Capstone Project - Hotel Reviews Sentiment Analysis and Topic Modeling

A. Project Name

Hotel Reviews Sentiment Analysis and Topic Modeling

B. Project Summary

Sentiment Analysis (SA) helps to automatically and meaningfully discover hotel customers' satisfaction from their shared experiences and feelings on social media. Several studies have been conducted to improve the precision of SA in the hospitality industry, which vary in data preprocessing techniques, feature representation, sentiment classification levels, and models, and they use different datasets. Such variations are worthy of attention and monitoring. Despite the importance of SA in hospitality and tourism, review studies identifying gaps and suggesting future research directions are limited.

The project aims to analyze customer sentiment in hotel reviews using natural language processing (NLP) and machine learning techniques. By classifying reviews into positive and negative sentiments and extracting insights through topic modeling, the goal is to provide valuable feedback for hotel businesses. Customer reviews are a crucial source of information, helping hotels improve their services and enhance customer satisfaction.

The project involves cleaning and preprocessing textual data, conducting exploratory data analysis (EDA), generating word clouds to visualize key positive and negative words, applying sentiment analysis models, and using LDA and NMF techniques for topic modeling. The dataset includes hotel reviews, reviewer details, and metadata such as scores, review dates, and tags.

C. Deliverables of the Project

Data Preprocessing:

a. Text cleaning, tokenization, stopword removal, and lemmatization.
b. Handling missing values and outliers.

c. Parsing reviews into positive and negative sentiments.

Exploratory Data Analysis (EDA):

1. Understanding the distribution of reviewer scores.
2. Visualizing sentiment distribution across hotels, locations, and review counts.

Visualizations:

1. Word clouds for positive and negative reviews.
2. Distribution plots of sentiments and reviewer scores.

Sentiment Analysis Models:

1. Logistic Regression, Naive Bayes, and Support Vector Machines for sentiment classification.
2. Potential experimentation with deep learning models (RNN or LSTM) if applicable.

Topic Modeling:

1. Applying LDA and NMF to identify major themes in customer feedback.

Evaluation:

1. Evaluating models using accuracy, precision, recall, and F1 score.
2. Cross-validation to prevent overfitting and underfitting.

D. Resources

a. **Source**:Kaggle: Sentiment Analysis with Hotel Reviews : https://www.kaggle.com/code/jonathanoheix/sentiment-analysis-with-hotel-reviews/notebook#Conclusion
b. **Software**: Python libraries—Pandas, Scikit-learn, NLTK, spaCy, WordCloud, Gensim (for LDA), and NMF.

E. Milestones

a. Problem Definition: Identifying the need for sentiment analysis of hotel reviews.
b. Data Collection: Gathering a large dataset of hotel reviews.
c. Data Preprocessing: Cleaning and preparing the text data for analysis.
    d. Exploratory Data Analysis (EDA): Visualizing trends and extracting insights from reviews.
e. Feature Engineering: Creating features from the text data for machine learning models.
f. Modeling: Applying machine learning models for sentiment classification.
g. Topic Modeling: Extracting common themes from the reviews using LDA and NMF.
h. Model Evaluation: Assessing the performance of models using standard metrics.
i. Report Writing: Summarizing findings and preparing the final project report.

F. Report

This report provides a step-by-step breakdown of the notebook, explaining the methodology and rationale behind each step. The main sections cover data preprocessing, exploratory analysis, model training, evaluation, and topic extraction.

a) Data Preprocessing

The dataset used for this project includes the following columns: Review_Date, Average_Score, Hotel_Name, Reviewer_Nationality, Positive_Review, Negative_Review, Reviewer_Score, and others. Key preprocessing steps include:

- **Missing Data Handling**: Managing missing values by either imputing or removing incomplete records.
- **Text Preprocessing for Sentiment Analysis**:
    1. **Lowercasing**: Uniformly converting text to lowercase.
    2. **Removing Punctuation and Special Characters**: Simplifying the text to retain only relevant words.
    3. **Tokenization**: Splitting reviews into individual tokens using NLTK's word tokenizer.

4. **Stopword Removal**: Removing common stopwords to retain sentiment-bearing words.
5. **Lemmatization**: Reducing words to their root form for better analysis.

b) Exploratory Data Analysis (EDA)

EDA is essential for understanding data patterns and trends. Key analyses include:

- **Distribution of Reviewer Scores**: Visualizing the distribution of reviewer scores to assess balance or skewness in sentiment.
- **Review Word Count Distribution**: Analyzing word count distributions to understand if review length correlates with sentiment.
- **Word Cloud Visualizations**:
    - **Positive Word Cloud**: Frequently occurring words like "clean," "comfortable," and "friendly" from positive reviews.
    - **Negative Word Cloud**: Common negative words such as "noisy," "small," and "dirty."

c) Sentiment Analysis

The core task of sentiment analysis involves:

1. **Vectorization of Text Data**: Using techniques like CountVectorizer and TF-IDF to convert text into numerical features.
2. **Splitting Data**: Dividing the dataset into training and test sets for model evaluation.
3. **Model Training**:
    - **Logistic Regression**: A baseline binary classification model.
    - **Naive Bayes**: A probabilistic model commonly used for text classification.
    - **Support Vector Machines (SVM)**: A model that aims to maximize the margin between sentiment classes.
4. **Model Evaluation**: Assessing model performance using accuracy, precision, recall, and F1-score. Cross-validation was used to prevent overfitting.

d) Topic Modeling

Two topic modeling techniques were used to uncover recurring themes in customer feedback:

- **Latent Dirichlet Allocation (LDA)**: Applied to discover latent topics such as “staff friendliness,” “cleanliness,” and “room size.”
- **Non-Negative Matrix Factorization (NMF)**: An alternative technique to extract themes like “small rooms,” “noisy environments,” and “friendly staff.”

e) Model Evaluation

Final model evaluations included:

- **Sentiment Models**: Evaluating classification performance through accuracy and cross-validation.
- **Topic Models**: Reviewing the coherence of topics by analyzing the most relevant words.

G. Conclusion

The notebook successfully implements various natural language processing and machine learning techniques for sentiment analysis and topic modeling. Insights from this project provide actionable recommendations for hotels to enhance customer satisfaction by addressing negative feedback and reinforcing positive experiences.

H. Future Scope

Some potential areas for future work include:

- **Deep Learning Models**: Exploring RNNs or transformer-based models for better sentiment classification.
- **Quantitative Evaluation**: Incorporating accuracy, precision, and recall metrics to evaluate the sentiment analysis model.
- **More Data**: Applying the models to a larger dataset or incorporating additional features like hotel pricing or amenities.