**Model Architecture for Toxic Comment Detection**

**1. Input Layer**

* **Text Input**: Raw comment text.
* **Tokenization**: Converts input text into tokenized format using AutoTokenizer from Hugging Face.
* **Tokenized Outputs**:
  + input\_ids: Encodes words into numerical IDs.
  + attention\_mask: Differentiates actual words from padding.

**2. RoBERTa Encoder (Backbone)**

* **Pretrained RoBERTa-base Model**:
  + Uses **BPE (Byte Pair Encoding) Tokenization**.
  + Has **12 Transformer layers** (self-attention & feed-forward layers).
  + Each layer has **12 attention heads**.
  + **Hidden size**: 768 dimensions.
  + **Total parameters**: ~125M.

**3. Classification Head**

* The **final hidden state** is passed to a **fully connected layer** for classification.
* **Output layer**:
  + **Linear layer** maps RoBERTa embeddings to class logits.
  + **Softmax activation** converts logits into probability distributions over classes.