	x		x	x	x		x	x	x	x	x		x	x	x		x		x	x	x
10	x	x	x	x	x		x		x	x	x		x	x	x		x	x	x	x	x		x	x	x		x	x	x	x	x		x	x	x		x		x	x	x
11	x	x	x	x	x		x		x		x		x	x	x		x	x	x	x	x		x	x	x		x	x	x	x	x		x	x	x		x		x	x	x
12	x		x	x	x		x		x		x		x	x	x		x	x	x	x	x		x	x	x		x	x	x	x	x		x	x	x		x		x	x	x
13	x		x	x	x		x		x		x		x	x	x		x	x	x	x	x		x		x		x	x	x	x	x		x	x	x		x		x	x	x
14	x		x	x	x		x		x		x		x	x	x		x	x	x	x	x		x		x		x	x	x	x	x		x	x	x		x		x		x
15	x		x	x	x		x		x		x		x	x	x		x	x	x	x	x		x		x		x	x	x	x	x		x		x		x		x		x
16	x		x	x	x		x		x		x		x	x	x		x	x	x	x	x		x		x		x		x	x	x		x		x		x		x		x
17	x		x	x	x		x		x		x		x	x	x		x	x	x	x	x		x		x		x		x		x		x		x		x		x		x
18	x		x	x	x		x		x		x		x		x		x	x	x	x	x		x		x		x		x		x		x		x		x		x		x
19	x		x	x	x		x		x		x		x		x		x		x	x	x		x		x		x		x		x		x		x		x		x		x
20	x		x	x	x		x		x		x		x		x		x		x		x		x		x		x		x		x		x		x		x		x		x

Figura 3.2: Diagrama com a análise de *clusters* e termos de alta frequência para a casa inteira (Guo et al., 2019)

Cinco *clusters* foram assim identificados, incluindo «host service», «facilities», «location and transportation», «cleanliness» e «living environment».

No caso do quarto privado, os autores realizaram a mesma análise para a casa inteira. Sendo que, para este caso foram identificados seis *clusters,* «host service», «location and transportation», «security and privacy», «facilities», «cleanliness», «value for money» e «living environment».

De acordo com Tussyadiah (2015), foi reconhecido que o preço influencia a satisfação da experiência do hóspede, mas Guttentag et al. (2017) defende totalmente o inverso. Para estes autores, os principais atributos das experiências são as comodidades e o espaço. A

autenticidade também tem sido destacada por vários estudos, nomeadamente por Guttentag et al. (2017). Além disso, Guttentag (2015) considerou a interação com os locais como parte da autenticidade, mas Tussyadiah (2015) e Tussyadiah e Pesonen (2015) posicionaram essa interação separadamente, como parte de um benefício social desfrutado com o uso do Airbnb.

Apesar dos vários pontos de vista, as dimensões comumente estabelecidas que formam a experiência Airbnb são o preço (Guttentag e Smith, 2017), localização (Tussyadiah e Zach, 2017), comodidades (Guttentag, 2015), limpeza (Bridges e Vásquez, 2018), interação do proprietário com o hóspede (Festila e Dueholm Müller, 2017) e passar tempo em bairros locais (Tussyadiah e Zach, 2017).

Para além da procura de maior autenticidade na experiência do hóspede no Airbnb, os autores observam mais as experiências turísticas e a questão de uma experiência ser memorável. De acordo com Pizam (2010), *"criar experiências memoráveis é a essência e a razão de ser da indústria hoteleira"* (p. 343). De acordo com Oh et al. (2007), uma experiência memorável *"leva a uma maior memória, ou seja, a recordar um acontecimento particular, que moldará de forma positiva a atitude do turista em relação ao destino"* (p. 123). Por fim, Tung e Ritchie (2011) estabeleceram a relação entre resultados extraordinários e memoráveis da experiência através da noção consequencial, ou seja, o fato de a experiência se tornar memorável se existir algum tipo de importância percebida a partir do resultado da viagem.

3.2 Identificação de comentários positivos e negativos

Para a identificação das falhas e necessidades descritas pelos hóspedes nos vários alojamentos é efetuada uma extração e análise aos comentários, por forma a serem identificados estes contratempos. Neste sentido, a categorização dos comentários como positivos e negativos é um ponto de partida para esta análise.

O estudo de Bridges e Vásquez (2018), focou-se na identificação de comentários positivos e negativos, analisando 400 comentários escritos por hóspedes entre março e setembro de 2015 de quatro cidades dos Estados Unidos da América (Portland, Albuquerque, Philadelphia e Atlanta), correspondendo a 24.130 palavras. Dividindo a amostra para cada cidade (100 comentários para cada), onde 50 dos comentários correspondem à perspetiva dos hóspedes relativamente aos alojamentos, com 18.539 palavras e com uma média de 75 palavras, variando entre os comentários mais curtos com cerca de 15 palavras e os mais longos com até quase 400 palavras. Por fim, os restantes 50 comentários correspondem à análise que efetuaram da perspetiva dos proprietários relativamente aos hóspedes, que não será dado enfoque nesta secção.

Para a análise em questão, os autores observaram os comentários escritos *online* e de forma manual procederam à categorização dos comentários positivos e negativos. Se-

guindo uma abordagem analítica do sentimento foram considerados como indicadores de um comentário positivo, o uso de adjetivos como, *great, wonderful*, entre outros. No que respeita aos comentários negativos, os adjetivos*, difficult, terrible,* etc*.,* bem como as negações, *not, un-, but*, foram igualmente considerados. Para além dos adjetivos positivos e negativos que identificam a polaridade de um comentário, existem outros recursos que foram explorados: como a intensificação de advérbios (por exemplo *very, seriously, definitely*) e a pontuação para ênfase. Posteriormente, os autores procederam à identificação de todas as menções de palavras num determinado domínio semântico utilizando o *software AntConc*[4], devido à abundância de comentários que abordavam o aspeto do conforto dos alojamentos (por exemplo, *cozy, coziness, comfy, comfortable, comfort*). Deste modo, os autores conseguiram ainda apurar através deste *software*, a frequência de palavras, bem como os padrões de palavras comumente coocorridos, verificando que 89,5% dos hóspedes mencionam o proprietário nos comentários do alojamento. Normalmente, os comentários referentes ao proprietário são positivos e tendem a abordar os seus comportamentos. No entanto, os autores verificaram que para além de um comentário global ser parcialmente negativo, não significa que quando se referem ao proprietário tenham uma crítica também negativa.

A maioria dos comentários (93%) foram categoricamente positivos em termos de linguagem.

Bridges e Vásquez (2018) ilustram vários aspetos interessantes nos exemplos seguintes.

Nos exemplos 1 e 2, os hóspedes tiveram uma boa experiência referindo o conforto e a interação social com o proprietário. No exemplo 3, os hóspedes referem a interação positiva com o proprietário, mas descrevem de forma negativa a experiência na casa.

Exemplo 1. *Our experience was great. The bedding was seriously comfortable, lodging cleverly appointed to meet all your needs and Paul a very considerate host. We would definitely recommend the 'Funky Pad'.*

Exemplo 2. *Josh did a fantastic job of making us feel welcome – he left us a personalized note, bottle of wine, and the house was in great shape when we arrived ….*

Exemplo 3. *Lane, who is the host, could not have been more pleasant. A wonderful guy. However, this rental is a terrible place… .*

Por fim, os autores identificaram que os comentários descritos de forma positiva se referem maioritariamente à limpeza do alojamento e à comunicação entre o hóspede e o proprietário.

No que concerne aos comentários negativos, os mesmos estão em minoria relativamente aos comentários positivos. Em geral, os autores identificaram que os comentários com avaliação negativa começam e terminam com comentários positivos, surgindo uma

[4] `https://www.laurenceanthony.net/software/antconc/`, acedido a 07-03-2019

reclamação no meio do comentário. No exemplo 4, os autores verificaram pelo comentário que o hóspede é compreensivo e cortês, indicando que, na maior parte das vezes do comentário que a estadia foi aceitável.

Exemplo 4. *The main floor bedroom bed was very comfortable, the upstairs 2 just okay; It was pretty hot each day and, while the upstairs room had ac, the main floor bedroom didn't [. . .] The upstairs bathroom is big, the main floor one is miniscule. Been on a cruise ship before? That kind of small. The kitchen could use a few more basics (paper towels, micro- wave) but is comfortable and appointed with IKEA everything [. . .].*

Foi ainda possível identificar pelos autores que a falta de conforto é o aspeto mais referido nos comentários negativos, de seguida a comunicação e por fim a limpeza.

Os autores verificaram que a análise em questão engloba mais comentários positivos do que negativos, pondo assim em causa a credibilidade dos mesmos. Embora, neste mesmo estudo, foram identificados pelos autores 19 comentários com a classificação *lukewarm* ("tépidos"), ou neutros. Na realidade, estes comentários não tem uma conotação negativa, nem tem uma conotação positiva. Os autores verificaram, que estes comentários são normalmente descritos quando o hóspede pretende comunicar uma avaliação não-positiva.

Com uma análise um pouco distinta da anterior, o estudo de Cheng e Jin (2019), aplicou técnicas de *text mining* na comparação entre a plataforma Airbnb e os hotéis tradicionais. O *dataset* foi retirado do *site Inside Airbnb*, com um total de 181.263 comentários escritos por hóspedes que relatam as suas experiências nos alojamentos Airbnb de Sydney, Austrália. No entanto, com a limpeza dos dados em idiomas diferentes de inglês, através do *software OpenRefine*[5] o *dataset* ficou reduzido a 170.124 comentários.

Foi utilizada a análise de sentimentos para identificar as opiniões positivas e negativas dos hóspedes, e para esta análise, os autores utilizaram um léxico e um método híbrido de aprendizagem automática, através do *software Leximancer*[6]. Um dos módulos do *Leximancer* é composto por listas de sentimentos positivos e negativos em inglês e itens de negação. Relativamente ao método híbrido de aprendizagem automática, este começa por remover termos óbvios do léxico e depois identifica outros termos relacionados dos dados.

Para a identificação de temas e conceitos da plataforma Airbnb, o *Leximancer* produz um mapa de calor como resultado final, onde os temas são codificados por cores (quente - frio) para indicar a proeminência dos temas, onde palavras que têm fortes significados semânticos são agrupadas.

De acordo com os resultados, os autores chegaram a quatro tópicos principais extraídos dos comentários escritos por ordem de importância da percentagem de documentos,

[5] `http://openrefine.org/`, acedido a 07-03-2019
[6] `https://info.leximancer.com/`, acedido a 07-03-2019

que incluem, localização (100%), comodidades (81%), proprietário (70%) e recomendação (18%).

Relativamente à localização e apesar de ser um dos tópicos mais importantes, os hóspedes relataram algumas experiências negativas associadas a este tópico (espaço do estacionamento, ruído à noite e ambientes inseguros), no entanto, foi tratado como um problema menor, desde que achassem que a localização do alojamento fosse apropriada, perto de pontos de interesse e próxima de transportes públicos. Sobre o tópico das comodidades, este está relacionado com o conforto do alojamento e as instalações do local, quanto aos comentários escritos, os hóspedes do Airbnb valorizaram os sentimentos de como se estivessem em casa, tendo sempre à sua disposição todos os utensílios necessários para uma boa experiência. Sendo que, os comentários negativos ocorreram quando o aspeto das instalações não era preciso, relativamente ao que foi descrito na página Airbnb pelo proprietário. Relativamente ao proprietário, são abrangidos diversos conceitos com conotação positiva, como a ajuda, a flexibilidade e a comunicação que os proprietários têm com os hóspedes. Com comentários negativos para este aspeto, foi indicado pelos hóspedes de que muitas das vezes são acusados pelos proprietários indevidamente. Os hóspedes referem que recebem algumas instruções que os proprietários solicitam que sejam seguidas na sua propriedade, no entanto, o hóspede manifesta a sua preocupação quando sente que estas indicações são demasiado rígidas. Sendo que, é também referido pelos hóspedes que por parte do proprietário está em falta as instruções para o uso de alguns serviços na propriedade. Em alguns dos comentários negativos identificados pelos autores, referem que a comunicação entre proprietário e hóspede não ser a mais flexível, e muitos dos hóspedes referem que tiveram de esperar muito tempo pelo proprietário para fazerem o próprio *check-in*.

Ju, Yongwook e Back, Ki-joon e Choi, Youngjoon e Lee, Jin-soo (2019) realizaram a análise de sentimentos para os alojamentos das cidades de Miami, Nova York, São Francisco e Chicago, Estados Unidos da América. Com um conjunto de dados com cerca de 16.430 comentários e 103 alojamentos, verificaram que a média de comentários por alojamento é de 160 e a média de palavras por comentário é de 53.

De acordo com a análise da alta frequência das palavras com o *software QDA Miner* 5[7], os autores identificaram os atributos mais relevantes para o hóspede (*host, room/house, location, neighborhood*). Por forma a restringirem o seu estudo, Ju, Yongwook e Back, Ki-joon e Choi, Youngjoon e Lee, Jin-soo (2019) realizaram a análise através do *software SentiStrength*[8] e focaram-se apenas em dois comentários, com a maior conotação positiva e conotação negativa respetivamente. Como este estudo tem como objetivo a identificação de atributos de qualidade do serviço do Airbnb para a satisfação do cliente, procederam à análise dos atributos encontrados no comentário positivo e negativo. Desta forma, no comentário positivo, verificaram que o proprietário, o conforto, a limpeza e a localização, são aspetos importantes para os hóspedes. Para os comentários negativos, os hóspedes

[7] `https://provalisresearch.com/news-events/qda-miner-5-released/`, acedido a 07-03-2019
[8] *`http://sentistrength.wlv.ac.uk/`*, acedido a 07-03-2019

descreveram a ausência de interação social com o proprietário.

Nas análises anteriores foi possível observar que existem mais comentários positivos do que negativos. Uma possível razão é apontada por Bridges e Vásquez (2018) que referiram no seu estudo que o Airbnb reserva o direito de não exibir comentários quando necessário, ou seja, quando exista algum comentário que ataque diretamente o proprietário, o Airbnb utiliza o seu direito de censurar o mesmo.

No estudo de Guo et al. (2019) (referido na Secção 3.1) utilizaram a análise de sentimentos no *software ROST CM 6.0*, sendo que as palavras emocionais são definidas pelo *CNKI-Net*[9], que inclui palavras emocionais positivas, como «satisfaction», «hospitality», «happy», «great», «warm» e «comfortable» e palavras emocionais negativas, como «noisy», «bad», «stressful», «sorry» e «insufficient». De acordo com Tussyadiah e Zach (2017) sugeriram que os comentários negativos dos alojamentos P2P devem ser analisados para fornecer aos proprietários diretrizes para melhorias. Deste modo, os autores apenas focaram-se nos sentimentos negativos dos hóspedes neste tipo de alojamentos, conforme se pode observar na Tabela 3.2. No geral, os hóspedes que ficaram em casas inteiras e em quartos privados partilharam menos emoções negativas do que positivas. No entanto, o atributo com os comentários mais negativos tanto para as casas inteiras como para quartos privados é o atributo *cleanliness*.

Attributes	**Keywords**	**Proportion of Negative Reviews (%)**	
		Entire House	**Private Room**
Cleanliness	sanitation	10.12%	13.20%
	bathroom	18.44%	32.43%
	room	6.19%	14.56%
Host service	host	7.55%	5.87%
	service	5.90%	3.49%
	aunt	1.63%	0.69%
	communication	2.73%	1.10%
Location and transportation	location	13.72%	8.20%
	transportation	8.72%	5.98%
	subway	4.50%	2.56%
	supermarket	2.28%	1.12%
	bus	2.50%	0.89%
Living environment	decoration	0.78%	2.23%
	layout	0.65%	1.03%
	environment	2.82%	5.26%
	district	0.67%	0.56%
Value for money	value for money	3.87%	0.39%
Facilities	facilities	7.23%	10.78%
	kitchen	17.40%	3.72%
Security and privacy	safety	2.47%	4.57%

Tabela 3.2: Comparação do sentimento negativo entre casas inteiras e quartos privados (Guo et al., 2019)

[9] `https://www.cnki.net`, acedido a 08-03-2019

Guo et al. (2019) ilustram alguns exemplos de comentários negativos (exemplo 5 e 6) e positivos (exemplo 7) nos exemplos seguintes.

Exemplo 5. *The space is really unacceptable. The quilt is musty and the house is old. The bathroom is completely leaky and the outside is wet. This experience was bad.*

Exemplo 6. *The kitchen is really a mess!*

Exemplo 7. *The room is very neat and beautiful, the decoration style is consistent with the pictures, the interior is very beautiful.*

No exemplo 5, os hóspedes referem o atributo *cleanliness*, indicando a má higiene da casa de banho. No exemplo 6 e para o atributo *facilities,* os hóspedes referem a desorganização da cozinha. Por fim, no exemplo 7, os hóspedes mostram-se positivos no atributo *living environment*, onde referem que as fotos do anúncio são precisas com a realidade. Os autores referem ainda que os comentários negativos são indicadores para o próprio proprietário poder melhorar certos aspetos no seu alojamento.

Outro estudo que engloba uma análise diferente do verificado anteriormente é a dos autores Sixto et al. (2013). Estes autores não se focaram em nenhum caso específico e simplesmente desenvolveram diversos modelos por forma a verificar qual o melhor modelo para este tipo de análises. Inicialmente, com técnicas de aprendizagem automática e processamento de linguagem natural (PLN), criaram um classificador baseado em modelos de entropia máxima para avaliar de forma positiva e negativa os comentários. Para desenvolver o classificador, os autores utilizaram a ferramenta *Stanford NLP Classifier*[10].

Seguindo a *framework* baseada em técnicas de *text mining* apresentada na Figura 3.3, os autores realizaram a análise a 1.000 comentários do setor hoteleiro.

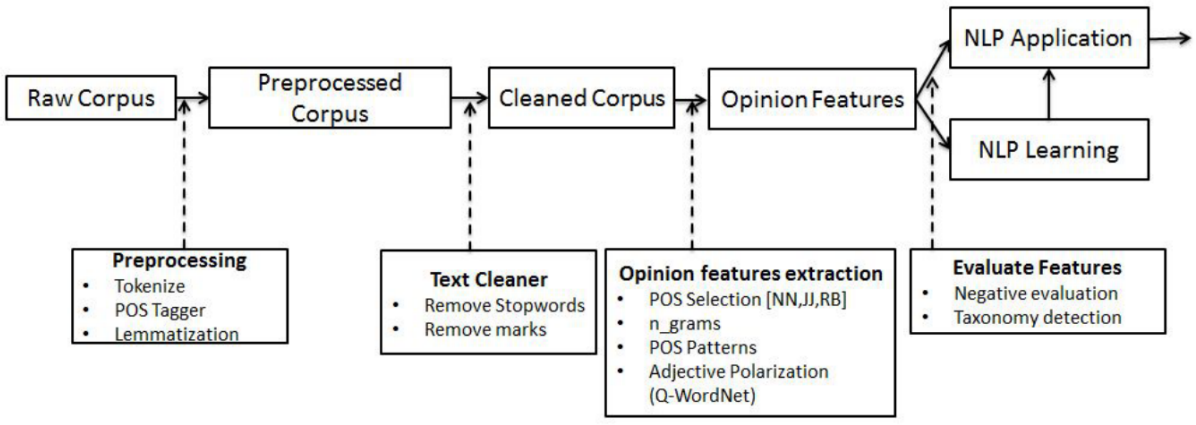


Figura 3.3: *Framework* utilizada no estudo de Sixto et al. (2013)

Como primeiro passo, realizaram a tokenização das frases por forma a analisarem as palavras. De seguida, limparam o texto de símbolos irrelevantes e procederam à remoção

[10] `https://nlp.stanford.edu/software/classifier.shtml`, acedido a 08-03-2019

de *stopwords*. Para poderem classificar as palavras com os lemas e as etiquetas *part-of-speech* (POS), os autores utilizaram o *Stanford POS Tagger*. Por fim, para estabelecerem a polaridade e verificarem se uma palavra ou frase é negativa ou positiva, foi utilizado o *Q-Wordnet*[11]. Gerando bigramas de etiquetas POS, relacionaram substantivos com adjetivos e a polaridade atribuída pelo *Q-Wordnet.*

Quanto aos resultados, descritos na Tabela 3.3, os autores verificaram que quando apenas foi utilizada a etiqueta POS, o valor da taxa de acerto *(accuracy)* do modelo diminuiu. No entanto, a adição da polaridade e dos *tokens* negativos melhorou significativamente o modelo, com 85,1% de taxa de acerto *(accuracy)*.

Components	**Accuracy** (%)	**Precision** (%)	**Recall** (%)	**F-score** (%)
Base classifier	79.8	79.6	78.6	79.9
Base + POS	79.5	79.4	78.5	78.7
Base + ClearText + POS	80.0	79.8	78.5	78.9
Base + ClearText	80.5	80.6	78.8	79.3
Base + ClearText + Polarity	83.0	82.8	81.0	81.6
Base + ClearText + Polarity + POS	83.2	82.7	81.9	82.2
Base + Polarity + Negative	82.9	80.6	79.4	79.6
Base + ClearText + Polarity + Negative	**85.1**	**85.7**	**83.6**	**84.0**

Tabela 3.3: Resultados da classificação dos comentários (Sixto et al., 2013)

3.3 Relação entre as perspetivas dos hóspedes e a oferta dos alojamentos

Segundo vários autores (Lehr, 2015; Cheng e Jin, 2019; Tussyadiah e Zach, 2017) e o próprio Airbnb, a relação encontrada entre a descrição do próprio alojamento e a experiência do hóspede passa pela classificação da precisão ou *accuracy* estabelecida no próprio alojamento. Através desta classificação, o hóspede tem a possibilidade de escolher entre uma a cinco estrelas, de acordo com a experiência que presenciou em determinado alojamento. No entanto, descrições completas e fotos reais do alojamento, normalmente dão origem a pontuações altas por parte dos hóspedes (Ju, Yongwook e Back, Ki-joon e Choi, Youngjoon e Lee, Jin-soo, 2019; Tussyadiah e Zach, 2017). Contudo, como se trata de uma análise muito específica da plataforma Airbnb, não existe trabalho relacionado que possa descrever esta secção. A única análise possível foi através da análise de sentimentos dos comentários negativos dos hóspedes, onde é descrito de seguida.

No estudo de Cheng e Jin (2019) foi possível identificar o relacionamento da experiência do hóspede com as ofertas do alojamento através de técnicas de *text mining*, mais propriamente a análise de sentimentos. Neste caso, pela categorização dos sentimentos negativos os autores conseguiram desvendar que os hóspedes escreveram comentários onde

[11] `http://wordnet.princeton.edu/`, acedido a 08–03-2019

referiram que as comodidades dos alojamentos não estavam conforme o referido pelos proprietários na página do alojamento. No excerto abaixo, pode ver-se o comentário negativo, relativamente à propriedade:

The listing was accurate except that I wasn't informed that the washing machine had blown up (which was a big problem and in- convenience for me). I really felt the terrace could have been cleaned so that it was usable – let's just say it did not look like this picture! Unfortunately, it's somewhat run down and in need of maintenance: the stove's gas rings don't all work and the doors are starting to fall off; the couch is stained and torn with legs falling off; there were broken lamps and various missing lightbulbs; a lot of non-working junk sitting around too. While food and drinks are generally not expected in the Airbnb experience, provision of these can be interpreted by Airbnb guests as a delight. Every little detail was thought of right down to bottles of water at your bedside. We only stayed there for one night but wish we could have stayed a lot longer.

Tradução: A descrição do alojamento era precisa, exceto que eu não fui informado de que a máquina de lavar estava avariada. Realmente senti que o terraço poderia ter sido limpo para que pudesse ser utilizado - digamos que não se parecia com a foto! Infelizmente, precisa de manuntenção! As bocas de gás do fogão não funcionam e as portas estão a cair; havia falta de lâmpadas e lâmpadas fundidas; muito lixo espalhado também. Geralmente não são esperados alimentos e bebidas na experiência do Airbnb, o fornecimento dessas pode ser interpretado pelos hóspedes do Airbnb como um brinde. Cada pequeno detalhe foi pensado e foram colocadas garrafas de água dos lados da cama. Ficamos lá apenas por uma noite, mas gostaria que pudéssemos ter ficado muito mais tempo.

3.4 Relação entre a classificação em estrelas e os comentários

A classificação em estrelas no Airbnb é obtida sobre o que o hóspede classifica para determinado alojamento, embora na realidade tal não significa que esta classificação esteja alinhada com os comentários. Neste sentido, os autores Tussyadiah e Zach (2017) (referidos na Secção 3.1) através do *software SPSS* procederam à análise de regressão das classificações em estrelas da experiência geral, que vão de uma a cinco estrelas como variáveis dependentes, mas associando os *clusters* de comentários como variáveis independentes ("serviço", "instalações", "localização", "acolhimento" e "conforto de um lar"), por forma a compreender melhor se os diferentes aspetos identificados anteriormente contribuem para a satisfação dos hóspedes. Estas classificações de experiência geral incluem a "precisão", o processo de "*check-in*", a "limpeza", a "comunicação", a "localização" e o "valor". Um pequeno desvio na classificação destes aspetos, pode ser uma indicação da falta de satisfação do hóspede: assim, estes resultados fornecem informações úteis sobre aspetos da estadia em alojamentos P2P que contribuem positiva ou negativamente.

Os autores procederam às análises de regressão, onde os dados em análise consistem

nos vários alojamentos com pelo menos um comentário pertencente aos cinco *clusters*. Relativamente aos resultados observados na Tabela 3.4, os mesmos mostram que uma maior proporção de comentários do grupo "acolhimento" contribui positivamente para as classificações mais elevadas em todos os aspetos, com exceção da classificação "localização". Por outro lado, uma maior proporção de comentários de "serviço" contribui negativamente para todas as classificações dos aspetos. Por forma a compreender o porquê de o grupo "serviços" ter um impacto negativo nas classificações, os autores analisaram esses comentários de modo a identificarem termos negativos que possam indicar um desempenho inferior. Os autores não conseguiram encontrar problemas relacionados com o "serviço", mas surgiu um padrão em que os hóspedes que não estão totalmente satisfeitos com a sua estadia devido às condições da limpeza, ruído, falta de comodidades ou problemas durante a sua estadia, tais como, perturbações, erros de comunicação, etc., tendem a enfatizar a rapidez com que os proprietários oferecem soluções para estes problemas, como por exemplo, o fato de responderem rapidamente a questões. Para conclusão, os autores sugeriram que os comentários relativos a serviços estão associados a classificações mais baixas, não necessariamente porque os proprietários têm um desempenho fraco no atendimento aos hóspedes, mas muitos hóspedes destacam os aspetos positivos dos serviços de acolhimento quando escrevem comentários, por forma a compensar a sua insatisfação com a limpeza, valores e, portanto, a permanência geral.

	Overall rating	Accuracy	Cleanliness	Check-in	Communication	Location	Value
Model							
R^2	.102	.038	.042	.042	.047	.098	.075
F (sig.)	7.838 (.00)	5.438 (.00)	6.077 (.00)	6.483 (.00)	6.779 (.00)	14.879 (.00)	11.218 (.00)
Independent Variables – Beta (sig.)							
Service	−.309 (.00)	−.133 (.02)	−.224 (.00)	−.191 (.00)	−.144 (.01)	−.153 (.01)	−.227 (.00)
Facility	n.s.	n.s.	n.s.	n.s.	n.s.	−.202 (.00)	n.s.
Location	.215 (.00)	.188 (.00)	.207 (.00)	.133 (.02)	.125 (.02)	.408 (.00)	.123 (.02)
Welcome	.298 (.00)	.113 (.02)	.108 (.03)	.153 (.00)	.184 (.00)	n.s.	.226 (.00)
Comfort	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Tabela 3.4: Análise de regressão dos *clusters* identificados e da classificação em estrelas (Tussyadiah e Zach, 2017)

O estudo de Fan e Khademi (2014) teve como objetivo predizer as classificações em estrelas dos restaurantes através dos comentários da plataforma *online* Yelp[12]. A motivação destes autores é eliminar as classificações em estrelas dada individualmente por cada hóspede e utilizar somente a classificação do restaurante, ao invés do que acontece na plataforma Airbnb.

O *dataset* é composto pela categoria "restaurantes" (4.243 restaurantes), escolhendo aleatoriamente 1000 restaurantes e 35.645 comentários para análise. Para avaliarem os resultados, os autores dividiram os dados em 90-10, onde para o conjunto de dados de treino, utilizaram os comentários e a classificação em estrelas do restaurante e para o teste, utilizaram os seus modelos para prever a classificação do restaurante, para consequentemente

[12]Plataforma *online* para pesquisa de restaurantes, locais a visitar, entre outras. Fonte: `https://www.yelp.pt`, acedido a 02-04-2019

compararem com a classificação real.

Inicialmente, Fan e Khademi (2014) dividiram a análise em três experiências (uma análise *baseline* e duas análises de *features*) importantes:

1. Analisaram os dados para calcular as palavras mais frequentes em todos os comentários dos restaurantes;

2. Através de *Part-of-Speech* (POS), procederam à descoberta de quais as palavras mais representativas. No entanto, com o decorrer da análise, os autores verificaram que os adjetivos eram mais vantajosos, isto porque são o tipo de palavras mais comumente utilizadas para descrever positividade ou negatividade.

Tendo extraído informação relevante dos adjetivos, os autores trataram desta situação como um problema de regressão. Os quatro modelos de regressão utilizados são *Linear Regression*[13], *Support Vector Regression*[14], *Support Vector Regression with normalized features* e *Decision Tree Regression*[15]. Como o objetivo dos autores é predizer a classificação em estrelas do restaurante utilizando o modelo de regressão, foi utilizada o *Root Mean Square Error*[16] para quantificar o erro, ao invés de se utilizar a precisão.

A Figura 3.4, mostra os resultados dos quatro modelos de regressão, de acordo com as experiências efetuadas.

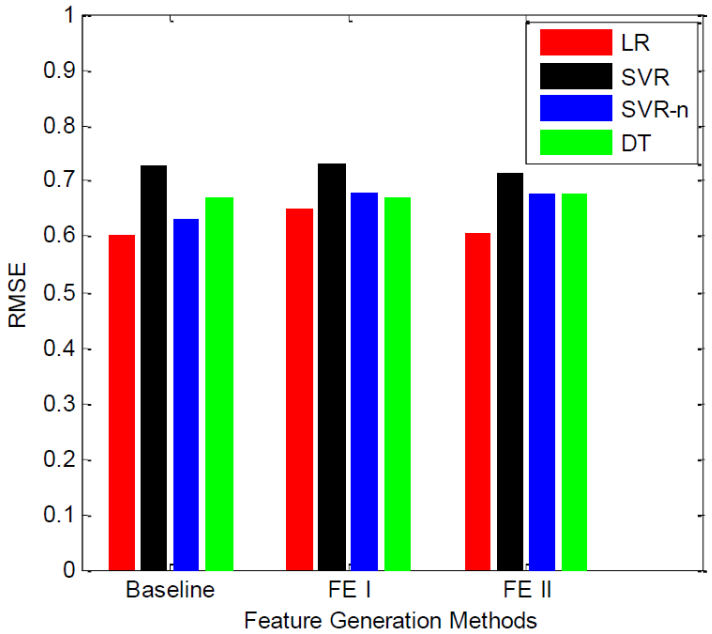


Figura 3.4: Resultados dos modelos para cada uma das experiências (Fan e Khademi, 2014)

[13] https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/#definition, acedido a 03-04-2019

[14] https://www.saedsayad.com/support_vector_machine_reg.htm, acedido a 03-04-2019

[15] https://www.saedsayad.com/decision_tree_reg.htm, acedido a 03-04-2019

[16] https://www.statisticshowto.datasciencecentral.com/rmse/, acedido a 03-04-2019

Relativamente aos modelos, os autores verificaram que a *Linear Regression* tem o melhor desempenho em geral. E inferiram que para os seus dados, as experiências utilizadas e a classificação podem estar linearmente correlacionadas.

3.5 Categorias relevantes de um *superhost* no Airbnb

Nesta última secção é possível verificar os métodos utilizados para analisar as categorias mais relevantes que um hóspede identifica num *superhost*.

O Airbnb implementou um programa de *superhost*, em que é atribuído a cada proprietário um ícone distintivo segundo determinados critérios de desempenho estabelecidos pela plataforma. Quando é conquistado este estado de *superhost* surge automaticamente um ícone distintivo nos seus anúncios e perfis para ser rapidamente identificável. Também, quando um hóspede faz uma pesquisa de um alojamento é possível no campo dos filtros selecionar os alojamentos cujo os proprietários são *superhosts*. De acordo com Liang et al. (2017), Gunter (2018) e Wang e Nicolau (2017), este ícone distintivo no perfil de cada *superhost* pode ser percebido pelos hóspedes como um sinal de qualidade do alojamento e compromisso do proprietário, resultando num número maior de reservas e aumento da receita para o proprietário. Isso demonstra que os hóspedes do Airbnb estão dispostos a pagar mais por um *superhost,* do que por um *host* regular. No entanto estes estatutos exigem que os perfis tendam a ser significativamente mais extensos do que os dos *hosts* regulares (uma média de 72,13 palavras comparado com 57,74 palavras de *hosts* regulares), segundo Ma et al. (2017). Por fim, com a prevalência do fenómeno da economia partilhada, há um número crescente de proprietários no Airbnb que gerem mais do que um alojamento. Possivelmente, a gestão de mais alojamentos, torna os proprietários mais experientes em termos de servir os hóspedes, mas de acordo com Xie e Mao (2017) pode prejudicar a qualidade do proprietário devido à capacidade de gestão mais restrita. Contudo, segundo Gunter (2018), os fornecedores comerciais do Airbnb são mais suscetíveis de obter o estatuto de *superhost*.

Os proprietários com este ícone distintivo tendem assim a interagir mais com os seus hóspedes e, normalmente, correspondem sempre às falhas e necessidades dos mesmos, oferecendo bons serviços de alojamento.

O estudo de Sun et al. (2019), com recurso a técnicas de *text mining*, como a classificação de texto e *clustering*, pretende identificar os comentários específicos de serviço escritos pelos hóspedes, comparando o *superhost* com o *host* regular. O *dataset* utilizado contém informação sobre todos os alojamentos Airbnb da cidade de Hangzhou, China. Após a remoção de duplicados, ficaram com um total de 43.584 comentários dos hóspedes, sobre 5.631 propriedades pertencentes a 2.669 proprietários, dos quais 447 são *superhosts*, correspondendo a 16,75%. Para garantir a qualidade dos dados, excluíram as observações com dados ausentes e focaram-se apenas nos comentários escritos em chinês (comentá-

rios em inglês, coreano e japonês foram removidos), obtendo uma amostra com 39.862 comentários. Depois de separar os comentários por pontuação, alcançaram 239.367 frases, totalizando 1.991.758 palavras e com uma média de 8,32 palavras por frase.

Para estudar os serviços específicos prestados pelos proprietários, foram realizadas duas etapas na análise destes comentários: utilizaram *Long Short-Term Memory* (LSTM) para separar os comentários relativos ao serviço do proprietário dos restantes (por exemplo localização do alojamento) e, posteriormente, é utilizado o *K-Means,* por forma a obter serviços específicos. O LSTM sendo uma rede neuronal recorrente comum que processa sequências, possui algumas modificações relativamente à *Recurrent Neural Network* (RNN). Esta para além de receber como entrada as *features* (palavras presentes no *corpus* de treino) processadas para representação vetorial, juntamente com uma matriz de pesos entra na camada oculta combinando informações, gerando assim uma saída para uma função de perda que calcula o quão longe ou perto o modelo está, de predizer uma saída correta. Nesta camada oculta existe um *loop*, ou seja, na próxima interação o modelo utiliza a informação que armazenou anteriormente e dá entrada, deste modo, vai usando as entradas anteriores para processar as sequências até ao final; o que numa rede com várias camadas ocultas, a camada oculta anterior está presente na seguinte camada oculta e assim sucessivamente. No entanto, o problema é que as RNNs tendem a esquecer o que aprenderam nas entradas anteriores, conforme cresce o número de camadas, este fenónemo dá-se o nome de *vanishing gradient.* O que torna este modelo ineficiente, para capturar informação contextual de sequências longas. Como tal foi utilizado o LSTM que através de um mecanismo específico nas camadas ocultas (boa capacidade de memória nas células) permite que o LSTM consiga se recordar das informações que armazenou mesmo depois de várias interações ocorrerem. A primeira camada de LSTM construída pelos autores é a camada de *word embeddings* (identificação da similaridade entre palavras e frases).

Graves (2012) adicionou mais um nível de verificação à abordagem LSTM, para uma melhor identificação sobre a utilidade da informação.

Para a aplicação do modelo descrito anteriormente (LSTM), os autores dividiram o *dataset* em treino e teste. A *accuracy* do modelo de treino foi de 99,12% e do modelo de teste foi de 98,61%. Após a aplicação deste modelo a todo o *corpus*, este passou para 14.046 comentários.

Quanto à especificação dos serviços, inicialmente foi utilizado um módulo que permite a segmentação dos comentários de serviços em chinês, e foram adicionadas ao dicionário palavras baseadas no Airbnb da cidade de Hangzhou. De seguida, os autores utilizaram a técnica TF-IDF para a remoção de *stopwords*. Após este último passo, os autores utilizaram o *K-Means*, para obter *clusters* por associação de similaridade entre objetos. Ou seja, dependendo da distância, quanto mais próximos os objetos, maior é a sua similaridade.

Inicialmente, definem-se *K* pontos como centros dos *clusters,* de acordo com certas regras (ou aleatoriamente). Calculando a distância entre cada objeto e o centro do *cluster*, os

pontos com a mínima distância do ponto central são classificados. Repete-se este processo várias vezes, reduzindo a área de distância ao ponto central, até que a distância dos objetos restantes seja coincidente com o centro de um determinado *cluster*.

Posteriormente, os autores utilizaram uma *WordCloud* para visualizar o *clustering* dos comentários em seis categorias diferentes:

1. Três refeições (pequeno-almoço, almoço ou jantar) ou lanches durante a noite;
2. Frutas, bebidas ou lanches;
3. Guias de viagem;
4. Transporte gratuito ou ajuda com a bagagem;
5. *Chats*;
6. Respostas ou comunicações.

Os autores procuraram saber que tipo de proprietário é preferencialmente escolhido pelos hóspedes, sendo que em média os *superhosts* estão mais presentes nos comentários relativos a serviços com 62% dos casos, enquanto os *hosts* regulares apenas chegam aos 38%. As três categorias em que os hóspedes dão mais ênfase aos *superhosts* são as categorias:

- Respostas ou comunicações;
- Frutas, bebidas ou lanches;
- Três refeições (pequeno-almoço, almoço ou jantar) ou lanches durante a noite.

Tendo em conta que este é um tema muito específico da plataforma, a literatura encontrada sobre esta análise e que utilize técnicas de *text mining* é muito reduzida. A maioria dos artigos encontrados referem-se a modelos econométricos, que não são considerados relevantes para o âmbito deste estudo.

4 *Modelação de comentários no Airbnb*

Neste capítulo é apresentada a abordagem realizada para efetuar a modelação de comentários no Airbnb. Inicialmente, este capítulo apresenta o problema que originou a análise deste estudo. De seguida, são apresentados os conjuntos de dados e o processamento efetuado para posteriormente ser utilizado na fase seguinte, a modelação. É nesta fase que serão demonstradas as estratégias ou técnicas utilizadas para responder a cada questão de investigação, com a ajuda de várias análises e modelos que foram desenvolvidos usando a linguagem de programação *Python*.

4.1 Problema

Os dados não estruturados têm tido um impacto significativo na sua quantidade e cada vez mais este tipo de informação gera conhecimento importante para a tomada de decisão nas organizações. Por forma a ser realizada a compreensão e leitura de grandes quantidades de dados não estruturados, devem ser utilizadas técnicas de recolha dos dados, para posteriormente serem pré-processados, garantindo a sua qualidade e tornando os mesmos mais adequados para determinado algoritmo, o que desta forma torna mais fácil a compreensão para o ser humano, que não tem capacidade para processar este tipo de informação em bruto. Por exemplo, os *social media* são uma fonte importante deste tipo de dados, em particular, devido ao crescimento dos comentários *online*. Os comentários *online* que para além de serem dados não estruturados, são na sua grande maioria fontes de informação textual ruidosa, pois contêm erros ortográficos, construções gramaticais incorretas, ou mistura de idiomas, e uma forma de lidar com estes aspetos é usar técnicas de *text mining*, de modo a extrair informação útil destes dados.

Para o presente estudo, foi recolhida informação do Airbnb acerca dos alojamentos em Lisboa, disponibilizada no *site Inside Airbnb*. O *Airbnb* trata-se de uma plataforma/aplicação *online*, que disponibiliza alojamentos de vários proprietários em vários locais por todo o mundo. Aquando de uma reserva num alojamento é possível ao hóspede poder partilhar a sua estadia ou experiência através de comentários *online* que se tornam disponíveis para todos os utilizadores desta plataforma. Estes comentários, cada vez mais em larga escala, refletem muitas vezes opiniões e sentimentos de experiências anteriores, ajudando, por um

lado, os proprietários de um determinado alojamento a reagirem mais rapidamente aos pedidos dos hóspedes, bem como a recolher informações quanto à perceção dos hóspedes em relação aos seus bens e serviços.

A análise destes comentários apoia o proprietário na compreensão das falhas e necessidades descritas pelos hóspedes, melhorando os níveis de serviço de cada alojamento disponível, levando também a um aumento da qualidade de serviço da plataforma Airbnb. Desta forma, os resultados obtidos através das técnicas de *text mining* ajudam não só o proprietário, mas também o hóspede e, ainda, a plataforma Airbnb.

4.2 Caraterização dos dados

A informação recolhida da plataforma *Inside Airbnb* (com dados de alojamentos e comentários até abril de 2019) refere-se a 22.242 alojamentos disponíveis na plataforma Airbnb para a cidade de Lisboa, Portugal, refletindo 769.636 comentários redigidos pelos hóspedes das experiências vividas após cerca de 8 milhões de reservas. Sendo o objetivo do presente estudo analisar a informação não estruturada, os dados recolhidos dizem respeito ao conteúdo dos comentários do hóspede e à descrição dos anúncios dos alojamentos dos proprietários. Sendo que, é igualmente importante efetuar a análise de dados estruturados da disponibilidade dos alojamentos durante um ano. Deste modo, serão analisadas três fontes de informação ou *datasets* no formato *CSV*, uma referente aos anúncios dos alojamentos, outra aos comentários dos hóspedes e outra referente à disponibilidade dos alojamentos. Na Figura 4.1 é possível verificar a relação entre elas.

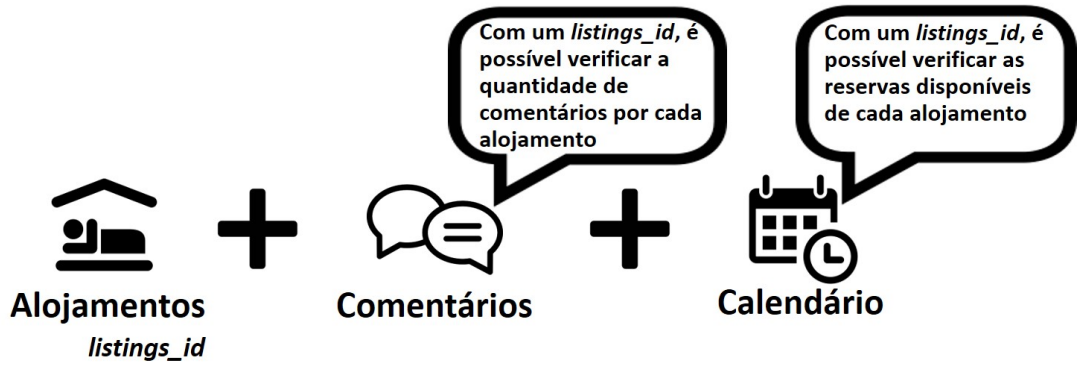


Figura 4.1: Esquema da relação entres os diferentes *datasets*

Na subsecção seguinte são abordadas e descritas as informações pertencentes a cada *dataset*.

4.2.1 Propriedades dos Alojamentos

Este *dataset* é constituído por 106 variáveis, sendo possível extrair informação detalhada de cada alojamento, desde o tipo de espaço: apartamento/casa inteira, quarto privado e quarto partilhado (variável *room_type*), descrição do alojamento (*description*), informação do próprio proprietário (variáveis *host_id, host_name, host_location, host_about, host_is_superhost*, entre outras), comodidades ou itens que o hóspede espera ter numa estadia confortável (variável *amenities*), respetivo *score* do alojamento atribuído por cada hóspede (0-100%), entre outras variáveis. No decorrer da análise, verificou-se que existiam várias variáveis que não eram relevantes para a análise, por exemplo *listing_url, scrape_id, last_scraped, experiences_offered, thumbnail_url, medium_url, picture_url*, etc., razão que levou a uma redução para apenas 29 variáveis, que se podem observar com mais detalhe no Anexo A, na Tabela A.1.

4.2.2 Propriedades dos Comentários

Este *dataset* é composto por apenas seis variáveis e dele é possível extrair a informação de todos os comentários relativos à estadia do hóspede em cada alojamento. As variáveis que compõem este *dataset* vão desde o comentário (variável *comments*), data do comentário *(*variável *date),* nome do hóspede *(*variável *reviewer_name),* bem como a identificação do alojamento no qual o hóspede teve a sua estadia, entre outras variáveis. Todas estas variáveis podem ser observadas no Anexo A, na Tabela A.2.

4.2.3 Propriedades do Calendário

O *dataset* calendário é formado por sete variáveis, em que, através destas é possível extrair a disponibilidade (variável *available*) de cada alojamento, bem como o preço por noite (variável *price*) e o mínimo e máximo de noites possíveis (variáveis *minimum_nights* e *maximum_nights*). Esta disponibilidade abrange um período de 2019-04-22 a 2020-04-21. No Anexo A, na Tabela A.3, é possível ver com mais detalhe a descrição de cada variável, bem como o seu tipo.

4.3 Preparação dos dados

A qualidade dos dados é fundamental para todo o processo de análise de dados, sendo possível medir propriedades dos dados de diferentes perspetivas. Deste modo, e apesar dos dados serem não estruturados, deve ser garantido que os mesmos são armazenados de forma adequada, em segurança e com consistência.

Esta fase tem como principal objetivo estruturar os dados iniciais num novo conjunto de dados final, por forma a serem utilizados na fase da modelação. Neste passo, é provável

que as tarefas da preparação de dados sejam realizadas várias vezes, e não num momento específico (Wirth e Hipp, 2000).

Os problemas que normalmente advêm da qualidade, como a inconsistência e a má qualidade dos dados, tendem a levar os decisores a tomar decisões erróneas ou a tirar conclusões imprecisas. No entanto, existem muitas maneiras de se poder medir a qualidade dos dados, com completude, conformidade, consistência, precisão, duplicidade e integridade.

Esta fase teve início no *dataset* dos alojamentos, comentários e, posteriormente calendário. Optou-se por fazer a preparação dos dados destes *datasets* em separado, possibilitando uma maior eficiência em qualquer análise que seja efetuada.

4.3.1 Dados dos Alojamentos

Neste *dataset*, o *corpus* foi alterado de acordo com os seguintes procedimentos, focando nas variáveis mais importantes a analisar. Após a menção destes pontos, será descrito cada procedimento detalhadamente.

Assim, a variável *description* segue as seguintes transformações*:*

- Remoção de valores omissos (NaNs);
- Substituição de carateres *n* e *r* por espaços em branco;
- Remoção dos dados em idioma diferente do inglês, dado ser a língua em que existem mais comentários (apresentado na secção seguinte);
- Remoção de *stopwords*;
- Remoção da pontuação e numeração;
- Lematização, com a biblioteca do *Python WordNetLemmatizer*[1]*;*
- Remoção de nomes de pessoas e palavras com apenas um carácter;
- Substituição de todas as palavras para palavras minúsculas.

No caso das variáveis *review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location* e *review_scores_value* apenas se procedeu a uma conversão para uma escala de 5 pontos.

Relativamente à variável *host_is_superhost* apenas se realizou a conversão para uma variável booleana (0 - Não; 1-Sim).

Um dos primeiros passos a investigar durante a análise da qualidade dos dados é a verificação se o *dataset* possui *missing values*. Os casos com valores omissos representam

[1] *http://www.nltk.org/_modules/nltk/stem/wordnet.html*, acedido a 14-08-2019

assim um importante desafio, sendo necessário definir uma estratégia para tratar destes valores, pelo que um valor ausente pode significar que os dados são imprecisos. No entanto, existem processos para proceder à correção destes valores, fazendo uma previsão sobre os valores omissos (identificando a percentagem de variáveis com valores nulos), eliminando-os ou substituindo-os. Neste caso, a implementação da estratégia passou por eliminar as entradas que continham estes valores. No Anexo A, na Tabela A.4, é possível ver os resultados obtidos.

No decorrer da análise verificou-se que principalmente na variável *description*, existiam frases que continham os carateres *\n* e *\r* (fim de linha). Estes carateres não faziam sentido estar na análise e foram retirados deste *dataset*. Assim, procedeu-se à substituição destes carateres por um espaço em branco. Após esta transformação, procedeu-se à tokenização das palavras, por forma a serem efetuadas determinadas tarefas como é o caso da remoção de *stopwords,* que correspondem a palavras comuns, tais como artigos definidos ou indefinidos e verbos auxiliares, incluindo *a*, *an*, *and*, *the*, etc., deste modo, foi executada a tarefa de remoção de *stopwords* para o idioma inglês. Dentro do mesmo processo, foi executado o código para lematização da variável *description* através da biblioteca *WordNetLemmatizer*. Outras das técnicas utilizadas, foram a remoção de nomes de pessoas, a remoção de palavras com apenas um carácter, a substituição de todas as palavras para palavras em letras minúsculas e ainda a remoção da pontuação.

Outra das situações foi a remoção de variáveis que não eram relevantes para a análise, tais como *listing_url, scrape_id, last_scraped, experiences_offered, thumbnail_url, medium_url, picture_url*, etc., razão que levou a uma redução deste *dataset* para apenas 29 variáveis. Com a remoção destas variáveis que não fazem sentido estar na análise, procedeu-se ainda à remoção de dados de alojamentos em idioma distinto de inglês. No Anexo A, na Tabela A.5 é possível verificar com detalhe os dez idiomas presentes na descrição dos alojamentos. O processo de deteção da língua foi realizado automaticamente com recurso à biblioteca *langdetect*.[2] Após realizados os passos descritos, o *dataset* ficou reduzido a 11.085 resultados no idioma inglês, devido a ser o idioma mais presente nas descrições dos alojamentos.

As restantes alterações cingem-se principalmente a conversões de variáveis. Para a análise das questões de investigação, foi necessário proceder à conversão das variáveis, *review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value* e *host_is_superhost*. A variável *review_scores_rating*, está definida como uma escala de 0 - 100 pontos, por sua vez as restantes variáveis (*review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location* e *review_scores_value*) estão definidas com uma escala de 0 - 10 pontos. Desta forma, para manter os dados coerentes com o que os proprietários e hóspedes geralmente observam no Airbnb, foram convertidas estas variáveis (*review_scores_accuracy,*

[2] `https://pypi.org/project/langdetect/`, acedido a 14-08-2019

review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location e *review_scores_value*) para uma escala de 5 pontos. Para a conversão da variável *host_is_superhost,* procedeu-se à alteração do tipo *string* para o tipo *boolean*.

A Tabela 4.1 indica as variáveis que foram adicionadas ao *dataset.*

Variáveis	**Descrição**	**Tipo**
cleaned_data_listings	Coluna com o pré-processamento realizado à variável *description*	Texto
review_length	Tamanho da coluna *cleaned_data_listings.*	Numérica

Tabela 4.1: Variáveis adicionadas ao dataframe *listings* após o processamento de texto

4.3.2 Dados dos Comentários

Conforme supramencionado, o pré-processamento dos dados no *dataset* que se aplicou à variável *description*, foi igualmente aplicado à variável *comments* dos dados dos comentários. Para além dessas alterações, foi ainda executado a remoção de comentários automáticos, com a descrição *This is an automated posting.*

O processo de preparação dos dados para o *dataset* dos comentários é muito idêntico ao processamento dos dados efetuado ao *dataset* dos alojamentos. Como primeiro passo, foram observados os valores omissos nas colunas. A única coluna que continha valores omissos era a variável *comments*, com 205 comentários vazios. Optou-se desta forma pela estratégia de remoção das linhas e não a substituição por outro valor, que devido à grande quantidade de comentários não iria invalidar a análise com a redução de cerca de duzentos comentários. Para a substituição dos carateres *\n*, *\r* na variável *comments*, foi aplicada a mesma estratégia utilizada para o *dataset* dos alojamentos. No entanto, ao contrário do que foi efetuado para estes, não foram retiradas variáveis, constatando-se que todas eram necessárias para a análise.

Para a agilização de todo este processo verificou-se a necessidade de obter um *dataset* com informação precisa e devido ao grande número de registos em diferentes idiomas (conforme Anexo A, na Tabela A.6) foram eliminados os comentários em idioma diferente do inglês por estes não representarem a maioria nos comentários dos hóspedes. O processo utilizado para este caso, foi o mesmo utilizado para o *dataset* dos alojamentos. Desta forma, o *dataset* ficou reduzido a 466.296 comentários.

Antes de se proceder a determinadas tarefas de pré-processamento dos dados observou-se que na variável *comments*, existiam determinados comentários que tinham carácter automático, como por exemplo: *The host canceled this reservation 241 days before arrival. This is an automated posting.* Por forma a poder verificar o que significa este

tipo de comentários, foi realizada uma pesquisa no *site* da comunidade Airbnb[3], onde se constatou que estes comentários surgem por diversas razões, tais como, o cancelamento por mútuo acordo, o hóspede não cumprir com as regras do alojamento, como é o caso de não ser permitida a entrada de animais de estimação, ou até mesmo trazer mais pessoas do que o permitido. Como estas frases não acrescentam valor para a análise, foram retiradas do *dataset*. Deste modo, o *dataset* passa de 466.296 para 461.534 comentários.

Concluída a tarefa anterior, procedeu-se à tokenização das palavras, por forma a serem efetuadas as tarefas de pré-processamento dos dados seguintes, como é o caso da remoção de *stopwords,* conforme abordado anteriormente no pré-processamento dos alojamentos. Com a execução desta tarefa verificou-se que existiam palavras comuns em português (na sua maioria a palavra «de») e ao analisar os comentários verificou-se que os hóspedes quando escrevem em inglês, tendem a escrever certas palavras em português (por exemplo, «pastéis de belém», «castelo de são jorge», etc). Deste modo, foi também executada a remoção de *stopwords* não só do idioma português, mas também de todos os idiomas. Dentro do mesmo processo, foi ainda executado o código para a lematização dos comentários. Posteriormente, foram realizados os passos seguintes, remoção de nomes pessoais, remoção palavras com apenas um carácter, substituição de todas as palavras para palavras minúsculas, etc.

Quanto à variável *date,* apenas se procedeu à conversão da mesma para *datetime,* podendo proceder deste modo a determinadas análises, que serão utilizadas na secção seguinte.

Na Tabela 4.2 podem ver-se as variáveis que foram adicionadas ao *dataset*.

Variáveis	**Descrição**	**Tipo**
cleaned_data_reviews	Coluna com o pré-processamento realizado à variável *comments*	Texto
review_length	Tamanho da coluna *cleaned_data_reviews.*	Numérica

Tabela 4.2: Variáveis adicionadas ao dataframe *reviews* após o processamento de texto

4.3.3 Dados Calendário

Para este *dataset* apenas foi realizada a remoção de valores omissos, sendo que o número de entradas eliminadas por valores omissos foi marginal nas variáveis *price*, *minimum_nights* e *maximum_nights*.

[3] `https://community.withairbnb.com/t5/Help/Host-canceled-question/td-p/456184`, acedido a 15-08-2019

4.4 Análise Exploratória

Para iniciar a estruturação dos dados não estruturados foi efetuada uma análise exploratória.

4.4.1 Análise dos Alojamentos

Realizando uma pequena análise ao *dataset* dos alojamentos, foi possível constatar que os alojamentos Airbnb na região de Lisboa têm vindo cada vez mais a aumentar significativamente, como se pode observar na Figura 4.2.

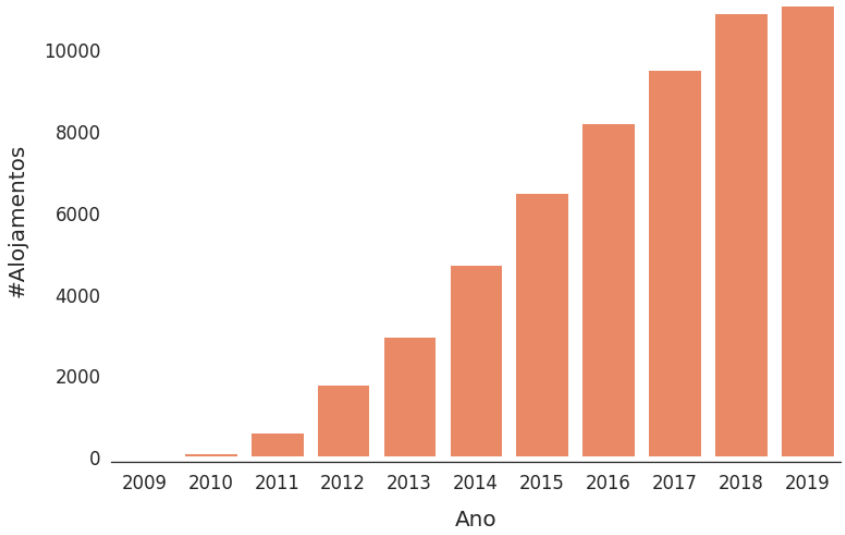


Figura 4.2: Alojamentos Airbnb na região de Lisboa ao longo dos anos

Como já referido em secções anteriores, o tipo de espaço no Airbnb refere-se ao apartamento/casa inteira, quarto privado e quarto partilhado. Na Figura 4.3 observa-se que o apartamento/casa inteira atinge os 81% (8.960 alojamentos), o quarto privado, 18% (2.013 alojamentos) e por fim o quarto partilhado, 1% (108 alojamentos). Destes espaços é importante verificar quais são os alojamentos (apartamento, barco, moradia, etc.) que fazem parte de cada tipo de espaço, existindo cerca de 30 alojamentos diferentes anunciados em Lisboa. Ainda nesta figura, é possível verificar quais são os tipos de alojamento que existem e em que quantidade, para determinado tipo de espaço. Surge assim, um *top* 8 dos tipos de alojamentos nas várias regiões de Lisboa: o apartamento é o alojamento com mais opção, com 7.504 apartamentos/casas inteiras, 1.389 quartos privados e 11 quartos partilhados. Por fim, o alojamento *bed & breakfast* com menos opção, apenas possui 5 apartamentos/-casas inteiras, 83 quartos privados e 17 quartos partilhados.

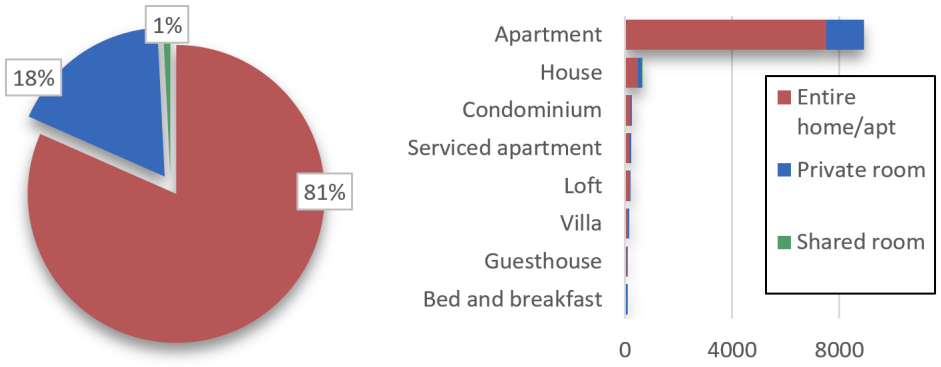


Figura 4.3: Distribuição dos tipos de espaço e tipos de alojamento

A Figura 4.4 mostra as palavras mais frequentes relativas à descrição do alojamento, mais propriamente das comodidades, através de uma *wordcloud* (utilizando a variável *amenities*) ou nuvem de palavras. As palavras tais como, «wifi», «kitchen» (cozinha e utensílios) e «cable TV» (tv por cabo), ocorrem muito frequentemente, sendo necessidades que o hóspede tendencialmente considera importantes num alojamento.

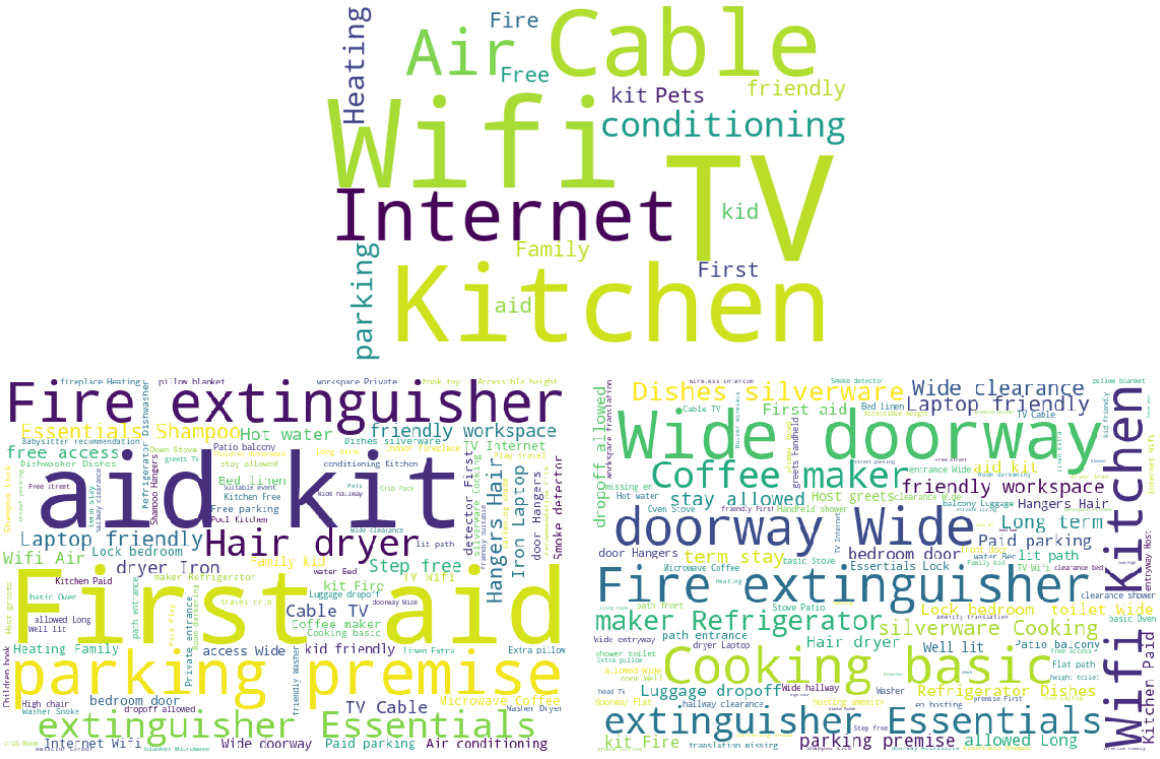


Figura 4.4: *WordClouds* das comodidades presentes em alojamentos (no topo) mais caras (esquerda) e mais baratas (direita)

Na mesma figura, é ainda possível observar as comodidades presentes em alojamentos mais caras e mais baratas. Nos alojamentos mais caros, é interessante verificar que palavras, tais como, «first aid» (primeiros-socorros), «aid kit» (*kit* primeiros-socorros) e

«fire extinguisher» (extintor de incêndio), ocorrem muito frequentemente, o que pode sugerir uma grande preocupação com a segurança do hóspede, por parte do proprietário. No que diz respeito aos alojamentos mais baratos, as palavras mais frequentes são «wifi», «kitchen» e «wide doorway» (portas largas).

Na Figura 4.5, estão reunidas todas as variáveis relativas ao *score* (*review_scores_location, review_scores_accuracy, review_scores_value, review_scores_communication, review_scores_checkin* e *review_scores_cleanliness*). Analisando cada variável, é possível perceber que a distribuição do *score* pelos vários alojamentos é elevada em todas as categorias. Estes valores, tendem a variar entre 4,5 a 5 pontos. Observando-se estas distribuições, constata-se que a pontuação igual ou inferior a 4 pode não ser uma boa classificação.

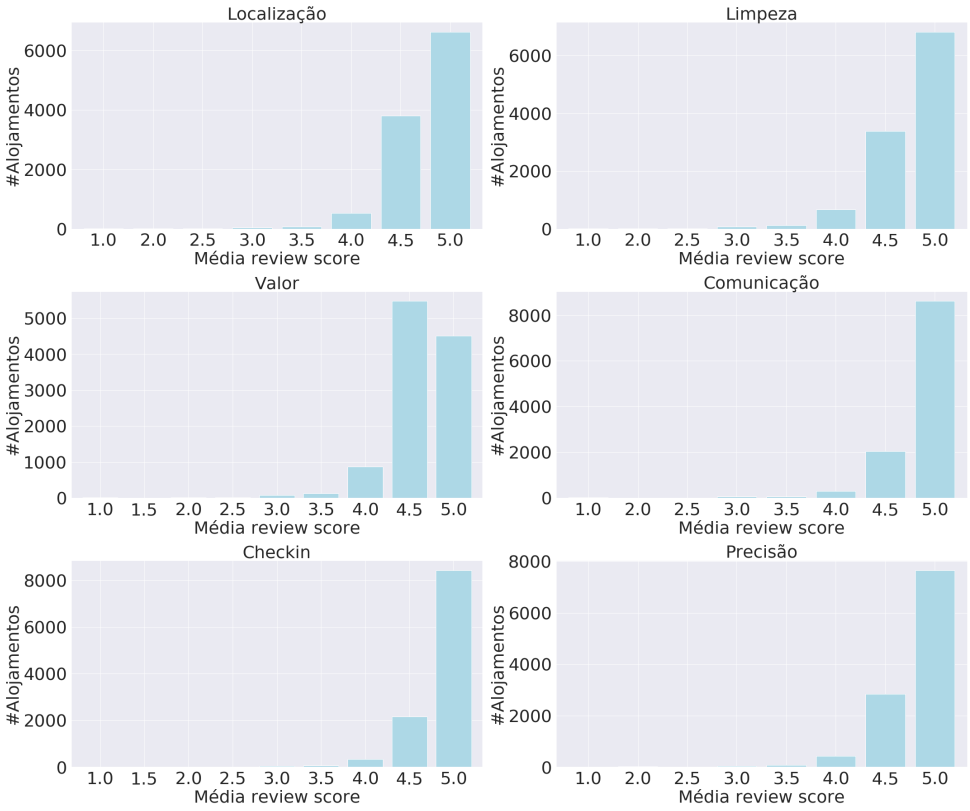


Figura 4.5: Frequência das variáveis *score*

Na Figura 4.6 e relativamente ao valor por noite, os apartamentos/casas inteiras apresentam um valor mais elevado que qualquer outro dos dois tipos de espaço. Verifica-se ainda uma pequena diferença entre o quarto privado e o quarto partilhado, sendo o quarto privado um pouco mais caro, devido a que neste o hóspede não tem de partilhar o quarto com mais ninguém, havendo maior privacidade. Ainda na análise da variável *price*, constatou-se que os alojamentos que pertencem a um *superhost* são mais caros do que os alojamentos dos *hosts* regulares. Tal acontece porque, tendencialmente os *superhosts* mostram outro tipo de cuidado para com o hóspede, seja ao nível de interação, como na

própria estadia.

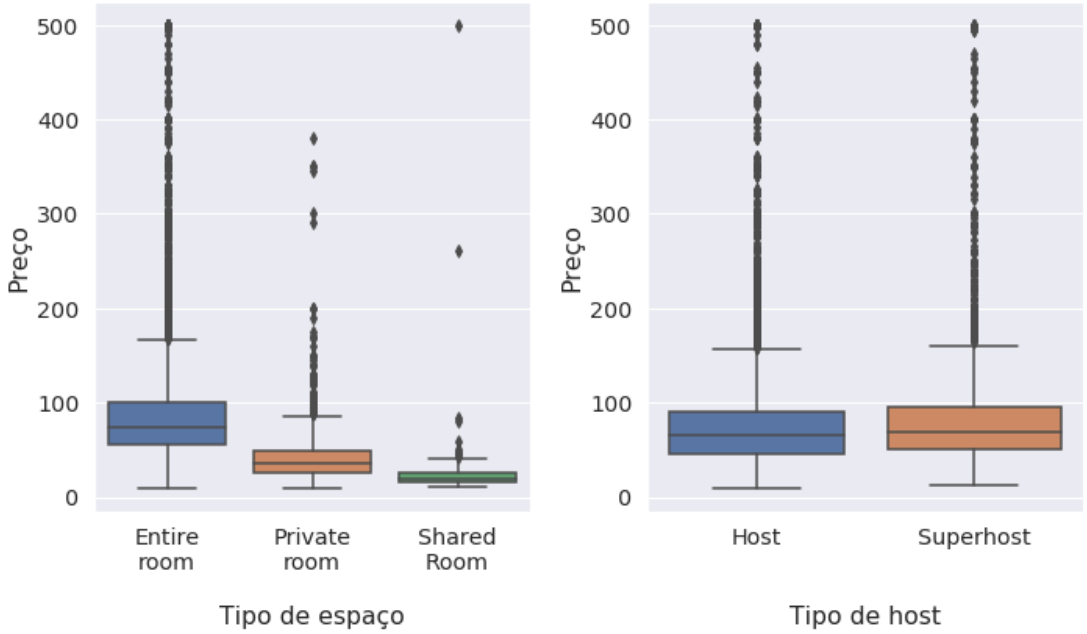


Figura 4.6: Diagramas de extremos e quartis, preço vs. tipos de espaço (esquerda), preço por noite vs. *host_is_superhost* (direita)

Conforme já referido anteriormente, os proprietários encontram-se divididos em dois grupos: os proprietários que são considerados como *superhosts* e os *hosts* regulares que não têm qualquer estatuto. Estes *superhosts*, como já mencionado, ganham este estatuto através de um programa implementado pelo próprio Airbnb, onde têm de cumprir certos requisitos para conseguirem este reconhecimento. Como se pode observar na Figura 4.7, para o caso de Lisboa, existem mais *hosts* regulares (6.308) do que *superhosts* (2.963), em que o mesmo se aplica aos seus alojamentos, com um total de 7.495 alojamentos geridos pelos *hosts* regulares e 3.590 alojamentos geridos pelos *superhosts*. Constatou-se nesta análise que existem muitas empresas a gerirem vários alojamentos, daí também haver mais alojamentos do que proprietários.

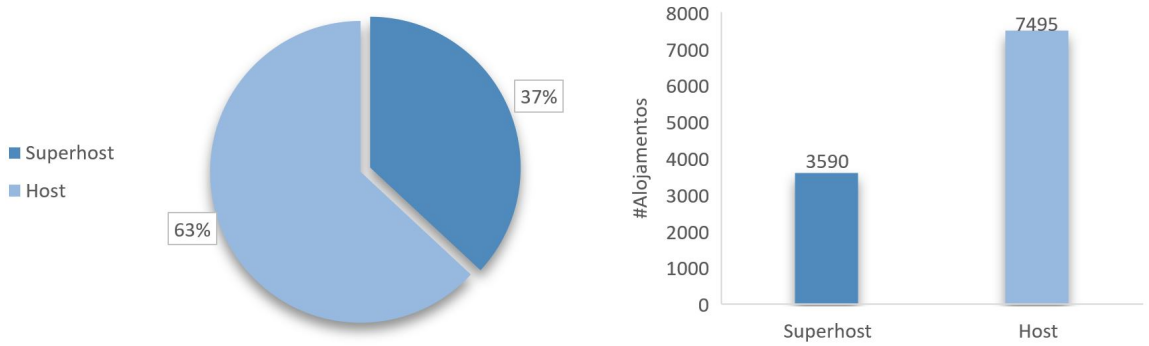


Figura 4.7: Análise dos *hosts* regulares e *superhosts* (esquerda) e número de alojamentos por proprietários (direita)

Analisando a Figura 4.8 de um *top* 10 das zonas de Lisboa que apresentam mais alojamentos, verifica-se que a zona de Santa Maria Maior é a que detém mais alojamentos tanto de *superhosts* como de *hosts* regulares, possivelmente, por ser uma freguesia que reúne as zonas do Centro Histórico de Lisboa.[4]

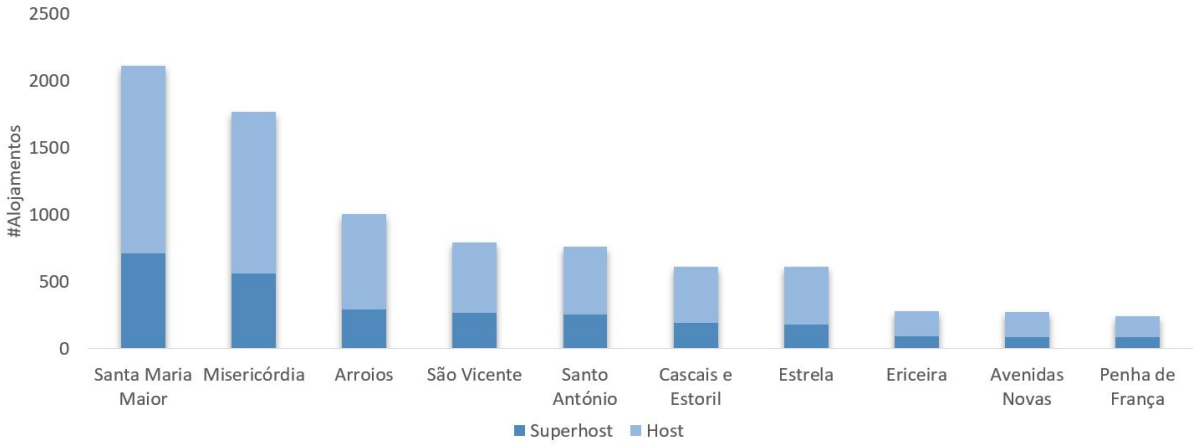


Figura 4.8: Distribuição dos alojamentos Airbnb dos superhosts/hosts por região

Quanto ao número de proprietários que foram ao longo dos anos aderindo a este conceito, pode-se constatar que no final de 2012 houve uma grande evolução nas adesões ao Airbnb, conforme Figura 4.9. Este crescimento pode dever-se a que novos proprietários vissem nesta plataforma uma nova oportunidade para mudar as suas vidas, uma vez que em 2012, Portugal atravessava uma crise económica. É também interessante observar que ao longo dos anos e, especificamente, no mês de agosto existe um aumento considerável da adesão ao Airbnb, por parte dos proprietários.

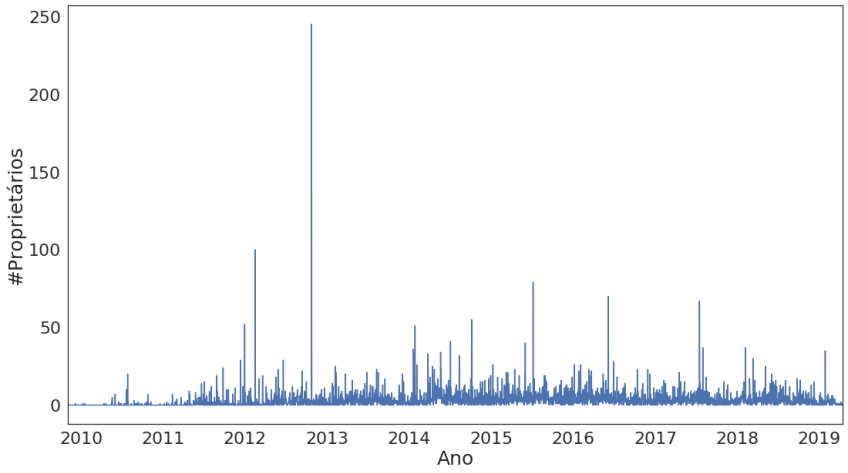


Figura 4.9: Distribuição do número de proprietários ao longo dos anos

Relativamente à disponibilidade dos alojamentos, na Figura 4.10 analisou-se o *dataset*

[4] `http://www.cm-lisboa.pt/municipio/juntas-de-freguesia/freguesia-de-santa-maria-maior`, acedido a 20-08-2019

calendário, onde é possível verificar a disponibilidade de cada alojamento. A análise realizada permitiu observar quais os meses do ano em que se observam mais reservas. Em Lisboa, os meses com mais reservas e logo com menos disponibilidade, são os meses de maio e junho, meses de primavera e verão. Os meses com mais disponibilidade são obviamente os meses de inverno.

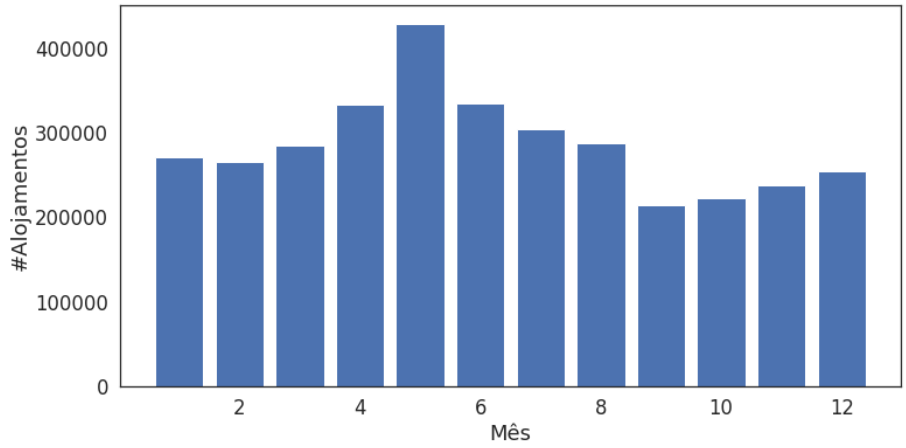


Figura 4.10: Número de alojamentos disponíveis por mês num ano

Todas as análises anteriores permitiram conhecer os dados dos alojamentos do Airbnb na região de Lisboa. De seguida, serão apresentadas as análises ao nível da extração de informação que se consegue obter através de técnicas de *text mining*. Deste modo, a variável escolhida para este passo é a variável *description* que consiste nas descrições dos anúncios que os proprietários colocam. Na Figura 4.11, é possível verificar a distribuição do número de palavras, onde o maior peso está entre as 150 e 200 palavras, em que até 200 palavras, existem 11.043 documentos. Relativamente a dados estatísticos, a média do número de palavras num documento é de 164, o máximo é de 226, e o mínimo é de 2 palavras.

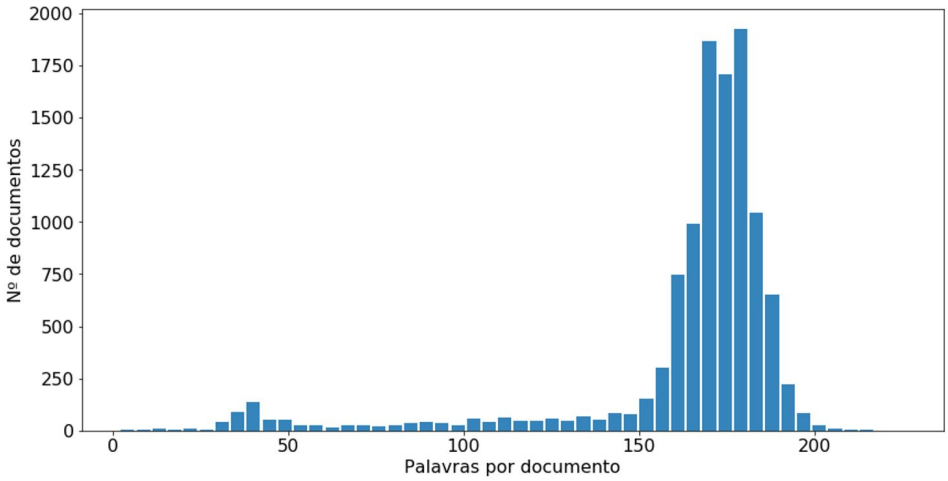


Figura 4.11: Distribuição do número de palavras da variável *description*

Após esta análise, é importante perceber quais são as palavras mais utilizadas pelos proprietários para descreverem os seus anúncios relativos às suas propriedades. Conforme a Tabela 4.3, as três palavras mais utilizadas são «apartment» (apartamento) (23.706), «lisbon» (lisboa) (17.390) e «room» (sala) (12.346).

Palavras	***Counter***
apartment (apartamento)	23.706
lisbon (lisboa)	17.390
room (sala)	12.346
bed (cama)	11.129
bedroom (quarto)	10.558
locate (localizar)	9.412
one (um)	9.072
kitchen (cozinha)	8.965
area (área)	8.432
bathroom (casa de banho)	7.449

Tabela 4.3: Frequência das palavras mais utilizadas pelos proprietários na descrição dos anúncios

4.4.2 Análise dos Comentários

Usando estratégias de extração de informação adequadas é possível identificar informação útil tanto para os proprietários de um determinado alojamento, como para um próximo hóspede. Para o proprietário, esta análise permite a possibilidade de uma rápida identificação de falhas ou necessidades que os hóspedes sentem e estão presentes neste tipo de alojamentos, não obstante, para os hóspedes ajuda na identificação de comentários antigos, que podem revelar experiências indesejavéis.

Com a enorme divulgação realizada para esta plataforma, as pessoas tendem a aderir mais a este conceito e a usufruir desta experiência, originando uma imensidade de comentários. Efetuando-se a análise do número de comentários ao longo dos anos, verificou-se que a partir de 2014 houve uma evolução mais acelerada dos comentários dos hóspedes, eventualmente porque até essa data os hóspedes não estavam tão suscetíveis a escreverem comentários ou, até mesmo porque, o número de utilizadores que usufruíam deste serviço ser menor. A análise da Figura 4.12, mostra essa evolução, o que por sua vez, leva a que surjam também comentários em diferentes idiomas. Com cerca de 40 idiomas presentes nos comentários dos hóspedes do Airbnb Lisboa, procedeu-se à observação dos cinco idiomas mais frequentes (*top* 5). O número um desta coleção com uma maior frequência de comentários (461.534), pertence ao idioma inglês. Em segundo lugar, neste *top* 5 surgiu o idioma francês (132.221), de seguida o português (47.722), depois o espanhol (46.980) e por fim, o alemão (30.791).

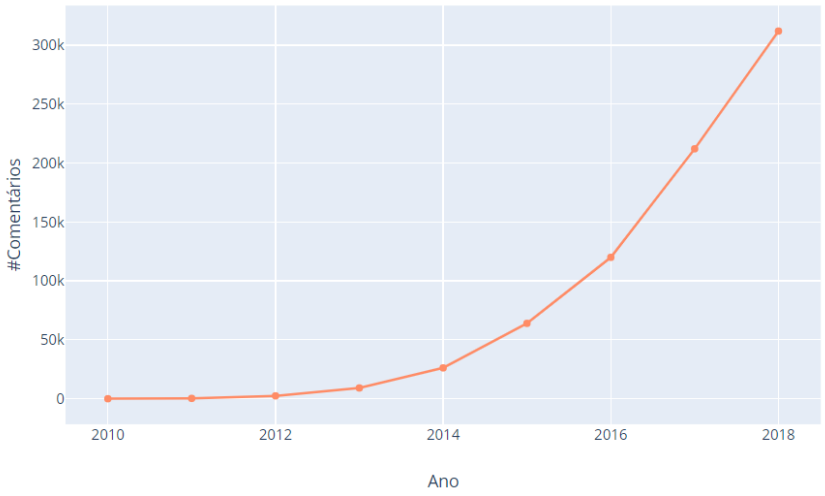


Figura 4.12: Novos comentários no Airbnb Lisboa desde 2010

Desta forma, apenas foi considerado o idioma inglês para todos os comentários a analisar, isto porque para além de ser o idioma mais utilizado nos comentários como se pôde constatar, é também o idioma selecionado para o *dataset* dos alojamentos. Na Figura 4.13 é possível observar as palavras mais representativas de cada idioma, utilizando a ferramenta *WordCloud*. Nesta análise, é possível detetar que para o idioma inglês as palavras que revelam uma maior representação são «stay» (ficar), «property» (propriedade), «apartment» (apartamento), «wonderful» (maravilhoso), «lisbon» (lisboa) e «clean» (limpeza), entre outras.

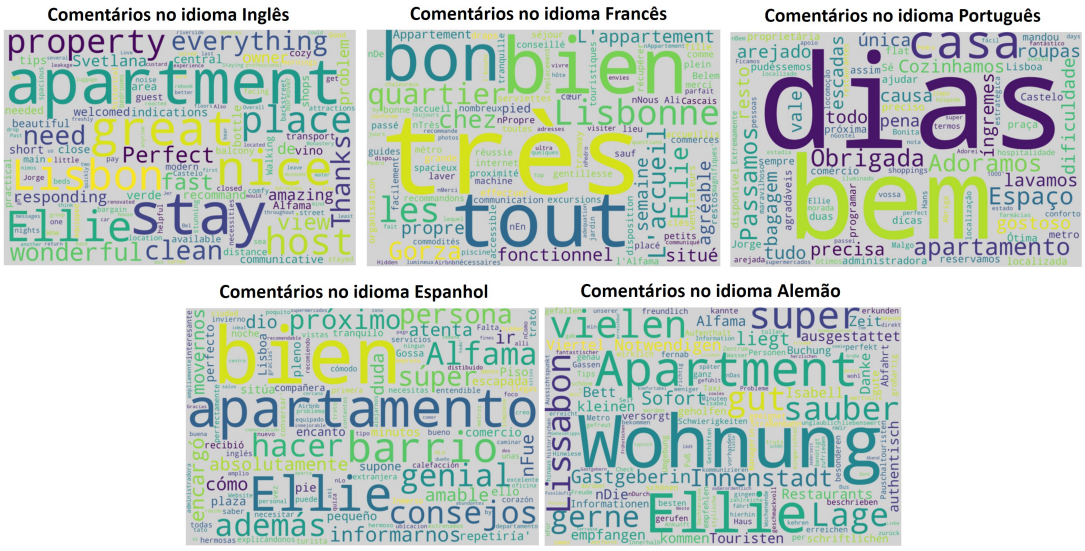


Figura 4.13: *WordCloud* do *Top* 5 de idiomas da variável *comments*

Quanto à distribuição do número de palavras, a Figura 4.14 mostra o número de palavras que em média um comentário apresenta, cerca de 10 a 50 palavras. Quanto ao número

mínimo e máximo de palavras, constatou-se que o mínimo de palavras num documento é de 1 e o máximo é de 1.109 palavras.

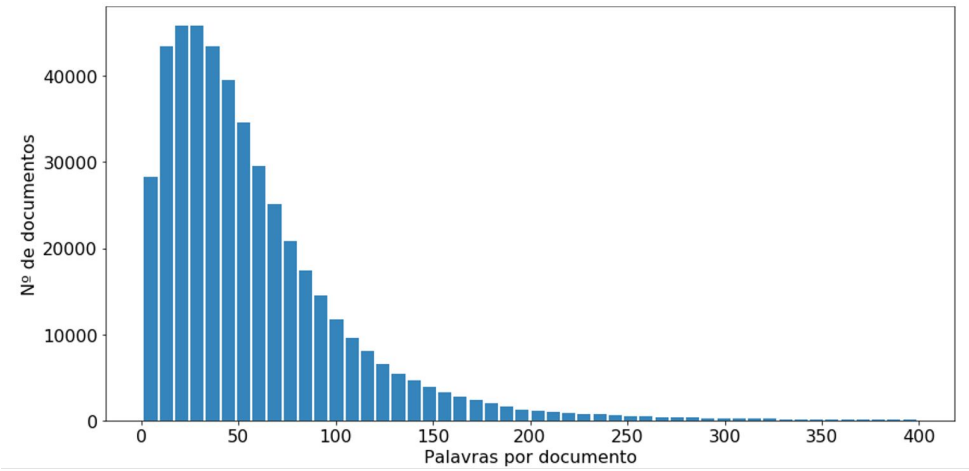


Figura 4.14: Distribuição do número de palavras dos comentários

Outro ponto importante do estudo é compreender quais as palavras mais utilizadas pelos hóspedes. Neste passo, optou-se por uma análise baseada na frequência das palavras do comentário, isto porque na fase de pré-processamento já tinha sido realizada a remoção de *stopwords*, pelo que não era relevante usar a técnica TF-IDF. Na Tabela 4.4, identificou-se o *top* 10 das palavras mais utilizadas nos comentários dos hóspedes e que são «apartment» (apartamento), «great» (ótimo), «stay» (ficar), «place» (lugar), entre outras. Constatou-se assim, que os hóspedes abordam mais as palavras relacionadas com os aspetos da estadia ao nível do alojamento.

Palavras	***Counter***
apartment (apartamento)	331.605
great (ótimo)	273.580
stay (ficar)	261.120
place (lugar)	227.164
us (nós)	192.100
location (localização)	179.449
lisbon (lisboa)	174.556
nice (agradável)	158.238
host (proprietário)	149.344
clean (limpeza)	142.994

Tabela 4.4: Frequência das palavras mais utilizadas pelos hóspedes

Dentro do mesmo tema, procedeu-se à identificação dos verbos e adjetivos mais utilizados. Desta forma, irá ajudar o proprietário na observação destas palavras, tendo conhecimento das palavras certas a utilizar na descrição do seu alojamento, levando a uma maior atenção por parte do hóspede e, possivelmente, a optar por esse alojamento. Através

da técnica *Part-Of-Speech Tagging (POS)* é atribuída uma *tag* ou etiqueta a cada palavra, de acordo com a sua definição e contexto, dependendo se se trata de um verbo, adjetivo, nome, etc. (Tussyadiah e Zach, 2017).

Para o caso dos verbos, as etiquetas[5] selecionadas são as seguintes:

1. Verbo VB (Forma base);
2. Verbo VBD (Pretérito perfeito ou passado);
3. Verbo VBG (Forma nominal - gerúndio);
4. Verbo VBN (Particípio passado);
5. Verbo VBZ (Presente na 3ª pessoa do singular).

De todas as etiquetas possíveis para a seleção dos verbos, que são um total de seis, optou-se apenas por cinco, devido a que a etiqueta em falta refere-se a verbos no presente, mas que não estejam na 3ª pessoa do singular. Assim, foi descartado esse tempo verbal porque os hóspedes tendem a referir-se neste caso ao proprietário sempre na 3ª pessoa do singular. Após esta seleção, é realizada a lematização de modo a obter a raiz da palavra em análise. No final destas fases realizou-se o desenvolvimento de uma *WordCloud*, como se pode constatar na Figura 4.15. Os verbos mais utilizados pelos hóspedes são «stay» (ficar), «locate» (localizar), «walk» (caminhar), «welcome» (acolher), «give» (dar), «find» (procurar), entre outros.

Figura 4.15: *WordCloud* dos verbos mais utilizados pelos hóspedes

Para além dos verbos, procedeu-se também à identificação dos adjetivos. O processo é idêntico ao anterior, à exceção das etiquetas selecionadas corresponderem a adjetivos, que são os seguintes:

[5] `https://pythonprogramming.net/part-of-speech-tagging-nltk-tutorial/`, acedido a 20-08-2019

1. JJ (Adjetivo);
2. JJR (Adjetivo, comparativo);
3. JJS (Adjetivo, superlativo).

A Figura 4.16 mostra os resultados obtidos.

Figura 4.16: *WordCloud* dos adjetivos mais utilizados pelos hóspedes

De acordo com todas as palavras já extraídas, é possível perceber que tendencialmente os hóspedes transmitem positivismo nos seus comentários e, segundo, Fan e Khademi (2014) verificaram que os adjetivos eram mais vantajosos, isto porque são o tipo de palavras mais frequentemente utilizadas para descrever positividade ou negatividade. Apesar da extração dos adjetivos dar uma ideia de positividade ou negatividade, uma análise mais sistemática baseada numa ferramenta automática será apresentada no capítulo 5.

4.4.3 Análise da extração textual - Alojamentos e comentários

Por fim, contabilizou-se o número de comentários, a média do *score*, o número total de palavras e, por fim, a média de palavras por comentário (Tabela 4.5). Estas análises foram organizadas por tipo de espaço. Por sua vez, realizou-se o *merge* de dois *datasets* (alojamentos e comentários). Deste modo, verifica-se que o tipo de espaço com o maior número de comentários são os apartamentos/casas inteiras (369.418) e com menos comentários, o quarto partilhado (2.402), no entanto, este para a média *review score* consegue ter uma melhor avaliação relativamente ao quarto privado. Este pequeno aumento surge apenas no seguimento de que os comentários do quarto partilhado são mais reduzidos e positivos, o que não significa que seja o melhor.

Na Tabela 4.5, também se pode constatar que os tipos de espaço em que os hóspedes escrevem os seus comentários com maior número de palavras em média são o apartamen-

to/casa inteira (58,10), de seguida o quarto privado (50,54) e, por fim, o quarto partilhado (42,21).

Tipos de espaço	**Nº de Comentários**	**Média *Review Score***	**Nº Total de Palavras**	**Média de Palavras por Comentário**
Apartamento/casa inteira	369.418	4,70	21.463.107	58,10
Quarto privado	73.102	4,64	3.694.438	50,54
Quarto partilhado	2.402	4,67	101.378	42,21
Total	444.922	4,67	25.258.923	50,28

Tabela 4.5: Número total de palavras por tipo de espaço

4.5 Modelação

Após a estruturação e pré-processamento dos dados tanto do *dataset* dos alojamentos, como dos comentários, nesta secção apresentar-se-á o tratamento do conjunto dos dados provenientes destes dois *datasets*.

Esta secção é assim dividida em quatro partes, abordando as estratégias e métodos utilizados, que permitem responder a cada questão de investigação.

4.5.1 Questão de Investigação nº 1: Através dos comentários dos hóspedes, quais são os aspetos mais relevantes da sua experiência? Tais como por exemplo a localização, a comunicação, a limpeza, etc.

O intuito desta questão é conseguir obter o tópico que os hóspedes mais abordam nos seus comentários. Este tipo de análise irá ajudar os proprietários a focalizar os seus alojamentos para aquilo que é mais importante para os hóspedes.

Para esta questão de investigação, não são necessárias todas as variáveis do *dataset* para a análise. Deste modo, procedeu-se apenas à seleção das variáveis *listing_id*, *review_id*, *comments* e *cleaned_data_reviews*. De seguida, é explicada a estratégia utilizada que se traduz nos passos seguintes:

1. Usando a variável *cleaned_data_reviews,* procedeu-se à modelação de tópicos, através da técnica *topic modelling* (modelação por tópicos), que consiste numa classe de algoritmos que organiza automaticamente grandes quantidades de texto, descobrindo tópicos ocultos ou temas que são discutidos num conjunto de documentos (Blei e Lafferty, 2009). Desenvolveu-se esta questão em dois tipos de abordagem de análise de tópicos: o *Latent Dirichlet Allocation* (LDA), em que a estrutura de tópicos ocultos

é inferida a partir de textos originais usando uma abordagem probabilística (Blei e Lafferty, 2009); e, o *Latent Semantic Analysis* (LSA), que se trata de uma técnica estatística que, dada uma matriz de documentos e termos, gera uma série de vetores que capturam a variância dentro da matriz original (Blake, 2011).

(a) Abordagem LDA: É criado um dicionário a partir do *corpus*, onde é atribuído um *id* a cada palavra nos documentos e por sua vez, são filtradas as palavras incomuns, de acordo com os seguintes parâmetros [6]:

- Filtrar as palavras que aparecem menos de 20 vezes;
- Filtrar as palavras que aparecem em mais de 50% de todos os documentos.

De seguida, é criado um *corpus* em que cada documento é representado pelos *id's* das palavras que o compõem e respetivas frequências. Primeiramente, foi criado um modelo com o número de tópicos com a maior coerência pela métrica *C_v* (referido no Capítulo 2) e 20 passagens pelo *corpus*.

2. Abordagem LSA: É criada a matriz *document-by-term,* em que cada coluna corresponde a cada termo que aparece em qualquer parte do *corpus* e uma linha para cada documento (cada comentário do hóspede é um "documento"). A estratégia utilizada para preencher esta matriz é através do cálculo da frequência dos termos de cada documento. Esta estratégia foi aplicada por forma a que as duas abordagens (LDA e LSA) pudessem utilizar igualmente o mesmo método, a frequência, de modo a que possa ser feita, caso necessário, uma comparação entre elas;

3. Realizou-se o cálculo da perplexidade para o modelo LDA e da coerência pela métrica *C_v* para as duas abordagens (referido no Capítulo 2);

4. Com a criação do modelo no primeiro passo, verificou-se a presença de palavras relacionadas a determinadas localizações de Lisboa, como por exemplo, Sintra, Cascais, Ericeira, etc. Nesse sentido, a remoção destas palavras não seria a melhor estratégia a adotar. Assim, antes de efetuar uma nova análise dos tópicos procedeu-se à substituição destas palavras pela palavra *location*, por forma a agrupar e a uniformizar os dados, reduzindo possíveis ambiguidades, sendo que, antes de uma nova execução, foram também eliminadas palavras consideradas como *stopwords*, que não tinham sido removidas no pré-processamento inicial, como é o caso de «many», «also», etc.;

5. Tendo em conta o passo anterior e consoante a melhor abordagem escolhida pelo seu valor da coerência é posteriormente verificado qual o melhor número de tópicos a utilizar para o estudo. Para o cálculo do melhor número de tópicos, é utilizada novamente a coerência, mas em função do número de tópicos, que permite medir o grau de semelhança semântica entre palavras com um alto peso no tópico. Estas

[6] `https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24`, acedido a 20-08-2019

medidas ajudam a distinguir entre tópicos que são semanticamente interpretáveis e tópicos que são artefactos de inferência estatística (Stevens et al., 2012);

6. Com a seleção do número de tópicos ótimo, analisaram-se os termos associados a cada tópico, atribuindo um nome ou categoria a cada um;

7. Por forma a encontrar os tópicos mais relevantes, determinaram-se os tópicos que estão associados a cada comentário. Com a distribuição dos tópicos pela coleção dos documentos, foi possível chegar ao tópico com o maior peso em cada comentário, e por sua vez com a soma destes tópicos chegar ao tópico com mais documentos, como é possível verificar o esquema na Figura 4.17

#Docs	Tópicos
Doc0	[(2, 0.048), (5, 0.213), (8, 0.121)]
Doc1	[(3, 0.167), (6, 0.034), (8, 0.056)]
Doc2	[(5, 0.255), (10, 0.155), (7, 0.129)]

#Docs	Tópico maior peso
Doc0	5
Doc1	3
Doc2	5

Tópicos	#Docs
5	2
3	1

Figura 4.17: Exemplo da identificação do tópico mais relevante

Finalizada a estratégia utilizada para a primeira parte da questão, é necessário responder aos restantes pontos:

4.5.1.1 Quais deles os hóspedes consideram mais positivos e mais negativos?

A resposta à questão anterior foi elaborada a partir da mesma estratégia considerada na questão principal. O desenvolvimento deste processo consistiu na elaboração de uma lista de palavras dos tópicos que definem as categorias (criadas no passo anterior). Para este passo foi necessário observar os termos que surgiram de cada tópico por forma a adicionar nesta lista de palavras, como por exemplo, a categoria *location* (localização), as palavras descritas são, «central» (central), «close» (perto), «nearby» (próximo), entre outras. Adicionalmente, para a determinação da polaridade dos sentimentos de cada categoria realizou-se o processo mostrado na Figura 4.18. Com a lista de palavras mencionada anteriormente, verifica-se se a palavra está contida no comentário, caso esteja, é analisado o contexto onde a palavra está inserida nos comentários para posteriormente ser calculado o *score* de polaridade do sentimento com a variável *score* agregado ou *compound*. Após este passo é devolvido o resultado da média do sentimento de determinada categoria, variância e desvio-padrão. Com essa média, é possível diferenciar quais as categorias com sentimentos positivos, negativos e neutros. A categoria é considerada positiva se o *score* do *compound* ou *score* agregado for superior ou igual a 0,5, neutra se o *score* estiver compreendido entre -0,5 e 0,5, e negativa se o *score* for inferior ou igual a -0,5.

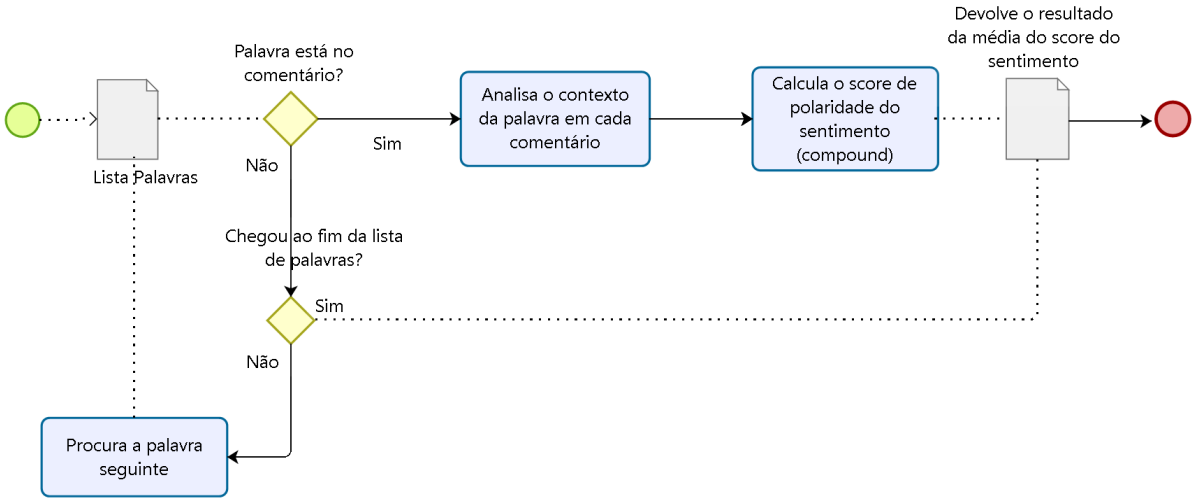


Figura 4.18: Esquema do processo da identificação do sentimento por categoria

4.5.2 Questão de Investigação nº 2: Será que os comentários negativos dos hóspedes chocam com as informações dos proprietários sobre os alojamentos?

Um dos objetivos do trabalho é compreender se os comentários negativos dos hóspedes chocam com as descrições colocadas pelos proprietários sobre os alojamentos. Os proprietários devem estar atentos a estes aspetos, pois um alojamento que coincida com as perspetivas do hóspede vai ao encontro das suas necessidades, sendo que um hóspede satisfeito com a sua experiência tende a refletir positivismo nos comentários.

As variáveis selecionadas para responder a esta questão de investigação são *listing_id*, *review_id*, *comments*, *room_type* e *review_scores_accuracy*. Neste caso, a estratégia adotada incide essencialmente nas variáveis *comments* e *review_scores_accuracy*. A única forma de perceber o que é que os hóspedes acham de determinada experiência, é pelos comentários e pelo *score accuracy*. Como já referido anteriormente, esta variável permite relacionar a descrição do próprio alojamento e a experiência do hóspede.

Numa escala de um a cinco estrelas, verificou-se que os alojamentos classificados com quatro estrelas são cerca de 70.000 registos e três estrelas com cerca de 50. Como já tinha sido referido anteriormente, os hóspedes tendem a dar pontuações altas independentemente dos seus comentários, deste modo, é incluído o *score* até três estrelas, por forma a verificar com maior rigor os exemplos encontrados para esta questão, bem como filtrados os comentários negativos.

Após este processo, é realizada a extração de *chunks* sintáticos. O método utilizado é o *noun phrase chunking* ou *NP-chunking*, que tem como objetivo identificar os sintagmas nominais. Este método possibilita a observação de uma lista de uma ou mais palavras coerentes, por forma a encontrar palavras-chave que possam ajudar na análise em questão,

neste caso encontrar palavras que salientem se os comentários negativos dos hóspedes chocam com as informações sobre os alojamentos. Na prática, prossegue-se com a identificação das partes constituintes das frases através de etiquetas *part-of-speech* (POS) relacionando estas partes com determinadas regras, que permitem dividir em sintagmas cada frase, aplicando as regras impostas à *priori* (Bird et al., 2009). A regra estabelecida para esta questão, é que o *chunk* é formado por um Determinante (DT) opcional, seguido por 0 ou mais adjetivos (JJ) e um nome (NN). Sendo que para este estudo foram utilizadas as seguintes etiquetas[7], para além das já mencionadas: WP (pronome pessoal), VBP (verbo, 3ª pessoa do singular), RB (advérbio), VBN (verbo, particípio passado), IN (preposição), NNP (nome próprio singular), NNPS (nome próprio plural), NNS (nome plural), JJR (adjetivo comparativo), JJS (adjetivo superlativo) e CC (conjunção coordenativa). Após este passo, são verificadas as ocorrências de cada um dos *chunks*, e conforme os *chunks* com ocorrências maiores, são mostrados exemplos de alojamentos Airbnb, com os comentários negativos vs. descrição do proprietário que contêm os referidos *chunks*.

4.5.2.1 Consoante o tipo de espaço, o tipo de constatações é diferente? Há evidências claras que isto muda consoante o tipo de espaço?

Para o desenvolvimento desta questão, foi utilizado o mesmo método elaborado na questão principal, no entanto com a particularidade de selecionar uma nova variável, *room_type*, com a análise por cada tipo de espaço (apartamento/casa inteira, quarto privado e quarto partilhado).

4.5.3 Questão de Investigação nº 3: O score associado aos indicadores espelha o que é descrito nos comentários?

O *score* é um meio de classificação que os hóspedes têm à sua disposição alinhado com os comentários para poderem partilhar a sua experiência. No entanto, constatou-se que estas duas vertentes não estão assim tão alinhadas, ou seja, os *scores* apresentaram valores altos para o que é descrito nos comentários, tanto de forma negativa como positiva. Esta questão aborda um método distinto do que foi estabelecido nas questões anteriores, tendo em conta que neste caso, é necessário predizer o *score* (*review_scores_rating*) de acordo com os comentários (*comments*). Este modelo passa por observar a variável resposta (*review_scores_rating*) efetuando a diferença entre o *score* atual e o *score* estimado através dos comentários. Com toda a preparação dos dados já efetuada, prosseguiu-se com o método de extração de *features* baseado no TF-IDF, semelhante à abordagem de *Bag of Words.* Após a extração de *features* utilizando o TF-IDF, dividiu-se a base em duas amostras: dados de treino e dados de teste, seguindo a proporção de 80-20. São também definidos os respetivos *alphas* e as taxas de aprendizagem (1, 0,1, 0,01, 0,001, 0,0001, 0,00001, 0,000001),

[7] `https://pythonprogramming.net/part-of-speech-tagging-nltk-tutorial/`, acedido a 29-08-2019

para a utilização da validação cruzada da amostra de treino, selecionando assim um dos três algoritmos diferentes, o *Stochastic Gradient Descent Regressor* (SGDR)[8] da biblioteca *scikit learn's, o XGBoostRegressor* [9] da biblioteca *xgboost* e o *CatBoostRegressor*[10] da biblioteca *catboost.* Posteriormente, com base nesses *alphas,* para o algoritmo SGDR, e taxas de aprendizagem, para os restantes algoritmos, calculam-se desta forma diferentes valores para os erros de validação e treino (*Mean squared error* (MSE)) (Erik e Poul, 1999). Através do menor erro de validação será selecionado o modelo e os seus parâmetros.

Por fim, é extraído um gráfico para análise que permite observar os *scores* previstos através dos comentários pelo algoritmo selecionado e os *scores* reais indicados pelos hóspedes.

4.5.4 Questão de Investigação nº 4: Relativamente aos alojamentos na cidade de Lisboa, os hóspedes escolhem tendencialmente os alojamentos cujos proprietários detêm o estatuto de *superhost*?

Conforme já mencionado, o Airbnb categoriza os proprietários em dois tipos, *hosts* regulares e *superhosts*. Os *superhosts* são classificados de forma distinta dos *hosts* regulares, através de determinados parâmetros, conforme já verificado anteriormente.

Para responder a esta questão de investigação, optou-se pela análise das variáveis *host_is_superhost* do *dataset* dos alojamentos e *available* do *dataset* calendário. O procedimento utilizado para esta resposta baseou-se na junção destes dois *datasets*, por forma a filtrar a informação dos alojamentos que pertencem aos *superhosts* e aos *hosts* regulares, tendo em conta o estado reservado do alojamento. Desta forma, é possível verificar tendencialmente a preferência dos hóspedes tanto pelos alojamentos dos *superhosts* como dos *hosts* regulares.

4.5.4.1 Existe variação pelo tipo de espaço?

A análise realizada para esta questão é idêntica à análise anterior, a diferença é que é adicionada uma nova variável, *room_type*, para poder identificar qual o tipo de espaço (apartamento/casa inteira, quarto privado e quarto partilhado) que o hóspede mais opta, tendo em conta os *hosts* regulares e os s*uperhosts*.

[8] SGD: o gradiente da perda é estimado para cada amostra em determinado momento, e o modelo é atualizado ao longo do processo com uma taxa de aprendizagem decrescente. Fonte: `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html`, acedido a 01-09-2019

[9] XGBoost: baseado numa árvore de decisão, é um algoritmo de *Machine Learning* que utiliza uma *framework* de *gradient bosting,* muito usado para previsão. Fonte: `https://xgboost.readthedocs.io/en/latest/`, acedido a 01-09-2019

[10] *CatBoost:* algoritmo idêntico ao *XGBoost,* no entanto ao contrário deste consegue lidar com caraterísticas categóricas. Fonte: `https://towardsdatascience.com/catboost-vs-light-gbm-vs-xgboost-5f93620723db`, acedido a 01-09-2019

4.5.4.2 Em comparação com os *hosts* regulares, quais as categorias mais abordadas nos comentários?

A resposta a esta questão está conforme supramencionado na análise da primeira questão de investigação, com a realização da modelação por tópicos. Embora o intuito desta seja consoante o *host* ou *superhost* perceber qual a categoria mais abordada nos comentários.

5 Análise dos Resultados

Neste capítulo são apresentados os resultados que permitem responder às questões de investigação apresentadas na Secção 1.3, bem como os resultados da análise de sentimentos para os comentários dos hóspedes, com o objetivo de ajudar os proprietários na compreensão e identificação dos comentários positivos e negativos, por forma a alcançarem o bom *feedback*.

5.1 Análise de Sentimentos - Comentários hóspedes

As palavras mais utilizadas pelos hóspedes ajudam na compreensão dos próprios proprietários de quais as palavras a utilizar nas informações dos alojamentos, por forma a melhorarem o seu estatuto e a classificação do seu alojamento. Sabendo que, estas palavras podem ter uma conotação positiva ou negativa, poderá ser aplicada a análise de sentimentos, para identificar os sentimentos associados. Nesta fase, optou-se por aplicar o módulo *vaderSentiment*, que integra a biblioteca NLTK, dado o seu bom comportamento em textos das redes sociais (Hutto e Gilbert, 2014). Para calcular os resultados de sentimentos (positivo, negativo, neutro e *compound*), utilizou-se a função *SentimentIntensityAnalyzer().* Após este passo, foram adicionados os resultados dos cálculos dos sentimentos ao *dataset*, onde cada variável pertence a uma métrica diferente, como se pode constatar nos exemplos apresentados na Tabela 5.2. Ao longo da análise, verificou-se que os hóspedes, em alguns casos, tendem a escrever comentários dando ênfase à pontuação e *emoticons,* o que igualmente influencia o cálculo do sentimento.

Pode verificar-se, através da tabela, que o *vader* classificou o sentimento dos comentários corretamente, como sendo positivos e negativos. No primeiro e segundo comentário, o *vader* identificou como os hóspedes se referem ao apartamento ser confortável e também ao proprietário de ser comunicativo. No último comentário é indicado que o apartamento não tem qualidade, que precisa de ser remodelado e que o valor do apartamento é excessivamente caro. Este passo ajuda os proprietários na identificação de falhas ou de necessidades dos seus alojamentos, sendo que as análises dos comentários negativos permitem melhorar certos aspetos e evitar que comentários com este tipo de conotação sejam redigidos por novos hóspedes, possibilitando *ratings* mais elevados.

comments	***neg***	***neu***	***pos***	***compound***
Nice and comfortable apartement in the city center with a very convenient location. Barbara is a great host, she was attentive to us and gave a lot of advises for the visits in Lisbon.	0,000	0,808	0,192	0,7845
Ellie is very nice and communicative. The apartment is great and clean in the beautiful Alfama area. We had all what we needed there :) Thanks!	0,000	0,500	0,500	0,9637
Antonio is a great host, however his apartment is very shoddy and needs redecorating. There isn't a bed but a futon bed which is very uncomfortable and on both nights we all came out in rashes. Given the negatives, it therefore doesn't offer value for money and is overly expensive for what it is.	0,112	0,843	0,045	-0,5918

Tabela 5.2: Exemplos extraídos do *dataset* com as respetivas métricas, tais como calculadas pelo *vaderSentiment*

Em termos de representação para todos os comentários, foi desenvolvido um gráfico para cada métrica, como observado na Figura 5.1, sendo possível verificar que os comentários dos hóspedes tendem a ser mais positivos do que negativos.

Para o caso do *score* do sentimento negativo, o *score* é de 0 na grande maioria dos comentários do *dataset*, ou seja, não tem qualquer conotação negativa, havendo umas oscilações mais negativas embora que muito reduzidas do *score* em cerca de 50.000 comentários. Isto significa, que tendencialmente os hóspedes ao escreverem os seus comentários com aspetos negativos tendem a enfatizar estes com aspetos positivos, como demonstrado no último exemplo da Tabela 5.2.

No *score* do sentimento neutro, são cerca de 60.000 os comentários que apresentam um *score* de 0,70, embora os restantes comentários apresentam uma variação entre 0,2 e 1. Este *score* é classificado de acordo com o tipo de comentários, cujo contexto é desconhecido, e ao verificar-se novamente o último exemplo da Tabela 5.2, pode-se constatar de que como se trata de um comentário com aspetos positivos e negativos, foi calculado com um *score* neutro bastante elevado (0,843).

Por fim, para o sentimento positivo, cerca de 50.000 comentários são considerados positivos, com um *score* de cerca 0,3 e os remanescentes entre 0 e 0,8.

O sentimento *compound* ou sentimento agregado (referido no Capítulo 2), é, para a maioria dos comentários, cerca de 250.000, considerados bastante positivos. O valor desta variável define se um comentário é positivo, negativo ou neutro, calculando a soma de todas as classificações do léxico, sendo estas normalizadas com os valores entre -1 e +1.

A Figura 5.2 apresenta uma distribuição dos sentimentos pelos comentários, em que os comentários positivos surgem como os mais frequentes e os negativos com menos frequentes.

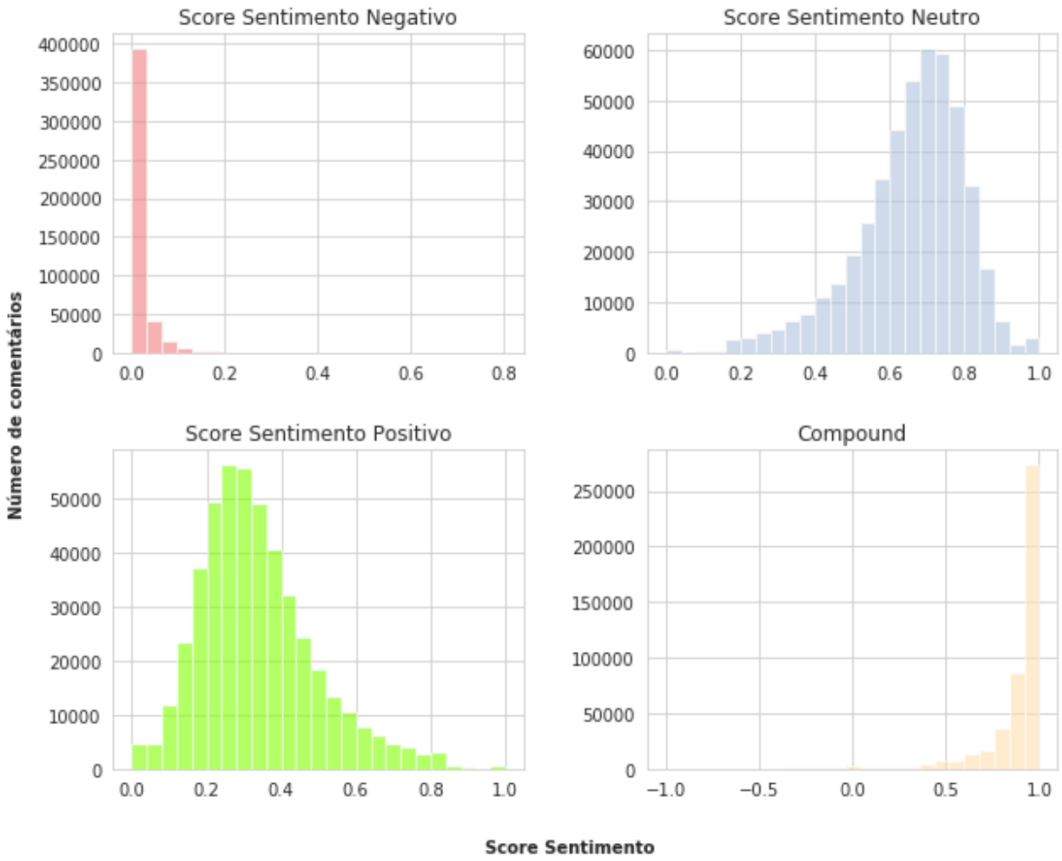


Figura 5.1: Análise de sentimentos dos comentários Airbnb para Lisboa

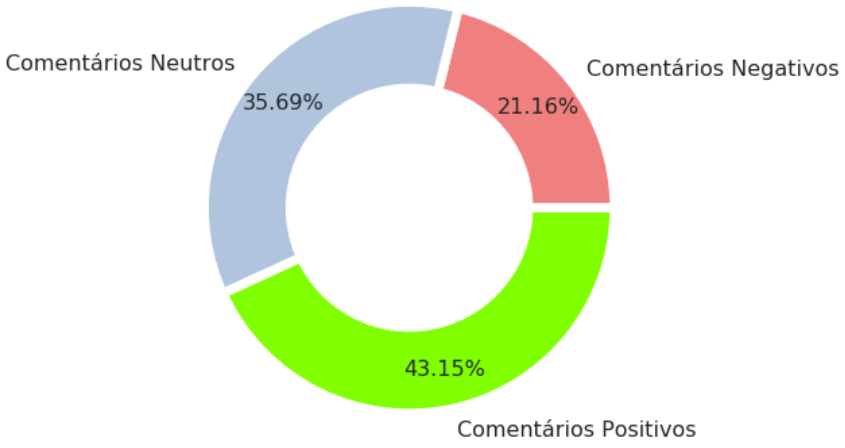


Figura 5.2: Gráfico circular da análise de sentimentos

A Figura 5.3 mostra que ao longo desta análise foi possível identificar que em média os hóspedes tendem a ser mais longos nos comentários positivos do que nos negativos. Tendencialmente, a razão para tal é que uma experiência satisfatória neste tipo de alojamentos leva a que os hóspedes se mostrem mais disponíveis e demonstrem o seu agrado ao escreverem os comentários de uma forma mais detalhada. No entanto, relativamente aos comentários negativos, os hóspedes estão insatisfeitos com a experiência e apenas apontam rapidamente as falhas, restringindo os detalhes.

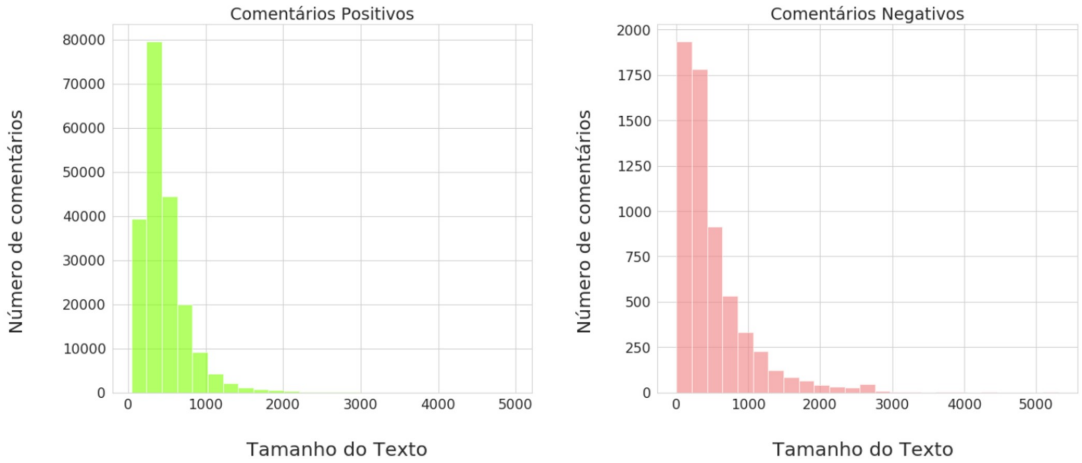


Figura 5.3: Distribuição para o comprimento dos comentários

Quanto a palavras positivas, na Figura 5.4 é apresentado um gráfico com as palavras mais frequentes nos comentários positivos, *top* 20. As palavras «apartment» (apartamento), «great» (ótimo), «stay» (ficar), «lisbon» (lisboa), «place» (lugar), «location» (location) e «clean» (limpeza), encontram-se entre cerca dos 80.000 (17,3%) e 200.000 (43,3%) comentários. É interessante verificar que estas palavras se referem a boas estadias em alojamentos limpos e bem localizados.

No caso das palavras negativas, é possível observar na Figura 5.4 que as palavras mais frequentes do *top* 20 são, «apartment» (apartamento), «place» (lugar), «stay» (ficar), «host» (proprietário), «location» (localização) e «room» (sala). Comparativamente com as palavras positivas, as palavras diferentes que aqui se destacaram foram «host» e «room». Isto sucede-se, porque tendencialmente os hóspedes escrevem comentários negativos relativos ao quarto não ter limpeza ou a cama do quarto ser desconfortável, e relativamente ao proprietário, o hóspede mencionar que este não é uma pessoa prestável ou a comunicação ser inexistente entre eles.

Com a identificação de todos os pontos anteriores, permite assim ao proprietário identificar mais rapidamente as falhas e necessidades dos hóspedes descritas nos comentários, por forma a que estes tenham melhores níveis de serviço para o(s) seu(s) alojamento(s).

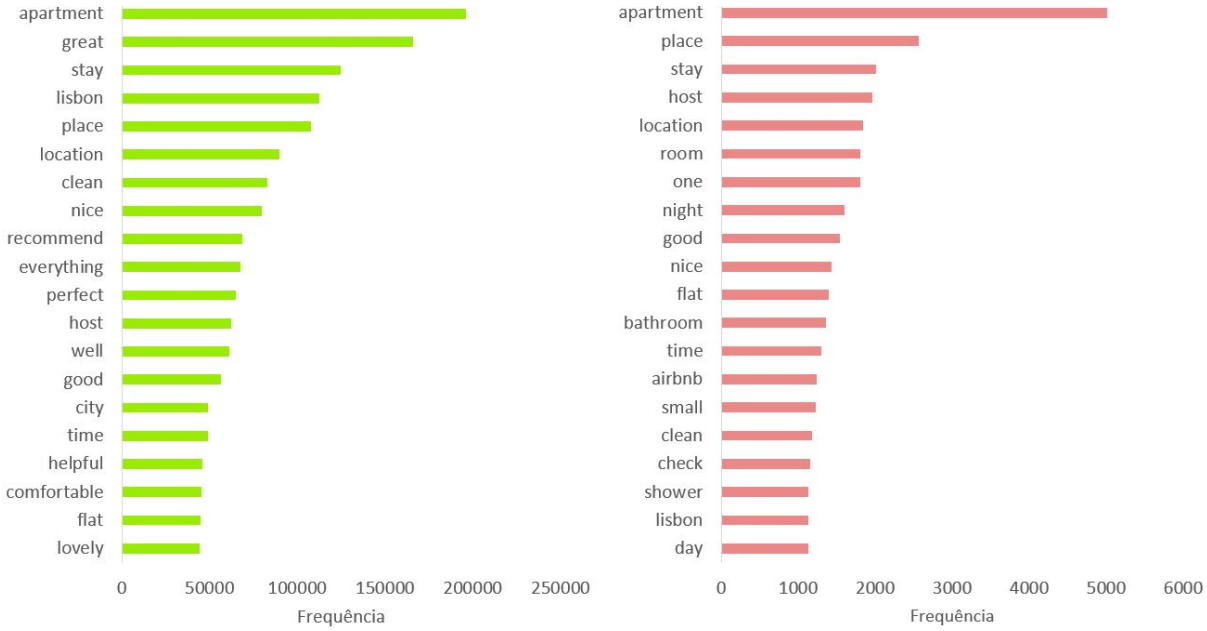


Figura 5.4: *Top 20* das palavras mais frequentes nos comentários positivos (esquerda) e nos comentários negativos (direita)

5.2 Questão de Investigação nº 1 - Resultados

Através dos comentários dos hóspedes, quais são os aspetos mais relevantes da sua experiência? Tais como por exemplo a localização, a comunicação, a limpeza etc.

Com vista a dar resposta a esta questão foi aplicada a modelação por tópicos. De acordo com os passos efetuados da fase da modelação (Secção 4.5), foram aplicadas as duas abordagens: LDA e LSA.

5.2.1 Modelação por tópicos - LDA e LSA

O primeiro passo a realizar para esta questão de investigação é a consideração da execução da melhor abordagem, LDA ou LSA, a ser implementada para a modelação dos tópicos. Para a escolha de uma das abordagens foi executado o *Coherence Model*[1], cujo valor utilizado indica a avaliação dos modelos de tópicos. Neste caso, o LDA obteve um valor de 55% de coerência para um conjunto de 14 tópicos, enquanto o LSA alcançou um valor de aproximadamente de 49% de coerência, com 1 tópico. Desta forma, procedeu-se com a abordagem LDA, devido à coerência mais elevada e aos termos de cada tópico definirem melhor o conceito do Airbnb.

[1] `https://radimrehurek.com/gensim/models/coherencemodel.html`, acedido a 05-09-2019

5.2.1.1 Tópicos LDA

De seguida, procedeu-se à análise dos termos de cada tópico. Com o pré-processamento dos dados realizado, procedeu-se novamente a uma nova execução da abordagem LDA. Por forma a selecionar o número de tópicos mais adequado (número de tópicos ótimo) foi calculada a coerência em função do número de tópicos, valores que são apresentados na Figura 5.5. Tendo em conta os valores calculados foram escolhidos 20 tópicos.

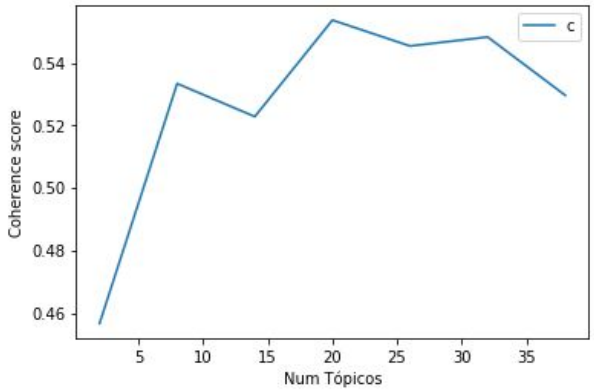


Figura 5.5: Coerência por número de tópicos no modelo LDA

Procedeu-se à identificação dos termos subjacentes de cada tópico. A Tabela 5.4, mostra os dez principais termos para cada tópico e o respetivo número de documentos associado.

5.2.1.2 Atribuição de categorias aos tópicos

Por forma a analisar corretamente esta questão de investigação foi necessário proceder à análise dos termos dos tópicos, para que posteriormente estes tópicos fossem aglomerados em várias categorias, consoante a similaridade entre eles. As categorias definidas tentaram seguir o padrão da classificação da experiência do hóspede, presente na plataforma. Este padrão é constituído pelos atributos que permitem classificar um alojamento no Airbnb: *Accuracy* (precisão), *Check-in* (chegada), *Cleanliness* (limpeza), *Communication* (comunicação), *Location* (localização) e *Value* (valor) no entanto, ao observar a análise, verificou-se que para além destes atributos houve necessidade de adicionar mais três categorias: *Amenities* (comodidades), *House* (casa), *Return & Recommendation* (voltar e recomendar), apresentadas na Tabela 5.3.

Categorias	Tópicos Finais
Accuracy	8: night, street, apartment, noise, great, place, noisy, quite, people, middle 18: best, place, experience, better, picture, exactly, describe, like, amaze, photo
Amenities	1: flat, machine, washing, conditioning, position, professional, fairly, air, brilliant, nespresso 12: car, parking, price, service, baby, expect, fix, park, find, bigger
Check-in	2: check, apartment, get, arrive, stay, time, host, late, make, checkin 17: wine, coffee, bottle, welcome, touch, left, arrival, food, fresh, portuguese
Cleanliness	11: nice, really, place, stay, clean, good, great, room, perfect, time
Communication	5: house, talk, apartment, brand, live, english, speak, lady, self, comfortable 19: give, tip, great, apartment, place, host, provide, local, stay, information
House	3: home, make, stay, feel, comfortable, apartment, like, felt, host, welcome 4: building, apartment, floor, stairs, top, hill, elevator, street, steep, old 7: room, bathroom, bed, kitchen, shower, good, bedroom, apartment, one, clean 9: view, beautiful, amaze, great, apartment, balcony, terrace, city, lovely, enjoy 13: family, space, pool, perfect, people, group, garden, area, property, large
Location	6: walking, restaurant, distance, close, apartment, public, great, bars, transport, attraction 10: walk, metro, station, close, minutes, apartment, restaurant, train, tram, bus 16: apartment, center, nice, city, need, locate, close, near, easy, metro
Return & Recommendation	14: back, come, stay, place, perfect, go, time, love, amaze, next 15: great, apartment, recommend, stay, place, host, highly, perfect, lovely, excellent
Value	0: apartment, good, value, awesome, luxury, money, hide, loft, helpfull, recommand

Tabela 5.3: Categorias atribuídas para os 20 tópicos

Deste modo, são nove as categorias definidas. Estas categorias surgiram consoante o que foi indicado por cada termo de determinado tópico, onde na seguinte descrição é possível verificar o significado de cada categoria:

Accuracy Nesta categoria os termos representam a coerência entre os comentários dos hóspedes e o que está descrito nos anúncios dos alojamentos. Como por exemplo, o alojamento na realidade não estar conforme as fotografias do anúncio do alojamento, ou até mesmo indicarem no anúncio que existe determinado equipamento e na reali-

dade não ter, etc.

Amenities Os termos aqui representados figuram as comodidades presentes em cada alojamento, como por exemplo, máquina de lavar, secador, forno, etc.

Check-in Esta categoria tem presente os termos que mostram como o proprietário dá as boas vindas ao hóspede. É a primeira impressão pessoal que o hóspede tem do proprietário.

Cleanliness Categoria que revela os termos que abordam as condições de limpeza de um alojamento.

Communication Nesta categoria, os termos abordam a comunicação acessível ou não com os proprietários e capacidade de resposta.

House Representa os termos referentes ao espaço de um alojamento (quarto, cozinha, sala, etc.), toda a área envolvente e o fato de os hóspedes se sentirem como se estivessem na sua própria casa.

Location Os termos indicam as vantagens e desvantagens da localização em termos de proximidade com outros pontos de interesse, caraterísticas da zona e se é ou não perto de transportes públicos e restaurantes.

Return & Recommendation São identificados termos que refletem o sentimento de voltar novamente ao alojamento e recomendar a experiência a outra pessoa.

Value Os termos representam o valor do alojamento, se compensa o preço e se está de acordo com a qualidade do mesmo.

Os tópicos mais relevantes de acordo com o número de documentos (quantos mais documentos, mais relevante é o tópico), são demonstrados na Tabela 5.4. O tópico 15 apresenta-se como o mais relevante, com 63.219 documentos, seguido do tópico 11, com 36.830 documentos. Como tal, e de acordo com a relevância dos tópicos, as categorias *Return & Recommendation* e *Location* são as mais relevantes.

Tópico	#Docs	Palavras
0	1.084	apartment, good, value, awesome, luxury, money, hide, loft, helpfull, recommend
1	194	flat, machine, washing, conditioning, position, professional, fairly, air, brilliant, nespresso
2	23.562	check, apartment, get, arrive, stay, time, host, late, make, checkin
3	19.106	home, make, stay, feel, comfortable, apartment, like, felt, host, welcome
4	3.902	building, apartment, floor, stairs, top, hill, elevator, street, steep, old
5	544	house, talk, apartment, brand, live, english, speak, lady, self, comfortable
6	9.251	walking, restaurant, distance, close, apartment, public, great, bars, transport, attraction
7	24.971	room, bathroom, bed, kitchen, shower, good, bedroom, apartment, one, clean
8	17.573	night, street, apartment, noise, great, place, noisy, quite, people, middle
9	7.862	view, beautiful, amaze, great, apartment, balcony, terrace, city, lovely, enjoy
10	28.832	walk, metro, station, close, minutes, apartment, restaurant, train, tram, bus
11	36.830	nice, really, place, stay, clean, good, great, room, perfect, time
12	649	car, parking, price, service, baby, expect, fix, park, find, bigger
13	4.366	family, space, pool, perfect, people, group, garden, area, property, large
14	11.836	back, come, stay, place, perfect, go, time, love, amaze, next
15	63.219	great, apartment, recommend, stay, place, host, highly, perfect, lovely, excellent
16	25.722	apartment, center, nice, city, need, locate, close, near, easy, metro
17	1.433	wine, coffee, bottle, welcome, touch, left, arrival, food, fresh, portuguese
18	5.341	best, place, experience, better, picture, exactly, look, like, photo, describe
19	21.040	give, tip, great, apartment, place, host, provide, local, stay, information

Tabela 5.4: Identificação dos termos dos 20 tópicos LDA

5.2.2 Análise de Sentimentos

Relativamente à segunda parte da questão de investigação, diz respeito aos sentimentos dos aspetos da experiência de um hóspede:

- **Quais deles os hóspedes consideram mais positivos e mais negativos?**

Quanto ao sentimento, observa-se na Tabela 5.5 a média, a variância e o desvio-padrão do sentimento (apenas da variável *compound* do *vaderSentiment*, abordagem referida na Subsecção 4.5.1.1) do hóspede dividido nas nove categorias. Os hóspedes tendem a ser mais positivos nas categorias *house, location* e *value*. Isto significa que os hóspedes se sentem satisfeitos com os alojamentos reservados, a localização e preço. No que concerne às categorias com o sentimento neutro, constatou-se que as categorias *amenities, accuracy e cleanliness* estão um pouco mais abaixo do esperado pelos hóspedes. Nestas categorias, os comentários tendem a referir a importância de determinada comodidade existir num alojamento, ou o fato de não existirem determinadas comodidades, ou seja, a proporção de comentários negativos e positivos para estas categorias estão bastantes equiparadas entre si. Quanto à variância, as categorias *amenities*, *cleanliness* e *value* apresentam valores mínimos, o que significa que, existe menos variação dos sentimentos nos comentários, logo mais próximos são os comentários entre si.

Categoria	Média	Variância	Desvio-Padrão
House	0,66	0,53	0,73
Location	0,53	0,38	0,69
Value	0,47	0,09	0,62
Check-in	0,26	0,15	0,49
Return & Recommendation	0,23	0,17	0,41
Communication	0,21	0,48	0,39
Cleanliness	0,12	0,08	0,28
Accuracy	0,10	0,24	0,30
Amenities	0,01	0,01	0,11

Tabela 5.5: Medidas estatísticas do sentimento por categoria

5.3 Questão de Investigação nº 2 - Resultados

Será que os comentários negativos dos hóspedes chocam com as informações dos proprietários sobre os alojamentos?

A solução desta questão divide-se em três diferentes análises relativas a cada tipo de espaço (apartamento/casa inteira, quarto privado e quarto partilhado). De um modo geral, de acordo com a informação extraída do *dataset* não existem muitos casos em que o hóspede choca negativamente com o que foi descrito pelo proprietário sobre o alojamento.

5.3.1 Análise *chunks* por tipo de espaço

- **Consoante o tipo de espaço, o tipo de constatações é diferente? Há evidências claras que isto muda consoante o tipo de espaço?**

De acordo com o descrito anteriormente (Secção 4.5.2), na Figura 5.6 é possível verificar os *chunks* mais utilizados para descrever estes casos dos apartamentos/casas inteiras.

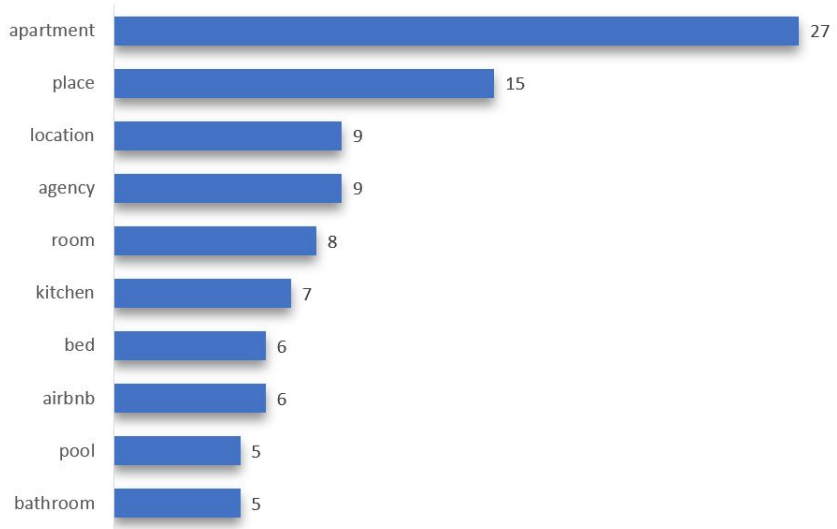


Figura 5.6: *Chunks* mais utilizados no caso dos apartamentos/casas inteiras

Num *top* 10, os *chunks* mais utilizados são «apartment» (apartamento), com cerca de 27 ocorrências, «place» (lugar) com 15 e «agency» (empresa) com 9. No entanto, surgiram outros *chunks* com menor frequência como «different apartment» (apartamento diferente) e «original booking» (reserva original). Nos exemplos abaixo, verifica-se com detalhe anúncios de alojamentos, onde estão presentes as descrições dos proprietários e os comentários que incluiem as palavras referidas na Figura 5.6. No Anexo A, na Figura A.1, é possível verificar mais um exemplo através da plataforma.

Exemplo 1 - Informação alojamento: *The space Apartment with 1 room, with 70 m2, 1 WC and a fully equipped kitchen for meals preparation, with kitchen utensils and home appliances (induction stove, hove, fridge, freezer, microwave, washing & dryer machine, toaster, electric kettle, iron, capsule coffee machine, pans,...), being also equipped with Air Conditioning. The room is equipped with one double bed, with the possibility to receive 2 extra guests in the living room, in a Sofa-Bed. Located in the center of Parque das Nações, the apartment is contemporary, with modern decoration and lots of natural light.*

Comentário hóspede: *The house was dusty the table smeared with grease. The kitchen did not have what was promised, no soap for dishwasher shampoo or gel. The ad said all this was included. The AC did not work it just blew air and was dirty. The fridge was dirty inside and the check in was horrible. I had to clean the apartment two separate times because of how dirty it was. (...) Furthermore at check in the guy checking me in had said on ge phone he could not hear me so he hung up on me, and did not call me back! I had to call him back. He also insisted on a 50 Euro deposit which was not outlined on airbnb. (...) He told me how to set up the tv and it did not work. I left a message for him to come fix it and nobody ever got back to me and I was never able to use the tv. Overall this was a terrible experience for airbnb. I am formally asking here for a partial refund.*

Exemplo 2 - Informação alojamento: *Apartment 206 is a charming apartment, located in the Baixa district. From here you can walk to Alfama, Chiado and Bairro Alto. Very good location in Lisbon. The apartment is on a 2nd floor (no elevator), has 50m2 and includes: - living room with dining and sitting areas, one double sofa bed (140cm x 200cm), flat screen TV and DVD - a kitchenette with electric stove, microwave, fridge with freezer, washing and dishwasher machines and utensils - one bedroom with double bed (140cm x 200cm) - a bathroom with bathtub. Cleaning and Maintenance - each flat is fully cleaned and prepared before your arrival - for stays longer than seven days we provide a mid-stay cleaning - when leaving, the tenant is obliged to leave the apartment in good condition.*

Comentário hóspede: *The place needs to be renovated for sure. There was smell of mould everywhere and marks of previous existing mould..especially in the be-*

droom. The walls were very dirty. The oven was not working. The washing machine was not working. The internet was not working for the first 4 days.

Resposta *host:* *(...) we will check what went wrong. About the internet unfortunately we had a problem with several apartments, that we did our best to solve quickly but we were dependent of the internet company to solve it.*

Nos exemplos apresentados, o hóspede mostra-se descontente com as condições de cada alojamento. Como é o caso de que as fotos do anúncio do alojamento mostram uma situação, mas posteriormente o hóspede na sua estadia verifica que não está conforme o exposto. Outro dos casos, são as descrições realizadas pelos proprietários, referindo as propriedades do alojamento e comodidades que este tipo de experiência oferece, mas que presencialmente, o hóspede deteta como uma falha, como é o exemplo de aparelhos e eletrodomésticos avariados e falta de certos utensílios.

Como esperado, os *chunks* utilizados para o quarto privado diferem um pouco do anterior tipo de espaço. O seu *top* 10, é demonstrado na Figura 5.7.

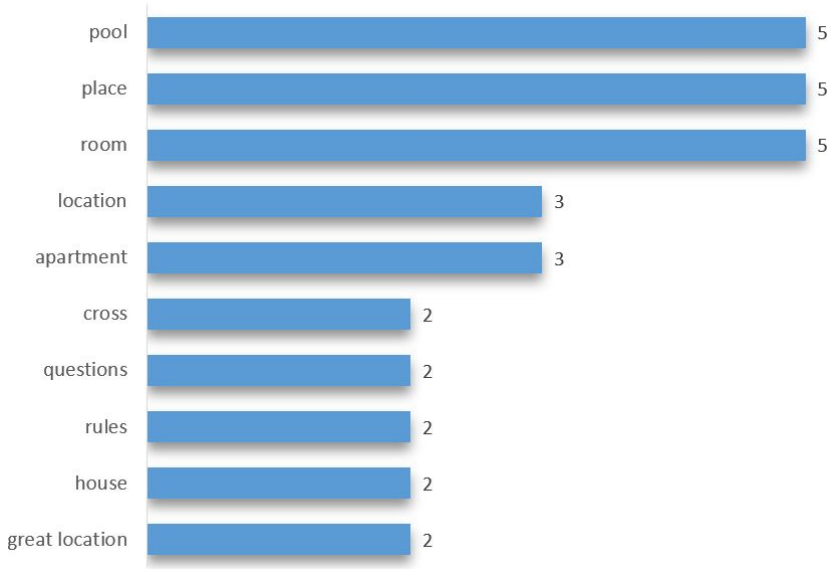


Figura 5.7: *Chunks* mais utilizados no caso dos quartos privados

Num *top* 10, os *chunks* mais utilizados são «pool» (piscina), «room» (sala) e «place» (lugar), com cerca de 5 ocorrências e «location» (localização) com 3 ocorrências. No entanto, surgiram outros *chunks* com menor frequência como «hear absolutely everything» (ouve-se absolutamente tudo) e «confused person» (pessoa confusa).

Nos exemplos seguintes, são apresentados alguns comentários que incluem estas palavras, também patentes na Figura A.2 do Anexo A.

Exemplo 3 - Informação alojamento: *Modern suites 5 minutes from Avenida da Liberdade, air conditioning, Netflix and wi fi with hot spot to use outside. Carefully deco-*

rated, the Smart Suites LX51 offer a bedroom with double bed and private bathroom. Guests have access to a co-working space, patio and seating area with Grab & Bistro area and Food Court where they can prepare light meals. Daily housekeeping service and breakfast served in the morning. The proximity to the subway guarantees access to the whole city. Apt In Lisbon's place is located in Lisbon, Portugal. Conveniently located a few meters from the accommodation you can find pharmacies and supermarkets as well as cafes, restaurants and bars.

Comentário hóspede: *Place wasn't at the location shown on the map or address. Luckily the cabbie was nice enough to help me find the place, else wouldn't have been able to. Room was decent, but a little tiny. Barely enough space to move around or leave luggage. Apt, if one is travelling alone with a small bag. Location was extremely central and couldn't be any better!*

Exemplo 4 - Informação alojamento: *An historical building recovered, with 4 modern and comfortable apartments, with 60m square each, holding 2 bedrooms, one double, the other single, one bathroom, living room and kitchen. These are fully equipped flats, suitable for 2 PAX.*

Comentário hóspede: *Centrally located, very well decorated (the room looks just like the photos) and comfortable but had to keep windows closed to keep out the noise at night which was too bad as they are lovely big windows and it wasn't too cold out. As beautiful as the space was it didn't look very nice after a few days of not being able to put away my clothes nor hang up coat, no cupboard nor hooks. I would have loved to have had a kettle to make myself tea as well. Lots of restaurants and bars around.*

Para o caso do quarto privado, os exemplos, demonstram que o hóspede se mostra igualmente descontente conforme descrito no tipo de espaço anterior, apesar da sua identificação das falhas e necessidades ser distinta, tratando-se de tipos de espaço diferentes. Como por exemplo, o fato de existir uma casa de banho partilhada, ou até mesmo o barulho sentido no alojamento.

Por fim, para o quarto partilhado não surgiram quaisquer contratempos relativamente a esta análise.

5.4 Questão de Investigação nº 3 - Resultados

O *score* associado aos indicadores espelha o que é descrito nos comentários?

Para dar resposta a esta questão, vamos utilizar vários algoritmos que permitem fazer regressão linear. A ideia será então utilizar o conteúdo dos comentários para modelar o score. Conforme referido na Secção 4.5.3, o primeiro passo inicia-se com a extração

de *features* usando o TF-IDF *Vectorizer*, posteriormente, é processada a validação cruzada para selecionar um dos três diferentes algoritmos de *machine learning*: *Stochastic Gradient Descent Regressor (SGDR), XGBoostRegressor* e *CatBoostRegressor* com diferentes taxas de aprendizagem e *alphas* (referidas na Secção 4.5.3). Cada um destes algoritmos observa a variável resposta (*review_scores_rating*, com um *score* de 0-100%) e faz a diferença com o *score* atual e o *score* estimado através dos comentários. De acordo com as várias taxas de aprendizagem e *alpha*, observa-se o menor erro de validação (MSE). Conforme Tabela 5.6. será escolhido o regressor e os parâmetros a utilizar na *pipeline* PLN.

Como se pode verificar pela Tabela 5.6 não existe proximidade entre os três algoritmos em termos de taxa de erro de validação (MSE) nos resultados de validação cruzada, esta taxa permite verificar o desempenho dos modelos em dados nunca antes vistos. No entanto, o algoritmo SGDR teve um desempenho ligeiramente superior na taxa de *alpha* de 0,0001 e foi o algoritmo selecionado para a análise.

Regressor	**Erros Treino (MSE)**	**Erros Validação (MSE)**
sgdr1	48,33	48,34
sgdr0,1	46,20	46,27
sgdr0,01	36,86	37,43
sgdr0,001	27,69	30,14
sgdr0,0001	25,20	29,44
sgdr1e-05	24,83	29,51
sgdr1e-06	24,81	29,52
cat1	11,55	50,99
cat0,1	16,13	31,99
cat0,01	28,70	33,48
cat0,001	1234,61	1237,22
cat0,0001	7124,83	7125,76
cat1e-05	8506,78	8506,89
cat1e-06	8659,01	8659,02
xgb1	9,65	46,01
xgb0,1	22,54	32,78
xgb0,01	1184,61	1187,04
xgb0,001	7035,06	7035,27
xgb0,0001	8414,40	8414,41
xgb1e-05	8566,40	8566,40
xgb1e-06	8581,75	8581,75

Tabela 5.6: Resultados da regressão para *SGDR, CatBoostRegressor* e *XGBoostRegressor*

Por fim, o *pipeline* TF-IDF *Vectorizer* foi ajustado para os dados, transformando-o para um algoritmo SGDR com uma taxa de *alpha* de 0,0001.

Na Figura 5.8, pode-se verificar a comparação dos dados reais dos primeiros duzentos alojamentos do conjunto de dados, bem como o *score* previsto de SGDR para os comentários. Contudo, quando o *score* real é abaixo de 75%, raramente a predição através dos comentários prevê o mesmo *score* que efetivamente foi dado. Neste caso, os hóspedes tendem a ser mais negativos a classificar o *score* da experiência, do que propriamente na

escrita do comentário.

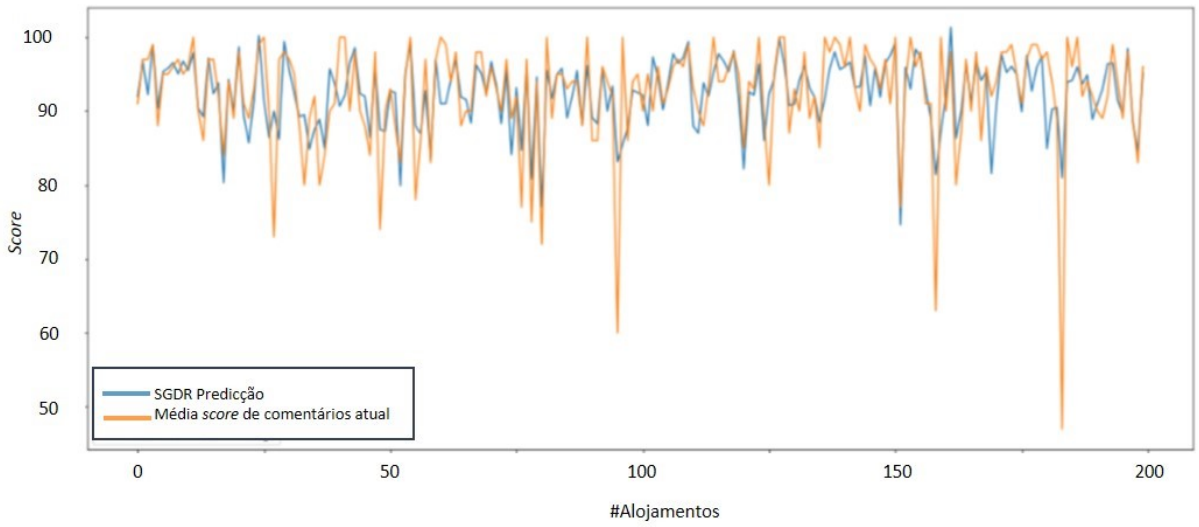


Figura 5.8: Previsão SGDR vs. Média do *score* real

5.5 Questão de Investigação nº 4 - Resultados

Relativamente aos alojamentos na cidade de Lisboa, os hóspedes escolhem tendencialmente os alojamentos cujos proprietários detêm o estatuto de *superhost*?

- **Existe variação pelo tipo de espaço?**

Analisando esta questões, foi possível verificar que, quanto às reservas nos alojamentos, num período de um ano, os hóspedes tendencialmente preferem os alojamentos que pertencem aos *hosts* regulares. Conforme Figura 5.9, é possível observar com 66% (1.084.704 reservas de alojamentos) os *hosts* regulares, contra 34% (548.641 reservas de alojamentos) dos *superhosts*.

Na Figura 5.9, é possível observar comparativamente a análise das reservas para cada tipo de espaço, dependendo do tipo de proprietário. Na mesma análise, para o *host* regular, foi verificado que quanto ao tipo de espaço, o hóspede tende a reservar os seus alojamentos nos apartamentos/casas inteiras (828.023), de seguida quarto privado (250.014) e por fim quarto partilhado (6.667). Constatou-se que os *hosts* alugam mais o tipo de casas inteiras e até mesmo o quarto privado, ao invés do que acontece para os alojamentos dos *superhosts,* em que a frequência de reservas para estes é bastante mais reduzida.

- **Em comparação com os *hosts* regulares, quais as categorias mais abordadas nos comentários?**

Relativamente a este ponto, foi necessário proceder à execução da modelação por tópicos. Nesta fase, procedeu-se à execução de duas abordagens: LDA e LSA.

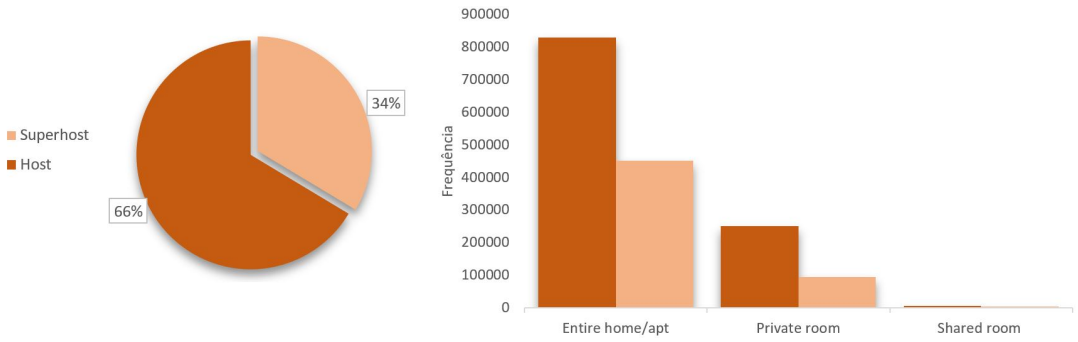


Figura 5.9: Preferência do hóspede por tipo de proprietário/espaço

5.5.1 Modelação por tópicos - LDA e LSA

Conforme já referido na Subsecção 5.2.1, o primeiro passo a realizar para esta questão de investigação é a consideração de qual das duas abordagens deve ser implementada. Desta forma, foi necessário calcular a configuração que maximiza a coerência do modelo. No caso da abordagem LDA, um conjunto de 14 tópicos apresenta uma boa coerência, sendo que para o LSA, a coerência escolhida baseou-se em 49% para 2 tópicos. Desta forma, devido à coerência, tópicos reduzidos e termos inconsistentes da abordagem LSA, procedeu-se com a abordagem LDA.

5.5.1.1 Tópicos LDA

De seguida, procedeu-se à análise dos termos de cada tópico. Conforme supramencionado na Secção 5.2.1.1, procedeu-se igualmente à substituição de determinados termos para o termo *location* e remoção de termos considerados como *stopwords*. Após estas alterações, procedeu-se novamente à execução da abordagem LDA. Por forma a selecionar o número de tópicos ótimo foi calculada a coerência em função do número de tópicos. Um total de 26 tópicos foi seleccionado, como se pode observar na Figura 5.10.

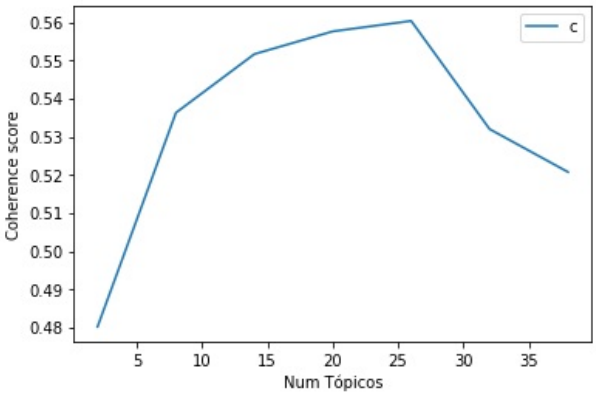


Figura 5.10: Coerência por número de tópicos no modelo LDA

5.5.1.2 Atribuição de categorias aos tópicos

A maior parte dos tópicos indicados por este modelo são idênticos aos dos resultados da primeira questão de investigação (20 tópicos). Sendo que, para este caso o modelo indicou mais 6 tópicos, por sua vez, as categorias são maioritariamente as mesmas do resultado anterior. Nesta análise, foram assim adicionadas duas categorias: *Host & Value* (proprietário & valor) e *Reservation* (reserva). Deste modo, observou-se pela Tabela 5.7 um total de 10 categorias.

Para além das categorias já mencionadas anteriormente, de seguida é efetuada a descrição das duas novas categorias encontradas para esta análise:

Host & Value Termos que referem os aspetos num proprietário, como por exemplo, simpático, prestativo, entre outras, e os termos que representam o valor do alojamento tendo em conta o proprietário, se compensa o preço e se está de acordo com a qualidade do mesmo.

Reservation Esta categoria indica os termos utilizados para a reserva de estadia num alojamento Airbnb.

Na Figura 5.11 pode-se observar os tópicos mais relevantes de acordo com o número de documentos. O tópico 22, está mais presente nos comentários para qualquer um dos proprietários, embora mais elevado nos *superhosts*. Sendo que, para os tópicos 0 e 25 a proporção dos comentários cujo alojamento pertence ao *host* regular estão mais elevados do que para os *superhosts*. Por fim, para os *superhosts*, os tópicos 5 e 13 são igualmente elevados, ao contrário do que acontece com os *hosts* regulares. Assim, os tópicos mais relevantes pertencem às categorias*: Return & Recommendation, Host & Value, Location* e *House*. É interessante verificar que com a análise do proprietário, a categoria *Host & Value* é mais relevante nos alojamentos cujo proprietário é um *host* regular, isto porque um *superhost* tendencialmente preza pelo bom serviço prestado ao hóspede.

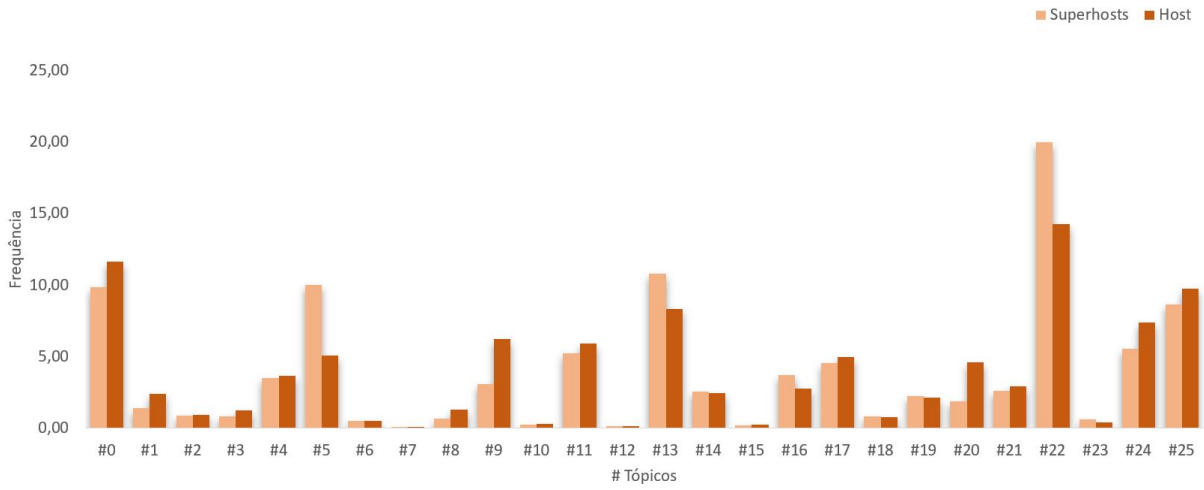


Figura 5.11: Proporção de comentários com um tópico como o tópico mais ponderado

Categorias	**Tópicos Finais**
Accuracy	8: night, street, noise, door, windows, open, sleep, hear, right, next 9: bit, little, place, great, small, quite, get, noisy, night, stay 15: picture, look, exactly, better, like, photo, expect, describe, sparkling, show
Amenities	10: easy, parking, access, car, check, park, expectation, free, make, find 12: apartment, nice, separate, bedding, curtain, situation, tasteful, man, cosy, soft 20: apartment, shower, air, water, hot, good, bathroom, cold, heater, wifi
Check-in	3: airbnb, meet, get, arrive, hide, time, show, wait, key, late 21: check, checkin, apartment, airport, flight, arrival, early, late, luggage, time 23: give, tip, welcome, wine, information, provide, local, bottle, warm, arrival
Cleanliness	24: really, nice, place, stay, good, enjoy, clean, everything, lot, perfect
Communication	4: need, always, apartment, great, quick, question, host, respond, answer, help
Host & Value	0: host, good, super, helpful, recommend, friendly, lovely, value, money, price
House	1: room, bathroom, space, kitchen, living, nice, rooms, big, two, bedroom 5: make, best, stay, place, experience, airbnb, go, one, sure, amaze 6: house, home, like, felt, feel, beautiful, hospitable, kind, friendly, comfortable 16: lovely, breakfast, enjoy, little, area, garden, morning, space, drink, pool 18: view, amaze, balcony, beautiful, terrace, river, city, top, stun, castle 19: bed, kitchen, apartment, need, floor, comfortable, machine, everything, stairs, comfy
Location	2: old, town, charm, part, heart, neighborhood, quiet, middle, right, street 11: walk, metro, station, minutes, close, train, bus, away, tram, city 17: restaurant, great, close, apartment, bars, shop, public, area, nearby, cafe 25: apartment, nice, city, locate, center, good, place, perfect, close, perfectly
Reservation	14: one, new, like, place, property, day, book, time, two, stay
Return & Recommendation	7: spot, wish, longer, sweet, worth, outstanding, blanket, seeing, tall, happily 13: stay, back, place, come, great, apartment, perfect, love, time, definitely 22: apartment, recommend, great, stay, highly, place, clean, host, definitely, stylish

Tabela 5.7: Categorias atribuídas para os 26 tópicos

6 Conclusões e Trabalho Futuro

Neste estudo foram analisadas várias abordagens para realizar a tarefa da modelação de comentários para o caso da plataforma Airbnb. Esta tarefa foi dividida em quatro questões de investigação, por forma a ser possível responder a cada uma.

Para a primeira questão, foram desenvolvidas duas abordagens de modelação por tópicos para a revelação das categorias mencionadas nos comentários pelos hóspedes. A abordagem que se destacou foi o LDA, com um valor de coerência de 55%. Com a escolha da abordagem LDA, realizou-se a identificação do número ótimo de tópicos, através da coerência. Com uma coerência de 55%, foram considerados 20 tópicos para a análise. Com a análise destes tópicos verificou-se que haviam tópicos muito idênticos entre si, tendo sido associados em categorias: *Accuracy, Amenities, Check-in, Cleanliness, Communication, House, Location, Return & Recommendation* e *Value*. Não obstante, pelo número de documentos as categorias mais relevantes pertencem ao *Return & Recommendation* e *Location,* sendo assim possível constatar que os hóspedes dão importância à localização do alojamento, estar perto de pontos de interesse e por gostarem tanto da experiência usufruída neste tipo de alojamentos, voltam e recomendam a outras pessoas.

Ainda relativamente à primeira questão, verificou-se que a maioria das categorias têm em média o sentimento do *score* agregado (*compound*) neutro, com valores entre os 0,01 (*Amenities*) e 0,47 (*Value*), à exceção das categorias *House* e *Location,* cujo sentimento agregado é o mais positivo, com 0,66 e 0,53. Neste caso, para as categorias com um sentimento mais neutro, pode servir como um alerta para os proprietários agirem rapidamente por forma a melhorar este tipo de situações que está sob a responsabilidade dos mesmos. Segundo o Airbnb[1] e para o caso das *Amenities*, os proprietários devem incluir as comodidades mais básicas, mas também adicionar detalhes ao alojamento e garantir que de uma estadia para outra as comodidades estão conformes, isto porque existem muitos comentários de que estas estão danificadas ou simplesmente não funcionam e estes pequenos contratempos trazem comentários negativos por parte dos hóspedes.

Para a segunda questão, foi desenvolvida a técnica de *chunks,* por forma a ser capaz de extrair a informação dos comentários negativos que chocam com as descrições dos alojamentos. São poucos os casos encontrados, mas para o caso do apartamento/casa inteira,

[1] https://community.withairbnb.com/t5/Airbnb-Updates/How-to-be-a-successful-Airbnb-host-Setting-up-your-space/td-p/989234, acedido a 10-09-2019

os termos considerados foram, «apartment» (apartamento), «place» (lugar), «agency» (empresa), «different apartment» (apartamento diferente), entre outros. Sendo que para o caso do quarto privado, os *chunks* foram distintos, como por exemplo, «pool» (piscina), «room» (sala), «location» (localização), «hear absolutely everything» (ouve-se absolutamente tudo), entre outros. Tratando-se de tipos de espaço distintos é esperado que as respetivas análises também o sejam, isto porque um hóspede que reserva um espaço inteiro, não tem a presença do proprietário que é como acontece no caso do quarto privado, o que por si só neste caso pode levar a discordâncias com o próprio proprietário mais facilmente. Embora tanto num caso como no outro na realidade existem inconformidades entre o que está descrito no anúncio ou até mesmo as fotografias não coincidirem com a realidade. Sendo que, existem certos casos em que o hóspede pode não ler ou perceber corretamente o anúncio e escrever desta forma comentários negativos. Estas reclamações servem de advertência aos futuros hóspedes e/ou como sugestões para o proprietário melhorar a situação. Nestes casos, o proprietário tem a possibilidade de melhorar a descrição do seu anúncio, colocar fotos mais apelativas descrevendo o alojamento como ele é na realidade e também responder aos comentários dos hóspedes por forma a corrigir o que foi descrito por eles, caso se aplique.

Relativamente à terceira questão, foi desenvolvido um modelo capaz de prever o *score* baseado nos comentários. O modelo em questão, recebe um conjunto de dados onde o *score* já se encontra atribuído pelos hóspedes, onde é assim calculada a diferença entre este e o *score* estimado. O modelo que se destacou na previsão do *score* foi o *Stochastic Gradient Descent Regressor* (SGDR), que obteve menor erro de validação (MSE) de 29,44. Nos duzentos alojamentos do conjunto de dados selecionado, o *score* previsto de SGDR através dos comentários e o *score* real tiveram um comportamento idêntico, sendo que, quando o *score* real é abaixo de 75%, raramente a predição através dos comentários prevê o mesmo *score* que efetivamente foi dado. Nesta situação, o *score* real não espelha corretamente o que é descrito nos comentários, existindo uma disparidade entre eles, isto porque o hóspede tende a ser mais negativo na classificação do alojamento, do que propriamente no comentário.

Por fim a quarta questão, foi dividida em duas partes. A primeira parte, focou-se na análise de preferência do hóspede pelos alojamentos de cada proprietário e a segunda parte na análise das categorias mais abordadas nos comentários, relativamente a cada proprietário. Ao realizar a análise de um ano de reservas dos alojamentos, verificou-se que os alojamentos cuja propriedade são dos *hosts* regulares detêm a maioria das reservas com 66% (onde a preferência persiste nos apartamentos/casa inteira com 828.023 reservas) contra 34% dos *superhosts*. Conforme indicado anteriormente, os alojamentos dos *superhosts* tendem a ser mais caros que os dos *hosts* regulares, ao que os hóspedes tendem mais facilmente a reservar os seus alojamentos. Sendo que quanto ao tipo de espaço, apesar do apartamento/casa inteira ser mais caro, os hóspedes se sentem mais confortáveis, do que ter o proprietário na própria casa. Na segunda parte da questão, e em modo de comparação com

os proprietários, a análise das categorias mais abordadas nos comentários seguiu a abordagem de modelação por tópicos. Com uma coerência de 56%, foram selecionados 26 tópicos para a abordagem LDA. As categorias selecionadas foram: *Accuracy, Amenities, Check-in, Cleanliness, Communication, Host & Value, House, Location, Reservation* e *Return & Recommendation*. A principal categoria mais abordada foi o *Return & Recommendation* que está mais presente nos comentários para qualquer um dos proprietários, embora mais elevado nos *superhosts*. Sendo que, existem certas categorias que também tiveram algum destaque para o *host* regular, como é o caso de *Host & Value* e *Location,* já para os *superhosts*, a categoria que também prevaleceu foi a de *House*. A categoria principal *Return & Recommendation* remete a uma indicação de que os hóspedes tendem a ficar satisfeitos com este tipo de estadia voltando mais vezes ao mesmo alojamento e até recomendarem o mesmo. No entanto esta análise é mais elevada para o *superhost*, com 19,98 dos casos, devido ao seu estatuto distinto que preza um pouco mais pelos seus hóspedes.

Contudo, os resultados obtidos nestas análises podem ser melhorados. Desta forma, foram identificados alguns possíveis desenvolvimentos a realizar em trabalho futuro:

- Análise com todos os idiomas disponíveis nos comentários;
- No processo de modelação dos tópicos, identificar e testar a análise de *clusters*. Explorar métodos de modelação de tópicos não paramétricos como por exemplo o *Hierarchical Dirichlet Process* (HDP) e comparar os resultados com os métodos paramétricos LDA já efetuados para a análise;
- Para a análise de sentimentos, selecionar outro método de classificação de sentimentos, como por exemplo do sistema *TextBlob*, e comparar com os resultados do *sentimentVader;*
- Na análise da coerência dos comentários com a descrição do anúncio do alojamento, utilizar bigramas e trigramas, para a identificação de frases e palavras-chave;
- No processo de previsão do *score* através dos comentários, identificar e testar outros modelos de regressão.

Outras análises:

1. Comparar toda esta análise com outra cidade europeia, por forma a perceber as diferenças que existem entre regiões distintas;
2. Recentemente, a plataforma Airbnb, tem também a possibilidade de reservar experiências. Por norma, são atividades que uma pessoa organiza, por exemplo *tours* com guia turístico, ou até mesmo uma aula de cozinha. Através dessas experiências, o pretendido seria realizar o mesmo tipo de análise realizado para os alojamentos, possibilitando a descoberta de novos tópicos e novas análises;

3. Criação de um sistema de recomendações de alojamentos, dependendo do que já foi reservado anteriormente;

4. Implementação de um *dashboard* que permite ao proprietário observar o estado atual ao nível do detalhe da página do anúncio de cada alojamento.

Referências Bibliográficas

Albinsson, P. e Yasanthi Perera, B. (2012). Alternative Marketplaces in the 21st Century: Building Community through Sharing Events. *Journal of Consumer Behaviour*, 11:303–315.

Azevedo, Ana e Santos, M. F. (2014). KDD , SEMMA and CRISP-DM : A parallel overview. (January 2008).

Bellegarda, J. (2000). Large vocabulary speech recognition with multispan statistical language models. *IEEE Transactions on Speech and Audio Processing*, 8:76–84.

Bird, S., Klein, E., e Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Black, H. G. e Kelley, S. W. (2009). A storytelling perspective on online customer reviews reporting service failure and recovery. *Journal of Travel & Tourism Marketing*, 26(2):169–179.

Blake, C. (2011). Text mining. *Annual Review of Information Science and Technology*, 45(1):121–155.

Blei, D., Carin, L., e Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6):55–65.

Blei, D. M. e Lafferty, J. D. (2009). Topic models. In *Text Mining*, pages 101–124. Chapman and Hall/CRC.

Blei, D. M., Ng, A. Y., e Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bower, B. J. L. e Christensen, C. M. (1995). Disruptive Technologies : Catching the Wave. *Harvard business review*, (February).

Braz, L. M., Ferreira, R., Dermeval, D., e Douglas, V. (2009). Aplicando Mineração de Dados para Apoiar na Tomada de Decisão. (May 2015).

Bridges, J. e Vásquez, C. (2018). If nearly all Airbnb reviews are positive, does that make them meaningless? *Current Issues in Tourism*, 21(18):2057–2075.

Camilla, V. (2011). Complaints online : The case of TripAdvisor. *Journal of Pragmatics*, 43(6):1707–1717.

Cansoy, M. e Schor, J. (2016). Who Gets to Share in the "Sharing Economy": Understanding the Patterns of Participation and Exchange in Airbnb. *Unpublished Paper, Boston College*, pages 1–28.

Cascia, M. L., Sethi, S., e Sclaroff, S. (1998). Combining Textual and Visual Cues for Content-Based Image Retrieval on the World Wide Web. In *Proceedings of the IEEE Workshop on Content - Based Access of Image and Video Libraries*, CBAIVL '98, pages 24–28, Washington, DC, USA. IEEE Computer Society.

Cheng, M. e Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, 76:58–70.

Chu, R. K. e Choi, T. (2000). An importance-performance analysis of hotel selection factors in the Hong Kong hotel industry: A comparison of business and leisure travellers. *Tourism Management*, 21(4):363–377.

Dolnicar, S. e Otter, T. (2003). Which Hotel attributes Matter? A review of previous and a framework for future research. *Faculty of Commerce - Papers*, 1.

Erik, H. e Poul, T. (1999). A statistical test for the mean squared error. *Journal of Statistical Computation and Simulation*, 63(4):321–347.

Fan, M. e Khademi, M. (2014). Predicting a business star in yelp from its reviews text alone.

Festila, M. e Dueholm Müller, S. (2017). The Impact of technology-mediated consumption on identity: the case of Airbnb. *Paper Presented at the Proceedings of the 50th Hawaii International Conference on System Sciences*, pages 54–63.

Filieri, R., Alguezaui, S., e Mcleay, F. (2015). Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth. *Tourism Management*, 51:174–185.

Gaikwad, S. V. (2014). Text Mining Methods and Techniques. *International Journal of Computer Applications*, 85(17):42–45.

Graves, A. (2012). Long short-term memory. In *Supervised sequence labelling with recurrent neural networks*, pages 37–45. Springer.

Gunter, U. (2018). What makes an Airbnb host a superhost? Empirical evidence from San Francisco and the Bay Area. *Tourism Management*, 66:26–37.

Guo, Y., Barnes, S. J., e Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, 59:467–483.

Guo, Y., Wang, Y., e Wang, C. (2019). Exploring the salient attributes of short-term rental experience: An analysis of online reviews from chinese guests. *Sustainability*, 11(16):1–19.

Guttentag, D. (2015). Airbnb : disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism*, 18(12):1192–1217.

Guttentag, D., Smith, S., Potwarka, L., e Havitz, M. (2017). Why Tourists Choose Airbnb: A Motivation-Based Segmentation Study. *Journal of Travel Research*, 57(3):342–359.

Guttentag, D. A. e Smith, S. L. (2017). Assessing Airbnb as a disruptive innovation relative to hotelsSubstitution and comparative performance expectations. *International Journal of Hospitality Management*, 64(2017):1–10.

Hutto, C. e Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Ivanova, P. (2017). A contemporary overview of the application of collaborative consumption in tourism. *Business Management/Biznes Upravlenie*, 2017(2):73–86.

Jennings, G. e Weiler, B. (2006). Mediating meaning: Perspectives on brokering quality tourist experiences. *Quality Tourism Experiences*, page 57–78.

Joshi, B., Macwan, N., Mistry, T., e Mahida, D. (2018). Text Mining and Natural Language Processing in Web Data Mining. *2nd International Conference on Current Research Trends in Engineering and Technology*, 4(5):392–394.

Ju, Yongwook e Back, Ki-joon e Choi, Youngjoon e Lee, Jin-soo (2019). Exploring Airbnb service quality attributes and their asymmetric effects on customer satisfaction. *International Journal of Hospitality Management*, 77(August 2018):342–352.

Karim, M. e Das, S. (2018). Sentiment analysis on textual reviews. *IOP Conference Series: Materials Science and Engineering*, 396:1–7.

Karlsson, L., Kemperman, A., e Dolnicar, S. (2017). May I sleep in your bed? Getting permission to book. *Annals of Tourism Research*, 62:1–12.

Kumar, L. e Bhatia, P. K. (2013). Text Mining: Concepts, Process and Applications. *Journal of Global Research in Computer Science*, 4(3):36–39.

Lauer, T. e Deng, X. (2007). Building online trust through privacy practices. *International Journal of Information Security*, 6(5):323–331.

Lehr, D. (2015). An Analysis of the Changing Competitive Landscape in the Hotel Industry Regarding Airbnb. *Graduate Master's Theses, Capstones, and Culminating Projects*, 188:1–76.

Liang, S., Schuckert, M., Law, R., e Chen, C.-C. (2017). Be a "Superhost": The Importance of Badge Systems for Peer-to-peer Rental Accommodations. *Tourism Management*, 60:454–465.

Lutz, C. e Newlands, G. (2018). Consumer segmentation within the sharing economy: The case of Airbnb. *Journal of Business Research*, 88:187–196.

Ma, X., Hancock, J. T., Lim Mingjie, K., e Naaman, M. (2017). Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2397–2409.

Mankad, S., Han, H. S., Goh, J., e Gavirneni, S. (2016). Understanding Online Hotel Reviews Through Automated Text Analysis. *Service Science*, 8(2):124–138.

Moon, H., Miao, L., Hanks, L., e Line, N. D. (2019). Peer-to-peer interactions: Perspectives of Airbnb guests and hosts. *International Journal of Hospitality Management*, 77:405–414.

Moufahim, M. (2013). User-Generated Brands and Social Media : Couchsurfing and AirBnb. *Contemporary Management Research*, 9(1):85–90.

Oh, H., Fiore, A., e Jeong, M. (2007). Measuring Experience Economy Concepts : Tourism Applications. *Journal of Travel Research*, 46(2):119–132.

Ozanne, L. e Ballantine, P. (2010). Sharing as a Form of Anti-Consumption? An Examination of Toy Library Users. *Journal of Consumer Behaviour*, 9:485–498.

Pande, V. C. e Khandelwal, A. (2014). A Survey Of Different Text Mining Techniques. *IBMRD's Journal of Management & Research*, 3(1):125–133.

Paulauskaite, D., Morrison, A. M., Powell, R., e Coca-Stefaniak, J. A. (2017). Living like a local : Authentic tourism experiences and the sharing economy. *International Journal of Tourism Research*, 19(6):619–628.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pine II, B. e Gilmore, J. (1998). Welcome to the experience economy. *Harvard business review*, 76:97–105.

Pizam (2010). Creating memorable experiences. *International Journal of Hospitality Management*, 29(3):343.

Pradhan, L., Zhang, C., e Bethard, S. (2016). Towards extracting coherent user concerns and their hierarchical organization from user reviews. In *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, pages 582–590. IEEE.

Redhu, S., Srivastava, S., Bansal, B., e Gupta, G. (2018). Sentiment Analysis Using Text Mining : A Review. *International Journal on Data Science and Technology*, 4(2):49–53.

Resnick, P. e Varian, H. R. (1997). Recommender Systems. *Commun. ACM*, 40(3):56–58.

Schmidt, G. e Druehl, C. (2008). When Is Disruptive Innovation Disruptive? *Journal of Product Innovation Management*, 25:347–369.

Sixto, J., Almeida, A., e López-De-Ipiña, D. (2013). Analysing customers sentiments: An approach to opinion mining and classification of online hotel reviews. In *International Conference on Application of Natural Language to Information Systems*, pages 359–362.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., e Buttler, D. (2012). Exploring Topic Coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, number July, pages 952–961. Association for Computational Linguistics.

Sumathy, K. L. (2013). Text Mining : Concepts , Applications , Tools and Issues – An Overview. *International Journal of Computer Applications*, 80(4):29–32.

Sun, N., Liu, D., Zhu, A., Chen, Y., e Yuan, Y. (2019). Do Airbnb's "Superhosts" deserve the badge? An empirical study from China. *Asia Pacific Journal of Tourism Research*, 24(4):296–313.

Talib, R., Hanif, M. K., Ayesha, S., e Fatima, F. (2016). Text Mining:Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications*, 7(11):414–418.

Tan, A.-H. et al. (1999). Text Mining : The state of the art and the challenges Concept-based. volume 8, pages 65–70.

Teubner, Timm and Saade, Norman and Kawlitschek, Florian and Weinhardt, Christof (2016). It's only pixels, badges, and stars: On the economic value of reputation on Airbnb. *Australasian Conference on Information Systems*.

Thet, T. T., Na, J. C., e Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6):823–848.

Tsytsarau, M. e Palpanas, T. (2016). Managing Diverse Sentiments at Large Scale. *IEEE Transactions on Knowledge and Data Engineering*, 28(11):3028–3040.

Tung, V. W. S. e Ritchie, J. B. (2011). Exploring the essence of memorable tourism experiences. *Annals of Tourism Research*, 38(4):1367–1386.

Tussyadiah, I. (2015). An exploratory on drivers and deterrents of collaborative consumption in travel. *Information & Communication Technologies in Tourism 2015*, (December):817–830.

Tussyadiah, I. P. e Park, S. (2018). When guests trust hosts for their words: Host description and trust in sharing economy. *Tourism Management*, 67(August):261–272.

Tussyadiah, I. P. e Pesonen, J. (2015). Impacts of Peer-to-Peer Accommodation Use on Travel Patterns. *Journal of Travel Research*, 55(8):1022–1040.

Tussyadiah, I. P. e Zach, F. (2017). Identifying salient attributes of peer-to-peer accommodation experience. *Journal of Travel & Tourism Marketing*, 34(5):636–652.

Vale, D. S. (2018). Lisboa: População, alojamento e acessibilidade. (December):227–244.

Vorhies, W. (2016). CRISP-DM – a Standard Methodology to Ensure a Good Outcome [online]. Disponível em: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome [Consultado a: 2019-09-01].

W. Lehman, E. e Sztompka, P. (2001). Trust: A Sociological Theory. *Contemporary Sociology*, 30:418.

Wang, D. e Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. *International Journal of Hospitality Management*, 62:120–131.

Whye Teh, Y., Jordan, M., e J. Beal, Matthew & M. Blei, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Williamson, S., Wang, C., Heller, K., e Blei, D. (2009). Focused Topic Models. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*, pages 1–4.

Wirth, R. e Hipp, J. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39.

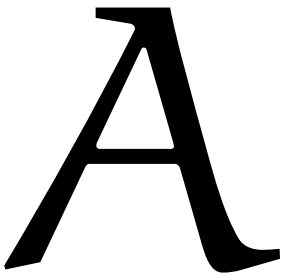
Xie, K. e Mao, Z. (2017). The Impacts of Quality and Quantity Attributes of Airbnb Hosts on Listing Performance. *International Journal of Contemporary Hospitality Management*, 29(9):2240–2260.

Xu, X., Wang, X., Li, Y., e 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.

Ye, Q., Law, R., e Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28:180–182.

Zervas, G., Proserpio, D., e Byers, J. W. (2013). The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. *Journal of marketing research*, 54(5):1–36.

Anexos



Tabelas

Figuras

[1] Siglas dos idiomas retiradas da fonte: `https://detectlanguage.com/languages`

[2] Siglas dos idiomas retiradas da fonte: `https://detectlanguage.com/languages`

Variáveis	**Descrição**	**Tipo**
id	Identificador único de cada alojamento	Metadata
name	Nome de cada alojamento	Metadata
description	Descrição textual de cada alojamento	Texto
neighborhood_overview	Breve descrição da zona onde se encontra o alojamento	Texto
host_id	Identificador único do proprietário	Metadata
host_name	Nome do proprietário	Metadata
host_since	Data que o proprietário ingressou na plataforma Airbnb	Data
host_is_superhost	Informação de que o proprietário é ou não um *superhost*	Categórica
host_total_listings_count	Total de alojamentos por cada proprietário	Numérica
neighbourhood_cleansed	Zona onde o alojamento está localizado	Texto
city	Cidade em que está situado cada alojamento	Categórica
latitude	Coordenadas de localização de cada alojamento	Numérica
longitude	Coordenadas de localização de cada alojamento	Numérica
property_type	Tipo de propriedade de cada alojamento	Categórica
room_type	Tipo de espaço de cada alojamento (apartamento/casa inteira, quarto privado, quarto partilhado)	Categórica
amenities	Informação das comodidades de cada alojamento	Categórica
price	Preço da reserva de cada alojamento	Numérica
number_of_reviews	Número de comentários que cada alojamento disponibiliza	Numérica
first_review	Data do primeiro comentário de cada alojamento	Data
last_review	Data do último comentário de cada alojamento	Data
review_scores_rating	*Score* de 0-100% associado a cada alojamento, agrega os *scores* associados abaixo, atribuídos pelos hóspedes	Numérica
review_scores_accuracy	Score de 1-10 na categoria da precisão em que a descrição do alojamento reflete exatamente o que é na realidade	Numérica
review_scores_cleanliness	*Score* de 1-10 na categoria da limpeza	Numérica
review_scores_checkin	*Score* de 1-10 na categoria de check-in. Chegada do hóspede ao alojamento	Numérica
review_scores_communication	*Score* de 1-10 na categoria da comunicação. Comunicação antes e durante a estadia	Numérica
review_scores_location	*Score* de 1-10 na categoria da localização	Numérica
review_scores_value	*Score* de 1-10 na categoria do valor, em que existe uma boa relação qualidade/preço	Numérica
calculated_host_listings_count	Número de alojamentos por proprietário	Numérica
reviews_per_month	Média dos comentários por mês de cada alojamento	Numérica

Tabela A.1: Descrição e tipo das variáveis do *dataset Listings*

Variáveis	Descrição	Tipo
listing_id	Identificador único dos alojamentos	Metadata
id	Identificador único dos comentários dos hóspedes	Metadata
date	Data em que foi publicado o comentário, no formato AAAA/MM/DD	Data
reviewer_id	Identificador único dos hóspedes	Metadata
reviewer_name	Nome dos hóspedes	Metadata
comments	Conteúdo do comentário escrito pelo hóspede	Texto

Tabela A.2: Descrição e tipo das variáveis do *dataset Reviews*

Variáveis	Descrição	Tipo
listing_id	Identificador único de cada alojamento	Metadata
date	Data no formato AAAA/MM/DD, com a disponibilidade de cada alojamento até ao ano seguinte	Data
available	Caso esteja o alojamento disponível, surge um «t», senão surge um «f»	Categórica
price	Preço por noite de cada alojamento	Numérica
minimum_nights	Mínimo de noites que o alojamento permite	Numérica
maximum_nights	Máximo de noites que o alojamento permite	Numérica

Tabela A.3: Descrição e tipo das variáveis do *dataset* Calendário

Variáveis	Valor
id	0,00
name	0,10
description	0,83
neighborhood_overview	33,53
host_id	0,00
host_name	0,00
host_since	0,00
host_is_superhost	0,00
host_total_listings_count	0,00
neighbourhood_cleansed	0,00
city	0,57
latitude	0,00
longitude	0,00
property_type	0,00
room_type	0,00
amenities	0,00
price	0,00
number_of_reviews	0,00
first_review	17,45
last_review	17,45
review_scores_rating	18,30
review_scores_accuracy	18,36
review_scores_cleanliness	18,33
review_scores_checkin	18,40
review_scores_communication	18,36
review_scores_location	18,39
review_scores_value	18,40
calculated_host_listings_count	0,00
reviews_per_month	17,45

Tabela A.4: Percentagem de valores omissos

Lisboa

Entire apartment
6 guests 2 bedrooms 3 beds 1 bath

The space

About your Apartment
This beautiful apartment located in the heart of downtown Lisboa has 60m2 and is inserted in a building without elevator. The apartment consists of two bedrooms and one bathroom with shower. Both bedrooms have 2 single beds each but one of them can be converted into a double bed upon request. The living room has a sofa bed big enough for two people to sleep in, a TV with international channels and a dining space with a table for six. The balcony has a table and two chairs and becomes the ideal place to enjoy a glass of wine after a day of city tour. The kitchen, spacious and fully equipped with everything you need to cook your favorite meals. The apartment has wi-fi connection throughout. Upon request and with a cost we can provide a baby cot.

Gianluca
January 2019

Very good position and clean, but the apartment was not the same as in the pictures, there was no balcony, was cold and the agency was 1 hour late for the check-in.

Response from
Thank you for your review, Gianluca. We agree that the location of the apartment is undoubtedly, one of its best features and we are glad to know that your overall experience was positive. However, we would like to apologise for moving you into a different property. In fact, we believed that as the other apartment was located in the same building and had one more bathroom and bedroom, you would be more comfortable. As you never complained, we truly believed have made a very convenient change. We should also apologise for our colleague's delay on the check-in day and we'll make sure to verify the heating situation in the apartment. It was a pleasure for us having you as our guest! You will always be welcome.

January 2019

Figura A.1: Exemplo de comentário vs. descrição alojamento para casa inteira

Lisboa

Private room in apartment
2 guests 1 bedroom 1 bed 4 shared baths

The space

We offer you a beautiful double bedroom with good ambience and natural light in one of the best neighbourhoods of Lisbon.

Agustina
June 2017

Its a cheap product for a small amount of money.walls are made of durlock so youcan hear absolutleyeverything. The bed is very bad, get the metal things (Website hidden by Airbnb) back. Bathroom shared with 4 more rooms, so always occupied and not clean.

Response from
Dear Augustina, We have only had very good reviews about this room and its features, so I am surprised regarding your feedback. Furthermore, we are very clear in our advertising that there is one shared bathroom for 4 private bedrooms. All the best,

June 2017

Figura A.2: Exemplo de comentário vs. descrição alojamento para quartos privados

Idiomas[1]	***Counter***
Inglês (en)	11.085
Português (pt)	1.818
Francês (fr)	188
Alemão (de)	40
Espanhol (es)	24
Italiano (it)	10
Russo (ru)	8
Holandês (nl)	2
Chinês (zh-cn)	2
Dinamarquês (da)	1

Tabela A.5: *Top* 10 dos idiomas identificados na variável *description*

Idiomas [2]	***Counter***
Inglês (en)	461.534
Francês (fr)	132.221
Português (pt)	47.722
Espanhol (es)	46.980
Alemão (de)	30.791
Italiano (it)	14.255
Holandês (nl)	8.770
Russo (ru)	4.436
Coreano (ko)	3.367
Chinês (zh-cn)	2.237

Tabela A.6: *Top* 10 dos idiomas identificados na variável *comments*