Statistics and Application 统计学与应用, 2025, 14(4), 34-46
Published Online April 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.144086

Hans汉斯

基于RoBERTa的新闻评论可解释性情感分析

张月月[1]，刘立佳[2]

[1]燕山大学理学院，河北 秦皇岛
[2]燕山大学信息科学与工程学院(软件学院)，河北 秦皇岛

收稿日期：2025年3月2日；录用日期：2025年3月21日；发布日期：2025年4月2日

摘 要

在数字化信息高速传播的当下，实现对新闻评论精准且具备可解释性的情感分析，对舆情洞察与舆论引导极为关键。本研究运用RoBERTa模型，结合多尺度卷积神经网络(MSCNN)，深度剖析新闻评论中的情感倾向。同时，采用局部可解释的模型无关解释方法(LIME)，实现对模型预测结果的深度解释，可视化展示模型在处理新闻评论时对词汇和短语的关注重点，为模型决策提供清晰依据。实验结果表明，RoBERTa-MSCNN在新闻评论情感分析任务上取得了更优的性能，准确率达到83.34%，精确率为82.6%，召回率为84.67%，F1值提升至83.62%。同时，可解释性分析为用户理解模型输出提供了清晰的视角，有助于新闻媒体更有效地进行舆情监测与引导，为相关领域的研究与应用提供了有力支持。

关键词

RoBERTa，卷积神经网络，可解释性，情感分析

Interpretable Sentiment Analysis of News Commentary Based on RoBERTa

Yueyue Zhang[1], Lijia Liu[2]

[1]School of Science, Yanshan University, Qinhuangdao Hebei
[2]School of Information Science and Engineering (School of Software), Yanshan University, Qinhuangdao Hebei

Received: Mar. 2nd, 2025; accepted: Mar. 21st, 2025; published: Apr. 2nd, 2025

Abstract

In the current era of high-speed digital information dissemination, accurate and interpretable emotional analysis of news comments is crucial for public opinion insight and guidance. In this study, RoBERTa model and multi-scale convolutional neural network (MSCNN) are used to analyze the emotional tendency of news commentary. At the same time, Local Interpretable Model-agnostic

文章引用: 张月月, 刘立佳. 基于 RoBERTa 的新闻评论可解释性情感分析[J]. 统计学与应用, 2025, 14(4): 34-46.
DOI: 10.12677/sa.2025.144086

Explanations (LIME) is used to realize the in-depth interpretation of the model prediction results and visually display the model's focus on words and phrases when processing news comments, providing a clear basis for the model's decision-making. The experimental results show that RoBERTa-MSCNN has achieved superior performance in the task of sentiment analysis of news comments. Its accuracy rate reaches 83.34%, the precision rate is 82.6%, the recall rate is 84.67%, and the F1 score has been increased to 83.62%. At the same time, interpretability analysis provides a clear perspective for users to understand the model output, helps news media to monitor and guide public opinion more effectively, and provides strong support for research and application in related fields.

Keywords

RoBERTa, Convolutional Neural Network, Interpretability, Sentiment Analysis

1. 引言

随着互联网的普及，新闻评论成为公众表达观点的重要渠道，其情感分析对舆情监测和舆论引导具有重要意义。然而，新闻评论语言多样、情感复杂，传统方法难以准确捕捉其深层语义和情感倾向。尽管预训练语言模型(如 RoBERTa)在自然语言处理任务中表现出色，但其对局部情感特征的提取能力有限，且模型可解释性较差。

为此，本文提出一种基于 RoBERTa-MSCNN 的新闻评论可解释性情感分析方法。该方法结合 RoBERTa 的语义理解能力和 MSCNN 的特征提取能力，实现对评论情感的精准分析。同时，引入 LIME 增强模型的可解释性。LIME 通过在输入文本中生成局部扰动样本并观察模型输出的变化，识别对预测结果影响最大的关键词语或短语，从而提供直观的解释。实验表明，该方法在公开数据集上表现优异，并结合 LIME 生成了可解释的情感分析结果，为新闻评论情感分析提供了新的解决方案，具有重要的应用价值。

2. 相关研究

新闻评论情感分析是自然语言处理(NLP)领域的一个重要分支，其目标是从新闻评论中提取出用户的情感倾向。随着互联网的快速发展，新闻评论数据呈现爆炸式增长，如何高效、准确地分析这些数据成为了一个重要的研究课题。近年来，深度学习技术，特别是预训练语言模型的出现，为情感分析领域带来了新的机遇和挑战。

2.1. 情感分析方法研究现状

宁益民[1]提出传统的情感分析方法主要分为基于情感词典和基于机器学习两大类。惠调艳等[2]为深度挖掘商品显性和隐性属性特征，提出了融合词典-TextCNN-Word2Vec 的在线评价细粒度情感分析模型。Athindran 等[3]利用朴素贝叶斯分析 Twitter 和 Facebook 上的客户评论，并对客户情绪进行分析，为企业提供关于提高产品质量的见解和决策。Vanaja 和 Belwal [4]采用朴素贝叶斯与支持向量机对亚马逊客户评论进行情感分析。基于情感词典的方法通过匹配预定义的情感词来计算文本的情感倾向，但其难以处理语义复杂性和上下文依赖性[5]。基于机器学习的方法则利用特征工程和分类算法进行情感分析，但其性能严重依赖于特征提取的质量。

近年来，深度学习方法，特别是基于卷积神经网络(CNN)和循环神经网络(RNN)的模型，在情感分析任务中取得了显著成果。Kamruzzaman 等[6]比较了 6 种集成模型的情感分析性能，其中包括传统集成模型和神经网络集成模型，神经网络集成模型分别由 CNN、RNN、LSTM、GRU 组成，实验结果表明神经网络模型效果更优。Janardhana 等[7]提出一种由卷积神经网络(CNN)与循环神经网络(RNN)组成的混合模型，然后使用 GloVe 预训练模型表示文本，对电影评论数据进行情感分析，精确度较高，达到 84%。CNN 能够有效捕捉局部特征，而 RNN 则擅长处理序列依赖关系，两者在情感分析中占据重要地位。

2.2. 预训练语言模型研究现状

预训练语言模型，如 BERT、RoBERTa 等，通过在大规模语料库上进行预训练，学习到丰富的语义表示，在各种 NLP 任务中取得了突破性进展。RoBERTa 作为 BERT 的改进版本，通过调整训练策略和超参数，进一步提升了模型的性能。Tan 等[8]设计了一种融合 RoBERTa 与长短期记忆网络(LSTM)的情感分析方法，将 RoBERTa 的自关注和动态掩蔽能力与 LSTM 捕获编码文本中远程依赖关系的能力相结合，分析 Twitter 等社交评论，并显示出较高的准确率。杨明[9]提出了一种基于并行 CNN 与双向门控循环单元神经网络(BiGRU)的新闻情感分析模型 BERT-CBA。实验证明，提出的 BERT-CBA 模型比其他模型具有更高的精度。预训练语言模型的出现为新闻评论情感分析提供了新的解决方案，但其计算复杂度高、可解释性差等问题仍需解决。

2.3. 可解释性技术研究现状

随着深度学习模型的复杂性不断增加，模型的可解释性变得越来越重要。可解释性技术旨在帮助用户理解模型的决策过程，提高模型的透明度和可信度。陈谦提出目前现存的可解释性方法主要分为两大类，一类是事后可解释性方法，另一类是基于注意力机制的可解释性方法。常用的可解释性技术包括注意力机制、LIME、SHAP 等。林奕欧[10]提出基于改进预训练语言模型、注意力机制多输入、自解释生成网络的文本情感分析建模方法对模型决策行为加以语义形式的理解和解释。蒋建洪和李梦欣[11]采用轻量级梯度提升机(LGBM)对企业负面事件下微博用户极端情感进行建模，并使用 SHapley 加性解释(SHAP)对特征变量进行可视化展示，解决网络口碑负面化的问题提供理论依据。Jain R 等[12]利用情感感知词典与情感推理器(VADER)和 LIME 相结合的方法对社交媒体的评论进行情感分析，结果以热图和条形图进行可视化展示。

新闻评论情感分析是一个充满挑战的研究领域，随着深度学习技术和可解释性技术的不断发展，该领域的研究将不断深入，为舆情分析、情感计算等应用提供更加强大的技术支持。本文提出的基于 RoBERTa-MSCNN 模型和 LIME 的可解释性情感分析方法，将为新闻评论情感分析提供新的思路和方法。

3. 模型结构

本文提出基于 RoBERTa-MSCNN 的情感分析模型，通过对新闻评论文本进行情感分类，从而获取文本情感倾向。该模型采用 RoBERTa 作为预训练模型，将新闻评论文本转化为词向量，并将生成的词向量输入 MSCNN 中，从而获取深层次的信息。模型结构见图 1。

3.1. RoBERTa 模块

RoBERTa 是一种基于 Transformer 架构的预训练语言模型，其主要围绕掩码语言模型和自注意力机制展开。输入原始文本被分词为子词或单词，并转换为词嵌入。同时，添加位置编码以保留序列的顺序信息。Transformer 编码器是多层的，且每层主要由多头注意力机制和前馈神经网络(FFN)两子层组成。每个子层的输出通过残差连接和层归一化进行训练。

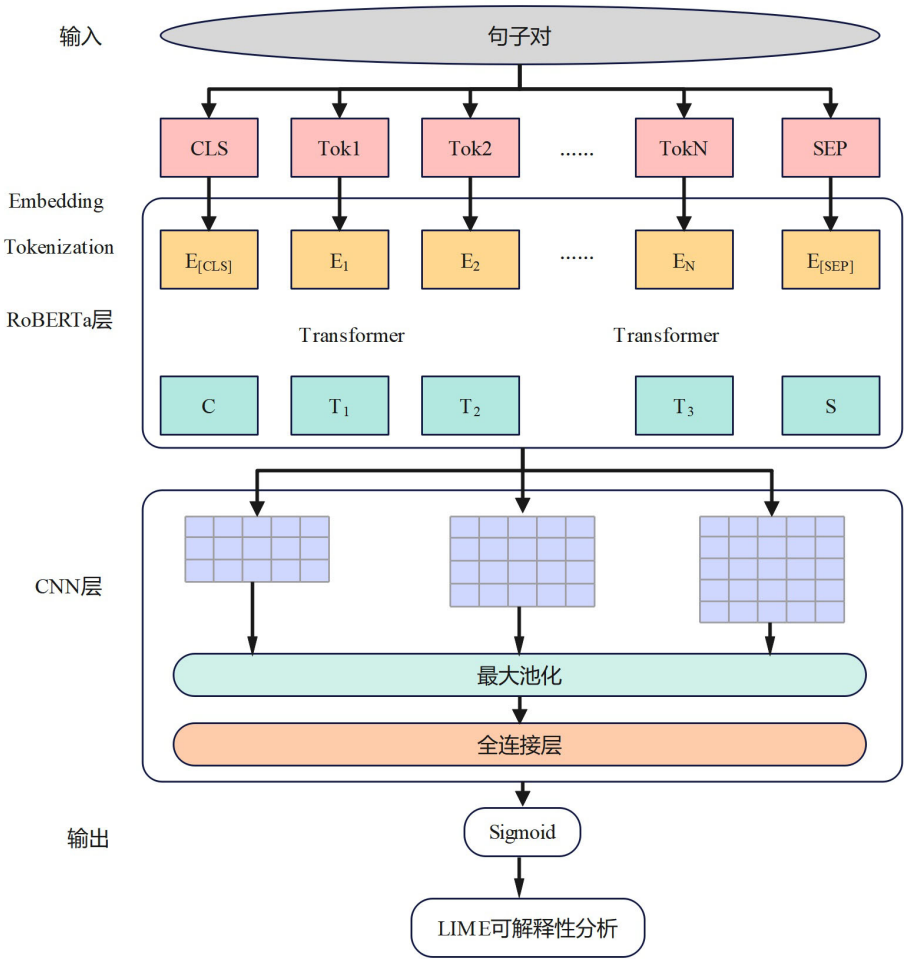

Figure 1. Interpretable sentiment analysis model structure based on RoBERTa-MSCNN
图 1. 基于 RoBERTa-MSCNN 的可解释性情感分析模型结构

掩码语言模型是通过从输入文本中随机选择一定比例的单词进行掩码处理，替换为[MASK]、随机单词，或保持不变。与 BERT 不同的是，RoBERTa 引入了动态掩码机制，即在每个训练 epoch 中为输入序列生成不同的掩码模式，使模型更具有泛化能力。同时，RoBERTa 移除了 BERT 中的下一句预测任务，专注于掩码语言模型任务，提高了模型的性能。通过多层 Transformer 编码器处理输入序列，生成每个位置的上下文表示。对于被掩码的位置，模型输出一个概率分布，预测最可能的单词。利用损失函数计算预测单词与真实单词之间的差异，反向传播优化模型参数。

自注意力机制用于捕捉输入序列中单词之间的依赖关系。首先计算注意力分数，使用 Softmax 函数将注意力分数归一化为概率分布。使用归一化的注意力分数对值向量进行加权求和，得到每个单词的上下文表示。使用多个注意力头并行计算，捕捉不同子空间的特征。将多个头的输出拼接并线性变换，得到最终表示。

3.2. MSCNN 模块

MSCNN 模型是通过使用不同大小的卷积核从输入特征中提取多尺度的信息，并行地捕捉文本中从局部到全局的多粒度情感特征，以理解情感表达的整体语义环境。由于新闻评论的文本长度是参差不齐的，MSCNN 模块能够自适应地处理不同长度的文本，同时，降低不规范文本的影响，可以提高模型鲁棒

性。由于新闻涉及多个领域，则评论可能具有不同的词汇、表达方式，MSCNN 模块可以通过调整不同尺度卷积核的参数，来适应这些差异，学习到不同文本中情感表达的独特规律，增强模型泛化能力。

MSCNN 模块包含输入层、多尺度卷积层、池化层、融合层和输出层。其结构如图 2 所示。输入层是将经过 RoBERTa 处理后的特征表示传递到卷积层。多尺度卷积层使用不同大小的卷积核捕捉不同长度的特征，本文使用了三种不同大小的卷积核，分别是$(3,768)$、$(4,768)$和$(5,768)$，且每种卷积核的数量均为 256 个。对每个卷积层的输出应用双曲正切激活函数 tanh 进行非线性变换。tanh 函数如下。

$$\tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{1}$$

对激活后的输出使用批量归一化层进行归一化处理，使模型训练更加稳定。池化层是对每个卷积层的输出经过最大池化操作。最后，融合层将三个卷积层的输出在通道维度上进行拼接，并传递到 Dropout 层。经过 Dropout 处理后的特征通过两个全连接层映射得到最终的分类结果。

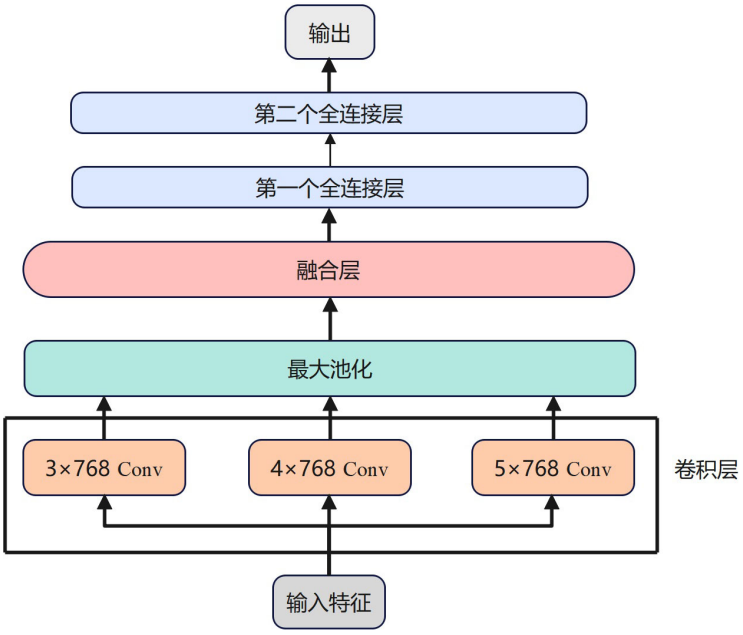

Figure 2. Multi-scale CNN structure
图 2. 多尺度 CNN 结构

3.3. LIME 模块

本文提出的 RoBERTa-MSCNN 模型是深度学习模型，属于“黑盒”模型，采用 LIME 对模型进行局部解释，可以提高模型预测的可解释性。LIME 是通过在局部用简单的可解释模型近似复杂的黑盒模型，利用加权最小二乘法和正则化技术求解可解释模型的系数，来解释模型在特定实例上的预测结果。

LIME 可以找到一个可解释的模型 g，使得在原始样本 x_0 的局部邻域内，g 能够通过最小化一个损失函数很好地近似黑盒模型 f。损失函数计算公式为：

$$\min_{g} L\left(f,g,\pi_{x_0}\right)+\Omega(g) \tag{2}$$

其中，$L\left(f,g,\pi_{x_0}\right)$ 是衡量 g 与 f 在基于权重分布 π_{x_0} 的局部样本上的差异程度的损失函数。$\Omega(g)$ 是对可解释模型 g 的复杂度惩罚项，用于避免模型过拟合。

4. 实验方法与结果分析

4.1. 实验环境

本文实验在 Windows 系统中完成，操作系统为 Windows 11，64 位，CPU 型号为 AMD Ryzen 7 7840HS。编程工具：Pycharm，编程语言：python。

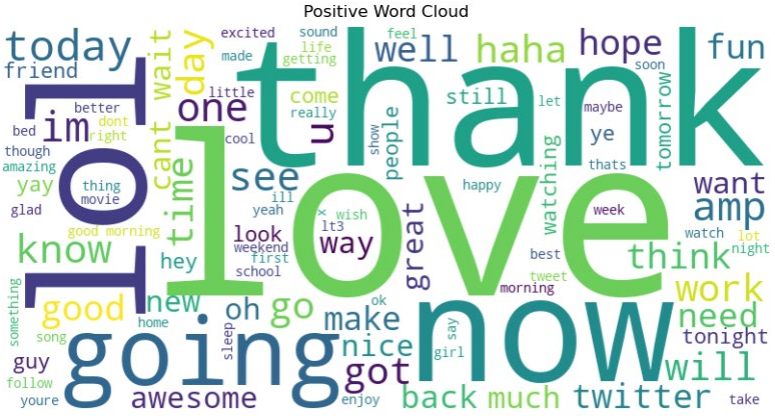

Figure 3. Positive word cloud
图 3. 积极词云

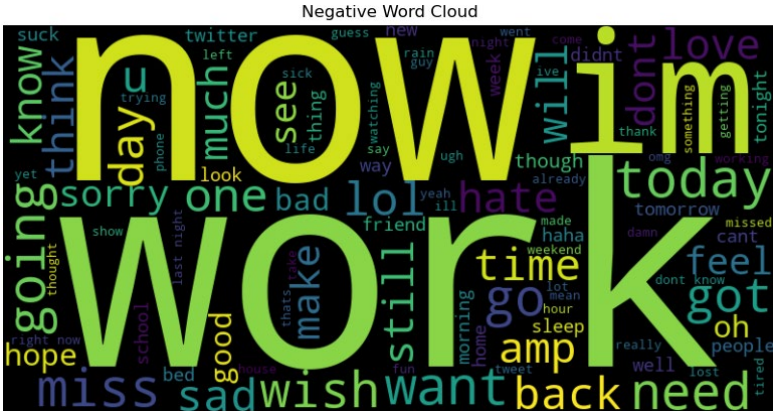

Figure 4. Negative word cloud
图 4. 消极词云

4.2. 实验数据集及数据预处理

本文研究以新闻评论文本为主体对象，采用由 Twitter API 提取的推文数据集 Sentiment140 [13]为实验数据集，推文内容涵盖了各种主题和话题，如娱乐、政治、科技、生活等，使得数据集具有较高的多样性和代表性。该数据集由 160 万条推文组成，该数据集只有两种情感极性，即消极和积极，适合做二分类的情感分类任务。如图 3 积极词云显示，该数据集中“love”、“thank”和“well”等词都能代表积极的情绪；如图 4 消极词云显示，“sorry”、“bad”和“hate”等词都能代表消极的情绪。

为让数据更满足后续的情感分析任务，要进行数据预处理。首先，清洗数据，去除数据中包含的特殊符号，如标点、表情、HTML 标签，同时去除推文中的停用词、URL 和用户名；其次，进行数据采样。由于数据集过大，为减少处理时间和计算资源损耗，进行随机采样，即随机选取 5 万条评论，实验采用 8:1:1 比例将数据划分为训练数据集、验证数据集和测试数据集。

4.3. 评价标准

本文采用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1 值 4 个指标判断模型优劣。准确率表示模型预测正确的样本占总样本的比例。精确率表示模型预测为正类的样本中，实际为正类的比例。召回率表示实际为正类的样本中，模型正确预测为正类的比例。F1 值是精确率和召回率的调和平均值，用于平衡两者。它们的计算基于混淆矩阵，混淆矩阵包含以下四个关键值：TP：模型预测为正类，实际类别也是正类的数量；FP：模型预测为正类，实际类别是负类的数量；TN：模型预测为负类，实际类别也是负类的数量；FN：模型预测为负类，实际类别是正类的数量。

各指标的计算公式分别为：

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

4.4. 实验超参数选取与优化

考虑到文本最大长度(Maxlen)设置过大会导致内存占用和训练时间加长，也为最大程度保留文本信息，本文把文本最大长度(Maxlen)设置为 128。为使训练更稳定，训练时每个批次包含的样本数量设置为 8，验证时每个批次包含的样本数量设置为 4。学习率过大，模型可能会跳过最优解，导致无法收敛；学习率过小，模型收敛速度会非常缓慢。由于本文使用的预训练模型(RoBERTa)已经在大规模数据上进行了训练，参数已经接近最优，使用较小的学习率可以避免对预训练参数进行过大的调整，从而保持模型的稳定性，所以使用 Adam 优化器，学习率设为 0.00001，权重衰减系数设为 0.00001 以使模型对参数进行自适应调整。

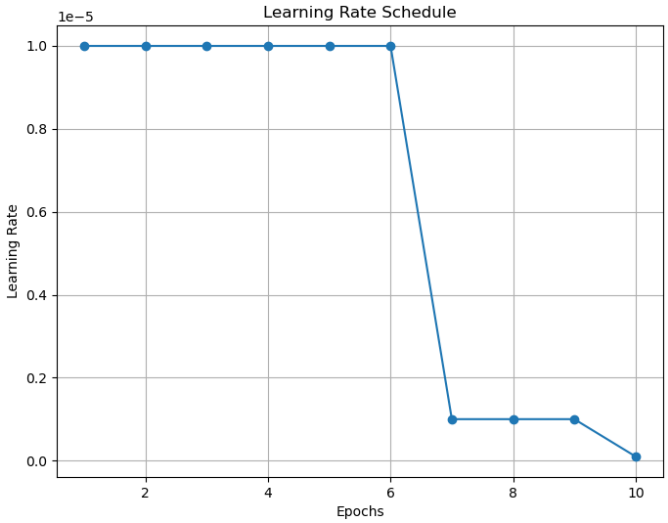

Figure5. Learning rate adjustment strategy diagram
图 5. 学习率调整策略图

图 5 展示了模型训练过程中学习率随训练轮次的变化情况。从图中可以看出，在第 1 轮到第 6 轮训练期间，学习率保持在固定值 1.0，没有发生变化。从第 6 轮到第 7 轮，学习率出现了急剧下降，从 1.0 骤降至接近 0.08。这是为了在模型初步收敛后，减小学习率以避免参数更新步长过大而错过最优解，帮助模型更精细地调整参数。在第 7 轮到第 10 轮，学习率继续缓慢下降，从接近 0.08 逐步降低到接近 0。这种缓慢下降的策略可以让模型在训练后期更加平稳地收敛，进一步优化模型参数，提高模型的泛化能力。

本文属于情感二分类问题，需要判断文本是积极情感(类别 1)还是消极情感(类别 0)。交叉熵损失可以根据模型预测的每个类别的概率与真实标签来计算损失，更好地优化模型。模型的输出通常是每个类别的得分(logits)，通过 Softmax 函数可以将这些得分转换为概率分布。交叉熵损失基于概率分布进行计算，它使模型为真实类别输出更高的概率，为其他类别输出更低的概率。

首先利用 Softmax 函数把每个类别的得分(logits)转化为概率分布 $p=\left[p_1,p_2,\cdots,p_C\right]$，其公式为：

$$p_i=\frac{e_{z_i}}{\sum_{j=1}^{C}e_{z_j}},\ i=1,2,\cdots,C \tag{7}$$

式中，z_i 是模型输出的每个类别的得分(logits)。此公式将每个类别的得分(logits)转变为概率值，保证所有概率值之和为 1。

单样本的交叉熵损失 L 的计算公式是：

$$L=-\sum_{i=1}^{C}y_i\log\left(p_i\right) \tag{8}$$

C 是类别数量，y_i 是一个独热编码向量(one-hot)，表示样本的真实类别。如果样本属于类别 i，则 $y_i=1$，否则 $y_i=0$。p_i 是模型预测的样本属于类别 i 的概率(经过 softmax 后的输出)。

经过实验，过多的训练轮数可能会导致模型过拟合，过少的训练轮数可能会导致模型欠拟合，导致模型分类性能下降，所以训练轮数设为 10。虽然卷积核数量与提取的特征量成正比，但本文的卷积核数量设置在 1024 或 512 时，会增加模型的复杂度和计算量，导致模型过拟合，所以本文卷积核数量设为 256。

为避免模型过拟合的发生，在上述模型设置的基础上引入数据增强技术。噪声注入是通过在文本中随机插入、删除或替换字符来增加数据的多样性，从而提高模型的泛化能力，减少过拟合。根据本文对数据增强的需要，噪声水平设为 0.1，即在文本中约 10%的字符会被进行噪声操作。同时，为增强过拟合优化效果，增加模型的丢弃率为 0.5，即在训练过程中，每个神经元有 50%的概率被随机丢弃。用于模型微调的超参数如表 1 所示。

Table 1. Hyperparameter sets for model fine-tuning
表 1. 用于模型微调的超参数集

超参数	值
大型语言模型(LLMs)	BERT (bert-base-uncased) RoBERTa (roberta-base)
卷积神经网络(CNN)	多尺度卷积神经网络(MSCNN)
优化器	Adam, AdamW
损失函数(L)	交叉熵损失
训练的轮数	5, 10, 15
卷积核数量	128, 256, 512
丢弃率	0.2, 0.5
学习率	0.00001, 0.000001

4.5. 实验结果

图 6 展示了模型在训练过程中多个评估指标随训练轮次的变化情况。其中，训练损失整体呈下降趋势，从初始的约 0.525 逐渐降低到接近 0.35，说明模型在训练集上不断学习，对数据的拟合能力逐渐增强。验证损失在前期有一定波动，在第 6 轮左右达到峰值后呈下降趋势并趋于平稳，最终略低于训练损失。这表明模型没有出现严重的过拟合现象，在验证集上也能较好地泛化。

训练和验证准确率从较低水平开始，随着训练轮次增加而持续上升，最终接近 84%。验证准确率起始较高，在训练过程中有一些波动，但总体保持在较高水平，最终略低于训练准确率，接近 83%。这说明模型在训练过程中对新数据也有较好的预测能力，不过后期训练准确率超过验证准确率，可能有轻微过拟合倾向。

训练召回率开始时较低，在训练过程中逐步上升，在第 10 轮达到约 0.84 。验证召回率前期波动较大，在第 6 轮达到约 0.89 的峰值后下降，第 10 轮时约为 0.85。验证召回率的波动可能与验证集数据特点或模型在不同阶段对正例的识别能力变化有关。

训练 F1 值从约 0.74 逐步上升到接近 0.85。验证 F1 值前期相对平稳，略有波动，后期被训练 F1 值超过，最终约为 0.84。F1 值综合了精确率和召回率，其变化趋势与准确率和召回率的变化相关。

模型在训练过程中性能不断提升，虽有轻微过拟合迹象，但整体表现良好，在训练集和验证集上都有不错的效果。

图 7 是模型的混淆矩阵图，反映了模型预测结果与真实标签之间的关系。模型的评估指标通过混淆矩阵图可以计算得出。TN 为左上角的 2040，TP 为右下角的 2127；FP 为右上角的 448；FN 为左下角 385。代入公式(3)、(4)、(5)、(6)可得模型准确率为 0.8334，精确率为 0.8260，召回率为 0.8467，F1 值为 0.8362。整体来看，该模型在正负情感分类效果较好，但仍存在部分误判情况，后续将通过优化模型、调整超参数或扩充数据集等方式进一步提升性能。

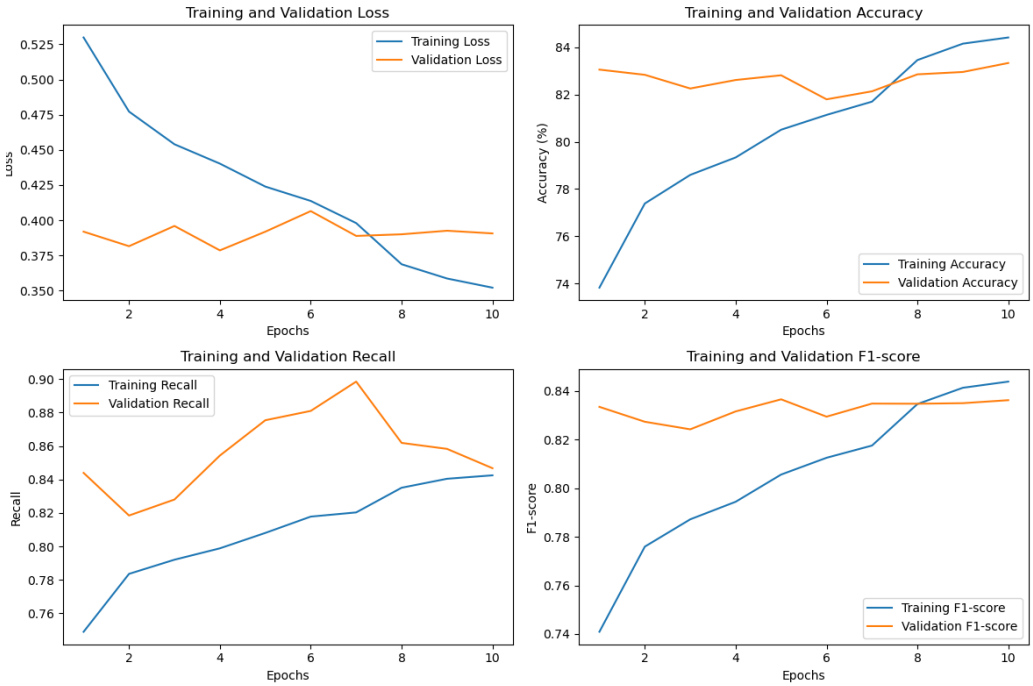

Figure 6. The variation of evaluation metrics with the number of training epochs
图 6. 评估指标随训练轮次的变化情况

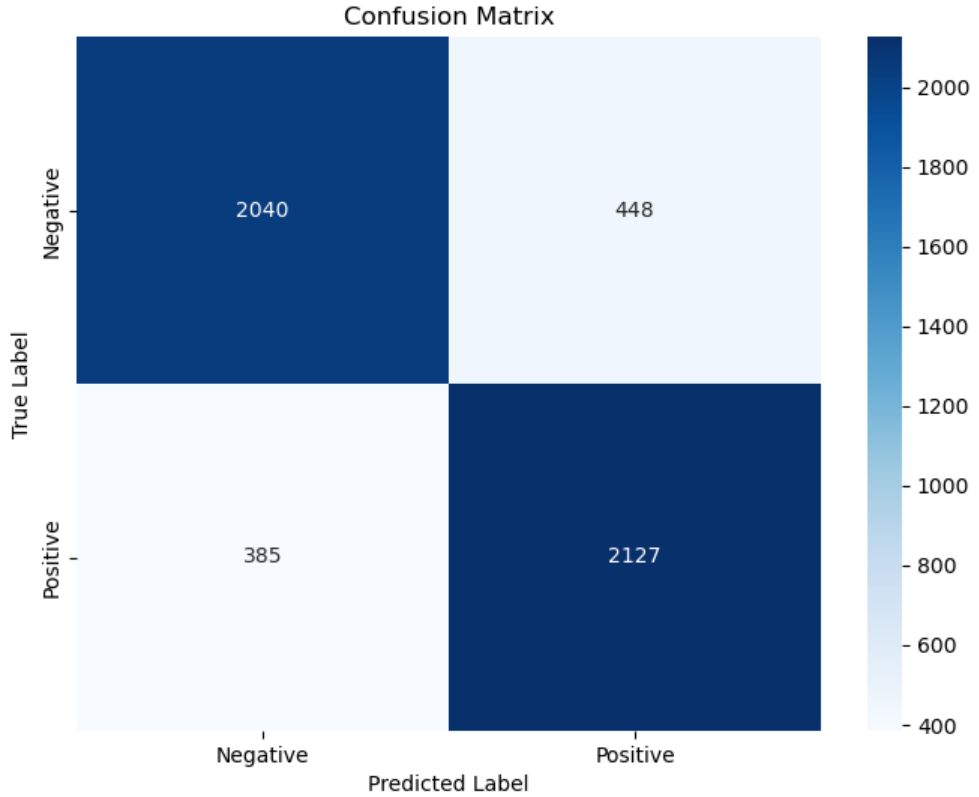

Figure 7. The variation of evaluation metrics with the number of training epochs
图 7. 评估指标随训练轮次的变化情况

本文在检查模型对预测错误的文本时发现，“Well maybe next time my dreams will come true.”等具有微妙情感变化文本的情感类别无法被判断准确。模型可能只看到了“dreams will come true”这样相对积极的表达，而忽略了“maybe next time”所传达出的一种无奈、不确定和当下的失落感。又如，“This was supposed to be a good weekend.”被预测为积极情感类评论，“was supposed to be”通常暗示实际与预期不符，可能带有失望情绪。但模型也许在分析时没有充分理解或考虑到这一隐含意义，单纯从“good weekend”这个表述出发，将其判定为积极。

模型可能更侧重于表面的词汇含义，而没有深入挖掘整体语境所传达的复杂情感。模型没有考虑到这条评论所处的新闻事件或话题背景，孤立地对句子进行了分析，导致无法准确判断情感。人类语言中的情感表达常常是微妙的，以上评论虽然有期望，但更多的是基于当下梦想未实现的一种感慨，并非单纯的积极情感。模型可能难以识别这种积极与消极混合情感中的主导因素。

针对模型弱点，可以收集更多类似包含惊讶、意外等复杂情绪的文本数据，以及不同情感强度的积极文本数据，让模型学习到更多样化的情感表达，提高对语义细微差别的识别能力。此外，将情感分类进一步细化，比如分为强积极、弱积极、惊讶积极等类别，使模型在预测时能够给出更准确的情感判断。同时，在训练模型时可以引入一些辅助信息，如用户的历史数据、社交网络背景等，帮助模型更好地理解用户的情感和意图。

4.6. 对比试验

本文选择 Word2Vec-MSCNN、BERT-MSCNN、RoBERTa-BiLSTM 作为对照模型，在相同数据集上与本文模型 RoBERTa-MSCNN 进行比较。结果如表 2 所示。

Table 2. Comparison of different models on the Sentiment140 dataset
表 2. 不同模型在 Sentiment140 数据集上的比较

模型	Accuracy	Precision	Recall	F1
Word2Vec-MSCNN	0.5245	0.5542	0.2749	0.3657
BERT-MSCNN	0.8144	0.8144	0.8144	0.8143
RoBERTa-BiLSTM	0.8225	0.8225	0.8225	0.8225
RoBERTa-MSCNN	0.8334	0.8260	0.8467	0.8362

可以看出，在相同的硬件条件和数据集下，本文模型性能相较 3 个对照模型均有不同程度的提升。其中，模型 Word2Vec 各项指标都相对较低。由于 Word2Vec 是生成静态词向量的模型，每个词对应固定向量，无法根据上下文动态调整语义，导致模型对情感倾向判断失误，模型整体性能欠佳。而 BERT-MSCNN 各项指标都达到 0.81 以上，说明模型能较好地理解文本情感语义，并通过卷积神经网络提取关键特征进行准确分类，在情感分析任务上表现良好。RoBERTa-MSCNN 相比 BERT-MSCNN 有提升，这得益于 RoBERTa 更优的预训练和 BiLSTM 对文本序列信息的有效处理，使模型能更精准地分析文本情感。本文提出的模型在所有模型中表现最佳，各项指标领先。因为 RoBERTa 提供了高质量的语义表示，而 MSCNN 又能有效挖掘不同尺度下的关键情感特征，两者协同作用，大幅提升了模型的情感分类能力。

4.7. 可解释性分析

本文利用 LIME 生成可解释性图，图 8 展示了模型对文本情感倾向的预测概率。消极的预测概率为 0.98，积极的预测概率为 0.02，说明模型判断这段文本的情感倾向为消极的可能性非常大。右侧有两个列表，“Negative”和“Positive”，分别展示了文本中不同词汇对消极和积极情感的贡献程度。其中，“annoying”对消极情感的贡献值为 0.18，“amazing”对积极情感的贡献值为 0.11，且该文本中词汇对消极情感的贡献程度更大，说明模型判断这段文本的情感倾向为消极的可能性非常大。下方的文本中，不同颜色高亮显示了不同词汇。橙色高亮的词汇对积极情感有一定贡献，蓝色高亮的词汇对消极情感贡献更大。结合评论的上文信息可知，虽然文本中提到了手机显示方面的优点，但电池续航差这一缺点使整体情感倾向偏向消极。

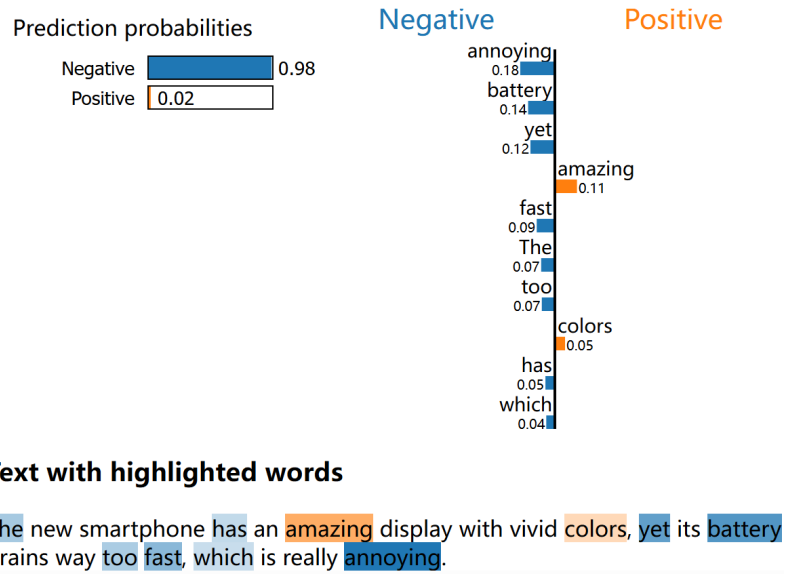

Figure 8. An explanation diagram of LIME for the classification results of the model
图 8. LIME 对模型分类结果的解释图

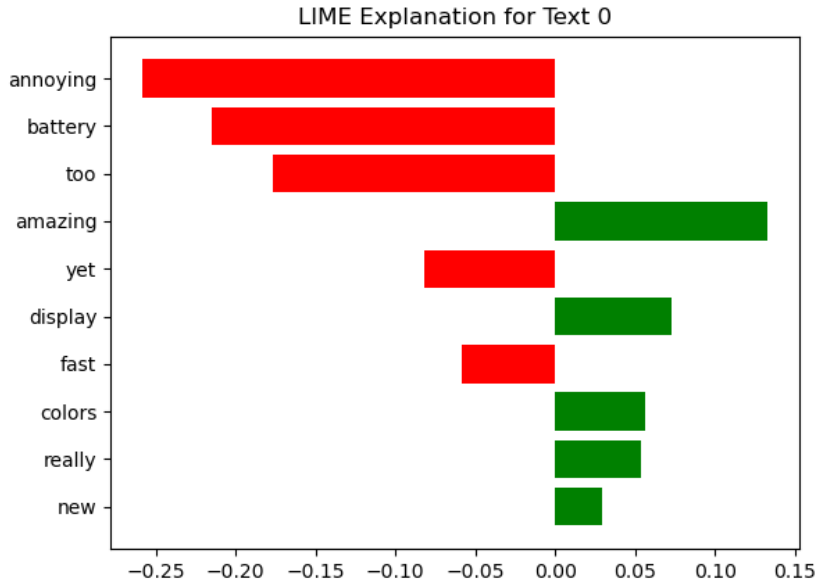

Figure 9. Examples of the contribution degree of words
图 9. 词汇贡献程度示例

为展示不同词汇对模型预测结果的影响，利用LIME计算出不同词汇对文本情感极性判断的重要性。模型在进行情感判断时，会综合考虑这些词汇的影响。结合这些词汇的贡献程度，可以一定程度上理解模型做出预测的依据。图 9 显示“annoying”(恼人的)的红色条形最长，说明它对文本被判断为负面情感的贡献最大；“amazing”(令人惊叹的)的绿色条形最长，对文本被判断为正面情感的贡献最大。同时，红色条形比绿色条形更长，表明负面情感对模型预测结果的影响越大。

5. 结语

本研究成功构建了基于 RoBERTa-MSCNN 的新闻评论可解释性情感分析模型，有效融合了预训练语言模型和多尺度卷积神经网络的优势。借助 RoBERTa 对上下文语义的深度理解能力，结合 MSCNN 对文本局部特征的捕捉能力，该模型在情感分析任务中展现出了较高的准确性和鲁棒性。

在实验过程中，通过合理调整卷积核的大小和数量，优化了模型对不同尺度语义特征的提取能力。同时，采用数据增强技术和合理的超参数设置，提高了模型的泛化能力，减少了过拟合现象的发生。

在可解释性方面，利用 LIME 解释器对模型的决策过程进行了可视化分析，揭示了模型在情感分类过程中对关键特征的关注情况。这不仅增强了模型的透明度，也为用户理解模型的决策依据提供了有力支持。

然而，本研究仍存在一些不足之处。首先，数据集比较单一，只能做二分类的情感分析。缺少对细粒度文本的情感分析，以此验证模型分类能力。其次，模型的复杂度较高，在实际应用中需要进一步优化以提高推理速度。此外，可解释性方法虽然取得了一定的效果，LIME 主要关注局部样本的解释，只能对单个或少数样本的预测结果进行解释，难以提供模型在整个数据集上的全局行为和决策边界的理解。

总之，本研究为新闻评论的情感分析提供了一种有效的方法，未来将继续致力于模型的优化和改进，以推动自然语言处理技术在实际应用中的发展。

参考文献

[1] 宁益民. 基于深度学习的新闻评论情感分析研究[D]: [硕士学位论文]. 北京: 北京邮电大学, 2023.

[2] 惠调艳, 王智, 何振华, 等. 基于词典-TextCNN-Word2Vec 组合模型的在线评价细粒度情感分析[J]. 情报理论与实践, 2025, 48(2): 168-177.

[3] Srivats Athindran, N., Manikandaraj, S. and Kamaleshwar, R. (2018) Comparative Analysis of Customer Sentiments on Competing Brands Using Hybrid Model Approach. 2018 3*rd International Conference on Inventive Computation Technologies* (*ICICT*), Coimbatore, 15-16 November 2018, 348-353. https://doi.org/10.1109/icict43934.2018.9034283

[4] Vanaja, S. and Belwal, M. (2018) Aspect-Level Sentiment Analysis on E-Commerce Data. 2018 *International Conference on Inventive Research in Computing Applications* (*ICIRCA*), Coimbatore, 11-12 July 2018, 1275-1279. https://doi.org/10.1109/icirca.2018.8597286

[5] 程顺正, 胡楠. 基于 BERT 模型的短视频平台评论信息情感分析研究[J]. 辽宁科技学院学报, 2024, 26(6): 30-33.

[6] Kamruzzaman, M., Hossain, M., Imran, M.R.I. and Bakchy, S.C. (2021) A Comparative Analysis of Sentiment Classification Based on Deep and Traditional Ensemble Machine Learning Models. 2021 *International Conference on Science & Contemporary Technologies* (*ICSCT*), Dhaka, 5-7 August 2021, 1-5. https://doi.org/10.1109/icsct53883.2021.9642583

[7] Janardhana, D.R., Vijay, C.P., Swamy, G.B.J. and Ganaraj, K. (2020. Feature Enhancement Based Text Sentiment Classification Using Deep Learning Model. 2020 5*th International Conference on Computing*, *Communication and Security* (*ICCCS*), Patna, 14-16 October 2020, 1-6. https://doi.org/10.1109/icccs49678.2020.9277109

[8] Tan, K.L., Lee, C.P., Anbananthen, K.S.M. and Lim, K.M. (2022) RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis with Transformer and Recurrent Neural Network. *IEEE Access*, **10**, 21517-21525. https://doi.org/10.1109/access.2022.3152828

[9] 杨明. 基于 BERT-CBA 方法的中文新闻情感分类[J]. 长江信息通信, 2022, 35(5): 170-172.

[10] 林奕欧. 深度情感分析模型的建模方法和可解释性研究[D]: [硕士学位论文]. 成都: 电子科技大学, 2021.

[11] 蒋建洪, 李梦欣. 融合 LGBM 和 SHAP 的企业负面事件下微博用户极端情感可解释模型[J]. 数据分析与知识发现, 2024, 8(7): 103-117.

[12] Jain, R., Kumar, A., Nayyar, A., Dewan, K., Garg, R., Raman, S., *et al.* (2023) Explaining Sentiment Analysis Results on Social Media Texts through Visualization. *Multimedia Tools and Applications*, **82**, 22613-22629. https://doi.org/10.1007/s11042-023-14432-y

[13] Go, A. Bhayani, R. and Huang, L. (2009) Twitter Sentiment Classification Using Distant Supervision. CS224N Project Report, Stanford.