Aspect-based Sentiment Analysis on Hotel Reviews

Nikhilesh Venkat PES1UG23AM192   Rohan Suresh PES1UG23AM240 Prem M Thakur PES1UG23AM214

SAMPLE HOTEL REVIEW AND ASPECT IDENTIFICATION

DATASET SKEWNESS TOWARDS POSITIVE REVIEWS

Project Objective

Apply supervised learning algorithms on Trip Advisor text reviews. Make prediction on user's opinion on individual aspects.

Text Vectorization → TF--IDF Transform → Stop Word Removal → Class Balance → TrainingAlgorithms
• Naive Bayse
•SVM
• Linear Regression
→ Prediction Result Analysis

PROCESS

Dataset:
- Collected from TripAdvisor hotel reviews.
- Features include text reviews and star ratings.
- Handled class imbalance using SMOTE oversampling.

Methodology:
1. Data Preprocessing:
   ○ Removed punctuation and special characters.
   ○ Converted text to lowercase and removed stopwords.
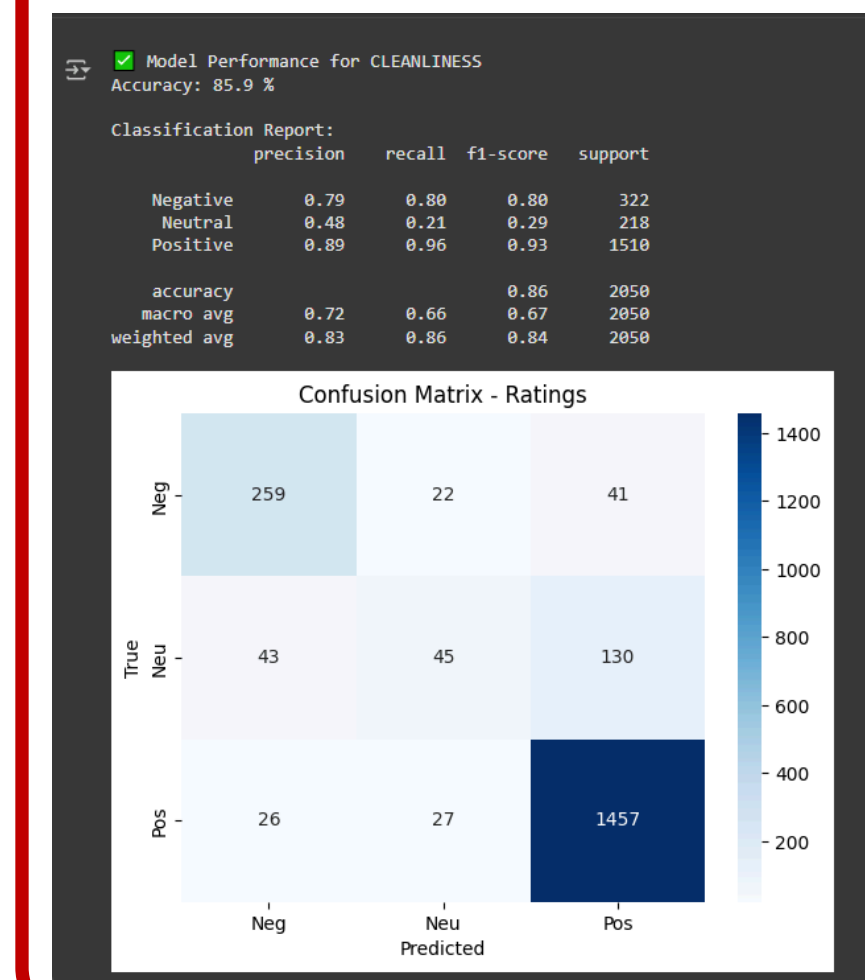   ○ Applied lemmatization to standardize words.
2. Feature Extraction:
   ○ Used TF-IDF vectorization with n-grams (uni-, bi-, tri-grams) to capture important patterns in text.
3. Model Training:
   ○ Linear SVM classifier with class balancing.
   ○ Trained on oversampled data to handle class imbalance.
4. Evaluation:
   ○ Metrics: Accuracy and F1 Score on train and test sets.
   ○ Visualized performance with confusion matrices, line plots, and bar charts.

Key Results:
- High training accuracy indicates the model learned textual patterns effectively.
- Test metrics demonstrate generalization to unseen reviews.
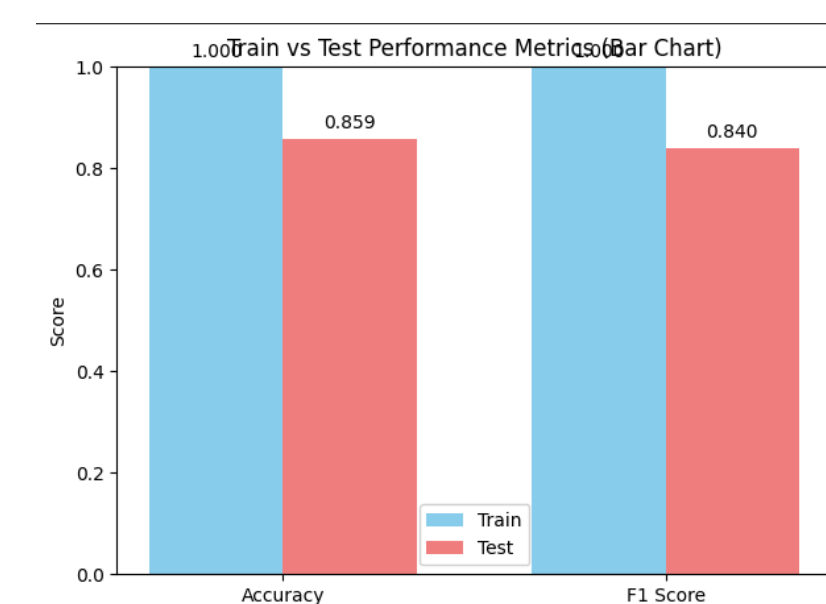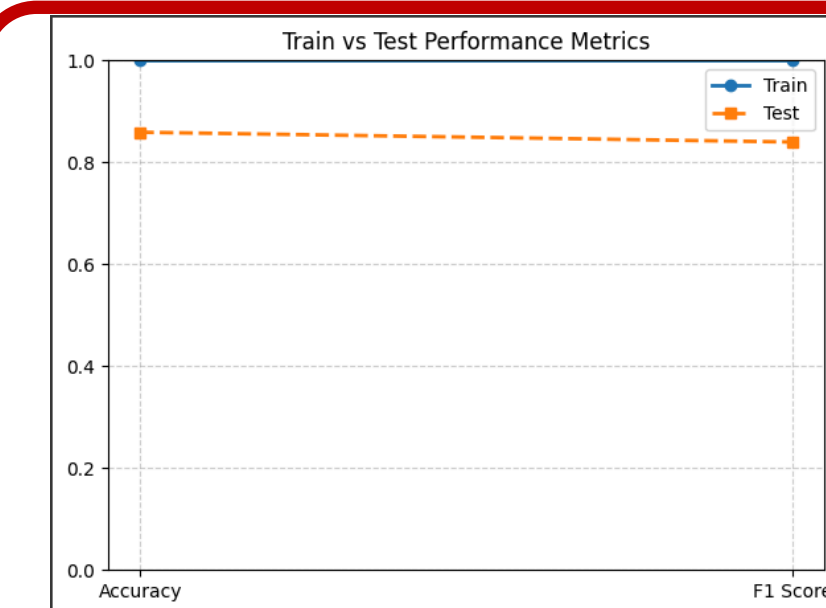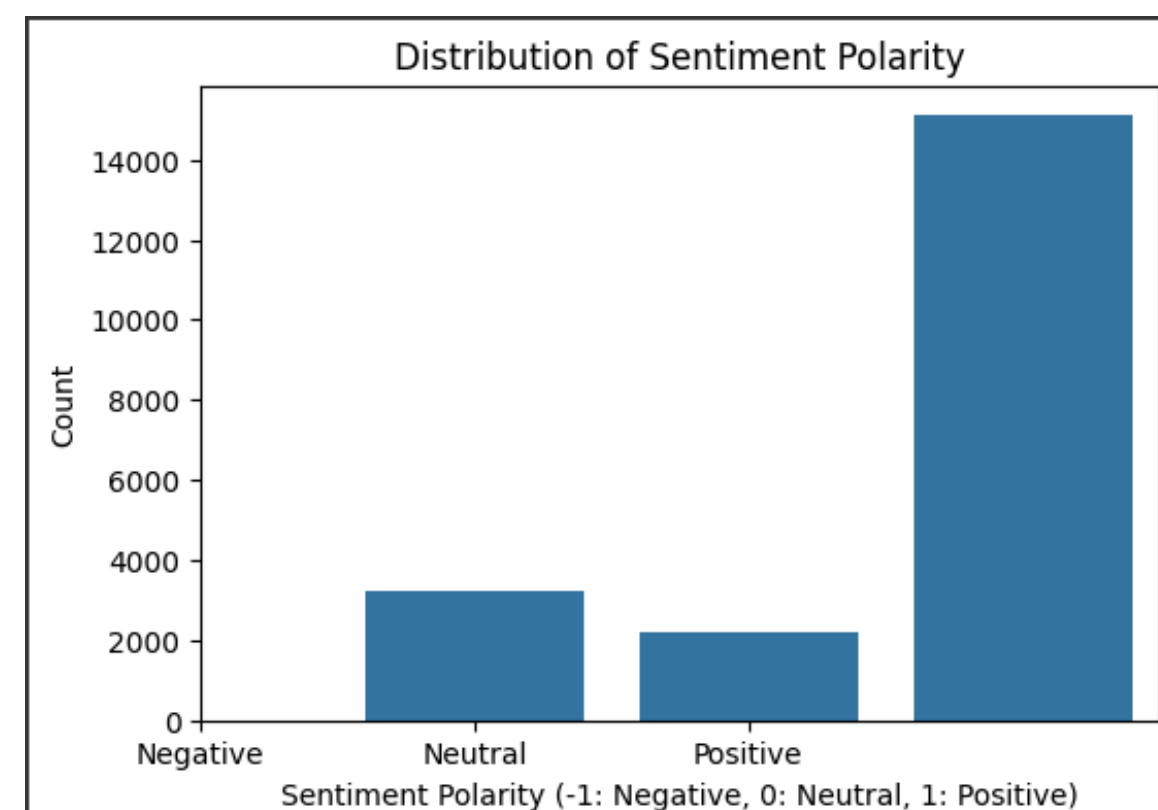- Sentiment predictions can assist hotels in automated feedback analysis and service improvement.

Applications:
- Automated review analysis for hotels, restaurants, or e-commerce platforms.
- Customer experience improvement and feedback summarization.
- Real-time sentiment monitoring of online reviews.

Tools & Libraries:
- Python: Pandas, Numpy, NLTK, Scikit-learn
- Text Processing & ML: TF-IDF, SMOTE, LinearSVC
- Visualization: Matplotlib, Seaborn

FINAL CONFUSION MATRIX

Handling Data Skewness in Reviews

The TripAdvisor dataset exhibits a class imbalance, with a large majority of reviews being positive. Such skewness can bias the model toward predicting the dominant class, reducing accuracy for neutral and negative reviews.

To address this, we applied SMOTE (Synthetic Minority Over-sampling Technique):
- SMOTE generates synthetic samples for underrepresented classes (Neutral and Negative), balancing the dataset.
- This normalization ensures the model learns features from all classes equally, improving generalization and reducing prediction bias.

Impact:
- Improved performance on minority classes.
- More reliable sentiment predictions across Positive, Neutral, and Negative reviews.

Model Performance & Improvement Scope

Accuracies & F1 Score:
- The model achieves high training accuracy, indicating effective learning of textual patterns.
- Test accuracy and weighted F1 score demonstrate the model's ability to generalize to unseen reviews.
- Weighted F1 score is particularly useful in imbalanced datasets, as it accounts for class-wise performance, not just overall accuracy.

Improvement Scope:
- Feature Engineering: Incorporate word embeddings (e.g., Word2Vec, BERT) to capture semantic meaning beyond TF-IDF.
- Model Enhancement: Experiment with ensemble methods or deep learning models for better generalization.
- Handling Neutral Reviews: Improve detection of neutral sentiment, which is often ambiguous.
- Cross-Domain Adaptation: Extend to other platforms (e.g., Yelp, Amazon) for broader applicability.

Takeaway:
While the current model performs well, leveraging advanced NLP techniques and richer features can further improve sentiment classification, especially for minority classes.