**Project Proposal: Named Entity Recognition (NER) for Burmese Language**

1. **Problem Statement**

With the rapid advancement of Artificial Intelligence (AI), especially through the rise of Large Language Models (LLMs), **Natural Language Processing (NLP)** has become a foundational component of intelligent systems. Among core NLP tasks, **Named Entity Recognition (NER)** plays a critical role in understanding and extracting structured information from unstructured text.

NER enables applications such as:

* **Information Extraction** – Pulling key details like names, dates, and locations from documents
* **Document Classification** – Enriching metadata and automating categorization
* **Search Engine Optimization** – Improving relevance in user queries
* **Recommendation Systems** – Personalizing results based on extracted context
* **Social Media Monitoring** – Tracking sentiment, trends, or named mentions

In essence, NER identifies and classifies entities into predefined categories such as **PERSON, ORGANIZATION, LOCATION, DATE**, and more. For example:

“Apple Inc. was founded by Steve Jobs in California.”  
→ "Apple Inc." → ORGANIZATION, "Steve Jobs" → PERSON, "California" → LOCATION

* 1. **Challenges in Burmese NER**

While many libraries (e.g., SpaCy, genism, Flair) offer strong support for English and other high-resource languages, **NER for Burmese remains underdeveloped** due to several key limitations:

* Absence of large-scale annotated corpora
* No standard labelling schema or annotation guidelines
* Limited availability of pretrained models or NLP tools
* Complex linguistic structure
* High context-dependency and ambiguous terms

For instance, in the Burmese sentence:

"ရန်ကုန်မြို့ရှိ ရန်ကုန်တက္ကသိုလ်တွင် တက္ကသိုလ်ကျောင်းသားများ စုပေါင်းဆန္ဒပြခဲ့သည်။"

The phrase "ရန်ကုန်မြို့" is a **LOCATION**, while "ရန်ကုန်တက္ကသိုလ်" is an **ORGANIZATION**. Despite sharing the token "ရန်ကုန်", a model must learn to distinguish between these based on context, a task which is non-trivial in low-resource settings.

* 1. **Why it matters**

This lack of NER support severely limits the development of NLP applications for the Burmese-speaking community. Imagine a law firm in Myanmar needing to review thousands of case documents. Without NER, searching for a client name or event date becomes a **manual, time-consuming, and error-prone** process.

By contrast, a well-trained NER system can automatically extract, categorize, and highlight named entities—**saving time, reducing mistakes, and enabling scalable data analysis**.

NER for Burmese is not just a technical gap—it is a **linguistic and societal need**. As Myanmar becomes more digitally connected, foundational tools like NER are essential for enabling advancements in **legal tech, public policy, education, journalism, and digital governance**.

#### ****Objective****

The primary objective of this project is to **explore, design, and evaluate a Burmese NER system** using a range of **deep learning models**, **transformers and transfer learning techniques**. The study will address the following key goals:

1. **Develop a Burmese NER dataset** by defining a labelling schema suitable for Burmese language structure
2. **Experiment with different NER architectures**
3. **Investigate the effectiveness of transfer learning** from multilingual models in the context of low-resource Burmese
4. **Analyse the challenges** in Burmese NER, including ambiguity, tokenization, and inconsistent spelling
5. **Compare performance metrics** (e.g., Precision, Recall, F1-score) across different models and techniques
6. **Identify Limitations and Future Directions:** analyse the strengths and weaknesses of the developed models. This will involve identifying specific challenges (e.g., handling complex word boundaries, rare entity types) and outlining a clear roadmap for future research and development, including potential improvements in data annotation, model architectures, and training techniques.

While the immediate output is not a deployable library, this project will **serve as a research foundation** for future Burmese NLP tools. It will also **help inform best practices** for developing NER systems for other low-resource languages with similar linguistic challenges.

1. **Dataset**

The dataset for this Burmese NER project is built from **multiple publicly available sources**, including:

* **BBC Burmese** news articles
* **Official government press releases**
* **Entertainment and sports websites**
* **Manually curated documents from online forums and blogs**

This results in a **custom corpus of over 50,000 Burmese sentences**. To strengthen the dataset, I have also included **20,000 sentences from the open-source corpus published by the [Asian Language Treebank project]**, bringing the **total dataset size to approximately 70,000+ sentences**.

For the annotation process, I will use the **Label Studio** tool to manually label entities across the dataset. The following entity types will be annotated:

|  |  |  |
| --- | --- | --- |
| **Entity Type** | **Description** | **Example (Burmese)** |
| LOCATION | Geographic locations like cities, countries, rivers, lakes | ရန်ကုန်မြို့၊ မြန်မာနိုင်ငံ |
| PERSON | Names of individuals | ဒေါ်အောင်ဆန်းစုကြည် |
| ORGANIZATION | Companies, institutions, or agencies | ရန်ကုန်တက္ကသိုလ်၊ WHO |
| DATETIME | Specific dates, months, or years | ဇူလိုင် ၁၂၊ ၂၀၂၅ |
| NUMBERS | Quantities, measurements, monetary values | ၁၀၀၀ ကျပ်၊ ၁၅ ဦး |
| TITLE | Official or professional titles or movie or book title | ဝန်ကြီးချုပ်၊ ဒေါက်တာ ၊ ဦး၊ဒေါ်၊ ပါမောက္ခ |

All tokens that do not fall under the above categories will be annotated with the **“O” (Outside)** label, following the standard **BIO tagging format** (e.g., B-ORG, I-ORG, O).

This annotated dataset will serve as the backbone for training, validation, and evaluation of multiple NER models.

* 1. **Annotation Scheme**

This project follows standard **NER tagging formats** to label entities in the dataset:

* **BIO (Begin-Inside-Outside)**  
  Each token is tagged using the BIO format:
  + B-XXX: The token is at the **beginning** of an entity of type XXX
  + I-XXX: The token is **inside** (but not the first token) of an entity of type XXX
  + O: The token is **outside** any named entity

Example (ORGANIZATION entity):

ရန်ကုန် တက္ကသိုလ် တွင် → B-ORG I-ORG O

**BIOES (Begin-Inside-Outside-End-Single)** (optional, for future extension)  
This is a more fine-grained version of BIO used in some deep learning models: For this project, **BIO** format will be used primarily for annotation and training. However, **BIOES** support is considered for model experimentation depending on how model performance is affected.

For this project, **BIO** format will be used primarily for annotation and training. However, **BIOES** support is considered for model experimentation depending on how model performance is affected.

### **Expectation**

With the prepared dataset, I plan to evaluate the NER models using standard performance metrics. **Accuracy and ROC-AUC** will be used as a general metric for overall model behaviour, while **Precision, Recall, and F1-score** will be computed specifically for each entity type to measure the model's ability to correctly identify and classify named entities.

I aim to achieve **at least 80% accuracy**, with an expected **F1-score above 70%** for major entities such as **PERSON, ORGANIZATION, and LOCATION**. However, I anticipate that **LOCATION** and **ORGANIZATION** entities might be slightly **overrepresented or imbalanced**, as the dataset contains more examples from news and government sources. This may cause slight bias in model predictions toward these categories.

The evaluation will help identify:

* Which entity types perform best
* Where the model struggles (e.g., rare or ambiguous entities)
* What different architectures (e.g., BiLSTM vs Transformers vs Naïve Bayes) affect performance

The outcome of this evaluation will inform possible improvements in **data balancing, entity definitions, or model fine-tuning** for future iterations.