基于 DistilBERT 的 Twitter 文本情感分类实验报告

课程名称：人工智能基础

姓名：李恒　**学号：**25803090195

2025 年 12 月 6 日

摘要

本实验旨在利用预训练语言模型 DistilBERT 对 Twitter 上的英文文本进行六分类的情感识别（Anger, Disgust, Fear, Joy, Sadness, Surprise）。实验主要对比了两种策略：一是将 DistilBERT 作为固定的特征提取器，配合 Scikit-Learn 中的传统分类模型；二是对 DistilBERT 进行端到端的全参数微调（Fine-tuning）。实验结果显示，特征提取方法的最高准确率为 62.3%（RidgeClassifier），而微调方法的准确率最高可达 94.1%。此外，实验重点探究了学习率对微调效果的决定性影响，分析了高学习率导致“模型坍塌”的原因，并通过对抗样本测试揭示了模型在处理反讽及隐晦表达时的局限性。

关键词： DistilBERT；情感分析；微调；Transformer；对抗样本

1 引言

文本情感分析是自然语言处理（NLP）领域的核心任务之一。随着深度学习的发展，基于 Transformer 架构的预训练模型（如 BERT）在各类 NLP 任务中取得了显著效果。DistilBERT 作为 BERT 的轻量化蒸馏版本，在保留了 BERT 97% 性能的同时，减少了 40% 的参数量并提升了 60% 的推理速度，非常适合资源受限场景下的应用。

本实验基于 Hugging Face 生态系统，使用 CARER 数据集进行训练。实验的核心目标不仅是构建高准确率的情感分类器，更在于探究不同训练策略（特征提取 vs 微调）以及超参数设置对 Transformer 模型性能的影响，并深入分析模型在面对复杂语境时的表现。

2 方法

2.1 数据预处理与标记化

模型无法直接处理原始文本，需将其转换为数值向量。实验首先尝试了字符级标记化（Character Tokenization）以理解底层原理。

任务 1 实现：创建 `token2idx` 字典

为了将文本中的每个字符映射为唯一整数，我们构建了如下字典：

```
unique_chars = sorted(set(tokenized_text))
token2idx = {char: idx for idx, char in enumerate(unique_chars)}
```

Listing 1: 字符级标记化实现

尽管字符级标记化实现简单，但缺乏语义信息。因此，后续实验主要采用 DistilBERT 自带的 WordPiece 分词器进行子词（Subword）级别的标记化。

2.2 特征提取 (Feature Extraction)

在特征提取实验中，我们冻结 DistilBERT 的参数，仅将其作为编码器提取文本的隐藏层表示。具体的，我们提取最后一层中对应 `[CLS]` 标记的向量（维度 768）作为句子的特征表示。

任务 2 实现：提取隐藏层 [CLS] 特征

在实现过程中，为了解决 `Dataset.map` 处理批次数据时可能出现的格式类型不一致问题（List vs Tensor），我们增强了函数的鲁棒性，显式进行了类型转换：

```
def extract_hidden_states(batch):
    # Place model inputs on the GPU
    # 增加 torch.tensor() 转换以防止输入为 list 时调用 .to() 报错
    batch = {k:torch.tensor(v).to(device) for k,v in batch.items() if k in
   ["input_ids", "attention_mask"]}
```

```
    # Extract last hidden states
    with torch.no_grad():
        outputs = model(**batch)
    # Return vector for [CLS] token
    return {"hidden_state": outputs.last_hidden_state[:, 0].cpu().numpy()}
```

Listing 2: 增强鲁棒性的特征提取函数

出现格式问题是因为，尽管在调用

```
emotions_hidden = emotions_encoded.map(extract_hidden_states, batched=True)
```

之前已经将张量格式设置为 torch，

```
emotions_encoded.set_format("torch",
                            columns=["input_ids", "attention_mask", "label"])
```

但是因为在 map 的过程中传入了 batched=True，所以还是会直接从底层读取数据而不是用 torch 的切片方法读取，格式不是 torch，只能强制转换。

也正因如此，后面在训练模型之前，必须再次调用 set_format 修改格式，否则无论如何调整超参数，都会出现模型坍塌，模型根本无法正确读取数据，没有学习到任何有价值的东西，导致模型预测时全部选择 Joy。

3 实验结果与分析

3.1 基于特征提取的分类模型对比

利用提取的 768 维特征向量，我们训练了多种 Scikit-Learn 分类器。实验结果如表 1 所示。

表 1: 基于特征提取的不同分类模型性能对比

模型名称	准确率 (Accuracy)	备注
RidgeClassifier	**0.6230**	**训练速度最快，效果最佳**
RidgeClassifierCV	0.6205	带交叉验证
SGDClassifier	0.5545	随机梯度下降
PassiveAggressiveClassifier	0.5490	被动攻击算法
LogisticRegression	0.5470	逻辑回归
Perceptron	0.5340	感知机

分析：从实验结果可以看出，**RidgeClassifier**（岭回归分类器）取得了 62.3% 的最高准确率，显著优于传统的逻辑回归和感知机模型。这主要归因于 DistilBERT 预训练输出的 768 维 [CLS] 向量在高维空间中已经具备了一定的线性可分性。Ridge 分类器通

过引入 L2 正则化项，有效限制了模型系数的大小，防止了在高维稀疏特征上的过拟合，从而表现出更强的泛化能力。相比之下，SGDClassifier 和 Perceptron 对于噪声较为敏感，且在未调整步长和正则项等参数的情况下，难以在预训练数据集中找到最优解。

然而，该方法的整体性能存在明显瓶颈，只能获得约 63% 的准确率上限。根本原因在于**“语义鸿沟”**：作为特征提取器的 DistilBERT 参数被冻结，其输出的是通用的语言学表征，而非针对本次六分类情感任务优化的特征空间。对于“Joy”（喜悦）和“Love”（爱）这类在通用语义上极度接近的类别，固定的特征提取器无法动态调整嵌入空间以扩大各个情感类之间的距离，导致分类器难以通过简单的线性或非线性边界进行有效区分。

3.2 微调策略与参数调优

我们对 DistilBERT 进行了全参数微调，重点研究了学习率（Learning Rate）对训练效果的影响。不同参数下的实验数据如表 2 所示。

表 2: 微调超参数对模型性能的影响 (Batch Size=64, Epochs=2)

Learning Rate	**Weight Decay**	**Accuracy**	**F1 Score**	**现象描述**
1e-3	0.01	0.3520	0.1833	**模型坍塌**
1e-5	0.01	0.8640	0.8504	欠拟合，收敛慢
2e-5	0.01	0.9245	0.9245	效果优异
5e-5	**0.01**	**0.9410**	**0.9411**	**最佳效果**
8e-5	0.01	0.9390	0.9388	开始过拟合
1e-5	0.001	0.8600	0.8602	欠拟合，收敛慢

分析：

模型坍塌 (Mode Collapse)：实验中最显著的现象是当学习率设为 `1e-3` 时，模型准确率停滞在 0.352。这一数值恰好等于数据集中样本数最多的类别（Joy）的占比。从优化理论角度分析，Transformer 模型的预训练权重位于损失函数曲面的一个特定局部极小值区域。对于微调任务，`1e-3` 的学习率相对于预训练参数而言过大，导致参数更新的步长超出了预训练权重的“安全区域”。

这种剧烈的梯度更新破坏了模型在预训练阶段习得的句法和语义知识。此时，模型实际上已经失去了理解文本的能力，退化为一个随机初始化网络。为了最小化全局损失，模型收敛到了统计层面上的“安全解”，即无视输入内容，始终预测出现频率最高的类别。这就是准确率锁定在多数类占比（35.2%）的根本原因。

lr 最佳区间：实验验证了 `2e-5` 至 `5e-5` 是微调 Transformer 的最佳学习率区间。在此数量级下，梯度更新既足以让模型参数适配特定任务的分布，又保留了预训练模型通用的语言表示能力。当学习率过小（如 `5e-6`）时，模型收敛极慢且易陷入鞍点；而当学

习率稍大（如 8e-5）时，验证集 Loss 的反弹表明模型已开始在训练数据上过拟合，泛化能力受损。

正则化系数：Weight_Decay 对模型效果基本没有显著影响。

3.3 正误分析

通过分析验证集中 Loss 较高（分类错误）和 Loss 较低（分类正确）的样本，我们发现：

- **分类正确原因**：样本中包含显著的情感强特征词，如"happy", "terrified"，且句子结构简单。模型很容易分析句子结构，抓住强烈、明显的情感特征词做出判断。
- **分类错误原因**：

  标签模糊：一些情感之间的界限并不是那么明显，也可能会有交集。如 Surprise 常被误判为 Joy 或 Fear，这是因为惊讶往往伴随着后续的正向或负向情绪，单纯的惊讶很难界定。

  隐晦表达：使用修辞而非直白情绪词，例如"The silence was deafening"（震耳欲聋的沉默）。这种语义含量很高的句子难以被模型学习到并判断准确。

  人工标注：一条推文中可能包含多种情绪，人工标注只有一种，导致模型困惑；人工标注本身就有错误，相当于用来衡量模型能力的标准就出现了问题，模型虽然预测正确但是被评为错误。

3.4 对抗样本测试与鲁棒性分析

为了评估模型在面对复杂语言现象时的鲁棒性，我们利用上面总结出的“隐晦表达”错因，构建了 5 条对抗样本，涵盖反讽、长距离依赖、歧义及常识推理等场景。实验结果（表 3）显示模型在部分样本上表现出了局限性。

表 3: 对抗样本预测结果分析

对抗样本内容	预测结果	真实情感	结果分析
”Great, my car broke down again. Just what I needed.”	**Sadness**	Anger	**部分正确**。模型成功忽略了反语词”Great” 的字面含义，捕捉到了负面情绪，但将“故障”理解为悲伤而非反讽语境下的愤怒。
”I cried my eyes out when I saw the ending, it was the most beautiful thing.”	**Joy**	Joy	**预测正确**。模型成功通过注意力机制将重心放在了”beautiful thing” 上，克服了前半句”cried” 的负面干扰。
”I am not exactly sure that I would call this a happy experience.”	**Joy**	Sadness	**预测错误**。模型未能处理”not exactly sure... would call” 这一复杂的否定结构，直接被句末的”happy” 误导。
”Fine.”	**Joy**	Anger	**预测错误**。在缺乏上下文时，模型倾向于将中性词按字面积极义处理，无法识别被动攻击语气。
”The clown smiled at the child from the dark sewer grate.”	**40%Anger** **35%Joy**	Fear	**严重混淆**。模型对”dark sewer” (Anger) 和”smile” (Joy) 产生了特征冲突，且因缺乏外部常识，完全未预测出”Fear”。

结果分析：实验结果表明，经过微调的 DistilBERT 模型具备了一定的上下文注意力机制。例如在样本 2 中，尽管出现了强烈的负面词”cried”，模型仍能结合后半句做出正确判断，这证明模型已经超越了简单的词袋匹配。

然而，样本 3 和样本 5 暴露了模型的显著缺陷：

对复杂否定的盲区：在样本 3 中，面对“弱否定词 + 长距离修饰 + 强情感词”的结构，模型的注意力机制失效，忽略了否定逻辑，直接被”happy” 吸引。

常识推理缺失：样本 5 的结果最具启发性。模型预测出 40% 的 Anger（可能源于 sewer/dark）和 35% 的 Joy（源于 smile），显示出模型在预测时内部出现了特征冲突。模型未能预测出 Fear，是因为它只学习了词汇共现的统计规律，而缺乏“下水道里的小丑是危险的”这一外部世界常识（World Knowledge）。

4 总结

本实验基于 DistilBERT 模型，系统地探究了 Twitter 文本情感分类任务中不同训练范式的有效性与局限性。实验结果确凿地证明，将预训练模型与下游任务进行端到端的全参数微调，其准确率达到了 94.1%，显著优于基于静态特征提取的传统机器学习方法（62.3%）。这一巨大的性能鸿沟揭示了通用语言表征与特定情感分类任务之间存在显著的语义差异，而微调过程能够有效地重塑特征空间以适配特定领域的分布。

在微调机制的探索中，实验深入验证了超参数敏感性对模型收敛的决定性影响。特别是针对“模型坍塌”现象的复现与分析表明，过大的学习率（如 1e-3）会引发灾难性遗忘，导致模型退化为仅依赖先验概率的多数类猜测器。实验确认了 2e-5 至 5e-5 为微调 Transformer 的最佳学习率区间，在此区间内模型能有效平衡预训练知识的保留与新任务特征的学习。

尽管模型在标准测试集上表现优异，但对抗样本的测试结果暴露了其深层语义理解能力的不足。虽然模型展现出了一定的上下文捕捉能力，能够克服部分词汇干扰（如在转折句中正确判断情感），但在面对复杂的长距离否定结构以及涉及外部常识推理的场景（如“下水道的小丑”）时，模型依然表现出对浅层词汇特征的过度依赖，导致特征冲突或判断失误。综上所述，未来的研究应致力于引入对抗训练或结合外部知识库，以弥补预训练模型在逻辑推理与常识理解层面的短板，从而构建更加鲁棒的情感分析系统。