Final Report – Amharic E-commerce NER + Vendor Scorecard for EthioMart

Date: June 24, 2025
Prepared by: Mesfin Mulugeta Wetere

1. Project Overview

EthioMart seeks to centralize e-commerce activity from Telegram into a structured platform, enabling smarter business insights and financial services like micro-lending. This project builds a Named Entity Recognition (NER) system to extract key entities (Product, Price, Location) from Amharic-language Telegram messages.

Beyond entity extraction, the system powers a Vendor Analytics Engine, offering a data-driven scorecard to assess vendors' suitability for small business loans.

2. Data Collection & Ingestion

A custom Telegram scraper was implemented using Telethon. It captured:
- Text, media, views, timestamps, and message metadata
- From 7 major Telegram vendors, including @Shageronlinestore, @ZemenExpress, and others

✅Over 1000+ messages were scraped and stored in telegram_data.csv.

3. Text Preprocessing

- Cleaned Amharic punctuation (፡, ።), whitespace, and noise
- Resulting in a clean dataset (telegram_data_cleaned.csv) for labeling and modeling

4. NER Data Labeling

- 70 manually labeled messages in CoNLL format
- Entities: Product, Price, Location with BIO tagging scheme

Example:

Skechers	B-Product
OD	I-Product
Price	B-PRICE
3000	I-PRICE

5. Model Comparison & Selection – Task 4

We experimented with multiple pre-trained models for NER in Amharic.

Model	F1 Score	Precision	Recall	Training Time	RAM Usage	Notes
XLM-Roberta-base	0.96	0.97	0.96	3 hr 55 min	High	Best accuracy, reliable, robust
rasyosef/bert-tiny-amharic	0.002	0.01	0.001	3 mins	Very Low	Severely underfits
masakhane/afroxlmr-large-ner-masakhaner	-	-	-	Not trained	Exceeded Colab RAM	Needs enterprise GPUs

✅Conclusion:
- XLM-Roberta-base was selected for production.
- It provides the best balance of performance and feasibility.

6. Vendor Analytics Engine

✅Metrics Computed Per Vendor: Posts/Week, Average Price (ETB), Lending Score

Vendor	Posts/Week	Avg Price (ETB)	Lending Score
Sheger online-store	52.24	0	36.57
Zemen Express®	43.48	0	30.43
NEVA COMPUTER®	9.55	0	6.68
መነሻዬ	6.40	0	4.48
EthioBrand®	10.61	0	7.42
ልዩ ኢቃ	42.17	0	29.52
HellooMarket	13.83	0	9.68

7. Key Insights

- Most Active Vendors: Sheger online-store, Zemen Express®, ልዩ ኢቃ
- Underperformers: NEVA COMPUTER®, መነሻዬ
- NER Gap: Price extraction failed due to inconsistent formatting in Telegram posts

8. Recommendations

1. Micro-Lending: Prioritize Sheger online-store, Zemen Express®, and ልዩ እቃ
2. NER Improvement: Enhance price detection with hybrid regex + NER models
3. Vendor Engagement: Encourage vendors to include prices clearly in their posts
4. Deployment Next: Develop this as an API or web dashboard for EthioMart

9. Conclusion

This report delivers a complete pipeline from data scraping to entity extraction to vendor analytics for micro-lending support.

✅Selected Model: XLM-Roberta-base — providing excellent performance for Amharic NER.
✅Vendor Scorecard: Offers actionable insights for EthioMart's financial decision-making.

□ Project Repository: https://github.com/5237-mests/Amharic-E-commerce-Data-Extractor