Introduction

Byte Pair Encoding (BPE) is an efficient subword tokenization technique used in NLP to break words into smaller subword units. This project focuses on implementing BPE from scratch to process **Roman Urdu text**, allowing effective vocabulary compression and improved handling of unknown words.

The model was trained using 100 Roman Urdu diary entries and later tested on 14-day diary entries to analyze its effectiveness in real-world scenarios. The goal was to reduce vocabulary size while ensuring that unknown words were properly tokenized into meaningful subwords.

Implementation Approach

The implementation of BPE followed a structured approach, ensuring that the tokenization process was optimized for Roman Urdu text:

1. **Training Phase:**
    - Extracted all unique characters from the dataset to initialize the vocabulary.
    - Applied BPE merging rules iteratively to combine frequently occurring character pairs into subwords.
    - Continued the merging process until the vocabulary was reduced to 1000 subwords while maintaining word structure.
    - Assigned unique IDs to each subword for encoding and decoding purposes.
    - Stored the trained BPE model, including vocabulary, merge rules, and token mappings.
2. **Testing Phase:**
    - The trained BPE model was applied to 14-day diary entries to test its tokenization performance.
    - Sentences were encoded into token IDs using the subword vocabulary.

- The encoded sequences were decoded back to text to verify if the tokenization preserved meaning.

Conclusion

The BPE model successfully **processes Roman Urdu text** by efficiently breaking words into subword units, ensuring that a fixed vocabulary size of 1000 subwords is maintained. This approach helps in text compression, better handling of unseen words, and improved generalization.

By limiting vocabulary size and applying subword tokenization, BPE provides a balance between efficient storage, meaningful representation, and flexibility in handling new words. The model's effectiveness proves **BPE as a powerful technique** for NLP applications involving Roman Urdu text.