AirBSet: Um Conjunto de Dados com Imóveis Brasileiros do Airbnb e respectivas Avaliações

Jonatas Freire[1], Luis Henrique Ferreira Costa[2], Carina F. Dorneles[3], Michele A. Brandão[2]

[1]Instituto Federal de Minas Gerais - Campus Sabará, MG - Brasil

[2]Universidade Federal de Santa Catarina - Florianópolis, SC - Brasil

[3]Instituto Federal de Minas Gerais - Campus Ribeirão das Neves, MG - Brasil

jonatasfreire1993@hotmail.com,luishenrique30102005@yahoo.com
carina.dorneles@ufsc.br, michele.brandao@ifmg.edu.br

Abstract. *Airbnb is an online platform with more than 6.6 million listings and 1.4 billion guests from different locations. As this indicates the presence of so many users, this platform generates a large volume of data that various applications can use. Therefore, this work presents AirBSet, a data set with Brazilian properties and respective valuations. This data set is described and characterized to facilitate its use in other studies.*

Resumo. *O Airbnb é uma plataforma online com mais de 6,6 milhões de anúncios e 1,4 bilhões de hóspedes originados de diferentes localidades. Diante de um número tão acentuado de usuários, essa plataforma gera um grande volume de dados que podem ser utilizados em diferentes aplicações. Por isso, este trabalho apresenta AirBSet, um conjunto de dados com imóveis brasileiros e respectivas avaliações. Aqui, esse conjunto de dados é descrito e caracterizado com o intuito de facilitar seu uso em outros estudos.*

1. Introdução

O Airbnb[1] é uma plataforma online, estabelecida em 2008, possui mais de 6,6 milhões de anúncios ativos em mais de 220 países e regiões. Em dezembro de 2022, essa plataforma recebeu mais 1,4 bilhões de hóspedes[2]. Por esse volume de usuários, o Airbnb tem sido alvo de muitos estudos. Por exemplo, [Jain et al. 2021] utilizam diferentes dados dessa plataforma para quantificar e acompanhar mudanças em determinadas regiões. Já [Arefieva et al. 2023] propõem uma nova interface personalizada de recomendação e exploração de destinos que permite aos usuários encontrar seu próximo destino turístico.

Dentre as diversas funcionalidades e recursos presentes no Airbnb, essa plataforma fornece aos proprietários a oportunidade de listar suas propriedades e oferece aos viajantes uma grande seleção de acomodações exclusivas e acessíveis. O impacto transformador do Airbnb na indústria de viagens, na economia compartilhada e na economia global é evidente em sua ampla popularidade e nos vários estudos acadêmicos que examinam sua influência [Ding et al. 2023]. Além do impacto econômico, o Airbnb gera

[1]Airbnb: https://www.airbnb.com.br/

[2]Sobre o Airbnb: https://news.airbnb.com/br/about-us/

Tabela 1. Conjuntos de dados do Airbnb de cidades brasileiras.

Nome	Principais atributos	Cidade
EDA: Airbnb in Brazil	Imóvel, anfitrião, bairro, quantidade mínima de noites para reservar, revisões e disponibilidade	Rio de Janeiro
Inside Airbnb	Imóveis, revisões dos imóveis, calendário e bairro	Rio de Janeiro
[Machado et al. 2022]	Url da acomodação, nome e tipo de acomodação, quantidade de hóspedes, preço diário de aluguel, identificador de um super anfitrião, bairro e a cidade da acomodação, avaliação das acomodações, comentários e lingua utilizada nos comentários	Florianópolis

dados de imóveis, usuários (proprietários e viajantes), cidades, entre outros, que podem ser utilizados para análises de outros impactos gerados pelo uso dessa plataforma, por exemplo, no turismo [Jordan et al. 2023] e na qualidade de vida da vizinhança de imóveis anunciados no Airbnb [Mody et al. 2021].

Nesse contexto, este trabalho apresenta o AirBSet[3], um conjunto de dados com imóveis Brasileiros do Airbnb e respectivas avaliações. Vale destacar que o AirBSet possui imóveis de 10 cidades consideradas empreendedoras e 10 cidades identificadas como turísticas. As cidades São Paulo e Florianópolis estão em ambos os ranqueamentos, por isso, o conjunto de dados proposto possui anúncios de imóveis de 18 cidades diferentes.

Assim, este artigo está organizado conforme segue. A Seção 2 descreve os trabalhos relacionados. A Seção 3 descreve as principais etapas para construção do AirBSet. A Seção 4 caracteriza o AirBSet e a Seção 5 apresenta uma aplicação desse conjunto de dados. Já a Seção 6 descreve as principais limitações e desafios do AirBSet. Finalmente, a Seção 7 detalha as considerações finais.

2. Trabalhos Relacionados

Conforme mencionado anteriormente, o Airbnb tem sido utilizado em diferentes estudos com variadas finalidades, por exemplo, análise de impactos na vizinhança e na economia e busca por destinos turísticos [Arefieva et al. 2023, Ding et al. 2023, Jain et al. 2021, Jordan et al. 2023, Mody et al. 2021]. Para realizar esses estudos, dados distintos foram extraídos do Airbnb, por exemplo, [Arefieva et al. 2023] consideram descrições de 6.851 experiências de um total de 69 destinos europeus mais populares. Já [Jain et al. 2021] consideram dados estruturados (por exemplo, número de imóveis e número de revisões) e não estruturados (por exemplo, comentários sobre os imóveis feitos pelos usuários) das cidades de Nova Iorque, Los Angeles e Londres. Em um estudo diferente, [Jordan et al. 2023] consideram dados de turismo sobre a ilha de São Miguel, no arquipélago dos Açores, e compara com dados de turismo do Airbnb.

Por outro lado, [Ding et al. 2023] focam em analisar as publicações do Scopus[4] para entender a evolução de pesquisas que tem como alvo o Airbnb, ou seja, esse trabalho não utiliza um conjunto de dados extraído da plataforma. No trabalho de [Mody et al. 2021], apesar de ter o Airbnb como foco de estudo, as análises realizadas para entender como essa plataforma impacta na qualidade de vida da vizinhança não utiliza diretamente dados do Airbnb, mas sim um questionário de pesquisa com indivíduos.

[3]AirBSet disponível em `https://zenodo.org/record/8101910`

[4]Scopus: `https://www.scopus.com`

Tabela 2. Top 10 cidades empreendedoras em 2021 e turísticas em 2022.

Top 10 cidades		
Posição	**Empreendedoras**	**Turísticas**
1º	São Paulo - São Paulo	Rio de Janeiro - Rio de Janeiro
2º	Florianópolis - Santa Catarina	São Paulo - São Paulo
3º	Curitiba - Paraná	Gramado - Rio Grande do Sul
4º	Vitória - Espírito Santo	Ubatuba - São Paulo
5º	Belo Horizonte - Minas Gerais	Porto Seguro - Bahia
6º	Porto Alegre - Rio Grande do Sul	Florianópolis - Santa Catarina
7º	São José dos Campos - São Paulo	Fortaleza - Ceará
8º	Osasco - São Paulo	Natal - Rio Grande do Norte
9º	Joinville - Santa Catarina	Porto de Galinhas - Pernambuco
10º	Cuiabá - Mato Grosso	Campos do Jordão - São Paulo

Os resultados revelam que os impactos positivos têm um efeito direto mais forte no apoio dos residentes ao Airbnb e ao turismo em geral.

É importante destacar que foram encontrados poucos trabalhos que disponibilizam dados do Airbnb sobre o Brasil, conforme mostra a Tabela 1. Em particular, foram encontrados dois conjuntos de dados da cidade do Rio de Janeiro, EDA: AirBnb in Brazil[5] e Inside Airbnb[6] e um da cidade de Florianópolis [Machado et al. 2022]. Os três possuem informações muito parecidas sobre os imóveis e revisões/comentários. Uma explicação para isso é que esses são os dados comumente disponibilizados pelo Airbnb. Neste trabalho, também consideramos informações similares a esses três conjuntos de dados, mas o diferencial é termos dados de 18 cidades brasileiras distintas.

3. Metodologia

Esta seção apresenta as principais etapas para a construção do conjunto de dados AirBSet, conforme mostra a Figura 1. Essas etapas consistem na seleção das cidades a terem os anúncios coletados, seleção das datas para coleta dos anúncios, coleta de dados do Airbnb e organização e armazenamento desses dados.

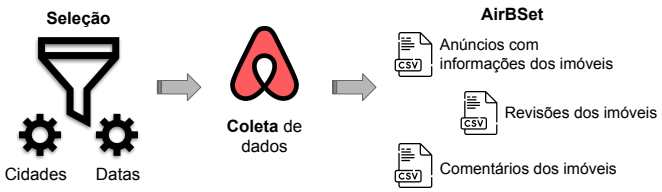


Figura 1. Principais etapas da metodologia para criação do AirbSet.

Seleção das cidades. No Airbnb, há anúncios de imóveis localizados em mais de 100 mil cidades ao redor do mundo[7], o que dificulta uma coleta total desses anúncios. Por isso, neste trabalho, foram selecionados anúncios de imóveis pertencentes a um top 10 cidades brasileiras consideradas empreendedoras e um top 10 cidades brasileiras apontadas como

[5]EDA: AirBnb in Brazil: https://www.kaggle.com/code/leonardorferreira/eda-airbnb-in-brazil

[6]Inside Airbnb: http://insideairbnb.com/get-the-data/

[7]Sobre o Airbnb: https://news.airbnb.com/br/about-us/

Tabela 3. Descrição das informações coletadas sobre os imóveis e respectivos atributos.

Sobre o Imóvel	Atributos
Anúncio	id, url, informação, avaliação, local, preço e comodidades
Revisão	id, limpeza, exatidão do anúncio, comunicação, localização, check-in e custo benefício
Comentário	id, id_quarto, nome, comentário, data e id_usuário

turísticas. A Tabela 2 apresenta o ranqueamento dessas dez cidades brasileiras empreendedoras de acordo com estudo conduzido pelo Enap (Escola Nacional de Administração Pública)[8] em 2021 e turísticas segundo o IBGE (Instituto Brasileiro de Geografia e Estatística)[9] em 2022.

Seleção das datas de coleta. O Airbnb é uma plataforma que possibilita apenas a busca por imóveis selecionando datas posteriores ao dia corrente. Portanto, a coleta foi realizada selecionando feriados do ano de 2023 após o dia 01 de junho de 2023. Assim, os dados dos imóveis foram coletados para dois feriados, de Corpus Christi (de 08 de junho de 2023 até 11 de junho de 2023) e Natal (de 24 de dezembro de 2023 até 25 de dezembro de 2023).

Coleta de dados do Airbnb. Após a seleção das cidades e períodos a serem coletados, utilizou-se o pacote do Python chamado Selenium[10], esse mesmo pacote também foi utilizado para coletar dados em [Silva et al. 2021]. Essa biblioteca automatiza interações em um navegador web e, assim, simula um usuário acessando à plataforma Airbnb e buscando por um imóvel na cidade e período passados como parâmetro. Em seguida, os dados retornados sobre os imóveis e respectivas avaliações são armazenados em arquivos no formato CSV (Comma Separated Values).

Figura 2. Estrutura das revisões nos anúncios de imóveis no Airbnb.

Organização e armazenamento dos dados. Cada arquivo CSV foi guardado em pastas separadas, uma pasta por cidade, ou seja, um total de 18 pastas, e duas pastas para os feriados, uma para cada feriado, dentro da pasta da cidade. Ao todo, o AirBSet possui 108 arquivos CSV. A Tabela 3 apresenta os atributos presentes em cada arquivo coletado sobre os anúncios, revisões e comentários. É importante destacar que as revisões são um agregado de notas nos anúncios, ou seja, são notas calculadas para um imóvel com base nas revisões dos usuários para tal imóvel. A Figura 2 mostra como as revisões são apresentadas nos anúncios e os atributos coletados são referentes a essas características apresentadas na imagem. Já os comentários são por usuário, portanto, um anúncio possui vários comentários.

[8]Muda o ranking de melhores cidades para empreender no Brasil: `bit.ly/43A3WMG`

[9]IBGE confirma atividade turística como importante indutora da economia brasileira: `bit.ly/3Cju7Lq`

[10]Selenium: `https://pypi.org/project/selenium/`

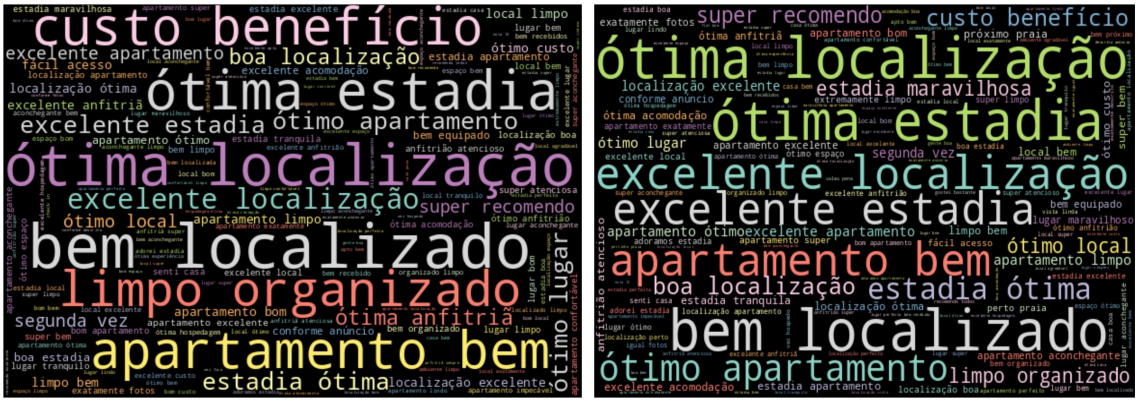
Tabela 4. Estatísticas sobre os anúncios, revisões e comentários presentes no AirBSet.

Tipo	Cidade	# de Anúncios = # de Revisões	# de Comentários
Turística e Empreendedora	São Paulo	344	1639
	Florianópolis	347	1623
	Total	691	3.262
	Média	346	1.631
Empreendedora	Curitiba	349	1.672
	Vitória	314	1.441
	Belo Horizonte	363	1.731
	Porto Alegre	277	1.376
	São José dos Campos	115	517
	Osasco	276	1.315
	Joinville	258	1.291
	Cuiabá	305	1.537
	Total	2.257	10.880
	Média	282	1.360
Turística	Gramado	283	1.289
	Ubatuba	348	1.479
	Porto Seguro	137	580
	Rio de Janeiro	365	1.622
	Fortaleza	341	1.623
	Natal	167	772
	Porto de Galinhas	274	1.206
	Campos do Jordão	251	1.043
	Total	1.825	9.614
	Média	228	1.201
Total Geral	—	4.773	23.756
Média Geral	—	265	1.319

4. Caracterização do AirBSet

Esta seção apresenta uma breve caracterização dos dados presentes no AirBSet. Em particular, a Tabela 4 descreve a quantidade de anúncios, revisões e comentários coletados para cada cidade, agrupados por cidades consideradas turísticas e empreendedoras, apenas empreendedora e apenas turística. As cidades estão ordenadas de acordo com o ranqueamento apresentado na Tabela 2. Como coletamos informações para dois feriados, cuidamos para que anúncios presentes nos dois períodos coletados não fossem contabilizados duas vezes. É importante destacar que a quantidade de anúncios e revisões são iguais pelo fato das revisões serem gerais para cada anúncio.

De maneira geral, na Tabela 4, é possível observar que as cidades consideradas empreendedoras e turísticas (São Paulo e Florianópolis) possuem mais anúncios, revisões e comentários em relação às cidades apenas empreendedoras ou apenas turísticas. Também é possível notar que as cidades empreendedoras juntas possuem cerca de 432 anúncios a mais que as cidades turísticas. Isso pode ser justificado pelo grande porte dessas cidades que acabam sendo turísticas também. Além disso, das cidades empreendedoras, São José dos Campos é a que possui menor quantidade de anúncios e revisões (115) e comentários (517), e Belo Horizonte é a que possui maior quantidade de anúncios e revisões (363) e comentários (1.731). Já para as cidades turísticas, Porto seguro é a que possui menor quantidade de anúncios e revisões (137) e comentários (580), e Rio de Janeiro é a que possui maior quantidade de anúncios e revisões (365), mas possui um comentário (1.622) a menos que Fortaleza (1.623).



(a) Empreendedora (b) Turística

Figura 3. Nuvem de palavras dos comentários das cidades (a) empreendedoras e cidades (b) turísticas.

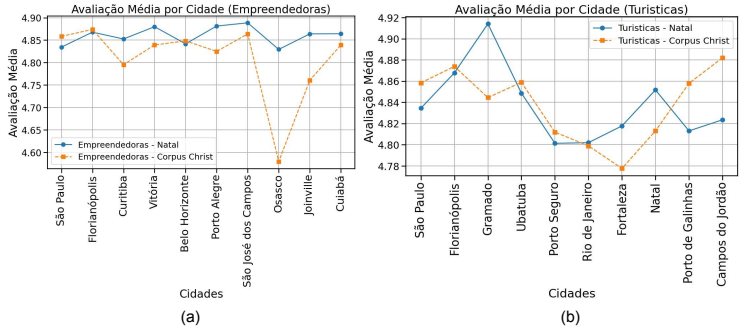


Figura 4. Avaliação média dos usuários do Airbnb por cidade (a) empreendedoras e (b) turísticas.

Além de dados numéricos, o AirBSet possui também dados textuais advindos dos comentários feitos pelos usuários. As Figuras 3(a) e 3(b) apresentam as nuvens de palavras dos 150 termos mais frequentes nos comentários coletados das cidades empreendedoras e turísticas, respectivamente. É possível perceber que em comentários de cidades empreendedoras aparecem com frequência os termos “custo benefício”, “limpo organizado”, “ótima localização”, “bem localizado” e “ótima estadia”. Já em relação às cidades turísticas, aparecem com frequência os termos “ótima localização”, “ótima estadia” e “bem localizado”. Ou seja, é possível perceber que os comentários possuem muitos termos positivos em relação aos imóveis.

De forma complementar, as Figuras 4(a) e 4(b) mostram a média das avaliações recebidas por cidades empreendedoras e turísticas, separadas para os feriados de Corpus Christi e Natal. É possível perceber que os imóveis da cidade de Osasco possuem a menor média de avaliação para as cidades empreendedoras, e Fortaleza é a que possui menor média para as cidades turísticas. Já as Figuras 5(a) e 5(b) mostram o preço médio de cada cidade por feriado separadas por empreendedoras e turísticas, respectivamente. É possível perceber que para as cidades empreendedoras, Osasco possui o maior preço médio por imóvel no feriado de Corpus Christi (aproximadamente R$275,00) e Florianópolis também teve um pico nesse mesmo feriado (por volta de R$215,00). Para Vitória, o pico foi no feriado de Natal, com preço médio de R$200,00. Por outro lado, as cidades turísticas

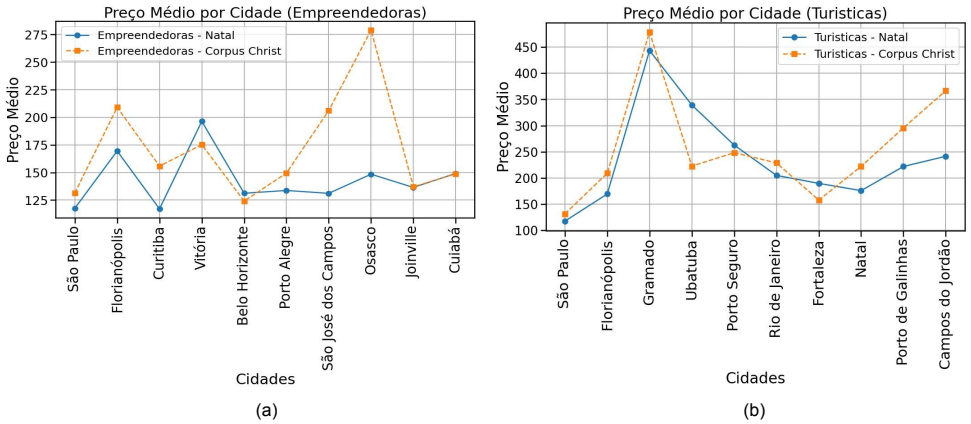


Figura 5. Preço médio dos imóveis do Airbnb por cidade (a) empreendedoras e (b) turísticas.

com preço médio mais elevado são Gramado (preço médio acima de R$450,00 em Corpus Christi e por volta de R$450,00 no Natal), Ubatuba (preço médio por volta de R$350,00 no Natal)e Campos do Jordão (preço médio acima de R$350,00 em Corpus Christi).

5. Aplicações

Uma aplicabilidade direta do AirBSet é comparar os dados de anúncios entre cidades com perfil mais empreendedor, cidades turísticas e, em alguns casos, cidades que apresentam características de ambos. Essa análise permite identificar tendências de aumento nos valores e no número de anúncios durante datas comemorativas ou eventos importantes para cada cidade. Outras possíveis aplicações são:

Análise de sentimento nos comentários. Uma aplicação é verificar se os sentimentos expressos nos comentários estão de acordo com a nota do anúncio. Também identificar características relevantes sobre o imóvel na perspectiva dos usuários.

Análise da relação entre os diferentes atributos coletados. O AirBSet possui informações sobre preços, notas sobre limpeza, comunicação, localização, entre outros. Esses atributos podem ser utilizados em um estudo estatístico para verificar a relação entre eles.

Integração do AirBSet com outros conjuntos de dados. Como o AirBSet inclui informações sobre a localização dos imóveis, é possível cruzar esses dados com outros sobre tal localização. Assim, é possível obter informações, por exemplo, sobre a segurança na vizinhança se for possível cruzar com dados da polícia ou sobre eventos próximos se realizar o cruzamento com dados de turismo.

6. Desafios e Limitações

A API (*Application Programming Interface* ou Interface de Programação de Aplicação, em português) do Airbnb[11] fornece limitação de coleta de apenas 300 anúncios por conta. Por isso, foi necessário desenvolver um coletor, conforme descrito na Seção 3. Entretanto, esse coletor não possibilita a coleta de dados de anúncios em datas retroativas ao dia da coleta. Assim, a coleta de datas retroativas é um desafio que gera uma limitação no

[11]API do Airbnb: `https://www.airbnb.com/partner`

AirBSet de não conter esse tipo de informação. Por exemplo, não foi possível obter dados do período da pandemia de COVID-19 e, assim, não é possível uma comparação entre os períodos de locação no Airbnb antes, durante e após esse acontecimento. Outro desafio é obter informações sobre os usuários do Airbnb, tanto de anunciantes quanto de revisores. Após diversas análises nos dados, foi possível coletar o campo identificador do usuário, mas o conjunto de dados descrito neste trabalho ainda não tem informações mais detalhadas sobre esses usuários.

7. Conclusões

Este trabalho apresentou o AirBSet, um conjunto de dados com imóveis brasileiros e respectivas avaliações de 18 cidades brasileiras, divididas em turísticas e empreendedoras. Esse conjunto de dados foi descrito e caracterizado com o intuito de facilitar seu uso em outros estudos. A análise dos dados revelou que em cidades empreendedoras, muitos comentários estão relacionados à limpeza e organização, já em cidades turísticas, os comentários estão mais relacionados à localização. Já a breve análise das revisões e preços revelam que os preços variam mais de acordo com o feriado do que as revisões. Também descrevemos possíveis aplicações, bem como os desafios e possíveis limitações.

Em trabalhos futuros, planeja-se a coleta de mais períodos, incluindo feriados e dias úteis. Isso possibilitará o uso dos dados para um estudo mais amplo, principalmente, comparando cidades empreendedoras e turísticas. Também planeja-se coletar informações sobre os usuários do Airbnb.

Agradecimentos. Trabalho parcialmente financiado pelo IFMG, CNPq e FAPEMIG.

Referências

Arefieva, V., Egger, R., Schrefl, M., and Schedl, M. (2023). Travel bird: A personalized destination recommender with tourbert and airbnb experiences. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1164–1167.

Ding, K., Niu, Y., and Choo, W. C. (2023). The evolution of airbnb research: A systematic literature review using structural topic modeling. *Heliyon*, page e17090.

Jain, S., Proserpio, D., Quattrone, G., and Quercia, D. (2021). Nowcasting gentrification using airbnb data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21.

Jordan, E. J., Vieira, J. C., Santos, C. M., and Huang, T.-Y. (2023). Do residents differentiate between the impacts of tourism, cruise tourism, and airbnb tourism? *Journal of Sustainable Tourism*, 31(2):265–283.

Machado, A. C. et al. (2022). Análise e correlação de dados: um estudo de caso usando o airbnb e o tripadvisor em florianópolis.

Mody, M., Suess, C., and Dogru, T. (2021). Does airbnb impact non-hosting residents' quality of life? comparing media discourse with empirical evidence. *Tourism Management Perspectives*, 39:100853.

Silva, M. O., Scofield, C., and Moro, M. M. (2021). Pportal: Public domain portuguese-language literature dataset. In *Anais do III Dataset Showcase Workshop - SBBD*, pages 77–88. SBC.