Lingo Blend [Low Prep.] (250 Points)

NJACK-ML

India stands as one of the most linguistically diverse countries globally, boasting over 1,600 languages and dialects. In this problem statement, we present two distinct subtasks of increasing complexity, each addressing various aspects of translation.

## Subtask 1: Language Classification [100]

Participants are given 500 sentences, each potentially in Bengali, Hindi, Punjabi, Tamil, or Telugu. The challenge is to identify the language by predicting its label, as outlined in the following table:

| **Language** | **Label** |
| --- | --- |
| Bengali | 0 |
| Hindi | 1 |
| Punjabi | 2 |
| Tamil | 3 |
| Telugu | 4 |

### Evaluation Metrics

Macro-weighed f1-score, as implemented in scikit-learn, will be considered for evaluation.

## Subtask 2: English Translation [100]

Participants are provided with 500 different unlabeled sentences. The task involves developing an AI model that translates these sentences into English.

### Evaluation Metrics

The average score calculated using BLEU, ROUGE-1 will be used for evaluation. (BLEU + ROUGE-1 + 2\*ROUGE-2 + ROUGE-L) / 5.

# Judging Criteria

1. The performance of the model will carry a total of 200 marks. Each subtask has its own metrics and different points associated with itself. The final score will be the aggregate weighted score of each subtask.
2. The presentation carries a total of **50** marks and the teams will be judged on the originality of their approach, depth of research and their presentation.

# Rules

1. For this task, participants are free to employ any Machine Learning (ML) or Deep Learning (DL) model or technique (Max parameters allowed: 600M). However, blatant copying or the use of third-party translation APIs is strictly prohibited and will result in disqualification.
2. Additional data usage is also prohibited.
3. This dataset deals with Bengali, Hindi, Punjabi, Tamil and Telugu. In the provided dataset, the following files are made available-
   1. train.csv- Contains 1500 sentences across 6 languages to be used as training data for each of the three subtasks.
   2. ‘language translation’ and ‘english translation’- Folders containing files on which the subtask is to be performed, along with an example submission file.
4. In case of any ambiguity or confusion feel free to reach out.
5. You cannot convert back the code-mixed languages to their native script.

# Submission

1. Submissions must include both the submission files (example submission files are included in the dataset) and the code used for generation, ensuring the reproducibility of the submission (to be verified).
2. Also a brief 10-min presentation should be prepared and presented for final evaluation.

# Resources:

* Dataset:- <https://drive.google.com/drive/folders/1MgFU9-6nYeCL-cxymCWYlQQJanmamFAu>
* Translation Starter’s Guide:-

<https://huggingface.co/docs/transformers/tasks/translation>

* Relevant Paper on Code Mixed Language:-

<https://aclanthology.org/W18-3817.pdf>

# Contact Details:

Archit :- +91 75092 28601

Poornash :- +91 94433 54972

Atharva :- +91 8668255873