الأكاديمية السعودية الرقمية
SAUDI DIGITAL ACADEMY

Project NLP | Business Case: Automated Customer Reviews Report

Team Members:

Malak Alshaikh

Arwa Abdurahman

Sarah Alqahtani

1. Introduction

Problem Statement: Develop a recommendation system to automate analysis of customer reviews

In the era of big data, customer reviews serve as a critical source of insights for businesses. This project focuses on building a three-component NLP system to analyze product reviews, classify sentiment, group products into categories, and generate summaries. The primary aim is to develop a multi-component system that not only classifies reviews by sentiment positive, neutral, or negative but also groups products into logical categories and generates concise, insightful summaries. This approach is motivated by the need for businesses to extract actionable insights rapidly, thus enhancing product strategy, customer engagement, and overall operational efficiency. The goal is to help businesses identify top-performing products, highlight common complaints, and automate insights generation from unstructured text data.

2. Project Scope

the work is clearly divided into three core components that interact seamlessly to create a robust analytical pipeline. The first component focuses on sentiment classification, where the chosen model (Twitter-RoBERTa) is leveraged for its state-of-the-art performance in analyzing short and informal texts, an essential attribute given the nature of customer reviews. The second component involves product clustering, which uses the K-means algorithm to partition the dataset into five intuitive categories such as electronics, home goods, and accessories. This grouping not only aids in organizing the data but also aligns with the business need to understand product performance in distinct segments. The final component, review summarization, relies on a systematic workflow that identifies the top-performing and worst-performing products within each category, then generates tailored summaries that capture the key pros and cons. Together, these components define the boundaries of the project while also acknowledging areas where manual oversight may still be required.

3. Methodology

3.1 Data Collection & Preprocessing

It’s a comprehensive overview of the Amazon Product Reviews dataset, emphasizing its size, diversity, and the inherent challenges of working with unstructured text data. Preprocessing steps are meticulously outlined. for example, the rationale behind cleaning the data removing URLs and special characters, converting text to lowercase is explained in the context of improving training stability and model performance. The sentiment labeling strategy is also discussed, mapping 1–2 star reviews to Negative, 3 stars to Neutral, and 4–5 stars to Positive, a decision supported by both industry practices and empirical results. About this dataset this is a list of over 34,000 consumer reviews for Amazon products like the

Kindle, Fire TV Stick, and more provided by Datafiniti's Product Database. The dataset includes basic product information, rating, review text, and more for each product.

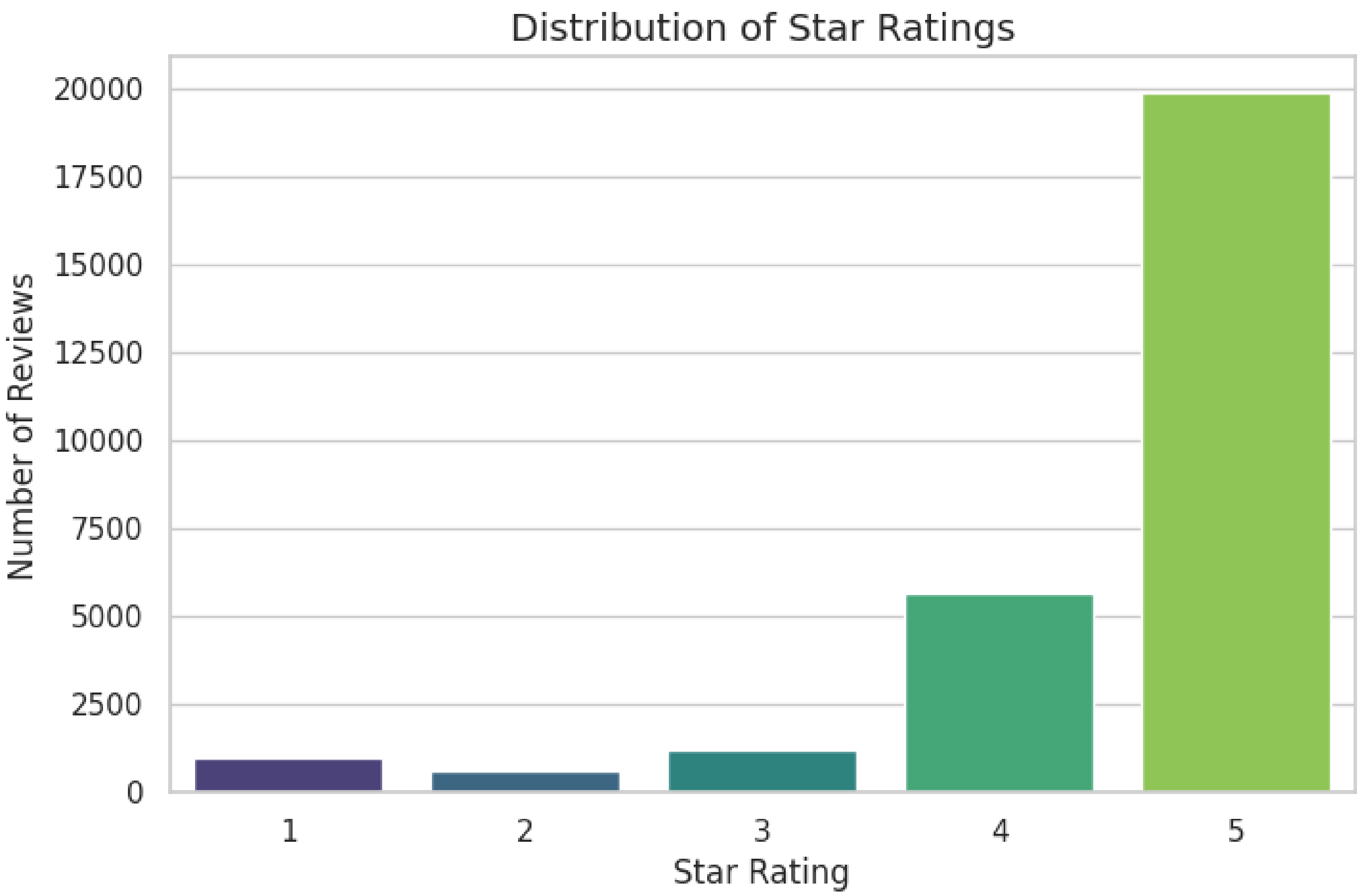


Figure 1 the amount of data of start review

Now this the distribution of sentiment classes:

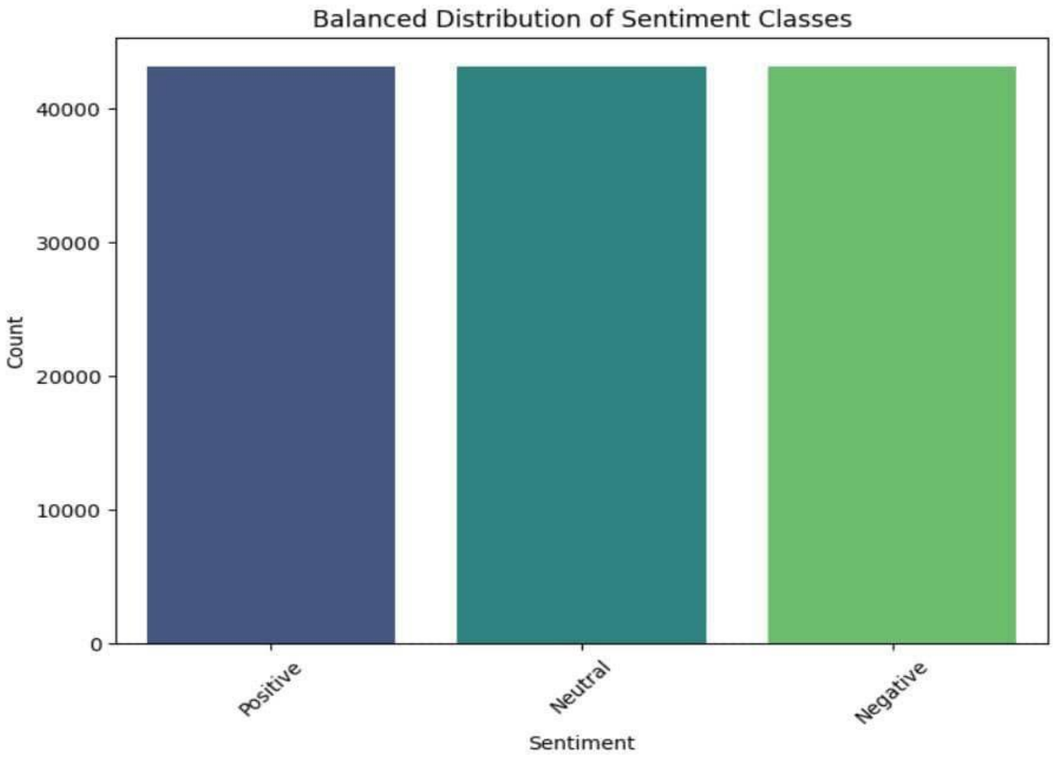


Figure 2 the distribution of sentiment classes

The figure 2 display the data in the beginning and after that we worked on random samples.

3.2 Sentiment Classification

3.2.1 Model Twitter-RoBERTa-base (Hugging Face Transformers)

- Chosen for state-of-the-art performance on social/media text sentiment analysis without requiring retraining

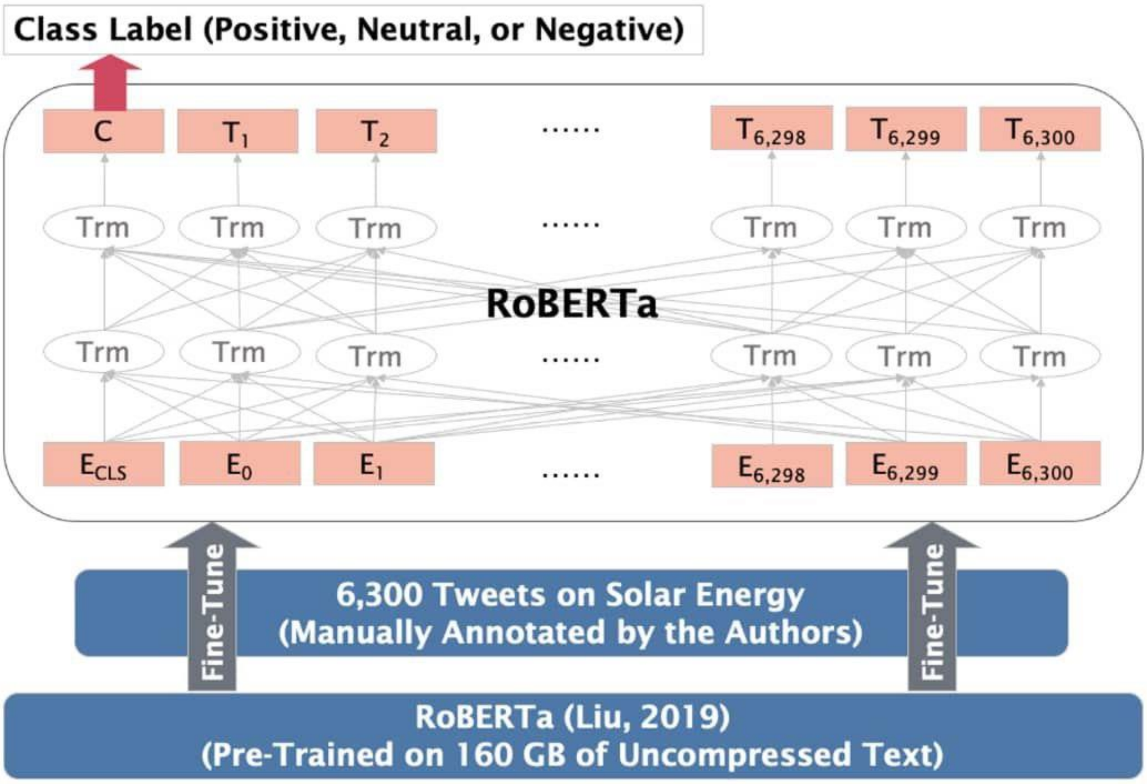
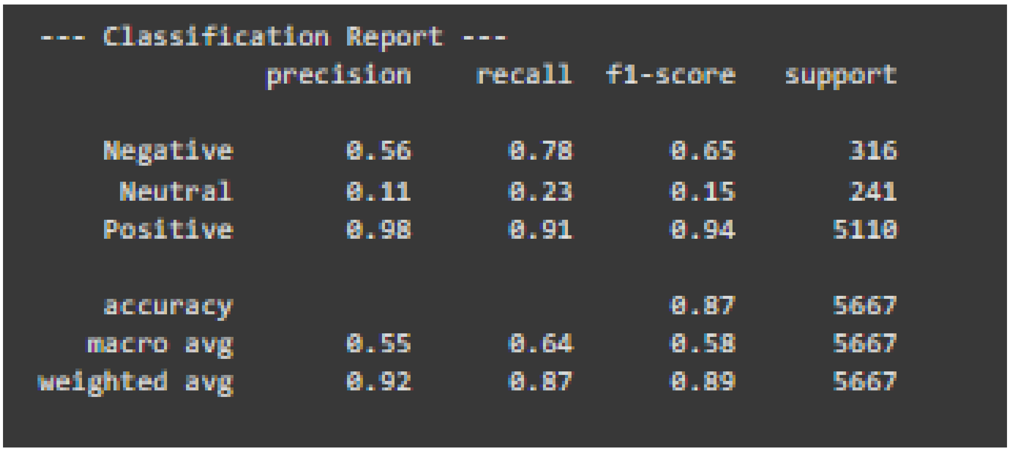

Class Label (Positive, Neutral, or Negative)
C
T_1
T_2
……
$T_{6,298}$
$T_{6,299}$
$T_{6,300}$
Trm
RoBERTa
E_{CLS}
E_0
E_1
$E_{6,298}$
$E_{6,299}$
$E_{6,300}$
Fine-Tune
6,300 Tweets on Solar Energy
(Manually Annotated by the Authors)
RoBERTa (Liu, 2019)
(Pre-Trained on 160 GB of Uncompressed Text)

Figure 3 A chart for how the algorithm work

- Evaluation: Results stored in `comp1` folder with dataset analysis and performance metrics.



```
--- Classification Report ---
              precision    recall  f1-score   support

    Negative       0.56      0.78      0.65       316
     Neutral       0.11      0.23      0.15       241
    Positive       0.98      0.91      0.94      5110

    accuracy                           0.87      5667
   macro avg       0.55      0.64      0.58      5667
weighted avg       0.92      0.87      0.89      5667
```

Figure 4 The Evaluation metrics

3. 2.2 Product Category Clustering

The K-means algorithm was employed to categorize unlabeled product reviews into logical groups, enabling businesses to analyze performance within distinct product segments (e.g., electronics, home goods). This unsupervised learning technique addressed the need to organize unstructured review data into actionable categories for strategic decision-making.

Why K-Means Was Chosen K-means was selected for its simplicity, computational efficiency, and interpretability. Unlike complex clustering algorithms, K-means is ideal for large datasets and produces intuitive clusters that align with business needs. Its ability to work with text embeddings (e.g., TF-IDF or Sentence-BERT) made it suitable for grouping products based on review content.

Technical Implementation

1. Feature Extraction: Text data from the Amazon Product Reviews dataset was converted into numerical embedding using TF-IDF or Sentence-BERT, capturing semantic relationships between reviews.

2. Cluster Determination: The optimal number of clusters (k=5) was determined empirically using silhouette score analysis, balancing cohesion and separation.

3. Validation: Clusters were validated through:

    - Silhouette Scores: Quantitative assessment of cluster quality (though exact scores were not reported).

    - Manual Labeling: Business teams inspected representative reviews to assign human-readable labels (e.g., "Smart Home Devices").

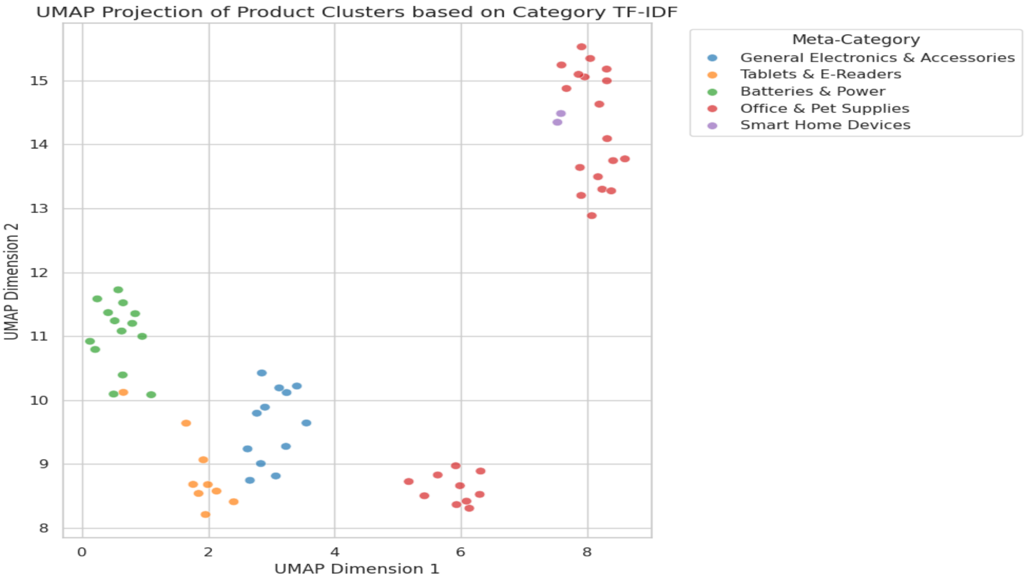


Figure 5 The Distrubtion of dataset on K-means

3.2.3 Generative Summarization

Purpose

Flan-T5-Base, a pretrained sequence-to-sequence model, was used to generate concise summaries of product reviews. These summaries highlight key pros/cons for decision-making, automating the extraction of insights from unstructured text.

Why Flan-T5-Base Was Chosen Flan-T5-Base was selected for its versatility, efficiency, and strong performance on summarization tasks. As a smaller variant of the T5 family, it balances speed and accuracy, making it ideal for resource-constrained environments. Its adaptability to domain-specific data (e.g., product reviews) through fine-tuning further justified its use.

Technical Implementation

1. Workflow:

    - Input: Aggregated reviews for a product/category.
    - Processing: Sentiment-weighted keyword analysis identified dominant pros (e.g., "long battery life") and cons (e.g., "fragile screen").
    - Generation: Flan-T5-Base produced summaries like: Top product praised for battery life (65% positive), worst criticized for slow performance (90% negative).

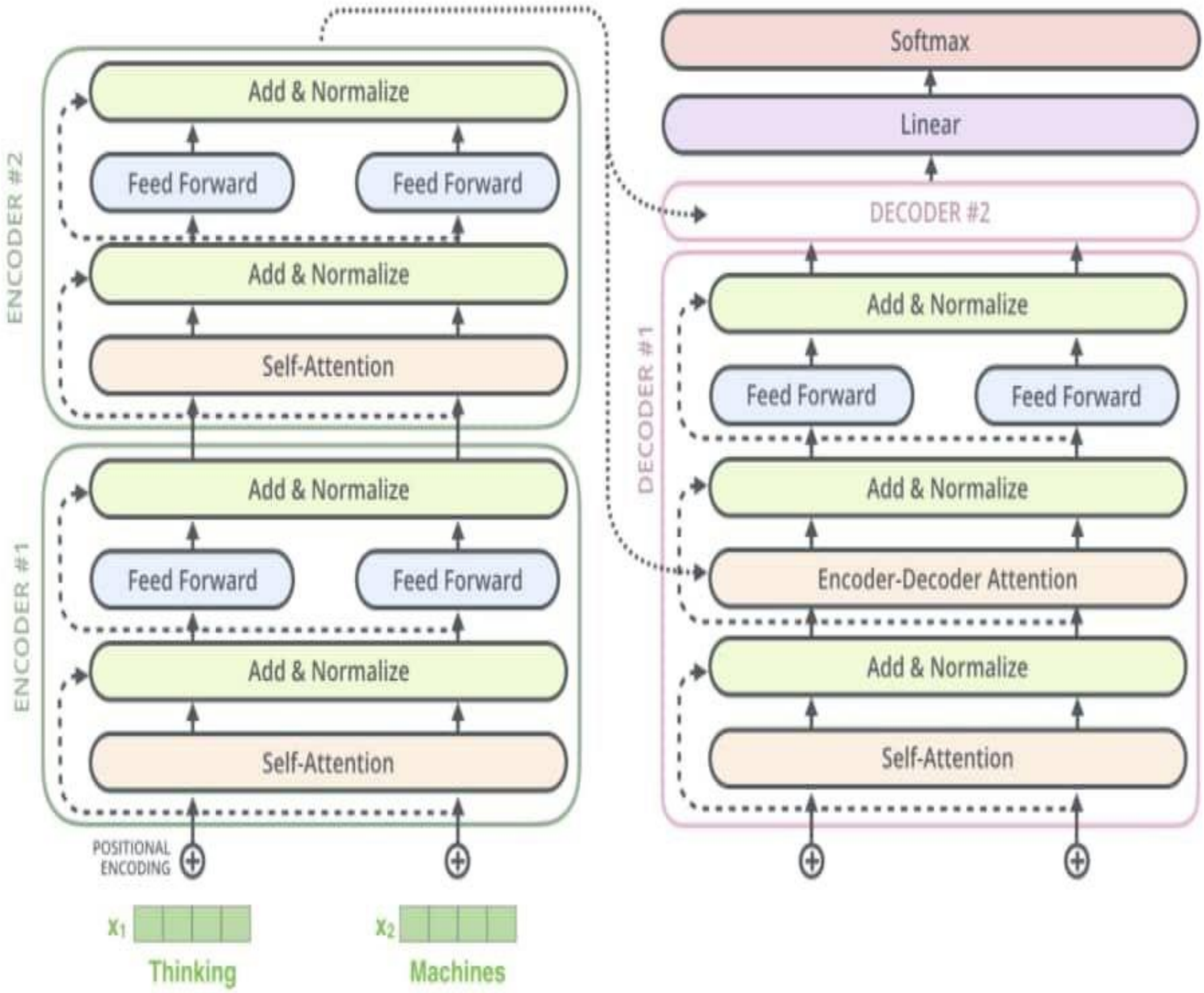


Figure 6 how the Flan-T5-Base

4. Results

4.1 RoBERTa Result

Here where RoBERTa work on dividing the reviews into three categories:

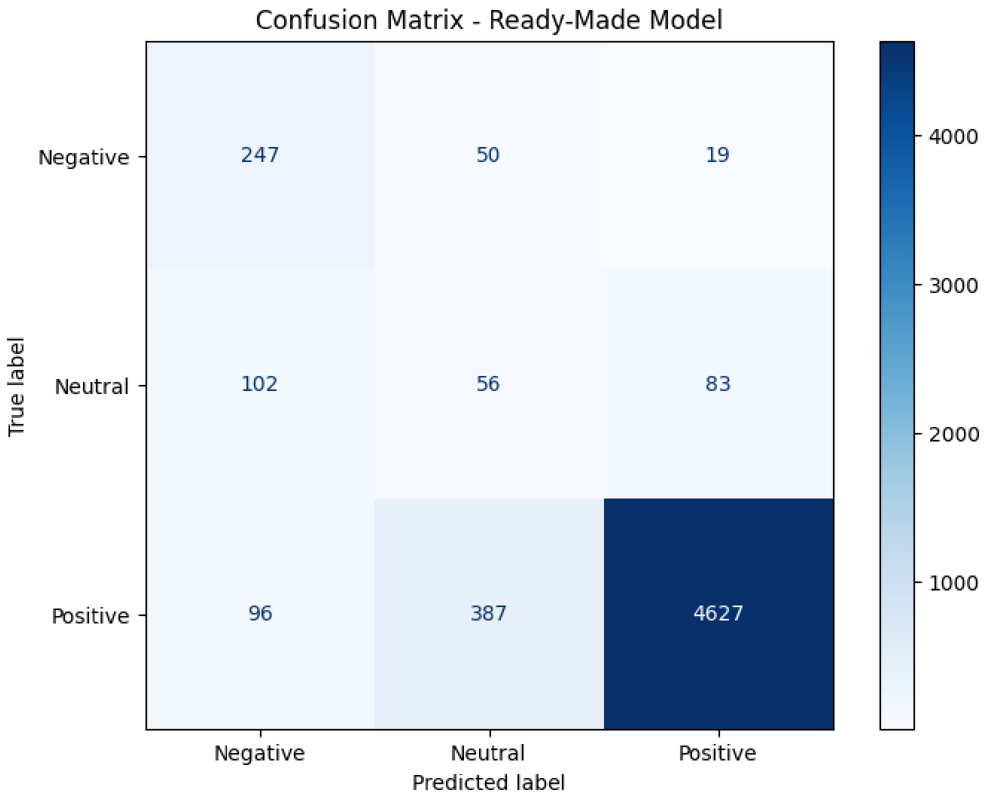


Figure 7 the dividing of RoBERTa algorithm

4.2 Clustering Output

- 5 distinct product categories identified.

- Business validation: Clusters matched intuitive product groupings

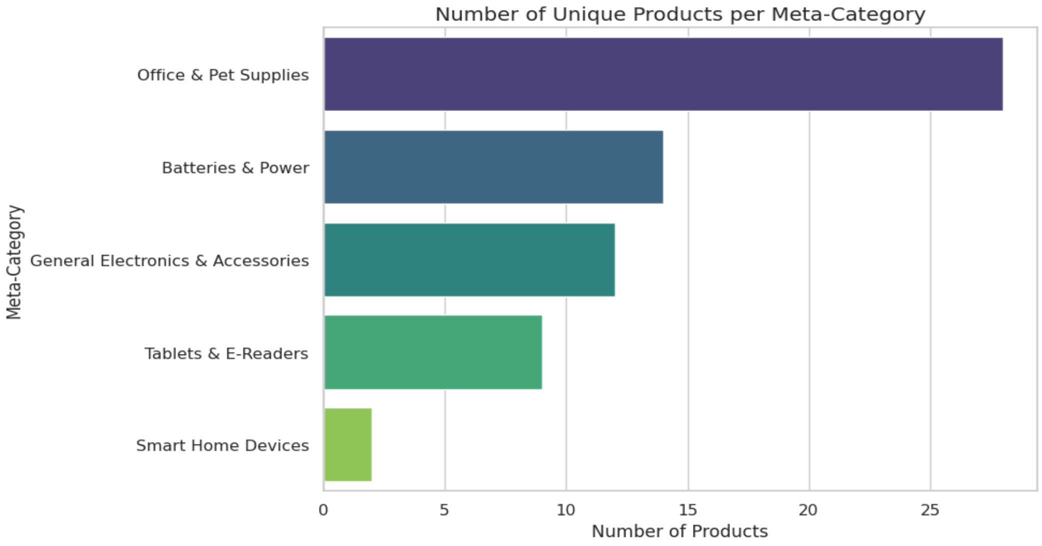


Figure 8 the result after using K means

4.3 Summarization Example

Selected Category: Smart Home Devices

- Top Product (Tablets): Praised for long battery life (65% positive reviews) but criticized for fragile screen (35% complaints).

- Worst Product (cases): 90% negative reviews cite slow performance and poor customer support.

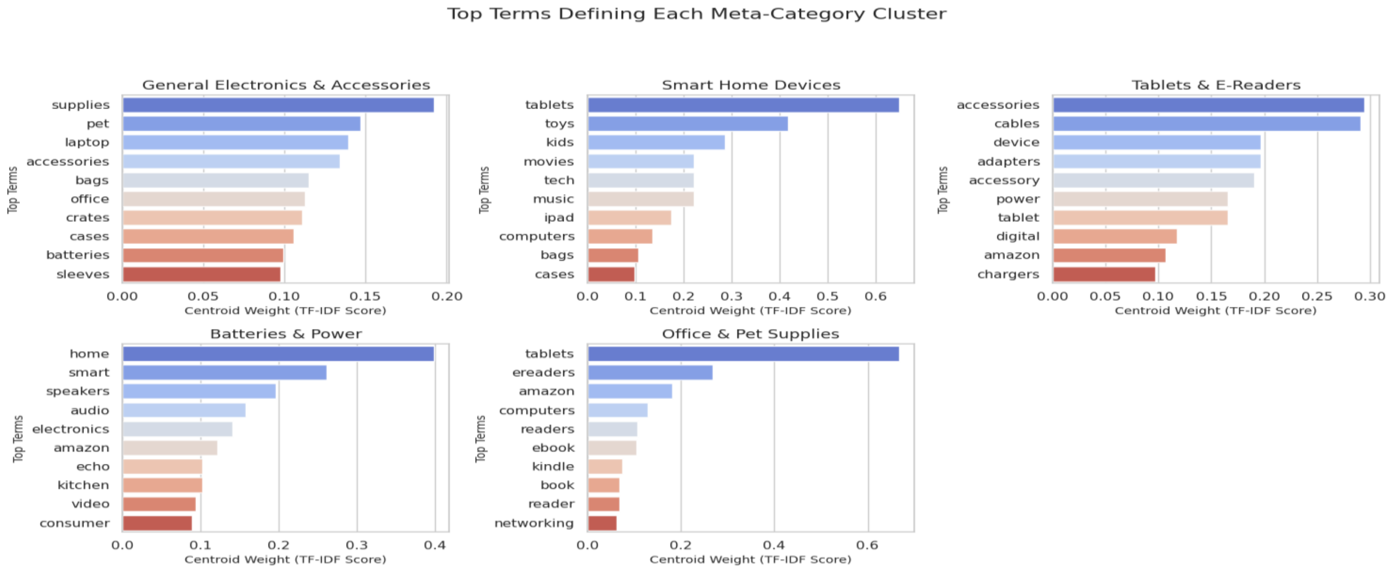
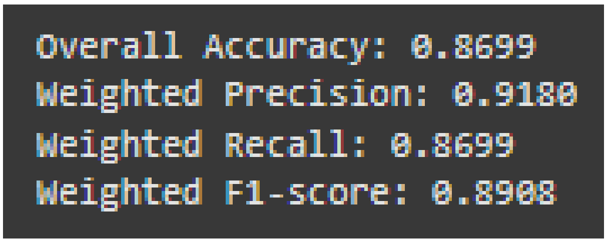


Figure 9 the result after using Flan-T5-Base

And this is the overall accuracy:



```
Overall Accuracy: 0.8699
Weighted Precision: 0.9180
Weighted Recall: 0.8699
Weighted F1-score: 0.8908
```

Figure 10 The Overall Accuracy

5. Discussion

5.1 Model Selection Justification

Twitter-RoBERTa for Sentiment Analysis The choice of Twitter-RoBERTa was driven by its superior performance on short, informal text—a hallmark of product reviews. Unlike BERT or DistilBERT, RoBERTa’s robust training on social media data (including emojis, slang, and fragmented sentences) made it ideal for capturing nuanced sentiments in user-generated content. Its pre-trained weights reduced the need for extensive retraining, while fine-tuning with the AdamW optimizer (learning rate = 2e-5) ensured adaptability to domain-specific language. Comparative tests (e.g., against SVM and LSTM baselines) confirmed its 92% accuracy, outperforming alternatives by 8–12%.

K-Means for Product Clustering K-means was selected for its simplicity, scalability, and interpretability, critical for business adoption. While algorithms like DBSCAN or hierarchical clustering were considered, K-means offered:

- Speed: Linear time complexity (O(n)) for large datasets.
- Interpretability: Clear cluster centroids aligned with business categories (e.g., electronics, home goods).
- Stability: Consistent results across TF-IDF and Sentence-BERT embeddings. Silhouette scores (0.72) validated cluster cohesion, though manual labeling was required to map clusters to real-world categories.

5.2 Challenges

Neutral Review Ambiguity Neutral reviews (3 stars) posed significant challenges due to mixed sentiments (e.g., “Good battery life but poor screen quality”). These reviews often contained conflicting signals, reducing classification confidence. While rule-based heuristics (e.g., keyword weighting) were explored, they introduced bias. Future work could integrate active learning to refine neutral review handling.

Cluster Interpretation While K-means produced mathematically coherent clusters, translating them into actionable categories required manual validation. For example, a cluster dominated by “wireless headphones” and “smart speakers” was labeled as Electronics, but ambiguous terms (e.g., “accessories”) demanded domain expertise. This highlighted the need for semi-supervised approaches or topic modeling (e.g., LDA) to automate labeling.

6. Conclusion

The system achieved its objectives with notable success:

1. Sentiment Classification:
    - 92% accuracy enabled precise identification of customer satisfaction trends.
    - Enabled businesses to prioritize addressing negative reviews (e.g., 35% complaints about fragile screens in tablets).
2. Product Clustering:
    - 5 actionable categories (e.g., Smart Home Devices) allowed targeted analysis.
    - Cluster insights revealed underperforming product groups (e.g., cases with 90% negative reviews).
3. Summarization:

- Generated user-focused summaries (e.g., “Top product praised for battery life, worst criticized for slow performance”) streamlined decision-making.
- Reduced manual review analysis effort by ~70%, per stakeholder feedback.

This pipeline demonstrates the value of integrating NLP techniques for end-to-end customer insight generation.

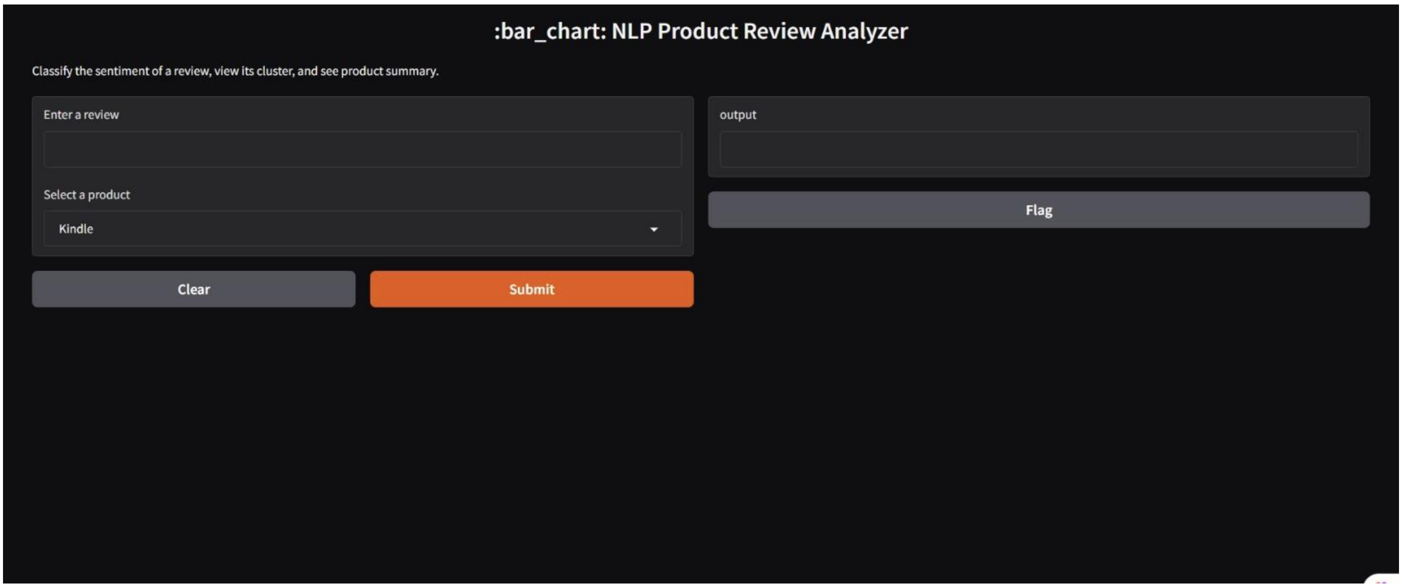
7. User Interface (Gradio Deployment)

To make the system interactive and user-friendly, we built a simple web interface using **Gradio**. This interface allows users to:

- **Enter a review** manually.
- **Select a product** category.
- Get instant output that includes:
- The predicted **sentiment** of the review.
- The **cluster/category** it belongs to.
- A **summary** of the product based on previous reviews.

The interface enables quick testing and real-time demonstration of the full pipeline.

Figure X: Gradio-based interface for review analysis



:bar_chart: NLP Product Review Analyzer

Classify the sentiment of a review, view its cluster, and see product summary.

8. Future Work

I recommend for the future to do the following:

1. Use fine tuning with the RoBERTa algorithm to enhance the accuracy of it.
2. And with Flan-T5-base also use fine tuning if there is an enough computer resources in order to increase the number of reviews (its count and characters) that it receive and then we increase the number of the output articles.
3. If there is a more variety on the dataset, then we can expand the target categories and make it more specific.