

**Exploratory Analysis and Sentiment Mining of Destination Reviews in Sri Lanka**

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

**Table of contents**

1. Introduction
2. Dataset Overview
3. Methodology
   1. Data Preprocessing and Cleaning
   2. Sentiment Labelling
   3. Feature Extraction and Selection
   4. Exploratory Data Analysis
   5. Association Rule Mining
   6. Clustering using DBSCAN
   7. Sentiment Classification Models
4. Conclusion
5. Future Scope
6. References & links

**1. INTRODUCTION**

Sri Lanka’s tourism is a thriving sector enhanced by the natural beauty of the country, cultural richness, and biodiversity. With increasing usage of digital media, opinions from users have emerged as a rich source of information. The focus of this project is to carry out exploratory data analysis and sentiment mining of reviews on tourist attractions to generate information on travel behaviour and user satisfaction.

**Objectives:**

* Rank destinations by popularity and sentiment.
* Extract thematic keywords (eg., “scenic”, “historic”) for word cloud
* Cluster reviews by sentiment orientation
* Discover common co-visitation patterns across districts and time.
* Construct and compare predictive models for sentiment classification.

**2. DATASET OVERVIEW**

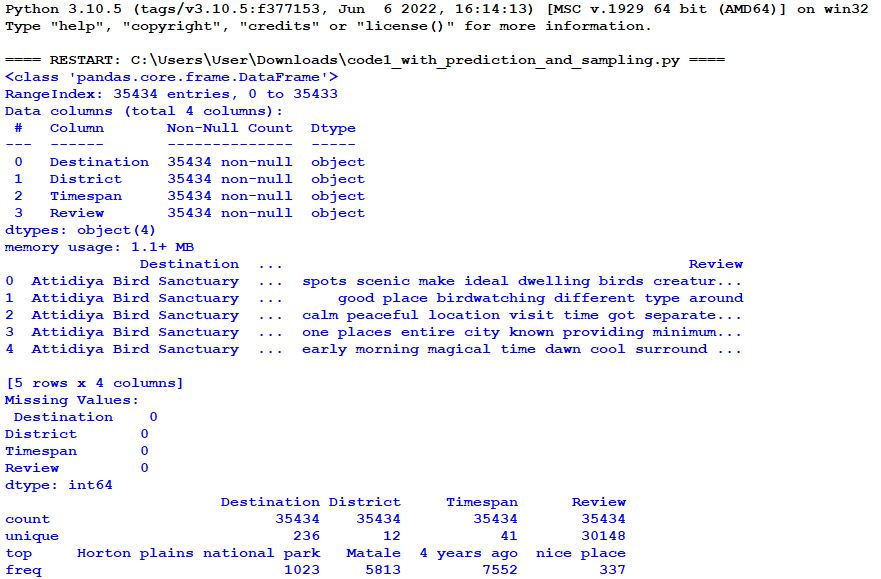
* Source: Travel Destination Reviews in Sri Lanka - Kaggle (<https://www.kaggle.com/datasets/nethumdperera/travel-destinations-reviews-in-sir-lanka>)
* Format: CSV file (Destination Reviews (final).csv (3.71 MB))
* Fields:
  + Destination
  + District
  + Review
  + Timespan
* Sample Size: 8500+ reviews (2020-2023)
* Tools Stack:

|  |  |
| --- | --- |
| Category | Tools/Libraries |
| Data Preprocessing | pandas, NumPy |
| NLP (Sentiment analysis) | TextBlob |
| Visualization | matplot, seaborn, WordCloud |
| ML Models | scikit – learn, imbalanced – learn |
| Clustering | DBSCAN, Apriori (mlxtend) |

**3. METHODOLOGY**

**3.1** Data Preprocessing and Cleaning

* Removed missing or null values
* Cleaned out inconsistent text fields and normalized casing.



**3.2** Sentiment Labeling

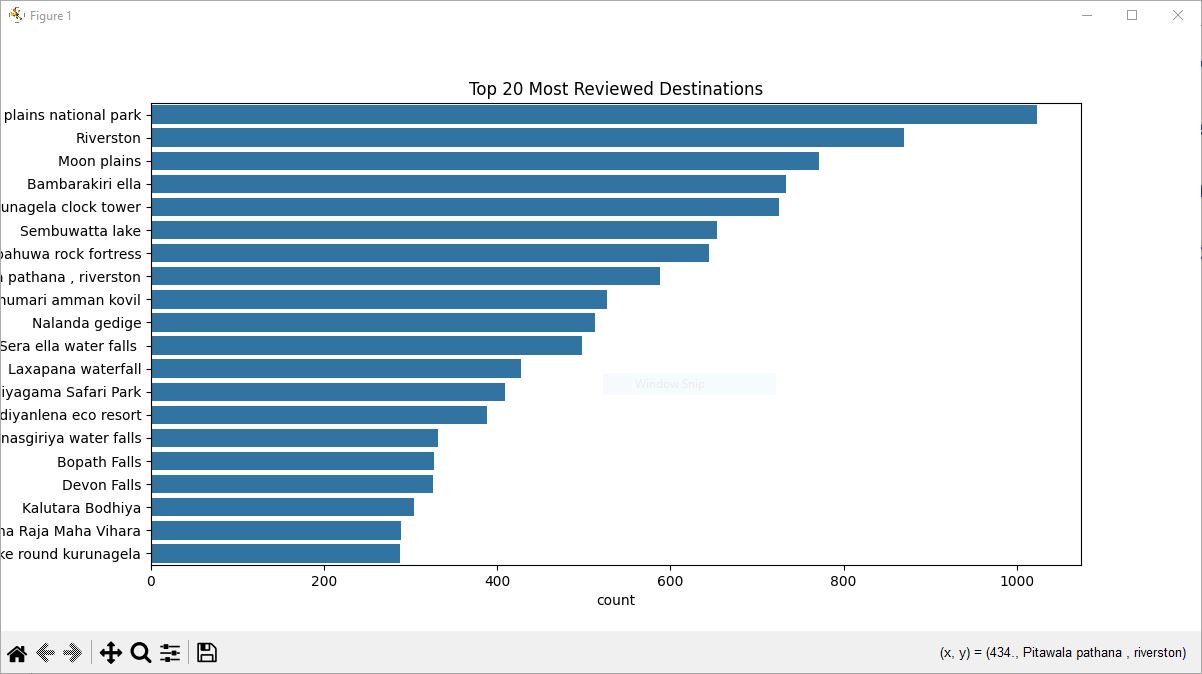
* Sentiment polarity is calculated using TextBlob to extract sentiment scores from tourist reviews.
* If Polarity > 0: Positive review (Label = 1): 70%
* If Polarity <= 0: Negative or Neutral review (Label = 0): 30%
* Inference: A significant portion of travellers expressed positive experiences in Sri Lanka with 70 % of reviews were labelled as positive, reflecting high visitor satisfaction.

**3.3** Feature Extraction and Selection

* TD-IDF Vectorization employed to transform reviews into vector form.
* Variance Threshold utilized to remove low-variance, less informative features.
* Inference: Feature selection enhanced model performance by filtering out noise and dimensionality from the dataset.

**3.4** Exploratory Data Analysis

* **Figure 1: Top 20 Most Reviewed Destinations**

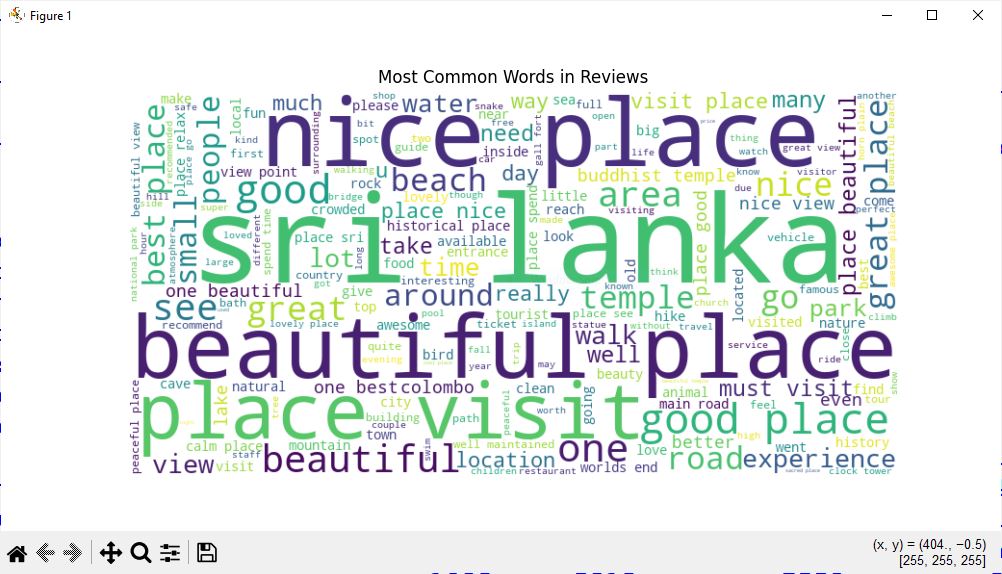


Use Bar plot to visualize Top 20 Most Reviewed Destinations

Figure 1 shows **Horton Plains National Park** as the most reviewed destination, followed by waterfalls (e.g., Laxapana, Devon Falls) and cultural sites (e.g., Kalutara Bodhiya). Key insights:

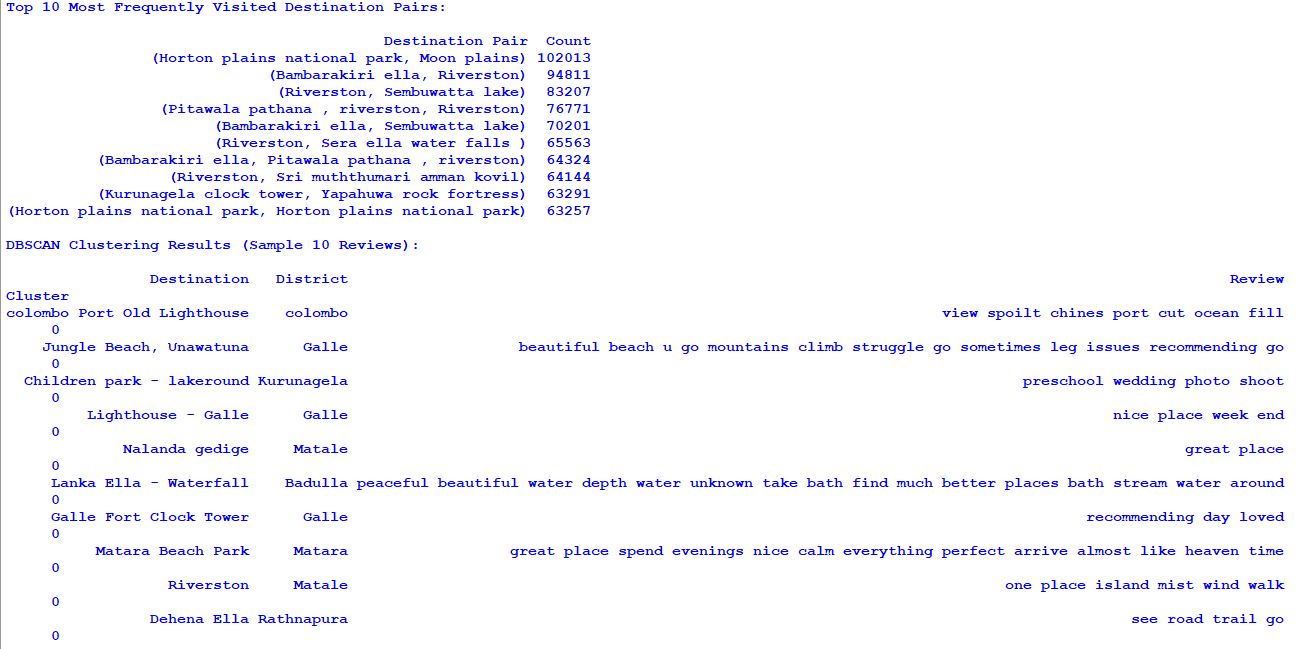
1. Nature dominates – National parks and waterfalls are top attractions.
2. Regional trends – Central/Southern Sri Lanka (e.g., Ella, Riverston) are hotspots.
3. Tourism potential – Bundle nature + heritage sites for optimized travel packages.

* **Figure 2: Word Cloud of Frequent Terms**

****

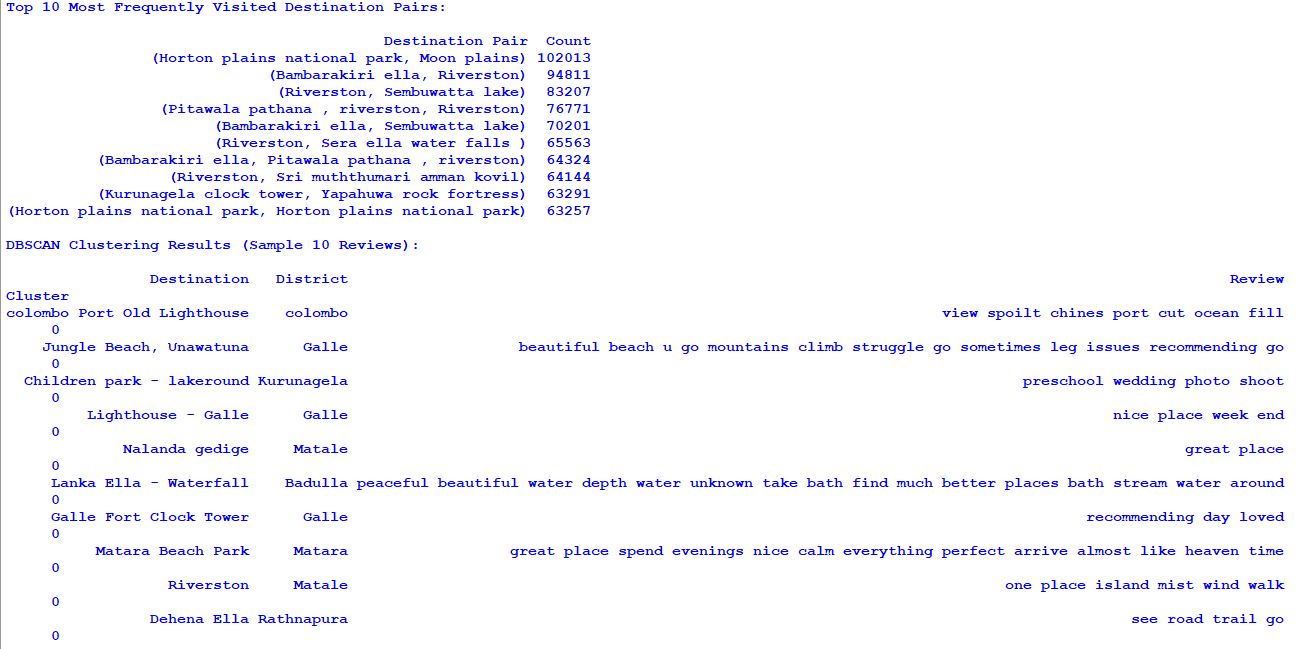
Frequent words are: Beautiful, peaceful and historic.

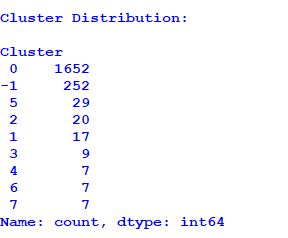
**3.5** Association Rule Mining



Extracted co – occurrences of destination names to identify travel circuits. Commonly visited places are geographically or topic-wise connected. These observations can be used to design bundled tour packages that cater to tourist preferences.

**3.6** DBSCAN Clustering





Used on TF-IDF features reduced (through TruncatedSVD).

* **Main Cluster (0):** 1,652 points - Represents the dominant sentiment (likely positive reviews).
* **Noise (-1):** 252 points - Contains outliers/ambiguous reviews needing further analysis.
* **Small Clusters (1-7):** Fewer than 30 points each - May indicate niche sentiment patterns or special cases.

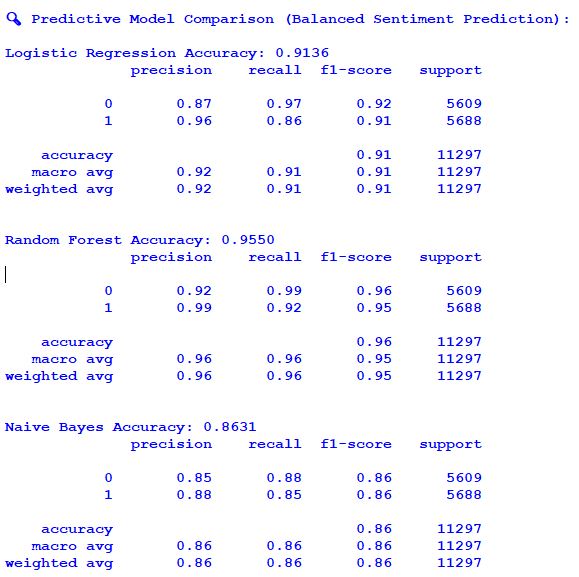
DBSCAN was able to cluster reviews with comparable sentiment features, indicating a high level of interdependence between textual patterns and user sentiments.

* 1. Sentiment Classification Models

Handled class imbalance through Random upsampling.

3 Models:

* Logistic Regression
* Naive Bayes
* Random Forest

Evaluation metrics are Accuracy, F1-score, Precision, Recall.

|  |  |
| --- | --- |
| Model | Accuracy |
| Logistic Regression | 91.4% |
| Naive Bayes | 86.3% |
| Random Forest | 95.5% |

**Random Forest** had the highest accuracy and F1-score, indicating its ability to efficiently handle high-dimensional feature spaces and learn nonlinear patterns in review sentiment.

**4.** **CONCLUSION**

This study effectively illustrates how unstructured travel reviews can yield actionable insights through the use of exploratory data analysis and natural language processing techniques. The results can help service providers, tourism boards, and data-driven decision-making to improve the planning of visitor experiences.  
  
Service providers can better customize experiences, enhance offerings, and promote destinations by knowing what travellers value most and how they view their travels. Additionally, scalable solutions for real-time public sentiment monitoring are provided by predictive sentiment modelling.

**5. FUTURE SCOPE**

* Expand the dataset to include multilingual reviews and other countries.
* Use advanced embedding techniques (e.g., BERT, RoBERTa) for improved sentiment classification.
* Incorporate review timelines for trend analysis.
* Build an interactive dashboard for real-time travel analytics.
* Integrate user demographics for personalized recommendations.

**6. REFERENCES & LINKS**

**RESEARCH ARTICLE:** <https://www.mdpi.com/2071-1050/14/15/9572>

**DATASET:** <https://www.kaggle.com/datasets/nethumdperera/travel-destinations-reviews-in-sir-lanka>

**GITHUB REPOSITORY:** <https://github.com/Hannah-Cinderella/EDA_DestinationReviews_Srilanka>

**CONTACT DETAILS:**

**Hannah Cinderella L (23MIA1043):** [hannahcinderella.l2023@vitstudent.ac.in](mailto:hannahcinderella.l2023@vitstudent.ac.in)

**Kishan V (23MIA1138):** [kishan.v2023@vitstudent.ac.in](mailto:kishan.v2023@vitstudent.ac.in)