AI Engineering – LLM Text Summarization Assignment

1. Introduction

Text summarization is an important Natural Language Processing (NLP) task that aims to condense large volumes of textual data into concise and meaningful summaries. With the rapid growth of user-generated content such as app reviews, manual analysis becomes inefficient and error-prone.

This project focuses on building an **LLM-based abstractive text summarization pipeline** to summarize real-world user reviews of the Infloso application. The solution leverages a pretrained transformer model to generate coherent summaries while handling noisy, informal, and mixed-language text.

2. Dataset Exploration (STEP 2)

What was done

The provided dataset consists of **100 user reviews** collected from an app marketplace. Each review is a free-form text written by users sharing their experience with the Infloso platform.

Dataset characteristics

- Unstructured textual data
- Informal language with spelling mistakes
- Mixed English and Hinglish
- Presence of emojis and special characters
- Strong sentiment polarity (very positive or very negative)

Why this analysis was necessary

Understanding the dataset helped identify challenges such as noise, language variability, and sentiment imbalance. These factors influenced the preprocessing strategy and motivated the use of an LLM-based abstractive summarization approach rather than traditional extractive methods.

3. Data Preprocessing (STEP 3)

What was done

The following preprocessing steps were applied:

- Converted all text to lowercase
- Removed URLs and special characters
- Removed emojis and unnecessary symbols

- Normalized extra whitespaces
- Preserved sentence structure and stopwords

Why this approach was chosen

LLM-based models rely on contextual understanding, so aggressive preprocessing (such as stopword removal) was avoided. The goal was to reduce noise while maintaining semantic meaning. This ensured that the summarization model received clean but context-rich input.

4. Summarization Strategy

Approach

An **abstractive summarization** approach was selected using a pretrained Large Language Model.

Model Used

- **facebook/bart-large-cnn**
- Transformer-based encoder–decoder architecture
- Pretrained on large-scale summarization datasets

Why this model

- Designed specifically for summarization tasks
- Capable of generating human-like summaries
- Handles long and noisy text effectively
- Industry-standard model for document summarization

5. Sentiment-Aware Summarization

What was done

The reviews were heuristically divided into:

- **Positive reviews**
- **Negative reviews**

Separate summaries were generated for each group and later combined into a balanced final summary.

Why this was done

Direct summarization of the entire dataset can bias the output toward longer or more frequent sentiments. By separating reviews based on sentiment, the final summary fairly represents both positive feedback and user complaints, resulting in a more informative and trustworthy summary.

6. Results and Analysis

6.1 Dataset Summary

The summarization pipeline was executed on a dataset containing **100 user reviews** related to the Infloso application.
The sentiment distribution after preprocessing was:

- **Positive reviews:** 70
- **Negative reviews:** 30

This imbalance reflects real-world user feedback, where satisfied users tend to leave more reviews, but negative feedback often contains more detailed complaints.

6.2 Generated Balanced Summary

To ensure fair representation, sentiment-aware summarization was applied. Separate summaries were generated for positive and negative reviews and then combined into a balanced output.

Positive Summary Analysis

The positive summary highlights:

- High user satisfaction with influencer–brand collaboration features
- Appreciation for paid barter campaigns and monetization opportunities
- Positive feedback on reminders, deadlines, and collaboration workflow
- Overall perception of Infloso as a valuable tool for content creators

This indicates that the application performs well in terms of its core value proposition for influencers.

Negative Summary Analysis

The negative summary captures recurring issues such as:

- Problems connecting social media accounts (Instagram and YouTube)
- Poor customer support and lack of response to collaboration requests
- Misleading campaign eligibility requirements
- Delayed payments and lack of transparency

These issues point toward usability and support-system limitations rather than core functionality problems.

6.3 Important Review Extraction

In addition to summarization, the system identified the **Top 5 most important reviews** using TF-IDF-based importance scoring.

These reviews represent:

- Critical user complaints with high informational value
- Detailed feedback on UX, backend issues, and payment failures
- Highly positive reviews explaining why users find the app valuable

This step ensures that decision-makers can directly access impactful individual reviews alongside the generated summary.

6.4 Evaluation Metrics

The quality of summarization was evaluated using the **ROUGE-L metric**, which measures the longest common subsequence between generated summaries and reference text.

- **Average ROUGE-L Score: 0.0359**

Interpretation

The ROUGE score is relatively low due to:

- Absence of a human-written gold reference summary
- Highly abstractive nature of the generated summaries
- Informal and noisy review text

Therefore, **qualitative evaluation** was prioritized over purely numerical metrics.

6.5 Qualitative Evaluation

Manual inspection confirms that:

- The summaries are coherent and readable
- Key themes from both positive and negative reviews are preserved
- The balanced summary avoids sentiment bias
- Generated content accurately reflects user experience

Despite a low ROUGE score, the summaries successfully capture actionable insights, making them valuable for real-world analysis.

6.6 Output Artifacts

The following output files were generated:

- `final_summary.txt` – Contains the balanced positive and negative summaries
- `top_5_reviews.csv` – Stores the most informative reviews
- Google Collab notebook – Complete reproducible pipeline

6.7 Overall Result Discussion

The results demonstrate that the proposed LLM-based summarization pipeline:

- Effectively handles noisy real-world text
- Produces sentiment-aware, balanced summaries
- Extracts high-value user feedback
- Is scalable and deployable for app review analysis

This validates the practical applicability of transformer-based summarization models for customer feedback analysis.

7. Top 5 Reviews Selection

What was done

The top 5 most representative reviews were selected using **TF-IDF-based importance scoring** instead of random or length-based selection.

Why this method

TF-IDF identifies reviews containing informative and distinctive terms. This ensures that the selected reviews capture diverse and critical user experiences rather than repetitive or generic feedback.

(Paste the Top 5 Reviews output here)

8. Evaluation Strategy

Evaluation Method

- **Qualitative analysis**: Manual inspection of summary coherence and coverage
- **ROUGE-L metric**: Used as a reference-based evaluation metric

Observations

The ROUGE score is relatively low due to the absence of a gold reference summary and the highly abstractive nature of the task. Therefore, qualitative evaluation was prioritized, which confirmed that the generated summary accurately reflects user sentiment and key issues.

9. Architecture and Hyperparameters

Architecture

- Transformer Encoder–Decoder (BART)
- Pretrained weights from Hugging Face

Key Hyperparameters

- Maximum summary length: 120–140 tokens
- Minimum summary length: 40–50 tokens
- Beam search size: 3–4

- Chunk size: 500 words
- CPU inference used for stability

10. Future Improvements

Potential Model Improvements

- Fine-tuning the summarization model on domain-specific app review data
- Using multilingual models to better handle Hinglish and regional language reviews
- Implementing aspect-based summarization for UI, payments, and support issues

Future Research Directions

- Fake review detection using LLM embeddings
- Sentiment-aware hierarchical summarization
- Real-time summarization for continuously incoming reviews

11. Conclusion

This project demonstrates a complete LLM-based text summarization pipeline applied to real-world noisy data. By combining preprocessing, sentiment-aware summarization, evaluation, and thoughtful review selection, the solution provides meaningful insights into user feedback and showcases practical AI engineering skills.