Wids-2

Monday, 5 January 2026 9:11 PM

Train-Test split:
A model cannot be tested on the same data we train it on ,so we need new data
Train: To change the parameters using the backpropagation(these are not done by humans)
Validation: To change the hyperparameters (how well will the model learn {like the learning rate and all}){these were set initially before the training and are changed later on according to the purpose u need them for}
Overfitting:
When the model memorizes the training data instead of learning from it
Then the training loss decreases while the validation loss increases
Validation loss increases when the model stops generalizing and starts memorizing
The local minima in the validation loss curve is the best model
Stratified split:
For justification of the training and testing
Text → tokens → token IDs → model
This mapping is learned during pretraining
DistilBERT uses word Piece tokenization: That is the words are divided into subwords {not characters or the fullwords }

Old way	New way
DistilBERT is frozen, only a small classifier is trained	Both are not frozen
Load the DistilBERT tokenizer	Same
Encode the sentence into tokens	same
Pass these tokens into the tokenizer to get the vector{matrix}	Here the DistilBERT encoder
Extract the CLS embedding	The CLS will get automatically extracted
Train the external classifier	Train the classification head
n	n

Learnings:
The .json files can be of 2 formats [{},{}] and {}{}. For the 2nd the extra lines=True argument needs to be passed
.loc[..] for accessing a particular row

Dynamic padding is padding to the length of the maximum word while the static padding is padding to the maximum_lenght set no matter how small the overall/one text is