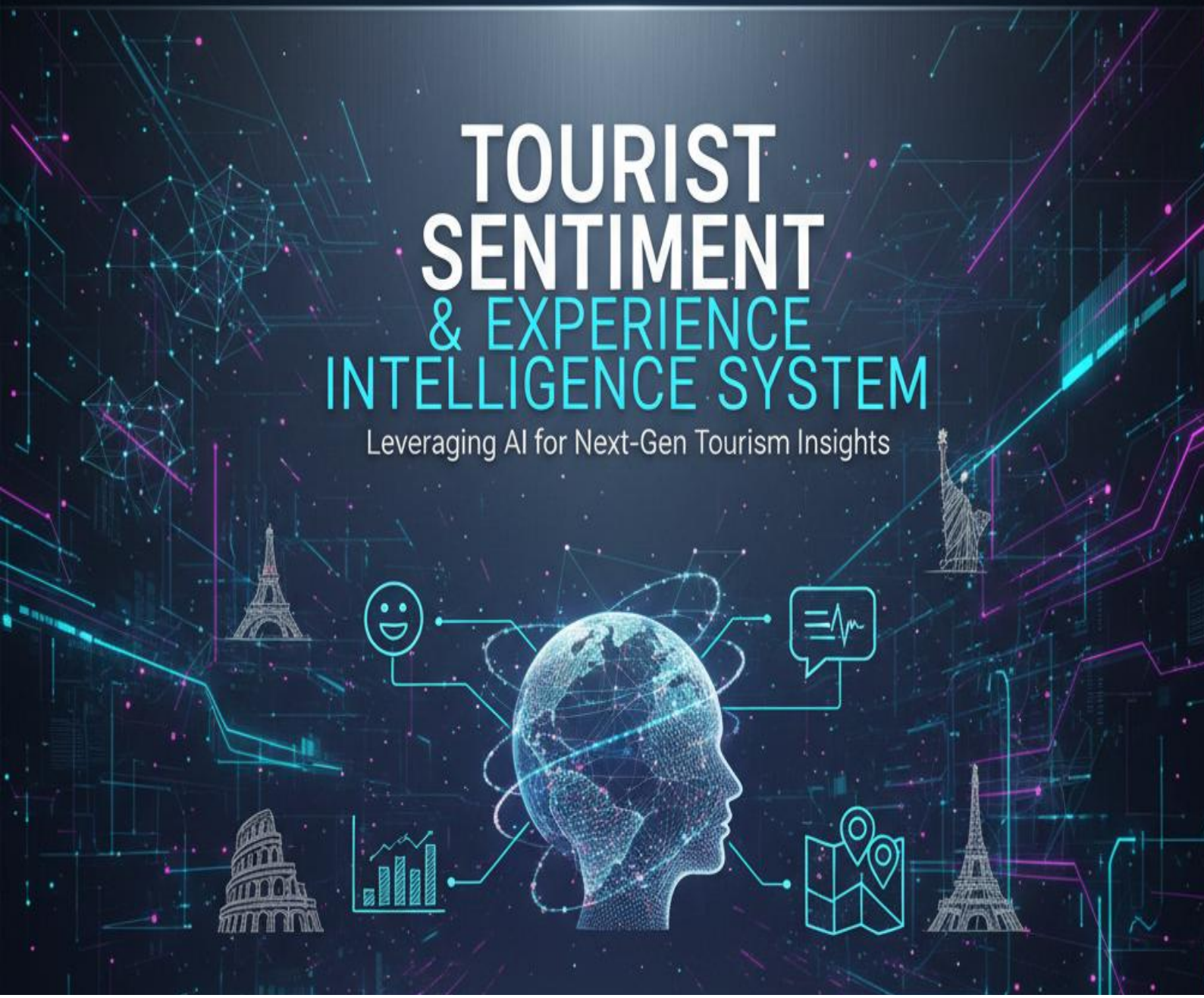

Title: Decoding Tourist Voices: A Data-Driven Analysis

A capstone project on Text analysis & Web scrapping

Presented by: Satyaprakash Behera

(24MBMB14)

School of Management studies

Abstract

The rapid growth of digital platforms has significantly increased the availability of user-generated travel content, particularly in the form of reviews, social media posts, and online discussions. These textual records contain valuable insights into tourist experiences, satisfaction levels, pain points, and evolving expectations. However, due to the unstructured nature and massive volume of such data, traditional analysis methods remain inadequate.
This project presents an end-to-end Natural Language Processing (NLP) pipeline for extracting, processing, analyzing, and visualizing tourist experiences across major Indian travel destinations. The work integrates web scraping, text preprocessing, sentiment analysis, keyword extraction, emotion detection, topic modeling, and issue identification. An interactive Streamlit dashboard consolidates all results for intuitive interpretation. The findings demonstrate that NLP can effectively decode experiential patterns, uncover hidden issues, and support data-driven decision-making in the tourism sector.

1. Introduction

Tourism plays a pivotal role in India's economy, contributing substantially to employment, cultural exchange, and overall socio-economic development. With the emergence of digital platforms, tourists increasingly share their experiences online, providing rich qualitative data that reflect their emotions, expectations, and personal interactions with destinations.
Despite the importance of this information, tourism stakeholders often lack a systematic method to analyze these unstructured textual sources. Traditional surveys and manual assessment techniques are susceptible to bias, limited in scope, and incapable of handling large-scale data.

This project addresses this gap by employing modern NLP techniques to automatically analyze tourist experiences. It proposes a computational framework capable of extracting insights from thousands of text entries and transforming them into structured, decision-friendly outputs.

2. Problem Definition

Although abundant tourist data is available online, the following challenges hinder the extraction of meaningful insights:

1. **Unstructured Text Data**
   Tourist reviews vary in writing style, grammar, length, language, and intent.
2. **Hidden Emotions and Sentiments**
   Simple positive/negative classifications are insufficient to interpret nuanced tourist emotions.
3. **Inability to Identify Underlying Themes**
   Text often contains multiple topics (transport, safety, weather, service quality, cultural interactions), making manual topic detection difficult.
4. **Lack of Systematic Issue Monitoring**
   Tourism departments require early identification of common grievances—scams, overpricing, unsafe environments, cleanliness issues—but these are hidden in lengthy texts.
5. **Absence of a Unified Dashboard**
   Even when insights are extracted, they are rarely presented in an integrated, interactive format.

Thus, the primary objective of this research is to create a robust NLP-driven analytics system that can:

- **Collect real-world tourist data**
- **Clean and prepare text**
- **Apply multiple NLP models**
- **Visualize findings**
- **Provide destination-wise intelligence**

3. Literature Context (Brief)

Text mining has been widely applied in domains such as marketing, healthcare, and social behavior analysis. Studies show that sentiment analysis and topic modeling can greatly enhance understanding of user-generated content. VADER and NRCLex have proven effective for social-media-like text. RAKE is frequently used in unsupervised keyword extraction. NMF and TF-IDF combinations provide coherent topic clusters for short texts.

This project adapts and integrates these proven techniques into a unified system tailored to Indian tourist data.

4. Methodology

The methodology follows a multi-stage approach:

4.1 Data Collection

Data was collected through web scraping of travel forums, blogs, and user-uploaded comments. Python libraries such as requests, BeautifulSoup, and tqdm facilitated extraction of approximately 400+ review entries from tourists across popular Indian destinations such as Goa, Kerala, Manali, Ladakh, Darjeeling, Agra, Jaipur, Delhi, and others.
The scraped dataset contained fields including source, destination, review text, date, and link references.

4.2 Data Preprocessing

To ensure compatibility with NLP models, several preprocessing operations were applied:

- **Lowercasing**
- **Removal of HTML tags and URLs**
- **Removal of punctuation and numeric characters**
- **Stopword elimination using NLTK**
- **Lemmatization for vocabulary normalization**
- **Whitespace trimming and special character handling**

A cleaned dataset was generated and stored as: tourist_experience_cleaned.csv

5. NLP Use Cases and Experiments

Each use case represents a specific analytical technique applied independently.

5.1 Use Case 1: Sentiment Analysis

Problem:
Tourism departments must understand whether travellers express positive, neutral, or negative experiences. Manual sentiment classification is slow and inconsistent.

Methodology:
The VADER sentiment analyzer was chosen due to its suitability for informal, social-media-style text.

Implementation:

- **Each cleaned review was passed to the VADER model.**
- **Sentiment polarity was computed.**
- **Scores were categorized into Positive, Negative, or Neutral classes.**
- **Destination-wise sentiment distributions were generated.**

Results:
The majority of reviews exhibited positive sentiment, particularly for natural destinations such as Goa, Kerala, Ladakh, and Manali.
Moderate negative sentiment was observed in urban regions like Delhi and Agra, primarily due to traffic, crowds, weather, or safety concerns.
The analysis produced detailed sentiment charts and structured CSV outputs for further modeling.

5.2 Use Case 2: Keyword Extraction

Problem:
Tourists highlight specific attractions or issues in their reviews. Identifying recurring keywords helps understand the most discussed aspects of each destination.

Methodology:
The RAKE algorithm (Rapid Automatic Keyword Extraction) was applied to extract meaningful phrases.

Implementation:

- **Reviews were aggregated by destination.**

- **RAKE extracted multi-word keywords.**
- **Destination-wise keyword lists were stored and visualized.**

Results:
Distinct patterns were identified:

- **Goa: beach, nightlife, shack, food, sunset**
- **Kerala: backwaters, houseboat, greenery, peaceful**
- **Manali: snow, trekking, mountains, adventure**

These results offer concrete insight into what tourists value at each destination.

5.3 Use Case 3: Emotion Detection

Problem:
Sentiment polarity alone cannot capture nuanced emotional responses such as joy, fear, anger, or anticipation.

Methodology:
Emotion classification was performed using the NRCLex model, which maps words to eight emotional categories.

Implementation:
Each review was processed to assign a dominant emotion.

Results:
Most locations reflected predominant emotions of joy and trust, indicating satisfaction.
However, Delhi and Kashmir showed increased fear and anger, most likely due to issues like safety concerns, weather instability, and traffic intensity.

This level of emotional granularity provides actionable psychological insight into traveller experiences.

5.4 Use Case 4: Topic Modeling

Problem:
Tourist reviews include conversations covering multiple aspects of travel.

Identifying these themes helps stakeholders understand what topics dominate overall tourist discourse.

Methodology:
A combination of TF-IDF vectorization and Non-Negative Matrix Factorization (NMF) was used to extract coherent topics.

Implementation:

- **Preprocessed text was converted into TF-IDF matrices.**
- **NMF extracted six dominant topics.**
- **Topics were interpreted and labeled.**
- **Each review was assigned a dominant topic label.**

Results:
Six interpretable topics emerged:

1. **Travel essentials & logistics**
2. **Cultural experience & food**
3. **Attractions & general tourism**
4. **Hotels & seasonal travel**
5. **Solo travel & safety**
6. **Events & group activities**

Topic distribution revealed the thematic concerns of tourists across destinations and allowed structured theme classification.

5.5 Use Case 5: Issue Identification

Problem:
Tourism authorities often struggle to identify recurring pain points mentioned by travelers.

Methodology:
A keyword-dictionary-based approach was used, mapping reviews to seven issue categories (safety, crowd, cleanliness, price, service, traffic, weather).

Implementation:

- **Each review was scanned for issue-related terms.**
- **Reviews were tagged with detected issues.**
- **Aggregate issue frequencies were generated.**

Results:
The most frequent issues were:

- **Safety concerns**
- **Overcrowding**
- **Cleanliness problems**
- **Traffic and weather-related difficulties**

This categorization directly supports tourism improvement planning and policy changes.

6. Dashboard Development

An interactive Streamlit dashboard was developed to present:

- **Sentiment charts**
- **Emotion maps**
- **Keyword listings for each destination**
- **Topic distribution graphs**
- **Issue summaries**
- **A geospatial sentiment map using geopy**

The dashboard enables tourism stakeholders to explore insights in real time with filters and visual tools.

7. Discussion

The combination of multiple NLP techniques enabled a comprehensive analysis of tourist perspectives. The results confirmed the power of automated text analysis in identifying both positive experiences and structural issues

within popular tourist destinations. By merging sentiment, emotion, topic, and issue outputs, a more holistic understanding of tourism data emerged.

The system also demonstrates scalability, allowing future integration of multilingual datasets, additional destinations, or real-time social media feeds.

8. Conclusion

This project successfully developed a unified NLP-driven analytical system for monitoring and understanding tourist experiences across India. The system collected real text from online platforms, cleaned and structured the data, applied multiple analytical methods, and consolidated insights into an intuitive dashboard. The findings reveal strong tourist satisfaction in natural destinations, but highlight recurring concerns in urban areas.
The project showcases the effectiveness of NLP in tourism research and decision-making, offering a foundation for advanced tourism intelligence platforms.

9. Future Scope

To enhance the system further, the following extensions are recommended:

1. **Multilingual Support for Hindi, Bengali, Tamil, Malayalam, etc.**
2. **Deep Learning Models like BERT/RoBERTa for emotion and sentiment analysis.**
3. **Real-Time API Integration for Twitter, Reddit, and Instagram.**
4. **Predictive Analytics to forecast tourist satisfaction based on text patterns.**
5. **Deployment as a public dashboard or enterprise SaaS product.**
6. **Geo-tagged Reviews enabling cluster-level tourism insights.**