Artificial Intelligence for Detecting Cultural Heritage Issues on Overtourism Literature: A Topic Modeling Application

Laura Claudia Verdesca, Elisabetta Ronchieri(✉), and Alessandro Costantini

INFN CNAF, Viale Berti Pichat 6/2, Bologna, Italy
elisabetta.ronchieri@cnaf.infn.it

Abstract. Cultural heritage requires protection for preservation. Artificial Intelligence technologies present innovative methods to extract valuable insights from digital literature and understand complex themes in literary data.

This study applies artificial intelligence methods to cultural heritage literature in order to detect overtourism impacts on the preservation of cultural artifacts and historical sites.

Six different topic modeling techniques-Latent Dirichlet Allocation, Latent Semantic Indexing, Hierarchical Dirichlet Process, Non-Negative Matrix Factorization, Structural Topic Modeling, and Correlated Topic Modeling-have been used to determine topics.

The analysis reveals that overtourism is a multifaceted issue that significantly impacts cultural heritage. Many topics overlap across different models, such as overtourism and its impact, and cultural heritage and identity.

Keywords: Artificial Intelligence · Overtourism · Cultural Heritage · Topic Modeling

1 Introduction

Overtourism is a pressing global phenomenon characterized by an excessive number of tourists in popular destinations, leading to negative impacts on local communities, environments, and cultural heritage. As defined by [1], overtourism occurs when the volume of tourists exceeds the capacity of the infrastructure to accommodate them, resulting in strained resources and a decreased quality of life for residents. Cities, such as Venice, Barcelona, and Amsterdam, have experienced significant challenges due to the influx of tourists, leading to social conflicts, environmental degradation, and economic disparities [2].

Given the growing relevance of overtourism in tourism studies, it is crucial to conduct a comprehensive literature review to understand the underlying issues and trends with respect to cultural heritage. This study aims at identifying

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
O. Gervasi et al. (Eds.): ICCSA 2025 Workshops, LNCS 15886, pp. 51–69, 2026.
https://doi.org/10.1007/978-3-031-97576-9_4

the key themes, challenges, and solutions presented in the existing research in order to develop effective management strategies. Furthermore, synthesizing the literature provides a solid foundation for future research and policy-making in sustainable tourism.

To systematically review the literature on overtourism, various topic modeling techniques have been employed, including Latent Dirichlet Allocation (LDA) [3], Latent Semantic Indexing (LSI) [4], Hierarchical Dirichlet Process (HDP) [5], Non-negative Matrix Factorization (NMF) [6], Structural Topic Modeling (STM) [7], and Correlated Topic Modeling (CTM) [8]. These techniques uncover latent themes and patterns in the literature, facilitating a deeper understanding of how overtourism is conceptualized and addressed in scholarly work. Furthermore, using multiple models ensures that the identified themes are consistent across different artificial intelligence (AI) technologies.

The `Gensim` Python library has been used to implement the LDA, LSI, HDP, and NMF models, while `Seaborn` and `Matplotlib` Python libraries have been employed for data visualization. In addition to Python, the R programming language has been essential in our analysis: the `stm` package in R has supported the implementation of the Structural Topic Model (STM), while the `topicmodels` package has been used to apply the Correlated Topic Model (CTM).

2 Dataset Definition and Preprocessing Phase

To develop the raw dataset, a systematic search has been conducted by using the keyword "overtourism" across two prominent databases: Scopus and Web of Science. This process has yielded 439 articles from Scopus and 471 articles from Web of Science. After removing duplicates, a comprehensive dataset of 698 articles has been defined. Further refinement has involved the removal of articles without abstracts, resulting in a final dataset of 662 articles suitable for analysis.

The earliest articles on overtourism in our dataset date back to 2017, highlighting that this issue has gained considerable attention in the past decade (see Fig. 1). The growing number of studies from a variety of journals underscores the multidisciplinary nature of overtourism, involving fields, such as environmental science, economics, urban planning, and social studies. This diversity of perspectives has enriched our understanding and has emphasized the complex, interconnected challenges of managing tourism sustainably.

A preprocessing pipeline has been defined to prepare the dataset for topic modeling analysis and includes several key steps for standardizing and cleaning the text data. First, all text has been converted into lowercase, ensuring consistency across words regardless of capitalization. Then, we have removed links, accented characters, and punctuation, reducing noise by eliminating irrelevant content and ensuring uniformity. Digits and single-character tokens have also been removed to further minimize extraneous data. Contractions have been expanded (e.g., "don't" has been transformed into "do not") to maintain semantic integrity. Next, tokenization has split the text into individual words or "tokens", enabling more precise text analysis. Stopwords, including both standard and

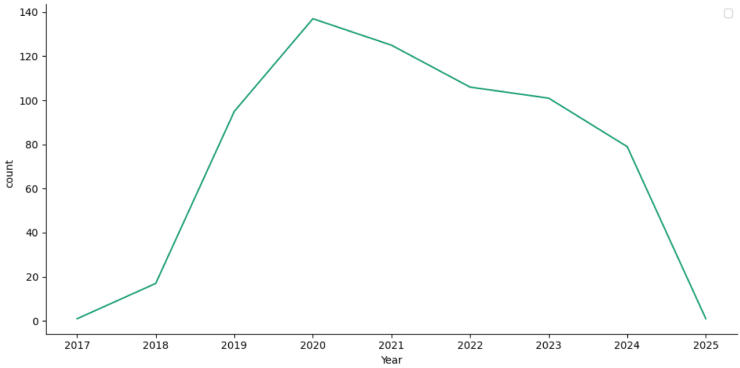

Fig. 1. Number of published papers per year - line chart showing the count papers from 2017 to 2025. The count increases from 2017, peaks around 2020, and then gradually declined towards 2025.

domain-specific terms, have been removed to concentrate on the core content. Finally, lemmatization has reduced words to their root forms, enhancing coherence by grouping variations of the same word. Together, these preprocessing steps have refined the text corpus, setting the stage for reliable and interpretable topic modeling results.

2.1 Initial Analysis

With the computations of unigrams and bigrams, we have identified the occurrence terms that have prominently emerged in the dataset:

- ***tourism***: A central term, as the concept of overtourism directly examines the negative aspects of excessive tourism in popular locations.
- ***destination***: Refers to the targeted places for tourism, often locations suffering from overtourism due to their appeal and limited capacity.
- ***heritage***: Indicates areas rich in cultural or natural heritage that are often highly vulnerable to degradation from overcrowding.
- ***negative impact***: Describes the adverse outcomes of overtourism on the environment, local infrastructure, and community well-being.
- ***sustainable tourism***: Highlights the need for tourism that balances economic benefits with environmental preservation and social equity, a solution often suggested for overtourism.
- ***tourism development***: Refers to the planning and growth of tourism infrastructure, which, if unregulated, can contribute to overtourism.
- ***community***: Denotes the local population whose lives and environment are affected by tourism influxes.
- ***quality of life***: Reflects the well-being of local residents, which can be negatively impacted by overtourism through increased congestion, noise, and other stressors.

– ***local resident***: Refers to inhabitants of tourism hotspots who experience both the benefits and downsides of tourism, often advocating for measures to curb excessive tourism.

3 Topic Modeling Application

Each selected topic modeling technique has been applied to the preprocessed data to extract the appropriate themes and patterns. The quantitative evaluation of the models has been performed by computing the coherence score that measures how interpretable the topics are to humans: the higher the coherence score, the closer the words within a topic are. We have considered the UMass [9], and CV [10] coherence measures for LDA, LSI, HDP and NMF, while the semantic coherence measure [7], closely related to the UMass measure, for STM and CTM.

According to the training time, LSI has been the fastest model followed by NMF. LDA has requested a longer training time, while HDP has taken a moderate amount of time. STM and CTM have the highest training times due to their complexity.

3.1 LDA-Identified Topics in Overtourism Literature

Latent Dirichlet Allocation (LDA) have been applied to identify the major topics in the literature related to overtourism. The optimal number of topics have been identified by plotting the coherence score for models computed with topic numbers ranging from 1 to 15. The highest coherence score has been achieved with 8 topics, as the most suitable number as shown in Fig. 2.

Below there are the main themes, their associated key terms, and a description of each topic.

Topic 1 - Tourismphobia, Challenges to Local Identity and Heritage

– **Key terms:** *enterprise, business, tourismphobia, cultural, historical, resident, identity, sustainable*
– **Description:** This topic emphasizes the cultural and identity impacts of tourism, including the concept of "tourismphobia" and its challenges to local identity and heritage preservation.

Topic 2 - Sustainable and Innovative Tourism Strategies

– **Key terms:** *innovative, involvement, hospitality, tackle, strategy, sustainable, postcovid, stakeholder*
– **Description:** This topic focuses on the development of innovative and sustainable tourism strategies, particularly in the context of post-COVID-19 recovery efforts.

Topic 3 - New Smart Technology in Managing Events and Crises

– **Key terms:** *event, smart, host, crisis, covid, technology, sustainability, framework*

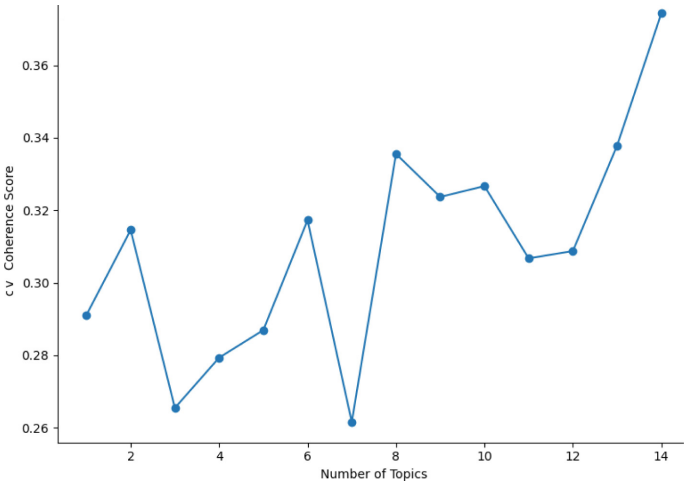

Fig. 2. CV coherence score for different numbers of topics identified with LDA - line chart showing the relationship between the number of topics and the coherence score.

– **Description:** This topic explores how technology and smart solutions contribute to managing events and crises, particularly within tourism contexts.

Topic 4 - Environmental Challenges in Tourism

– **Key terms:** *water, transport, climate, cruise, adaptation, season, traffic, heritage, unesco*
– **Description:** This topic discusses environmental issues in tourism, including the impacts on water resources, climate change, traffic, and UNESCO heritage sites.

Topic 5 - Urban Tourism Impact on Social Experiences and Communities

– **Key terms:** *city, resident, destination, tourist, urban, social, development, community*
– **Description:** This topic centers on urban tourism's social impacts, including resident experiences, community development, and the effects of increased tourist presence.

Topic 6 - Cruise Tourism and Environmental Impact

– **Key terms:** *cruise, water, port, venice, degrowth, policy, pollution, climate*
– **Description:** This topic focuses on the environmental implications of cruise tourism, especially concerning pollution, climate impact, and policy considerations in sensitive areas like Venice.

Topic 7 - Tourist Attraction Management and Resilience

- **Key terms:** *located, innovation, attraction, destination, resilience, service, transport, public*
- **Description:** This topic examines the management of tourism attractions with an emphasis on resilience, innovation, and public infrastructure to support sustainable tourism.

Topic 8 - Tourism Capacity and Management Techniques

- **Key terms:** *learning, technique, carrying capacity, tool, municipality, island, indicator, capacity*
- **Description:** This topic discusses strategies for managing tourism capacity, focusing on techniques and tools to ensure sustainability in areas with limited resources.

Wordclouds and barplots for the best 10 words in the topics we have unveiled are displayed in Figs. 3 and 4.



Fig. 3. Wordclouds visualization with eight topics identified with LDA - each topic contain a cluster of words varying in size, indicating their relevance.

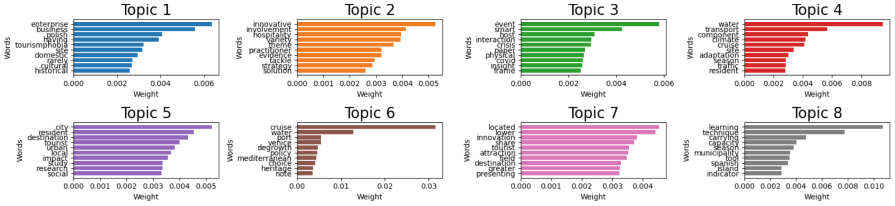

Fig. 4. Barchart plots for the best 10 words of each topic identified with LDA - the charts illustrate the weight of each word within its respective topic.

3.2 LSI-Identified Topics in Overtourism Literature

Latent Semantic Indexing (LSI) has been applied to identify main topics in the overtourism literature. The optimal number of topics have been determined by plotting the coherence score for models computed with topic numbers ranging from 1 to 15. The highest coherence score has been achieved with 6 topics, as the most suitable number as shown in Fig. 5.

Below there are the identified themes, their key terms, and a brief description of each topic.

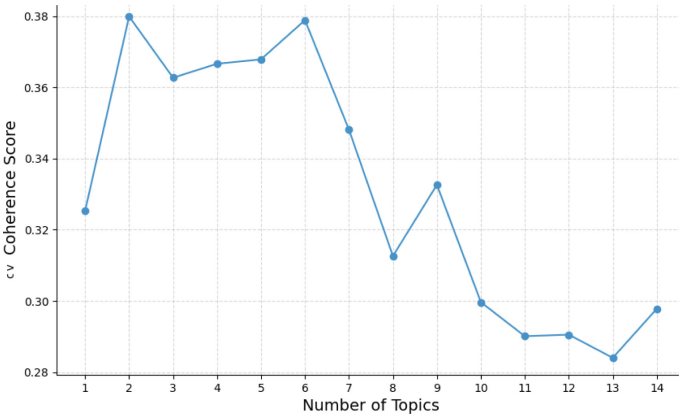

Fig. 5. CV coherence score for different numbers of topics identified with LSI - line chart showing the relationship between the number of topics and the c v coherence score.

Topic 1 - General Impact of Tourism and Development of a More Sustainable Approach

- **Key terms:** *impact, analysis, sustainability, model, policy, development, approach*
- **Description:** This topic broadly examines tourismâĂŹs impact, focusing on sustainability and the role of policies and models in guiding a sustainable approach within tourism studies.

Topic 2 - Urban Tourism and Resident Impact

- **Key terms:** *tourist, city, resident, impact, urban, social, development*
- **Description:** This topic tmphasizes the effects of tourism on urban areas and local residents, particularly the social impacts and challenges faced by cities with high levels of tourist activity.

Topic 3 - Urban Growth, Conflict, and Sustainability

- **Key terms:** *city, urban, growth, conflict, sustainable, problem, social*
- **Description:** This topic focuses on urban growth issues related to tourism, such as conflicts, sustainability, and social challenges emerging from increased tourism in city environments.

Topic 4 - Heritage and Cultural Tourism

- **Key terms:** *heritage, cultural, city, resident, development, pandemic*
- **Description:** This topic examines cultural and heritage tourism, discussing interactions with local development and community, as well as potential impacts stemming from the pandemic.

Topic 5 - Tourism Management and Development Models

- **Key terms:** *sustainable, model, approach, management, effect*
- **Description:** This topic covers tourism management strategies, focusing on sustainable approaches and development models aimed at managing tourism's impact on destinations.

Topic 6 - Community and Social Effects of Tourism

- **Key terms:** *community, effect, resident, social, impact, management*
- **Description:** This topic discusses the social and community-level effects of tourism, with an emphasis on local resident experiences and community-based management practices.

Wordclouds and barplots for the best 10 words in the topics we have unveiled are displayed in Figs. 6 and 7.

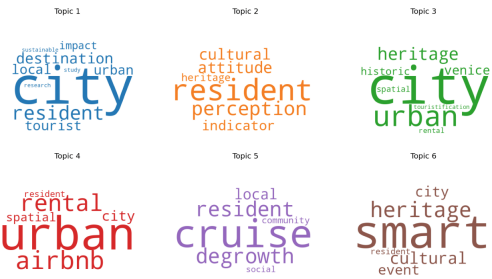

Fig. 6. Wordclouds for topics identified with LSI

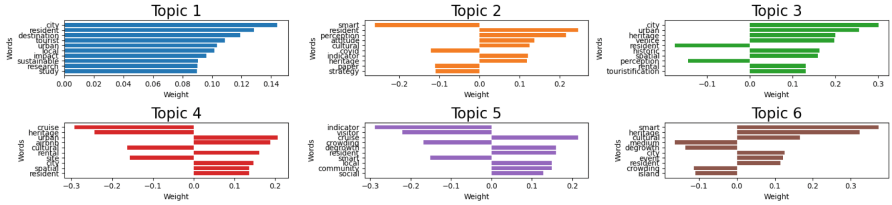

Fig. 7. Barplots for the best 10 words of each topic identified with LSI

3.3 HDP-Identified Topics in Overtourism Literature

The Hierarchical Dirichlet Process (HDP) model has revealed several key themes in the overtourism literature. In the following, each topic has been summarized with its key terms and description.

Topic 1 - Overview of Tourism and Local Impact

- **Key terms:** *tourist, destination, study, city, local, cultural, resident, impact, research*
- **Description:** This topic provides a broad overview of tourism, focusing on the relationships between tourists, local residents, and destinations, as well as the social and cultural impacts of tourism.

Topic 2 - Tourism Conditions and Quality Issues

- **Key terms:** *condition, impact, highlighted, suggests, quality, domestic, growing, tourist*
- **Description:** This topic emphasizes tourism conditions and challenges, including quality issues, the rise of domestic tourism, and increasing demands on tourism infrastructure.

Topic 3 - Tourism Patterns, Pricing, and Resident Dynamics

- **Key terms:** *resident, tourist, phenomenon, pattern, growth, attraction*
- **Description:** This topic discusses tourist behavior patterns, including pricing dynamics, tourism growth, and the interaction between tourists and local residents.

Topic 4 - Historical and Resilience-Focused Tourism

- **Key terms:** *position, context, historic, resilience, component, externality, stage*
- **Description:** This topic focuses on historical and resilience-oriented tourism, including the impact of tourism on heritage sites and efforts to manage externalities in historic areas.

Topic 5 - Tourism-Related Environmental Issues and Planning Concerns

- **Key terms:** *problem, decline, pollution, planning, model, cultural*
- **Description:** This topic examines environmental issues tied to tourism, including pollution, the decline of certain tourist locations, and the importance of strategic planning and modeling for sustainable management.

Topic 6 - Inclusivity in Tourism Development

- **Key terms:** *reaction, building, opportunity, inclusive, interview, evidence, historical*
- **Description:** This topic highlights inclusivity in tourism development, considering opportunities, stakeholder reactions, and community involvement based on historical evidence.

Topic 7 - Tourismphobia and Protection of Local Interests

- **Key terms:** *tourismphobia, planning, aspect, destination, protect, owner*

– **Description:** This topic focuses on "tourismphobia", or negative attitudes toward excessive tourism, and strategies to protect local interests and manage destinations effectively.

Topic 8 - Data Analysis and Innovation in Tourism Research

– **Key terms:** *analyse, data, innovative, primary, experiment, statistical*
– **Description:** This topic explores data-driven methods and innovation in tourism research, with a focus on data analysis, statistical techniques, and experimental approaches.

Wordclouds and barplots for the best 10 words in the topics we have unveiled are displayed in Figs. 8 and 9.



Fig. 8. Wordclouds for topics identified with HDP

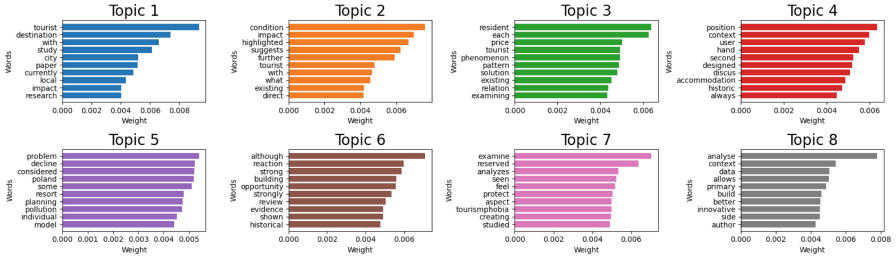

Fig. 9. Barplots for the best 10 words of each topic identified with HDP

3.4 NMF-Identified Topics in Overtourism Literature

The Non-Negative Matrix Factorization (NMF) model has uncovered several key themes within the overtourism literature. The optimal number of topics has been determined by plotting the coherence score for models computed with topic numbers ranging from 1 to 15. The highest coherence score was achieved with 10 topics, as the most suitable number as shown in Fig. 10.

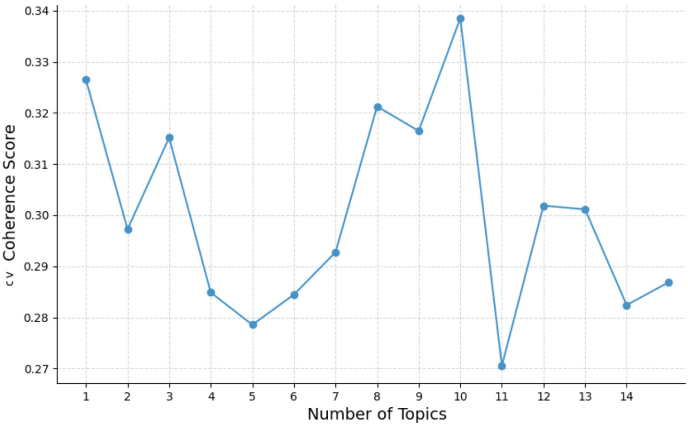

Fig. 10. CV coherence score for different numbers of topics identified with NMF

In the following, there is a summary of each topic, highlighting its key terms and description.

Topic 1 - Tourism Design, Experience, and Practical Implications

- **Key terms:** *experience, design, hospitality, visitor, methodology, practitioner*
- **Description:** This topic addresses the design and practical aspects of tourism, focusing on visitor experiences, industry practices, and methodological approaches.

Topic 2 - Urban Tourism and Housing Impact

- **Key terms:** *urban, conflict, gentrification, touristification, housing, congestion*
- **Description:** This topic focuses on the impacts of tourism in urban areas, particularly conflicts related to gentrification, housing shortages, and urban congestion.

Topic 3 - Environmental Issues and Rural Tourism

- **Key terms:** *medium, policy, rural, news, discourse, planning, behavior*
- **Description:** This topic examines media and policy discourse surrounding tourismâĂŹs environmental impact, particularly in rural settings.

Topic 4 - Cruise Tourism, Degrowth, and Technological Governance

- **Key terms:** *cruise, degrowth, smart, port, governance, technology*
- **Description:** This topic centers on cruise tourism, with a focus on degrowth strategies, technology integration, and governance.

Topic 5 - Seasonal Demand, Carrying Capacity, and Climate Concerns

- **Key terms:** *season, capacity, climate, crowding, park, demand*
- **Description:** This topic discusses seasonal demand, capacity management, and climate-related issues, particularly in high-traffic tourist areas.

Topic 6 - Post-Pandemic Tourism and Innovation

- **Key terms:** *sharing, economy, pandemic, sustainability, innovation, covid*
- **Description:** This topic explores the impacts of the COVID-19 pandemic on tourism, highlighting the sharing economy and sustainable innovations.

Topic 7 - Resident Attitudes and Cultural Impact of Tourism

- **Key terms:** *site, resident, cultural, survey, effect, crowding*
- **Description:** This topic addresses resident attitudes towards tourism, cultural impacts, and issues related to crowding and community responses.

Topic 8 - Heritage and Mass Tourism at UNESCO Sites

- **Key terms:** *heritage, cultural, unesco, site, venice, mass*
- **Description:** This topic focuses on heritage tourism and the challenges of managing mass tourism at UNESCO World Heritage Sites.

Topic 9 - Short-Term Rentals, Regulation, and Urban Impact

- **Key terms:** *rental, airbnb, regulation, pandemic, historic, population*
- **Description:** This topic examines the impact of short-term rentals on urban areas, with a focus on regulatory responses and effects on historic sites.

Wordclouds and barplots for the best 10 words in the topics we have unveiled are displayed in Figs. 11 and 12.



Fig. 11. Wordclouds for topics identified with NMF

3.5 STM-Identified Topics in Overtourism Literature

The Structural Topic Model (STM) analysis has revealed several key themes in the overtourism literature. A diagnostic of the model has been run with a range of topics between 2 and 10, including Held-out Likelihood, Residuals, and Semantic Coherence. The best balance between these metrics have been found at 6, as shown in Fig. 13.

Each topic is detailed below, including its focus, key words, and an explanation.

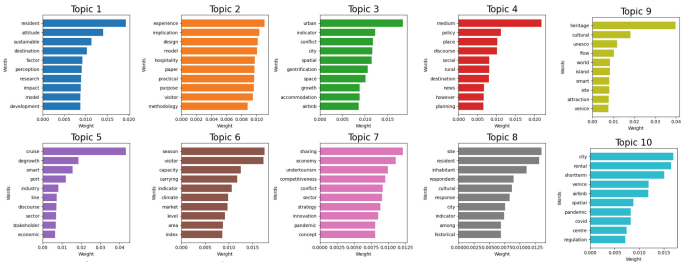

Fig. 12. Barplots for the best 10 words of each topic identified with NMF

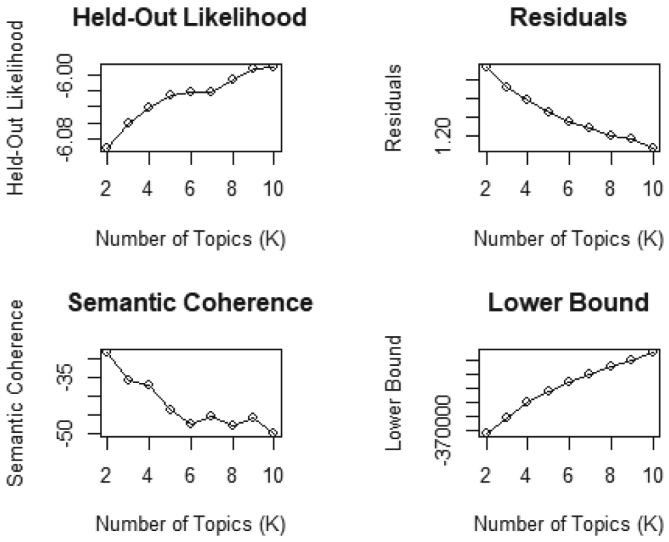

Fig. 13. Diagnostic values for different numbers of topics identified with STM

Topic 1 - Overtourism Analysis

- **Key Words:** *overtourism, tourism, urban, social, public policy, conflict, research, methodology, discourse, stakeholder, narrative, medium, analysis, interaction, space, network*
- **Description:** This topic explores the phenomenon of overtourism in urban settings, emphasizing social dynamics, conflicts, and public policy implications that arise when tourism reaches unsustainable levels. It highlights methodologies, discourse, and stakeholder engagement in urban tourism research.

Topic 2 - Sustainable Tourism: Overtourism and Undertourism

- **Key Words:** *overtourism, undertourism, sustainable development, innovation, economy, literature review, problem, concept, scientific, enterprise, solution, sharing, business, research, model*
- **Description:** This topic focuses on sustainable tourism practices, particularly regarding overtourism and undertourism. It discusses economic factors, innovation, and the literature addressing these challenges, highlighting sustainable development strategies for balanced tourism growth.

Topic 3 - Cultural Heritage and Tourism

- **Key Words:** *tourist, visitor, heritage, cultural, site, cruise, environmental analysis, UNESCO, traffic, attraction, location, natural, flow, park, season, index*
- **Description:** This topic examines the relationship between cultural heritage sites and tourism, focusing on visitor interactions and the environmental impacts. It emphasizes the importance of managing cultural and natural heritage, especially UNESCO sites, to preserve them for future generations.

Topic 4 - Community Engagement in Sustainable Tourism

- **Key Words:** *tourism, sustainable, destination, stakeholder, community, strategy, management, governance, resilience, planning, practice, finding, smart, approach, industry, crisis*
- **Description:** This topic discusses the role of local communities in sustainable tourism practices, with a focus on stakeholder involvement, governance strategies, and resilience. It explores how communities can address tourism challenges through smart planning and collaborative efforts.

Topic 5 - Urban Tourism Impacts on Housing

- **Key Words:** *city, tourism, urban, spatial, negative impact, rental, gentrification, pandemic, citizen, housing, short-term rental, touristification, concentration, analysis, life*
- **Description:** This topic examines the impacts of tourism in urban environments, particularly short-term rentals, gentrification, and the effects of the pandemic. It highlights the negative consequences of tourism concentration and discusses strategies to maintain balanced community dynamics.

Topic 6 - Relationship Between Tourists and Residents

- **Key Words:** *tourist, resident, perception, overtourism, crowding, satisfaction, attitude, survey, indicator, influence, significant, relationship, questionnaire, research, effect, local*
- **Description:** This topic explores residents' perceptions of overtourism and its effects on communities. It focuses on attitudes toward crowding, satisfaction, and public perception, using surveys to assess local views on tourism-related changes.

Wordclouds for the best 10 words in the topics we have unveiled are displayed in Fig. 14.



Fig. 14. Wordclouds for topics identified with STM

3.6 CTM-Identified Topics in Overtourism Literature

The following topics have been identified from the Correlated Topic Model (CTM) analysis, each providing insights into different aspects of overtourism and its impacts. The best number of topics is 6 based on the same metrics of STM, shown in Fig. 13.

Topic 1 - Overtourism Dynamics

- **Key Words:** *tourism, overtourism, destination, city, tourist, local, result, sustainable, effect, development, resident, analysis, research, covid, impact, future, negative, model, pandemic*
- **Description:** This topic examines the phenomenon of overtourism, where the volume of tourists overwhelms local resources and infrastructure. It explores the implications for urban environments, sustainability, and local residents, including the impact of events such as the COVID-19 pandemic on tourism patterns.

Topic 2 - Overtourism Impact on Local Communities

- **Key Words:** *study, social, overtourism, research, urban, tourist, finding, approach, paper, impact, local, policy, public, literature, area, sustainability, article, destination, community, factor*
- **Description:** This topic focuses on the social aspects of overtourism, examining how increased tourist activity affects local communities. It highlights the importance of understanding social dynamics, public policies, and community engagement for sustainable tourism management.

Topic 3 - Visitor Analysis in Tourism

- **Key Words:** *tourism, tourist, visitor, phenomenon, study, destination, analysis, flow, well, urban, area, impact, approach, model, result, research, future, sustainable, city, event*
- **Description:** This topic investigates tourist behaviors and their interactions with destinations, covering the impact of visitor flow on urban areas. It includes demographic insights, motivations, and overall tourist experience, helping to understand and manage tourism effectively.

Topic 4 - Sustainable Tourism Development

- **Key Words:** *tourism, destination, study, sustainable, research, based, management, development, space, indicator, growth, article, data, analysis, model, community, factor, economic, local, approach*
- **Description:** This topic covers sustainable tourism practices aimed at minimizing environmental impacts while maximizing social and economic benefits for communities. It discusses strategic management approaches and stakeholder roles in achieving sustainable tourism growth.

Topic 5 - Community Impact and Management

- **Key Words:** *impact, study, resident, community, tourism, negative, level, social, purpose, development, analysis, stakeholder, value, economic, perception, data, management, factor, result, finding*
- **Description:** This topic explores tourism's impact on local communities and the strategies used to manage these changes. It highlights the role of stakeholders in addressing social impacts and emphasizes community involvement in decision-making to address tourism challenges.

Topic 6 - Overtourism Research and Analysis

- **Key Words:** *overtourism, study, tourism, approach, research, paper, problem, resident, development, local, management, analysis, social, finding, concept, term, many, economic, data, result*
- **Description:** This topic focuses on research related to overtourism and its impacts, including methodologies, frameworks, and indicators to measure tourism effects. The aim is to inform policymakers and stakeholders about the challenges and mitigation strategies for overtourism.

Wordclouds for the best 10 words in the topics we have unveiled are displayed in Fig. 15.

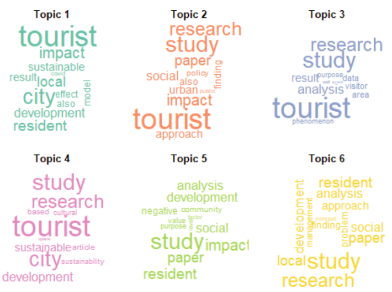

Fig. 15. Wordclouds for topics identified with CTM

4 Discussion

The analysis of tourism-related themes has revealed several prominent topics that consistently emerge across various models (see Tables 1 and 2).

Table 1. Number of topics per modeling technique

Techniques	Number of Topics	Techniques	Number of Topics
LDA	8	NMF	10
LSI	6	STM	6
HDP	8	CTM	6

Table 2. Most common topics over modeling techniques

Topics	Modeling Techniques
Overtourism and its impact	LSI, HDP, STM, CTM
Cultural Heritage and Identity	LSI, HDP, NMF, STM
Sustainable Tourism Practices	LDA, LSI, HDP, NMF, STM CTM
Urban Dynamics and Challenges	LDA, LSI, HDP, STM, CTM
Community Perspective and Resident Impact	LSI, HDP, NMF, CTM
Economic Factors and Development	LSI, STM, HDP, NMF
Visitor Experience and Engagement	LDA, LSI, NMF, STM, CTM

Below there are the key themes:

1. **Overtourism and Its Impacts:** This theme highlights the social dynamics, community impact, and management strategies associated with overtourism. It emphasizes the negative consequences such as crowding, resident dissatisfaction, and the strain on local infrastructure, underscoring the urgency of addressing these issues.
2. **Cultural Heritage and Identity:** Many topics underscore the importance of cultural heritage and the identity of local communities affected by tourism. The relationship between tourism and the preservation of cultural and natural heritage, especially in relation to UNESCO sites, is a recurring concern that calls for protective measures and responsible tourism practices.
3. **Sustainable Tourism Practices:** Across all analyses, there is a strong emphasis on sustainable tourism strategies. This includes discussions around innovative solutions to mitigate tourismâĂŹs adverse effects, the role of stakeholders, and governance approaches that aim to achieve sustainable development while benefiting both tourists and local communities.
4. **Urban Dynamics and Challenges:** Urban tourism poses significant challenges, with several topics addressing the issues faced by cities due to tourism.

Concerns such as short-term rentals, gentrification, and social dynamics within urban environments highlight the necessity for comprehensive urban planning and management strategies to foster a balance between tourism and resident needs.

5. **Community Perspectives and Resident Impact:** The analysis frequently explores the perceptions and attitudes of residents toward tourism. It delves into how tourism impacts local communities, including social implications, quality of life, and the balance between tourism benefits and the challenges it brings, revealing the need for inclusive decision-making processes.
6. **Economic Factors and Development:** Economic aspects of tourism, particularly development challenges and the sharing economy, are frequently discussed in relation to overtourism. This theme examines the economic implications of tourism growth on local communities, emphasizing the importance of balancing economic benefits with sustainable practices that do not compromise local cultures and environments.
7. **Visitor Experience and Engagement:** Several topics address the experiences of tourists themselves, exploring their interactions with cultural sites. The significance of visitor satisfaction, and how it relates to the cultural and environmental context, is a key area of focus, suggesting that positive tourist experiences can enhance community support for tourism initiatives.

The analysis reveals that overtourism is a multifaceted issue that significantly impacts cultural heritage, community dynamics, and sustainable tourism practices. One of the critical insights is the interdisciplinary nature of tourism research; themes such as sustainability, cultural impact, and community perspectives are deeply interconnected. This interrelationship suggests that effective tourism management must consider multiple facets of tourism, recognizing that changes in one area can have ripple effects on others.

Moreover, the findings indicate a pressing need for informed policy-making and robust governance frameworks that incorporate the diverse experiences and needs of all stakeholders involved in tourism. Policymakers must engage with local communities, tourism businesses, and visitors to develop strategies that address the complex challenges posed by overtourism while ensuring the long-term viability of tourism as an economic driver.

Additionally, the recurring issues associated with overtourism and its management highlight areas ripe for future research. There is a significant opportunity to explore innovative solutions that support sustainable tourism while preserving local cultures and environments. Future studies could focus on the development of best practices for stakeholder engagement, community resilience, and sustainable tourism models that benefit both residents and tourists alike.

5 Conclusions

The application of topic modeling techniques successfully identified underlying themes within the reviewed literature. Many topics overlapped across different models, reflecting the complex and multidisciplinary nature of the subject.

The analysis underscores the need for comprehensive strategies and collaboration among stakeholders to address the challenges of overtourism effectively. By fostering a tourism model that prioritizes sustainability, respects cultural heritage, and considers the perspectives of local communities, it is possible to create a more equitable and responsible tourism industry that serves the interests of both visitors and residents.

In the future, a classifier will be developed to categorize new research publications into distinct thematic topics. The generated topic probability distributions from the modeling techniques will be used as input features for various classifiers. BERT-based embeddings will be also used to capture nuanced semantic information from abstracts.

Acknowledgments. The authors thank INFN for having funds this research with the Summer Student Internship Program.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Doxey, G.V.: A causation theory of visitor-resident irritants: methodology and research inferences. In: Proceedings of 6th Annual Conference on Travel and Tourism Research Association, pp. 195–198 (1975)
2. Goeldner, C.R., Ritchie, J.: Tourism: Principles. John Wiley & Sons, Practices, Philosophies (2016)
3. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 601–608 (2001)
4. Papadimitriou, C.H., Tamaki, H., Raghavan, P., Vempala, S.: Latent semantic indexing: a probabilistic analysis. In: Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, PODS '98, pp. 159–168 (1998)
5. Lee, D., Seung, H.: Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–791 (1999)
6. Teh, Y., Jordan, M., Beal, M., Blei, D.: Hierarchical Dirichlet processes. Mach. Learn., 1–30 (2006)
7. Roberts, M., et al.: Structural topic models for open ended survey responses. Am. J. Polit. Sci. **58**(4), 1064–1082 (2014)
8. Blei, D., Lafferty, J.: A correlated topic model of science. Ann. Appl. Stat. **1** (2007)
9. Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 262–272 (2011)
10. Roder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408 (2015)