DOI:10.16652/j.issn.1004-373x.2020.01.020

基于改进深度学习算法的文本极性智能判断方法研究

宋思晗，王兴芬，杜惠英
（北京信息科技大学 信息管理学院，北京 100192）

摘 要：为了解决传统的文本极性智能判断方法判断结果准确率和召回率普遍较低的问题，基于改进深度学习算法研究一种新的文本极性智能判断方法。在CNN结构基础上设计一种新的深度学习算法模型，模型由输入层、输出层、采集层、连接层、卷积层五部分构成。使用该模型对文本进行智能判断，判断过程共有五步，分别是文本预处理、情感词提取、表情符号提取、感情倾向值计算和情感最终倾向值分析。为检测所提方法的有效性以及优越性，与传统判断方法进行实验对比，结果表明，基于改进深度学习算法的文本极性智能判断方法判断的准确率和召回率更高，发展空间更广阔。

关键词：文本极性；智能判断方法；算法模型设计；有效性检测；深度学习算法；文本预处理
中图分类号：TN911.1-34；TP393 **文献标识码**：A **文章编号**：1004-373X(2020)01-0076-04

Research on text polarity intelligent judgment method based on improved deep learning algorithm

SONG Sihan, WANG Xingfen, DU Huiying
(School of Information Management, Beijing Information Science & Technology University, Beijing 100192, China)

Abstract: The accuracy and recall rate of traditional text polarity intelligent judgment methods both are generally low. In view of the above, a new method of text polarity intelligent judgment is studied based on improved deep learning algorithm. A new deep learning algorithm model is designed based on the CNN structure. The model consists of five parts: input layer, output layer, acquisition layer, connection layer and convolution layer. This model is used for text intelligent judgment. The judgment process is devided into five steps: text preprocessing, emotion word extraction, expression symbol extraction, emotion tendency value calculation and emotion final tendency value analysis. In order to test the effectiveness and superiority of the proposed method, an experimental comparison with the traditional judgment method was performed. The results show that the judgemental accuracy and recall rate of the text polarity intelligent judgment method based on the improved deep learning algorithm is higher, and the development space is broader.

Keywords: text polarity; intelligent judgment method; algorithm model design; effectiveness detection; deep learning algorithm; text pre-processing

0 引 言

随着互联网技术的进步，网络成为人们工作生活必不可少的组成部分。据2018年市场调查显示，我国互联网的发展速度已经处于世界前列，互联网在全国的普及率高达61.3%，网民规模达到了8.25亿[1]。近年来，随着移动互联网的不断普及，网络服务范围得以最大化推广，大众生活方式也得以改变[2]。

人机智能是一种新型技术，在智能识别和智能判断中发挥着重要作用，将人机智能融入到文本极性智能判断中，可以大大提高判别算法的工作效率。在机器学习研究中，深度学习算法有着很大的发展空间，这种起源于人工神经网络的学习算法可以模拟人的大脑对事物进行分析、解释文本、辨别声音[3]。深度学习算法不需要监督，它可以在低层特征中不断组合，再根据高层特征和属性特征找到数据的分布特征，从而完成文本分层、预测、判断等工作[4]。

本文基于改进深度学习算法研究了一种文本极性智能判断方法，在卷积神经网络(CNN)的基础上进行优

收稿日期：2019-06-18 **修回日期**：2019-07-11
基金项目：国家自然科学基金面上项目：网络零售交易风险动态评估及预警研究(71571021)

化，重新训练学习数据，采用隐式特征抽取的方式从训练数据中学习。该判别方法可以达到细粒度标记水准，将被判别文本清晰明确地分成非常消极、消极、中性、积极、非常积极五个层次[5]。

本文设计的改进深度学习算法采用了局部权值共享的特殊结构，能够更好地处理语音文本和图像文本，在布局上与生物神经网络十分相似。多维向量输入使判断过程不需要重建数据，降低工作复杂度[6]。为了更好地检测所设计的文本极性智能判断方法的有效性，本文以微博热门话题作为样本数据进行实验，通过准确率、召回率的比较实验，对比改进模型与普通的CNN、RNN模型。

1　改进深度学习算法模型建立

结合已有的CNN、LSTM、多层CNN、Bi-LSTM-CRF等结构，建立了一种新型深度学习算法网络结构。该神经网络结构共包括输入层、输出层、采集层、连接层、卷积层五部分，改进神经网络结构图如图1所示。

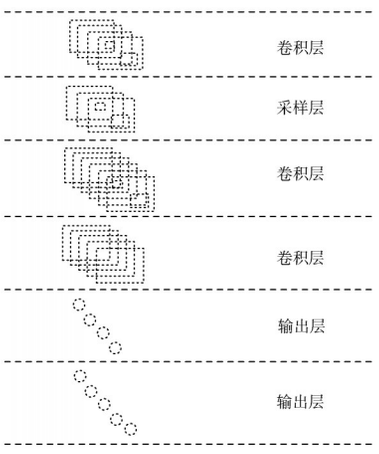

图1　改进神经网络结构图

改进神经网络中，每层之间的变换都涉及一次特征提取，提取后的层由多个二维平面组成，这些二维平面统称为特征映射图。在输入层中输入原始文本，多次提取原始文本数据。本文采用的计算方式为二次计算，即使输入的原始数据有很大的形变，二次计算也能够较好地计算出结果[7]。

改进神经网络结构中卷积层和子采样层都是独立工作的，卷积层工作过程如图2所示。

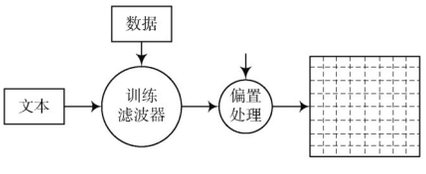

图2　卷积过程图

观察图2可知，卷积层会利用训练滤波器对输入的数据和文本进行卷积、偏置处理，从而得到卷积层[8]。卷积层将最初的输入文本编程为不同的网格，每个网格都记录着不同的特征数据，便于进行后续工作。

子采样过程如图3所示。

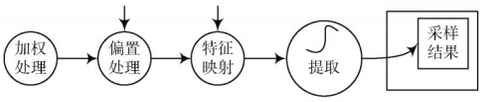

图3　子采样过程图

将邻域的4个像素汇集到一起求和，集合成一个像素后，进行加权处理和偏置处理，通过激活函数缩小特征映射图，缩小后的特征映射图可以被直接提取，耗费成本低[9]。

卷积运算和采样运算都能够强化文本特征，降低噪音。

连接层是以隐含状态存在的，能够连接上一层和下一层，在连接层中设置了权重向量和偏置向量，输入数据经过加权处理和偏置处理后得到一个新的数值，该数值最终会被传给sigmoid函数。

输出层具有分类功能，通过回归曲线计算输入文本属于各种类别的概率。

将本文建立的改进深度学习算法模型应用到文本极性智能判断中，选取文本中的小部分区域在神经网络最低层次中输入，依次滤波处理和加权处理，直至确定文本信息最显著的特征。为确保识别的一致性，每个映射上使用的权值都是相等的，随着逐层输出，网络参数会变得越来越少，最后会出现唯一的不变性特征[10]。文本也可以直接以网格方式输出，不需要重建数据，工作方式较为简单。

2　基于改进深度学习算法的文本极性智能判断方法

利用前文建立的深度学习算法改进模型对文本进行极性智能判断，分析文本中的情感词和语义规则，判断流程图如图4所示。

分析图4可知，本文研究的文本极性智能判断方法共分为五步：

1）对提取出来的文本数据进行预处理，通过Java工具提炼所有的分词。

2）构建情感词典，将情感词典与文本中的数据进行匹配，如果情感词典中不包含文本数据中的关键词，则要重新设定阈值，计算情感极性。

3）通过表情词典提炼文本中的表情符号，如果文本中不包含表情符号，则直接进入下一步。

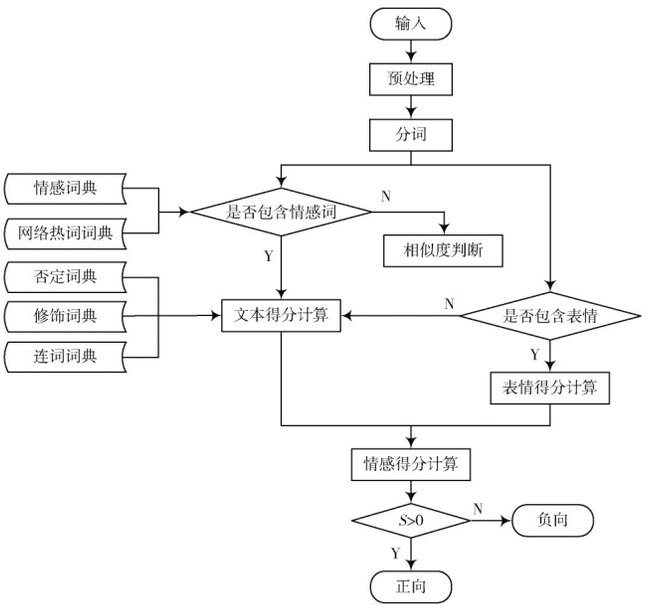

图4 判断方法流程图

4）同时使用否定词典、修饰词典和连接词典计算出文本的感情倾向值。

5）利用加权算法对上述步骤进行求值，得到最终的情感倾向值S，如果$S>0$，则判断该文本方向为正向；如果$S<0$，则判断该文本方向为负向。

2.1 文本数据提取与预处理

2.1.1 文本数据提取

文本数据提取采用网络爬虫提取方式，所有的目标网站和关键字需要自定义[11]。文本数据信息量大，一些文本数据还需要登录，普通爬虫难以直接提取数据，本文利用Python设计了一种新的爬虫，能够模拟登录用户ID，本文设计的爬虫为scrapy爬虫，获取文本信息的流程图如图5所示。

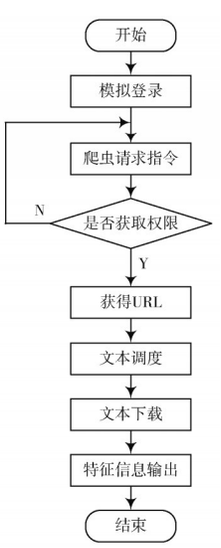

图5 文本数据提取流程

本文加入了1 000个关键词组成关键词数据库，使爬虫能够更快地获取信息。

2.1.2 文本预处理

通常爬虫得到的文本都会含有噪声信息，如果直接对其进行判断，准确度会大大降低，因此需要对文本数据进行预处理[12]。预处理主要从三个方面进行：繁体字处理；无效链接处理；交互信息处理。

虽然绝大多数的文本信息都是简体字，但是也有部分文本信息为繁体字，影响后续的分词判断、情感词判断、权重处理等操作，所以有必要将文本中的繁体字转化成简体字。很多文本中可能会存在无效链接，对于智能判别毫无帮助，在整体处理之前，要将没有用的链接剔除。通常只有少量文本含有交互信息，这些交互信息对于实际判别没有任何帮助，需要去除。

2.2 文本中情感词提取

在文本中，情感词是十分重要的组成部分，提取情感词对于文本判断有着重要意义。每一段文本中的信息都要与情感词典进行匹配，如果能够在情感词典中匹配到相应的信息，则只需要记录下极性和强度值即可；如果不能匹配到对应的词语，则需要利用语义相似度计算方法计算出每个词汇的情感倾向，设定固定阈值[13]。

情感词典中的词被划分到五个类别中，分别为非常消极、消极、中性、积极、非常积极，结构如图6所示。

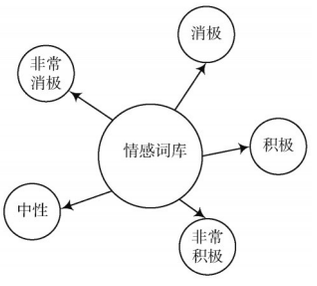

图6 情感词典

图6中的情感词典是经过多次提炼和反复匹配的，包括了大量能够表达情感的词语，但是也有部分情感词难以在情感词典中匹配到，所以需要利用语义相似度方法计算文本中词汇的情感倾向值。设定文本中的词语为x，被对比的词语为y，假设词语x可以解释成m个义项，则每个义项就可以用$x_1,x_2,\cdots,x_m$来表示，假设词语y有n个义项，则每个义项就可以用$y_1,y_2,\cdots,y_n$来表示，词语x和词语y每个义项的最大相似度计算公式如下：

$$\mathrm{Sim}(x,y)=\max[\mathrm{Sim}(x_i,y_i)] \tag{1}$$

利用可变参数λ计算出义项原相似度：

$$\mathrm{Sim}(x_1,y_1)=\frac{\lambda}{\lambda+d(x_i,y_i)} \tag{2}$$

将每个义项原值进行相似度计算，通过计算平均值差，得到最终的情感值计算结果。

2.3　语义规则与表情符号判断

每一个文本句子都会有自己的语义规则，不同的语义规则将句子划分为不同的种类，情感倾向通常通过修饰副词表现出来，修饰强度不同，情感倾向也不同。如果句子中加入了否定词语，那么情感的极性也会完全发生改变，例如未加否定词语之前，该句子表达的为“绝对肯定”，加入了否定词后，该句子想要表达的意思就变成了“绝对否定”。例如“我非常喜欢明星A”表达的是自己对A明星的绝对喜爱之情，在加入否定词后，就会变成“我非常不喜欢明星A”，表达的是对某个明星的绝对厌恶之情，这是两种完全不同的感情。

修饰程度副词可以分为6级，代表性词语如表1所示。

表1　程度副词分类表

程度	代表词
1级	有些、丝毫、相对……
2级	略微、稍微、或多或少……
3级	更加、较为、越来越……
4级	很、十分……
5级	过分、何止……
6级	极其、完全、绝对……

除了情感词外，本文设定的判断方法也会对表情符号进行判断，因为判断过程比较简单，所以本文不做研究。

3　验证实验

3.1　实验数据

为了检测本文研究的基于改进深度学习算法的文本极性智能判断方法的实际工作效果，与传统判断方法进行对比，从具有明确情感信息的30 000条微博数据中随机选取正向情感的微博和负向情感的微博各10 000条进行实验。微博中文本信息示例如表2所示。

表2　微博文本信息示例

情感判别	微博文本示例
正向	感谢大家对本次列车的支持，我们将竭诚为每一位旅客服务，如果各位旅客有宝贵的意见或建议，可以发给我们的相关工作人员，或者直接@我们哦~
负向	从来没有见过这么垃圾的列车，不仅晚点，服务态度还不好，晚点了多久也不通知，这种列车可以去死了，真是恶心!!!

3.2　实验评判标准

本文将准确率和召回率作为评价指标，将判断正确的正向情感微博文本记为TP，判断错误的正向情感微博文本记为TN，判断正确的负向情感微博文本记为FP，判断错误的负向情感微博文本记为FN。

正向类别的微博文本准确率计算公式为：

$$P_{pos}=\frac{TP}{TP+FP} \tag{3}$$

正向类别的微博文本召回率计算公式为：

$$R_{pos}=\frac{TP}{TP+FN} \tag{4}$$

负向类别的微博文本准确率计算公式为：

$$P_{neg}=\frac{TN}{TN+FN} \tag{5}$$

负向类别的微博文本召回率计算公式为：

$$R_{neg}=\frac{TN}{TN+FP} \tag{6}$$

3.3　实验结果与分析

根据上述参数和评价标准进行实验，设定α为判断后的准确率。不同α值下的文本分类准确率如图7所示。

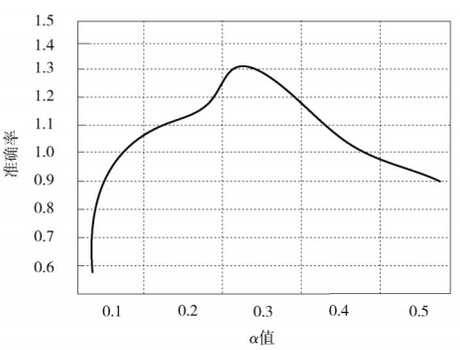

图7　不同α值下的文本分类准确率

观察图7可知，当α值达到0.3时，准确率最高。选用传统判断方法和本文判断方法对同一文本进行判断，对比两种方法的准确率和召回率，实验结果对比如表3所示。

表3　实验结果对比表

方法	正向		负向	
	准确率	召回率	准确率	召回率
传统方法	0.215	0.253	0.258	0.204
本文方法	0.298	0.301	0.312	0.299

综上所述，本文研究的判断方法相较于传统方法在准确率和召回率方面均有很大程度的提高，对于关键词的提取也十分准确，即使在文本表达复杂的情况下，也能够快速准确地做出智能性判断。

4　结　语

本文基于改进深度学习算法提出一种新的文本极性智能判断方法，该方法将传统的情感词典匹配法和语义相似度计算法结合到一起，同时构建了新的情感词典。本文设计的判断方法不需要多次对数据进行标记，具有实时判断能力。

（下转第85页）

6 结　论

本文在城市复杂道路场景下，提出一种基于Haar-like和时空信息的交通状态区域提取算法，该算法可以分车道地提取各交通状态区域，得到细化精准的各交通状态情况。将该算法与基于帧差法的交通状态检测和基于车辆检测的交通状态检测进行对比实验，结果表明该算法在复杂交通情形下仍然有着较高的准确率。

注：本文通讯作者为巨永锋。

参　考　文　献

[1] 美国交通研究委员会.道路通行能力手册[M].北京：人民交通出版社，2007：116-189.

[2] 北京交通委员会.国外交通拥堵定义指标简介[EB/OL]. [2012-12-28]. http://www.bjjtw.gov.cn/gzdt/yjzl/201212/t20121228_70291.htm.

[3] WONG G C K, WONG S. A multi-class traffic flow model—an extension of LWR model with heterogeneous drivers [J]. Transportation research part A: policy & practice, 2002, 36(9): 827-841.

[4] 计文平，郭宝龙，丁贵广.基于虚拟线圈的光流法车流量检测[J].计算机仿真，2004(1)：109-110.

[5] TAN H, ZHAI Y, LIU Y, et al. Fast anomaly detection in traffic surveillance video based on robust sparse optical flow [C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016). [S.l.]: IEEE, 2016: 1976-1980.

[6] 郝毫刚，陈家琪.基于五帧差分和背景差分的运动目标检测算法[J].计算机工程，2012，38(4)：146-148.

[7] SENGAR S S, Mukhopadhyay S. A novel method for moving object detection based on block based frame differencing [C]// 3rd International Conference on Recent Advances in Information Technology. [S.l.]: IEEE, 2016: 467-472.

[8] 郭怡文，袁飞虎.基于背景差分的多车道车流量检测系统[J].电光与控制，2010，17(9)：90-93.

[9] JOUDAKI S, SUNAR M S B, Kolivand H, et al. Background subtraction methods in video streams: a review [C]// 2015 4th International Conference on Interactive Digital Media. [S.l.]: IEEE, 2016: 53-60.

[10] LAM C T, GAO H, NG B. A real-time traffic congestion detection system using on-line images [C]// 2017 IEEE 17th International Conference on Communication Technology. [S.l.]: IEEE, 2017: 1548-1552.

[11] VIOLA P, JONES M J. Robust real-time face detection [J]. International journal of computer vision, 2004, 57(2): 137-154.

作者简介：薛飞杨(1994—)，男，陕西韩城人，硕士，研究方向为交通检测技术。

巨永锋(1962—)，男，陕西周至人，教授，博士生导师，主要研究方向为交通控制与管理、智能测控技术及应用。

宋永超(1990—)，男，山东威海人，博士，研究方向为交通多对象检测。

(上接第79页)

虽然具备上述优点，但本文提出的判断方法仍然有一部分需要深入研究，如网络新词的判断，以及如何更好地搜寻到文本中表达关键信息的词汇，希望在后续的研究中能够得以解决。

参　考　文　献

[1] 马胜蓝.基于深度学习的文本检测算法在银行运维中应用[J].计算机系统应用，2017，26(2)：184-188.

[2] 朱国进，沈盼宇.基于深度学习的算法知识实体识别与发现[J].智能计算机与应用，2017，7(1)：17-21.

[3] 刘江玉，李天剑.基于深度学习的仓储托盘检测算法研究[J].北京信息科技大学学报(自然科学版)，2017，32(2)：78-84.

[4] 左艳丽，马志强，左宪禹.基于改进卷积神经网络的人体检测研究[J].现代电子技术，2017，40(4)：12-15.

[5] 吕淑宝，王明月，翟祥，等.一种深度学习的信息文本分类算法[J].哈尔滨理工大学学报，2017，22(2)：105-111.

[6] 喻一梵，乔晓艳.基于深度学习算法的正负性情绪识别研究[J].测试技术学报，2017，31(5)：398-403.

[7] 廖健，王素格，李德玉，等.基于增强字向量的微博观点句情感极性分类方法[J].郑州大学学报(理学版)，2017，49(1)：39-44.

[8] 徐嵩，李玉峰.最大效益准则下基于分配公平性的CSGC改进算法[J].电子设计工程，2017，25(5)：97-102.

[9] 陈江昀.一种基于深度学习的新型小目标检测方法[J].计算机应用与软件，2017，34(10)：227-231.

[10] 李翌昕，马尽文.文本检测算法的发展与挑战[J].信号处理，2017，33(4)：558-571.

[11] 邹煜，刘兴旺.基于深度学习手写字符的特征抽取方法研究[J].软件，2017，38(1)：23-28.

[12] 蒋兆军，成孝刚，彭雅琴，等.基于深度学习的无人机识别算法研究[J].电子技术应用，2017，43(7)：84-87.

[13] 冯通.基于深度学习的航空飞行器故障自助检测研究[J].计算机仿真，2015，32(11)：119-122.

作者简介：宋思晗(1992—)，男，山东曲阜人，硕士，主要研究方向为自然语言处理。

王兴芬(1968—)，女，山东平度人，博士，教授，主要研究方向为Web安全、电子商务、大数据分析与管理创新。

杜惠英(1982—)，女，福建泉州人，博士，副教授，主要研究方向为移动互联网、电子商务、大数据消费者行为。