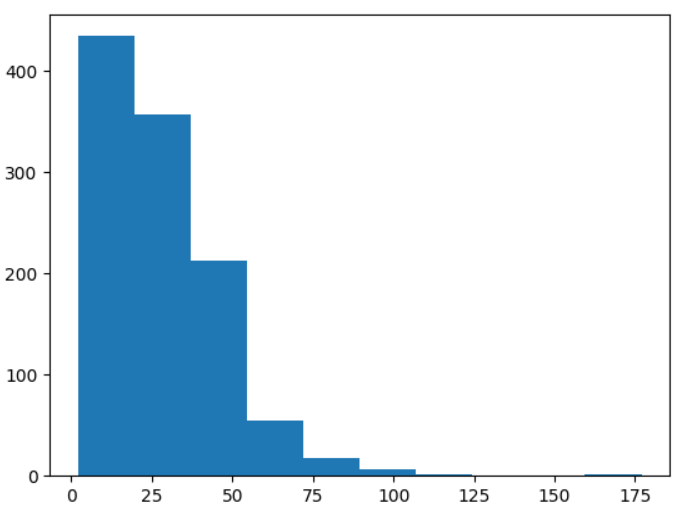
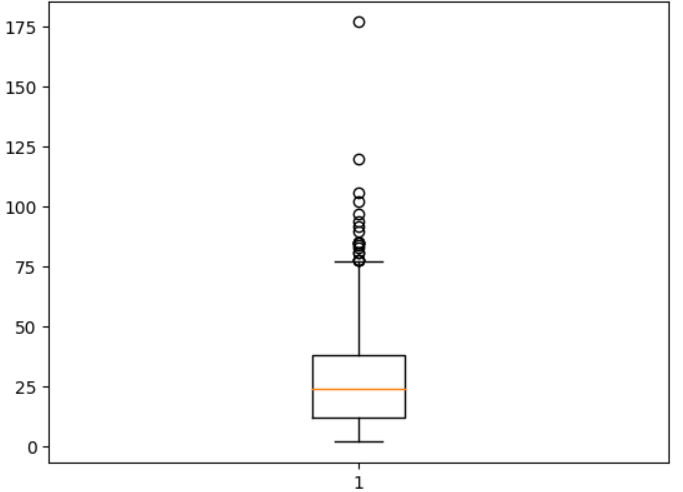
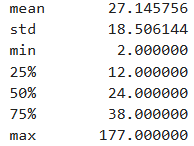
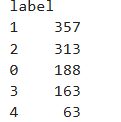
最大文本长度参数的取值





四分之三的数据文本长度在40以内，只有少数数据文本较长，因此不适合将最大文本长度设置过大。考虑到bert模型对[PAD]填充具有一定的鲁棒性，不想丢失太多长文本的信息，考虑定在50左右，多尝试。具体的确定需要专门查阅一下多数文本较短但存在长文本的论文。

各标签样本量

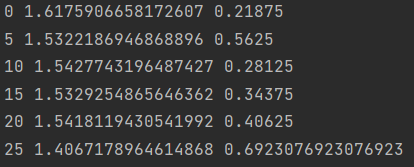


存在类间样本不平衡的情况，主要是科学解释水平达到4的学生较少。

解决措施：①SMOTE过采样；②类别权重

预训练的语言模型在大量数据上训练过，一般已经学习到了少数类样本的有用特征，所以应该不用太担心这个问题。

在训练集上的结果：



在测试集上的结果：

