Sentiment Analysis on Amazon Reviews

Introduction

The goal of this project was to perform sentiment analysis on Amazon reviews. Sentiment analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker.

Data Preprocessing

The data used in this project was a CSV file containing Amazon reviews. The pandas library was used to load the data into a DataFrame. The data was then preprocessed using several steps:

1. **Label Encoding**: The 'Sentiment' column, which contained the target labels, was transformed from categorical data to numerical data using LabelEncoder from sklearn.preprocessing.
2. **Text Normalization**: The 'reviewText' column, which contained the review texts, was normalized. This involved converting the text to lowercase, removing URLs, replacing non-word characters with spaces, and removing leading and trailing spaces.
3. **Vectorization**: The normalized texts were then converted into a matrix of token counts using CountVectorizer from sklearn.feature_extraction.text.
4. **TF-IDF Transformation**: The count matrix was transformed into a TF-IDF representation using TfidfTransformer from sklearn.feature_extraction.text.

Model Training and Evaluation

Two models were trained on the preprocessed data:

1. **Multinomial Naive Bayes**: A MultinomialNB model was trained on the data. The number of mislabeled points out of the total number of points was printed to evaluate the model's performance.
2. **Stochastic Gradient Descent Classifier**: A SGDClassifier model was also trained on the data. The model's performance was evaluated in the same way as the MultinomialNB model.

Confusion matrices were generated for both models to provide a more detailed view of their performance. The confusion matrix is a specific table layout that allows visualization of the performance of an algorithm. Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class (or vice versa).

Conclusion

This project demonstrated how to perform sentiment analysis on Amazon reviews using two different models. The models were evaluated based on the number of mislabeled points and confusion matrices. Future work could involve trying different models, tuning the parameters of the current models, and performing more detailed exploratory data analysis.