Entity Recognition (NER) for Amharic Telegram Messages

A Blog-style Report

Introduction

Named Entity Recognition (NER) is a key component of Natural Language Processing (NLP) that identifies and classifies entities in text, such as names, locations, organizations, and products. This project focuses on extracting entities from Amharic Telegram messages. Given Amharic's low-resource status in NLP, this project leverages advanced multilingual models like **XLM-Roberta** to achieve robust results.

This report outlines the data preparation process, model selection criteria, fine-tuning procedures, evaluation metrics, and interpretability analysis for the NER task.

---

1. Data Preparation

Data Source

The raw dataset comprises Telegram messages collected from Amharic channels. The dataset includes noisy, user-generated text with varying structures and entity types such as:

- **Products** (e.g., "አልጋ"),
- **Prices** (e.g., "100 ብር"),
- **Locations** (e.g., "አዲስ አበባ").

Data Cleaning

- **Preprocessing Steps**:
    - Removed URLs using regex patterns (e.g., `http\S+` or `www\S+`).
    - Tokenized Amharic text using a custom tokenizer that respects the script's unique grammar.
    - Removed redundant symbols, emojis, and stopwords.
- **Challenges**:
    - Amharic text often contains overlapping entities, and spacing inconsistencies complicated tokenization.

Labeling Data

Labeled entities were manually annotated in CoNLL format, a standard format for NER tasks. Example:

አዲስ B-LOC
አበባ I-LOC
በትክክለኛ O
ዋጋ O

The dataset was split into training (80%), validation (10%), and test (10%) sets.

---

2. Model Selection

Given the project's multilingual and low-resource requirements, the following models were considered:

1. **XLM-Roberta**:
    - Multilingual transformer model supporting Amharic.
    - Excels in handling low-resource languages.
    - High computational cost.
2. **mBERT (Multilingual BERT)**:
    - Pretrained on multiple languages.
    - Moderate accuracy for Amharic but slower inference speed.
3. **DistilBERT**:
    - Lightweight and faster.
    - Suitable for deployment on devices with limited computational resources.

Criteria for Selection:

- **Performance**: Accuracy and F1-score on validation set.
- **Speed**: Inference time per message.
- **Scalability**: Ability to handle noisy, unstructured Amharic text.

Decision: XLM-Roberta was selected for fine-tuning due to its superior performance despite higher computational costs.

---

3. Fine-Tuning Procedure

Environment Setup

- **Google Colab** with GPU enabled for faster training.
- Libraries: Hugging Face Transformers, Datasets, and PyTorch.

Steps:

1. **Dataset Loading**:
    - The labeled dataset in CoNLL format was converted to Hugging Face's `datasets` format for easier handling.
    - Tokenized text using XLM-Roberta's tokenizer, aligning labels to tokens.
2. **Training Configuration**:
    - **Learning Rate**: `5e-5`.
    - **Batch Size**: `16`.
    - **Epochs**: `5`.
    - **Evaluation Strategy**: Validation at the end of each epoch.
3. **Training**:
    - Hugging Face's `Trainer` API was used to streamline the fine-tuning process.
    - Model checkpoints were saved at regular intervals.
4. **Evaluation**:
    - The model was evaluated on the validation set using metrics like **Precision**, **Recall**, and **F1-Score**.
    - Predictions were post-processed to handle overlapping entities.

Results:

Metric	Value (%)
Precision	94
Recall	93
F1-Score	93.5

---

4. Model Comparison

To validate the choice of XLM-Roberta, we compared it with other models:

Model	Precision (%)	Recall (%)	F1-Score (%)	Inference Time
XLM-Roberta	94	93	93.5	0.4s/text

mBERT	88	87	87.5	0.5s/text
DistilBERT	90	89	89.5	0.3s/text

Insights:

- XLM-Roberta provided the best performance.
- DistilBERT was faster but less accurate.
- mBERT struggled with Amharic's complex morphology.

---

5. Model Interpretability

To ensure transparency and trust in the system, interpretability tools were applied:

1. **SHAP (SHapley Additive exPlanations)**:
    - Showed feature contributions for each token in identifying entities.
    - Example: SHAP highlighted አዲስ as a key indicator for a **Location (LOC)** entity.
2. **LIME (Local Interpretable Model-agnostic Explanations)**:
    - Generated perturbed text samples to analyze model predictions.
    - Insights into ambiguous cases (e.g., በአቅራቢያ being classified as both **Price** and **Location**).

Challenges Identified:

- Overlapping entities (e.g., "አዲስ አበባ" contains both **Location** and **Product** in some contexts).
- Ambiguous tokens (e.g., ትክክለኛ could refer to both product quality and price).

---

6. Conclusion

This project demonstrated the potential of advanced multilingual models like XLM-Roberta in handling low-resource languages such as Amharic for NER tasks. The model achieved an impressive F1-score of 93.5%, outperforming alternatives like mBERT and DistilBERT.

Interpretability analysis highlighted areas for improvement, such as better handling of ambiguous and overlapping entities.

Recommendations:

- Explore custom embeddings tailored for Amharic.
- Experiment with hybrid models that combine rule-based and deep learning approaches.
- Enhance interpretability further with visual tools for end-user understanding.
