NLP | Automated Customer Reviews

Hessa-Shatha-Ghadah

overview of the project:

- Data reprossesing
- Larger Data
- Sentiment Classification
- clustring
- Web App Deployment

PROJECT GOALS

Classify customer reviews as positive, negative, or neutral.

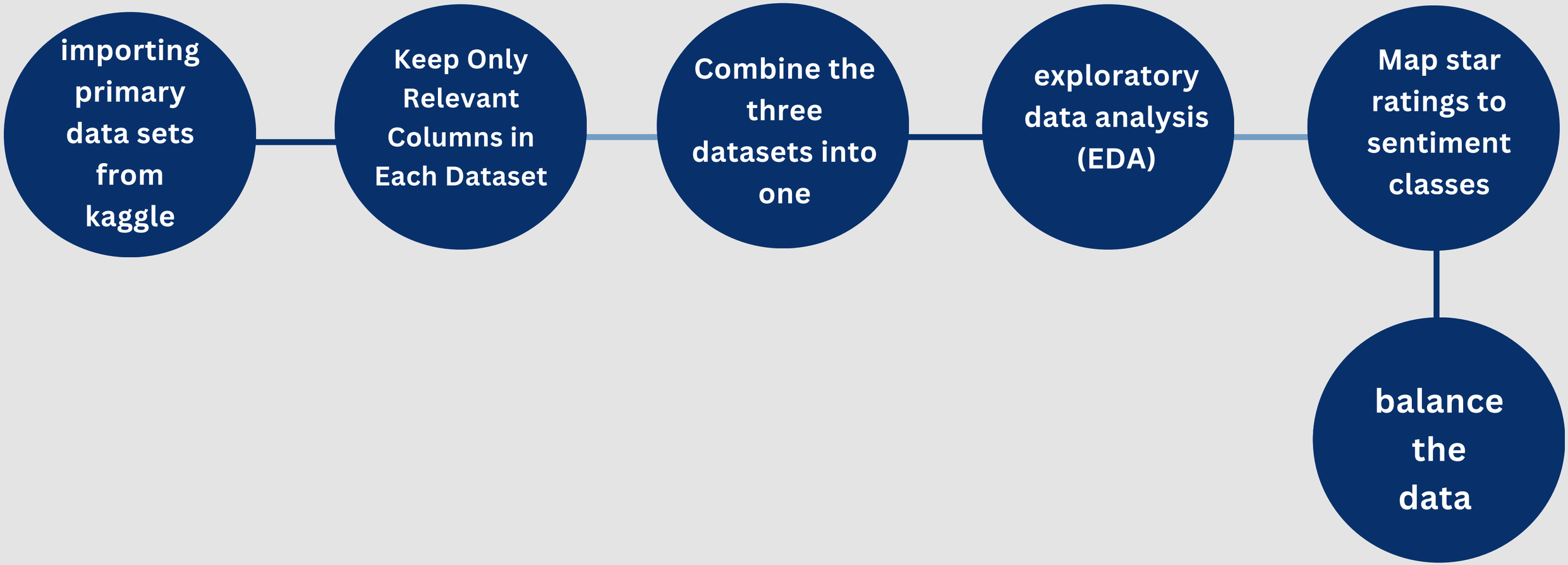
Cluster reviews into product categories

Generate blog-style summaries using Generative AI.

Deploy the system as a user-friendly web app

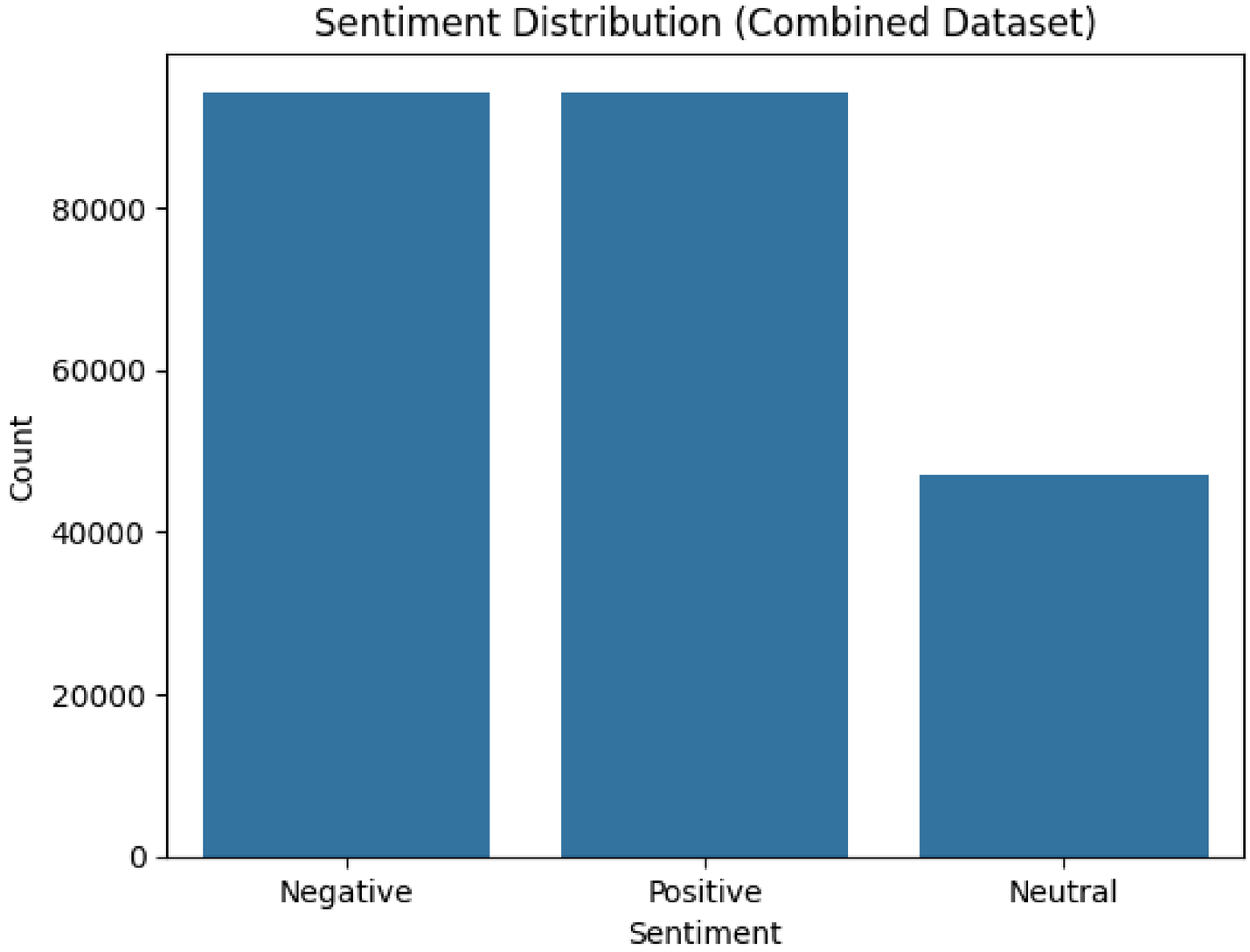
Data explore & Preprocessing

primary Data Processing:

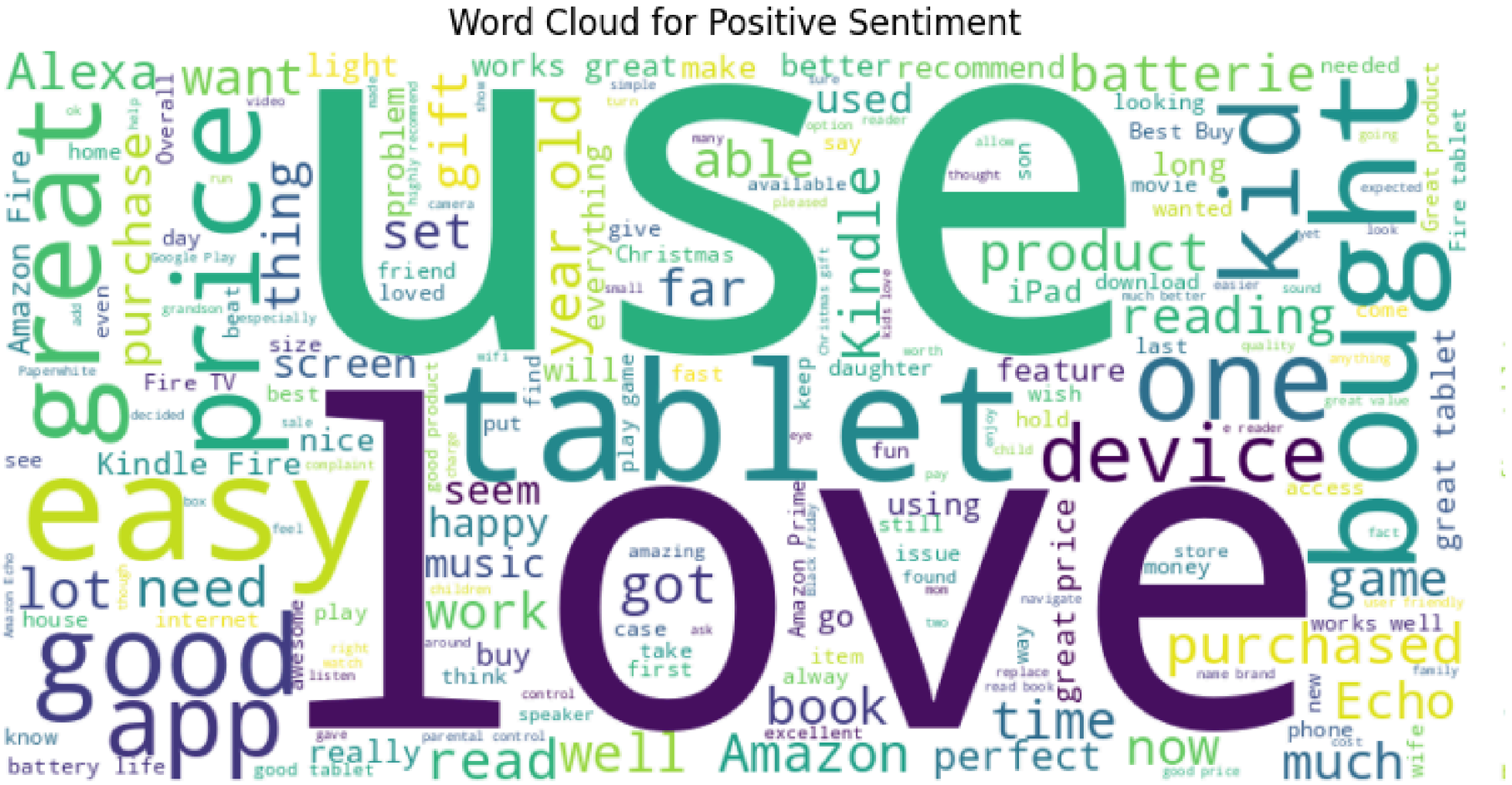


Correlation Matrix :

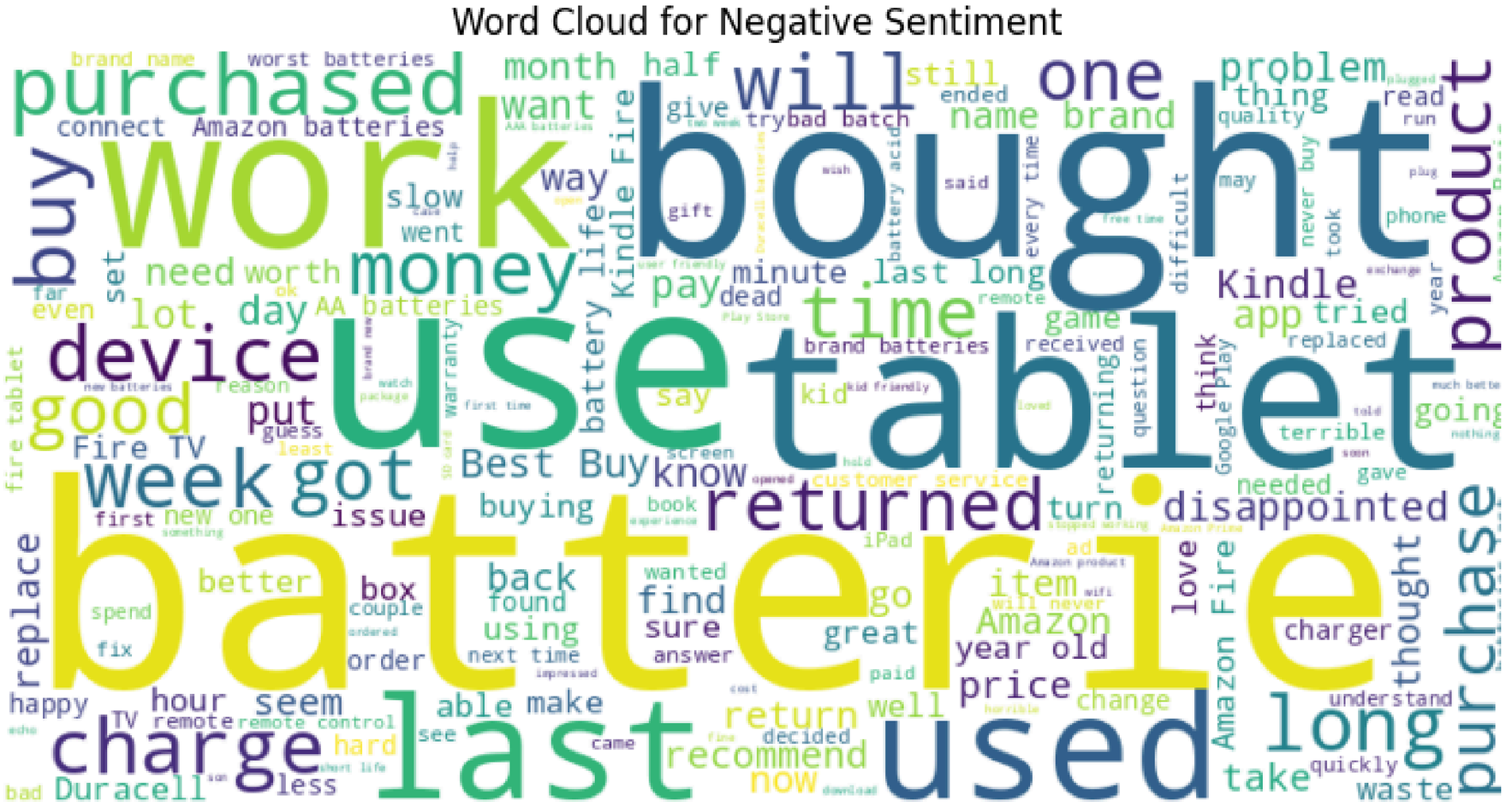
explore sentiment:



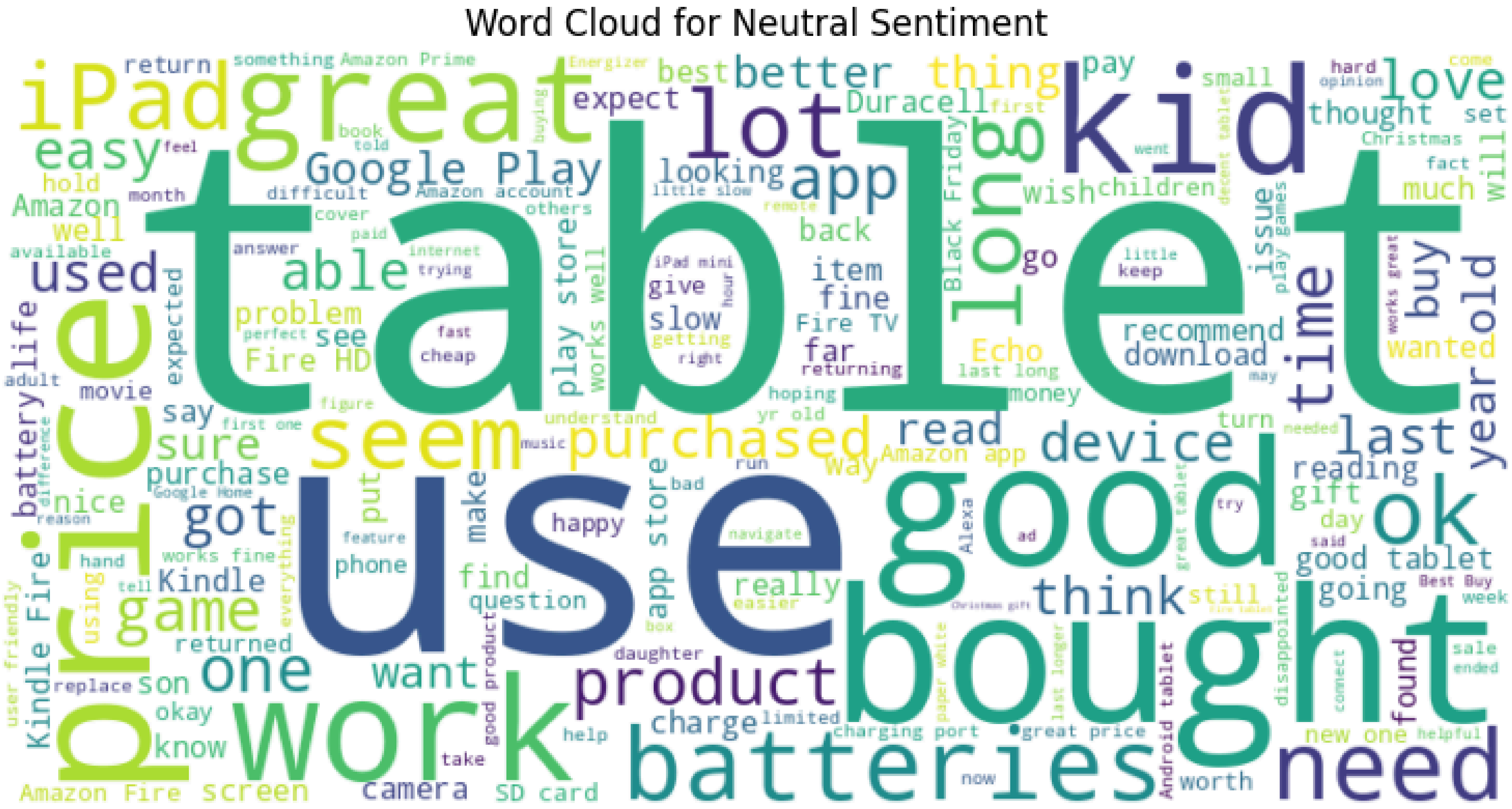
word cloud(positive):



word cloud(negative):

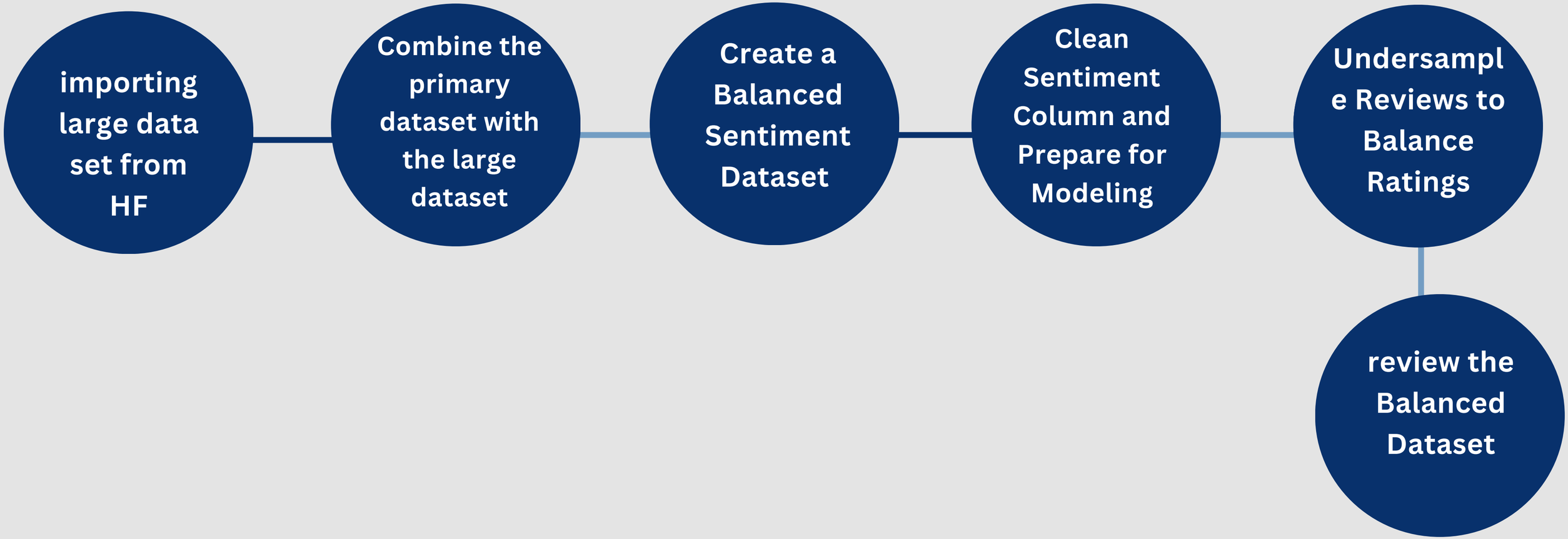


word cloud(neutral):

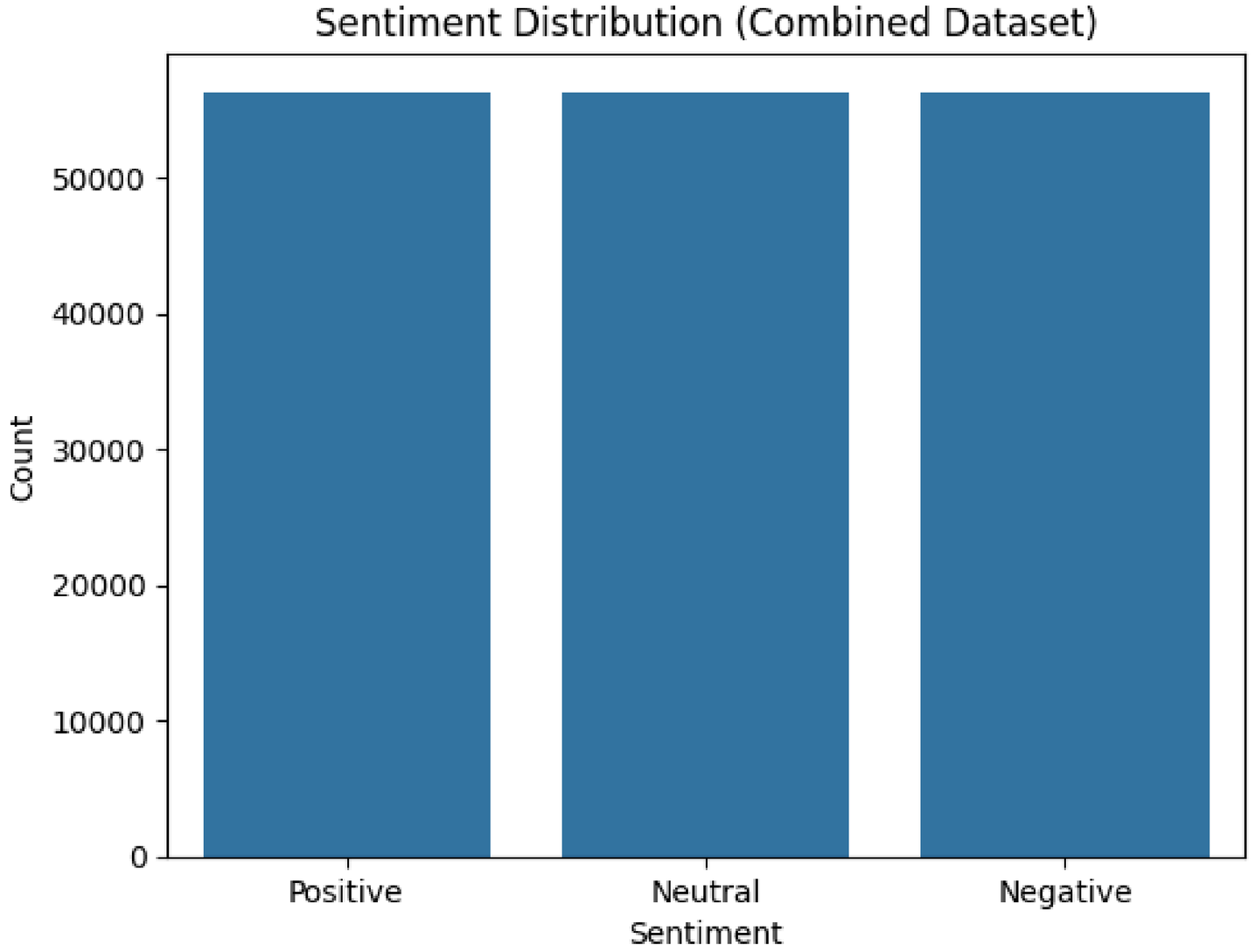


Larger Data

primary Data Processing:



explore sentiment:



Sentiment Classification :

Sentiment Classification: Model Comparison

This section presents a comparative analysis between two transformer-based models used for sentiment classification on Amazon product reviews:

- Model A: DistilRoBERTa (uses the **larger** Amazon Reviews Dataset)
- Model B: distilroberta-base (uses the **Primary** Dataset: Amazon Product Reviews)

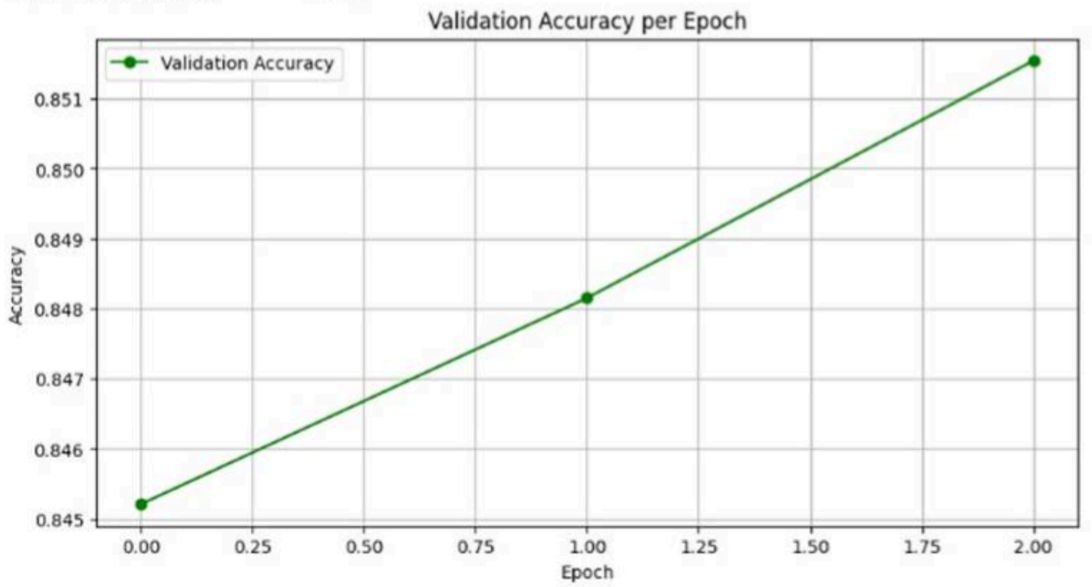
-The comparison focuses on each model's efficiency and accuracy, based on the size and diversity of the dataset.

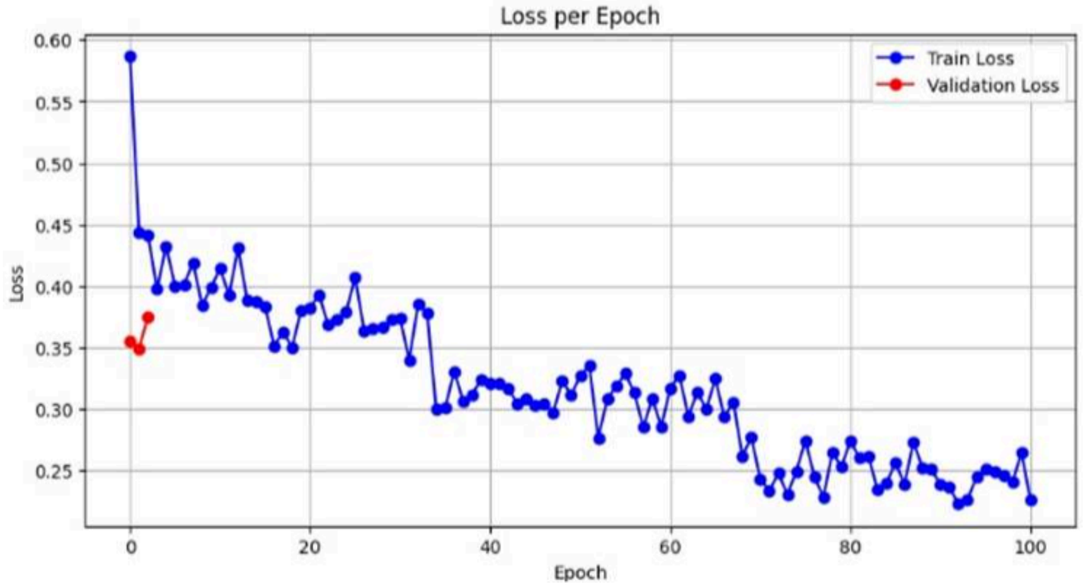
Model A: DistilRoBERTa

DistilRoBERTa is a smaller and faster update of the RoBERTa model. It still gives good results. It was trained on a big Amazon reviews dataset, which includes many product types and a lot of labeled reviews.

Because it saw more different examples during training, this model can understand new or unbalanced data better. It is expected to work well even with data it has not seen before.

Model A: DistilRoBERTa –Performance:



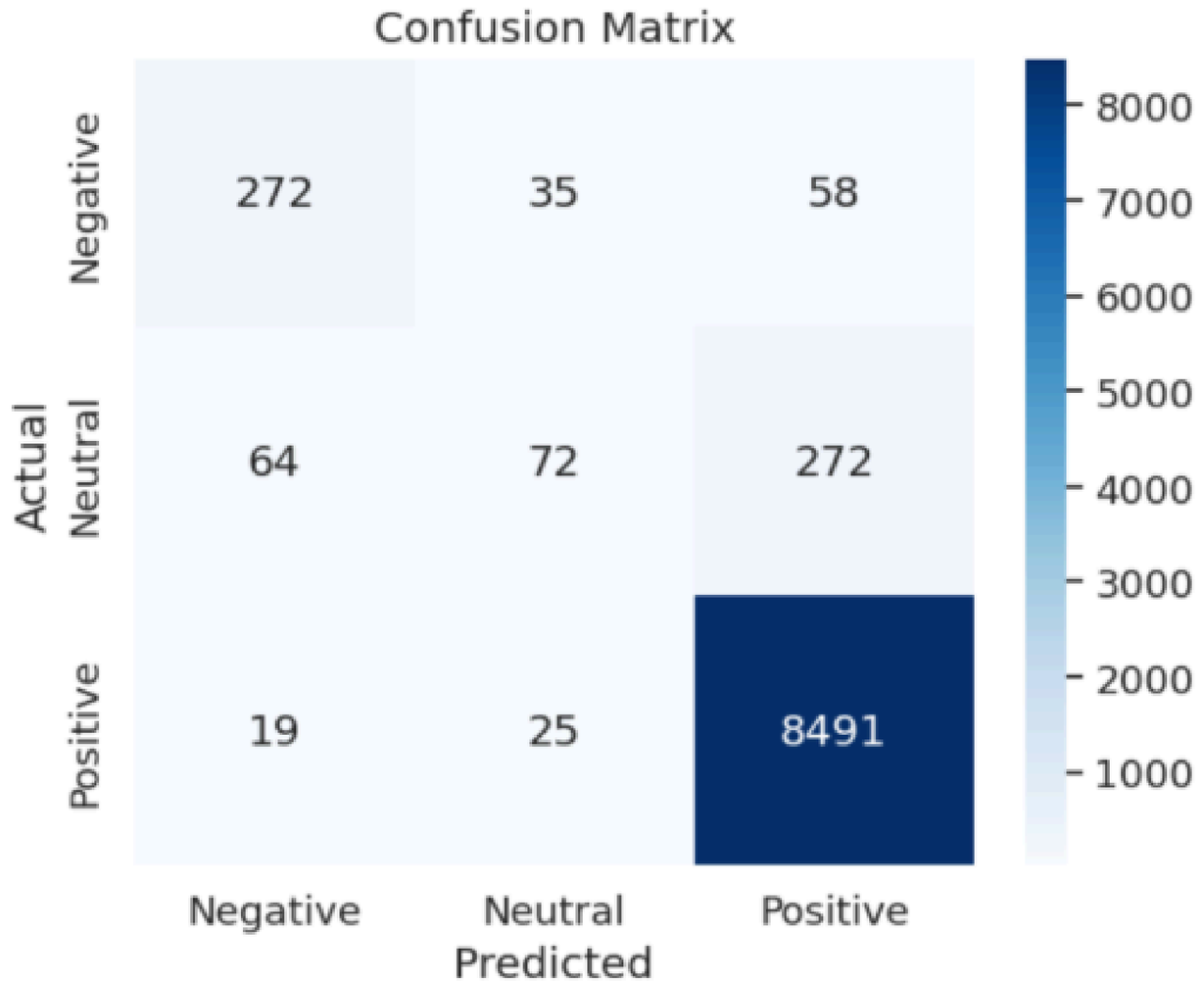


Model B: distilroberta-base

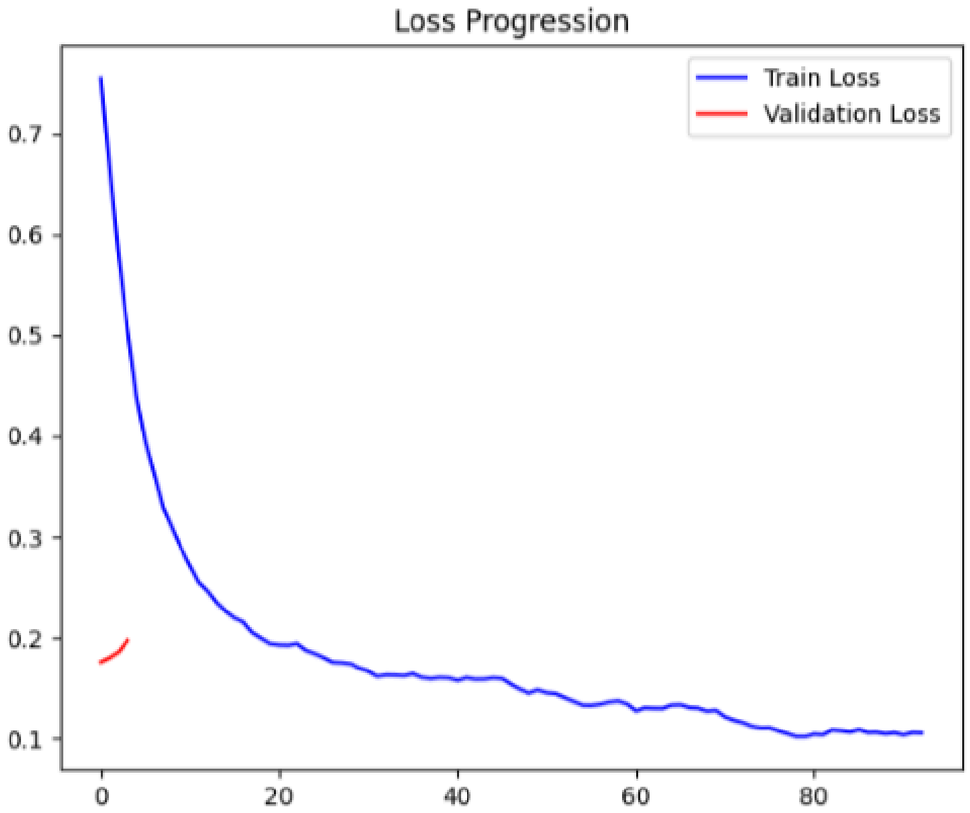
- This model was trained on a smaller Amazon reviews dataset. It has three main files with customer reviews. Even though it’s smaller, the data is more focused and well-organized. This helps the model learn better from clear and good-quality data.

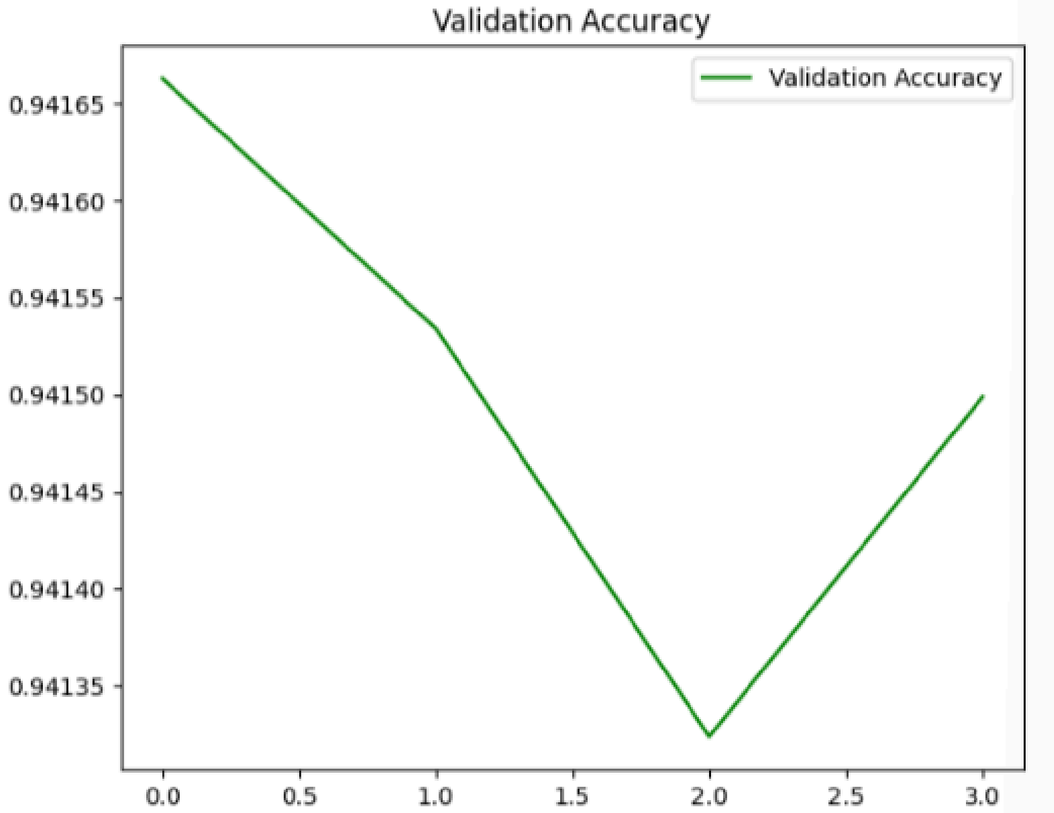
- This model is good for tasks where we want to understand feelings (sentiment) in a specific topic, not general ones.

Model A: distilroberta Confusion Matrix:

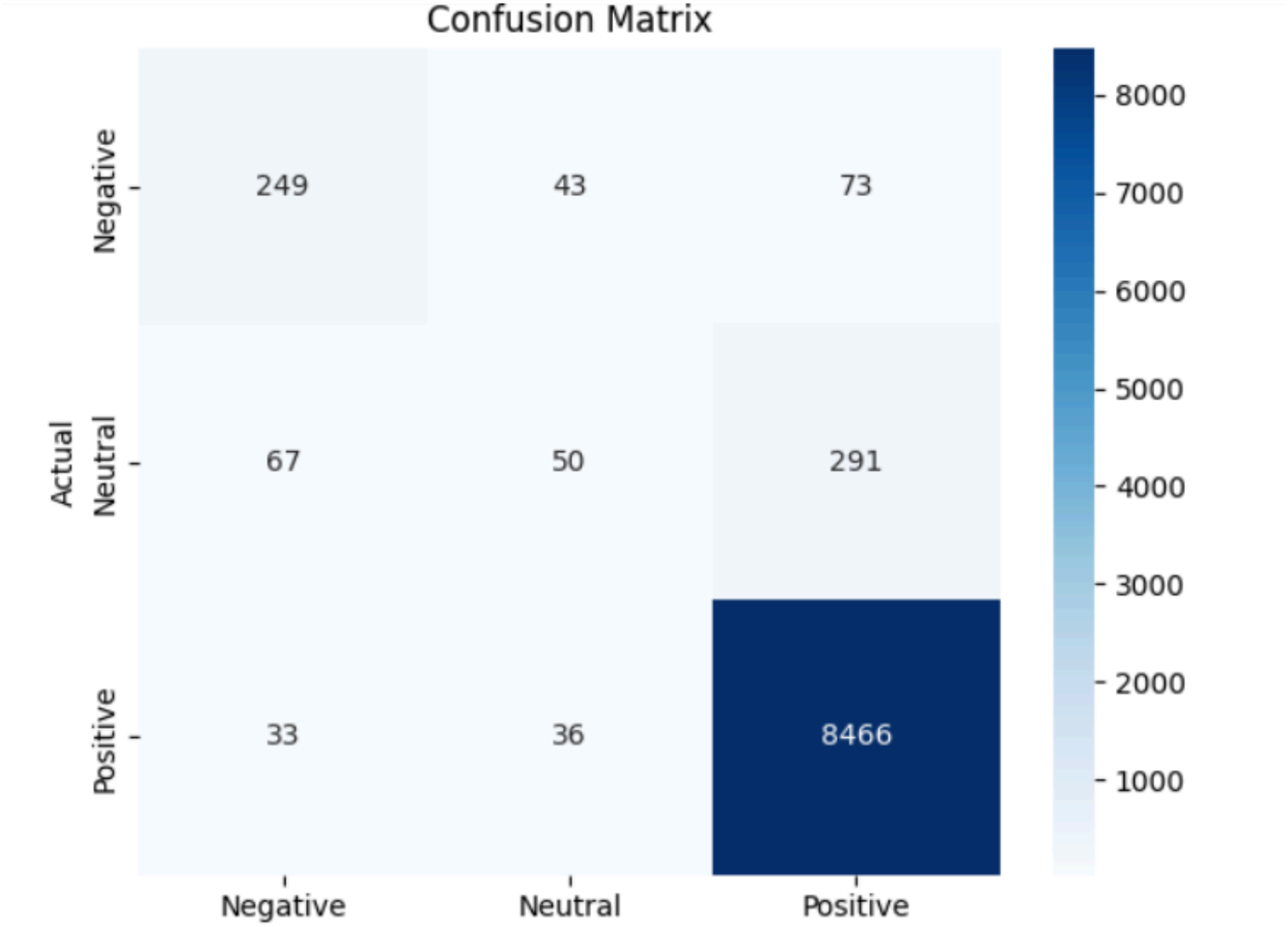


Model B: distilroberta-base-Performance:





Model B: distilroberta-base Confusion Matrix:



Model Comparison Table:

Metric	Model A: DistilRoBERTa uses the larger Dataset	Model B: distilroberta-base uses the Primary Dataset
Accuracy	85.15%	94%
Precision	Negative 81% Neutral 65% Positive 98%	Negative 71% Neutral 39% Positive 96%
Recall	Negative 85% Neutral 58% Positive 99%	Negative 68% Neutral 12% Positive 99%
F1-Score	Negative 83% Neutral 61% Positive 98%	Negative 70% Neutral 19% Positive 98%
Training Time	Longer	Faster

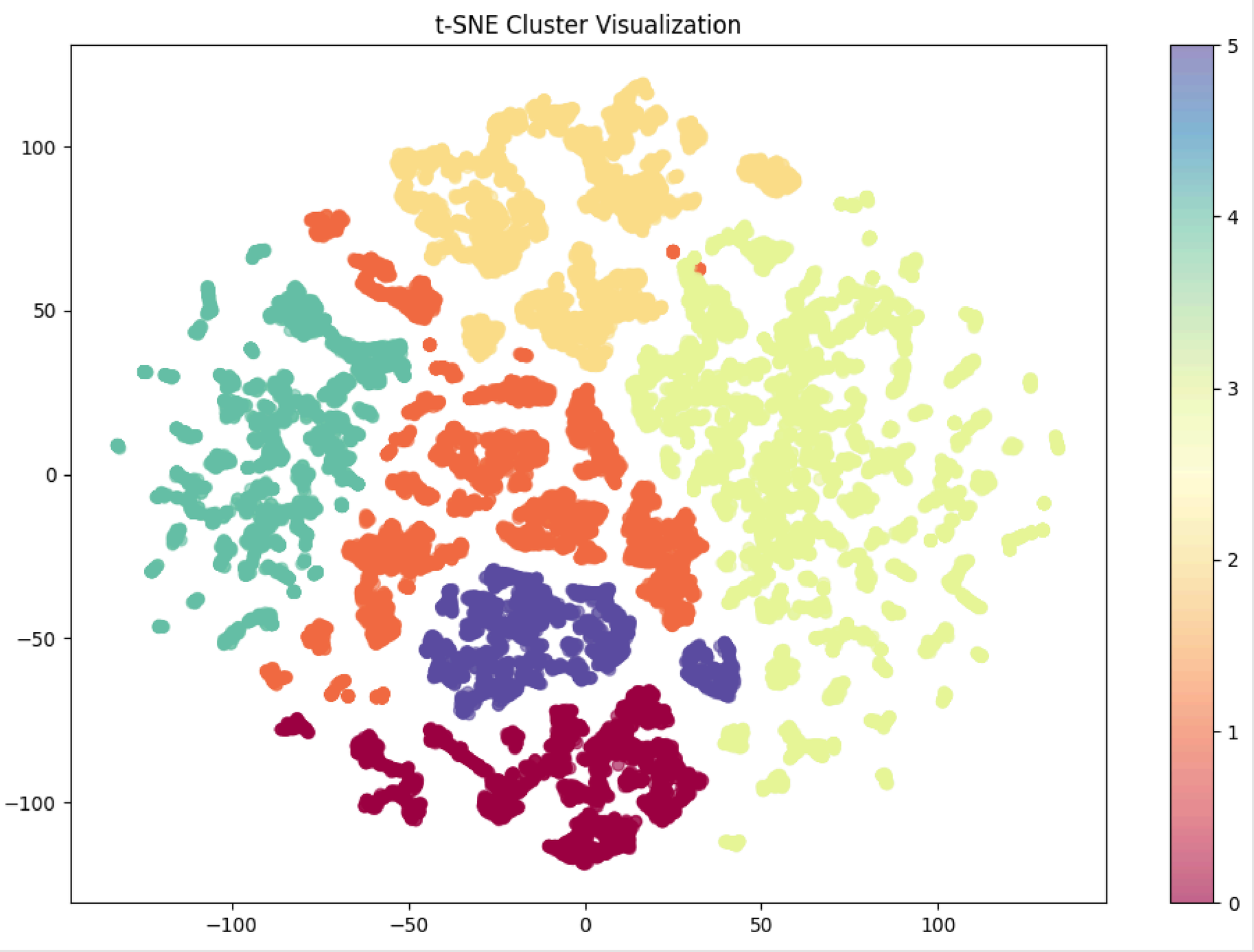
CLUSTRING

clustring

to simplify product categorization, we built a clustering model that groups detailed product types into 4–6 broader meta-categories. The process involved:

- Data Cleaning & Preprocessing: Removed noise and extracted meaningful phrases using spaCy.
- Embedding Generation: Used the MiniLM sentence transformer to convert text into numerical vectors.
- Dimensionality Reduction: Applied UMAP to reduce embedding size for faster and more efficient clustering.
- Clustering: Used K-Means, tested multiple cluster counts (4–6), and evaluated results using silhouette and Davies-Bouldin scores. A refinement step improved cluster quality.
- Visualization & Analysis: Visualized clusters using t-SNE and extracted top keywords per cluster while ignoring generic terms.

This model effectively groups products by semantic similarity, supporting better search, recommendations, and category management.


t-SNE Cluster Visualization
100
50
0
−50
−100
−100
−50
0
50
100
5
4
3
2
1
0

clustring

```
Top meaningful words in Cluster
0
['tablet', 'display', 'offer', 'wifi', 'special_offer', 'include', 'display_wifi', 'hd_display', 'tabletsamazon', 'tabletscomputer']
Top meaningful words in Cluster
1
['tablet', 'display', 'offer', 'special_offer', 'tabletsamazon', 'kindle', 'home', 'alexa', 'ebook', 'offer_fire']
Top meaningful words in Cluster
2
['tablet', 'kid', 'case', 'edition', 'display', 'kidproof', 'fire_kid', 'wifi', 'display_wifi', 'kidproof_case']
Top meaningful words in Cluster
3
['aaa', 'alkaline', 'performance', 'alkaline_battery', 'amazonbasics', 'household', 'count', 'baby', 'care', 'personal']
Top meaningful words in Cluster
4
['performance', 'packaging', 'alkaline', 'amazonbasics', 'household', 'baby', 'count', 'care', 'householdcamcorder', 'batteriescamera']
Top meaningful words in Cluster
5
['tablet', 'display', 'offer', 'alexa', 'tangerine', 'special_offer', 'tabletstabletsall', 'tabletsamazon', 'gb_tangerine', 'offer_fire']
```

Summarization Model:

deployment :

Deployment

🧩 Environment

- Python 3.10 with virtual environment
- Key libs: transformers, sentence-transformers, scikit-learn, pandas, torch, gradio, openai, nltk

💻 Local Setup

- Deployed with Gradio
- Run via: python app.py
- UI includes:
  - Classification tab
  - Clustering tab
  - GPT-4 summarizer tab

🌐 API & Files

- OpenAI GPT-4 via secure API key
- Supports .csv and .parquet file uploads

DEPLOYMENT :

THANK YOU !