DOI：10.11830/ISSN.1000-5013.201803011

表情符向量化算法

吴晨茜，陈锻生

（华侨大学 计算机科学与技术学院，厦门 361021）

摘要： 为了更加客观准确地判断微博的情感倾向，提出表情符向量化算法.首先，该算法将初始化表情符向量从随机产生改进为包含表情符语义信息的向量；然后，用随机产生的负向样本提高泛化能力.通过定性和定量分析可知：该算法能够保留表情符的语义信息；相对于忽略表情符的纯文本情感分析，在微博文本中融入表情符信息的微博情感分析能够提高微博情感分类的精度.

关键词： 表情符；表情符向量；卷积神经网络；情感分析；微博

中图分类号： TP 520.60 **文献标志码：** A **文章编号：** 1000-5013(2019)03-0399-06

Emoticon Vectorization Algrorithm

WU Chenxi, CHEN Duansheng

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract: In order to judge the emotional orientation of Weibo more objectively and accurately, an emoticon-vectorization algorithm is proposed. Firstly, the initialization emoticon vector isimproved from random generation to a vector containing emoticon semantic information; Secondly, the randomly generated negative samples are used to improve the generalization performance. Through qualitative and quantitative analysis, the algorithm can preserve the semantic information of emoticons. Compared with the plain text sentiment analysis that ignores emoticons, sentiment analysis of Weibo incorporating emoticon information in Weibo text can improve the accuracy of Weibo sentiment classification .

Keywords: emoticon; emoticon vector; convolutional neural network; sentiment analysis; Weibo

微博自 2007 年进入中国以来，在短时间内迅速崛起，用户数量激增.庞大的用户群产生的微博信息中含有大量的带有情感色彩的主观性文本.早期，文本情感分析的方法是基于情感词典的方法，如 Hu 等[1]借助 wordNet 词汇语义网构造情感词典；另一种文本情感分析的方法是基于机器学习的方法，如 Pang 等[2]使用传统机器学习方法提取文本特征，并对文本进行情感分类.深度学习在特征的抽取和模型的建立上更具有优势，文献[3-5]提出基于卷积神经网络的文本情感分析模型.表情符的快速发展引起学者们对自然语言处理的关注.林振扬[6]对表情符号所代表的文化涵义进行研究.张艳晖[7]提出融合表情符号和微博新词的文本情感分析算法.Davidov 等[8]充分利用 Twitter 中的标签和表情符，利用 KNN 分类器设计一个情感分类框架.随着对微博情感分析研究的深入，专家们对表情符越来越重视.谭文芳[9]阐述了表情符号的形成过程和在网络中的影响力等.Wang 等[10]通过聚类的方式得出表情符的含义和表情符的使用场景.Yang 等[11]通过分析词与表情符之间的关系，构建一个表情符情感词典.Jiang 等[12]抽取表情符的特征向量，构建一个表情符空间以判断其情感极性.张仰森等[13]结合情感词

收稿日期： 2018-03-09

通信作者： 陈锻生(1959-)，男，教授，博士，主要从事数字图像处理与模式识别的研究.E-mail：dschen@hqu.edu.cn.

基金项目： 国家自然科学基金资助项目(61370006)；福建省科技计划重点资助项目(2015H0025)

和表情符的特点,提出级联式的情感分类器模型.刘宝芹等[14]基于表情符号建立六维情绪向量,并应用于微博情感分类.目前,针对表情符的研究主要存在以下两个问题:1)针对表情符对微博情感分析的作用及表情符本身的研究还不够深入;2)表情符包含的丰富语义信息还没有被充分利用.为解决上述两个问题,本文提出表情符向量化算法.

1 表情符向量化

1.1 表情符号的选择和分类

微博中最受欢迎的表情符当属emoji表情符,如图1所示.目前,emoji表情符已经普遍应用于各大社交网络当中,其中一部分表情符具有明确的情感倾向,如😍具有积极情感倾向,而😩具有消极情感倾向.为保证实验结果,需要先对表情符进行选择和分类.1)收集大量的微博数据,统计每个表情符被使用的频次,将表情符按照频次的高低进行排序.2)选择前200个频次高的表情符,组织5名研究人员对这200个表情符进行情感分类.分类类别为积极、消极和中性.3)选择3名及3名以上研究人员标记相同的积极和消极表情符作为文中的研究对象.5名研究人员对部分表情符反应的分类情况,如图2所示.根据分类情况,选择使用频次高且具有明显倾向的53个表情符进行研究,如表1所示.表1中:积极表情符有28个;消极表情符有23个.

图1 微博中广泛使用的表情符

Fig. 1 Emoticons widely used in microblogs

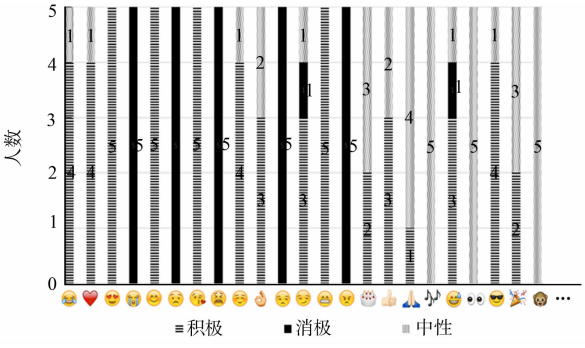

图2 表情符的分类情况

Fig. 2 Classfication of emotion

表1 表情符及其情感倾向

Tab. 1 Emoticon and its sentiment orientation

情感倾向	表情符
积极	😂 ❤️ 😍 😉 ☺️ 👌 😊 😌 😏 👍 😎 ✌️ 😁 😜 😋 👏 😆 😃 😄 😔 😝 😇 😛 😚 😑 😗 😘 😅
消极	😠 😒 😩 😔 😢 💔 😳 😞 😱 😡 😫 😤 😓 😰 😖 😣 👎 😨 😕 😧 😯 😵 😐

1.2 词向量训练

word2vec(https://code.google.com/p/word2vec/)是Google在2013年开源的一款将文本中的字词转化为计算机可以理解的数据形式的工具.它从未标记的训练集中无监督地学习词语间隐藏的信息,得到的词向量能够保留词语间句法和语义关系.

表情符向量化算法需借助word2vec工具,而采用word2vec工具则需要优质的语料库.维基百科语料库具有质量高、覆盖广等特点,已经在数据挖掘、自然语言处理等领域中发挥重要作用.因此,选择维基百科的中文语料库进行词向量的训练,训练结束后可得到词向量.词向量训练的流程图,如图3所示.

图3 词向量训练流程图

Fig. 3 Flowchart of training word vector

1.3 样本集构造

目前,针对表情符的研究还比较少,并没有一个合适的数据集能够将表情符转化为向量形式.因此,

文中借鉴 Eisner 等[15]的方法，根据表情符向量化算法的实际需要，构造一个表情符样本集. 样本由表情符图片、表情符名称和表情符描述性语句 3 部分组成. 表情符图片指的是文中选择的使用率较高，且具有明确情感极性的 53 个表情符，如😎.

表情符名称是样本的第 2 个组成元素. 通过查看微博源代码的方式可以发现，微博上的表情符都是以[XX]的文本格式存在，[XX]中的内容不仅是对表情符含义的简单描述，而且是唯一的. 因此，文中将[XX]中的内容作为表情符的名称；如😎是以[酷]的形式存在，😎这个表情符号的名称为“酷”.

full emoji list(http://www.unicode.org/emoji/charts/full-emoji-list.html)是一个表情符列表，其中详细记录了每个 emoji 表情符的编码和英文描述性短语等内容. 将英文描述性短语翻译成中文就构成样本中的第 3 个组成部分，即描述性语句. 通过这种方法构建样本集中的正向样本，如{😎，酷，戴着墨镜的笑脸}.

为保证实验结果的准确性，提高泛化能力，还需构建负向样本. 负向样本与正向样本的不同之处在于其描述性语句是通过随机产生的. 具体来说，随机产生 4～6 个中文词汇，将这些随机产生的中文词汇按顺序串联形成的短句作为与表情符号不相符的描述性语句，虽然此短句在语义和语法上不成立，但这并不影响构造负向样本的初衷. 如随机产生的 4 个词汇分别为“研究”、“前”、“大学”和“主要”，将这 4 个词汇依次连接形成“研究前大学主要”作为与表情符不相符的描述性语句，所以此时的负向样本为{😎，酷，研究前大学主要}. 部分表情符样本，如表 2 所示.

表 2　部分表情符样本

Tab. 2　Sample set of emoticon

表情符	正向样本(表情符，名称，相符的描述语句)	负向样本(表情符，名称，不符的描述语句)
😎	{😎，酷，戴着墨镜的笑脸}	{😎，酷，研究前大学主要}
❤️	{❤️，心，红色的心}	{❤️，心，苦无对策传话升上来}
😍	{😍，色，带爱心的笑脸}	{😍，色，贴膜二年制震裂存人玩伴}
😊	{😊，开心，微笑的脸}	{😊，开心，长殇冷衫石常大卫分明}
☺️	{☺️，热情，满意的脸}	{☺️，热情，理士码负平曲侯高吻}
👌	{👌，OK，可以的手势}	{👌，OK，本级两寸专力溶质}

1.4　表情符向量化算法步骤

提出的表情符向量化算法有以下 4 个具体步骤.

1) 初始化表情符向量 $\boldsymbol{x}_i$. 每个样本中包含表情符的名称，将表情符名称所对应的词向量 $\boldsymbol{w}_{\text{name}}$ 作为表情符向量的初始向量，如果表情符名称是未登录词，则随机初始化表情符向量. 即表情符向量 $\boldsymbol{x}_i=\boldsymbol{w}_{\text{name}}$. 表情符名称是对表情符含义的简单描述. 因此，初始的表情符向量已经包含一部分表情符的语义信息，这将更有利于表情符向量的形成.

2) 构造描述向量 $\boldsymbol{v}_j$. $\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_N$ 是一组词向量序列，分别对应样本中描述性语句中的词序列. 将这些词向量相加作为表情符的描述向量，即描述向量 $\boldsymbol{v}_j=\sum_{k=1}^{N}\boldsymbol{w}_k$. 描述向量实际上是描述性语句中各个词语对应词向量的和，它综合了描述性语句中所有词语的句法和语义信息.

3) 建立数学模型. 表情符向量 $\boldsymbol{x}_i$ 与描述向量 $\boldsymbol{v}_j$ 的点积可以表示这两个向量之间的相似性. 用 sigmoid 函数对表情符向量 $\boldsymbol{x}_i$ 与描述向量 $\boldsymbol{v}_j$ 的相似概率建模，即

$$P(\boldsymbol{y})=h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}_j)^{y}(1-h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}_j))^{1-y},\qquad h(x)=\frac{1}{1+\mathrm{e}^{-x}}. \tag{1}$$

4) 计算表情符向量 $\boldsymbol{x}_i$. 数据集 $D=\{(\boldsymbol{v}_j,\boldsymbol{y}_{i,j})\mid \boldsymbol{v}_j\in\mathbf{R}_n,\boldsymbol{y}_{i,j}\in\{0,1\}\}$ 由每个描述性向量 $\boldsymbol{v}_j$ 组成，当描述性语句 j 与表情符 i 相符合时，$\boldsymbol{y}_{i,j}$ 值为 1；否则，为 0. 对数据集 D 中的描述向量 $\boldsymbol{v}_j$ 计算式(1)的对数损失函数，其对数损失函数为

$$\begin{aligned}L(i,j,\boldsymbol{y}_{i,j})&=-\ln P(\boldsymbol{y})=-\ln\prod_{i,j}h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}_j)^{\boldsymbol{y}_{i,j}}(1-\mathrm{h}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}_j))^{1-\boldsymbol{y}_{i,j}}=\\&-\sum_{\mathrm{i,j}}\boldsymbol{y}_{i,j}\ln h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}_j)-\sum_{i,j}(1-\boldsymbol{y}_{i,j})\ln(1-h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}_j)).\end{aligned} \tag{2}$$

使用梯度下降算法，寻找最佳的 $\boldsymbol{x}_i$，从而得到表情符向量. 文中得到的表情符向量是一个 300 维向

量,样本集中的每一个表情符有一个对应的表情符向量,3个表情符的表情符向量为

$$V(😃)=\begin{pmatrix}1.157\,428\,145\,408\\-0.012\,668\,944\,895\\1.208\,384\,037\,017\\\vdots\\-1.370\,171\,785\,354\\0.379\,211\,932\,420\\0.851\,672\,232\,151\end{pmatrix}_{300\times1},\quad V(😎)=\begin{pmatrix}-0.022\,049\,268\,707\\1.809\,485\,912\,322\\-0.993\,784\,904\,479\\\vdots\\1.717\,911\,839\,485\\0.618\,068\,456\,649\\-1.126\,184\,701\,919\end{pmatrix}_{300\times1},\quad V(😭)=\begin{pmatrix}-0.442\,550\,122\,737\\0.277\,557\,462\,458\\1.277\,621\,030\,807\\\vdots\\1.352\,854\,490\,280\\-0.611\,459\,255\,218\\2.289\,661\,645\,889\end{pmatrix}_{300\times1}.$$

2 实验结果与分析

2.1 定性分析

word2vec能够将词汇映射成高维向量空间中的一个点,表情符向量化算法借助word2vec工具,能够将表情符映射到相同的向量空间中,且该向量空间中两个点的距离可以衡量两个元素之间的相似性.其中,点$X(x_1,x_2,\cdots,x_n)$和点$Y(y_1,y_2,\cdots,y_n)$之间的距离公式为

$$l=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}.$$

文中将词向量和通过表情符向量化算法得到的表情符向量映射到同一个向量空间中,并计算每个元素之间的距离,通过观察与表情符最相近的表情符和词汇定性分析表情符向量化算法的正确性和可行性.以4个目标表情符作为案例,在向量空间中,分析与目标表情符距离最近的5个表情符和5个词汇,如表3所示.

表3 向量空间中与目标表情符距离最近的表情符和词汇

Tab.3 Nearest emoticons and vocabulary in vector space from target emoticons

目标表情符	表情符	词汇
😃	😃😊😁😄☺😆	愉快、脸、丰满、欢迎、放松
😔	😔😭😟😞😢☹	担心、哀伤、生气、混乱、激怒
😚	😔😍😘☺😙😂	接吻、情绪、拥抱、情感、心痛
😜	😔😝😛😉😚😋	活泼、快乐、爱玩、讥讽、累赘

由表3可知:与目标表情符距离最近的5个表情符都与目标表情符有着相同的情感极性,而且它们的语义信息也很类似.以目标表情符😜为例,该表情符最突出的是“吐舌头”这一动作,且面部表情较为俏皮,而与该表情符距离最近的5个表情符中有3个表情符含有“舌头”这一元素,一个表情符含有俏皮的意味.由此可以推断,相似的表情符在表情符向量空间中距离较近,而且表情符向量保留表情符原本的语义信息.

由表3还可知:与表情符最相似的词汇并不是描述表情符号的词汇,而是与表情符号表达相同或相似语义的词汇.表情符向量化算法虽然从表情符号的描述性语句入手,但是取得的向量化表示与描述性词汇相关性低,这从侧面反映该算法的合理性和可行性.

2.2 定量分析

2.2.1 基于卷积神经网络的分类模型 卷积神经网络中发挥重要作用的是卷积层和池化层.卷积层能够提取出输入数据中大量的局部特征和语义组合.池化层是在卷积层的基础上,对局部特征和语义组合进行选择,过滤掉一些不重要的局部特征和可置信低的语义组合.多个卷积层和池化层的交替叠加,可以将文本数据中高度抽象的特征提取出来,提高情感分类的精度.

基于卷积神经网络分类模型的示意图,如图4所示.图4中:d为维度.该分类模型的输入是一个句子矩阵,句子矩阵由句子中所有词对应的词向量依次连接形成的.卷积层中使用窗口长度h不同的卷积过滤器作用于输入矩阵中所有长度为h的相邻词向量上,以提取输入层的局部特征.池化层对提取出的局部特征进行筛选,将池化层的输出用全连接的方式连接到最后一层的输出结点上,利用softmax分类器进行微博情感分类.

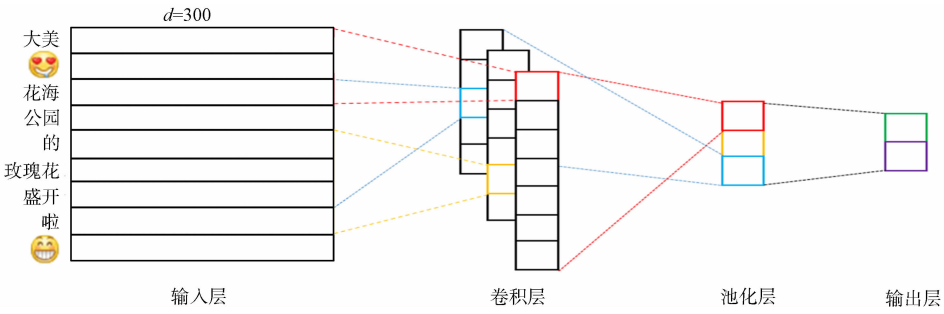

图 4　基于卷积神经网络分类模型

Fig. 4　Classification model based on convolution neural network

2.2.2　实验数据和模式　实验在 Linux 12.04 环境下进行，GPU 为 NVIDIA Quadro P4000，内存为 32 G DDR4，使用 Python 2.7 和 Theano 0.9 框架. 以 NLPCC2014 任务 1 提供的语料作为实验数据. 该语料将微博分为生气、厌恶、害怕、高兴、喜欢、伤心和惊喜 7 类. 文中将高兴、喜欢和惊喜划分为积极类别；将生气、厌恶、害怕和伤心划分为消极类别. 将积极类别和消极类别进行二元情感分类，使用基于卷积神经网络的句子分类模型，通过以下 3 种不同的模式进行对比实验.

1）模式 1：word2vec＋CNN. 仅考虑微博语料中的微博文本，剔除表情符号，将实验数据中包含的所有词对应的词向量作为基于 CNN 句子分类模型的输入，进行微博情感分类实验.

2）模式 2：emoji2word＋word2vec＋CNN. 在微博环境中，表情符号是以[XX]的文本形式存在的. 因此，先将实验数据中的所有表情符号转化为对应的文本，形成一个纯文本的实验数据；再在纯文本实验数据的基础上训练出词向量，并将词向量作为基于卷积神经网络的句子分类模型的输入.

3）模式 3：emoji2vec＋word2vec＋CNN. 将表情符向量和词向量连接起来构成句子矩阵，并作为基于卷积神经网络的句子模型的输入进行情感分类.

采用上述 3 个模型在两个语料上进行实验分析，语料 1 是 NLPCC2014 已标注的共 45 423 条微博（包含 2 906 条带表情符的微博），其中，积极数据和消极数据分别占全体数据的 52.5％和 47.5％. 语料 2 是 NLPCC2014 中已标注且带有表情符的 2 906 条微博，其中，积极数据和消极数据分别占全体数据的 61.2％和 38.8％.

2.2.3　实验结果与分析　实验以正确率（η）、召回率（ϕ）、F 值和准确率（ψ）作为实验的评价指标，具体的实验结果，如表 4 所示.

表 4　3 种模式的评价指标

Tab. 4　Evaluation indicators of three models

模式	类别	语料 1				语料 2			
		η/％	ϕ/％	F/％	ψ/％	η/％	ϕ/％	F/％	ψ/％
模式 1	积极	73.4	79.1	76.1	80.4	78.3	72.7	75.4	73.1
	消极	70.3	81.7	75.6	80.4	75.4	74.3	84.7	73.1
模式 2	积极	74.1	80.5	77.2	81.7	74.1	71.3	72.2	70.3
	消极	72.5	77.4	74.9	81.7	76.0	73.3	74.6	70.3
模式 3	积极	76.9	83.4	80.0	83.5	80.7	77.8	79.2	78.4
	消极	77.2	82.0	83.5	83.5	76.2	75.9	76.1	78.4

由表 4 可知：在语料 1，2 中，模式 3 的正确率、召回率和准确率都高于模式 1. 这说明将表情符转化为向量并且将其作为特征引入后，情感分类器的性能得到一定程度的提高.

由表 4 还可知：通过对比模式 1，2 的实验结果，在语料 2 中，将表情符号转化为文字后，其准确率略微降低，其主要原因是由表情符转化而来的文字并不能完全替代表情符在微博中包含的语义信息；而在语料 1 中，其准确率却略有提升，可能的原因是语料 1 中包含文字较多，微博整体的情感倾向对表情符的依赖性并不强. 因此，从实验结果的不确定性可以看出，单纯的将表情符转化为文字的做法并不适用于微博情感分类领域. 通过对比模式 1，3 的实验结果，融合表情符特征的基于卷积神经网络分类模型能够提高微博情感分类的准确率. 因此，相对于忽略表情符的纯文本情感分析，在微博文本中融入表情符

向量的微博情感分析,在一定程度上能够提高微博情感分类的精度,说明表情符向量化算法在判断微博情感倾向中可发挥重大作用.

3 结束语

由于表情符不仅自身具有情感倾向,而且对微博的整体情感倾向也有影响,因此,提出表情符向量化算法,通过提取表情符的特征,将表情符号转化为向量形式,让表情符与词汇一样能够在情感分析领域中灵活应用.通过定性分析可知,表情符向量化算法保留了表情符的语义信息.通过定量分析可知,表情符向量能够提高情感分类器的性能.

参考文献:

[1] HU Mingqing,LIU Bing. Mining and summarizing customer reviews[C]// Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle: ACM,2004:168-177. DOI:10.1145/1014052.1014073.

[2] PANG Bo,LEE L,VAITHYANATHAN S. Thumbs up?: Sentiment classification using machine learning[C]// Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics,2002:79-86. DOI:10.3115/1118693.1118704.

[3] SEVERYN A, MOSCHITTI A. Unitn: Training deep convolutional neural network for twitter sentiment classification[C]// Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver: Association for Computational Linguistics,2015:464-469.

[4] 李松如,陈锻生.采用循环神经网络的情感分析注意力模型[J].华侨大学学报(自然科学版),2018,39(2):252-255. DOI:10.11830/ISSN.1000-5013.201606123.

[5] SANTOS C N D,GATTI M. Deep convolutional neural networks for sentiment analysis of short texts[C]// the 25th International Conference on Computational Linguistics. Ireland:COLING,2014:69-78.

[6] 林振扬.网络表情符号的符号学阐释[J].美术大观,2016(2):128. DOI:10.3969/j.issn.1002-2953.2016.02.067.

[7] 张艳辉.融合表情符号的微博文本倾向性分析[D].济南:山东师范大学,2015.

[8] DAVIDOV D,TSUR O,RAPPOPORT A. Enhanced sentiment learning using Twitter hashtags and smileys[C]// COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics. Beijing: Association for Computational Linguistics,2010:241-249.

[9] 谭文芳.网络表情符号的影响力分析[J].求索,2011(10):202-204.

[10] WANG Hao,CASTANON J A. Sentiment expression via emoticons on social media[C]// Proceedings of the 2015 IEEE International Conference on Big Data. Washington D C: IEEE Press,2015:2404-2408. DOI:10.1109/BigData.2015.7364034.

[11] YANG Changhua,LIN H Y,CHEN H H. Building emotion lexicon from weblog corpora[C]// Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Stroudsburg: Association for Computational Linguistics,2007:133-136. DOI:10.3115/1557769.1557809.

[12] JIANG Fei,LIU Yiqun,SUN Jiashen,*et al*. Microblog sentiment analysis with emoticon space model[J]. Journal of Computer Science and Technology,2015,30(5):1120-1129. DOI:10.1007/s11390-015-1587-1.

[13] 张仰森,孙旷怡,杜翠兰.一种级联式微博情感分类器的构建方法[J].中文信息学报,2017(5):183-189. DOI:10.3969/j.issn.1003-0077.2017.05.025.

[14] 刘宝芹,牛耘,张景.基于统计数据的微博表情符分析及其在情绪分析中的应用[J].计算机工程与科学,2016,38(3):577-584. DOI:10.3969/j.issn.1007-130X.2016.03.027.

[15] EISNER B,ROCKTÄSCHEL T,AUGENSTEIN I,*et al*. Emoji2vec: Learning emoji representations from their Description[C]// The Fourth International Workshop on Natural Language Processing for Social Media. Austin: Association for Computational Linguistics,2016:48-54. DOI:10.18653/v1/W16-6208.

(**编辑:** 李宝川 **责任编辑:** 陈志贤 **英文审校:** 吴逢铁)