- Fortunately, there is a class of multilingual transformers that come to the rescue. Like BERT, these models use masked language modeling as a pretraining objective, but they are trained jointly on texts in over one hundred languages.
- By pretraining on huge corpora across many languages, these multilingual transformers enable zeroshot cross-lingual transfer. This means that a model that is fine-tuned on one language can be applied to others without any further training! This also makes these models well suited for "code-switching," where a speaker alternates between two or more lan- guages or dialects in the context of a single conversation.
- One of the first multilingual transformers was mBERT, which uses the same architecture and pretraining objective as BERT but adds Wikipe- dia articles from many languages to the pretraining corpus. Since then, mBERT has been superseded by XLM-RoBERTa (or XLM-R for short), so that's the model we'll consider in this chapter.
- Multilingual transformers involve similar architectures and training procedures as their monolingual counterparts, except that the corpus used for pretraining consists of documents in many languages
- The RoBERTa part of the model's name refers to the fact that the pretraining approach is the same as for the monolingual RoBERTa models. RoBERTa's developers improved on several aspects of BERT, in particular by removing the next sentence prediction task altogether.
- XLM-R also drops the language embeddings used in XLM and uses SentencePiece to tokenize the raw texts directly.
- So far we have treated tokenization as a single operation that transforms strings to integers we can pass through the model. This is not entirely accurate, and if we take a closer look we can see that it is actually a full processing pipeline that usually consists of four steps (Normalization,Pretokenization,Tokenizer model,Postprocessing)
- we now understand that Senten- cePiece adds and <\s> instead of [CLS] and [SEP] in the postprocessing step
- The SentencePiece tokenizer is based on a type of subword segmentation called Unigram and encodes each input text as a sequence of Unicode characters
- d the fact that many languages, like Japanese, do not have whitespace characters. Another special feature of SentencePiece is that whitespace is assigned the Unicode symbol U+2581, or the _ character, also called the lower one quarter block character. This enables SentencePiece to detokenize asequence without ambiguities and without relying on language-specific pretokeniz- ers
- n, NER is often framed as a token classification task
- when we call transformer model it comes with body and head separated
- Evaluating a NER model is similar to evaluating a text classification model, and it is common to report results for precision, recall, and F1 -score.
- we can see that German, French, and Italian achieve similar performance in the all category, suggesting that these languages are more similar to each other than to English
-