Home Assignment 03

Text Mining Summer Semester 2025

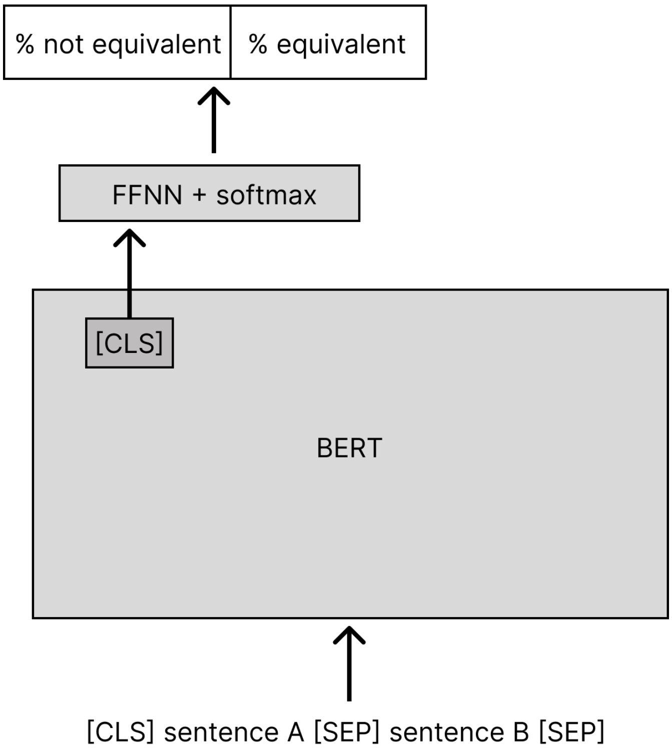
Giulia Luongo 10076102
Mohammad Sukri 10062921
Samer Sakor 10074501

1 MRPC – Solution Design Task

Paraphrase identification is the task of predicting whether two given sentences have the same meaning. It is a binary classification problem: the model outputs a label 1 if the sentences are equivalent, 0 if they are not.

Architecture BERT is well suited for this task. It is a Transformer encoder model that can be pre-trained on next sentence prediction, which helps it to understand the relationships between sentence pairs.

Each input consists of two sentences, A and B, concatenated with the special token [SEP] in between. A [CLS] token is added at the beginning, and its final embedding represents the summary of the information of the entire input sequence. During fine-tuning, a feedforward classification layer is added on top of the [CLS] embedding. This layer produces logits for two classes ('equivalent' or 'not equivalent'), which are then converted to probabilities using a softmax function.



Training The model is trained using the Microsoft Research Paraphrase Corpus (MRPC), which consists of pairs of sentences labelled as ’equivalent’ (1) or ’not equivalent’ (0). To feed data into the model, the two sentences are tokenized and concatenated with a [SEP] token.

For binary classification, the appropriate loss function is Cross-Entropy (if using softmax over two output classes).

Training typically runs for a few epochs for the fine-tuning rather than the full pretraining schedule (approximately 40 epochs). To select the best model, a validation set is used to monitor model performance after each epoch. Early stopping can be applied, meaning that training is stopped when the performance of the validation set stops improving.

Evaluation The model is evaluated using a test set of sentence pairs that were not seen during training. This enables us to assess its ability to generalize to new examples.

For evaluation, we can use accuracy, which is the proportion of correctly predicted labels, and it’s a standard metric for binary classification.