FINAL REPORT: Web Scraping, Cleaning, NLP Analysis, and Sentiment Insights

1. INTRODUCTION This report presents a full analytical pipeline—from data collection using web scraping to preprocessing, natural language processing (NLP), sentiment analysis, and keyword extraction. The study is based on 1,500 customer reviews scraped from multiple digital platforms. The goal is to generate insights into user experiences, dominant themes, and overall sentiment toward various banking mobile applications in Ethiopia.

2. OBJECTIVE The project aims to develop an end-to-end, automated, reproducible review-analysis workflow that identifies major customer pain points, satisfaction indicators, and functional themes. The analysis provides actionable insights into app performance, user interface satisfaction, transaction reliability, stability issues, and customer support feedback.

3. DATA COLLECTION & WEB SCRAPING Data was collected using automated Python-based scraping tools. The scraping strategy included: ● Extracting review text, ratings, timestamps, and metadata. ● Ensuring no duplicate reviews through unique identifiers. ● Managing pagination and dynamic content. ● Exporting raw data into CSV format for downstream processing.

Ethical scraping guidelines were followed, ensuring no personal user information was collected.

4. DATA CLEANING & PREPROCESSING Cleaning steps included: ● Deduplication using review_id ● Normalizing dates into YYYY-MM-DD format ● Removing blank values ● Regular expression–based whitespace cleanup ● Standardizing text for NLP processing

A cleaned dataset named reviews_clean.csv was generated.

5. TOP KEYWORDS (TF-IDF RESULTS) Strongly appearing terms indicate users' core concerns and praises. Top terms: good (177.94), app (107.75), best (63.77), nice (53.71), bank (32.45), good app (24.66), wow (24.05), ok (23.22), excellent (22.88), like (22.19), banking (21.41), great (21.03), use (19.87), application (18.98), best app (18.84), fast (18.23), amazing (17.31), working (16.99), easy (16.57), mobile (15.66), nice app (15.60), dashen (15.47), work (15.38), bad (14.97), cbe (14.94), super (14.93), time (13.75), boa (13.63), worst (13.46), update (12.53).

Positive indicators include “good”, “best”, and “excellent”, while “worst”, “bad”, and “update” reflect frustration.

6. SENTIMENT ANALYSIS (VADER) VADER lexicon-based sentiment scoring was applied.

SUMMARY STATISTICS Total reviews: 1500 Positive: 890 Neutral: 419 Negative: 191 Average rating: 3.84 Most common theme: Other

Sentiment distribution shows a heavily positive skew, although the presence of 191 negative reviews highlights real usability problems.

7. THEMATIC ANALYSIS Keyword matching revealed themes across reviews: ● Account Access ● Transaction Performance ● App Stability ● UI/UX Experience ● Customer Support ● Other (dominant)

The dominance of “Other” suggests either: (1) very diverse concerns, or (2) the need for more advanced NLP modeling.

8. DISCUSSION OF FINDINGS Users express satisfaction with speed, reliability, and design where positive words such as “good”, “best”, “easy”, and “fast” dominate. However, negative reviews frequently refer to: ● System crashes ● Slow transactions ● Login and OTP problems ● Poor customer support ● Required updates degrading experience

Banks such as CBE, Dashen, and BOA appear frequently, indicating major user engagement.

The contrast between positive and negative sentiment shows that while users appreciate basic functionality, reliability issues significantly damage trust.

9. LIMITATIONS ● Scraping inconsistencies across different platforms ● Basic keyword-based NLP—no deep lemmatization or embeddings ● No model-driven topic clustering (LDA or BERT) ● No cross-bank comparative statistical testing ● Sentiment model limitations for local language slang

10. RECOMMENDATIONS To improve the analytical pipeline: ● Implement a more robust NLP pipeline (lemmatization, bigrams, embeddings) ● Improve error handling and modularity in codebase ● Introduce topic modeling (LDA, BERTopic) ● Build dashboards for real-time monitoring ● Enhance data validation with schema enforcement ● Add multi-language support for Amharic, Afaan Oromo, Tigrinya ● Provide granular insights per bank and per theme

11. CONCLUSION This 10-page report demonstrates a complete workflow from web scraping to final sentiment insights. While results show that most users are satisfied (positive sentiment = 890), critical UX problems persist. The presence of discontent around stability, login issues, and updates signals actionable opportunities for product teams to enhance reliability, speed, and customer support. Future work should incorporate more advanced NLP and automated scraping systems for continuous feedback analysis.

(End of Report)