基于BERT+PET方式文本分类介绍

学习目标

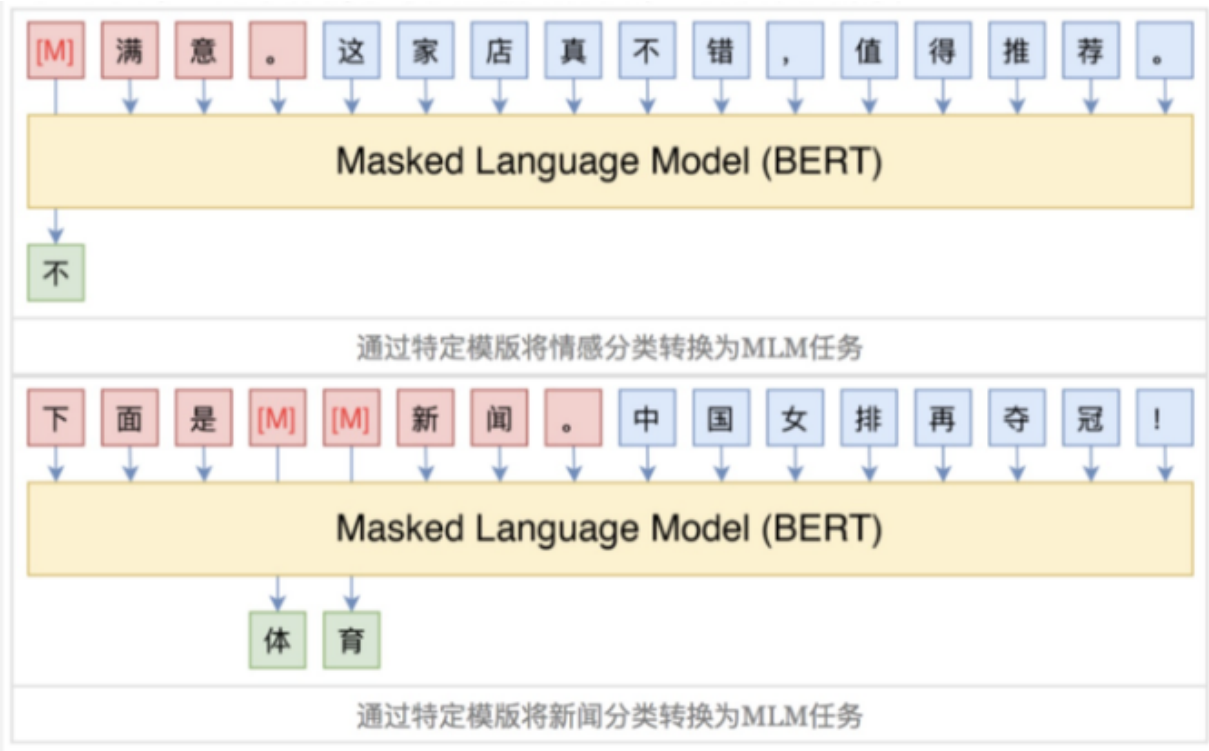
- 理解PET方式的思想
- 安装项目必备的工具包
- 了解基于BERT+PET方式实现文本分类的整体项目架构

1 项目介绍

- 本章我们将以"电商平台用户评论数据"为背景，基于BERT+PET（硬模版）方法实现评论文本的准确分类

2 PET回顾

- PET（PatternExploiting Training）的核心思想是：根据先验知识人工定义模版，将目标分类任务转换为与MLM一致的完形填空，然后再去微调MLM任务参数。



> 图中示例1: 情感分类任务（好评还是差评），原始文本:这家店真不错,值得推荐。PET模板:[MASK]满意。Label:不/很。标签词映射（Label Word Verbalizer）：例如如果`[MASK]`预测的词是“不”，则认为是差评类，如果是“很”，则认为是好评类。

> 图中示例2:新闻分类任务（多分类），原始文本：中国女排再夺冠！PET模版：下面是[MASK][MASK]新闻，Label：体育/财经/时政/军事

- PET方式实现过程：将模版与原始文本拼在一起输入预训练模型，预训练模型会对模板中的mask做预测，得到一个label
- PET方式的特点：
  - 优点：
    - 人工模版，释放预训练模型知识潜力
    - 不引入随机初始化参数，避免过拟合
    - 较少的样本就可以媲美多样本的传统微调方式
  - 缺点：
    - 人工模板稳定性差，不同模板准确率可相差近20个百分点
    - 模板表示无法全局优化

3 环境准备

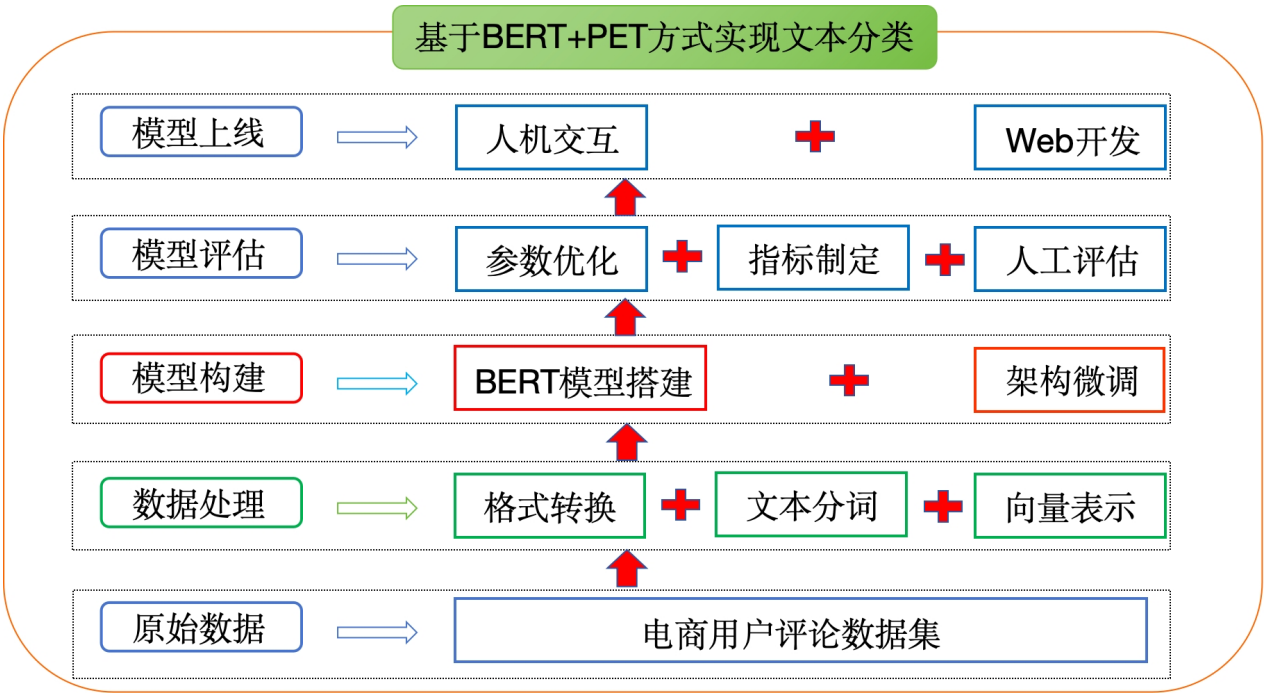
本项目基于 torch+ transformers 实现，运行前请安装相关依赖包：

- torch
- transformers==4.22.1
- datasets==2.4.0
- evaluate==0.2.2
- matplotlib==3.6.0
- rich==12.5.1
- scikit-learn==1.1.2
- requests==2.28.1

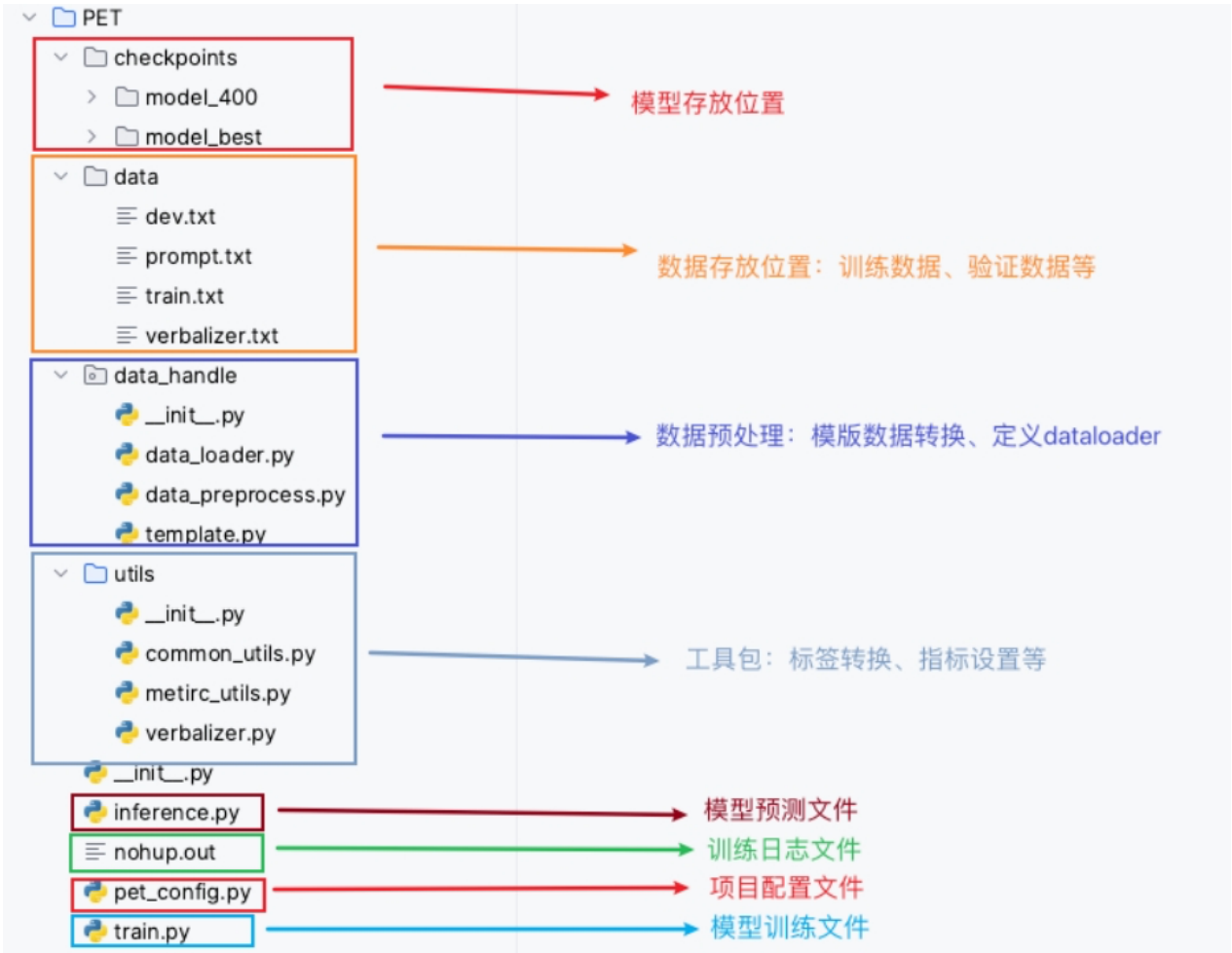
---

4 项目架构

项目架构流程图：



项目整体代码介绍：



小结总结

本章节主要介绍了项目开发的背景及意义，明确了项目的整体架构，并对项目中整体代码结构进行了介绍。