Hugging Face learnings summary

**What is Hugging Face?**

I would say Hugging Face is a collection of pre-trained/untrained ML models and Datasets to train models. These models and datasets can be used via a library in python.

**Transformers**

These are models which understands context and predict outputs accordingly. They comprise an encoder and a decoder. The encoder translates the text into numbers. These numbers not only represent a specific word but the word put into a specific context. The decoder can be fed the encoder output, or it can be used standalone. But a decoder doesn’t take into consideration the context.

Different types of transformer models

1. Encoder only model
2. Decoder only model
3. Encoder-Decoder model

**Tokenizer**

A tokenizer converts the entire text input into tokens which can later be translated to numbers using the model dictionary. I learned 3 types of tokenization methods.

1. Word-based tokenizers: Each word is a token. (Disadvantage: Too large dictionary of words)
2. Character-based tokenizers: Each character is a token. (Disadvantage: Loses meaning)
3. Subword tokenizer: Compound/Large words are divided into fragments while simple/small words are not. (Advantage: Relatively small dictionary still keeping the meaning)

At last I learned how fully train our own Data set using the Trainer API