

基于机器学习挖掘天猫评论情感倾向

摘要

随着电子商务的快速发展，用户评论成为衡量商品质量和服务水平的重要依据。传统的好评率指标无法深入反映用户的真实情感，因此本文基于机器学习技术对天猫平台的评论数据进行了情感分析。我们首先通过 Jieba 分词和 TF-IDF 特征提取对评论文本进行了特征向量提取，并结合网格调参分别使用了逻辑回归、决策树、随机森林和深度学习神经网络等模型对情感倾向进行了分类。我们发现逻辑回归和全连接神经网络表现出较高的分类准确率。之后，为了充分利用数据集信息以及进一步提升模型性能，我们采用了分类别训练的改进方案，即分别针对 10 个商品类别训练和推测。经过实验发现，改进使得逻辑回归和全连接神经网络的准确率进一步提高。

关键词：情感分析，机器学习，逻辑回归，随机森林，深度学习，分词，TF-IDF

1 引言

随着电子商务的快速发展，用户对商品的评价变得越来越重要[1]。对于平台电商如天猫、京东等来说，用户评论不仅仅是产品质量的直接反馈，更是平台改进用户体验的重要依据。然而，传统的好评率指标过于粗糙，无法深入反映用户的真实感受。因此，对评论数据的情感倾向进行深入挖掘，可以帮助商家更清晰地了解用户的需求和不满之处，从而针对性地改进商品和服务。

我们基于机器学习技术，针对天猫平台上的评论数据进行了情感倾向的分析。我们首先收集了部分商品的评论信息，并对其进行了情感标签的标注。然后，通过使用 Jieba 分词工具将评论数据转换为文档向量，最终使用机器学习方法构建了一个情感倾向预测模型。该模型不仅能够帮助商家更精准地识别用户的满意度，还能进一步挖掘评论数据中的隐藏信息，为提升用户体验提供有力支持。

2 相关工作

情感分析，也被称为意见挖掘，是自然语言处理（NLP）的一个重要分支，旨在识别、提取和量化文本中的主观信息[2]。电商平台的普及使得用户在购物后的评论已经成为商品质量和服务水平的关键反馈渠道。评论文本中蕴含了丰富的用户情感信息，而如何有效地挖掘这些信息成为了当前研究的热点之一。在天猫、京东等大型电商平台上，情感分析技术已经被广泛应用于满意度调研和用户体验优化。

早期的情感分析方法主要依赖基于词典的技术，这种方法通常通过预先构建的情感词典来对文本中的情感词进行匹配和打分[3]。虽然这种方法简单直观，并且在处理小规模数据时效果良好，但它存在几个显著的缺陷，例如，由于词典方法依赖于预先定义的情感词表，对于新出现的词语、俚语或者特定领域的术语无法有效处理，导致情感识别的准确性降低。此外它还无法充分利用大规模数据，因为词典方法的本质是规则匹配，难以从大规模数据中学习到深层次的情感模式和上下文关联，导致在处理

复杂的情感表达时显得力不从心。

随着数据规模的扩大和文本内容的多样化，基于机器学习的情感分析方法逐渐成为主流[4]。这些方法不再依赖于固定的词典，而是通过从数据中学习特征和模式，构建分类模型，自动化地对用户评论进行情感倾向预测。与传统方法相比，基于机器学习的情感分析自适应能力强，因为机器学习模型可以通过训练数据不断更新和优化，对于新词汇、俚语和领域特定的术语具有更好的适应能力。并且，机器学习还能够通过学习数据中的上下文信息，识别出复杂的情感模式，从而提高分析的准确性[5]。

目前，情感分析领域广泛采用的机器学习方法包括逻辑回归、随机森林、决策树、支持向量机等。逻辑回归作为一种经典的线性模型，具有较好的解释性和较快的训练速度，适用于大规模的文本数据处理。随机森林和决策树模型则通过集成学习的方法，提高了模型的鲁棒性和预测精度。神经网络模型，尤其是深度神经网络，通过多层非线性变换，能够捕捉到文本中的深层次情感特征。

为了进一步提升模型的性能，我们还进行了网格搜索优化，通过调整超参数来寻找最佳模型配置。这些优化使得我们的模型在准确性和鲁棒性方面都有了显著的提升，能够更好地帮助商家理解用户反馈，为商品和服务的改进提供科学依据。

3 实验过程

3.1 实验数据集描述

实验中使用的数据集汇集了用户在天猫平台上对多种商品类别的评论信息，是进行情感分析建模的重要数据来源。数据集中包含了 62,763 条记录，覆盖了 10 个商品种类和用户反馈信息。

在这些数据中，每条记录都包括三个关键字段。其中，*cat* 字段表示商品的类别。数据集涉及的商品类别有十种，分别是书籍、平板、手机、水果、洗发水、热水器、蒙牛、衣服、计算机和酒店。其次，*label* 字段是需要学习的对象，用于标注用户评论的情感倾向。这一字段采用了二分类标注，1 表示正面评论，0 表示负面评论。最后，*review* 字段包含了用户对商品的详细文本评论，将作为实验的特征向量。这些评论反映了用户在购买和使用商品后的真实感受，是情感分析的核心数据来源。

需要注意的是，在整个数据集中，虽然 *label* 标签无一缺失，但在 *review* 字段中有一条记录存在缺失，并且其标签是 1。这条数据我们在数据预处理阶段并不会删除，因为我们认为，如果用户对于产品或服务不满意，应该会对产品进行负面评价，而空评价我们将默认为用户是比较满意的。

在数据集的划分上，我们按照 7:1:2 的比例将数据集随机分为训练集、验证集和测试集。其中训练集包含 45,189 条记录，验证集 5,021 条，测试集 12,553 条。数据集的正负样本总体较为均衡，正样本有 31,721 条，负样本有 31,042 条，并且我们确保在训练集、验证集和测试集中的正负样本比例也保持了平衡性。

3.2 数据预处理

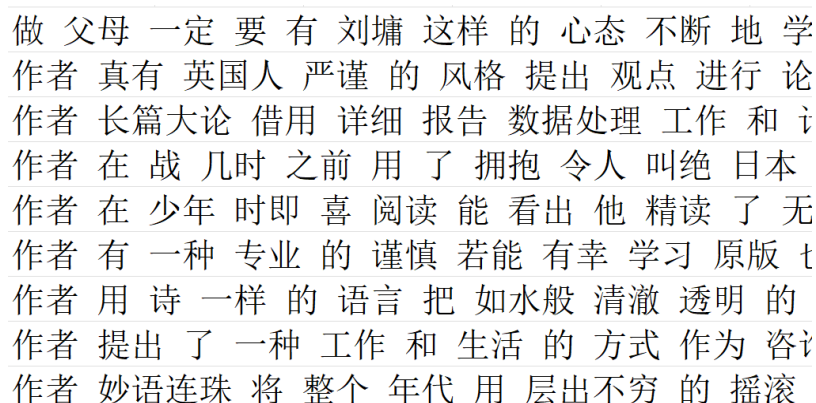
数据预处理阶段，我们的主要目的是对原始数据进行清洗、分词、特征工程等操作，最终为模型训练提供高质量的输入数据。

3.2.1 数据清洗

在进行分词之前，我们首先使用了正则表达式对评论文本进行了初步的过滤。这一步的目的是去除评论中可能包含的非中文字符，如数字、标点符号和英文单词等。这些字符在情感分析中往往不具备实际意义，反而可能增加数据的噪声，因此有必要在处理前将它们移除。

在文本数据的清洗阶段，我们采用了 Jieba 分词工具，这是一款广泛应用于中文自然语言处理的分词库。分词就是将连续的文本字符串拆分为独立的词语或短语的过程，这是由于中文和英文的语言区别导致的，因为中文不像英文那样一个单词就是一个单元，中文往往几个字组合在一起才是一个单元，所以直接对整段文字进行分析往往会导致模型无法准确理解文本的含义。分词是中文处理中的一个关键步骤。通过分词，我们可以将一个句子拆解成独立的词汇单元，从而为后续的特征提取和模型训练提供基础。分词的效果见图 1。

完成初步的字符过滤后，我们进入了分词的关键步骤。Jieba 分词工具通过一种基于词频统计和动态规划的方法，将连续的中文字符序列切分成有意义的词语。在这一过程中，Jieba 会依据内置的词典和上下文信息，判断出最合理的切分方式。例如，将“我喜欢自然语言处理”拆分为“我/喜欢/自然语言处理”，从而为每个词汇分配独立的意义。



做 父 母 一 定 要 有 刘 墉 这 样 的 心 态 不 断 地 学
作 者 真 有 英 国 人 严 谨 的 风 格 提 出 观 点 进 行 论
作 者 长 篇 大 论 借 用 详 细 报 告 数 据 处 理 工 作 和 作
作 者 在 战 几 时 之 前 用 了 拥 抱 令 人 叫 绝 日 本
作 者 在 少 年 时 即 喜 阅 读 能 看 出 他 精 读 了 无
作 者 有 一 种 专 业 的 谨 慎 若 能 有 幸 学 习 原 版 作
作 者 用 诗 一 样 的 语 言 把 如 水 般 清 澈 透 明 的
作 者 提 出 了 一 种 工 作 和 生 活 的 方 式 作 为 咨
作 者 妙 语 连 珠 将 整 个 年 代 用 层 出 不 穷 的 摇 滚

图 1. 句子的分词

然而，文本中并不是所有的词汇对于情感分析都具有重要意义。某些词汇虽然频繁出现，但并不携带实际的情感信息，例如“的”、“了”、“和”等。这些词汇被称为停用词，它们的存在可能会干扰模型的判断，降低分析的准确性。因此，在分词的基础上，我们进一步过滤掉了这些停用词。为了确保停用词过滤的有效性，我们选择了百度公司提供的停用词表。这一停用词表经过长期优化，列出了大量在实际应用中无实际意义的常见词汇，通过使用这一词表，我们能够更精确地保留对情感分析有用的词汇，减少噪声数据的影响。

做 父母 刘墉 心态 学习 补充 新鲜血液 一颗 年轻
作者 真有 英国人 严谨 风格 提出 观点 论述 论证
作者 长篇大论 借用 详细 报告 数据处理 工作 计算
作者 战 拥抱 令人 叫绝 日本 战败 会 美军 占领 没
作者 少年 时即 喜 阅读 精读 无数 经典 一个 庞大
作者 一种 专业 谨慎 若能 有幸 学习 原版 也许 会
作者 诗 语言 如水般 清澈 透明 思想 娓娓道来 一个
作者 提出 一种 工作 生活 方式 咨询 界 元老 提出
作者 妙语连珠 年代 层出不穷 摇滚 巨星 故事 紧紧

图 2. 分词、停用词过滤后的词语列表

经过分词和停用词过滤后，我们得到了一个干净且有意义的词语列表，如图 2。这一列表中的词语已经被优化为适合情感分析的形式，为后续的特征提取和模型训练提供了高质量的输入数据。这一过程不仅提高了文本处理的效率，还极大地增强了模型在分析中文文本时的准确性和鲁棒性。

3.2.2 特征工程

接下来，我们使用了 $TF-IDF$ 技术将清洗后的文本数据转换为数值特征向量。 $TF-IDF$ 是一种广泛应用的文本特征提取方法，它通过计算词语在文档中的相对重要性，将文本数据转换为可供机器学习模型处理的数值格式。在特征提取过程中，我们设置了 max_df 和 min_df 参数，用于控制词语在文档集中的出现频率。具体来说， max_df 用于过滤掉在大多数文档中都出现的词语，而 min_df 则用于过滤掉在极少数文档中出现的低频词语。这样处理的结果是，我们获得了一组稀疏的 $TF-IDF$ 特征向量，其中的值表示每个词语在特定评论中的重要性。

在完成特征向量的生成后，我们将其与对应的情感标签合并，生成了一个包含特征向量和标签的完整数据集。这一数据集将作为后续模型训练和评估的输入，帮助我们构建和优化情感分析模型。

3.3 情感分析模型的构建

在本次实验中，我们构建了多个情感分析模型来对天猫评论数据进行分类。模型的选择包括了传统的逻辑回归模型、决策树模型、随机森林模型以及全连接神经网络。我们将通过实验得出最适合情感分析任务的最佳模型。

3.1 逻辑回归

逻辑回归是一种广泛应用于分类任务的线性模型，尽管其名称中含有“回归”一词，但实际上它用于解决二分类问题。逻辑回归的基本思想是通过特征的线性组合来估计事件发生的概率，然后将该概率映射到类别标签上。其核心是使用逻辑函数

（Sigmoid 函数）将线性组合的结果映射到 0 到 1 之间，表示某一事件属于某个类别的概率。

逻辑回归模型的公式为：

$$p(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}}$$

其中, $(\beta_0, \beta_1, \dots, \beta_n)$ 是模型的参数, (x_1, x_2, \dots, x_n) 是输入的特征。输出的概率值通过设置阈值 (通常为 0.5) 来判断该样本属于正类 ($(y = 1)$) 还是负类 ($(y = 0)$)。

在逻辑回归中, 通过极大似然估计来拟合模型参数。模型的目标是最大化以下对数似然函数:

$$\mathcal{L}(\beta) = \sum_{i=1}^m [y_i \log(p(y_i = 1|x_i)) + (1 - y_i) \log(1 - p(y_i = 1|x_i))]$$

这里, $\mathcal{L}(\beta)$ 表示对数似然函数, y_i 是样本的实际标签, $p(y_i = 1|x_i)$ 是预测概率。通过最大化该函数, 模型能够找到最优的参数。

逻辑回归的一个常见变体是添加正则化项, 以防止模型过拟合。正则化可以限制模型参数的大小, 确保模型不会对训练数据过于拟合。常用的正则化类型包括 *L1* 正则化和 *L2* 正则化。*L1* 正则化鼓励稀疏解, 使得一些参数的值为 0, 而 *L2* 正则化则倾向于缩小参数的整体规模。

实验中, 我们通过网格搜索方法调整模型的超参数, 包括正则化的类型 (*L1* 正则化和 *L2* 正则化) 以及正则化强度 (参数 *C* 的值)。*L1* 正则化可以使某些特征的权重归零, 从而自动进行特征选择, 而 *L2* 正则化则更倾向于平滑权重, 避免过拟合。表格 1 是网格搜索的超参数空间。

表格 1. 三种模型网格调参的超参数

模型		参数	取值集合
逻辑回归		<i>penalty</i> <i>C</i>	[<i>L1</i> , <i>L2</i>] [0.1, 0.3, 0.5, 0.7, 0.9, 1.0]
决策树		<i>max_depth</i>	[1, 11, 21, ..., 91]
随机森林	随机搜索	<i>n_estimators</i>	[1, 51, 101, ..., 451]
		<i>max_depth</i>	[1, 51, 101, 151]
	网格搜索	<i>n_estimators</i>	[200, 230, 260]
		<i>max_depth</i>	[100, 130, 160, 190]

3.2 决策树

决策树模型是一种基于树结构的分类算法, 其基本思想是通过对数据的特征进行逐步划分, 最终将数据分配到不同的类别中。决策树的每个节点表示一个特征的测试, 分支代表测试结果, 叶节点对应分类标签。通过不断地对特征进行选择和划分, 决策树能够形成一个明确的决策路径, 从根节点到叶节点的路径即为该模型对某个样本的分类决策过程。

由于我们使用 *sklearn* 机器学习库, 因此决策树模型的划分规则默认使用了基尼系数来表示。基尼系数是衡量数据集不纯度的一种指标, 值越低表示数据集越纯净。决策树在每个节点上选择使基尼系数减少最多的特征进行划分, 这样能够逐步最大化样

本的纯度，即每个叶节点尽可能包含属于同一类别的样本。

基尼系数的公式如下：

$$\text{基尼系数}(S) = 1 - \sum_{i=1}^n p_i^2$$

其中， S 是数据集， p_i 是第 i 类样本在数据集 S 中的比例， n 是类别的总数。基尼系数用于衡量某一特征划分后的数据纯度。通过选择基尼系数最小的特征进行划分，模型能够最有效地减少分类的不确定性，最终形成分类的决策路径。

在构建决策树模型的过程中，一个常见的问题是过拟合，即模型在训练数据上表现得非常好，但在新数据上的表现却很差。为了解决这个问题，可以通过控制树的最大深度或设置最小样本数来停止树的生长，防止模型过于复杂。

因此在本实验中，我们使用网格搜索对 `max_depth` 进行了最优参数寻找，具体的搜索的范围是从 1 到 100 的整数步进值。通过交叉验证的方式，我们评估了不同深度下决策树模型的性能，并选择了最能平衡准确性和泛化能力的参数组合。表格 1 展示了决策树模型的超参数搜索空间。

3.3 随机森林

随机森林模型是一种基于集成学习的算法，它通过构建多个决策树模型并综合它们的预测结果来提高分类的准确性和稳定性。随机森林的基本思想是利用“森林”中多个不同的“树”进行投票表决，从而提高模型的泛化能力和抗噪性。相比于单一决策树，随机森林能够有效地降低过拟合的风险，并在处理高维数据时表现出色。

随机森林的每棵树都是在数据的不同子集上训练得到的，这些子集是通过自助法从原始数据中随机抽取的。在每棵树的节点分裂过程中，随机森林会从所有特征中随机选择一个子集，并在该子集上选择最佳分裂特征。这种随机化的特性使得随机森林能够降低决策树之间的相关性，从而进一步提升模型的稳定性和预测性能。

随机森林模型的公式为：

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x)$$

其中， T 表示森林中决策树的数量， $f_t(x)$ 是第 t 棵决策树对输入 x 的预测，最终的预测结果是所有决策树预测的平均或多数投票结果。

在本实验中，由于随机森林涉及的参数较多且训练时间较长，因此我们首先通过随机网格搜索来缩小参数搜索空间，随后使用精细的网格搜索来确定最优参数组合。

在实验的开始，我们进行了随机搜索，与网格搜索相比，随机搜索并未尝试所有参数值，而是从指定的分布中采样固定数量的参数设置。我们对两个主要参数进行了调优：树的数量 `n_estimators` 和树的最大深度 `max_depth`。这样做的目的是通过较少的计算资源快速找到一个较好的参数组合。

在随机搜索得到初步的最优参数后，我们进行了更精细的网格搜索。在这一阶段，我们进一步缩小参数范围，并针对较小的参数空间进行详尽的搜索，以找到最优的树数量和最大深度。表格 1 展示了随机森林模型的超参数搜索空间。

3.4 全连接神经网络

除了三种常用的机器学习模型，我们还使用 Pytorch 搭建了一个简单的全连接神经网络来学习情感分析。全连接神经网络的优势在于其灵活性和强大的特征学习能力，特别是在处理非线性数据时表现出色。通过多个隐藏层的设计，网络能够捕捉到文本数据中潜在的复杂情感关系，从而在情感分析任务中表现出良好的效果。

全连接神经网络的基本结构包括输入层、多个隐藏层以及输出层。输入层接收经过预处理的文本特征，隐藏层则通过线性变换和激活函数将输入映射到更高维度的特征空间，以便提取出更有用的特征。输出层通常使用 *Sigmoid* 激活函数，将最终输出的数值映射为二分类任务中的概率值。

模型的公式可以表示为：

$$y = \sigma(W_n(\dots \sigma(W_2 \sigma(W_1 x + b_1) + b_2) \dots) + b_n)$$

其中， (W_1, W_2, \dots, W_n) 是权重矩阵， (b_1, b_2, \dots, b_n) 是偏置， σ 是激活函数（通常为 *ReLU* 或 *Sigmoid*）。通过反向传播算法，网络能够自动调整权重和偏置以最小化损失函数。

在本实验中，我们搭建的多层全连接神经网络包括四个全连接层，每层之后都包含一个 *Dropout* 层实现随机失活，用于防止过拟合。通过将输入数据逐层传递，网络能够逐步提取更抽象的特征，从而提高分类的准确性。

模型训练进行了 300 个周期，我们使用了二分类交叉熵损失函数和 Adam 优化器。学习率设置为 $1e-5$ ，训练过程中我们每经过一个周期都会在验证集上进行测试，并根据验证集的准确率保存最优模型。

3.4 模型评价标准

在本任务中，我们使用了几种常用的评价标准：混淆矩阵、准确率、ROC 曲线和 AUC。

3.4.1 混淆矩阵和准确率

混淆矩阵是二分类或多分类问题中常用的评价指标，它通过记录模型的预测结果与真实结果的对比，直观展示模型的分类表现。对于二分类问题，混淆矩阵是一个 2×2 的矩阵，表示四种可能的分类结果：

$$\begin{bmatrix} \text{True Positive (TP)} & \text{False Positive (FP)} \\ \text{False Negative (FN)} & \text{True Negative (TN)} \end{bmatrix}$$

其中，*TP* 为模型正确预测为正例的样本数量，*FP* 为模型错误地预测为正例的负类样本数量，*FN* 为模型错误地预测为负例的正类样本数量，*TN* 为模型正确预测为负例的样本数量。

混淆矩阵提供了分类错误的具体分布情况，能够帮助我们更好地理解模型的不足之处，例如模型是否偏向于某一类或对某类样本的误分类情况较为严重。

在混淆矩阵下，准确率是最常用的分类评价标准之一，它表示模型预测正确的样本占总样本的比例。准确率的公式为：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

准确率直观反映了模型在所有样本中预测正确的比例。虽然准确率是一个重要的评价标准，但在类别不平衡的情况下，单独依赖准确率可能会导致误导，但本任务中正负样本均衡，因此准确率能够很好地反应模型的精度。

3.4.2 ROC 曲线和 AUC

ROC 曲线是衡量分类模型性能的另一重要工具，特别适合用来评价二分类问题中的模型表现。ROC 曲线展示了模型的假正率（ FPR ）与真正率（ TPR ）之间的关系。 TPR 与 FPR 的公式如下：

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

ROC 曲线通过在不同阈值下计算 TPR 和 FPR 来绘制曲线。曲线的左上角（ TPR 接近 1， FPR 接近 0）代表理想的分类器，即正确分类的能力强、错误分类的概率小。

AUC 是 ROC 曲线下的面积，用来量化模型的整体性能。AUC 的取值范围为 0 到 1，值越接近 1，说明模型的区分能力越强；而 AUC 为 0.5 则表示模型的分类能力相当于随机猜测。AUC 的优点在于它不受类别不平衡的影响，因此即使在类别不均衡的数据集中，AUC 仍能提供较为可靠的模型性能评价。

4 实验结果

在本次实验中，我们对逻辑回归、决策树、随机森林以及全连接神经网络四种模型进行了调参和实验评估。通过对每个模型进行网格搜索或随机搜索，我们找到了各自的最优参数，并使用这些参数对模型进行了最终评估。

首先，在所有模型的超参数调优过程中，我们主要以准确率作为模型性能的评价指标。通过对不同的参数组合进行搜索，我们得到了每个模型在验证集上的最优超参数组合。表格 2 汇总了各模型的最优超参数配置以及对应的准确率表现。

逻辑回归作为一种线性模型，在处理大规模文本分类任务时展现了出色的性能。通过网格搜索，我们发现使用 L2 正则化且正则化强度参数 $C = 1.0$ 时，逻辑回归模型达到了最佳表现。最终，逻辑回归模型在测试集上的准确率达到 87.89%，为四种模型中最优模型。这一结果表明，尽管逻辑回归是一个相对简单的模型，但在情感分析任务中仍然能够很好地捕捉文本中的关键信息，从而实现高效的分类。

表格 2. 以准确率为指标下三种模型网格调参的最优超参数

模型		参数	最优参数值
逻辑回归		<i>penalty</i>	<i>L2</i>
		<i>C</i>	1.0
决策树		<i>max_depth</i>	91
随机森林	随机搜索	<i>n_estimators</i>	251
		<i>max_depth</i>	151
	网格搜索	<i>n_estimators</i>	190
		<i>max_depth</i>	230

表格 3. 三种模型最优超参数下的准确率

模型	最优参数下的准确率（%）
决策树	80.66
随机森林	85.84
逻辑回归	87.89
全连接神经网络	87.65

决策树模型在本次实验中表现相对较差。通过网格搜索，我们找到了最优的树深度参数 *max_depth* = 91 。在这一参数配置下，决策树模型在测试集上的准确率仅为 80.66% 。

随机森林模型作为一种集成学习方法，通过结合多棵决策树的预测结果，显著提高了分类的准确性。在随机搜索和精细网格搜索的帮助下，我们最终确定了最优参数组合：在随机搜索阶段，树的数量 *n_estimators* 为 251，最大树深度 *max_depth* 为 151；而在网格搜索阶段，树的数量为 190，最大树深度为 230。最后在这一最优参数配置下，随机森林模型在测试集上达到了 85.84% 的准确率。

全连接神经网络模型由于其强大的特征学习能力，在处理复杂数据时展现了优异的性能。在本次实验中，我们仅仅设计了一个包含多层全连接层的神经网络，通过使用交叉熵损失函数和 Adam 优化器进行训练。在训练中，网络训练迅速，损失下降，最终神经网络模型在测试集上实现了87.65%的准确率，这一水平明显高于决策树和随机森林。值得一提的是，这个神经网络并没有参与各个参数的调整，其训练时间远远低于其他几个模型，并且准确率较高。

综上所述，通过本次实验的结果可以看出，逻辑回归和全连接神经网络在准确率上占据了优势，分别达到了87.89%和87.65%。

5 模型的改进

在本次情感分析任务中，我们使用了一个通用的模型对所有类别的商品评论进行情感倾向的预测。然而，这种方法并不完全合理，因为在训练过程中，我们忽略了商品类别这一关键信息。事实上，商品类别本身可能包含重要的信息，能够帮助模型更好地区分评论中的情感倾向。例如，不同商品类别的用户使用不同的表达方式或关注不同的产品特性，因此直接将所有类别的评论放入同一个模型中，可能会导致模型无法有效捕捉每个类别中的特定情感模式。

为了改进模型性能，我们考虑了以下方法。由于数据集中的商品类别较少（共 10 种），我们可以选择针对每个类别单独训练一个模型。具体来说，我们为每种商品类别训练并保存一个独立的模型权重。在测试时，根据评论所属的商品类别，选择相应的模型进行推断。这种方式能够更好地利用商品类别信息，避免类别间差异对模型的干扰。

在改进方案中，我们通过在 `main` 函数中调用训练模块，分别为每个类别的数据集训练模型并保存相应的最优权重。测试时，根据类别标签调用相应的模型权重进行预测。该方案虽然在逻辑上与原先的单一模型存在较大区别，但能够更有效地提升分类准确率，特别是在不同类别的评论风格存在较大差异的情况下。

此外，虽然类别较少时可以单独训练多个模型，但如果类别数量较多，这种方法可能不再适用。此时，可以考虑将类别标签作为特征加入输入数据中，模型将能够根据类别自动学习不同类别之间的情感特征差异。这种方法能够在训练时保留类别信息，并避免训练过多的模型。

由于之前的实验中我们可以看到逻辑回归和神经网络的效果明显更好，因此在这里我们分别使用了逻辑回归和神经网络进行优化。在实验中，我们针对 10 个商品类别进行了分类别的训练和测试。每个类别分别训练并保存了最优模型，测试时根据类别选择相应模型进行推断。优化前后的模型精度指标见表格 4，ROC 曲线分别见图 3，图 4。

表格 3. 三种模型最优超参数下的准确率

模型	准确率（%）	AUC	F1
逻辑回归改进前	88.44	0.9512	0.8844
逻辑回归改进后	89.01	0.9551	0.8904
神经网络改进前	87.76	0.9471	0.8771
神经网络改进后	89.34	0.9529	0.8918

通过这一改进，模型在不同商品类别上的表现更加精准，进一步提升了整体准确率。虽然分类别训练增加了计算成本，但其在准确性上的提升证明了该改进的有效性。

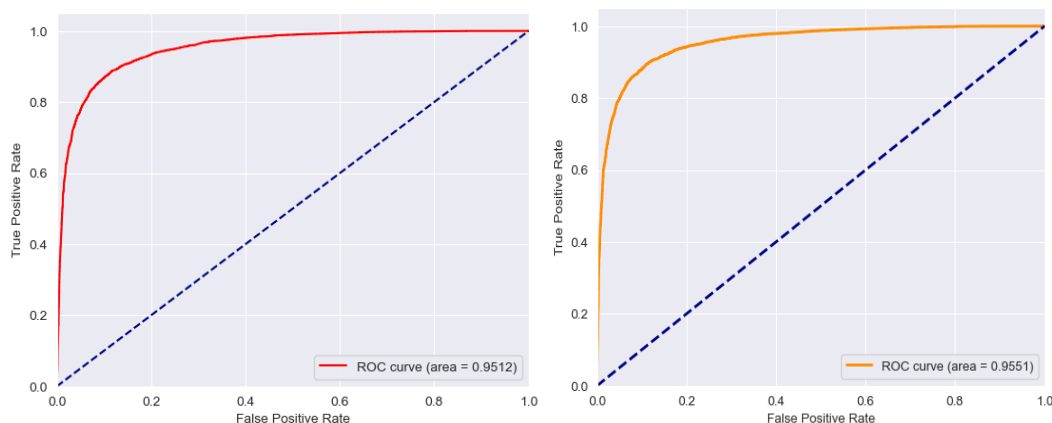


图 3. 逻辑回归优化前后 ROC 曲线对比

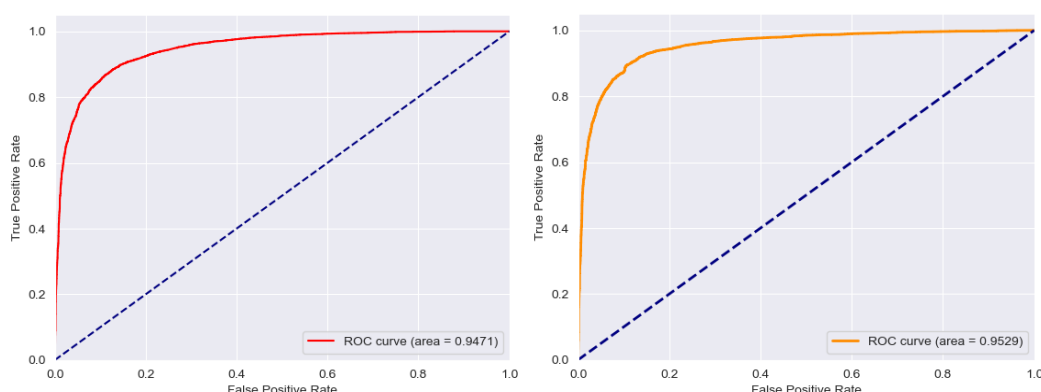


图 4. 神经网络优化前后 ROC 曲线对比

6 结语

在本次情感分析任务中，我们首先使用了停用词表和 Jieba 工具对评论文本进行了过滤和分词，并通过 $TF-IDF$ 方法将文本转换为数值特征，作为模型的输入。

之后，我们系统地构建并评估了四种经典的机器学习与深度学习模型，包括逻辑回归、决策树、随机森林以及全连接神经网络。在此基础上，我们对各个模型进行了训练和调参，最终得到了每种模型的最优超参数配置，并在测试集上进行了全面的评估。实验结果显示，逻辑回归和全连接神经网络在分类准确率上表现尤为出色，验证了这些模型在情感分类任务中的有效性。

在模型的改进方面，我们发现将所有商品类别统一使用一个模型进行预测并不完全合理。商品类别本身包含了重要的信息，这些信息可以帮助模型更好地区分评论中的情感倾向。因此，我们引入了分类别训练的改进方案。通过为每个商品类别分别训练模型，并在推断时依据类别调用对应的模型权重，改进后的方案有效提升了模型的准确性。这种方式充分利用了类别信息，避免了类别间的差异对模型预测结果的影响。

然而，分类别训练虽然提高了模型的准确性，但也带来了计算开销的增加。在实际应用中，如果商品类别数量较多，将类别信息作为特征输入模型是一种更加高效的解决方案。模型可以通过学习不同类别的特征差异来自动适应不同类别的评论，从而在不增加模型数量的情况下实现更高效的推理。

通过本次实验，我们不仅探索了多种模型在情感分析任务中的表现，还通过引入类别信息改进了模型的效果。随着数据规模和复杂度的增加，未来的情感分析任务中可以继续探索更多复杂的模型架构与特征工程技术，以进一步提升模型的性能和应用价值。

参考文献

- [1] Lackermair, G., Kailer, D., & Kanmaz, K. (2013). Importance of online product reviews from a consumer's perspective. *Advances in economics and business*, 1(1), 1-5.
- [2] Saberi, B., & Saad, S. (2017). Sentiment analysis or opinion mining: A review. *Int. J. Adv. Sci. Eng. Inf. Technol*, 7(5), 1660-1666.
- [3] Hardeniya, T., & Borikar, D. A. (2016). Dictionary based approach to sentiment analysis-a review. *International Journal of Advanced Engineering, Management and Science*, 2(5), 239438.
- [4] Ahmad, M., Aftab, S., Muhammad, S. S., & Ahmad, S. (2017). Machine learning techniques for sentiment analysis: A review. *Int. J. Multidiscip. Sci. Eng*, 8(3), 27.
- [5] Nguyen, H., Veluchamy, A., Diop, M., & Iqbal, R. (2018). Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches. *SMU Data Science Review*, 1(4), 7.