Interim Report: Amharic E-Commerce NER Project

1. Data Collection Summary

- **Channels Scraped:**

  ZemenExpress,
  helloomarketethiopia, FashionTera,
  modernshoppingcenter, Shewabrand

- **Total Messages:**
  220 messages
  50 messages per major channel
  20 messages from smaller channels

- **Sample Raw Message:**

"ላፕቶፕ 2500 ብር በቦሌ - የቤት ዕቃዎች በ1000 ብር"

2. Preprocessing Steps

1. Cleaning

Removed emojis, phone numbers, links, and unnecessary punctuation.

2. Text Normalization

Standardized price formats and added spacing (e.g., "2500ብር" → "2500 ብር").

3. Example Output

Before: "💥ላፕቶፕ 2500ብር📞0912345678"

After: "ላፕቶፕ 2500 ብር"

3. Labeling Progress

Labeled Samples: 30

(5 labeled messages per channel)

Entity Distribution:

PRODUCT: 45%

PRICE: 30%

LOCATION: 25%

Example Labeled Sentence:

"ላፕቶፕ 2500 ብር በቦሌ"
Tags: B-PRODUCT I-PRODUCT B-PRICE I-PRICE B-LOC

4. Challenges & Next Steps

Challenges Encountered:

Mixed Amharic-English tokens in the same sentence

Inconsistent price formats (e.g., "በ2000", "2000 ብር", "በ2000 ብር")

Limited labeled data

Immediate Next Steps:

Label 50+ additional samples for balanced training data

Begin training initial NER model (Model v1)

Planned Timeline:

Labeling Completion: [Insert date]

Model v1 Training: [Insert date]

Evaluation & Feedback: [Insert date]

Prepared by: Yalembrhan Kelayneh
Date: June 23, 2025
Project: 10Academy – Week 4 Challenge