Machine Learning & Natural Language Processing in Tourism

Course: AI and Tourism – MIT-AI @ Gandaki University

Bidur Devkota, PhD
GCES Pokhara

December 24, 2025

Agenda

Introduction to ML & NLP in Tourism

Theory Definitions

- **Machine Learning (ML):**
  - Subfield of AI
  - Uses experience (data) to improve performance
  - Makes predictions without explicit programming
- **Natural Language Processing (NLP):**
  - Translates human language into computable data
  - Enables computers to learn about the world from text

Relevance to Tourism

- Tourism requires extensive coordination
- Understanding individual traveler needs is crucial
- ML & NLP enable "Smart Tourism"
- Process big data from sensors and UGC
- Provide personalized recommendations in real-time

Illustration: Phygital Experience

Digital footprints (reviews, social media) → **ML analysis** → **Predict physical needs** at destination

Key Insight

Tourism industry generates massive structured and unstructured data perfect for ML & NLP applications

What is Machine Learning?

Arthur Samuel's Definition (1959)

"Field of study that gives computers the ability to learn without being explicitly programmed."

Formal Definition (Tom Mitchell)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Tourism Application

- **E:** Historical booking data and reviews
- **T:** Predicting tourist arrivals
- **P:** Mean Absolute Percentage Error (MAPE)

ML Paradigms: The Four Learning Types

1. Supervised Learning

- **Input:** Labeled data (X, y)
- **Goal:** Learn mapping $f : X \to Y$
- **Tasks:** Classification, Regression
- **Tourism Use:** Price prediction, Customer classification

3. Semi-Supervised Learning

- **Input:** Mixed labeled/unlabeled data
- **Goal:** Leverage all available data
- **Tourism Use:** Review classification with limited labels

2. Unsupervised Learning

- **Input:** Unlabeled data X
- **Goal:** Discover hidden patterns
- **Tasks:** Clustering, Dimensionality Reduction
- **Tourism Use:** Customer segmentation, Topic modeling

4. Reinforcement Learning

- **Input:** Environment feedback
- **Goal:** Maximize cumulative reward
- **Formulation:** Learn policy $\pi : S \to A$
- **Tourism Use:** Dynamic pricing, Route optimization

Supervised Learning Algorithms

Algorithm	Mathematical Formulation	Tourism Use Case
Linear Regression	$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$	Tourist spending prediction
Logistic Regression	$P(y = 1\|x) = \frac{1}{1+e^{-(\beta_0+\beta x)}}$	Booking conversion prediction
Support Vector Machine	$\min_{w,b} \frac{1}{2}\|w\|^2$ s.t. $y_i(w \cdot x_i + b) \geq 1$	Superhost classification
Random Forest	$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$	Demand forecasting
XGBoost	$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$	Arrival prediction
Multi-layer Perceptron	$a^{(l+1)} = \sigma(W^{(l)}a^{(l)} + b^{(l)})$	Deceptive review detection

Unsupervised Learning Algorithms

K-Means Clustering

Objective:

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Use: Destination grouping, Tourist segmentation

Latent Dirichlet Allocation (LDA)

Generative Process:

$$p(w, d) = p(d) \sum_{z} p(w|z)p(z|d)$$

Use: Discovering activity preferences from itineraries

DBSCAN (Density-Based)

Core point: MinPts in ϵ-neighborhood **Formulation:** Density-connected clusters **Use:** Spatial clustering of tourist trajectories

Principal Component Analysis

$$\Sigma = \frac{X^T X}{n-1}, \quad \Sigma v = \lambda v$$

Use: Feature extraction from reviews

Predictive Modeling: Theory & Applications

Theory: Time Series Analysis & Regression

- Forecasts tourism demand using historical patterns
- Identifies statistical relationships between variables
- Predicts future patterns: arrivals, expenditures, preferences
- Combines traditional data with digital sources

Tourist Arrivals Prediction

Key Challenge: Seasonal variations, external factors **Data Sources:**

- Historical arrival data
- Economic indicators
- Web search data (Google Trends)
- Social media mentions
- Flight bookings data

Tourist Preferences Prediction

Key Challenge: Evolving preferences, personalization **Data Sources:**

- Past bookings
- Reviews and ratings
- Social media activity
- Browsing history
- Demographic information

Research: Tourism Forecasting with Internet Data

Title: "Review of tourism forecasting research with internet data", 2020
Objective and Scope

- Conducted a systematic review of **47 published studies** to evaluate the use of Internet big data in tourism and hotel demand forecasting .
- Analyzed research published between **2012 and 2019** across four data categories: search engines, web traffic, social media, and multi-source data .

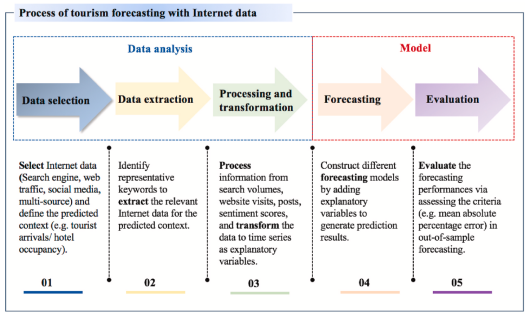


Fig. 3. Process of tourism forecasting with Internet data.

Research: Tourism Forecasting with Internet Data (cont...)

Title: "Review of tourism forecasting research with internet data", 2020
Key Data Sources

- **Search Engine Data (53%):** The most dominant source, primarily leveraging Google Trends and Baidu Index for structured time-series analysis .
- **Social Media & Web Traffic:** Increasing use of unstructured data (reviews/photos) and DMO website visits to capture tourist sentiment and intent .

Methodologies and Results

- **Model Dominance: Time series** (ARIMA , SARIMA) and **econometric models** (Gravity Model) remain the most common, while **Artificial Intelligence (AI), Machine Learning and Deep Learning** are growing in popularity .
- **Accuracy:** Internet data consistently serves as an important supplement to traditional data, significantly **improving forecasting accuracy** .
- **Emerging Trends:** Multi-source data and hybrid models are expected to provide more robust and timely (daily/hourly) forecasts in the future .

Case Study:Survey-based methodology to measure how tourists' feelings and expectations were influenced by the UGC

Journal: "User-Generated Content Sources in Social Media: A New Approach to Explore Tourist Satisfaction", 2018
Objectives

- Analyze how different **User-Generated Content (UGC)** sources influence tourist satisfaction through the formation of expectations and perceptions .
- Examine the impact of **strong-tie** (family/friends), **weak-tie** (strangers), and **tourism-tie** (organizations) sources on destination characteristics.

Methodology

- **Survey:** 375 valid responses from tourists in **Valencia, Spain** who used social media for pre-travel information.
- **Framework:** Used **Structural Equation Modeling (SEM)** to test relationships between UGC sources, expectations, and real perceptions of **Core Resources** (culture, history) and **Supporting Factors** (safety, hospitality).

Case Study:Survey-based methodology to measure how tourists' feelings and expectations were influenced by the UGC (Condt ...)

Journal: "User-Generated Content Sources in Social Media: A New Approach to Explore Tourist Satisfaction", 2018
Key Findings

- **Indirect Effect:** UGC sources influence satisfaction indirectly by shaping expectations, which tourists later compare with real experiences.
- **Source Variance:** Strong and tourism ties influence both core and supporting expectations; however, **weak-tie sources (strangers) only influence expectations of supporting factors**.
- **Assimilation Theory:** Confirmed that tourists adjust their perceptions to match pre-travel expectations formed on social media .

Managerial Implications

- Social media sets **expectations** (through reviews, posts, influencers).
- Social media affects not only the **decision to buy**, but also **how satisfied** people feel afterward.
- **DMOs** must maintain active social media presences (photos/videos) as they are the most trusted source for factual core resource information .

Research: Discovering Activity Preferences

Journal of Tourism Management: "Discovering implicit activity preferences in travel itineraries by topic modelling , 2019"

Objective and Methodology

- Aimed to uncover **latent activity preferences** (e.g., dining, shopping, sightseeing) hidden within complex travel itineraries .
- Analyzed a large-scale dataset (**Foursquare/Twitter**) of **12,446 daily itineraries** constructed from the venue check-ins of 4,077 international tourists .
- Applied **Latent Dirichlet Allocation (LDA)** to discover the hidden semantic structures (topics) of travel behavior .

Key Findings

- **Thematic Itineraries:** Identified travel themes: T1 (Sightseeing/Monuments), T5 (Shopping/Clothing Stores), and T24 (Theme Parks).
- **Pervasiveness of Dining:** Food-related activities appeared with high probability across most topics, reflecting its essential role in tourism.
- **Regional Preferences:** Notable differences were found between groups; e.g., Japanese tourists preferred T1 (Sightseeing), while Thai tourists preferred T2 (Dining/Shrines) .

Research: Discovering Activity Preferences (condt...)

Journal: "Discovering implicit activity preferences in travel itineraries by topic modelling , 2019"

Business Implications

- Provides a framework for **targeted marketing** and the development of travel packages with an appropriate mixture of activities .

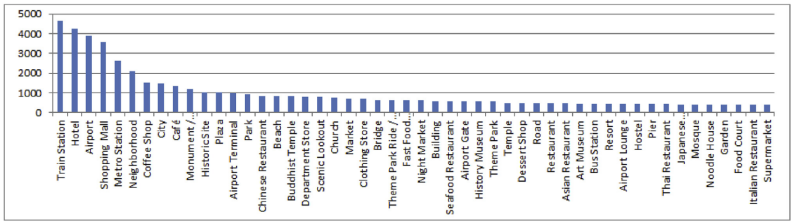


Fig. 2. Most frequent venue categories.

Research: Discovering Activity Preferences (condt...)

Journal: "Discovering implicit activity preferences in travel itineraries by topic modelling , 2019"

Identified 24 distinct topics like :

- **T1 (Sightseeing):** The most popular topic, characterized by visits to **Plazas, Churches, Monuments, and Landmarks** .
- **T2 (Japan Dining):** Centered on dining at **Ramen Restaurants** and visiting **Shrines** .
- **T3 (Taiwan Dining):** Highlights visits to **Chinese Restaurants, Noodle Houses, and Night Markets**.
- **T5 (Pure Shopping):** Focused heavily on **Department Stores, Clothing Stores, and Boutiques** .
- **T7 (Thailand Dining):** Features **Thai Restaurants, Flea Markets, and Night Markets**.
- **T10 (Transit):** Primarily related to travel infrastructure like **Airports and Train Stations**.
- **T24 (Entertainment):** Specifically identified for visits to **Theme Parks**.

Application Areas in Tourism

Analogy: ML as a Specialized Telescope

- **Wide-angle lens (Clustering):** See the "big picture" - where people are going
- **High-magnification lens (Neural Networks):** Understand "fine print" - what they are saying
- **Remember context:** Focus on words while understanding full sentence meaning

Non-NLP Applications

- Spatial structure analysis
- Movement pattern mining
- Host classification
- Demand forecasting
- Price optimization
- Anomaly detection

NLP Applications

- Sentiment analysis
- Topic modeling
- Review classification Chatbot development
- Content generation
- Multilingual processing

Research: Superhost Classification (Supervised)

Journal: "Value proposition operationalization in peer-to-peer platforms using machine learning", 2021

- The researchers employed **machine learning(like SVM)** to identify key variables that determine **Airbnb Superhost status**.
- More than 250 variables from 5136 listings were analyzed in the Canary Islands region.

Objectives:

- Identify variables contributing to **"Superhost"** status
- Operationalize value proposition for sharing economy
- Provide data-driven insights for hosts

Key Findings:

- **Communication**(e.g., reviews and responses) and **value factors** (ratings, cancellation policy) are most important; **property features** matter least.
- **Guest activity**, esp **review count**, is the strongest predictor.
- **High ratings, many reviews, and low cancellations** distinguish top hosts.
- **Geographical location** has little impact on Superhost status.

Research: Spatial Clustering of Trajectories (Unsupervised)

Journal: "Spatial structures of tourism destinations: A trajectory data mining approach leveraging mobile big data", 2020

- Analyzed **mobile roaming data** from **116,807 international travelers** in three **South Korean cities**.
- Applied **DBSCAN** to identify **spatial hot spots** and **SPADE** to extract **frequent travel routes**.

Key Findings

- **Highly concentrated demand**: visitors covered only a **small share of destinations**.
- **Polycentric structures** with **multiple centers** rather than single hubs.
- Distinct **movement patterns**: **circular loops**, **linear chains**, and **radiating hubs**.

Planning Implications

- Enables **collaborative destination management** across **interconnected hot spots**.
- Supports **infrastructure optimization**, **smart mobility**, and **dynamic recommendations**.

Research: Deceptive Review Detection

Journal: "Identification of the deceptive reviews in the hospitality sector", 2019
Objectives

- Develop an automatic machine learning tool to distinguish between **deceptive and non-deceptive reviews** in the hospitality sector .
- Analyze how **sentiment polarity** (positive vs. negative) influences the effectiveness of review classification .

Methodology

- Analyzed a dataset of **1,600 annotated reviews** (800 honest, 800 deceptive) from 20 popular Chicago hotels .
- Used a **review-centric approach** focused on content analysis and **unique attributes** selected via ANOVA and Turkey's HSD test .
- Compared six classifiers, including k-NN, Logistic, Random Forest, and **Support Vector Machines (SVM)** .

Research: Deceptive Review Detection (Condt ...)

Journal: "Identification of the deceptive reviews in the hospitality sector", 2019
Key Findings

- **SVM** was the most effective classifier, maintaining high accuracy while reducing the number of attributes from 918 to 134 .
- **Deceptive positive reviews** often emphasize hotel meals and location, while **non-deceptive** ones focus on global experiences and feelings .
- **Deceptive negative reviews** focus on tangible issues (e.g., broken items), whereas **non-deceptive** complaints emphasize bad experiences and feelings .

Managerial Implications

- Enables review site operators to implement **automatic detection systems** to maintain user trust.
- Helps hotel managers distinguish **authentic negative reviews** to prioritize genuine customer service recovery .

Sentiment Analysis: Theory & Approaches

Sentiment Analysis: Computational study of people's attitudes and emotions toward entities. Measures emotional valence (positive to negative) in text.

Lexicon-Based Approaches

- Uses pre-defined sentiment dictionaries
- **VADER:** Rule-based, understands social media context
- **SentiWordNet:** Assigns positivity/negativity scores
- **Pros:** Fast, no training needed
- **Cons:** Limited context understanding

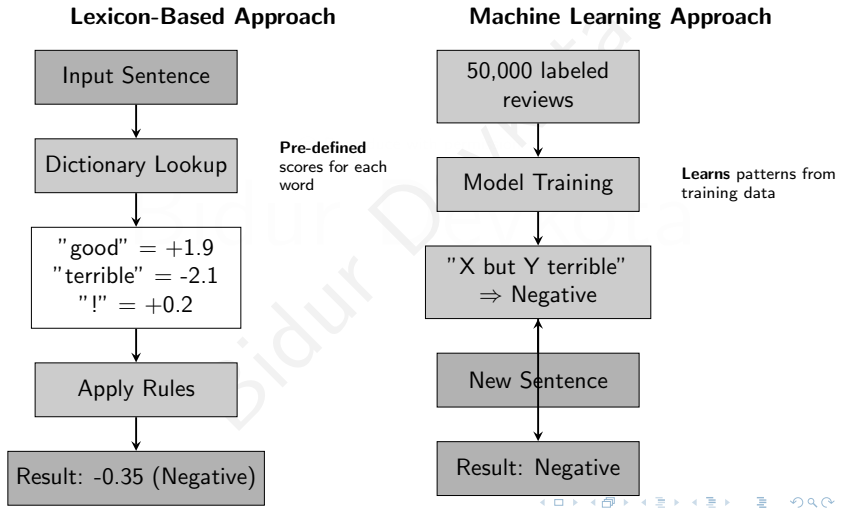
Machine Learning Approaches

- Trains on labeled samples
- **Traditional ML:** SVM, Naive Bayes, Random Forest
- **Deep Learning:** RNN, CNN, BERT, Transformers
- **Pros:** Context-aware, adaptable
- **Cons:** Requires labeled data, computationally intensive

Example Sentence

"The movie was good, but the ending was terrible!"

Visual Comparison: How Each Method Works



Research: Topic Modeling of Complaints

Journal: "What do hotel customers complain about? Text analysis using structural topic model", 2019

Objectives

- Identify the primary causes of hotel customer dissatisfaction through a comparison of positive and negative reviews .
- Examine how types of complaints systematically vary across different hotel grades (star classes).

Methodology

- Analyzed a balanced corpus of **27,864 Tripadvisor reviews** for 315 hotels in New York City.
- Applied the **Structural Topic Model (STM)**, which incorporates document metadata (sentiment and hotel grade) into the topic generation process.
- Determined an optimal **30-topic model** to ensure semantic coherence and exclusivity.

Research: Topic Modeling of Complaints (Condt ...)

Journal: "What do hotel customers complain about? Text analysis using structural topic model", 2019

Key Findings

- Identified **10 distinct negative topics**, including severe service failure, dirtiness, booking and cancellation, room type, and overcharging.
- **High-end Hotels:** Complaints are primarily driven by **intangible service failures** and pricing issues.
- **Low-end Hotels:** Dissatisfaction is mainly rooted in **tangible facility issues** and cleanliness standards.
- **Universal Issues:** Booking/cancellation and room type discrepancies impact dissatisfaction across all hotel grades.

Managerial Implications

- Encourages the development of automated monitoring platforms using STM to track dynamic customer feedback.
- Enables managers to prioritize interventions based on hotel grade, such as improving service processes for luxury brands or core facility maintenance for budget brands .

Future Directions (2025+)

Technological Advances & Trends

- **Multimodal AI:** Analyze text, images, audio, and video
- **Generative AI:** Personalized content creation
- **Explainable AI:** Transparent decision-making
- **Federated Learning:** Privacy-preserving collaboration
- **Edge Computing:** Real-time on-device processing
- **Democratization:** No-code AI tools for SMEs
- **Regulation:** AI governance frameworks
- **Collaboration:** Open data and model sharing

Industry Applications

- Real-time crisis response
- Sustainable tourism optimization
- Hyper-personalized experiences
- Predictive infrastructure maintenance
- Autonomous tour planning

Challenges & Ethical Considerations

Technical Challenges

- **Data Quality:** Inconsistent, noisy, biased data
- **Integration:** Legacy systems, data silos
- **Multilingual:** 50+ languages, cultural nuances
- **Real-time Processing:** Speed and scalability
- **Model Maintenance:** Concept drift, retraining

Ethical Considerations

- **Privacy Protection:** GDPR, CCPA compliance
- **Transparency:** Clear data usage policies
- **Fairness:** Regular bias audits and mitigation
- **Accountability:** Human oversight and responsibility
- **Sustainability:** Environmental impact of AI

Algorithmic Challenges

- **Context Understanding:** Sarcasm, idioms, ambiguity
- **Bias Mitigation:** Fairness across demographics
- **Explainability:** Black box model interpretation

Conclusion

1. AI transform tourism from **reactive** to **predictive** & **prescriptive**
2. Start with clear **business problems**, not just technology
3. **Data quality** and **domain knowledge** are critical success factors
4. **Ethical implementation** builds sustainable competitive advantage
5. **Human + AI** collaboration delivers optimal results
6. Continuous **learning** and **adaptation** are essential

Thank You!

Questions & Discussion