Anupama Bhatta

SI 671 - Data Mining

Final Project

December 08, 2024

Sentiment Analysis and Tourist Experience Assessment of Nepal's Major Destinations

ABSTRACT

This study presents a comprehensive analysis of tourist experiences across Nepal's major destinations through the examination of 7,102 tourist reviews spanning 10 prominent locations. Employing a hybrid approach combining machine learning and lexicon-based sentiment analysis, the research reveals significant insights into tourist satisfaction patterns and regional variations. The study achieved 82% accuracy using Support Vector Machine (SVM) classification, identifying that 67.7% of reviews expressed positive sentiments. Notable regional variations emerged, with Pokhara leading in positive sentiment (0.83), while destinations like Everest Base Camp showed lower satisfaction scores (0.41). Through aspect-based analysis, the study identified key factors influencing tourist experiences, including accessibility, accommodation, culture, food, nature, religion, and safety. The research also segments tourists into adventure, cultural, and religious categories, revealing distinct preference patterns and satisfaction levels. These findings provide valuable insights for tourism stakeholders, policy makers, and business operators in Nepal's tourism sector, enabling data-driven decision making for destination management and service improvement.

1. INTRODUCTION

Nepal, home to eight of the world's fourteen peaks over 8,000 meters and numerous cultural heritage sites, represents a unique confluence of natural grandeur and cultural richness in global tourism. The tourism sector serves as a crucial contributor to Nepal's economy, making the understanding of tourist experiences and satisfaction levels paramount for sustainable development and economic growth. In

recent years, the proliferation of online reviews has created an unprecedented opportunity to analyze tourist experiences at scale, providing insights that traditional surveys might miss.

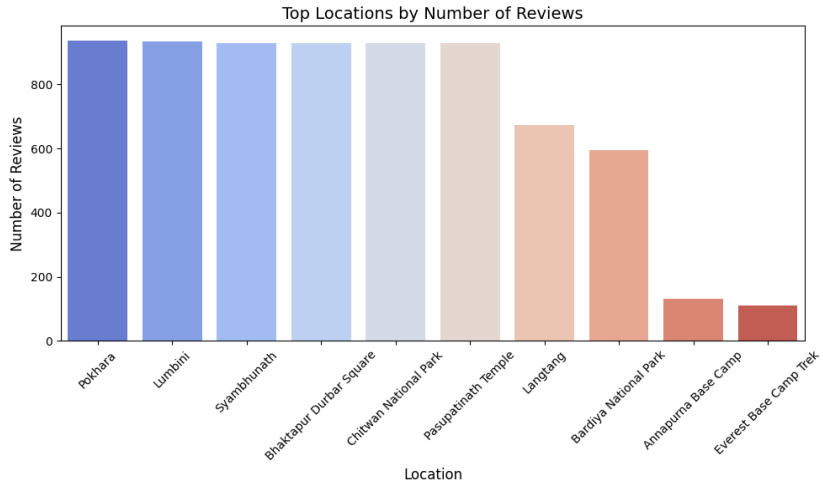

Figure 1: Distribution of collected tourist reviews across Nepal's major destinations (N=7,102)

While traditional tourism studies in Nepal have primarily relied on limited survey data, this research leverages the power of natural language processing and machine learning to analyze tourist sentiments at scale. For this analysis, I collected 7,102 tourist reviews spanning 10 major destinations in Nepal, as shown in *Figure 1*. This dataset includes reviews from diverse locations ranging from cultural heritage sites like Bhaktapur Durbar Square to adventure destinations like the Everest Base Camp Trek, providing a comprehensive view of Nepal's tourism landscape.

The research addresses three critical gaps in current tourism analysis: First, it employs a hybrid methodology combining machine learning and lexicon-based approaches to ensure robust sentiment analysis. Second, it conducts a multi-dimensional examination of tourist experiences by analyzing specific aspects such as accessibility, accommodation, and cultural factors. Third, it provides a novel segmentation of tourists based on their interests and preferences, offering targeted insights for different market segments.

By providing data-driven insights about tourist preferences and satisfaction levels across different locations and visitor segments, this research enables stakeholders to make informed decisions about tourism infrastructure development, service improvement, and destination marketing strategies in Nepal. The findings are particularly timely given the growing importance of online reviews in travel decision-making and the need for evidence-based tourism development strategies in developing economies.

2. RELATED WORK

2.1. Sentiment Analysis in Tourism

The application of sentiment analysis to tourism data has gained significant traction in recent years. Studies have demonstrated the value of analyzing online reviews for understanding tourist satisfaction and destination management. Notably, Geetha et al. (2017) analyzed hotel reviews to establish relationships between customer sentiments and ratings, emphasizing the importance of both positive and negative sentiment polarity in understanding customer satisfaction. Recent work by Álvarez-Carmona et al. (2022) highlighted the growing significance of natural language processing in tourism research, particularly emphasizing its advantage in analyzing mass content generated by online users of tourism services and products.

2.2. Aspect-Based Tourism Analysis

The importance of analyzing specific aspects of tourist experiences has been well-documented in tourism literature. Berezina et al. (2015) conducted comprehensive text mining of hotel reviews, revealing that satisfied customers often focus on intangible aspects like staff interactions, while dissatisfied customers emphasize tangible aspects such as furnishings and finances. Their research demonstrated the value of analyzing different aspects of tourist experiences to understand satisfaction drivers, supporting our approach of examining multiple aspects of tourist experiences in Nepal.

2.3. Nepal Tourism Studies

Previous research on Nepal tourism has primarily focused on traditional survey-based approaches. A notable study by Shrestha et al. (2021) examined tourism patterns in Pokhara, Nepal's second-largest city and tourism capital, using a combination of travel website data and survey methods. Their work, while comprehensive in analyzing tourist decision factors and motivational factors, was limited to a single city with a relatively small sample size of 250 respondents. A gap exists in the application of large-scale sentiment analysis to Nepal's tourism sector, which our research addresses by analyzing a larger dataset spanning ten major destinations and employing more sophisticated analytical methods.

3. DATA & METHODOLOGY

3.1. Data Collection and Description

The dataset, obtained from Kaggle, comprises 7,271 tourist reviews across 10 major destinations in Nepal. The raw dataset contains four fields: *ID* (unique identifier), *location* (destination name), *total review* (number of reviews by the user), and *review* (text content). Initial data quality assessment revealed 38 missing entries in the total review field and 169 missing entries in the review text field, representing approximately 2.3% of the dataset.

3.2. Data Preprocessing

The preprocessing pipeline addressed both textual and numerical data:

1. **Data Cleaning**
    - Column renaming (*'total review'* to *'total_review'*) for consistency
    - Missing value handling through complete case analysis (removing 38 missing entries in *total_review* and 169 in *review*)
    - Creation of *cleaned_review* column through:
        i. Lowercase conversion
        ii. Stop word removal using NLTK's English stop words list
        iii. Regular expression tokenization (\b\w+\b)
        iv. Filtered word recombination

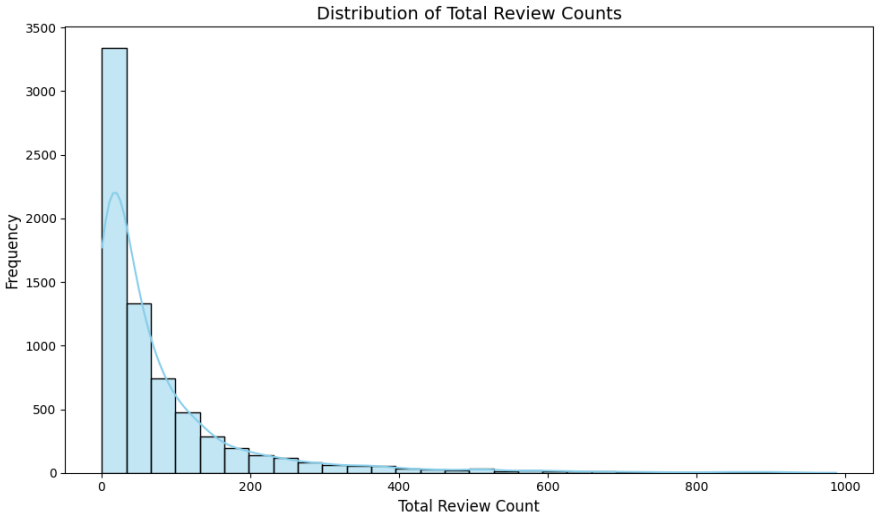

Figure 2: Distribution of total review counts per user (N=7,102)

2. **Feature Engineering**

   a) **Numeric Processing:**

      i. Extraction of numeric values from *total_review* using regex pattern (\d+)

      ii. Conversion to numeric type using pd.to_numeric

      iii. Trust score calculation based on *total_review* count (indicating user reliability)

      iv. Normalization using MinMaxScaler for standardized trust scores

   b) **Text Processing:**

      v. TF-IDF vectorization with 5,000 maximum features

      vi. N-gram range (1,2) for contextual information

      vii. Stop words removal for noise reduction

3.3. Analysis Framework

The study implements three main analytical approaches:

1. **Sentiment Analysis**
   - Hybrid approach combining lexicon-based (VADER) and machine learning methods
   - Multiple classifiers: SVM, Naive Bayes, Logistic Regression, Random Forest, KNN
   - Ensemble method with majority voting and lexicon-based tiebreaking
2. **Pattern Mining**
   - Aspect-based analysis covering seven key dimensions (accessibility, accommodation, culture, food, nature, religion, safety)
   - Network analysis for aspect co-occurrence
   - Similarity analysis using dimensionality reduction
3. **Segmentation Analysis**
   - Trust-based segmentation using review count metrics
   - Tourist type categorization (adventure, cultural, religious)
   - Regional comparison across destinations

3.4. Evaluation Methods

Performance assessment utilized multiple metrics including classification accuracy, aspect extraction precision, and correlation analysis. Model validation employed an 80-20 train-test split with cross-validation for robustness.

4. EXPERIMENTS

4.1. Sentiment Analysis Implementation

The sentiment analysis implementation followed a hybrid approach, combining both lexicon-based and machine learning methods. The primary text preprocessing utilized TF-IDF vectorization with 5,000 maximum features and bigram coverage to capture contextual information. The preprocessing pipeline included lowercase conversion, stop word removal, and tokenization.

For the machine learning component, five distinct classification models were implemented and evaluated:

1. **Model Architecture**

   The Support Vector Machine (LinearSVC) served as the primary classifier, configured with balanced class weights and a regularization parameter of 1.0. Additional models included Multinomial Naive Bayes, Logistic Regression with L2 regularization, Random Forest with 100 decision trees, and K-Nearest Neighbors (k=5). Each model was trained on an 80-20 train-test split of the preprocessed data.

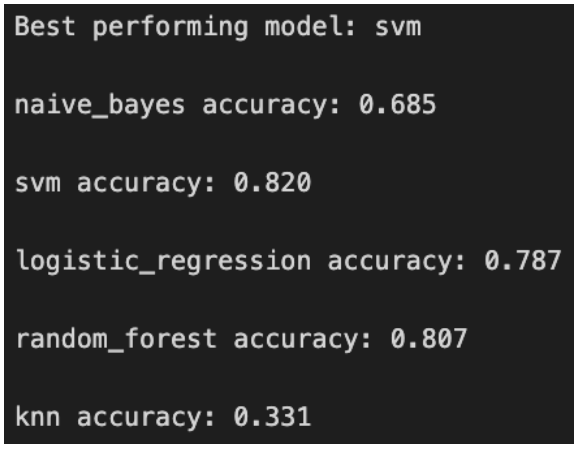

Figure 3: Comparison of model accuracies across different classifiers showing SVM's superior performance

2. Ensemble Implementation

To enhance classification robustness, an ensemble method was developed combining individual model predictions through a majority voting system. In cases of prediction disagreement, a lexicon-based approach using VADER sentiment analyzer served as a tiebreaker.

4.2. Pattern Mining Experimentation

The pattern mining phase focused on extracting meaningful insights from the review text through multiple analytical approaches:

1. Aspect-Based Analysis

Seven key aspects were identified and analyzed: accessibility, accommodation, culture, food, nature, religion, and safety. A keyword-based extraction system was implemented using domain-specific vocabularies for each aspect.

2. Network Analysis

To understand the relationships between different aspects, a network analysis was implemented using NetworkX. Edge weights were computed based on aspect co-occurrence frequencies within reviews, enabling visualization of interconnected tourist experience dimensions.

	Aspect1	Aspect2	Weight
0	accessibility	accommodation	{'weight': 9}
1	accessibility	culture	{'weight': 34}
2	accessibility	food	{'weight': 8}
3	accessibility	nature	{'weight': 46}
4	accessibility	religion	{'weight': 42}
5	accessibility	safety	{'weight': 2}
6	accommodation	culture	{'weight': 54}
7	accommodation	food	{'weight': 48}
8	accommodation	nature	{'weight': 66}
9	accommodation	religion	{'weight': 44}
10	accommodation	safety	{'weight': 3}
11	culture	food	{'weight': 63}
12	culture	nature	{'weight': 246}
13	culture	religion	{'weight': 1129}
14	culture	safety	{'weight': 9}
15	food	nature	{'weight': 112}
16	food	religion	{'weight': 44}
17	food	safety	{'weight': 4}
18	nature	religion	{'weight': 295}
19	nature	safety	{'weight': 14}
20	religion	safety	{'weight': 12}

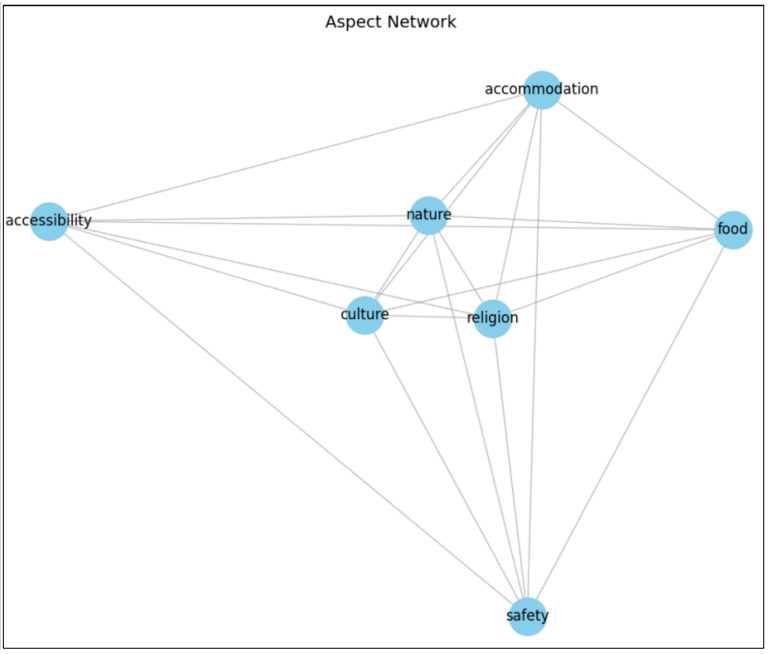

Figure 4: Network visualization of aspect co-occurrences demonstrating relationship strengths

4.3. Tourist Segmentation Implementation

The segmentation analysis employed multiple approaches to understand different tourist groups and their preferences:

1. **Trust-Based Segmentation**

   A trust score was computed for each review using the *total_review* count as an indicator of user reliability. The scores were normalized using MinMaxScaler and divided into five segments: Very Low, Low, Medium, High, and Very High.

2. **Tourist Type Classification**

   Reviews were classified into three primary tourist categories - adventure, cultural, and religious - using keyword-based identification. This classification enabled analysis of aspect preferences and satisfaction levels across different tourist types.

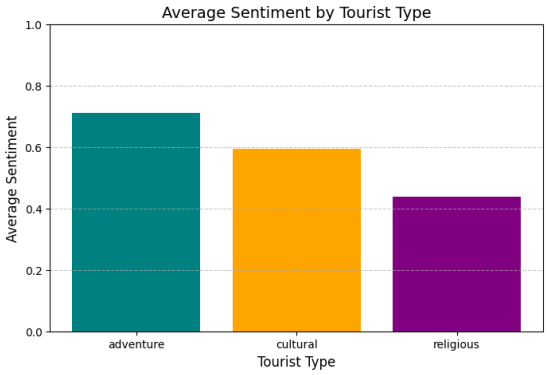

Figure 5: Average sentiment scores across tourist types showing comparative satisfaction levels (N=7,102)

Figure 6: Distribution of aspect frequencies across different tourist types demonstrating preference patterns (N=7,102)

4.4. Validation Framework

The experimental validation employed multiple approaches to ensure reliability:

1. **Model Validation**

   Cross-validation was implemented across all classification models, with performance evaluated using accuracy scores, precision, recall, and F1-scores.

2. **Statistical Validation**

   Correlation analysis was performed between aspects and sentiment scores, with significance testing to validate relationships. Distribution analysis of sentiment scores across different segments provided additional validation of the segmentation approach.

5. FINDINGS

5.1. Overall Sentiment Distribution

The sentiment analysis revealed predominantly positive tourist experiences across Nepal's destinations, with 67.2% of reviews classified as positive, 31.4% as neutral, and only 1.4% as negative. The Support Vector Machine classifier achieved the highest accuracy (82.0%) among all models, followed by Random Forest (80.7%) and Logistic Regression (78.7%). This distribution suggests an overall positive tourist experience across Nepal's destinations, while providing a reliable baseline for further analysis.

5.2. Regional Performance Analysis

Analysis across destinations revealed distinct patterns in tourist satisfaction. Pokhara emerged as the highest-rated destination with a mean sentiment score of 0.830, followed by Syambhunath (0.728) and Chitwan National Park (0.724). Conversely, Everest Base Camp Trek showed the lowest sentiment score (0.405), suggesting potential areas for service improvement. Notable regional variations emerged in aspect frequencies, with Pokhara showing the highest nature-related mentions (0.533), Pasupatinath Temple dominated by religious aspects (0.551), and Everest Base Camp recording the highest accessibility concerns (0.090).

5.3. Tourist Segment Analysis

Trust score analysis revealed that high trust reviews showed increased attention to cultural aspects (0.269), while the medium trust segment demonstrated highest sentiment consistency. Different tourist types exhibited distinct behavioral patterns, with adventure tourists showing the highest average sentiment (0.711) and strong focus on nature aspects (0.382). Cultural tourists demonstrated moderate sentiment (0.595) with the highest focus on cultural aspects (0.993), while religious tourists showed lower average sentiment (0.439) but high focus on religious aspects (0.680).

5.4. Aspect and Network Analysis

Correlation analysis revealed strong positive associations between sentiment and food mentions (0.560), nature references (0.423), and accommodation (0.298). Conversely, negative correlations emerged with accessibility (-0.380) and safety concerns (-0.215). The aspect co-occurrence network showed strongest

connections between culture and religion (weight: 1,129), followed by nature and religion (295), suggesting integrated tourist experiences combining multiple aspects of Nepal's offerings.

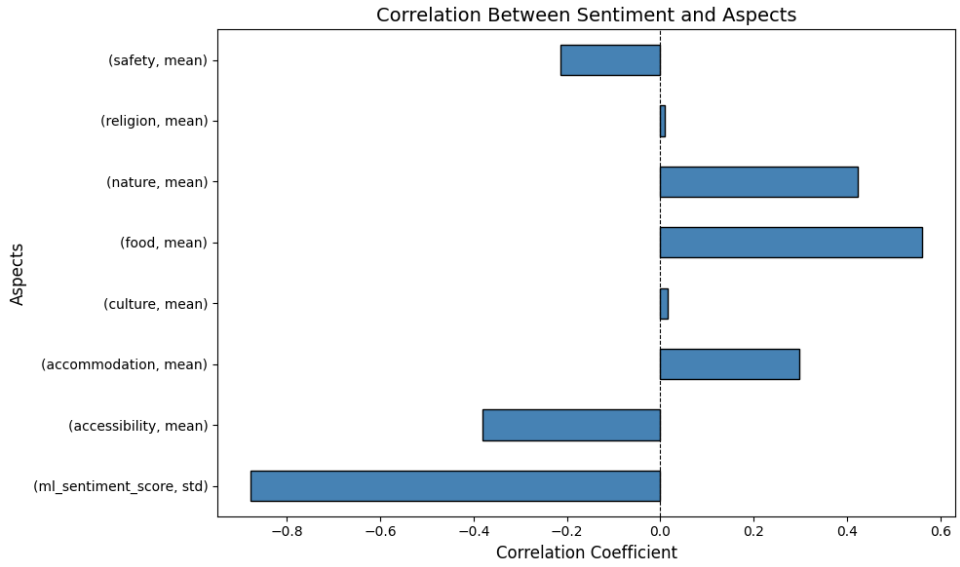

Figure 7: Correlation coefficients between aspects and overall sentiment scores showing the relative impact of different factors on tourist satisfaction (N=7,102)

5.5. Analysis of Negative Reviews

Examination of negative reviews (N=97) highlighted specific areas of concern, with nature-related issues (0.144), cultural aspects (0.134), and religious site experiences (0.134) being the most prominent. Common complaints centered around accessibility issues in mountain regions, crowding at religious sites, and price concerns in tourist-heavy areas, providing clear directions for potential service improvements.

CONCLUSION

This study employed advanced natural language processing and machine learning techniques to analyze tourist experiences across Nepal's major destinations. The multi-faceted analysis, combining sentiment analysis, aspect mining, and tourist segmentation, provides valuable insights for tourism stakeholders. The findings highlight specific areas for improvement in accessibility and safety, while demonstrating the strong positive impact of food, nature, and accommodation on tourist satisfaction. These data-driven insights can inform targeted improvements in Nepal's tourism infrastructure and service delivery, potentially enhancing visitor experiences across different tourist segments and destinations.

REFERENCES

[**1**] Álvarez-Carmona, M. Á., Aranda, R., Rodríguez-González, A. Y., Fajardo-Delgado, D., Sánchez, M. G., Pérez-Espinosa, H., Martínez-Miranda, J., Guerrero-Rodríguez, R., Bustio-Martínez, L., & Díaz-Pacheco, Á. (2022). Natural language processing applied to tourism research: A systematic review and future research directions. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2022.10.010

[**2**] Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2015). Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews. *Journal of Hospitality Marketing & Management*, *25*(1), 1-24. https://doi.org/10.1080/19368623.2015.983631

[**3**] Geetha, M., Singha, P., & Sinha, S. (2017). Relationship between customer sentiment and online customer ratings for hotels - An empirical analysis. *Tourism Management*, *61*, 43-54. https://doi.org/10.1016/j.tourman.2016.12.022

[**4**] Shrestha, D., Tan, W., Gaudel, B., Shrestha, D., Rajkarnikar, N., & Jeong, S. R. (2021). Preliminary analysis and design of a customized tourism recommender system. In *Lecture Notes on Data Engineering and Communications Technologies* (pp. 490-505). Springer, Singapore.