www.cya.unam.mx/index.php/cya

Contaduría y Administración
Accounting & Management

Contaduría y Administración 70 (1), 2025, 169-183

Utilizing large datasets: Evaluating tourists' views on sustainable tourism in the Mekong Delta via Tripadvisor feedback

Evaluando las opiniones de los turistas sobre el turismo sostenible en el Delta del Mekong a través de las valoraciones de TripAdvisor

Dang TD*

Eastern International University, Vietnam

Received September 26, 2023; accepted January 10, 2024
Available online January 11, 2024

Abstract

In sustainable tourism studies, this research focuses on two primary inquiries concerning the Mekong Delta, a region in Vietnam known for its ecological diversity and potential as a green tourism benchmark. The study first evaluates the most effective machine learning model in forecasting travelers' emotions. Then, it determines the prominent elements of eco-tourism reflected in traveler feedback. Using a collection of 3,532 Tripadvisor reviews related to the "Mekong Delta Tour from HCM City - Discover the Delta's Charms," sentiment analysis and topic modeling techniques are utilized to decode tourists' perspectives on the sustainable tourism methods of the region. Preliminary outcomes indicate a significant inclination towards eco-friendly tourism initiatives. Additionally, key sustainability areas of interest and concern emerge through topic analysis. Such insights offer practical guidance for tourism stakeholders and decision-makers, emphasizing the role of data-driven strategies in guiding tourism towards environmental harmony, particularly in sensitive habitats like the Mekong Delta.

JEL Code: C55, C38, L83
Keywords: Mekong Delta; green tourism; sentiment analysis; topic modeling; sustainable practices

*Corresponding author.
E-mail address: doan.dang@eiu.edu.vn (Dang TD).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

http://dx.doi.org/10.22201/fca.24488410e.2025.5227

Resumen

En los estudios de turismo sostenible, esta investigación se centra en dos investigaciones principales relacionadas con el delta del Mekong, una región de Vietnam conocida por su diversidad ecológica y su potencial como punto de referencia del turismo verde. El estudio evalúa primero el modelo de aprendizaje automático más eficaz para pronosticar las emociones de los viajeros. Luego, determina los elementos destacados del ecoturismo reflejados en los comentarios de los viajeros. Utilizando una colección de 3.532 reseñas de Tripadvisor relacionadas con el "Tour por el Delta del Mekong desde la ciudad de Ho Chi Minh - Descubra los encantos del Delta", se utilizan análisis de sentimientos y técnicas de modelado de temas para decodificar las perspectivas de los turistas sobre los métodos de turismo sostenible de la región. Los resultados preliminares indican una inclinación significativa hacia iniciativas de turismo ecológico. Además, a través del análisis de temas surgen áreas clave de interés y preocupación en materia de sostenibilidad. Estos conocimientos ofrecen una orientación práctica para las partes interesadas y los tomadores de decisiones en el turismo, enfatizando el papel de las estrategias basadas en datos para guiar el turismo hacia la armonía ambiental, particularmente en hábitats sensibles como el delta del Mekong.

Código JEL: C55, C38. L83
Palabras clave: Delta del Mekong; turismo verde; análisis de sentimientos; modelado de temas; prácticas sostenibles

Introduction

Sustainability is a pillar of modern practices in the evolving landscape of global tourism (Robinson, Martins, Solnet, & Baum, 2019). With destinations navigating the complex interplay between economic vitality and ecological preservation, understanding traveler perceptions becomes a cornerstone of crafting holistic tourism experiences (Kiráľová, 2019). Central to this understanding is the vast swath of user-generated content (UGC) on platforms like Tripadvisor (Saydam, Olorunsola, Avci, Dambo, & Beyar, 2022). However, manually interpreting such voluminous data remains a herculean task, underscoring the need for advanced analytical methods.

In recent years, the importance of sustainable tourism has grown exponentially, both as a research area and as a practical endeavor. This study delves into understanding travelers' sentiments and topical trends related to green tourism and sustainable development. In the digital age, UGC, mainly reviews and feedback, has become a rich reservoir of insights (Saydam et al., 2022). Extracting meaningful information from extensive text data is challenging. However, with the advent of Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL), this task has become more feasible. NLP is a branch of artificial intelligence that focuses on the interaction between computers and humans through natural language, aiming to read, decipher, and understand human languages in a valuable manner (Otter, Medina, & Kalita, 2020). For this study, NLP facilitates the extraction of sentiment and topics from textual data. While ML, a subset of NLP, employs algorithms to find patterns

or regularities in data, DL uses neural network architectures to model and process complex patterns in large datasets (Bigne, Ruiz, Cuenca, Perez, & Garcia, 2021; Otter et al., 2020). In the context of current research, ML and DL are pivotal in discerning underlying sentiments and emergent topics in tourism reviews. This paper utilizes these advanced techniques, giving readers an in-depth understanding of how technology aids in unraveling intricate patterns in the realm of sustainable tourism.

The Mekong Delta, recognized for its ecological bounty and cultural vibrancy, provides a fitting backdrop for this investigation (Quang, Nguyen, Vo, & Nguyen, 2022). The current research delves into a dataset crawled from Tripadvisor in July 2023, using Python and the Selenium library, encompassing reviews for the "Mekong Delta Tour from HCM City - Discover the Delta's Charms". Through this dataset, the study aims to address two pivotal questions:

(1) Which ML model demonstrates the most reliable performance for predicting the sentiments of travelers to the Mekong Delta?

(2) Which aspects or themes of green tourism in the Mekong Delta emerge as prominent in traveler reviews?

By seeking answers to these questions, this study aspires to provide a data-driven roadmap for stakeholders in the Mekong Delta's tourism sector, highlighting areas of opportunity and improvement in green tourism.

Literature review

Green tourism and sustainable development

Green tourism, often interlinked with sustainable tourism, focuses on traveling in a manner that minimizes environmental impact and promotes conservation (Butler, 1999). Butler (1999) delineated that sustainable development in tourism is not just about ecology but also includes economic viability and socio-cultural aspects—however, a significant element of green tourism zeroes in on energy-saving measures. Gössling, Peeters, and Scott (2013) highlighted the energy intensity of the tourism sector, emphasizing the need for sustainable energy solutions (Higham, Cohen, Peeters, & Gössling, 2013). Solar-powered accommodations, energy-efficient transport modes, and the use of biofuels have been recognized as imperative for sustainable tourism growth. While efforts are underway to integrate energy-saving measures in tourism, consistent and expansive application remains an opportunity (Bojanic & Warnick, 2020).

User-generated content and tripadvisor platform

The digital age has ushered in the UGC era, where consumers actively produce and share content, particularly reviews, photos, and opinions about their experiences (Saydam et al., 2022). Many studies emphasized the pivotal role of UGC in shaping prospective travelers' decision-making processes (Bigne et al., 2021; Otter et al., 2020; Verma & Yadav, 2021). Among the myriad platforms for UGC, Tripadvisor stands out as a dominant player in the tourism sector, accumulating many reviews, ratings, and traveler photos (Kar & Dwivedi, 2020). The credibility and influence of reviews on TripAdvisor have been underscored in several studies, rendering it a vital tool for travelers and service providers in the tourism industry (Rita, Ramos, Borges-Tiago, & Rodrigues, 2022).

Sentiment analysis and topic modeling studies in tourism

The burgeoning field of data analytics has made inroads into the tourism sector, with sentiment analysis and topic modeling emerging as potent tools for analyzing UGC (Rita et al., 2022). Sentiment analysis investigates the emotional undertone within texts, essential for assessing travelers' satisfaction or dissatisfaction (Rita et al., 2022; Verma & Yadav, 2021). On the other hand, topic modeling, as elaborated by Blei, Ng, and Jordan (2003), aids in unearthing dominant themes or topics from extensive textual data (Blei, Ng, & Jordan, 2003). Several studies have applied these techniques to tourism, offering insights into traveler preferences, pain points, and trends (Mishra, Urolagin, Jothi, Neogi, & Nawaz, 2021; Vayansky & Kumar, 2020). However, their application specifically to green tourism remains sparse. Given the rising prominence of sustainability in global tourism, this gap signifies a rich avenue for exploration. By focusing on sentiment analysis and topic modeling within the context of green tourism, this study aims to contribute to this under-researched domain.

Methodology

Study process and datasets description

The research was built upon a sturdy dataset extracted from Tripadvisor. Leveraging web scraping method, this study employed Python and the Selenium library to collect data (Ali, Omar, & Soulaimane, 2022; Kar & Dwivedi, 2020). The dataset pertains to the "Mekong Delta Tour from HCM City - Discover

the Delta's Charms" and comprises 3,532 records. Each record boasts myriad attributes that reflect the reviews' quantitative and qualitative aspects. The procedure of study is described in detail in Fig. 1.

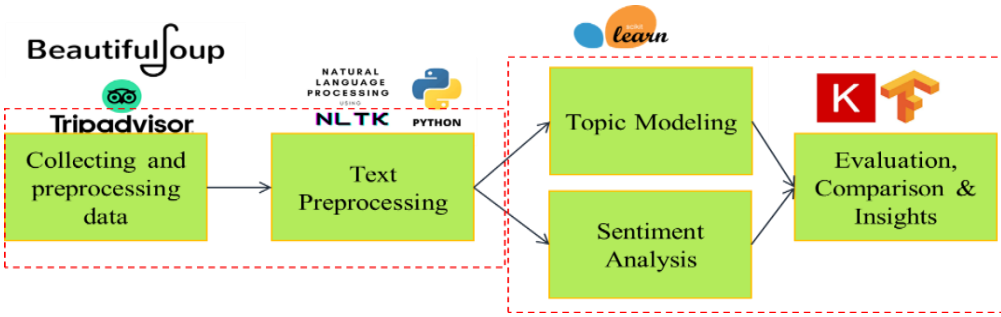

Figure 1. Summary of the research procedure

Preprocessing steps

Given the intricate nature of textual data, preprocessing is indispensable for clarity and accuracy in subsequent analyses. Bird, Klein, and Loper (2009) accentuate the significance of such steps in natural language processing (Bird, Klein, & Loper, 2009). The study harnessed the Natural Language Toolkit (NLTK) — a standout Python library for natural language processing. Subsequently, the study constructed a robust preprocessing pipeline with several strategic operations:

(1) Textual Standardization: All text was converted to lowercase, providing uniformity and eliminating case-related discrepancies.

(2) Tokenization: The corpus was segmented into individual tokens, laying the groundwork for granular analysis.

(3) Lemmatization: Words were streamlined to their canonical forms, fostering semantic consistency.

(4) Stopword Removal: Common stopwords, despite their prevalence, were removed, given their limited analytical value.

(5) Multi-word Consolidation: Certain multi-word phrases were fused into single token representations, preserving their inherent context.

(6) Token Filtering: Tokens with fewer than three characters were discarded to focus on contextually richer terms.

After optimizing this pipeline, its implementation on the current dataset resulted in the 'PreprocessedText' column, representing the reviews' distilled and enhanced version. To conclude the

preprocessing, the conditioned text was numerically structured using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. This method modulates token significance, penalizing frequent terms if they are widespread across the dataset.

The carefully preprocessed text is appropriately structured for sentiment analysis and topic modeling. The sentiment analysis reveals the text's emotional tone, while the topic modeling unravels embedded themes. The comprehensive preprocessing ensures the dataset delivers robust results for both analytical dimensions. The visualization of the most popular words is presented in Fig. 2.

Figure 2. Word cloud from the travelers' comment

Machine learning models for sentiment analysis and topic modeling

Sentiment Analysis

Sentiment analysis seeks to unearth the sentiment embedded within textual data. Rita et al. (2022) emphasize that this method is paramount in understanding UGC data. The analysis employed:

- Decision Tree: A flowchart-like model used for classification and regression, where nodes represent features, branches depict decision rules, and leaves show outcomes (Quinlan, 1986).
- Multinomial Naive Bayes: Suitable for classification with discrete features, especially text classification, it calculates probabilities based on the assumption of feature independence (McCallum & Nigam, 1998)
- Logistic Regression: A statistical algorithm primarily for binary classification, it determines relationships between variables using a logistic function (Hosmer Jr, Lemeshow, & Sturdivant, 2013).
- Gradient Boosting: An ensemble method that iteratively builds on decision trees, reducing errors by adding trees that correct the mistakes of previous ones (Friedman, 2001).

- This study employed advanced DL techniques to optimize text classification capabilities. The approach resulted in the development of the "ConvBiGRUAttention Classifier". This architecture amalgamates the convolutional layers, adept at extracting local features, with bidirectional GRU units, facilitating the capture of contextual dependencies. An attention mechanism is also integrated, enabling the model to focus on salient portions of the text, thereby enhancing interpretability. Consequent dense layers finalize the classification. Preliminary results indicate that this composite structure offers significant promise in text analytics (Aguiar, Krawczyk, & Cano, 2023; Alsaeedi & Khan, 2019).

In this study, the chosen methodology for information analysis centers around sentiment analysis, a crucial approach for deciphering UGC data. A diverse set of traditional ML techniques, including Decision Tree, Multinomial Naive Bayes, Logistic Regression, and Gradient Boosting, were employed to ensure a ccomprehensive examination. Additionally, the study investigation incorporated advanced DL techniques, culminating in the development "ConvBiGRUAttention Classifier." This innovative architecture integrates convolutional layers for local feature extraction, bidirectional GRU units for capturing contextual dependencies, and an attention mechanism for focusing on salient text portions. The amalgamation of these elements enhances sentiment analysis's interpretability and overall classification performance. The current study's selection of sentiment analysis over alternative methodologies is justified by its efficacy in unveiling sentiments within textual data. Preliminary results indicate promising outcomes, emphasizing the value of the study composite approach in text analytics.

Topic Modeling

Latent Dirichlet Allocation (LDA) is a generative probabilistic model introduced by Blei et al. in 2003 (Blei et al., 2003) that is commonly used in e-commerce and tourism research (Ali et al., 2022). LDA posits that every document contains a blend of topics and specific words often represent a particular topic. This method produces a collection of words for every topic and the likelihood of each word's occurrence, indicating the significance of every word to that topic (El-Kassas, Salama, Rafea, & Mohamed, 2021). Furthermore, LDA suggests that documents can encompass multiple topics. The likelihood of a document on a particular topic can be gauged using LDA. Its applications span numerous domains, including document classification, content recommendation, and thematic trend analysis (Mishra et al., 2021).

Experiment result and discussion

This study assessed the performance of four traditional ML classifiers (Decision Tree, Multinomial Naive Bayes, Logistic Regression, Gradient Boosting) and a DL model called ConvBiGRUAttention Classifier

to meet the dual research goals. Subsequently, the study implemented the LDA methodology to distill salient topics within the travel reviews for theme extraction. The computational processes were conducted on a system running Windows 10, equipped with an Intel Core i9 processor, 32 GB RAM, and a GeForce RTX 3080 graphics card with 10 GB VRAM. Python 3.8, along with a Jupyter Notebook environment and relevant Python packages for data preprocessing and machine learning tasks, was utilized for these tasks.

Data preprocessing and model evaluation

This study's rigorous extraction process resulted in a dataset of 3,532 unique entries. Each entry is characterized by ten distinct attributes, providing a comprehensive understanding of visitor sentiments and experiences on the Mekong Delta Tour. The attributes capture diverse traveler profiles, spanning from families to solo adventurers, delineate their geographical origins, and trace the temporal contours of their visits, encompassing year, month, and day. The dataset is enriched with quantitative measures, such as the aggregate of contributions and likes, and qualitative components, including concise summaries and detailed reviews (see Table 1)

Notably, in this study, two attributes have undergone detailed refinement: "PreprocessedText", which is derived from the attribute "FullReview" and "Sentiment" which is based on the calculated "Rating". This enriched and meticulously curated dataset presents a significant resource, offering potential insights into visitor sentiments, evolving trends, and preferences associated with the Mekong Delta Tour.

Table 1
Summary attributes of the dataset

ID	Attributes	Data Types	Range/Sample
1	TypeTravellers	object	[unknown, Couples, Family, Friends, Solo]
2	Country	object	[Los Angeles, CA5 contributions, 1 contributio...
3	Year, Month, Day	float64	[2017.0, 2023.0]; [1.0, 12.0]; [1.0, 1.0]
4	TotalContributions	int64	[0, 3282]
5	Like	int64	[0, 14]
6	ShortReview	object	[Awesome one-day tour to the Mekong Delta, Mek...
7	FullReview	object	[Pleasantly surprised by the quality and profe...
8	Rating	float64	[1.0, 5.0]
9	Sentiment	int64	[0, 2]
10	PreprocessedText	object	[pleasantly surprised quality professionalism ...]

Source: own summary from the dataset

Model evaluation

For comparison purposes, some evaluation metrics have been considered. Model performance was assessed using accuracy, precision, recall, and F1-score metrics (see Table 2). The research used these

metrics to evaluate six ML and DL algorithms on the hotel reviews dataset, providing insights into their predictive power and areas for improvement (Hossin & Sulaiman, 2015; Ullah, Marium, Begum, & Dipa, 2020).

Table 2
Classification performance metrics

Performance Metric	Description	Formula
Confusion Matrix	It is a 2x2 table that includes True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) metrics.	
Accuracy	It measures the proportion of correct predictions made by the model.	Accuracy = (TP + TN) / (TP + TN + FP + FN)
Precision	It measures the fraction of positive predictions.	Precision = TP / (TP + FP)
Recall	It measures the fraction of actual positive instances correctly predicted by the model.	Recall = TP / (TP + FN)
F1-score	The harmonic means of precision and recall.	F1-score = 2 * (precision * recall) / (precision + recall)

Source: Hossin & Sulaiman (2015)

Sentiment analysis

When scrutinizing the performance metrics of diverse computational models on current dataset, it becomes paramount to underscore both accuracy and the F1-score. Given the dataset's imbalanced nature, the F1-score—a composite measure melding precision and recall—provides a nuanced lens through which model robustness can be ascertained (Vujović, 2021). Across the negative (Neg), neutral (Neu), and positive (Pos) sentiment classes, the Decision Tree exhibited F1-scores of 0.22, 0.26, and 0.97, respectively, with an accuracy of 0.95. The Multinomial Naive Bayes, Logistic Regression, and Gradient Boosting all recorded similar F1-scores in the range of 0.00 to 0.98 and an accuracy of 0.96. The innovative ConvBiGRUAttention Classifier demonstrated F1-scores of 0.48, 0.50, and 0.98 across the three sentiment classes, matching the top models with an accuracy of 0.96. While models like Logistic Regression showed zero precision and recall for negative and neutral classes, the ConvBiGRUAttention Classifier presented a more even performance across all sentiment classes (details see Table 3). This suggests its potential superiority in achieving a consistent trade-off between precision and recall, making it a strong contender for future sentiment classification endeavors.

Table 3
Results of the model's evaluation

Model	Accuracy	Precision (Neg, Neu, Pos)	Recall (Neg, Neu, Pos)	F1-score (Neg, Neu, Pos)
Decision Tree	0.95	0.33, 0.24, 0.97	0.17, 0.29, 0.98	0.22, 0.26, 0.97
Multinomial Naive Bayes	0.96	0.00, 0.00, 0.96	0.00, 0.00, 1.00	0.00, 0.00, 0.98
Logistic Regression	0.96	0.00, 0.00, 0.96	0.00, 0.00, 1.00	0.00, 0.00, 0.98
Gradient Boosting	0.96	0.40, 0.50, 0.97	0.11, 0.29, 0.99	0.17, 0.37, 0.98
ConvBiGRUAttention Classifier	0.96	0.53, 0.37, 0.99	0.44, 0.76, 0.97	0.48, 0.50, 0.98

Source: Own calculate from ML and DL models

Topic modeling

Through the LDA analysis, a spectrum of topics emerged, characterizing the sentiments and experiences of travelers. Here is a detailed breakdown of the significant topics uncovered:

Topic #1: reccomend, exceeded, expectation, delivered, thao, mei, engaged, theo, wealth, case

Topic #2: tour, trip, guide, day, great, recommend, experience, nice, mekong, good

Topic #3: dao, polite, genuine, path, ton, context, mango, overview, phenomenal, kindness

Topic #4: age, atmosphere, voice, amusing, talkative, impressive, affordable, punctual, nguyen, expectation

Topic #5: tour, great, guide, day, steven, mekong, recommend, river, local, dave

Topic #6: tour, guide, great, good, day, trip, recommend, mekong, lot, really

Topic #7: vegan, toan, aka, commercial, grew, played, true, cung, teach, heap

Topic #8: commentary, outstanding, detailed, sightseeing, musician, support, traveled, recomended, booked, luxury

Topic #9: cung, trip, alot, lân, thanks, tour, thoroughly, hang, uncle, guide

Topic #10: boat, tour, coconut, ride, lunch, trip, day, island, great, fruit

Through an examination of the ten topics generated via LDA, a condensation to three comprehensive themes becomes apparent:

Theme 1: Customer Experience & Satisfaction:

Topic #1, Topic #2, Topic #5, Topic #6, and Topic #9 collectively emphasize travelers' positive reactions to their experiences. Words like 'recommend', 'great', 'thanks', and 'exceeded' underpin the general satisfaction of tourists. This theme also touches upon the quality of the tours and day activities and the overall sentiment of recommending these experiences to others.

Theme 2: Personalized and Authentic Experiences:

Topic #3, Topic #4, Topic #7, and Topic #8: These topics delve into the more personal touchpoints of the travel experience. Mentioning individual names might imply tour guides or individuals who left a significant mark on travelers. The contrast drawn between genuine versus commercial experiences suggests that tourists value authenticity. The detailed sightseeing, luxurious experiences, and dietary preferences catered for (like 'vegan') hint at a tailored and genuine tour experience.

Theme 3: Local Exploration & Sustainable Undertones:

Topic #10: Emphasizes hands-on local experiences, including boat rides, gastronomic adventures, and island visits. The frequent mention of 'Mekong' showcases the popularity of river-based activities. Though not overtly mentioned, the nature-centric undertones and the appreciation for local and organic experiences suggest an underlying appreciation for sustainable tourism practices.

By focusing on these three overarching themes above, the experiences and sentiments captured by the LDA topics can be concisely represented, offering a clearer understanding of what travelers value and cherish in their trips.

Conclusion

This study addresses two key research questions about travelers' sentiments and topical trends in green tourism and sustainable development. The study analysis sheds light on tourism reviews' prevailing sentiments and emergent topics by leveraging traditional ML techniques and state-of-the-art deep learning models.

Key findings

Study sentiment analysis, harnessed from traditional models like Decision Tree, Multinomial Naive Bayes, Logistic Regression, Gradient Boosting and the advanced ConvBiGRUAttention Classifier model, underscored the predominant positive sentiment among travelers. The F1-score, a balanced metric of precision and recall, showed the ConvBiGRUAttention Classifier as the most well-rounded model, especially for negative and neutral sentiments, with scores of 0.48 and 0.50, respectively. This positivity can be inferred as a testament to the inherent quality and satisfaction of sustainable tourism practices (Mariani & Baggio, 2022). Furthermore, the topic modeling via the LDA method provided a granular look into the specific aspects that resonate with travelers. Notably, among the various topics, there was an apparent inclination towards experiences that emphasized local exploration, authenticity, and possibly eco-friendly undertones, signaling the appreciation and value tourists place on sustainable and green tourism. In conclusion, through the prism of ML and DL, this study illuminated the contours of sentiment

and thematic predilections in the green tourism context, paving the way for more informed and sustainable tourism practices.

Implications

From an academic perspective, this research enriches the expanding corpus of literature on green tourism and sustainable development. By harnessing the capabilities of the ML and DL models, the study has elucidated profound insights from UGC data (Verma & Yadav, 2021). This study serves as a conduit, bridging the previously existing chasm between unprocessed tourist sentiments and cogent, structured insights, thereby emphasizing the pivotal role of big data analytics in propelling sustainable tourism research forward.

On a practical front, tourism providers and stakeholders, particularly those operating in the Mekong Delta and analogous eco-touristic regions, can derive actionable intelligence from study's findings. A pronounced incentive exists to emphasize authentic, ecologically conscious, and localized experiences (Ginzarly, Srour, & Roders, 2022). Such initiatives increase tourist satisfaction and bolster sustainable tourism's objectives (Quang et al., 2022). Tourism managers aiming to enhance travelers' perceptions of green destinations should focus on a few key strategies. Firstly, set clear expectations for travelers and then strive to exceed them for a lasting positive impression. Offering engaging guided tours with knowledgeable tour guides can deeply enrich the experience. Incorporating local elements, such as a Mekong River ride or sampling local fruits, adds authenticity to the journey. Catering to niche interests, providing detailed sightseeing commentary, and emphasizing punctuality and affordability can further elevate traveler satisfaction. Regularly collecting and implementing feedback ensures consistent quality and improvement.

Limitations and future work

The current study, though insightful, has limitations. A key drawback is the dataset's reliance on reviews from a single platform, potentially introducing inherent biases (Vujović, 2021). Additionally, sentiment analysis faced challenges in identifying subtle sentiments, and the topics derived from LDA could benefit from enhanced human interpretation. To advance, incorporating reviews from diverse platforms will deepen and broaden insights. Implementation of advanced DL models, such as transformer-based architectures, may refine analytical precision (Otter et al., 2020).

Acknowledgments

The author thanks Eastern International University, Binh Duong Province, Vietnam, for generously supporting this research.

References

Aguiar, G., Krawczyk, B., & Cano, A. (2023). A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. Machine learning, 1-79.

Ali, T., Omar, B., & Soulaimane, K. (2022). Analyzing tourism reviews using an LDA topic-based sentiment analysis approach. MethodsX, 9, 101894.

Alsaeedi, A., & Khan, M. Z. (2019). A study on sentiment analysis techniques of Twitter data. International Journal of Advanced Computer Science and Applications, 10(2), 361-374.

Bigne, E., Ruiz, C., Cuenca, A., Perez, C., & Garcia, A. (2021). What drives the helpfulness of online reviews? A deep learning study of sentiment analysis, pictorial content and reviewer expertise for mature destinations. Journal of Destination Marketing & Management, 20, 100570.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit: " O'Reilly Media, Inc.".

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Bojanic, D. C., & Warnick, R. B. (2020). The relationship between a country's level of tourism and environmental performance. Journal of Travel Research, 59(2), 220-230.

Butler, R. W. (1999). Sustainable tourism: A state-of-the-art review. Tourism geographies, 1(1), 7-25.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. Expert systems with applications, 165, 113679.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Ginzarly, M., Srour, F. J., & Roders, A. P. (2022). The interplay of context, experience, and emotion at World Heritage Sites: A qualitative and machine learning approach. Tourism Culture & Communication, 22(4), 321-340.

Higham, J., Cohen, S. A., Peeters, P., & Gössling, S. (2013). Psychological and behavioural approaches to understanding and governing sustainable mobility. Journal of Sustainable Tourism, 21(7), 949-967.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398): John Wiley & Sons.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.

Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research–Moving away from the "What" towards the "Why". International Journal of Information Management, 54, 102205.

Kiráľová, A. (2019). Sustainable tourism marketing strategy: Competitive advantage of destination. In Sustainable tourism: Breakthroughs in research and practice (pp. 183-206): IGI Global.

Mariani, M., & Baggio, R. (2022). Big data and analytics in hospitality and tourism: a systematic literature review. International Journal of Contemporary Hospitality Management, 34(1), 231-278.

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. Paper presented at the AAAI-98 workshop on learning for text categorization.

Mishra, R. K., Urolagin, S., Jothi, J. A. A., Neogi, A. S., & Nawaz, N. (2021). Deep learning-based sentiment analysis and topic modeling on tourism during Covid-19 pandemic. Frontiers in Computer Science, 3, 775368.

Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. IEEE transactions on neural networks and learning systems, 32(2), 604-624.

Quang, T. D., Nguyen, H. V., Vo, T. V., & Nguyen, M. H. (2022). Tour guides′ perspectives on agrotourism development in the Mekong Delta, Vietnam. Tourism and Hospitality Research, 14673584221089733.

Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1, 81-106.

Rita, P., Ramos, R., Borges-Tiago, M. T., & Rodrigues, D. (2022). Impact of the rating system on sentiment and tone of voice: A Booking. com and TripAdvisor comparison study. International Journal of Hospitality Management, 104, 103245.

Robinson, R. N., Martins, A., Solnet, D., & Baum, T. (2019). Sustaining precarity: Critically examining tourism and employment. Journal of Sustainable Tourism, 27(7), 1008-1025.

Saydam, M. B., Olorunsola, V. O., Avci, T., Dambo, T. H., & Beyar, K. (2022). How about the service perception during the COVID-19 pandemic: an analysis of tourist experiences from user-generated content on TripAdvisor. Tourism Critiques: Practice and Theory, 3(1), 16-41.

Ullah, M. A., Marium, S. M., Begum, S. A., & Dipa, N. S. (2020). An algorithm and method for sentiment analysis using the text and emoticon. ICT Express, 6(4), 357-360.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. Information Systems, 94, 101582.

Verma, S., & Yadav, N. (2021). Past, present, and future of electronic word of mouth (EWOM). Journal of Interactive Marketing, 53, 111-128.

Vujović, Ž. (2021). Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications, 12(6), 599-606.