Automated Customers Reviews
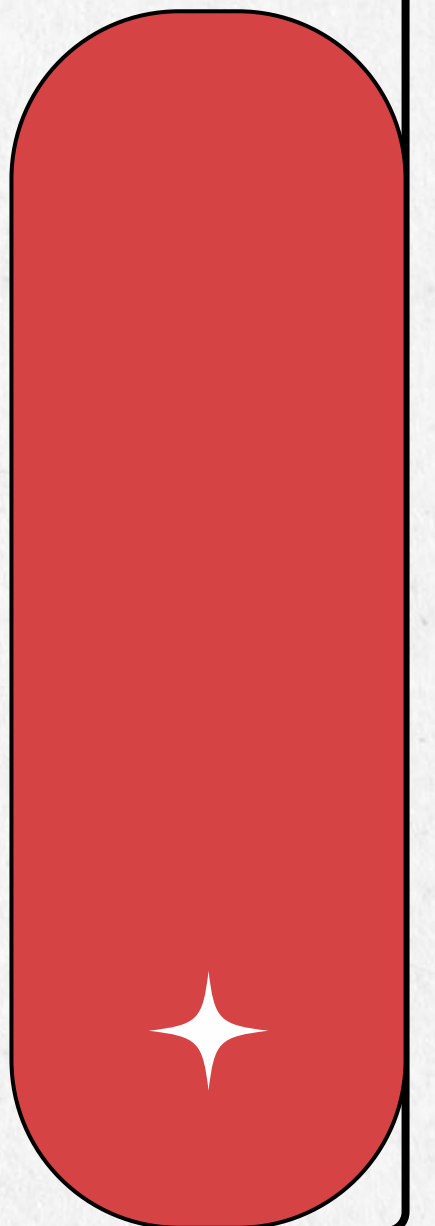
BY : ELAF ALKHALDI

Table of content

1- Project objective.

2- Preprocessing (Map Star Ratings to Sentiment Classes)

3- Classification with pre-trained models.

4- Product Category Clustering.

5- Summarize reviews using generative AI.

6- Deploy via Flask.

Project objective

The objective of this project is to develop a comprehensive product review system utilizing NLP models. The system will classify customer reviews, cluster them based on product categories, and generate recommendation articles by summarizing feedback using generative AI techniques.

Preprocessing

Two product review datasets were merged into a single file named **merged_reviews.csv**. The combined data was loaded using the Pandas library for further processing. A mapping function was implemented to convert numerical star ratings into categorical sentiment labels: ratings of 1 or 2 were categorized as Negative, a rating of 3 as Neutral, and ratings of 4 or 5 as Positive. Based on this mapping, a new column named **sentiment** was added to the dataset.

The dataset was then split into training and testing sets, with 80% of the data allocated for training and 20% for testing, ensuring reproducibility by setting a fixed random state.
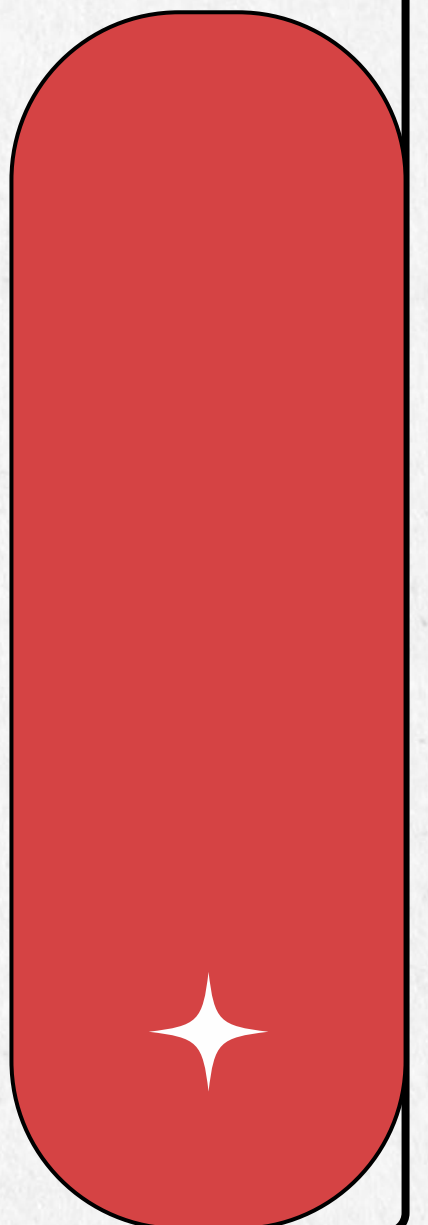
Classification with RoBERTA

The code uses **RoBERTa**, a transformer-based model designed to improve performance on natural language understanding tasks by training longer and on larger datasets compared to BERT. It was fine-tuned to classify customer reviews into three sentiment classes: Negative, Neutral, and Positive.

To address class imbalance in the dataset, **class weights** were computed and incorporated into the loss function, ensuring the model does not favor the majority class during training.
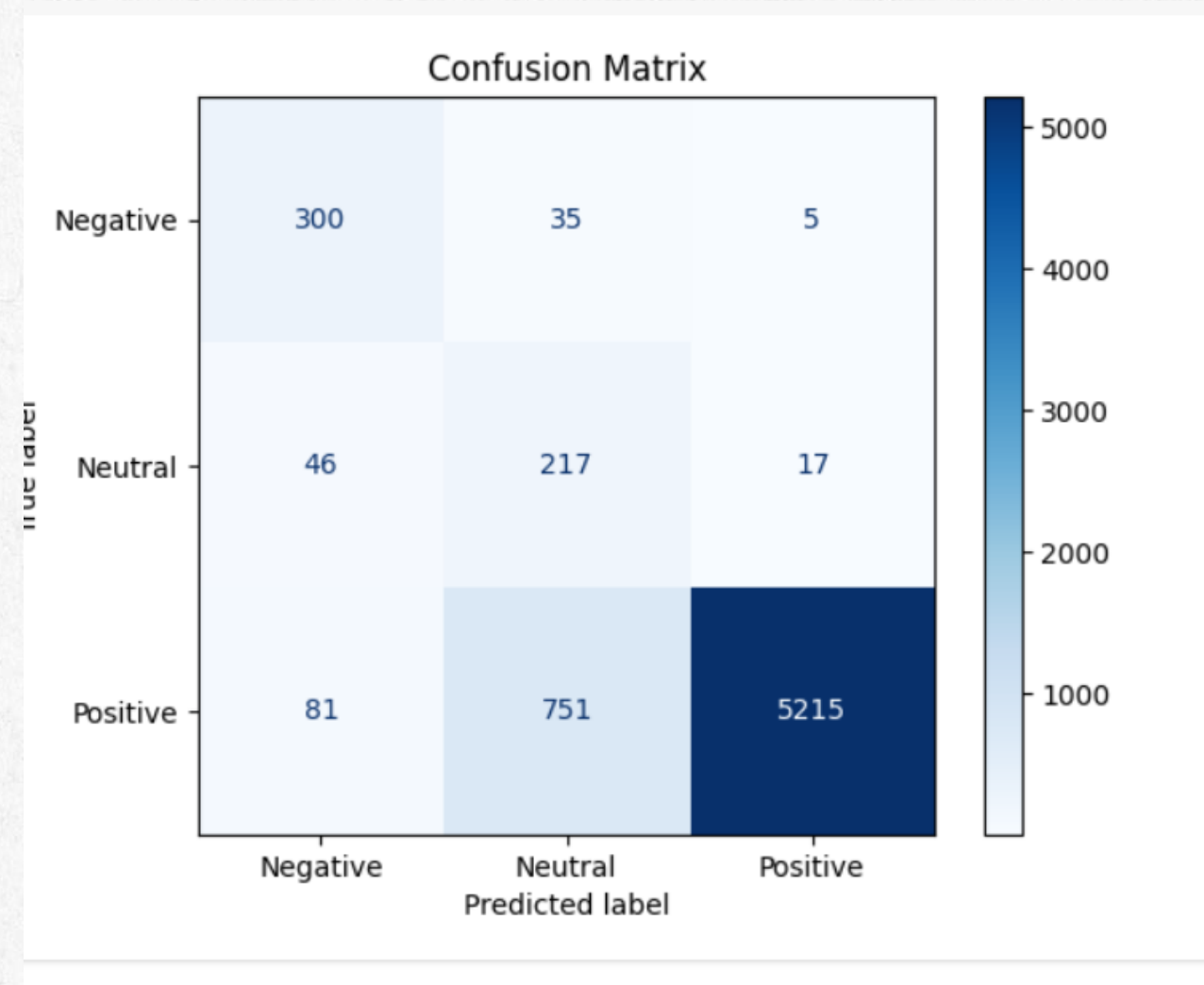
After training, the model achieved its results by evaluating on the test set and generating a detailed classification report showing the precision, recall, and F1-score for each sentiment class.
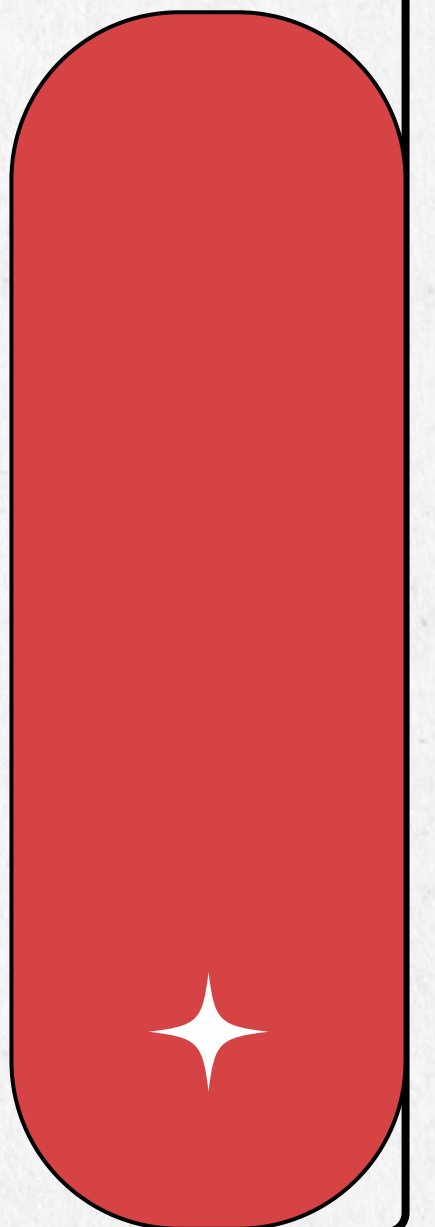
Training result

CONFUSION MATRIX

PRECISION, RECALL, F1-SCORE AND ACCURACY

|	precision	recall	f1-score	support
Negative	0.82	0.80	0.81	340
Neutral	0.37	0.76	0.49	280
Positive	0.99	0.95	0.97	6047
accuracy			0.93	6667
macro avg	0.73	0.83	0.76	6667
weighted avg	0.96	0.93	0.94	6667

Accuracy: 0.9307

Product category clustering

This code uses the **RoBERTa** model to extract semantic embeddings from product names loaded from a CSV file. It applies **KMeans clustering** to group similar products into 5 clusters. Each product is also categorized manually based on keywords into types like "Ebook Readers," "Batteries," "Accessories," and others. After clustering, the code identifies the top frequent words in each cluster and generates **word clouds** to visualize common themes.
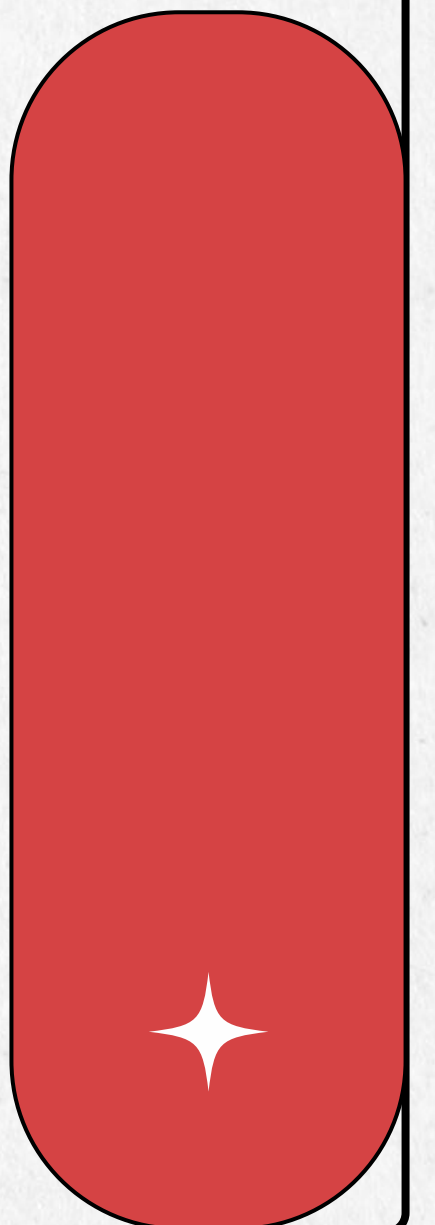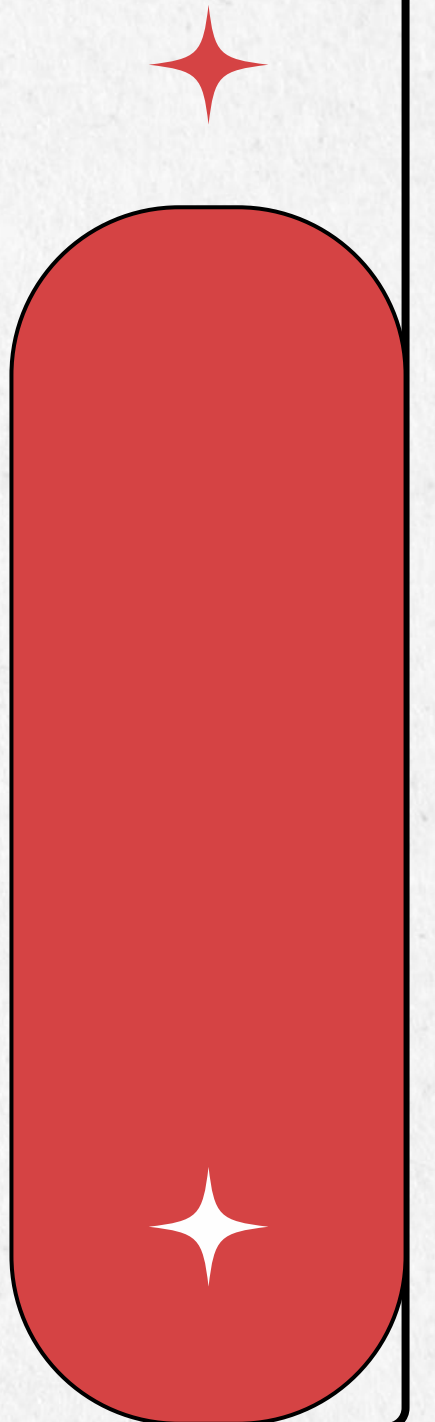
Summarize reviews using generative AI

The code processes a dataset of product reviews by selecting key columns and removing missing entries. It uses a pre-trained BART model for summarization. Reviews are grouped by product category, and for each category, the top three products (based on highest ratings) and the worst product (based on lowest rating) are selected. It summarizes reviews and highlights complaints for these products, then generates a final article summarizing the key points for each category.

DEMO

ANY QUESTIONS?

Thank you!