DOI:10.11918/j.issn.0367-6234.201809214

基于用户关联度的半监督情感分析模型

金志刚,杨 洋

(天津大学 电气自动化与信息工程学院,天津 300072)

摘 要: 随着信息技术与社交媒体的不断发展,用户情感分析在舆情监控、信息预测、产品评价上发挥着越来越重要的作用.然而,监督学习手工标签获取困难,无监督学习缺少标签的引导,因此本文基于社会学理论建立了半监督的情感分析模型,该模型主要分为标签添加和情感分析两部分.标签添加部分首先基于情感一致性和情感传染性两种被认可的社会学理论建立UR-S模型,然后通过用户关联度和文本相似度进行改进,建立TRS-SAT模型,增加标签数量.情感分析部分将TRS-SAT模型与卷积神经网络结合,通过卷积神经网络挖掘特征集合与情感分析标签之间的深层次联系,构建半监督学习模型改善情感分析性能.实验表明,本文提出的基于用户关联度和深度学习的半监督情感分析模型,与半监督的支持向量机模型相比,准确率、召回率、F值分别提升11.40%、5.90%、8.65%;与卷积神经网络模型相比,分别提升4.12%、4.17%、4.14%,均有较好的表现.由此证明,该模型能够为舆情分析与用户决策提供良好的理论基础,具有创新性和实用性.

关键词: 用户关联度;半监督学习;深度学习;卷积神经网络;情感分析;文本相似度

中图分类号: TP391 **文献标志码:** A **文章编号:** 0367-6234(2019)05-0050-07

A semi-supervised short text sentiment analysis model based on social relationship strength

JIN Zhigang, YANG Yang

(School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: With the development of information technology and social media, user sentiment analysis tends to play an increasingly important role in public opinion monitoring, information prediction and product evaluation. However, collecting sufficient manual sentiment labels in supervised learning is still difficult and costly, and unsupervised learning is lack of label guidance. Therefore, a semi-supervised sentiment analysis model based on sociological theory is established in this paper, which is mainly divided into two parts: label addition and emotion analysis. First, a UR-S (User Relationship using Social relations) model was built, which was inspired by sentiment consistency and emotional contagion. Then a TRS-SAT (Text Relationship Strength using Social relations, user Attribute and Text similarities) model based on UR-S model and add labels was established. Finally, the TRS-SAT model and CNN (convolutional neural network) were combined to construct SA-SRS-CNN (Sentiment Analysis using Social Relationship Strength and Convolutional Neural Network) model. The model uses CNN to mine the deep connection between the feature set and the emotional labels to improve the emotional performance. Experiments show that the accuracy, recall, and the F value of the proposed model increased by 11.40%, 5.90% and 8.65%, respectively compared with SVM, and increased 4.12%, 4.17%, and 4.14%, respectively compared with CNN, which suggests that the model is innovative and practical and can provide a good theoretical basis for public opinion analysis.

Keywords: social relationship strength; semi-supervised learning; deep learning; convolutional neural network; sentiment analysis; text similarity.

在信息时代的大潮下,微博、twitter等社交媒体蓬勃发展,情感分析已经成为自然语言处理的重要方面,在电商平台的评论信息分析、社交媒体平台用户的评论导向等方面有重要的现实价值.

情感分析的传统方法是基于词典和机器学习的方法[1-2].TURNEY等[3]使用无监督学习模型通过分类的词典将文本分类,CHINSHA等[4]使用基于规则、依存关系和评价词典的无监督模型,虽然不需训练数据,但领域依赖性强.BHUSHAN等[5]建立基于文本间相似性的机器学习模型.ABDI等[6]对使用最广的几种特征选择技术和机器学习分类器在情感分析中的表现进行了性能研究.然而,传统的词典与

收稿日期: 2018-09-30
基金项目: 国家自然科学基金项目(71502125)
作者简介: 金志刚(1972—),男,教授,博士生导师
通信作者: 金志刚,zgjin@tju.edu.cn

机器学习的方法无法解决社交短文本自身词汇稀疏性、语法随意性与热词性导致的问题，深度学习逐渐成为主流. HUSSAIN 等[7]基于深度信念网络通过深度学习算法进行文本情感分析. KONATE 等[8]证明了单层卷积神经网络(convolutional neural network，CNN)的深度学习模型相比于机器学习表现更好. 金志刚等[9]基于卷积神经网络结合表情符号建立情感分析模型. WU 等[10]提出了结合规则和深度学习的混合无监督方法. 另一方面，监督学习手工标签获取困难需要专家制定规则，进行人工标注，无监督学习又缺少标签引导，效果往往不能令人满意，因此半监督学习是一种有效的解决方式. KIM[11]提出了改进的半监督维数约简框模型，保留特征提取优点解决情感分析缺点. WANG 等[12]提出将 K-means 算法融合进 CNN，实现半监督学习的文本情感分析.

同时，网民们在社交平台上通过短文本发表观点分享生活，相似的爱好与观念将网友们聚集成个性化的社交网络. 有研究表明，社交理论对社交网络短文本情感分析有着一定的指导作用，可提高预测效果. 抽取社交网络特征进行文本情感分析[13]已取得了较好效果. HU 等[14]基于线性回归、社会关系提出了 MSA 图正则化模型. WANG 等[15]定义用户到用户主题包含度并构建其稀疏网络. XIAO 等[16]量化共同邻居的依赖关系，分析结构空间中的用户相关性. XIA 等[17]充分利用词语关系，使用基于主题图的模型实现多领域应用. 卢桃坚等[18]利用微博-微博关系构建基于图的半监督分类器，连接标记和未标记数据. SHI 等[19]给出了基于 CNN 的多特征情感分析模型，肖云鹏等[20]分析社交网络中用户属性和关系数据，发现了用户关系建立的关键因素. 徐志明等[21]定义了用户关系强度，并给出了基于各种用户属性信息的计算方法. WEI 等[22]研究了用户在社交网络情感传播中的作用. 白朔天等[23]通过多任务回归网络挖掘方法，分析社交媒体用户人格和网络行为的关联模式.

目前在社交媒体短文本情感分析方面，使用半监督算法的研究者鲜少利用社交网络进行研究；而少数通过社交关系进行情感分析的研究又大都未采用半监督算法且对用户关系的衡量不够细致. 因此，本文一方面充分利用社交网络，通过用户关联度建立文本关系模型，使有标签和无标签的文本建立联系形成聚类，通过有标签的文本标注一部分无标签文本，从而增加标签数量，另一方面通过 word2vec 训练大规模互联网语料库，学习词组的高维向量表示，作为中文词汇高效的数学表示模型，有利于加速后续深度学习模型训练的收敛，结合卷积神经网络建立半监督的深度学习模型，为舆情监控、信息预测提供基础.

1　基于用户关联度的半监督情感分析模型

1.1　模型框架

本文建立基于用户关联度的半监督情感分析模型(sentiment analysis using social relationship strength and convolutional neural network，SA-SRS-CNN)，主要分为标签添加和情感分析两部分. 标签添加部分首先通过用户间的关注关系，基于社交理论构建 0-1 分布的用户关联度模型(user relationship using social relations，UR-S)，然后通过用户背景属性和文本相似度改进 UR-S 模型，构建文本-文本关系模型(text relationship strength using social relations，user attribute and text similarities，TRS-SAT)，实现有标注和无标注数据的关联，通过有标签的文本标注一部分无标签文本，从而增加标签数量. 情感分析部分，通过 word2vec 训练大规模互联网语料库，学习词组的分布式高维向量表示，作为中文词汇高效的数学表示模型，有利于加速后续深度学习模型训练的收敛. 结合 CNN 构建基于用户关联度和深度学习的半监督情感分析模型，实现短文本情感分析. 其流程图见图 1.

该模型的特点在于，结合用户关系、用户背景属性、文本相似度与卷积神经网络，将监督学习改为半监督学习. 其实现流程如下：

1) 预处理原始微博文本并计算文本相似度矩阵；

2) 根据用户背景属性和用户关注关系计算用户关联度，并根据用户关联度和文本相似度建立 TRS-SAT 模型，计算文本-文本关系；

3) 基于文本-文本关系，实现有标注和无标注数据的关联，通过有标签的文本标注一部分无标签文本，从而增加标签数量，将所有的有标签文本作为卷积神经网络的输入语料集；

4) 使用 word2vec 工具训练大规模互联网语料库，学习词组的高维向量表示；

5) 通过 word2vec 计算微博文本的词向量表示，若微博文本中的词组在 4)中存在，直接使用其结果，否则，通过 word2vec 随机初始化；

6) 通过卷积与池化运算，捕获并筛选局部特征，训练微博文本情感分类器，实现情感分析.

提出模型的两个主要构成，即 TRS-SAT 模型和 CNN 模型.

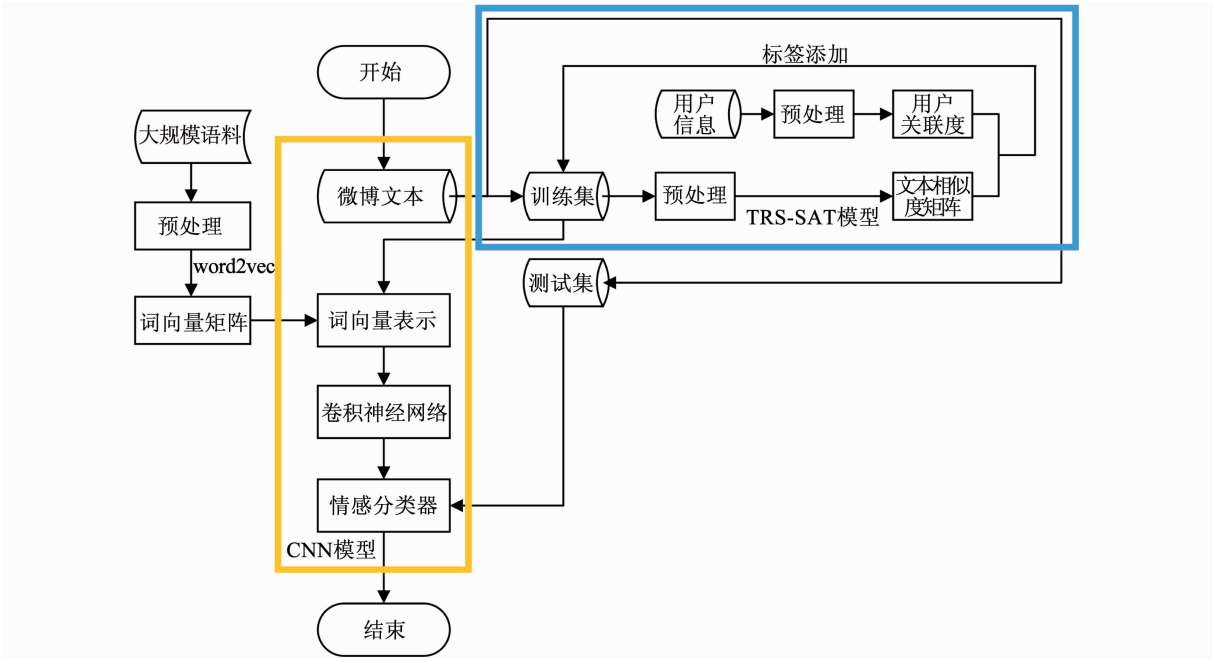


图 1 SA-SRS-CNN 模型流程图

Fig. 1 Flow chart of SA-SRS-CNN model

1.2 改进用户关联度的文本关系模型

提出的改进用户关联度的文本关系模型(text relationship strength using social relations, user attribute and text similarities, TRS-SAT)是由基于社交理论的 UR-S 模型拓展而来,在其基础上,引入用户属性和文本相似度将 0-1 分布的用户关联度转化为文本-文本的关系,构建 TRS-SAT 模型. 该模型为 SA-SRS-CNN 模型的第一部分,用于增加标签.

1.2.1 用户关联度模型

用户关联度模型即 UR-S 模型通过关注关系,基于情感一致性、情感感染性两种被社会学家所认同的社交理论,构建 0 - 1 分布的用户关系. 情感一致性,指相同作者发表的微博比随机采样的微博更有可能在情感极性上保持一致;情感感染性,意指好友间发表的微博更有可能在情感极性上保持一致. 定义用户基于情感一致性、情感感染性产生的关系强度为用户关联度.

用户-文本矩阵 $\boldsymbol{U}\in\boldsymbol{R}^{d*n}$,$d$ 是语料库中用户的数量,n 为文本的个数. 若用户 u_i 发表微博 t_j,则 $\boldsymbol{U}_{ij}=1$,否则$\boldsymbol{U}_{ij}=0$,见图 2 (a). 用户-用户关系矩阵 $\boldsymbol{F}\in\boldsymbol{R}^{d*d}$ 中,若用户 u_i 关注用户 u_j,则$\boldsymbol{F}_{ij}=1$,否则 $\boldsymbol{F}_{ij}=0$,见图 2 (b).

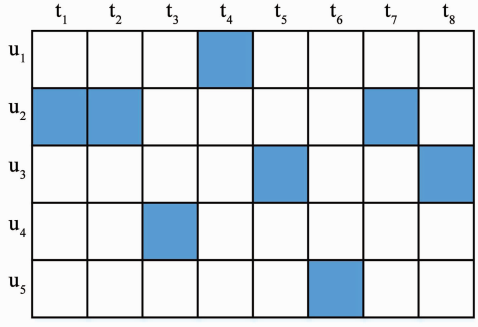
则依据情感感染性和情感一致性有:

$$\boldsymbol{A}_{\mathrm{sc}}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}, \tag{1}$$

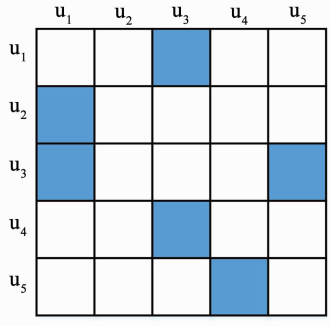
$$\boldsymbol{A}_{\mathrm{ec}}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{U}. \tag{2}$$

式中:$\boldsymbol{A}_{\mathrm{sc}}$为情感一致性关系矩阵,$\boldsymbol{A}_{\mathrm{ec}}$为情感感染性关系矩阵,$\boldsymbol{U}$ 为用户-文本矩阵,F 为用户-用户关系矩阵. 在对称矩阵$\boldsymbol{A}_{\mathrm{sc}}$中,元素$\boldsymbol{A}_{\mathrm{sc}_{ij}}=1$ 表明,微博 t_i 和 t_j 是同一用户发表,根据情感一致性,这两条对应的微博,更有可能表达相似情感. 在非对称矩阵$\boldsymbol{A}_{\mathrm{ec}}$中,元素$\boldsymbol{A}_{\mathrm{ec}_{ij}}=1$ 表明,微博 t_i 和微博 t_j 的作者是朋友,这两条微博表达的情感有更强的相似性. $\boldsymbol{A}_{\mathrm{s}}$ 结合$\boldsymbol{A}_{\mathrm{sc}}$和$\boldsymbol{A}_{\mathrm{ec}}$给出用户关联度 $\boldsymbol{A}_{\mathrm{s}}$ 的表达式为

$$\boldsymbol{A}_{\mathrm{s}}=\boldsymbol{A}_{\mathrm{sc}}+\boldsymbol{A}_{\mathrm{ec}}. \tag{3}$$



(a) 矩阵$\boldsymbol{U}$



(b) 矩阵$\boldsymbol{F}$

图 2 矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{F}$

Fig. 2 Matrices $\boldsymbol{U}$ and $\boldsymbol{F}$

1.2.2　改进用户关联度的文本关系模型

虽然 UR-S 模型给出了用户关联度，但是考虑到用户对每个好友的信任强度不同（如用户 1 同时受到用户 3 和用户 2 的影响，但是用户 3 和用户 2 的倾向不同），简单的二元关系已不能满足对社交网络研究的需要. 在 UR-S 模型的基础上引入用户属性和文本相似度，构建基于用户关联度的文本－文本关系模型：TRS-SAT 模型.

在 UR-S 模型的基础上，结合本实验获得的数据将用户信息进行度量，度量标准见表 1. 构建用户属性矩阵 $\boldsymbol{S}$. 其中 S_1、S_2、S_3 分别表示位置信息、教育信息与性别信息. 并改进 A_{ec} 定义式为

$$\boldsymbol{A}_{ec} = \boldsymbol{U}^{T}\boldsymbol{S}\circ\boldsymbol{F}\boldsymbol{U}. \tag{4}$$

式中：$\boldsymbol{S}$ 为用户属性矩阵，符号° 表示逐元素的 Hadamard 乘积.

表 1　用户相似性度量及其方法

Tab. 1　Measure and the method of user similarity

符号	取值
S_1	省、市一致，$S_1=1$；只有省一致，$S_1=\frac{2}{3}$；其他 $S_1=0$
S_2	教育等级、学校一致，$S_2=1$；只有教育等级一致，$S_2=\frac{2}{3}$；其他 $S_2=0$
S_3	性别一致，$S_3=1$；不一致，$S_3=0$

另一方面，通过文本相似度将 UR-S 模型中 0－1分布的用户关系转化为非二元的文本关系. 由于短文本具有不规范性和随意性，在词项文档矩阵分词前首先进行预处理实现降维. 利用向量空间模型与 TF-IDF 算法构建文本内容特征矩阵，通过余弦相似度计算文本相似度矩阵 $\boldsymbol{M}$，得到任意两条语料的相似度.

在 TF-IDF 算法中，矩阵中每个元素的值代表相应列文本对应行上的单词元素的权重，定义：

$$T_{tf} = \frac{a}{b}, \tag{5}$$

$$I_{idf} = \lg\frac{N}{d}, \tag{6}$$

$$T_{tf-idf} = T_{tf}\cdot I_{idf}. \tag{7}$$

式中：a 为该文本中该词项出现数量，b 为本文件词项总数量，N 为语料库中语料的总数量，d 为含有所要计算单词的语料数. 余弦相似度是计算两个向量的夹角余弦值也是判断向量相似度的重要方法，公式为

$$\cos\theta = \frac{\sum_{1}^{n}\boldsymbol{A}_i\boldsymbol{B}_i}{\sqrt{\sum_{1}^{n}\boldsymbol{A}_i^2}\sqrt{\sum_{1}^{n}\boldsymbol{B}_i^2}}. \tag{8}$$

式中：向量 $\boldsymbol{A}=(A_1,A_2,\cdots,A_n)$，向量 $\boldsymbol{B}=(B_1,B_2,\cdots,B_n)$. 在此基础上构建 TRS-SAT 模型：

$$\boldsymbol{A}_{st} = (\boldsymbol{A}_{sc}+\boldsymbol{A}_{ec})\circ\boldsymbol{M}, \tag{9}$$

$$\boldsymbol{A}_{st} = (\boldsymbol{U}^{T}U+\boldsymbol{U}^{T}\boldsymbol{F}*\boldsymbol{S}\boldsymbol{U})\circ\boldsymbol{M}. \tag{10}$$

式中 $\boldsymbol{A}_{st}$ 为文本与文本之间的关联强度，用来衡量文本与文本关系. 通过与 $\boldsymbol{M}$ 的 Hadamard 乘积将用户间关系转化为文本间关系.

1.3　基于卷积神经网络的情感分析模型

SA-SRS-CNN 模型的第二部分是结合 word2vec 的基于卷积神经网络的情感分析模型，流程图见图 3.

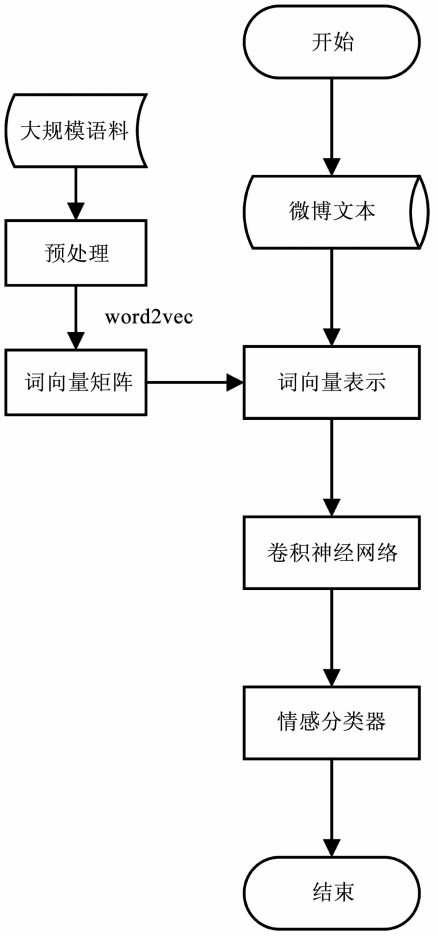


图 3　结合 word2vec 的 CNN 模型流程图

Fig. 3　Flow chart of CNN model combined with word2vec

word2vec 是从大量文本中以无监督学习的方式学习语义知识的模型，其本质就是将单词从原先所属的空间映射到新的多维空间中. 通过学习文本，用词向量的方式表征词的语义信息；通过嵌入空间，将语义上相似的单词映射到距离相近的地方. 本文在 Word2vec 中采用 Skip-gram 模型，计算输入 word 的 input vector 与目标 word 的 output vector 之间的余弦相似度，并进行 softmax 归一化.

word2vec 首先训练大规模互联网语料，学习词组的高维向量表示. 再通过 word2vec 计算微博文本的词向量表示，若微博文本中的词组在大规模互联

网语料中存在，直接使用其结果，否则，通过 word2vec 随机初始化. 用词向量表示预处理后的文本作为 CNN 的输入，构建情感分类器. 本文借鉴 Kim 等[24]使用的 CNN 模型，该模型虽然不是第一次提出将卷积神经网络用于文本分类，但是给出了多个变体和调参过程，是卷积神经网络用于文本分类的开山之作，该模型结构图见图 4.

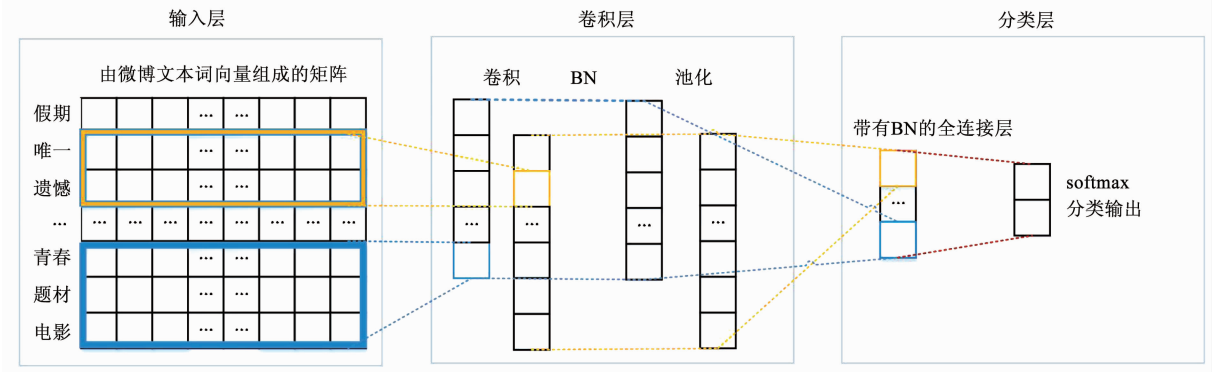


图 4 卷积神经网络结构

Fig. 4 Structure of CNN model

该模型分为输入层、卷积层和分类层三部分.

输入层是 $r \cdot u$ 维度的文本词向量矩阵，r 为每条文本的特征词组数，u 由 word2vec 决定. 卷积层首先通过长度为 h 的卷积核 w 卷积词向量矩阵. 然后通过 BN (Batch Normalization)算法进行归一化提升训练速度，再通过最大值池化进行降维，并将特征数量一致化. 分类层通过 BN 算法防止数据分布改变，并通过 softmax 层计算分类概率. 卷积过程为

$$t_i = f(w * s_{i:i+h-1} + b). \tag{11}$$

式中：$s_{i:i+h-1}$ 为由第 i 个词组至第 $i+h-1$ 个词组组成的连续文本片段，$*$ 为卷积运算，b 为偏置项，f 为非线性激活函数. 分类概率计算公式为

$$P_j = P(y = j \mid X, b) = \frac{e^{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_j + b_j}}{\sum_{i=1}^{L} e^{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_i + b_i}}. \tag{12}$$

式中：P_j 为该文本属于第 j 类的概率，$\boldsymbol{X}$、$\boldsymbol{W}$、b_i、b_j 分别为分类层的输入、权值矩阵、偏置项的第 i 元素和偏置项的第 j 元素，L 为类别数量.

2 实验与分析

2.1 数据集和评价指标

实验数据集分为两部分：word2vec 训练语料和微博数据集.

使用搜狗实验室整理的新闻数据集作为 word2vec 训练语料，包含 2 706 229 条新闻语料和 565 345 个词组. 微博数据集中出现在词向量集合中的词组使用 word2vec 的计算结果，否则随机初始化.

微博数据集：由于并未发现同时具有用户背景信息、用户–用户关系、用户–文本关系的公开数据集，自行采集约 13 000 条微博文本，并对其中 9 000 条微博文本进行积极和消极的标注，其中积极类微博 3 467 条，消极类微博 5 533 条. 通过本文模型的标签添加部分后，获得有标签的微博文本 9 873 条. 微博文本示例见表 2.

表 2 微博文本示例

Tab. 2 Examples of Weibo text

积极文本	嘻嘻谢谢我的大熊带我来看后来的我们，爱你爱你.	后来的我们，成绩不错小小的鼓励一下林更新.	今天去看后来的我们演的真好！
消极文本	看了后来的我们，联系不到他，手机没电胃病又犯了，很难受.	为什么相爱的人却总是不能在一起.	后来的我们看到后面哭得根本停不下来，最终输给了现实.

本文采用的评价指标是准确率(A_{accuracy})、召回率(R_{recall})和 F 值：对于给定的测试数据集，TP 为将积极文本分类为积极的数量，TN 为将消极文本分类为消极的数量，FP 为将积极文本分类为消极文本的数量，FN 为将消极文本分类为积极文本的数量. 其计算公式为

$$A_{\text{accuracy}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{13}$$

$$R_{\text{recall}} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{14}$$

$$F = \frac{2A_{\text{accuracy}}R_{\text{recall}}}{A_{\text{accuracy}} + R_{\text{recall}}}. \tag{15}$$

2.2 实验设计

对比实验共 4 组. 为验证文本相似度和社交关系对情感分析的促进效果，对文献[24]中 Kim 等提出的基于 CNN 的社交网络短文本情感分析模型进

行改进，引入 word2vec 训练词向量和添加批量归一化算法（batch normalization，BN）进行归一化，记为 CNN 模型，作为对比；为进一步验证将社交关系转化为非二元分布对情感分析的促进效果，与主模型去除改进用户关联度的 CNN 模型进行对比，记为 WS-CNN；为验证将社交关系与卷积神经网络进行结合构建文本情感分析的优越性，与传统的支持向量机模型对比，记为 SVM 模型；为证社交关系引入的普遍适用性，SVM 模型与通过社交关系改进的半监督 SVM 模型对比，记为 SS-SVM 模型. 将主模型记为 SA-SRS-CNN，模型设置见表 3.

表 3　模型类别设置

Tab. 3　Settings of model types

模型名称	半监督	深度学习	社交关系	社交强度
SA-SRS-CNN 模型	有	有	有	有
WS-CNN 模型	有	有	有	无
CNN 模型	无	有	无	无
SS-SVM 模型	有	无	有	有
SVM 模型	无	无	无	无

对于主模型，首先使用训练过大规模语料的 word2vec 计算词向量表示；然后计算文本相似度，结合用户关联度构建 TRS-SAT 模型；再依据 TRS-SAT 模型增加标签作为 CNN 的输入，最后通过卷积与池化，捕获并筛选局部特征，训练情感分类器. 在 CNN 部分，使用 Adadelta 算法实现学习率的自动更新，反向传播算法训练模型，随机梯度下降算法求解模型. 其参数选取是在小批量数据集上交叉验证后确定的，实验参数具体设置见表 4.

对于 SS-SVM 模型与 SVM 模型中所用到的支持向量机模型，设置类型为 C-SVC，核函数选择 RBF 核函数，参数 c 和 g 通过交叉验证（cross-validation）和网格搜索（grid-search）得到最优，其中 $c = 1$，$g = 0.005$.

表 4　实验参数设置

Tab. 4　Settings of experimental parameters

参数名	参数设置
词向量长度/个	250
卷积单元数量/个	100
迷你批长度/条	30
迭代次数/次	25
分类层输入节点数/个	40
Adadalta 衰减参数	0.9
卷积核长度/个	2,3,4,5

2.3　实验结果和分析

每组实验均采用十折交叉验证，各组实验结果见表 5.

表 5　实验结果

Tab. 5　Experimental results

模型	SA-SRS-CNN	WS-CNN	CNN	SS-SVM	SVM
准确率	0.755 3	0.733 2	0.725 4	0.678 0	0.632 7
召回率	0.764 1	0.742 3	0.733 5	0.721 5	0.692 7
F 值	0.759 6	0.737 7	0.729 4	0.699 1	0.661 3
准确率提升率/%		3.02	4.12	11.40	18.56
召回率提升率/%		2.94	4.17	5.90	10.06
F 值提升率/%		2.97	4.14	8.65	14.37

由表 5 可知，随着 CNN 模型、社交关系、社交强度对模型的不断改进，模型的分类准确率逐步提高. 将本文提出的 SA-SRS-CNN 模型与 SS-SVM 模型对比，情感分析的准确率、召回率、F 值分别相对提升 11.40%、5.90%、8.65%；与 CNN 模型相比，分别相对提升 4.12%、4.17%、4.14%；结合 WS-CNN 模型可知，其中改进的用户关联度分别贡献 3.02%、2.94%、2.97%. 在基于社交理论的 UR-S 模型的基础上，引入用户属性和文本相似度将 0－1 分布的用户关系转化为量化的文本关系；构建的 TRS-SAT 模型，通过用户关联度和文本相似度同时保证标签添加的效率和准确度，为 SA-SRS-CNN 模型的半监督提供基础，实现标签增加提升情感分析的准确率与效率.

对比 SS-SVM 模型与 SVM 模型，情感分析准确率、召回率、F 值分别提升相对值 7.16%、4.16%、5.72%. 进一步证明 TRS-SAT 模型通过社交关系、文本相似度和用户属性增加标签，能够提升情感分类器性能，优化情感分析效果，具有普遍适用性.

综上可得出结论，本文提出的 SA-SRS-CNN 模型通过用户关联度、文本相似度实现半监督对微博情感分析改善效果显著；利用词向量计算文本的语义特征，利用卷积神经网络挖掘特征集合与情感标签间的深层次关联，能够提升情感分类器性能. 由此可见，该对比实验充分验证本文提出的 SA-SRS-CNN 模型有良好的性能.

3　结　论

为解决监督学习大量的标签获得困难和社交媒体短文本具有的词汇稀疏性、语法随意性、热词性导致的问题，本文基于用户社交关系、用户背景属性、文本相似度构建 TRS-SAT 模型，并结合 CNN 模型，构建 SA-SRS-CNN 模型. 本文充分利用社交网络，建立用户关联度模型和基于用户关联度模型的文本关系模型增加标签数量，结合 CNN 实现半监督的深度学习. 通过对比实验将本文提出的 SA-SRS-CNN 模型与 SS-SVM 模型进行对比，情感分析准确率、召回率、F 值分别提升 0.077 3、0.042 6、0.060 5，相对提升 11.40%、5.90%、8.65%.

该模型有助于提升情感分析准确率可归结为以

下3点:1)基于社会学理论,充分挖掘社交网络的隐含关系,通过社交关系添加标签数量,使得大量的无标签文本得以利用;2)在社交关系基础上,通过文本相似度进行补充校正,进一步提高标签添加的准确率;3)基于深度学习,通过CNN算法,实现对文本语义和标签联系的深层次挖掘,解决短文本自身特点导致的问题. 因此,与SS-SVM模型和CNN模型相比,本文提出的基于用户关联度的半监督情感分析模型不仅提高情感分析准确率,改善社交媒体短文本的情感分析性能,也证明了深度学习和社交网络相结合的方法在未来自然语言处理领域的可行性与重要价值. 同时,本文提出的SA-SRS-CNN模型也存在一些不足:1)建立用户关联度模型需要大量的用户关系数据,在一些用户关系不易获取或者用户间关系较少的情况下,用户关联度矩阵A过于稀疏,能够添加标签数目较少,并且增加了计算复杂度;2)越来越多的用户在发表评论时使用颜文字等由符号构成的表情传递情感,本文提出的SA-SRS-CNN模型并未对这类表情进行处理,忽视了部分情感信息.

参考文献

[1] HU M, LIU B. Mining and summarizing customer reviews[C]// Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, Washington, USA: DBLP, 2004: 168. DOI:10.1145/1014052.1014073

[2] 柳位平, 朱艳辉, 栗春亮, 等. 中文基础情感词词典构建方法研究[J]. 计算机应用, 2009, 29(10): 2875
LIU Weiping, ZHU Yanhui, LI Chunliang, et al. Research on building Chinese basic semantic lexicon[J]. Journal of Computer Applications, 2009, 29(10): 2875

[3] TURNEY P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[C]//Meeting on Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2002: 417. DOI: 10.3115/1073083.1073153

[4] CHINSHA T, JOSEPH S. A syntactic approach for aspect based opinion mining[C]//Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing. [S. l.]: IEEE, 2015: 24. DOI:10.1109/ICOSC.2015.7050774

[5] BHUSHAN S, DANTI A. Classification of compressed and uncompressed text documents[J]. Future Generation Computer Systems, 2018. DOI:10.1109/icosc.2015.7050774

[6] ABDI A, SHAMSUDDIN S, HASAN S. Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment[J]. Expert Systems with Applications, 2018: 109. DOI: 10.1016/j.eswa.2018.05.010

[7] HUSSAIN S, KEUNG J, KHAN A, et al. Implications of deep learning for the automation of design patterns organization[J]. Journal of Parallel and Distributed Computing, 2018: 117. DOI:10.1016/j.jpdc.2017.06.022

[8] KONATE A, DU Ruiying. Sentiment analysis of code-mixed Bambara-French social media text using deep learning techniques[J]. Wuhan University Journal of Natural Sciences, 2018, 23(3): 237. DOI:10.1007/s11859-018-1316-z

[9] 金志刚, 胡博宏, 张瑞. 基于深度学习的多维特征微博情感分析[J]. 中南大学学报(自然科学版), 2018, 49(5): 1135. DOI:10.11817/j.issn.1672-7207.2018.05.015
JIN Zhigang, HU Bohong, ZHANG Rui. Analysis of Weibo sentiment with multi-dimensional features based on deep learning[J]. Journal of Central South University (Science and Technology), 2018, 49(5): 1135. DOI:10.11817/j.issn.1672-7207.2018.05.015

[10] WU Chuhan, WU Fangzhao, WU Sixing, et al. A hybrid unsupervised method for aspect term and opinion target extraction[J]. Knowledge-Based Systems, 2018: 148. DOI:10.1016/j.knosys.2018.01.019

[11] KIM K. An improved semi-supervised dimensionality reduction using feature weighting: Application to sentiment analysis[J]. Expert Systems with Applications, 2018: 109. DOI:10.1016/j.eswa.2018.05.023

[12] WANG Z, MI H, ITTYCHERIAH A. Semi-supervised clustering for short text via deep representation learning[J]. Proceedings of the 20th SIGNLL Conference on Computational Natural Language. Berlin, Germany: Association for Computational Linguistics, 2016: 31. DOI:10.18653/v1/K16-1004

[13] HUX, TANGJ, GAOH, et al. Unsupervised sentiment analysis with emotional signals[C]//International Conference on WorldWide Web. New York: 2013:607. DOI:10.1145/2488388.2488442

[14] HU X, TANG L, TANG J, et al. Exploiting social relations for sentiment analysis in microblogging[J]. WSDM, 2013: 537. DOI: 10.1145/2433396.2433465

[15] WANG Zhiqiang, LIANG Jiye, LI Ru. Exploiting user-to-user topic inclusion degree for link prediction in social-information networks[J]. Expert Systems with Applications, 2018: 108. DOI: 10.1016/j.eswa.2018.04.034

[16] XIAO Yunpeng, LI Xixi, LIU Yuanni, et al. Correlations multiplexing for link prediction in multidimensional network spaces[J]. Science China Information Sciences, 2018, 61(11): 112103. DOI:10.1007/s11432-017-9334-3

[17] XIA L, WANG Z, CHEN C, et al. Research on feature-based opinion mining using topic maps[J]. The Electronic Library, 2016, 34(3): 435. DOI:10.1108/EL-11-2014-0197

[18] LU T. Semi-supervised microblog sentiment analysis using social relation and text similarity[C]// 2015 International Conference on Big Data and Smart Computing. [S. l.]: IEEE, 2015: 194. DOI: 10.1109/35021bigcomp.2015.7072831

[19] SHI Shumin, ZHAO Meng, GUAN Jun, et al. Multi-features group emotion analysis based on CNN for Weibo events[J]. 2017 (cii). DOI:10.12783/dtcse/cii2017/17275

[20] 肖云鹏, 杨光, 刘宴兵, 等. 一种基于最大熵原理的社交网络用户关系分析模型[J]. 电子与信息学报, 2017, 39(4): 778. DOI:10.11999/JEIT160605
XIAO Yunpeng, YANG Guang, LIU Yanbing, et al. Social relationship analysis model based on the principle of maximum entropy[J]. Journal of Electronics & Information Technology, 2017, 39(4): 778. DOI:10.11999/JEIT160605

[21] 徐志明, 李栋, 刘挺, 等. 微博用户的相似性度量及其应用[J]. 计算机学报, 2014, 37(1): 207
XU Zhiming, LI Dong, LIU Ting, et al. Measuring similarity between microblog users and its application[J]. Chinese Journal of Computer, 2014, 37(1): 207

[22] HUANG W, WANG Q, CAO J. Tracing public opinion propagation and emotional evolution based on public emergencies in social networks[J]. International Journal of Computers Communications & Control, 2018, 13(1): 129. DOI:10.15837/ijccc.2018.1.3176

[23] 白朔天, 袁莎, 程立, 等. 多任务回归在社交媒体挖掘中的应用[J]. 哈尔滨工业大学学报, 2014, 46(9): 100. DOI:10.11918/j.issn.0367-6234.2014.09.017
BAI Shuotian, YUAN Sha, CHENG Li, et al. Application of multi-task regression in social media mining[J]. Journal of Harbin Institute of Technology, 2014, 46(9): 100. DOI:10.11918/j.issn.0367-6234.2014.09.017

[24] KIM Y. Convolutional neural networks for sentence classification[EB/OL]. 2014-08-25. https://arxiv.org/abs/1408.5882

(编辑 苗秀芝)