LMA Mini Project Mid Report (2024701016)

Corpus Statistics (Tokenized)

Per-language statistics:

Language	Sentences	Tokens	Share of total	Types(unique ids)
English	85,560,609	1,867,232,562	51.58%	37,999
Telugu	94,877,782	1,312,757,994	36.26%	22,716
Sindhi	11,600,789	439,998,236	12.15%	30,864

Overall statistics:

- **Total sentences : 192,039,180**
- **Total tokens : 3,619,988,792**
- **Total no of tokens : > 3.6 billion**

Vocab language distribution :

- **English: 20555 (32.12%)**
- **Telugu : 22716 (35.49%)**
- **Sindhi : 8228 (12.86%)**
- **Other : 12501 (19.53%)**

  Total vocab = 64k

Tokenizer used: Byte Pair Encoding(BPE) Tokenizer

Work done so far

Phase 1: Data Collection and Preprocessing

1) Data sourcing & extraction

- Pulled English (Wikipedia + FineWeb), Telugu (multiple web sources,AI4Bharat,wiki), and Sindhi (Wikipedia + Cultura JSONL) corpora. [from hugging face]
- Converted Parquet/JSONL → clean UTF-8 TXT (one sentence per line) with safe, streaming scripts

2) Normalization & cleaning

- **Unicode normalization** (NFC/NFKC), URL/email/handle removal, digit stripping where appropriate.
- **Language filtering:**
    - Telugu: keep-only \u0C00–\u0C7F, drop stray non-Telugu tokens, handle ZWJ/ZWNJ; custom sentence split on ।॥.!?.
    - Sindhi (Arabic script): keep-only \u0600–\u06FF plus ؟-?!, sentence split on ! ? . ۔ ؟.
    - English: NLTK sentence tokenizer; normalize curly quotes; remove pure-symbol lines.

Phase 2: Tokenizer Training

- Implemented a Byte Pair Encoding (BPE) tokenizer using the Hugging Face `tokenizer` library.
- Training corpus: merged English, Telugu, and Sindhi cleaned corpora (~3B tokens).
- Vocabulary size fixed at 64,000 subword units.
- Added special tokens `[PAD], [UNK], <s>, </s>` for sequence handling.
- Set `min_frequency=2` to discard ultra-rare subwords.

Timeline for future phases

Phase 3: Model Pretraining (Aug 29 - Sept 4)

- Preparing train and test splits for pretraining by balancing with appropriate ratios for each language.
-
- Pretrain qwen model on the collected corpus:.
    - Hidden size, layers, heads sized to resources .
    - Set up training environment (HuggingFace + Accelerate/DeepSpeed).
    - Build dataloader to stream tokenized shards.
    - Train on full 3B-token mixture.
    - Save periodic checkpoints.
    - Track validation perplexity per language (EN/TE/SD).
    - Hyperparameter tuning for better performance.

Phase 4: Finetuning on Reasoning Tasks (Sep 6 – Sep 9) :

- Fine-tune pretrained LM on Task1 Sentimental Analysis
- Fine-tune pretrained LM on Task2 Main Clause Prediction
-

Phase 5: Evaluation and Analysis (Sept 10-11)

- Evaluate pretrained and fine-tuned models using appropriate metrics (e.g., perplexity, accuracy).
- Perform detailed error analysis and observations of good and bad performance
- Create any visualizations,plots if required

Report (Sep 12 – Sep 14) :

- Draft final report:
    - Data pipeline (cleaning, balancing, tokenizer).
    - Model setup (arch, hyperparams, training curve).
    - Evaluation (perplexity, accuracy, token mix).
    - Analysis (where it performs well, where it fails).