Report on Sentiment Analysis in Amazon Customer Reviews - Oorja

Task Definition

Objective: The primary goal of this project is to perform binary sentiment classification on Amazon customer reviews to accurately distinguish between positive and negative sentiments expressed by customers.

Significance: This analysis helps in understanding customer preferences and sentiments, which can significantly impact business strategies, product improvements, and customer service enhancements. Accurate sentiment analysis provides a more nuanced understanding of consumer behaviour and can guide marketing and product development decision-making processes.

Dataset Insights

Preparation Process:

- Source: The dataset was sourced from "Amazon product data," curated by Dr. Julian McAuley from UCSD. It focuses on reviews from the Kindle store spanning from May 1996 to July 2014.
- Labeling: Reviews were labeled as negative for ratings of 1-3 and positive for ratings of 4-5.
- Undersampling: Due to a large imbalance in the dataset (84.4% positive), positive reviews were undersampled to match the number of negative reviews, ensuring a balanced dataset for unbiased model training.
- Preprocessing: The text was cleaned using HTML parsing, normalization of contractions, padding of punctuation, and tokenization. Significant emphasis was placed on negation handling and the use of n-grams to enhance the model's ability to understand context and nuances in language.

Training Summary

Steps to Fine-Tune the Model:

- Feature Extraction: Implemented a Bag of Words model using tokenization, normalization, and TF-IDF weighting to convert text into numerical vectors.
- Model Selection: Three models were compared: Multinomial Naive Bayes, Logistic Regression, and Linear Support Vector Classification.
- Hyperparameter Tuning: Basic feature selection was performed by removing features appearing only once. The models were fine-tuned with varying preprocessing techniques, including the addition of negation handling and n-grams.

Evaluation Results

Metrics and Analysis:

- Accuracy Achieved: The LinearSVC model demonstrated the highest accuracy, especially when both negation handling and n-grams were used, achieving an accuracy of up to 87.31% on the test set.
- Comparison: All models showed improvement with the addition of complex preprocessing steps. The comparative analysis highlighted that while all models benefited from enhanced preprocessing, LinearSVC consistently outperformed the others across different setups.

Future Improvements

Suggestions for Enhancing the Model:

1. Advanced Preprocessing: Further refine preprocessing techniques, perhaps by exploring more sophisticated methods of handling negations and utilizing contextual embeddings like Word2Vec or BERT for feature extraction.
2. Deeper Feature Selection: Implement more granular feature selection techniques to optimize the model's performance, reducing overfitting and improving generalization.
3. Algorithm Exploration: Experiment with more complex algorithms, such as deep learning models like LSTM or GRU, which might capture sequential dependencies in text more effectively.
4. Parameter Optimization: Conduct more extensive hyperparameter tuning, including exploring different kernel types for SVC and regularization strengths for Logistic Regression.
5. Cross-Validation: Implement k-fold cross-validation to ensure the model's robustness and reliability across different subsets of data.
6. Real-Time Analysis: Develop a pipeline for processing and classifying reviews in real-time, allowing for dynamic updates to sentiment analysis as new reviews are posted.

By addressing these areas, the project can further enhance the accuracy and utility of sentiment analysis, providing more actionable insights for businesses based on customer feedback.