Padding

The Problem of Variable-Length Sequences:

- **Different Lengths:** In natural language processing, sentences (or documents) often have different lengths.

- **Machine Learning Requirements:** Machine learning models typically require input data to have a **fixed** size. They expect data to be organized in matrices or tensors, where each row represents a sample and each column represents a feature.

- **Parallel Processing:** Having fixed-length sequences allows for efficient parallel processing, which is crucial for training large models.

Definition: Padding is the process of adding special tokens (usually "0" or a special <PAD> token) to the end of shorter sequences to **make them all the same length** as the longest sequence in the dataset.

Purpose: To ensure that all input sequences have the same length so that they can be processed in batches by machine learning models.

Why Padding is Necessary:

Batching: Most machine learning models are trained using batches of data. All sequences within a batch must have the same length. Padding ensures that this is the case.

Matrix/Tensor Operations: Machine learning models perform matrix and tensor operations. These operations require inputs to have consistent dimensions.

Recurrent Neural Networks (RNNs): RNNs can handle variable-length sequences, but they often benefit from padding, especially when using techniques like batching or attention.

Transformers: Transformer models also require padding for efficient batch processing.