

EthioMart Data Extraction Interim Report

Report Date: 2025-07-20

Data Collection Summary

Total Messages: 146

Vendors: 5

Avg Views: 1268.6986301369864

With Images: 146

Labeling Summary

Labeled Messages: 50

Preprocessing Steps:

1. Telegram message scraping using Telethon
2. Image OCR with Tesseract (Amharic)
3. Text cleaning and normalization
4. Rule-based entity labeling for NER training