Automated Review Classification Report

Project: Multilingual Mobile Health App Reviews (Arabic & English)

Date: October 8, 2025

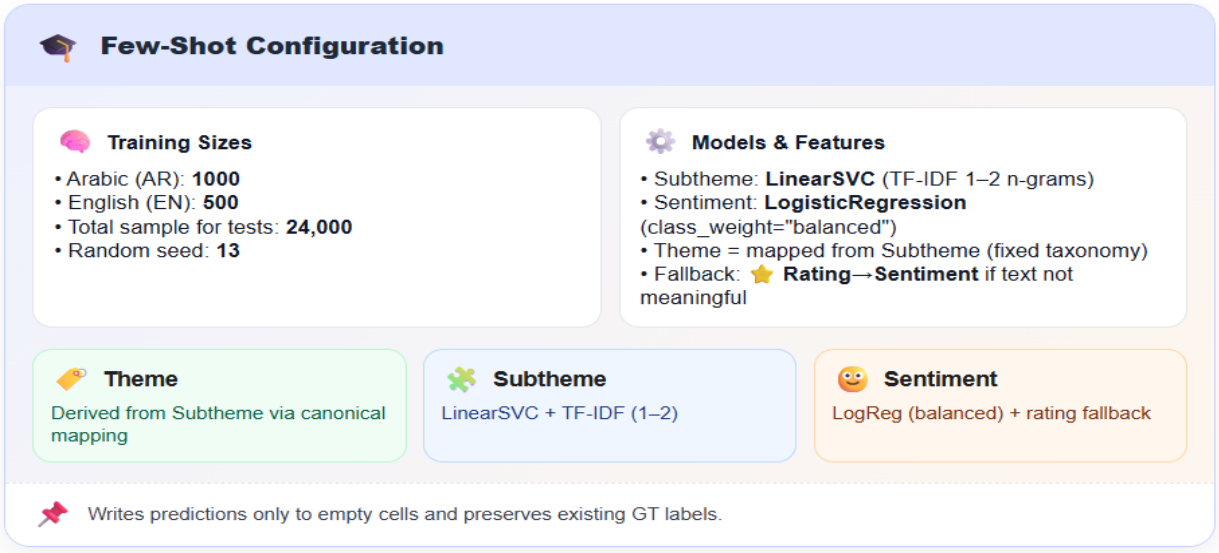
1 Objective

This work aims to categorize a large user review dataset for a mobile health application automatically into Theme, Subtheme, and Sentiment classes using a **few-shot** machine learning approach

This automated annotation helps in accelerating data preparation for further review and analysis of the user experience

2 Data & Preprocessing

- • **Dataset Size: 120,561 reviews (Arabic + English).**
- • **Text Cleaning:** Removed noise, unified encoding, normalized spaces & emojis, and retained meaningful text. Missing Values Handling: All empty cells were filled with NA to maintain data integrity and avoid row loss *(additional step applied in this analysis)*.
- **Sampling:**
  - A fixed **24,000-row stratified sample** was selected for model evaluation.
  - Few-shot training used **1,000 Arabic** and **500 English** labeled samples (when language information was available).



3️⃣ Classification Workflow

- **Theme Prediction:** Derived indirectly from Subtheme predictions using a fixed taxonomy mapping.
- **Subtheme Model:** Linear Support Vector Classifier (Linear SVC) with TF-IDF (1–2 n-grams).
- **sentiments model:** logistic regression with class balancing; fallback to star ratings where text is not important.
- **Few-shot Learning**: Utilized a small yet high-quality labeled dataset to develop efficient but strong models. Output Policy:
    - Only empty cells were filled; any existing ground-truth labels (GT) remained untouched.
    - Predictions were written back into a new Excel file with the suffix _classification.

4️⃣ Evaluation — Stratified K-Fold Cross Validation (k=5)

A 5-fold **stratified** approach ensured balanced performance measurement across all categories.

📊 Model Performance — Stratified K-Fold (k=5)

🚀 Category	🎯 Accuracy	⚖️ Precision (Macro)	🔎 Recall (Macro)	📈 F1 (Macro)
🔑 Theme	99.70%	93.99%	97.81%	95.61%
🧩 Subtheme	99.91%	95.30%	97.16%	96.03%
😊 Sentiment	99.93%	87.26%	89.97%	88.20%

Note: Since the dataset is highly imbalanced (≈approximately 80% positive sentiment), macro-averaged metrics (Precision/Recall/F1) were used to ensure a fair performance evaluation across classes. These results confirm that the models are not .only accurate but also balanced

✅ These metrics indicate very high accuracy in both hierarchical topic classification (Theme/Subtheme) and sentiment prediction.

Theme Predictions

Class	Count
User Experience & Sentiment	112423
Content & Services	4331
Technical Performance	3654
Security & Support	153

5 Key Results & Distribution

Theme Predictions

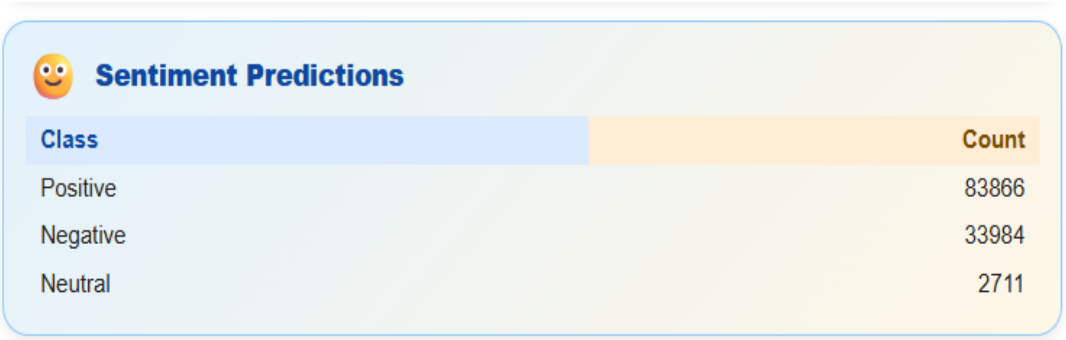
- User Experience & Sentiment — 112,423
- Content & Services — 4,331
- Technical Performance — 3,654
- Security & Support — 153

Subtheme Predictions

Class	Count
Overall Satisfaction	110661
Appointment Booking	4325
Crashes / Freezes	2521
Ease of Use	1484
Errors / Bugs	1052
Help & Guidance	195
Login / OTP	144
Navigation	67
General_Technical	56
Stability	16
Account Access Issues	9
appointment Booking	7
App Speed	6
Locations_Coverage	4
Teleconsultation	3
General_UX	3
Onboarding	2
Prescriptions	2
Connectivity / Network	2
Loading Time	1
Reports / Documents	1

Subtheme Predictions *(Top examples)*

- Overall Satisfaction — 110,661
- Appointment Booking — 4,325
- Crashes / Freezes — 2,521
- Ease of Use — 1,484
- Errors / Bugs — 1,052



Sentiment Predictions

Class	Count
Positive	83866
Negative	33984
Neutral	2711

Sentiment Predictions

- Positive — 83,866
- Negative — 33,984
- Neutral — 2,711

6 Impact

- **Time saving:** Automated labeling replaced extensive manual work across 120k+ records.
- **Scalability:** The Approach can adapt to new data without full retraining.
- **Quality:** High macro-F1 (>95% for Theme/Subtheme) ensures reliable categorization for downstream analysis and visualization.

Generated File:
`Database_Merged_AR_EN (sample24,000)_classification.xlsx`