文档说明

一、实现目标

本项目旨在实现对评语的自动化多标签分类，主要目标是通过对原始评语数据进行清洗、分句、结构化和抽样，生成适合后续建模的数据集；随后，利用种子词典和句子 BERT 模型，对无标签数据进行自动多标签弱监督标注，生成训练集；最后，基于 BERT 模型进行多标签分类训练，并对模型效果进行详细评估和可视化展示。

二、主要思路与流程

1. 数据预处理

目标：将原始 Excel 格式的评语数据，经过清洗、分句、结构化、去重和抽样，生成标准化的训练/标注数据。

主要步骤：

数据加载：读取原始评语 Excel 文件。

列合并：将“评语”和“不足与建议”两列合并为“整体评语”。

分句处理：对“整体评语”进行智能分句，支持多种编号和标点格式。

无关列删除：去除与建模无关的列，保留核心信息。

句子展开：将每条评语的分句展开为多行，每行一句。

空值清理：去除无效或空的句子。

数据抽样：随机抽取部分数据用于人工标注或验证。

文本去重：对分词结果文件进行去重，保证词表唯一性。

输出：标准化的 Excel 数据文件和去重后的分词文本。

2. 弱监督自动标注

目标：利用大语言模型生成的“种子句”，通过句子 BERT 模型对无标签数据进行自动多标签标注，生成训练集。

主要步骤：

加载种子句：读取每个每个类别的种子句子。

加载原始数据：读取待标注的训练数据。

句子编码：用 Sentence-BERT 对所有句子和种子样例进行向量化。

相似度计算：计算每个句子与各类别种子句子的余弦相似度。

自动标注：相似度超过阈值则赋予对应标签，实现弱监督多标签标注。

结果保存：输出自动标注后的训练集。

评估与可视化：对测试集进行同样推理，输出多标签分类报告，并生成各类别 precision/recall/f1-score 和预测数量分布的可视化图表。

输出：自动标注的训练集和评估可视化结果。

3. 多标签分类模型训练与评估

目标：基于 BERT 模型对自动标注后的数据进行多标签分类训练，并对模型效果进行详细评估和可视化。

主要步骤：

数据加载：读取自动标注后的训练集和测试集。

数据编码：对文本进行 BERT 分词和编码，标签转为多标签向量。

模型训练：采用 BERT+FocalLoss 进行多标签分类训练，缓解类别不平衡。

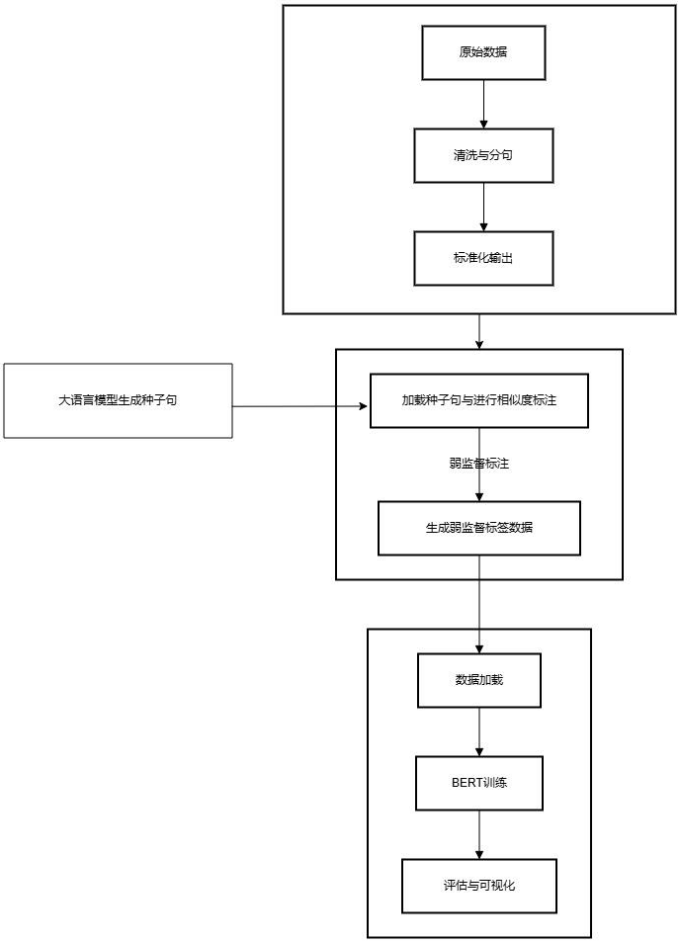
阈值优化：在验证集上自动寻找最佳分类阈值。

模型评估：在测试集上推理，输出多标签分类报告。

评估可视化：生成各类别 precision/recall/f1-score 和预测数量分布的可视化图表，便于直观分析模型表现。

输出：训练好的模型、预测结果、详细评估报告和可视化图表。

三、流程图



四、总结

本项目实现了从原始数据的清洗、弱监督自动标注到深度学习多标签分类的完整流程，充分体现了数据处理与智能建模的有机结合。通过引入先进的 sentence-BERT 和 BERT 模型，结合种子句的弱监督标注策略以及 FocalLoss 损失函数，有效提升了多标签分类任务的自动化水平和分类准确性。整个流程不仅大大减少了人工标注的工作量，还增强了模型对复杂评语的理解和判别能力。同时，项目在模型评估阶段引入了多种可视化手段，能够直观展示各类别的分类效果和模型表现，为后续的分析和优化提供了有力的数据支持和决策依据，进一步推动了评语文本智能处理的实用化和高效化。