Interim Report: Amharic E-commerce Data Extractor

Author: Abenezer M. Woldesenbet
Date: June 24, 2025
GitHub Repo: https://github.com/abeni505/week4-amharic-ecommerce-extractor/

Introduction - Project Overview and Understanding

This project aims to develop a sophisticated data extraction engine for EthioMart, a platform designed to centralize Telegram-based e-commerce in Ethiopia. The core objective is to transform unstructured, messy Telegram posts into a structured, queryable database. This is achieved by fine-tuning a Named Entity Recognition (NER) model to automatically identify key business entities—specifically Product Names, Prices, and Locations—from Amharic text.

My understanding is that this structured data is the foundational component for a larger FinTech engine. This engine will ultimately analyze vendor activity and customer engagement to create a "Vendor Scorecard," enabling EthioMart to identify promising vendors who are strong candidates for micro-lending opportunities. This interim report details the progress made in the initial data collection and preparation phases.

Methodology - What You Did, How You Did It, and Your Results

The initial phase of the project focused on building a robust data pipeline to source and prepare data for model training. The methodologies and tools employed were chosen to ensure a reproducible and scalable workflow.

Methodologies & Tools Used:

- Programming Language: Python 3.10
- Virtual Environment: Python venv for dependency management.
- Data Ingestion: Telethon library for asynchronous communication with the Telegram API.
- Data Handling: pandas library for data manipulation and storage.
- Development Environment: JupyterLab for interactive coding and analysis.
- Data Annotation: Manual labeling using the CoNLL (Conference on Natural Language Learning) format.

Steps Taken & Results:

Task 1: Data Ingestion and Preprocessing

The first step was to programmatically collect a large corpus of real-world e-commerce posts. A Python script was developed using Telethon to connect to the Telegram API and scrape messages from five prominent Ethiopian e-commerce channels. The script successfully collected text, metadata (timestamps, view counts), and stored the cleaned data in a unified CSV file.
Results: A total of 5,673 messages were collected. The distribution across the channels is as follows:

Channel Name	Number of Messages Scraped
Shageronlinestore	1,845
ethioamazon	1,231
gebeyamart	952
ethioonlinemarket1	890
onlinemarketethiopia	755
Total	5,673

This dataset serves as the primary source of raw data for the entire project, directly contributing to our goal of building a comprehensive e-commerce hub.

Task 2: Data Labeling for NER

To train a supervised NER model, a "gold standard" or "ground-truth" dataset is required. A random sample of 50 messages was selected from the scraped data for manual annotation. The labeling was performed using the B-I-O (Beginning, Inside, Outside) tagging scheme in the CoNLL format. This format is standard for token classification tasks and is readily parsable by modern NLP frameworks like Hugging Face.

The chosen entities for this initial phase are PRODUCT, PRICE, and LOCATION.
Example of a Labeled Sentence:
Original Message: Adjustable Posture Corrector ዋጋ፦ ✅ 800 ብር
CoNLL Formatted Labels:

Sample 1:
Adjustable B-PRODUCT
Posture I-PRODUCT
Corrector I-PRODUCT
ዋጋ፦ O
✅ O
800 B-PRICE
ብር I-PRICE

This meticulously labeled file, labeled_data.conll, is the most critical asset produced to date, as it will directly serve as the training and evaluation data for our NER mo

Challenges & Solutions – What Issues Came Up, How You Addressed Them

Challenge 1: Inconsistent Message Formatting

- Problem: Telegram posts are highly unstructured, containing a mix of Amharic, English, emojis, and varied formatting for prices and locations.
- Solution: During the labeling phase, a strict internal guideline was developed. For example, currency symbols and words like "·ብር" were consistently included as part of the PRICE entity (I-PRICE). Emojis and non-essential punctuation were designated as O (Outside). This consistency is crucial for model performance.

Challenge 2: Ambiguity in Entities

- Problem: Identifying the precise boundaries of entities was challenging. For instance, determining if a location name was one word or multiple words required careful consideration.
- Solution: The B-I-O tagging scheme was instrumental in solving this. By using B-LOC for the beginning of a location and I-LOC for subsequent parts (e.g., ስሪ ኤም ሲቲ ሞል), we can teach the model to recognize multi-word entities.

Future Plan - What's Left and How You Plan to Finish

The foundational data work is now complete. The project will now transition into the machine learning phase.

Remaining Tasks:

1. Task 3: Fine-Tune NER Model: Use the labeled_data.conll file to fine-tune a pre-trained transformer model (e.g., xlm-roberta-base) for Amharic NER.
2. Task 4: Model Comparison & Selection: Experiment with and compare multiple models (e.g., mBERT, bert-tiny-amharic) based on performance metrics (F1-score, precision, recall) to select the optimal model for production.
3. Task 5: Model Interpretability: Use tools like SHAP or LIME to understand and explain the model's predictions, ensuring transparency and building trust in the system.
4. Task 6: FinTech Vendor Scorecard: Develop the final analytics engine to process the NER outputs and post metadata, calculating a "Lending Score" for each vendor.

Conclusion - Summary of Current Progress and Confidence Going Forward

To date, this project has successfully established a data collection pipeline and produced a high-quality, manually labeled dataset, which are the essential prerequisites for the subsequent machine learning tasks. The initial challenges in data handling have been effectively addressed, providing a solid foundation for the next steps.

I am highly confident in my ability to complete the remaining tasks. The path forward is clear, and the foundational work completed ensures that the focus can now shift entirely to building and evaluating a powerful NER model to achieve the project's ultimate goal of creating a smart FinTech engine for EthioMart.