计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：融合语法规则的双通道中文情感模型分析
作者：邱宁佳，王晓霞，王鹏，王艳春
收稿日期：2020-05-29
网络首发日期：2020-09-21
引用格式：邱宁佳，王晓霞，王鹏，王艳春．融合语法规则的双通道中文情感模型分析[J/OL]．计算机应用．
https://kns.cnki.net/kcms/detail/51.1307.TP.20200918.1728.004.html

Journal of Computer Applications ISSN 1001-9081
计算机应用 CODEN JYIIDU http://www.joca.cn

文章编号:2020050723 doi:10.11772/j.issn.1001-9081.2020050723

融合语法规则的双通道中文情感模型分析

邱宁佳，王晓霞，王鹏*，王艳春

（长春理工大学 计算机科学技术学院,长春 130022）

（*通信作者电子邮箱 wpeng@cust.edu.cn）

摘 要: 针对使用中文文本进行情感分析时，忽略语法规则关系，会降低分类准确率的问题。提出一种混合语法规则的双通道中文情感分类模型 CB_Rule（Grammar Rules of CNN and Bi-LSTM）。首先设计语法规则提取出情感倾向更加明确的信息，再利用卷积神经网络（CNN）的局部感知特点提取出语义特征；然后考虑到规则处理时可能忽略上下文的问题，使用双向长短时记忆网络（Bi-LSTM）提取包含上下文信息的全局特征，并对局部特征进行融合补充，完善 CNN 模型情感特征倾向信息；最后将完善后的特征输入到分类器中进行情感倾向判定，完成中文情感模型的构建。在中文电商评论文本数据集上将所提模型与 R-Bi-LSTM，SCNN(syntactic rules for convolutional neural network)模型进行对比，实验结果表明，所提模型在准确率上分别提高了 3.89%、0.63%，表明了该 CB_Rule 模型具有很好的分类效果。

关键词: 情感分析；语法规则；特征融合；卷积神经网络；双向长短时记忆网络

中图分类号:TP391 **文献标志码**：**A**

Analysis of double-channel Chinese emotion model integrating grammar rules

QIU Ningjia, WANG Xiaoxia, WANG Peng*, WANG Yanchun

(School of Computer Science and Technology, Changchun University of Science and Technology, Changchun 130022, China)

Abstract: Concern the problem that ignoring the relationship between grammatical rules will reduce the accuracy of classification when using Chinese text for sentiment analysis, CB_Rule (Grammar Rules of CNN and Bi-LSTM) was proposed, which is a method double-channel Chinese emotion classification model based on mixed grammar rules. First, Grammar rules are designed to extract more explicit information about emotional tendencies, and then the semantic features are extracted by using the local perceptual features of the convolutional neural network. After that, considering the problem of ignoring the context when processing rules, Bi-directional short-term memory network (Bi-LSTM) is used to extract global features containing contextual information, and the local features are fused and supplemented, so that the CNN model emotional feature orientation information is improved. Finally, the improved features are input into the classifier to judge the affective tendency, and the Chinese affective model is constructed. Compare the proposed model with R-Bi-LSTM, SCNN (syntactic rules for convolutional neural network) model on the Chinese e-commerce review text data set, the experimental results show that the accuracy of the proposed model is increased by 3.89% and 0.63% respectively, indicating this the CB_Rule model has a good classification effect.

Keywords: sentiment analysis; grammar rule; feature fusion; Convolutional Neural Network(CNN); Bidirectional Long Short-Term Memory(Bi-LSTM)

0 引言

近年来对文本进行情感分析成为了自然语言处理领域的重要分支，进行有效的情感分析能够帮助用户及时掌握所在领域的情绪动态。传统的文本情感分类方法主要为基于情感词典与基于机器学习的方法。在基于情感词典的研究方法上，Araque 等人[1]使用语义相似性度量与嵌入式表示结合使用，该模型表明了词汇的选择对跨数据集性能有影响。Zhang 等人[2] 提出了一种基于情感词典的方法，解决了中文文本情感分析问题。Xu 等人[3]提出的基于扩展情感词典方法对评论文本的情感识别具有一定的可行性和准确性。此外，对于情感词典跨数据集的适用性问题，Hung[4]提出了一种新颖的基于上下文词汇概念质量和上下文词典质量口碑质量分类模型，

收稿日期: 2020-05-29; 修回日期: 2020-08-17; 录用日期: 2020-08-24。

基金项目：吉林省科技发展计划技术攻关项目(20190302118GX)

作者简介: 邱宁佳(1984-)，男，河南南阳人，讲师，博士，CCF 会员，主要研究方向为数据挖掘、算法分析、机器学习、自然语言处理；王晓霞(1996-)，女，贵州遵义人，硕士研究生，主要研究方向为数据挖掘、机器学习、自然语言处理；王鹏(1973-)，男，内蒙古包头人，教授，博士，CCF 会员，主要研究方向为数据挖掘；王艳春（1964-），女，黑龙江鸡西人，副教授，硕士，主要研究方向为智能计算、数据挖掘。

并在 IMDB 和 hotels.com 数据集上验证了该模型的显著性。Khoo 等人[5]也提出了新的通用情感词典 Wee Kim Wee School of Communication and Information (WKWSCI)，将其与常用的五种情感词典进行比较后也取得了不错的分类成绩。在基于机器学习的研究方法上， Singh 等人[6]利用了朴素贝叶斯, J48, BFTree 和 One Rule(OneR)四种机器学习分类器对 IMDB 电影评论数据集进行了实验，对比分析了四种分类器的各自性能。Anggita[7]使用粒子群优化算法(particle swarm optimization, PSO)优化了朴素贝叶斯和支持向量机（support vector machines, SVM），提高了原算法的分类精度。对产品评论进行情感分类时, Tama V O 等人[8]采用了朴素贝叶斯算法得到了 80.48%的分类准确性。但基于情感词典的分类过分依赖于构建的情感词典，通用性不强，基于机器学习的方法通常需依赖复杂的特征过程，且人工标注成本较高。

随着深度学习在不同情感分析领域取得了优异成绩，现已成为文本情感分析的主流技术。陈珂等人[9]利用多通道卷积神经网络（Multi-Channels Convolutional Neural Networks, MCCNN）模型使其从多方面的特征表示学习输入句子的情感信息。Long 等人[10]将双向长短时记忆网络(Bidirectional Long Short-Term Memory, Bi-LSTM)与多头注意力机制相结合对社交媒体文本进行情感分析，克服了传统机器学习中的不足。Kai[11]、李洋[12]等人、赵宏等人[13]将卷积神经网络（Convolutional Neural Network，CNN）与 Bi-LSTM 融合起来，解决了现有情感分析方法特征提取不充分的问题，并分别通过实验表明了该融合模型在实际应用中具有较大的价值。同时，Wang 等人[14]研究了树形结构的区域 CNN-BiLSTM 模型，提供了更细粒度的情感分析,在不同语料库上都取得了不错的分类效果。同时，为了充分发挥语法规则在中文文本中的重要性，学者们尝试将其融入到神经网络中，卢强等人[15]将语法规则与 Bi-LSTM 相融合，何雪琴等人[16]则与 CNN 相融合，通过设置对比试验，各自都在不同数据集上取得了更高的分类效果。

针对上述研究现状，本文融合语法规则构建双通道中文情感模型，首先设计语法规则对文本进行预处理，以保留情感倾向更明显的文本，然后使用 CNN 的强语义特征提取能力在不同窗口大小得到粒度不同的局部情感特征，同时为了弥补语法规则处理时可能忽视上下文信息问题的不足，利用 Bi-LSTM 挖掘到文本时间跨度更大时的语义依赖关系，获取到包含上下文信息的全局特征，最后将融合后的局部特征与全局特征使用分类器对文本进行情感分类。

1 中文语法规则的构建

为了解决因中文文本语义多样化而导致 CNN 情感特征提取困难的问题，本文考虑首先设计语法规则文本进行初步情感信息清洗，降低文本语义复杂性，从而获取到情感倾向更加明确的文本信息，再使用 Word2Vec 模型进行训练得到规则特征向量后作为 CNN 的输入。通过中文文本语法规则研究发现，文本中的情感倾向词所在句直接表达了作者正面或负面情感，总结词则表明了文本的中心思想，直接影响了句子的情感倾向，而转折词则实现前后情感反转的作用，其中转折词分为两类,甲类转折词所在句带有明显的情感倾向，乙类转折词则起到过渡作用，所在句的内容不能够表达文本的情感倾向，其情感倾向常表现在余下语句中。故为了充分发挥情感倾向词、总结词、转折词在文本情感倾向信息提取中的作用，本文将依据数据集进行这三类词的提炼汇总，设计出三类情感词典：EmoTendencyWords 情感倾向词词典、SumWords 总结词词典、TurnWords 转折词词典，然后根据这三类词典对中文文本进行语法规则设定，以获取情感倾向更加明确的信息，以方便 CNN 在训练时情感倾向特征的获取。规则设定如下，其中，W 表示整个评论文本，W_i 表示文本中的各个分句，定义该评论文本的分句集合为 $\{W_1, W_2, \mathbf{L}, W_n\}$，$W$¢则表示经语法规则处理后的文本。

规则 1：若评论文本W 中通过匹配 EmoTendencyWords 情感倾向词词典，存在情感倾向词，则直接提取情感倾向词所在的分句W_i，然后根据情感倾向词词典直接判定评论文本W 的正负面；

其中，当文本中出现多个情感倾向词时，参照文献[16]提出的“主题词+直接分类法”进行该文本的情感倾向判定，通过主题词判定该情感倾向词是否有效，若无效则舍弃。具体方法为：首先根据数据集设定好种子主题词，利用 Word2vec 工具文本将文本转换为词向量表示 $w_i = \{s_{i1}, s_{i2}, \mathbf{L}, s_{ik}\}$，然后计算词向量之间的欧氏距离来判断该分句与种子主题词之间相似度，阈值范围以内则为相关主题，表示该情感倾向词有效。最后统计有效正负面情感倾向词个数并比较，正面个数多则该文本W 情感倾向为积极，反之则为消极；相似度计算公式如式（1）所示。

$$sim(w_1, w_2) = \sqrt{\sum_{j=1,\mathbf{L},k} (s_{1j} - s_{2j})^2} \qquad (1)$$

规则 2：若评论文本W 中无情感倾向词，则与 SumWords 总结词词典进行匹配，若存在某总结词，则直接提取总结词后的分句W_i。若文本中出现多个总结词，为提高分类效率，默认只提取第一个总结词以后的分句W_i；

规则 3：若评论文本W 中无情感倾向词与直接分类词，则与 TurnWords 转折词词典进行匹配，若存在甲类转折词，则直接提取该转折词之后的所有分句$\{W_i, W_{i+1}, \mathbf{L}, W_n\}$；若存在乙类转折词，则忽略该转折词所在分句$W_i$，提取评论其他内容$\{W_1, W_2, \mathbf{L} W_{i-1}, W_{i+1}, \mathbf{L}, W_n\}$。

规则 4：若评论文本W 均不属于上述三种情况，则直接保留原文本内容W 。

利用语法规则提取情感倾向语句流程如图 1 所示。

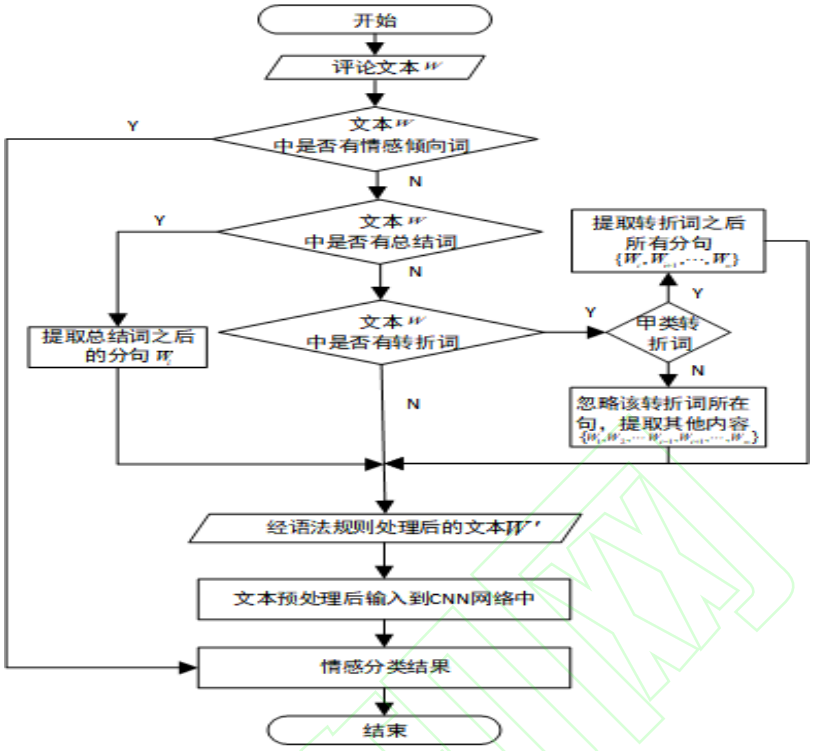

图 1　语法规则提取情感倾向语句流程图

Fig. 1　Flow chart of grammatical rule extraction sentiment sentence

文本 W 经语法规则处理后会先后得到四种情况：（1）直接根据情感倾向词得到文本的情感分类结果；（2）得到含有总结词的分句；（3）得到判断甲乙类转折词的分句或；（4）得到原文本。如此处理后得到的文本 W'，大部分比起原文本更加简短且具有明显的情感倾向，大大降低中文文本的语义多样化，从而解决了输入到 CNN 后训练时因文本语义复杂而导致的特征提取困难问题。

2　双通道神经网络

2.1　CNN 通道

CNN 拥有局部感知与参数共享两大特点，每个神经元只需对局部进行感知，且在局部连接中，每个神经元的参数都是一样的，进行卷积操作时实际上是提取一个个的局部信息。因此对于规则处理后的文本 W'，使用 CNN 模型能够有效的提取出局部特征。CNN 通道模型结构图如图 2 所示。

在 CNN 模型训练中，由图 1 得到经语法规则处理后的评论文本 W'，然后使用 jieba 分词得到文本序列为 $x=\{x_1,x_2,\cdots,x_n\}$，其中 $x_n \in R$，通过词嵌入技术 Word2Vec 得到整个文本序列的词向量句子表示如公式（2）所示，其中，x_i 表示 w_i 对应的词向量，$\oplus$ 为拼接操作。

$$X = x_1 \oplus x_2 \oplus \cdots \oplus x_n \tag{2}$$

将 X 作为卷积层的输入，通过大小为 $r * k$ 的滤波器提取出不同位置的局部特征，计算公式如式（3）。

$$h_i = f\left(w \cdot x_{i:i+k-1} + b\right) \tag{3}$$

其中，w 为卷积核，k 为卷积核尺寸，$x_{i:i+k-1}$ 为 i 到 $i+k-1$ 个词组成的特征向量，b 为偏置项。故通过卷积层后得到输出 $h=[h_1,h_2,\cdots,h_{n-k+1}]$。由于卷积核共享存在着特征提取不充分的问题，需通过增加多个卷积核来弥补，通过固定参数的训练方法得到 CNN 卷积窗口分别为 3、4、5 时会分类效果更好，故经过卷积操作后本文的卷积输出为 h_3、h_4、h_5。

然后对于卷积层的每一个输出向量 h 与 Bi-LSTM 提取出的全局特征 h_{blstm} 进行注意力池化操作以提取出更能够表达情感倾向的特征。其中，注意力池化是通过计算局部特征与全局特征之间的相似性，相似性越高将给该局部特征分配更大的权重。计算公式如式（4）（5）所示。

$$e_i = sim(h, h_{blstm}) \tag{4}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{i=1}^{n} \exp(e_i)} \tag{5}$$

其中，函数 $sim()$ 通过余弦函数计算局部特征与全局特征之间的相似度，α_i 为权重。计算出权重后，最终的局部特征表示 h_{cnn} 由式（6）得到。

$$h_{cnn}=\sum_{i=1}^{n}\alpha_i h \tag{6}$$

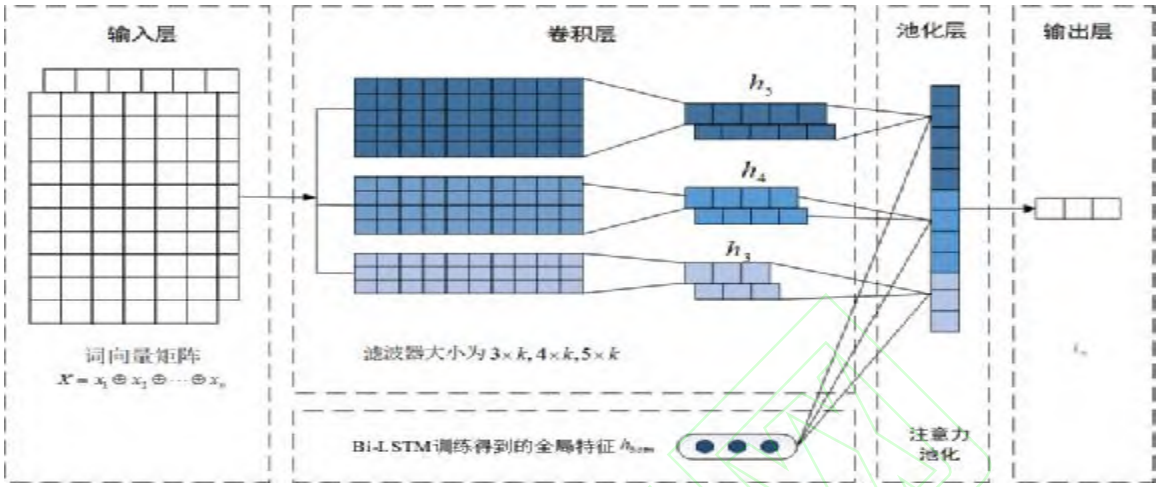

图 2 CNN 通道模型结构

Fig.2 CNN channel model structure

2.2 Bi-LSTM 通道

由于经语法规则处理后的评论文本 W 可能会省略掉部分文本，从而导致了上下文信息缺失的问题，因此使用 Bi-LSTM 模型来获取包含上下文信息的全局特征。模型结构图如图 3 所示。

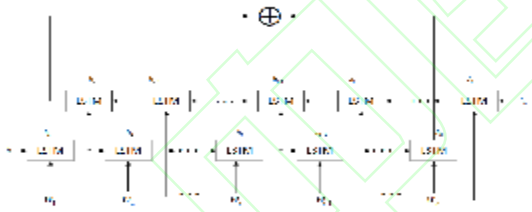

图 3 Bi-LSTM 通道模型结构

Fig. 3 Bi-LSTM channel model structure

将未经语法规则处理的文本经过 jieba 分词以后利用 Glove 工具训练得到词向量，作为 Bi-LSTM 模型的输入。其原理就是首先构建基 于语料库的词共现矩阵，然后根据共现矩阵与 Glove 模型学习词向量。Glove 模型综合了隐形语义分析（LSA）和 Word2Vec 模型的优点，在高效清晰的表达文本语义的同时考虑了全局的文本信息。设第 i 个单词的 n 维词向量表示为 $v_i=\{w_1,w_2,\cdots,w_n\}$，将每个单词向量 V 结合起来形成句子的矩阵表示形式 $R^{s\times n}$，其中，每一行是每个单词对应的词向量权重。设 s 代表单词总数，若确定了词向量权重的维度大小，则该矩阵的列也将确定，n 代表词语维度。令 $v_i\in R^n$ 表示第 i 个词的 n 维词向量，则长度为 s 的文本表示为：

$$V_{1:s}=v_1\oplus v_2\oplus\cdots\oplus v_s \tag{7}$$

经过将单词转换为词向量，则句子中的每个词的词向量拼接最终组成了词向量矩阵，既 $V\in R^{s\times n}$。接着将词向量从前后两个方向输入模型，设定 w、u、v、v'为 Bi-LSTM 模型的权重，当前单元输入为 x_t，前一单元输入为 h_{t-1}，后一单元输入为 h_{t+1}。由公式（8）得到上文的情感倾向特征，由公式（9）得到下文的情感倾向特征，最终由公式（10）得到了包含上下文信息的全局特征 h_{blstm}。

$$\overrightarrow{h_t}=f(w x_t+u\overrightarrow{h_{t-1}}) \tag{8}$$

$$\overleftarrow{h_t}=f(w x_t+u\overleftarrow{h_{t+1}}) \tag{9}$$

$$h_{blstm}=g(v\overrightarrow{h_t}+v'\overleftarrow{h_t}) \tag{10}$$

综上，由双通道神经网络得到了文本的局部 h_{cnn} 与全局特征 h_{blstm}，并将两者作为本文提出的 CB_Rule 模型特征融合的输入，以增强分类器中情感特征的全面性，从而提高情感分类精度。

3 融合语法规则的双通道中文情感分析模型

虽然经语法规则处理后的文本能够使 CNN 获取到情感倾向更加明确的局部特征 h_{cnn}，但也存在因语法规则而存在的忽略上下文信息的问题，考虑使用 Bi-LSTM 提取出的全局特征来作为局部特征忽略问题的弥补，所以本文将其与

Bi-LSTM 提取的全局特征 h_{blstm} 融合起来。融合公式如公式（11）所示。

$$h = h_{cnn} \oplus h_{blstm} \quad (11)$$

融合即将 h_{cnn} 与 h_{blstm} 拼接在一起，一同作为全连接层的输入，并引入 dropout 机制，这样能有效避免模型对部分特征产生依赖，从而发生过拟合现象，最后将其输入到 softmax 分类器中。流程图如图 4 所示。

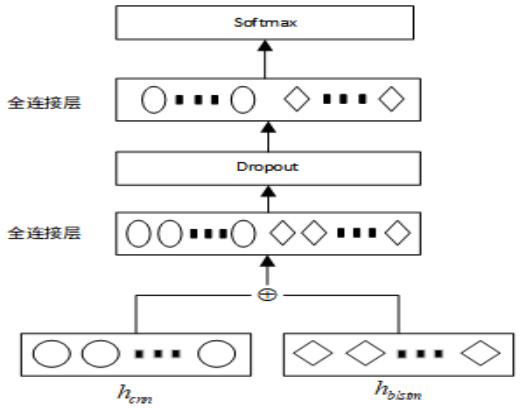

图 4　双通道特征融合流程图

Fig. 4　Flowchart of dual-channel feature fusion

特征融合充分利用了 CNN 较强的文本特征提取能力，又发挥了 Bi-LSTM 对时间序列信息强大的记忆力，最终能够让分类器得到的情感倾向特征 h 更加全面，最后通过分类器得到中文文本情感分类类别，分类公式如公式（12）所示。

$$y = soft\max(W_h \cdot h + b_h) \quad (12)$$

其中，W_h 为权重矩阵，b_h 为偏置，y 为情感类别。

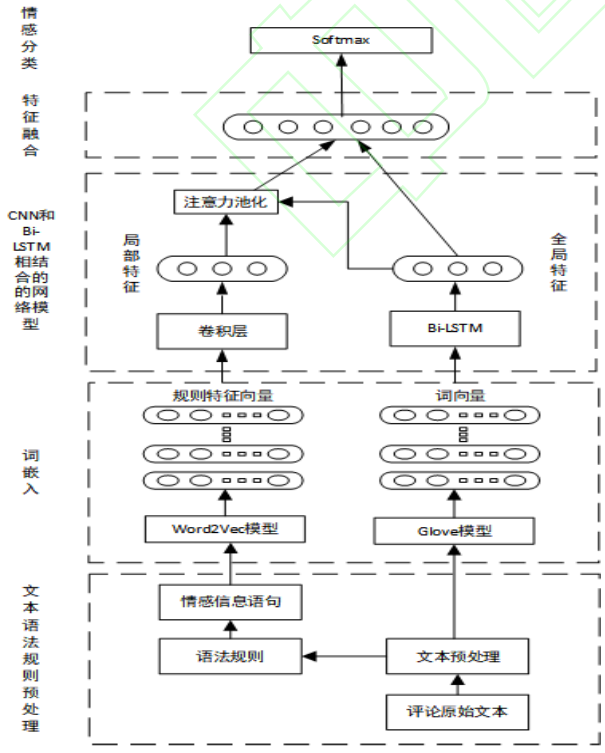

图 5　融合语法规则的双通道神经网络模型

Fig.5　A dual-channel neural network model incorporating grammatical rules

同时本文将利用反向传播算法来训练模型，通过最小化交叉熵得到的损失函数来优化模型，如公式（13）所示。

$$LOSS = -\sum_{i=1}^{n}\sum_{j=1}^{c} p_i \log(y_i) + \lambda \|Q\| \quad (13)$$

c 为情感类别数量，n 为句子数量，p_i 为实际类别，y_i 为预测类别，λ 为 L_2 正则化权重，Q 包含了 CNN 和 Bi-LSTM 中的所有权重及偏置项。

综上，融合语法规则的双通道神经网络模型如图 5 所示，其构建过程主要分为三步：

（1）将文本预处理后的数据依据设定的语法规则获取到文本的情感信息语句，然后通过 Word2Vec 词嵌入工具转换成规则特征向量，再将规则特征向量输入到 CNN 模型中。同时将未经规则处理的文本经过 Glove 工具转换成词向量，输入到 Bi-LSTM 模型中；

（2）在 CNN 和 Bi-LSTM 相结合的神经网络模型中，CNN 模型提取出文本的局部特征，其中将使用注意力池化的方法来提取出 CNN 卷积层的局部特征，以此来判断哪些特征能够包含更多的情感信息。而 Bi-LSTM 则用来提取出文本的全局特征；

（3）将双通道神经网络模型输出的局部特征与全局特征进行融合后，输入到分类器中进行情感分类。

4　实验与结果分析

4.1　实验数据

本文所采用的实验数据为情感分析开源数据集 online_shopping_10_cats 电商购物评论，数据对象有书籍、平板、手机、水果等十个类别，电商评论情感标签分为两类[0,1]，积极评论情感标签为 1，消极评论情感标签为 0。共 62272 条数据，其中正向评论 31351 条，负向评论 31421 条，数据集具体数据分布见表 1。实验数据的训练集与测试集比例设置为 8:2。

表 1 数据集数据分布

Tab.1 Data set data distribution

类别	正向评论数	负向评论数
书籍	2100	1751
平板	5000	5000
手机	1163	1158
水果	5000	5000
计算机	1996	1996
酒店	5000	5000
蒙牛牛奶	992	1041
热水器	100	475
洗发水	5000	5000
衣服	5000	5000

4.2 实验参数设置

本文融合神经网络模型中 CNN 部分的参数及值如表 2 所示，Bi-LSTM 部分的参数及值如表 3 所示。

表 2 卷积神经网络模型参数

Tab. 2 parameters of CNN

参数	值
词向量维度	100
卷积核大小	3、4、5
隐藏层大小	128
激活函数	RELU

表 3 双向长短时记忆网络模型参数

Tab. 3 Parameters of BiLSTM model

参数描述	值
词向量维度	100
Bi-LSTM 层数	2
隐藏层大小	128
学习率	0.001
优化函数	Adam

针对不同数据集所需的情感词典不同，根据上文的规则设定，由本文实验所用的电商评论数据集得到的三类情感词典的部分关键词如下：

正面情感倾向词：推荐、值得、值、物超所值、强烈推荐、性价比高、性质量不错、五星、好评、给力、满意等。

负面情感倾向词：不推荐、不值得、质量不行、性价比低、差评、不满意、失望、别买、一星、不值等。

总结词：总的来说 、总之、总的感觉、总体、在我看来、综上所诉、个人认为、反正、个人建议、整体等。

甲类转折词：但是、但、可是、却、不过、然而、所以、因此等。

乙类转折词：只是、只不过、但就是、而且、就是、虽然、如果等。

由电商评论数据集提取出的种子主题词有产品、快递、价格、质量、性能、包装、客服、外型等，种子主题词与待定主题之间相似度阈值范围设为 0.8。

4.3 评价指标

本文将采用准确率 Acc，召回率 $\mathrm{Re}call$ 以及 $F1$ 值作为实验评价指标，其他符号表示如表 4 所示。

表 4 分类类别混淆矩阵

Tab.4 Classification category confusion matrix

真实结果	分类结果	
	属于情感类别 X	不属于情感类别 X
属于情感类别 X	r	t
不属于情感类别 X	s	z

准确率 Acc 表示测试集所有样本都正确分类的概率，计算公式如式（14）。

$$Acc=\frac{r}{r+t} \tag{14}$$

召回率 $\mathrm{Re}call$ 表示测试集分类结果中某情感标签中的真实类别占所有真实类别的比例，计算公式如式（15）所示。

$$\mathrm{Re}call=\frac{r}{r+s} \tag{15}$$

F1 值是一个准确率 $Acc(A)$ 与召回率 $\mathrm{Re}call(R)$ 综合性能的指标，对两者加权调和计算得到最终分类效果，即

$$F1=\frac{2\acute{}\ A\acute{}\ R}{A+R} \tag{16}$$

4.4 实验结果分析

4.4.1 语法规则可行性分析

本文随机选取电商评论中的 10000 条数据，验证上文提出的语法规则对 CNN 分类结果的影响，其中，r1-r3 为文中第一章提出的前三个规则，CNN 参数设置见表 2。实验结果如表 5 所示。

表 5 语法规则对 CNN 模型分类结果表

Tab. 5 Grammar rules on the CNN model classification result table

模型	Acc	Recall	F1
CNN	0.893	0.881	0.887
CNN+r1	0.920	0.918	0.919
CNN+r1+r2	0.924	0.926	0.925
CNN+r1+r2+r3	0.929	0.931	0.930

由表 5 实验结果发现，本文根据情感倾向词、总结词、转折词设定的语法规则能够有效提升 CNN 模型的分类精度，其中规则 1 对模型的分类结果影响最大，F1 值较 CNN 模型提升了 3.48%，表明情感倾向词对分类结果的影响比重高于总结词与转折词，通过提取有效情感倾向词更能够促进文本情感分类效果。虽然使用规则 2 与规则 3 的提升效果没有规则 1 明显，但总体上都促进了 CNN 模型的最终分类精度。证明了通过语法规则处理能够得到更加明确的情感倾向信息，进而帮助 CNN 提取到语义特征，提高分类精度。

同时，本文还将语法规则应用到了机器学习算法 SVM 上，并与 CNN 模型进行对比实验，同样随机选取 10000 条数据，设定批大小 batch 大小为 64，迭代次数 iteration 为 157 次，数据集训练轮数 epoch 为 15 次。结果见图 6。

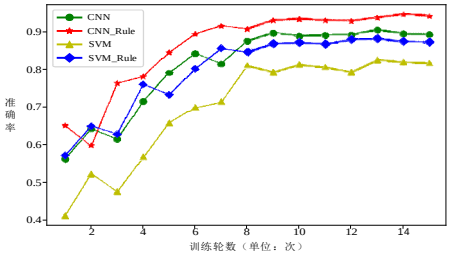

图 6 应用语法规则效果图

Fig. 6 Effect diagram of applied grammar rules

从图 6 发现，语法规则应用到机器学习算法 SVM 与传统 CNN 模型上，分类准确率都得到了显著的提升。其准确率都随着 epoch 的增加而增加， CNN、CNN_Rule、SVM、SVM_Rule 最终的准确率稳定在 89%、93%、80%、86%左右，进一步有效验证了改语法规则的可行性。

4.4.2 融合语法规则的双通道模型分类精度

为解决双通道模型特征融合时出现的过拟合现象，使用 10000 条电商评论文本作为实验数据，在模型全连接层加入 Dropout，并通过实验对比了 Dropout 值在 0.1-1 之间的准确率变化，最终选择 0.5 的 Dropout 的最适值，实验结果如图 7 所示。

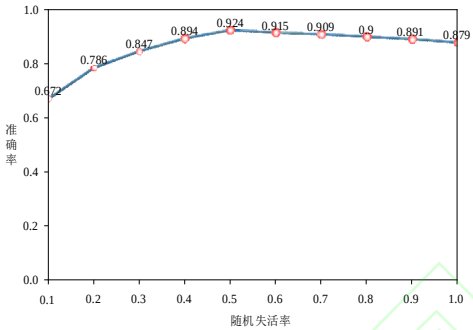

图 7 Dropout 参数值对模型性能的影响图

Fig. 7 Diagram of the influence of Dropout parameter value on model performance

为了验证本文提出的 CB_Rule 模型性能，在相同实验环境下使用表 1 数据进行实验，并根据图 7 的实验结果选取 Dropout 为 0.5。首先分别利用 Word2Vec 和 Glove 向量化工具将评论文本转换成矩阵向量，再构造单一的 CNN，Bi-LSTM 模型以及双通道模型 CNN_BLstm 模型与 CB_Rule 进行对比实验，使用 AUC 值作为情感分类效果的评价指标，ROC 曲线如图 8 所示。

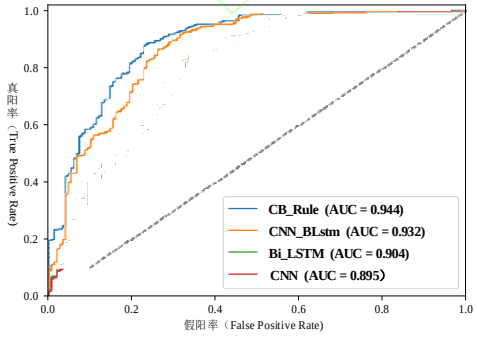

图 8 CB_Rule 模型与其他分类模型 ROC 图

Fig.8 ROC diagram of CB_Rule model and other classification models

由图8可知，Bi_LSTM的AUC值比CNN模型高出0.9%，说明在中文情感分类任务中，上下文信息影响着分类结果，所以仅使用融合规则的 CNN 模型进行情感分类时，就容易忽略了上下文信息，造成模型分类性能下降。双通道 CNN_BLstm 模型的 AUC 值较传统的单 Bi-LSTM、单 CNN 模型分别高出 3.7%、2.8%，究其原因，CNN 模型具有的局部感知与参数共享使其关注的是局部语义特征的提取，而较少考虑到上下文信息；反之，Bi-LSTM 由于其对时间序列的超强记忆功能，通过正反向 LSTM 传播得到了上下文信息，但也忽略了局部语义特征在中文情感分析中的重要性。这再次说明了将 CNN 提取出的局部特征与 Bi-LSTM 提取的全局特征融合起来对情感分类效果有着显著的影响。同时，将语法规则融入到双通道模型中时，CB_Rule 模型的 AUC 值又比双通道 CNN_BLstm 模型提高了 1.27%，验证了将语法规则融入其中更有助于情感特征的获取，提升神经网络分类效果。

4.4.3 CB_Rule 模型性能对比

为了验证本文提出的规则融合模型的情感分类性能，将本文提出的 CB_Rule 模型与文献[12-16]提出的模型在表 1 数据集上进行对比实验，其中各个模型的 CNN 与 Bi-LSTM 皆按照表 2、表 3 进行参数设置，实验结果如表 6 所示。文献[12]提出的 L- BiLSTM_CNN 模型将 CNN 提取的局部特征与 Bi-LSTM 提取的全局特征特征融合后使用分类器进行情感分类。Z-BiLSTM_CNN 为文献[13]构建的 BiLSTM 和 CNN 的串行混合模型，首先利用 Bi-LSTM 提取上下文特征，再对上下文特征进行局部特征提取，最后使用分类器进行分类。R-Bi-LSTM 为文献[15]提出的融合语法规则的 Bi-LSTM 模型，并采用 Glove 工具进行词向量训练。而 SCNN 则为文献[16]提出的融合句法规则和 CNN 的旅游评论情感分析模型，但词向量训练采用 word2vec 模型。

表 6 CB_Rule 模型与其他模型分类结果对比

Tab. 6 Comparison of classification results of CB_Rule model and other models

模型	准确率
CB_Rule	0.951
L- BiLSTM_CNN	0.942
Z-BiLSTM_CNN	0.916
R-Bi-LSTM	0.914
SCNN	0.945

由表 6 可知，本文所提出的 CB_Rule 模型的准确率优于其他模型。将 CNN 提取出的局部特征和 Bi-LSTM 提取的全局特征进行融合时，L-BiLSTM_CNN 模型的准确率明显高于 Z-BiLSTM_CNN 模型，说明直接并行提取出特征进行融合的效果优于串行提取出后再进行特征融合，故而本文采用了不同的词向量处理工具对文本数据并行处理。同时，本文提出的将语法规则融入到双通道模型在准确率上较 R-Bi-LSTM，SCNN 模型高出了 3.89%、0.63%，则进一步验证了 CB_Rule 模型在情感分类效果上的有效性。

5 结语

进行情感研究对当今社会意义重大，本文针对传统的 CNN 与 Bi-LSTM 这类情感分类模型所存在的问题，提出了融合语法规则的双通道中文情感分析模型，将语法规则融入到 CNN 中，训练得到更具有情感倾向的局部特征，同时为了解决语法规则处理后出现的忽略上下文信息问题，利用 Bi-LSTM 对之进行补充改进，最后将提取出的特征进行融合，将其输入到分类器中提高情感分类精度。在电商评论文本数据集上设计了语法规则的可行性分析、融合语法规则的双通道模型的分类精度以及 CB_Rule 模型性能对比等实验，验证了本文提出的 CB_Rule 模型具备良好的情感分类效果。

参考文献

[1] ARAQUE O, ZHU G, IGLESIAS C A, et al. A semantic similarity-based perspective of affect lexicons for sentiment analysis[J]. Knowledge Based Systems, 2019,165: 346-359.

[2] ZHANG S, WEI Z, WANG Y, et al. Sentiment analysis of chinese micro-blog text based on extended sentiment dictionary[J]. Future Generation Computer Systems, 2018,81: 395-403.

[3] XU G, YU Z, YAO H, et al. Chinese text sentiment analysis based on extended sentiment dictionary[J]. IEEE Access, 2019,7: 43749-43762.

[4] HUNG C. Word of mouth quality classification based on contextual sentiment lexicons[J]. Information Processing and Management, 2017, 53(04): 751-763.

[5] KHOO C S, JOHNKHAN S B. Lexicon-based sentiment analysis: comparative evaluation of six sentiment lexicons[J]. Journal of Information Science, 2018, 44(04): 491-511.

[6] SINGH J , SINGH G , SINGH R . Optimization of sentiment analysis using machine learning classifiers[J]. Human-centric Computing and Information Sciences, 2017, 7(01):1-12.

[7] ANGGITA,SHARAZITA D,IKMAH. Algorithm comparation of naive bayes and support vector machine based on particle swarm optimization in sentiment analysis of freight forwarding services[J].Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi) ,2020,4(02):362-369.

[8] TAMA V O , SIBARONI Y , ADIWIJAYA. Labeling analysis in the classification of product review sentiments by using multinomial naive bayes algorithm[J]. Journal of Physics: Conference Series, 2019,1192(01):12036.

[9] 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(05):945-957.(CHEN K,LIANG B,KE W D,et al. Chinese micro-blog sentiment analysis based on multi-channels convolutional neural networks[J]. Journal of Computer Research and Development, 2018,55(05):945-957.)

[10] LONG F, ZHOU K, OU W H. Sentiment analysis of text based on bidirectional LSTM with multi-hhead attention[J]. IEEE Access, 2019,7: 141960-141969.

[11] KAI S. Word attention-based BiLSTM and CNN ensemble for chinese sentiment analysis[J]. Computer Science and Application 2020, 10(02), 312-324.

[12] 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.(LI Y,DONG H B. Text sentiment analysis based on feature fusion of convolution neural network and bidirectional long short-term memory network[J]. Journal of Computer Applications, 2018,38(11):3075-3080.)

[13] 赵宏,王乐,王伟杰.基于 BiLSTM-CNN 串行混合模型的文本情感分析[J].计算机应用,2020,40(01):16-22. (ZHAO H,WANG L,WANG W J. Text sentiment analysis based on serial hybrid model of bi-directional long short-term memory and convolutional neural network[J]. Journal of Computer Applications, 2020,40(01):16-22.)

[14] WANG J, YU L, LAI K R, et al. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2020,28: 581-591.

[15] 卢强,朱振方,徐富永,等.融合语法规则的 Bi-LSTM 中文情感分类方法研究[J].数据分析与知识发现,2019,3(11):99-107.(LU Q,ZHU Z F,XU F Y,et al. Chinese sentiment classification method with Bi-LSTM and grammar rules [J].Data Analysis and Knowledge Discovery ,2019,3(11):99-107.)

[16] 何雪琴,杨文忠,吾守尔·斯拉木,等.融合句法规则和 CNN 的旅游评论情感分析[J].计算机工程与设计,2019,40(11):3306-3312.(HE X Q,YANG W Z, WUSHOUER S L M,et al.Sentiment analysis of tourist reviews combined with syntactic rules and CNN[J]. Computer Engineering and Design, 2019,40(11):3306-3312.)

This work is partially supported by Jilin Province Science and Technology Development Plan (No.20190302118GX).

QIU Ningjia, born in 1984, Ph. D., lecturer. His research interests include data mining, algorithm analysis, machine learning,natural language processing.

WANG Xiaoxia, born in 1996, M. S. ,candidate. Her research interests include data mining, machine learning, natural language processing.

WANG Peng, born in 1973, Ph. D., professor. His research interests include data mining.

WANG Yanchun, born in 1964, M. S. master. Her research interests include data mining, smart computing.