基于深度学习算法的藏文微博情感计算研究

孙本旺[1]，田　芳[2]
（1. 青海大学 计算机技术与应用系，青海 西宁 810016；
2. 青海大学 信息化技术中心，青海 西宁 810016）

摘　要：针对藏文文本情感计算研究，将 CNN-LSTM 深度学习模型引入到藏文微博情感计算，弥补了少数语言自然语言处理研究的缺乏，对藏文研究具有一定的推动作用。针对藏文语料的不公开，通过藏文同反义情感词典对标注好的藏文微博语料中情感词汇的同反义词进行替换，进一步扩充了藏文微博语料，以适合深度学习对大数据语料的要求。藏文微博分词后，利用 Word2vec 工具训练出藏文微博词向量模型，提高特征向量对文本深层次语义信息的表达；然后，将训练好的词向量和对应的情感倾向标签直接引到由卷积层、池化层、LSTM 层、全连接层等构成的 CNN-LSTM 模型，在每一层的输出做归一化处理；最后经过 Softmax 分类器对藏文微博进行情感倾向分类，并与 LSTM 以及传统的情感词典做了实验对比。结果表明，该算法获得了较好的分类效果。

关键词：深度学习；藏文微博；词向量；情感计算
中图分类号：TP391.1　　**文献标识码**：A　　**文章编号**：1673-629X(2019)10-0055-04
doi：10.3969/j.issn.1673-629X.2019.10.012

Research on Tibetan Micro-blog Affective Computation Based on Deep Learning Algorithm

SUN Ben-wang[1]，TIAN Fang[2]
（1. Department of Computer Technology and Applications，Qinghai University，Xining 810016，China；
2. Information Technology Center，Qinghai University，Xining 810016，China）

Abstract：Aiming at the study of Tibetan text emotion calculation，the CNN-LSTM deep learning model is introduced into Tibetan micro-blog emotion calculation，which makes up for the lack of research on minority language natural language processing，and has certain impetus to Tibetan studies. For the non-disclosure of Tibetan corpus，the Tibetan and the anti-sense sentiment dictionary are used to replace the antonyms of the emotional vocabulary in the Tibetan micro-blog corpus，further expanding the Tibetan micro-blog corpus to meet the requirements of deep learning to big data. After the Tibetan micro-blog’ word segmentation，the Word2vec tool is used to train the Tibetan micro-blog’ word vector model to improve the expression of the deep vector semantic information of the feature vector. Then，the trained word vector and the corresponding emotional tendency label are directly introduced into the CNN-LSTM model consisting of convolutional layer，pooling layer，flatten layer，LSTM layer，and the output at each layer will be batch normalization. Finally，the Softmax Classifier is used to affect the Tibetan micro-blog. Compared with LSTM and traditional sentiment lexicon，it shows that the proposed algorithm achieves better classification effect.

Key words：deep learning；Tibetan micro-blog；word vector；emotional calculation

0　引　言

随着互联网技术的成熟和发展，藏族网民的数量越来越多，微博等成为藏民对社会热点关注和情感表达的平台。藏族网民在网络上发表意见、表达情感已成为一种日常习惯，由此产生了大量的藏文情感信息，其中的信息包含各种各样的情感特征。因此，如何通过复杂的信息抓取分析藏民的情感变化，便成为一项极为重要的研究课题。

近来年，深度学习模型已经广泛应用于文本分类。文中将 CNN-LSTM 深度学习算法模型引入藏文文本

收稿日期：2018-12-03　　**修回日期**：2019-04-09　　**网络出版时间**：2019-06-26
基金项目：国家自然科学基金（61461045）；青海省科技计划项目（2016-ZJ-743）
作者简介：孙本旺（1990-），男，硕士研究生，研究方向为自然语言处理；田　芳，博士，教授，通信作者，研究方向为自然语言处理、语义关系抽取、本体自动构建等。
网络出版地址：http://kns.cnki.net/kcms/detail/61.1450.TP.20190626.0833.046.html

情感分析领域,对于推动藏文文本情感分析研究具有十分重要的意义。

1 相关研究

2006 年 Geoffrey Hinton[1-2] 等首次提出了深度信念网络(deep belief network,DBN)深度学习算法的思想,并以其较强的学习能力和最大限度提取特征的特点,成为其后深度学习算法的主要框架。随着深度学习技术的发展,之后出现了堆栈自编码[3]、卷积神经网络(convolution neural networks,CNN)[4]、长短时记忆网络(long short-term memory,LSTM)[5]等深度学习模型。

麻省理工学院的 Picard 教授最早提出了情感分析的概念。Picard 教授在 1995 年发表了论文《Affective Computing》[6],并在两年后在此基础上撰写了的有关情感计算的最早同名论著[7]。Richard Socher 等提出深度递归自编码算法,在中文[8]和英文[9]的情感分析中,都取得了不错的结果。Socher R 等将 Matrix-Vector 融入循环神经网络(RNN)模型来学习逻辑命题和运算符含义,对电影评论的情感标签进行分类[10]。Tang D 等通过卷积神经网络和循环神经网络相结合的算法进行情感分析,自动推荐适当的表情符号,取得了优异的成效[11]。B. Sun 等提取了声学特征、lbptop、密集 SIFT 和 CNN-LSTM 特征,用 LSTM 和 GEM 模型来识别电影人物的情感[12]。J Huang 等提取其他声学音频特征集、外观特征和深层视觉特征作为补充特征。每种特征类型分别使用长时记忆递归神经网络(LSTM-RNN)进行训练,而且用于每个维的情感预测,要分别考虑注释延迟和时间池[13]。宋梦姣结合双向 LSTM 和卷积神经网络构建的 CNN-LSTM 模型在情感计算的性能上有所提升,在此模型的基础上又设计了使用注意力机制的 CNN-BLSTM-Attention 模型;注意力机制能帮助模型得到含有注意力概率分布的语义编码,有效突出文本中对情感分析任务更关键的词语,在文本情感分类任务上取得了更高的准确率[14]。焦晨晨提出基于横向卷积和纵向卷积相结合的卷积神经网络(HV_CNN),结合动态卷积神经网络(DCNN)的网络模型[15]。

在藏文情感分析方面,闫晓东等通过人工方法构建了一个全面、高效的极性词典,包括基础词词典、否定词词典、双重否定词词典、程度副词词典以及转折词词典,并提出了基于极性词典的藏语文本句子情感分析方法[16]。张俊等通过借鉴中文微博情感分析中比较常见的基于统计的方法和基于词典的方法对藏文微博进行情感分析,实验结果表明基于藏文词典的藏文微博情感分析的准确率明显高于基于 TF-IDF 的藏文微博情感分析的准确率[17]。杨志根据藏文微博的行文特征,提出了基于情感词典与机器学习算法多特征融合的藏文微博情感分类方法[18]。袁斌针对藏文微博中存在的藏汉混排问题,提出了一种基于语义空间的藏文微博情感表示方法。该方法通过句法树实现了语义向量化,提高了情感特征中的语义成分,并解决了多语言混合文本处理问题[19]。李苗苗提出了藏文文本情感分析的词语级、句子级、篇章级三层框架,提出了利用情感词典和规则集分析藏文句子情感的一种方法,采用 SVM 算法对篇章级进行情感分析[20]。普次仁等将藏文分词后,把深度领域内的递归自编码算法引入到藏文情感分析中,以更深层次提取语义情感信息,有监督地训练输出层分类器以预测藏文语句的情感倾向[21]。

2 藏文微博的情感倾向分析方法

情感分析首先要对藏文微博数据进行预处理:去除 xml 和@ 符号,去停用词等,将单词条内容处理成单行数据。然后对藏文微博进行分词,文中主要结合人工和情感词典进行藏文分词。最后利用规则和统计的方法进行情感计算。

2.1 基于情感词典的方法

由于藏文情感词典都不公开,也没有统一标准用于藏文情感分析的藏文情感词典,故使用文中自动构建的藏文情感词典。该藏文情感词典总词量达 27 361 个,包括程度副词 220 个、基础情感词(积极 10 670 个、消极 10 402 个、中性 5 711 个)、停用词 385 个,相比其他藏文情感词典多了双重否定词。

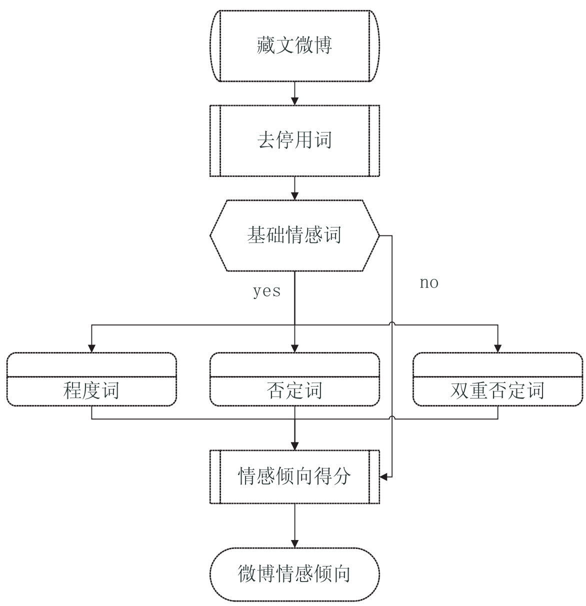

图 1 基于情感词典的情感计算流程

基于情感词典的藏文微博情感分析的方法主要用于实验结果对比。通过微博中情感词或情感短语的权

值叠加计算来判断某条微博的情感倾向。如果微博包含转折词,取转折词后面的部分微博进行情感计算,还要考虑微博中的程度词和否定词等。情感计算流程如图 1 所示。

2.2　基于 CNN-LSTM 模型的方法

微博文本向量化为文本处理提供了基础。结合 CNN 和 LSTM 的模型特点,提出了 CNN-LSTM 算法模型。该模型以 CNN 的第三层输出作为 LSTM 第一层的输入,在每一层的输出都做归一化处理。该模型既能保留 CNN 对文本的全局度量,又能保留 LSTM 对文本的上下深层语义信息,挖掘出更深层次的语义关系,取得了较好的分类效果。

2.2.1　Word2vec 词向量

神经网络的输入需要将藏文微博语料映射成为向量,Word2vec 使用的模型分为 CBOW 和 Skip-gram,文中使用 Skip-gram 模型实现词向量化,最终得到词向量字典。

Skip-gram:是用中心词来预测周围的词。在 Skip-gram 中,会利用周围的词的预测结果情况,使用 GradientDecent 不断调整中心词的词向量,最终所有的文本遍历完毕之后,也就得到了文本所有词的词向量。每个词在作为中心词时,都要进行 *K* 次的预测、调整,这种多次的调整会使得词向量相对更加准确。

2.2.2　CNN-LSTM 模型

CNN 可以保留文本的全局度量特征,但无法解决文本上下文的长期依赖问题和上下文语义关系问题。而 LSTM 具有学习长期上下文记忆依赖的能力,能有效利用和记忆很宽范围的上下文语义关系。结合两者的结构特点,文中构建 CNN-LSTM 模型用于藏文微博的情感计算。

CNN-LSTM 的网络层包括卷积层、Batch Normalization 层、池化层、时序层、输出层,如图 2 所示。

卷积层:经过词向量表达的藏文微博文本为一维数据,文中利用三层一维卷积,抽取藏文微博的局部特征,经过卷积核运算产生微博文本特征。

Batch Normalization 层:作用在每层卷积层之后。不仅极大提升了训练速度,收敛过程大大加快,还能增加分类效果,类似于 Dropout 的一种防止过拟合的正则化表达方式,所以不用 Dropout 也能达到相当的效果;另外调参过程也简单多了,对于初始化要求没那么高,而且可以使用大的学习率。

池化层:采用 max-pooling,池化层作用在每层卷积层和 Batch Normalization 层之后,是一种非线性降维的方法。用来缩减输入数据的规模进行特征映射层,此阶段保留 *K* 个最大的信息,保留了全局的序列信息。

时序层:将两层 LSTM 作为文中模型的时序层。其能够解决远距离上下文依赖特性关系、存储和挖掘出上下文深层语义信息。

输出层:采用 Softmax 分类器。

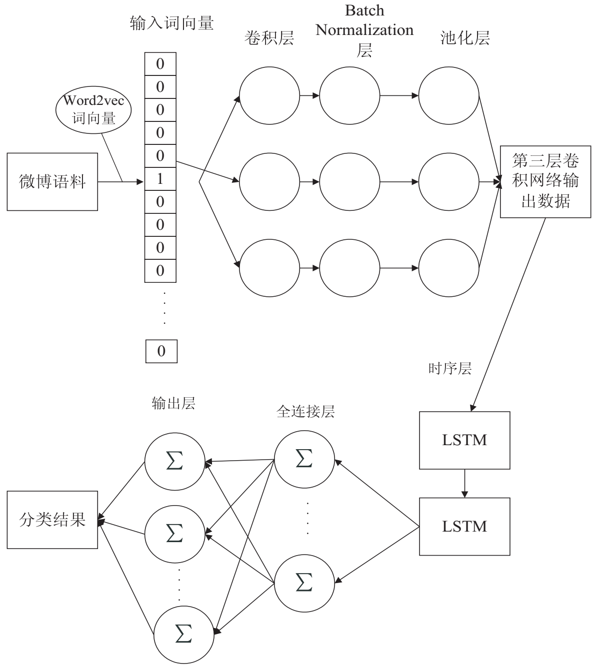

图 2　CNN-LSTM 网络模型结构

3　实验结果分析

利用标注好的藏文微博语料,经过微博中词语的同反义词替换来扩充语料,增加的语料基本满足了深度学习对数据量的需求。为了验证算法的准确性,对基于情感词典,LSTM 和 CNN-LSTM 的深度学习算法进行藏文微博情感倾向分析进行对比。深度学习模型 LSTM 和 CNN-LSTM 激活函数为 softsign,优化函数为 Adam(学习速率为 0.01)同样的语料库,结果如图 3 ~ 图 5 所示。

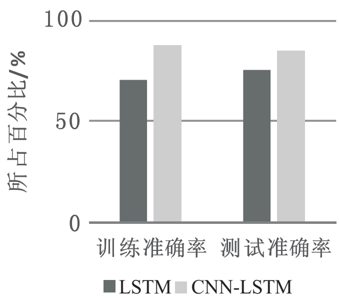

图 3　LSTM 和 CNN-LSTM 准确率对比

从图 3 可以看出,CNN-LSTM 比单独的 LSTM 模型的测试准确率高约 10.2%,训练准确率高约 18.3%。CNN-LSTM 模型能够保证每条微博的全局结构不变,又能挖掘出更深层次语义信息结构,所以其训练测试率都比较优异。

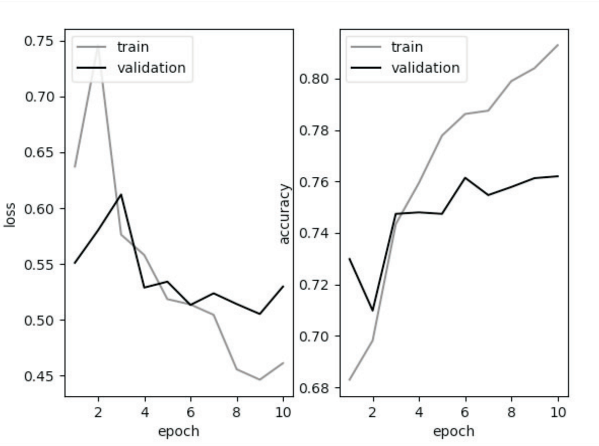

图 4　LSTM 的 loss 和 accuracy 趋势变化

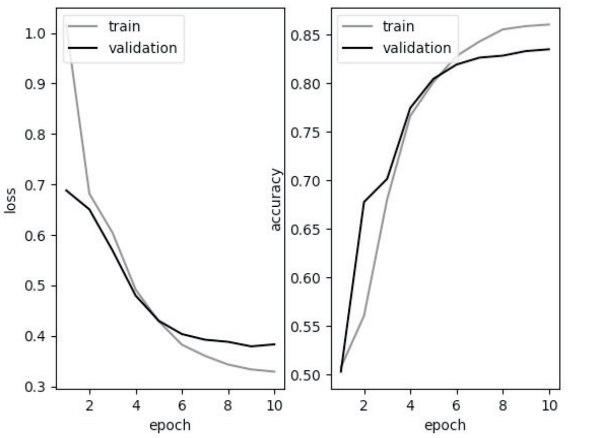

图 5　CNN-LSTM 的 loss 和 accuracy 趋势变化

从图 4 和图 5 得出,CNN-LSTM 模型的训练集损失率下降比较平稳,训练集的准确率又能稳定上升,此模型相对其他算法模型具有良好的稳定性。

接着将基于藏文情感词典、LSTM 和 CNN-LSTM 的准确率进行对比,如表 1 所示。

表 1　分类准确率对比

算法	准确率/%
情感词典	71.32
LSTM	75.31
CNN-LSTM	85.11

从表 1 可以看出,基于 CNN-LSTM 情感分类比 LSTM 模型高 10.2%。卷积神经网络注重于对全局的度量,RNN 侧重于每一相邻信息的重构,而 LSTM 要比传统 RNN 对文本深层语义信息的处理更加有效。模型能够保证每条微博的全局度量,又能挖掘出更多的深层次语义信息,做出更精准的情感分类。

4　结束语

文中将深度学习算法的 CNN-LSTM 模型引入到藏文的情感倾向分析。同时,研究了藏文微博中情感倾向分类的 LSTM、CNN-LSTM 等方法,对于每个微博情感特征,训练分类器,不同情感分类具有不同的判别能力。CNN-LSTM 利用卷积层和 LSTM 层融合网络来处理情感特征,保留文本的全局度量又能挖掘出更深层次的语义关系,取得了较好的分类效果。此外,该模型也存在一定的不足,如藏文语料分词困难等,这些还有待进一步研究。

参考文献:

[1]　HINTON G E,SALAKHUTDINOV R R. Reducing the dimensionality of data with neural network[J]. Science,2006,313(5786):504-507.

[2]　HINTON G E,OSINDERO S,TEH Y W. A fast learninal algorithm for deep belief nets[J]. Neural Computation,2006,18(7):1527-1554.

[3]　LAROCHELLE H,BENGIO Y,LOURADOUR J,et al. Exploring strategies for training deep neural networks[J]. Journal of Machine Learning Research,2009,10:1-40.

[4]　HAYKIN S,KOSKO B. Gradient-based learning applied to document recognition[C]//IEEE congress on evolutionary computation. Brazil:IEEE,2009:306-351.

[5]　HOCHREITER S,SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[6]　LISETTI C L. Affective computing[J]. Pattern Analysis & Applications,1998,1(1):71-73.

[7]　PICARD R W. Affective computing:challenges[J]. International Journal of Human-Computer Studies,2003,59(1-2):55-64.

[8]　梁　军,柴玉梅,原慧斌,等. 基于深度学习的微博情感分析[J]. 中文信息学报,2014,28(5):155-161.

[9]　SOCHER R. Recursive deep learning for natural language processing and computer vision[D]. California:Stanford University,2014.

[10]　SOCHER R,HUVAL B,MANNING C D,et al. Semantic compositionality through recursive matrix-vector spaces[C]//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Jeju Island:ACL,2012:1201-1211.

[11]　TANG Duyu,QIN Bing,LIU Ting. Document modeling with gated recurrent neural network for sentiment classification[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. [s. l.]:[s. n.],2015:1422-1432.

[12]　SUN Bo,WEI Qinglan,LI Liandong,et al. LSTM for dynamic emotion and group emotion recognition in the wild[C]//Proceedings of the 18th ACM international conference on multimodal interaction. Tokyo,Japan:ACM,2016:451-457.

[13]　HUANG Jian,LI Ya,TAO Jianhua,et al. Continuous multimodal emotion prediction based on long short term memory recurrent neural network[C]//Proceedings of the 7th annual

(下转第 99 页)