Predicting Tourist Growth in Saudi Arabia: Trends and Impact on Vision 2030

Technical presentation

Overview:

1- Workflow Overview

2- Impact of Tourist Volume on Service Quality Perceptions

3- Predicting Spending Based on Length of Stay and Key attributes

4- Language sentiment analysis on google maps reviews

5- Cultural Heritage and Modern Attractions Comparative Analysis of Tourist Reviews

Overview:

For each:

- Algorithms and Rationale
- Model Evaluation
- Findings on Large Language Models
- Future Work

workflow overview

- Data Collection
- Data Preprocessing
- Model Training
- Model Evaluation

Impact of Tourist Volume on Service Quality Perceptions in Saudi Arabia’s Tourist Sites

Data Collection:

- Gathered data related to tourist volume from the primary data, including city-based and category-based features (e.g., city_Riyadh, categories_Resort hotel).
- Service Quality: totalscore or stars was used as the target variable.

Data Preprocessing:

- Handling Missing Values: Applied mean imputation to handle missing data in both features and target variables.
- Feature Selection: Chose relevant features that impact tourist volume and service quality.
- Data Splitting: Split data into 80% training and 20% testing to evaluate model performance on unseen data.

Algorithm Selection

Linear Regression:

- **Why Chosen:** provides a baseline model,Simple, interpretable model for establishing a baseline and testing linear relationships between tourist volume and service quality.
- **Limitations:** Assumes linearity and doesn't perform well for non-linear relationships.

Decision Tree:

- **Why Chosen**: Captures non-linear relationships and works well with both numerical and categorical data. Provides interpretable rules for decision-making.
- **Limitations**: Prone to overfitting, especially when the tree is too deep.

Random Forest:

- **Why Chosen:** An ensemble method that improves upon decision trees by averaging predictions from multiple trees. Handles overfitting and generalizes better, robust and suitable for complex, real-world data with non-linearities.
- **Limitations:** Less interpretable than a single decision tree.

1. **Model Training:**
    - Trained Linear Regression, Decision Tree, and Random Forest models.
2. **Model Evaluation:**

R-squared (R^2):

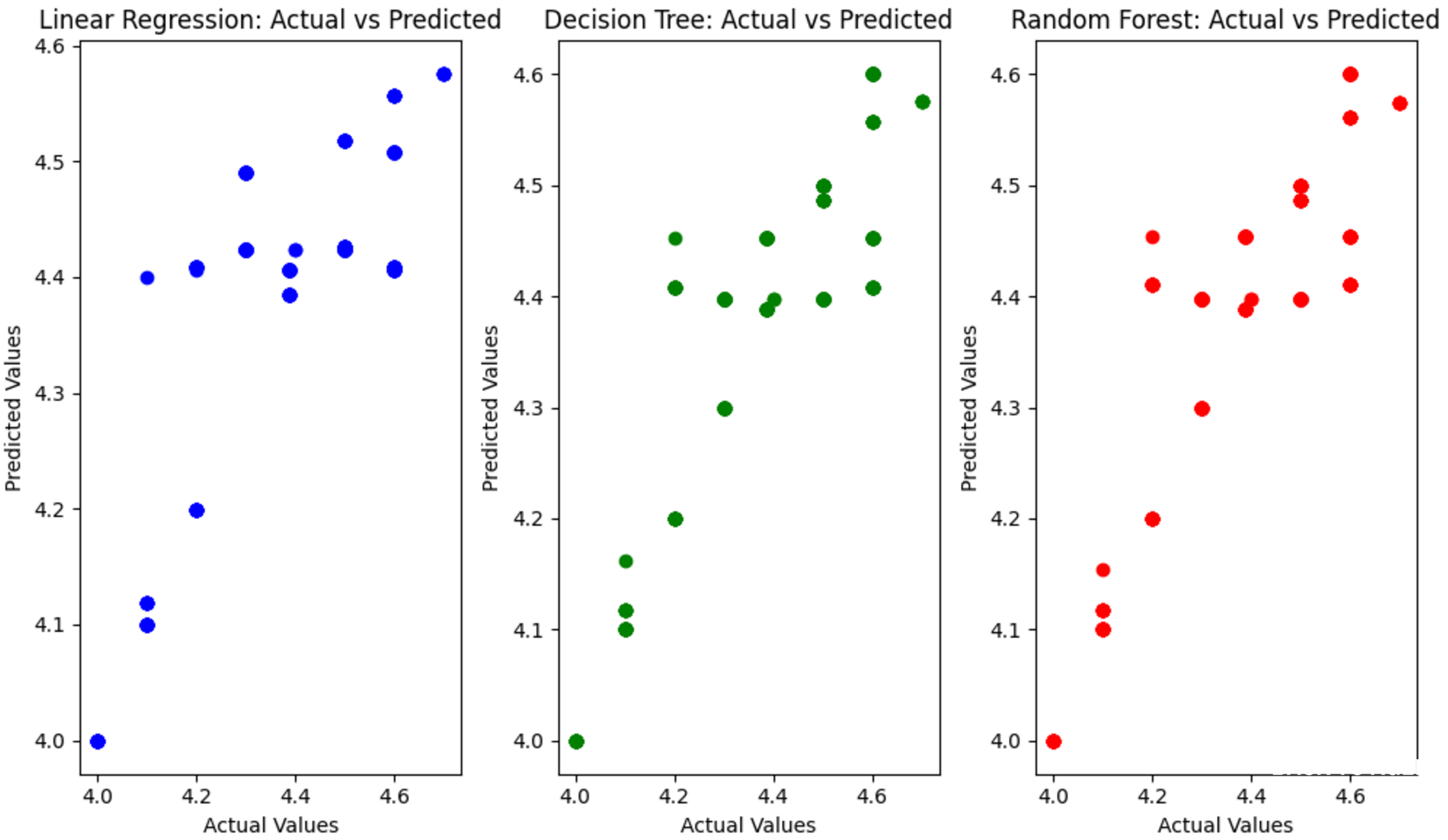
- Measures the proportion of variance explained by the model.
- **Why Chosen**: Provides an understanding of how well the model fits the data. Ranges from 0 (poor fit) to 1 (perfect fit).
- **What it Reveals:** A higher R^2 indicates a model that explains more of the variance in service quality.

2.**Model Evaluation:**

Mean Squared Error (MSE):

- Measures the average squared difference between actual and predicted values.
- **Why Chosen:** MSE penalizes large errors more heavily and gives insight into how far off the predictions are from actual values.
- **What it Reveals:** A lower MSE indicates better prediction accuracy and less error.

	Model	R-squared	MSE
0	Linear Regression	0.54	0.02688
1	Decision Tree	0.71	0.00839
2	Random Forest	0.71	0.00840


Linear Regression: Actual vs Predicted
Decision Tree: Actual vs Predicted
Random Forest: Actual vs Predicted
Predicted Values
Actual Values
4.0
4.1
4.2
4.3
4.4
4.5
4.6

Final decision

while both **Decision Tree** and **Random Forest** are performing similarly in terms of MSE and R^2, **Random Forest is the better model** because it provides better **generalization**, more **reliable predictions**, and greater **robustness**, making it more suited for real-world applications, reduces the risk of overfitting, and offers the advantage of feature importance. Despite the small differences in MSE and R^2, Random Forest is the better choice for long-term deployment and when looking for a model that performs well on unseen data.

Predicting Spending Based on Length of Stay and Key attributes

1. **Model Training:**
    - Trained Linear Regression, and Random Forest models

and Gradient Boosting.

1. **Model Evaluation:**

Mean Absolute Error (MAE):

- MAE quantifies the average absolute differences between the predicted and actual values, providing a straightforward evaluation of the model's prediction accuracy.

- **Why Chosen**: MAE is a reliable metric because it is less influenced by outliers, offering a balanced measure of the model's overall performance.

- **What it Reveals:** A lower MAE means the model produces smaller and more consistent errors, demonstrating its ability to predict accurately.

2. **Model Evaluation:**

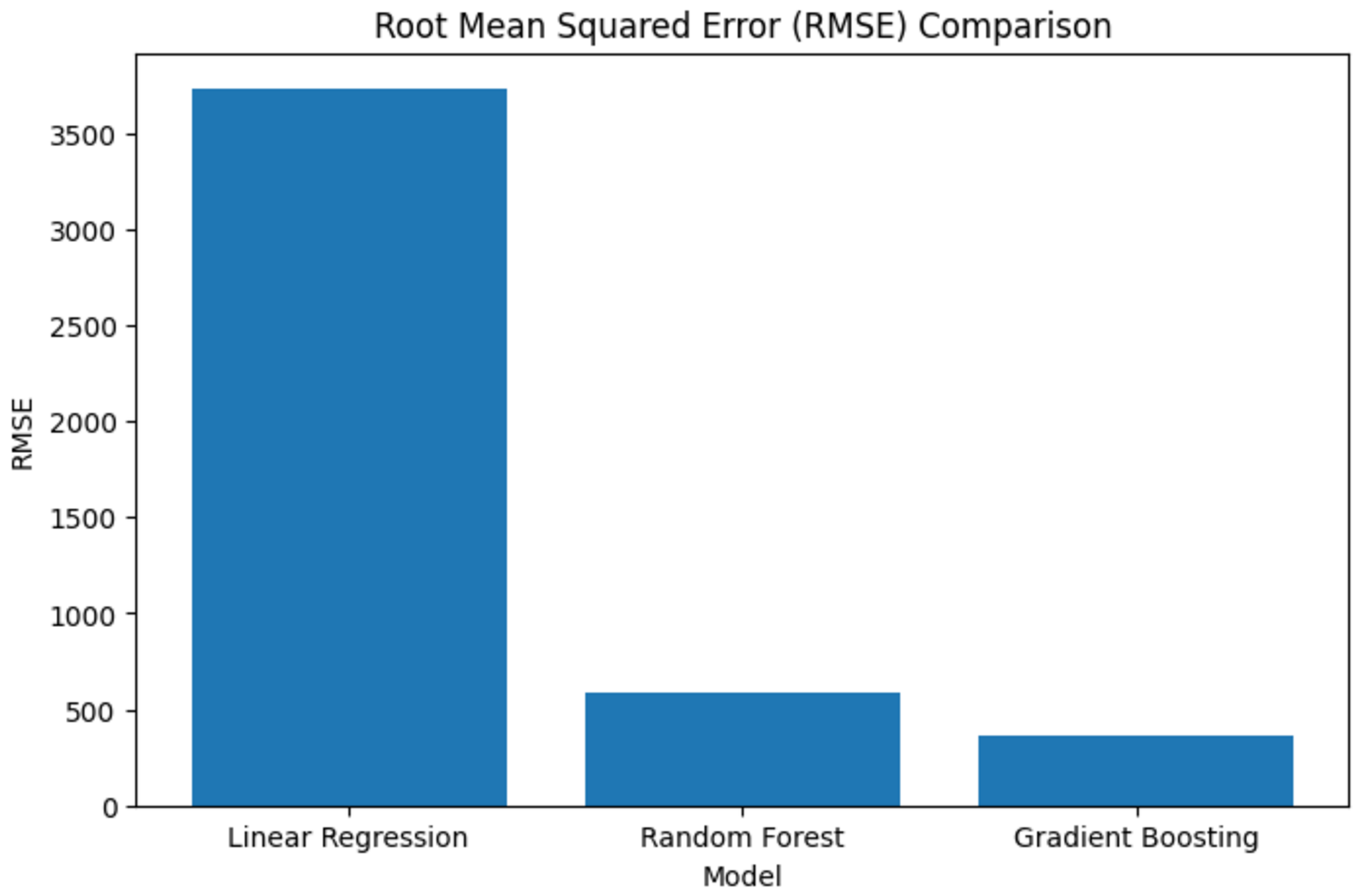
Root Mean Squared Error (RMSE):

- RMSE calculates the square root of the average squared differences between predicted and actual values, emphasizing larger prediction errors more heavily.

- **Why Chosen**: RMSE is highly effective for identifying models that handle large deviations poorly, as it penalizes significant errors more than MAE.

- **What it Reveals:** A smaller RMSE signifies higher prediction accuracy and better alignment between the model's predictions and actual outcomes.

	Model	MAE	RMSE
0	Linear Regression	2742.97	3730.43
1	Random Forest	494.82	582.18
2	Gradient Boosting	322.54	361.07


Root Mean Squared Error (RMSE) Comparison
RMSE
3500
3000
2500
2000
1500
1000
500
0
Linear Regression
Random Forest
Gradient Boosting
Model

Final decision

Gradient Boosting was selected as the most suitable model due to its outstanding accuracy and ability to handle complex data relationships. This model demonstrated the lowest error metrics (MAE and RMSE) among all competitors, making it the most reliable in predicting tourist spending. Its capability to effectively capture non-linear relationships, such as those involving the purpose of visit and length of stay, highlights its suitability for intricate datasets. Furthermore, **Gradient Boosting's** iterative approach, which progressively learns from previous prediction errors, enhances its robustness and adaptability to various data scenarios. These qualities establish Gradient Boosting as the **optimal choice** for accurate and reliable predictive tasks.

technical hurdles

Missing Data:

- Challenge: Handling missing values in the dataset.
- Solution: Applied mean imputation to handle missing numerical data, ensuring a more complete dataset for training.

Computational Complexity:

- Challenge: Training Random Forest models can be computationally expensive, especially with a large dataset.
- Solution: We used a smaller number of estimators initially and could scale it further depending on the data size.

Future Improvements and Expansions

- **Feature Engineering**: Adding new features and improving existing ones.
- **Real-Time Data Updates:** Ensuring the model adapts to new data.
- **Model Interpretability:**Making the model's decisions transparent and understandable.
- **Explore External Data Sources:** Incorporating external data sources can enrich the model and improve its predictive power.
- **Business Integration:** Incorporating the model into real-world decision-making processes.

The Role of Language in Tourist Satisfaction Analyzing Google Maps Reviews from International vs. Local Visitors

The Role of Language in Tourist Satisfaction

- performing sentiment analysis on google maps reviews
- counting the frequency of each language for each star rating.
- analyzing the frequency of various categories across different languages
- Topic Modeling Using Latent Dirichlet Allocation (LDA)

1. **performing sentiment analysis on google maps reviews**

1. performing sentiment analysis on google maps reviews

Detecting the Language of Each Review

we specifically focuses on two languages:

- Arabic
- English

Applying Language-Specific Sentiment Models

Depending on the detected language, a specific sentiment analysis model is applied: SBERT for english reviews and BERT for arabic reviews

Storing Sentiment and Language Data

For each review, the following data is stored:

- Detected Language: Either "Arabic" or "English."
- Sentiment: Either "positive" or "negative."

1. performing sentiment analysis on google maps reviews

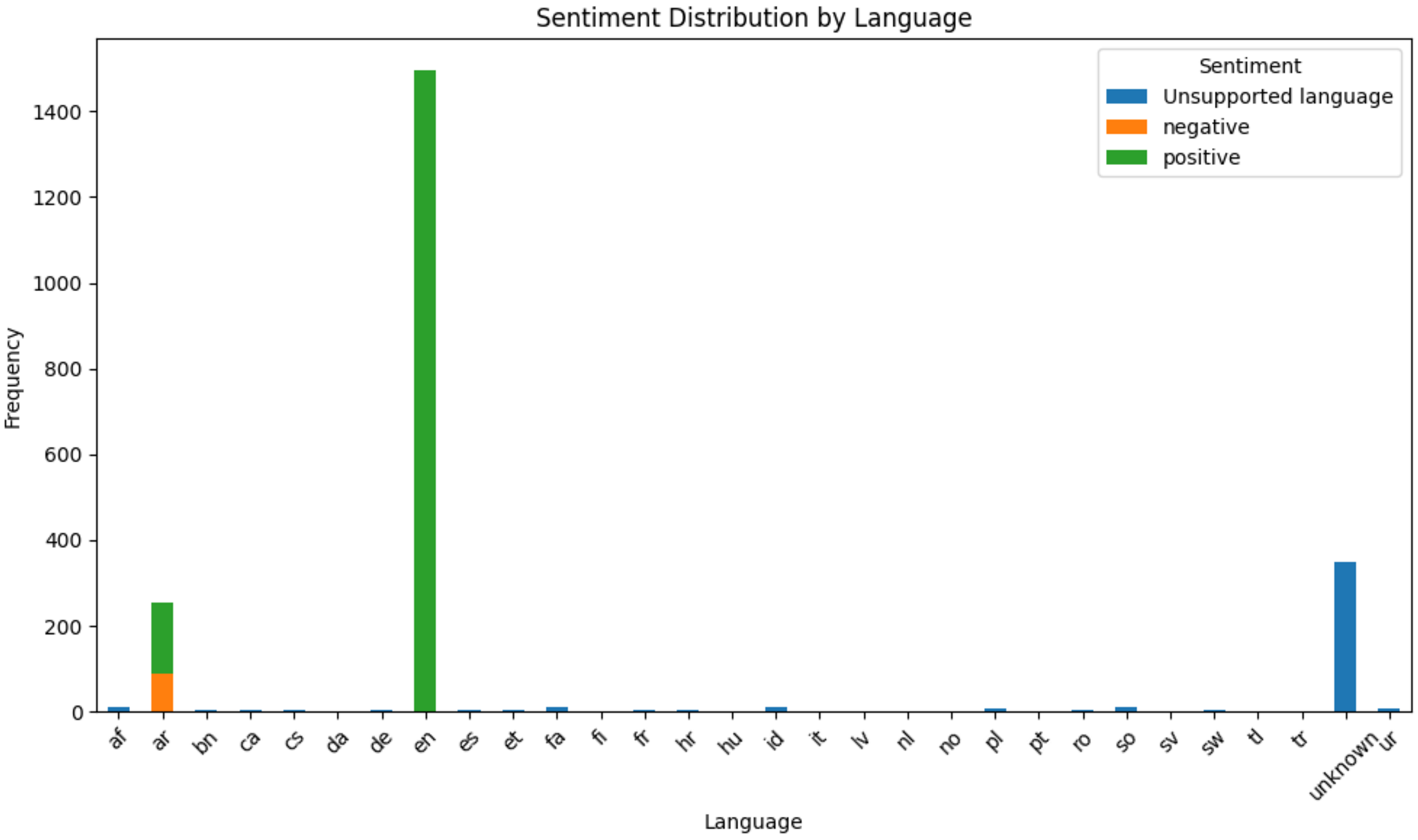
Grouping Sentiment by Language

- The data is grouped by language to calculate the frequency of postive and negative sentiments
- This results in a distribution of sentiments for each language.

Visualizing the Sentiment Distribution

A stacked bar chart is created to display the sentiment analysis results:

- X-axis: Languages ("Arabic" and "English").
- Y-axis: frequency of reviews.
- Stacked Bars: Shows "positive" and "negative" sentiments



Sentiment Distribution by Language
Sentiment
Unsupported language
negative
positive
Frequency
0
200
400
600
800
1000
1200
1400
af
ar
bn
ca
cs
da
de
en
es
et
fa
fi
fr
hr
hu
id
it
lv
nl
no
pl
pt
ro
so
sv
sw
tl
tr
unknown
ur
Language

Findings:

- based on the chart results, the number of positive sentiments in English reviews is higher compared to other languages.

- English-speaking users tend to leave more positive feedback on the platform.

- Arabic reviews, on the other hand, exhibit a combination of both positive and negative sentiments, suggesting a more mixed response from Arabic-speaking users.

- This highlights a contrast in the sentiment trends between the two language groups.

Applicability:

- Customer Feedback Analysis: Helps businesses understand customer sentiment across different languages, enabling targeted responses and improvements.
- Social Media Monitoring: Similar sentiment analysis can be applied to social media reviews, helping brands identify public sentiment and refine their communication strategies.
- Content Moderation: Can be applied to flag inappropriate or harmful reviews or comments.

2. **counting the frequency of each language for each star rating.**

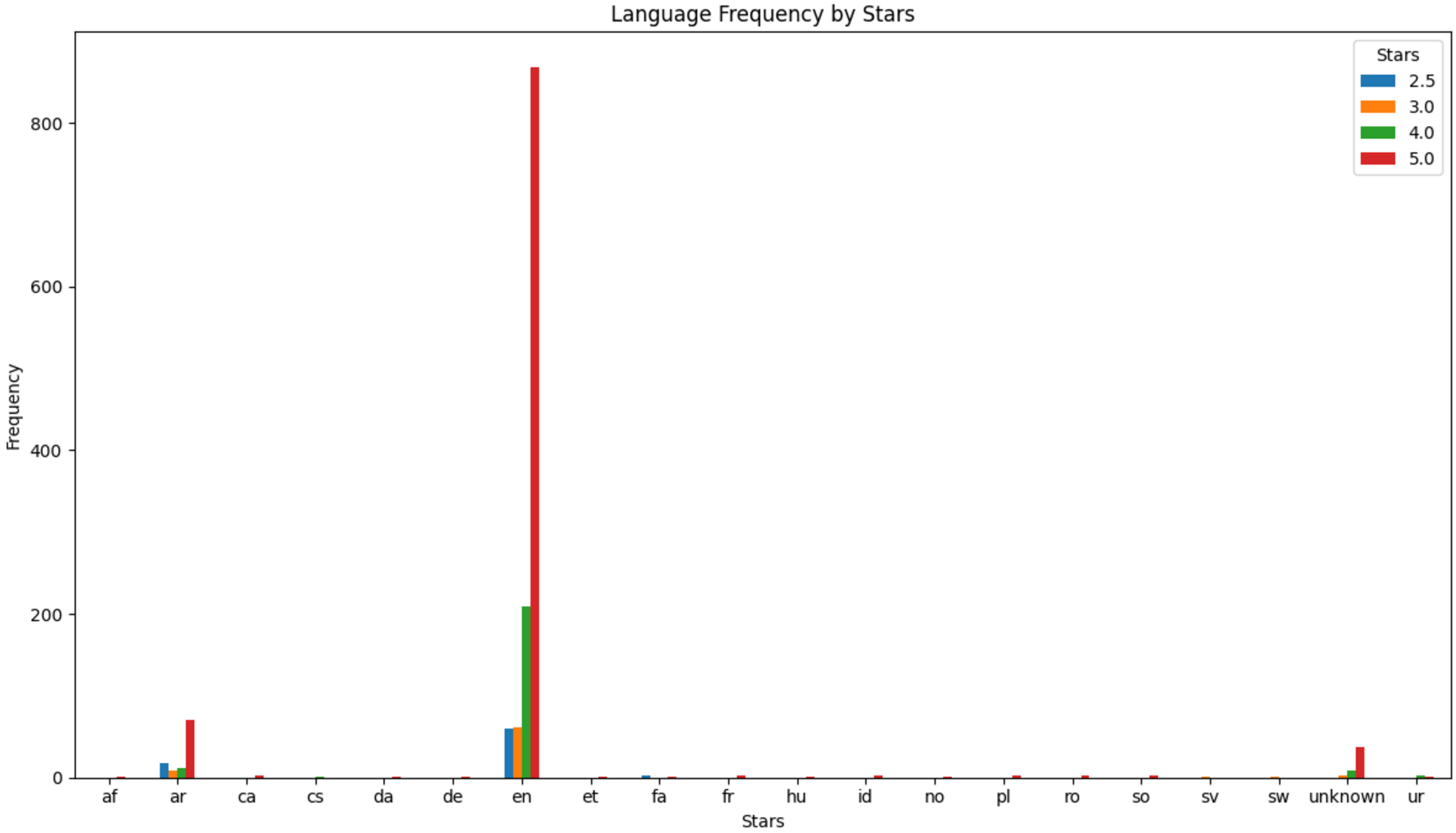
2. **counting the frequency of each language for each star rating.**

Grouping Reviews by Language and Star Ratings

counting the frequency of each language for each star rating

Visualizing the Data by stacked bar chart

- X-axis: Star ratings (1 to 5 stars).
- Y-axis: Frequency of reviews.
- Stacked Bars: Each bar shows the frequency of reviews for each star rating, segmented by language.


Language Frequency by Stars
Stars
2.5
3.0
4.0
5.0
Frequency
800
600
400
200
0
af
ar
ca
cs
da
de
en
et
fa
fr
hu
id
no
pl
ro
so
sv
sw
unknown
ur
Stars

Findings:

- English has the highest frequency of reviews overall, with the majority of reviews receiving a 5-star rating, followed by 4 stars. Ratings of 3.0 and 2.5 are close in frequency.

- Arabic reviews also have a significant presence, with 5 stars being the most common rating, followed by 2.5 stars, then 4 stars, and finally 3 stars.

- Additionally, reviews in "unknown" languages show a similar rating distribution, with most reviews rated 5 stars, followed by 4 stars and 3 stars.

Findings:

- This suggests that while the language could not be detected, the rating trends closely resemble those of both Arabic and English reviews, with a preference for higher ratings.
- This indicates that a larger proportion of the reviews are written in English compared to other languages, suggesting that international visitors might contribute more to the review data.
- The bar chart effectively highlights this trend by showing a higher count of English reviews in the plot.

Applicability:

- Trend Detection: Analyzing the distribution of ratings across languages can reveal trends, such as the preference for higher ratings in certain regions.
- Market Segmentation: Identifies which language groups are more likely to leave high ratings, enabling tailored customer engagement strategies.
- Geographical Insights: By understanding language and rating trends, businesses can determine which geographical regions require more attention or have a higher level of user engagement.

3. **analyzing the frequency of various categories across different languages**

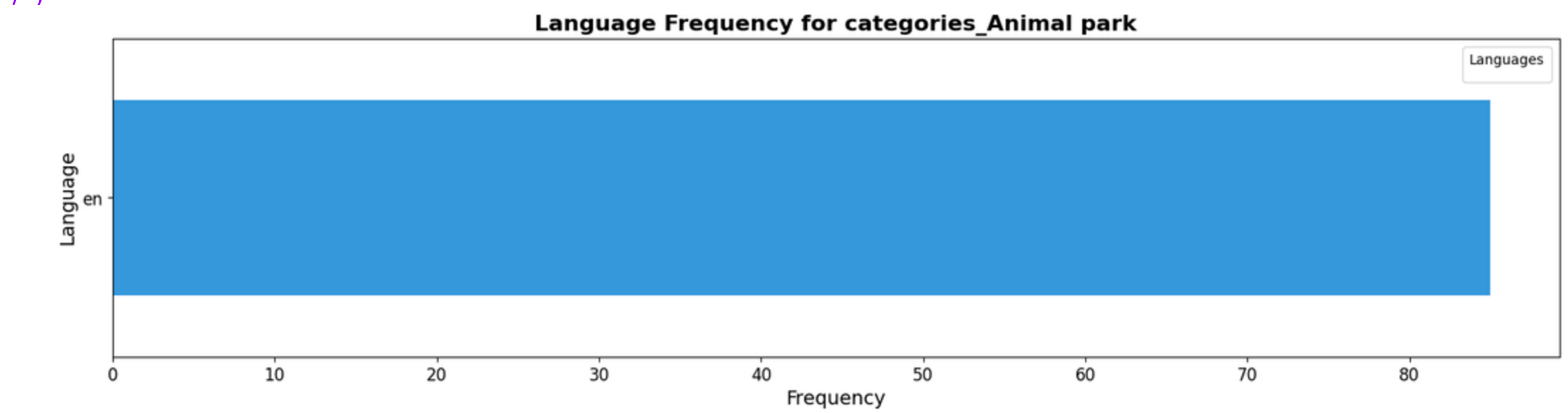
3. analyzing the frequency of various categories across different languages

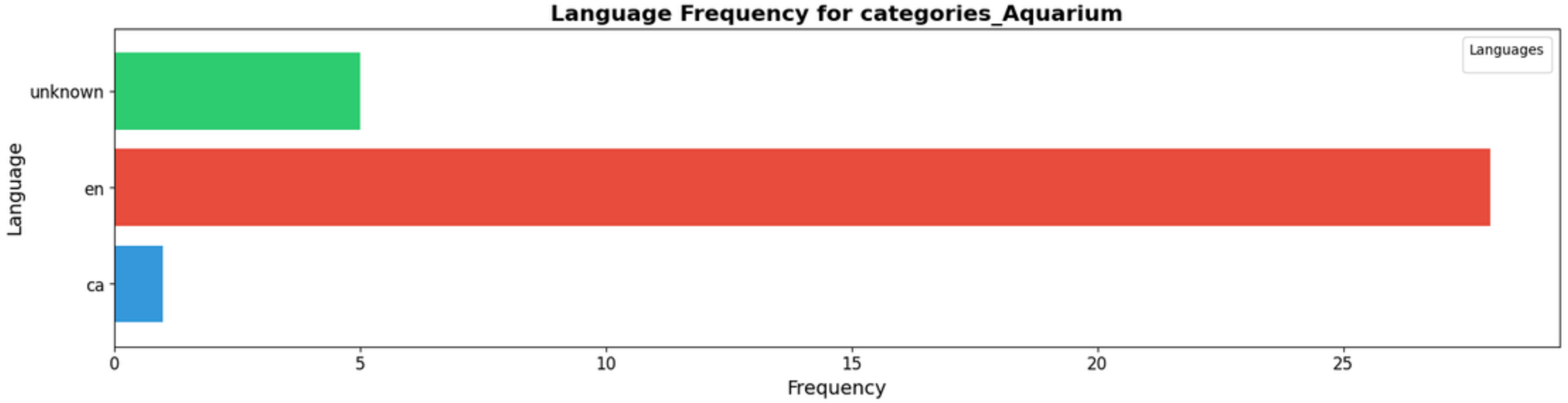
Identifying Category Columns (those starting with "categories_")

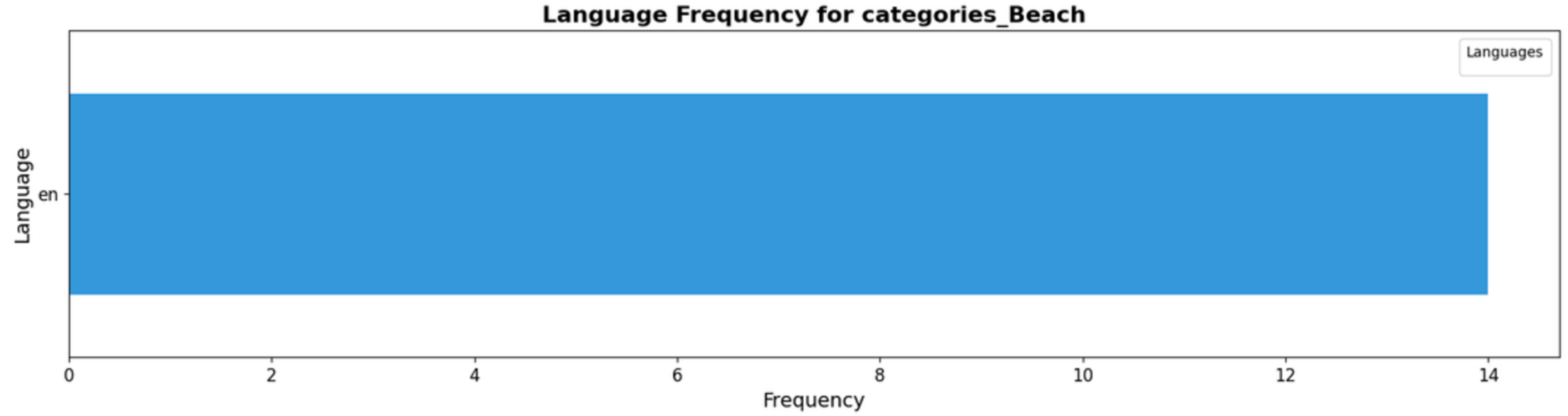
calculating how often each category occurs for each language (such as English, Arabic, etc.)

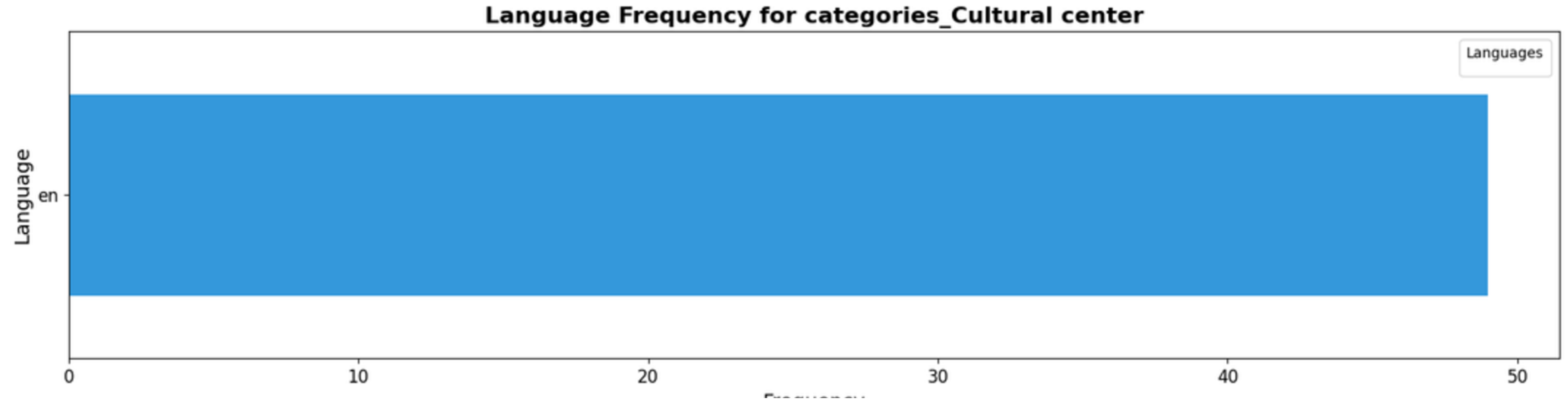
Visualizing the Data by horizontal bar chart

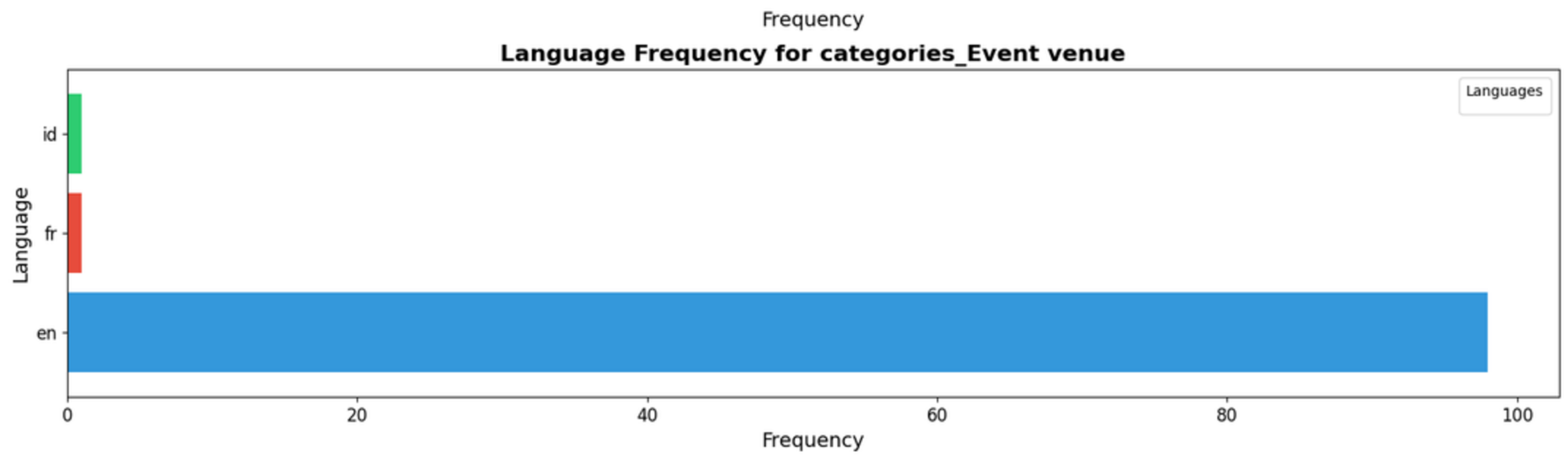
A horizontal bar chart is generated for each category, displaying the frequency of each language's occurrences.

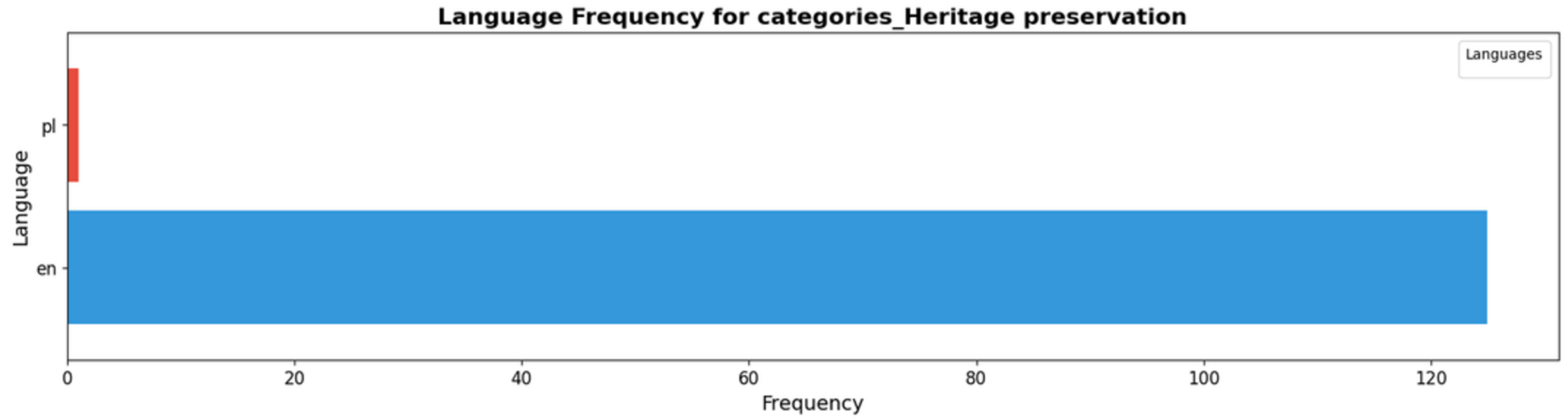

Language Frequency for categories_Animal park
Languages
Language
en
0
10
20
30
40
50
60
70
80
Frequency

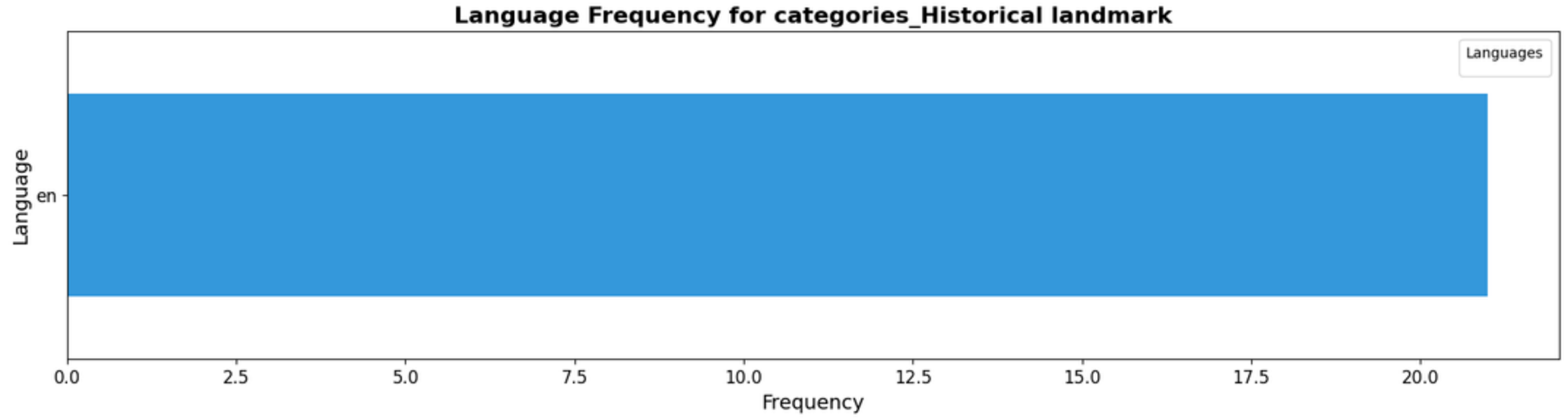

Language Frequency for categories_Aquarium
Languages
Language
unknown
en
ca
0
5
10
15
20
25
Frequency

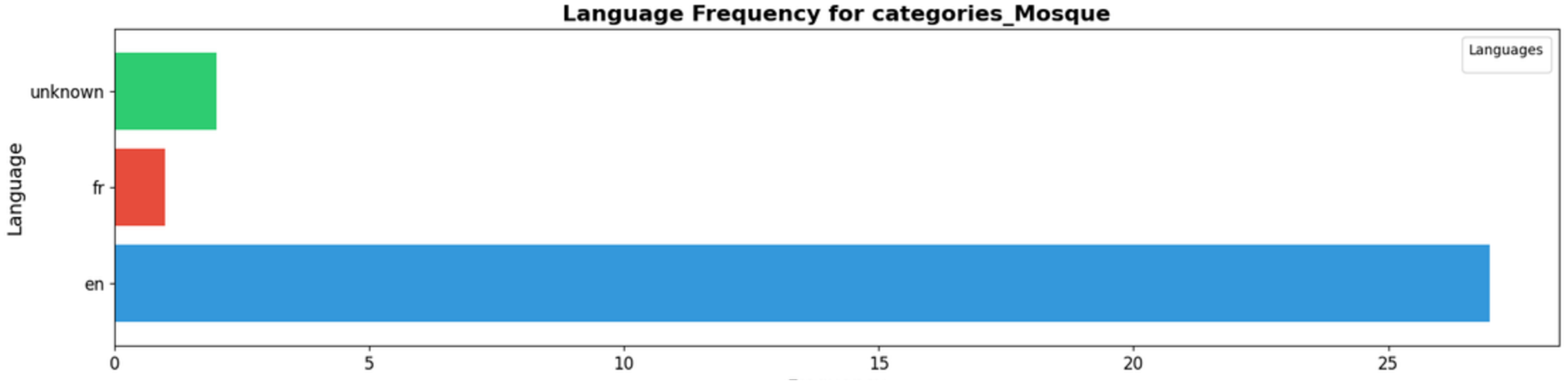

Language Frequency for categories_Beach
Languages
Language
en
0
2
4
6
8
10
12
14
Frequency

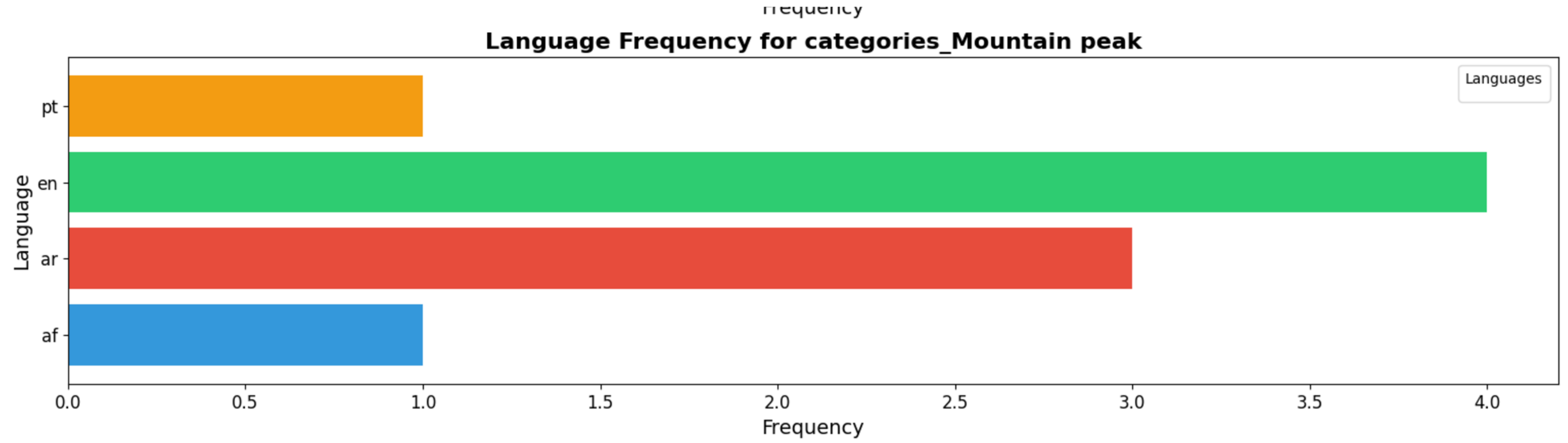

Language Frequency for categories_Cultural center
Languages
Language
en
0
10
20
30
40
50

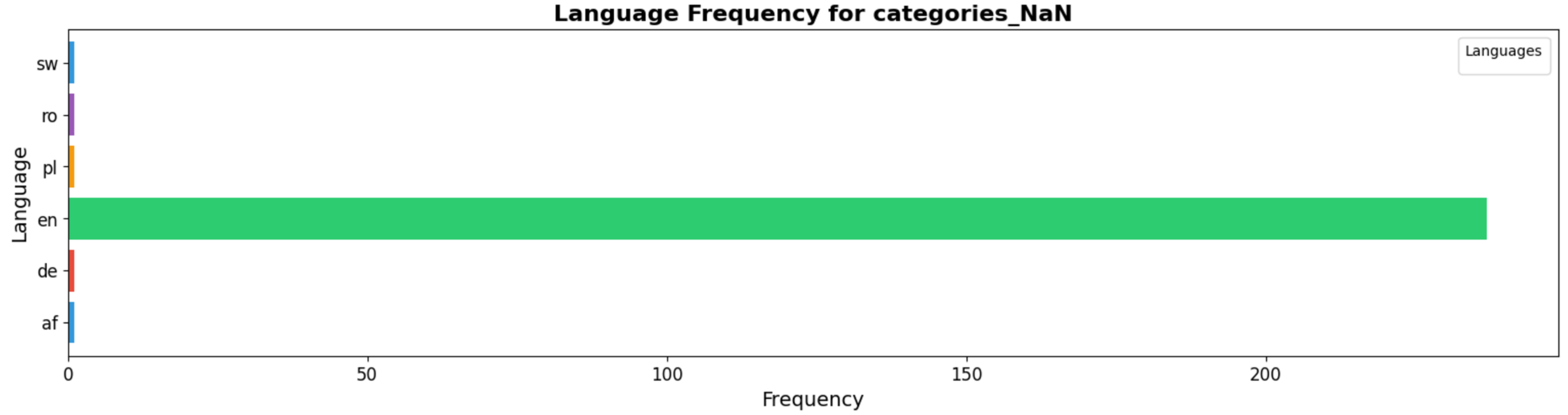

Frequency
Language Frequency for categories_Event venue
Languages
Language
id
fr
en
0
20
40
60
80
100
Frequency

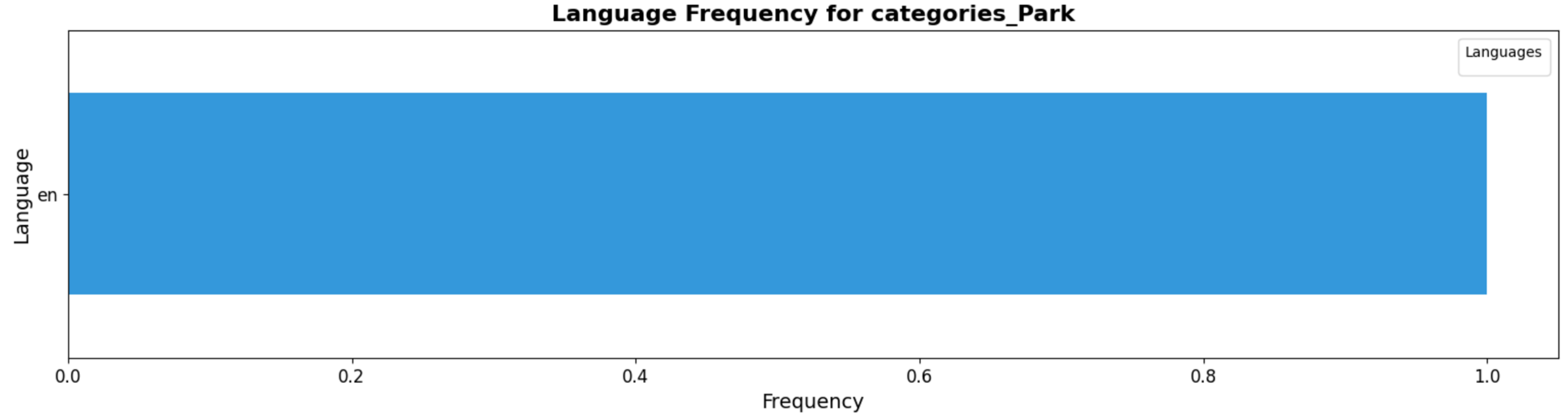

Language Frequency for categories_Heritage preservation
Languages
Language
pl
en
0
20
40
60
80
100
120
Frequency

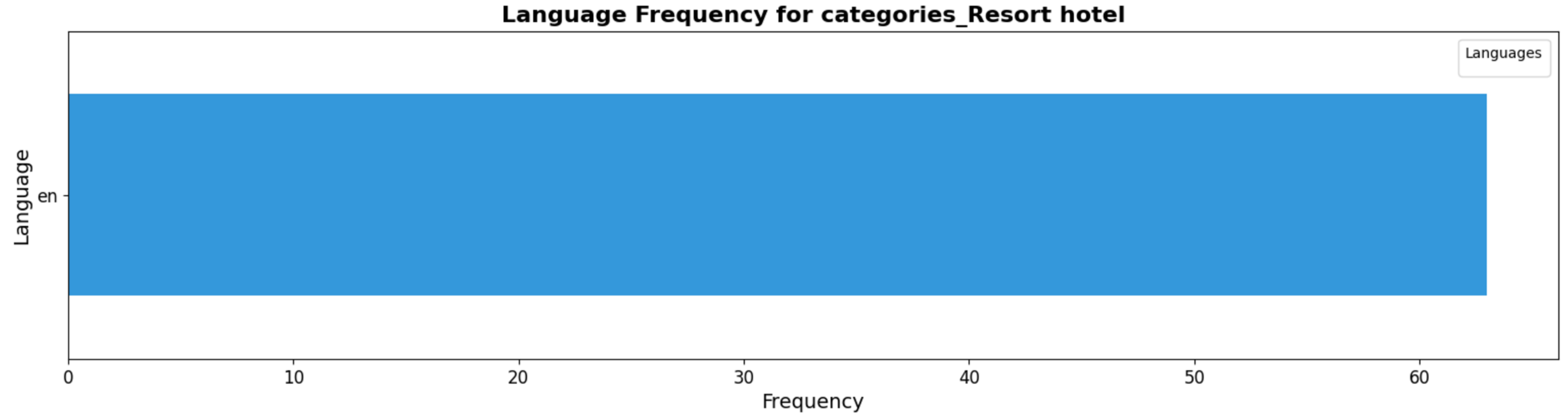

Language Frequency for categories_Historical landmark
Languages
Language
en
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Frequency

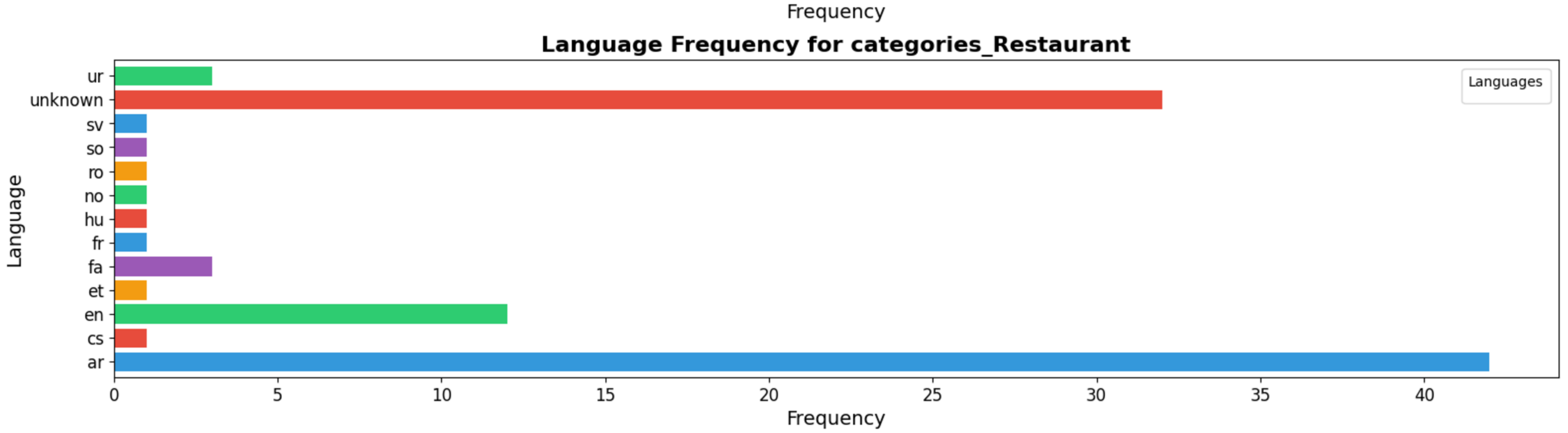

Language Frequency for categories_Mosque
Languages
Language
unknown
fr
en
0
5
10
15
20
25

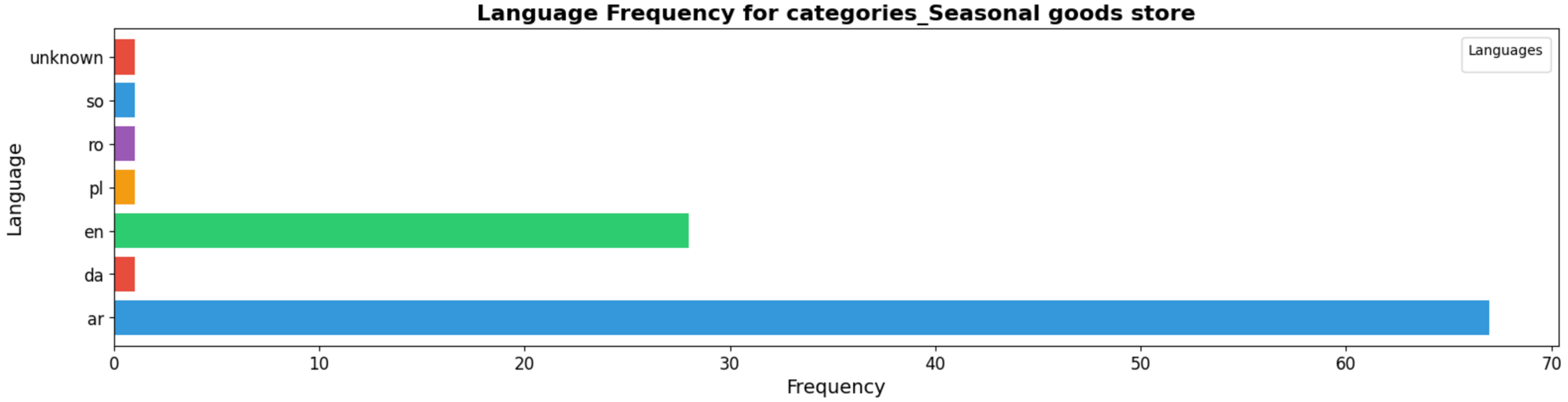

Frequency
Language Frequency for categories_Mountain peak
Languages
Language
pt
en
ar
af
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Frequency

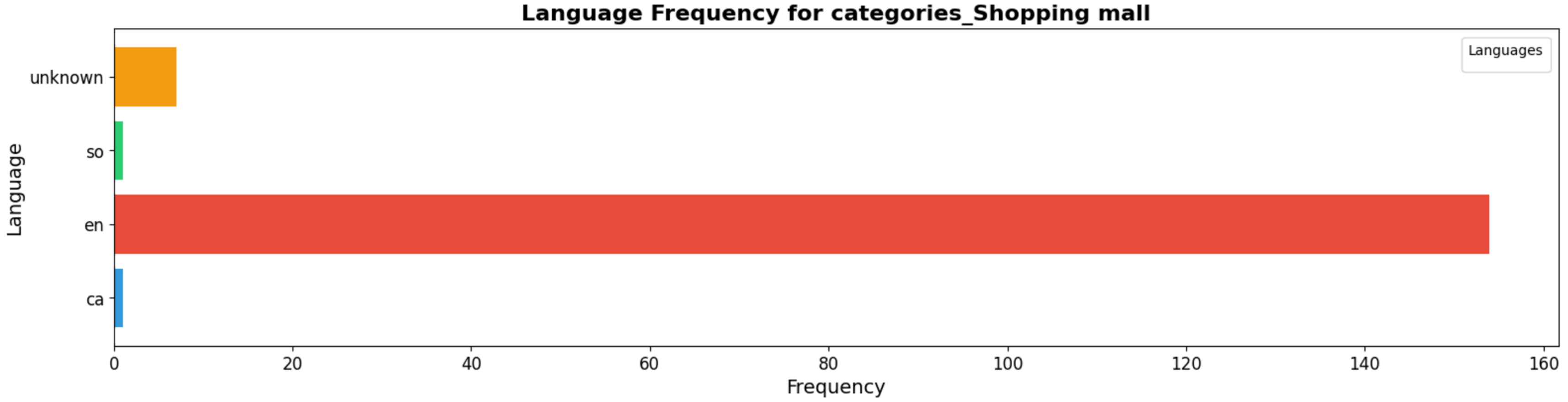

Language Frequency for categories_NaN
Languages
Language
sw
ro
pl
en
de
af
0
50
100
150
200
Frequency

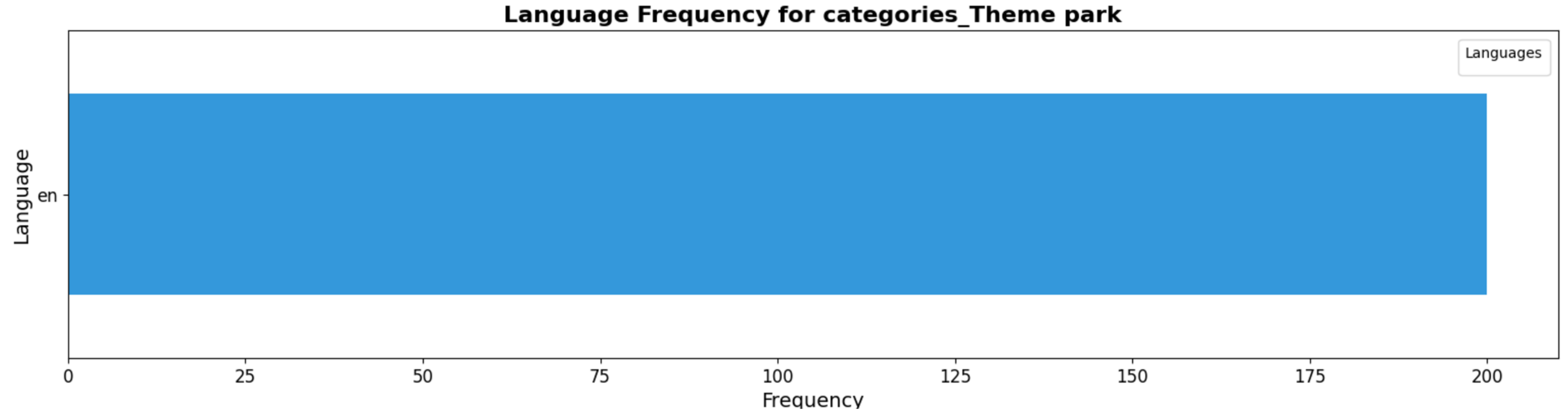

Language Frequency for categories_Park
Languages
Language
en
0.0
0.2
0.4
0.6
0.8
1.0
Frequency

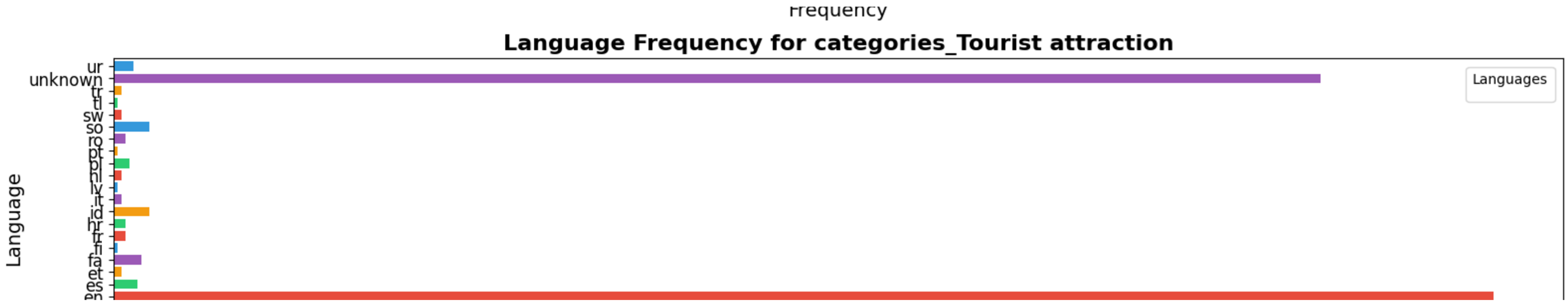

Language Frequency for categories_Resort hotel
Languages
Language
en
0
10
20
30
40
50
60
Frequency


Frequency
Language Frequency for categories_Restaurant
Languages
Language
ur
unknown
sv
so
ro
no
hu
fr
fa
et
en
cs
ar
0
5
10
15
20
25
30
35
40
Frequency


Language Frequency for categories_Seasonal goods store
Languages
Language
unknown
so
ro
pl
en
da
ar
0
10
20
30
40
50
60
70
Frequency


Language Frequency for categories_Shopping mall
Languages
Language
unknown
so
en
ca
0
20
40
60
80
100
120
140
160
Frequency


Language Frequency for categories_Theme park
Languages
Language
en
0
25
50
75
100
125
150
175
200
Frequency


Frequency
Language Frequency for categories_Tourist attraction
Language
Languages
ur
unknown
tr
tl
sw
so
ro
pt
pl
nl
lv
it
id
hr
fr
fi
fa
et
es
en

Findings:

- In Animal Park and Aquarium, English has the highest frequency, with Catalan and unknown languages following.
- Beach, Cultural Center, Historical Landmark, Park, and Resort Hotel categories all show English as the most frequently used language.
- Event Venue sees English as the most common language, followed by French and Indonesian.
- For Heritage Preservation, English is the most frequent, with Polish reviews appearing less often.
- In Mosque, English dominates, followed by unknown languages and French.

Findings:

- Mountain Peak has English as the highest frequency, with Arabic second, and Portuguese and Afrikaans tied.

- NaN and Shopping Mall categories both have English as the most frequent language, followed by unknown languages, with Catalan (ca) and Somali (so) tied for lower frequencies in Shopping Mall.

- In Restaurant, Arabic reviews are most frequent, followed by unknown languages, English, and French and Urdu tied, with lower frequencies in Swedish (sv), Slovak (sk), Romanian (ro), Norwegian (no), Hungarian (hu), Estonian (et), and Czech (cs).

- For Seasonal Goods Store, Arabic leads, followed by English, with Slovak (sk), Romanian (ro), Polish (pl), and Danish (da) appearing the least.

- In Theme Park and Tourist Attraction, English again has the highest frequency. For Tourist Attraction, unknown languages and Arabic complete the top three.

Findings:

- Overall, English is the predominant language across most categories, with Arabic showing strong representation in specific categories such as Restaurant and Seasonal Goods Store.

- This distribution highlights language patterns, showing that English and Arabic are often preferred languages for reviews across a range of venues, while certain categories also see moderate representation from other languages.

Applicability:

- Optimizing User Experience: By analyzing language preferences in various categories (e.g., English for Theme Parks, Arabic for Seasonal Goods Store), businesses can enhance user experience by localizing their platforms, websites, or services in the most commonly spoken languages for each category.

- Language-Specific Analysis: Analyze how language distribution varies across different categories to gain insights into user demographics and preferences.

- Recommendation Systems: Personalize recommendations by integrating language preferences for specific categories (e.g., recommending restaurants in Arabic for Arabic-speaking users).

4. **Topic Modeling Using Latent Dirichlet Allocation (LDA)**

4. Topic Modeling Using Latent Dirichlet Allocation (LDA)

Preprocessing the Text

- Text data is prepared using TF-IDF vectorization to assign weights to words, focusing on unigrams (single words) and bigrams (two-word combinations)

Applying LDA for Topic Extraction

- Latent Dirichlet Allocation (LDA) is applied to the vectorized text to identify 15 topics.
- Each topic is described by its top words, which are used to create descriptive labels that summarize the topics.

Assigning Topics to Documents

- Each document is assigned the topic with the highest probability, and two new columns are added to the dataset:
- Topic Assignment: The ID of the assigned topic.
- Topic Label: A descriptive label derived from the top words of the topic.

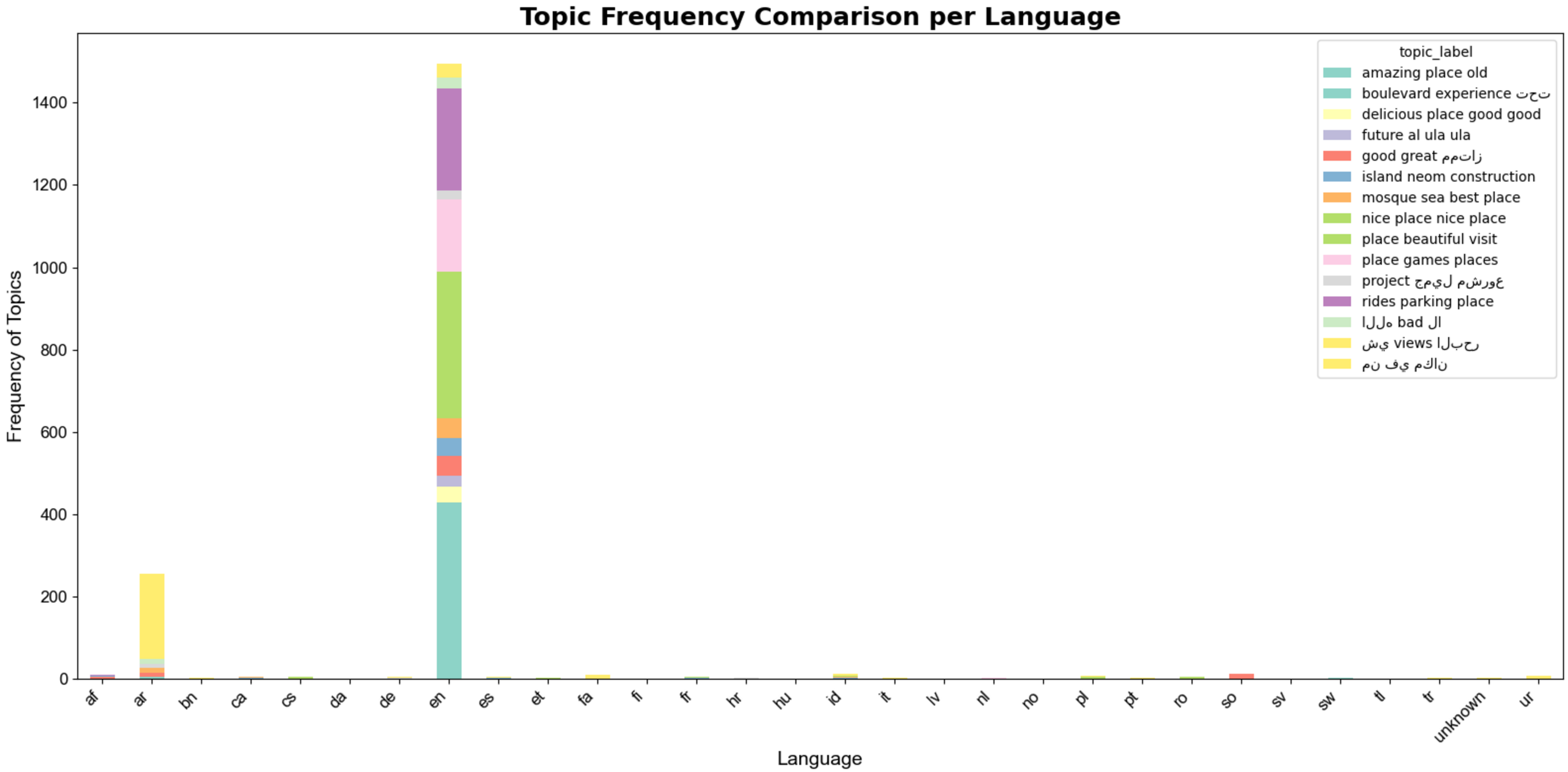
4. Topic Modeling Using Latent Dirichlet Allocation (LDA)

Calculating Topic Frequencies by Language

- The dataset is grouped by language, and the frequency of each topic is calculated for every language in the dataset.

Visualizing Topic Distribution

- A stacked bar chart is created to show how the topics are distributed across different languages.
- Special rendering is applied to handle Arabic characters properly for correct display in the visualization.



Note: That Matplotlib doen't suppot Arabic font

Matplotlib not suppoting Arabic font

- A technical hurdle encountered was Matplotlib's lack of native support for Arabic fonts, which affects the correct rendering of Arabic characters in visualizations. Even though we chose one that supports arabic

- for readablility you can see the topic bellow:
  ['place beautiful visit' 'rides parking place' 'amazing place old'
  'delicious place good good' 'place games places' 'nice place nice place'
  'boulevard experience اللّٰه' 'من في مكان' 'تحت bad لا'
  'mosque sea best place' 'island neom construction' 'good great ممتاز'
  ' شي views البحر' 'project جميل مشروع' 'future al ula ula']

Findings:

- The highest frequency topics in English are "nice place," "boulevard experience," "rides parking place," and "place games places."
- These topics highlight popular aspects of reviews, focusing on the overall appeal of the location, the experience along a boulevard, parking availability for rides, and the presence of game areas, such as amusement parks or entertainment zones.
- In arabic, the highest frequency topic is "من في مكان"

Applicability:

- User Preference Analysis: Identifies popular review aspects, enabling businesses to focus on customer priorities (e.g., parking or rides).
- Content Personalization: Provides insights to tailor recommendations based on user interests (e.g., amusement parks).
- Sentiment and Experience Correlation: Links frequent topics to customer sentiment, aiding in service improvement.

Cultural Heritage and Modern Attractions Comparative Analysis of Tourist Reviews in Saudi Arabia.

1. **Define cultural and modern attraction categories**

Cultural Heritage Columns

Cultural center

Heritage preservation

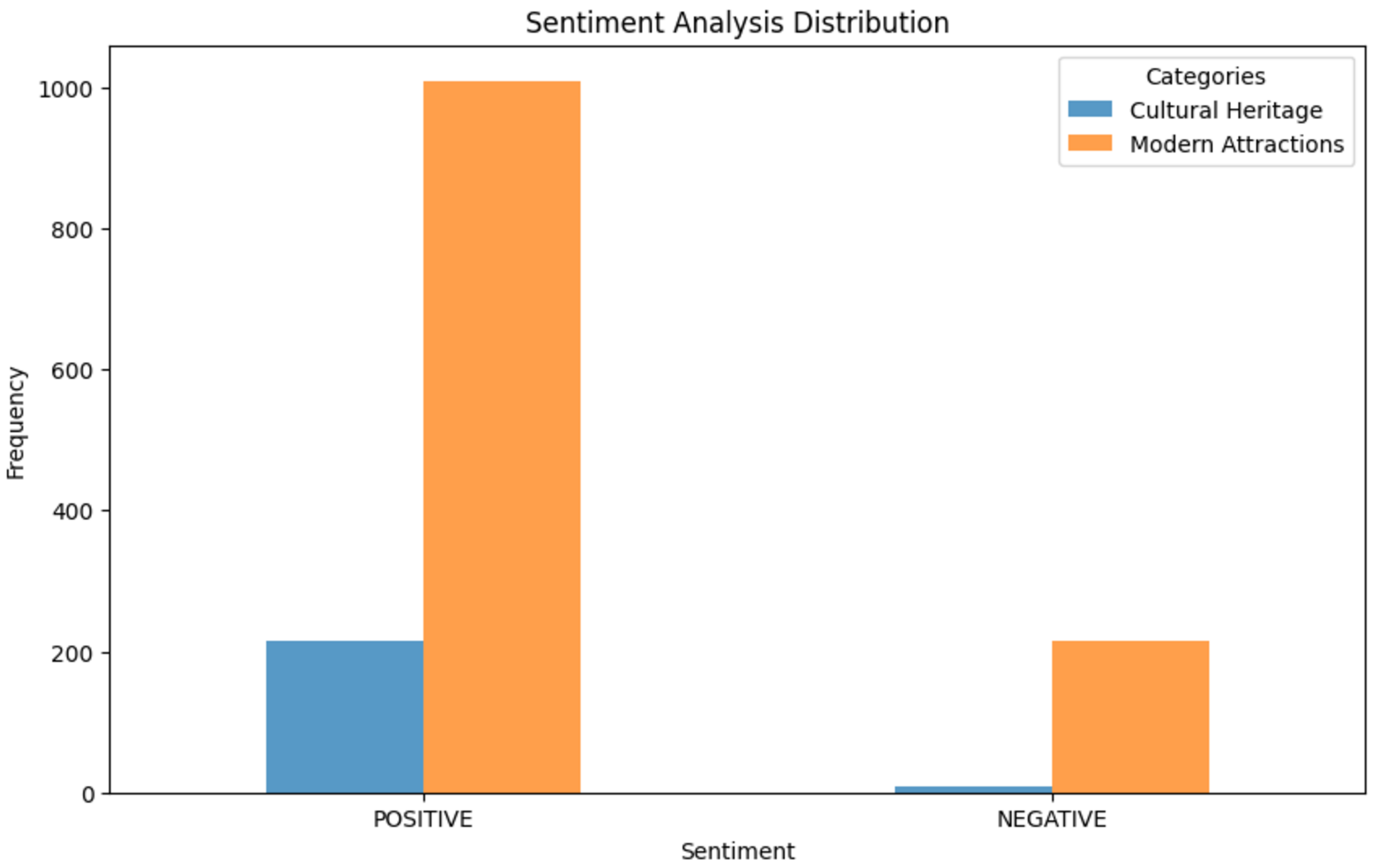
Historical landmark

Mosque

Modern Attraction Columns

- Shopping mall
- Theme park
- Tourist attraction
- Restaurants

2.Sentiment Analysis for review.


Sentiment Analysis Distribution
Categories
Cultural Heritage
Modern Attractions
Frequency
1000
800
600
400
200
0
POSITIVE
NEGATIVE
Sentiment

Findings:

The sentiment distribution for Cultural Heritage and Modern Attractions reveals distinct patterns:

- Modern Attractions:
  - Dominated by POSITIVE reviews (~1000).
  - A noticeable number of NEGATIVE reviews, higher than Cultural Heritage.
- Cultural Heritage:
  - Fewer POSITIVE reviews than Modern Attractions.
  - Almost no NEGATIVE reviews, indicating higher positivity or fewer total reviews.

Overall:

- Modern Attractions garnered more reviews, skewed positively but with greater variability.
- Cultural Heritage shows consistent positivity and minimal negativity.

3. **Negative Review Analysis**

Price Concerns

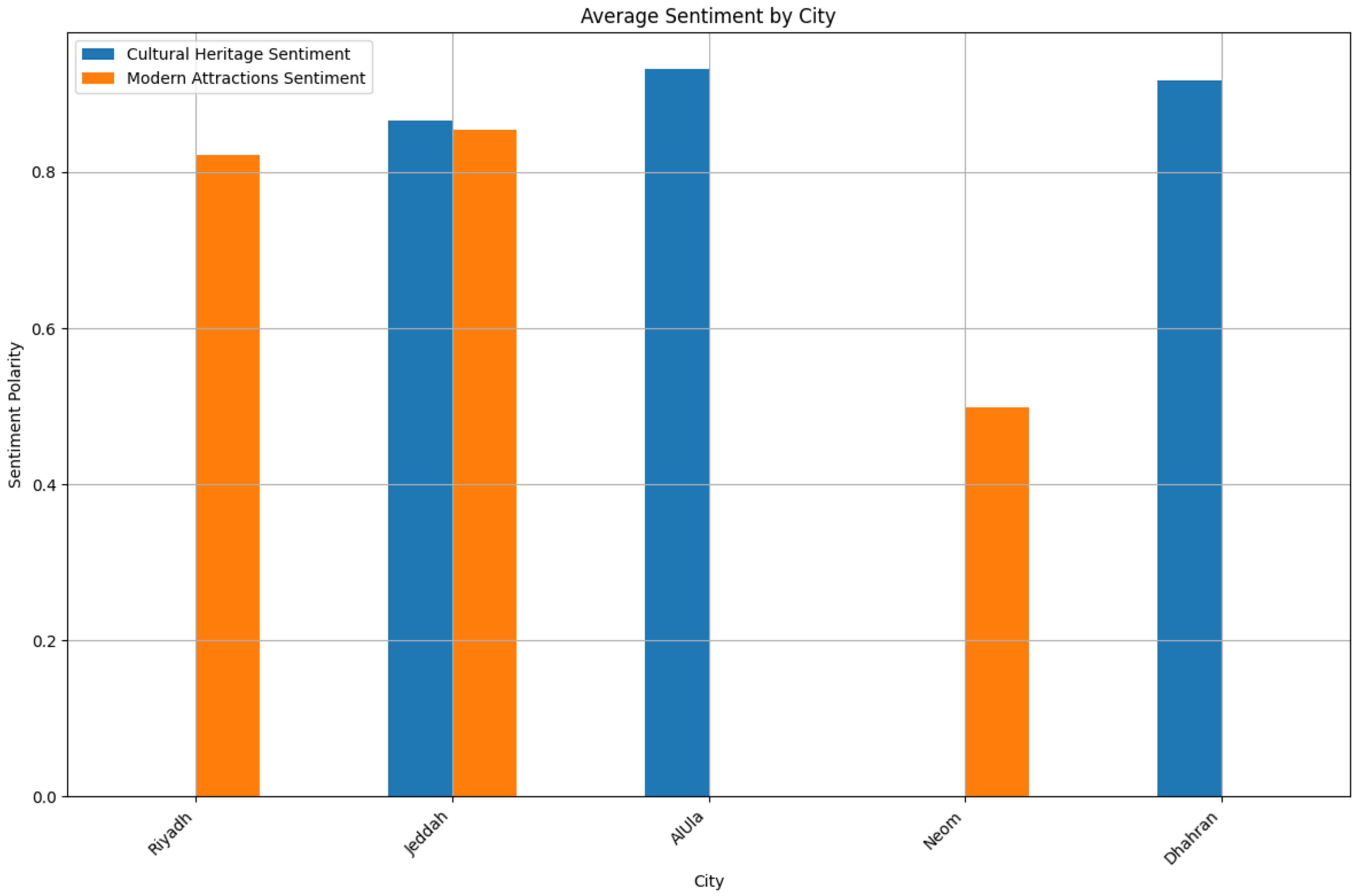
Operational Issues

Accessibility Challenges

Findings:

- **Modern Attractions:**
- Negative reviews highlight overpricing, navigation challenges, and poor management practices.
- Example: "50 SAR entrance fee...not worth it."
- Despite the appealing ambiance, issues with value-for-money and service quality diminish visitor satisfaction.

3. **Negative Review Analysis**



Average Sentiment by City
Cultural Heritage Sentiment
Modern Attractions Sentiment
Sentiment Polarity
0.0
0.2
0.4
0.6
0.8
Riyadh
Jeddah
AlUla
Neom
Dhahran
City

Findings:

- **Riyadh:**

Modern Attractions outperform Cultural Heritage in sentiment, indicating tourists are more satisfied with modern sites.

- **Jeddah:**

Cultural Heritage receives the highest sentiment score across all cities, reflecting strong tourist appreciation. Modern Attractions perform well but slightly trail Cultural Heritage.

- **AlUla:**

Cultural Heritage is the sole category reviewed, earning a high sentiment score, showcasing tourist satisfaction with its historical significance.

Findings:

- **Neom:**

Modern Attractions are the only reviewed category, receiving lower sentiment scores compared to other cities, suggesting mixed feedback or limited reviews.

- **Dhahran:**

Cultural Heritage alone is reviewed and scores well, indicating tourist satisfaction with its cultural offerings.

- entrance fee...not worth it."
- Despite the appealing ambiance, issues with value-for-money and service quality diminish visitor satisfaction.