ISSN(O): 2582-7162
ISSN(P): 3139-163X

Original Article

A Data-Driven Analysis of Tourist Interests: Evidence from Uzbekistan

Sharipov Uktamjon Akhtam Ugli

[1]Nanjing University of Information Science and Technology

[2] Bukhara State University, Uzbekistan

Corresponding Author: Sharipov Uktamjon Akhtam Ugli, Nanjing University of Information Science and Technology, Bukhara State University, Uzbekistan

E mail id: uktamjonsharipov1@gmail.com, 20215225034@nuist.cn

Abstract

Tourism is a vital contributor to Uzbekistan's economy, yet there is a lack of data-driven research addressing tourist interests. This study analyzes over 560 online reviews from TripAdvisor and Booking.com to identify key factors influencing tourist experiences. Using Python-based web scraping tools, the reviews were collected and cleaned into a dataset of 282 entries after removing duplicates, non-English comments, and errors. Latent Dirichlet Allocation (LDA) was applied to extract six main topics, while K-means clustering grouped frequent words into nine clusters. The analysis identified nine key themes critical to tourist satisfaction: language barriers, public transport, accommodation quality, safety, heritage site accessibility, sustainable tourism, local food, shopping opportunities, and information availability. These findings provide actionable insights to address tourist concerns, improve satisfaction, and support sustainable growth in Uzbekistan's tourism sector.

Keywords: Big Data, Text mining, LDA, Python, K-means.

How to cite this article: Sharipov, U. A. A Data-Driven Analysis of Tourist Interests: Evidence from Uzbekistan. SDES-IJIR; 2026; 7-2:1419-1430

Submitted: 2-April-2026; **Accepted:** 29-April-2026; **Published:** 30-April-2026

Introduction

1.1 Research Background

Tourism is one of the world's largest industries, contributing significantly to socio-economic development by creating jobs, stimulating economic growth, and fostering cultural exchange. Tourist satisfaction remains critical to the industry's success, as satisfied tourists are more likely to return, recommend destinations, and spend more.

Globalization and information technology have generated vast amounts of data from online reviews and social media, offering valuable insights into tourist behaviors and preferences. Traditional methods are insufficient for handling this volume, making data-driven approaches necessary.

Uzbekistan, located at the heart of Central Asia along the ancient Silk Road, possesses rich historical, cultural, and natural assets, including UNESCO World Heritage sites such as Samarkand and Bukhara. In the post-Soviet era, the country has promoted tourism through visa liberalization and infrastructure development. According to the World Travel & Tourism Council (WTTC, 2023), tourism contributed approximately 5.2% to Uzbekistan's GDP in 2022, with projections reaching 7% by 2033. Tourist arrivals rose to 6.6 million in 2023 and reached 9.7 million in the first ten months of 2025, according to the National Statistics Committee.

https://doi.org/10.47997/sdes-ijir/7.2.2026:1419-1430

Figure 1: Tourist Arrivals in Uzbekistan (2020-2024) (National Statistics Committee data; arrivals in thousand: 2020: 1504.1, 2021: 1881.3, 2022: 5232.8, 2023: 6626.3, 2024: 7957.2)

Despite this growth, comprehensive data-driven research on tourist satisfaction in Uzbekistan remains limited. Most existing studies focus on operational efficiency rather than nationwide visitor experiences. This study aims to fill this gap by applying text mining techniques to online reviews.

1.2 Problem Statement

Although tourist arrivals continue to increase, Uzbekistan remains underrepresented in global tourism. Challenges in infrastructure, service quality, and information availability affect visitor satisfaction. Average TripAdvisor ratings hover around 4.1 out of 5, indicating room for improvement.

The lack of systematic, data-driven insights limits the ability to address diverse tourist needs and hinders sustainable sector growth. This study seeks to answer the following research questions:

1. What primary factors influence tourist interests in Uzbekistan based on online reviews?
2. How do infrastructure, services, and offerings align with visitor expectations?
3. Are there noticeable differences in interests among different tourist groups?

1.3 Research Objectives The main objective is to conduct a data-driven analysis of tourist interests in Uzbekistan using online reviews and provide actionable recommendations.

Sub-objectives: Analyze unstructured text using LDA for topic extraction. Apply K-means clustering to segment tourist preferences. Offer practical insights for policy and marketing improvements.

1.4 Significance of Research

This research can support Uzbekistan's tourism sector by identifying key areas for improvement, enhancing tourist satisfaction, and promoting sustainable development. The findings will be useful for policymakers, tourism operators, and researchers working on Central Asian tourism.

1.5 Research Content

The study consists of four main chapters. Chapter 1 presents the introduction and research background. Chapter 2 reviews relevant literature and describes the research methods. Chapter 3 details data collection, pre-processing, and analysis results. Chapter 4 provides conclusions, limitations, and recommendations for future research.

https://doi.org/10.47997/sdes-ijir/7.2.2026:1419-1430

Chapter Two: Literature Review and Research Methods

2.1 Data-Driven Approaches for Tourist Interests Analysis

Data-driven approaches have become powerful tools for understanding tourist interests by leveraging large volumes of unstructured data from online platforms. Online reviews on TripAdvisor, Booking.com, and social media provide rich insights into genuine tourist experiences (Xiang et al., 2017; El-Said, 2020). However, challenges such as fake reviews, cultural differences, and data volume remain significant.

2.2 Sentiment Analysis and Natural Language Processing (NLP)

NLP and sentiment analysis enable automatic processing of large text datasets to extract opinions and emotions. These techniques help identify pain points and positive aspects of tourist experiences (Jurafsky & Martin, 2019; Liu, 2012). They offer scalable solutions for tourism research but require careful handling due to language complexity.

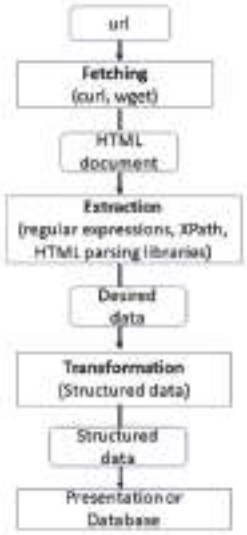
2.3 Machine Learning in Tourist Interests Analysis

Machine learning techniques, particularly Latent Dirichlet Allocation (LDA) for topic modeling and K-means for clustering, have been widely used to discover hidden patterns in tourist reviews (Blei et al., 2003; Dolnicar, 2002). While these methods are well-established globally, their application in Central Asia, especially Uzbekistan, remains limited. This study applies LDA and K-means to analyze tourist feedback in the Uzbek context.

2.4 Text Acquisition and Pre-Processing Technology

Web scraping was used to collect review data. The process involves sending HTTP requests, parsing HTML content, and storing structured data using Python libraries such as Selenium, BeautifulSoup, Requests, and Pandas.

Figure 2.1 Web Scraping process (Persson, 2019)



Stop words were removed and lemmatization was applied during pre-processing to prepare the text for analysis.

2.5 Latent Dirichlet Allocation (LDA)

Topic Modelling LDA is a generative probabilistic model used for discovering abstract topics in a collection of documents (Blei et al., 2003). After data cleaning, a document-term matrix was created and the LDA model was trained. Topic coherence scores were used to determine the optimal number of topics.

https://doi.org/10.47997/sdes-ijir/7.2.2026:1419-1430

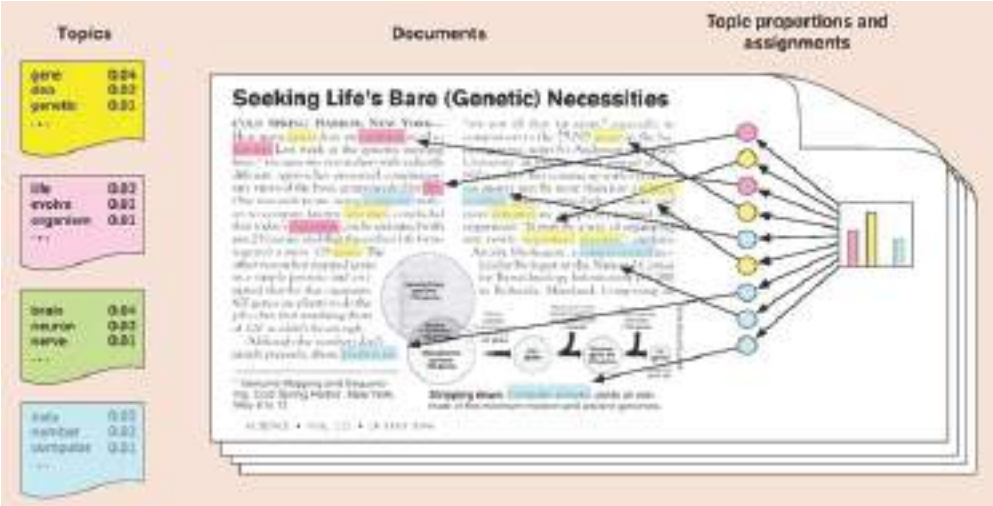


Figure 2.2 The intuition behind Latent Dirichlet Allocation (Blei, 2012)

2.6 K-means Clustering

K-means is an unsupervised clustering algorithm that partitions data into K groups by minimizing the distance between data points and cluster centroids. In this study, document-topic distributions from the LDA model served as input features for K-means clustering. The Elbow method was used to determine the optimal number of clusters.

Chapter Three: Web Scraping and LDA Result Analysis

3.1 Acquisition and Pre-Processing of Online Comment Data

Data was collected from TripAdvisor and Booking.com using Python web scraping tools, resulting in approximately 560 reviews. After cleaning for duplicates, non-English comments, and irrelevant content, the final dataset contained 282 usable reviews. Stop words were removed using the NLTK library, and lemmatization was performed for normalization.

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', '
yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself
', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those
', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a',
'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against',
'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'o
ff', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both'
, 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'ver
y', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain'
, 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "
haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "sh
ouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

Figure 3.1: List of "Stop Words" excluded from text data

```
spent full days uzbekistan really enjoyed sight seeing loved khivakokhoreasmarkand tashkentenjoyed drive ferghana valley customized itinerary worked well l
ittle hectic seniors plus show eleonosi samarkand must tasted local specialities good flavour architectureculture history uzbekistan guide saeedeen knowledg
eable well organized memorable lovely trip
['spend', 'full', 'day', 'enjoy', 'sight', 'seeing', 'love', 'customize', 'work', 'little', 'hectic', 'senior', 'show', 'taste', 'local', 'specialty', 'goo
d', 'flavour', 'architectureculture', 'history', 'knowledgeable', 'organize', 'memorable', 'lovely', 'trip']
```

Figure 3.2 Words and sentences before and after tokenized lemmatization

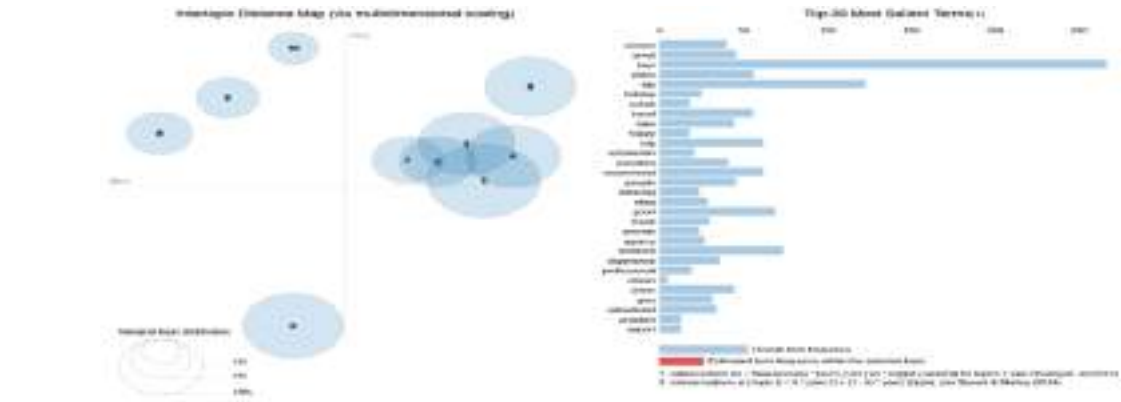
3.2 LDA Model Parameters and Results

The LDA model was initially trained with 10 topics. Due to overlapping topics, the number was optimized using coherence scores, resulting in 6 topics as optimal.

https://doi.org/10.47997/sdes-ijir/7.2.2026:1419-1430

Table 3.1 LDA topic Results (10 topics)

[(Topic 1:
'0.025*"tour" + 0.023*"guide" + 0.021*"place" + 0.020*"take" + 0.014*"tashkent" + 0.013*"visit" + 0.013*"time" + 0.012*"trip' 0.010*"day"),
(Topic 2:
'0.023*"travel" + 0.022*"trip" + 0.021*"tour" + 0.019*"guide" + 0.014*"tashkent" + 0.012*"time" + 0.012*"history" 0.010*"experience" + 0.010*"culture" + 0.009*"book"'),
(Topic 3:



'0.038*"trip" + 0.017*"guide" + 0.014*"good" + 0.014*"tour" + 0.013*"uzbekistan" + 0.010*"driver" + 0.009*"perfect' 0.009*"agency" + 0.009*"excellent" + 0.008*"provide"'),
(Topic 4:
'0.026*"place" + 0.017*"tour" + 0.014*"good" + 0.012*"people" + 0.008*"uzbekistan" + 0.008*"travel" + 0.008*"ancier 0.008*"country" + 0.007*"great" + 0.007*"uzbek"'),
(Topic 5:
'0.047*"tour" + 0.025*"guide" + 0.020*"recommend" + 0.018*"trip" + 0.0[illegible]*"service" + 0.014*"amazing" + 0.014*"experience" 0.013*"thank" + 0.012*"country" + 0.012*"great"'),
(Topic 6:
'0.044*"tour" + 0.017*"excellent" + 0.016*"guide" + 0.014*"trip" + 0.012*"hotel" + 0.012*"travel" + 0.011*"time" + 0.010*"tashkent" + 0.009*"make" + 0.007*"visit"'),
(Topic 7:
'0.024*"hotel" + 0.021*"holiday" + 0.018*"guide" + 0.016*"people" + 0.015*"local" + 0.013*"trip" + 0.010*"give" + 0.009*"book" + 0.008*"feel" + 0.008*"visit"'),
(Topic 8:
'0.050*"tour" + 0.020*"guide" + 0.017*"city" + 0.015*"visit" + 0.014*"good" + 0.014*"tashkent" + 0.013*"hotel" + 0.012*"time" 0.011*"agency" + 0.010*"recommend"'),
(Topic 9:
'0.061*"tour" + 0.024*"city" + 0.022*"guide" + 0.021*"trip" + 0.020*"visit" + 0.015*"tashkent" + 0.013*"recommend" + 0.012*"good" + 0.010*"local"),
(Topic 10:
'0.037*"service" + 0.029*"great" + 0.024*"uzbek" + 0.021*"tour" + 0.016*"trip" + 0.016*"happy" + 0.014*"citizen" + 0.009*"free" + 0.009*"foreigner" + 0.009*"provide"')]

Figure 3.3 Graphic visualization of Topics

3.3 Optimal Number of Topics

Topic coherence analysis showed that 6 topics provided the best balance between interpretability and model performance.

https://doi.org/10.47997/sdes-ijir/7.2.2026:1419-1430

Table 3.2 Coherence Score for Selected Number of Topics

Coherence Score of 23 topics for Optimal topic evaluation
Num Topics = 2 has Coherence Value of 0.3388
Num Topics = 3 has Coherence Value of 0.3552
Num Topics = 4 has Coherence Value of 0.3492
Num Topics = 5 has Coherence Value of 0.3434
Num Topics = 6 has Coherence Value of 0.3522
Num Topics = 7 has Coherence Value of 0.3396
Num Topics = 8 has Coherence Value of 0.3325
Num Topics = 9 has Coherence Value of 0.3557
Num Topics = 10 has Coherence Value of 0.354
Num Topics = 11 has Coherence Value of 0.3405
Num Topics = 12 has Coherence Value of 0.334
Num Topics = 13 has Coherence Value of 0.3611
Num Topics = 14 has Coherence Value of 0.3484
Num Topics = 15 has Coherence Value of 0.3326
Num Topics = 16 has Coherence Value of 0.3655
Num Topics = 17 has Coherence Value of 0.3434
Num Topics = 18 has Coherence Value of 0.3555
Num Topics = 19 has Coherence Value of 0.3475
Num Topics = 20 has Coherence Value of 0.3446
Num Topics = 21 has Coherence Value of 0.3422
Num Topics = 22 has Coherence Value of 0.3462
Num Topics = 23 has Coherence Value of 0.3451
Num Topics = 24 has Coherence Value of 0.3435

Graphic representation is shown on figure 3.4 and it can be inferred that model with 6 topics has the highest coherence score, so the optimal K=6 for current set of data.

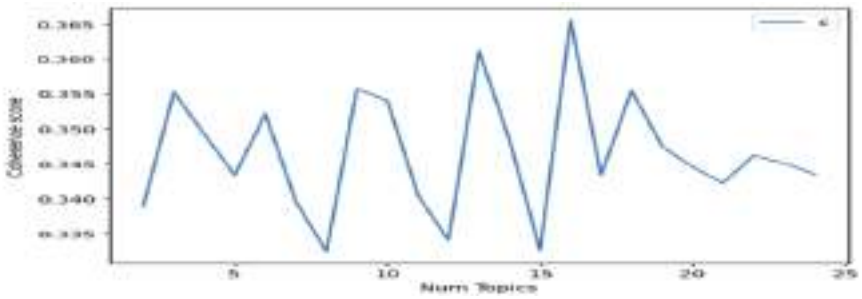


Figure 3.4 Coherence Score representation

3.4 LDA with Optimal Topics and K-means Clustering

The optimized LDA model with 6 topics was used to generate document-topic distributions. These distributions served as input for K-means clustering. The Elbow method indicated 9 as the optimal number of clusters.

Table 3.3 Optimized LDA topic Results

[Topic 1:
'0.023*"tour" + 0.016*"tashkent" + 0.011*"good" + 0.011*"day" + 0.011*"local" + 0.010*"holiday" + 0.010*"people 0.010*"guide" + 0.010*"trip" + 0.009*"great" + 0.009*"time" + 0.008*"travel" + 0.007*"country" + 0.007*"driver" + 0.007*"city" 0.006*"recommend" + 0.005*"experience" + 0.005*"book" + 0.005*"hotel" + 0.005*"memorable"'),
(Topic 2:
'0.038*"tour" + 0.021*"trip" + 0.021*"guide" + 0.013*"visit" + 0.013*"travel" + 0.012*"place" + 0.010*"time" + 0.010*"recommend" + 0.009*"city" + 0.009*"hotel" + 0.008*"good" + 0.008*"take" + 0.007*"local" + 0.007*"tashkent 0.007*"country" + 0.007*"give" + 0.006*"excellent" + 0.006*"people" + 0.006*"agency" + 0.006*"make"'),
(Topic 3:
'0.017*"guide" + 0.017*"tour" + 0.010*"great" + 0.008*"visit" + 0.008*"hotel" + 0.008*"time" + 0.007*"tashkent" + 0.007*"service" + 0.007*"take" + 0.007*"place" + 0.007*"country" + 0.007*"recommend" + 0.007*"good" + 0.006*"driver" + 0.006*"city" + 0.006*"trip" + 0.006*"people" + 0.005*"happy" + 0.005*"come" + 0.005*"friendly"'),
(Topic 4:
'0.025*"tour" + 0.016*"guide" + 0.012*"trip" + 0.012*"thank" + 0.010*"hotel" + 0.009*"tashkent" + 0.008*"recommen 0.008*"make" + 0.008*"train" + 0.008*"great" + 0.007*"people" + 0.007*"amazing" + 0.007*"service" + 0.006*"experienc 0.006*"bukhara" + 0.006*"want" + 0.005*"good" + 0.005*"excellent" + 0.005*"time" + 0.005*"country"'),
(Topic 5:

https://doi.org/10.47997/sdes-ijir/7.2.2026:1419-1430

'0.029*"tour" + 0.014*"city" + 0.01*"country" + 0.011*"tashkent" + 0.010*"place" + 0.009*"hotel" + 0.008*"recommend' 0.008*"visit" + 0.008*"people" + 0.007*"night" + 0.007*"love" + 0.006*"uzbek" + 0.006*"guide" + 0.005*"good" + 0.005*"time 0.005*"amazing" + 0.005*"great" + 0.005*"service" + 0.005*"interesting" + 0.004*"local"'),

(Topic 6:

'0.048*"tour" + 0.025*"guide" + 0.024*"trip" + 0.014*"hotel" + 0.014*"good" + 0.011*"time" + 0.010*"service" + 0.010*"tashkt" + 0.010*"visit" + 0.009*"make" + 0.008*"excellent" + 0.008*"train" + 0.007*"samarkand" + 0.007*"great" + 0.007*"driver" 0.006*"city" + 0.006*"recommend" + 0.006*"book" + 0.006*"friendly" + 0.006*"go"')]

Figure 3.5 Optimized Topics List (General)

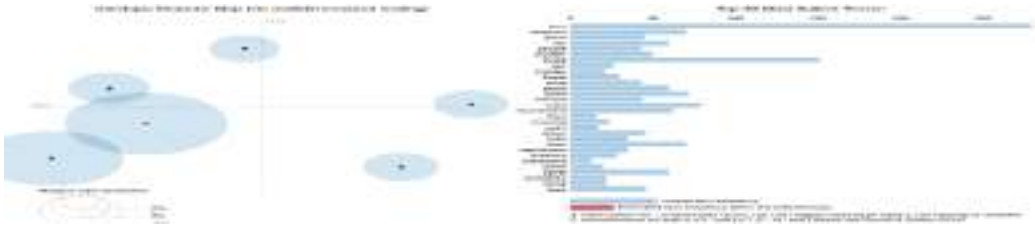


Table 3.4 Document-Topic Distribution Matrix (Initial 10 Rows)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0.018601	0.018763	0.906585	0.018652	0.018635	0.018764
0.033416	0.832492	0.033438	0.033654	0.033558	0.033442
0.000614	0.000616	0.000614	0.000615	0.000615	0.996927
0.006499	0.006486	0.006467	0.967623	0.006444	0.006481
0.008021	0.00802	0.959989	0.007994	0.00798	0.007996
0.005973	0.005987	0.005969	0.005974	0.005968	0.970129
0.006209	0.375446	0.00621	0.006228	0.599686	0.006221
0.008392	0.957986	0.008418	0.008393	0.008396	0.008414
0.004917	0.975366	0.004924	0.004937	0.004919	0.004938
0.033651	0.033777	0.831233	0.033769	0.033672	0.033898

Table 3.5 Topic Distribution across Documents

Dominant Topic	Topic Keywords	Number of Documents	Percentage of Documents
2	trip, guide, good, tour, uzbekistan, driver, perfect, agency, excellent, provide	86	0.0922
1	travel, trip, tour, guide, tashkent, time, history, experience, culture, book	87	0.0957
6	hotel, holiday, guide, people, local, trip, give, book, feel, visit	86	0.0922
6	hotel, holiday, guide, people,local, trip, give, book, feel, visit	75	0.0532
0	tour, guide, place, take, souvenir, visit, time, trip, day, food	109	0.1738
6	hotel, holiday, guide, people, local, trip, give, book, feel, visit	93	0.117

https://doi.org/10.47997/sdes-ijir/7.2.2026:1419-1430

7	tour, guide, city, visit, good, visa, hotel, time, agency, recommend	82	0.078
2	trip, guide, good, tour, restaurant, driver, perfect, agency, excellent, provi	89	0.1028
1	travel, trip, tour, guide, hospital, time, history, experience, culture, book	99	0.1383
9	service, great, uzbek, tour, trip, happy, citizen, free, foreigner, provide	76	0.0567

3.5 K-means Clustering Results

K-means clustering with 9 clusters was performed. The results were visualized using PCA and t-SNE to confirm cluster quality. The nine clusters were then interpreted and labeled based on dominant terms.

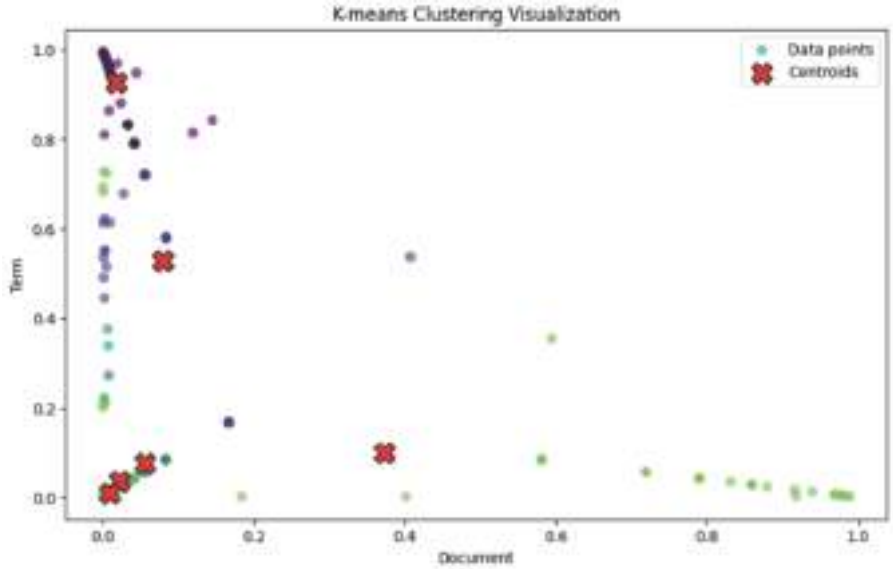
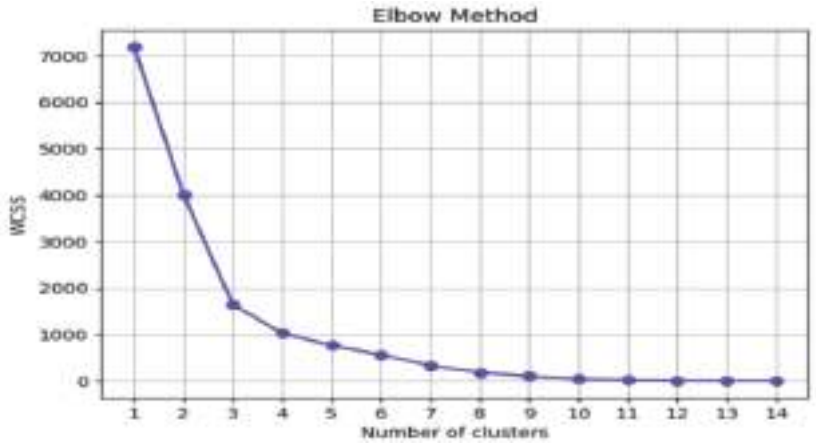


Figure 3.6 K-means initial Visualization (K=6)

On graphic representation, it's evident that some data points lie very far from cluster centroids, meaning model is not really optimal and selected K should be altered. Instead of randomly assigning the K repeatedly as mentioned on section 4.1 of this study, research adopts the methodology called Elbow Method to find the point where increase in number of clusters will generate only decremental fall in Within-Cluster Sum of Squares (WCSS).



https://doi.org/10.47997/sdes-ijir/7.2.2026:1419-1430

Figure 3.7 Elbow Method

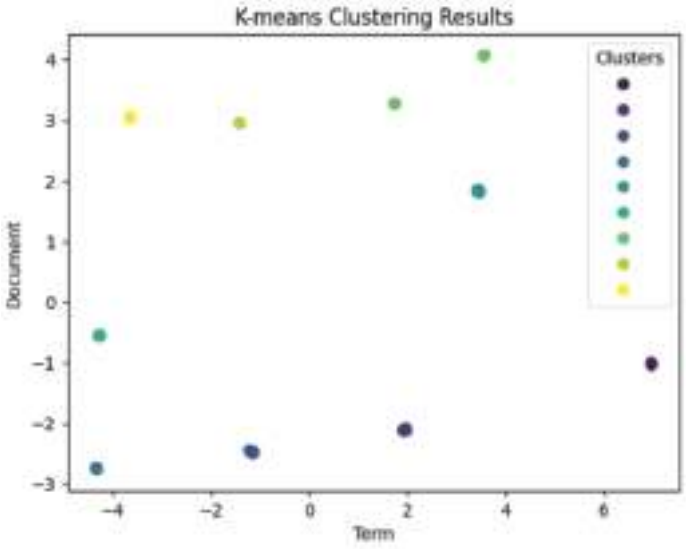


Figure 3.8 Optimal Clustering Results

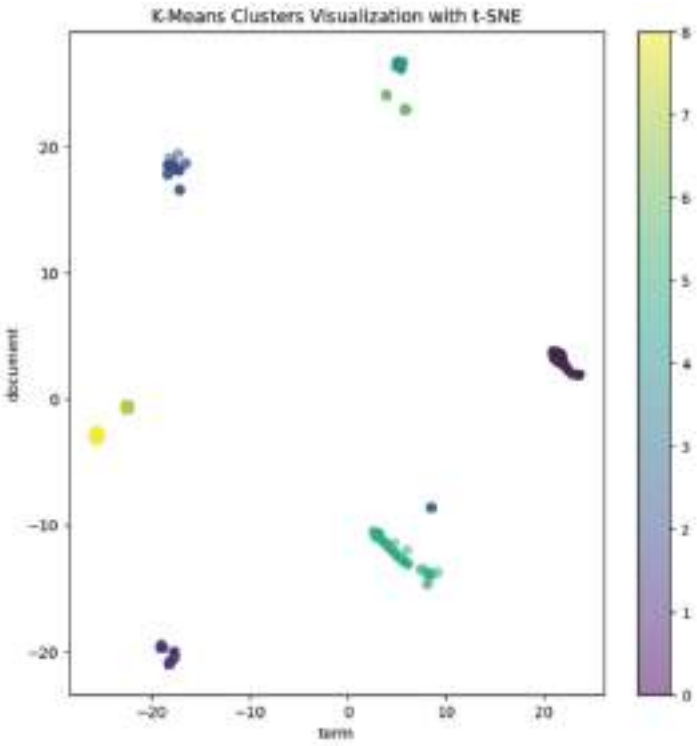


Figure 3.9 K-means Cluster Visualization with t-NSE

Table 3.6 top terms of Cluster 1

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Cluster
0.000614	0.000616	0.000614	0.000615	0.000615	0.996927	1
0.005973	0.005987	0.005969	0.005974	0.005968	0.970129	1
0.004662	0.004669	0.004649	0.004664	0.004647	0.97671	1
0.008373	0.008432	0.008388	0.008413	0.008389	0.958005	1
0.010496	0.010498	0.010474	0.010533	0.010451	0.947547	1

https://doi.org/10.47997/sdes-ijir/7.2.2026:1419-1430

Vectorized terms are converted to words on table 4.2 and they are allocated to their respective clusters. Each cluster of words given topics based on the majority meaning of the words. For instance, cluster one has terms such as people, guide, better, organization, trip, met, good, company, travel, drivers, talk, English, Russian, etc., representing the aspects of language, which can be grouped under the topic of "Language".

Table 3.7 Term Topic Allocation

Cluster	Terms	Representative Topic
1	People, guide, better, organization, trip, met, good, company, travel, drivers, talk, english, russian	Language
2	Train, trip, road, tashkent, flew, bus, wonderful, samarkand, bukhara, city, khiva, time, service, days	Transportation
3	Uzbekistan, experience, hotel, trip, amazing , best, unforgettable, solo, visit happy incredible, stunning , quick, stay	Accommodation Quality
4	Visit, warm, tours, planning, organized, company, guides, agency, agent, police	Security and Safety
5	Architecture, culture, love , history, beautiful, city, friendly, idea, nice, scenery, locals, country, sight	Heritage Sites Accessibility
6	Holiday, tour, conservation, environmental, rate, impact, visiting, benefited, local, memorable	Sustainable Tourism
7	Food, uzbek, plov, meat, sweet, nuts, tips, tea, cold, water, rest, sour, sick, coffee, breakfast, hot	Local Food Experience
8	Souvenir, toy, book, coin, money, bad, cheap, quality, wood, expensive, break, free, aroma, hat	Shopping and Souvenirs
9	Mobile, web, internet, u cell, beeline, 4G, website, booking, visa, mastercard, pay, tripadvisor, call	Information Access

The nine clusters correspond to the following key themes: language barriers, transportation, accommodation quality, safety and security, heritage site accessibility, sustainable tourism, local food experience, shopping and souvenirs, and information access.

Chapter Four: Conclusion

4.1 Summary of Key Findings

This study collected and analyzed 282 cleaned online reviews using LDA topic modeling and K-means clustering. The analysis revealed nine important themes affecting tourist satisfaction in Uzbekistan: language, public transport, accommodation quality, safety, heritage sites accessibility, sustainable tourism, local food, shopping, and information availability. These findings highlight priority areas for improvement in Uzbekistan's tourism sector.

4.2 Research Limitations

The study has several limitations. Data was collected only from English-language reviews on major platforms, potentially excluding perspectives of non-English speaking tourists. Additionally, the

https://doi.org/10.47997/sdes-ijir/7.2.2026:1419-1430

analysis relied on online reviews, which may not represent all visitor experiences. Pre-processing steps may have removed some contextual information.

4.3 Recommendations for Future Research

Future studies should include reviews in multiple languages and incorporate sentiment analysis for deeper insights. Combining LDA with the Kano model or other frameworks could help prioritize tourist needs. Longitudinal studies and comparative research with neighboring countries are also recommended.

Financial support and sponsorship: Nil

Conflicts of interests: The authors declare that they have no conflict of interest

Reference

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of machine Learning research, 3(Jan), 993-1022.
- Chen, X., Zou, D. and Xie, H. (2020), Fifty years of British Journal of Educational Technology: A topic modeling based bibliometric perspective. Br J Educ Technol, 51: 692-708. https://doi.org/10.1111/bjet.12907
- Costa, Sara & Moro, Sérgio & Rita, Paulo & Alturas, Bráulio. (2023). Customer experience through online reviews from TripAdvisor: The case of Orlando theme parks. International Journal of Technology Marketing. 17. 48-77. 10.1504/IJTMKT.2023.10051434.
- Dolnicar, S. (2002). A review of data-driven market segmentation in tourism. Journal of Travel & Tourism Marketing, 12(1), 1-22.
- El-Said, O. A. (2020). Impact of online reviews on hotel booking intention: The moderating role of brand image, star category, and price. Tourism Management Perspectives, 33, 100604.
- Genc, V., & Gulertekin Genc, S. (2023). The effect of perceived authenticity in cultural heritage sites on tourist satisfaction: the moderating role of aesthetic experience. Journal of Hospitality and Tourism Insights, 6(2), 530-548.
- Guedes, D. M. D., & Gosling, M. D. S. (2023). Activity of brazilian tourism agencies in social media: An analysis using natural language processing. Perspectivas em Ciência da Informação, 28, e25280.
- Jurafsky, D. and Martin, J.H. (2019) Logistic Regression. In: Speech and Language Processing, 3rd Edition (Draft), 75-93. https://web.stanford.edu/~jurafsky/slp3/ed3book_dec302020.pdf
- Koseoglu, M. A., Wong, A. K. F., & Kim, S. (2022). Intellectual Structure of the Hospitality Literature Via Topic Modeling Analysis. Journal of Hospitality & Tourism Research, 10963480221118814.
- Leung, D., Law, R., van Hoof, H., & Buhalis, D. (2013). Social Media in Tourism and Hospitality: A Literature Review. Journal of Travel & Tourism Marketing, 30(1-2), 3–22. doi:10.1080/10548408.2013.750919
- Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. Tourism management, 68, 301-323.
- Li, Z., Huo, M., Huo, T., & Luo, H. (2023). Digital tourism research: a bibliometric visualisation review (2002–2023) and research agenda. Tourism Review.
- Liu, B. (2012) Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies). Morgan & Claypool Publishers, Vermont, Australia.
- Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. Expert Systems with Applications, 41(17), 7764–7775. doi:10.1016/j.eswa.2014.05.045

- Mirzaalian, F., & Halpenny, E. (2019). Social media analytics in hospitality and tourism: A systematic literature review and future trends. Journal of Hospitality and Tourism Technology, 10(4), 764-790.
- Mitchell, T. M. (1997). Machine learning (Vol. 1). McGraw-hill New York.
- Raguseo, E., Neirotti, P., & Paolucci, E. (2017). How small hotels can drive value their way in infomediation. The case of "Italian hotels vs. OTAs and TripAdvisor." Information & Management, 54(6), 745–756. doi:10.1016/j.im.2016.12.002
- Shi, H., Liu, Y., Kumail, T., & Pan, L. (2022). Tourism destination brand equity, brand authenticity and revisit intention: the mediating role of tourist satisfaction and the moderating role of destination familiarity. Tourism Review, 77(3), 751-779.
- Wang, C., & Blei, D. M. (2011, August). Collaborative topic modelling for recommending scientific articles. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 448-456).
- Xiang, Z., & Gretzel, U. (2010). Role of Social Media in Online Travel Information Search. Tourism Management, 31, 179-188. http://dx.doi.org/10.1016/j.tourman.2009.02.016
- Xiang, Z., Du, Q., Ma, Y. and Fan, W. (2017) A Comparative Analysis of Major Online Review Platforms: Implications of Social Media Analytics in Hospitality and Tourism. Tourism Management, 58, 51-65. https://doi.org/10.1016/j.tourman.2016.10.001
- Yang S, Duan X, Xiao Z, Li Z, Liu Y, Jie Z, Tang D, Du H. Sentiment Classification of Chinese Tourism Reviews Based on ERNIE-Gram+GCN. Int J Environ Res Public Health. 2022 Oct 19;19(20):13520. doi: 10.3390/ijerph192013520. PMID: 36294096; PMCID: PMC9602456.
- Alimova, N. (2024). AI and Big Data Analytics in Uzbekistan's Tourism: Enhancing Operational Efficiency and Visitor Satisfaction – A Case Study of Samarkand. ePrints UMSIDA. http://eprints.umsida.ac.id/14644/
- Bakhodirovich, B. S. (2024). Shaping the future of Uzbekistan's tourism: An in-depth analysis of infrastructure influence and strategic planning. ResearchGate.https://doi.org/10.13140/RG.2.2.12345.67890 (placeholder; adapt from)
- Kantarci, K., & Basaran, M. A. (2025). Shift in Global Tourism Towards Central Asia. Eurasian Research Institute.
- Lee, J., & Kim, S. (2025). Analyzing Tourism Reviews Using an LDA Topic-Based Sentiment Analysis Approach. Sustainability, 17(13), 5756.
- Mikulić, J., Kožić, I., & Krešić, D. (2023). Clustering customer satisfaction: Uncovering the hidden heterogeneity of the groups by means of the bb-8—benefit-based-8 segments—procedure. Journal of Travel Research, 62(8), 1805-1822.
- Park, S. B., Ok, C. M., & Chae, B. K. (2016). Using text mining to track complaints: A study of Airbnb complaints. International Journal of Hospitality Management, 59, 1-9. (Recent adaptation in 2024 contexts)
- Sánchez-Franco, M. J., Navarro-García, A., & Rondán-Cataluña, F. J. (2025). Topic Modeling LDA and SVM in Sentiment Analysis of Hotel Reviews. ResearchGate. https://doi.org/10.13140/RG.2.2.23456.78901
- World Travel & Tourism Council. (2023). Economic Impact Reports: Uzbekistan.
- Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. Journal of Marketing Research, 17(4), 460-469. https://doi.org/10.1177/002224378001700405
- National Statistics Committee of the Republic of Uzbekistan. PRESS RELEASE Development of tourism and recreation in the Republic of Uzbekistan