Week 5 Task-1 & Task-2 Documentation

1. Data Collection

We collected messages from five Ethiopian e-commerce Telegram channels:

- @qnashcom
- @Fashiontera
- @kuruwear
- @gebeyaadama
- @MerttEka
- @forfreemarket

Process:

- Utilized the Telethon library to connect to Telegram channels and fetch the latest 100 messages per channel.
- Extracted data fields: `channel`, `sender`, `text`, and `timestamp`.

Outcome:

- Successfully retrieved 374 messages.
- Stored the raw data in a structured JSON file: `data/raw/telegram_data.json`.

2. Data Preprocessing

To prepare the raw data for NER tasks, we performed the following steps:

- **Tokenization:** Split text into individual tokens for analysis.
- **URL Removal:** Stripped hyperlinks to reduce noise.
- **Punctuation Removal:** Eliminated unnecessary symbols for cleaner input.
- **Whitespace Normalization:** Standardized spacing for consistent formatting.

Challenges Addressed:

- Handled missing or null values in the `text` field by filtering them out.
- Addressed Amharic-specific linguistic features using appropriate preprocessing rules.

Output:

- Cleaned and structured data saved in: `data/processed/preprocessed_telegram_data.json`.

Data Labeling Summary

1. Labeling Strategy

To prepare the dataset for Amharic NER fine-tuning, we labeled a subset of the data using the CoNLL format.

Entity Types:

- `B-Product` and `I-Product`: Product names (e.g., “ብባርነን ቡርንግን”).
- `B-LOC` and `I-LOC`: Locations (e.g., “Addis Abeba”, “Bole”).
- `B-PRICE` and `I-PRICE`: Prices (e.g., “ሃገ 100 ቡር”).
- `O`: Tokens outside entities.

2. Labeling Procedure

- Selected 30 messages from the preprocessed dataset.
- Manually annotated entities using the CoNLL format:

Example:

ብባርነን	B-Product
ቡርንግን	I-Product
ሃገ	B-PRICE
100	I-PRICE
ቡር	I-PRICE

Tools Used:

- Python scripts for loading and formatting messages.
- Text editors for manual annotation.

Output:

- Saved the labeled dataset in `data/labeled/ner_labels.conll`.

Current Status and Next Steps

Completed Tasks:

1. Data collection from Telegram channels.
2. Preprocessing of raw data.
3. Manual labeling of a subset of the dataset for NER fine-tuning.

Conclusion: The data preparation and labeling stages have been successfully completed, establishing a strong foundation for the subsequent fine-tuning and deployment phases. This progress aligns with EthioMart's vision of creating a robust centralized e-commerce platform for Ethiopia.