Source - wikipedia

TripAdvisor Hotel Review Data Analysis Report

This project aimed to investigate the relationship between TripAdvisor reviews and corresponding ratings. By employing a combination of Natural Language Processing (NLP) techniques and machine learning models, the study sought to understand the extent to which review text could predict or explain rating variations.

Dataset

Hotels play a crucial role in traveling and with the increased access to information new pathways of selecting the best ones emerged.
With this dataset, consisting of 20k reviews crawled from Tripadvisor, you can explore what makes a great hotel and maybe even use this model in your travels

Source- zenodo

Libraries Used

- **TensorFlow and PyTorch:** Deep learning frameworks for model development.
- **scikit-learn:** Machine learning library for model training and evaluation.
- **NLTK:** Natural Language Toolkit for text preprocessing and analysis.
- **matplotlib and seaborn:** Data visualization libraries for creating informative plots.
- **pandas:** Data manipulation and analysis library for handling the dataset.

Data Preprocessing

- **Text Cleaning:** Removing noise, stop words, and irrelevant information from the review text.
- **Tokenization:** Breaking down text into individual words or tokens.
- **Feature Extraction:** Converting text data into numerical representations (e.g., tokenization and index).

Data Visualization

- **Word Clouds:** Visual representation of the most frequent words in the review data to identify prominent themes and keywords.
- **Pie Charts:** Graphical display of the distribution of categorical variables (e.g., rating distribution, sentiment distribution).
- **Bar Charts:** Comparison of categorical or numerical data (e.g., frequency of different rating categories, sentiment scores across different hotel chains).
- **Joint Plots:** Exploration of relationships between numerical variables (e.g., correlation between review length and rating).

NLP Techniques

- **Sentiment Analysis:** Determining the overall sentiment (positive, negative, neutral) of a review.
- **Topic Modeling:** Identifying the main themes or topics discussed in the reviews.
- **Aspect-Based Sentiment Analysis:** Analyzing sentiment towards specific aspects of the hotel (e.g., service, cleanliness, location).

BERT Model Application

- **Fine-tuning BERT:** Training the model on your TripAdvisor dataset to enhance performance.
- **Text Classification:** Categorizing reviews based on ratings or predefined classes (e.g., excellent, good, average, poor).
- **Sentiment Analysis:** Improving sentiment analysis accuracy using BERT's contextual understanding.
- **Aspect-Based Sentiment Analysis:** Identifying sentiment towards specific aspects (e.g., "food was excellent").

Potential Insights

- **Sentiment-Rating Correlation:** Understanding how sentiment directly impacts the rating.
- **Topic-Rating Correlation:** Identifying topics that strongly influence ratings (e.g., positive reviews about cleanliness, negative reviews about service).
- **Aspect-Rating Correlation:** Determining which hotel aspects contribute most to high or low ratings.
- **Review Evolution:** Analyzing how review characteristics (length, complexity, sentiment) change over time and their relationship to rating trends.

Key Findings (Potential)

- **Correlation between review and rating:** Quantifying the relationship between positive/negative sentiment and higher/lower ratings.
- **Impact of specific aspects:** Determining which aspects (e.g., service, cleanliness) have the strongest influence on ratings.
- **Identifying sentiment drivers:** Understanding the key factors that contribute to positive or negative reviews.
- **Predicting ratings:** Building a model to predict ratings based on review text.

Conclusion

The analysis revealed a limited correlation between TripAdvisor reviews and corresponding ratings. While reviews offer valuable insights into guest experiences and identify areas for hotel improvement, they proved insufficient as standalone predictors of overall ratings. The data suggests that factors beyond textual content significantly influence the assigned rating, indicating a complex interplay of variables.

While hotels can undoubtedly benefit from meticulous review analysis to enhance their services, relying solely on reviews to assess overall quality or performance may be misleading. A comprehensive evaluation necessitates consideration of additional metrics and contextual factors.