DOI:10.3969/j.issn.1000-1565.2021.02.014

集成局部和全局关键特征的文本情感分类方法

柴变芳[1],杨蕾[1],王建岭[2],李仁玲[2]
(1.河北地质大学 信息工程学院,河北 石家庄 050031;2.河北中医学院 图书馆,河北 石家庄 050200)

摘 要:融合卷积神经网络(convolutional neural network,CNN)和双向长短期记忆网络(Bi-directional long short-term memory,BiLSTM)的情感分析模型(CNN_BiLSTM)是一个流行的模型,其学习文本的局部特征和全局特征实现情感分类,但是忽略了特征对分类结果的重要程度,且没充分利用词语间的特征,导致分类准确率不高.提出一种集成基于多卷积核的卷积神经网络和注意力双向长短期记忆网络特征的文本情感分类方法(MCNN_Att-BiLSTM),其集成局部和全局的重要特征作为文本语义特征,该特征进而用于训练文本情感分类器 XGBoost(eXtreme gradient Boosting).该方法基于注意力机制的 BiLSTM 提取对分类影响大的全局关键特征,基于多卷积核的 CNN 获得更全面的词语间特征,为集成分类器准备了有效分类的特征.实验结果表明,该模型具有更好的情感分类准确率,与 CNN_BiLSTM 模型相比,在 IMDB 数据集上准确率提升了 1.75%,在 txt-sentoken 数据集上准确率提升了 1.67%,在谭松波-酒店评论数据集上准确率提升了 3.81%.

关键词:情感分析;CNN;BiLSTM;XGBoost;特征融合
中图分类号:TP391 **文献标志码**:A **文章编号**:1000-1565(2021)02-0201-11

Text sentiment classification approach with integrated local and global prominent features

CHAI Bianfang[1], YANG Lei[1], WANG Jianling[2], LI Renling[2]
(1. College of Information Engineering, Hebei GEO University, Shijiazhuang 050031, China;
2. Library, Hebei University of Chinese Medicine, Shijiazhuang 050200, China)

Abstract: The model combining convolutional neural network (CNN) with Bi-directional long short-term memory (BiLSTM) feature (CNN_BiLSTM) is popular for sentiment analysis. It takes into account the local and global features to realize sentiment classification of the text. However, it ignores the importance of the features for the classification results, and does not make full use of features between words, which result in low classification efficiency and accuracy. Thus, a model based on a CNN that integrates multiple convolution kernels and a bidirectional long-term and short-term memory network of attention (MCNN_Att-BiLSTM) for

收稿日期:2020-03-12
基金项目:国家自然科学基金资助项目(81473773);河北省自然科学基金资助项目(F2019403070);河北省教育厅重点项目(ZD2020175)
第一作者:柴变芳(1979—),女,山西运城人,河北地质大学教授,博士,主要从事机器学习、复杂网络分析研究.
E-mail:chaibianfang@163.com
通信作者:李仁玲(1973—),女,河北巨鹿人,河北中医学院副研究馆员,主要从事数据检索和大数据分析研究.
E-mail:w3365@126.com
王建岭(1973—),男,河北巨鹿人,河北中医学院教授,主要从事网络安全和大数据挖掘研究.
E-mail:wang_jl@126.com

text sentiment classification is proposed. It integrates local and global prominent features as semantic features of the text, which are used as inputs to train a XGBoost (eXtreme gradient Boosting) classifier to realize text sentiment classification. It utilizes the attention mechanism based on the BiLSTM to fully captain the global prominent features that affect the classification results largely. In addition, it utilizes the CNN with multi-convolution kernel to obtain more comprehensive inter-word features. Experimental results show that the model is better compared to the CNN_BiLSTM model, which improves the accuracy rate by 1.75% on the IMDB dataset, and 1.67% on the txt-sentoken dataset, and 3.81% on the Tan Songbo-hotel review dataset.

Key words: sentiment analysis; CNN; BiLSTM; XGBoost; feature fusion

随着社交媒体的快速发展,微博、微信等网络平台逐渐进入公众的生活且带来了极大的便利.人们足不出户就能知道社会热点问题及现象,因此越来越多的人不只是浏览信息,而是经常发表自己的情感态度及观点看法.通过对这些信息进行情感分析,可以帮助国家政府部门正确引导社会舆论方向.目前,情感分析技术已经广泛应用在舆情监控、商业决策、信息预测等领域[1].

自从 Nasukawa 等[2]在 2003 年提出情感分析概念以来,便受到了越来越多研究者的关注.当前主要的研究方法有 3 类:1)基于词典和规则的方法[3-4].该类方法根据经验提取文本中的情感词,然后按照特定的规则对文本进行打分,最后根据分值判断文本的情感极性.2)基于经典机器学习的方法[5-6].此类方法通过特征工程获得每个训练文档的特征,然后基于训练集学习情感分类模型,新文档利用该模型实现情感预测.3)基于深度学习的方法[7-11].该类方法不需要特征工程,以情感分类任务为目标利用深度学习模型自动提取特征,进而训练情感分类神经网络模型.

近年来,一些研究者大量使用基于词典和规则的方法进行情感分类.文献[3]提出了一种利用单词的统计特征创建文本分类中特征空间的表达方法.文献[4]在已有的情感词典基础上,通过 LDA 模型从语料中提取主题词来扩展特定领域词典,并在多个领域数据集上进行应用且取得不错效果.该类方法适用的语料范围较广,但灵活度不高,分类结果过于依赖情感词典.随着机器学习的发展,基于经典机器学习方法在文本情感分析任务中得到了广泛的应用.文献[5]提出了一种基于表情符号的文本自动标注方法.首先从文本中筛选出情感倾向明显的表情符号作为训练集,然后用机器学习的方法训练分类器,最后在人工标注的测试集中验证并取得了较高的准确率.文献[6]使用集成机器学习技术来提高所提出方法的效率和可靠性,同时将支持向量机与决策树合并,在准确性方面提供了更好的分类结果.此类方法虽然提高了分类准确率,但泛化能力较差,需要大量数据且难以充分挖掘文本中词语更深层次的语义信息.

近来,深度学习在情感分析任务中也取得了一些成果.Le 等[7]提出了一种具有不同尺寸卷积核和多类型池化的 CNN 用于文本分类.陈珂等[8]提出了一种多通道卷积神经网络,通过将词向量特征、情感词特征和位置特征进行组合形成不同通道,再使用 CNN 进行分类,最终获得了比普通卷积神经网络更好的性能.文献[7]和[8]使用卷积神经网络模型,利用词向量、词语位置信息等来获取词语间深层次的情感信息,但对于较长的文本,该方法不能记忆长距离的全局情感信息,只能提取文本的局部特征,导致分类准确率不高.Irsoy 等[9]使用循环神经网络(recurrent neural network,RNN)及长短期记忆网络(long short-term memory,LSTM)为情感分类构建了深度学习模型.李洋等[10]提出了一种 CNN 和 BiLSTM 特征融合模型,有效提高了文本分类准确率.文献[9]和[10]使用循环神经网络模型及其各种变体,充分考虑了文本序列的前后依赖关系,但该模型认为文本中各部分信息对分类结果的影响相同,忽略了情感词语相比于普通词语对情感倾向影响更大.Araque 等[11]提出了一种结合词语浅层特征和深层特征的模型,充分利用词语的多方面情感信息进行分类并取得了很好的效果.该文献将深度学习自动提取的特征和传统方法手动提取的特征相结合,既充分考虑词语间深层次信息,又通过传统方法提取更准确的情感信息.

由于集成分类器在提高效率和准确率方面具有明显的优势,一些研究者也将其应用在情感分析任务中.

苏兵杰等[12]采用 XGBoost 算法对网络上的商品评论进行情感分析,通过对数据集中的训练集提取特征,利用 XGBoost 算法训练分类器获得情感分类模型.龚维印等[13]提出了一种基于卷积神经网络和 XGBoost 的文本分类模型 CNNs-XGB.首先利用 word2vec 对预处理后的数据进行词向量表示,其次利用多尺寸卷积核的卷积神经网络进行数据特征提取,最后利用 XGBoost 对深度提取的特征进行分类处理.

为了全面考虑文本中词语表达的情感信息及各部分情感信息的重要程度,提出一种集成多卷积核的卷积神经网络和注意力双向长短期记忆网络模型(MCNN_Att-BiLSTM)的情感分类方法,提高分类准确率.利用注意力机制为 BiLSTM 获取的全局特征分配 1 个权重向量,充分考虑各部分特征对分类结果的影响大小;利用多种不同大小的卷积核,获取词语间更全面的局部信息;然后融合这 2 部分特征,利用集成分类器 XGBoost 进行情感分类,提高分类的准确率和效率.

1　集成关键特征的情感分类模型

提出一个利用深度特征实现文本情感分类的模型 MCNN_Att-BiLSTM.针对融合 CNN 和 BiLSTM 特征的情感分类方法没有充分考虑词语间情感信息的问题,利用多卷积核的卷积神经网络(MCNN)模型实现局部特征的提取,在 CNN 中使用多种大小的卷积核,以词为单位,将提取的不同词语间特征进行拼接,从而获得更准确的情感信息.针对 CNN_BiLSTM 模型未考虑提取的上下文特征对情感分类结果的影响程度问题,利用 Att-BiLSTM 实现全局关键特征提取,在 BiLSTM 基础之上增加注意力层,为隐层输出特征赋予 1 个权重,可获得不同特征对情感分类影响程度的量化值.为了进一步提高情感分类的准确率,融合基于 MCNN 和 Att-BiLSTM 特征,输入集成分类器 XGBoost 实现文本情感分类.基于 MCNN_Att-BiLSTM 模型的情感分类模型训练流程如图 1 所示.首先对文本进行数据处理,将文本中的每个词语利用嵌入技术进行向量表示,然后利用 MCNN 和 Att-BiLSTM 模型学习文本特征,最后送入集成分类器 XGBoost 训练分类模型.新来一个待预测文本,利用此流程可得其情感分类结果.

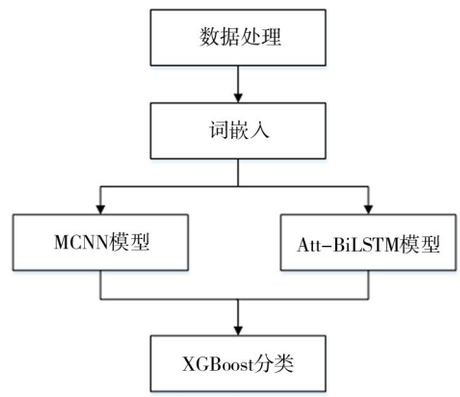


图 1　MCNN_Att-BiLSTM 模型流程

Fig.1　MCNN_Att-BiLSTM model flow chart

1.1　数据处理

情感分类模型的输入为大量的积极和消极评论的文本数据.首先对读取的每个文本进行分词处理(英文数据通过空格进行分词,中文数据使用 jieba 库进行分词),将每个文本数据转换为词语的集合.为了提高效率,需要去除其中的停用词,即一些没有特定含义的词语,如英文中的“the”,“is”,“a”等,中文中的“一些”、“一个”等.

文本分词完成后,根据词语在文本数据中出现的频率的大小生成文本词典.然后将每个文本中的词语和词典中的词序号一一对应形成一种映射关系,使每个文本由词语的集合转换为每个词语在词典中位置序号

的集合.最后将所有文本打乱顺序,按照8∶2的比例划分训练集和测试集.

1.2 词嵌入

在文本情感分析任务中,计算机不能识别英文单词或中文词语,因此需要将数据通过编码向量化.最简单的一种编码方式为one-hot编码[14],为每个文本定义一个由0和1组成的一维向量,长度为文本词典中词语的数量,每维表示对应词语是否在文本中出现,出现为1,否则为0.虽然one-hot编码可以将词语转化为向量,但是形成的向量非常稀疏且维度巨大,不能保留文本的语义关系,使得分类效果并不准确.

为了将词语转化为向量且保留文本语义关系,本文选择Tensorflow框架中的嵌入方式.首先初始化一个符合均匀分布的-1到1的向量矩阵,行数为文本词典中词语的数量,列的数量为词语编码的维度数量;然后根据数据处理后的位置序号集合在初始化的向量矩阵中查找每个词语序号对应的向量,使每个文本由词序号集合转化为由词向量构成的向量矩阵.最后将所有的向量矩阵根据batch_size划分批次,分批输入2类神经网络模型中进行训练.

1.3 多卷积核的卷积神经网络模型

多卷积核的卷积神经网络在处理文本时的模型如图2所示.该模型包括输入层、多核卷积层、池化层及全连接层4部分.

1) 输入层.文本矩阵$\boldsymbol{S}_j=\{V(W(1)),\cdots,V(W(m))\}$.其中$V(W(i))\in R^K$代表矩阵$\boldsymbol{S}_j$中第$i$个$K$维词向量,$\boldsymbol{S}_j\in R^{m\times K}$,$m$代表文本矩阵$\boldsymbol{S}_j$中的词语数量.

2) 卷积层.利用多种大小为$r\times K$的卷积核提取文本矩阵$\boldsymbol{S}_j$的局部特征,具体公式如下:

$$c_i=f(F\cdot V(W(i:i+r-1))+b), \tag{1}$$

其中,c_i为文本矩阵卷积后的第i个局部特征;F为$r\times K$的卷积核,b为偏置向量;f为激活函数;$V(W(i:i+r-1))$代表$\boldsymbol{S}_j$中从i到$i+r-1$共r行向量.在文本矩阵$\boldsymbol{S}_j$中,卷积核以步长1从上到下滑动,最终得到局部特征向量集合C如下:

$$C=\{c_1,c_2,\ldots,c_{h-r+1}\}, \tag{2}$$

其中,h表示文本中词的个数,即文本矩阵中的宽.

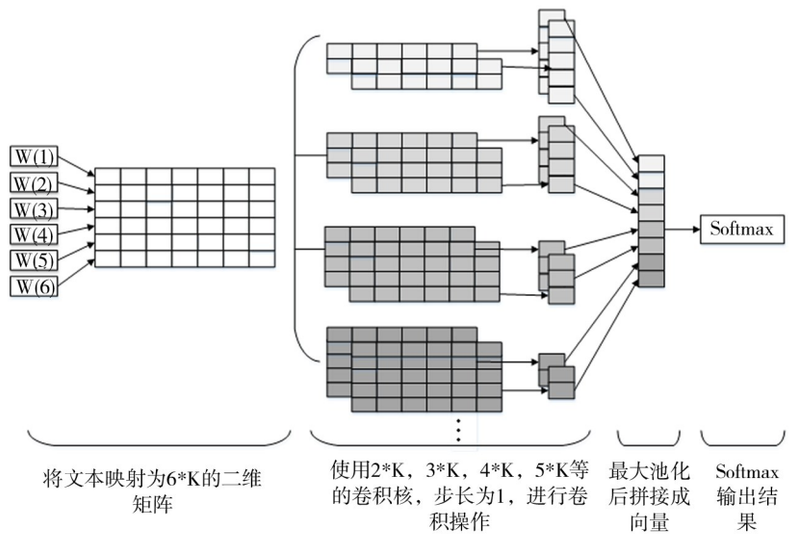


图2 MCNN网络结构

Fig.2 MCNN network structure

3) 池化层.采用最大池化的方法提取局部特征中值最大的特征.

$$d_i=\max C. \tag{3}$$

4) 全连接层.将所有池化后的特征进行组合得到向量

$$\boldsymbol{U}=\{d_1,d_2,\ldots,d_n\},\tag{4}$$

其中,n 表示池化层得到的特征向量数.

最后将 $\boldsymbol{U}$ 输入 softmax 分类器中得到分类结果.模型利用训练数据中的标签,通过反向传播算法进行参数优化.

1.4　基于注意力机制的 BiLSTM 模型

由于 RNN 适用于处理序列数据,所以被广泛应用于自然语言处理任务中,但 RNN 存在梯度消失和梯度爆炸问题.LSTM 模型[15]利用门机制控制每一个 LSTM 单元记忆历史信息和当前输入的信息,保留重要的特征,丢弃不重要特征.LSTM 单元的门机制表达式如图 3 所示.

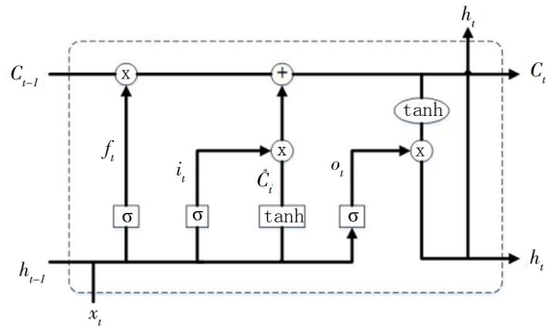


图 3　LSTM 单元模型

Fig.3　LSTM unit model

1) 遗忘门

$$f_t=\sigma(\boldsymbol{W}_f\cdot[h_{t-1},x_t]+b_f),\tag{5}$$

其中,$\boldsymbol{W}_f$ 为权重矩阵,$[h_{t-1},x_t]$表示把前一单元的隐层输出和当前的输入拼接成一个向量,$\boldsymbol{b}_f$ 为偏置向量,σ 是 sigmoid 函数.

2) 输入门

$$i_t=\sigma(\boldsymbol{W}_i\cdot[h_{t-1},x_t]+\boldsymbol{b}_i),\tag{6}$$

$$\tilde{\boldsymbol{C}}_t=\tanh(\boldsymbol{W}_C\cdot[h_{t-1},x_t]+\boldsymbol{b}_C),\tag{7}$$

其中,i_t 表示当前输入要更新的部分值,$\tilde{\boldsymbol{C}}_t$ 表示单元状态中新的候选值向量,$\boldsymbol{W}_i$ 和 $\boldsymbol{W}_C$ 分别为权重矩阵,$\boldsymbol{b}_i$ 和 $\boldsymbol{b}_C$ 分别为偏置向量,σ 为 sigmoid 函数.

3) 更新单元状态

$$C_t=f_t\cdot C_{t-1}+i_t\cdot\tilde{\boldsymbol{C}}_t,\tag{8}$$

其中,C_t 为当前单元状态,C_{t-1}为前一时刻单元状态.

4) 输出门

$$o_t=\sigma(\boldsymbol{W}_0\cdot[h_{t-1},x_t]+b_0),\tag{9}$$

其中,o_t 表示输出值,$\boldsymbol{W}_0$ 为权重矩阵,b_0 为偏置项.

5) LSTM 单元最终输出

$$h_t=o_t\cdot\tanh(C_t).\tag{10}$$

在文本情感分析任务中,标准的 LSTM 模型可根据前面词语推测后面词语的情感信息,后面词语可根据前面词语的语义得到更准确的语义表示.双向长短期记忆网络(BiLSTM)模型[16]在正向的 LSTM 基础上增加反向 LSTM,从 2 个方向收集词语情感语义表示.为了进一步得到更准确的词语语义,增加注意力机制,使每个词语语义表示由其与各个词语表示的相关性来确定.该相关性对应一个权重向量,最终词语表示由权

重向量与 LSTM 层输出的隐含特征相乘再加上偏置向量得到,其通过不断学习进行优化.Att-BiLSTM 模型结构如图 4 所示.

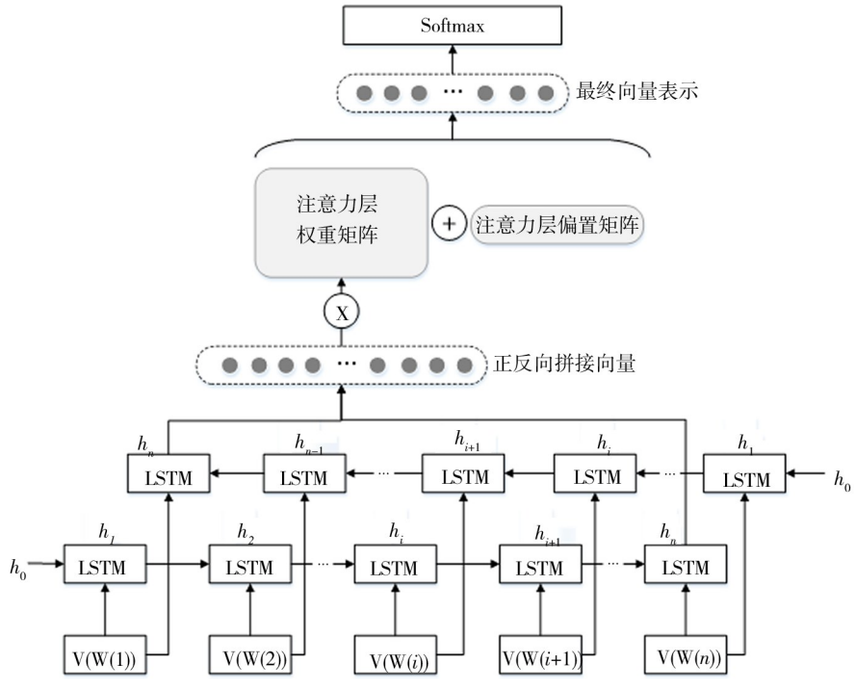


图 4 Att-BiLSTM 模型

Fig.4 Att-BiLSTM model

1.5 XGBoost 模型

陈天奇在 2014 年首次提出 XGBoost 算法[17],它是一种传统的 Boosting 方法,也是一种提升树模型.该算法将许多树模型集成在一起,形成一个强分类器,其中树模型多为 CART 回归树模型.在 CART 回归树模型中使用二叉树,通过信息增益函数确定最优的划分属性.

XGBoost 算法的核心思想是通过不断添加新的回归树,不断进行特征分裂进而生长一棵树[18].每次添加 1 棵树即学习 1 个新函数用来拟合上次预测的残差.在预测一个新样本的分数时,根据该样本的特征,在每棵树中对应 1 个叶子节点,每个叶子节点对应 1 个分数,最后将每棵树对应的分数相加即为该样本的预测值.

$$\hat{y}_i=\sum_{k=1}^{K}f_k(x_i),f_k\in F, \tag{11}$$

其中 $F=\{f(x)=w_{q(x)}\}(q:R^m\to T,w\in R^T)$,表示 $q(x)$CART 结构,即 x 在某一个 CART 树中叶子节点的位置信息;$W_{q(x)}$ 表示输入 x 在某棵 CART 树中的分数;T 为树中叶子节点数;f_k 表示 1 个 CART 树,由树的结构 q 和叶子节点的权重 w 确定.

1.6 模型训练

网络模型将 MCNN 提取的特征和 Att-BiLSTM 提取的特征进行整合为$[v,\tilde{v}]$,作为文本数据的最终表示结果.通过全连接层及 softmax 分类器得到预测的情感类别

$$\hat{y}=\mathrm{softmax}(\boldsymbol{W}_v[v,\tilde{v}]+\boldsymbol{b}_v), \tag{12}$$

其中,$\boldsymbol{W}_v$ 为权重矩阵,$\boldsymbol{b}_v$ 为偏置向量.同时利用反向传播算法训练模型,优化参数,使用 dropout 机制避免过拟合,通过最小化交叉熵调整模型参数,具体公式如下:

$$loss=-\frac{1}{D}\sum_{i}y_i\log\hat{y}_i, \tag{13}$$

其中,D 为训练数据集的大小,y_i 为第 i 类的实际标签,$\hat{y}_i$ 为第 i 类的预测标签.

经过网络模型训练后得到最终的文本特征,将该特征输入 XGBoost 分类器中进行分类获得最终分类结果.同时使用一阶导数和二阶导数最小化下列目标函数,优化模型参数

$$L(\phi)=\sum_{i}l(\hat{y}_i,y_i)+\sum_{k}\Omega(f_k), \tag{14}$$

其中,$\Omega(f)=\gamma T+\frac{1}{2}\lambda\|w\|^2$;$l$ 为自定义的可微凸函数,本实验中使用均方误差;$\hat{y}_i$ 为预测标签;y_i 为真实标签;$\sum_k\Omega(f_k)$为正则化项,降低过拟合的风险;T 表示叶子节点数;w 表示叶子节点分数;γ 和 λ 分别控制 CART 树的个数和叶子节点的分数值.

本文模型的整体流程如图 5 所示.

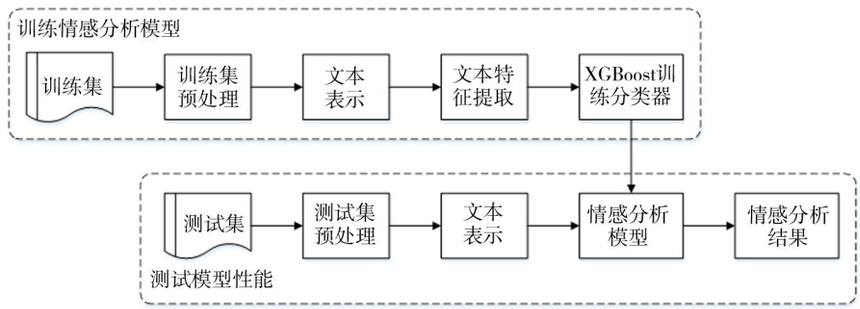


图 5　模型整体流程

Fig.5　Overall flow chart of the model

2　实验

2.1　实验设置

2.1.1　实验数据

实验使用 3 个数据集:第一,Keras 内部集成的 IMDB 影评英文数据集,实验使用其中的 12 500 条正面评论和 12 500 条负面评论进行训练和测试;第二,由 Pang 和 Lee 在 2004 年 ACL 会议上使用的 txt-sentoken 英文数据集[19],包含 1 000 条正面评论和 1 000 条负面评论;第三,谭松波-酒店评论中文数据集,包含 7 000条正面评论和 3 000 条负面评论.

2.1.2　数据集划分

数据经过分词、去停用词及向量化处理后按照 80%和 20%的比例划分训练集和测试集.具体划分如表 1 所示.

表 1　数据集划分

Tab.1　Data set partition

数据集名称	训练集		测试集	
	正面/条	负面/条	正面/条	负面/条
IMDB	10 000	10 000	2 500	2 500
txt-sentoken	800	800	200	200
酒店评论	5 600	2 400	1 400	600

2.1.3 实验参数

实验参数的选取直接影响实验结果,根据固定参数法,本文分别比较了最大句子长度为 100、120、150 个词;将卷积神经网络的卷积核大小由{3,4,5}扩充到{2,3,4,5,6,7},卷积核数分别取 64 和 128 进行比较;dropout 对比了 0.5、0.7 和 0.8,BiLSTM 层数比较了 1 层和 2 层,网络批次大小(batch_size)对比了 100 和 128 对实验结果的影响.通过以上等参数的对比,发现在各参数取表 2 的参数值时,模型准确率较好.

表 2 模型实验参数

Tab.2 Model experiment parameters

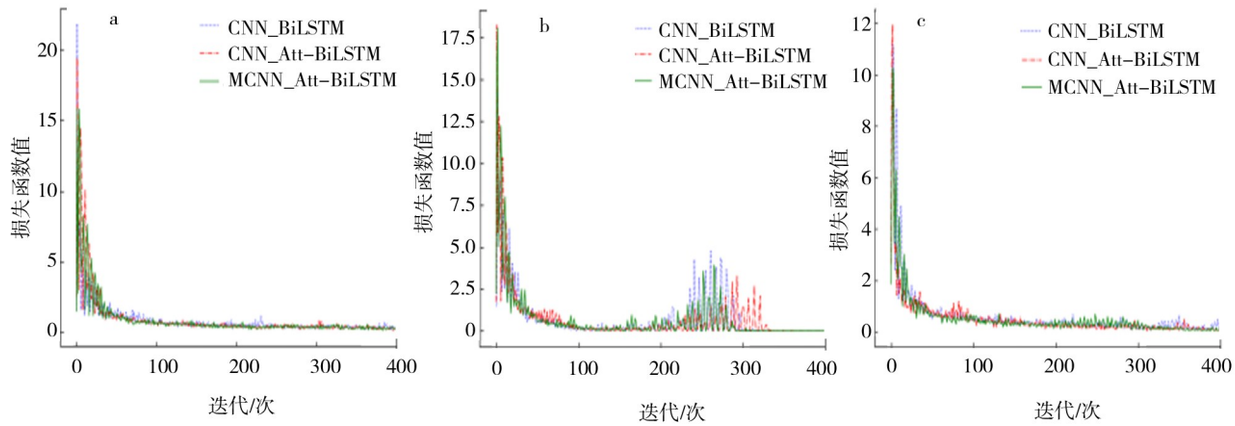
数据参数			网络参数						
词向量维度	词典大小	句子最大长度	CNN 卷积核大小	CNN 卷积核数量	CNN 池化方法	BiLSTM 层数	BiLSTM 隐层大小	dropout	batch_size
128	3 000	120	{2,3,4,5,6,7}	64	max pooling	2	128	0.5	128

2.2 实验结果分析

为了验证各个模型的分类性能,将融合 CNN 和 BiLSTM 特征的模型(CNN_BiLSTM)与本文改进的融合 CNN 和 Att-BiLSTM 特征的模型(CNN_Att-BiLSTM)、融合 MCNN 和 Att-BiLSTM 特征后输入 XGBoost 分类器的模型(MCNN_Att-BiLSTM)在 3 个数据集上进行了测试.

各个模型在 3 个数据集 1 次实验的损失函数如图 6 所示,其中横轴表示迭代数,纵轴表示训练集上的损失函数值.由实验数据可知,在多次迭代后,损失函数值波动较小且趋于稳定.在 IMDB 数据集上,CNN_BiLSTM 模型损失函数值趋向于 0.387,CNN_Att-BiLSTM 模型损失函数值趋向于 0.318,MCNN_Att-BiLSTM 模型损失函数值趋向于 0.267.在 txt-sentoken 数据集上,3 个模型的损失函数值比较接近,都趋向于 0.000 1.在酒店评论数据集上,CNN_BiLSTM 模型损失函数值趋向于 0.127,CNN_Att-BiLSTM 模型损失函数值趋向于 0.032,MCNN_Att-BiLSTM 模型损失函数值趋向于 0.071.

从上述数据和图 6 中可以看出,CNN_Att-BiLSTM 模型和 MCNN_Att-BiLSTM 模型的损失函数比 CNN_BiLSTM 模型收敛速度快,且损失函数值低于 CNN_BiLSTM 模型.每个模型在 3 个数据集上都取得了很好的收敛效果.



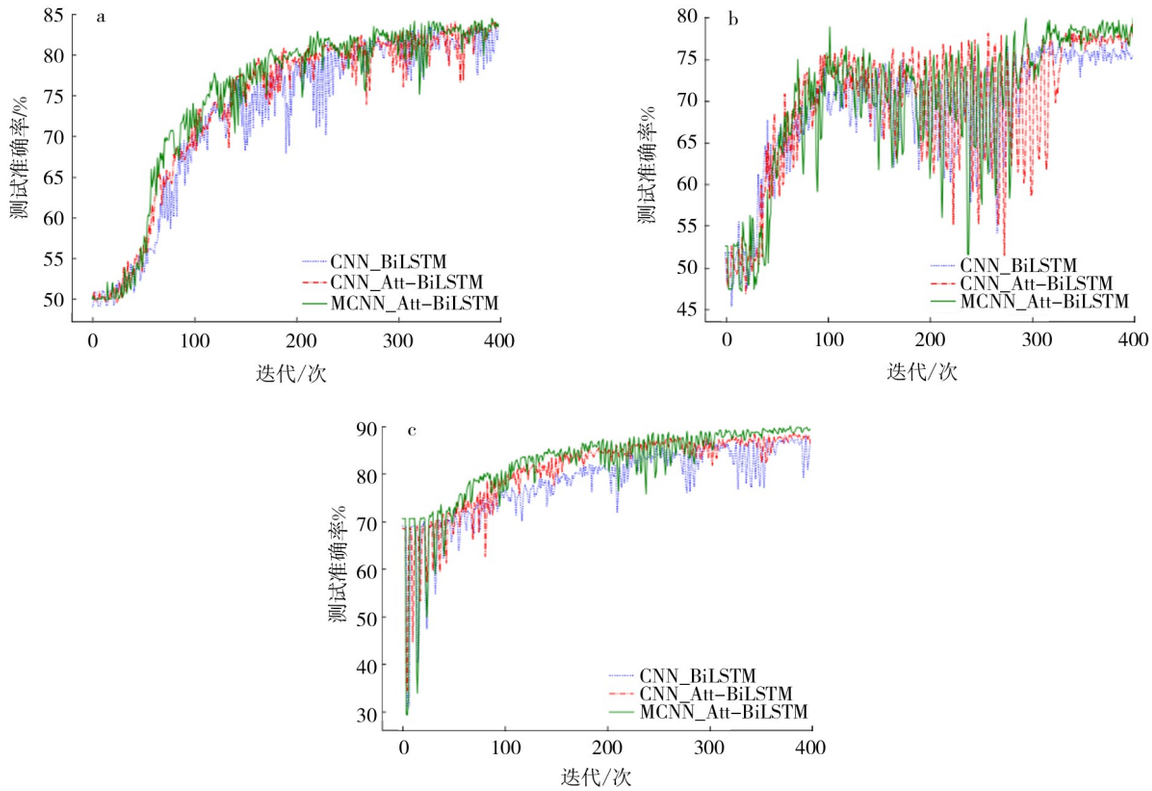
a.IMDB;b.txt-sentoken;c.谭松波-酒店评论.

图 6 各数据集的损失函数

Fig.6 Loss of each data set

各个模型在 3 个数据集上的准确率结果如图 7 所示,其中横轴表示迭代数,纵轴表示测试集准确率.从图 7 中可以看出 CNN_Att-BiLSTM 模型和 MCNN_Att-BiLSTM 模型的准确率比 CNN_BiLSTM 模型波动较小.经过计算 CNN_Att-BiLSTM 模型和 MCNN_Att-BiLSTM 模型准确率均高于 CNN_BiLSTM 模型.

本文各模型在 3 个数据集的上的准确率、召回率、F 值分别如表 3、表 4、表 5 所示.从表中可以发现,本文改进的 MCNN_Att-BiLSTM 模型在准确率、召回率及 F 值方面均优于其他模型.在准确率方面,增加注意力机制的 CNN_Att-BiLSTM 模型比 CNN_BiLSTM 模型在 IMDB 数据集上准确率提升了 0.70%,在 txt-sentoken 数据集上提升了 0.66%,在酒店评论数据集上提升了 2.89%.增加卷积核和注意力机制的 MCNN_Att-BiLSTM 模型比 CNN_BiLSTM 模型在 IMDB 数据集上准确率提升了 1.75%,在 txt-sentoken 数据集上提升了 1.67%,在酒店评论数据集上提升了 3.81%.



a.IMDB;b.txt-sentoken;c.谭松波-酒店评论.

图 7　各模型的准确率

Fig.7　Accuracy of each model

表 3　各个模型在测试集上的准确率

Tab.3　Accuracy of each model on test set

模型	准确率/%		
	IMDB	txt-sentoken	酒店评论
CNN_BiLSTM	81.34	76.45	84.89
CNN_Att-BiLSTM	82.04	77.11	87.78
MCNN_Att-BiLSTM	83.09	78.12	88.70

表4 各个模型在测试集上的召回率

Tab.4 Recall rate of each model on test set

模型	召回率/%		
	IMDB	txt-sentoken	酒店评论
CNN_BiLSTM	77.0	76.5	83.5
CNN_Att-BiLSTM	80.1	78.1	86.7
MCNN_Att-BiLSTM	80.9	79.6	87.5

表5 各个模型在测试集上的 F 值

Tab.5 F value of each model on the test set

模型	F 值/%		
	IMDB	txt-sentoken	酒店评论
CNN_BiLSTM	76.8	76.4	84.0
CNN_Att-BiLSTM	78.6	78.0	86.9
MCNN_Att-BiLSTM	79.8	78.7	87.0

在召回率和 F 值方面,融合 CNN 和 BiLSTM 特征模型在 IMDB 数据集上召回率达到 77.0%,F 值达到 76.8%,在 txt-sentoken 数据集上召回率达到 76.5%,F 值达到 76.4%,在酒店评论数据集上召回率达到 83.5%,F 值达到 84.0%.而增加注意力机制的 CNN_Att-BiLSTM 模型考虑了特征对分类结果的不同影响,使得召回率和 F 值都得到了提升,在 IMDB 数据集上召回率达到 80.1%,F 值达到 78.6%,在 txt-sentoken 数据集上召回率达到 78.1%,F 值达到 78.0%,在酒店评论数据集上召回率达到 86.7%,F 值达到 86.9%.本文模型 MCNN_Att-BiLSTM 通过增加卷积核和注意力机制,不仅提取了词语间全面的情感信息,而且还考虑了特征对分类结果的不同影响,召回率得到了进一步的提高.在 IMDB 数据集上召回率达到 80.9%,F 值达到 79.8%,在 txt-sentoken 数据集上召回率达到 79.6%,F 值达到 78.7%,在酒店评论数据集上召回率达到 87.5%,F 值达到 87.0%.

3 结束语

本文在 CNN_BiLSTM 模型基础上,改进了 2 种模型分别为 CNN_Att-BiLSTM 模型和 MCNN_Att-BiLSTM 模型.2 种模型能够充分考虑词语间局部特征和上下文特征中的重要程度,同时提高了训练速度.在接下来的工作中,将研究词语语义特征对分类结果的影响,即将语义特征和词向量进行融合,通过上述模型实现情感分类.

参 考 文 献:

[1] 李然,林政,林海伦,等.文本情绪分析综述[J].计算机研究与发展,2018,55(1):30-52. DOI:10.7544/issn1000-1239.2018.20170055.

[2] NASUKAWA T, YI J. Sentiment analysis: capturing favorability using natural language processing[Z].The international Conference on Knowledge Capture, Sanibel Island,2003. DOI:10.1145/945645.945658.

[3] 韩彤晖,杨东强,马宏伟.一种利用情感词统计信息构造文本特征表示的方法[J].计算机应用研究,2018,36(7):2087-2092. DOI:10.19734/j.issn.1001-3695.2018.01.0035.

[4] CHEN L, YANG Y C. Emotional speaker recognition based on i-vector through Atom Aligned Sparse Representation[Z].IEEE International Conference on Acoustics, Speech and Signal Processing,Vancouver,BC, Canada. DOI:10.1109/icassp.2013.6639174.

[5] 阳庆玲,郑志伟,邱佳玲,等.基于表情符号的文本情感分析研究[J].现代预防医学,2019,46(9): 1537-1540.

[6] RATHI M, MALIK A, VARSHNEY D, et al. Sentiment analysis of tweets using machine learning approach[Z].Eleventh International Conference on Contemporary Computing (IC3), Noida,2018. DOI:10.1109/ic3.2018.8530517.

[7] LE H T, CERISARA C, DENIS A. Do convolutional networks need to be deep for text classification[Z].The AAAI Conference on Artificial Intelligence,New Orleans,Louisiana,2018.

[8] 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. DOI:10.7544/issn1000-1239.2018.20170049.

[9] IRSOY O, CARDIE C. Deep recursive neural networks for compositionality in language[J]. Advances in neural information processing systems,2014,3:2096-2014.DOI:10.5555/2969.2969061.

[10] 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. DOI:10.11772/j.issn.1001-9081.2018041289.

[11] ARAQUE O, CORCUERA-PLATAS I, SÁNCHEZ-RADA J F, et al. Enhancing deep learning sentiment analysis with ensemble techniques in social applications[J]. Expert Syst Appl, 2017, 77: 236-246. DOI:10.1016/j.eswa.2017.02.002.

[12] 苏兵杰,周亦鹏,梁勋鸽.基于 XGBoost 算法的电商评论文本情感识别模型[J].物联网技术,2018,8(1): 54-57. DOI:10.16667/j.issn.2095-1302.2018.01.015.

[13] 龚维印,王力.基于卷积神经网络和 XGBoost 的文本分类[J].通信技术,2018,51(10): 2337-2342. DOI:10.3969/j.issn.1002-0802.2018.10.012.

[14] OKADA S, OHZEKI M, TAGUCHI S. Efficient partition of integer optimization problems with one-hot encoding[J]. Sci Rep, 2019, 9(1): 1-12. DOI:10.1038/s41598-019-49539-6.

[15] 於雯,周武能.基于 LSTM 的商品评论情感分析[J].计算机系统应用,2018,27(8): 159-163. DOI:10.15888/j.cnki.csa.006483.

[16] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification[Z]. The 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016. DOI:10.18653/v1/p16-2034.

[17] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system[Z]. The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco,2016. DOI:10.1145/2939672.2939785.

[18] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system[Z]. The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2016. DOI:10.1145/2939672.2939785.

[19] PANG B, LEE L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts[Z]. The 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain, 2004. DOI: 10.3115/1218955.1218990.

[20] 蔡林森,彭超,陈思远,等.基于多样化特征卷积神经网络的情感分析[J].计算机工程,2019,45(4): 169-174,180. DOI:10.19678/j.issn.1000-3428.0050338.

(责任编辑:孟素兰)