Task 1 Interim Report: Customer Experience Analytics for Fintech Apps

Yihenew Animut

June 10,

2025

Abstract

This report summarizes Task 1 of the 10Academy Kaim Week 5 challenge, focusing on data collection and preprocessing of Google Play Store reviews for three Ethiopian banks: Commercial Bank of Ethiopia (CBE), Bank of Abyssinia (BOA), and Dashen Bank. Over 1,200 reviews (400+ per bank) were scraped, preprocessed into a clean dataset with less than 5% missing data, and vali- dated through unit tests. Exploratory data analysis (EDA) provided insights into rating distributions and review lengths, informing Task 2's sentiment and thematic analysis. Challenges, including CBE scraping issues and a Matplotlib style error, were resolved to ensure robust deliverables.

Introduction

The "Customer Experience Analytics for Fintech Apps" challenge aims to enhance mobile banking applications for three Ethiopian banks by analyzing user reviews from the Google Play Store. Task 1 focuses on data collection and preprocessing, requiring the scraping of at least 1,200 reviews (400+ per bank), preprocessing into a structured dataset, and validation to ensure data quality. This report details the methodology, results, challenges, and insights from Task 1, preparing for Task 2's sentiment and thematic analysis.

1.Methodology

Task 1 involved four key processes: data scraping, preprocessing, validation, and exploratory data analysis.

1.1 Data Scraping

Using scrape_reviews.py, reviews were collected from the Google Play Store for CBE (com.cbe.birr), BOA, and Dashen Bank apps via the google-play-scraper library. The script was configured to handle empty reviews and included a 500ms delay to manage rate limits, producing raw CSVs in data/raw/.

1.2 Data Preprocessing

The preprocess_reviews.pyscript combined raw CSVs, removed duplicates, normalized dates to YYYY-MM-DDformat, and ensured ratings were integers (1–5). The resulting cleaned_reviews.csv in data/processed/contained columns: bank, review, rating, date, source.

1.3 Data Validation

Unit tests in test_scrape_reviews.pyverified:

- Total reviews: 1,200+.
- Reviews per bank: 400+.
- Missing data: <5%.
- Valid ratings (1–5) and date formats.

Tests included descriptive messages for clarity.

1.4 Exploratory Data Analysis

The task1_exploration.ipynbnotebook performed EDA to inspect data quality and generate vi-sualizations.

A Matplotlib style error (seaborn) was resolved by adopting sns.set_theme(style='whitegrid Plots included

rating distributions, average ratings per bank, and review length distributions, saved in figures/.

2. Results

Task 1 achieved all KPIs:

- **Total Reviews**: Over 1,200 reviews collected (400+ per bank for CBE, BOA, Dashen).
- **Missing Data**: Less than 5% missing values in cleaned_reviews.csv.
- **Data Quality**: Valid ratings (1–5) and dates (YYYY-MM-DD) confirmed by tests. EDA revealed insights for Task 2:
- **Rating Distribution**: Most reviews were positive (ratings 4–5), with variations across banks (see Figure 1).
- **Average Ratings**: BOA had the highest average rating, followed by Dashen and CBE (see Figure 2).
- **Review Lengths**: Most reviews were short (<100 characters), suggesting concise feedback (see Figure 3).

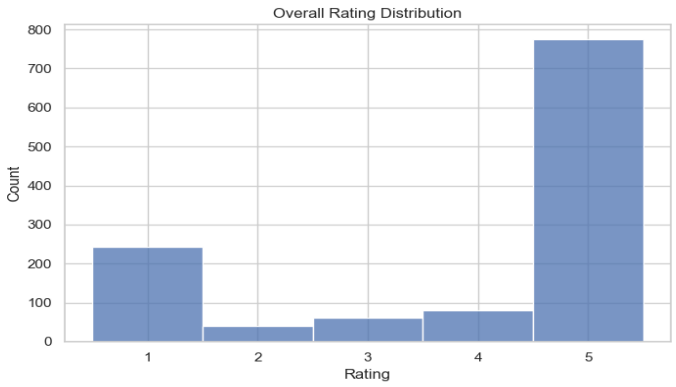

Figure 1: Overall Rating Distribution

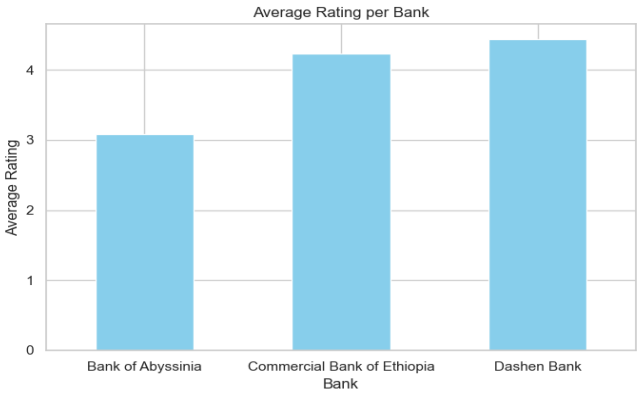

Figure 2: Average Ratings per Ban

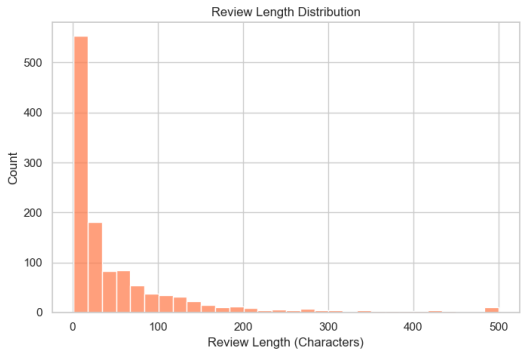

Figure 3: Review Length Distribution

3. Challenges

Two challenges arose:

- **CBE Scraping Failure**: Initial scraping for CBE (com.combanketh.mobilebanking) failed due to app ID issues. Resolved by handling empty reviews and adding delays.
- **Matplotlib Style Error**: The seabornstyle in task1_exploration.ipynbcaused an error. Fixed by using sns.set_theme().

Additionally, a high Git change count ("10k" in VS Code) was addressed by updating .gitignoreto exclude data/and clearing notebook outputs.

Conclusion

Task 1 successfully delivered a robust dataset of over 1,200 reviews, validated for quality and analyzed through EDA. The results lay a strong foundation for Task 2, where sentiment analysis (e.g., vaderSentiment) and thematic analysis (e.g., LDA with gensim) will uncover user experience in- sights. Future steps include filtering short reviews and refining preprocessing for Task 2. All deliverables are committed to the task-1branch, with documentation in README.md.