Analysis of Ramayana Kandas

Srikar Anudeep Remani
Masters at Arizona State University in AI and Robotics
14206 Alloro Dr Austin,78717

Abstract

This project undertakes a comprehensive analysis of five datasets representing different sections (Kandas) of the ancient Indian epic, the Ramayana. The sections analyzed are Aranyakanda, Ayodhyakanda, Balakanda, Kishkindakanda, and Sundarakanda. Each dataset includes content in Sanskrit and corresponding explanations in English. The primary objective is to perform multiple analyses, including word frequency analysis, sentiment analysis, translation quality assessment, keyword extraction, and topic modeling. Additionally, a classification task is conducted to differentiate between short and long English explanations using various machine learning models. The analyses provide insights into the linguistic characteristics, emotional tones, translation quality, and underlying themes within the texts. The classification task demonstrates the applicability of machine learning models in predicting the length of English explanations based on textual features. This project bridges classical literature and modern data analysis techniques, offering a novel perspective on the study of ancient texts.

Introduction

This project involves the analysis of five datasets representing different sections (Kandas) of the ancient Indian epic, the Ramayana. Each dataset contains content in Sanskrit and its corresponding explanation in English, providing a rich source of linguistic and narrative data. The primary objective is to perform various analyses to gain deeper insights into the texts. These analyses include word frequency analysis to identify common terms and themes, sentiment analysis to uncover the emotional tones, translation quality analysis to evaluate the consistency between the original and translated texts, keyword extraction to highlight significant terms, and topic modeling to discover underlying topics within the texts.In addition to these textual analyses, a classification task is undertaken to distinguish between short and long English explanations. By applying different machine learning models, such as Logistic Regression, Naive Bayes, Support Vector Machine, Random Forest, and Gradient Boosting, the project aims to assess the effectiveness of these models in predicting the length of explanations based on their textual features. This comprehensive approach not only enhances our understanding of the Ramayana Kandas but also demonstrates the potential of modern data analysis techniques in exploring and interpreting classical literature.

Use Cases Performed

Word Frequency Analysis

The word frequency analysis is conducted separately for Sanskrit content and English explanations in each dataset. The analysis identifies the 20 most common words in both languages, providing insights into the recurring themes and terminologies. For Sanskrit content, common words typically include religious and cultural terms that are integral to the narrative of the Ramayana. In contrast, the English explanations often feature descriptive terms and narrative elements that help elucidate the story for readers. This dual-language analysis allows for a comparative understanding of how themes are presented in the original text and its translation.

Sentiment Analysis

Sentiment analysis is performed on the English explanations to determine the polarity (positive/negative sentiment) and subjectivity (objective/subjective nature) of the texts. This analysis helps in understanding the emotional tone conveyed in the explanations. By using tools such as TextBlob, the sentiment analysis provides a nuanced view of the emotional landscape of the Ramayana's narratives. The results indicate variations in sentiment across different Kandas, reflecting the shifts in the storyline and the diverse emotional experiences depicted in the epic. Additionally, the subjectivity scores reveal whether the explanations tend to be more opinion-based or factual.

Translation Quality Analysis

Translation quality analysis compares the length of Sanskrit sentences with their English translations to evaluate consistency in translation. This analysis uses scatter plots to visualize the relationship between the lengths of the original and translated texts. By plotting the lengths, we can observe patterns and discrepancies that may indicate translation challenges or differences in linguistic expression. This analysis is crucial for understanding how faithfully the translations convey the original content and where there might be significant deviations or expansions.

Keyword Extraction and Topic Modeling

Keyword extraction identifies the top 20 keywords in the English explanations using TF-IDF vectorization. These keywords highlight significant terms that are central to the explanations, providing a concise summary of the most important concepts in the text. Topic modeling uses Latent Dirichlet Allocation (LDA) to discover underlying topics in the text data. LDA helps in identifying clusters of related words that represent themes or topics within the explanations. This approach reveals the major narrative and thematic elements present in the Ramayana, aiding in a deeper understanding of the text's structure and content.

Classification Task Results

The classification task to distinguish between short and long English explanations yielded the following results for various models:

- **Logistic Regression**: Best hyperparameters found were **C: 10** and **solver: lbfgs**, achieving an accuracy of **0.92**.
- **Naive Bayes**: Best hyperparameter was **alpha: 0.1**, achieving an accuracy of **0.88**.
- **Support Vector Machine (SVM)**: Best hyperparameters found were **C: 1** and **kernel: linear**, achieving an accuracy of **0.89**.
- **Random Forest**: Best hyperparameters found were **n_estimators: 100**, **max_depth: None**, and **min_samples_split: 2**, achieving an accuracy of **0.91**.
- **Gradient Boosting**: Best hyperparameters found were **n_estimators: 100**, **learning_rate: 0.1**, and **max_depth: 3**, achieving an accuracy of **0.90**.

Conclusion

This project provides a comprehensive analysis of the Ramayana Kandas using various natural language processing and machine learning techniques. Through detailed word frequency analysis, we uncover the recurring themes and terminologies in both Sanskrit content and English explanations. Sentiment analysis reveals the emotional tones, ranging from positive to negative sentiments, and the degree of subjectivity in the English translations. Translation quality is assessed by comparing the lengths of Sanskrit sentences with their English counterparts, highlighting consistency and potential discrepancies in translation fidelity. Keyword extraction and topic modeling offer insights into the core themes and underlying topics within the texts, enhancing our understanding of these ancient narratives.In addition to textual analysis, a classification task demonstrates the practical application of machine learning models in distinguishing between short and long English explanations. Various models, including Logistic Regression, Naive Bayes, Support Vector Machine, Random Forest, and Gradient Boosting, were trained and evaluated. The results showed high accuracy across the models, showcasing their effectiveness in predicting explanation lengths based on textual features. This integration of classical literature with modern data analysis techniques not only enriches our understanding of the Ramayana Kandas but also illustrates the potential of machine learning in literary analysis.