

Abstract

SENTIMENT ANALYSIS - Amazon Customer Reviews

Objective:

The primary aim of this project was to develop a machine learning model capable of classifying customer reviews into distinct sentiment categories: Positive, Neutral, and Negative. The model is trained on a dataset containing reviews and star ratings for various products on Amazon.

Methodology:

The project employed Natural Language Processing (NLP) techniques for text cleaning and feature extraction. The cleaned reviews were transformed into numerical form using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. The class imbalance issue was addressed by oversampling the minority classes. A Logistic Regression model was trained on the resampled dataset and evaluated using metrics such as accuracy, precision, recall, and F1-score.

Results:

The trained model achieved an overall accuracy of approximately 75.975.9. While the model demonstrated high precision and recall for 'Positive' reviews, it showed significant improvement in identifying 'Negative' and 'Neutral' reviews after resampling, with recall rates of (63%\ and (53%\ respectively.

Conclusions:

The project successfully demonstrated the applicability of machine learning algorithms in sentiment analysis of customer reviews. Although the model showed promising results, especially after addressing class imbalance, further improvements are still possible through hyperparameter tuning, feature engineering, and the application of more complex algorithms.

Dataset Overview

- **Number of Rows:** 128,845
- **Number of Columns:** 11