Análise Comparativa de Plataformas de Mídias Sociais Turísticas: um Estudo de Caso de Booking e TripAdvisor para o Município de Santarém, Pará

Marlisson Jean Amaral Aleixo
Universidade Federal do Oeste do Pará
Santarém, Pará, Brasil
marlissonamaral@gmail.com

Gabriele de Sousa Araujo
Universidade Federal do Oeste do Pará
Santarém, Pará, Brasil
gabimitusa@gmail.com

Fernando Almeida do Carmo
Universidade Federal do Oeste do Pará
Santarém, Pará, Brasil
tinhofernando44@gmail.com

Antonio Fernando Lavareda Jacob Junior
Universidade Estadual do Maranhão
São Luís, Maranhão, Brasil
antoniojunior@professor.uema.br

Fábio Manoel França Lobato
Universidade Federal do Oeste do Pará
Santarém, Pará, Brasil
fabio.lobato@ufopa.edu.br

ABSTRACT

The tourism industry is one of the most prominent sectors in analyzing user-generated content, considering the amount of data on social media platforms dedicated to this subject. The analysis of this data can provide insights to better understand customer opinions about companies' products and services, supporting decision-making processes. This work presents a study on the Booking and TripAdvisor platforms. To this end, a comparative analysis of user-generated content on tourism social media was conducted, focusing on hotel data from Santarém, Pará, Brazil, an important tourist city in the Amazon region. The collected data went through the pre-processing stage for adaptation and cleaning, where exploratory analysis, gender analysis, and topic modeling and sentiment analysis techniques were applied. The results show no significant difference in the genres that most comment on TripAdvisor and that both platforms have similar topics. The insights obtained can guide the public sector towards better governance and improving business processes.

KEYWORDS

Turismo, Mídia Social, Modelagem de Tópicos, Booking, TripAdvisor

1 INTRODUÇÃO

De acordo com o Ministério do Turismo do Brasil, a participação do setor turístico na economia nacional representou 8,1% do PIB em 2018[1]. De maneira geral, o turismo é um dos serviços que mais contribuem para o crescimento econômico de um país, colaborando com o desenvolvimento regional e empregabilidade [1]. Isto é ainda mais imperativo em países de dimensões continentais como o Brasil, principalmente, pela grande diversidade em cultura e belezas naturais [1]. Portanto, é de grande importância compreender os pontos fortes e fracos de uma região com base na percepção dos turistas [2].

Neste cenário, a cidade de Santarém, Pará, Brasil, localizada no coração da Floresta Amazônica, merece atenção, uma vez que sua economia é dependente do turismo. Segundo a Secretaria Municipal de Turismo, o setor injetou 176 milhões na economia local em 2018, representando um crescimento de 27% em relação ao ano anterior[2]. As estatísticas apresentadas são de 2018, considerando o cenário pré-pandemia, uma vez que nos anos de 2019 a 2021, ainda sem vacinação plena, o turismo foi um dos setores econômicos mais afetados, com uma queda de 97% no turismo internacional ocasionado pelas restrições impostas sobre viagens [3]. Alter do Chão é um grande exemplo de destino turístico em Santarém, que encanta e impressiona muitos visitantes. A vila de Alter do Chão recebeu o prêmio da categoria "Melhor destino turístico nacional" em 2021, premiação realizada pela UPIS de turismo[3], o que certamente influencia na economia regional.

Historicamente, o turismo sofreu mudanças significativas com avanço da internet, principalmente na forma como as empresas vendem e entregam seus produtos e serviços [4]. Desta forma, as Tecnologias de Informação e Comunicação surgem e impulsionam o crescimento das áreas turísticas, mudando a forma como os turistas se relacionam com estes serviços [5]. Consequentemente, as mídias sociais surgem como uma nova forma de comunicação on-line entre usuários, permitindo que os clientes troquem informações, propiciando o chamado boca a boca eletrônico [6].

Em sua maioria, consumidores procuram e compartilham informações e opiniões sobre determinado produto, serviço ou destino. Assim, as mídias sociais, como plataformas colaborativas, tornam os seus usuários, além de consumidores, geradores de conteúdo, dos quais outras pessoas tomam como base para escolher um produto ou serviço [7]. Ainda, essas mídias influenciam positivamente na fase de pré-compra, auxiliando o cliente no processo de tomada de decisão da compra [8]. Nesse panorama, plataformas on-line do setor de turismo são grandes candidatas a serem explorados por partes interessadas, uma vez que fornecem uma quantidade considerável

[1] https://www.gov.br/turismo/pt-br/assuntos/noticias/cresce-a-participacao-do-turismo-no-pib-nacional

[2] https://g1.globo.com/pa/santarem-regiao/noticia/2019/02/11/turismo-em-santarem-cresce-em-2018-e-injeta-r-176-milhoes-na-economia-aponta-estudo.ghtml

[3] https://santarem.pa.gov.br/noticias/turismo/alter-do-chao-e-eleita-o-melhor-destino-turistico-nacional-1vit8b

de opiniões de usuários sobre experiências em hotéis, restaurantes e atrações turísticas [9]. Estas aplicações são extremamente importantes para a publicidade das empresas, pois influenciam diretamente na sua imagem [10].

Entretanto, as empresas ligadas ao turismo não conseguem realizar a análise manual dessas mídias sociais devido ao grande volume de dados disponíveis. Neste sentido, técnicas de mineração de texto têm sido amplamente utilizadas para obter *insights*. Tais técnicas tornam possível transformar os dados coletados destas plataformas em conhecimento por meio do reconhecimento de padrões em conjuntos de dados [6, 11].

Para esse estudo, as plataformas Booking e TripAdvisor foram consideradas, devido à sua relevância para o estudo de caso (Alter do Chão). Além disso, o uso de várias plataformas com a perspectiva de diferentes públicos nos permite comparar dados distintos e validar os resultados obtidos, o que é encorajado por [12]. Dado o contexto e os conceitos apresentados, este trabalho visa realizar uma análise comparativa dos dados gerados pelos usuários nas mídias sociais de turismo com foco nos dados hoteleiros da cidade de Santarém. Como impacto do estudo, têm-se a compreensão da percepção dos turistas, identificando falhas e potencialidades, o que agrega valor às empresas do setor e traz vantagens competitivas, além de ter o potencial de auxiliar na construção de políticas públicas efetivas.

O restante deste artigo está organizado da seguinte forma. Trabalhos relacionados são discutidos na seção 2. Na Seção 3 são descritos os Materiais e Métodos usados nos experimentos. Os resultados são discutidos na Seção 4. Por fim, as conclusões e sugestões de trabalhos futuros são apresentadas na Seção 5.

2 TRABALHOS RELACIONADOS

Diversos trabalhos na literatura investigam o processo de tomada de decisão de clientes por meio de relatos de experiências de outros consumidores em mídias sociais. [13] mostra que o TripAdvisor exibe informações detalhadas e dados que podem ser usados no planejamento de viagens. Contudo, poucos hotéis estão gerenciando ativamente sua reputação no site. Em [14], é utilizada uma classificação baseada em *Naïve Bayes* para identificar polaridades em comentários de consumidores. De maneira similar, [15] utiliza uma abordagem de modelagem de tópicos em conjunto com *Naïve Bayes*, para indicar a probabilidade da presença de cada termo em comentários positivos e negativos. É importante destacar que esta abordagem está bem estabelecida no estado da prática. No entanto, apenas o sentimento/polaridade é insuficiente para fornecer *insights* informativos a fim de apoiar o processo de tomada de decisão.

Em [16] buscou-se compreender o processo de tomada de decisão em duas plataformas de reclamações on-line por meio de técnicas de mineração de textos. É importante ressaltar que a análise sentimentos não se aplica a reclamações, pois o sentimento é negativo *per se*. A análise realizada pelos autores permitiu identificar as principais divergências entre os grupos de usuários, tipos de soluções, confiabilidade da empresa e principais tópicos de reclamações. Os autores também usaram métodos para medir a legibilidade da reclamação, modelagem de tópicos com *Latent Dirichlet Allocation (LDA)* e análise exploratória de dados.

Estudos investigam os fatores-chaves e critérios de avaliações on-line que afetam a escolha de hotéis por parte do viajante. Em [17] é empregado o algoritmo de *Word2vec* para identificar padrões na seleção de hotéis entre cinco tipos de viajantes na plataforma TripAdvisor. De fato, esse processo é fundamental para identificar as semelhanças entre os diferentes públicos que acessam as plataformas de turismo. Apesar da relevância do tema, há uma escassez de pesquisas direcionadas à análise comparativa entre essas diferentes personas. Além disso, poucos estudos expandem essas análises para mais de uma plataforma.

[18] analisou os padrões extraídos de avaliações de restaurantes na plataforma TripAdvisor. Por meio de técnicas de mineração de texto e modelagem de tópicos, foi possível identificar quais gêneros de clientes possui as críticas mais negativas, os padrões de comentários e o relacionamento entre os tópicos identificados. A análise de sentimento foi realizada utilizando a biblioteca *Polyglot* em Python. Os autores usaram a *Non Negative Matrix Factorization* (NMF) para modelagem de tópicos, dada a sua eficiência para tarefas de mineração de texto em documentos curtos - conforme discutido em [19]. De forma semelhante, o trabalho de [20] utilizou a biblioteca Python TextBlob para a rotulação automática de sua base de dados a partir da pontuação de polaridade obtida. Os dados rotulados foram divididos em dois subconjuntos, treino e teste, para que fosse feita a avaliação de desempenho de 4 algoritmos de classificação, posterior utilização na criação de um *framework* para análise de sentimentos. Os autores questionam e analisam a rotulação automática, devido a classificações errôneas. Ainda assim, a maioria dos dados foram classificados corretamente. É de se considerar que a partir da rotulação automática com bibliotecas como TextBlob e Polyglot pode-se obter resultados satisfatórios. Outros trabalhos utilizam técnicas baseadas em modelagem de tópicos para identificar os principais assuntos presentes nos comentários.

Em [21], os autores abordam a relação entre os gêneros dos consumidores e os principais tópicos presentes nos comentários. Com isso, foi possível obter os principais tópicos de avaliação, como atendimento, localização, cardápio e precificação, bem como os tópicos de comentários por gêneros. Considerar o gênero é bastante interessante para os tomadores de decisão, no entanto, as plataformas de turismo não disponibilizam esta informação. [21] fez uma anotação manual seguindo a estratégia *human-in-the-loop* [22]. No trabalho de [2], foi feita uma análise nas tendências de tópicos ao longo dos anos, permitindo identificar aspectos crescentes ou decrescentes no interesse dos clientes. Esse tipo de análise é útil para os tomadores de decisão avaliarem a evolução do mercado, bem como detectar tendências; no entanto, requer uma grande quantidade de dados distribuído ao longo do tempo.

Maneiras de como dar suporte às marcas no gerenciamento de comentários negativos em plataformas on-line vêm sendo abordadas na literatura. Em [23], os autores apresentam uma aplicação que utiliza técnicas baseadas em *Explainable Sentiment Analysis*, que pode ser utilizada por analistas de marketing no entendimento das atitudes dos clientes em relação às marcas. O objetivo do trabalho foi fornecer a essas partes interessadas a compreensão do por que determinado sentimento é previsto, o que raramente é explorado por plataformas que enfrentam crises de mídias sociais. Esse suporte pode ser eficaz ao fornecer as palavras mais importantes que abordam o sentimento dos clientes em comentários individuais.

Em geral, muitos autores utilizam técnicas de mineração de texto para identificar padrões e, até mesmo, usar as classificações para

sistemas de recomendações, como proposto em [24]. O foco desta pesquisa está em auxiliar e agregar valor a partir do conhecimento extraído, visando que as instituições interessadas melhorem seus produtos e serviços.

3 MATERIAIS E MÉTODOS

O *framework* adotado neste estudo foi inspirado em [25], que realizou um estudo em comentários de notícias sobre COVID-19. O fluxo de análise é apresentado na Figura 1.

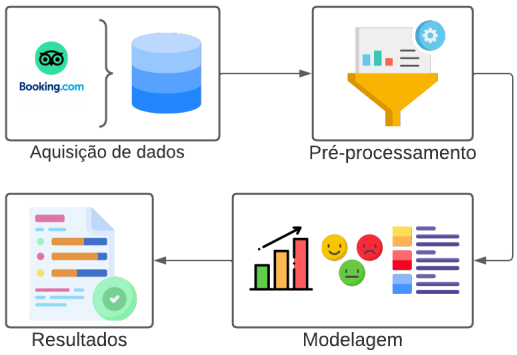


Figura 1: Etapas do *framework* experimental adotado no estudo.

Conforme disposto na Figura 1, o *framework* experimental é composto por quatro fases, a saber: i) Aquisição e organização dos dados; ii) Preparação e pré-processamento; iii) Modelagem e validação dos resultados; iv) e Interpretação dos resultados. As subseções a seguir apresentam cada etapa em detalhes.

3.1 Aquisição e características dos dados

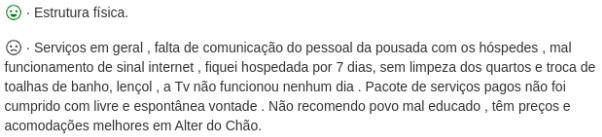
Dois *Web Crawlers* foram desenvolvidos na linguagem Python para coletar dados das plataformas sob escrutínio. As ferramentas *Selenium* e *Beautifulsoup 4* foram escolhidas devido a familiaridade dos autores com elas. Além disso, estas ferramentas estão entre as mais utilizadas. Destarte, possuem documentação atual e forte suporte da comunidade de desenvolvedores, tornando o processo de desenvolvimento mais célere e eficiente. Como mencionado anteriormente, o Booking e o TripAdvisor foram escolhidos por sua popularidade e grande número de consumidores on-line [2, 21, 26]. Ademais, de acordo com a plataforma Similar Web[4], eles estão no *ranking* dos dez sites mais visitados na categoria de viagens e turismo no Brasil e no mundo. Os atributos coletados e seus respectivos tipos de dados são apresentados na Tabela 1. Alguns itens foram removidos de ambas as plataformas, de modo a dispor apenas de dados semelhantes.

Cada plataforma analisada possui suas características, como por exemplo, a forma de dar nota ao hotel e expressar, por meio de uma avaliação escrita, sua experiência. O Booking, conforme pode ser visto na Figura 2, permite ao usuário produzir as avaliações escritas e informar se foram positivas ou negativas. A nota dada pelos usuários aos hotéis da plataforma Booking, podem variar entre 1 e 10, contemplando casas decimais (e.g., 8,4). A disponibilidade das métricas de avaliações no Booking facilita no processo de utilização das técnicas de Análise de Sentimento e favorece a análise do processo de Modelagem de Tópicos.

[4]https://www.similarweb.com/pt/top-websites/brazil/category/travel-and-tourism/

Tabela 1: Dados coletados da plataforma TripAdvisor e Booking.

Atributos	Tipo de dado
Nome do hotel	Textual
Nota do hotel	Numérico
Titulo do comentário	Textual
Comentário	Textual
Nota de usuário	Numérico
Data da hospedagem	Numérico
Tipo de Viagem	Textual


· Estrutura física.

· Serviços em geral , falta de comunicação do pessoal da pousada com os hóspedes , mal funcionamento de sinal internet , fiquei hospedada por 7 dias, sem limpeza dos quartos e troca de toalhas de banho, lençol , a Tv não funcionou nenhum dia . Pacote de serviços pagos não foi cumprido com livre e espontânea vontade . Não recomendo povo mal educado , têm preços e acomodações melhores em Alter do Chão.

Figura 2: Exemplo de comentário no Booking.

De maneira distinta, no TripAdvisor as notas dadas por usuários podem variar entre 1 e 5 e lhe é permitido realizar uma única avaliação geral.

3.2 Pré-processamento

A etapa de pré-processamento é uma das mais importantes no processo de análise de dados, sobre tudo quando se trata de dados textuais. Nessa etapa foi realizada a limpeza e padronização das avaliações, que consistem em remover ruídos/inconsistências, que possam prejudicar a confiabilidade das análises subsequentes [27]. A primeira tarefa consistiu na tradução de todas as avaliações para o português brasileiro, utilizando a biblioteca *Python-translate*, visto que o processo de coleta considerou avaliações de todos os idiomas disponíveis na plataforma. Convém destacar que aproximadamente 90% do comentários já estavam escritos na língua portuguesa. Posteriormente, foi feita a remoção de duplicatas que eventualmente ocorreram no processo de aquisição de dados.

Os passos seguintes incluíram conversão para caixa baixa e as remoções de: i) pontuações, caracteres especias e acentuação(quando necessários); ii) números; iii) emojis; e iv) *stopwords* (palavras que possuem pouco valor semântico). Por fim, foi feita a tokenização (transformação do texto em lista de palavras) [27]. A Tabela 5 apresenta um exemplo dos passos utilizados no pré-processamento.

3.3 Sumário de dados

A Tabela 3 apresenta um resumo dos principais atributos das bases coletadas. A contagem dos atributos foi realizada após a etapa de remoção de duplicas. A média de *tokens* para o Booking, diz respeito a média para cada tipo de comentário, respectivamente positivo e negativo.

Tabela 2: Exemplo das etapas de pré-processamento.

Método	Saída
Original	Solicitei as filmagens e me falaram que não viram nada. Mistério...😤
Conversão para caixa baixa	solicitei as filmagens e me falaram que não viram nada. mistério...😤
Remoção de pontuação, acentuação e caracteres especiais	solicitei as filmagens e me falaram que viram nada misterio 😤
Remoção de números e emojis	solicitei as filmagens e me falaram que nao viram nada misterio
Remoção de *stopwords*	solicitei filmagens falaram viram misterio

Tabela 3: Sumário dos dados obtidos nas plataformas

Atributo	TripAdvisor	Booking
Total de hotéis	55	17
Total de usuários	2537	1184
Total de avaliações	2.973	2475
Média de tokens por comentário	72	13/14

3.4 Classificação de Gênero

A literatura e o estado da prática aponta como salutar a análise considerando gênero a fim de entender como os serviços de turismo influenciam o comportamento do consumidor, uma vez que cada gênero observa o ambiente e dá significado ao turismo em diferentes perspectivas [28].

Uma vez que as plataformas não disponibilizam tal informação, foi realizada a classificação de gênero de forma manual pelos autores. Para isso, foi considerado o nome do autor do comentário, sendo que cada um foi classificado como masculino, feminino ou desconhecido. Mesmo que não haja uma norma para a geração de nomes próprios masculinos e femininos, ainda assim é seguido alguns preceitos para nomes comuns a partir do gênero gramatical [29].

Para situações onde não foi possível definir o gênero por meio do nome, foram conduzidas verificações no perfil dos usuários. Neste ponto utilizaram-se características que fosse possível identificar seu gênero, caso contrário era classificado como desconhecido [18, 21]. Dessa maneira, um usuário com nome "Luciano" seria classificado Masculino, "Denise" como Feminino e "Gmz" como Desconhecido.

3.5 Análise de Sentimentos

A análise de sentimentos é um subcampo da Mineração de Texto que analisa a opinião que as pessoas expressam sobre entidades, serviços ou produtos, buscando identificar a polaridade expressa, por exemplo, positivo, negativo ou netro [30]. No presente trabalho a biblioteca Polyglot foi utilizada, tal como em [21]. O Polyglot utiliza dicionários léxico para sua classificação e cobre 136 idiomas incluindo o português brasileiro. Por ser tratar de uma biblioteca de classificação automática, o *ground truth* foi feito em algumas avaliações selecionadas aleatoriamente para que fosse possível validar os resultados da classificação de sentimentos [31].

Como mencionado anteriormente, os dados utilizados para aplicação do processo de análise de sentimento não passaram pela etapa de remoção de acentuação no pré-processamento. Além disso, não foi aplicado a análise nos dados da plataforma Booking, pois é entendido que a plataforma permite ao usuário fazer diferentes avaliações, positiva e negativa.

3.6 Modelagem de tópicos

Para determinar os tópicos mais aparentes (ou seja, os principais aspectos dos comentários coletados), algoritmos não supervisionados aplicados em textos foram utilizados [32, 33]. O processo de modelagem foi conduzido em três etapas. A primeira consistiu na extração dos tópicos, a qual foi seguida pela classificação manual feita pelos pesquisadores, uma vez que a interpretação dos tópicos ainda depende do parecer humano para relacionar as palavras e seus significados [34].

Para a extração de tópicos foi utilizado o algoritmo de modelagem não supervisionado *Latent Dirichlet Allocation* (LDA), sendo este um dos mais utilizados para esse tipo de análise conforme [35–37]. Testamos diferentes configurações (número de tópicos e número de palavras por tópico) para alcançar o resultado mais satisfatório, que foi analisado qualitativamente pelos pesquisadores. O método consistiu em utilizar o modelo de representação *Bag-of-Words*, como entrada do algoritmo LDA, e o *Term-Frequency Inverse Document Frequency* (TF-IDF) como esquema de pesos para a contagem e filtragem dos termos da base de dados [38].

A anotação foi realizada por dois analistas experientes com o setor turístico e com avaliação de modelagem de tópicos. Os rótulos foram discutidos até chegar a um consenso, um método também usado por [39].

4 RESULTADOS

No total, 3.945 revisões foram coletadas da plataforma TripAdvisor e 4.220 da Booking. Após a etapa de pré-processamento, o total de avaliações resultou em 2.973 e 2.475, respectivamente. Uma análise exploratória das avaliações dos usuários foi realizada para ambas as plataformas e são mostradas nas Figuras 3 e 4.

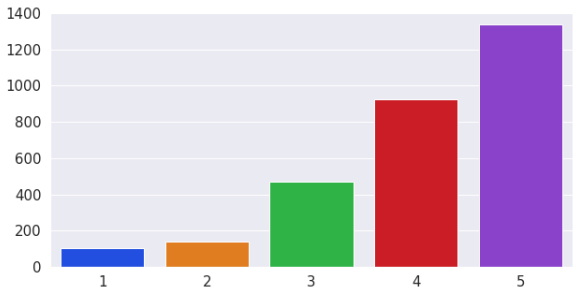


Figura 3: Notas de usuários do TripAdvisor.

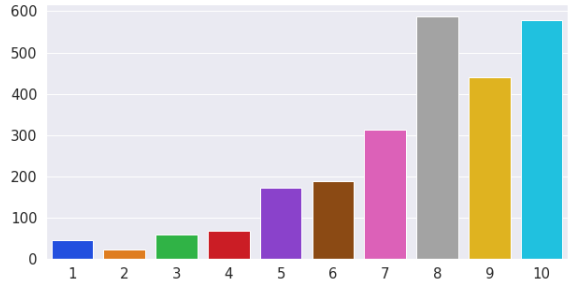


Figura 4: Notas de usuários do Booking.

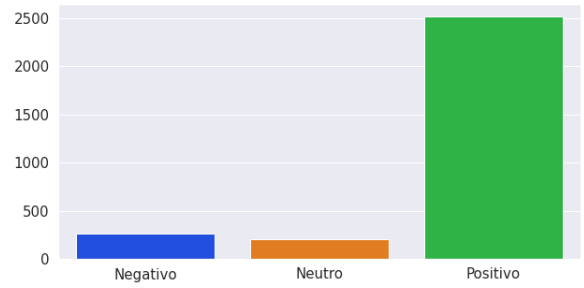


Figura 5: Quantidade de comentários por sentimento da plataforma TripAdvisor.

Analisando as Figuras 3 e 4 é possível perceber que cada plataforma tem seu próprio parâmetro de notas de avaliação, conforme discutido anteriormente. Destaca-se que a classificação por nota dada por usuários no TripAdvisor varia de 1 à 5, e no Booking de 1 à 10. Para as notas do Booking, aplicamos o arredondamento em razão aos valores decimais permitidos pela plataforma. Se considerarmos as notas 4 e 5 no TripAdvisor como exclusivamente positivas, observamos a discrepância em relação à distribuição das demais e notamos a tendência a comentários positivos. Podemos considerar o mesmo para as notas do Booking, considerando 8, 9 e 10 como indiscutivelmente positivas. Ambas as plataformas tendem a ter mais comentários positivos.

No caso do TripAdvisor, a quantidade de comentários classificados com nota 4 e 5 representam 76% do total de ocorrências. Podemos considerar o mesmo para as notas 8, 9 e 10 do Booking, as quais representam aproximadamente 64% das ocorrências. Parte dessa informação já se encontra disponível nas plataformas e são consideradas relevantes, principalmente, para os consumidores no processo de tomada de decisão, contudo para os gestores somente isso não é suficiente [21].

Comparamos a distribuição da classificação dos usuários com a análise dos sentimentos. A Figura 5 mostra a predominância de comentários positivos no TripAdvisor. Como o Booking permite que seus usuários expressem em duas revisões os pontos positivos e negativos de sua experiência, a aplicação da ferramenta de análise de sentimentos foi desnecessária. Entretanto, foi possível comparar a quantidade de comentários positivos e negativos coletados da plataforma.

A diferença no valor do Booking não é tão significativa quanto ao TripAdvisor; ainda assim deve ser considerada, pois a plataforma induz seus usuários a pensar nos aspectos negativos da acomodação para avaliar. Neste cenário, é notável a discrepância no número de revisões positivas em relação às demais, corroborando o resultado apresentado nas Figuras 3 e 4, o que torna perceptível que ambas as plataformas analisadas tendem a ter comentários positivos ou são utilizadas principalmente para expressar opiniões positivas sobre hotéis e seus serviços em geral.

Em relação a gênero, as Figuras 6 e 7 apresentam a quantidade de comentários classificados como masculino, feminino e desconhecido do TripAdvisor e Booking, respectivamente. Conforme explanado na seção anterior, foram classificados como desconhecidos quando o nome do usuário em questão não tornava possível sua identificação, como por exemplo , “macpilot78”,“gdammski” e “TrailBlazer634730”. Os resultados sugerem que não há diferença significativa quanto ao gênero, em relação a quem mais realiza comentários na plataforma TripAdvisor, de maneira contraria ao Booking. Essa análise se faz útil para que os gestores do hotéis possam traçar estratégias que atendam seu principal publico alvo.

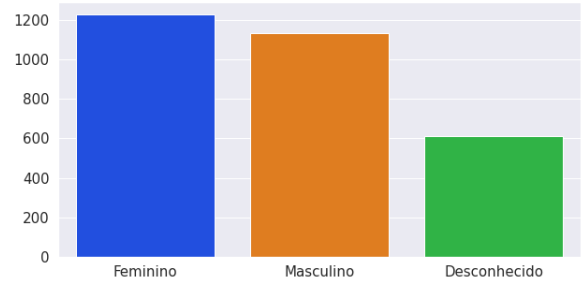


Figura 6: Quantidade de comentários do TripAdvisor por gênero.

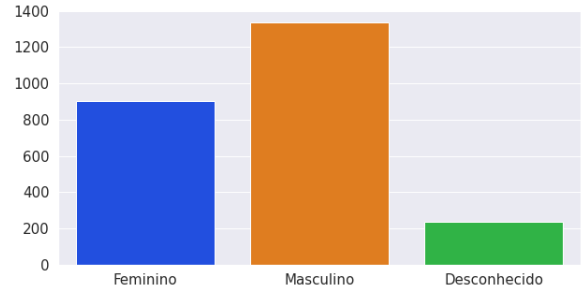


Figura 7: Quantidade de comentários do Booking por gênero.

Também, avaliamos a distribuição temporal dos comentários para ambas as plataformas (Figuras 8 e 9). Com relação aos dados

do TripAdvisor (Figura 8), podemos observar um aumento quase contínuo no número de comentários de julho a novembro e um aumento isolado em janeiro. Essa tendência, também, é vista na plataforma Booking (Figura 9). A forte semelhança entre as duas distribuições corrobora demonstra que os usuários tendem a fazer os *reviews* logo após a estadia - o pico nos meses reflete os meses de maior fluxo turístico na região.

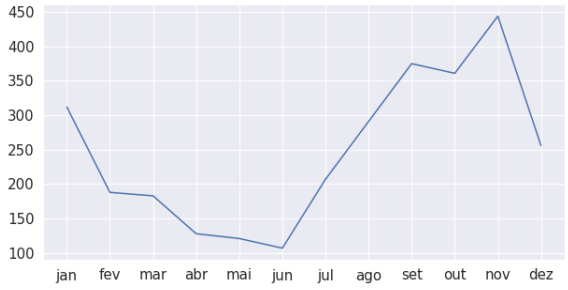


Figura 8: Quantidade de comentários por mês no TripAdvisor.

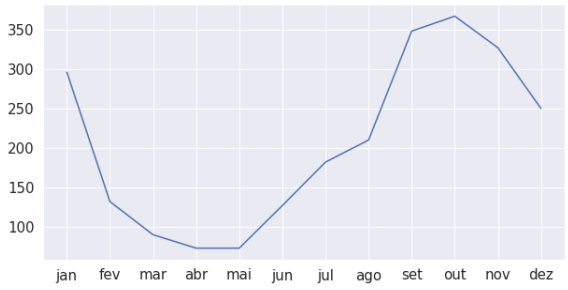


Figura 9: Quantidade de comentários por mês no Booking.

A festa regional conhecida com Sairé ocorre anualmente no mês de setembro na vila de Alter do Chão[5], o que pode justificar ser o segundo maior mês em quantidade de avaliações em ambas as plataformas. Outro ponto que pode explicar esse aumento é o período de secas na região, onde as praias estão mais aparentes, e que atrai muitos turistas que gostam de aproveitá-la em sua melhor forma[6].

Quanto aos experimentos conduzidos para a modelagem de tópicos, optou-se por utilizar 4 tópicos e 8 termos. Neste cenário, as Tabelas 4 e 5 apresentam os tópicos mais aparentes das duas plataformas pesquisadas.

Os principais tópicos identificados nos comentários da plataforma TripAdvisor foram "café da manhã", "alter do chão", "localização", e "quarto", o que leva a considerar atenção dos clientes

[5]https://g1.globo.com/pa/santarem-regiao/noticia/2021/03/18/semc-e-liderancas-de-alter-do-chao-definem-datas-de-realizacao-do-saire-de-2021-a-2024.ghtml
[6]https://g1.globo.com/turismo-e-viagem/descubra-o-brasil/noticia/2021/01/04/alter-do-chao-no-para-tem-apenas-duas-estacoes-no-ano-e-pode-oferecer-praia-ou-floresta-alagada-a-depender-da-epoca.ghtml

Tabela 4: Tópicos da plataforma TripAdvisor.

Tópico	Termos
Café da manhã	café manhã quarto bom simples sucos frutas cama
Alter do Chão	pousada alter café manhã passeios lugares super chão
Localização	café bom manhã quartos localização atendimento cidade estadia
Quarto	quarto manhã ar condicionado café dia chão banheiro

Tabela 5: Tópicos da plataforma Booking.

Tópico	Termos
Preço	valor booking reserva diária pagamento lanche estadia pagar
Quarto	banheiro quarto chuveiro água quente ar condicionado cheiro
Vista	pousada praia ronaldo rio tapajos vista estadia alter
Café da manhã	café manhã quarto localização bom atendimento cama limpeza

para os adjetivos como "bom", "simples" e "super" a respeito do café da manhã e da localização do hotel. Para a plataforma Booking os tópicos identificados foram: "preço", "quarto", "vista", e "café da manhã". Podemos destacar o tópico café da manhã como um dos mais importantes por apresentar-se em ambas as plataformas. Também, é possível notar a presença de termos iguais em grande parte dos tópicos, como: "atendimento", "ar", "condicionado", "limpeza", "quente" e "quarto", dando a entender que existe uma relação entre eles e aparentam ser termos com forte presença nas avaliações dos usuários, que refletem ao nível de serviço do hotel.

5 CONCLUSÕES

O setor turístico está intrinsicamente ligado às mídias sociais. A mineração de opinião nos permite realizar a modelagem comportamental de usuários, evidenciando as necessidades dos clientes e segmentando o mercado. Existem diversas plataformas de mídias sociais dedicadas ao turismo, que traz à tona dois desafios: um grande volume de conteúdo gerado pelo usuário, que é desestruturado por natureza, e a multicanalidade. Visando enfrentar esses desafios, apresentamos neste trabalho uma análise comparativa de avaliações de hotéis nas plataformas Booking e TripAdvisor. Usamos um destino turístico brasileiro em constante crescimento, Alter do Chão, localizado no coração da Floresta amazônica.

Atento ao estado da prática e ao estado da arte, aplicamos técnicas de mineração de texto com Modelagem de Tópicos e Análise de sentimento em conjunto com a análise exploratória de dados. Os resultados obtidos nos permitem concluir que i) a comparação de plataformas gera *insights* para instituições de hospitalidade, ii) é possível compreender as percepções dos turistas, e iii) as análises agregam valor às instituições para impulsionar este setor. Os

insights obtidos podem orientar o setor público em melhor governança, bem como melhorar o processo das empresas.

Os *insights* gerados por esta pesquisa contribuíram para o estudo do gerenciamento de Pequenas e médias empresas (PMEs), agregando no entendimento de como os consumidores respondem aos serviços das PMEs. Com isso, concluiu-se que há a escassez de habilidades de gestão e P&D (Pesquisa e Desenvolvimento) nas PMEs, em se adequarem à evolução tecnológica e aproveitarem os recursos fornecidos pelo meio digital para agregar valor à marca e fidelizar seus clientes.

Como desdobramento direto desta pesquisa, destaca-se que foi desenvolvido uma oficina com o intuito de melhorar o letramento digital e comunicacional das PEMs localizadas na vila de Alter do Chão (Santarém, Pará), cidade usada como estudo de caso deste trabalho e é destacado como um centro turístico em Santarém. Portanto, esse projeto gera impacto social e econômico na região por meio da inclusão digital, assim, estimulando a atividade turística e impulsionando o desenvolvimento local e regional.

Nosso estudo possui algumas limitações. Em relação a análise computacional, utilizamos uma abordagem *Term-Frequency Inverse Document Frequency* (TF-IDF), contudo pretendemos investigar uma abordagem de representação *word-embeddings*. Além disso, usamos apenas LDA para modelagem de tópicos. A partir desta perspectiva, em trabalhos futuros, pretendemos incorporar *word-embeddings* para analisar os dados em um nível semântico. Ainda, gostaríamos de testar outros algoritmos de modelagem de tópicos, como *Non Negative Matrix Factorization* (NMF) e *Latent Semantic Analysis* (LSA), também implementar métodos explicativos para análise de sentimento baseada em aspectos. Passando para a avaliação qualitativa, gostaríamos de estender nossa validação para proprietários/associações de hotéis e colocar a Secretaria de Turismo na linha. Dessa forma, com intuito de avançar nas análises e resultados de dados gerados pelos usuários, pretende-se incluir novas fontes de dados ampliando ainda mais o escopo de plataformas abordadas. Ademais, se torna interessante para o estudo, analisar os diferentes tipos de perfis de usuários que comentam e identificar relações entre perfis que atribuem notas baixas/altas ou expressam elogios/críticas. Por fim, gostaríamos de agregar dados de redes sociais como Instagram e Twitter.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - DT - 308334/2020; pela Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA) - PRONEM-FAPESPA/CNPq nº 045/2021. Agradecemos também aos revisores(as) pelas sugestões que muito auxiliaram na melhora do trabalho.

REFERÊNCIAS

[1] Carla Regina Ferreira Freire Guimarães and Cauê Bomfim Morano. Revisão sistemática de trabalhos acadêmicos sobre turismo e emprego no brasil, entre os anos de 2010-2020. *RITUR-Revista Iberoamericana de Turismo*, 10(2):123–135, 2020.

[2] Carla Marcolin, JoÃ Luiz Becker, Fridolin Wild, Giovana Schiavi, and Ariel Behr. Business Analytics in Tourism: Uncovering Knowledge from Crowds. *BAR - Brazilian Administration Review*, 16, 00 2019. ISSN 1807-7692.

[3] United Nations World Tourism Organization UNWTO. New data shows impact of covid-19 on tourism as unwto calls for responsible restart of the sector, 6 2020.

[4] Mabel Simone de Araújo Bezerra Guardia, Marília Barbosa Gonçalves, and Sergio Ramiro Guardia. As mídias sociais no marketing turístico: Um estudo sobre seu uso na promoção do roteiro seridó. 2012.

[5] André Riani Costa Perinotto, Janaina Cavalcante Farias Camarço, Solano De Souza Braga, and Marina Furtado Gonçalves. Perceptions on services in ceará-brazil luxury hotels registered on tripadvisor. *Journal of Global Scholars of Marketing Science*, 0(0):1–20, 2021.

[6] Fábio Lobato, Marcia Pinheiro, Antonio Jacob Junior, Olaf Reinhold, and Ádamo Santana. Social crm: Biggest challenges to make it work in the real world. pages 221–232, 01 2017. ISBN 978-3-319-52463-4.

[7] Austin Rong-Da Liang. Consumers as co-creators in community-based tourism experience: Impacts on their motivation and satisfaction. *Cogent Business & Management*, 9(1):2034389, 2022. doi: 10.1080/23311975.2022.2034389. URL https://doi.org/10.1080/23311975.2022.2034389.

[8] Sujin Song and Myongjee Yoo. The role of social media during the pre-purchasing stage. *Journal of Hospitality and Tourism Technology*, 7:84–99, 2016.

[9] Ana Valdivia, Maria Luzon, and Francisco Herrera. Sentiment analysis in tripadvisor. *IEEE Intelligent Systems*, 32:72–77, 01 2017.

[10] Maria Teresa Borges-Tiago, Carolina Arruda, Flavio Tiago, and Paulo Rita. Differences between tripadvisor and booking. com in branding co-creation. *Journal of Business Research*, 123:380–388, 2021.

[11] Wendel Silva, Ádamo Santana, Fábio Lobato, and Márcia Pinheiro. A methodology for community detection in twitter. In *Proceedings of the International Conference on Web Intelligence*, pages 1006–1009, 2017.

[12] Gustavo Almeida, Isabelle Guimarães, Antonio Jacob Jr., and Fábio Lobato. Fontes de dados gerados por usuários: quais plataformas considerar? In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 25–36, Porto Alegre, RS, Brasil, 2020. SBC.

[13] Peter O'Connor. Managing a hotel's image on tripadvisor. *Journal of Hospitality Marketing & Management*, 19:754–772, 10 2010.

[14] Rachmawan Adi Laksono, Kelly Rossa Sungkono, Riyanarto Sarno, and Cahyaningtyas Sekar Wahyuni. Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes. In *2019 12th International Conference on Information Communication Technology and System (ICTS)*, pages 49–54, 2019.

[15] Viriya Taecharungroj and Boonyanit Mathayomchan. Analysing tripadvisor reviews of tourist attractions in phuket, thailand. *Tourism Management*, 75: 550–568, 2019.

[16] Gustavo de Sousa, Isabelle Guimarães, Antonio Jacob Jr, and Fábio Lobato. Análise comparativa das principais plataformas de reclamações online: implicações para análise de mídia social em negócios. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 154–165, Porto Alegre, RS, Brasil, 2020. SBC.

[17] Le Wang, Xiao kang Wang, Juan juan Peng, and Jian qiang Wang. The differences in hotel selection among various types of travellers: A comparative analysis with a useful bounded rationality behavioural decision support model. *Tourism Management*, 76:103961, 2020. ISSN 0261-5177. doi: https://doi.org/10.1016/j.tourman.2019.103961. URL https://www.sciencedirect.com/science/article/pii/S0261517719301591.

[18] Luiz Carlos Fernandes, Jorge Silva, Antonio Jacob, and Fábio Lobato. An extensive analysis of online restaurant reviews: a case study of the amazonian culinary tourism. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, 2020.

[19] Yong Chen, Hui Zhang, Rui Liu, Zhiwen Ye, and Jianying Lin. Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems*, 163:1–13, 2019.

[20] Kudakwashe Zvarevashe and Oludayo O. Olugbara. A framework for sentiment analysis with opinion mining of hotel reviews. In *2018 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–4, 2018. doi: 10.1109/ICTAS.2018.8368746.

[21] Luiz F Junior, Jorge Silva Junior, and Fábio Lobato. Um olhar sobre turismo gastronômico: Um caso no tripadvisor. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 519–530. SBC, 2020.

[22] Doris Xin, Litian Ma, Jialin Liu, Stephen Macke, Shuchen Song, and Aditya Parameswaran. Accelerating human-in-the-loop machine learning: Challenges and opportunities. In *Proceedings of the second workshop on data management for end-to-end machine learning*, pages 1–4, 2018.

[23] Douglas Cirqueira., Fernando Almeida., Gültekin Cakir., Antonio Jacob., Fabio Lobato., Marija Bezbradica., and Markus Helfert. Explainable sentiment analysis application for social media crisis management in retail. In *Proceedings of the 4th International Conference on Computer-Human Interaction Research and Applications - WUDESHI-DR,*, pages 319–328. INSTICC, SciTePress, 2020. ISBN 978-989-758-480-0.

[24] Mehrbakhsh Nilashi, Sarminah Samad, Azizah Abdul Manaf, Hossein Ahmadi, Tarik A Rashid, Asmaa Munshi, Wafa Almukadi, Othman Ibrahim, and Omed Hassan Ahmed. Factors influencing medical tourism adoption in malaysia: A dematel-fuzzy topsis approach. *Computers & Industrial Engineering*, 137:106005, 2019.

[25] Lucas Rodrigues, Ana Prado, and Fábio Manoel França Lobato. Pandemia de covid-19 no brasil: uma análise sobre notícias e comentários de usuários. *Culturas*

Midiáticas, 16:26–26, 2022.

[26] Eva Martin-Fuentes, Carles Mateu, and Cesar Fernandez. Does verifying uses influence rankings? analyzing booking. com and tripadvisor. *Tourism Analysis*, 23(1):1–15, 2018.

[27] Douglas Cirqueira, Márcia Fontes Pinheiro, Antonio Jacob, Fábio Lobato, and Ádamo Santana. A literature review in preprocessing for sentiment analysis for brazilian portuguese social media. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 746–749, 2018.

[28] Jessica Mei Pung, Ryan Yung, Catheryn Khoo-Lattimore, and Giacomo Del Chiappa. Transformative travel experiences and gender: A double duoethnography approach. *Current Issues in Tourism*, 23(5):538–558, 2020.

[29] Eduardo Tadeu Roque Amaral and Márcia Sipavicius Seide. *Nomes próprios de pessoa: introdução à antroponímia brasileira.* Editora Blucher, 2020.

[30] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[31] Geandreson Costa, Danielle Couto, Antonio Jacob Junior, and Fábio Lobato. Feminismo e redes sociais online: uma análise de tweets sobre o dia internacional da mulher. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 169–180, Porto Alegre, RS, Brasil, 2022. SBC. doi: 10.5753/brasnam.2022.223334. URL https://sol.sbc.org.br/index.php/brasnam/article/view/20526.

[32] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, apr 2012. ISSN 0001-0782.

[33] Vivek Kumar Rangarajan Sridhar. Unsupervised topic modeling for short texts using distributed representations of words. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 192–200, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1526. URL https://aclanthology.org/W15-1526.

[34] Benjamin Garner, Corliss Thornton, Anita Luo Pawluk, Roberto Mora Cortez, Wesley Johnston, and Cesar Ayala. Utilizing text-mining to explore consumer happiness within tourism destinations. *Journal of Business Research*, 139: 1366–1377, 2022.

[35] Yue Guo, Stuart J Barnes, and Qiong Jia. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism management*, 59:467–483, 2017.

[36] Yuyan Luo, Jinjie He, Yu Mou, Jun Wang, and Tao Liu. Exploring china's 5a global geoparks through online tourism reviews: A mining model based on machine learning approach. *Tourism Management Perspectives*, 37:100769, 2021.

[37] Rui Esteves, Fernando Paulo Belfo, and Antonio Trigo. Most valued factors in rural tourism: An analysis of portuguese customer comments on a booking platform. 2021.

[38] Jorge Silva Junior, Rafael Rossi, and Fábio Lobato. A lyric-based approach for brazilian music knowledge discovery: Brazilian country music as a case study. 10 2020.

[39] Huangxiong Qi and Rucong Mo. Exploring customer experience of smart hotel: A text big data mining approach. In *E3S Web of Conferences*, volume 251, page 01034. EDP Sciences, 2021.