Project Proposal: Burmese News Article Topic Classification Using Deep Learning

*Submitted by **Ei Ei Khaing***
Deep Learning Course / MMDT
9 Aug 2025

Problem Statement

In recent years, Myanmar's news industry has shifted rapidly toward digital platforms. Every day, hundreds of new articles are published in Burmese, covering topics from politics to sports. While this explosion of content is great for readers, it also creates a real challenge, finding relevant news quickly. Most platforms still rely on people to manually assign categories, which takes time and often leads to inconsistency.

An automatic classification system for Burmese news articles can change that. It can save countless hours, help readers discover what matters to them, and make news archives far easier to manage. But here's the catch, Burmese is a low-resource language in NLP. There are few high-quality datasets, tokenization is tricky due to the script's complexity, and off-the-shelf tools rarely work well.

This project takes on that challenge by building a deep learning model to sort Burmese news into **five main categories**:

- Business
- Crime
- Entertainment
- Politics
- Sports

The goal is simple: accurate, consistent, and fast categorization, helping both media organizations and everyday readers.

Input and Output

Input:
A Burmese-language news article or sentence. The text will first be cleaned, tokenized, and normalized to ensure consistency.

Output:
One predicted label --- Business, Crime, Entertainment, Politics, or Sports --- along with a confidence score showing how sure the model is.

Dataset

Type: Multi-class Burmese text classification dataset.

Sources:

1. **Web Crawling** from Burmese news websites:
    - ✓ https://www.mdn.gov.mm/my/local-sports

- ✓ https://burma.irrawaddy.com/
- ✓ https://www.popularmyanmar.com/
- ✓ https://www.frontiermyanmar.net/mm/
- ✓ https://www.bnionline.net/mm
- ✓ https://news-eleven.com/

2. **GitHub Repositories:**
   - ✓ https://github.com/ayehninnkhine/MyanmarNewsClassificationSystem/tree/master/resources/datasets2
   - ✓ https://github.com/aungkhantmyat/Burmese-News-Topic-Classification/tree/main

Category Distribution:

Category	Count
Business	203
Crime	205
Entertainment	205
Politics	204
Sports	210

Total records: **1,027 articles**

The data will be cleaned to remove HTML tags, ads, and symbols. Burmese script normalization will handle font variations. If needed, data augmentation will help the model generalize better.

Expected Performance

With a balanced dataset, the aim is to achieve:

- **Accuracy**: 50% or higher
- **Macro F1-Score**: 0.50 or higher

These targets are realistic for a dataset of this size in a low-resource language, while still setting a clear bar for quality. The model's performance will be tested on unseen data to ensure it works beyond the training set.

Motivation

This project isn't just about machine learning, it's about solving a real problem for Burmese media. Readers should be able to quickly find news they care about, journalists should spend less time on repetitive sorting, and news archives should be easy to search.

For me, it's also about contributing to Burmese NLP research. Many global NLP breakthroughs never reach low-resource languages like Burmese. By building and sharing a clean dataset and trained model, this work can be a stepping stone for future tools, from automatic summarizers to sentiment analyzers.

In short, this project will:

- ✓ Help people access relevant Burmese news faster.
- ✓ Reduce the workload for editors and news portals.
- ✓ Add to the limited pool of Burmese NLP resources.
- ✓ Lay the groundwork for bigger, multilingual classification systems in the future.