计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(11):205−210 [doi: 10.15888/j.cnki.csa.006645]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深度神经网络的微博文本情感倾向性分析①

钮成明, 詹国华, 李志华

(杭州师范大学 信息科学与工程学院，杭州 311121)
通讯作者: 詹国华, E-mail: ghzhan@hznu.edu.cn

摘　要: 随着新型社交媒体的发展, 作为传播网络舆论的重要媒介, 微博已然成为挖掘民意的平台. 自然语言处理技术可以从微博文本中提取有效情感信息, 为网络舆情监控、预测潜在问题及产品分析等提供科学的决策依据. 为了克服现有的浅层学习算法对复杂函数表示能力有限的问题, 本文尝试融合深度学习的思想, 提出基于 Word2Vec 和针对长短时记忆网络改进的循环神经网络的方法进行中文微博情感分析. 在两万多条中文标注语料上进行训练实验, 实验数据与 SVM、RNN、CNN 作对比, 对比结果证明, 本文提出的情感分析模型准确率达到了 91.96%, 可以有效提高微博文本情感分类的正确率.

关键词: 中文微博; 情感分析; 深度学习; 长短时记忆网络; 词向量

引用格式: 钮成明,詹国华,李志华.基于深度神经网络的微博文本情感倾向性分析.计算机系统应用,2018,27(11):205–210. http://www.c-s-a.org.cn/1003-3254/6645.html

Chinese Weibo Sentiment Analysis Based on Deep Neural Network

NIU Cheng-Ming, ZHAN Guo-Hua, LI Zhi-Hua

(School of Information Science and Engineering, Hangzhou Normal University, Hangzhou 311121, China)

Abstract: With the development of new social media, Weibo, as an important media for the dissemination of public opinion, has become a platform for the excavation of public opinion. Natural Language Processing technology can extract effective emotional information from Weibo texts, and provide scientific decision-making basis for monitoring network public opinion, forecasting potential problems, and product analysis. In order to overcome the limitation of the existing shallow learning algorithm for complex function expression, this study attempts to integrate the idea of deep learning, and puts forward an improved recurrent neural network based on Word2Vec and long-term memory network to analyze Chinese Weibo emotion. In the more than 20 000 Chinese corpus of training experiment, the experimental data with SVM, RNN, and CNN are compared, comparison results show that the emotion analysis model proposed in this study reaches the accuracy rate of 91.96%, thus it can effectively improve the accuracy of the Weibo text sentiment classification.

Key words: Weibo; sentiment analysis; deep learning; long short-term memory network; word vectors

2017 年 11 月, 新浪发布了第三季度财报, 财报指出, 截至 2017 年 9 月, 微博每月有 3.76 亿活跃用户, 日常约 1.65 亿活跃用户, 继续保持持续稳定的增长. 随着日益火热的微博, 海量的信息在微博迅速传播. 这些看似琐碎的信息却富含观点、倾向和态度, 蕴含了庞大的社会价值和商业价值, 以辨别微博文本中的情感倾向性, 具有重大的现实意义[1].

对文本情感分析国外已经取得了不少研究成果,

① 收稿时间: 2018-04-08; 修改时间: 2018-04-27; 采用时间: 2018-05-15; csa 在线出版时间: 2018-10-24

但是国内由于起步晚, 使得对微博短文本的研究成果较少. 目前常用的浅层文本情感分析方法: 一是基于情感词典的文本情分析, 二是基于传统的机器学习方法. 第一种分析方法的效果依赖于情感词典的好坏, 而构造一个好的情感词典需要耗费大量的人力和物力, 且在长难句、无情感词情感倾向明显的情况下, 这种情感分析方法就会失效. 基于监督学习的机器学习的方法分类效果比人工效果要好, 但是机器学习方法的关键之处在于特征组合的选取以及分类器的选取, 往往要耗费大量精力, 同时, 机器学习算法对复杂函数表示能力有限, 未深入考虑文本内部含义连接, 从而影响了分类的准确. 鉴于此问题, 本文研究提出了融合深度学习的方法建立模型, 研究中文微博情感分析.

1 相关工作

深度学习起源于对人工神经网络 (ANN) 的研究[2], 它能够通过模拟人类大脑的结构,提取外部输入的复杂数据的特征, 准确、高效地自动学习知识,因此能够高效地解决问题[3]. 传统的基于词典的方法与机器学习算法也能够解决文本的情感倾向性分析工作. 文献[4]对比了 Naive Bayes、ME 和 SVM 这三种机器学习算法对文本进行情感分类的效果. 随着深度学习的研究深入, 深度学习也被研究者们成功地应用与自然语言处理领域. 如: 文献[5]提出一种获取词向量的基于神经网络语言模型, 统筹考虑局部文本信息与全局文本信息的神经网络从而更加快速准确的学习到词语表征, 从而可获得文本主题, 结果表明比单独使用局部文本或单独使用全局文本的效果要好; 文献[6]首先基于循环神经网络 (RNN) 和卷积神经网络 (CNN) 训练了文本的向量, 最终使用普通的人工神经网络 ANN 进行序列短文本分类, 实验证明了添加顺序信息可以提高预测质量; 文献[7]用递归神经网络对 Twitter 数据进行了分析, 证明了深度学习应用在文本情感分析领域的可行性; 文献[8]设计了一个基于注意力的 LSTM 网络, 用于跨语种情感分类, 它能有效的将情感信息从资源丰富的语言适应于资源贫乏的语言, 并且有助于提高情感分类的性能. 文献[9]提出了一个统一的 CNN-RNN 模型,用于视觉情感识别. 该体系结构利用 CNN 的多个层次, 在多任务学习框架内提取不同级别的特征, 并提出了一种双向 RNN, 将在 CNN 模型中不同层次的学习特征整合在一起, 从而大大提高了分类性能. 一系列的研究表明, 深度学习在解决此类问题更具有优势.

2 算法描述

2.1 基于 Word2Vec 的情感词的向量转换

自然语言处理任务是为了让计算机使用人类语言, 而其中词向量的任务是要用数字化来表示语言, 从而转变成计算机能够理解的语言. Word2Vec 就是这样一款用于产生词向量的相关模型, 能够将自然语言转换成数值形式, 即词嵌入 (Word Embbeding)[10]. 这个模型既可以在海量的词典和大规模的数据集上进行高效的训练, 得到的结果又能够较好的度量词与词之间的相似度. Word2Vec 这一开源工具用到了 CBOW 模型和 Skip-gram 模型这两个模型 (见图 1)[11].

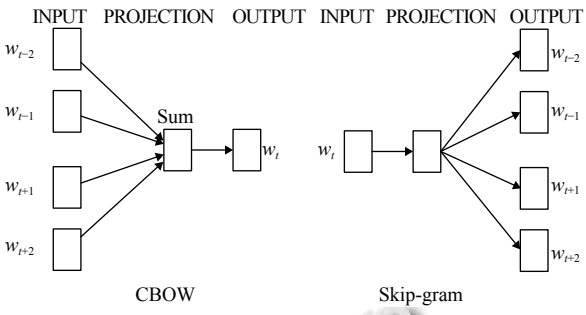

图 1 CBOW 模型和 Skip-gram 模型

由图 1 可见, CBOW 模型和 Skip-gram 模型都分别含有输入层, 投影层, 输出层这三层. CBOW 模型是依据当前要预测的词 w_t 的上下文的词 w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2} 来预测当前词 w_t; Skip-gram 模型是依据当前词 w_t, 预测 w_t 的上下文 w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}. 本文使用的是 Word2Vec 的 Skip-gram 模型, 根据 Hierarchical Softmax 进行设计的. 下面以样本 (w,$context(w)$) 为例进行分析.

已知当前词 w, 需要对其上下文 $Context(w)$ 中的词进行预测, 因此目标函数如下:

$$L = \sum_{w \in C} \log p(context(w)|w) \tag{1}$$

其中 $p(context(w)|w)$ 构造的条件概率, 定义如下:

$$p(context(w)|w) = \prod_{u \sigma Context(w)} p(u|w) \tag{2}$$

$$p(u|w)=\prod_{j=2}^{l^u} p\left(d_j^u|V(w),\theta_{j-1}^u\right) \tag{3}$$

其中,

$$p\left(d_j^u|V(w),\theta_{j-1}^u\right)=\left[\sigma\left(V(w)^{\mathrm{T}}\theta_{j-1}^u\right)\right]^{d_j^u} \times\left[1-\sigma\left(V(w)^{\mathrm{T}}\theta_{j-1}^u\right)\right]^{d_j^u} \tag{4}$$

将式 (4) 依次带回前面各式, 得到对数似然函数的具体表达式, 即 Skip-gram 模型的目标函数:

$$L=\sum_{w\in C}\log\prod_{u\sigma Context(w)}\prod_{j=2}^{l^u}\left\{\left[\sigma\left(V\left(w\right)^{\mathrm{T}}\theta_{j-1}^u\right)\right]^{d_j^u} \cdot\left[1-\sigma\left(V\left(w\right)^{\mathrm{T}}\theta_{j-1}^u\right)\right]^{d_j^u}\right\} \tag{5}$$

接着用随机梯度上升法对其优化, 每取一个样本 $(context(w)|w)$ 就要对目标函数中的相关参数进行刷新.

θ_{j-1}^u 的更新公式可写为:

$$\theta_{j-1}^u=\theta_{j-1}^u+\eta[1-d_j^u-\sigma(V\left(w\right)^{\mathrm{T}}\theta_{j-1}^u\rceil\times V(w) \tag{6}$$

$V(w)$ 的更新公式为:

$$V(w)=V(w)+\eta\sum_{u\sigma Context(w)}\sum_{j=2}^{l^u}\frac{\partial L(w,u,j)}{\partial V(w)} \tag{7}$$

通过对 Word2Vec 的训练, 可以把中文文本内容转化成空间向量运算, 词向量的维度与隐层节点数致, 从而可以通过向量的相似度判断词语相似度.

2.2 基于 LSTM 改进的循环神经网络模型

为了实现中文微博情感分析, 本文提出了基于 Word2Vec 和针对 LSTM 改进的神经网络模型的网络结构. 该模型是循环神经网络 RNN 的变体, RNN 结构如图 2 所示.

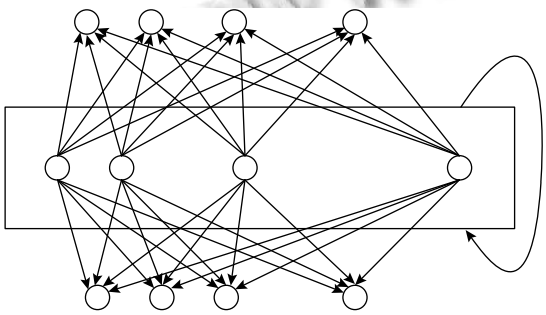

图 2 循环神经网络的简化结构

RNN 带有环结构, 能够记住序列前的信息, 因此文本序列的处理更适合具有记忆功能的 RNN 来处理, 典型的 RNN 如图 3 所示.

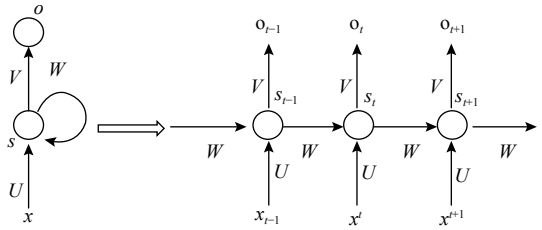

图 3 典型的 RNN

图 3 中, x 为输入层, o 为输出层, s 为隐藏层, V、W、U 分别为输入层、隐层与输出层的权重, t 为第 t 次的计算次数, 其中计算第 t 次的隐含层状态时为:

$$S_t=f(U*X_t+W*S_{t-1}) \tag{8}$$

从而实现了当前隐层计算结果与当前的输入与上一次的隐层计算结果相关, 达到了记忆功能的目的. 但是由于 RNN 自身的局限性, 使它不能学习到距离相对较远的相关文本信息, 且会导致梯度爆炸或者梯度消失的问题, 因而会影响分类效果[12]. 鉴于此, 本文的提出 LSTM 改进的神经网络模型模型, 将 RNN 的每个隐藏层全部替换成具有记忆功能的 cell, 它在处理时间序列和语言文本序列十分有优势; 且相比于 CNN 的输入十分固定, 它对非定长数据的输入表示更加灵活, 能很好的学习到距离相对较远的信息, 从而很好的解决了上述的问题[13].

原始的 LSTM 神经网络模型是由一个个神经元组成的, 每个神经元都相当于一个记忆细胞 (cell), 细胞彼此循环连接. 每个细胞都有输入门、遗忘门和输出门三个门. 具体结构如图 4 所示[14]. 改进的 LSTM 循环网络在各个门进行计算时, 接受细胞状态输入. 一个 Cell 由输入门、遗忘门和输出门以及一个 cell 单元组成, 门使用一个 Sigmoid 激活函数, Sigmoid 激活函数将权重设置为 0 到 1 之间的值,而输入门和细胞状态通常会使用 tanh 来转换.

首先计算输入门的激活值 i_t 和细胞状态在 t 时刻的输入变换$\widetilde{C}_t$:

$$i_t=\sigma W_i x_t+U_i h_{t-1}+b_i \tag{9}$$

$$\widetilde{C}_t=\tanh(W_c x_t+U_c h_{t-1}+b_c) \tag{10}$$

其中, x_t 为当前的输入向量, h_t 为当前的隐藏层向量. b_i是偏置, U_i 是输入权重, W_i是遗忘门的循环权重, 式中$\delta(*)$是 sigmoid 激活函数函数: $\delta(x)=\dfrac{1}{1+e^{-x}}$.

其次, 计算遗忘门记忆细胞在时刻 t 的激活值f_t:

$$f_t = \sigma W_f x_t + U_f h_{t-1} + b_f \tag{11}$$

其中, b_f 是偏置, U_f 是输入权重, W_f 是遗忘门的循环权重. 计算得到了输入门的激活值 i_t, 以及遗忘门的激活值 f_t, 和候选值 $\widetilde{C}_t$, 就可以计算得到记忆细胞在时刻 t 的状态更新 C_t;

$$C_t = i_t * \widetilde{C}_t + f_t * C_{t-1} \tag{12}$$

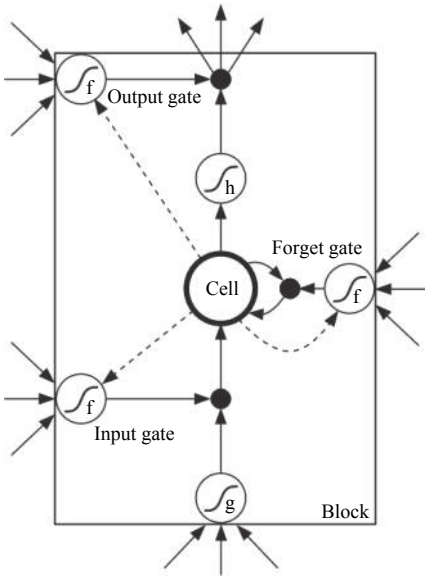

图 4 LSTM 循环网络“细胞”框图

随着细胞的状态更新, 就可以计算出输出门的激活值 O_t, 以及它的输出 h_t:

$$O_t = \sigma W_o x_t + U_o h_{t-1} + b_o \tag{13}$$

$$h_t = O_t * \tanh(C_t) \tag{14}$$

其中, b_o 是偏置, U_o 是输入权重, W_o 是遗忘门的循环权重. X_t 是当前的输入向量, h_t 是当前的隐藏层向量, h_t 包含了所有 LSTM 细胞的输出.

本文模型的整体流程如图 5 所示: 首先对输入的预处理之后的句子进行特征提取, 即向量化, 将句子转换为词向量表示之后再输入改进的 LSTM 转换成句向量, 最后经过深度神经网络分类器, 输出最终结果. 本模型设计了三层网络, 含有一个隐层, 由于增加隐层节点数比增加隐层数的训练效果更好且更容易实现. 为了避免节点过少引发的网络性能差与节点过多引发的训练时间过长以及易陷入局部极小值而达不到最优, 综合考虑下用节点删除法与扩张法确定隐层节点数为 128. 模型在选择和决策时参考了上一时刻隐层的状态, 保证了记忆功能. 模型权值调整时, 前后时间带来的影响能够同时作用, 保证了训练出来的模型具有较长时间的记忆功能.

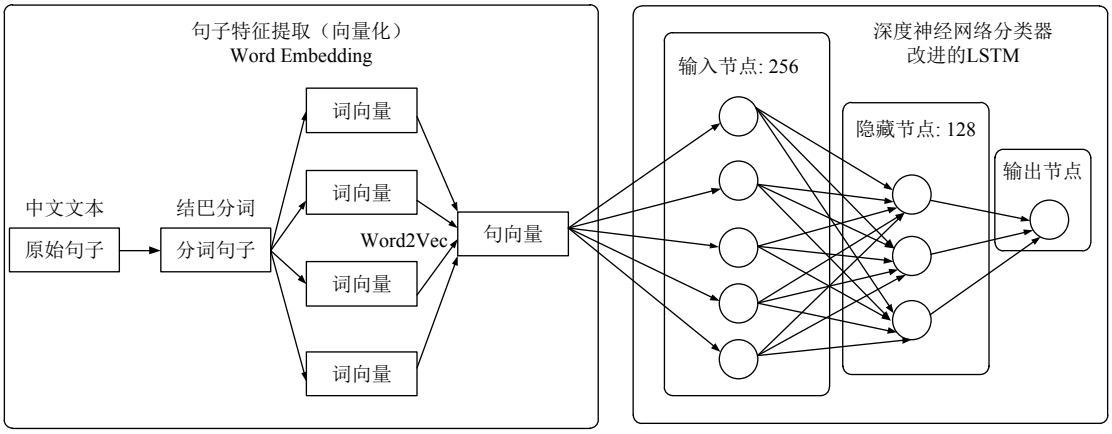

图 5 微博情感分析深度学习模型

3 实验

3.1 数据收集与预处理

由于本文实验使用了 Word2Vec 进行特征提取, 而 Word2Vec 训练时, 样本量越多, 词量越大, 结果可靠性越高; 且本文的模型是基于监督学习的, 因此需要已经分好类的句子越多越好. 为了实验的可靠性, 从网上搜集了两万多条中文标注语料, 包含 6 个不同话题. 进过去重与清理空数据, 最终选用了 21 107 条语料, 其中包含积极情感的为 10 428 条, 消极情感的为 10 679 条, 以保证在样本量有限的情况下, 词量尽可能多. 将语料大致按照 3:1 的比例划分为训练集与预测集, 其中随机抽取 15 000 条作为训练语料, 剩下的 6105 条作

为测试语料.

3.2 模型参数设定

本文所提到的模型超参数包括优化函数 (optimizer)、学习速率 (rate)、迭代次数 (epochs) 等. 为了找到最优超参数, 本文选择固定其他参数, 改变可变参数的大小进行试验.

(1) 优化函数

由于本文所使用数据属于文本挖掘领域, 用于训练的数据是稀疏数据且需要训练的是一个复杂的深度网络, 因此在选取优化函数上, 选择如下几种具有自适应性的优化函数.

从图 6 可知. 当优化函数为 RMSprop 时, 模型准确率约为 91.53%, 准确率达到最高. 从原理上来看, 四种超参数都是类似的, 因此本文模型单纯的选取实验结果最好的 RMSprop 作为优化函数.

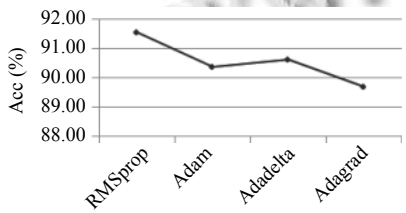

图 6 优化函数对比试验结果

(2) 学习速率

在优化模型时, 学习速率的选取相当重要, 过大则容易导致震荡. 过小则会使得收敛过慢. 根据学习速率选取策略不断尝试,

从图 7 可知, 随着学习速率的改变, 模型的准确率保持在 91% 左右, 在 0.5 时达到最高点, 随后稍微降低. 出现这种结果, 分析原因为当学习效率为 0.1 及以下时, 长时间算法无法收敛, 优化效率降低; 当学习效率达到 0.5 以上时, 每次迭代不会减少代价函数的结果, 甚至会越过最优值.

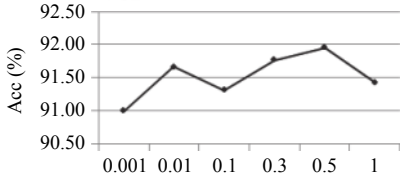

图 7 学习速率试验结果

(3) 迭代次数

迭代次数是对训练集整个训练一遍的次数, 随着迭代的次数增加, 模型逐渐逼近最优, 但当迭代次数超过一定的范围则易产生过拟合, 导致模型淡化能力下降.

由图 8 可知, 随着迭代次数的增加, 模型准确率逐渐增加, 当迭代次数为 20–30 之间时, 准确率趋于稳定, 且效果不错.

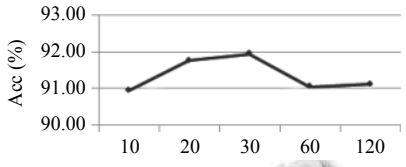

图 8 迭代次数试验结果

3.3 实验结果与分析

本文使用的深度学习框架是 Keras, 超参数值的选择是本模型的关键, 在控制变量的情况下, 重要参数的选择如表 1 所示, 能使最终结果达到最佳.

表 1 重要参数选择

参数名称	参数值
损失函数 (Loss)	msle
优化器 (Optimizers)	RMSprop
学习速率 (Rate)	0.5
迭代次数 (Epochs)	30

为了避免实验过程产生过拟合, 用不同的测试对象和训练对象做交叉比对进行训练. 加载预训练模型在训练过程中的准确率和损失率曲线如图 9、图 10 所示, 这样得到的训练集的准确率为 99.84%, 测试集准确率约为 91.96%.

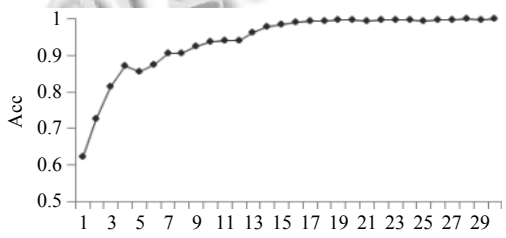

图 9 模型准确率

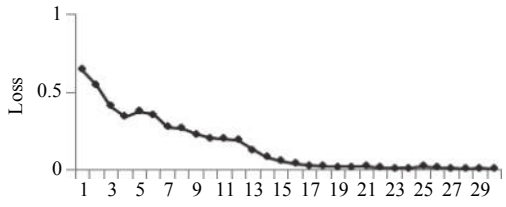

图 10 模型损失率

由图 9, 图 10 分析得出, 本文的模型损失率稳步下降, 训练集的准确率接近 100%, 在训练数据迭代到第

20 次时, 准确率与损失率趋于稳定. 然而, 达到零亏损的模型并不是十分好的事情, 应该特别注意防止模型过拟合.

3.3 模型对比实验

为了证明本模型的有效性, 本文做了一组对比实验, 使用了相同数据集, 分别与 SVM(采用 RBF 核)、CNN, RNN 等模型对微博文本的情感分析效果作对比. 主要参数选用相同的数值, 具体见表 1, 实验结果如图 11 所示. 本文提出的模型得到了最好的准确率.

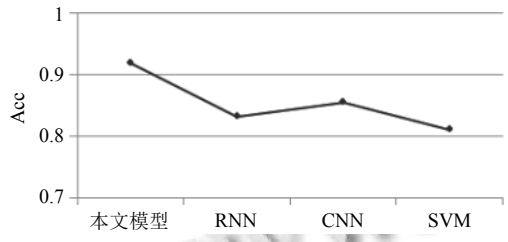

图 11 模型对比结果

由图 11 可知, 相同条件下本文提出的模型准确率高达 91.96%, 明显高于其他模型. 分析原因: 1) SVM 模型主要基于人工经验进行特征提取, 有很大的主观性; 2) CNN 更加擅长处理图像数据, 且注重全局的模糊感知; 3) RNN 则注重相邻近位置的信息, 但对于距离稍远的相关文本信息就不能够很好的学习到, 容易导致梯度爆炸或者梯度消失[15]. 为了克服上诉问题, 本文提出的方法既有效解决了特征提取的难点, 又避免了学习不到较远距离信息的弱点, 因此能够达到更好的效果.

4 结论与展望

本文融合深度学习的思想, 基于 Word2Vec 进行特征提取, 用 LSTM 改进的循环神经网络的方法实现了中文微博情感分析, 且得到了较以往方法更好的测试准确率, 证实本文提出的模型可以很好的解决短文本情感分析问题. 实验中尤其要注意过拟合的问题. 另外本文提出的模型训练时间消耗过长, 后续将在提高模型的训练速度做进一步探索.

参考文献

1 郭义超, 樊红. 基于中文文本分析的微博情感地图的制作. 计算机系统应用, 2017, 26(2): 25–29. [doi: 10.15888/j.cnki.csa.005594]
2 孙志远, 鲁成祥, 史忠植, 等. 深度学习研究与进展. 计算机科学, 2016, 43(2): 1–8.
3 Guo LL, Ding SF. Research progress on deep learning. Computer Science, 2015, 42(5): 28–33.
4 Pang B, Lee L, Vaithyanathan S. Thumbs up?: Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA. 2002. 79–86.
5 Huang EH, Socher R, Manning CD, *et al*. Improving word representations via global context and multiple word prototypes. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea. 2012. 873–882.
6 Lee JY, Dernoncourt F. Sequential short-text classification with recurrent and convolutional neural networks. arXiv: 1603.03827, 2016.
7 Goller C, Kuchler A. Learning task-dependent distributed representations by backpropagation through structure. Proceedings of International Conference on Neural Networks (ICNN-96). Washington, DC, USA. 1996.
8 Zhou XJ, Wan XJ, Xiao JG. Attention-based LSTM network for cross-lingual sentiment classification. Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Austin, TX, USA. 2016.
9 Zhu XG, Li L, Zhang WG, *et al*. Dependency exploitation: A unified CNN-RNN approach for visual emotion recognition. Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017). Melbourne, Australia. 2017.
10 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. Lake Tahoe, NV, USA. 2013. 3111–3119.
11 Mikolov T, Le QV, Sutskever I. Exploiting similarities among languages for machine translation. arXiv: 1309.4168, 2013.
12 Bayer J, Osendorfer C, van der Smagt P. Learning sequence neighbourhood metrics. Villa AEP, Duch W, Érdi P, *et al*. Artificial Neural Networks and Machine Learning-ICANN. Berlin, Heidelberg: Springer, 2012. 531–538.
13 Jozefowicz R, Zaremba W, Sutskever I. An empirical exploration of recurrent network architectures. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France. 2015.
14 Graves A. Supervised sequence labelling with recurrent neural networks. Berlin Heidelberg: Springer, 2012.
15 Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization. arXiv: 1409.2329, 2015.