Entity Extraction for Amharic E-commerce Telegram Channels using LLM Fine-Tuning

This project focuses on building a real-time data ingestion and entity extraction pipeline for Amharic messages from Ethiopian e-commerce Telegram channels. The system leverages fine-tuned Large Language Models (LLMs) to identify key business entities such as product names, prices, and locations. The extracted information is used to populate a centralized platform for EthioMart, aiming to streamline e-commerce activities in Ethiopia by consolidating decentralized Telegram channels into a unified hub. The project also includes handling Amharic-specific linguistic features and evaluating model performance for Named Entity Recognition (NER).

Project directory strucutres

The repository is organized into the following directories:

`.github/workflows`: Contains GitHub workflow configurations for continuous integration.

`.vscode`: Configuration files for the Visual Studio Code editor.

`fonts`: Contains files related to analyzing the most common Amharic words in telegram e-commerce messages.

`notebooks`: Jupyter notebooks that were used for data exploration, preprocessing, and labeling tasks.

`scripts`: Python scripts used for data scraping, preprocessing, and implementing the NER labeling logic.

`tests`: Contains test cases for different parts of the project.

Installation Instructions

To run the project locally, follow these steps:

PROF

Clone the Repository:

```
git clone https://github.com/epythonlab/amharic-telegram-ecommerce-entity-extraction.git
```

```
cd amharic-telegram-ecommerce-entity-extraction
```

Set up the Virtual Environment:

```
python3 -m venv .venv
source .venv/bin/activate # For Windows: .venv\Scripts\activate
```

Install Dependencies:

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