Title: Sentiment Analysis of Hotel Reviews Using Machine Learning Techniques

Abstract

In the era of digitalization, online reviews play a crucial role in influencing consumer decisions, particularly in the hospitality industry. This paper presents a machine learning-based sentiment analysis model designed to classify hotel reviews into positive, neutral, and negative sentiments. The model employs natural language processing (NLP) techniques, including text preprocessing, feature extraction using TF-IDF, and classification using Support Vector Machines (SVM). The performance of the model is evaluated using cross-validation and confusion matrix metrics, demonstrating its effectiveness in understanding customer sentiments.

1. Introduction

The hospitality industry is heavily reliant on customer feedback, which is often expressed through online reviews. Understanding the sentiment behind these reviews can provide valuable insights for hotel management, enabling them to enhance customer satisfaction and improve service quality. This research aims to develop a robust sentiment analysis model that can accurately classify hotel reviews based on their sentiment.

2. Literature Review

Sentiment analysis has gained significant attention in recent years, with various approaches being proposed. Traditional methods include rule-based systems and lexicon-based approaches, while more recent advancements leverage machine learning and deep learning techniques. Studies have shown that models such as SVM, Naive Bayes, and neural networks can effectively classify sentiments in textual data. This research builds upon these findings by implementing an SVM-based model for hotel review sentiment analysis.

3. Methodology

3.1 Data Collection

The dataset used in this study consists of hotel reviews obtained from TripAdvisor. Each review is accompanied by a rating, which is used to categorise the sentiment into three classes: positive (ratings 4-5), neutral (rating 3), and negative (ratings 0-2).

3.2 Data Preprocessing

The preprocessing steps include:

- **Text Cleaning**: Removal of special characters, numbers, and conversion to lowercase.
- **Tokenization**: Splitting the text into individual words.
- **Stopword Removal**: Filtering out common words that do not contribute to sentiment.
- **Lemmatization**: Reducing words to their base form to ensure uniformity.

3.3 Feature Extraction

The TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer is employed to convert the preprocessed text into numerical features suitable for machine learning algorithms. The model is limited to the top 5000 features to reduce dimensionality and improve computational efficiency.

3.4 Model Training

The SVM classifier with a linear kernel is selected for sentiment classification. The model is trained using a training set created by splitting the dataset into 80% for training and 20% for testing. Hyperparameter tuning is performed using GridSearchCV to optimise the model's performance.

3.5 Model Evaluation

The model's performance is evaluated using cross-validation scores and a confusion matrix. The classification report provides precision, recall, and F1-score metrics for each sentiment class.

4. Results

4.1 Model Performance

The model achieved an average cross-validation score of approximately 85%, indicating a strong ability to classify sentiments accurately. The confusion matrix visualises the model's predictions against the actual sentiments, highlighting areas of misclassification.

4.2 Visualisations

Visualisations, including rating distribution and word clouds for each sentiment, provide additional insights into the data. The word clouds illustrate the most frequently used words in positive, neutral, and negative reviews, revealing common themes and sentiments expressed by customers.

5. Discussion

The results demonstrate that the SVM-based sentiment analysis model effectively classifies hotel reviews, providing valuable insights for hotel management. The model's ability to identify customer sentiments can aid in decision-making processes, allowing hotels to address negative feedback and enhance positive experiences.

6. Conclusion

This research presents a machine learning-based approach to sentiment analysis of hotel reviews, utilising NLP techniques and SVM classification. The model's performance indicates its potential as a tool for understanding customer sentiments in the hospitality industry. Future work may explore the integration of deep learning techniques and the application of the model to other domains.

7. References

- [1] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.
- [2] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.
- [3] Joachims, T. (1999). Support Vector Machines. In: Advances in Kernel Methods - Support Vector Learning. MIT Press.