Interim Report – Building an Amharic E-commerce NER System for EthioMart

Date: June 22, 2025

1. Project Overview

EthioMart aims to centralize e-commerce activity from multiple Telegram channels into a single platform. This project supports that goal by building a Named Entity Recognition (NER) system to extract business-relevant entities (like product names, prices, and locations) from Amharic messages.

Our system transforms unstructured Telegram messages into structured, machine-readable data — laying the foundation for vendor analytics and micro-lending decisions.

2. Data Collection & Ingestion

I implemented a custom **Telegram scraper** using `Telethon`, capable of extracting messages (text + media) from multiple public e-commerce Telegram channels, such as:

- `@Shageronlinestore`
- `@ZemenExpress`
- `@nevacomputer`
- `@meneshayeofficial`
- `@ethio_brand_collection`
- `@Leyueqa`
- `@helloomarketethiopia`

Each message was stored in `telegram_data.csv`, including fields like:

- Channel name and username
- Message ID and timestamp
- Text content
- Downloaded media (image paths)

> ✅ Over 1000+ messages were collected across 7 channels.

3. Text Preprocessing

To prepare the data for NER labeling:

- Normalized Amharic punctuation (፡, ፡) and spacing
- Removed noise (emojis, extra whitespace, non-Amharic characters)
- Stored clean messages in `telegram_data_cleaned.csv`

This cleaned version is used for labeling and downstream model training.

4. Manual Data Labeling for NER

A sample of **70 messages** was manually labeled using the **CoNLL format**, with entity types:

- `Product`
- `Price`
- `Location`

The final labeled file (`ner_data.conll`) is ready for fine-tuning. Each token is labeled with BIO scheme (e.g., `B-Product`, `I-PRICE`, `O`).

> Example snippet: Skechers B-Product OD I-Product Price B-PRICE 3000 I-PRICE

5. Deliverables Summary (So Far)

Deliverable	Status
Telegram scraper	✅ Implemented
Data ingestion notebook	✅ Complete
Preprocessed text data	✅ Saved
CoNLL labeling (NER dataset)	✅ Done
GitHub Repo	🔗 Included

6. Next Steps (Task 3–6)

- Fine-tune transformer models (XLM-Roberta, Bert-Amharic)
- Evaluate performance on validation set
- Apply SHAP/LIME to interpret predictions
- Develop vendor analytics engine for lending decisions

Author: Mesfin Mulugeta Wetere

https://github.com/5237-mests/Amharic-E-commerce-Data-Extractor KAIM Week 4 | 10 Academy AI Mastery Program