Amharic NLP Projects: Instructions and Guidelines

🌍 1. Amharic Text Classification (e.g., News or Social Media Posts)

📌 Objective:

Build a machine learning model that classifies Amharic text (e.g., news headlines or Facebook posts) into predefined categories such as Politics, Sports, Entertainment, etc.

🛠️ Tools:

- Python

- scikit-learn (Naive Bayes, Logistic Regression)

- Tokenization (basic whitespace or regex-based)

- Optional: fastText for embedding-based classification

✅ Instructions:

1. Collect Dataset

- Scrape or gather a dataset of Amharic news or social media posts.

- Each entry should be labeled with a category.

Example

```
text, label
"የኢትዮጵያ አርሶ አደሮች በገበሬ ቀን ደስታ አመሰገኑ", "Politics"
```

2. Preprocess the Text

- Normalize text.

- Tokenize using whitespace or regex.

- Remove stopwords.

3. Feature Extraction

- Use TfidfVectorizer or CountVectorizer.

- Optionally use fastText embeddings.

4. Model Training

- Train Naive Bayes or Logistic Regression.

- Evaluate with accuracy, precision, recall, F1.

5. Save and Test

- Save the model using joblib.

- Test classification with new input.

📌 Guidelines:

- Use consistent labels.

- Ensure enough examples (200+ per class).

- Use cross-validation.

🔤 2. POS Tagger for Amharic (Tiny Version)

📌 Objective:

Tag words in a sentence with POS like noun, verb, adjective, etc.

🛠️ Tools:

- Python

- Option 1: Rule-based (dictionary + suffix)

- Option 2: BiLSTM (TensorFlow or PyTorch)

✅ Instructions:

Option 1: Rule-Based POS Tagger

1. Create a lexicon with common words and tags.

2. Define suffix rules for unknown words.

3. Loop through the sentence and tag each word.

- Example:

```
lexicon = {
    "ሰው": "NOUN",
    "ይሄዳል": "VERB",
    "ትልቅ": "ADJ"
}
```

Option 2: BiLSTM Tagger (Advanced)

- Prepare word-tag formatted data.

- Train using a BiLSTM model.

📌 Guidelines:

- Keep the tagset small (NOUN, VERB, ADJ, etc.)

- Focus on short, simple sentences.

🗣️ 3. Amharic Chatbot (Rule-based)

📌 Objective:

Build a chatbot that handles greetings, weather, and simple questions.

🛠️ Tools:

- Python

- if/else logic or dictionaries

- Optional: Rasa for expansion

✅ Instructions:

1. Define Use Cases (e.g., greetings, time, weather).

- Example user intents:
  - Greeting: "ሰላም", "እንዴት ነህ?"
  - Weather: "ዛሬ የአየር ንብረት እንዴት ነው?"

2. Create a dictionary of input-response pairs.

```
responses = {
    "ሰላም": "ሰላም! እንኳን ደህና መጡ!",
    "እንዴት ነህ": "ደህና ነኝ። አንተስ?",
```

3. Write input handling logic (e.g., match with regex).

4. Improve with regex patterns.

📌 Guidelines:

- Start small (10–20 intents).

- Normalize input (punctuation, casing).

- Expand to include slot-filling later.

Deliverables.

1. Project work report explaining your approaches, methods, and results
2. Working (runnable) code

Deadline: May 24,2025