



北京航空航天大学
BEIHANG UNIVERSITY

机器学习团队作业

所选题目 垃圾邮件分类

姓	名	刘嘉明
学	号	23371007

2025 年 6 月 25 日



1 小组分工

刘嘉明	陈加玺	陈柏玺	赵敏智	杨明和
填写实验报告 编写代码	编写代码, 调试	编写代码	编写代码	填写实验报告

2 模型选择与原理

在本次作业中，我们主要尝试了三个方案，分别是朴素贝叶斯，SVM 和 Transformer 模型，最终我们选择了 Transformer 模型的输出结果作为提交答案。

1. 朴素贝叶斯：朴素贝叶斯的本质是贝叶斯公式，基本假设是“条件独立”，但事实上，“条件独立”这一原则往往不成立，在垃圾邮件识别的任务中尤为明显。因此，朴素贝叶斯方法的实验结果并不理想，正确率仅有 83%，我们最终也没有选取此方法。

2. SVM：支持向量机是一个既简单又高效的模型。实现步骤主要分为预处理，特征选取，建模，评估这四步。首先，我们在读入文本内容后并没有急于训练，而是选择了预处理的手段，通过正则表达式匹配，将所有文本转为小写，消除大小写差异；去除 HTML 标签；替换数字、网址、邮箱、货币符号；使用 Porter Stemmer 算法进行词形归并（如"running"→"run"）。之后，我们采用 TF-IDF 向量化方法，基于训练集建立 5000 维的特征空间，通过 TfidfVectorizer 计算词频-逆文档频率，并把所有文本转换为高维稀疏特征矩阵。然后，使用线性 SVM 分类器，寻找最优超平面最大化正负样本间隔。最后一步，评估模型在验证集上的正确率并对测试集进行预测。

3. Transformer：深度学习方法，通过预训练-微调范式实现了更加精准的垃圾邮件分类，使准确率从 SVM 的 98% 提升到了 99%。该方案采用 DistilBERT 作为预训练模型，在保持 97% 原始 BERT 性能的同时显著提升了推理效率。模型加载阶段先从本地读取数据，然后初始化包含 2 个输出类别的分类头。数据预处理环节通过 Hugging Face 的 Tokenizer 实现文本到数字 ID 的转换，采用 WordPiece 算法进行子词切分。训练配置采用 AdamW 优化器，进行多次重复训练-预测，选择 F1 分数作为主要评估指标。



3 模型调参与中间结果

1. TF-IDF + SVM 模型调参与中间结果:

参数	调优范围	最优值	分析
max_features	[1000, 5000, 10000]	5000	维度越高, 模型越复杂, 但可能过拟合
ngram_range	[(1,1), (1,2), (1,3)]	(1,2)	2-gram 捕捉短语
stop_words	[None, "english"]	None	保留停用词有助于垃圾邮件识别
C	[0.01, 0.1, 1]	0.1	增强正则化, 防止过拟合
class_weight	[None, "balanced"]	"balanced"	解决数据不平衡问题

2. DistilBERT (Transformer) 模型调参与中间结果

参数	调优范围	最优值	分析
pretrained_model	["bert-base", "distilbert"]	"distilbert"	速度更快
num_labels	2	2	二分类
learning_rate	[1e-5, 2e-5, 5e-5]	2e-5	微调需小学习率避免破坏预训练知识
batch_size	[8, 16, 32]	16	平衡显存和梯度稳定性
num_train_epochs	[2,3,4,5.....]	3	重复轮数
max_length	[64, 128, 256]	128	截断长文本, 保留关键信息
padding	固定	固定	统一输入长度

4 实验结果

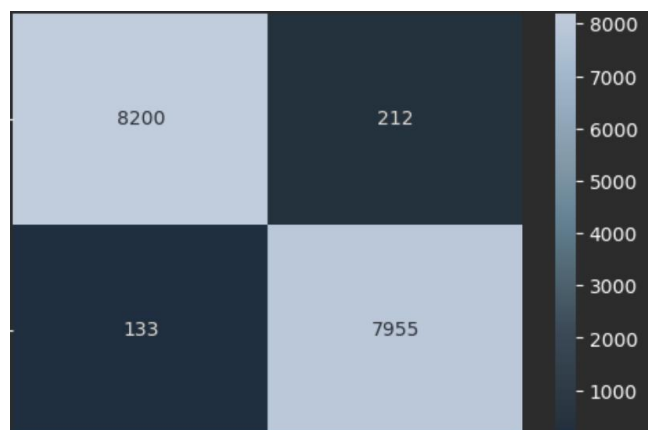
1. SVM: 计算了训练集准确率、验证集准确率、查准率、查全率、F1 分数等指标，并绘制了混淆矩阵。最后对测试集进行预测，并将结果保存在 csv 文件中。

训练准确率: 98.33%

验证准确率: 97.91%

验证集分类报告:

	precision	recall	f1-score	support
正常邮件	0.98	0.97	0.98	8412
垃圾邮件	0.97	0.98	0.98	8088
accuracy			0.98	16500
macro avg	0.98	0.98	0.98	16500
weighted avg	0.98	0.98	0.98	16500



2. Transformer: 同样，每轮都计算训练集准确率、验证集准确率、查准率、查全率、F1 分数等指标，最终选取 F1 最大者对应的模型作为测试模型，对测试集进行预测，并将结果保存在 csv 文件中。

Tracking run with wandb version 0.20.1

Run data is saved locally in /content/wandb/run-20250624_172035-6ekcr0nn

Syncing run ./results to Weights & Biases (docs)

View project at <https://wandb.ai/picnight-beihang-university/huggingface>

View run at <https://wandb.ai/picnight-beihang-university/huggingface/runs/6ekcr0nn>

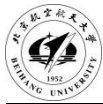
[8439/8439 27:32, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.018200	0.039204	0.989600	0.989401	0.988595	0.990208
2	0.001000	0.049311	0.991000	0.990792	0.993842	0.987760
3	0.040600	0.039849	0.993200	0.993070	0.992261	0.993880

在验证集上的最终评估结果:

[313/313 00:18]

```
{'eval_loss': 0.039848655462265015, 'eval_accuracy': 0.9932, 'eval_f1': 0.9930697105584998, 'eval_precision': 0.9922606924643584, 'eval_recall': 0.9938800489596084, 'eval_runtime': 18.9446, 'eval_samples_per_second': 263.927, 'eval_steps_per_second': 16.522, 'epoch': 3.0}
```



5 参考文献

[机器学习--SVM（支持向量机）垃圾邮件分类器 svm 分类器训练详细步骤-CSDN 博客](#)

[【论文阅读】Attention is all you need \(Transformer\) attention is all you need 原文-CSDN 博客](#)

[模式分类与应用-贝叶斯垃圾邮件分类 hewlett-packard 实验室收集的邮件 数据集 -CSDN 博客](#)