

文章编号: 2095-2163(2024)02-0187-04

中图分类号: TP391.1

文献标志码: A

# 基于星级权重和双向长短期记忆神经网络的 豆瓣网影评情感分析

吴俊杰, 陈佳俊

(太原师范学院 计算机科学与技术学院, 山西 榆次 030619)

**摘要:** 应用神经网络模型对影评数据进行预测时,会出现预测结果跟观众本意不相符的情况,本文提出了一种基于星级权重和双向长短期记忆网络的模型,全面考虑星级评分和上下文信息,还能对长距离依赖关系进行有效建模。实验结果表明,该模型有效改善了预测结果与观众本意不相符的情况,使得情感分析预测的结果具有更准确的分类精度。

**关键词:** 星级权重; 影评; 神经网络; 情感分析

## Sentiment analysis of Douban movie reviews based on star rating weight and bidirectional LSTM neural network

WU Junjie, CHEN Jiajun

(College of Computer Science and Technology, Taiyuan Normal University, Yuci Shanxi 030619, China)

**Abstract:** When applying neural network models to predict movie review data, discrepancies can occur between the predicted results and the actual intentions of the audience. This paper proposes a model based on star-level weighting and bidirectional long short-term memory networks, which comprehensively considers star ratings and contextual information, and can effectively model long-distance dependency relationships. The experimental results show that this model significantly improves the match between predicted results and audience intentions, thereby enhancing the accuracy of sentiment analysis predictions.

**Key words:** star rating weight; movie review; neural network; sentiment analysis

## 0 引言

随着互联网上电影评论的兴起,人们可以更轻松地分享自己的看法和评价。然而,网上评论不断扩大,用户需要花费大量时间和精力来筛选和阅读评论,以获取有关电影的有效信息。对这些海量的情感文本进行数据分析,可使用户有更好的体验感。

最初的文本情感分析主要是基于情感词典,陈晓东<sup>[1]</sup>用扩展的情感倾向点互信息算法,设立了一个中文微博领域的情感词典,这种方法的局限性在于其依赖于手工构建的特征。由于上述方法完成效率和质量不高,人们开始应用深度学习构建网络模型进行文本分类工作<sup>[2]</sup>。其中最具代表性的是利用卷积神经网络(Convolutional Neural Networks, CNN)进行关系分类。但CNN不适合学习长距离的语义信息<sup>[3]</sup>。基于深度学习神经网络的方法大多

采用word2vec进行词向量化,但由此获得的都是静态的文本特征,舍弃了大量位置信息<sup>[4]</sup>。递归神经网络(Recursive Neural Network, RNN)无法处理长依赖的情况<sup>[5]</sup>;相比之下,长短记忆循环神经网络(Long Short Term Memory, LSTM)可以有效处理长距离依赖关系,而且能够存储更多信息,Tai等<sup>[6]</sup>提出了一种树形长短期记忆网络模型,在语义关联和情感分类方面取得了良好成果;Baziotis等<sup>[7]</sup>将注意力机制引入LSTM中,在SemEval-2017 Task4 for Twitter的情感分析中取得了良好结果。

综合以上研究成果,本文提出了一种基于星级权重和双向长短期记忆网络的模型来进行豆瓣网影评文本情感分析,在自行爬取的豆瓣网数据集上对不同模型进行多组对比试验,实验结果表明该模型具备更好的分类准确率。

**作者简介:** 陈佳俊(1999-),男,硕士研究生,主要研究方向:软件开发。

**通讯作者:** 吴俊杰(1974-),男,硕士,副教授,硕士生导师,主要研究方向:数据挖掘与人工智能处理。Email: mygodrue@gmail.com

收稿日期: 2023-11-07

## 1 相关模型简介

传统的循环神经网络模型无法捕捉长距离的语义联系,并且难以保留较早时刻的信息。在参数训练过程中,梯度会逐步减小直至消逝,限制了序列数据的长度。为了解决这个问题,长短时记忆网络(LSTM)引入了输入门*i*、输出门*o*、遗忘门*f*和记忆单元以解决梯度问题和信息遗忘的困扰,能够更有效地处理长序列数据,提高了循环神经网络在实际任务中的性能。LSTM网络结构如图1所示。

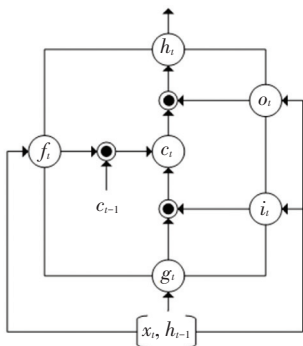


图1 LSTM网络结构图

Fig. 1 Structure of LSTM

遗忘门*f*选定最后一刻要从记忆单元中遗忘的信息,输出值介于0和1之间,计算公式(1):

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f) \quad (1)$$

其中, $x_t$ 代表当前时刻的输入; $h_{t-1}$ 代表前一时刻的隐藏状态; $c_{t-1}$ 代表前一时刻的记忆单元; $W_{fx}$ 、 $W_{fh}$ 、 $W_{fc}$ 是相应输入的权重矩阵; $b_f$ 是遗忘门的偏置项; $\sigma$ 是sigmoid激活函数。

在每个当前时间步骤中,输入门*i*可以使用公式(2)~公式(4)更新记忆单元中的信息:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i) \quad (2)$$

$$g_t = \tanh(W_{gx}x_t + W_{gh}h_{t-1} + W_{gc}c_{t-1} + b_g) \quad (3)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot g_t \quad (4)$$

其中, $\tanh$ 是双曲正切激活函数; $f_t$ 和*i*分别指 $c_{t-1}$ 和 $g_t$ 的权值; $b_i$ 和 $b_g$ 是偏置项。

输出门*o*用来确定LSTM网络的输出值,公式(5)和公式(6):

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}c_{t-1} + b_o) \quad (5)$$

$$h_t = o_t \cdot \tanh(c_t) \quad (6)$$

其中, $W_{ox}$ 是 $x_t$ 和 $o_t$ 的连接权重; $W_{oh}$ 是 $h_{t-1}$ 和 $o_t$ 的连接权重; $W_{oc}$ 是 $c_{t-1}$ 和 $o_t$ 的连接权重; $b_o$ 是偏置项。

为了能够更全面地处理上下文信息和长距离依

赖的任务,本文引入双向长短期记忆神经网络(BiLSTM),即在LSTM的基础上加入了一个反向的结构,以增加对序列信息的全面性和上下文的理解。

## 2 影评情感分析模型构建

前期准备工作中遇到了小样本语料的向量稀疏问题。本文将分词处理后的每个词输入到Fastest中文预训练词向量模型中,得到每个词的词向量,解决小样本语料的向量稀疏问题。

### 2.1 构建神经网络

针对影评领域模型构建存在的问题以及当前影评情感分析方法的缺点,本文提出了一种基于星级权重和双向长短期记忆网络的神经网络。现有针对影评领域的模型很少关注到星级特征,本文模型融合了星级特征,从影评本身出发进行研究,加入星级权重层得出更好的情感分类结果,并改善部分影评情感偏向不明显的状况;通过权重层修正预测分类结果,提高最终情感分类精度,模型结构如图2所示。

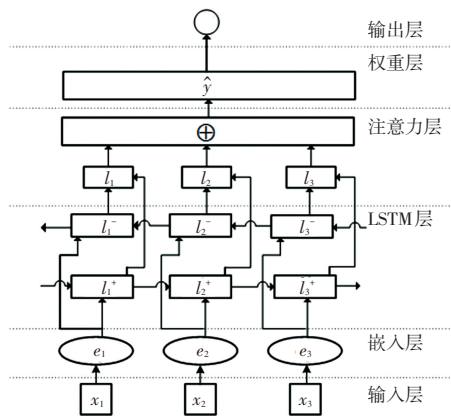


图2 基于星级权重和双向长短期记忆网络的神经网络结构

Fig. 2 The neural network architecture based on star rating weights and bidirectional LSTM

基于星级权重和双向长短期记忆网络模型包含6个组件:

输入层:将句子输入到该模型;

嵌入层:将每个单词映射为低维向量;

LSTM层:本文使用双向长短期记忆网络(BiLSTM)获取高级特征,该网络包括两个子网络,分别进行前向和后向交互。 $l_i$ 表示正向部分 $\vec{l}_i$ 和反向部分 $\overleftarrow{l}_i$ 的拼接结果,公式(7):

$$l_i = [\vec{l}_i \oplus \overleftarrow{l}_i] \quad (7)$$

注意力层:生成一个权重向量,并通过乘以权重向量将不同时间片的单词级特征合并成一个句子级

特征向量;

权重层: 将特征向量与星级权重进行融合;

输出层: 结合句子级特征向量和星级特征得出分类结果;

本文利用 softmax 分类器来预测句子的标签  $\hat{q}$ , 分类器以隐藏状态  $h^*$  作为输入, 公式(8):

$$\hat{q} = \text{softmax}(W_i h^* + b_i) \quad (8)$$

其中,  $b_i$  为偏置项。

再提取星级权重特征对  $\hat{q}$  进行修正, 得到最终分类, 修正公式(9):

$$S = \hat{q} + \frac{|S_i - W_i|}{W_i} \quad (9)$$

其中,  $S_i$  表示未加入星级权重层之前得出的概率值,  $W_i$  表示临界值。

### 2.2 星级权重分配算法

豆瓣网的影评星级分为 5 类: 五星、四星、三星、二星、一星。五星和一星有较大的话语权, 在星级权重分配的时候, 影评本身内容和星级之间的权重分配还是以影评本身为主, 星级主要作为一个辅助修正的特征点, 这样既能有更好的情感分析的准确率, 也能改善影评情感倾向不明显的问题。在星级权重分配的过程中五星和一星为一大类, 四星和二星为一中类, 三星单独为一小类。以五星为例, 从抽样影评中筛选出符合本研究的五星好评, 再通过公式(10) 计算出每条的情感因子值, 最后取均值作为该星级的权重值构建星级权重矩阵, 见表 1。

$$sen_i = \frac{|S_i - W_i|}{W_i} \quad (10)$$

表 2 部分影评情感倾向预测实验结果

Table 2 Results of movie review sentiment prediction

用户	影评	星级	正向情感概率值	负向情感概率值
1 加油	“一个人如果没有了爸爸妈妈该怎么办”	五星	0.417 5	0.582 5
2 热爱生活了吗	“或许只有武汉知道那场战争到底有多苦”	四星	0.404 3	0.595 7
3 咸咸的土豆片	“不要害怕眼前的黑暗, 因为总有人会点亮炬火”	五星	0.612 1	0.382 9
4 Seamoon	“和儿子一起看。我们总要不断面对十字路口, 这时就需要正确的方向和坚定的信念。”	三星	0.565 4	0.434 6

将本文的模型与 CNN (Convolutional Neural Networks)、RNN (Recurrent Neural Networks)、LSTM (Long Short Term Memory)、BiLSTM (Bi-directional Long Short-Term Memory) 4 种模型作对比, 实验结果见表 3。CNN 的情感分类效果较差,  $F1 - score$  为 85%; RNN 和 LSTM 的  $F1 - score$  可达到 87%,

表 1 星级权重矩阵

Table 1 Star level weight matrix

	五星	四星	三星	二星	一星
正向	+0.3	+0.17	+0.06	-0.17	-0.3
负向	-0.3	-0.17	-0.06	+0.17	+0.3

## 3 实验

### 3.1 数据预处理

本实验爬取了豆瓣网影评数据作为数据集, 先利用正则表达式对影评进行处理, 去除不符合语法和逻辑的影评, 再去掉停用词。数据集共包括 14 980 条电影评论数据, 训练集 11 017 条, 测试集 3 963 条。根据实际影评内容, 每条数据都直接被标记为正向评价或者负向评价。情感标签分为两类, 正向评价为 1, 负向评价为 0。

### 3.2 实验设置

实验平台配置: 操作系统 Windows 10, Intel i5-7300HQ, 8G 内存, GPU: NVIDIA GeForce, RTX 1050 ti 8G 显存; 深度学习框架 PyTorch; 编程语言 Python; 实验训练的参数为: 词向量维度 50、批处理尺寸 64、隐藏层节点数 128、训练迭代周期 10、学习率为 0.01。

### 3.3 实验结果及分析

本文采用精确度 ( $P$ )、召回率 ( $R$ ) 和  $F1 - score$  作为评价指标。基于星级权重和双向长短期记忆神经网络模型在正向样本和负向样本上的平均精确度达到了 90.21%, 平均召回率为 90.38%, 表明该模型能够有效地对影评的情感倾向进行准确的识别。部分影评情感倾向预测实验结果见表 2。

BiLSTM 的  $F1 - score$  达到了 91%。本文提出的基于星级权重和双向长短期记忆神经网络模型引入了 BiLSTM 结构和星级权重层, 能更有效地捕获上下文语义信息和星级特征信息, 因此情感分析效果最佳,  $F1 - score$  显著提高。