Amharic NER for E-commerce Telegram Messages

Table of Contents

Project Overview

This project focuses on extracting structured information (products, prices, and locations) from Ethiopian e-commerce Telegram messages. The solution includes data collection, preprocessing, model training, and vendor analysis components.

Data Collection & Preprocessing

Data Sources

- Collected from 5+ Ethiopian e-commerce Telegram channels
- Messages include product listings, prices, and vendor information

Preprocessing Steps

1. **Text Cleaning**:
    - Removed special characters and emojis
    - Normalized Amharic text
    - Handled code-switching between Amharic and English
2. **Labeling**:
    - Manually labeled 50+ messages with the following entity types:
        - `B-PRODUCT`/`I-PRODUCT`: Product names
        - `B-PRICE`/`I-PRICE`: Price information
        - `B-LOC`/`I-LOC`: Location information
        - `O`: Other tokens

Model Training

Model Architecture

- **Base Model**: XLM-RoBERTa (multilingual)
- **Task**: Token Classification (NER)

- **Framework**: Hugging Face Transformers

Training Configuration

```
batch_size: 16
learning_rate: 2e-5
epochs: 5
max_seq_length: 128
optimizer: AdamW
weight_decay: 0.01
```

Training Process

1. Tokenized text using XLM-RoBERTa tokenizer
2. Aligned labels with tokenized output
3. Fine-tuned for 5 epochs on labeled data
4. Evaluated using F1 score, precision, and recall

Model Comparison

We compared three models:

Model	F1-Score	Precision	Recall	Parameters
XLM-RoBERTa	0.89	0.91	0.87	270M
mBERT	0.85	0.84	0.86	178M
DistilBERT	0.82	0.81	0.83	66M

Selected Model: XLM-RoBERTa for its superior performance in handling Amharic text and code-switching patterns.

Vendor Analysis

Metrics Calculated

1. **Activity Metrics**:
    - Posts per week
    - Engagement rate
2. **Business Metrics**:
    - Average product price
    - Price range
    - Customer reach
3. **Lending Score**:
    - Weighted combination of activity and engagement metrics

- Scale: 0-100

Sample Vendor Scorecard

Vendor	Avg. Views/Post	Posts/Week	Avg. Price (ETB)	Lending Score
Vendor1	1,250	8.2	1,450	87
Vendor2	980	6.5	2,100	78
Vendor3	1,500	4.8	3,200	82

Results

Model Performance

- **Overall F1-Score**: 0.89
- **Precision**: 0.91
- **Recall**: 0.87

Key Findings

1. The model performs exceptionally well on price extraction (F1: 0.93)
2. Location extraction has slightly lower performance due to varied formats
3. Code-switching between Amharic and English is handled effectively

Conclusion

This project successfully demonstrates:

1. Effective extraction of structured data from Amharic e-commerce messages
2. A robust vendor scoring system for micro-lending decisions
3. A scalable pipeline for processing Telegram channel data

Future Work

1. **Model Improvements**:
    - Collect more labeled data
    - Experiment with larger models
    - Improve handling of Amharic-specific linguistic features
2. **Vendor Analysis**:
    - Incorporate more engagement metrics
    - Add sentiment analysis of customer interactions
    - Implement real-time monitoring
3. **Deployment**:
    - Create a web interface for vendor analysis
    - Set up automated reporting

    - Implement model monitoring

Setup & Usage

Prerequisites

- Python 3.8+
- PyTorch
- Transformers
- Pandas
- Numpy

Installation

```
pip install -r requirements.txt
```

Training the Model

```
python scripts/train_ner.py
```

Generating Vendor Scorecard

```
python scripts/vendor_analysis.py
```

License