- Building Classification twitter text using BERT
- quick recap on how to import data using its URL only
- there is unbalanced dataset on sklearn using to fix the imbalance dataset https://imbalanced-learn.org/stable/references/index.html#api
- Transformer models have a maximum input sequence length that is referred to as the maximum context size. For applications using DistilBERT, the maximum context size is 512 tokens
- quick recap on tokenization and its type
- The basic idea behind subword tokenization is to combine the best aspects of charac- ter and word tokenization
- The main distinguishing feature of subword tokenization (as well as word tokenization) is that it is learned from the pre- training corpus using a mix of statistical rules and algorithms
- The ## prefix in ##izing and ##p means that the preceding string is not whitespace; any token with this prefix should be merged with the previous token when you convert the tokens back to a string
- To tokenize the whole corpus, we'll use the map() method of our DatasetDict object
- First, the text is tokenized and represented as one-hot vectors called token encodings. The size of the tokenizer vocabulary determines the dimension of the token encod- ings, and it usually consists of 20k–200k unique tokens. Next, these token encodings are converted to token embeddings, which are vectors living in a lower-dimensional space.
- Although the code in this book is mostly written in PyTorch, Transformers pro- vides tight interoperability with TensorFlow and JAX. This means that you only need to change a few lines of code to load a pretrained model in your favorite deep learn- ing framework!
- steps on how to fine tune the transformer model on hugging face
- building classifier using pytorch and show how data might be misleading