

# Interim Report – B5W4: Amharic E-commerce Data Extractor

**Submitted by:** Yamlak Negash

**Project:** EthioMart NER System – Interim Submission

**Date:** June 22, 2025

---

## Executive Summary

The EthioMart initiative aims to centralize e-commerce activity happening through Telegram in Ethiopia by aggregating content from various vendor channels. The key challenge is extracting structured information such as product names, prices, and locations from unstructured Amharic messages. This interim report presents progress on Tasks 1 and 2: data ingestion/preprocessing and CoNLL-style manual labeling, forming the foundation of a Named Entity Recognition (NER) system tailored for Amharic e-commerce.

---

## Data Description and Source

Data for this project was obtained from active Ethiopian e-commerce Telegram channels and compiled into a CSV file ([telegram\\_data.csv](#)). The primary column of interest is [message](#), containing unstructured Amharic promotional content. The initial dataset consists of:

- Raw product descriptions
- Mentions of locations (e.g., delivery areas)
- Price information in birr

**Metadata fields** included:

- Message text
- (Where available): timestamps, post ID, sender name

## Preprocessing Steps:

- Normalized spacing and removed non-Amharic punctuation/symbols
- Filtered out empty or meaningless rows
- Added a new column `cleaned_text` that preserves useful Amharic content
- Tokenization considered Amharic-specific word boundaries (later used in labeling)

Cleaned data was saved in: `data/preprocessed_data.csv`

---

## Explanation of Labeling Process

To create training data for the NER system, 50 sample messages were selected from the cleaned dataset and labeled manually using the **CoNLL format**. The goal was to tag the following entities:

- `B-Product`, `I-Product`: Product name phrases
- `B-PRICE`, `I-PRICE`: Monetary amounts
- `B-LOC`, `I-LOC`: Locations (cities, neighborhoods)
- `O`: Tokens that don't belong to any entity

### Labeling Process:

1. Randomly sampled diverse messages with at least one identifiable entity.
2. Tokenized text manually in Amharic, respecting phrase boundaries.
3. Applied standard IOB tagging (Begin, Inside, Outside).
4. Saved output in `data/labeled_data.conll` following CoNLL standards.

Example:





ጥጋወ B-PRICE

1000 I-PRICE  
ገቢ I-PRICE  
እና O  
ምርቱ B-Product  
ሻይ I-Product  
በ B-LOC  
አዲስ B-LOC  
አበባ I-LOC

This labeled dataset will serve as input for fine-tuning a transformer-based model (e.g., XLM-R, mBERT) in the next stage.

---

## Summary of Progress

-  Completed ingestion and preprocessing of 1000+ Telegram messages
  -  Cleaned and saved usable text for NER tasks
  -  Labeled 50 messages with key entities in CoNLL format
  -  Ready for token alignment and model fine-tuning in next phase
- 

## Next Steps

- Fine-tune pre-trained language models (XLM-R, AfroXMLR) using Hugging Face Transformers
  - Evaluate performance with F1-score and interpret results with SHAP/LIME
  - Generate a vendor activity scorecard for micro-lending suitability based on structured outputs
- 

**End of Interim Report**