文章编号:1007 - 2985(2024)01 - 0024 - 06

基于卷积神经网络的特定目标文本情感分析模型*

叶海燕
(巢湖学院计算机与人工智能学院,安徽 巢湖 238000)

摘　要:在特定目标文本情感分析过程中,文本序列分类受标注方式的限制,导致分析结果的准确率和召回率较低.为了解决这个问题,构建了基于卷积神经网络的特定目标文本情感分析模型(文本分析模型).根据情感差异分析特定目标文本序列,在输入层将文本特征矩阵作为卷积神经网络语言模型的输入数据,拼接成词性序列矩阵;分段池化捕获文本序列不同的关键特征,并分类处理提取到的特征向量;加入 dropout 机制完成特定目标文本情感分类,确定文本中每个词的重要度信息,实现特定目标文本情感分析.实验结果表明,文本分析模型的准确率高于 84%,召回率最大值为 87%,能够有效实现特定目标文本情感分析.

关键词:卷积神经网络;特定目标;dropout 机制;文本情感
中图分类号:TP391.1　　**文献标志码**:A　　**DOI**:10.13438/j.cnki.jdzk.2024.01.005

人们通过社交网络平台评论热点话题时,会产生大量的具有情感色彩的文字信息,根据这些信息可以看出人们的偏好.由于文字信息数量迅速增长,人工标注方式不能及时地获取用户偏好,因此有学者设计了不同的算法来解决情感分析问题.例如,王卫红等[1]构建了基于 Spark 的情感分析集成算法,该算法通过 Spark 分布式计算方式实现情感分析集成,再结合机器学习算法完成特定目标文本情感分析;刘振宇等[2]构建了基于主动学习和多种监督学习的分析集成算法,该算法融合主动学习技术,利用训练分类器获取综合投票结果,以此分类情感信息.考虑到文本分析时过多地进行人工标注,会使得复杂情感分类行为的泛化能力受到限制,导致分析结果不精准,而卷积神经网络[3]可以将包含不同情感色彩的文本当作一维图像,通过捕捉邻近词之间的关联实现更精准的分类.因此,为了进一步优化特定目标文本情感分析效果,笔者拟设计一种基于卷积神经网络的特定目标文本情感分析模型(简称“文本分析模型”),通过神经网络模型的训练与输出,实现特定目标文本情感分析.

1　基于卷积神经网络语言模型的特定目标文本序列分析

1.1 卷积神经网络语言模型的构建

卷积神经网络语言模型如图 1 所示.从图 1 可知,每个词都是通过矩阵转化为词向量形式.该矩阵是由词汇的维数和词典(词汇的总数)构成的.

卷积神经网络可以划分为输入层、隐层和输出层.与一般的神经网络相似,卷积神经网络的隐层输出

* **收稿日期**:2023 - 02 - 24
基金项目:安徽省质量工程省级教学研究一般项目(2020JYXM1253);安徽省省级教学示范课程(2020SJJXSFK1720);安徽高校自然科学研究重点项目(KJ2020A0681)
作者简介:叶海燕(1983—),女,安徽巢湖人,巢湖学院计算机与人工智能学院讲师,硕士生,主要从事物联网工程、情感计算和深度学习研究.

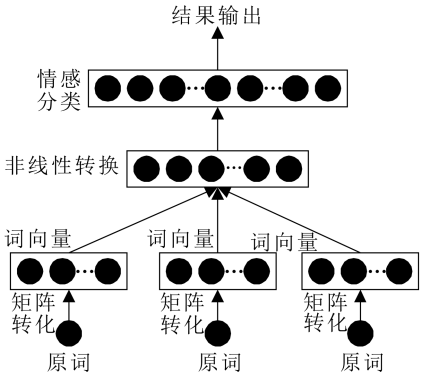

图 1　卷积神经网络语言模型

Fig. 1　Convolutional Neural Network Language Model

结果(y_1)可表示为

$$y_1=\tan(\lambda_1+\omega_1 x).$$

其中:ω_1 为从输入层到隐层的权重;λ_1 为隐层偏置项;x 为输入数据.对 y_1 进行非线性转换,并将其输入到下一层[4].

输出层利用 softmax 分类器将输出的数值转化成一个可能的数值[5],具体描述为

$$y_2=\lambda_2+\omega_1 x+\omega_2 y_1.$$

其中:y_2 为输出层利用 softmax 分类器将输出数值转化后的结果;ω_2 为从隐层到输出层的权重;λ_2 为输出层偏置项.

通过不断地迭代,直至网络整体的错误最少[6],再经过不断地优化与更新,获得与特定目标文本原词对应的词向量.

1.2 特定目标文本序列分析

利用卷积神经网络语言模型可进行特定目标文本序列分析.首先,对一个静态词向量矩阵进行迭代处理[7-9].在这个矢量中,每一个词向量都能得到最靠近它的语义的 k 个近邻向量[10].余弦相似性是一种常见语义测量方法.它是先计算静态词向量矩阵中的词向量与其他词向量之间的相似性,再按余弦相似度的递减顺序,将最前面的 k 个词向量组合起来,即词向量的 k 个邻近向量[11].然后,根据情感差异对 k 个词向量的邻近向量进行重新排列,使其排列顺序愈靠前愈好[12].最后,计算词向量 $\boldsymbol{\gamma}_i$ 与其邻居词向量 $\boldsymbol{\gamma}_j$ 的情感差异 $D(\boldsymbol{\gamma}_i,\boldsymbol{\gamma}_j)$,计算公式为

$$D(\boldsymbol{\gamma}_i,\boldsymbol{\gamma}_j)=\left|c(\boldsymbol{\gamma}_i)-c(\boldsymbol{\gamma}_j)\right|,$$

其中 $c(\boldsymbol{\gamma}_i)$,$c(\boldsymbol{\gamma}_j)$ 分别为词向量 $\boldsymbol{\gamma}_i$,$\boldsymbol{\gamma}_j$ 的情感.$D(\boldsymbol{\gamma}_i,\boldsymbol{\gamma}_j)$ 越小,说明差异越小,则让 $\boldsymbol{\gamma}_i$ 靠近 $\boldsymbol{\gamma}_j$;反之,说明差异越大,则让 $\boldsymbol{\gamma}_i$ 远离 $\boldsymbol{\gamma}_j$.

2　文本分析模型的构建与训练

2.1 模型的构建

考虑到文本类型的特征和句中情绪取向的变化对句子的情绪类型有很大的影响,因此,为了获取不同特定目标文本序列特征,采用分段池化构建分段单元,如图 2 所示.

由图 2 可知,将基于卷积神经网络文本序列特征提取输出结果分为多个卷积向量,对这些卷积向量进行最大池化处理,由此可以获取相关序列特征.拼接这些特征,通过全连接层实现特定目标文本情感分析[13].文本分析模型构建的详细步骤如下:

(ⅰ)构建二维矩阵.将特定目标文本转换为矩阵形式后输入到卷积神经网络中,再将其与卷积层相结合[14].首先,分割并标记具有维度为 n 的语句,再将它们的词类与词性结合起来,通过 Word2vec 转化成相应的词向量形式.然后,在特定目标文本序列中选择维度为 n 的单词,将其拼接成词性序列矩阵 $\boldsymbol{E}$,即

$$\boldsymbol{E}=(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n),$$

其中 $\boldsymbol{e}_n$ 为词性序列矩阵 $\boldsymbol{E}$ 中维度为 n 的元素.因为每个词性序列矩阵都存在多个词向量,所以构成的

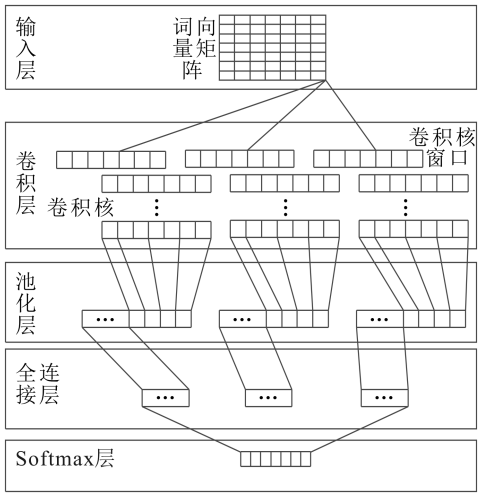

图 2　基于卷积神经网络文本序列特征提取

Fig. 2　Text Sequence Feature Extraction Based on Convolutional Neural Network

矩阵就是二维矩阵[15].

(ⅱ)卷积运算.在卷积运算过程中,利用 $h*n$ 卷积核来抽取高层次的文本序列特征,其中 h 为要滑动的相邻词数量.卷积操作是用卷积法对输入的特征矩阵进行卷积运算[16].通过卷积法进行特征映射,能够有效地改善模型的分类性能.

(ⅲ)分段池化处理.由于在卷积运算中易出现非线性降维的问题,因此在卷积运算过程中加入一个池操作,通过分段池化处理得到一维特征矩阵.将卷积通道统一化,以最大值作为向量特征,用公式表示为

$$a_{\text{pool}}=\max(a_i).$$

其中:a_{pool} 为分段池化处理结果;a_i 为第 i 个特定文本情感特征序列,$i=1,2,\cdots,I$.

(ⅳ)情感分类.分段池化处理后,在每次训练时加入 dropout 机制来剔除一部分神经元,再将原有文本信息输入到分类器中,进行情感分类.这样可以较好地筛选出对分类有影响的特征.情感分类结果(m)的计算公式为

$$m=\frac{1}{1+\exp(-\beta x)},\tag{1}$$

其中 β 为所有分类参数.

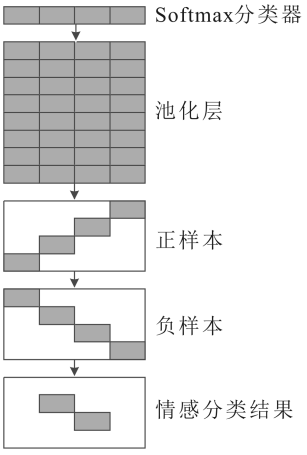

图 3 情感分类模型

Fig. 3 Emotion Classification Model

情感分类模型如图 3 所示.由图 3 可知,使用 Softmax 分类器与池化层相连,分类处理正样本和负样本问题,并结合(1)式能够确定分类结果.

(ⅴ)输出结果.采用注意力模型生成文本分析模型.文本分析模型的第 i 个特定文本量化结果用 δ_i 表示,$\delta_i=\sum_{i=1}^{I}p_i r_i$.其中:$r_i$ 为第 i 个特定目标文本,$r_i=f_\delta(x_i,t_i)$;p_i 为第 i 个特定目标文本中词与输出结果之间的相关性,

$$p_i=\frac{\exp(r_i)}{\sum_{i=1}^{I}\exp(r_i)}.$$

这里 t_i 为第 i 个特定目标文本中句子对应的标签,f_δ 为隐层的前向网络.

2.2 模型的训练

在卷积神经网络中,当文本数据量很大时,采用梯度下降算法会出现存储量不足、训练迭代率低等问题;利用随机梯度下降法可以快速地进行训练,但是资料太少会导致过度拟合.批量随机梯度下降算法在提高模型学习效率的同时,还能减少每个迭代方差的影响,因此利用批量随机梯度下降算法训练模型.训练目的是使模型的损耗函数

$$\psi=-\sum_{i=1}^{I}\hat{y}_i\ln y_i$$

最小.其中:ψ 为模型训练后的输出结果;$\hat{y}_i$ 为句子中第 i 个特定目标文本情感实际类别;y_i 为第 i 个特定目标文本分析结果为正确的类型.在模型训练过程中,设定学习速率以防止过度拟合,设定动量以防止在最低处出现失真.在输出结果与预期值不一致的情况下,按照各个层的传送次序,采用批量随机梯度下降算法进行调整,以减小最终的特定目标文本情感分析误差,从而实现基于卷积神经网络的特定目标文本情感分析.

3 实验部分

3.1 实验数据准备

为了验证文本分析模型的有效性,选用有关特定目标文本情感挖掘的外卖评论语料[17]作为基础数据

进行实验.该语料规模为 1 500 篇,分为 4 个子集,本实验筛选其中 5 000 个数据,正样本和负样本数据各 2 500个.语料样例见表 1.

表 1　外卖评论语料样例

Table 1　Sample Food Delivery Review Corpus

积极评论	消极评论
饭菜很好吃,很干净,而且味道真的很不错!	为什么饭菜有一股馊了的味道,以后我不会再买了.
非常好吃的外卖,太可口了,明天还会订的!	我肯定不会再订了,味道太差了.
很干净,出单速度也很快,出现送错的问题也及时采取了退款措施,服务超级好!	我觉得这个饭菜与价格不匹配,根本不值这个钱.
在网上搜到她家外卖销售量超级高,订了之后果然好吃,下次还会再订的!	老板,味道不太好啊,让厨师少放盐.

按照(1)式划分情感词典部分,示例见表 2.

表 2　情感词典部分示例

Table 2　Example of Sentiment Dictionary Section

正样本	正确分类得分	负样本	错误分类得分
很好吃	2.878 788	馊味	−3.287 834
很干净	1.968 377	不会	−2.098 374
很不错	2.093 818	味道差	−3.123 944
可口	2.293 874	不匹配	−3.094 857
速度快	2.394 056	不值	−2.938 475
还会	2.349 760	不太好	−2.384 755

3.2 实验指标

采用准确率(A)、召回率(R) 作为评估指标:

$$A=\frac{P_{\mathrm{T}}}{P_{\mathrm{T}}+P_{\mathrm{F}}},R=\frac{P_{\mathrm{T}}}{P_{\mathrm{T}}+N_{\mathrm{F}}}.$$

其中:P_{T} 表示特定目标文本中情感分析为正确的正样本;P_{F} 表示特定目标文本中情感分析为错误的负样本;N_{F} 表示特定目标文本中情感分析为错误的正样本.

3.3 实验结果

将文本分析模型与基于 Spark 的情感分析集成算法[1]、基于主动学习和多种监督学习的分析集成算法[2]进行对比实验,精确度和召回率结果如图 4,5 所示.

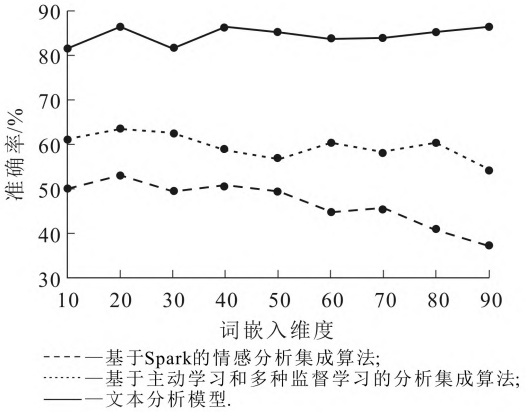

图 4　3 种方法的准确率

Fig. 4　Accuracy of Three Methods

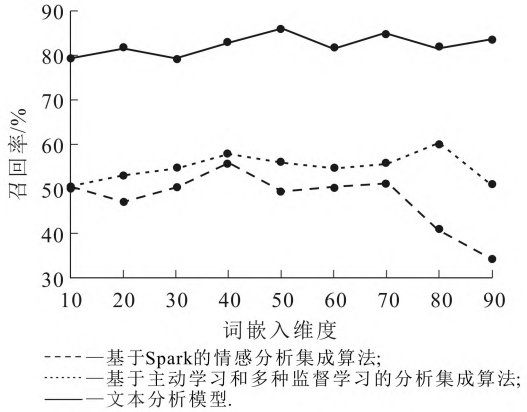

图 5　3 种方法的召回率

Fig. 5　Recall Rates of Three Methods

由图 4 可知:基于 Spark 的情感分析集成算法的准确率最低,随着词嵌入维度的增加,准确率整体呈下降趋势,当词嵌入维度为 20 时,准确率最高为 59%;基于主动学习和多种监督学习的分析集成算法的准确率其次,随着词嵌入维度的增加,准确率变化幅度较小,当词嵌入维度为 20 时,准确率最高为 68%;文本

分析模型的准确率最高,随着词嵌入维度的增加,准确率变化幅度较小,当词嵌入维度为90时,准确率最高为87%.

由图5可知:基于Spark的情感分析集成算法当词嵌入维度在40时,召回率最高为55%;基于主动学习和多种监督学习的分析集成算法当词嵌入维度在80时,召回率最高为62%;文本分析模型当词嵌入维度为50时,召回率最高为87%.

4 结语

为了提高文本分析的准确率,笔者采用卷积神经网络的文本分类方法构建了新的文本情感分析模型,并利用该模型对文字进行编码,得到相应的文本特征.以准确率、召回率作为评估指标,将文本分析模型与基于Spark的情感分析集成算法和基于主动学习和多种监督学习的分析集成算法进行对比实验,结果表明,文本分析模型具有较高的准确率和召回率,说明该模型更可行、有效.然而,现实生活中的情感信息往往不仅通过文本表达,还通过图片、视频等多种形式表达.图像信息包含了丰富的视觉特征,能够传达更加具体和直观的情感信息,而不同的人对于同一件事物的感受在文本和图像中的表达也可能不同.因此,在后续工作中,笔者将深入探讨文本与图像信息相结合的情感分析问题,并寻求一种更合适的分析方法,确保情感分析不再局限于文本.

参考文献:

[1] 王卫红,金凌剑.基于Spark的情感分析集成算法[J].浙江工业大学学报,2020,48(4):405-410;434.

[2] 刘振宇,李钦富,杨硕,等.一种基于主动学习和多种监督学习的情感分析模型[J].中国电子科学研究院学报,2020,15(2):171-176.

[3] 汤凌燕,熊聪聪,王媛,等.基于深度学习的短文本情感倾向分析综述[J].计算机科学与探索,2021,15(5):794-811.

[4] LU Xinxin,ZHANG Hong.An Emotion Analysis Method Using Multi-Channel Convolution Neural Network in Social Networks[J].Computer Modeling in Engineering & Sciences,2020,125(1):281-297.

[5] WANG Qingqing,LUO Jianlin,SONG Jianwen.Emotion Analysis Method for Elderly Living Alone Based on CNN-BGRU Neural Network[J].International Journal of Wireless and Mobile Computing:IJWMC,2021,20(4):352-362.

[6] HAJEK PETR,BARUSHKA ALIAKSANDR,MUNK MICHAL.Fake Consumer Review Detection Using Deep Neural Networks Integrating Word Embeddings and Emotion Mining[J].Neural Computing and Applications,2020,32(23):17259-17274.

[7] DEBORAH ANGEL S.MIRNALINEE T T,MILTON RAJENDRAM S.Emotion Analysis on Text Using Multiple Kernel Gaussian[J].Neural Processing Letters,2021,53(2):1187-1203.

[8] 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(1):16-22.

[9] 王家乾,龚子寒,薛云,等.基于混合多头注意力和胶囊网络的特定目标情感分析[J].中文信息学报,2020,34(5):100-110.

[10] 袁勋,刘蓉,刘明.融合多层注意力的方面级情感分析模型[J].计算机工程与应用,2021,57(22):147-152.

[11] 孙佳慧,韩萍,程争.基于知识迁移和注意力融合的方面级文本情感分析[J].信号处理,2021,37(8):1384-1391.

[12] 杨书新,张楠.融合情感词典与上下文语言模型的文本情感分析[J].计算机应用,2021,41(10):2829-2834.

[13] 李文亮,杨秋翔,秦权.多特征混合模型文本情感分析方法[J].计算机工程与应用,2021,57(19):205-213.

[14] 李辉,黄钰杰.基于多头自注意力和并行混合模型的文本情感分析[J].河南理工大学学报(自然科学版),2021,40(1):125-132.

[15] 杨长利,刘智,鲁明羽.双通道混合神经网络的文本情感分析模型[J].计算机工程与应用,2020,56(11):124-128.

[16] 韩开旭,黎永壹,邱桂华,等.基于分段卷积神经网络的文本情感极性分析[J].计算机仿真,2020,37(6):361-364;378.

[17] 胡胜利,张丽萍.基于ALBERT-CNN的外卖评论情感分析[J].现代信息科技,2022,6(10):157-160.

Sentiment Analysis Model of Specific Target Text Based on Convolutional Neural Network

YE Haiyan

(College of Information and Artificial Intelligence Engineering, Chaohu University, Chaohu 238000, Anhui China)

Abstract: In the process of emotion analysis of specific target text, text sequence classification is limited by the labeling method, resulting in low accuracy and recall of analysis results, so a model of emotion analysis of specific target text based on convolution neural network is constructed. Specific target text sequences are analysed based on emotional differences, and the text feature matrix is used as input data to construct a convolutional neural network language model in the input layer, concatenating it into a part of speech sequence matrix. Segmented pooling captures different key features of text sequences and classifies and processes the extracted feature vectors. A dropout mechanism is added to complete sentiment classification of specific target texts, determine the importance information of each word in the text, and achieve sentiment analysis of specific target texts. The experimental results show that the accuracy of the model is higher than 84%, and the maximum recall rate is 87%, which can effectively achieve the emotional analysis of specific target text.

Key words: convolution neural network; specific objectives; dropout mechanism; text emotion

(上接第 23 页)

[14] CHOUHAN J, RAI M, SABRI M S. A Literature Survey of Different Data Encryption Algorithm for Secure Health Care System in Cloud[J]. SSRN Electronic Journal, 2021, 10(3): 1 - 7.

[15] XUAN Chunqing. Design of Wireless Sensor Network Data Acquisition System via Health Sensor Based on Symmetric Encryption Algorithm[J]. Journal of Testing and Evaluation: A Multidisciplinary Forum for Applied Sciences and Engineering, 2023, 51(1): 278 - 290.

Electronic File Uploading Encryption Method Based on Chaotic Sequence

XU Debin[1,2]

(1. Office of Hefei Vocational and Technological College, Hefei 230012, China; 2. School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: In order to improve the transmission security of electronic files and the security of user information, a chaotic sequence based encryption method for uploading electronic files is studied. The formal scheme is used to decompose the steps of uploading electronic files, and the form of uploading is defined uniformly in the interactive process. Logistic mapping is used to represent chaotic mode. Given the interval of chaotic sequence, the file encryption key stream is randomly generated. On the premise of ensuring the complete decryption of electronic files, the replacement process is constructed by setting ciphertext replacement rules, adjusting ciphertext order according to reversible relationship, realizing the encryption and uploading of electronic files, and completing the method design. The experiment simulates the whole process of uploading electronic files based on the configuration of optical fiber network. The chaotic sequence method can generate 128 bit encryption key within 10 s and complete 498 groups of electronic files with memory size of 1.5 GB within a specified time.

Key words: ciphertext replacement; electronic records; chaotic sequence; logistic mapping

(责任编辑　向阳洁)