VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

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

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:

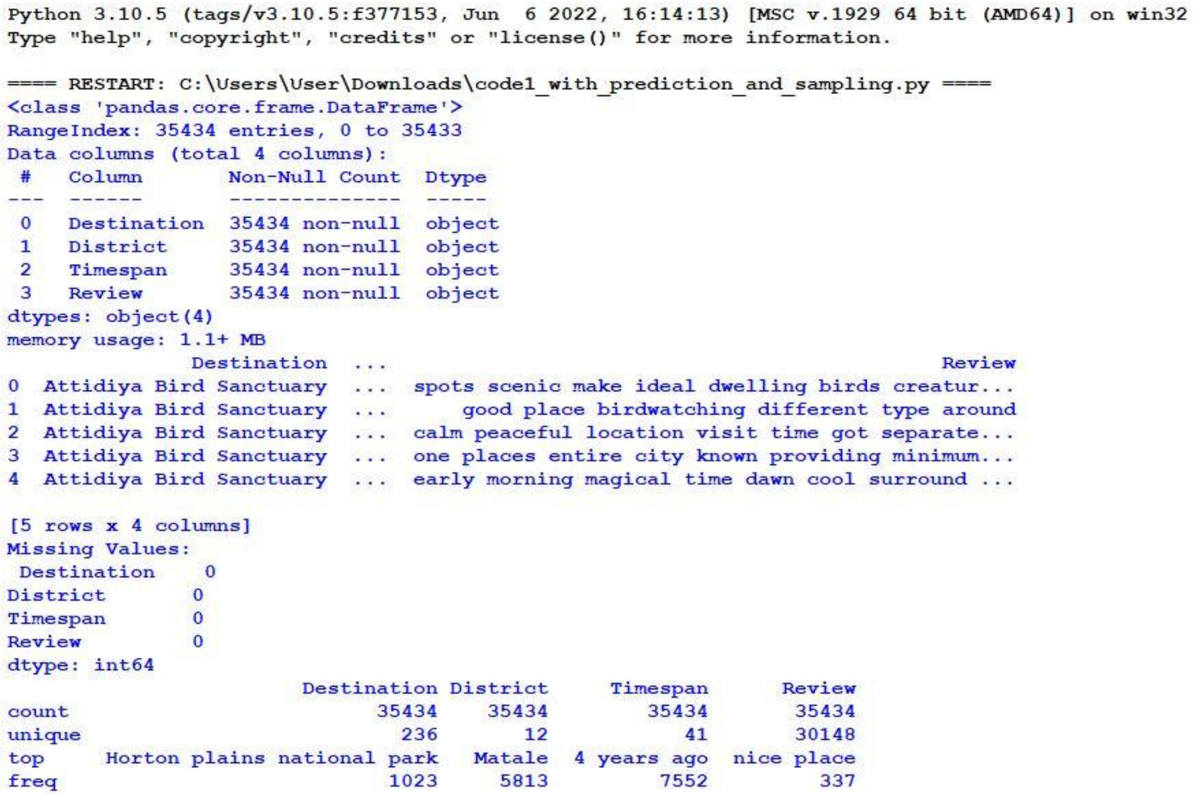
Category	Tools/Libraries
Data Preprocessing	pandas, NumPy
NLP (Sentiment analysis)	TextBlob
Visualization	matplot, seaborn, WordCloud
ML Models	scikit – learn, imbalanced – learn
Clustering	DBSCAN, Apriori (mlxtend)

---

3. METHODOLOGY

3.1 Data Preprocessing and Cleaning

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



```
Python 3.10.5 (tags/v3.10.5:f377153, Jun  6 2022, 16:14:13) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.

==== RESTART: C:\Users\User\Downloads\code1_with_prediction_and_sampling.py ====
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35434 entries, 0 to 35433
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   Destination  35434 non-null  object
 1   District     35434 non-null  object
 2   Timespan     35434 non-null  object
 3   Review       35434 non-null  object
dtypes: object(4)
memory usage: 1.1+ MB
               Destination  ...                                             Review
0  Attidiya Bird Sanctuary  ...  spots scenic make ideal dwelling birds creatur...
1  Attidiya Bird Sanctuary  ...      good place birdwatching different type around
2  Attidiya Bird Sanctuary  ...  calm peaceful location visit time got separate...
3  Attidiya Bird Sanctuary  ...  one places entire city known providing minimum...
4  Attidiya Bird Sanctuary  ...  early morning magical time dawn cool surround ...

[5 rows x 4 columns]
Missing Values:
 Destination    0
District       0
Timespan       0
Review         0
dtype: int64
                         Destination District     Timespan      Review
count                          35434    35434        35434       35434
unique                           236       12           41       30148
top     Horton plains national park    Matale  4 years ago  nice place
freq                            1023     5813         7552         337
```

3.2 Sentiment Labeling

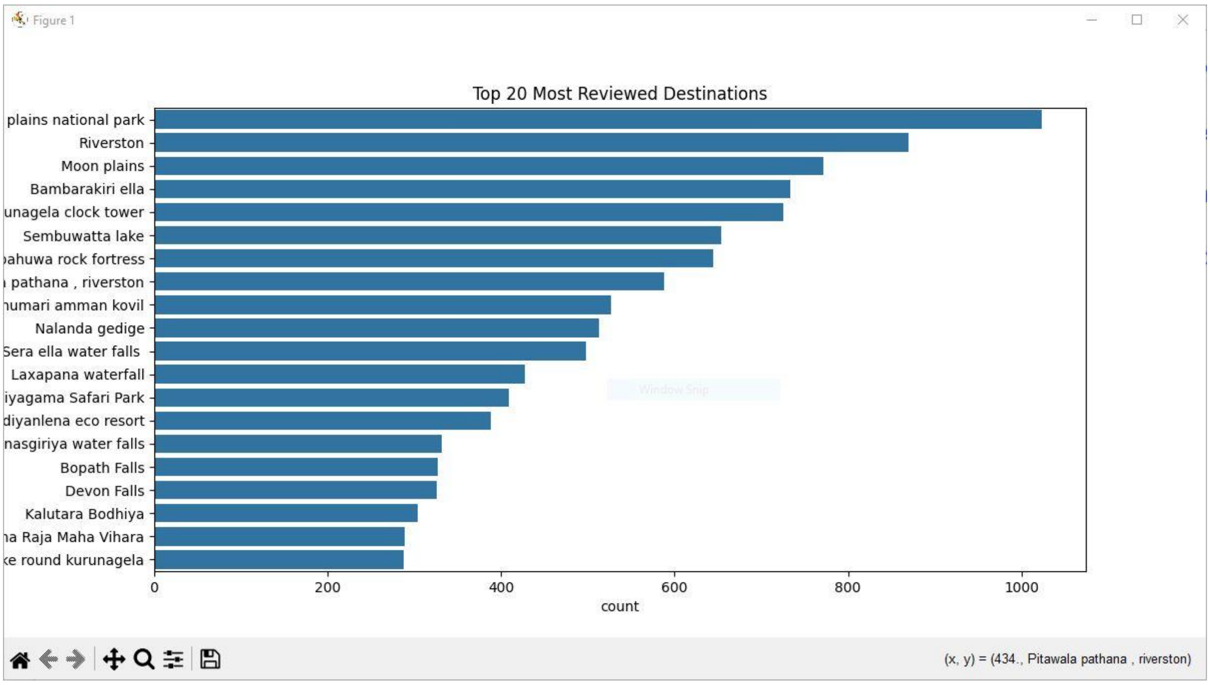
- Sentiment polarity is calculated using TextBlob to extract sentiment scores from tourist reviews.
- If Polarity > 0: Positive review (Label = 1): 70%
- If Polarity <= 0: Negative or Neutral review (Label = 0): 30%
- Inference: A significant portion of travellers expressed positive experiences in Sri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

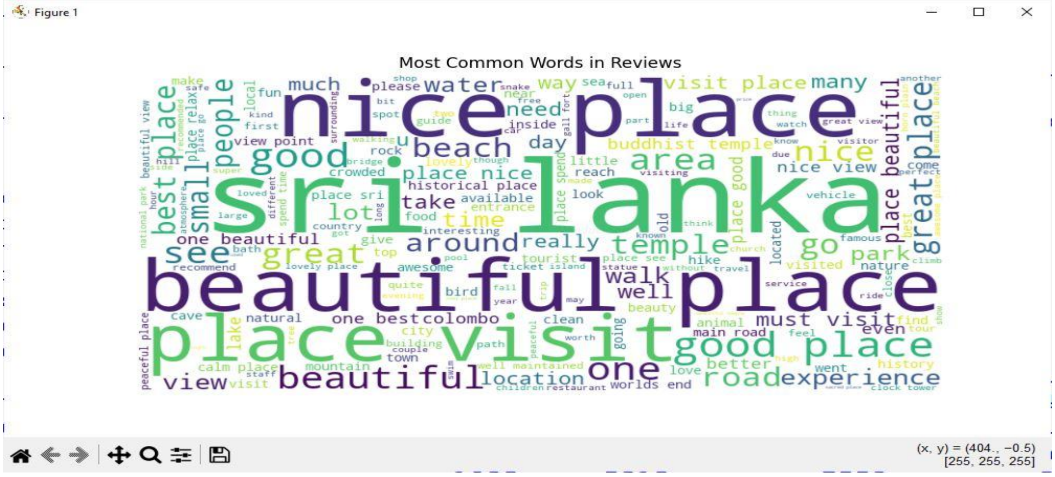


Use Bar plot to visualize Top 20 Most Reviewed Destinations

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:

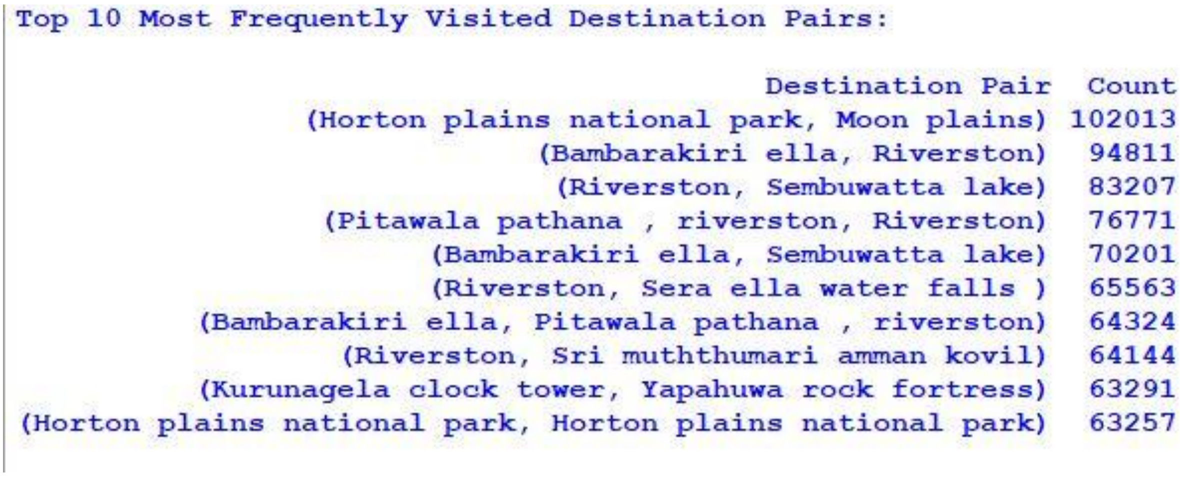
1. Nature dominates – National parks and waterfalls are top attractions.
2. Regional trends – Central/Southern Sri Lanka (e.g., Ella, Riverston) are hotspots.
3. Tourism potential – Bundle nature + heritage sites for optimized travel packages.

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



Frequent words are: Beautiful, peaceful and historic.

3.5 Association Rule Mining

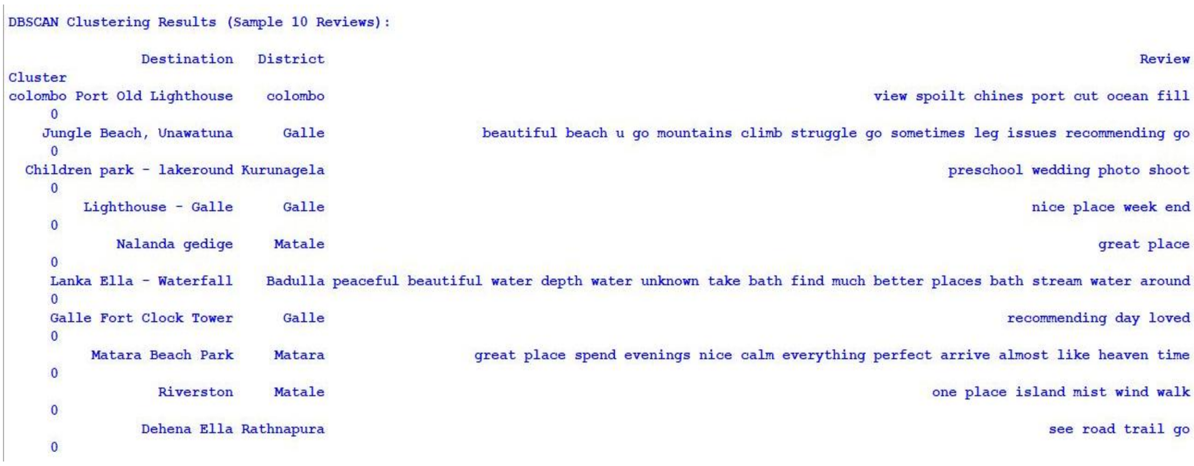


```
Top 10 Most Frequently Visited Destination Pairs:

                                          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
```

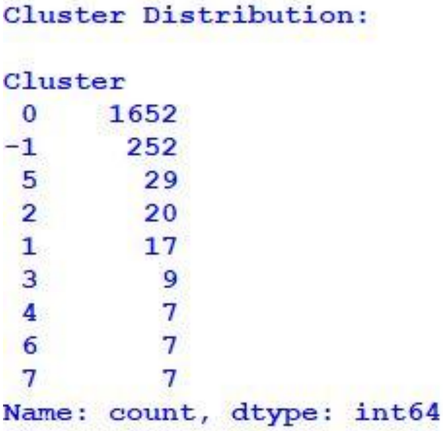
Extracted co – occurrences of destination names to identify travel circuits.
Commonly visited places are geographically or topic-wise connected. These observations can be used to design bundled tour packages that cater to tourist preferences.

3.6 DBSCAN Clustering



```
DBSCAN Clustering Results (Sample 10 Reviews):

                Destination   District                                                                                                  Review
Cluster
colombo Port Old Lighthouse    colombo                                                                  view spoilt chines port cut ocean fill
     0
    Jungle Beach, Unawatuna      Galle                  beautiful beach u go mountains climb struggle go sometimes leg issues recommending go
     0
  Children park - lakeround Kurunagela                                                                           preschool wedding photo shoot
     0
         Lighthouse - Galle      Galle                                                                                     nice place week end
     0
             Nalanda gedige     Matale                                                                                             great place
     0
     Lanka Ella - Waterfall    Badulla peaceful beautiful water depth water unknown take bath find much better places bath stream water around
     0
     Galle Fort Clock Tower      Galle                                                                                  recommending day loved
     0
          Matara Beach Park     Matara                 great place spend evenings nice calm everything perfect arrive almost like heaven time
     0
                  Riverston     Matale                                                                         one place island mist wind walk
     0
                Dehena Ella Rathnapura                                                                                       see road trail go
     0
```



```
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
```

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.

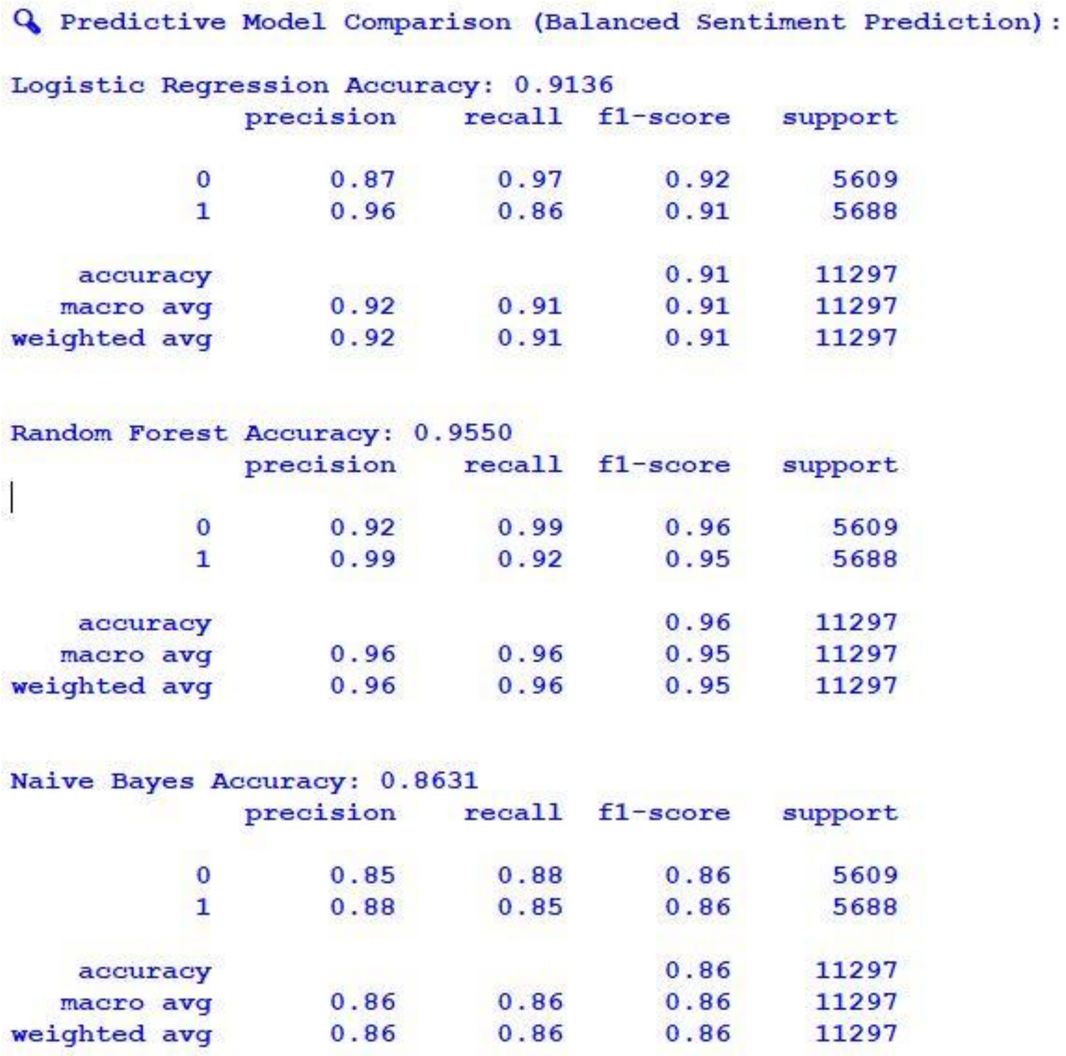
3.7 Sentiment Classification Models

Handled class imbalance through Random upsampling.

3 Models:

- Logistic Regression
- Naive Bayes
- Random Forest

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



```
🔍 Predictive Model Comparison (Balanced Sentiment Prediction):

Logistic Regression Accuracy: 0.9136
              precision    recall  f1-score   support

           0       0.87      0.97      0.92      5609
           1       0.96      0.86      0.91      5688

    accuracy                           0.91     11297
   macro avg       0.92      0.91      0.91     11297
weighted avg       0.92      0.91      0.91     11297


Random Forest Accuracy: 0.9550
              precision    recall  f1-score   support

           0       0.92      0.99      0.96      5609
           1       0.99      0.92      0.95      5688

    accuracy                           0.96     11297
   macro avg       0.96      0.96      0.95     11297
weighted avg       0.96      0.96      0.95     11297


Naive Bayes Accuracy: 0.8631
              precision    recall  f1-score   support

           0       0.85      0.88      0.86      5609
           1       0.88      0.85      0.86      5688

    accuracy                           0.86     11297
   macro avg       0.86      0.86      0.86     11297
weighted avg       0.86      0.86      0.86     11297
```

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

EDA POSTER

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.

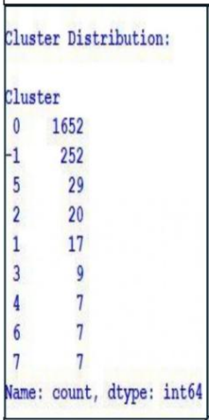
SCOPE OF THE PROJECT

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

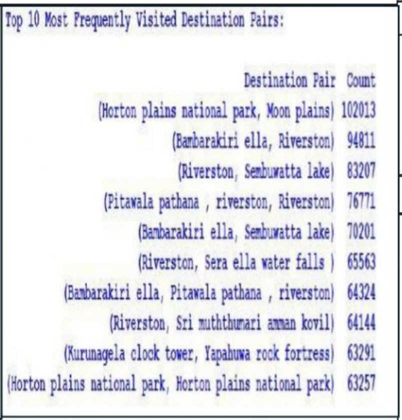
Step 1: Data Cleaning & Sentiment Extraction (TextBlob)
Step 2: TF-IDF Vectorization (converts text to numerical features) & Variance Threshold Feature Selection (removes low-variance features)
Step 3: EDA — Visualize Top 20 Reviewed Destinations, Word Frequency (Word Cloud), and print frequent destination 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

CONTACT DETAILS:

Hannah Cinderella L (23MIA1043): hannahcinderella.l2023@vitstudent.ac.in

Kishan V (23MIA1138): kishan.v2023@vitstudent.ac.in