SAUDI DIGITAL ACADEMY

Himah Digital Bootcamps - AI Bootcamp

Business Case

Automated Customer Reviews Analysis using AI

2025

Represent By

Team Members

- AHMED ALQARNI
- AMAL ALGHTANI
- HANAN ALNBHANI

Group Number: 1

AGENDA

1. Introduction
2. Data Understanding
3. Data Preprocessing
4. Review Classification
5. Model Evaluation
6. Product Category Clustering
7. Review Summarization
8. Deployment
9. Challenges & Solutions
10. Team Organization
11. Q&A Session

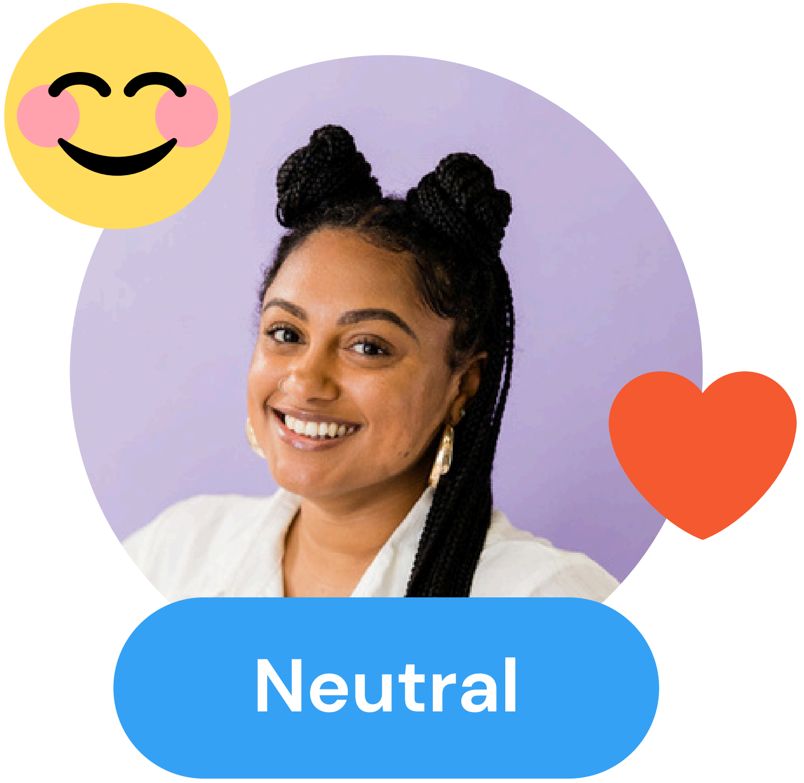
Introduction

Natural Language Processing (NLP): is a field of AI that enables machines to analyze and understand human language like sentiment analysis and text classification.

Main Objective of the Project

Problem: There are thousands of customer reviews online, and analyzing them manually is inefficient.

Goal: To build a system using (NLP) to classify, cluster, and summarize reviews.



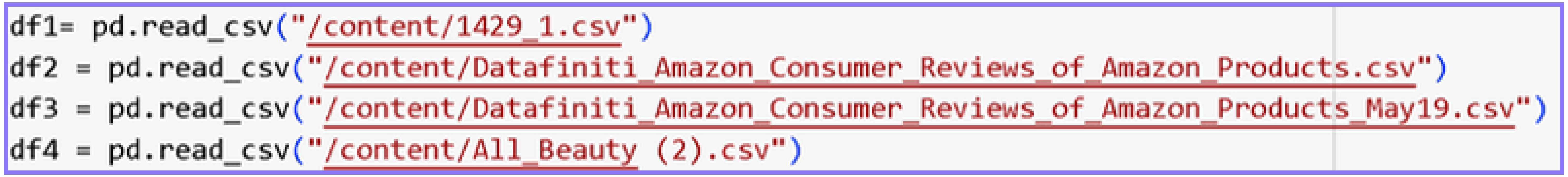
Understand Dataset

- PRIMARY DATASET: AMAZON PRODUCT REVIEWS
- LARGER DATASET: AMAZON REVIEWS DATASET

Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products_May19

Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products

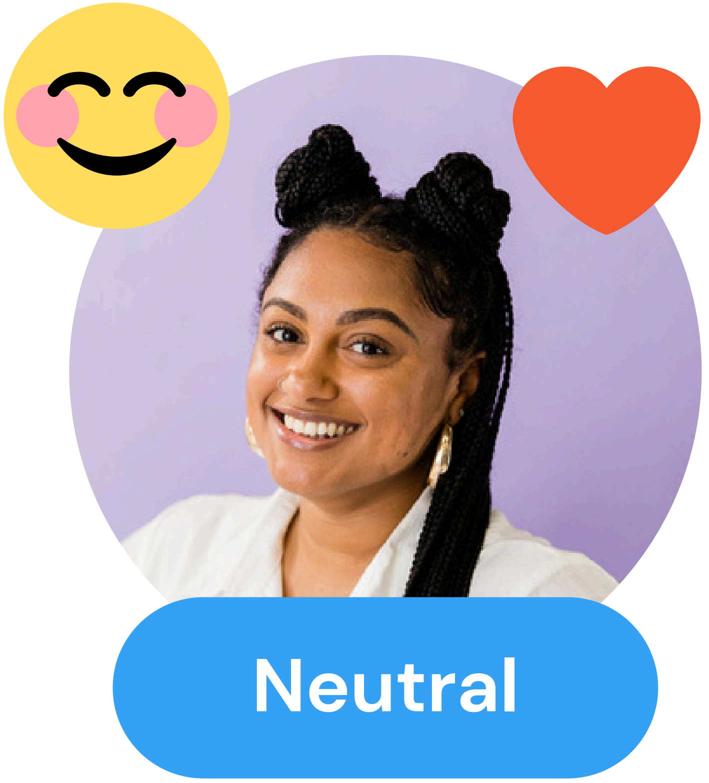
1429_1



```
df1= pd.read_csv("/content/1429_1.csv")
df2 = pd.read_csv("/content/Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products.csv")
df3 = pd.read_csv("/content/Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products_May19.csv")
df4 = pd.read_csv("/content/All_Beauty (2).csv")
```

1. REVIEW CLASSIFICATION

Positive

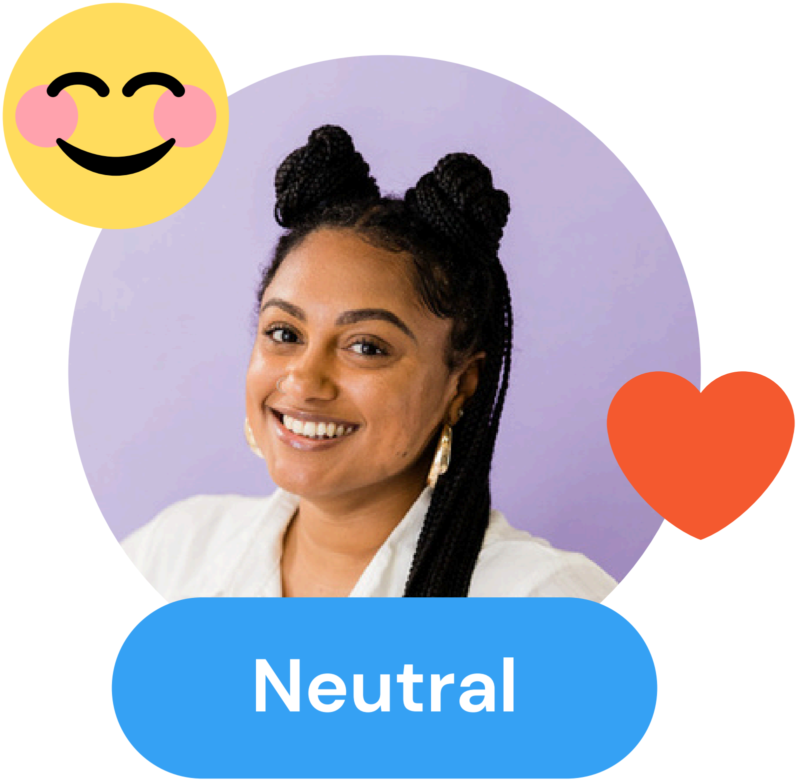


Neutral

1. REVIEW CLASSIFICATION

Preprocessing Steps:

- Removing stopwords and irrelevant tokens
- Filtering out symbols and special characters
- Handling missing or incomplete ratings
- Converting star ratings to sentiment classes:
- ⭐ 1–2 → Negative
- ⭐ 3 → Neutral
- ⭐ 4–5 → Positive



Data Preparation & Pre-processing

1

Merge The Data

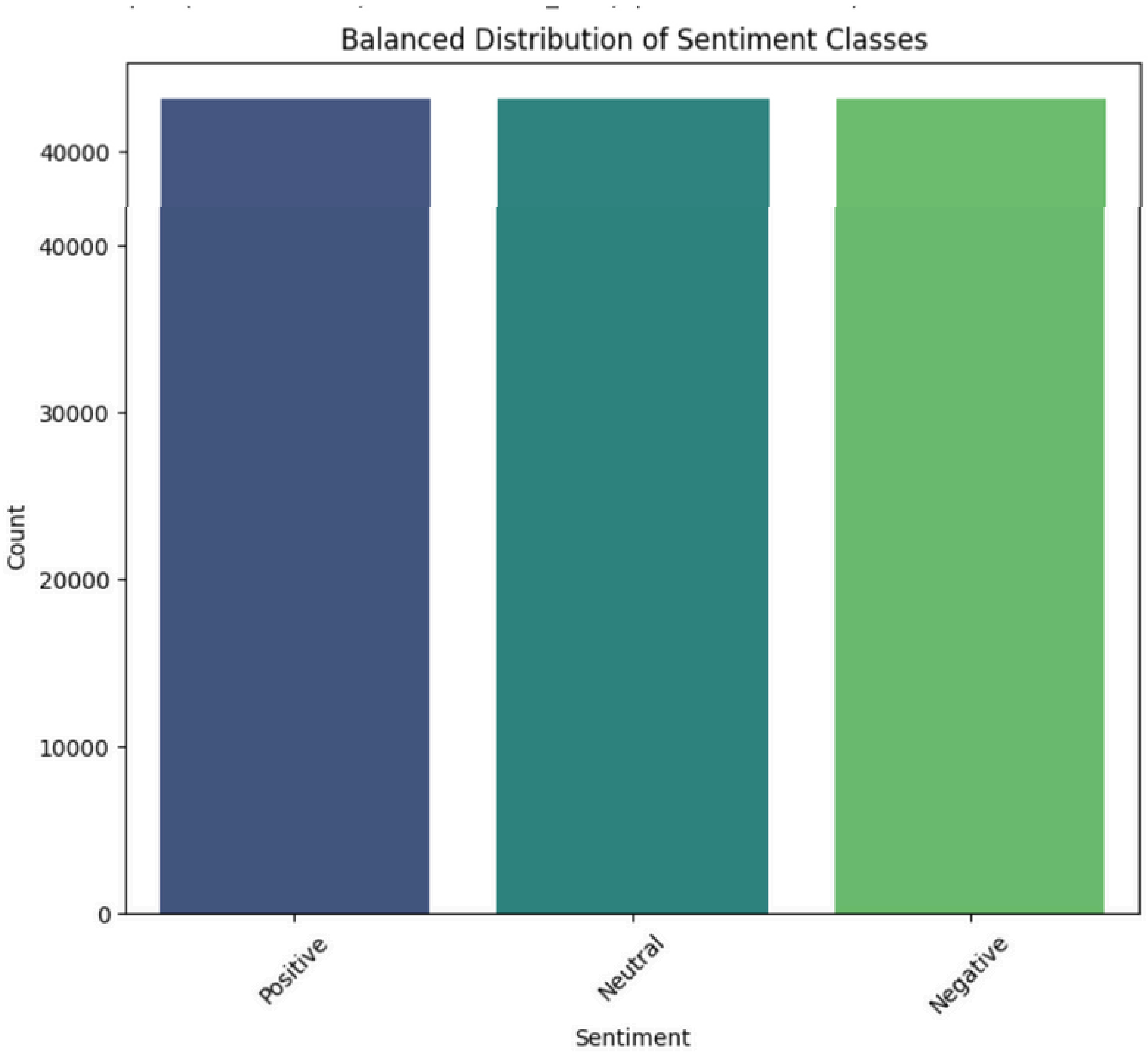
2

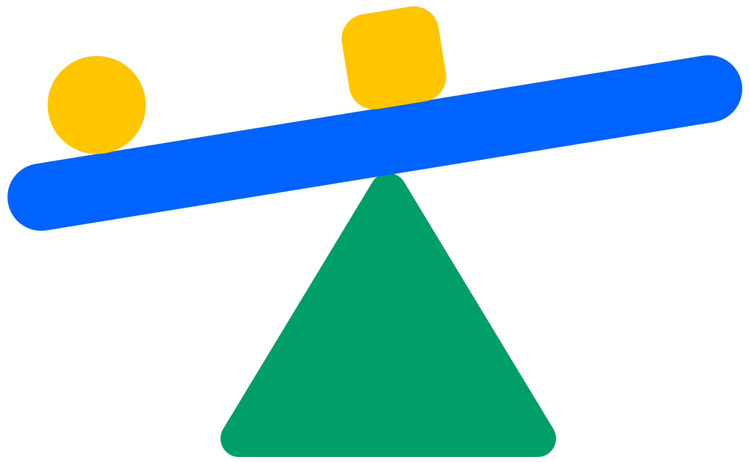
Keep only the text
and rating columns

3

Missing values
duplicates
Text cleaning

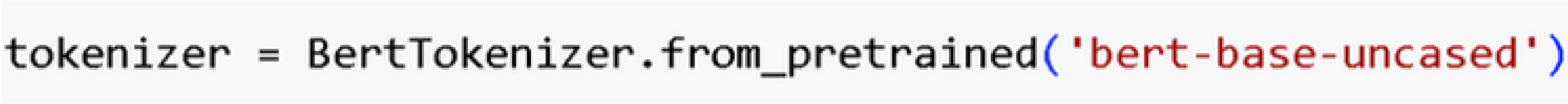
Balance categories with the highest number of positive reviews



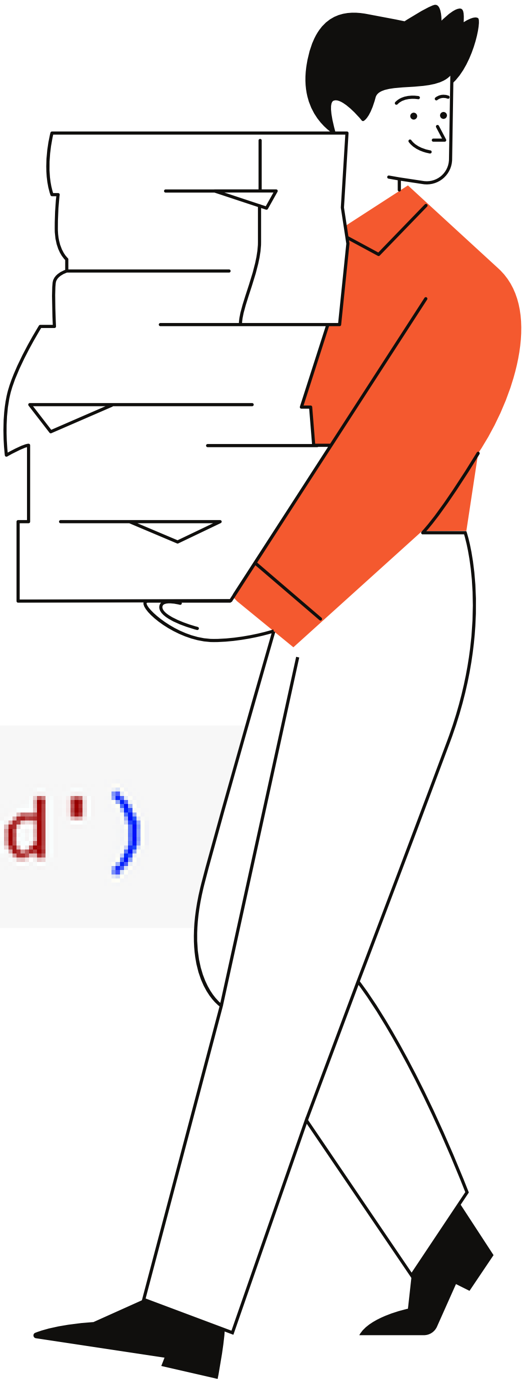


Model Training

Split data into features (X) and targets (y)

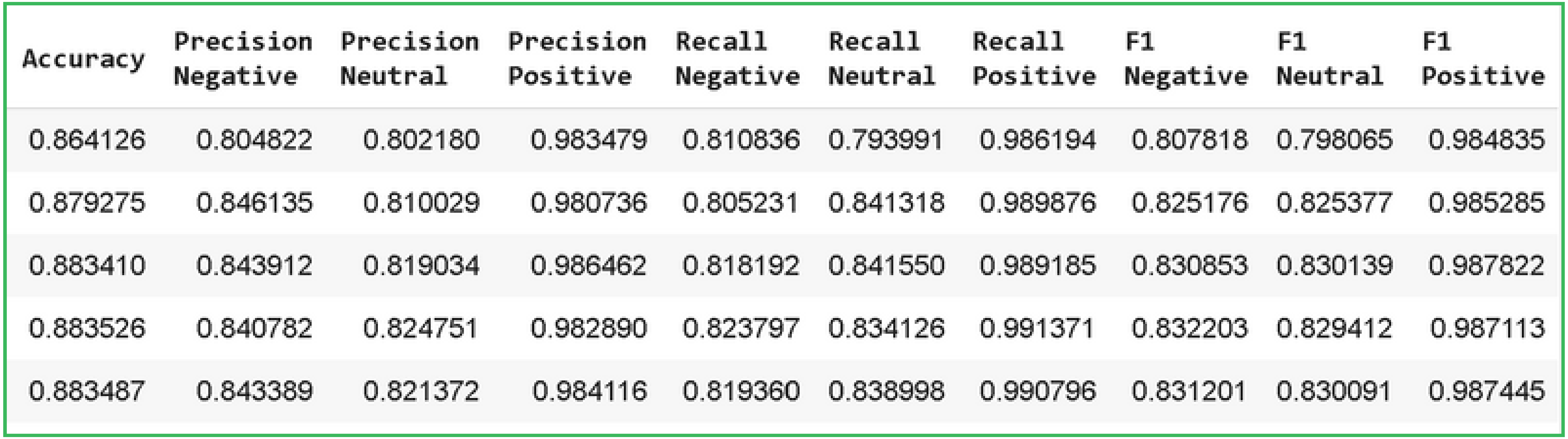


```
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
```



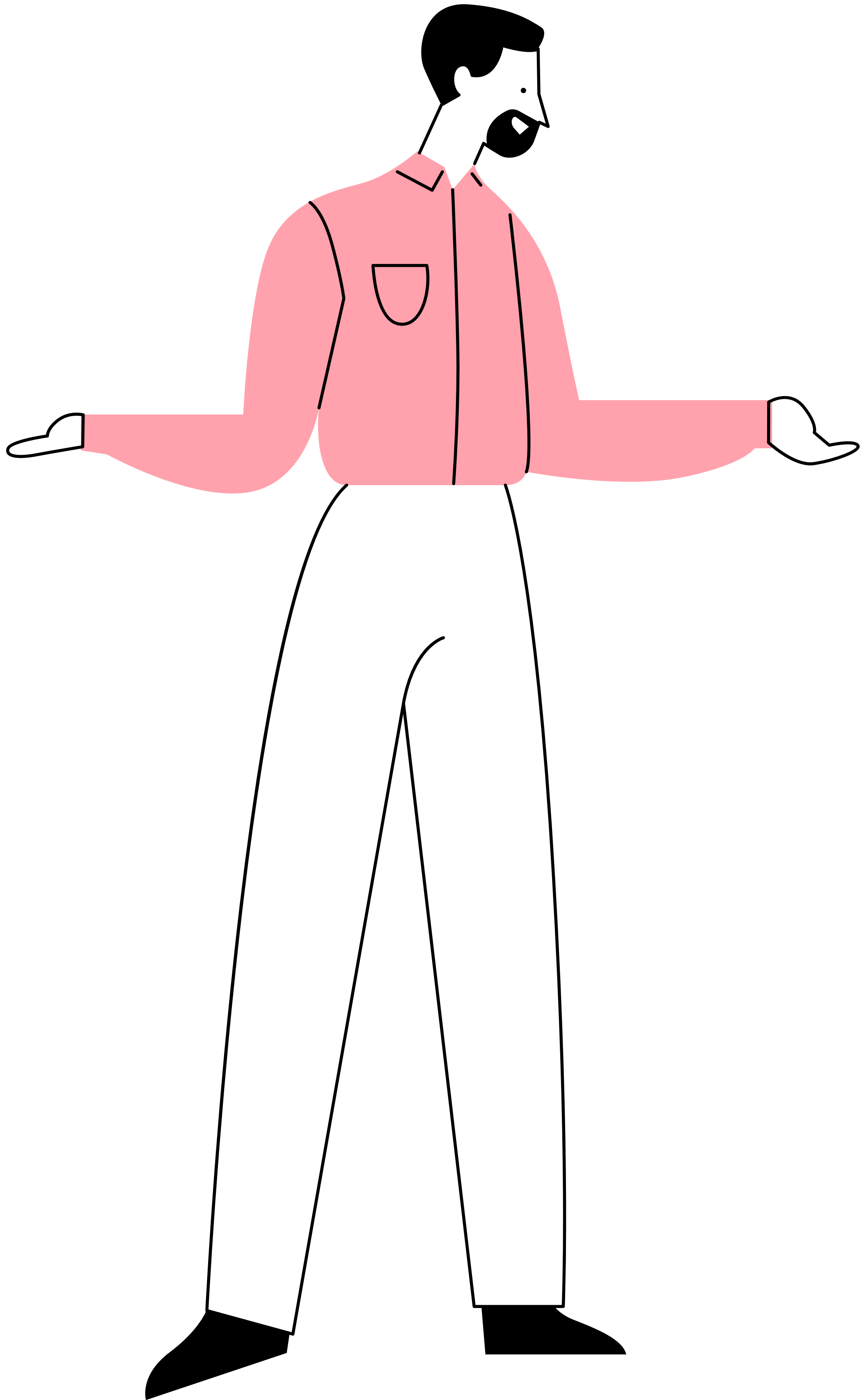
Models Evaluation

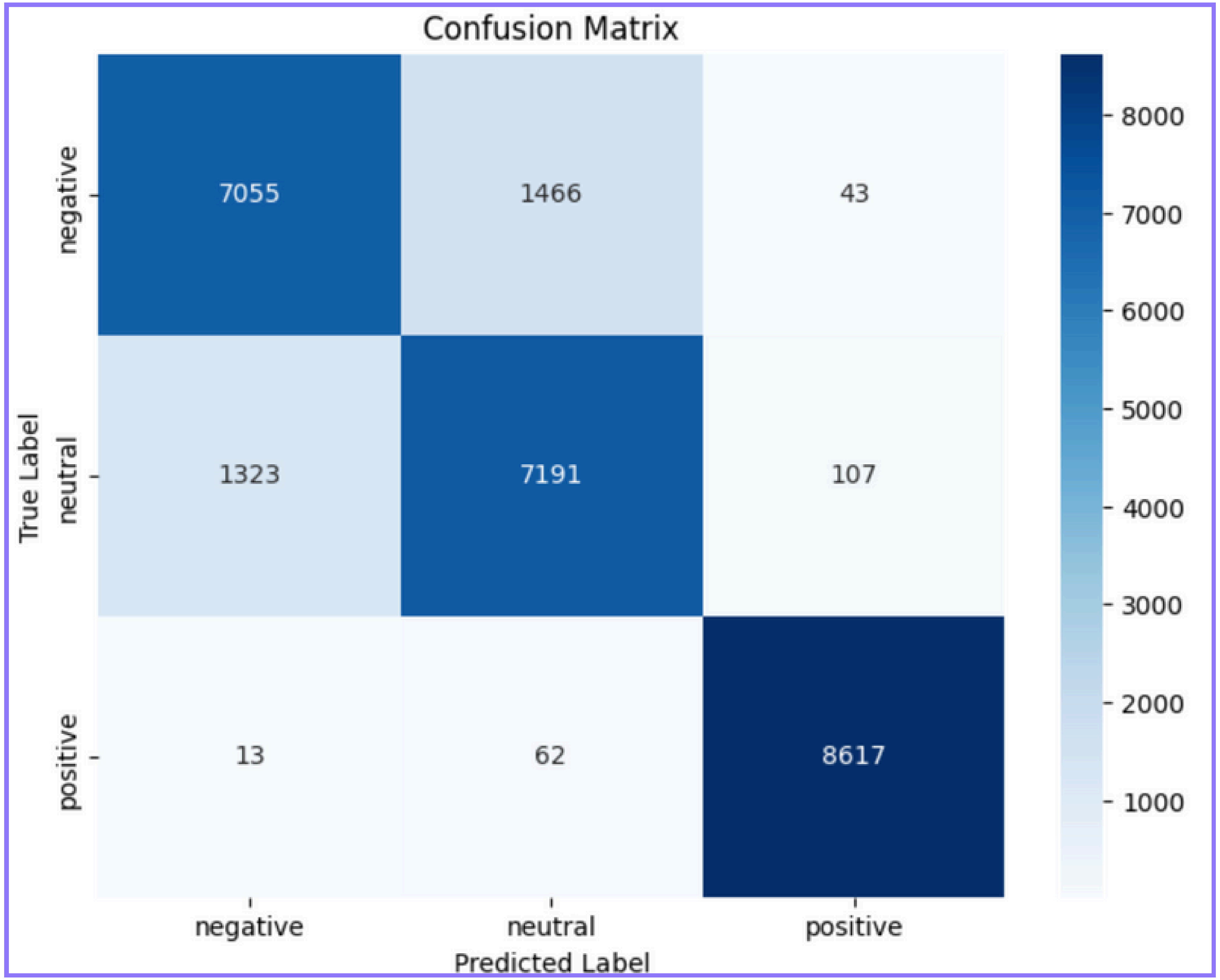
Calculate the scales for each category



Accuracy	Precision Negative	Precision Neutral	Precision Positive	Recall Negative	Recall Neutral	Recall Positive	F1 Negative	F1 Neutral	F1 Positive
0.864126	0.804822	0.802180	0.983479	0.810836	0.793991	0.986194	0.807818	0.798065	0.984835
0.879275	0.846135	0.810029	0.980736	0.805231	0.841318	0.989876	0.825176	0.825377	0.985285
0.883410	0.843912	0.819034	0.986462	0.818192	0.841550	0.989185	0.830853	0.830139	0.987822
0.883526	0.840782	0.824751	0.982890	0.823797	0.834126	0.991371	0.832203	0.829412	0.987113
0.883487	0.843389	0.821372	0.984116	0.819360	0.838998	0.990796	0.831201	0.830091	0.987445

Confusion matrix

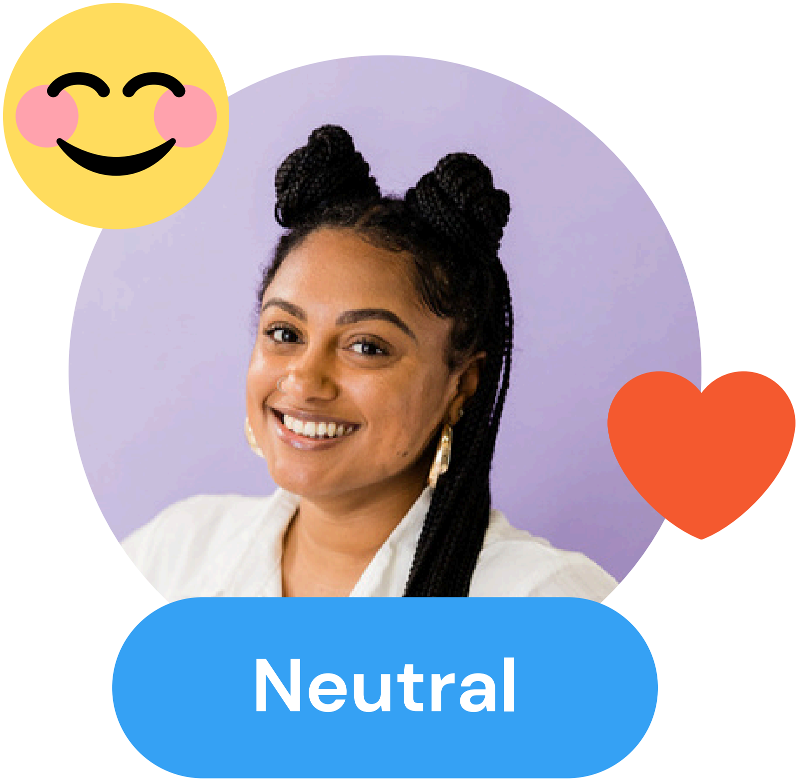




Confusion Matrix

True Label \ Predicted Label	negative	neutral	positive
negative	7055	1466	43
neutral	1323	7191	107
positive	13	62	8617

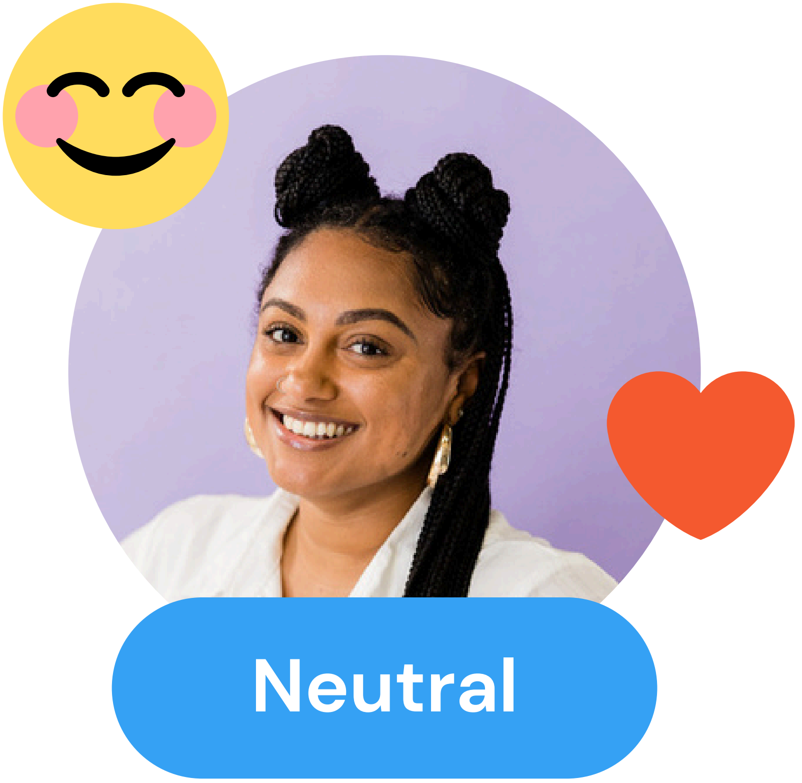
2. PRODUCT CATEGORY CLUSTERING



2. PRODUCT CATEGORY CLUSTERING

Preprocessing Steps:

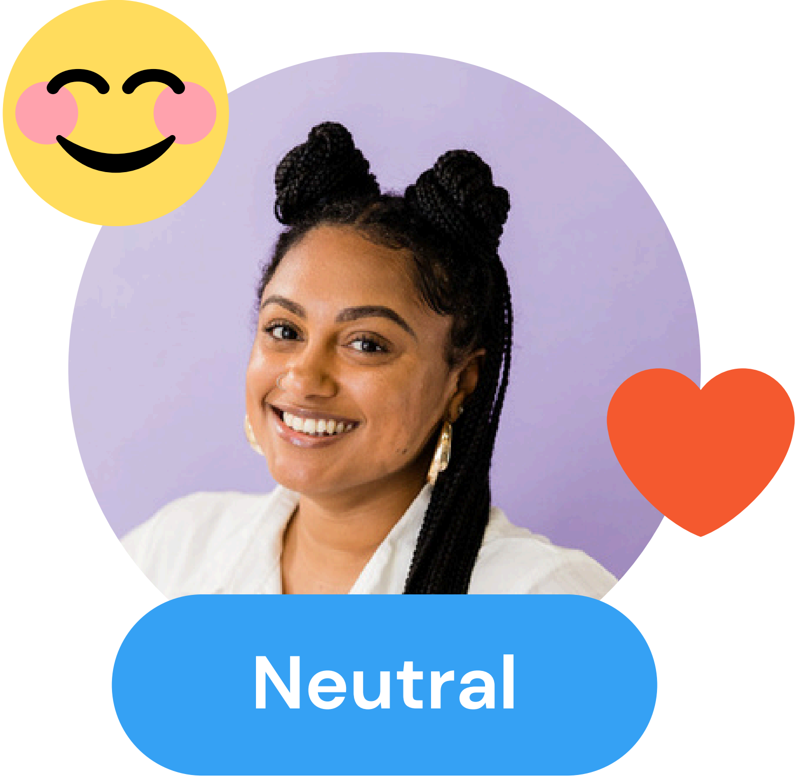
- Removing stopwords and irrelevant tokens
- Filtering out symbols and special characters
- Text Normalization
- super_clean Function
- Replace text between words with a space
- Remove spaces
- Remove common or simple words



2. PRODUCT CATEGORY CLUSTERING

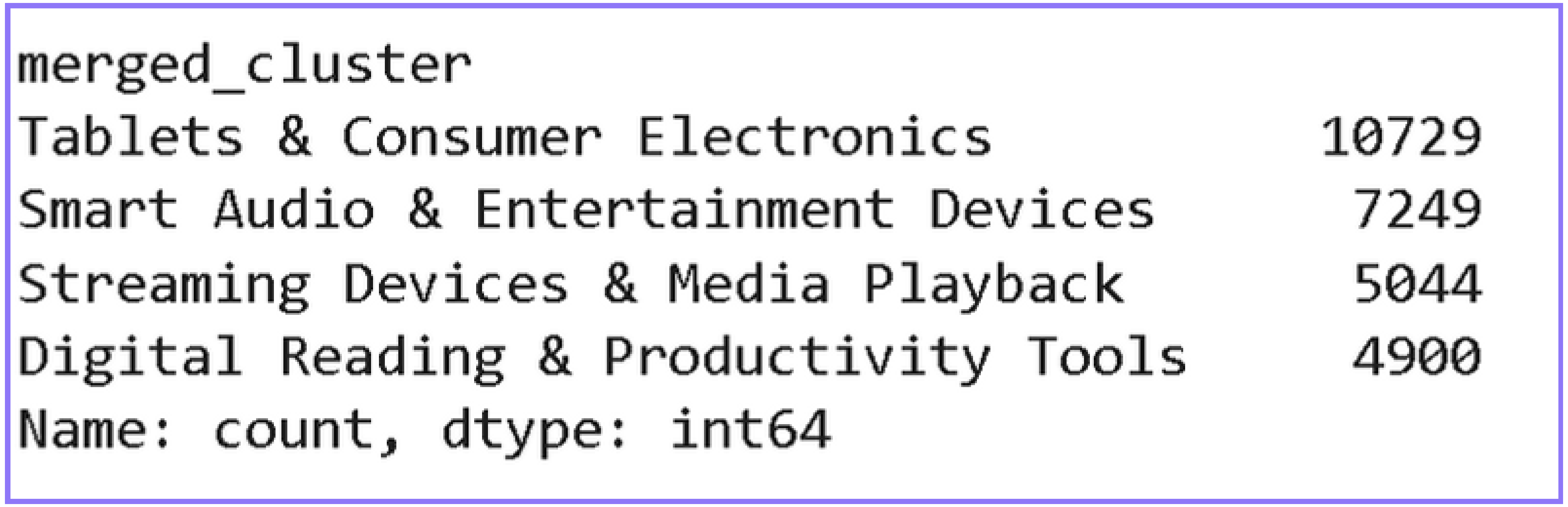
Solution Steps:

- Select the categories - reviews.text
- Embeddings using (intfloat/e5-small-v2)
- Chose number of cluster: 5
- Used unsupervised learning techniques : K-Means.
- Merged clusters
- Evaluate clusters based on Top words.

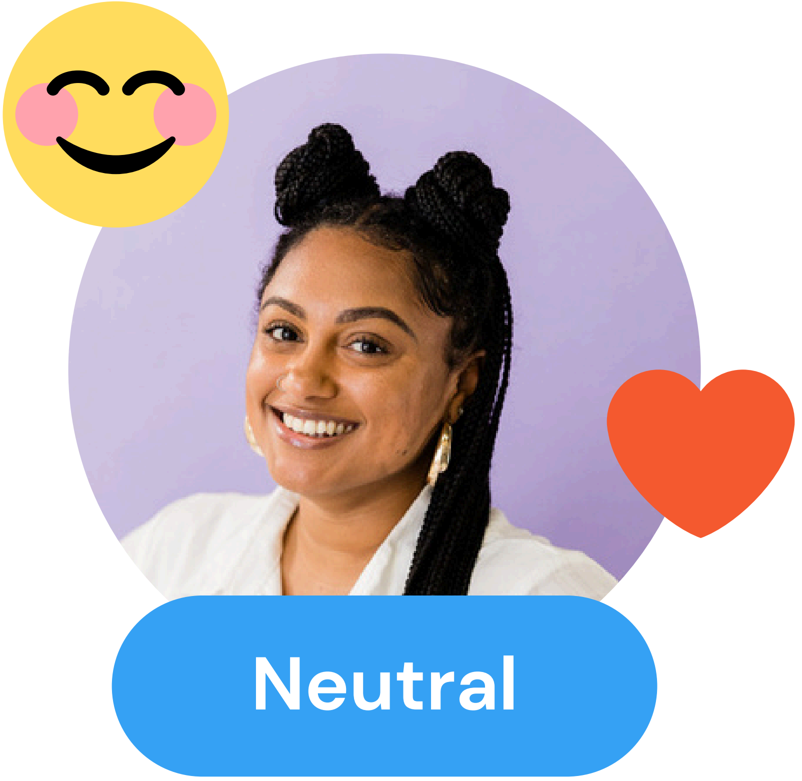


2. PRODUCT CATEGORY CLUSTERING

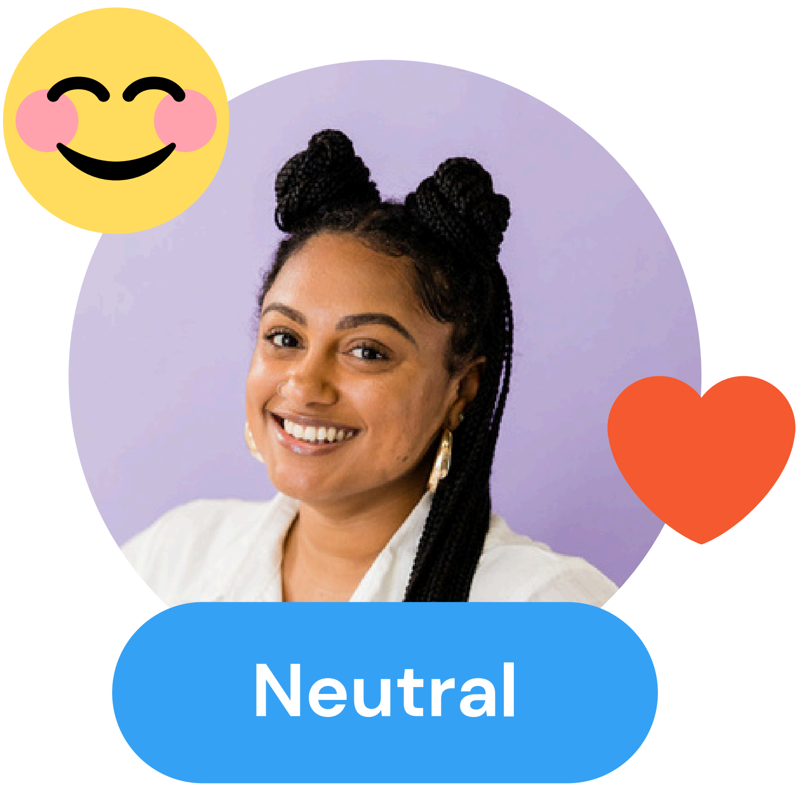
Names Of The Clusters After Merge



```
merged_cluster
Tablets & Consumer Electronics             10729
Smart Audio & Entertainment Devices         7249
Streaming Devices & Media Playback          5044
Digital Reading & Productivity Tools        4900
Name: count, dtype: int64
```



3. REVIEW SUMMARIZATION



3. REVIEW SUMMARIZATION

1

BART

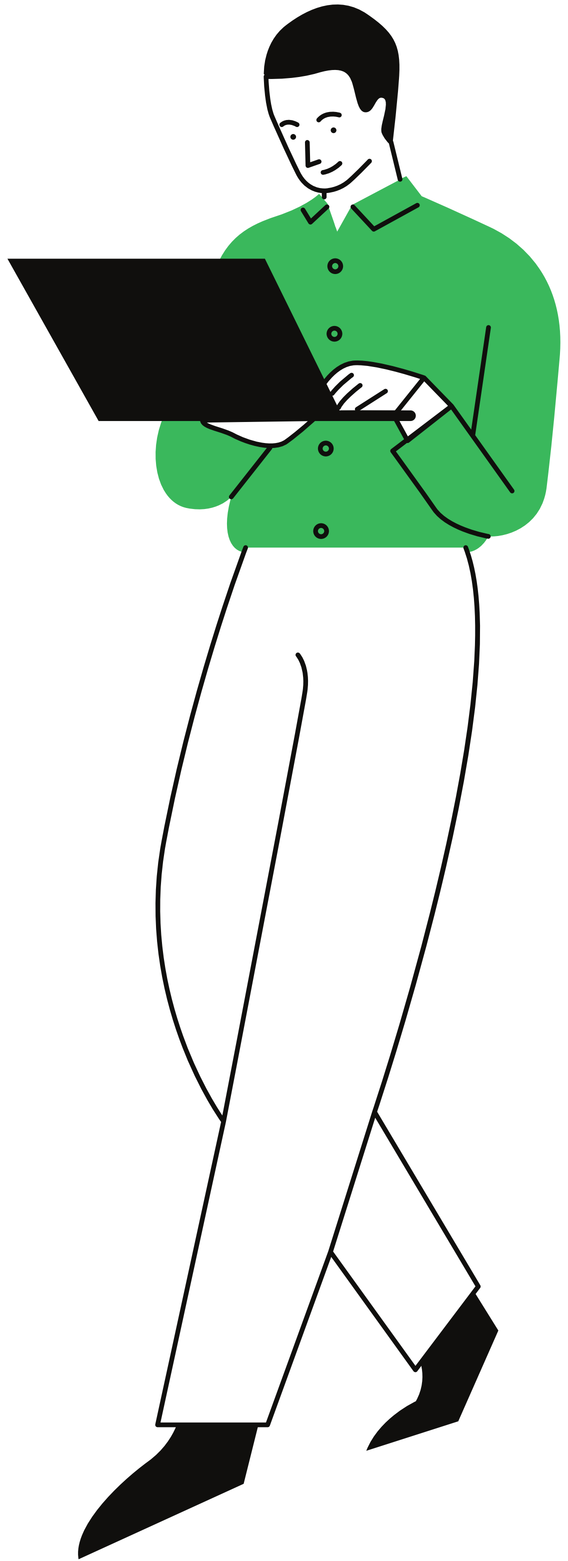
2

T5

3

GPT-3.5

COMPARE OUTPUTS OF ALL THREE MODELS

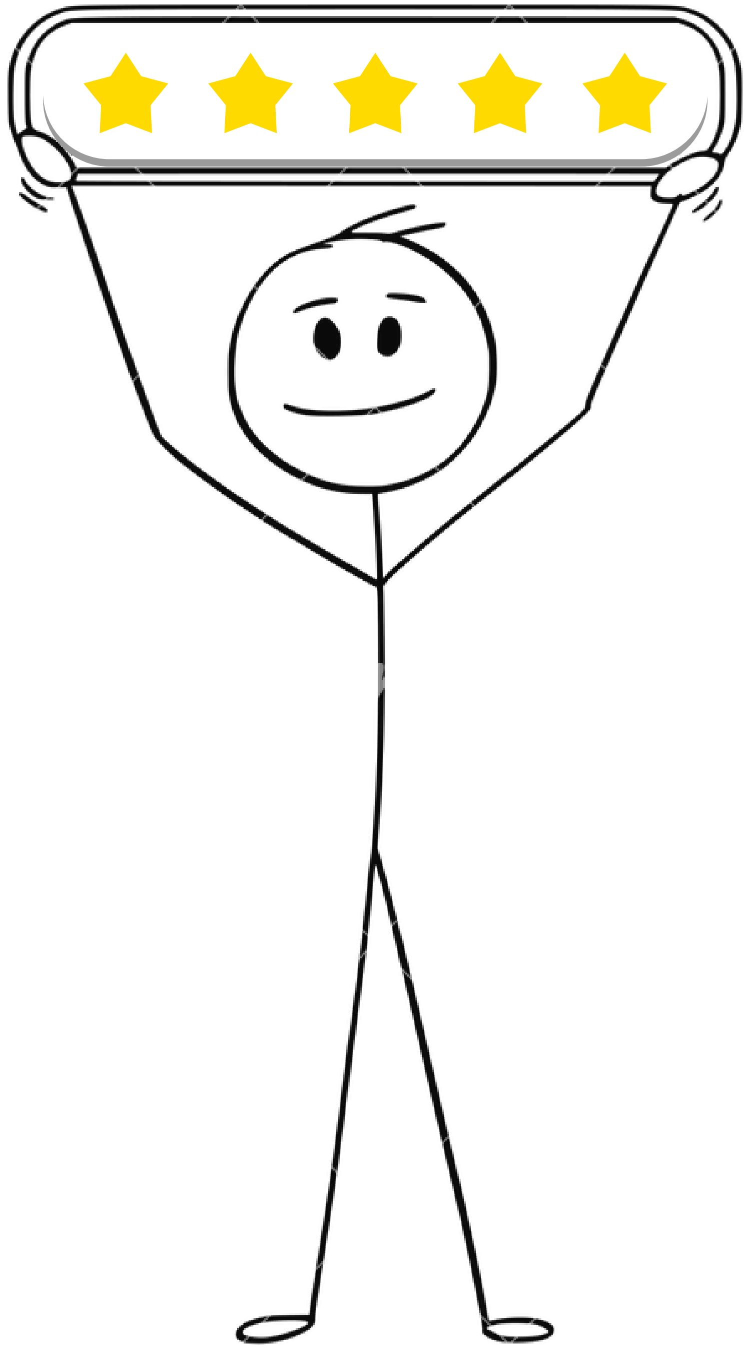


3. REVIEW SUMMARIZATION

Best Model Selection

The best-performing model is

“GPT-3.5”



CHALLENGES & SOLUTIONS

Challenges	Solutions
1. Dataset Imbalance	Use rebalancing techniques such as oversampling or under sampling.
2. Large-scale review datasets (like Amazon Reviews) require high computational resources and can slow down training and testing.	Use lightweight models like BERT to reduce resource consumption.
3. Difficulty in choosing the appropriate number of groups for clustering	Use criteria like Silhouette Score and Elbow Method to automatically select the optimal number.

DEPLOYMENT

😊 ❤️ ⭐ ✨ 👍 👏 👌

To make our project accessible and user-friendly, we deployed the models using Gradio, an open-source Python library that allows for quick and interactive demo. 🚀

gradio

CONCLUSION

😊 ❤️ ⭐ ✨ 👍 👏 👌

Our project proves how NLP can turn massive customer reviews into clear, smart, and useful insights.

We built a system that classifies, clusters, and summarizes reviews helping both businesses and users make better decisions.

This is just the beginning of what AI can do for customer experience! 🚀

DEMO

https://ae1942335d54fb783d.gradio.live

The product arrived on time and the packaging was fine, but the quality didn’t match my expectations. It stopped working properly after just two weeks, and customer support wasn’t helpful at all.

Absolutely love this product! The battery lasts forever, it’s super easy to use, and it works even better than I expected. Totally worth the price!"

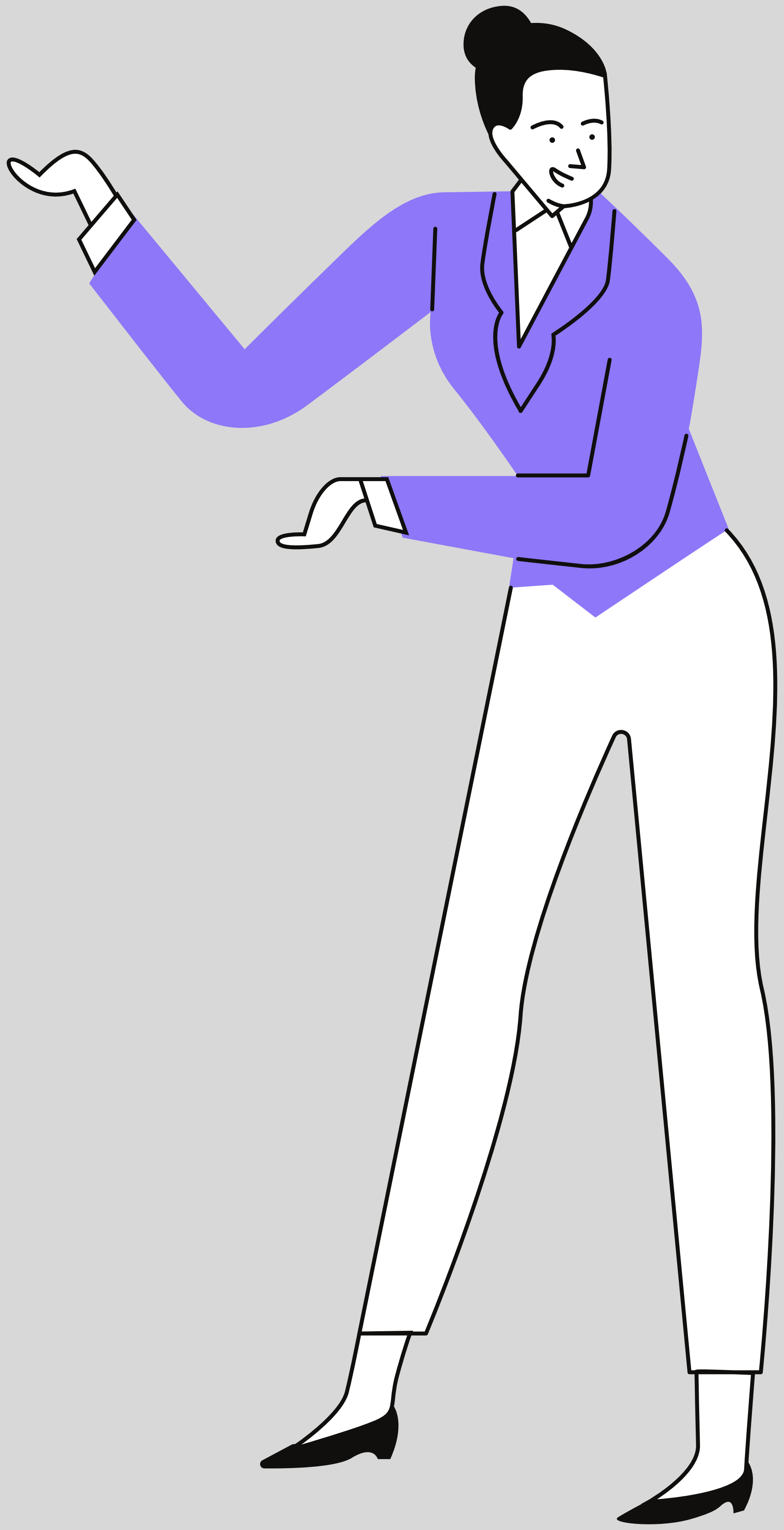
The product works as described. Nothing extraordinary, but it does the job. Delivery was okay and setup was straightforward.

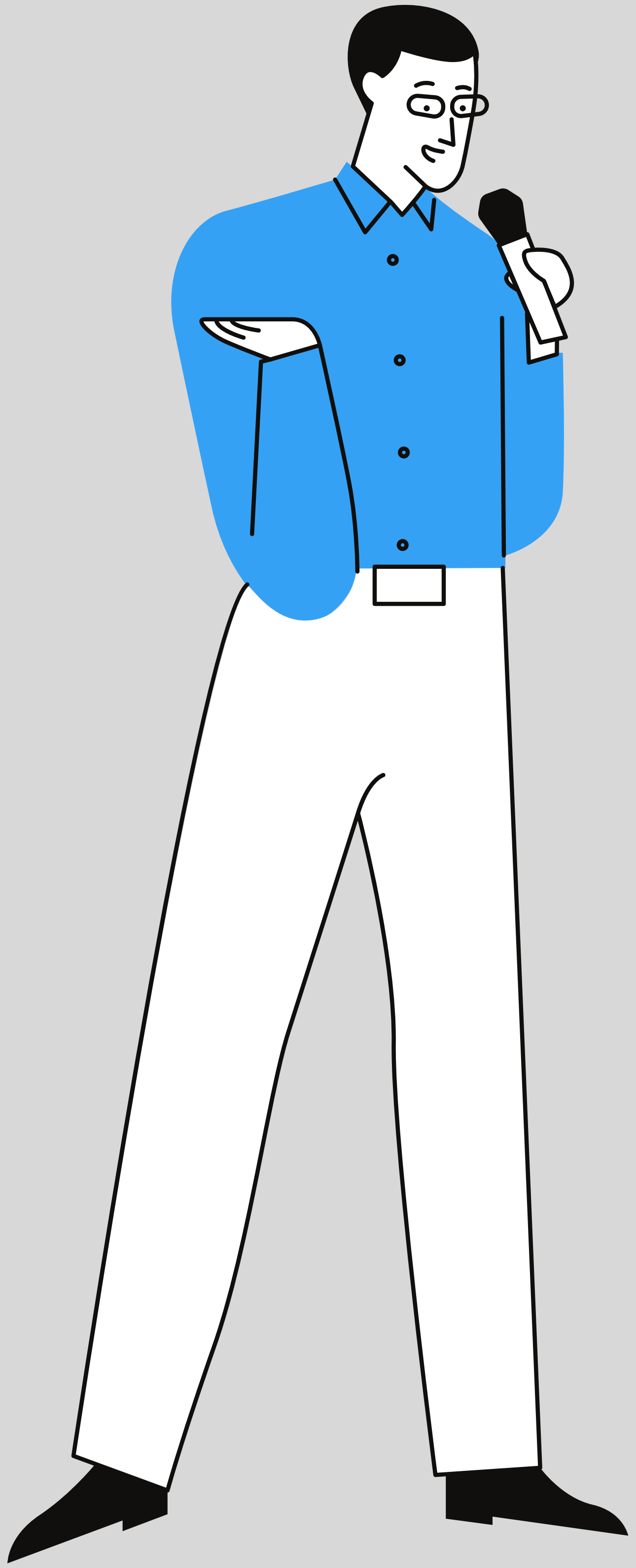
Organizing Labor Division in Our Team Strategies & Implementation

Ahmed Alqarni	Review Classification
Amal alghtani	Product Category Clustering
Hanan Alnbhani	Review Summarization

Thank you

For your kind listening





ANY Questions?