DOI:10.19651/j.cnki.emt.2005570

基于 Bi-LSTMA-CNNA 的线上评论情感分析模型*

占 妮
(华中师范大学 物理科学与技术学院 武汉 430070)

摘 要：基于深度学习的文本情感分析是目前自然语言处理研究的重要方向，在卷积神经网络、双向长短期记忆网络的基础上提出一种性能优于前面两种算法的情感分析算法。改进后的情感分析算法结合传统的深度学习结构，将双向长短期记忆网络、卷积神经网络以及注意力机制相结合，其中双向长短期记忆网络与注意力机制相结合的部分主要用来提取全局特征，并对目标词重点关注；卷积神经网络与注意力机制相结合的部分主要用来提取局部重要特征；最后将两部分特征相融合再进行分类。实验结果表明在对线上评论情感分析时，CNN 模型的 F1 为 0.793 9、Bi-LSTM 的 F1 为 0.795 9、Bi-LSTM-Attention 的 F1 为 0.799 8、Bi-LSTMA-CNNA 的 F1 为 0.802 8；因此改进后的模型性能优于其他 3 个模型。

关键词：双向长短期记忆网络；卷积神经网络；注意力机制
中图分类号：TP183；TP391.1　　**文献标识码**：A　　**国家标准学科分类代码**：520.2020

Online comment sentiment analysis based on Bi-LSTMA-CNNA

Zhan Ni
(College of Physical Science and Technology, Central China Normal University, Wuhan 430070, China)

Abstract: Text emotion analysis based on deep learning is an important direction of natural language processing research at present. Based on CNN and Bi-LSTM proposes a emotion analysis algorithm with better performance than the previous two algorithms. The improved emotion analysis algorithm combines the traditional deep learning structure and combines the Bi-LSTM, CNN and Attention mechanism. The part combining the Bi-LSTM and Attention mechanism is mainly used to extract global features and focus on target words. The part of CNN and Attention mechanism is mainly used to extract local important features. Finally, the characteristics of the two parts are integrated and then classified. The results of the experiment showed that when analyzing the emotions of online comments. The F1 of CNN model was 0.793 9, the F1 of Bi-LSTM was 0.795 9, the F1 of Bi-LSTM-Attention was 0.799 8, and the F1 of Bi-LSTMA-CNNA was 0.802 8. Therefore, the performance of the improved model is better than the other three models.

Keywords: Bi-LSTM; CNN; Attention mechanism

0 引　　言

随着计算机与互联网的飞速发展，电商等网上销售模式应运而生。淘宝、京东等网上购物平台已经成为全民消费的主要场所，新浪微博已成为主要的信息交互和社交平台。海量的数据出现在微博、淘宝、京东等各种网站上，这些信息包含了发表者对评论对象的情感信息和主管观点，其他的用户也习惯从别人的评论信息里获取对自己有价值的信息来辅助自己做决策。因此线上评论的情感分析已经成为自然语言处理的重要研究方向之一[1]。

近年来深度学习技术被广泛的应用在自然语言处理领域。Kim[2]用卷积神经网络(CNN)对电影评论进行情感分类，Kalchbrenner 等[3]用卷积神经网络解决了 Twitter 的极性判别问题，陈珂等[4]用多通道的卷积神经网络对文本进行情感分类，这类基于深度学习技术的算法模型在情感分析领域取得了比基于情感词典进行情感分类的传统分类器更好的效果。

CNN 和 LSTM 在针对文本进行情感分析时有不同的特点，CNN 捕获空间特征的能力较强，但是其输入间是完全独立的，在学习句子相关性上有所欠缺，容易丢失文本的重要信息；LSTM 可以处理句子间的依赖关系，但是训练所需时间较长，容易产生梯度消失问题。

收稿日期：2020-12-20
* 基金项目：中央高校基本科研业务费(CCNU20TS010)资助

针对上述问题，本文提出了改进后 Bi-LSTMA-CNNA 模型。利用 CNN 和 LSTM 的优点，将其分别与注意力机制模型相结合，CNNA 区域主要用于局部特征的提取，捕捉重要的视角词，通过注意力机制调整视角词的比重。Bi-LSTMA 区域主要提取句子的全局语义特征并提高情感词的权重。CNNA 区域与 Bi-LSTMA 区域相结合达到句子重要特征的全覆盖，从而提高模型的性能。在从慕课网爬取的线上教学评论数据集上，将 CNN 模型、Bi-LSTM 模型、Bi-LSTM＋Attention 模型与改进后的 Bi-LSTMA-CNNA 模型相比较，Bi-LSTM＋Attention 模型取得了更好地分类效果，从而验证了改进后的模型在性能上优于其他 3 个。

1　相关介绍

1.1　CNN

卷积神经网络（convolutional neural networks，CNN）主要应用在图像处理领域[5]，通过对图像进行卷积运算，提取图像中的深层次特征[6]。近年来 CNN 也常用于文本分类中，其处理过程与图像处理十分类似。训练过程可以分为以下 4 步：

1）输入处理。主要目的是建立词向量；在数据集语料的基础上利用 Word2vec 神经网络语言模型将词汇训练成计算机可以理解的低维度的稠密向量[7]，如果一个句子是由 n 个词组成，每个词转换为维度为 k 的词向量，那么 CNN 网络的输入就是一个 n 行 k 列的二维矩阵[8]。

2）卷积层。卷积层的作用是用来提取句子的特征；通过不同大小的卷积核在从上到下的滑动进行卷积操作。

3）池化层。池化层主要是将卷积后的结果映射到同一维度，实质是对特征做进一步提取，将最重要的特征选取出来[9]。常采用最大池化法，取出特征映射中的最大值作为最重要的特征[10]。

4）全连接层和 Softmax 层。主要作用是对池化后的结果进行非线性转换，然后进行分类操作，最后输出句子的情感极性。

1.2　Bi-LSTM

循环神经网络（recurrent neural network，RNN）是一种可以处理序列关系的特殊神经网络。传统的神经网络输入层其各个输入神经元之间是相互独立的，但文本是序列型的数据，这些序列型的数据往往都具有时序上的关联性[11]。因此对文本进行分类时各个输入之间不能是相互独立的。RNN 可以对序列中的每个元素执行相同的处理，并结合 RNN 处理前面序列得到的结果进行输出。但在实际运用过程中 RNN 会对比较久以前的数据不敏感，会丢失之前的信息，出现梯度丢失的问题。为解决这个问题，学者们提出了长短时记忆网络（long short-term memory，LSTM）[12]。LSTM 网络单元结构如图 1 所示，LSTM 为解决 RNN 存在的问题，使用输入门、遗忘门、输出门 3 种门来保持和控制信息。输入门用于控制网络当前输入数据流入记忆单元的是多少，即有多少输入信息可以被保留下来；当输入新的信息时，模型若需要遗忘旧的信息，此时就需要遗忘门来完成，来控制哪些信息要保留，哪些信息要遗忘；输出门用于反映当前时刻信息被输出的程度。具体计算公式如下：

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i\boldsymbol{x}_t + \boldsymbol{U}_i\boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \tag{1}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f\boldsymbol{x}_t + \boldsymbol{U}_f\boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \tag{2}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o\boldsymbol{x}_t + \boldsymbol{U}_o\boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \tag{3}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tanh(\boldsymbol{W}_c\boldsymbol{x}_t + \boldsymbol{U}_c\boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \tag{4}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \tag{5}$$

式中：$\boldsymbol{x}_t$ 为输入矩阵；$\boldsymbol{h}_{t-1}$ 为上个时间的隐藏单元向量；$\boldsymbol{W}_i$、$\boldsymbol{W}_f$、$\boldsymbol{W}_o$ 和 $\boldsymbol{W}_c$ 为输入的词向量 $\boldsymbol{x}_t$ 对应的系数矩阵；$\boldsymbol{U}_i$、$\boldsymbol{U}_f$、$\boldsymbol{U}_o$ 和 $\boldsymbol{U}_c$ 为隐藏状态；$\boldsymbol{h}_{t-1}$ 为对应的系数矩阵；$\boldsymbol{b}_i$、$\boldsymbol{b}_f$、$\boldsymbol{b}_o$ 和 $\boldsymbol{b}_c$ 为偏置向量；σ 是激活函数；$\boldsymbol{h}_t$ 为 t 时刻的输出[13]。

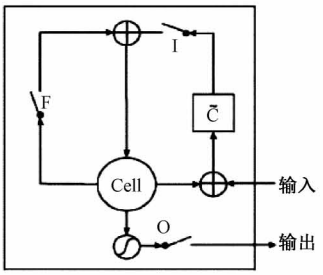

图 1　LSTM 网络单元结构

LSTM 通过训练可以学到记忆哪些信息，遗忘哪些信息，能够更好的捕捉到较长距离的依赖关系。但 LSTM 只能处理单向的时间序列，为此有些学者提出了双向长短期记忆神经网络[14]（Bi-LSTM）。

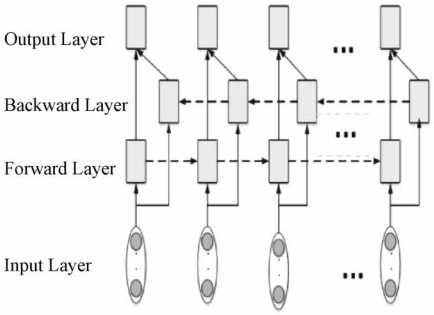

图 2　Bi-LSTM 网络单元结构

Bi-LSTM 结构中 Forward 层从时刻 1 到时刻 t 正向计算一遍，得到并保存每个时刻向前隐含层的输出。Bi-LSTM 网络单元结构如图 2 所示，Backward 层从时刻 t 到时刻 1 反向计算一遍，得到并保存每个时刻向后隐藏层的输出。最后在每个时刻结合 Forward 层和 Backward 层相应时刻的输出结果得到最终的输出[15]。因此 Bi-LSTM 能够很好的捕捉双向语义依赖。

1.3　注意力机制

注意力机制是借鉴了人本身的选择性注意力机制，即人通常会关注比较重要的特征[16]。近年来注意力机制被

广泛的应用在自然语言处理领域，注意力机制可以选择重要的输入数据生成输出数据；在情感分析中每个词语在句子中的重要性都不相同，利用注意力机制可以让传统模型更多关注句子中有情感色彩的词语，从而使模型达到很好的分类效果。

2　Bi-LSTMA-CNNA 模型

本节主要是对提出的 Bi-LSTMA-CNNA 模型进行详细的分析。Bi-LSTMA-CNNA 模型的结构如图 3 所示。

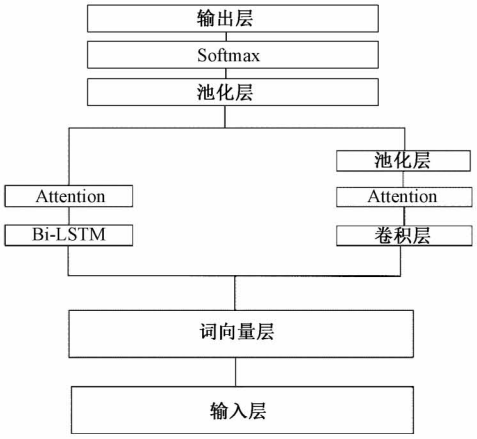

图 3　Bi-LSTMA-CNNA 模型结构

1）输入层：采用的数据集是基于线上课程的用户评论，在将它输入模型前须对整理好的数据集做一些预处理。首先把每个用户的评论分行放置，一行代表一个完整的评论，即每一行作为一个样本的输入；然后数据集分为训练集和测试集分开保存；最后利用 jieba 分词器进行分词，词与词之间用空格隔开。

2）词向量层：目的是使用密集向量来表示词汇和文档的一类方法。常用的是 Google 于 2013 年提出的 Word2Vec 来进行词向量的训练，将句子、词语转换成词向量格式[17]。Word2Vec 是浅层神经网络模型，有两种创建词向量的方法：CBOW 和 Skip-gram。CBOW 模型是将一个词的上下文作为输入来计算这个词出现的概率。Skip-gram 模型利用的原理与 CBOW 相反，是将目标词作为输入来计算前后出现几个词的概率[18]。本文模型用的是 Word2Vec 的 Skip-gram 模型训练数据。将输入层的数据通过 Word2Vec 模型训练，得到相应的词向量。

3）Bi-LSTMA 区域：是由 Bi-LSTM 模型和注意力机制模型组成。首先通过双向 LSTM 网络提取全局特征，再经过注意力机制模型重新给特征加权。

4）CNNA 区域：是由 CNN 模型和注意力机制模型组成。Attention 模型主要放在池化之前，词向量通过卷积层提取出特征后送入注意力机制模型重新加权，再进行池化降低维度，提取主要特征。

5）输出层：主要是将 Bi-LSTM 区域提取出的特征和 CNNA 区域提取出的特征进行融合后；然后进行最大池化操作，对关键信息进行获取，对不重要信息做出舍弃，得到特征向量；最后将特征向量输入到 Softmax 分类器得到分类结果。

3　实验结果与分析

3.1　数据集

本文实验中的数据集来自慕课网上的课程评论。利用爬虫技术从慕课网上爬取了南京师范大学形势政策课、北京大学思想道德基础与法律修养等 10 门课程，共 11 824 条数据。整个数据分为训练集和测试集两大类，主要用来训练和测试模型；其中训练集有数据 10 510 条，测试数据 1 314 条。并为每条数据设置了一个情感标签，情感标签分为积极的、中立的、消极的 3 类。

通过爬虫技术爬取的数据比较凌乱，会有一些特殊字符，如图 4 所示。在使用该数据前，需要对异常数据进行处理，保证数据的标准化、规范化。对数据的处理主要采用以下 3 个方法：1）对重复的汉字、重复的字母进行剔除；2）删除数据中的非文本信息；3）利用 jieba 分词器对数据进行切分和去除停用词。处理后的数据如图 5 所示。

课程1：形势与政策_南京师范大学_中国大学MOOC_慕课		
看完后，我的心久久不能平静!构思新颖，从	xzit20181510141	发表于 2019-01-06
我认为在高校开设《形势与政策》非常重要，这是对我们大学生进行形势政策教育	NJMU18601312	发表于 2018-05-26
要是没那么突然截止就好了	NJNU21160836	发表于 2019-01-08
本课程让我更深刻、更全面、更真实的了解到国内外现在面临的形势和施行的政	jsnu18090142	发表于 2018-11-04
本身对国家形势与政策这一块是非常迷茫的	NJMU18600614	发表于 2018-06-01
错过时间了	NJNU21161023	发表于 2018-11-13
《形势与政策》是德育课程中的重要组成部	njmu18605046	发表于 2018-06-25
给个证书吧	mooc1526830699600	发表于 2018-06-27
课程全看完不显示没学习通好	善良巨帅的吴健	发表于 2020-03-03
所以说，我的证书怎么还么下来？	月mooc289	发表于 2018-06-30
政治是门博大精深的学问,本身对国家形势	njmu18600291	发表于 2018-06-08
讲的真的非常好。系统，形象，结合国内国际，历史未来，使我们能对祖国的发展	mooc1525743915067	发表于 2018-06-18

图 4　处理前的数据

学到很多知识　2
语速有点慢　1
老师们都很棒　0
我学到了很多关于国际的形势，对国家事务有了更多的认知　2
生活就是心理学很好研究好生活研究好心理　0
课程很好很棒　0
我觉得非常好　0
非常不错啦　0
课程内容很丰富，比想象的更好　0
很好很好很好很好很好很好很好很好好好好　0
学到很多心理学知识，相信对我自己会有所帮助　2
老师讲的很好，事例都是通俗易懂的　0
课程授课方式很好，老师讲得很生动形象。通过这门课，我对当前世界的政治局势有了更深的了解。　0

图 5　预处理后的数据

3.2　实验参数设置

本文提出了 4 个模型，分别是 CNN 模型、Bi-LSTM 模型，Bi-LSTM＋Attention 模型、Bi-LSTMA-CNNA 模型。将整理好的数据集分别输入 4 个模型，验证模型的输入效

果。每个模型的具体参数如表 1 所示。

表 1　模型参数

名称	Dropout	Epoch	Batch	Pad	Learning	Hidden
CNN	0.5	20	128	32	0.001	0
BLSTM	0.5	20	128	32	0.001	128
BLSTMA	0.5	20	256	32	0.001	128
BLSTMA-CNNA	0.5	20	256	32	0.001	128

3.3　实验结果与分析

如表 2 所示模型的性能指标主要给出召回率(Recall)、准确率(Precision)、F1。Recall 是用来衡量类别的查全率，Precision 是用来衡量类别的查准率，F1 是查全率和查准率的综合，以及对它们的偏向程度。因此将 F1 作为模型性能评价的主要指标。在同一数据集以 F1 作为衡量标准的情况下，4 个模型的性能排序如下：Bi-LSTMA-CNNA＞Bi-LSTM＋Attention＞Bi-LSTM＞CNN。综上可得改进后的 Bi-LSTMA＋CNNA 模型在一定程度上提高了模型的性能。

表 2　4 个模型在数据集上的实验结果

模型	准确率	召回率	F1	Support
CNN	0.762 1	0.843 2	0.793 9	1 314
BLSTM	0.763 0	0.834 6	0.795 9	1 314
BLSTMA	0.760 7	0.844 0	0.799 8	1 314
BLSTMA-CNNA	0.763 6	0.847 0	0.802 8	1 314

4　结　　论

本文主要是对文本的情感分析进行了研究，基于神经网络模型在分类任务上取得了较好的结果，Bi-LSTM 能够很好的考虑到文本上下文间的联系；本文在卷积神经网络和双向长短时记忆网络的基础上提出一种改进后的新模型，因为文本中每个词语的重要程度不同，改进的模型引入了注意力机制模型。实验结果表明，改进后的 Bi-LSTMA-CNNA 模型在一定程度上性能优于 CNN 模型、Bi-LSTM 模型、Bi-LSTM ＋ Attention 模型。本文提出的 Bi-LSTMA-CNNA 模型虽然在线上教学评论的数据集下取得不错的结果，但是模型的适应能力还需要验证，在接下还需要将模型应用在不同的数据集上。

参考文献

[1]　郭豪. 基于微博旅游信息的中文关键词提取与分析研究[D]. 天津：天津大学，2018.

[2]　KIM Y. Convolutional neural networks for sentence classification[C]. Proceedings of the 2014 Conferrence on Empirical Methods in Natural Language Processing, Stroudsburg, PA: ACL, 2014: 1746-1751.

[3]　KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A convolu-tional neural network for modelling sentences[C]. Pr-oc of the 52nd Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA: ACL, 2014: 655-665.

[4]　陈珂，梁斌，柯文德，等. 基于多通道卷积神经网络的中文微博情感分析[J]. 计算机研究与发展，2018，55(5)：945-957.

[5]　曲景影，孙显，高鑫. 基于 CNN 模型的高分辨率遥感图像目标识别[J]. 国外电子测量技术，2016，35(8)：45-50.

[6]　张珊，于留宝，胡长军. 基于表情图片与情感词的中文微博情感分析[J]. 计算机科学，2012(S3)：146-148.

[7]　樊振，过弋，张振豪，等. 基于词典和弱标注信息的电影评论情感分析[J]. 计算机应用，2018，38(11)：3084-3088.

[8]　段宇翔，张仰森，张益兴，等. 基于 LSTM-CNNS 情感增强模型的微博情感分类方法[J]. 北京信息科技大学学报(自然科学版)，2019，34(6)：1-7.

[9]　祝元勃. 基于深度学习与自注意力机制的情感分类方法研究[D]. 西安：西安理工大学，2019.

[10]　朱军，刘嘉勇，张腾飞，等. 基于情感词典和集成学习的情感极性分类方法[J]. 计算机应用，2018，38(S1)：95-98.

[11]　景春臻. 基于深度学习的情感分类系统的研究与实现[D]. 北京：北京邮电大学，2019.

[12]　汪贝贝. 基于 Seq2Seq 模型的自动文本摘要技术研究与实现[D]. 武汉：华中科技大学，2019.

[13]　蔡道悦，刘斌. 基于区域 BLSTM 和 CNN 的情感分析模型[J]. 计算机工程与设计，2019，40(8)：2361-2365，2395.

[14]　周瑛，刘越，蔡俊. 基于注意力机制的微博情感分析[J]. 情报理论与实践，2018，41(3)：89-94.

[15]　刘路. 基于序列到序列模型的答案生成方法研究[D]. 武汉：武汉科技大学，2019.

[16]　陈巧红，于泽源，孙麒，等. 基于注意力机制与 LSTM 的语音情绪识别[J]. 浙江理工大学学报(自然科学版)，2020，43(6)：815-822.

[17]　黄欣欣，年梅，胡创业，等. 基于卷积神经网络的虚假评论检测[J]. 计算机时代，2019(11)：41-45.

[18]　陈炳丰，郝志峰，蔡瑞初，等. 基于 AWCRF 模型的微博情感倾向分类方法[J]. 计算机工程，2017，43(7)：187-192.

作者简介

占妮，硕士研究生，主要研究方向为自然语言处理之情感分析。

E-mail：277078213@qq.com