ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.19, No.35, December 2023

E-mail: eduf@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于CNN-LSTM深度学习技术的知乎文本情感分析

刘飞生,魏超

(广州松田职业学院,广东 广州 51000)

摘要:知乎平台作为中国主要的知识共享社区,承载着海量信息,因此对其进行情感分析具有重要的现实意义。本研究旨在结合卷积神经网络(CNN)与长短时记忆网络(LSTM)技术,实现对知乎平台上大量文本数据的情感分析。本文研究并验证了CNN和LSTM技术在文本情感分析中的性能,通过融合两种技术有效地提升了情感分类的精度。实验结果表明,融合了CNN与LSTM的模型在在多个领域中情感分类方面呈现出更优异的表现,从而验证了其显著的有效性和潜力。

关键词:文本情感分析;知乎;卷积神经网络;长短时记忆;循环神经网络

中图分类号:TP18 **文献标识码:**A

文章编号:1009-3044(2023)35-0020-03

开放科学(资源服务)标识码(OSID):

DOI:10.14004/j.cnki.ckt.2023.1843

0 引言

在信息时代的浪潮下,社交媒体和网络平台扮演着不可或缺的角色,为人们提供了一个广泛的信息交流和表达平台。知乎作为中国领先的知识分享社区,吸引了4亿多的注册用户,涵盖了广泛的话题、问题和观点。随着用户规模的不断扩大,知乎平台所积累的海量文本信息数据变得丰富和多样。这些文本数据承载着用户对于各类话题的观点、情感以及态度,其深层次的情感信息对于理解用户需求、产品改进以及舆情分析具有重要意义。

随着社交媒体信息的爆炸性增长,对于海量文本数据的高效处理和情感分类需求变得更加迫切[1]。传统的文本分析方法往往受限于特征工程的复杂性和规模效应,难以满足大规模数据的处理要求。近年来,深度学习技术的兴起为文本情感分析带来了崭新的可能性。通过构建复杂的神经网络结构,深度学习能够自动地从原始文本数据中提取高层次的语义特征,从而实现高效准确的情感分类[2]。

深度学习在文本情感分析领域的应用已经取得了显著的成果,其中卷积神经网络(Convolutional Neural Networks, CNN)和长短时记忆网络(Long Short-Term Memory, LSTM)等技术在文本分类任务中表现出色。CNN在图像处理中的成功应用启发了研究人员将其扩展到文本领域,其卓越的特征提取能力对于捕捉文本的局部特征非常有效[3];LSTM作为一种适用于序列数据的循环神经网络,能够捕捉文本的时序信息,对于情感分析尤为重要[4]。

本文旨在借助深度学习技术,通过CNN和LSTM技术进行融合来探索并解决知乎平台上海量文本数据的情感分类问题。本文将深入研究并验证CNN和LSTM技术在文本情感分析中的表现,进一步探讨它们融合的优势和潜力,以期为社交媒体情感分析领域的研究和应用提供有力的支持。

1 实验环境及数据处理

1.1 实验环境

本文的实验环境如下所示:

- 操作系统:CentOS Linux 7.0;
- 开发环境:TensorFlow;
- 通用库:numpy、scikit-learn、scipy、nltk。

TensorFlow作为深度学习框架,为本文的实验提供了稳健的基础。通过numpy、scikit-learn、scipy和nltk等通用库的支持,我们能够方便地进行数据处理、特征提取和模型评估。

1.2 实验数据集

本文实验数据集分为2个主要部分:

1) 知乎网站数据集

通过Python爬虫技术,笔者采集了丰富的知乎用户评论和帖子数据,作为情感分析的基础数据集。这些数据涵盖了多个领域和话题,涉及科技、文化、娱乐等多个领域,为我们的研究提供了丰富多样的文本素材。

2) 新华社新闻数据集

引用中文新闻数据集,通常被称为“新华社”数据集,作为补充实验数据。这个数据集包含了大量的中文新闻文本,覆盖了不同的新闻领域和主题。这样的数据集在情感分析任务中能够提供更多的文本样本,

收稿日期:2023-08-25

作者简介:刘飞生(1990—),江西赣州人,高级工程师,博士研究生,研究方向为人工智能、深度学习;魏超(1971—),辽宁葫芦岛人,副教授,高级工程师,硕士研究生,研究方向为智能制造。

丰富了研究数据。

1.3 数据预处理

为了准备数据,采取以下步骤进行数据预处理:

1) 中文分词

使用中国科学院计算所开发的中文分词软件包NLPIR进行中文分词。NLPIR不仅提供了高效准确的中文分词功能,还能进行词性标注、命名实体识别以及用户词典的支持。这有助于将文本数据转化为更加适合模型处理的词汇序列。

2) 文本清洗与停用词去除

在分词完成后,笔者进行了文本清洗,包括去除特殊字符、标点符号和无意义的空白符。此外,还剔除了停用词,这些停用词通常不携带太多情感信息,但会占据文本中的空间。

经过以上数据预处理步骤,得到了分词、清洗且剔除了停用词之后的文本数据,为接下来的特征提取和模型训练做好了准备。

2 融合模型设置

为了充分发挥CNN和LSTM的优势,在参考Ombabi[5]的研究成果基础上,本文针对中文语境提出了一种融合方法,将它们结合起来进行知乎平台文本情感分析。以下是该融合方法的详细步骤:

2.1 文本表示

首先,利用预训练的词向量模型(如Word2Vec或GloVe)将原始文本转换为词向量表示。这些词向量能够有效地捕捉词汇之间的语义关系,为后续的模型提供有意义的输入。

2.2 卷积操作

将词向量表示输入一层卷积神经网络中,该网络由多个卷积核和池化操作构成。卷积核在捕捉不同大小的局部特征方面表现出色,而池化操作则有助于减少数据的维度,同时提取关键特征。

2.3 LSTM建模

卷积层的输出被馈送至一个双向LSTM层。双向LSTM能够同时捕捉文本的前向和后向信息,从而更好地理解文本的上下文语境。LSTM层的输出被连接在一起,并通过全连接层进行情感分类。

2.4 融合模型的训练优化

在模型训练过程中,使用以下实验参数设置来优化融合模型:

- 优化算法:选用常用的Adam优化算法,以最小化交叉熵损失函数。
- 学习率:初试学习率设定为一个较小的值0.001,通过实验验证找到合适的学习率调度策略,如学习率衰减。
- 批大小:批处理大小对模型训练速度和稳定性具有影响,进行批大小的调整和实验。
- Dropout:在全连接层和LSTM层中引入Dropout层,以减少过拟合风险。
- Epochs:设置合适的训练迭代次数,避免过拟合或欠拟合情况的出现。

2.5 实验参数设置

为了验证模型性能,笔者设计了一系列实验,包括单独使用CNN、单独使用LSTM以及融合CNN与LSTM模型的情况,最终参数如表1所示。

表1 实验模型参数设置

模 型	卷积核大小	LSTM隐藏层单元数	学习率	批大小	Dropout率
单独CNN	3、4、5	–	0.001	64	0.5
单独LSTM	–	128	0.001	32	0.3
融合模型	3、4、5	64	0.001	64	0.4

通过对不同模型的参数设置进行调整,旨在获得最佳性能,以便在知乎平台文本情感分析任务中取得更准确的结果。

3 模型训练及评估

3.1 对比实验

在基于单独使用CNN、LSTM以及融合CNN与LSTM三种方法的基础上,笔者选择了知乎平台上不同领域内容,包括“美食”“台风”和“科技”,进行了模型训练及评估,其中涉及80%的训练数据和20%的测试数据。

3.2 实验评估指标

在评估性能时使用标准评估指标进行验证,使用accuracy准确度、precision精密度(又称精度)、sensitivity灵敏度(又称召回率)、specificity特异性、F-Score综合评估指标这5个参数进行性能评估,其值可以使用混淆矩阵及对应公式来确定。

表2 混淆矩阵表

	实际为真	实际为假
预测为真	真阳性(TP)	假阳性(FP)
预测为假	假阴性(FN)	真阴性(TN)

$$准确度Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$精度Precision = \frac{TP}{TP + FP} \quad (2)$$

$$召回率Sensitity = \frac{TP}{TP + FN} \quad (3)$$

$$特异性Specificity = \frac{TN}{TN + FP} \quad (4)$$

$$F1 = \frac{P*R}{P + R} \quad (5)$$

3.3 实验结果与分析

作者对知乎数据集和新华社数据集上进行了一系列的实验，针对“美食”“台风”“科技”三个特定领域进行了情感分类性能评估。实验结果如表3所示。

表3　实验评估性能

实验方法	研究领域	准确度	精度	召回率	特异性	F-1分数
单独CNN	美 食	0.87	0.88	0.86	0.90	0.87
	台 风	0.78	0.81	0.76	0.85	0.78
	科 技	0.92	0.91	0.93	0.94	0.92
单独LSTM	美 食	0.85	0.86	0.84	0.88	0.85
	台 风	0.76	0.79	0.74	0.82	0.76
	科 技	0.91	0.90	0.92	0.93	0.91
CNN与LSTM融合模型	美 食	0.90	0.91	0.89	0.92	0.90
	台 风	0.82	0.85	0.80	0.87	0.82
	科 技	0.93	0.92	0.94	0.95	0.93

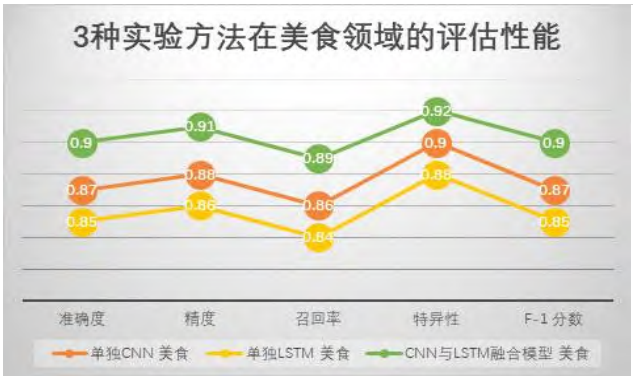

图1　3种方法在美食领域的评估性能

通过对表3中的美食研究领域进行数据分析，评估性能如图1所示。经过对比结果可知，本文采用的CNN与LSTM融合模型的准确度、精度、召回率、特异性、F-1分别为0.90、0.91、0.89、0.92、0.90，要比单独的CNN模型和单独的LSTM模型取得的评估性能更好。说明CNN与LSTM融合模型在知乎平台的“美食”领域上的评估预测率更好。

图2　3种方法在台风领域的评估性能

通过对表3中的台风研究领域进行数据分析，评估性能如图2所示。经过对比结果可知，本文采用的CNN与LSTM融合模型的准确度、精度、召回率、特异性、F-1分别为0.82、0.85、0.8、0.87、0.82，要比单独的CNN模型和单独的LSTM模型取得的评估性能要好。说明CNN与LSTM融合模型在知乎平台的“台风”领域上的评估预测率更好。

图3　3种方法在科技领域的评估性能

通过对表3中的科技研究领域进行数据分析，评估性能如图3所示。经过对比结果可知，本文采用的CNN与LSTM融合模型的准确度、精度、召回率、特异性、F-1分别为0.93、0.92、0.94、0.95、0.93，要比单独的CNN模型和单独的LSTM模型取得的评估性能更好。说明CNN与LSTM融合模型在知乎平台的“科技”领域上的评估预测率更好。

综合以上可知，在3个不同领域中，融合了CNN与LSTM的模型在情感分类上表现出了更高的准确率。这表明通过将卷积神经网络和长短时记忆网络相融合，能够更好地捕捉文本中的特征和上下文信息，从而提高情感分类的准确性。

4 结束语

本文以知乎这个信息丰富的社交平台为对象，探索了一种有效的文本情感分析方法。通过结合卷积神经网络(CNN)和长短时记忆网络(LSTM)的融合模型，在不同领域的情感分类任务中取得了令人满意的成果，证实了融合模型在情感分析任务中的潜力。然而，鉴于实验设备条件的限制，本文未能在深层次上探索CNN与LSTM的融合，未来的研究可以考虑利用更强大的计算资源，进一步挖掘模型的潜力。

参考文献：

[1] 杜昌顺.面向细分领域的舆情情感分析关键技术研究[D].北京:北京交通大学,2019.
[2] 邓钰.面向短文本的情感分析关键技术研究[D].成都:电子科技大学,2021.
[3] KIM Y.Convolutional neural networks for sentence classification[EB/OL]. 2014: arXiv: 1408.5882. https://arxiv. org/abs/1408.5882.pdf
[4] HOCHREITER S,SCHMIDHUBER J.Long short-term memory [J].Neural Computation,1997,9(8):1735-1780.
[5] OMBABI A H,OUARDA W,ALIMI A M.Deep learning CNN - LSTM framework for Arabic sentiment analysis using textual information shared in social networks[J].Social Network Analysis and Mining,2020,10(1):1-13.

【通联编辑：唐一东】