STEP 1

For each dataset put all the data in one text file.

Create files dataset1.txt – dataset10.txt

Find a good text editor put all in one text file.

STEP 2

Use BERT-type tokenizer from Hugging Face and produce the same stuff but tokenized.

Put BERT-tokenized data in files: dataset1\_tokenized.txt-dataset10\_tokenized.txt

BERT-tokenizer for all-in-one file.

STEP 3

Postponed balancing.

Format the tables:

Sentence -> category.

Format for model training: dataset\_v1.txt

Write Python file to import a tokenizer.

Learn how to use it (tokenizer).

Write the rest of the file.

Read 10 datasets -> file loading.

How to apply tokenizer to each file and how to write it.

STEP 4

Write Python file to do the splitting

Split the .py

Do the statistics how much data for each set- full distribution.

Report the distribution.