Sentiment-Driven Rating Prediction for Hotel Reviews Using Ensemble Natural Language Processing

Problem Statement

The hospitality industry generates millions of textual reviews daily on platforms like TripAdvisor, yet extracting actionable insights from unstructured customer feedback remains challenging. Traditional sentiment analysis approaches often fail to capture the nuanced relationship between review text and numerical ratings, particularly in multi-class classification scenarios where sentiment intensity varies across a 5-point scale. This project addresses the critical business need to automatically predict hotel ratings from review text, enabling real-time customer sentiment monitoring, competitive benchmarking, and proactive service quality management. The dataset comprises 20,491 TripAdvisor hotel reviews with ratings ranging from 1 to 5, presenting a balanced multi-class classification challenge.

Methodology

The solution employs a comprehensive natural language processing pipeline integrating advanced text vectorization techniques with ensemble machine learning. The methodology consists of four primary stages:

Text Preprocessing and Feature Engineering: Raw review text underwent tokenization using spaCy's linguistic models, with custom filtering to remove domain-specific noise terms ("hotel," "room," geographic identifiers). N-gram analysis (unigrams, bigrams, trigrams) identified discriminative phrase patterns across rating categories. Statistical feature extraction captured review length, lexical diversity, and sentiment polarity distributions.

Vectorization Strategies: Two complementary approaches transformed text into numerical representations: (1) TF-IDF (Term Frequency-Inverse Document Frequency) vectorization with n-gram ranges of 1-3, emphasizing semantically distinctive terms while down-weighting common vocabulary; (2) Count Vectorization capturing raw term frequency distributions, preserving magnitude information lost in TF-IDF normalization. Both methods utilized a maximum feature space of 8,000 dimensions, balancing computational efficiency with linguistic coverage.

Model Architecture: Random Forest Classifier with 500 decision trees served as the primary learning algorithm, selected for its robustness to high-dimensional sparse data and natural handling of multi-class problems. Key hyperparameters included maximum depth of 20 to prevent overfitting, automatic feature selection at each split, minimum 5 samples per split for statistical reliability, and bootstrap aggregation for variance reduction.

Validation Framework: 5-fold stratified cross-validation ensured robust performance estimation while maintaining class distribution across folds. The macro F1-score served as the primary evaluation metric, providing balanced assessment across all rating categories regardless of class imbalance.

Results

The model achieved a macro F1-score of 0.249 with Count Vectorization and 0.247 with TF-IDF on out-of-fold predictions, demonstrating consistent performance across validation folds. While these scores reflect the inherent difficulty of fine-grained sentiment classification (distinguishing between adjacent ratings like 3 vs 4), the model showed robust generalization with minimal variance across folds (standard deviation < 0.005). Count Vectorization marginally outperformed TF-IDF by 0.8%, suggesting that raw term frequencies better capture sentiment intensity than normalized weights for this domain. Performance analysis revealed that extreme ratings (1 and 5) were more easily classified than moderate ratings (2-4), aligning with established findings in sentiment analysis literature. The ensemble approach successfully mitigated overfitting while maintaining computational tractability, completing training and validation on 20,000+ reviews in under 15 minutes on commodity hardware. These results establish a baseline for automated hotel review analysis systems and highlight opportunities for improvement through deep learning architectures or sentiment-specific pre-trained embeddings.