Amazon Customer Reviews: Sentiment Analysis

Abstract

The rapid growth of e-commerce platforms has resulted in a plethora of customer reviews that serve as valuable resources for both businesses and consumers. However, manually analyzing these reviews for sentiment is a huge task. This project aimed to automate the sentiment analysis process of customer reviews by using Natural Language Processing (NLP) and Machine Learning techniques.

I used a version of the popular Amazon Review dataset on Kaggle which was significantly reduced in size to: ~128000 rows and 11 columns. Using this dataset consisting of customer reviews and star ratings, the study employs preprocessing techniques to clean and transform the textual data into a format suitable for machine learning. Two classic classifiers—Logistic Regression and Naive Bayes—are used to predict the sentiment of the reviews as "Positive," "Neutral," or "Negative."
To manage the effects of class imbalance in the dataset, a resampling technique, specifically Random Over Sampling, is used in the model training pipeline. Hyperparameter tuning is carried out using GridSearchCV, offering a systematic way to optimize the models using 3-fold cross-validation.

Initial results show that both models achieve moderate success in classifying positive reviews but struggle to accurately predict neutral and negative sentiments. The best-performing Logistic Regression model achieves an overall accuracy of 76%, although with compromised precision and recall for the minority classes. In contrast, the Naive Bayes model has an overall accuracy of 72%, suffering from similar limitations.

Future work could include more advanced resampling techniques, feature engineering, and the application of more complex machine learning models to improve the classification performance across all sentiment classes.

-S.Saini