Word2vec-CNN-Bilstm 短文本情感分类

王立荣

(福建省泉州华侨职业中专学校 福建 泉州 362000)

摘 要 使用传统的神经网络的短文本分类算法对其进行情感分类易出现定位误差等问题。为了解决对短文本情感分类时存在的定位误差，本文通过将词向量模型（Word2vec）、双向长短时记忆网络模型（BiLSTM）以及卷积神经网络（CNN）按照一定的框架进行组合，提出了 Word2vec-CNN-BiLSTM 的短文本情感分类模型。Word2vec-CNN-BiLSTM 模型采用对预处理后的文本进行向量化表示来提取文章特征向量，并在神经网络层进行双向语义捕捉实现文本的情感分类。实验结果显示 Word2vec-CNN-BiLSTM 的短文本情感分类模型有效解决了对短文本分类出现的情感分类定位误差问题。

关键词 神经网络；情感分类；词向量；短文本

中图法分类号 TP181 TP391.1 DOI:10.16707/j.cnki.fjpc.2020.01.003

Word2vec-CNN-Bilstm Based Short-text Sentiment Classification

WANG Lirong

(Fujian Province Quanzhou Huaqiao Secondary Vocational School, Quanzhou, China, 362000)

Abstract Traditional neural network-based short-text classification algorithms for sentiment classification is prone to positioning errors. In order to solve this problem, the Word Vector Model (Word2vec), Bidirectional Long-term and Short-term Memory networks(BiLSTM) and convolutional neural network (CNN) are combined according to a certain framework, and a short-text sentiment classification model of Word2vec-CNN-BiLSTM is proposed. The Word2vec-CNN-BiLSTM model uses the vectorized representation of the pre-processed text to extract the feature vector of the article and performs bidirectional semantic capture at the neural network layer to achieve sentiment classification of the text. The experimental results show that the short-text sentiment classification model based on the Word2vec-CNN-BiLSTM effectively solves the positioning error of sentiment classification for short-text classification.

Keywords Neural Networks; Sentiment Classification; Word Vectors; Short-text

1 引言

随着互联网的普及，人们使用互联网的频率越来越高，产生了大量的用户信息。信息以文本、图片、视频等多种方式进行广泛传播。在众多的传播方式中，文本仍旧是人们产生和获取信息的重要方式，且这些文本通常具有简短的特点。将短文本进行情感分类有利于对短文本信息进行精准定位分类，更好地完成用户推送服务，提高用户体验，方便管理。近几年来，文本情感倾向性分析的研究开始得到诸多学者的关注，逐渐成为国内外研究的热点。随着神经网络的兴起，其相关算法在文本分类中展现出了较高的分类效果。国内外也有很多关于文本的情感分类的研究，文本情感检测和分类通过采用心理模型将单词、句子和文档映射到一组情感。该任务已经从纯粹以研究为导向的主题逐渐发展为在各种应用程序中扮演的角色，包括对话系统、智能代理、精神障碍的临床诊断或社交媒体挖掘等。由于应用程序种类繁多，因此领域和文本差异的集合也很大。现在实验的对象开始应用于网络

王立荣，男，1978年生，主要研究领域为软件开发。E-mail:94450634@qq.com。

上的一些评论类短文本，例如商品评价和微博评论等。针对这些短文本，也相继涌现了诸多的研究，王丽亚等将注意力模型引入到神经网络中增强神经网络[1]，计算每个时序的权重；崔争艳对传统机器学习的算法在文本分类上的分类效果进行了对比和研究[2]；孙学琛等将半监督学习的思想引入到文本分类中[3]，大大提高了分类的效果。

在本文中，考虑短文本信息前后的依赖关系，在传统的递归神经网络模型和 LSTM 模型中，信息只能向前传播，导致当前时刻的状态仅取决于该时刻之前的信息。为了使每时每刻都包含上下文信息，采用结合了双向递归神经网络（BiRNN）模型和 LSTM 单元的 BiLSTM 来捕获上下文信息。对于词向量处理采用 Word2Vec 为相似单词实现相似向量。Word2Vec 将彼此相关的单词映射到在高维空间中彼此靠近的点，通过学习相邻词的密集向量来预测单词，在共享相似上下文的词的同时也共享语义。本文通过组合 CNN-Word2vec 模型结构和 BiLSTM 模型结构，完成文本情感分类任务。利用情感分析结果作为样本，实现模型参数的训练。并在多个数据集上实验 Word2vec-CNN-BiLSTM 验证其有效性，实现文本的情感分类准确率提高。

2 相关理论

2.1 文本的情感分类

情感分类任务主要有两种类型：二元分类（情感极性的粗粒度分类）和多类别分类（细粒度分类多个类别）。先前的大多数研究工作都集中在二元分类上，即正面和负面。但是，显示更多详细信息的多类分类系统通常具有更多的实际意义。例如，如果已知用户的特定情绪状态，则将更准确地、更少烦人地推送商业广告。进一步了解用户的当前感受也将有助于社交网络网站营造一种更加温暖和友好的氛围。因此本文采用三分类标准，即正面、中立以及反面。

文本中的情感处理的研究和应用还处于非常初级的阶段。自然语言固有的含糊性和微妙性是使这项任务变得非常具有挑战性的众多因素中的一些，尤其是在社交网络环境中，短文本通常是不完整或不连贯的。对于传统的情感分类和多标签情感分类，研究人员已经提出了各种机器学习方法。现有的大多数系统将这个问题作为文本分类问题来解决。最近，深度学习模型已被用于文本的情感分类，系统能够更加精确地从原始数据中提取高级特征。

2.2 Word2vec词向量表示

Word2vec 模型是在 Log-Bilinea 及 NNLM 两个模型的基础上由 Tomas 等人开发的工具[4]。Word2vec 可以将词从高维空间分布式映射到低维空间且保留了词向量之间的位置关系，从而解决了向量稀疏和语义联系两个问题。Word2vec 分为 CBOW(continuous bag-of-words)和 Skip-gram 两种方式。本文主要是基于 Skip-gram 方法进行词向量处理，因为它可以在大型数据集上产生更准确的结果。

词向量表示是指在对词语按照其表达的意思进行分类时，对词语进行向量化处理。在现实中，词语所表达的意思通常是向多个方向发散的，因此需要将词语映射为到多维向量。这样做一方面解决了语意的多方向发散问题，另一方面多维向量能够使用较小的数字来表征词语。Word2vec 类似于自动编码器，它在向量中编码每个单词，但是 word2vec 不会像受限玻尔兹曼机那样通过重构来针对输入的单词进行训练，而是针对与输入语料库中与它们相邻的其他单词进行训练。

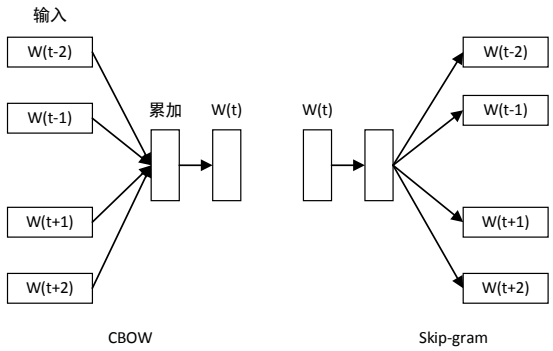

图 1 Word2vec 模型

2.3 CNN模型

CNN 是一种带有卷积结构的前馈神经网络，利用卷积神经网络进行局部特征向量提取。采用梯度下降法最小化损失函数对网络中的权重参数逐层反馈调节，通过迭代训练提高网络的精度。本文采用的卷积神经网络主要由输入层、卷积层、池化层、全连层以及输出层组成[5]。

卷积层：将通过嵌入层输出的每个句子矩阵进行卷积操作。

$$S = f(wZ + b)$$

其中 w 为权重矩阵，b 为偏置向量，Z 为词向量矩阵，S 为经过卷积操作后的特征矩阵。

池化层：对特征矩阵 S 进行下采样，求解局部值的最优解。在本文的实验中采用的是 MaxPooling 函数，表示如下：

$$M = \max(s_1, s_2, \cdots s_{n-g+1}) = \max\{S\}$$

其中 $s_i \in S, i = 1,2\cdots n-g+1$，$n$ 为词数，g 为卷积核的尺寸。

全连层：将池化后的 M_i 向量连接成向量 Q

$$Q = \{M_1, M_2, \cdots M_n\}$$

将连续高阶窗口 Q 作为 BiLSTM 的输入。

2.4 BiLSTM双向长短时记忆网络模型

LSTM 全称是 Long Short-Term Memory，它是 RNN（Recurrent Neural Network）的一种[6]。BiLSTM 是 LSTM 的改进，它将前向隐藏层与后向隐藏层相结合，可以系统地、有选择地使用之前和之后的信息。LSTM 由于其设计的特点，非常适合用于对时序数据建模，例如文本数据。BiLSTM 是 Bi-directional Long Short-Term Memory 的缩写，是由前向 LSTM 与后向 LSTM 组合而成。BiLSTM 在自然语言处理任务中常被用来处理上下文信息。通常，将词的表示组合成句子的表示或者相加的方法没有考虑到词语在句子中的前后顺序[7]。使用 LSTM 模型可以更好地捕捉到较长距离的依赖关系。因为 LSTM 通过训练过程可以学到记忆哪些信息和遗忘哪些信息。但是利用 LSTM 对句子进行处理存在无法编码从后到前的信息的问题，因此本文通过加入 BiLSTM 模型将更加精确[8]。

由图 2 可知，LSTM 网路层主要包括记忆单元 c、输入门 i、遗忘门 f 以及输出门 o。其中，遗忘门 f 和输入门 i 用来控制记忆单元的输入信息，决定之前时刻单元状态 c_{t-1} 和当前时刻网络输入 x_t 的保留比例。输出门 o 用来控制单元状态 c_t 输入到当前时刻输出值 h_t 的信息量。

具体过程如下：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_i)$$
$$\widetilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$
$$C_t = f_t \circ C_{t-1} + l_t \circ \widetilde{C}_t$$
$$o_t = \sigma(w_0[h_{t-1}, x_t] + b_0)$$
$$h_t = o_t \circ \tanh(c_t)$$

在本实验中，采用双向 LSTM 来捕捉文本信息中的前后文本信息关系，BiLSTM 神经网络最后输出 h_n 由 2 个单向、反向的 LSTM 输出结果拼接得到的值[9]，过程如下：

$$H_n = h' \oplus h$$

在上面的一系列公式中，σ 是逻辑 sigmoid 函数，$\circ$表示按元素相乘；W_i、W_f、W_o、W_c 是权重矩阵；b_i、b_f、b_o、b_c 是偏移量。除了隐藏单元 h_t 之外，LSTM 还包括输入门 i_t、忘记门 f_t、输出门 o_t、输入调制门和存储器单元 c_t。其网络层结构如图 2 所示[10]。

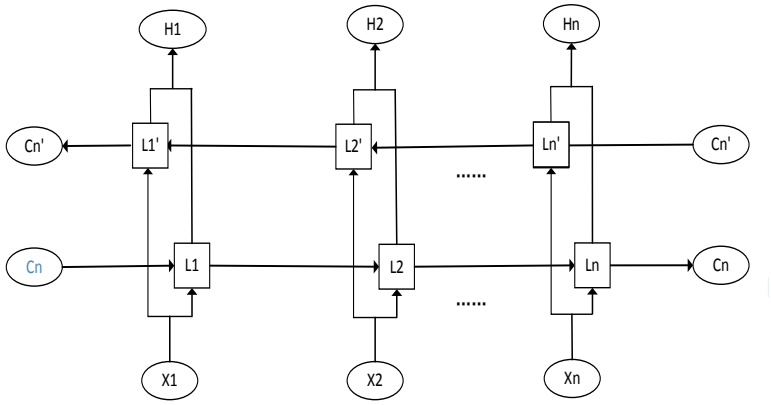

图 2 BiLSTM 模型

3 Word2vec-CNN-BiLSTM 组合模型

使用 Word2Vec 将文章进行分析存在断章取义的问题，使得对文章进行情感分类产生较大的误差。通过引入 BiLSTM 捕捉较长距离的双向依赖关系[11]。利用 BiLSTM 编码从后到前的信息使对文本的情感分类的准确度得到提高[12]。使用 Word2Vec 将文章进行分析的同时使用 BiLSTM 捕捉较长距离的双向依赖关系。利用 BiLSTM 编码从后到前的信息，在更细粒度的分类时，对于褒义、中性的贬义的三分类任务更加注意情感词、程度词、否定词之间的交互，进行统计从而分析出文本的情感倾向。

本文提出一种基于 Word2vec 的 BiLSTM-CNN 混合神经网络模型，通过采用 BiLSTM 与 CNN 构建混合神经网络模型，并利用该模型进行短文本的情感分类。所构建的组合模型主要由以下几个模块所组成：输入层、词嵌入层、CNN 层、双向 BiLSTM 层以及 Dense 层组成。模型主要分为以下几个步骤：

步骤 1：对短文本数据进行预处理，去掉短文本数据中的停用词、低频词。然后进行分词处理，本文实验中实验 jieba 分词进行中文分词。

步骤 2：用 Word2vec 对维基百科的训练资料进行训练，得到所需的词向量。将预处理完的短文本数据导入 Word2vec 中获取文本的词向量表示。

步骤 3：将生成的词索引通过词嵌入层转换成 CNN 卷积网络的输入，然后添加池化层结构，并将经过池化操作后的数据输入到 BiLSTM 网络层。

步骤 4：最后加入全连层和分类器。本实验将数据分成三类：正面（positive）、反面（negative）、中立（neutral），并计算评估指标。本实验从预测得分和预测的准确率来评估模型的分类效果。

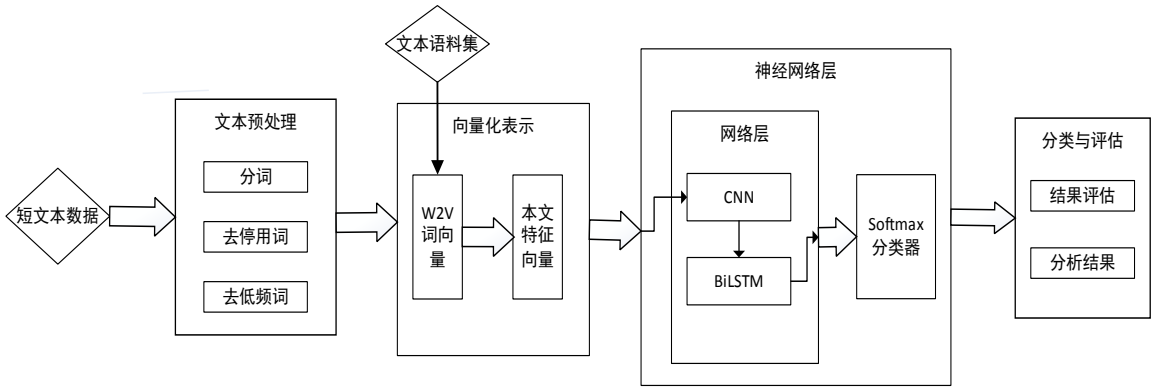

图 3 模型组合

4 实验分析

4.1 实验环境

本文实验环境如下：操作系统为 windows 10 ，CPUIntel Core i5-8250U ， GPU 为 GeForce GTX1050，内存大小为 8GB，开发环境为 tensorflow 1.14.0，开发工具使用的是 PyCharm。

4.2 实验数据

本文利用 16873 条关于当当网、酒店、电脑、蒙牛牛奶、手机的评价作为实验数据。经分析处理，数据集中涉及 3 种情感类型的分类，即正面、中立、反面三种情感类型，并按 8 ： 2 的比例划分为训练集和测试集。数据统计方法参考张俊飞等人提出的方法，具体数据如表 1 所示。

表 1 样本统计

数据集	样本类别	样本数量
训练集	positive	6443
	negative	3479
	neutral	6951
测试集	positive	1590
	negative	876
	neutral	1752

4.3 模型参数

实验参数的合理设置对实验结果有直接影响。参数调整过程中使用固定参数的方法，分别对 LSTM 层单元数、丢弃率等参数进行了对比实验。当准确率不再上升时停止训练，以避免过拟合、不

收敛等问题，同时可以加快学习速度，提高调参效率。BiLSTM 模型的 LSTM 层单元数参数的值为 64，丢弃率（dropout）的参数值为 0.2，批尺寸的参数值为 64，Epoch 的参数值为 10，优化器（optimizer）的参数值为 adam。

表 2 实验模型参数

参数	值
词向量维度	200
词向量训练模型	Skip-Gram
批尺寸	64
迭代次数（Epoch）	20
优化器（optimizer)	Adam
LSTM 层单元数	32
丢弃率（dropout）	0.5
卷积核大小	3
学习率	0.05

4.4 实验结果

实验是通过将利用 16873 条评价作为实验数据，数据集中涉及 3 种情感类型的分类。在进行数据统计分析时，分类任务的常用评价标准有准确率(precision)。本文还比较了不同模型下的损失值的变化情况，根据准确率的大小和损失值的波动进行模型对比和参数对比，最终确定模型在短文本分类的优越性以及最佳的参数设定。准确率的计算如下：

$$Acc = \frac{1}{N}\sum_{i=1}^{N}\left|y_i = \hat{y}_i\right|$$

其中，$\hat{y}_i$ 表示 x_i 的预测标签，y_i 表示 x_i 的实际标签，N 表示测试集的大小。

通过实验，对比在不同的迭代次数（epoch）下的实验结果，发现正面和负面的评价所占比例较高，中立的评价所占比例较低。实验数据分布如表 3 所示。

表 3 epoch 数值统计

epoch	损失值	准确率/%
10	0.29	90.2
20	0.37	91.8
30	0.41	90.9
40	0.45	90.9

对比表 3 可以发现：当 epoch 为 20 时，准确率最高，为 91.0%；当 epoch 的值继续增加时，损失值上升，准确率下降，模型出现过拟合。

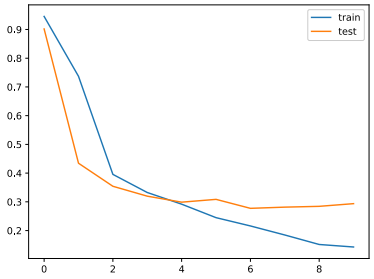

图 4 epoch=10

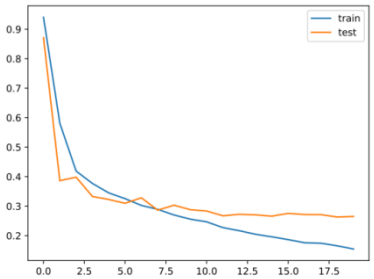

图 5 epoch=20

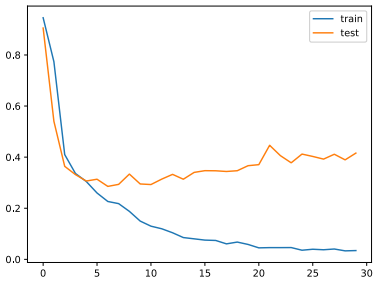

图 6 epoch=30

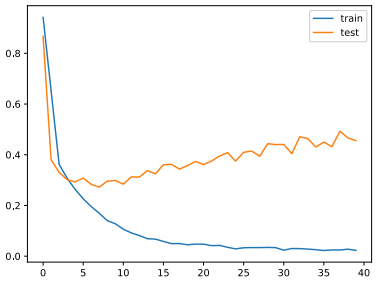

图 7 epoch=40

为寻找 Word2vec-CNN-BiLSTM 模型合适的迭代次数，本文在实验过程中分别将参数 epoch 设置为 10、20、30、40。在图 4 至图 7 系列中，图 4 为

在迭代次数 epoch 为 10 时损失函数的变化；图 5 为在迭代次数 epoch 为 20 时损失函数的变化；图 6 为在迭代次数 epoch 为 30 时损失函数的变化；图 7 为在迭代次数 epoch 为 40 时损失函数的变化；train 表示训练集；test 表示测试集。利用 python 绘图模块 matplotlib 绘制得到。对比四个损失函数变化趋势得到：当迭代次数为 10 时，损失值仍呈下降趋势；当迭代次数为 20 时，损失值较低，较为稳定；当迭代次数为 30、40 时，损失值又回升。这说明 epoch=20 时分类效果最优。

4.5 对比实验分析

为验证本文提出的 Word2vec-CNN-BiLSTM 组合模型的分类效果，分别将文本组合模型与单一模型进行实验对比。对比模型为 LSTM 模型、CNN 模型、TEXTCNN 模型、CNN_LSTM 模型。具体实验结果如表 4 所示。

表 4 各分类模型分类准确率

分类模型	损失值	预测准确率/%
LSTM	0.27	90.9
CNN_LSTM	0.33	91.5
CNN	0.42	91.3
TextCNN	0.53	88.9
Bi_LSTM	0.33	91.2
本文模型	0.27	91.8

（1）由表 4 可得本文提出的组合模型的准确率比传统单模型 LSTM、CNN、BiLSTM 模型更高，损失值更低。这证明相对于单一结构神经网络，混合网络模型的表现更好。理论上随着模型深度的增加，由于参数增加的模型表达能力也会更加优秀。针对在训练过程中出现的过拟合现象，本实验通过调整 Dropout 的大小和采用 L2 正则化处理的方法进行平衡。

（2）将本文 Word2vec-CNN-BiLSTM 模型与其他组合模型相比：相比于 CNN-LSTM 模型损失值减少 0.06，准确率提高约 0.3%；损失值比 TextCNN 模型减少 0.22，准确率提高约 2.9%。3 种模型的准确率从低到高依次排序为 TEXTCNN、CNN-LSTM、Word2vec-CNN-BiLSTM，证明本文 Word2vec-CNN-BiLSTM 模型分类效果较优。

5 结论与展望

本文采用基于 Word2Vec 的 CNN-BiLSTM 组合模型对短文本的情感分类问题进行实验研究。对比其他经典的深度学习的模型，组合模型展现了较高的分类效果。本文通过采用 Word2Vec，用高维向量表示词语，将相近意思的词语放在相近的位置。通过训练大量的的语料获得词向量，解决“一义多词”的问题。同时使用 BiLSTM 捕捉较长距离的双向依赖关系，利用 BiLSTM 编码从后到前的信息。实验中发现使用 Word2vec-CNN-BiLSTM 模型分析出文本的情感倾向将会变得更容易、更准确，大幅度提高短文本情感分类的准确率。

参 考 文 献

[1] 王丽亚,刘昌辉,蔡敦波,赵彤洲,王梦.基于CNN-BiLSTM网络引入注意力模型的文本情感分析.武汉工程大学学报,2019,41(04):386-391

[2] 崔争艳.中文短文本分类的相关技术研究[硕士学位论文]. 河南大学, 开封,2011

[3] 孙学琛,高志强,全志斌,等.基于半监督学习的短文本分类方法.山东理工大学学报（自然科学版）, 2012, 26(1):1-4

[4] 张谦,高章敏,刘嘉勇.基于Word2vec的微博短文本分类研究.信息网络安全, 2017(1):57-62

[5] 罗帆,王厚峰.结合RNN和CNN层次化网络的中文文本情感分类.北京大学学报(自然科学版),2018,54(03):459-465

[6] 白静,李霏,姬东鸿.基于注意力的BiLSTM-CNN中文微博立场检测模型.计算机应用与软件, 2018(3):266-274

[7] 沈竞.基于信息增益的LDA模型的短文本分类.重庆高教研究, 2011, 30(6):64-66

[8] 李洋,董红斌.基于CNN和BiLSTM网络特征融合的文本情感分析.计算机应用, 2018, 38(11):29-34

[9] 和志强,杨建,罗长玲.基于BiLSTM神经网络的特征融合短文本分类算法.智能计算机与应用, 2019, 9(02):29-35

[10] 张俊飞,毕志升,吴小玲.基于词向量Doc2vec的双向LSTM情感分析.计算机与数字工程, 2018, 46(12):10-14,24

[11] 龚琴,雷曼,王纪超,王保群.基于注意力机制的卷积-双向长短期记忆模型跨领域情感分类方法.计算机应用,2019,39(08):2186-2191

[12] 和志强,杨建,罗长玲.基于BiLSTM神经网络的特征融合短文本分类算法.智能计算机与应用, 2019, 9(02):29-35