# 基于卷积神经网络的多维特征微博文本情感分析*

余 鹏 田 杰
（深圳供电局有限公司电力科学研究院 深圳 518000）

**摘 要** 以word2vec工具进行词向量运算，根据短文本语义特征，采用卷积神经网络模型提取出深度抽象特征，再对分类器进行训练来实现情感分类的目的。分析基于卷积神经网络的多维特征微博文本情感，通过F值和准确率来衡量实际分类效果，分析结果表明：相对于机器学习模型，该微博情感分析模式使情感分析和F值准确率依次增大了0.1060与0.1320。采用卷积神经网络和多维度文本特征分析方法可以有效提升微博情感分析的效果。

**关键词** 情感分析；卷积神经网络；微博文本；表情字符
**中图分类号** TP391 **DOI**：10.3969/j.issn.1672-9722.2020.09.032

# Analysis of Emotion of Micro-blog Based on Convolution Neural Network

YU Peng TIAN Jie
（Institute of Electric Power Science, Shenzhen Power Supply Bureau Co., Ltd., Shenzhen 518000）

**Abstract** The word2vec tool is selected to complete the operation of word vectors, semantic features are extracted from the short text, deep abstract features are extracted from the convolutional neural network model, and then the classifier is trained to achieve the purpose of emotion classification. By analyzing the multi-dimensional characteristics of micro-blog emotions based on convolutional neural network, the actual classification effect is measured by F value and accuracy. The analysis results show that, compared with the machine learning model, this micro-blog emotion analysis model increases the accuracy of emotion analysis and F value by 0.1060 and 0.1320 respectively. Convolutional neural network and multi-dimensional text feature analysis can effectively improve the effect of microblog emotion analysis.

**Key Words** sentiment analysis, convolutional neural networks, Weibo text, emoticons
**Class Number** TP391

## 1 引言

目前，各类互联网应用技术快速发展，产生了贴吧、论坛、网站等多种信息交流平台，人们可以通过多种渠道发表自己的评论或与他人分享照片及各类感兴趣的事物等，也可以从这些网络渠道收集资讯与新闻等信息[1~5]，这些情况都促进了各类社交软件的快速发展。其中，新浪微博属于一个建立在用户关系基础上进行信息传递、分享、收集的数据处理平台，使传统社交网络的交互模式发生了显著变化，使用户可以更加快速、方便地获取各类所需的信息，因此在短时间内新浪微博就成为了一个具有极高人气的新媒体社交平台[6~8]。大部分互联网用户可以利用微博作为热点事件的获取来源，并对事态进行实时关注及发表自己的观点；还有一些部门在微博上建立了自己的官方号，通过官方微博来及时发布一些事件的实时进展，便于迅速澄清事实以及快速回应民众的各项需求。由此可见，如何选择合适的方法来准确分析通过微博平台发布的内容已经成为一项重要的事情[9~10]。现阶段，大量学者都对Twitter社交平台开展了深入研究。例如，采用传统方式对Twitter进行情感分析时通常都是利用词典情感分析的方式进行处理，可以利用包含情感词性、否定等副词等来描述句子情感状态[11~12]。

* **收稿日期**：2020年3月8日，**修回日期**：2020年4月10日
**作者简介**：余鹏，男，硕士，高级工程师，研究方向：输变电技术。

很少有关于中文短文本方面的情感分析，一般都是先提取出文本的情感特征，再利用分类学习算法来实现情感分析的功能。各类网络用语与互联网新词也不断变化，这使得词典的维护与更新过程也变得更加困难[13]。现阶段，还很少有学者在中文微博情感分析方面使用深度学习模型[14~15]。

## 2　方案设计

本文构建得到了一种新的具有多维特征的微博情感分析制度。从图 1 中可以看到对多维特征微博情感进行分析的具体流程。通过分析微博文本多维度特征可以发现主要包含了情感和语义共两方面的特征。其中，语义特征需要采用无监督学习的方法来完成大规模语料的训练，同时计算出词组对应的词向量。利用词向量将词组映射至高维空间中，从而实现向量化转变的过程，接着利用高维空间内的词向量余弦距离表示词组相似性，也同时包含了不同词组间的更深层语义关系。在微博文本中可以从表情字符中分析出具体的情感特征。采用一定的方法提取得到微博文本中的所有表情字符并完成转换过程，之后再利用随机向量化的方式进行词向量匹配，建立由微博文本构成的特征集合。通过卷积与池化的算法获取局部特征并筛选得到局部特征，以上述各项特征作为情感分类器输入，并对微博文本的情感分类器训练。

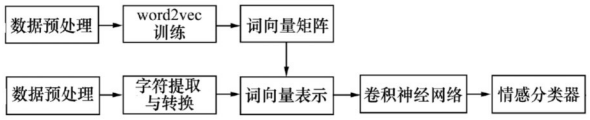


图 1　微博文本情感分析流程图

## 3　卷积神经网络模型

### 3.1　模型概况

从图 2 中可以看到，本文构建的卷积神经网络包含了三部分内容。其中，第 1 部分属于输入层，都是通过词组构成的词向量形式来表达。对数据进行预处理后可以得到一条包括 $n$ 个词组的微博文本 $s$，将其表示为 $\{x_1, x_2, \cdots, x_n\}$，这些元素基本都是文本语义特征词，当表情字符出现在微博内容中时，则会形成和该表情相匹配的情感特征词组。可以发现，各词组 $x_i$ 都可以被看成是一个 $d$ 维向量，此时可通过 word2vec 工具训练获得。其中，1 条微博文本对应 1 个 $n\times d$ 矩阵。第 2 部分为卷积层，需通过卷积与池化运算来实现。在模型中可以设定各个长度的卷积核来获取所需的特征，可以将各卷积核 $w$ 都看成 $h\times d$ 矩阵，以 $h$ 表示卷积核的长度。采用卷积核实现对微博文本的卷积过程，具体表达式如下：

$$t_i = f(w * s_{i:i+h-1} + b)$$

上式中，$s_{i:i+h-1}$ 代表介于第 $i$ 个词组到 $i+h-1$ 个词组之间的连续片段；$f$ 代表非线性激活函数 ReLU；* 是卷积运算符号；$t_i$ 代表第 $i$ 个卷积特征。通过卷积核 $w$ 来实现对微博文本的卷积过程，由此得到特征集合 $T=\{t_1, t_2, \ldots, t_{n-h+1}\}$，考虑到各个微博文本在词组数量方面存在明显差异，而且每个卷积核长度也存在一定的差异，从而导致特征集合长度也发生变化。此模型是以最大值池化操作的过程进行计算，从特征集合中选择最大值来表示特征值。

$$\hat{t} = \max(\{t_1, t_2, \ldots, t_{n-h+1}\})$$

可以利用池化操作来确保在不同长度的卷积核条件小得到具有相同长度的特征向量。分类层位于第 3 部分，类别概率计算结果如下：

$$P_j = P(y=j|X, b) = \frac{e^{W_j X + b_j}}{\sum_{i=1}^{L} e^{W_i X + b_i}}$$

上式的 $P_j$ 代表第 $j$ 类概率；$X$ 对应分类层输入，$W$ 对应权值矩阵；$b_i$ 与 $b_j$ 依次代表偏置项 $b$ 第 $i$ 和第 $j$ 个元素；$L$ 是类别数量。

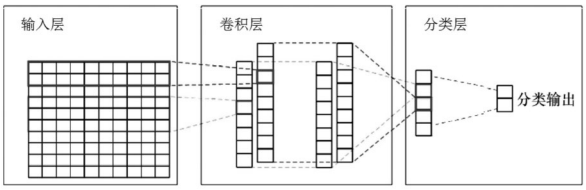


图 2　卷积神经网络控制示意图

### 3.2　正则化

该模型是经过归一化运算获得的网络层。第一类通过卷积运算实现单个卷积核的卷积特征归一化，之后再实施池化运算；第二类是加入归一化层后再计算激活值，具体表达式如下所示：

$$z = f(WX + b)$$

采用归一化层得到激活值的计算公式是

$$z = f(g(WX))$$

上式的 $g$ 代表归一化函数。因为归一化运算不会对偏置产生影响，所以把偏置项去除。

## 4　基于微博表情字符的情感特征

用户可以通过微博平台获得多种不同的默认表情符，从而更加形象地体现自己想要表达的感想。从图 3 中可以看到微博为用户提供的所有默认表情符。选择传统自然语言分析博客等各类语

料时,只对文本自身含义进行分析与信息提取,对文本进行预处理时只对文本信息间过滤,同时将所有网页链接以及各类特殊字符全部删除,由此便会引起缺少微博文本情感的结果。

图3 表情库

进行数据预处理时,应提取图像标签并对其进行转换,再把转换获得的表情字符插入对应微博文本的位置,之后再采用方括号标注上述表情字符,从而有效区分微博包含的各类表情和文本信息。把经过处理的表情字符通过随机初始化的模式使其转化为相应的词向量,并跟语义特征形成一致的状态,由此完成情感特征以及语义特征之间的融合。

## 5 测试分析

### 5.1 数据集

在数据集中含有 word2vec 训练语料与微博数据。先对新闻正文内容进行分词,以此构建 word2vec 工具对数据进行训练,再把词向量的长度设定在 $d$=300。完成以上训练后,可以生成共 51.2685 万个词组。再以随机初始化方式处理词向量集合不包含的词组。由于微博数据集模型需处理表情字符,不能得到公开数据集,所以采用自行采集的过程得到 1 万条左右的微博文本,再利用人工标注的方法分类得到消极与积极共二种类型,各样本见表 1。

表1 积极文本和消极文本示例

| 积极文本 | 消极文本 |
|---|---|
| 你的美无法形容了<br>威风轻轻起,穷好喜欢你 | 电视剧能把人感动哭了<br>爱人得病了,无法为其分担 |
| 这个主人公获得了属于自己的爱情 | 这个主人公应公殉职,好让人难受 |
| 努力拼搏后实现了自己的美好愿望 | 努力的过程中承受了太多的委屈 |

### 5.2 结果与讨论

总共进行了 3 组情感测试。其中,第 1 组采用本文的微博情感分析模式,测试参数设置见表 2。第 2 组是没有感情特征的建立在卷积神经网络基础上的模型,是根据本文模型把所有表情字符除去后再将其表示为不含情感特征的模型。第 3 组是以情感词典为基础建立的机器学习模型。通过 F 值和准确率来衡量实际分类效果。以 $P_{acc}$ 表示准确率,$P_{prec}$ 表示精确率,$P_{recall}$ 表示召回率,以上各值与 $F$ 值可以通过如下式子进行计算:

$$P_{acc}=\frac{T_P+T_N}{T_P+F_P+T_N+F_N}$$

$$P_{prec}=\frac{T_P}{T_P+F_P}$$

$$P_{recall}=\frac{T_P}{T_P+F_N}$$

$$F=\frac{2P_{prec}\cdot P_{recall}}{P_{prec}+P_{recall}}$$

以上各组测试都通过 10 折交叉计算来完成。从表 3 中可以看到各个测试的结果。

表2 测试参数设置

| 参数 | 取值 |
|---|---|
| 卷积核维度 | 1~5 |
| 单元数量/个 | 150 |
| 节点数/个 | 60 |
| 迭代数 | 50 |

表3 测试结果

| 模型 | 准确率 | 精确率 | 召回率 | F 值 |
|---|---|---|---|---|
| 机器学习 | 0.7025 | 0.6886 | 0.7386 | 0.7186 |
| 无情感 | 0.8306 | 0.8412 | 0.8024 | 0.8226 |
| 有情感 | 0.8516 | 0.8588 | 0.8426 | 0.8506 |

通过分析表 3 可以发现,相对于机器学习模型,本文构建的微博情感分析模式使情感分析和 F 值准确率依次增大了 0.1060 与 0.1320。根据以上结果可知,综合运用多维度文本特征与卷积神经网络方法更有助于进行微博情感分析。

## 6 结语

利用 word2vec 工具处理词向量,根据短文本内容获取语义特征,并将微博文本中的表情字符作为情感特征,由此构建得到特征集合;通过卷积神经网络模型提取深度抽象特征,同时训练分类器实现情感分类的过程。通过 F 值和准确率来衡量实际分类效果,相对于机器学习模型,本文构建的微博情感分析模式使情感分析和 F 值准确率依次增大了 0.1060 与 0.1320。采用卷积神经网络和多维度文本特征分析方法可以有效提升微博情感分析

的效果。

## 参 考 文 献


[1] TABOADA M, BROOKE J, TOFILOSKI M, et al. Lexiconbasedmethods for sentiment analysis[J]. ComputationalLinguistics, 2011, 37(2): 267-307.

[2] THELWALL M, BUCKLEY K, PALTOGLOU G. Sentimentstrength detection for the social web[J]. Journal of the AmericanSociety for Information Science & Technology, 2012, 63(1):163-173.

[3] SEVERYN A, MOSCHITTI A. Twitter sentiment analysis withdeep convolutional neural networks[C]// The International ACMSIGIR Conference. Santiago, Chile: ACM, 2015: 959-962.

[4] TANG Duyu, WEI Furu, QIN Bing, et al. Coooolll: A deeplearning system for twitter sentiment classification [C]//Proceedings of the 8th International Workshop on SemanticEvaluation. Dublin, Ireland: ACL, 2014: 208-212.

[5] KIM Y. Convolutional neural networks for sentenceclassification[J]. Eprit Arxiv, 2014-08-25.

[6] ZHANG Xiang, ZHAO Junbo, LECUN Y. Character-levelconvolutional networks for text classification[C]// Advances inNeural Information Processing Systems. New York, USA:Curran Associates, Inc., 2015: 649-657.

[7] 黄发良, 冯时, 王大玲. 基于多特征融合的微博主题情感挖掘[J]. 计算机学报, 2017, 40(4): 872-888.
HUANG Faliang, FENG Shi, WANG Daling. Emotion mining of microblog themes based on multi-feature fusion [J]. Journal of computer science, 2017, 40(4): 872-888.

[8] 张冬雯, 杨鹏飞, 许云峰. 基于word2vec 和SVMperf 的中文评论情感分类研究[J]. 计算机科学, 2016, 43(S1): 418-421,447.
ZHANG Dongwen, YANG Pengfei, XU Yunfeng. Classification of Chinese comment emotions based on word2vec and SVMperf [J]. Computer Science, 2016, 43(S1): 418-421,447.

[9] 黄仁, 张卫. 基于word2vec 的互联网商品评论情感倾向研究[J]. 计算机科学, 2016, 43(S1): 387-389.
HUANG Ren, ZHANG Wei. Research on emotional tendency of Internet commodity reviews based on word2vec [J]. Computer Science, 2016, 43(S1): 387-389.

[10] COLLOBERT R, WESTON J, BOTTOU L, et al. Naturallanguage processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(1): 2493-2537.

[11] 黄梦莹,张晓滨. 融合CHI与信息增益的情感文本特征选择[J]. 西安工程大学学报, 2018, 32(06): 713-717.
HUANG Mengying, ZHANG Xiaobin. Selection of emotional text features integrating CHI and information gain [J]. Journal of Xi'an University of Technology, 2008, 32(06):713-717.

[12] 林江豪,顾也力,周咏梅. 基于表情符号的情感词典的构建研究[J]. 计算机技术与发展,2019, 22(06):1-5.
LIN Jianghao, GU Yili, ZHOU Yongmei. Research on the construction of emoticons based emotion dictionary [J]. Computer Technology and Development, 2019, 22(06):1-5.

[13] 王景中,庞丹丹. 基于L-STM 模型的中文情感分类[J]. 计算机工程与设计,2018,39(11):3438-3443.
WANG Jingzhong, PANG Dandan. Classification of Chinese emotions based on l-stm model [J]. Computer Engineering and Design, 2008,39(11):3438-3443.

[14] 尹光花,刘小明,张露. 基于LSTM 特征模板的短文本情感要素分析与研究[J]. 电子科技,2018,31(11): 38-41,46.
YIN Guanghua, LIU Xiaoming, ZHANG Lu. Analysis and research of emotion elements in short text based on LSTM feature template [J]. Electronic Technology, 2008,31(11):38-41,46.

[15] 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[J]. 计算机系统应用,2018,27(11):205-210.
NIU Chengming, ZHAN Guohua, LI Zhihua. Analysis of emotional tendency of micro-blog based on deep neural network [J]. Computer System Application, 2008, 27(11):205-210.