Sentiment Analysis of Amazon Reviews Using Logistic Regression

By Anna Fenner

Abstract

This project aims to perform sentiment analysis on a dataset of 4,000,000 Amazon reviews. We developed logistic regression models to classify the sentiment of reviews as positive or negative, comparing the effectiveness of CountVectorizer and TF-IDF vectorization methods. The CountVectorizer model achieved a slightly higher precision score (0.8856) compared to the TF-IDF model (0.8830). The findings highlight the potential of these models in accurately classifying sentiments, providing a foundation for further improvement using advanced models like random forests and LSTM neural networks.

Why

Understanding customer sentiment through reviews is crucial for businesses seeking to enhance their products and services. This project focuses on performing sentiment analysis on a large dataset of Amazon reviews. By classifying reviews as positive or negative using logistic regression models, we aim to provide insights that can help businesses make data-driven decisions. The project also compares the performance of two vectorization methods, CountVectorizer and TF-IDF, to determine the most effective approach for sentiment classification.

Data

This dataset from Kaggle.used for this project comprises of 4,000,000 Amazon reviews, with 400,000 designated for testing. The data was sourced from a Kaggle dataset. Each review is labeled as positive (1) or negative (0). The text data underwent preprocessing, including tokenization, lowercasing, removal of punctuation and stop words, and stemming.

Preprocessing Steps

1. Tokenization: Two vectorization methods were applied to the preprocessed text data:
    a. CountVectorizer: Converts text data into a matrix of token counts.
    b. TF-IDF Vectorizer: Converts text data into a matrix of term frequency-inverse document frequency values.
2. Lowercasing: Converting all words to lowercase.
3. Removing punctuation and special characters.
4. Removing stop words: Filtering out common words that do not contribute to sentiment.
5. Stemming: Reducing words to their root forms.

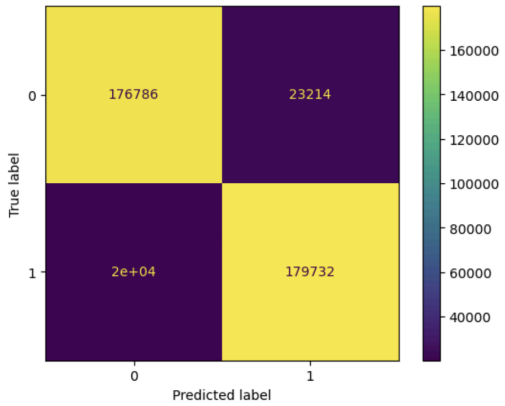
Models and Evaluation

Logistic regression models were trained using both vectorized datasets. Precision was chosen as the evaluation metric to minimize false positives.

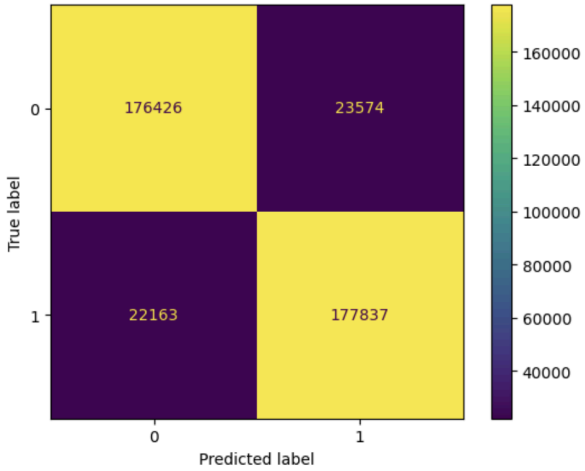
Results

The analysis focused on evaluating the performance of logistic regression models trained with different vectorization methods.

1. Model 1: Logistic Regression with CountVectorizer
    a. Precision Score: 0.8856
    b. Confusion Matrix:



2. Model 2: Logistic Regression with TF-IDF Vectorizer
    a. Precision Score: 0.8830
    b. Confusion Matrix:



The slightly higher precision of the CountVectorizer model suggests it is more effective at minimizing false positives. These results demonstrate the effectiveness of both models in accurately classifying the sentiment of Amazon reviews.

Conclusion

In conclusion, the logistic regression models trained for sentiment analysis on Amazon reviews effectively classified positive reviews with high precision. Moving forward, the project will explore the implementation of more advanced models, such as random forests and LSTM neural networks, to further enhance sentiment analysis capabilities. These models can potentially capture more intricate patterns in the text data, improving overall performance and providing deeper insights into customer sentiment.