基于深度学习的微博评论情感分析

孙维泽[1]，种晨熙[1]，刘卉滢[1]，杜　骁[1]，纪胜谦[1]，李国文[3]，崔子践[2]，娄颜超[2*]

（1 喀什大学计算机科学与技术学，新疆　喀什　844000；2 喀什大学电子与通信工程学院，新疆　喀什　844000；3 喀什大学物理与电气工程学院，新疆　喀什　844000）

摘要：深度学习在短文本情感分析中提高了准确率等指标，但常因数据量小导致模型鲁棒性差、泛化能力差，易出现训练数据表现好的假象。针对此问题，提出基于清华大学中文新闻语料库 Word2vec 预训练模型和 BiLSTM-MHA 模型对短文本情感分析的处理方法。首先通过 Word2vec 预训练模型将文本信息转化成特征向量，然后利用 BiLSTM 提取特征，通过多头注意力机制增强关键特征。最后，将权重分配后的信息传入 softmax 分类器进行情感分类。本模型在 12 万微博评论上的准确率为 85.97%，表明其在处理大量文本数据时可保持较高准确率。

关键词：清华大学中文新闻语料库；Word2vec 预训练模型；双向长短时记忆网络；文本情感分析；多头注意力机制；NLP

收稿时间：2026 年 3 月 6 日　　**中图分类号：**029

通讯作者：*娄颜超，喀什大学电子与通信工程学院

Weibo Comment Sentiment Analysis Based on Deep Learning

Sun Weize[1], Chenxi Zhong[1], Huiying Liu[1], Xiao Du[1], Shengqian Ji[1], Guowen Li[3], Zijian Cui[2], Yanchao Lou[2*]

（1School of Computer Science and Technology, Kashi University, Kashi, 844000, China;2School of Electronics and Communications Engineering,, Kashi University, Kashi, 844000, China;School of Physics and Electrical Engineering,, Kashi University, Kashi, 844000, China）

Abstract: Deep learning has improved accuracy and other metrics in short text senti- ment analysis, but often suffers from poor model robustness and generalization ability due to small data volumes, leading to the illusion of good performance on training data. To address this issue, a method for short text sentiment analysis based on the Word2vec pre-trained model from Tsinghua University's Chinese news corpus and the BiLSTM- MHA model is proposed. First, the Word2vec pre-trained model is used to transform text information into feature vectors, then BiLSTM is utilized to extract features, and the multi-head attention mechanism is employed to enhance key features. Finally, the weighted information is passed into the softmax classifier for sentiment classification. The model achieves an accuracy of 85.97% on 120, 000 Weibo comments, indicating its ability to maintain high accuracy when processing large amounts of text data.

Key words: Tsinghua University Chinese News Corpus Word2vec pre-trained model; Bidirectional Long Short-Term Memory network; Text Sentiment Analysis; Multi-Head Attention mechanism; NLP

0　引言

随着互联网的迅速发展，大众对社会热点等社会事实的评论越来越多，大量短文本涌现。对短文本进行情感分析，并提取出有价值的信息

是学术界普遍关注的问题。情感分析也称为意见挖掘，旨在研究人们对某些实体的情感。文本情感分析是自然语言处理（Natural Language Processing，NLP）领域的一个重要分支，广泛应用于网络舆情分析和不同内容推荐等方面[1]。目前根据情感分析使用方法的不同，将其划分为基于情感词典、基于传统机器学习和基于深度学习的情感分析方法[2]。

基于情感词典的分析方法中，情感词典的构建可以依赖于已有的基础情感词典，目前被广泛认可的基础情感词典有：知网 HowNet 情感词典、台湾大学的中文情感极性词典 NTUSD 等，可以利用这些已有的情感词典构建短文本的情感词典；Wu 等将多部情感词典和语义规则集进行结合，该方式有效提高了对短文本的情感分析效果[3]。

基于传统机器学习的分析方法中，Pang 等首次将支持向量机、朴素贝叶斯等算法对电影评论数据进行了情感分类；Cao 等提出的 BGRU 模型效果优 CNN 和 BLSTM 等模型，并提出考虑将注意力模型、语言学知识结合到该模型中，且结构较为简单、训练速度较快，其训练速度是 BLSTM 的 1.36 倍；Li 等提出 UC-JS 模型，使得微博情感极性的分析效果均有显著提升[4-6]。

基于深度学习的情感分析方法中，Li 等提出一种卷积神经网络和双向长短时记忆（BiLSTM）特征融合的模型，该模型有效避免了传统循环神经网络梯度消失或梯度弥散问题，提高了特征融合模型在文本分类的准确率；Zheng 等提出卷积记忆神经网络模型，该模型既能够有效提取短文本局部最优特征，又能够解决远距离的上下文依赖，在短文本情感分析上优于 CNN、LSTM 和 SVM 模型；Liu 等提出一种基于 BERT 和 BiLSTM 的文本情感分类模型；Cao 等提出一种借助并行时序卷积网络（TCN）获得全面文本特征，并结合注意力机制的模型；Wu 等提出基于字向量表示方法并结合 Self-Attention 和 BiLSTM 的中文短文本情感分析算法；Yue 等在词嵌入层引入预训练词向量，在特征提取层使用 BiLSTM 提取上下文特征；Ren 等提出融合 TCN 和改进 BiLSTM（TCN-BiALSTM）的短文本情感分析算法，各项指标均达到 92% 以上[7-14]。

特别是将深度学习的方法引入情感分析任务后，情感分类的准确率越来越高，但普遍存在数据量较小的问题，易导致模型鲁棒性和泛化能力不足。基于此，本文提出基于 THUCNews Word2vec 和 BiLSTM-MHA 的微博评论情感分析模型，利用 2.19GB 的 THUCNews 语料库训练 Word2vec 模型，建立词向量，并用近 12 万条微博评论作为 BiLSTM-MHA 的训练数据集，从而有效缓解数据量不足的问题。

1 文本情感分析的相关理论基础

1.1 文本向量化

文本向量化是利用深度学习的方法对文本情感分析的前提，使用词向量的方法表示被分析的文本。在英文文本中，一个单词是语义的最小单元。与英文文本不同，中文文本的语义最小单元并不是单个汉字，而是由字构成的“词”。将文本信息划分成词更有利于模型对文本情感的理解。例如，将“您今天也带阿婆阿公去看猫宝宝了吧有爱心的乖宝宝赞悦馨是啊野猫把猫宝宝生在我家阳台上啦”按照字划分为“您 / 今 / 天 / 也 / 带 / 阿 / 婆 / 阿 / 公 / 去 / 看 / 猫 / 宝 / 宝 / 了 / 吧 / 有 / 爱 / 心 / 的 / 乖 / 宝 / 宝 / 赞 / 悦 / 馨 / 是 / 啊 / 野 / 按照词划分则为“您 / 今天 / 也 / 带 / 阿婆 / 阿公 / 去 / 看 / 猫宝宝 / 了 / 吧 / 有 / 爱心 / 的 / 乖宝宝 / 赞悦馨 / 是 / 啊 / 野猫 / 把 / 猫宝宝 / 生 / 在 / 我家 / 阳台 / 上 / 啦”，显然按照词划分更能表达文本的信息，符合中文表达习惯。早期词划分后的文本表示方法可以大致分为基于向量空间模型和 one-hot 两种方式。前者中的向量维度受词典中词数的影响，词数过多会引发维数灾难，进而导致计算代价过大；后者虽然简单，但由于词语之间独立编码，严重忽略了词与词之间的语义相关性。在文本处理问题上，由于上述方法本身的局限性，并不适用于短文本情感分析。目前流行的词向量转化方法有 Word2vec、GloVe 和 BERT。本文利用 Google 于 2013 年提出的 Word2vec 模型实现词向量化，可以有效缓解维数灾难、计算代价过大以及忽略语义相关性等问题。通过对 Word2vec 模型的训练，得到训练好的词向量，再用词向量对待训练的文本进行词向量化，实现将文本转化为指定维度的特征向量组。此外，Word2vec 对词的预测并不局限于第 n-1 个词，而是以窗口大小为 k 计算位于窗口

中心词出现的概率，使得预测更加全面。

1.2 双向长短时记忆网络

长短时记忆网络（LSTM）在一定程度上解决了循环神经网络（RNN）的梯度爆炸和梯度消失问题，但与 RNN 相同的是，LSTM 状态的传输方向也是从前往后，先天的结构缺陷导致 LSTM 无法获取从后往前方向的信息，只能提取到每个词的前文信息，而后文信息无法充分利用，如图 1 所示。短文本中词语的情感表达与上下文密切相关，因此解决短文本情感分析问题需要使用 BiLSTM。

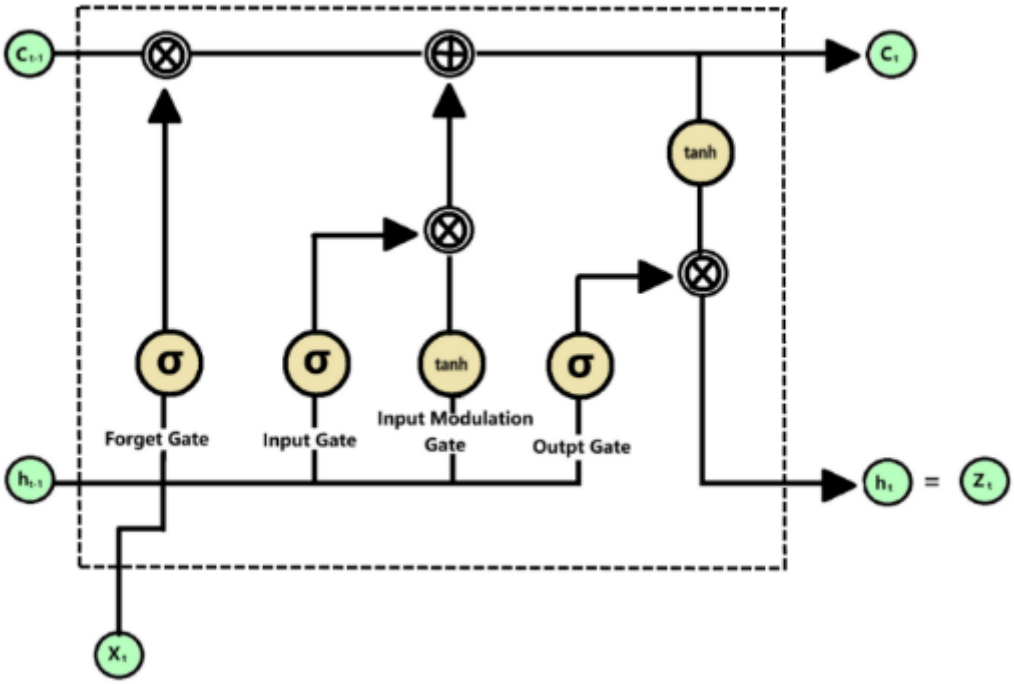

图 1 LSTM 神经元结构图

设每个时间步的输入为 xt，隐藏状态为 ht，记忆单元为 ct，LSTM 的门控结构可表示为：

$$ft = \sigma（Wf\,xt + Uf\,ht-1 + bf）$$

$$it = \sigma（Wixt + Uiht-1 + bi）$$

$$ot = \sigma（Woxt + Uoht-1 + bo）$$

其中 σ（）表示 Sigmoid 激活函数。记忆单元与隐藏状态更新为：

$$c-t = tan\,h（Wcxt + Ucht-1 + bc）$$

$$ct=ft \odot ct-1+it \odot c-t$$

$$ht = ot \odot tan\,h（ct）$$

其中 ⊙ 表示 Hadamard 乘积。

BiLSTM 由一个前向 LSTM 和一个后向 LSTM 构成，对同一输入序列从不同方向进行处理。设前向隐藏状态为 $\overrightarrow{h\,t}$，后向隐藏状态为 $\overleftarrow{h\text{-}t}$，则 BiLSTM 的输出为：

$$h\,t = concat(\overrightarrow{h\,t}, \overleftarrow{h-t})$$

1.3 多头注意力机制

BiLSTM 本质上是按照顺序处理输入数据，较早出现的数据经过长距离传递后会逐渐丧失影响力，难以捕捉长距离的依赖关系，如图 2 所示。多头注意力机制可以直接计算输入序列任意两个特征之间的相似度，更好地捕捉全局依赖关系；同时允许所有位置上的计算同时进行，大幅提升训练效率，改善模型的并行化能力。

给定输入序列矩阵 X，首先通过线性变换得到查询（Query）、键（Key）和值（Value）：

$$Q=XWQ，K=XWK，V=XWV$$

缩放点积自注意力机制定义为：

$$Attention(Q,K,V) = softmax(\frac{QKT}{\sqrt{dk}})V$$

其中 dk 为键向量维度。

多头注意力机制（Multi-Head Attention，MHA）

设置多个头并行计算：

$$MultiHead(Q, K, V) = Concat(head1, \cdots, headh) W^{O}$$

$$headi = Attention(QWiQ, KWiK, VWiV)$$

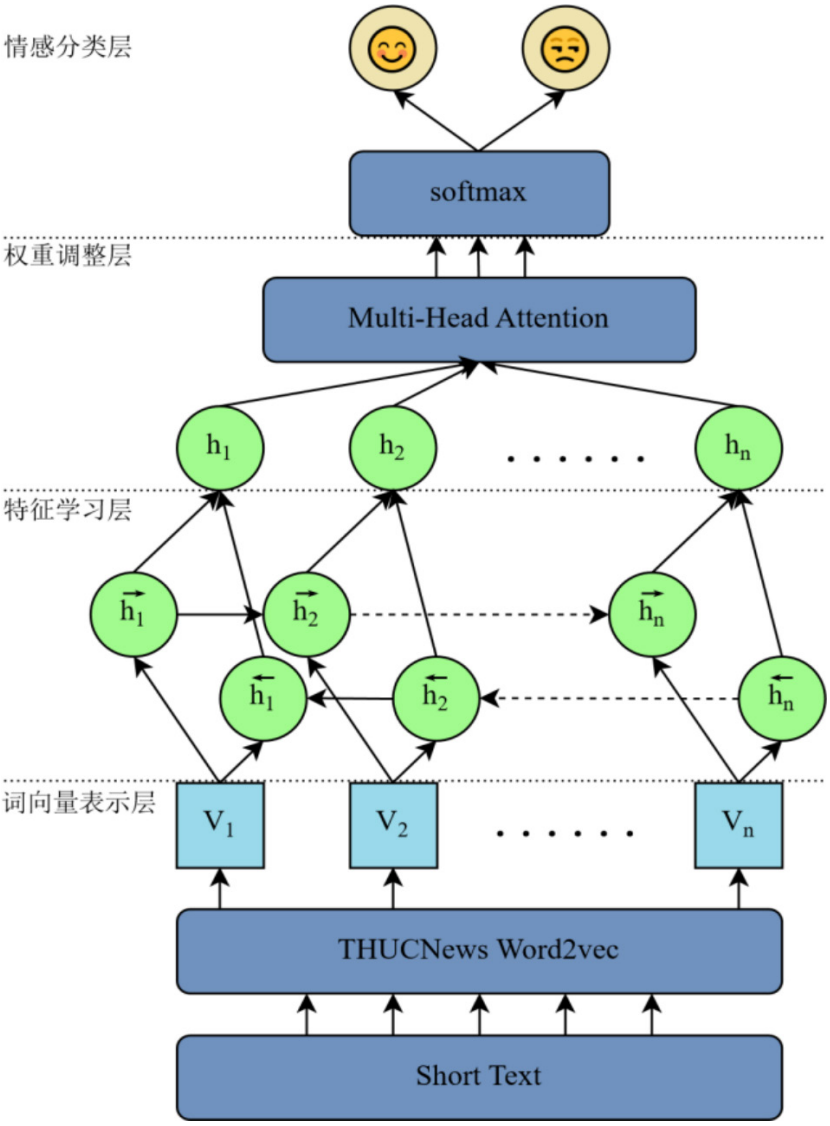

图 2 可视化多头注意力机制

2 模型与实验设计

2.1 实验数据及数据处理

THUCNews 数据集分为十四类新闻，共计约 74 万篇，每篇文本较短。为便于集中处理，首先编写脚本将同类新闻整合到同一个文本文件中，对文本进行清洗和分词，然后使用 Word2vec 训练 200 维词向量。对于 12 万条微博评论数据，采用类似的预处理流程，并对评论进行人工标注，通过对词向量求平均等方式实现评论级特征表示，具体推导和公式同前述。

2.2 实验环境与可调参数

实验环境包括 Windows 11 操作系统、12th Gen Intel（R）Core（TM）i7-1165G7 处理器、16 GB 内存以及 Intel（R）Iris（R）Xe Graphics 显卡，开发环境采用 MATLAB 2023b 与 Python 3.12。Word2vec 和 BiLSTM-MHA 模型的主要超参数设置为：词向量维度 200、窗口大小 5、最小词频 5，BiLSTM 隐藏层神经元数 32、层数 2，学习率 0.01，Dropout 比例 0.2，Batch size 为 50 等，具体数值与前文表格一致。

3 实验结果与分析

为全面评估模型分类性能，本文选取准确率、精确率、召回率、F1 值和特异性等指标，并给出 TP、TN、FP、FN 的数学定义及其对应公式。通过测试集准确率与损失函数随迭代次数变化的曲线可以看出，模型在约 1200 次迭代后基本收敛，准确率稳定在 85%－86% 左右，损失函数约收敛于 0.2，表明模型具有良好的稳定性和收敛性，如图 3、图 4 所示。

将 BiLSTM-MHA 与 k-NN、NBM、DT 等传统统计方法，以及 MLP、LSTM、CNN、BiLSTM 等深度学习模型进行对比实验。结果显示，在多个评价指标上 BiLSTM-MHA 均明显优于对比模型：与表现最好的 LSTM 模型相比，Accuracy 提升约 2.86%，Precision 提升约 2.72%，Recall、F1-score 和 Specificity 分别提升约 2.61%、2.70% 和 2.63%；与 BiLSTM 基线模型相比，Accuracy 提升约 2.74%，Precision 提升约 2.62%，Recall、F1- score 和 Specificity 分别提升约 2.49%、2.58% 和 2.51%。这些结果证明，多头注意力机制有效增强了 BiLSTM 对关键信息的建模能力。

此外，本文使用近 12 万条微博评论作为数据集，与其他文本情感分析研究相比，数据规模更大，且不同的模型得到的准确率与鲁棒性不同，如图 5 所示。实验结果表明，在大规模短文本情感分析任务下，所提出的 BiLSTM-MHA 模型仍能保持较高的准确率和较好的鲁棒性。

4 讨论与结束语

本文提出了一种基于 THUCNews Word2vec 和 BiLSTM-MHA 的大数据量微博评论情感分析模型。在文本向量化阶段，采用 Word2vec 将短文本转化为特征向量，有效缓解了维数灾难、计算代价过大以及忽略语义相关性等问题；在特征提取阶段，引入 BiLSTM 并结合多头注意力机制，大幅提升了模型对全局信息和长距离依赖的捕捉能力，如表 1 所示。

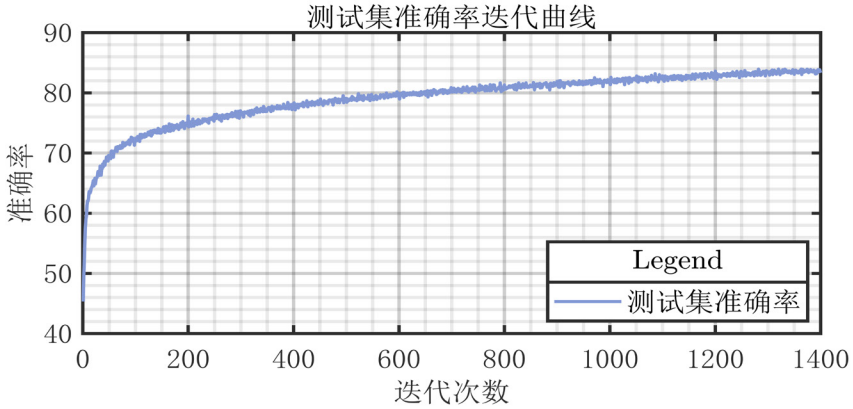

图 3 测试集准确率迭代曲线

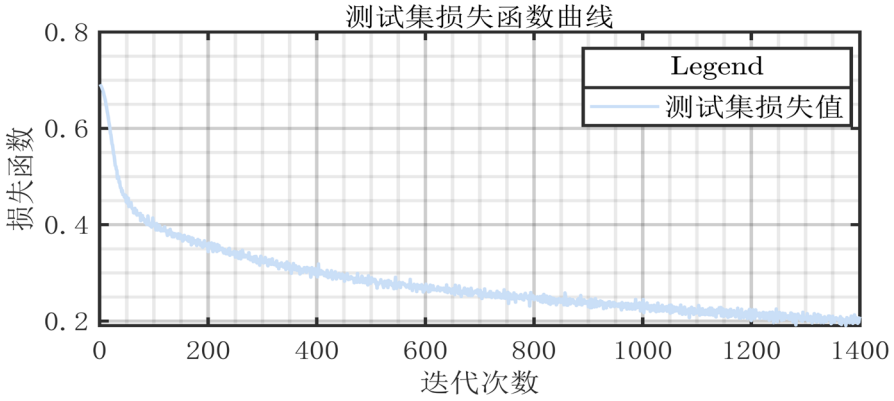

图 4 测试集损失函数曲线

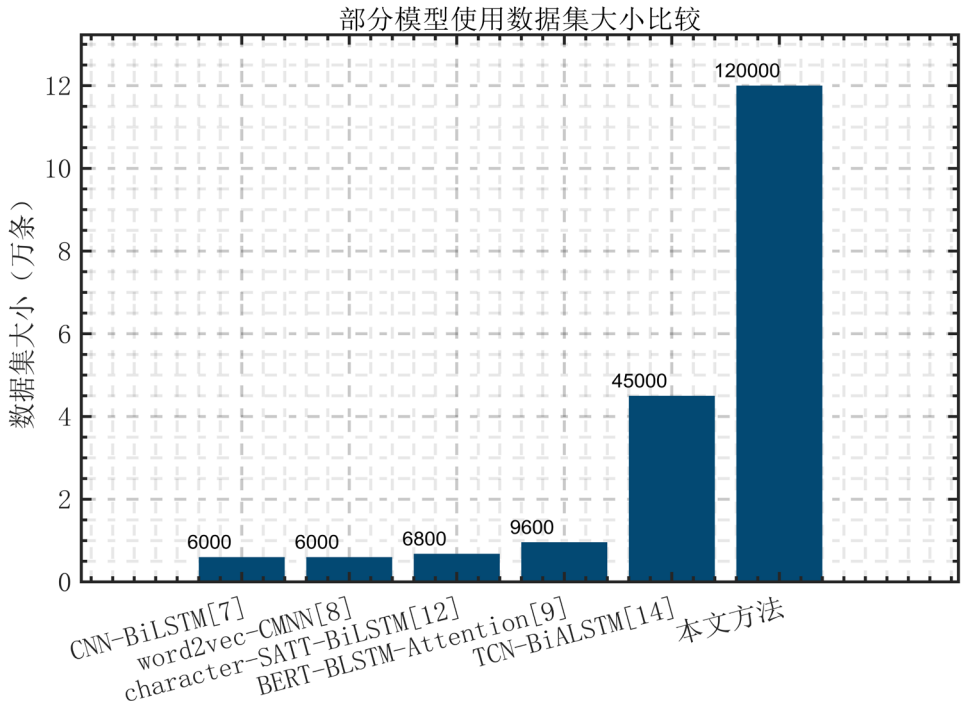

图 5 部分模型使用数据集大小比较

表 1 加入 MH A 前后结果对比表 %

Method	Accuracy	Precision	Recall	F1-score	Specificity
BiLSTM	83.23	83.23	83.24	83.23	83.24
BiLSTM-MHA	85.97	85.85	85.73	85.81	85.75

实验结果表 2 所示，相比传统统计方法和不含注意力机制的深度学习模型，BiLSTM- MHA 在准确率、精确率、召回率、F1 值和特异性等评价指标上均有显著提升；测试集准确率和损失函数曲线表明，模型在训练过程中具有良好的稳定性和收敛性。结合近 12 万条微博评论这一大规模数据集，进一步验证了模型在实际短文本情感分析任务中的有效性与鲁棒性。

表 2 加入 MH A 前后结果对比表 %

模型名称	Accuracy	Precision	Recall	F1-score	Specificity
k-NN	68.08	68.23	68.08	68.01	68.08
NBM	68.42	69.62	68.43	67.93	68.43
DT	66.38	66.38	66.38	66.38	66.38
RF	76.39	76.40	76.39	76.39	76.39
MLP	81.16	81.22	81.17	81.16	81.17
LSTM	83.11	83.12	83.12	83.11	83.12
GRU	82.72	82.74	82.73	82.72	82.73
CNN	80.96	81.17	80.93	80.92	80.93
BiLSTM-MHA	85.97	85.85	85.73	85.81	85.75

需要指出的是，本文主要聚焦于情感极性的二分类任务，尚未对情感强度进行更精细的建模。

未来工作将进一步考虑引入多级情感标签或连续情感强度刻画，对不同强度和细粒度情感进行建模与分析；同时，可探索在更大规模、多领域语料上的迁移与适配，以进一步提升模型在复杂真实场景中的适用性。

参考文献：

[1] Xing Fang, Justin Zhan. Sentiment analysis using product review data [J]. Journal

of Big Data, 2015, 2:5. DOI:10.1186/s40537-015-0015-2.

[2] Dai L, Liu B, Xia Y, et al. Measuring semantic similarity between words using HowNet [C] // 2008 International Conference on Computer Science and Information Technology. IEEE, 2008: 601 - 605.

[3] 吴杰胜, 陆奎. 基于多部情感词典和规则集的中文微博情感分析研究[J]. 计算机应用与软件, 2019, 36(09): 93 - 99.

[4] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment Classification using Ma- chine LearningTechniques [J]. Empirical Methods in Natural Language Processing, 2002: 79 - 86.

[5] 曹宇, 李天瑞, 贾真, 殷成凤. BGRU: 中文文本情感分析的新方法[J]. 计算机科学与探索, 2019, 13(06): 973 - 981.

[6] 李玉强, 黄瑜, 孙念, 李琳, 刘爱华. 基于性格情绪特征的改进主题情感模型[J]. 中文信息学报, 2020, 34(07): 96 - 104.

[7] 李洋, 董红斌. 基于 CNN 和 BiLSTM 网络特征融合的文本情感分析[J]. 计算机应用, 2018, 38(11): 3075 - 3080.

[8] 郑啸, 王义真, 袁志祥, 等. 基于卷积记忆神经网络的微博短文本情感分析[J]. 电子测量与仪器学报, 2018, 32(03): 195 - 200. DOI:10.13382/j.jemi.2018.03.028.

[9] 刘思琴, 冯胥睿瑞. 基于 BERT 的文本情感分析[J]. 信息安全研究, 2018, 6(03).

[10] 刘文秀, 李艳梅, 罗建, 等. 基于 BERT 与 BiLSTM 的中文短文本情感分析[J]. 太原师范学院学报, 2020, 19(4): 52 - 58.

[11] Cao Dong, Huang Yujie, Fu Yunbin. Text sentiment analysis based on parallel TCN model and attention model [C] // 2nd Symposium on Signal Processing Systems, 2020: 51 - 55.

[12] 吴小华, 陈莉, 魏甜甜, 等. 基于 Self-Attention 和 Bi-LSTM 的中文短文本情感分析[J]. 中文信息学报, 2019, 33(06): 100 - 107.

[13] Arias M E, Cochrane T A, Piman T, et al. Quantifying changes in flooding and habitats in the Tonle Sap Lake (Cambodia) caused by water infrastructure development and climate change in the Mekong Basin [J]. Journal of Environmental Management, 2012, 112: 53-66.

[14] Fotheringham A S, Yang W, Kang W. Multiscale GWR (MGWR) [J]. Annals of the AAG, 2017, 107(6): 1247-1265.

作者简介：孙维泽（2004-），男，汉族，山东济南人，喀什大学网络工程专业本科在读，主要研究方向为网络入侵检测、多模态；种晨熙（2004-），男，汉族，河南灵宝人，喀什大学网络工程专业本科在读，主要研究方向为网络安全、路由交换；纪胜谦（2004-），男，汉族，山东聊城人，喀什大学网络工程专业本科在读，主要研究方向网络架构、路由交换与园区网络优化；杜骁（2002-），女，汉族，山东青岛人，喀什大学计算机科学与技术专业本科在读，主要研究方向网络架构、网络安全；刘卉滢（2004-），女，汉族，山东临沂人，喀什大学计算机科学与技术专业本科在读，主要研究方向为自然语言处理、计算机技术；崔子践（2004-），男，满族，辽宁铁岭人，喀什大学通信工程专业本科在读，主要研究方向为无线通信技术、数字信号处理；李国文（2005-），男，汉族，河北张家口人，喀什大学电气工程及其自动化专业本科在读，主要研究方向为电力系统保护、网络安全。