CONIC SEMESP
24º Congresso Nacional de Iniciação Científica

CATEGORIA CONCLUÍDO

ANÁLISE DE AVALIAÇÕES TURÍSTICAS DO TRIPADVISOR UTILIZANDO SVM E REGRESSÃO LOGÍSTICA

Autor: Daniel Douglas Rodrigues

1. RESUMO

Este artigo propõe um estudo sobre os sentimentos dos turistas após visitarem atrações turísticas no Brasil. Utilizando uma base de dados do *Kaggle* com comentários de turistas, aplicamos algoritmos de análise de sentimentos baseados em Máquina de Vetores de Suporte e Regressão Logística, visando baixo custo computacional e alta precisão. O objetivo é extrair as métricas necessárias para melhorar a experiência dos futuros visitantes. Nossos resultados mostraram uma acurácia média de 96%, indicando que os modelos empregados foram eficazes em classificar os sentimentos expressos nas avaliações. Este resultado sugere que, embora a análise de sentimentos baseada em comentários tenha sido bem-sucedida, é fundamental considerar a representatividade dos dados e o balanceamento das classes. A pesquisa oferece uma percepção valiosa sobre as oportunidades e limitações atuais e serve como base para futuros estudos na área, que possam explorar outras abordagens ou fontes de dados para aprimorar a recomendação de atrações turísticas.

2. INTRODUÇÃO

O turismo é um setor altamente competitivo que busca constantemente aprimorar a experiência do visitante. As avaliações *online*, cada vez mais influentes no processo de decisão de compra, oferecem uma oportunidade única para compreender as percepções dos turistas e identificar áreas de melhoria. A análise de sentimentos, ao permitir a classificação automática de opiniões em positivas, negativas ou neutras, possibilita a extração de *insights* relevantes a partir de grandes volumes de dados textuais. Essa abordagem contribui para a identificação de pontos fortes e fracos dos destinos turísticos, permitindo a implementação de ações estratégicas para aumentar a satisfação dos visitantes e a competitividade dos destinos (Yae-Jie e Hak-Seom, 2022).

No Brasil, os turistas utilizam o *TripAdvisor*[1] para comparar facilmente acomodações, restaurantes e atrações, beneficiando-se de avaliações detalhadas de outros viajantes. Por exemplo, em uma notícia recente, as Cataratas do Iguaçu e Cristo Redentor foram destacados como os melhores lugares do mundo para se visitar considerando as avaliações dos turistas[2].

Tais comentários fornecem uma rica fonte de dados a ser explorada computacionalmente para melhorar a experiência dos viajantes. Técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM) podem ser aplicadas para analisar as avaliações e extrair informações sobre a qualidade dos serviços bem como detectar padrões de satisfação ou insatisfação. Com isso, é possível não apenas entender melhor as expectativas dos turistas, mas também desenvolver estratégias personalizadas para atender às suas necessidades, promovendo melhorias contínuas nos serviços oferecidos e, consequentemente, aumentando a competitividade dos destinos turísticos.

Neste sentido, o objetivo deste trabalho é aplicar técnicas de PLN e AM para analisar avaliações do *TripAdvisor* e extrair métricas relevantes que possam contribuir para a melhoria da experiência dos futuros visitantes. Enquanto a PLN permite extrair informações semânticas e sentimentais dos textos das avaliações, o AM possibilita a construção de modelos capazes de classificar automaticamente as opiniões dos visitantes em categorias como positivas, negativas e neutras. Ao aplicar essas técnicas a um conjunto de dados de avaliações de diversos destinos turísticos, busca-se desenvolver um sistema capaz de identificar padrões, tendências e pontos críticos nas opiniões dos consumidores.

Para a realização deste estudo, utilizou-se a base de dados de avaliações de diversos destinos turísticos brasileiros disponíveis no *Kaggle*[3]. A plataforma funciona como um *hub* de competições de ciência de dados onde estudantes e profissionais da área podem compartilhar experiências e resolver desafios em conjunto (Batista, 2022). A partir dessa base, foram aplicados algoritmos de AM, como Máquinas de Vetores de Suporte (SVM) e Regressão Logística, para classificar as opiniões dos turistas em categorias de sentimento (positivo, negativo e neutro). A performance

[1] TripAdvisor: https://www.tripadvisor.com.br/
[2] G1: Cataratas do Iguaçu e Cristo Redentor, segundo turistas: https://bit.ly/3YkoE3h. Acesso em 29/07/2024.
[3] Kaggle: https://www.kaggle.com/

desses modelos foi avaliada por meio de métricas como acurácia, precisão, revocação (sensibilidade) e *F1-score*. As análises estatísticas e a visualização dos resultados permitiram identificar os aspectos mais relevantes das avaliações e compreender os fatores que influenciam a satisfação dos turistas.

A escassez de dados rotulados é um desafio comum em diversas áreas da inteligência artificial, incluindo a análise de sentimentos. Conforme apontado por (Bassani *et al.*, 2022), a obtenção de grandes volumes de dados rotulados pode ser um obstáculo para a aplicação de modelos de AM. Neste trabalho, buscamos contribuir para a área ao investigar a aplicabilidade de técnicas de AM em um cenário com um conjunto de dados de tamanho moderado. Os resultados desta pesquisa podem oferecer *insights* valiosos para a comunidade científica e para profissionais da área de turismo, demonstrando a viabilidade da análise de sentimentos em contextos com recursos limitados.

O restante deste artigo está organizado como segue. A Seção 3 apresenta os objetivos da proposta deste estudo complementando com uma revisão bibliográfica. A metodologia utilizada na condução deste trabalho e o desenvolvimento realizado são discutidos nas Seções 4 e 5, respectivamente. A Seção 6 apresenta e discute os resultados encontrados. Por fim, a Seção 7 discorre sobre as principais conclusões, propostas de continuação deste trabalho e destaca as principais limitações.

3. OBJETIVOS

A análise de sentimentos em avaliações *online* se tornou uma ferramenta essencial para o setor de turismo. Através da mineração de dados, é possível identificar padrões e tendências nas opiniões dos clientes, auxiliando na tomada de decisões estratégicas. Como aponta Cagnacci (2022), a automação de processos é fundamental para analisar o crescente volume de dados gerados pelas plataformas *online*. Técnicas como o LDA, utilizadas por Cagnacci (2022), permitem classificar os dados em grupos hierárquicos, facilitando a identificação de diferentes níveis de satisfação dos usuários.

Uma revisão abrangente de Xu *et al*. (2022) destaca a diversidade de metodologias empregadas na análise de sentimentos, incluindo AM e análise léxica. Ao analisar avaliações de turismo na Indonésia, Alamanda *et al*. (2019) demonstram

como essa abordagem pode ser utilizada para criar mapas de prioridades para destinos turísticos. Esses estudos, assim como o presente trabalho, evidenciam a importância da análise de sentimentos para melhorar a experiência do cliente e a competitividade das empresas do setor. Nota-se que pesquisadores têm se interessado em utilizar técnicas de AM para lidar com comentários em plataformas *online*. A Tabela 1 resume as principais técnicas adotadas para a análise de sentimentos em textos, com foco no setor do turismo. Esses estudos, assim como o presente trabalho, buscam identificar os principais desafios e oportunidades da aplicação da análise de sentimentos no contexto das avaliações de turistas, contribuindo para a melhoria da tomada de decisão no setor.

Tabela 1. Resumo dos trabalhos envolvendo análise de sentimentos em plataformas digitais e as ferramentas utilizadas.

Autores	**Plataformas**	**Ferramentas e Algoritmos**
Cagnacci (2022)	Airbnb	Método LDA e *Naive Bayes*
Xu *et al*. (2022)	Facebook, Reddit, TripAdvisor e Twitter	*Machine Learning*, Analisador léxico e Híbrido
Alamanda *et al*. (2019)	Google Review e Instagram	Mineração de Dados, Limpeza e Análise de Sentimentos
Sari *et al*. (2022)	Instagram, Snapchat e Twitter	*Multiplayer Perceptron*, *Naive Bayes*, *Random Forest* e *Support Vector Machine*

4. METODOLOGIA

Entender o comportamento das avaliações de experiências de usuários em suas viagens requer uma perspectiva de análise de sentimento baseada em dados. Para isso, executamos uma metodologia em cinco passos, conforme resumido pela Figura 1. Em seguida, apresentamos todas as etapas da metodologia, desde a coleta e pré-processamento dos dados até a aplicação de algoritmos de aprendizado de máquina e a avaliação dos modelos construídos.

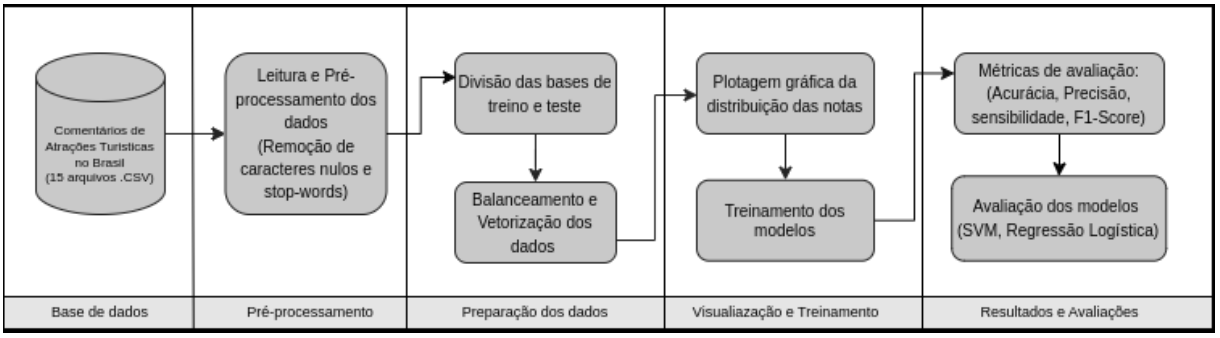


Figura 1. Fluxograma do processo algorítmico para análise de sentimentos.

5. DESENVOLVIMENTO

Base de Dados. Para a realização deste estudo, utilizou-se a base de dados *Comentários de Atrações Turísticas no Brasil* (Filho, 2023), que contém um vasto conjunto de avaliações de diversos destinos turísticos brasileiros. Tal base de dados está publicamente disponível na plataforma *Kaggle* e possui cerca de 70.516 comentários, classificados com avaliações que variam de 1 a 5. A alta quantidade de dados bem organizados desta base facilita a aplicação de modelos de AM. O *dataset* possui apenas três colunas: comentário, nota e data.

Pré-processamento. O segundo passo da metodologia consiste em preparar os dados para a análise de sentimentos. Os dados brutos obtidos do *Kaggle* foram submetidos a um processo de pré-processamento com o objetivo de torná-los adequados para a aplicação dos algoritmos de AM. Como apontado por (Ravi e Ravi, 2015), dados do mundo real são frequentemente ruidosos e não estruturados, exigindo um tratamento cuidadoso antes da análise. Para garantir a qualidade dos resultados, foram realizadas diversas etapas de limpeza e transformação dos dados, como a remoção de *stop words* e a normalização do texto. Além disso, a fim de diminuir a complexidade do problema e parametrizar os dados para a análise, também foi realizado o mapeamento das notas, categorizando-as em conceitos negativos (1 e 2), neutros (3) e positivos (4 e 5).

Preparação dos Dados. Para preparar os dados para o treinamento do modelo de AM, foi aplicada a técnica SMOTE (*Synthetic Minority Oversampling Technique*) para balancear as classes, conforme recomendado por Maione *et al*. (2020). A vetorização TF-IDF (*Term Frequency-Inverse Document Frequency*) foi utilizada para transformar as palavras em representações numéricas ponderadas,

permitindo o processamento pelos algoritmos. Seguindo as práticas padrão (Gholamy *et al*., 2018), os dados foram divididos em conjuntos de treinamento (70%) e teste (30%) para uma avaliação imparcial.

Em seguida, foi implementado o algoritmo de análise de sentimentos. Dentre as abordagens para essa tarefa (baseadas em léxico e híbridas) descritas por Maynard e Funk (2011), optou-se por utilizar algoritmos de classificação tradicionais e robustos, como a SVM e a Regressão Logística. Esses algoritmos têm se mostrado eficazes em diversas tarefas de classificação de textos, incluindo a análise de sentimentos.

Visualização e Treinamento. A fim de comparar a eficácia de diferentes abordagens na análise de sentimentos, foram escolhidos dois algoritmos de classificação amplamente utilizados na literatura: a SVM e a Regressão Logística. A SVM, conhecida por sua capacidade de encontrar os melhores hiperplanos de separação, é particularmente adequada para problemas de classificação não lineares, como aqueles encontrados na análise de sentimentos. Por outro lado, a Regressão Logística, devido à sua simplicidade e eficiência computacional, é uma excelente opção para classificar grandes volumes de dados. Uma análise mais aprofundada desses algoritmos pode ser encontrada em Faceli *et al*. (2011).

Resultados e Avaliações. Por fim, para avaliar a performance dos modelos, cada atração turística foi analisada individualmente, considerando que as características e os contextos de cada local podem influenciar significativamente as opiniões dos turistas. As métricas de acurácia, precisão, sensibilidade e *F1-score* foram utilizadas para avaliar a capacidade dos modelos em classificar corretamente os sentimentos expressos nas avaliações. Essas métricas, amplamente utilizadas na literatura de AM, fornecem uma visão abrangente do desempenho dos modelos. Para uma discussão mais detalhada sobre essas métricas, sugere-se consultar (Naidu *et al*., 2023). Os resultados obtidos com essas métricas serão apresentados e discutidos na próxima seção.

Para o desenvolvimento deste trabalho foi utilizada a linguagem de programação *Python* na versão 3.10.12. O *script* foi executado em um notebook baseado em nuvem chamado *Google Colab* devido à sua praticidade e suporte às bibliotecas necessárias. Além disso, uma das vantagens do *Colab* é a possibilidade

de acesso gratuito a GPUs (*Graphics Processing Unit*), o que permite a execução de tarefas paralelas em um menor tempo de execução (de Araújo Júnior, 2023).

6. RESULTADOS

Após a fase de treinamento, os modelos de SVM e Regressão Logística foram avaliados utilizando a biblioteca *scikit-learn*. Para a Regressão Logística, foi empregado o método *Grid Search* para otimizar o parâmetro de regularização C, sendo [0.001, 0.01, 0.1, 1, 10, 100], e definindo um número máximo de iterações igual a 1000. A fim de obter uma análise mais detalhada do desempenho dos modelos, foram geradas Matrizes de Confusão utilizando a biblioteca *matplotlib*. Essas matrizes permitem visualizar a distribuição das predições corretas e incorretas para cada classe de sentimento, auxiliando na identificação de padrões e na interpretação dos resultados. As Figuras 2 e 3 mostram as matrizes de confusão para os modelos de SVM e Regressão Logística, respectivamente.

Figura 2. Matriz de desempenho de Regressão Logística.

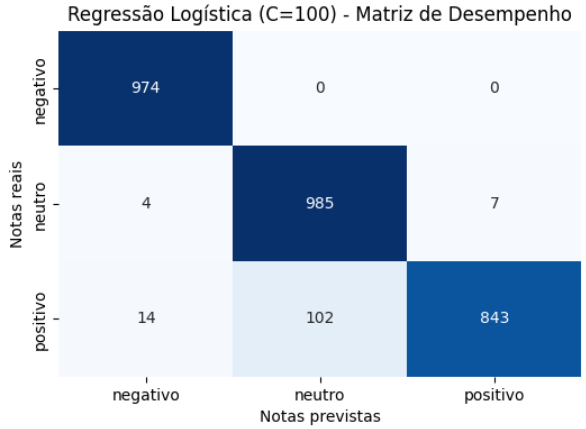
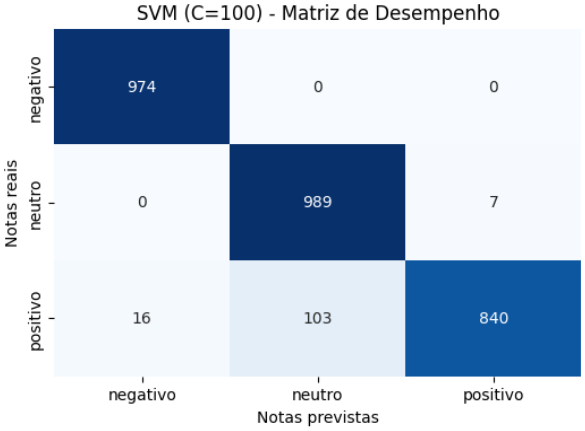


Figura 3. Matriz de desempenho de SVM.



Os resultados obtidos pelos modelos de classificação, apresentados na Tabela 2, indicam que as métricas de desempenho, como a acurácia, foram bastante satisfatórias. A acurácia média de 96% sugere que os modelos foram altamente eficazes em generalizar para os dados de teste. Esse resultado é consistente com estudos anteriores, como o de (Naidu *et al*., 2023), que também reportaram acurácias elevadas. A alta performance alcançada pelos modelos demonstra a eficácia das técnicas empregadas e a adequação dos hiperparâmetros selecionados.

Tabela 2. Desempenho dos Modelos.

	Acurácia	Precisão	Sensibilidade	F1-Score
Regressão Logística	95%	96%	96%	97%
SVM	96%	97%	96%	98%

A análise dos resultados por classe de sentimento revelou um desempenho superior para as avaliações positivas e neutras, o que pode ser atribuído à maior representatividade dessas classes na base de dados. Essa constatação corrobora as observações de (Bassani *et al*., 2022) e (Filho, 2023) sobre a importância do volume de dados para a construção de modelos de análise de sentimentos eficazes. As métricas de precisão, sensibilidade e F1-score, detalhadas nas Tabelas 3 e 4, reforçam essa tendência, evidenciando que os modelos tiveram um desempenho consistente nas classes mais representadas.

Tabela 3. Métricas isoladas por nota em Regressão Logística.

Notas	Precisão	Sensibilidade	F1-Score
Negativo	99%	89%	94%
Neutro	93%	100%	95%
Positivo	98%	99%	97%

Tabela 4. Métricas isoladas por nota em SVM.

Notas	Precisão	Sensibilidade	F1-Score
Negativo	99%	89%	93%
Neutro	92%	99%	95%
Positivo	98%	100%	99%

No entanto, é importante destacar que o balanceamento das classes desempenhou um papel crucial na melhoria dos resultados gerais. Futuras pesquisas podem explorar o impacto de diferentes técnicas de balanceamento e outros fatores, como a qualidade dos dados e a complexidade dos algoritmos, no desempenho dos modelos.

7. CONSIDERAÇÕES FINAIS

Neste estudo, exploramos a aplicação de técnicas de AM, especificamente a SVM e a Regressão Logística, para a análise de sentimentos em avaliações de turistas. Utilizamos uma base de dados extraída do *TripAdvisor*, que revelou um conjunto de dados abrangente e representativo para a tarefa. A análise focou na capacidade dos

modelos de classificar corretamente as avaliações em diferentes categorias de sentimentos, considerando a variedade e complexidade das opiniões dos usuários.

Com o balanceamento das classes, os modelos alcançaram uma acurácia satisfatória e consistente para diversas classes de sentimentos, demonstrando a eficácia das técnicas empregadas. Esses resultados destacam a importância do balanceamento de classes ao lidar com desafios como a complexidade da linguagem natural e a subjetividade dos sentimentos, que são inerentes à tarefa de classificação.

Limitações. Apesar dos esforços para realizar uma análise proficiente e robusta, ainda foram identificadas algumas limitações ao longo do processo. A base de dados utilizada apresenta baixa padronização e diversas *stop words* presentes nos comentários, o que pode impactar na consistência e qualidade dos resultados. Além disso, a complexidade da linguagem natural e a subjetividade dos sentimentos representam desafios contínuos, exigindo um cuidado especial na interpretação dos resultados e na escolha das técnicas de pré-processamento.

Trabalhos Futuros. Sugere-se como trabalhos futuros a exploração mais aprofundada das técnicas de SVM e Regressão Logística para aperfeiçoar o aprendizado máquina a fim de reconhecer um espectro mais amplo de emoções e identificar similaridades entre diferentes avaliações. Além disso, pode-se investigar a integração dessas técnicas com abordagens que detectem padrões e tendências nos sentimentos dos turistas, contribuindo para uma análise mais adaptável e perspicaz na interpretação das opiniões dos usuários.

8. FONTES CONSULTADAS

ALAMANDA, D. T.; RAMDHANI, A.; KANIA, I.; SUSILAWATI, W.; HADI, E. S. Sentiment analysis using text mining of Indonesia tourism reviews via social media. **International Journal of Humanities, Arts and Social Sciences**, p. 72–82, 2019.

BASSANI, C. N. O.; SAITO, P. T. M.; BUGATTI, P. H. Avaliação de abordagens semi-supervisionadas aplicadas a redes neurais convolucionais. In: Anais do XVI Brazilian e-Science Workshop. São Paulo: SBC, p. 1–8, 2022.

BATISTA, A. F. M.; SARAIVA, K. S. Pontos de divergência e complementaridade entre redes neurais em relação aos modelos autorregressivos. 2022. p. 9.

CAGNACCI, R. R. A influência da qualidade do serviço na hospitalidade domiciliar: um estudo do uso da Airbnb por turistas. 2022.

CONSONI, B. A importância do feedback. Fundação Educacional do Município de Assis – FEMA-Assis, p. 24, 2010.

DE ARAUJO JUNIOR, S. L. *et al*. Ferramenta para auxílio na rotulação de datasets para segmentação semântica. In: Anais do XX Congresso Latino-Americano de Software Livre e Tecnologias Abertas. São Paulo: SBC, 2023.

FACELI, K.; LORENA, A. C.; GAMA, J.; DE CARVALHO, A. C. P. L. F. Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina. LTC/Grupo Gen, 2011.

FILHO, J. S. Comentários de atrações turísticas no Brasil. Disponível em: https://www.kaggle.com/datasets/jeffersonsfilho/dadosatracoesturisticasbr/data. Acesso em: 5 jan. 2024.

GHOLAMY, A.; KREINOVICH, V.; KOSHELEVA, O. Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. Technical report, University of Texas at El Paso, 2018.

MAIONE, Camila *et al*. Balanceamento de dados com base em oversampling em dados transformados. 2020.

MAYNARD, D.; FUNK, A. Automatic detection of political opinions in tweets. In: Extended Semantic Web Conference. Berlin, Heidelberg: Springer, p. 88–99, 2011.

NAIDU, G.; ZUVA, T.; SIBANDA, E. M. A review of evaluation metrics in machine learning algorithms. In: Computer Science On-line Conference. Springer International Publishing, p. 15–25, 2023.

RAVI, K.; RAVI, V. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. **Knowledge-Based Systems**, v. 89, p. 14–46, 2015.

SARI, B. A.; ALKHALDI, R.; ALSAFFAR, D. *et al*. Sentiment analysis for cruises in Saudi Arabia on social media platforms using machine learning algorithms. **Journal of Big Data**, p. 21, 2022

XU, Q. A.; CHANG, V.; JAYNE, C. A systematic review of social media-based sentiment analysis: Emerging trends and challenges. In: Decision Analytics Journal. Springer, v. 3, p. 2772-6622, 2022

YAE-JI, K.; HAK-SEON, K. The impact of hotel customer experience on customer satisfaction through online reviews. **Sustainability**, v. 14, n. 2, p. 848, 2022.