Project NLP | Business Case: Automated Customer Reviews - Final report

1) Project Overview

The goal of this project is to develop a product review system.That can classify sentiment, cluster by product categories and write product summaries into recommendation articles.

Data used for this project is from Amazon reviews, specifically from the *Video_Games* category. Data can be found here: https://amazon-reviews-2023.github.io/

2) Approach

Data Preprocessing

Mapped star ratings to sentiment labels (1–2 = Neg, 3 = Neutral, 4–5 = Pos).

Used minimalized cleaning (normalize unicode, remove html tags, normalize space), to keep it transformer friendly

Analysed data, to find imbalance and check review lengths.

Review Classification Model

Splitted the data 70/15/15

Tokenised with 3 different transformer models
- **RoBERTa-base**
- **BERT-base-uncased**
- **DistilBERT-base-uncased**

For each model trained using different weights for neutral class for the imbalance, learning rates and epochs.And calculated precision, recall and f1-score.

Results were visualized on a test sample to see confusion matrix, per class scores and model vs model comparison.

In the end RoBERTa performed the best overall. See results

Clustering Model

It was first done on a sample set of data, to quickly see results and give an idea of clusters in the data set.

Data was first embedded using: all-MiniLM-L6-v2. Then PCA was used to do dimensionality reduction. After this MiniBatchKmeans was used for clustering the data. With k=6, and then labeling the clusters with top terms using TF-IDF. It was clear that 2 of 6 clusters were not categorical but sentimental. So these 2 clusters were merged to nearest clusters. Resulting in 4 categorical clusters. Clusters were inspected to find categories.

Summarization Model

The review clusters were analysed to find the top rated and worst products. Key words were used to prevent leaked products with good reviews to be found as the best product of the category.

Used the generative model GPT-4o-mini to create articles about the top 3 products, their main complaints and the worst product for each category. Using a zero shot approach with clear instructions.

facebook/bart-large-cnn and google/flan-t5-large transfer models were also tried. for making the summarization.

Deployment

Done on Hugging Face Space. On this space the generated review summaries can be found. On the top 3 products and key differences between them. Top complaints for each of those products.And the worst product in the category and why it should be avoided. There is also a classification tab in which a review can be uploaded, which will then be classified on sentiment by the roberta-base model. And also the found clusters from the data set can be seen.

See https://huggingface.co/spaces/DaanBooy/games_and_accessories_reviews

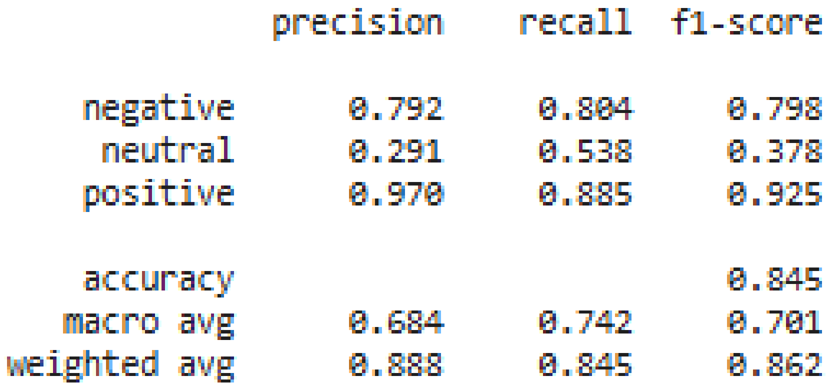
---

3) Results

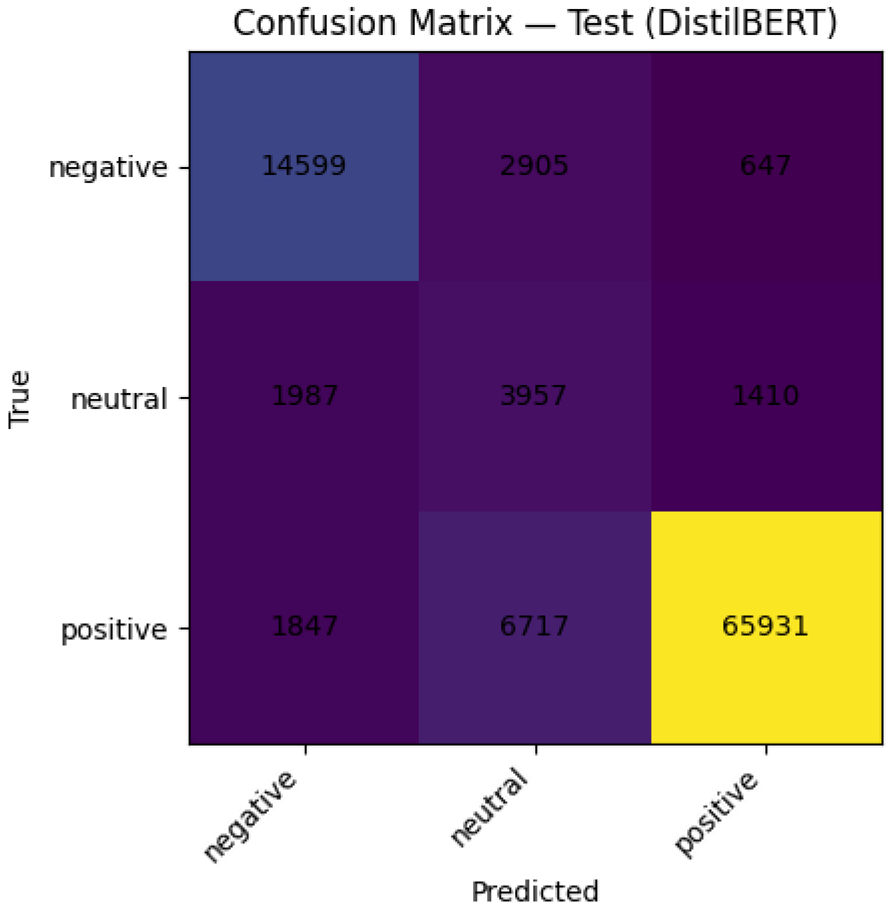
Sentiment classification:

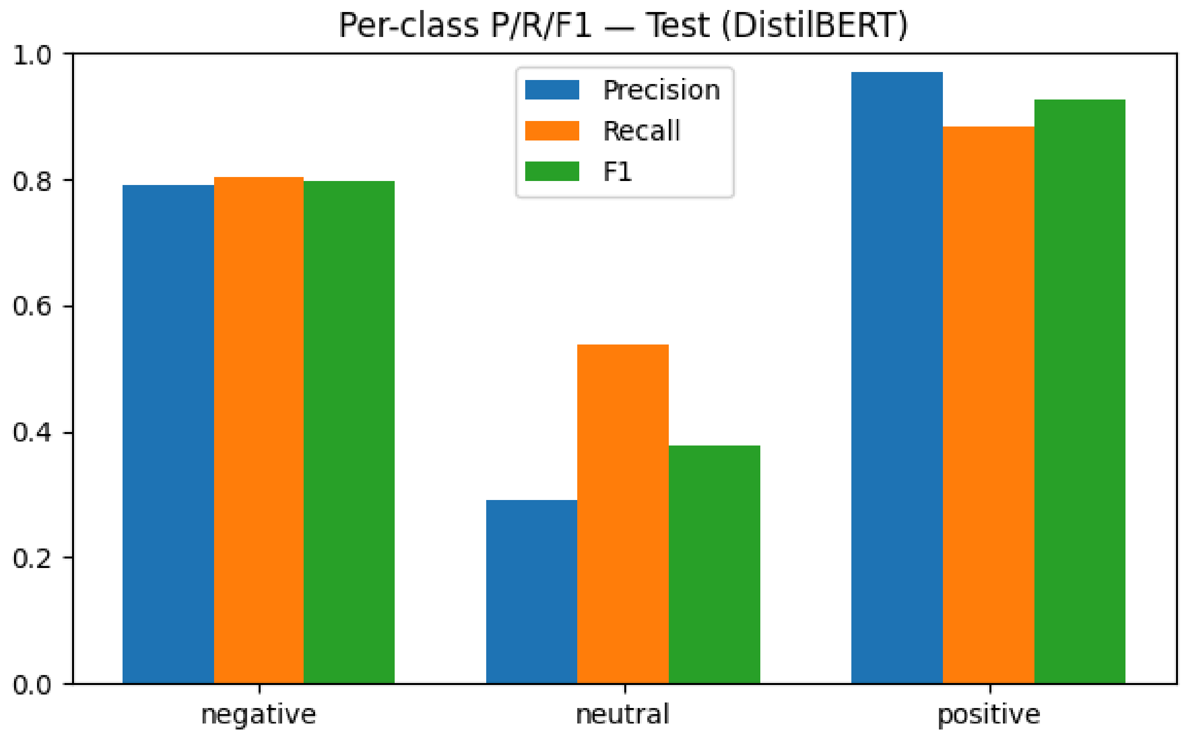
Distilbert-base-uncased

Scores:



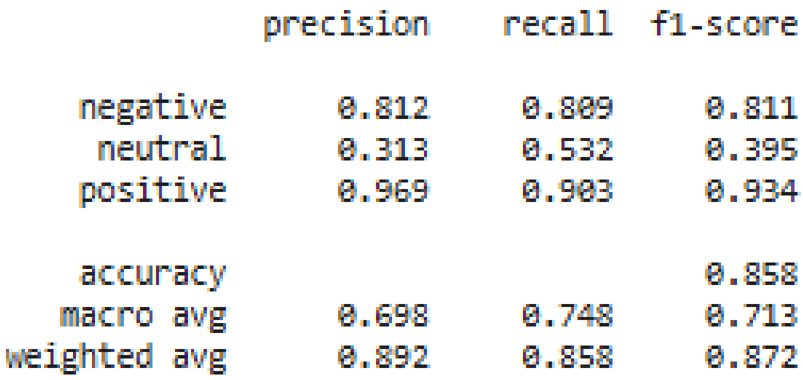
```
              precision    recall  f1-score

    negative      0.792     0.804     0.798
     neutral      0.291     0.538     0.378
    positive      0.970     0.885     0.925

    accuracy                          0.845
   macro avg      0.684     0.742     0.701
weighted avg      0.888     0.845     0.862
```



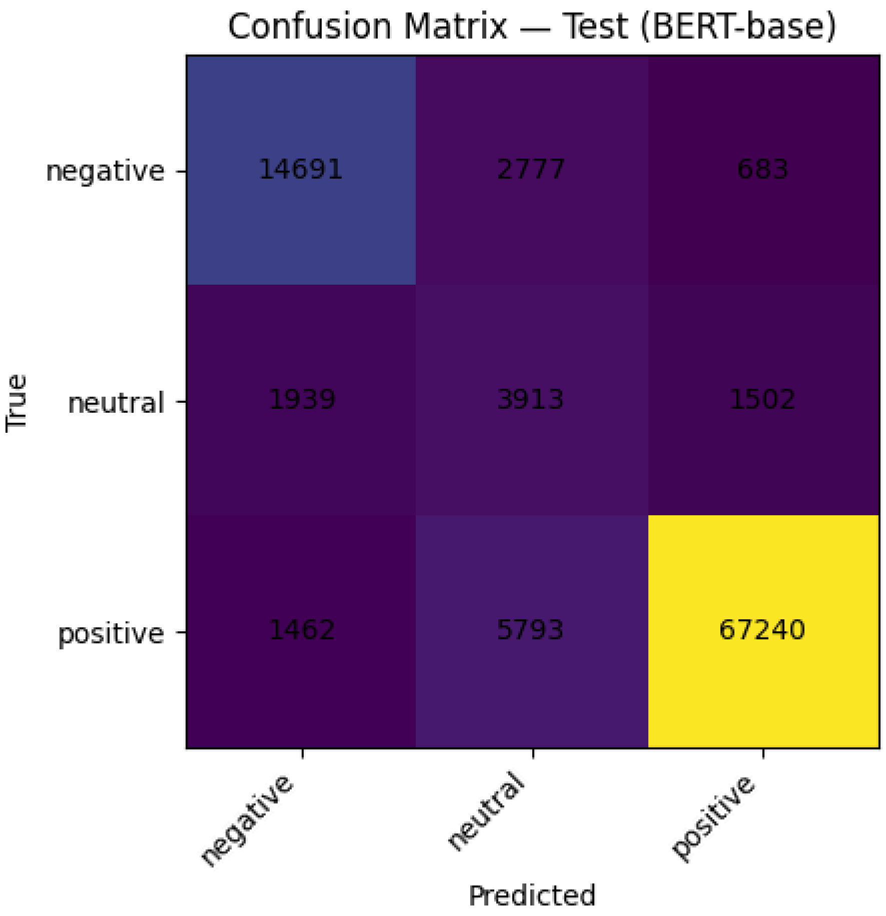


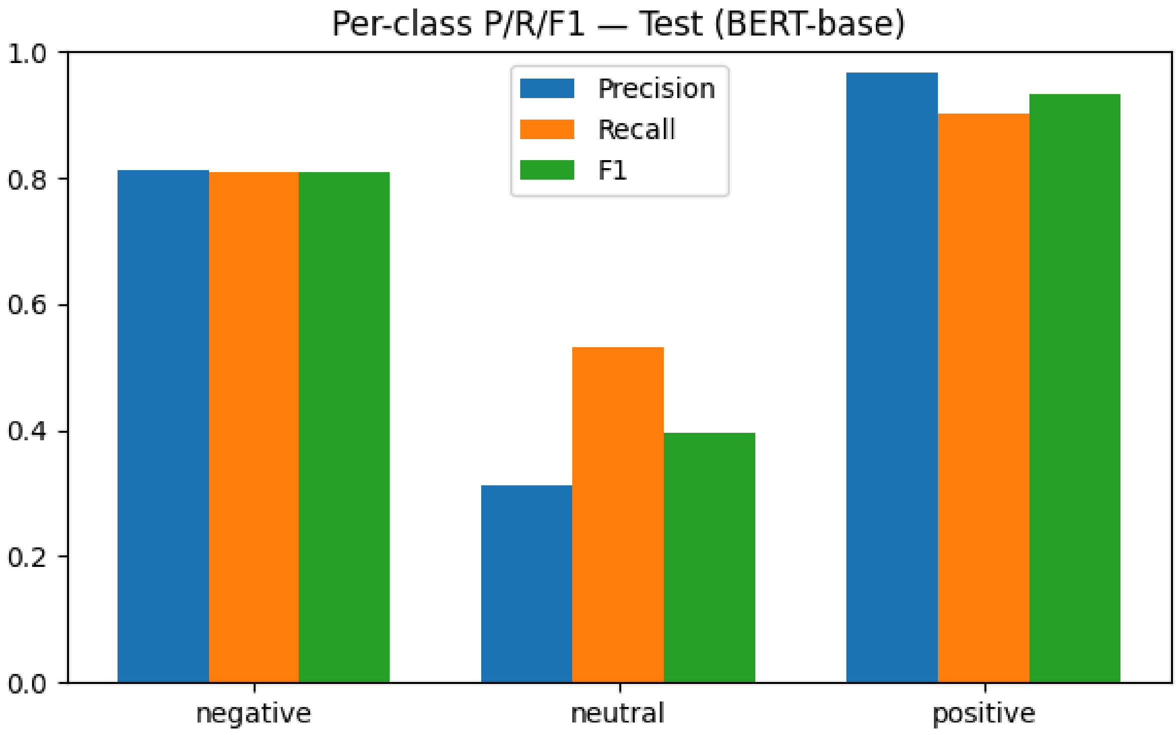
Bert-base-uncased

Scores:



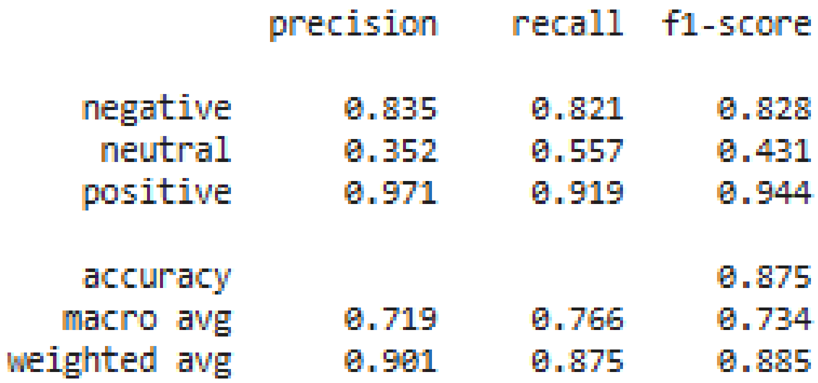
```
              precision    recall  f1-score

    negative      0.812     0.809     0.811
     neutral      0.313     0.532     0.395
    positive      0.969     0.903     0.934

    accuracy                          0.858
   macro avg      0.698     0.748     0.713
weighted avg      0.892     0.858     0.872
```


Confusion Matrix — Test (BERT-base)

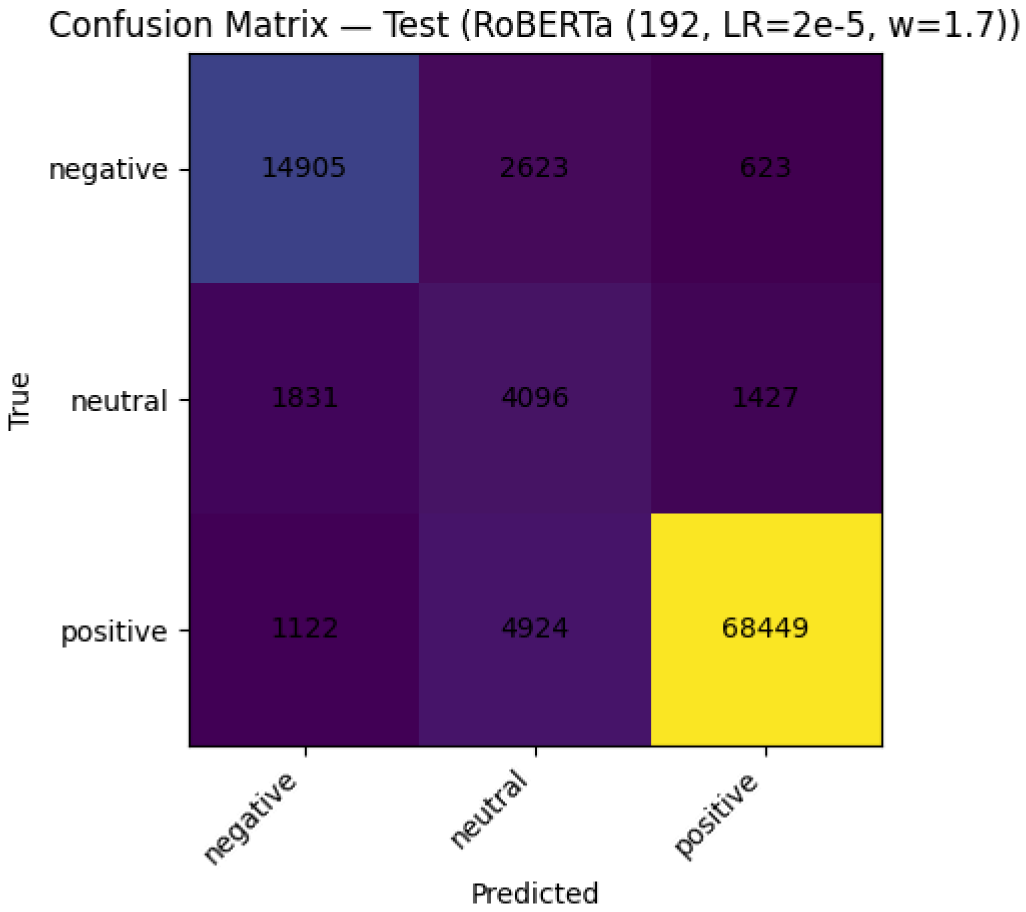

Per-class P/R/F1 — Test (BERT-base)

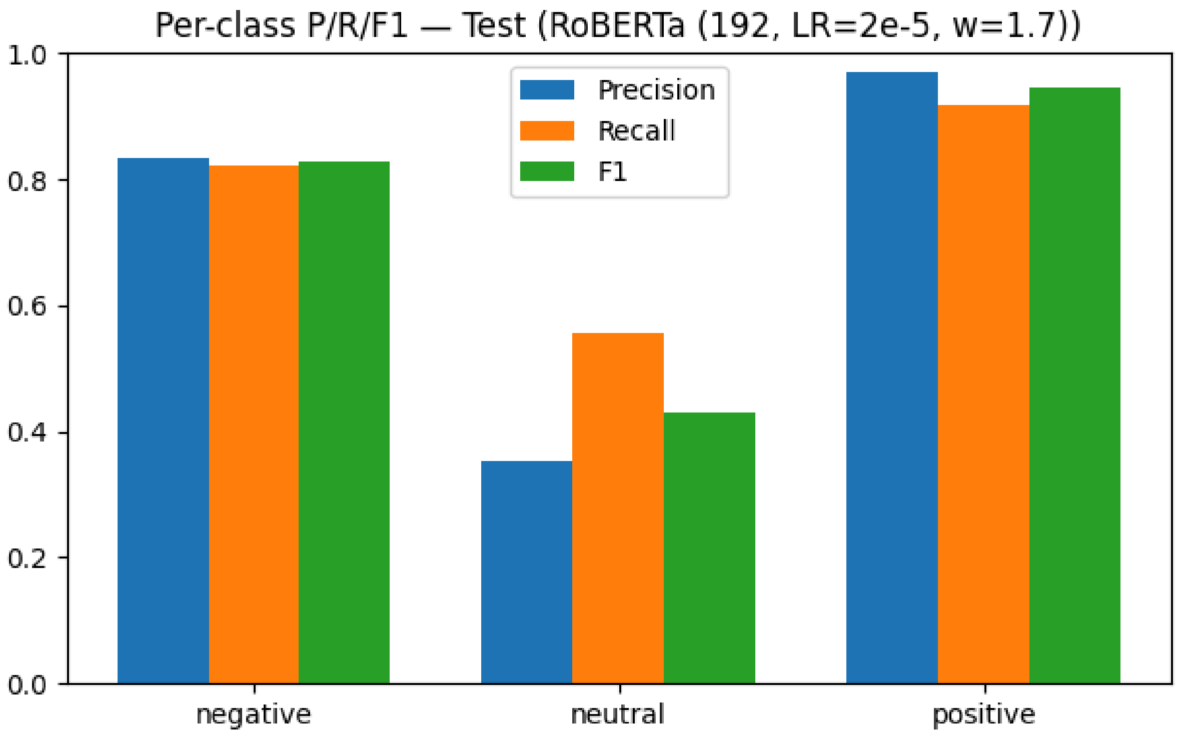
Roberta-base

Scores:

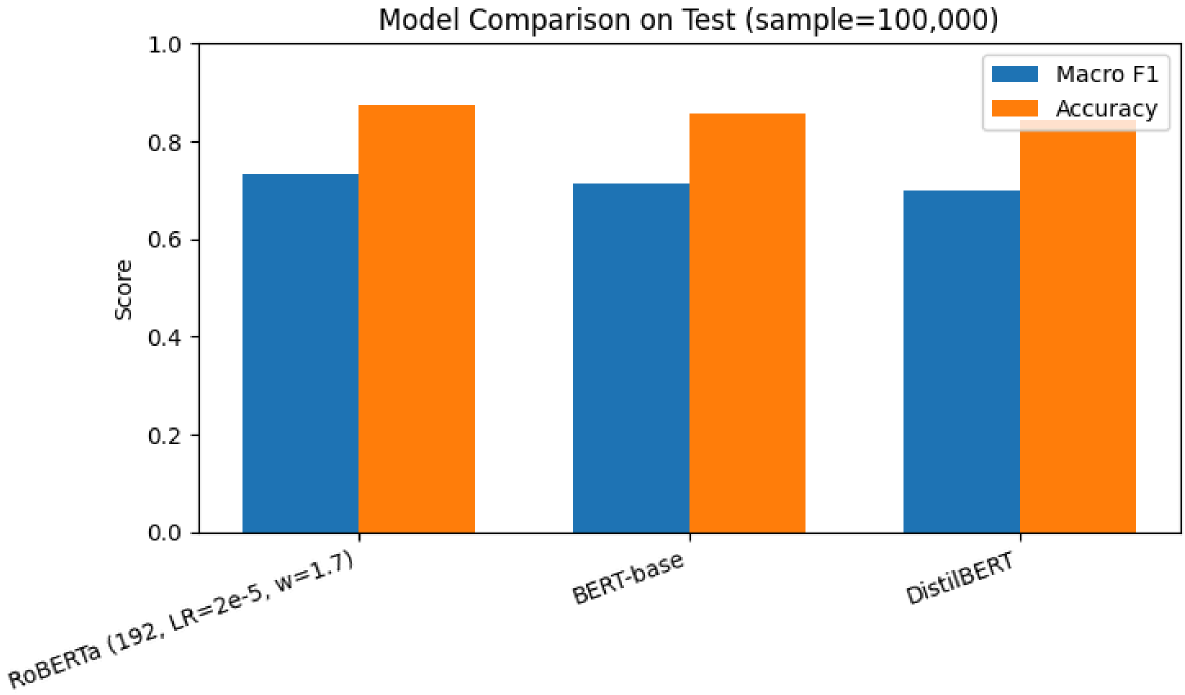


```
              precision    recall  f1-score

    negative      0.835     0.821     0.828
     neutral      0.352     0.557     0.431
    positive      0.971     0.919     0.944

    accuracy                          0.875
   macro avg      0.719     0.766     0.734
weighted avg      0.901     0.875     0.885
```





Model comparison:



Clustering:

The following 4 product category clusters were found:

- Keyboard and mice
- Headsets and audio
- Games
- Controllers

Summarization:

The best summarization result was found using GPT-4o-mini. Which produced a short article for each category stating the top 3 products and their flaws. The worst product and why it should be avoided. Full articles posted on the HuggingFace Space in the deployment.

Deployment:

See the deployment on:
https://huggingface.co/spaces/DaanBooy/games_and_accessories_reviews

---

4) Analysis

Sentiment Classification

- RoBERTa-base achieved the strongest overall results, with F1-scores of 0.944 (positive), 0.828 (negative), and 0.431 (neutral). Its macro average F1 was 0.734, outperforming the other models, confirming it was the most robust choice.

- BERT-base-uncased was not much worse with 0.934 (positive), 0.811 (negative), and 0.395 (neutral), for a macro F1 of 0.713.

- DistilBERT-base-uncased performed slightly worse than the others, with 0.925 (positive), 0.798 (negative), and 0.378 (neutral), reaching a macro F1 of 0.701.

Across all models, the neutral class remained the weakest (F1 between 0.37–0.43), reflecting the dataset imbalance and the difficulty of distinguishing neutral from weakly positive or negative reviews. Positive and negative classes did consistently achieve strong F1-scores above 0.80.

These results show that while all three models performed well on clear sentiment, RoBERTa provided the best balance across classes, making it the most suitable classifier for deployment. Improving neutral classification remains the main challenge and would likely benefit from additional data balancing or more tailored loss weighting.

Clustering

The MiniLM embeddings + PCA + MiniBatchKMeans (k=6) approach grouped reviews efficiently. After merging two sentiment-driven clusters, we obtained 4 clear product categories:

- Keyboards & Mice
- Headsets & Audio
- Games
- Controllers

The clusters align well with natural product groupings, though the initial emergence of sentiment-based clusters highlights that embeddings capture both category and opinion information.

Summarization

For each cluster, GPT-4o-mini generated well-structured articles that clearly highlighted the top 3 products, their main complaints, and the worst product. The outputs were fluent, coherent, and suitable for direct use in recommendation-style summaries. By contrast, BART and FLAN-T5 performed significantly worse: while they occasionally produced usable fragments, their outputs were often prompt repeats or random in coherent sentences. GPT-4o-mini was the most reliable and effective summarization model for this project.

Overall

- RoBERTa-base proved to be the strongest sentiment classifier, with the highest macro F1 (0.734) and consistently strong performance on positive (0.944) and negative (0.828) reviews.

- Neutral sentiment remained the hardest to classify, with F1-scores only 0.37–0.43 across all models, highlighting the need for better handling of class imbalance and ambiguous reviews.

- Clustering produced 4 clear product categories (Keyboards & Mice, Headsets & Audio, Games, Controllers), aligning well with the dataset's structure.

- GPT-4o-mini generated the best summaries and articles