基于 Transformer 的文本可读性评估模型开题报告

组员： 李叙言 2006500004

余子意 2006500006

1．项目背景：

随着教育领域对个性化学习的需求日益增长，能够准确评估文本可读性的技术变得越来越重要。CommonLit Readability Prize 是一个 Kaggle 竞赛，旨在利用机器学习和自然语言处理技术开发一个能够准确预测文本可读性的模型。本项目将基于 Transformer 架构开发一个文本可读性评估模型，以提高学生阅读能力并提升教育质量。

2．问题陈述：

https://www.kaggle.com/competitions/commonlitreadabilityprize/overview

本项目的主要任务是根据给定文本的内容预测其可读性得分。文本可读性评估面临的挑战包括文本的复杂性、语言特点和领域知识等。

3．目标和研究问题：

本项目的目标是开发一个基于 Transformer 的高性能文本可读性评估模型。具体研究问题包括：

a) 如何使用 Transformer 模型有效地提取文本特征以预测可读性？

b) 如何调整 Transformer 模型参数以提高预测准确性？

4．数据集描述：

竞赛分别提供了一个 train.csv 训练集和 test.csv 测试集。两个数据集中的特征变量为文本内容，train.csv 的目标变量是可读性得分。

5．研究方法：

本项目将采用以下研究方法：

a) 数据预处理：对文本进行清洗、标准化，消除噪声并准备数据以供模型使用。

b) 特征工程：使用预训练词嵌入或预训练 Transformer 模型（如 BERT、RoBERTa 等）对文本进行编码，以提取有意义的文本特征。

c) 模型选择：选择一个基于 Transformer 的架构，如 BERT、RoBERTad 等模型。

d) 模型训练：使用训练集对选定的 Transformer 模型进行训练，优化模型参数以提高预测准确性。

e) 模型测试：使用测试集评估模型的性能。

6．预期成果和影响：

项目完成后，预期将获得一个基于 Transformer 的高准确性文本可读性评估模型，有望在 CommonLit Readability Prize 竞赛中取得优异成绩。推动个性化学习资源推荐，为教育工作者和学生提供更精准的阅读材料建议，从而提高学生的阅读能力和学习效果。