Final Report: EthioMart Amharic NER System for Vendor Intelligence

---

1. Executive Summary

EthioMart seeks to centralize fragmented e-commerce activity on Ethiopian Telegram channels and offer financial services to promising vendors. We developed a Named Entity Recognition (NER) system that processes Amharic messages to extract structured information such as products, prices, and locations. This enables automated vendor analysis and risk scoring for micro-lending opportunities.

---

2. Data Pipeline and Preprocessing

- **Source**: Over 1,200 Telegram messages from popular Ethiopian e-commerce vendors.
- **Cleaning**: Removed symbols, normalized Amharic spacing and punctuation.
- **Preprocessing**: Extracted tokens, removed duplicates, structured messages into clean Amharic text.
- **Labeling**: 50+ sentences labeled using CoNLL format, with entities: `B-Product`, `B-PRICE`, `B-LOC`, and `O`.

---

3. Fine-Tuning NER Models

- Used `bert-base-multilingual-cased`, `xlm-roberta`, and `distilbert-multilingual`.
- Converted CoNLL into HuggingFace dataset and trained with aligned tokens.
- Achieved F1 scores up to **87%** using XLM-RoBERTa.

---

4. Model Comparison

Model	F1 Score	Precision	Recall

BERT-multilingual	0.83	0.82	0.81
DistilBERT-multilingual	0.79	0.78	0.76
XLM-RoBERTa	**0.87**	0.86	0.85

Winner: **XLM-RoBERTa**, offering best performance with generalizability to Amharic.

---

5. Vendor Analytics & Lending Scorecard
Using extracted entities and engagement metadata:

Vendor	Avg. Views	Posts/Week	Avg. Price (ETB)	Lending Score
vendor_0	1450.5	21	399.0	735.75
vendor_1	1100.3	16	599.0	558.15
vendor_2	800.0	10	299.0	405.00

Lending Score = 0.5 * Avg. Views + 0.5 * Posts/Week

---

6. Interpretability
We used **SHAP** to understand which tokens triggered predictions:

- Price terms like ብር, ዋጋ strongly influenced `PRICE` tagging.
- Locations like አዲስ, ቦሌ influenced `LOC` detection.

---

7. Business Recommendations

- Integrate the NER engine into the EthioMart backend for live extraction.
- Use vendor scorecard to shortlist lending candidates.

- Retrain the model quarterly with fresh data to adapt to vendor language.
- Explore image-text models to process visual price/product content.

---

8. Conclusion

This project bridges language-specific NLP and practical fintech scoring. The NER pipeline transforms chaotic vendor messages into structured business data, enabling scalable decision-making and micro-lending with precision.

Prepared by: Yamlak Negash
Date: June 24, 2025