CHIP

基于BERT与模型融合的短文本分类方法

团　队：wzm

汇报人：吴梓明

团队简介

吴梓明　华南理工大学　研究生三年级

- 贵在联通——“联创黔线”杯大数据应用创新大赛 1st
- 蚂蚁金服风险大脑：支付风险识别(内部赛) 1st
- DigSci 科学数据挖掘大赛 2019 3rd
- 同盾科技声纹识别建模大赛 4th
- 第三届融360天机智能金融算法挑战赛 5th
- 智源&计算所-互联网虚假新闻检测挑战赛 7th
- 美年健康AI大赛——双高疾病风险预测 9th

任务说明

任务描述

给定事先定义好的45种筛选标准类别和一系列中文临床试验筛选标准的描述句子，需返回每一条筛选标准的具体类别。

评价指标

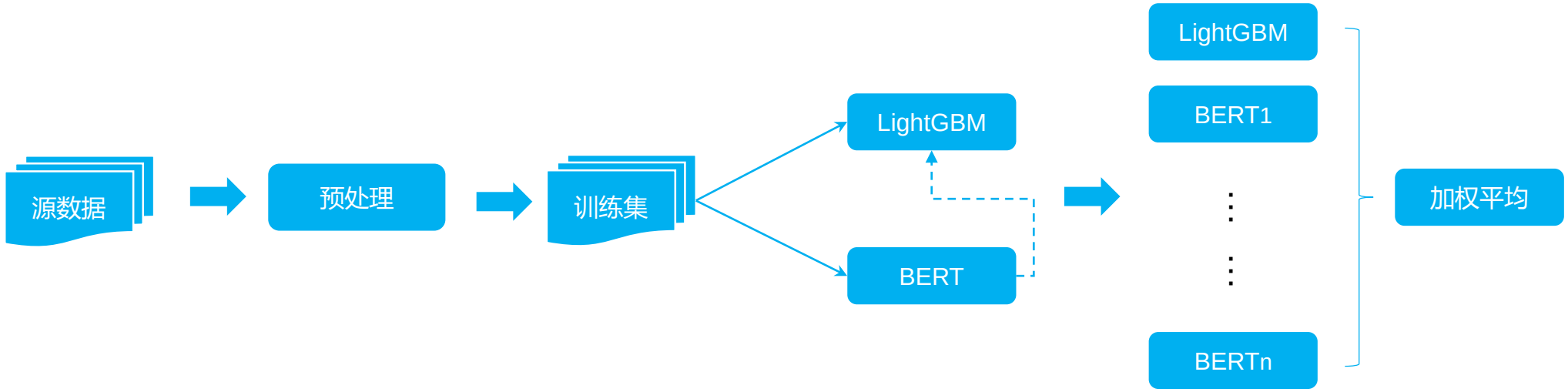
评价指标包括宏观准确率(Macro Precision)，宏观召回率(Macro Recall)，Average F1值。最终排名以Average F1值为基准。假设有n个类别，C_1，…… ……，C_i，…… ……，C_n。

$$\text{准确率}P_i = \frac{\text{正确预测为类别}c_i\text{的样本个数}}{\text{预测为}c_i\text{类的样本个数}}$$

$$\text{召回率}R_i = \frac{\text{正确预测为类别}c_i\text{的样本个数}}{\text{真实的}ci\text{类的样本个数}}$$

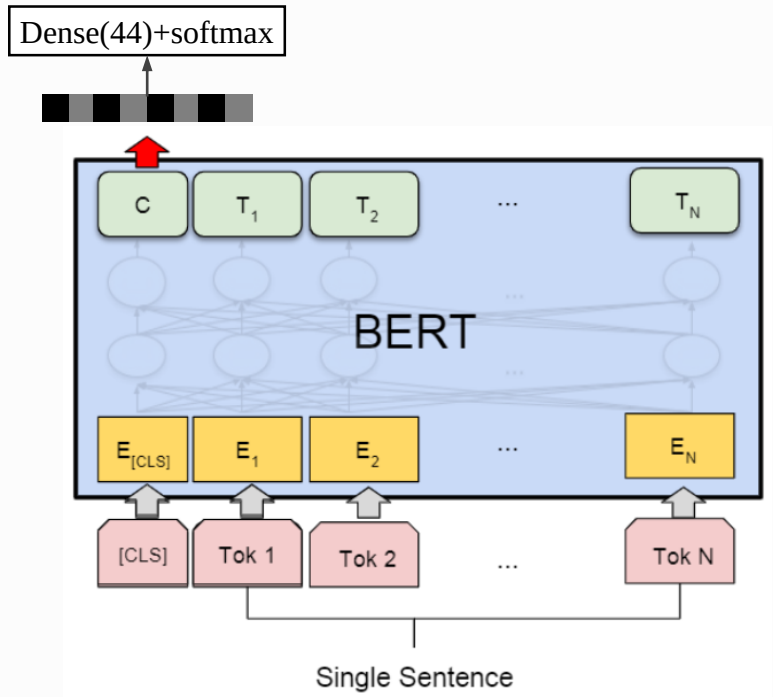
$$\text{Average F1} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\frac{2*P_i*R_i}{P_i+R_i}$$

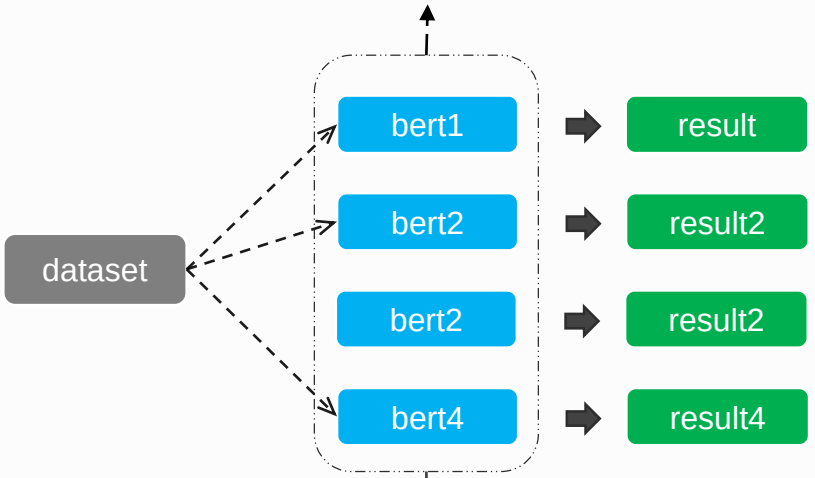
整体方案设计



BERT模型

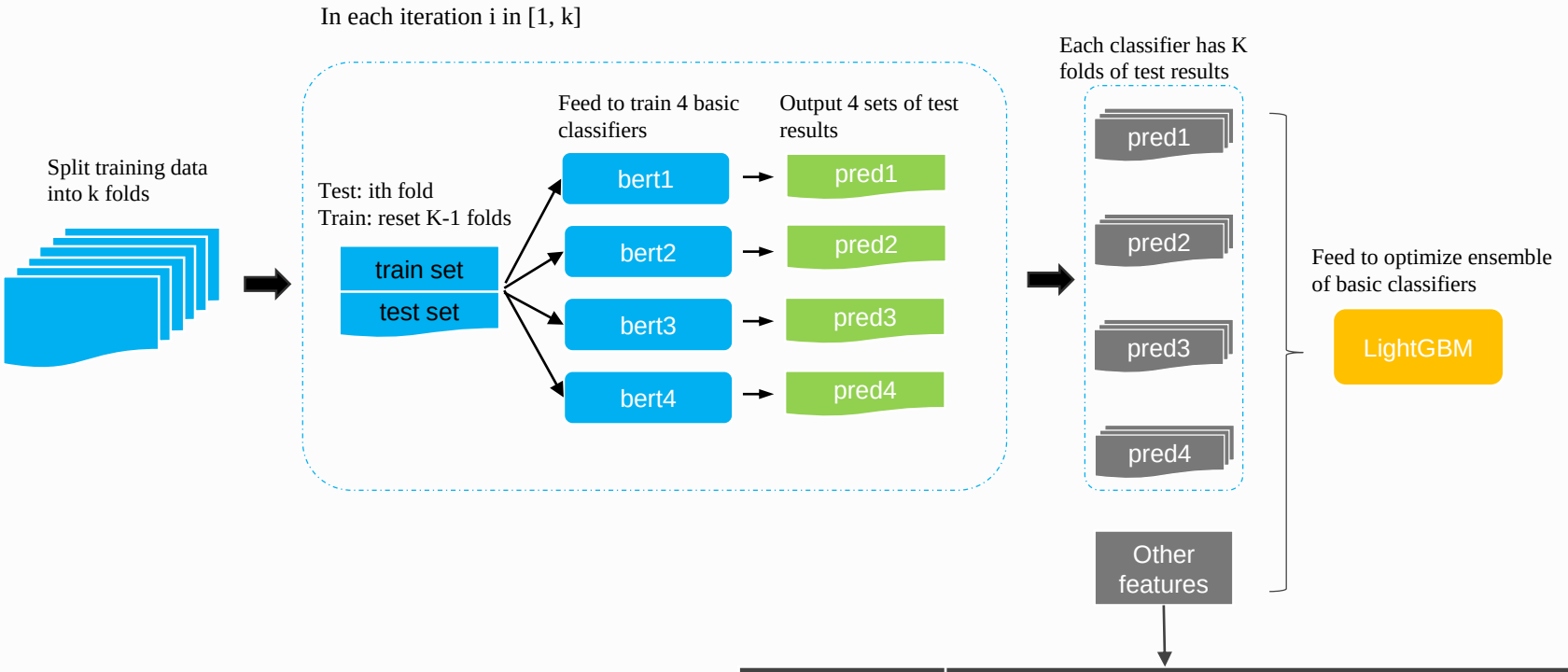
bert1: BERT-wwm-ext
bert2: RoBERTa-wwm-ext-large
bert3: BERT-Base
bert4: RoBERTa-zh-Large
wwm: 全词掩盖
RoBERTa: bert的变种





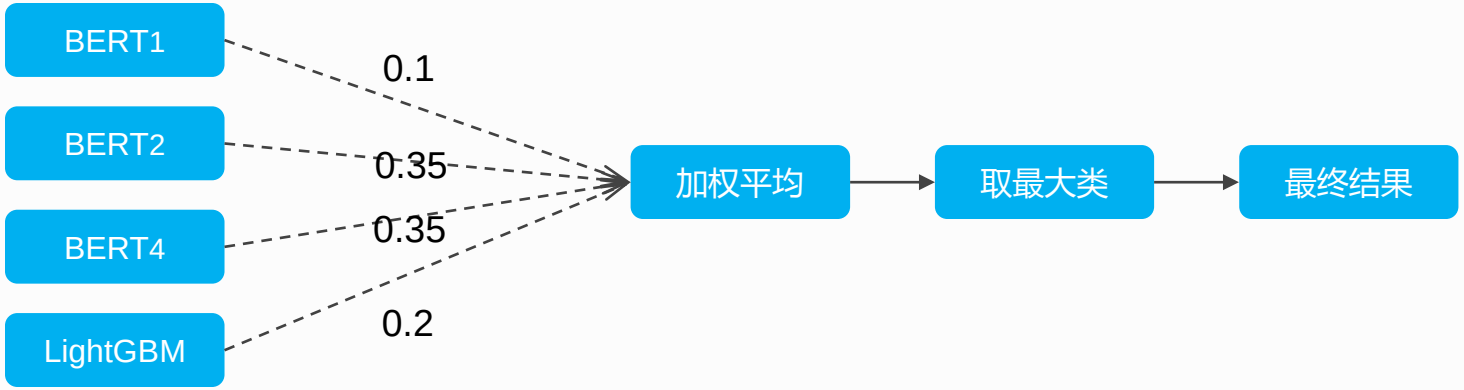
[1] Cui Y, Che W, Liu T, et al. Pre-Training with Whole Word Masking for Chinese BERT[J]. arXiv preprint arXiv:1906.08101, 2019.
[2] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
[3] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

LightGBM



特征	说明
LDA	对句子进行LDA主题模型的特征提取，共44维
len	句子的长度
digit_num	句子中包含的数字个数
op_num	句子中包含的比较符个数
alp_num	句子中包含的英文字母个数

集成



最终的结果是四个模型的结果的加权平均，然后取概率最大的类即可。

还有什么能做?

1

数据增强，伪标签（Pseudo-Labelling）

2

构造更多的特征，比如表征句子语义的特征

(sentence embedding)

3

修改损失函数，如：Focal Loss，

根据线下自定义每个类别的权重等

…

THANKS

团　队：wzm

汇报人：吴梓明