Project Proposal: Hate Speech Detection in Myanmar Language Using Deep Learning

1. Problem Statement

Hate speech in online platforms written in Myanmar language poses a growing threat to social stability and digital well-being. The need for automatic detection of such content is critical, especially for under-resourced languages like Myanmar where moderation tools are lacking.
Population: All Myanmar-language user-generated content posted on public digital platforms such as Facebook, YouTube comments, forums, and news article comment sections.
Sample Population:
Based on the **myHateSpeech** dataset (GitHub Repository), Myanmar hate speech can be categorized into the following **9 distinct classes**:

- ab: Abusive
- re: Religious hate
- ra: Racial or ethnic hate
- bo: Body shaming
- po: Political hate speech
- se: Sexually offensive
- le: Lethal speech
- ed: Educational hate speech
- no: No hate speech

Due to the limited time frame of **3 weeks**, this project will focus only on **3 classes**:

- any 2 hated speech and
- no: no hate speech

These two types of hate speech are prevalent in online Myanmar content and can be reliably collected and labeled within the given timeline. The target sample size is **1,500–2,000** labeled sentences, with a **balanced class distribution** between any 2 hated speech classes, and a control group of **"non-hate"** samples for contrast.

---

2. Input and Output

- **Input**: Raw Myanmar-language sentences (from user input)
- **Output**: Predicted class label: any 2 hated speech classes, or not-hate

Data Collection:

- Sources include from online social media or webpage
- Existing dataset reuse: The myHateSpeech dataset from GitHub https://github.com/ye-kyaw-thu/myHateSpeech will be partially reused. Only the relevant classes (2 hated speech classes) will be extracted for training and validation purposes.
- Synthetic data augmentation: If the dataset is insufficient for training a deep learning model, additional Myanmar languate hate speech and non-hate samples may be synthetically generated using AI to balance clas representation. These AI-generated samples will be carefully reviewed and manually verified to ensure linguistic accuracy and label quality.

Labeling Strategy:

- A subset of the data will be **manually labeled** according to the guidelines from the original dataset creators.

---

3. Dataset Type and Collection Method

- **Dataset Type**: Myanmar-language hate speech dataset (multi-class), focused on any 2 hated speech and non-hate.

- **Collection Method**:
- Reuse and filter from the open-source **myHateSpeech** dataset (focusing on 2 classes).

- Supplement with additional data collected via **web scraping** from public sources and

- Apply **synthetic data augmentation**, if needed, by generating Myanmar language samples using AI and manually verifying them.

---

4. Expected Performance

Given the focused classification of two hate speech types and one control class (non-hate), performance goals are:

- **Accuracy**: 60–85%

- **F1-score** (per class): 0.60 or higher

---

5. Motivation

This project is motivated by the increasing prevalence of hate speech in Myanmar's digital spaces and the limited availability of tools that can detect such content automatically.
By narrowing down to any 2 hated speech classes, this project will:

- Make meaningful contributions in a short timeline.

- Provide a starting point for scalable detection of all 9 classes in the future.

- Demonstrate practical use of **deep learning and NLP** methods in low-resource languages.