基于 IMDB 影评数据集的情感分类任务

一、实验简介和动机

本实验基于 IMDB 电影评论数据集，构建并对比了三种主流深度学习架构（TextCNN、BiLSTM+GloVe、DistilBERT）在情感二分类任务上的性能。IMDB 影评情感分类是自然语言处理领域的经典基准任务，要求模型从长篇英文文本中准确捕捉作者的情感倾向（正面/负面），对模型的语义理解、长距离依赖建模和泛化能力均有较高要求。

选择这一实验题目主要出于以下动机：第一，情感分类作为 NLP 入门任务，技术链条完整，便于完整地实践从数据预处理、模型搭建、训练评估到结果分析的整个流程；第二，不同模型架构（CNN、RNN、Transformer）在文本建模上的哲理差异显著，通过控制变量实验可以直观地比较其优势与局限，加深对深度学习模型选择的理解；第三，近年来预训练语言模型（如 BERT）在各类 NLP 任务中占据主导，我们希望通过引入轻量级的 DistilBERT 以及静态预训练词向量 GloVe，探究“预训练+微调”范式相比于传统随机初始化模型带来的提升幅度，并分析词向量质量与模型结构之间的交互关系。

实验的选题来源于 Final Project 的参考选题。在此基础上，我进行了以下实验：

①**系统化的超参数调优**：以 TextCNN 为对象，围绕嵌入维度、卷积核大小组合、卷积核数量、学习率和 Dropout 等关键参数设计了多组控制变量实验。通过逐次单变量调整并记录验证曲线与测试指标，明确了各参数对模型收敛、泛化的影响规律，最终筛选出 embedding_dim=300、filter_sizes=[3,4,5,6]、lr=0.001、dropout=0.5 的最优组合，测试准确率达 84.7%。

②**词向量初始化策略的对比实验**：在 BiLSTM 框架下，我们分别使用随机初始化词向量和 GloVe 6B.100d 预训练词向量进行训练。随机初始化时模型严重欠拟合，测试准确率仅约 55%，且完全无法收敛；引入 GloVe 后准确率大幅提升至 81.1%，训练过程平稳收敛。这一对比清晰地揭示了词向量质量对 RNN 类模型的决定性作用。进一步地，我们与基于动态上下文词向量的 DistilBERT（准确率 90.0%）形成“随机初始化—静态预训练—动态预训练”的完整递进对比，多维度论证了词表征在情感分类中的核心地位。

③**多模型统一对比框架**：为保证公平评估，所有模型均采用完全一致的预处理流水线（清洗、分词、词表、序列填充）和相同的训练/验证/测试划分。我们在该框架下对比了三大类经典架构：CNN 代表：TextCNN（基线 84.5%，最优 84.7%）；RNN 代表：LSTM（54.0%）、BiLSTM（55.1%）以及经 GloVe 优化的 BiLSTM（81.1%）；Transformer 代表：

DistilBERT（90.0%）。通过统一的基准平台，我们得以客观地分析不同模型结构的适用场景和性能瓶颈，避免了因数据预处理差异而引入的变量。

④交互式 Demo 开发：基于 Gradio 搭建了三模型实时对比的情感预测网页界面。用户可输入任意英文影评，界面将同步展示 TextCNN、BiLSTM+GloVe 以及 DistilBERT 的预测标签与置信度。这一 Demo 不仅直观验证了各模型的预测行为，也增强了实验成果的展示性与可交互性。

通过上述工作，我们不仅完成了课程要求的基本建模任务，更深入探讨了结构设计、词向量初始化、超参数选择对模型性能的综合影响，为后续更复杂的 NLP 任务积累了工程经验。

二、小组分工

本实验由本人独立完成，在实验设计、代码调试和报告撰写过程中适当借助 AI 辅助工具。具体工作涵盖以下方面：

数据预处理与词表构建：独立完成 IMDB 数据集的下载、清洗、分词、词频统计、词表构建，以及训练/验证/测试集划分，实现了可复用的预处理流水线。

TextCNN 模型与超参数实验：独自搭建 TextCNN 网络，设计并执行了嵌入维度、卷积核组合、学习率、Dropout 等关键参数的多组对照实验，记录实验结果并完成最优模型筛选。

循环神经网络建模与词向量对比：独立构建单层 LSTM、BiLSTM，以及引入 GloVe 预训练词向量的 BiLSTM 模型，完成训练与评估，对比分析了随机词向量与预训练词向量的性能差异。

DistilBERT 微调与性能评估：自主完成 DistilBERT 预训练模型的加载、微调、验证与测试，对比传统模型，展现预训练 Transformer 的优势。

交互式 Demo 开发：单独使用 Gradio 框架，将 TextCNN、BiLSTM+GloVe 和 DistilBERT 封装为三模型实时预测界面，实现输入影评后的多模型情感对比展示。

实验报告与 PPT 制作：独立撰写完整的实验报告，并制作课堂展示 PPT。

以上所有工作均由本人承担，无其他成员参与。

三、实验内容

（1）数据预处理

实验使用课程提供的IMDB影评数据集，包含训练集question2_train.csv（25000条）和测试集question2_test.csv（25000条），每条数据由评论文本text和情感标签label（0=负面，1=正面）组成，正负样本各占50%。预处理流程如下：

①文本清洗：原始影评中包含大量HTML标签（如
）、标点符号及大小写混杂。我们使用BeautifulSoup去除HTML标签，利用正则表达式去掉所有非字母字符，统一转换为小写，并合并多余空白，最终得到纯英文单词序列。

②分词与词表构建：以空格为分隔符对清洗后的文本进行分词。统计训练集中所有词的频率，仅保留出现次数≥5的词，以减少低频噪声并控制词表规模。同时加入两个特殊标记：<PAD>（索引0，用于序列填充）和 <UNK>（索引1，用于未登录词）。最终词表大小为30739。

③序列编码与填充：将每条影评的词序列映射为对应的索引序列，并统一长度max_len=200。长度不足者末尾补<PAD>，超出者截断前200词。

④数据集划分：原始训练集按8:2比例分层抽样，划分为训练集（20000条）和验证集（5000条），保证正负比例一致。测试集保持25000条不变，仅在最终评估时使用。所有模型共用此划分，确保公平对比。

（2）模型设计与实现

本实验共实现了三类深度学习模型，覆盖CNN、RNN和Transformer三种主要架构。

①TextCNN（卷积神经网络）

TextCNN是Kim (2014)提出的经典文本分类模型，核心思想是利用多尺寸卷积核在词向量矩阵上提取不同粒度的局部N-gram特征。其训练配置为：优化器Adam，损失函数交叉熵，默认学习率0.001，批大小64，训练5个epoch。其结构设计为：

嵌入层：将词索引映射为稠密向量，维度可在实验中调整（50/100/200/300）。

卷积层：使用多种尺寸的一维卷积核（滑动方向为序列方向），每种卷积核各100个，对嵌入矩阵做卷积操作。默认使用 [3,4,5] 三种尺寸，最优配置扩展为 [3,4,5,6]。

池化层：对每个卷积核输出的特征图做全局最大池化，提取最显著特征。

分类层：拼接所有池化结果，经Dropout正则化后送入全连接层，输出2分类logits。

②LSTM / BiLSTM（循环神经网络）

RNN 类模型按时间步顺序处理文本序列，能够建模长距离依赖关系，理论上更适合捕捉句子级的情感走向。单层 LSTM（全连接取最后一时间步）：嵌入层随机初始化，维度 100；单层单向 LSTM，隐藏层维度 128；取最后一个时间步的隐藏状态→Dropout→全连接层。BiLSTM（双向 LSTM）：在 LSTM 基础上增加反向序列读取，双向隐藏状态拼接后得到 256 维表示，其余结构与 LSTM 相同。

上述两个模型在随机初始化词向量时表现极差，因此进一步引入 GloVe 预训练词向量：BiLSTM + GloVe：嵌入层维度设为 100，权重使用 GloVe 6B.100d 预训练向量初始化，加载时与词表对齐，未匹配到的词随机初始化，嵌入权重允许微调，LSTM 采用单层双向，隐藏层 128，其余训练配置与 TextCNN 相同。

③DistilBERT（预训练 Transformer）

DistilBERT 是 BERT 的蒸馏压缩版，保留了 97%的性能同时参数量减少 40%。本实验将其作为 Transformer 架构的代表。DistilBERT 自带动态上下文词向量，在预训练阶段已习得丰富的语言知识，因此仅需少量微调即可适配下游任务。

加载与微调方式：使用 HuggingFace transformers 库加载 distilbert-base-uncased 预训练权重，在模型顶端添加 2 输出的分类头，采用 Trainer API 进行全参数微调：学习率 2e-5，批大小 16，最大序列长度 200，训练 2 个 epoch，每 epoch 评估验证集准确率，保留最佳模型。

（3）超参数实验设计（针对 TextCNN）

为系统探究超参数对模型性能的影响，我们以一组默认配置为基准，采用控制变量法逐次调整单个参数，进行多组对比实验。默认参数及搜索范围如下：

超参数	默认值	实验值
词向量维度 (embedding_dim)	100	50, 100, 200, 300
卷积核数量 (num_filters)	100	50, 100, 150, 200
卷积核大小组合 (filter_sizes)	[3,4,5]	[2,3,4], [3,4,5], [4,5,6], [3,4,5,6]
学习率 (lr)	0.001	0.0001, 0.0005, 0.001, 0.005, 0.01
Dropout	0.5	0.3, 0.5, 0.7, 0.8

每组实验固定训练 5 个 epoch，记录验证集最佳准确率和测试准确率。所有实验结果通过命令行参数脚本自动追加记录至 experiment_results.csv，并生成对应训练曲线和

混淆矩阵，便于后续汇总分析。

（4）交互式 Demo 开发

为直观展示不同模型的预测行为，我们使用 Gradio 框架搭建了三模型实时对比的情感预测 Web 界面。该 Demo 既可作为实验成果的可视化验证，也为课堂展示提供了直观的交互手段。

加载已保存的 TextCNN（best_combo）、BiLSTM+GloVe 和 DistilBERT 最优模型；编写统一预处理函数和推理接口；设计三标签页界面："三模型对比"标签页以表格形式同时展示三个模型对同一影评的预测标签和置信度；另外三个独立标签页分别展示各模型单独预测的结果；内置示例影评方便快速体验。TextCNN 和 BiLSTM 共享同一个词汇表，推理时需保持与训练阶段完全一致的清洗和编码流程；DistilBERT 使用其专属 tokenizer 进行编码。

（5）实验环境

硬件：CPU（无 GPU）

操作系统：Windows 11

编程语言：Python 3.10

深度学习框架：PyTorch（CPU 版本）

预训练模型库：HuggingFace Transformers

关键依赖：pandas, scikit-learn, matplotlib, seaborn, nltk, BeautifulSoup, tqdm, Gradio

四、实验结果分析

（1）整体性能对比

在统一的预处理和数据集划分下，各模型在 IMDB 测试集上的分类指标汇总如下表所示：

模型	测试准确率	Precision (macro)	Recall (macro)	F1-score (macro)
TextCNN (baseline)	84.50%	0.84	0.84	0.83
TextCNN (best_combo)	84.70%	0.85	0.85	0.85
LSTM (随机词向量)	54.00%	0.58	0.54	0.46
BiLSTM (随机词向量)	55.10%	0.55	0.55	0.55
BiLSTM + GloVe	81.10%	0.81	0.81	0.81
DistilBERT	90.00%	0.90	0.90	0.90

从表中可以看出：传统 CNN 模型（TextCNN）即使使用随机词向量，也能达到约 84.5% 的准确率，经过超参数调优后小幅提升至 84.7%，表现稳健。RNN 系列模型在随机初始化下几乎完全无法工作（准确率仅比随机猜测高约 4%），但引入 GloVe 预训练词向量后，BiLSTM 的准确率跃升至 81.1%，充分说明了词向量质量对 RNN 的重要性。预训练 Transformer（DistilBERT）以 90.0%的准确率大幅领先所有其他模型，同时精确率、召回率和 F1 值均达到 0.90，证明其强大的语义理解和泛化能力。

（2）超参数对 TextCNN 性能的影响

为探究各超参数对 TextCNN 的影响，我们进行了控制变量实验，结果总结如下：

①词向量维度（embedding_dim）

实验名称	embedding_dim	验证集最佳准确率	测试准确率
A1_embed50	50	0.8208	0.8234
A2_embed100	100	0.8510	0.8448
A3_embed200	200	0.8536	0.8423
A4_embed300	300	0.8584	0.8453

维度从 50 增至 100 时，准确率从 82.3%提升至 84.5%，继续增大到 200 或 300 时提升极小，但训练时间成倍增加。这说明 100 维已能充分编码语义信息，更高维度带来的边际收益有限。

②卷积核数量（num_filters）

实验名称	num_filters	验证集最佳准确率	测试准确率
B1_filters50	50	0.8332	0.8302
B2_filters100	100	0.851	0.8448
B3_filters150	150	0.8514	0.8421
B4_filters200	200	0.8450	0.8412

从 50 增加到 100 时，准确率从 83.0%升到 84.5%，再增加则略有下降，说明 100

个核在特征丰富度和过拟合风险间达到平衡。

③卷积核大小组合（filter_sizes）

实验名称	filter_sizes	验证集最佳准确率	测试准确率
C1_kernel234	[2,3,4]	0.8484	0.8414
C2_kernel345	[3,4,5]	0.8510	0.8448
C3_kernel456	[4,5,6]	0.8358	0.8357
C4_kernel3456	[3,4,5,6]	0.8536	0.8464

在[3,4,5]基础上增加 6-gram 核，准确率提升至 84.6%；仅用较大核[4,5,6]则降至 83.6%，说明多粒度特征互补有助于捕捉不同长度的情感短语。

④学习率（lr）

实验名称	lr	验证集最佳准确率	测试准确率
D1_lr00001	0.0001	0.7530	0.7562
D2_lr00005	0.0005	0.8150	0.8205
D2_lr0001	0.001	0.8510	0.8448
D4_lr0005	0.005	0.8386	0.7955
D5_lr001	0.01	0.8218	0.6856

学习率为 0.001 时性能最佳（84.5%），0.0001 时下降至 75.6%（欠拟合），0.01 时急剧下降至 68.6%（训练崩坏），表明学习率对收敛至关重要。

⑤ Dropout 比例（dropout）

实验名称	dropout	验证集最佳准确率	测试准确率
E1_dropout03	0.3	0.8588	0.8413
E2_dropout05	0.5	0.8510	0.8448
E3_dropout07	0.7	0.8294	0.8248
E4_dropout08	0.8	0.8034	0.7986

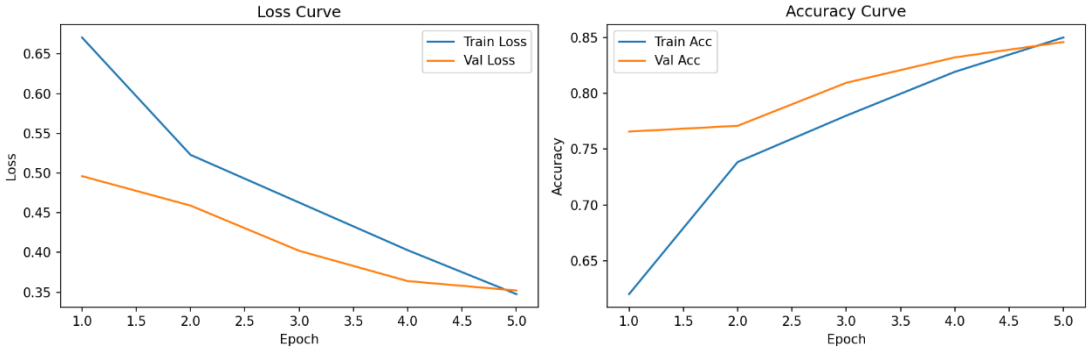
Dropout 为 0.5 时效果最好，0.8 时降至 79.9%，过强的正则化导致模型表达能力不足。

最终选取各参数的近似最优值组合（embedding_dim=300, filter_sizes=[3,4,5,6], lr=0.001, dropout=0.5）进行训练，得到测试准确率 84.7%的相对最优 TextCNN 模型。

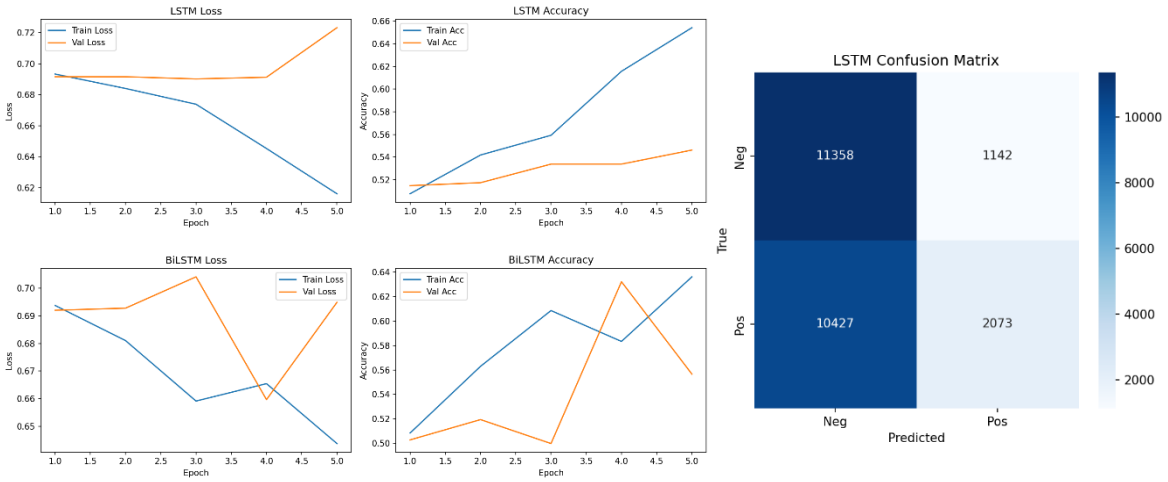
（3）训练过程可视化与收敛分析

我通过绘制训练/验证损失曲线和准确率曲线，直观地分析了模型的训练动态。

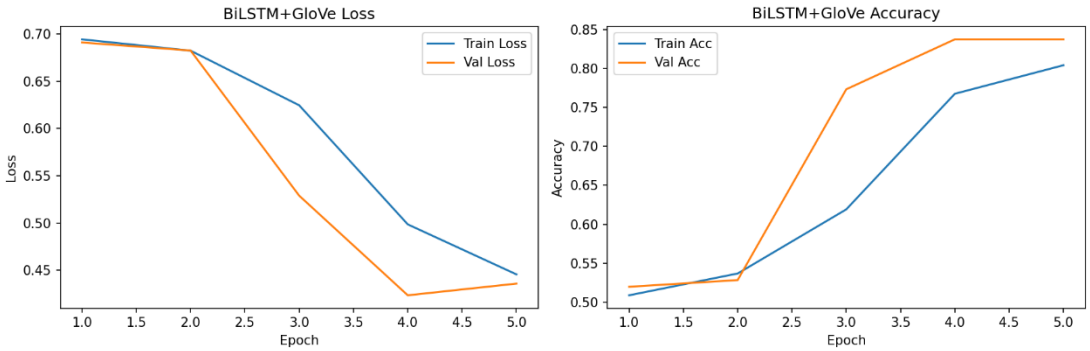
TextCNN：损失曲线平滑下降，验证损失与训练损失同步降低并逐渐趋平，在第 3 ~ 4 epoch 后验证损失不再明显下降，准确率稳定在 84%左右，显示模型**正常收敛，未出现严重过拟合。**



随机词向量的 LSTM/BiLSTM：训练损失缓慢下降，但验证损失在初始阶段即停滞或上升，准确率在 50% ~ 55% 剧烈震荡。这是典型的**欠拟合+不收敛**表现：模型容量不足以从随机词向量中学习文本规律，最终退化为“几乎全预测为负面”的策略（由混淆矩阵可以看出）。



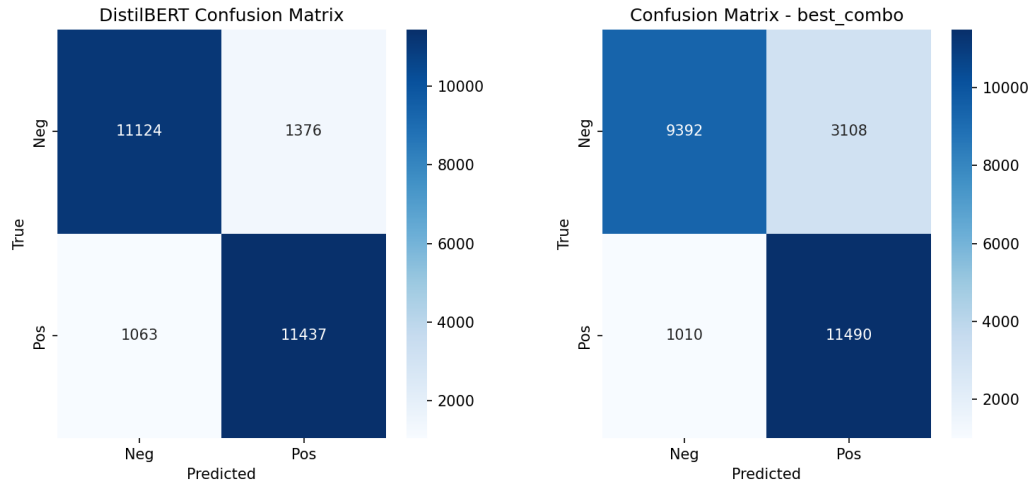
BiLSTM + GloVe：引入 GloVe 后，训练曲线变为平滑下降，验证准确率稳步上升至 84%，整体表现为**收敛**。



DistilBERT：仅 2 个 epoch 的微调，训练损失从 0.69 降至 0.25，验证准确率即达到 90.3%，**收敛迅速且泛化极佳**，展现了预训练模型的强大威力。

（4）混淆矩阵分析

我们绘制了各模型在测试集上的混淆矩阵。以 DistilBERT 为例，其对正面和负面样本的召回率分别为 0.91 和 0.89，误判分布均匀。TextCNN (best_combo) 的混淆矩阵则显示：正面召回率高达 0.92，但负面召回率仅 0.75，**模型偏向将部分负面影评误判为正面**。这与训练数据中可能存在的正面词汇在负面样本中偶发出现有关，也可能是模型对否定词（如“not good”）等局部特征的捕捉仍不够精确。BiLSTM+GloVe 的混淆矩阵两端较为平衡（0.79/0.84），但仍略逊于 DistilBERT。



（5）模型性能差异的原因分析

三类模型在测试准确率上呈现出从 55%到 90%的显著差异，其根本原因可归纳为词表示质量、模型结构特性以及预训练知识迁移三方面。

词向量质量对序列模型具有决定性影响：随机初始化的词向量未包含任何语义先验，LSTM/BiLSTM 在训练过程中难以从零习得有意义的单词表示，导致梯度更新失效，最终陷入几乎全预测为负面的退化解，准确率仅有 54%~55%。引入 GloVe 预训练词向量后，BiLSTM 的准确率大幅提升至 81.1%，说明对于依赖顺序读入单词的循环网络，高质量的静态词向量能够提供必要的语义起点，使模型得以正常收敛。相比之下，TextCNN 通过卷积核提取局部短语特征，即使词向量随机也可在训练中快速适配，因此其对词向量质量的依赖程度低于 RNN。

模型结构决定了特征提取的偏好与局限：TextCNN 利用不同尺寸的卷积核捕捉局部 N-gram 特征，与情感分析中起关键作用的情感短语高度吻合，因此能以较浅的结构取得 84.7%的准确率。BiLSTM 具备建模长程依赖的理论能力，但在实际训练中受序列填充带来的噪声和梯度衰减等因素影响，难以有效利用远距离语境，因此即便使用了 GloVe，其性能（81.1%）仍略低于 TextCNN。这表明在文本长度有限且填充较多的场景下，循环网络的长距离优势并不明显。DistilBERT 基于自注意力机制，允许每个词直接与全文任意位置的词进行交互，既保留了局部搭配信息，也能捕捉全局语义关系，结构上更加适合需要综合全文线索的情感分类任务。

预训练知识迁移使模型起点显著不同：DistilBERT 在大规模语料上经过了充分的预训练，其参数中已蕴含丰富的语言知识和上下文建模能力。在 IMDB 数据集上进行微调时，模型仅需调整分类头并小幅更新内部参数，即可快速达到 90.0%的准确率，收敛速度极快。而 TextCNN 和 BiLSTM 均从零开始学习，既缺乏词义先验，也不具备句法常识，只能完全依赖有限标注数据。这是导致 DistilBERT 性能大幅领先的核心原因。

更好的词表示、更适配任务的结构以及充分的预训练，是提升文本分类效果的三个关键维度。在本实验中，DistilBERT 凭借三者的共同优势取得了最佳性能；TextCNN 以简洁的结构和高效训练成为性价比最高的方案；BiLSTM 则通过引入 GloVe 证明了预训练词向量对于循环网络的重要性，但其结构本身的局限仍制约着进一步提升空间。

五、后续工作展望

引入注意力机制：在 BiLSTM 基础上添加自注意力层，有望进一步捕捉长距离依赖，缩小与 Transformer 的差距。

数据增强：针对否定句式、转折等复杂语义，可采用回译、同义词替换等方法扩充训练数据，提升模型鲁棒性。

模型集成：尝试对 TextCNN、BiLSTM 和 DistilBERT 的输出进行加权投票，或许能获得比单一模型更高的准确率。

错误案例分析：对 DistilBERT 仍然分错的样本进行人工分析，归纳其错误类型（如反讽、依赖外部知识等），为模型改进提供方向。

更大规模预训练模型：如果计算资源允许，可进一步微调 RoBERTa 或 DeBERTa 等更强模型，预期准确率更高。