Predicting honeymoon destinations using Reddit Datasets

Web API and NLP Project

Abirami Rajamanickam

Problem Statement

- Developed a classification model to categorize Reddit posts for couples seeking unique honeymoons

- Collected two subreddits from 'travel' and 'travel hacks' categories, and applied sentiment analysis to identify top-rated destinations within the 'travel' subreddit.

Data Collection

Data Collection: Collected subreddit posts from Reddit website using Web API and NLP

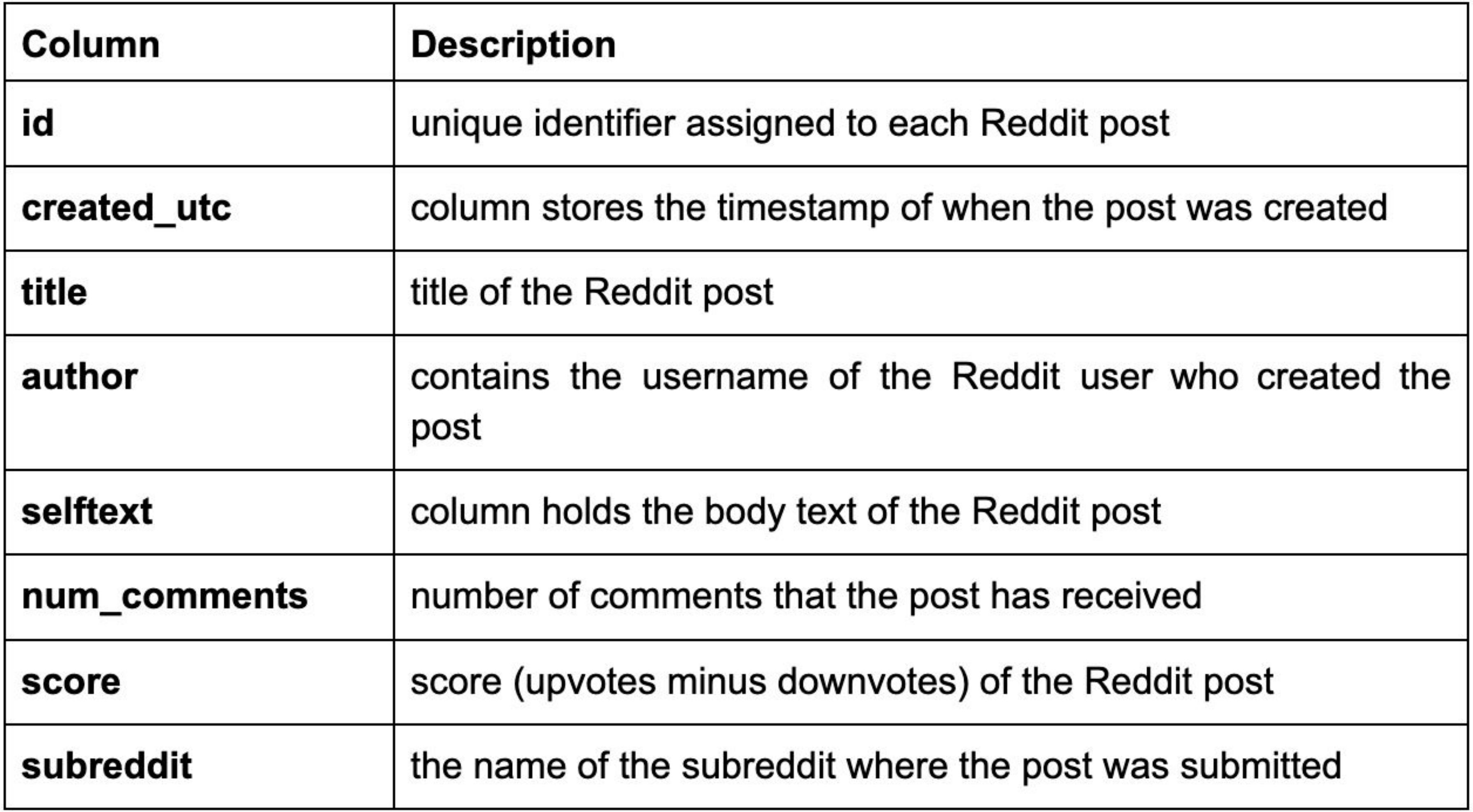
Post types: Collected post types like new,top and hot to gather wider data.

Source: Source Reddit website using PRAW Web API

Travel subreddit is a community on Reddit where users discuss topics related to travel.

Travel Hacks is a community on Reddit sharing tips, tricks, and strategies to make travel more affordable, efficient, and enjoyable.

Data Science Steps: Data Collection, Data Cleaning, Data Preprocessing, Data Modeling and Evaluation

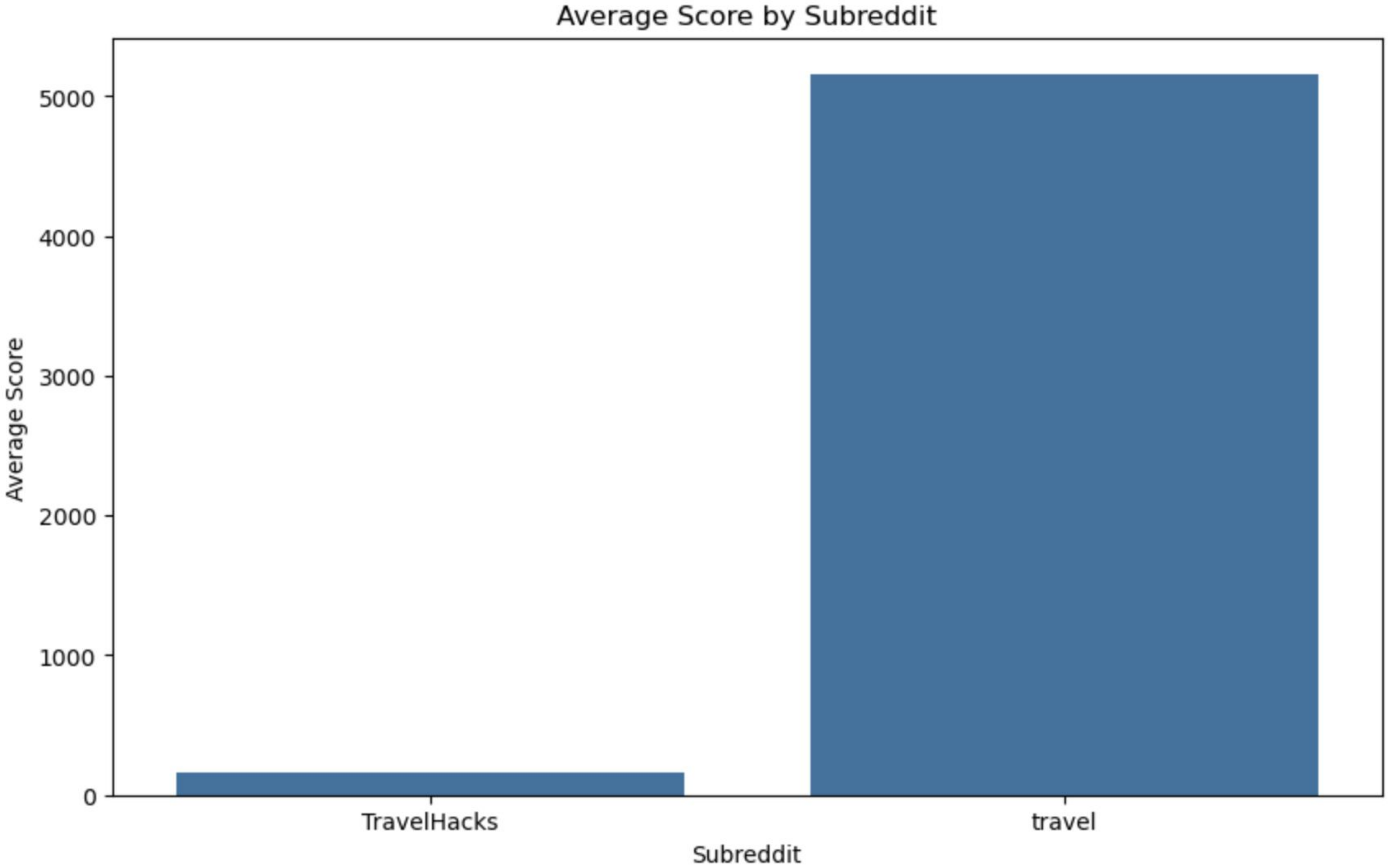


Column	Description
id	unique identifier assigned to each Reddit post
created_utc	column stores the timestamp of when the post was created
title	title of the Reddit post
author	contains the username of the Reddit user who created the post
selftext	column holds the body text of the Reddit post
num_comments	number of comments that the post has received
score	score (upvotes minus downvotes) of the Reddit post
subreddit	the name of the subreddit where the post was submitted

Exploratory Data Analysis

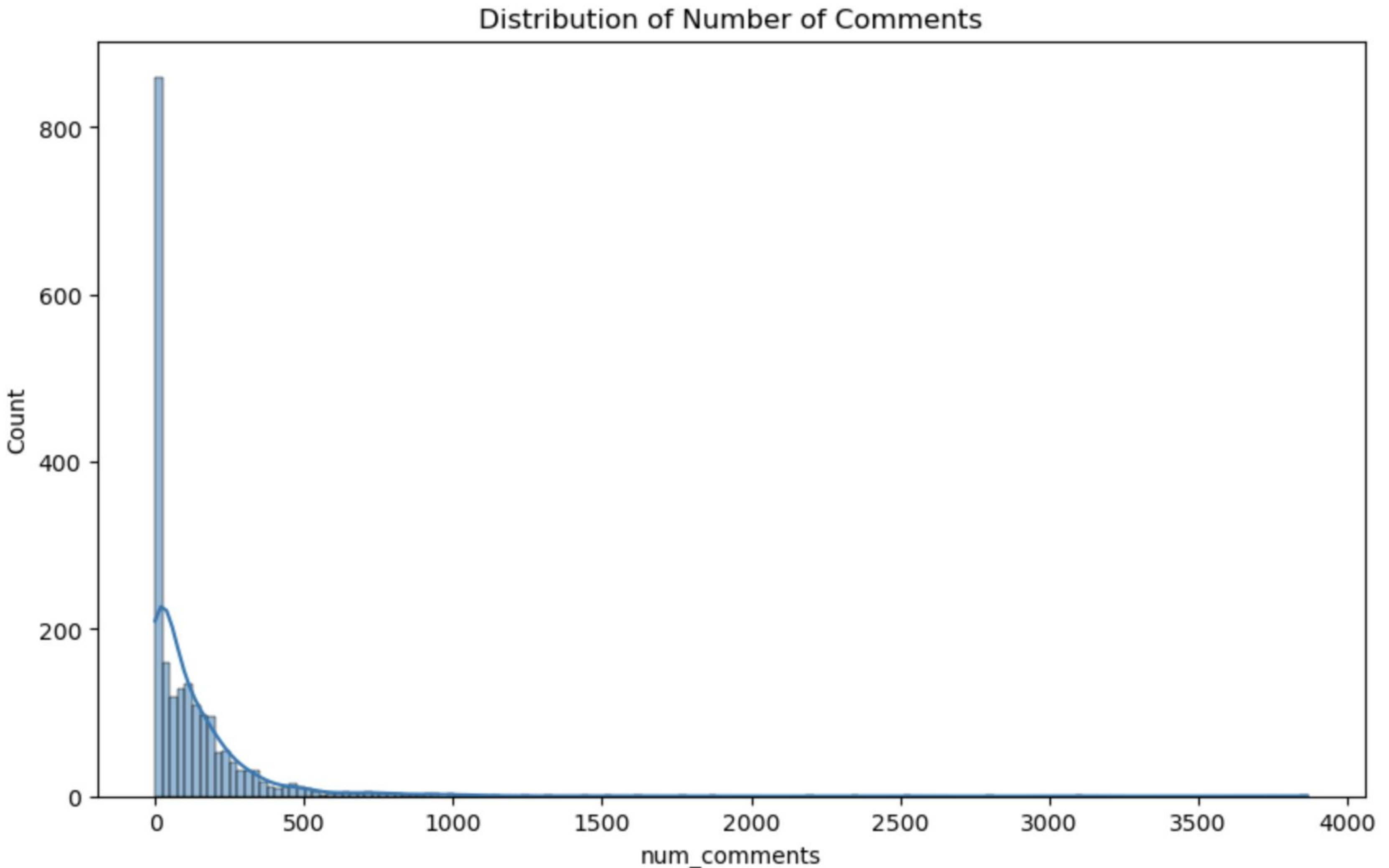
Performed Exploratory Data Analysis separately to identify statistical analysis of both travel and travel hacks subreddits

- Displayed sample data using df.head() method
- Performed summary statistics using df.describe
- Visualize the Distribution of features
- Explored correlation between features
- Text Data Exploration
- Tokenization, Lemmitizing
- Bigrams and Trigrams analysis
- Checks for balance of class/target feature
- Explored score and number of posts


Average Score by Subreddit
Average Score
5000
4000
3000
2000
1000
0
TravelHacks
travel
Subreddit

Preprocessing

- **Stop Words Removal**: Removed unnecessary words.
- **Tokenization**: Converted text into a processable format.
- **Baseline Model**: Implemented Logistic Regression with TF-IDF for initial predictions.
- **TF-IDF Vectorization:** Used TfidfVectorizer(stop_words='english', ngram_range=(1, 1)) to convert text to numerical vectors, capturing word importance.



Distribution of Number of Comments
Count
0
200
400
600
800
num_comments
0
500
1000
1500
2000
2500
3000
3500
4000

Sentimental Analysis

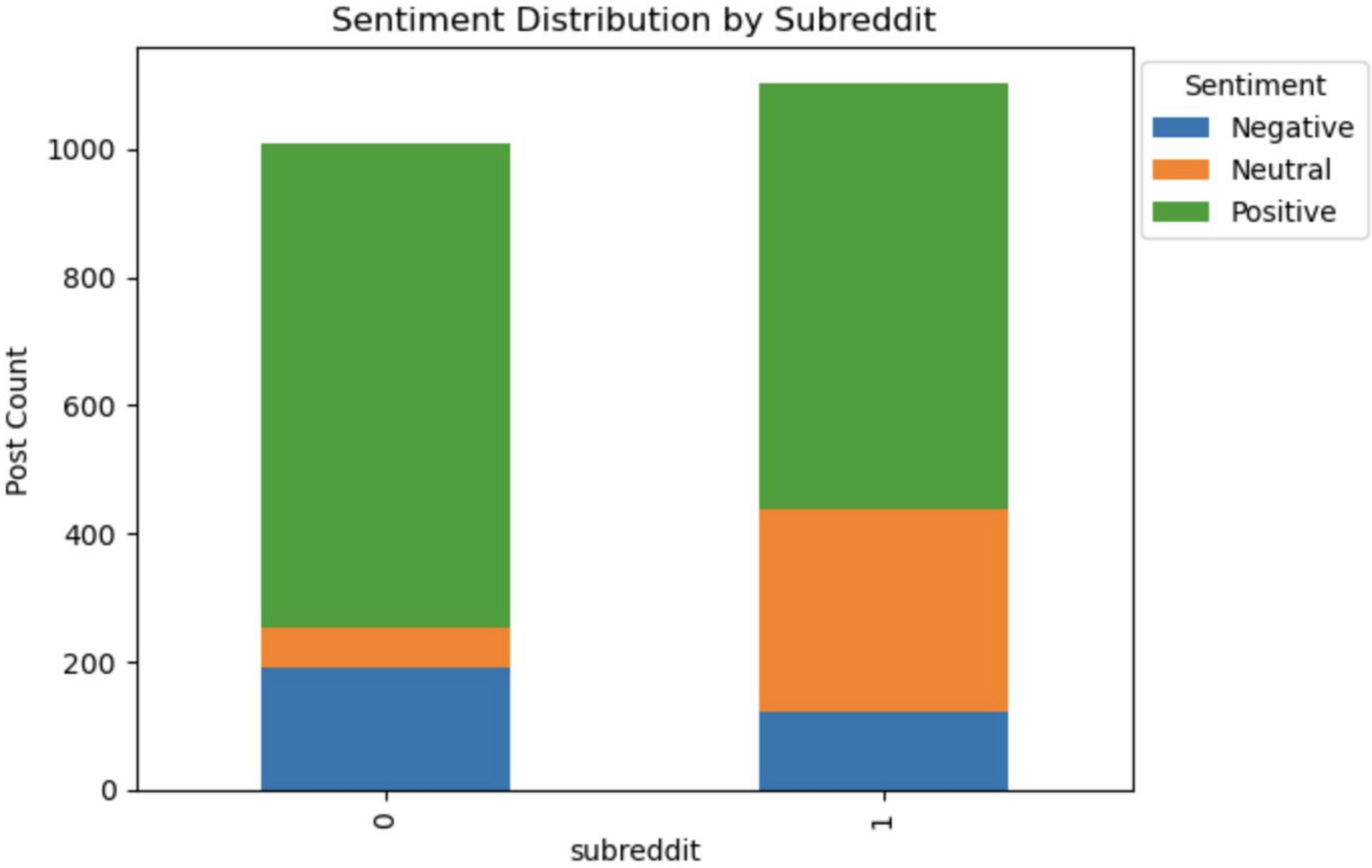
Sentiment Analysis: Applied VADER to score text sentiment.

Honeymoon Pattern Identification: Created a new column, is_honeymoon, mapped to 1 if the text contains "honeymoon," else 0.

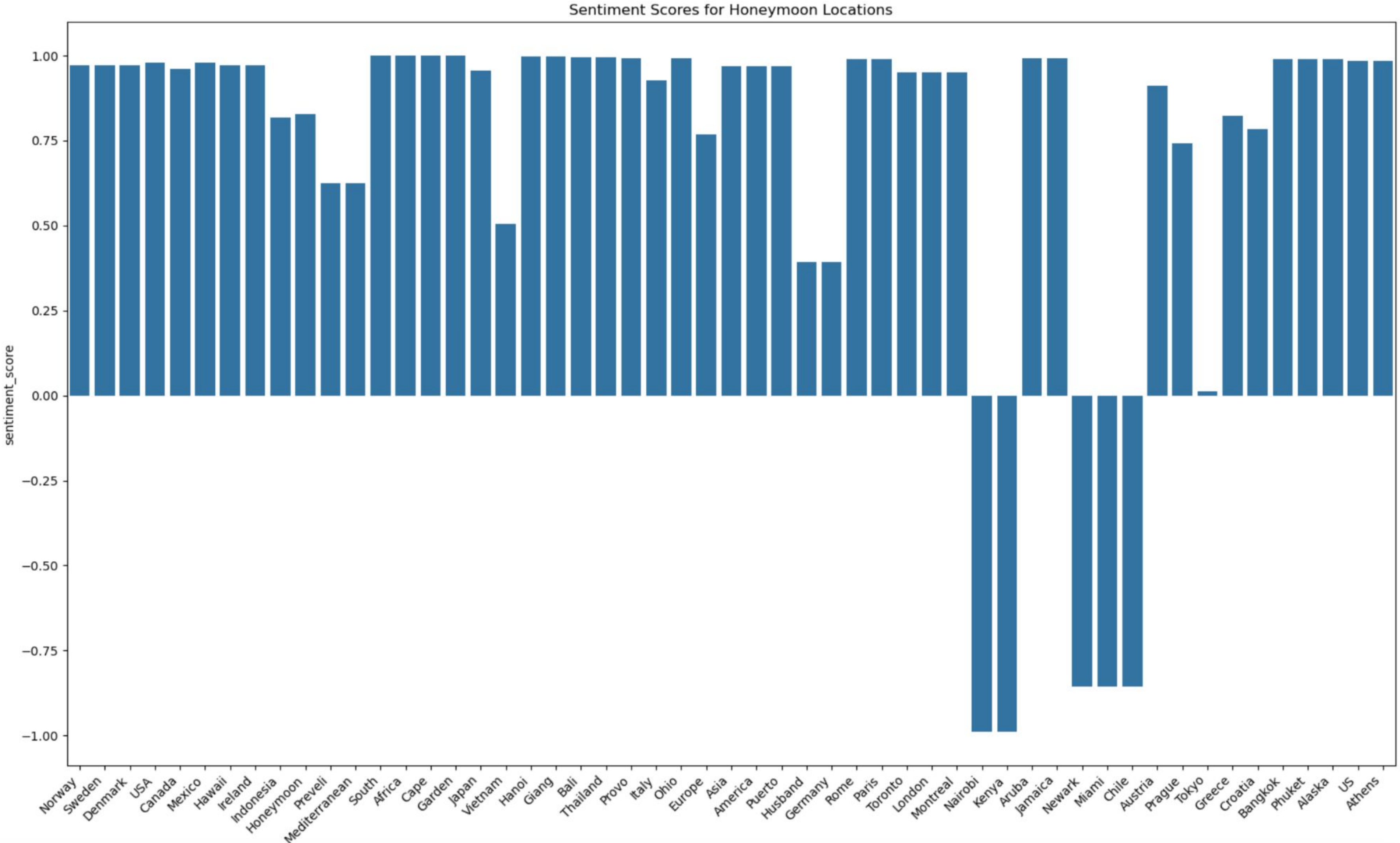
spaCy Installation: Installed spaCy (pip install spacy) for advanced NLP to recognize locations and places.

Place Identification: Cleaned and visualized place data, comparing location with sentiment score to predict positive reviews.

Model Finalization: Deployed an optimized ensemble model based on accuracy score using test data.


Sentiment Distribution by Subreddit
Post Count
0
200
400
600
800
1000
0
1
subreddit
Sentiment
Negative
Neutral
Positive

Models	Accuracy
Logistic Regression with TF-IDF	0.83649
Random Forest Classifier	0.80805
GridSearchCV and hyperparameter tuning	0.83175
An ensemble model	0.86255


Sentiment Scores for Honeymoon Locations
sentiment_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Norway
Sweden
Denmark
USA
Canada
Mexico
Hawaii
Ireland
Indonesia
Honeymoon
Preveli
Mediterranean
South
Africa
Cape
Garden
Japan
Vietnam
Hanoi
Giang
Bali
Thailand
Provo
Italy
Ohio
Europe
Asia
America
Puerto
Husband
Germany
Rome
Paris
Toronto
London
Montreal
Nairobi
Kenya
Aruba
Jamaica
Newark
Miami
Chile
Austria
Prague
Tokyo
Greece
Croatia
Bangkok
Phuket
Alaska
US
Athens

Recommendations

- Optimize model performance through hyperparameter tuning like grid search, Bayesian optimization, with k-fold cross-validation for better generalization.
- Group by continent, country, cities and other famous tourist spots
- Gather more data to include features to predict ‘Travel Mode’, ‘Mode of Staying’ and many more
- Enhance predictions by analyzing user preferences and trending destinations.

Thank you