重庆邮电大学

学生实验实习报告册

学年学期：2022-2023 学年(秋)学期

课程名称：行业大数据分析综合实践

学生学院：计算机学院/人工智能学院

专业班级：数据科学与大数据技术

学生学号：XXXXXX

学生姓名：XXXXXX

联系电话：XXXXXX

重庆邮电大学教务处印制

目　录

教师评阅记录表

【重要说明】

- 学生提交报告册最终版时，**必须包含此页**，否则不予成绩评定。
- 本报告册模板内容格式除确实因为填写内容改变了布局外，**不得变更其余部分的格式**，否则不予成绩评定。

报告是否符合考核规范	☑ 符合　☐ 不符合
报告格式是否符合标准	☑ 符合　☐ 不符合
报告是否完成要求内容	☑ 是　☐ 否
报告评语：	
报告成绩：	
评阅人签名（签章） 2023 年 1 月 8 日	

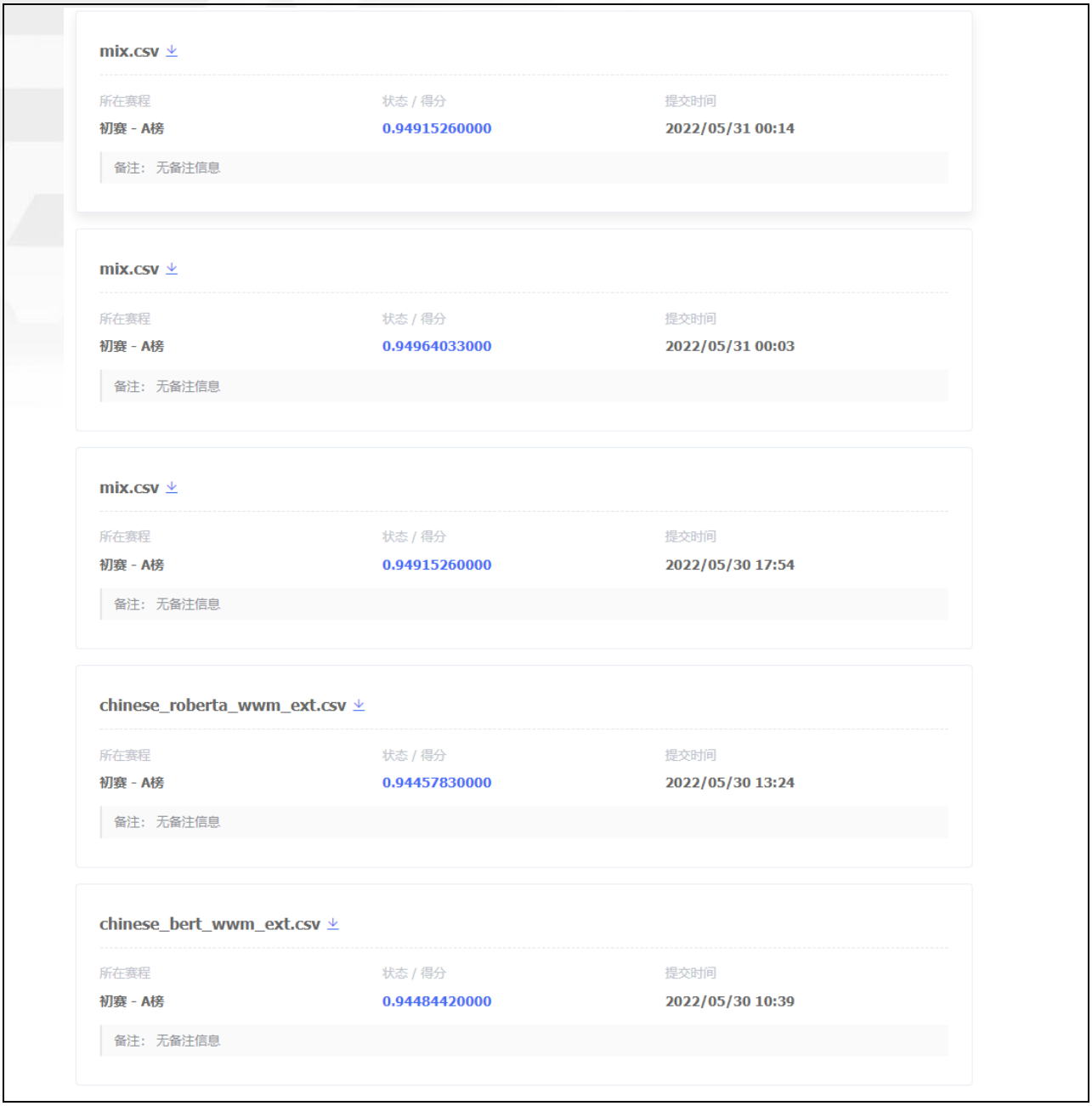
实验或实习报告

课程名称	行业大数据分析综合实践	课程编号	A2040870
开课学院	计算机学院/人工智能学院		
指导教师	XX		
实验实习地点	综合实验大楼 B517		

学号	姓名
*****	*****

技术方案（文字描述+图表说明为主，**不得大段粘贴代码**）、实验结果及分析（技术方案的书写可以适当参考 QQ 群文件中的“R3.技术方案撰写参考文档”）

一、线上测试成绩（给出最优成绩和线上测评 ID，截图给出完整的历史提交成绩记录）



mix.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.94915260000	2022/05/31 00:14

备注：无备注信息

mix.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.94964033000	2022/05/31 00:03

备注：无备注信息

mix.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.94915260000	2022/05/30 17:54

备注：无备注信息

chinese_roberta_wwm_ext.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.94457830000	2022/05/30 13:24

备注：无备注信息

chinese_bert_wwm_ext.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.94484420000	2022/05/30 10:39

备注：无备注信息

chinese_bert_wwm_ext.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.93586700000	2022/05/29 21:04

备注：无备注信息

chinese_bert_wwm_ext.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.93556080000	2022/05/26 13:50

备注：无备注信息

chinese_bert_wwm_ext.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.92924535000	2022/05/25 15:52

备注：无备注信息

chinese_bert_wwm_ext.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.93556080000	2022/05/25 13:45

备注：无备注信息

chinese_bert_wwm_ext.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.93493974000	2022/05/25 00:17

备注：无备注信息

chinese_bert_wwm_ext.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.93462470000	2022/05/24 18:09

备注： 无备注信息

chinese_bert_wwm_ext.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.93144214000	2022/05/24 16:41

备注： 无备注信息

mix_model.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.94736844000	2022/05/17 00:50

备注： bert-large, bert-wwm-ext-large, roberta-wwm-ext-large

mix_model.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	逻辑错误 查看日志	2022/05/17 00:49

备注： bert-large, bert-wwm-ext-large, roberta-wwm-ext-large

mix.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.93975910000	2022/05/16 10:28

备注： bert-wwm-ext, bert-base, roberta-wwm-ext-large

mix.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.94685990000	2022/05/16 10:14

备注： 无备注信息

chinese_roberta_wwm_ext_large.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.94004790000	2022/05/16 00:13

备注：无备注信息

mix.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.94457830000	2022/05/15 11:45

备注：五个bert模型融合

roberta_chinese_base.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.93689320000	2022/05/15 11:41

备注：无备注信息

ernie_1.0.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.91549290000	2022/05/15 00:13

备注：无备注信息

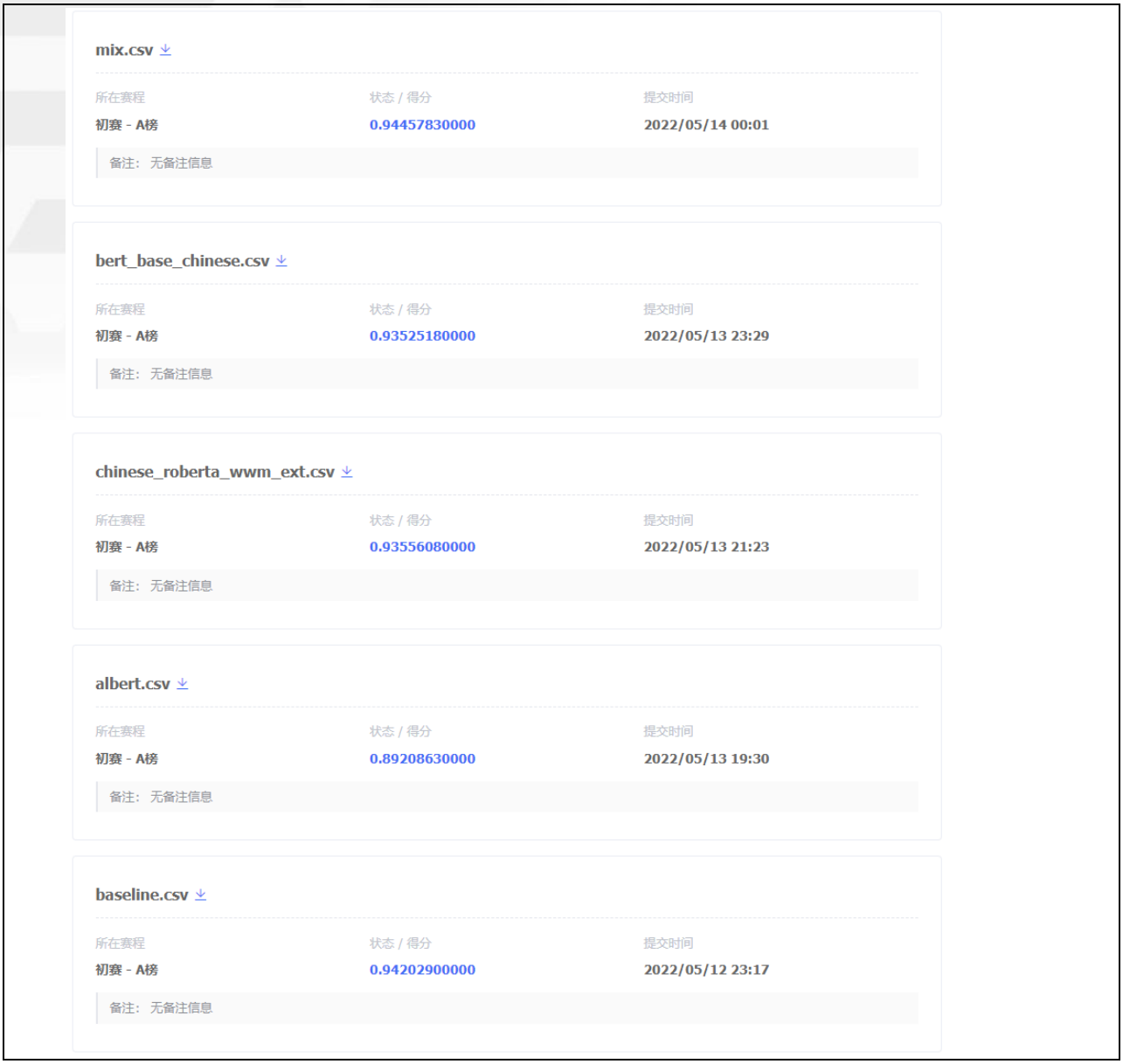
chinese_bert_wwm_ext.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.94230765000	2022/05/14 19:27

备注：无备注信息

chinese_bert_wwm.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.93462470000	2022/05/14 11:34



mix.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.94457830000	2022/05/14 00:01

备注：无备注信息

bert_base_chinese.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.93525180000	2022/05/13 23:29

备注：无备注信息

chinese_roberta_wwm_ext.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.93556080000	2022/05/13 21:23

备注：无备注信息

albert.csv

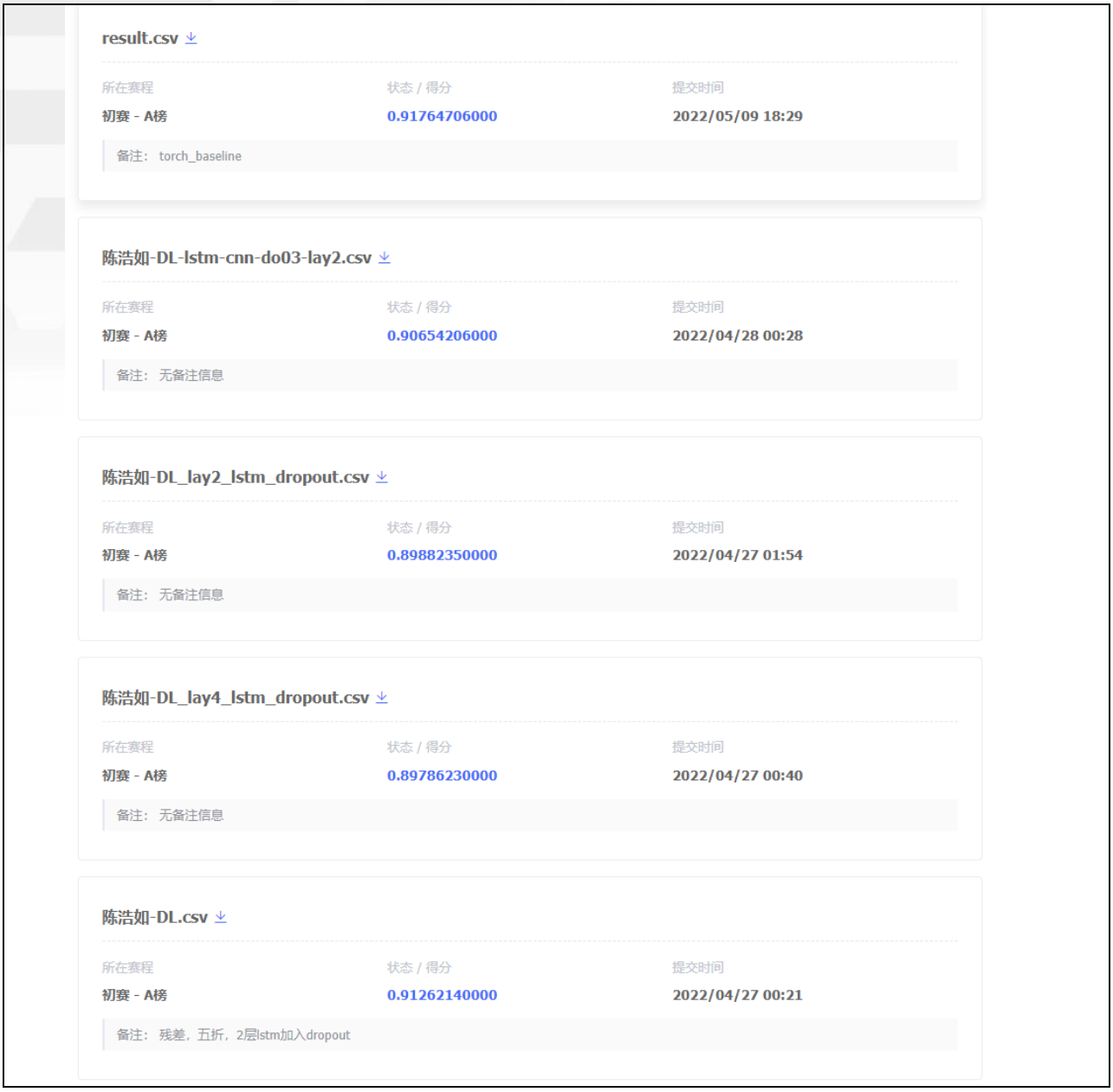
所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.89208630000	2022/05/13 19:30

备注：无备注信息

baseline.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.94202900000	2022/05/12 23:17

备注：无备注信息



result.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.91764706000	2022/05/09 18:29

备注： torch_baseline

陈浩如-DL-lstm-cnn-do03-lay2.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.90654206000	2022/04/28 00:28

备注： 无备注信息

陈浩如-DL_lay2_lstm_dropout.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.89882350000	2022/04/27 01:54

备注： 无备注信息

陈浩如-DL_lay4_lstm_dropout.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.89786230000	2022/04/27 00:40

备注： 无备注信息

陈浩如-DL.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.91262140000	2022/04/27 00:21

备注： 残差，五折，2层lstm加入dropout

陈浩如-DL.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.88834953000	2022/04/26 21:36

备注：五折 残差

陈浩如-DL_TextCNN.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.89929736000	2022/04/26 19:22

备注：五折，加入残差

result_LSTM.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.87439600000	2022/04/26 16:06

备注：LSTM+TextCNN

陈浩如-DL_TextCNN.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.91139245000	2022/04/25 01:25

备注：无备注信息

result_mix_model.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.88607590000	2022/04/25 00:50

备注：无备注信息

result_TFIDF_SVM.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.88383836000	2022/04/25 00:50

备注：五折交叉验证

result_mix_model.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.88721806000	2022/04/24 14:43

备注：无备注信息

result_TFIDF_XGBOOST.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.85641026000	2022/04/24 14:43

备注：无备注信息

result_BOW_SVM.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.87410920000	2022/04/21 17:21

备注：无备注信息

result_mix_model.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.89447240000	2022/04/21 17:21

备注：无备注信息

result_TFIDF_XGBOOST.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.85117495000	2022/04/21 17:20

备注：2500轮

result.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.90069280000	2022/04/17 23:47

备注： bert_baseline_max_len=150

result.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.91764706000	2022/04/17 23:26

备注： bert_baseline

陈浩如-DL_TextCNN.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.90547264000	2022/04/15 13:36

备注： 无备注信息

result_LSTM.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.88395065000	2022/04/13 19:20

备注： 无备注信息

result_RNN.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.89330024000	2022/04/13 14:31

备注： 无备注信息

result_TextCNN_1.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.90464544000	2022/04/12 19:40

备注： epcho = 50

result_TFIDF_SVM.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.91904765000	2022/04/05 16:42

备注： 关键词大法

result_TFIDF_SVM.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	逻辑错误 查看日志	2022/04/05 16:40

备注： 关键词大法

result_TFIDF_SVM.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.87191010000	2022/04/05 16:31

备注： 换了第二套参数

result_TFIDF_SVM.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.91041160000	2022/04/05 16:13

备注： TFIDF-SVM调参后

result_TFIDF_SVM.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.87218050000	2022/04/03 13:58

备注： 无备注信息

result_TFIDF_SVM.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.87153650000	2022/04/03 13:46

备注： 无备注信息

result_TFIDF_SVM.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.87531805000	2022/04/02 22:40

备注：无备注信息

result_TFIDF_SVM_linear.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.87531805000	2022/04/02 15:24

备注：无备注信息

result_BOW_SVM_linear_f1_917.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.88077855000	2022/04/02 14:50

备注：无备注信息

result_BOW_SVM.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.72514623000	2022/04/01 22:09

备注：无备注信息

result_Word2Vec.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.70050760000	2022/04/01 19:57

备注：无备注信息

result_TF-IDF_LogsiticR.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.85567015000	2022/04/01 16:55

备注：无备注信息

result.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.88118810000	2022/03/31 22:33

备注：随机种子2019

result.csv

所在赛程	状态 / 得分	提交时间
初赛 - A榜	0.88118810000	2022/03/31 22:26

备注：无备注信息

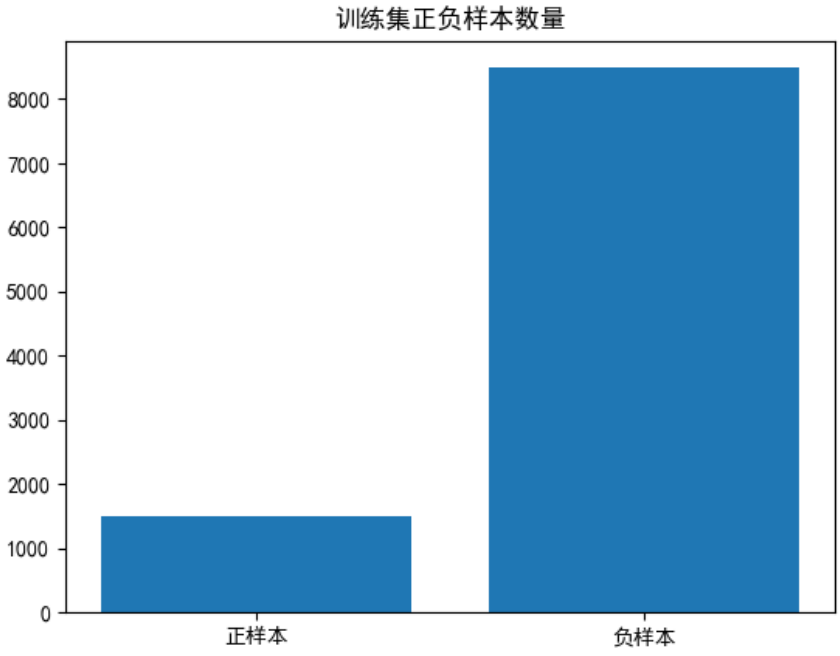
二、摘要

在新型餐饮模式的发展下，外卖行业兴起。但随着外卖行业的兴起，食品安全也不可忽视。本次文本分类任务针对外卖评论是否涉及食品安全进行二分类。本次二分类任务出于学习的目的，基于一万条训练样本，基于 BERT 对多种 BERTology 模型(Bert-wwm-ext、Bert-base、Roberta-wwm-ext、Roberta-wwm-ext-large 等等)做了尝试，同时在多次尝试修改下游任务后，得到了实验最优解，最后在算法优化上采用了 FGM 对抗训练任务，取得了线上成绩 F1=0.94964033 的成绩。

三、数据分析与数据预处理

3.1 数据可视化分析

(1)对训练集中正样本和负样本做可视化统计，输出如下



可见整个训练集中关于食品安全问题的评论还是占少数， 故在后续定义交叉熵损失函数的时候，

按样本比例调整权重输入。

(2)对训练集中每条正样本词集做交集，观察哪些词语在正样本中出现次数最多。

绘制词云图后发现，上述明显与食品卫生安全强相关的词语导致了标签为正。

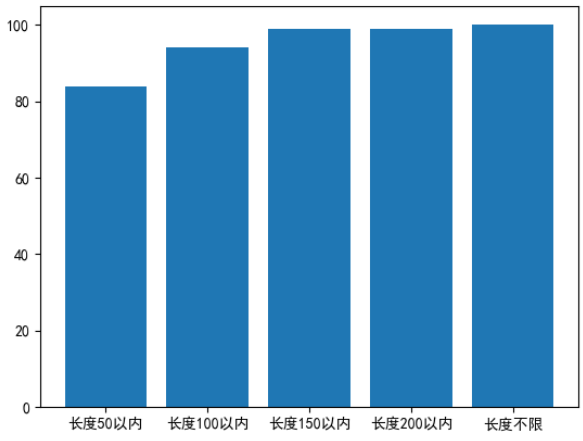
(3)判断这些高频食品安全词语在负样本中出现次数

```
cnt = 0
for i in neg['comment']:
    x = ' '.join(jieba.cut(i)).split()
    x = [y for y in x if y not in stop_word]
    ok = 0
    for j in x:
        if j in out:
            ok = 1
            print(j)
    if ok == 1:
        cnt += 1

print('食品安全高频词在负样本中出现%d次'%cnt)
```

```
发现
吐
发现
发现
发现
死
发现
死
食品安全高频词在负样本中出现108次

Process finished with exit code 0
```

可以看到这些高频词在负样本中极少出现，合理确定为特殊词语。

(4) 文本长度统计



可以看到将 max_len 设置为 200 即可，能代表绝大多数数据，并减小计算量。

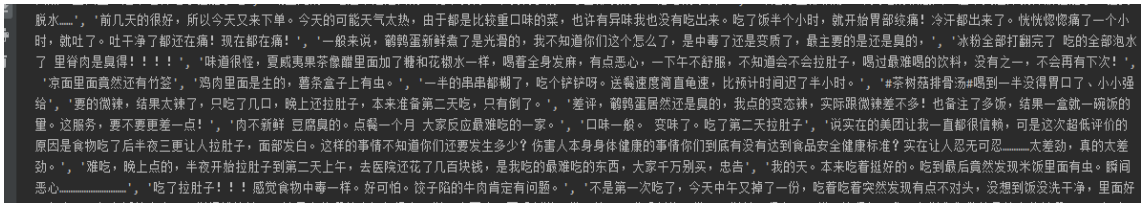
3.2 数据预处理

训练集和测试集的 comment 中存在大量表情、颜文字、图片标签、html 标签等无用数据。在调用 filter 函数通过正则匹配对这些文字过滤后，达到了较好的效果。

Filter 函数如下：

```
def filter(text):
    text = re.sub(r"[A-Za-z0-9\!\=\? \%\[\]\,\ (\) \>\<:&lt;\/#\. ----\_]", "", text)
    text = text.replace('图片', '')
    text = text.replace('\xa0', '')

    cleanr = re.compile('<.*?>')
    text = re.sub(cleanr, ' ', text)
    text = re.sub(r"\\【.*?】+|\\《.*?》+|\\#.*?#+|[.!/_,$&%^*()～<>+?@|:♥@~{}#]+|[——! \\\, 。=？、：“”‘’￥…… （）《》【】]",'',text)
    text = text.strip()
    return text
```

过滤前的文本如下，存在大量标点符号和无用的表情图。



过滤后文本就显得较为干净。

四、算法实现与算法优化

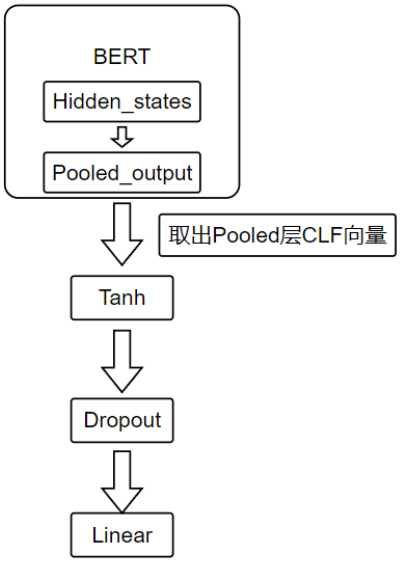
在读入预处理完毕的数据后，通过调用 transformers 库中的 BertTokenizer 的 batch_encode_plus()函数将其按照 BERT 所需的 input_ids、attention_mask、token_type_ids 和一系列参数处理好，再定义下游任务。

1. 下游任务改动(以 Bert-wwm-ext 为基础 Bert 模型)：

(1)直接调用 transformers 库中的 BertForSequenceClassification

Huggingface 的 api 可以直接输入参数，完成下游任务调优和 loss 输出以及梯度更新。

其网络结构图如下：



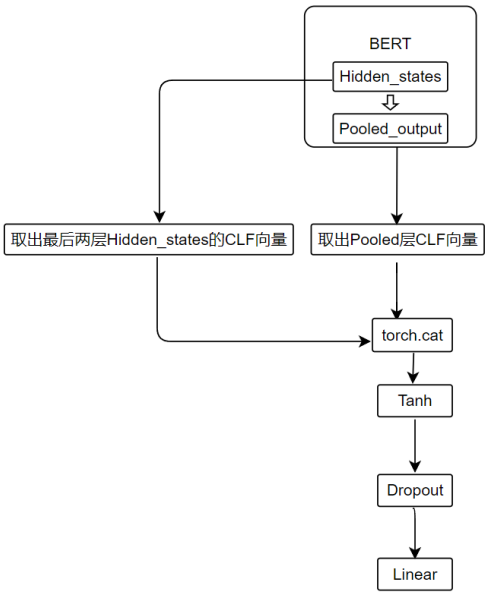
在内部损失函数是采用了 CrossEntropyLoss，且未对交叉熵设置样本权重。由于接入了 Tanh 激活函数，使得模型有了更佳的非线性拟合能力。

在最后接入 softmax 函数，对预测结果进行二分类判断。

最终得到线上测试 F1 分数为：0.93462470000。

(2)提取两层隐藏层向量和 pooled 层 CLF 向量的下游任务

在第一套下游任务的网络结构设计中，考虑到 Pooled_output 层能表达的语义信息较少，为增加更多的语义信息和数据维度,在阅读文献后选择了效果最佳的将 Pooled_output 层与最后两层 Hidden_state 的 CLF 向量取出拼接在一起。同时也加入 Tanh 激活函数，重新设计了自定义的下游任务网络结构。具体结构如下：



同时对交叉熵函数设定了样本比例权重。

```
nn.CrossEntropyLoss(weight=torch.from_numpy(np.array(10000/8500, 10000/1500)).float())
```

其中 10000 是训练样本总数，8500 是负样本个数，1500 是正样本个数，在设置了权重后，计算损失函数时会将正样本的错判惩罚力度增大。这样就能解决样本不平衡问题。

增加了语义信息后，F1 分数达到了 0.94230765000。

可以看到在改动下游任务且调整损失函数权重后得到了较高的 F1 值，后续尝试多种 BERTology 模型都会在此下游任务网络结构下讨论。

2. 不同的 BERTology 模型尝试

在对一系列 BERTology 模型做五折交叉验证测试和实验。采用在一折内只保留最好验证集 F1 分数的模型，同时对每个模型都单独调整参数。

得到下述对比表格。

模型	Batch_size	线下训练 F1	线上测试 F1
Bert-wwm-ext	16	0.937523	0.942307
albert	16	0.882301	0.892086
roberta-wwm-ext	16	0.928923	0.935561
bert-base-chinese	16	0.928152	0.935251
Bert-wwm	16	0.921501	0.934624
ernie_1.0	16	0.901352	0.915492
roberta-base	16	0.932013	0.936893
roberta-wwm-ext-large	4	0.942013	0.940047

3. 对抗训练 FGM

对上述最优 bert 模型进行线上成绩筛选，选出拥有最高成绩的三个 BERTology 模型，再加上 FGM 对抗训练进行训练。

FGM 对抗训练的原理是对 BERT 任务中的 word_embeddings 层进行对其梯度优化的恶意攻击，加入对 embeddings 的扰动，以增强模型的鲁棒性，本质是在深度学习网络的利用梯度改变参数的过程中加入一些恶意的攻击。

本项目对线上表现最优的三个模型加入了 FGM 对抗训练后，新的模型拥有更好的泛化能力。提升了接近一个百分点的线上成绩。

最终本项目分别训练的单模成绩如下：

五、模型融合

本项目模型融合共分两步：

1. 单个模型内部五折交叉验证

在经过模型内部五折交叉验证后，平均对测试集的 F1 分数能比验证集的要高上 5%左右。

2. 投票法融合多个 BERT 模型

保留线上测试成绩最高的 5 个 BERT 模型，在投票融合后得到最佳成绩：F1=0.94964033

六、实验总结

在本次实验中，我从零入门深度学习，并在 pytorch 和 keras 两种大深度学习框架下进行了一系列自己对深度学习网络的尝试和实验，对搭建自己的网络有了更深的理解，一定程度上学会了衡量一个网络性能优劣的推理方法。在此基础上，我通过本项目对多种不同的 NLP 任务进行了实践，在不断的思考和阅读文献中学习到了 NLP 数据竞赛的基础，打下了扎实的实践基础和理论基石。同时也学习到了一些 NLP 任务 trick 的使用方法，也意识到了一味往模型里加各种东西不一定能提升网络的性能。

最重要的收获是入门了 BERT 模型，以及一系列对 BERT 模型的操作。在以后的研究生生涯要更多地思考和学习，加强对 NLP 的理解和认识。