A glance of gastronomic tourism: A case on TripAdvisor

Luiz Carlos S. F. Junior[1], **Jorge L. F. Silva Junior**[1], **Fábio M. F. Lobato**[1,2]

[1]Instituto de Engenharia e Geociências – Universidade Federal do Oeste do Pará
Santarém, PA, Brasil
[2]Centro de Ciências Tecnológicas – Universidade Estadual do Maranhão
São Luís, MA, Brasil

{luizcarlossfjr,jorgeluizfigueira}@gmail.com,

fabio.lobato@ufopa.edu.br

Abstract. *Analyzing and extracting information from the User-Generated Content (UGC) has become a prominent research topic. These data contain information such as consumer opinions, classifications, and recommendations for products and services, being a rich information source for assisting purchase decision making. Many papers have been published on UGC related to tourism, in particular culinary tourism. However, when observing the state of the art, it was found that there is a lack of antecedents that address the analysis of online reviews of Brazilian restaurants. In this sense, this work's focus is to fill this gap through a case study of Santarém city. The results show that professionals in this segment can use these analyses to improve their services' experience.*

Resumo. *Analisar e extrair informações do Conteúdo Gerado pelo Usuário (UGC) tornou-se um tópico de pesquisa com crescente atenção. Esses dados contêm informações como opiniões dos consumidores, classificações e recomendações de produtos e serviços, sendo uma rica fonte de informação para auxiliar nas decisões de compras. Muitos artigos foram publicados sobre UGC relacionado ao turismo, em especial, o turismo gastronômico. No entanto, ao observar a literatura, verificou-se a escassez de antecedentes que abordem a análise de avaliações online de restaurantes brasileiros. Nesse sentido, o foco deste trabalho é preencher essa lacuna por meio de um estudo de caso da cidade de Santarém. Os resultados mostram que os profissionais desse segmento podem utilizar essas análises a fim de aprimorarem a experiência de seus serviços.*

1. Introdução

Em 2018, o turismo contribuiu com US$ 152,5 bilhões ao Produto Interno Bruto (PIB) brasileiro, sendo equivalente a 8,1% da receita de todos os bens, produtos e serviços produzidos no país nesse ano[1]. No município de Santarém (Pará, Brasil) a colaboração desse setor também é significativa. De acordo com dados de um estudo realizado pela Secretaria Municipal de Turismo da cidade, essa atividade injetou R$ 176 milhões na economia local em 2018, movimentando segmentos como restaurantes, hotéis, agências de viagens, bares e lojas de artesanato [G1 2019].

[1]Dados do Ministério do Turismo (2019)

O mercado deste setor sofreu mudanças profundas com a Internet, que fez surgir novas formas de consumo e distribuição de informações turísticas [Navío-Marco et al. 2018]. Muitas plataformas de mídias sociais permitem que os usuários compartilhem seus comentários, opiniões e experiências pessoais relacionadas à viagens [Xiang and Gretzel 2010]. Nesse contexto, apesar de existir uma vasta gama de plataformas que fornecem conteúdo gerado por turistas, o TripAdvisor[2] destaca-se como a maior comunidade de viajantes do mundo, fornecendo mais de 800 milhões de comentários de usuários e 8 milhões de negócios turísticos. Em especial, os comentários sobre restaurantes são úteis para o segmento tratado como turismo gastronômico [da Silva et al. 2019].

Essa modalidade de turismo possibilita o reconhecimento de valores relacionados a cultura de um determinado território, de modo que a gastronomia se transforme em produtos turísticos [Nistoreanu et al. 2018]. Segundo [da Silva et al. 2019], é crescente a participação dessa atividade na receita dos destinos turísticos, visto que é possível considerar a comida como um fator diferencial de um local.

Nesse panorama, [Lee et al. 2016] aponta que comentários online sobre restaurantes influenciam na tomada de decisão dos seus consumidores, sendo então de vital importância a análise dessas informações pelas empresas a fim de melhorarem seus serviços [Schmunk et al. 2013, Silva et al. 2017]. Nos últimos anos, com o crescimento no volume de dados e sua diversidade de formatos, surgiu um desafio para os gerentes de empresas do setor turístico: analisar o Conteúdo Gerado pelo Usuário, conhecido pelo seu acrônico em inglês UGC - *User-Generated Content* [Nistoreanu et al. 2018, Schmunk et al. 2013, Lobato et al. 2016].

Para contornar esse problema, técnicas de mineração de texto podem ser empregadas a fim de identificar padrões de comportamento e gerar *insights* que podem ajudar no processo de tomada de decisão [Talib et al. 2016, Zhao 2012, Schmunk et al. 2013]. Por exemplo, é possível identificar os principais tópicos presentes nos comentários fazendo o uso de modelagem de tópicos, como explorado nos trabalhos de [Marcolin et al. 2019] e [Santos et al. 2018]. Ademais, a utilização de análise de sentimentos é viável para identificar a polaridade de um comentário, assim como a previsão de gênero do autor de uma avaliação e a extração de regras podem ser análises úteis para identificar padrões, ambas as técnicas abordadas, respectivamente, em [Santos et al. 2018, Silveira et al. 2018, Silva Junior et al. 2020].

Por meio de uma revisão do estado da arte, notou-se uma escassez de antecedentes que exploram a extração de conhecimento de comentários das mídias sociais de restaurantes brasileiros. Nesse contexto, o presente trabalho busca analisar a extração de padrões dos comentários sobre restaurantes da plataforma TripAdvisor, realizando um estudo de caso do município de Santarém. Além disso, os trabalhos relacionados não abordam a correlação do gênero dos autores com os tópicos considerados mais importantes por eles, sendo este um dos diferenciais do presente estudo. Nesse contexto, técnicas de mineração de texto foram empregadas a fim de responder as seguintes Perguntas de Pesquisa (PP):

1. Qual é o sentimento predominante manifestado nos comentários de restaurantes da cidade de Santarém na plataforma TripAdvisor?

[2]https://tripadvisor.com/

2. Qual gênero tem mais comentários de restaurantes classificados como negativos?
3. Quais são os principais tópicos abordados nos comentários de restaurantes da cidade de Santarém na plataforma TripAdvisor? Há distinção de temáticas entre os gêneros masculino e feminino?

O restante deste artigo está organizado como segue. Nas Seções 2 e 3 serão abordados, os trabalhos relacionados e a metodologia utilizada para a condução desta pesquisa. Na Seção 4 são descritos os resultados. Por fim, na Seção 5 as conclusões e as perspectivas de trabalhos futuros são apresentadas.

2. Trabalhos Relacionados

Análises de UGC vêm sendo amplamente abordadas em diversos estudos devido as potencialidades de aplicação no processo de aperfeiçoamento de serviços e produtos [Lobato et al. 2016, Almeida et al. 2020, de Sousa et al. 2019]. Essas informações são ainda mais importantes para o setor do Turismo, cujo público considera como um método bastante confiável para auxiliar o processo de tomada de decisão [Yan et al. 2019, Narangajavana Kaosiri et al. 2019]. Nesse cenário, a maior parte dessas informações são constituídas por dados textuais, assim, uma abordagem adequada para se analisar essa grande quantidade de dados é o uso de técnicas de mineração de texto [Schmunk et al. 2013].

Entre a utilização dessas técnicas, destaca-se o uso de *Latent Semantic Analysis* (*LSA*) e *Latent Dirichlet Allocation* (*LDA*) para a tarefa de modelagem de tópicos, como descrito nos estudos de [Marcolin et al. 2019] e [Santos et al. 2018]. Ao analisar dados coletados das plataformas TripAdvisor, Airbnb, Couchsurfing e Booking, os autores obtiveram como resultado a segmentação do tipo de avaliação (como por exemplo: conforto, localização e experiência) bem como a identificação de principais problemas de serviço (exemplo: níveis de limpeza do banheiro).

Adicionalmente à análise de modelagem de tópicos, podem ser desempenhadas a análise de sentimento como observado nos trabalhos de Taecharungroj e Mathayomchan (2019), Silveira et al (2018). Em [Taecharungroj and Mathayomchan 2019], os autores utilizaram um classificador baseado em Naive Bayes e em [Silveira et al. 2018] foi utilizado as bibliotecas python NLTK e TextBlob para identificar a polaridade de revisões coletadas da plataforma TripAdvisor e Yelp. Como resultado, ambos os trabalhos apresentam a distribuição da opinião dos consumidores em relação ao serviço.

Durante a investigação em UGC, as análises podem ser combinadas e correlacionadas, como também complementadas com outras. [Taecharungroj and Mathayomchan 2019] combinam modelagem de tópicos com análise de sentimentos para obter os principais tópicos segmentados por polaridade. E em [Silveira et al. 2018] foi utilizada a biblioteca Python Guess-Gender para identificação dos gêneros dos autores das avaliações, a fim de determinar se há diferenças nos tópicos de acordo com o gênero.

Por fim, pode-se fazer ainda a identificação de padrões como observado nos trabalhos de Lee et al. (2016) e Ashish and Duggal (2015) ao investigarem o padrão de comportamento dos consumidores. Em [Lee et al. 2016], com auxilio do software Nvivo 10 foi empregada uma análise de avaliações de restaurantes disponíveis na plataforma

TripAdvisor, e em [Ashish and Duggal 2015], os autores avaliaram dados de perfis do Twitter de três maiores redes de pizzaria. Os pesquisadores observaram a existência de diferenças significativas entre as preferências gastronômicas dos usuários.

3. Metodologia

Nesta seção será apresentada a metodologia utilizada para a condução do presente estudo, detalhando as etapas de aquisição dos dados, pré-processamento e os métodos empregados para se realizar a descoberta de conhecimento nos comentários, conforme exposto na Figura 1 .

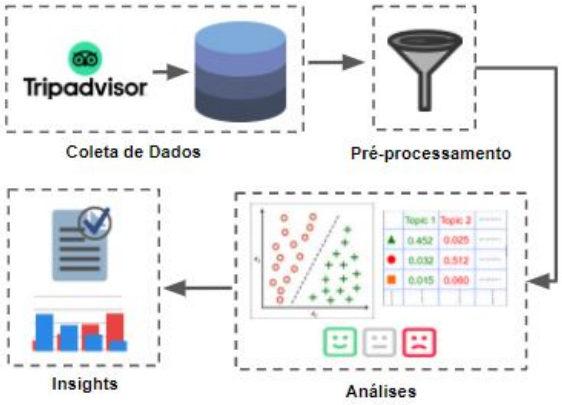

Figura 1. Visão geral da metodologia.

3.1. Aquisição dos dados

Conforme mencionado na introdução, escolheu-se a plataforma TripAdvisor como fonte de dados devido sua posição de destaque como a maior comunidade de viajantes do mundo. Ao todo foram coletados 3.881 avaliações de 186 restaurantes da cidade de Santarém. Dentre os dados extraídos, se enumeram: i) nome do restaurante, ii) nota do restaurante, iii) título do comentário, iv) nota do comentário, v) conteúdo do comentário e vi) nome do usuário. Para realizar esse processo, desenvolveu-se um *web crawler* na linguagem *Python* com o uso da biblioteca *BeautifulSoup*[3] e da *Application Programming Interface* (*API*) *Selenium*[4]. Destaca-se a necessidade de utilizar essa API por conta das interações com as páginas através do navegador (*Chromedriver*[5]) que são fundamentais para o carregamento das informações. Por fim, os dados foram armazenados em um arquivo no formato *Comma-Separated Values* (*CSV*), para posteriormente serem utilizados nas tarefas de análises.

3.2. Pré-processamento

Em se tratando de atividades de mineração de texto, uma tarefa que influencia a qualidade dos resultados é a preparação dos dados [Aggarwal 2018]. Essa etapa consiste no tratamento das informações a fim de remover inconsistências e melhorar a confiabilidade dos resultados de pesquisa [Rodrigues et al. 2020, Cirqueira et al. 2018]. Foi implementado

[3]https://www.crummy.com/software/BeautifulSoup/
[4]https://selenium-python.readthedocs.io/
[5]https://chromedriver.chromium.org/downloads

o pré-processamento através da linguagem *Python*, usando a biblioteca *NLTK* por possuir suporte para o idioma Português. Foram efetuadas: (i) conversão de caracteres para caixa baixa; (ii) remoção de acentuação; (iii) remoção de pontuação e caracteres especiais; (iv) remoção de números; (v) remoção de *stopwords*; (vi) remoção de emojis.

3.3. Análise de Sentimentos

A análise de sentimentos é frequentemente utilizada para extrair opiniões e emoções de diferentes fontes, tendo como principal uso a extração de polaridade de arquivos textuais [Marcolin et al. 2019, Yan et al. 2019, He et al. 2013], sendo aplicada, por exemplo, pelas empresas quando se deseja conhecer a taxa de aceitação de um novo produto [Santos et al. 2018]. Neste trabalho, a análise de sentimentos é utilizada para identificar a polaridade de um comentário, visto que a nota dada pelo usuário não pode ser utilizada para caracterizar esse atributo em todas as sentenças ou aspectos [Valdivia et al. 2019]. Para executar essa tarefa foi usada a biblioteca Polyglot [Chen and Skiena 2014], por possuir bons resultados em trabalhos anteriores no idioma Português [Rodrigues et al. 2020]. Pelo fato dessa ferramenta produzir um resultado numérico de -1 à 1, foi realizada uma categorização da polaridade (P), de modo a ser neutra quando $P = 0$, positiva quando $0 < P \leq 1$ e negativa quando $-1 \leq P < 0$.

3.4. Identificação de gênero

Para realizar a identificação de gênero do autor de um comentário, os pesquisadores conduziram uma classificação manual com base no nome usado pelo usuário, de modo a esse pertencer a um dos gêneros (masculino ou feminino) ou ser indefinido, quando o nome não permitir claramente uma rotulação. Convém destacar que em primeiro momento tentou-se utilizar ferramentas automáticas, contudo, a taxa de erro observada fez com quê esta abordagem fosse descartada. Isso ocorreu pois os usuários informam apelidos ao invés de nomes próprios.

3.5. Extração de regras

Existem diferentes maneiras de se extrair padrões de comentários. Uma delas é a representação com árvores de decisão, que é facilmente interpretada por humanos e possibilita entender como os documentos são classificados. Cada parte de uma árvore de decisão corresponde a uma regra [Aggarwal 2018]. As regras abrangem exemplos pertencentes principalmente ou exclusivamente a uma classe no conjunto de dados. Cada regra é composta por testes lógicos considerando termos e sua frequência [Freitas 2014]. Neste trabalho, o conjunto de regras foi obtida através do algoritmo de Árvore de Decisão do Scikit-learn [Pedregosa et al. 2011].

3.6. Modelagem de Tópicos

Para a modelagem de tópicos, foi utilizada a técnica *Non-Negative Matrix Factorization* (*NMF*) por ser mais eficiente para tarefas de mineração de textos curtos quando comparado à métodos tradicionais como *LDA* [Chen et al. 2019]. Esse modelo se baseia em um esquema de representação denominado *Bag-of-Words* (*BoW*), de modo que, no contexto do estudo, o conjunto de comentários é representado num espaço multidimensional, onde cada revisão passa a ser um vetor e cada palavra se torna uma característica. O peso utilizado para representar os valores dessas palavras na matriz de termos foi o

Term Frequency-Inverse Document Frequency (*TF-IDF*), em razão de trabalhos anteriores terem reportado bons resultados [Silva Junior et al. 2020, Vijayarani et al. 2015]. O *TF-IDF* mede a frequência de uma palavra no documento pelo inverso da frequência dos documentos.

O algoritmo *NMF* trata seus componentes como eixos de coordenadas, de modo que cada documento corresponde a um único ponto no espaço linear latente [Chen et al. 2019]. Assim, após executar a extração, é necessário gerar uma categorização dos resultados, sendo essa etapa realizada manualmente a partir de uma análise subjetiva dos autores [Chen et al. 2013]. Foram utilizadas diferentes opções de construções de tópicos e palavras (T-P), respectivamente: 3-5, 5-10, 10-5, 4-5, 4-8, 5-8. Visto que a escolha dessas representações são arbitrárias [Silva Junior et al. 2020, Rodrigues et al. 2020], adotou-se o melhor esquema que possibilitasse a rotulação dos tópicos pelos anotadores.

4. Resultados

Uma informação relevante que o TripAdvisor fornece é que cada comentário vem acompanhado de uma avaliação do estabelecimento de 0 à 5. Na Figura 2 é possível visualizar a quantidade de comentários por avaliação dada pelos usuários. A partir da análise da Figura 2 é possível notar uma tendência que os usuários tem em dar uma boa avaliação, de modo que as menores notas (10 e 20) juntas possuem apenas 127 ocorrências de um total de 3.881 avaliações. Apesar de serem utilizadas no cômputo da média do estabelecimento e este ser importante para a tomada de decisão de compras pelos consumidores, esta informação, sozinha, não é considerada tão relevante aos gestores. Por este motivo, o cruzamento das notas com o conteúdo se faz tão importante, pois permite a identificação de falhas em produtos, serviços e processos internos, por exemplo.

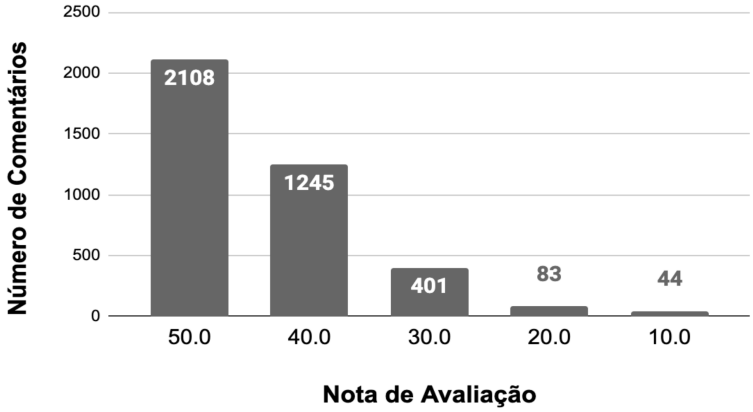

Figura 2. Distribuição de comentários por nota.

Após a etapa de pré-processamento os comentários foram submetidos à análise de sentimentos. A Figura 3 apresenta a distribuição da polaridade, na qual se percebe que há uma quantidade relevante de avaliações positivas e neutras. Diante desse cenário e da 1º PP, é possível concluir que o principal sentimento manifestado é positivo, tendo uma prevalência de 66,5%.

Tais avaliações incluem, em grande parte, elogios a comida, ambiente e atendimento, como é possível observar nos itens 1 e 2 da Tabela 1. Ademais, o conteúdo dos

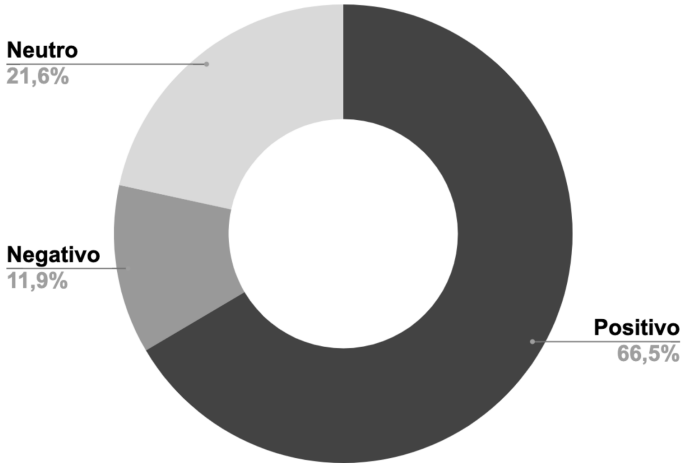

Figura 3. Análise de Sentimentos.

comentários negativos (11,9%) é composto por frustrações quanto a esses mesmos aspectos e ao preço, como os itens 3, 4 e 5 da Tabela 1.

Tabela 1. Exemplos de comentários positivos e negativos.

Item	Comentário pós pré-processamento
1	gostei quantidade grande sabores atendimento lugar limpo diferenciado vale pena conhece
2	ambiente maravilhoso decorado comida gostosa excelente local descontrair dia unica ressalva pessoal atendimento sao bons automatico certeza vez vier santarem nossa casa sera parada obrigatoria
3	lugar carissimo r kg sushi absurdo combos virem pedacos frescos salmao podre recomendo pior inimigo
4	tradicao local lanchonete parou tempo ambiente sujo atendentes indispostos tempo espera enoooorme h sanduiches mas ha concorrencia salva sabor comida
5	atendentes emburradas lanches ruins dar passada estiver precisando supermercado parte lanches lotada atendentes desorganizadas perca tempo fazendo lanche estabelecimento

Quanto a identificação de gênero, de um total de 3.811 nomes de usuários, 1.735 (44,7%) foram anotados como sendo do gênero masculino, 1.516 (39,1%) do feminino e 630 (16,2%) indefinido. Os comentários classificados como indefinido se justificam pela alta presença de ruído encontrada nos nomes e pela existência de pseudônimos, como por exemplo, *Dream508624* e *Y4979PGalinem*, sendo por isso impossível rotulá-los.

Nesse panorama, considerando apenas os comentários em que foi possível identificar o autor como pertencente a um dos gêneros e correlacionando esses dados com a análise de sentimentos dos comentários, houve uma ocorrência de 220 (12,7%) avaliações negativas no gênero masculino e 171 (11,3%) no feminino, conforme apresentados na Figura 4. Assim, é possível responder a 2º PP, na qual o gênero masculino obteve uma maior ocorrência de comentários negativos.

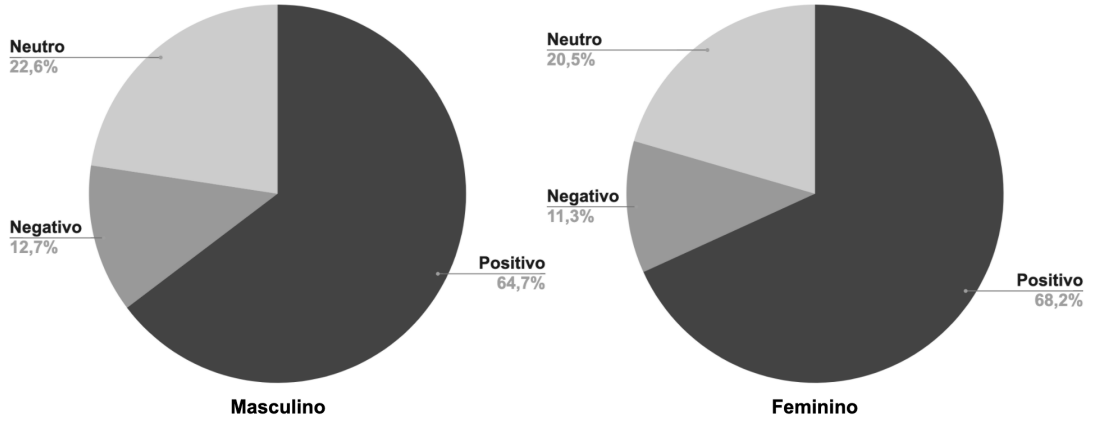

Figura 4. Correlação de sentimentos com gênero.

Com o intuito de entender como o padrão de um comentário positivo ou negativo é caracterizado, adicionalmente fez-se a extração de regras dos comentários das avaliações considerando a polaridade do sentimento como uma classe. Semelhante a análise conduzida em [Silva Junior et al. 2020], foi utilizado o algoritmo de Árvore de Decisão e, como entrada de dados, uma representação *BoW* dos comentários com esquema de peso binário.

O resultado dessa análise é um conjunto de regras descritoras apresentadas na Tabela 2. Vale ressaltar que foram extraídas as regras com maior valor de abrangência, ou seja, maior percentual de cobertura para cada grupo de comentários.

Tabela 2. Regras extraídas da Árvore de Decisão

Classe	Regras	Abrangência
Positivo	Ocorrência de: (vale a) pena, excelente, comer, demora, variado, boa, frito, caro, havia, aconchegante	51%
Negativo	Ausência de: (vale a) pena, excelente, delicioso, super, natureza, grande, visita, opções, maravilhoso	30%

Ao analisar as regras extraídas é possível observar que comentários, cujo o sentimento é tido como positivo, tendem apresentar a ocorrência de termos como "*pena*" que faz menção à "*vale a pena*", "*excelente*","*boa*", "*caro*" e "*aconchegante*". A partir destes termos é possível inferir aspectos como elogios à qualidade do serviço, enquanto os comentários de teor negativo demonstram ausência dos termos como "*vale a pena*", "*excelente*","*delicioso*", "*natureza*" e "*opções*", o que revela uma insatisfação com o cardápio e ambiente.

A partir da análise do melhor cenário para a modelagem de tópicos descrita na Seção 3, optou-se pela combinação de 4 tópicos e seus 5 principais termos. A tarefa de rotulação se baseou numa anotação manual dos pesquisadores, de modo que eles indicavam o tema que julgavam prevalente nos principais termos, seguido de um cruzamento entre essas afirmações para se chegar a um consenso. Os resultados dessa etapa são apresentados nas Tabelas 3, 4 e 5.

Ao analisar a Tabela 3 é possível responder a primeira parte da 3º PP, na

Tabela 3. Tópicos com base em todos os comentários.

Tópico	Principais Termos
Atendimento	atendimento ambiente excelente agradavel otimo
Localização	vista rio vale lugar pena
Cardápio	pratos sao restaurante pirarucu peixes
Precificação	boa comida preco opcao cidade

qual nota-se que os principais tópicos presentes nos comentários são atendimento, localização, cardápio e precificação. Isso corrobora parcialmente os resultados de [Silveira et al. 2018], que identificaram por meio de nuvens de palavras os termos comida, local e serviço como mais recorrentes em revisões desses estabelecimentos na plataforma Yelp, concluindo que as avaliações dos usuários giravam em torno dessas temáticas. Adicionalmente a isso, destaca-se ainda o preço como tópico considerado importante pelos consumidores.

A fim de responder a segunda parte da 3º PP são apresentadas as Tabelas 4 e 5.

Tabela 4. Tópicos de comentários do gênero masculino.

Tópico	Principais Termos
Atendimento	boa comida atendimento ambiente agradavel
Cardápio	pratos sao pirarucu restaurante peixes
Localização	vista rio lugar tapajos local
Satisfação	vale pena conferir conhecer experimentar

Tabela 5. Tópicos de comentários do gênero feminino.

Tópico	Principais Termos
Atendimento	atendimento comida ambiente otimo agradavel
Localização	vista lugar rio praia tapajos
Cardápio	pratos sao restaurante pirarucu peixe
Satisfação	vale pena boa conferir conhecer

Ao analisar os dados das Tabelas 4 e 5, percebe-se que não há uma distinção entre os tópicos ordenados por gênero, o que permite-nos inferir que: tratando-se de restaurantes, para o cenário analisado, não foram encontradas diferenças significativas entre os aspectos abordados por homens e mulheres.

5. Conclusões e Trabalhos Futuros

O turismo gastronômico tem sido cada vez mais propagado pois permite a ampliação da cadeia sustentável de desenvolvimento, por meio da participação de pequenos produtores, valorizando a cultura regional e por potencializar os atrativos locais. Como as mídias sociais permitiram a produção de um grande volume de dados sobre produtos e serviços do setor turístico, a exemplo de plataformas como AirBnB, Booking, TheFork e TripAdvisor, abordagens automáticas para a análise desses dados se fazem necessárias para auxiliar os gestores no processo de tomada de decisão e aprimoramento de serviços.

Este trabalho apresentou um estudo baseado no conteúdo gerado pelos usuários na plataforma TripAdvisor, conduzindo: análise de sentimentos, identificação de gênero do autor, extração de regras e modelagem de tópicos, de modo responder à três perguntas de pesquisa. Nesse contexto, os resultados do presente trabalho demonstraram que o conteúdo gerado pelo usuário é um fonte bastante rica para a extração de conhecimento relevante de avaliações online de restaurantes, tomando como base ainda o gênero do autor.

Como diferencial do estudo, destaca-se a consideração do gênero dos autores dos comentários nas análises. Além disso, como impactos dos achados de pesquisa, verificou-se que os resultados podem contribuir com o gerenciamento de empresas ligadas ao turismo gastronômico, auxiliando-as a melhor desenvolver seus produtos e serviços centrados no consumidor.

Como trabalhos futuros, pretende-se estender o escopo da pesquisa, considerando outras localidades e estabelecimentos. Vislumbra-se ainda uma articulação com secretarias de turismo para a avaliação de espaços públicos a fim de guiar a construção de políticas públicas centradas no cidadão.

Agradecimentos

Os autores gostariam de agradecer a Universidade Federal do Oeste do Pará (UFOPA), Pró-Reitoria de Ensino de Graduação (PROEN) e ao PET/FNDE/MEC pelo apoio ao desenvolvimento deste trabalho.

Referências

Aggarwal, C. C. (2018). *Machine learning for text*. Springer.

Almeida, G. R., Guimarães, I., Jacob Jr, A. F. L., and Lobato, F. M. F. (2020). Fontes de dados gerados por usuários: quais plataformas considerar? In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 25–36. SBC.

Ashish, D. and Duggal, S. (2015). Evaluation of websites using Balanced Scorecard (BSC) Approach in the Hotel Landscape in India. *Journal of Tourism*, 16(1):1–16.

Chen, Y. and Skiena, S. (2014). Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389.

Chen, Y., Zhang, H., Liu, R., Ye, Z., and Lin, J. (2019). Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems*, 163:1–13.

Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., and Ghosh, R. (2013). Leveraging multi-domain prior knowledge in topic models. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Cirqueira, D., Pinheiro, M. F., Jacob, A., Lobato, F., and Santana, Á. (2018). 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. IEEE.

da Silva, M. B. d. O., de Souza Moreira, M. C., de Souza, Á. G. R., de Oliveira Arruda, D., and Mariani, M. A. P. (2019). Gastronomia no tripadvisor: O que os turistas comentam sobre os restaurantes de bonito-ms? gastronomy on tripadvisor: What tourists comment about restaurants in bonito-ms-brazil? *ROSA DOS VENTOS-Turismo e Hospitalidade*, 11(4).

de Sousa, G. N., Almeida, G. R., and Lobato, F. (2019). Social Network Advertising Classification Based on Content Categories. In *International Conference on Business Information Systems*, pages 396–404.

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10.

G1 (2019). Turismo em Santarém cresce em 2018 e injeta R$ 176 milhões na economia, aponta estudo. Disponível em: `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`. Acesso em: 23 de Abril de 2020.

He, W., Zha, S., and Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3):464–472.

Lee, S., Ro, H., et al. (2016). The impact of online reviews on attitude changes: the differential effects of review attributes and consumer knowledge. *International Journal of Hospitality Management*, 56:1–9.

Lobato, F., Pinheiro, M., Jacob, A., Reinhold, O., and Santana, Á. (2016). Social crm: Biggest challenges to make it work in the real world. In *International Conference on Business Information Systems*, pages 221–232. Springer.

Marcolin, C., Becker, J. L., Wild, F., Schiavi, G., and Behr, A. (2019). Business analytics in tourism: Uncovering knowledge from crowds. *BAR-Brazilian Administration Review*, 16(2).

Narangajavana Kaosiri, Y., Callarisa Fiol, L. J., Moliner Tena, M. A., Rodriguez Artola, R. M., and Sanchez Garcia, J. (2019). User-generated content sources in social media: A new approach to explore tourist satisfaction. *Journal of Travel Research*, 58(2):253–265.

Navío-Marco, J., Ruiz-Gómez, L. M., and Sevilla-Sevilla, C. (2018). Progress in information technology and tourism management: 30 years on and 20 years after the internet-revisiting buhalis & law's landmark study about etourism. *Tourism Management*, 69:460–470.

Nistoreanu, B. G., Nicodim, L., and Diaconescu, D. M. (2018). Gastronomic tourism-stages and evolution. In *Proceedings of the International Conference on Business Excellence*, volume 12, pages 711–717. Sciendo.

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

Rodrigues, L., Junior, A., and Lobato, F. (2020). Notícias relacionadas a pessoas com deficiência: uma análise do conteúdo gerado pelos usuários em postagens de mídias sociais. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 811–822. SBC.

Santos, G., Santos, M., Mota, V. F., Benevenuto, F., and Silva, T. H. (2018). Neutral or negative? sentiment evaluation in reviews of hosting services. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, pages 347–354.

Schmunk, S., Höpken, W., Fuchs, M., and Lexhagen, M. (2013). Sentiment analysis: Extracting decision-relevant knowledge from ugc. In *Information and Communication Technologies in Tourism 2014*, pages 253–265. Springer.

Silva, W., Santana, Á., Lobato, F., and Pinheiro, M. (2017). A Methodology for Community Detection in Twitter. In *Proceedings of the International Conference on Web Intelligence*, pages 1006–1009.

Silva Junior, J., Rossi, R., and Lobato, F. (2020). Uma abordagem baseada em letras para a descoberta de conhecimento da música brasileira: o sertanejo como um estudo de caso. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 949–960, Porto Alegre, RS, Brasil. SBC.

Silveira, M. P., Xavier, W. Z., and Marques-Neto, H. T. (2018). Análises de dados de sistemas crowdsourcing: estudo de caso de avaliações de estabelecimentos realizadas no yelp. In *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*. SBC.

Taecharungroj, V. and Mathayomchan, B. (2019). Analysing tripadvisor reviews of tourist attractions in phuket, thailand. *Tourism Management*, 75:550–568.

Talib, R., Hanif, M. K., Ayesha, S., and Fatima, F. (2016). Text mining: techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, 7(11):414–418.

Valdivia, A., Hrabova, E., Chaturvedi, I., Luzón, M. V., Troiano, L., Cambria, E., and Herrera, F. (2019). Inconsistencies on tripadvisor reviews: A unified index between users and sentiment analysis methods. *Neurocomputing*, 353:3–16.

Vijayarani, S., Ilamathi, M. J., and Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1):7–16.

Xiang, Z. and Gretzel, U. (2010). Role of social media in online travel information search. *Tourism management*, 31(2):179–188.

Yan, L., Cha, N., Cho, H., and Hwang, J. (2019). Video diffusion in user-generated content website: An empirical analysis of bilibili. In *2019 21st International Conference on Advanced Communication Technology (ICACT)*, pages 81–84. IEEE.

Zhao, Y. (2012). *R and data mining: Examples and case studies*. Academic Press.