文章编号：1003-0530(2021)06-1066-09

基于全局特征图的半监督微博文本情感分类

方 澄[1] 李 贝[2] 韩 萍[1]
（1. 中国民航大学电子信息与自动化学院，天津 300300；2. 中国民航大学经济与管理学院，天津 300300）

摘 要：网络社交的流行与普及，使得微博等短文本区别于以往传统文章，具有了独有的文学表达形式和情感发泄方式，导致基于短文本的机器学习情感分析工作难度逐渐增大。针对微博短文本的语言表达新特性，爬取收集大量无情感标记微博数据，建立微博短文本语料库，基于全局语料库构建词与短文本的全局关系图，使用 BERT(Bidirectional Encoder Representations from Transformers)文档嵌入作为图节点的特征值，采用图卷积进行节点间的特征传递和特征提取。采样部分无情感标记微博数据进行人工标注，采用半监督机器学习方法结合全局关系图提高情感分类器的性能，实验表明通过无情感标记数据比例的增加，该方法可以更好地捕捉全局特征，提高情感分类的精度。在自建人工标记数据、COAE2014 数据集和 NLP&CC2014 数据集上进行了对比实验，实验结果表明该方法在精确率和召回率上均具有很好的表现。

关键词：微博文本；情感分析；图卷积；半监督
中图分类号：TP3 **文献标识码**：A **DOI**：10.16798/j.issn.1003-0530.2021.06.018

引用格式：方澄，李贝，韩萍. 基于全局特征图的半监督微博文本情感分类[J]. 信号处理，2021，37(6)：1066-1074. DOI：10.16798/j.issn.1003-0530.2021.06.018.

Reference format：FANG Cheng，LI Bei，HAN Ping. Semi-supervised microblog text sentiment classification based on global feature graph[J]. Journal of Signal Processing，2021，37(6)：1066-1074. DOI：10.16798/j.issn.1003-0530.2021.06.018.

Semi-supervised Microblog Text Sentiment Classification Based on Global Feature Graph

FANG Cheng[1] LI Bei[2] HAN Ping[1]
（1. College of Electronic Information and Automation，Civil Aviation University of China，Tianjin 300300，China；
2. College of Economics and Management，Civil Aviation University of China，Tianjin 300300，China）

Abstract：Online social networks have gradually become popular and popularization. A number of social networks such as microblog have formed a unique form of literary and emotional expression. Because the expression of microblog is different from the expression of traditional articles，the sentiment analysis research based on short-text machine learning has become more and more difficult. Aiming at the new features of Microblog short text language expression，we crawl and collect a large amount of non-emotionally labeled Microblog data，and build a Microblog short text corpus to create a global relationship graph between words and short texts. The BERT (Bidirectional Encoder Representations from Transformers) document embedding is used as the feature value of the graph node，and graph convolution is used for feature transfer and feature extraction between nodes. We manually annotate non-emotionally labeled Microblog data which sample from the whole Microblog short text corpus. A semi-supervised machine learning method combined with global relationship graph is proposed to improve the performance of sentiment classifier. Experiments show that by increasing the proportion of unmarked data，the method can better capture global features and improve the accuracy of sentiment classification. Comparative experiments

收稿日期：2021-03-22；修回日期：2021-04-13
基金项目：中央高校基本科研业务费专项资金(3122018C005)

are carried out on self-built artificial labeling data, COAE2014 data set and NLP&CC2014 data set. The experimental results show that the method has a good performance in accuracy and recall.

Key words: microblog text; sentiment analysis; graph convolutional network; semi-supervised

1　引言

随着网络技术和智能终端设备的不断发展演变,人们的生活习惯也随之发生改变,特别是人们的阅读方式和情感表达方式:从纸质文献到网页文章,再到个人博客,直至今天的即时短文本,阅读表达的篇幅越来越短,速度越来越快。微博作为一种即时短文本分享方式,成为了人们表达观点、分享新闻、宣泄情绪的一种重要途径。因此通过对微博文本的情感分析,可以挖掘用户情感、了解网络新闻热点以及进行舆情监控,及时有效的了解网络热点和舆情动向,对于市场营销、信息安全等领域均有重要意义。

自然语言处理的核心目标是使计算机像人一样可以读懂文本内容。相对于语音和视觉,自然语言处理需要对文本进行更高度的抽象化表示。传统的自然语言处理方法主要分为三类:1)基于语言学的处理方法;2)基于规则逻辑的处理方法;3)基于机器学习的处理方法。近些年随着深度学习技术在计算机视觉、语音识别等领域的重大突破,使得基于机器学习的自然语言处理主要采用深度学习方法,并取得了一些令人瞩目的成绩[1-6]。文本情感分析作为自然语言处理的一个重要分支,同样将深度学习技术视为最为重要的分析方法。文本情感分析又称为意见挖掘,是对人们带有主观感情色彩的观点、评论、态度等文本进行计算和推理的过程。CNN(Convolutional Neural Network)和 RNN(Recurrent Neural Network)是文本情感分析计算的两种主要网络,在这两种主干网络的基础上,研究者们扩展出了众多新型网络。Nal Kalchbrenner 等人[7]使用动态 K-Max 池构建动态卷积神经网络(DCNN),将句子进行语义建模,捕获语句中短距离和长距离间的关系特征,该网络在多个语句情感分析任务上都取得了非常出色的表现。Kai Sheng Tai 等人[8]将语法的结构因素加入到序列化 LSTM(Long Short-Term Memory)模型中,使得文本情感分析结合了语义分析和句法分析的特征。深度学习的注意力(Attention)机制借鉴了人类的注意力思维方式,可以更好的模拟表示人类对事物的理解能力,注意力机制与 CNN/LSTM 结合可以更好提高文本情感分析的准确率。Bahdanau 等人[9]将 Attention 机制应用于自然语言处理领域,在机器翻译任务上取得了很好的效果。随后 Google 团队[10]提出在机器翻译上大量使用自注意力(self-attention)机制来学习文本表示,使得注意力机制成为自然语言处理的热点。虽然注意力机制可以解决语句计算中的一些局限性问题,但人类在情感表达形式上,具有多样化表达的特点,因此需要更加复杂的模型表达方式。

在利用深度学习模型进行文本情感分析时,除了需要更丰富的网络模型表示外,情感分析的研究还需要大量丰富的词典资源支持。在词典资源上构建词向量语言模型,就是用一个固定长度的向量来表示文本实质。将词向量作为网络模型的输入,可以取得更好的文本分析效果。2013 年,Google 团队[11]提出 word2vec 词向量模型,该模型主要包含两种:skip-gram 模型和 CBOW(continuous bag of words)模型,word2vec 模型开启了词向量自然语言分析的新时代。但 word2vec 模型缺少对全局上下文的观测能力,不能表达语句间的全局特征。ELMo 模型[12]对 word2vec 模型进行了改进,用两个单向模型表示双向语言模型,但 ELMo 模型的单向模型只能注意到文本固定方向的表达即前向的词语表达,不能关注文本后面词语。BERT 模型[13]全称 Bidirectional Encoder Representation from Transformers,其创新性地使用了 MLM(Masked Language Model)策略,在预训练过程中真正实现了双向语言模型,且在多个自然语言任务测试集上取得了最佳成绩,验证了 BERT 模型的有效性。由于中文文本内部包含了更加丰富的语义信息,因此在中文文本情感分析领域,中文词向量语言模型对情感分析的准确性更加重要。Li yanran[14]等人提出了一种更加关注汉字部首信息的中文词向量方法,提升了文档分类的准确性。Chen xinxiong[15]等人针对中文词组含有丰富表达,但一些词语拆分后单个汉字没有语义的问题,提出

了多原型特征嵌入的方法,增强了词向量的表达能力,一定程度上提升了中文词向量的有效性。另外除了需要大量丰富的词典资源构建词向量模型,文本情感分析还需要大量的标注文本,但中文文本标注数据集的质和量都有待提高。

针对以上问题,本文提出一种使用 BERT 词向量特征的图卷积文本情感分析方法(BERT Graph Convolutional Network-- BGCN),更丰富的表达全局文本特征,主要创新点包括:(1)使用微博短文本构建一个全局异构中文表示图,用图来表示词、文本之间的关系;(2)采样微博短文本,进行情感极性标注,基于采样的有限标注数据集,采用半监督的方法对微博数据进行情感极性判断,通过无标签数据的特征传递提高分类的准确度,同时解决标注样本不足的问题。(3)使用图卷积神经网络对微博短文本进行情感分析,将 BERT 词向量作为特征,提高模型性能。通过大量实验对比表明,该方法可以提高微博情感分析的准确率和性能。

2 相关工作

2.1 图卷积神经网络

图模型可以刻画不同节点间的连接关系,对于图模型的研究常被用来解决分子结构分类、社区关系图谱、自然语言文本分类等相关问题。Franco 等人在文献[16]提出了图神经网络(Graph Neural Network,GNN)的概念,使得图神经网络得到了学术界更广泛的关注。2017 年 Kipf 和 Welling 在文献[17]中提出一种采用卷积操作的图卷积神经网络(Graph Convolutional Network,GCN)模型,该模型在自然语言处理的多个实际任务中达到了最好的结果表现,使得图卷积神经网络及其相关改进方法成为了目前深度学习领域一个热点的研究方向。图卷积神经网络就是使用一种类似卷积神经网络的通用范式来对图模型进行特征提取的方法。图通常定义为 $G=(V,E)$,由节点集 V 及连接节点的边集 E 组成,节点集 $V(|V|=n)$ 中每个节点 ν 具有 m 维的特征向量,所有图节点构成图特征矩阵 $\boldsymbol{L}_{(j)}$,$\boldsymbol{X}$ 中的一行 $\boldsymbol{x}_i \in R^m$ 表示对应节点 ν_i 的特征向量。由节点集 V 及边集 E 构成的图 G 可以用邻接矩阵 $\boldsymbol{A}$ 表示。图 G 中每个节点的度数构成度矩阵 $\boldsymbol{D}$,度矩阵 $\boldsymbol{D}$ 是一个对角矩阵,对角线上的元素为对应节点的连接边数。图卷积神经网络特征提取过程可以表示为:

$$\begin{cases} \boldsymbol{L}_{(j)}=\rho(\hat{\boldsymbol{A}}(\boldsymbol{L}_{(j-1)})\boldsymbol{W}_{(j-1)}) \\ \boldsymbol{L}_{(j-1)}=\rho(\hat{\boldsymbol{A}}(\boldsymbol{L}_{(j-2)})\boldsymbol{W}_{(j-2)}) \\ \vdots \\ \boldsymbol{L}_{(1)}=\rho(\hat{\boldsymbol{A}}(\boldsymbol{X})\boldsymbol{W}_0) \end{cases} \tag{1}$$

其中 $\hat{\boldsymbol{A}}=\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$ 为正则化对称邻接矩阵,由于在图 G 建立后,矩阵 $\boldsymbol{A}$ 和矩阵 $\boldsymbol{D}$ 为固定值,只依赖于图 G 的拓扑结构,因此图卷积神经网络特征提取过程只需要更新学习参数矩阵 $\boldsymbol{W}_i$。对于具有 j 层的图卷积神经网络,首先在第一层使用图节点构成的图特征矩阵 $\boldsymbol{X}$ 与正则化对称邻接矩阵 $\hat{\boldsymbol{A}}$ 及第一层的参数矩阵 $\boldsymbol{W}_0$ 相乘,并通过非线性激活函数 ρ(例如 ReLU 函数)进行激活,得到第一层的输出转移特征矩阵 $\boldsymbol{L}_{(1)}$。$\boldsymbol{L}_{(1)}$ 将作为第二层图卷积神经网络的输入与第二层的参数矩阵 $\boldsymbol{W}_1$ 进行特征转移计算。通过不断的进行特征传递,第 j 层的图卷积将最终得到转移特征矩阵 $\boldsymbol{L}_{(j)}$。一层图卷积神经网络将使节点获得其直连邻接点的特征,而 j 层的图卷积神经网络将使节点获得其 j 步可达节点的特征。

3 BGCN 模型微博情感分析方法

本文提出的方法整个过程由三部分组成:数据处理、全局图构建、图卷积分类网络。算法流程如图 1 所示。

方法首先将下载的全量微博文本数据进行清洗处理,将处理后的全部文本数据构建全局异构图,异构图包含两类节点(词节点和短文本节点),边的权重使用全局的词共现特征来获取。然后再利用 BERT 模型获取的词向量表示来得到不同节点的初始特征矩阵。最后将微博全局异构图输入 GCN 网络[18],采用半监督的方法进行模型训练,提取短文本节点的高层情感语义特征,最后将提取的特征进行分类,得到情感倾向。下面详细介绍各个部分的具体实现方法。

3.1 数据集获取与处理

目前可用的中文短文本情感分析公开数据集多为中文商品评价数据集,但微博数据不同于商品评论数据,微博中文情感表达有其自身的特点,因此没有足够的训练样本用于微博情感分析模型训练,

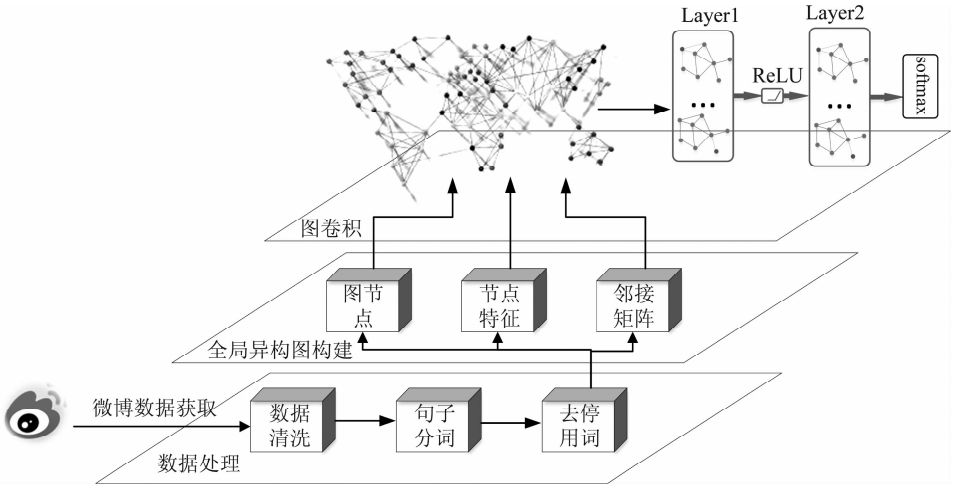

图 1　算法流程图

Fig. 1　Algorithm flowchart

商品评论数据训练的模型也不能做简单的模型迁移,需要针对微博数据制作大量的情感分析测试样本数据。为此本文通过大量的人工标注,制作高质量的可用于微博情感分析的测试数据集,将获取的微博文本采用人工标注的方式,对其中的 15000 条进行情感二分类标注。为了减少人为原因造成的样本标注偏差,将标记人员分为三组,每组分别独立对 15000 条数据进行情感标注,最后再对三组的标记结果进行综合评定,得到最后准确的文本标注结果。

15000 条标注数据作为半监督模型训练的标签数据,而异构图的构建则使用网络爬取的全部微博短文本数据集。在构建全局异构图前需要对微博数据进行预处理,首选需要对爬取的微博文本进行数据清洗,将微博文本中包含的广告、标签、代码和注释等文本情感无关信息进行去除。对于去除情感无关信息后的中文微博语句再进行分词处理。jieba 分词工具是一个基于统计词典的中文分词工具,具有分词准确、安装方便、功能强大的特点,本文使用 jieba 分词工具对中文微博进行分词处理。分词处理后的微博文本会存在大量停用词,停用词是指在中文语句中一类没有意义的词,比如“的”、“吧”等等。文本中如果存在大量的停用词,会对语句中有效信息造成噪音干扰,因此需要将其剔除或替换。同时针对微博文本中常存在表情、颜文字(符号表情)、网络用语等的特点,本文对常用停用词表进行了扩充,构建了一个包括网络用语和表情、颜文字(符号表情)的停用词表,在常用停用词表的基础上增加了新的无意义的网络热词,并在此基础上增添了对表情和颜文字的相应处理,依据表情和颜文字意思用对照表词语进行替换。

3.2　构建基于全局特征的异构图

为了使用图卷积神经网络挖掘微博文本的全局特征,将全量微博数据(包括 15000 条已标注微博数据和所有未标记数据)构建成一个大的全局文本异构图,如图 2 所示。异构图包含单词节点和短文本节点两类不同图节点,通过短文本节点和全局单词节点的关联特征传递,图卷积神经网络可以获得每个短文本节点的更深层次的语义表达。异构图的节点数为 $|V|=(|S|+|C|)$,其中 $|S|$ 为所有微博数据中的独立微博短文本数,$|C|$ 为所有语句进行清洗分词后的不重复词语数。由于异构图中包含两类不同性质的节点,因此异构图可以有三种不同性质的边:单词节点-短文本节点边、单词节点-单词节点边、短文本节点-短文本节点边。短文本节点与短文本节点之间定义为没有直连边,短文本节点间的关系通过单词节点的特征传递得到,异构图中只存在带权重的单词节点-短文本节点边和带权重的单词节点-单词节点边。某一微博短文本节点 s_i 和某一个单词节点 c_i 是否存在边,依赖于 c_i 是否在 s_j 中出现。如果 s_j 分词后含 i 个词,得到 s_j 的词集合为 $C_j=\{c_1,c_2,\cdots c_i\}$,$C_j\in C$,则 $c_1,c_2,\cdots c_i$ 节点分别和 s_j 存在一条带权重的无向边,假设 s_j=“今天是国庆节,祝祖国繁荣昌盛。”,分词含“今天”、“国庆节”、“祝”、“祖国”、“繁荣”和“昌盛”共 6 个词,

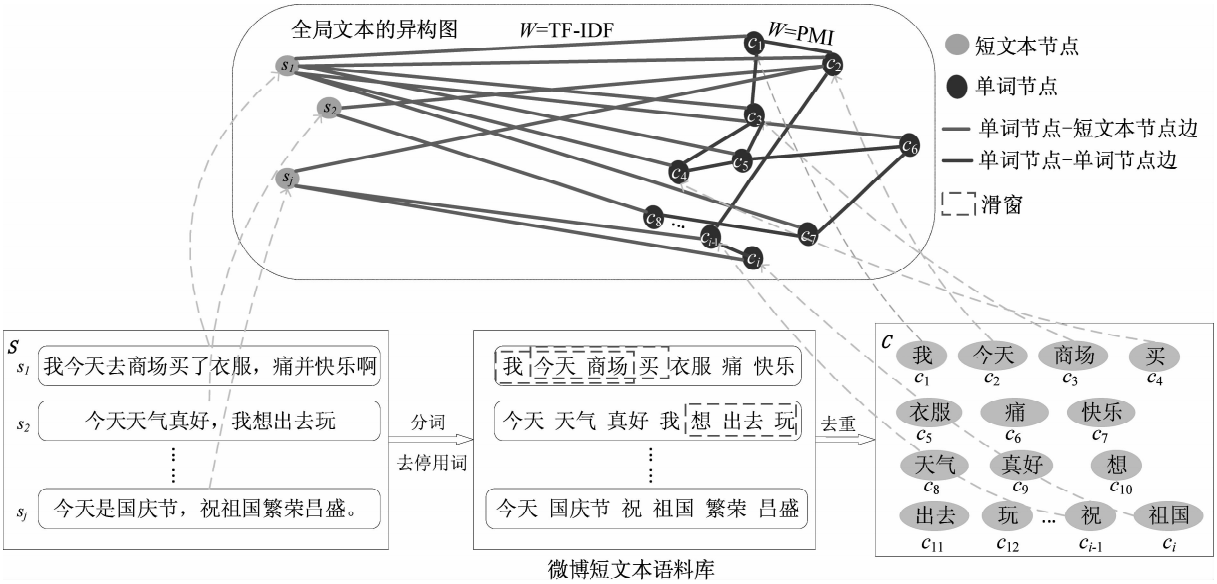

图 2 文本异构图构建示意图

Fig. 2 Schematic diagram of text heterogeneous graph construction

所以这里 $i=6$，得到 s_j 的词集合 $C_j=\{$“今天”，“国庆节”，“祝”，“祖国”，“繁荣”，“昌盛”$\}$。边的权重由 TF-IDF（Term Frequency-Inverse Document Frequency，词频-逆文档频率）算法计算得到。TF-IDF 算法是一种常用的文本关键词特征发现统计方法，包括词频（TF）和逆文档频率（IDF）两部分。TF-IDF 算法的目的是为了发现单词在文本中的重要程度，单词的重要程度随着其在文本中出现的次数成正比，而同其在语料库中出现的频率成反比，即同词频（TF）成正比而同逆文档频率（IDF）成反比。词频和逆文档频率的计算如公式（2）和（3）。

$$\mathrm{tf}_{ij}=\frac{n_{ij}}{\sum_k n_{kj}} \tag{2}$$

$$\mathrm{idf}_{ij}=\log\frac{|S|}{|\{j:c_i\in s_j\}|+1} \tag{3}$$

公式中 n_{ij} 表示单词 c_i 在短文本 s_j 中出现的次数，$\sum_k n_{kj}$ 是短文本 s_j 中所有单词出现次数的总和，$|S|$ 表示语料库中短文本的总数，$|\{j:c_i\in s_j\}|$ 表示所有包含单词 c_i 的短文本数量总和，为了避免分母为零值，将分母进行加 1 处理。

单词节点与单词节点是否存在边及其边权重的大小由两点间的相关性确定。步长设置为 1，即滑窗每次移动一个词的位置，统计两个词在同一个滑窗出现的次数，用于计算两个词的 PMI 值（Pointwise Mutual Information，点互信息）。PMI 值量化表示了两个词的相关程度，两个单词 c_i，c_j 的 PMI 值计算如公式（4）所示：

$$\mathrm{PMI}\ (c_i,c_j)=\log\frac{p(c_i,c_j)}{p(c_i)p(c_j)} \tag{4}$$

$p(c_i,c_j)$ 是单词 c_i，c_j 联合出现的概率，$p(c_i,c_j)$ 等于单词 c_i 和 c_j 在同一滑窗出现的总次数除以语料库中的滑窗总数。$p(c_i)$ 和 $p(c_j)$ 是单词 c_i，c_j 边缘概率，等于单词出现在滑窗的次数除以语料库中的滑窗总数。若 c_i 和 c_j 的 PMI 值大于设定的阈值，则在两点之间建立权重边且以该 PMI 值作为边的权重。

窗口大小 m 依据微博语句表达的统计特征设定，对微博语料库中的每条微博长度进行频次统计，如图 3 所示。

对不同的窗口大小进行了测试实验，分别选取出现频率最高的文本长度、平均文本长度、频率最高文本长度的 1/2 等滑窗大小进行实验，实验表明选取频数出现最高的文本长度 6 的 1/2（即 $m=3$）作为窗口大小，可以达到最好的实验结果。

3.3 图卷积文本分类

图卷积文本分类网络由输入层，特征提取层，分类层组成。输入层的输入包括邻接矩阵 $\boldsymbol{A}$、度矩阵 $\boldsymbol{D}$ 以及节点特征 $\boldsymbol{X}$。邻接矩阵 $\boldsymbol{A}$、度矩阵 $\boldsymbol{D}$ 通过构建的全局文本异构图得到。节点特征 $\boldsymbol{X}$ 本文通

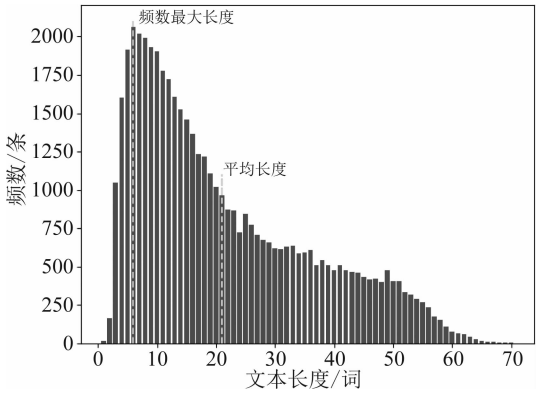

图 3　文本长度频次图

Fig. 3　Statistical graph of text length

过 BERT 自然语言模型产生。词向量可以通过对自然语言的建模将词组映射到实数多维向量，将多维向量作为后续自然语言处理任务的底层词嵌入，输入到高层任务，可以提升任务处理的性能。传统的 Word2vec 词向量模型由于模型简单，特别是对于中文中的一词多义不能有很好的表示，因此在一些任务上表现不佳。而一词多义现象在微博短文本中显得尤为突出。BERT 模型采用双向注意力机制，可以丰富向量词嵌入的表示，有效改善一词多义的表达效果。使用 BERT 模型分别训练得到词和短文本的向量表示，并对向量进行全局池化，产生节点特征矩阵 $\boldsymbol{X}$，作为图卷积的输入。

邻接矩阵 $\boldsymbol{A}$、度矩阵 $\boldsymbol{D}$ 以及节点特征 $\boldsymbol{X}$ 输入到特征提取层的图卷积网络，进行特征抽取，产生短文本的特征嵌入，最后将产生的特征嵌入送入到文本分类层，分类层使用 softmax 函数进行文本情感分类预测。大量实验[17-18]表明图卷积在使用少数层时就可以达到很好的效果，本文的图卷积也采用两层的网络机构。模型训练采用交叉熵作为损失函数。训练过程采用半监督的方式，损失函数只计算 15000 条已标注微博数据的损失函数值，并对整体的模型权重系数进行更新调整，特征传递则是在整个全局异构图内进行，因此使用全局异构图对短文本的情感进行分类，特征嵌入加入了无标签数据的全局特征，可以提升情感分类的准确率。

4　实验分析

4.1　实验数据集

实验共用了三组数据集，实验数据文本样例如表 1 所示。

（1）自建微博短文本数据集：网络爬取的真实微博短文本数据集 55000 条，将数据集分为两个部分：一部分是用于半监督中无标签数据的微博文本 40000 条，另一部分是人工标注的有标签数据 15000 条。

（2）COAE2014 数据集：COAE2014 数据集是国内权威的中文倾向性分析评测数据集，由中国中文信息学会信息检索专业委员会提供，致力于探索中文倾向性分析的新技术、新方法，任务涉及主客观分析、情感极性分析、评价对象抽取以及搭配抽取等方面。本文选用了数据集中的任务 4—微博观点识别（在给定的微博集合中，判别每个句子的情感倾向性）的数据集。该数据集共 4 万条微博，其中 7000 条已知情感标签。

（3）NLP&CC2014 数据集：数据来源于 2014 年 Natural Language Processing & Chinese Computing 会议任务 2—基于深度学习的情感分类，包含正负训练样本各 5000 条，测试样本 2500 条。

表 1　数据样本实例

Tab. 1　Data samples

数据集	标签	数据样例
自建数据集	负向情感	航班频频延误，乘客发飙拦截飞机，你怎么看？被延误的心情可想而知，太闹心了
	正向情感	每一个机组的平安到达、每一次航班的安全起飞，都离不开你们辛勤工作的汗水
COAE 2014 数据集	负向情感	XX 手机怎么那么破啊，充满电只接电话打电话都用不了一天，而且没电自动关机充电还开不了机。
	正向情感	如有理财观念，现在还有机会购买这种理财产品；中国平安推出的《金裕人生》是不错的选择哦！
NLP&CC 2014 数据集	负向情感	可悲啊，怎么就买不到正版的呢?！强烈建议卓越 严格控制货源，从源头杜绝 盗版！
	正向情感	挺不错。老狼还是没有什么变化，听他的歌永远能让人想起自己的纯真年代

对数据集进行数据清洗、句子分词、去除停用词和表情和颜文字词语替换等数据预处理操作后，得到的处理结果如表 2 所示。预处理后的数据作为最终模型训练的数据集。

4.2　模型参数

本文的实验环境使用 Intel CPU 1.70GHz，8GB

内存和 windows10 系统,使用 Python36 和 Tensorflow 深度学习开源框架实现。主要参数如表 3 所示。

表 2 预处理结果对比

Tab. 2 Comparison of preprocessing results

原数据	预处理结果
一直跟个脑抽的人住一起,我不是被气爆😷就是被逼疯	脑抽 人 住在一起 不是 气爆 **气死** 逼疯
周二开始工作啦! 可怕的拖延症认真对待每件你该去完成的工作!!!! 💪💪	周二 工作 可怕 拖延症 认真对待 每件 去 完成 工作 **加油 加油**
午休啊(ノ ^)ノ ┻┻中午单位的电话一概不接	午休 **掀 桌子** 中午 单位 电话 一概 不接
早安(^O^)! 亲们(●✿∀✿●),新的一天又开始了!	早安 **开心** 亲们 **可爱** 新 一天 开始

表 3 模型参数设置

Tab. 3 The parameter settings of model

参数	值
Epoch(迭代轮数)	32
Optimizer(优化器)	Adam
Learning_rate(初始学习率)	0.02
Dropout_rate	0.5
Hidden_unit(隐藏层单元数)	64
权重初始化	Random Initialization

4.3 评估标准

本文分别采用精确率(Precision)、召回率(Recall)、F1 值作为模型的评价指标,其计算公式具体如下:

$$\text{Presion}=\frac{\text{TP}}{\text{TP}+\text{FP}}\times 100\% \tag{5}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}\times 100\% \tag{6}$$

$$\text{F1}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}\times 100\% \tag{7}$$

其中,TP 表示正样本中预测正确的样本数,FN 表示正样本中预测错误的样本数,TN 表示负样本中预测正确的样本数,FP 表示负样本中预测错误的样本数。通过 F1 值的高低来评估模型的性能是否优于其他模型。

4.4 结果分析

为了验证本文半监督算法的有效性,对标记数据在整个训练数据集中的不同占比进行分别实验,采用未标记数据与标记数据比例分布分别为 1/2、1/1、2/1、3/1、4/1 的数据集进行训练,实验结果如图 4 所示。

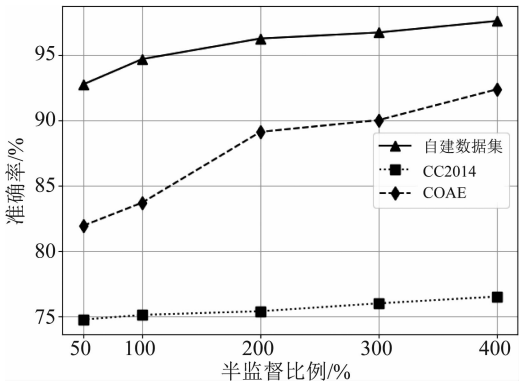

图 4 不同比例标记数据的结果

Fig. 4 Results of labeled data at different scales

由图 4 可以看出随着无标记数据的增加,模型在三个数据集上的分类准确率都不断提高,说明了本文的半监督算法可以从无标签数据中提取有用的全局特征,仅利用少量的有标签数据就可以对微博情感进行较准确的分类。

除了验证本文半监督算法的有效性外,还将本文方法模型与以下深度学习模型进行了性能比较:

1)CNN。采用文献[19]中的传统卷积神经网络模型,卷积层中使用卷积核为 5 的 CNN 提取文本特征。

2)LSTM。采用文献[20]中的模型,将词嵌入作为模型输入,学习语义特征。

3)Bi-LSTM[21]。与 LSTM 一样输入词嵌入信息,学习结合上下文的语义特征。

4)BM-ATT-BiLSTM[21]。目前公开文献能查到的针对 COAE2014 数据集和 NLP&CC2014 数据集上的最好结果。

算法模型同时在 3 组数据集上进行了对比试验,结果如表 4 所示。

从表 4 的实验结果可以看出,本文提出的结合 BERT 和图卷积的 BGCN 在不同的数据集上均取得了最好的情感分类效果,其中自建数据集在 F1 分数上取得了 97.54% 的好结果。对比实验选择的 CNN、LSTM、Bi-LSTM 模型,本文算法在自建数据集上分别有了 9.03、8.09 和 3.39 个百分点的精确率提升,F1 分数也比 Bi-LSTM 模型提高了 4.1 个百分点。在 COAE2014 数据集和 NLP&CC2014 数据集

上比 BM-ATT-BiLSTM 模型也有了一定的提升，说明 BGCN 模型能更好地捕捉微博文本的语义特征和情感信息，证明了无标记数据语义特征在情感分类任务中的有效性。

表 4　不同模型分类结果对比

Tab. 4　Comparison of classification results of different models

模型	指标	NLP&CC 2014	COAE 2014	自采集数据
CNN	P	70.40	87.56	88.53
	R	66.31	87.25	82.39
	F	68.29	87.27	82.41
LSTM	P	72.23	92.67	89.47
	R	72.56	91.52	88.10
	F	72.39	92.09	88.67
Bi-LSTM	P	73.92	94.00	94.17
	R	71.17	95.82	92.89
	F	72.52	94.91	93.44
BM-ATT-BiLSTM	P	**75.57**	95.74	–
	R	70.69	95.34	–
	F	73.05	95.54	–
BGCN（本文）	P	73.21	**95.76**	**97.56**
	R	**75.20**	**95.69**	**97.52**
	F	**74.19**	**95.71**	**97.54**

5　结论

在微博情感分析任务中，本文在分析现有基于深度学习的情感分类技术的基础上给出了一种针对微博短文本的全局文本异构图构建方法，该方法利用所有的微博数据构建全局异构文本图，全局异构文本图分为单词节点和短文本节点，点与点之间是带权重的边，反应了不同节点的相关性。基于构建的全局异构文本图提出了一个结合 BERT 词向量特征和图卷积的半监督微博文本情感分类方法，通过 BERT 模型进行语义建模，最后通过图卷积网络完成半监督的文本分类。该方法充分利用了语料库的全局特征，量化词语之间的紧密程度，分析词语对句子情感的影响，从而有效地提取文本的全局高层特征。实验结果验证了该算法在中文短文本情感分类任务上的有效性。下一步工作将重点研究如何融合更多的特征进行情感信息丰富、模型的跨图学习以及该算法在多分类情感分析下的表现。

参考文献

[1] YUE Lin, CHEN Weitong, LI Xue, et al. A survey of sentiment analysis in social media[J]. Knowledge and Information Systems, 2019, 60(2): 617-663.

[2] 李卫疆, 伊靖. 基于扩展特征矩阵和双层卷积神经网络的微博文本情感分类[J]. 计算机应用与软件, 2019, 36(12): 150-155.
LI Weijiang, YI Jing. Weibo text sentiment classification based on extended feature matrix and double-layer convolution neural network[J]. Computer Applications and Software, 2019, 36(12): 150-155. (in Chinese)

[3] YANG Min, JIANG Qingnan, SHEN Ying, et al. Hierarchical human-like strategy for aspect-level sentiment classification with sentiment linguistic knowledge and reinforcement learning[J]. Neural Networks, 2019, 117: 240-248.

[4] 张仰森, 郑佳, 黄改娟, 等. 基于双重注意力模型的微博情感分析方法[J]. 清华大学学报(自然科学版), 2018, 58(2): 122-130.
ZHANG Yangsen, ZHENG Jia, HUANG Gaijuan, et al. Microblog sentiment analysis method based on a double attention model[J]. Journal of Tsinghua University (Science and Technology), 2018, 58(2): 122-130. (in Chinese)

[5] MAIPRADIT R, HATA H, MATSUMOTO K. Sentiment classification using N-gram inverse document frequency and automated machine learning[J]. IEEE Software, 2019, 36(5): 65-70.

[6] 陈培新, 郭武. 融合潜在主题信息和卷积语义特征的文本主题分类[J]. 信号处理, 2017, 33(8): 1090-1096.
CHEN Peixin, GUO Wu. Document topic categorization combining latent topic information and convolutional semantic features[J]. Journal of Signal Processing, 2017, 33(8): 1090-1096. (in Chinese)

[7] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long

Papers). Baltimore, Maryland. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 655-665.

[8] TAI Kaisheng, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015: 1556-1566.

[9] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. Computer Science, 2014.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, 2017: 5998-6008.

[11] MIKOLOV T, CHEN KAI, CORRADO G, et al. Efficient estimation of word representations in vector space [J]. Computer Science, 2013.

[12] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018.

[13] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019.

[14] LI Yanran, LI Wenjie, SUN Fei, et al. Component-enhanced Chinese character embeddings[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015: 829-834.

[15] CHEN Xinxiong, XU Lei, LIU Zhiyuan, et al. Joint learning of character and word embeddings[C]//International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina: AAAI Press, 2015: 1236-1242

[16] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2009, 20(1): 61-80.

[17] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]//International Conference on Learning Representations. Toulon: [s.n.], 2017.

[18] YAO Liang, MAO Chengsheng, LUO Yuan. Graph convolutional networks for text classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 7370-7377.

[19] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 1746-1751.

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

[21] 廖经真. 基于深度学习的短文本情感分析[D]. 南昌: 江西财经大学, 2020.
LIAO Jingzhen. Short text sentiment analysis based on deep learning[D]. Nanchang: Jiangxi University of Finance and Economics, 2020. (in Chinese)

作者简介

方 澄 男,1980年生,天津人。中国民航大学讲师,主要研究方向为数据挖掘、大数据、计算机视觉等。
E-mail: cfang@cauc.edu.cn

李 贝 女,1993年生,四川绵阳人。中国民航大学硕士研究生,主要研究方向为自然语言处理、情感分析等。
E-mail: libei053146@163.com

韩 萍 女,1966年生,天津人。中国民航大学教授,主要研究方向为图像处理、模式识别等。
E-mail: hanpingcauc@163.com