# 文本分类实验报告

## 1. 实现细节

### 1.1 Log-linear 模型

本实验中的 Log-linear 模型由 **TF-IDF 特征提取器** 和 **逻辑回归分类器** 组成：

- **TF-IDF特征提取**：
  - 使用 1-gram 和 2-gram（即单词和词组特征）。
  - 最大特征数限制为 50,000。
  - 应用 sublinear TF（对频率取对数）变换。
  - 移除英文停用词。
- **逻辑回归分类器**：
  - L2正则化，惩罚系数设为 4.0。
  - 采用 `class_weight='balanced'` 以处理类别不平衡。
  - 最大迭代次数设置为 1000，随机种子为42，保证结果可复现。

**部分关键代码：**

```
Pipeline([
    ('tfidf', TfidfVectorizer(ngram_range=(1,2), max_features=50000, sublinear_tf=True, stop_wor
    ('lr', LogisticRegression(max_iter=1000, C=4.0, class_weight='balanced', n_jobs=-1, random_s
])
```

### 1.2 BERT 模型

本实验中的 BERT 模型基于 HuggingFace Transformers 进行微调：

- **模型**：bert-base-uncased
- **输入处理**：
  - 文本最大长度为128。
  - 自动填充（padding）与截断（truncation）。
- **训练参数**：
  - 训练轮数：3 epochs
  - batch size：8

- 学习率：2e-5
- 随机种子设为42，保证结果可复现。

**部分关键代码：**

```
TrainingArguments(
    output_dir='outputs/tmp',
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=3,
    learning_rate=2e-5,
    logging_steps=200,
    seed=42
)
```

# 2. 实验结果

| 模型 | 数据集 | 集合 | Accuracy | Macro-F1 | Micro-F1 |
|---|---|---|---|---|---|
| Log-linear | 20news | 训练集 | 0.9673 | 0.9701 | 0.9673 |
| | | 测试集 | 0.6949 | 0.6857 | 0.6949 |
| Log-linear | HoC | 训练集 | 0.9770 | 0.9827 | 0.9770 |
| | | 测试集 | 0.6333 | 0.6064 | 0.6333 |
| BERT | 20news | 训练集 | 0.8964 | 0.8913 | 0.8964 |
| | | 测试集 | 0.7059 | 0.6930 | 0.7059 |
| BERT | HoC | 训练集 | 0.8622 | 0.7795 | 0.8622 |
| | | 测试集 | 0.7500 | 0.6119 | 0.7500 |

# 3. 结果分析与讨论

## 准确率比较

- 在两个数据集上，BERT模型的测试准确率普遍高于Log-linear模型。
- 特别是在HoC数据集上，BERT的测试准确率达到0.75，而Log-linear模型仅为0.6333，显示出BERT更好的迁移与泛化能力。

## 过拟合情况

- Log-linear模型在训练集上表现出极高的准确率（训练集准确率接近97%），但测试集上准确率明显下降，存在过拟合现象。
- BERT模型虽然训练集准确率略低，但测试集准确率更高，说明其拥有更强的泛化能力。

## 数据集特性影响

- **20 Newsgroups** 数据集样本较大、语言多样，Log-linear模型提取的n-gram特征可以覆盖丰富的信息，因此能取得接近70%的测试准确率。
- **HoC**（健康文本分类）数据集样本少、专业术语多，传统Log-linear模型难以捕捉深层次语义，而BERT通过预训练在医学、科技等领域知识上有更好的迁移效果，因此在HoC上表现更优。

## Macro-F1 与 Micro-F1 差异

- 两个数据集中，Macro-F1指标均低于Micro-F1，特别是HoC数据集（0.6119 vs 0.75）。
- 说明类别分布存在不均衡，小类别难以模型学习到，从而拉低了Macro-F1分数。

## 模型能力比较

- Log-linear模型依赖于表层的统计特征，对数据量大、类别词汇丰富的任务较为友好，但对小样本、专业领域任务适应性较差。
- BERT模型可以利用上下文理解文本内容，对专业领域、小数据量任务有较强适应性，同时对类别不平衡问题也具有一定的鲁棒性。