Exploratory Analysis and Sentiment Mining of Destination Reviews in Sri Lanka

Hannah Cinderella L | Kishan V | Dr. Pattabiraman V | School of Computer Science Engineering

INTRODUCTION

Tourism in Sri Lanka is a thriving industry fueled by the country's scenic beauty and cultural heritage. With the increasing influence of digital platforms, tourists leave reviews that contain rich, unstructured data.

Project Objectives:

- Understand travel preferences via review analysis.
- Group destinations by sentiment using clustering.
- Recommend travel paths based on co-visitation patterns.
- Build predictive models to classify review sentiment.

SCOPE OF THE PROJECT

This project focuses on extracting meaningful insights from user - generated reviews of Sri Lankan tourist destinations. By leveraging Exploratory Data Analysis, sentiment mining, clustering, and predictive modelling, the project aims to:

- Identify popular tourist attractions based on review volume and sentiment.
- Classify tourist experiences into positive and negative groups using machine learning.
- Discover frequently co-visited destination pairs to enhance travel planning.
- Provide data-backed recommendations for tourism boards and travel service providers.

METHODOLOGY

Step 1: Data Cleaning & Sentiment Extraction (TextBlob)
Step 2: TF-IDF Vectorization (converts text to numerical features) & Variance Threshold Feature Selection (removes low-variance features)
Step 3: EDA — Visualize Top 20 Reviewed Destinations, Word Frequency (Word Cloud), and print frequent destination pairs using association-style logic
Step 4: DBSCAN Clustering on TF-IDF + SVD-Reduced Features
Step 5: Association Rule Mining — Use co-occurrence of destinations per time span to extract frequent destination pairs
Step 6: Predictive Modeling — Train and compare: Logistic Regression, Naive Bayes, Random Forest
Step 7: Evaluation using Accuracy & F1-score
Step 8: Resampling Techniques — Apply Random Oversampling to balance class distribution

```
Cluster Distribution:

Cluster
 0    1652
-1     252
 5      29
 2      20
 1      17
 3       9
 4       7
 6       7
 7       7
Name: count, dtype: int64
```

Figure 1: DBSCAN Clustering

```
Top 10 Most Frequently Visited Destination Pairs:

                                        Destination Pair  Count
                (Horton plains national park, Moon plains) 102013
                                (Bambarakiri ella, Riverston)  94811
                                (Riverston, Sembuwatta lake)  83207
                (Pitawala pathana , riverston, Riverston)  76771
                                (Bambarakiri ella, Sembuwatta lake)  70201
                                (Riverston, Sera ella water falls )  65563
        (Bambarakiri ella, Pitawala pathana , riverston)  64324
                                (Riverston, Sri muththumari amman kovil)  64144
                        (Kurunagela clock tower, Yapahuwa rock fortress)  63291
(Horton plains national park, Horton plains national park)  63257
```

Figure 2: Association Rule Mining

RESULTS

Top 20 Destinations Identified: Horton Plains, Sigiriya, Ella, Nuwara Eliya, Mirissa, Galle Fort, Yala National Park, Anuradhapura, Polonnaruwa, and others.

Word Cloud showed frequent mentions of "beautiful", "nature", "peaceful", "waterfall", "temple", "historic".

Association Rules: "Horton Plains" often co-mentioned with "Bakers Falls", and "Nuwara Eliya" – indicating popular travel circuits.

"Galle" with "Mirissa" and "Unawatuna" for beach tourism

DBSCAN Clusters: 3 major clusters indicating positive, neutral, and negative sentiment zones.

Predictive Modeling:

Class Imbalance Handling:

- Applied upsampling to balance sentiment classes.

Model Performance

Model	Accuracy	F1 Score	Precision	Best performer in
Logistic Regression	91.4%	0.91	0.92	Precision ☑
Naive Bayes	86.3%	0.86	0.86	---
Random Forest	95.5%	0.95	0.96	All metrics ☑

CONCLUSION

Sentiment analysis reveals - **70% of reviews are positive.**

Top attractions are mainly **nature parks**, **waterfalls**, **beaches**, and **historical sites**.

Random Forest achieved best sentiment prediction **(95.5%)**

Co-visit patterns can help plan better tour packages.

DBSCAN effectively grouped sentiment-based clusters.

Feature extraction significantly improved predictive performance.

FUTURE SCOPE

Add confusion matrix & sentiment-wise feature importance.

CONTACT DETAILS

Hannah Cinderella L - hannahcinderella.l2023@vitstudent.ac.in

Kishan V - kishan.v2023@vitstudent.ac.in

GitHub Repository: https://github.com/Hannah-Cinderella/EDA_DestinationReviews_Srilanka

REFERENCES

Dataset: (https://www.kaggle.com/datasets/nethumdperera/travel-destinations-reviews-in-sir-lanka)

Libraries Used: pandas, scikit-learn, TextBlob, matplotlib, seaborn, imbalanced-learn

Tools: Jupyter Notebook, Python 3.9, WordCloud, DBSCAN, Apriori (mlxtend).

Research article: https://www.mdpi.com/2071-1050/14/15/9572