Interim Report: Amharic E-commerce Data Extractor (B5W4)

Submitted for the 10 Academy B5W4 Challenge

Daniel Shobe

Due: June 22, 2025

Executive Summary & Context

The Amharic E-commerce Data Extractor project aims to build a system for extracting structured product information from Telegram channels in Ethiopia, supporting the expanding e-commerce market. Task 1 involves scraping and preprocessing messages from 22 Telegram channels to create a dataset of products, prices, and locations. Task 2 requires manually labeling 30 messages in CoNLL format for Named Entity Recognition (NER) to identify entities such as products (e.g., PUMA SPIREX), prices (e.g., 4400 ብር), locations (e.g., ሜክሲኮ), and contact information (e.g., 0944222069). As of June 19, 2025, Task 1 is complete, and Task 2 has labeled 7 messages, with a semi-automated process underway to reach 30 by the deadline. This report outlines the data, processes, and challenges, aligning with project goals.

Description of Data & Sources

The dataset includes messages from 22 public Telegram channels focused on Ethiopian e-commerce, such as @Shageronlinestore and @sinayelj. These channels advertise products like footwear (SKECHERS QUANTUM FLEX), clothing (COTTON TISHERTS), bags (የሴቶች ቦርሳ), and electronics (HP ELITEBOOK). Each message contains:

- Text: Mixed Amharic and English, e.g., “PRICE :- 4400 ብር ... አድራሻ :- ሜክሲኮ”.
- Tokens: Word-level tokens generated using NLTK’s word_tokenize.
- Metadata: Channel name, message ID, and images stored in images/.

Raw data is stored in data/raw_messages.json, preprocessed in data/messages.json, and 30 selected messages are in data/selected_messages.json. Challenges included empty messages, mixed-language text, and inconsistent formatting, which were addressed by filtering invalid entries and standardizing tokenization.

Explanation of Process

Task 1: Data Ingestion and Preprocessing

- Scraping: Used Telethon to collect messages from 22 Telegram channels via API authentication, saving raw data to data/raw_messages.json with scripts/scrape_telegram.py.
- Preprocessing: Tokenized text using NLTK’s word_tokenize, removed emojis and special characters, and saved results in data/messages.json. Images were archived in images/.
- Selection: Selected 30 valid messages (non-empty text and tokens) using scripts/select_messages.py, ensuring diversity for Task 2.

Task 2: CoNLL Labeling

- Message Selection: Used scripts/select_messages.py to choose 30 messages, filtering out empty entries, stored in data/selected_messages.json.
- Labeling: Labeled 7 messages in CoNLL format, identifying entities: B-Product/I-Product (e.g., SKECHERS QUANTUM FLEX), B-PRICE/I-PRICE (e.g., 57000 birr), B-LOC/I-LOC (e.g., ድሬዳዋ አሸዋ ሚና ህንፃ), and B-CONTACT_INFO (e.g., httpstmeshewabrand). A rule-based script (scripts/auto_label_conll.py) generates initial labels, followed by manual review.
- Validation: scripts/validate_conll.py ensures the CoNLL file (samples/labeled_data.conll) has 30 messages with valid labels.

Challenges include multi-word entities (የሴቶች ቦርሳ), mixed-language text, and distinguishing events (HellooMarket) from products. The process combines manual and semi-automated labeling to meet the deadline.

Clarity & Professionalism

This report is structured with clear sections, using examples in Amharic (ሜክሲኮ, ቦርሳ) and English. It references scripts (scripts/auto_label_conll.py) and outputs (samples/labeled_data.conll) for transparency. The methodology addresses challenges like empty messages and language complexity. The project is on track to label 23 more messages by June 22, 2025, ensuring a complete submission.