计算机科学
COMPUTER SCIENCE

基于多特征融合的评论文本个性化情感分类新方法

王友卫, 刘奥, 凤丽洲

引用本文

王友卫, 刘奥, 凤丽洲. 基于多特征融合的评论文本个性化情感分类新方法[J]. 计算机科学, 2023, 50(11A): 221000217-7.

WANG Youwei, LIU Ao, FENG Lizhou. Multi-feature Fusion Based New Personalized Sentiment Classification Method for Comment Texts [J]. Computer Science, 2023, 50(11A): 221000217-7.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

方面级情感分析综述

Summarization of Aspect-level Sentiment Analysis

计算机科学, 2023, 50(6A): 220400077-7. https://doi.org/10.11896/jsjkx.220400077

基于知识蒸馏模型ELECTRA-base-BiLSTM的文本分类

Text Classification Based on Knowledge Distillation Model ELECTRA-base-BiLSTM

计算机科学, 2022, 49(11A): 211200181-6. https://doi.org/10.11896/jsjkx.211200181

基于不平衡数据与集成学习的属性级情感分类

Aspect-level Sentiment Classification Based on Imbalanced Data and Ensemble Learning

计算机科学, 2022, 49(6A): 144-149. https://doi.org/10.11896/jsjkx.210500205

基于交互注意力图卷积网络的方面情感分类

Interactive Attention Graph Convolutional Networks for Aspect-based Sentiment Classification

计算机科学, 2022, 49(3): 294-300. https://doi.org/10.11896/jsjkx.210100180

融合双重权重机制和图卷积神经网络的微博细粒度情感分类

Fine-grained Sentiment Classification of Chinese Microblogs Combining Dual Weight Mechanismand Graph Convolutional Neural Network

计算机科学, 2022, 49(3): 246-254. https://doi.org/10.11896/jsjkx.201200073

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000217

基于多特征融合的评论文本个性化情感分类新方法

王友卫[1]　刘　奥[1]　凤丽洲[2]
1 中央财经大学信息学院　北京 100081
2 天津财经大学统计学院　天津 300222
(ywwang15@126.com)

摘　要　现有的情感分类研究未能充分考虑用户个人历史评论中蕴含的个性特征对情感分类结果的影响,且未能综合考虑用户社会关系、个人属性、历史评论与当前评论等诸多因素的共同作用。为此,提出一种基于多特征融合的评论文本个性化情感分类新方法。首先,利用大量无标注的用户历史评论挖掘用户个性表达,结合用户历史评论和用户属性信息提取得到用户特征向量;然后,利用 node2vec 算法在获得图节点表示方面的优势对用户社会关系网络进行学习以得到用户的社会关系向量,并利用预训练的 word2vec 模型获得用户当前评论向量;最后,将用户特征向量、社会关系向量和有标注的当前评论向量输入全连接神经网络中进行训练以得到最终的分类模型。在从中文股吧爬取的真实数据集上的实验结果表明,与支持向量机、朴素贝叶斯、TextCNN、Bert 等典型方法相比,所提方法能够有效提高情感分类的准确率和 F_1 值,验证了其在改善情感分类表现方面的有效性。

关键词:情感分类;股票评论;社会关系;历史评论;全连接神经网络
中图法分类号　TP391

Multi-feature Fusion Based New Personalized Sentiment Classification Method for Comment Texts

WANG Youwei[1],LIU Ao[1] and FENG Lizhou[2]
1 School of Information,Central University of Finance and Economics,Beijing 100081,China
2 School of Science and Engineering,Tianjin University of Finance and Economics,Tianjin 300222,China

Abstract　Existing research on sentiment classification fails to fully consider the influence of personality characteristics contained in user's personal historical comments on the results of sentiment classification,and fails to comprehensively consider the combined effects of many factors such as user's social relations,personal attributes,historical comments and current comments. To this end,a new personalized method for sentiment classification of comment texts based on multi-feature fusion is proposed. First,the user's personality expressions is mined by using a great number of unlabeled user's historical comments,and the user's feature vector is extracted by combining user's historical comments and attribute information. Then,the advantages of the node2vec algorithm in obtaining the node representation of the graph are used to learn users' social relationship networks,so as to obtain the users' social relationship vectors,and the pre-trained word2vec model is used to obtain the user's current comment vector. Finally,the user's feature vector,social relationship vector and labeled current comment vector are entered into the fully connected classifier for training to obtain the final classification model. Experimental results on the real data set crawled from the Chinese stock page show that compared with typical methods such as support vector machine,naive Bayes,TextCNN,Bert,the proposed method can effectively improve the accuracy and F_1 value of sentiment classification,which verifies its effectiveness in improving sentiment classification performance.

Keywords　Sentiment classification,Stock comments,Social relations,Historical comments,Full connect neural network

随着信息化的发展,越来越多的投资者在论坛网站上通过发表评论来表达自己的观点。股吧中投资者的自身情感和观点极易受到其他投资者的影响,从而改变其投资交易行为。通过对股票评论的情感分析,可以帮助投资者更好地了解股票市场变化,给投资者提出相应的投资建议,减少投资风险。并且,还可以一定程度上预测股票的短线走势和市场波动,帮助企业提前做好应对措施以化解风险。因此,基于评论文本的情感分类在股票市场分析中具有重要的研究意义。

目前情感分类方法可分为基于情感词典的分类方法、基于机器学习的分类方法和基于深度学习的分类方法。基于

基金项目:国家自然科学基金(61906220);教育部人文社科项目(19YJCZH178);国家社科基金(18CTJ008);中央财经大学新兴交叉学科建设项目
This work was supported by the National Natural Science Foundation of China(61906220),Ministry of Education of Humanities and Social Science Project(19YJCZH178),National Social Science Foundation of China(18CTJ008) and Emerging Interdisciplinary Project of CUFE.
通信作者:刘奥(liuaohit@163.com)

情感词典的分类方法需要构建包含各种情绪词汇的情感词典，每个词汇有相应的数值化情感倾向，评论文本通过词汇匹配得到情感类别。Maqsood 等[1]利用情感词典 SentiWordNet 对从 tweet 的英文股票评论中提取出的 5000 个词汇进行情感标注，以此得到英文股票评论的情感词典。Chen 等[2]采用改进模拟退火算法对情感词典的词语分值进行优化，通过解决优化问题来提升情感词典性能，但仍存在词典情感判断不准确、词典构建复杂的问题。随着机器学习和深度学习的发展，相关的技术也被运用到股票评论的情感分类中。Alkubaisi 等[3]利用混合朴素贝叶斯的方法对股票评论进行情感分类，在 tweet 的股票评论数据集上达到 90.38%的准确率。Liu 等[4]提出了基于 Bert 预训练模型的 FinBert，模型通过大规模英文金融语料训练得到，解决了金融领域无预训练模型的问题。Cheng 等[5]提出了基于注意力机制的多通道 CNN 和 BiGRU 的情感分类模型，模型将 CNN 和 BiGRU 模型并行并结合注意力机制，以同时提取句子连续词的局部信息以及长文本的上下文信息，提高了模型文本特征的提取能力。但是，以上方法仅将现有机器学习或深度学习方法应用于金融及其他领域的文本，未能准确挖掘用户个性特征和社会关系对文本情感分析的影响。为此，Hu 等[6]利用情感一致性和情绪感染的社会学理论，提出了 SANT 方法（Sociological Approach to handling Noisy and short Texts），将用户间的社会关系融入对多噪音和短文本的情感分类中，实验研究证明了融入社会关系对情感分类的可行性和有效性。Liu 等[7]通过用户话题情感的一致性和社交关系的认可度（点赞和转发）来建立微博文本之间的情感关系，构建了半监督的情感分类模型，使模型减少了对训练数据集的依赖。Yang 等[8]利用用户与用户、用户与博文之间的关系构建图，并利用 LINE 算法得到含有社交关系信息的节点向量，通过含有注意力机制的神经网络将节点向量和评论向量融合，取得了较好的实验结果。

研究发现，目前的情感分类方法仍面临以下问题：(1)忽略了用户历史评论中蕴含的个性特征对情感分类结果的影响，未充分利用大量无标注历史评论数据对用户个性化情感特征进行建模。(2)缺乏综合考虑用户属性特征、用户历史评论特征、用户社交关系特征和当前评论文本内容特征的情感分类模型，限制了评论文本情感分类的准确性。为此，本文提出一种基于多特征融合的评论文本个性化情感分类新方法。主要贡献包括：(1)利用用户大量无标注的历史评论提取用户个性特征，结合用户属性特征构建用户特征向量。同时，利用 node2vec 算法能够灵活调整随机游走走向的优势，生成具有网络全局信息和局部信息的用户社会关系向量。(2)综合利用用户属性特征、用户历史评论特征、用户社交关系特征和当前评论文本内容特征，在保证模型分类结果个性化的同时考虑用户之间的相互影响，提高情感分类的准确性。

1 问题描述

定义 1（用户当前评论） 当前进行情感分类的用户评论文本，表示为 U_{ij}^{C}（$1\leqslant i\leqslant N$），代表用户 i 第 j 条当前评论。

定义 2（用户评论集） 用户一段时间内在平台上发表的所有评论，定义为 $U^{P}=\{U_i^{P}\}$（$1\leqslant i\leqslant N$），其中 $U_i^{P}=\{U_{ij}^{P}\}$（$1\leqslant j\leqslant C_i$，$C_i$ 为用户 i 具有的评论数），U_{ij}^{P} 表示第 i 个用户的第 j 条评论。

定义 3（社会关系） 用户间社会关系表示为 $G=(V,\boldsymbol{A})$，其中 $V=\{v_i\}$（$1\leqslant i\leqslant N$，$|V|=N$）为用户集合，$\boldsymbol{A}\in\boldsymbol{R}^{N\times N}$ 为用户的邻接矩阵，能够表达有向图 G 的结构信息，$\boldsymbol{A}_{ij}$ 的定义如式(1)所示：

$$\boldsymbol{A}_{ij}=\begin{cases}1, & \text{用户节点 } v_i \text{ 关注用户节点 } v_j\\ 0, & \text{其他}\end{cases} \tag{1}$$

定义 4（用户属性特征） 用户 i 的属性特征为 F_i，其中，$F_i=\{fun_{ij}\}$（$1\leqslant j\leqslant n$，n 为用户属性个数），fun_{ij} 表示用户 i 的第 j 个属性值，通常指用户的粉丝数、关注人数、影响力值等信息。

在此基础上，本文综合多种特征的影响，将情感分类函数定义为 $F(\boldsymbol{h}_{ij},\Omega)$。若 $F(\boldsymbol{h}_{ij},\Omega)=1$，则 U_{ij}^{C} 的情感类别为积极，否则，其情感类别为消极。其中，$\boldsymbol{h}_{ij}$ 为 U_{ij}^{C} 对应的多特征融合后的向量，Ω 为模型所需参数。为便于理解，本文相关符号定义如表 1 所列。

表 1 本文主要符号及含义

Table 1 Main symbols and their meanings

符号	含义
$U_i^{P}=\{U_{ij}^{P}\}$	第 i 个用户的所有评论
$U^{P}=\{U_i^{P}\}$	用户评论数据集
$word_{ijk}$	将 U_{ij}^{P} 分词、去停用词等操作后的第 k 个词
$\boldsymbol{h}_{ijk}^{W}\in\boldsymbol{R}^{m}$	$word_{ijk}$ 对应的词向量
$\boldsymbol{h}_{ij}^{S}\in\boldsymbol{R}^{m}$	U_{ij}^{P} 的句向量
$\boldsymbol{h}_i^{L}\in\boldsymbol{R}^{m}$	用户 i 的历史评论向量
$\boldsymbol{h}_i^{F}\in\boldsymbol{R}^{m+n}$	用户 i 的特征向量
F_i	用户 i 的属性集
$\boldsymbol{h}_i^{A}\in\boldsymbol{R}^{n}$	用户 i 的属性向量
U_{ij}^{C}	用户 i 的第 j 条当前评论
$\boldsymbol{h}_{ij}^{C}\in\boldsymbol{R}^{m}$	用户当前评论 U_{ij}^{C} 的句向量
V	G 的节点集合
$\boldsymbol{A}$	G 的邻接矩阵
$G=(V,\boldsymbol{A})$	用户间的社会关系网络
$\boldsymbol{h}_i^{R}\in\boldsymbol{R}^{m+n}$	用户 i 的关系向量
N	用户总数
M	用户社会关系网络的边数
m	词向量的维数
n	属性向量的维数

2 基于多特征融合的评论文本个性化情感分类新方法

本文方法主要分为 4 个步骤，具体过程如图 1 所示。

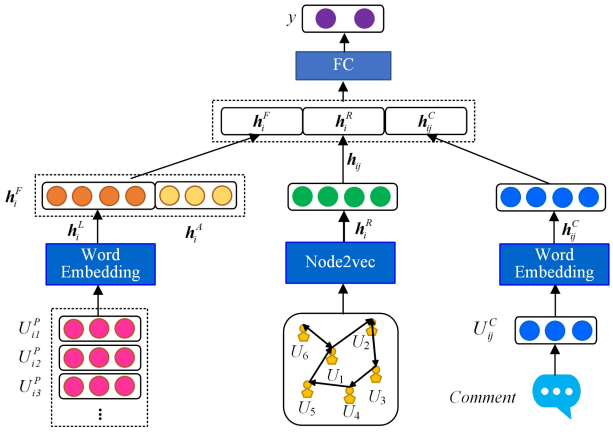


图 1 本文方法整体框架

Fig. 1 Overall framework of the proposed method

(1)用户特征向量生成，从用户评论集中抽取用户历史评论特征，结合用户属性特征形成用户特征向量；

(2)用户社会关系向量生成，利用 node2vec 算法对用户社会关系进行学习得到用户节点的社会关系向量；

(3)用户当前评论向量生成，利用预训练的 word2vec 模型提取当前评论文本语义以得到当前评论向量；

(4)多特征融合，将用户特征向量、社会关系向量、当前评论文本向量拼接得到融合后的向量，输入到全连接分类器中以得到情感分类结果。

2.1 用户特征向量生成

Jiang 等[9]采用分层的多头注意力机制，从多个角度挖掘用户和产品信息，使模型能更全面地获取用户和产品特征对情感分类的影响。Wang 等[10]根据评论与用户、产品信息之间的关联性建图，利用基于图卷积神经网络的模型学习用户和产品信息对评论的影响，从而提升情感分类性能。但上述模型仅利用用户带标注的当前评论来提取文本特征，未充分利用用户大量的无标注历史评论中蕴含的用户个性信息。因此，本文为每个用户构建个性化用户特征向量，其包含历史评论中蕴含的个性信息和粉丝数、访问量、影响力等属性信息，以此融合了用户的用词习惯、性格特点、态度偏好、影响力及活跃程度等个性化因素。具体过程如下。

2.1.1 用户历史评论向量生成

以用户 $i(1\leqslant i\leqslant N)$为例，其用户历史评论向量的生成过程如图 2 所示。

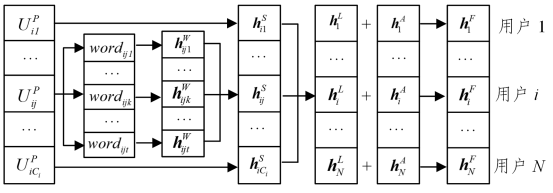


图 2 用户历史评论向量生成

Fig. 2 Generation of user's historical comment vector

图 2 中，C_i 为用户 i 具有的评论数，U_{ij}^{P} 表示第 i 个用户的第 j 条评论，$word_{ijk}$ 表示将 U_{ij}^{P} 分词、去停用词等操作后的第 k 个词，t 为该评论含有的词数，$\boldsymbol{h}_{ijk}^{\mathrm{W}}$ 表示 $word_{ijk}$ 对应的词向量，$\boldsymbol{h}_{ij}^{\mathrm{S}}$ 表示评论 U_{ij}^{P} 的句向量，$\boldsymbol{h}_i^{\mathrm{L}}$ 表示用户 i 的历史评论向量，$\boldsymbol{h}_i^{\mathrm{A}}$ 表示用户 i 的属性向量，$\boldsymbol{h}_i^{\mathrm{F}}$ 表示用户 i 的特征向量。“+”表示向量的拼接操作。

为获得用户 i 的第 j 个评论 U_{ij}^{P} 的句向量 $\boldsymbol{h}_{ij}^{\mathrm{S}}$，将 U_{ij}^{P} 所有词对应的词向量乘上该词相应的 tf-idf 值，以体现句子中每个词的重要程度，如式(2)所示。

$$\boldsymbol{h}_{ij}^{\mathrm{S}}=\frac{\sum_{k=1}^{t}\boldsymbol{h}_{ijk}^{\mathrm{W}}*\exp(T_k)}{t} \tag{2}$$

其中，exp()表示以 e 为底的指数函数，T_k 为词 $word_{ijk}$ 的 tf-idf 值，计算方法如式(3)－式(5)所示：

$$T_k=tf_k\times idf \tag{3}$$

$$tf_k=\frac{n_k}{t} \tag{4}$$

$$idf=\lg\frac{n_P}{n_{Pk}+1} \tag{5}$$

其中，tf_k 为词 $word_{ijk}$ 的词频，idf 为逆向文件频率。n_k 为词 $word_{ijk}$ 在本评论中出现的次数，n_p 为所有评论的个数，n_{pk} 为包含词 $word_{ijk}$ 的评论数。可见，T_k 越大，说明 $word_{ijk}$ 越重要，对评论句向量 $\boldsymbol{h}_{ij}^{\mathrm{S}}$ 的影响越大。

进一步地，按照式(6)计算得到用户 i 的历史评论向量 $\boldsymbol{h}_i^{\mathrm{L}}$ $(1\leqslant i\leqslant N, h_i^{\mathrm{L}}\in\boldsymbol{R}^m, m$ 为词向量维数)。

$$\boldsymbol{h}_i^{\mathrm{L}}=\mathrm{mean}(\boldsymbol{h}_{ij}^{\mathrm{S}}) \tag{6}$$

其中，mean()为向量按列求平均函数。

2.1.2 用户属性向量生成

首先，按照定义 4 提取 F_i 对应的属性值，以此得到一个 n 维向量 $\boldsymbol{V}_i=[v_{ij}](0\leqslant j<n)$。在此基础上，为避免不同维度跨度大小的影响，对该向量每维数据进行 Max-Min 归一化操作，以得到用户 i 的属性向量 $\boldsymbol{h}_i^{\mathrm{A}}=[\boldsymbol{h}_{ij}^{\mathrm{A}}](0\leqslant j<n)$，具体如式(7)所示。

$$\boldsymbol{h}_{ij}^{\mathrm{A}}=\frac{v_{ij}-v_{ij}^{\min}}{v_{ij}^{\max}-v_{ij}^{\min}} \tag{7}$$

其中，v_{ij} 表示单个原始数据，$\boldsymbol{h}_{ij}^{\mathrm{A}}$ 为数据归一化后的值，$v_{ij}^{\min}$ 为样本数据的最小值，$v_{ij}^{\max}$ 为样本数据的最大值。在此基础上，按照式(8)获得用户 i 的特征向量 $\boldsymbol{h}_i^{\mathrm{F}}$。

$$\boldsymbol{h}_i^{\mathrm{F}}=\boldsymbol{h}_i^{\mathrm{L}}\parallel\boldsymbol{h}_i^{\mathrm{A}} \tag{8}$$

其中，$\parallel$ 表示向量拼接操作，$\boldsymbol{h}_i^{\mathrm{F}}\in R^{m+n}$ 为用户 i 的特征向量。

2.2 用户社会关系向量生成

Hu 等[6]将用户间的社会关系信息融入对多噪音和短文本的情感分类中，并通过实验验证了融入社会关系对情感分类的可行性和有效性。Zou 等[11]不仅考虑用户间的直接关系，还利用拉普拉斯矩阵将用户隐含关系和话题间的相关性结合起来，使模型具有更好的情感分类性能。本文利用 node2vec 算法[12]对用户社会关系网络结构进行学习，通过控制节点序列产生过程中的游走趋向将更大范围的网络结构信息体现到节点向量中。假设 U_1－U_6 为股吧平台的实际用户，并且这些用户间的社会关系如图 3 所示，则由定义 3 知，U_1－U_6 对应的用户邻接矩阵如图 4 所示。

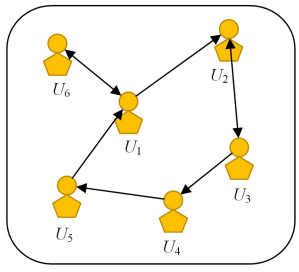


图 3 用户社会关系示意图

Fig. 3 Schematic diagram of users' social relationships

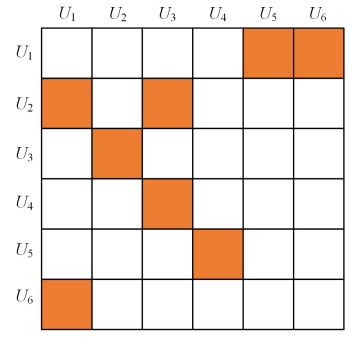


图 4 用户邻接矩阵示意图

Fig. 4 Schematic diagram of users' adjacency matrix

图 3 中，单向箭头和双向箭头分别表示用户间的单向关注和双向关注关系。图 4 中，颜色填充表示对应用户间存在

关注关系。针对用户社会关系 $G=(V,\boldsymbol{A})$，本文使用 node2vec 算法得到一系列节点序列，并通过 Skip-gram 模型训练获得用户社会关系向量集 $H_R=\{\boldsymbol{h}_i^R\}(1\leqslant i\leqslant N,\boldsymbol{h}_i^R\in R^{m+n})$。训练过程对应的目标函数如式(9)所示。

$$L=\frac{1}{N}\sum_{t=1}^{N}\sum_{-c\leqslant j\leqslant c,j\neq 0}\log p(w_{t+j}\mid w_t) \tag{9}$$

其中，N 为用户数量，c 为窗口大小，w_t 为第 t 个用户节点。$p(w_{t+j}\mid w_t)$ 的计算方法如式(10)所示：

$$p(w_{t+j}\mid w_t)=\frac{\boldsymbol{u}_{w_{t+j}}^{\mathrm{T}}v_{w_t}}{\sum_{i=1}^{|V|}\exp(\boldsymbol{u}_{w_i}^{\mathrm{T}}v_{w_t})} \tag{10}$$

其中，u_w 与 v_w 分别为节点 w 是背景节点和中心节点的向量表达。

2.3 基于多特征融合的情感分类

本文对用户评论数据集进行筛选，选取字长不超过 400 的用户评论进行标注，形成带有标注的用户当前评论数据 $U_{ij}^{\mathrm{C}}(1\leqslant i\leqslant N)$，对应的标签为 $\hat{y}_{ij}(\hat{y}_{ij}\in\{0,1\})$。进一步地，对当前评论进行分词、去停用词等操作，并利用预训练的 skip-gram 模型将各词语转换为词向量，再对词向量对应按列取平均，将用户当前评论 U_{ij}^{C} 转化为相应的句向量 $\boldsymbol{h}_{ij}^{\mathrm{C}}\in\boldsymbol{R}^m$。

在此基础上，本文对当前评论向量、用户特征向量和社会关系向量进行拼接，以此获得当前评论文本的最终向量表达，如式(11)所示。

$$\boldsymbol{h}_{ij}=[\boldsymbol{h}_{ij}^{\mathrm{C}}\parallel\boldsymbol{h}_i^{\mathrm{F}}\parallel\boldsymbol{h}_i^{\mathrm{R}}] \tag{11}$$

其中，$\boldsymbol{h}_{ij}\in\boldsymbol{R}^{3m+2n}$。进一步地，将 $\boldsymbol{h}_{ij}$ 输入到全连接层和 softmax 层中，以获得当前评论所对应的情感类别，如式(12)、式(13)所示。

$$\alpha=\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{h}_{ij}+\boldsymbol{b}_k \tag{12}$$

$$p_{ij}=\mathrm{softmax}(\alpha)=\frac{\exp(\alpha)}{\sum_{i=1}^{2}\exp(\alpha)} \tag{13}$$

其中，$p_{ij}\in\boldsymbol{R}^2$ 为当前评论 U_{ij}^{C} 的情感类别的概率分布结果，$\alpha\in\boldsymbol{R}^2$ 为全连接层的输出结果，$\boldsymbol{W}_1\in\boldsymbol{R}^{d1\times(3m+2n)}$，$\boldsymbol{W}_c\in\boldsymbol{R}^{dk\times dk-1}$，$\boldsymbol{b}_k\in\boldsymbol{R}^{dk}$，$2\leqslant c\leqslant 3$，$k$ 为全连接层数，$1\leqslant k\leqslant 3$，d_a 表示全连接层第 a 层的维度。

本文使用交叉熵函数作为模型分类损失函数，为避免过拟合，在损失函数中加入了 L2 正则化惩罚项，如式(14)所示。

$$L_{\mathrm{loss}}=-\sum_{i=1}^{N}\sum_{j=1}^{R_i}\hat{y}_{ij}\ln y_{ij}+\lambda\parallel\theta\parallel^2 \tag{14}$$

其中，L_{loss} 为损失函数值，$\hat{y}_{ij}$ 为 U_{ij}^{C} 的真实情感标签，y_{ij} 为预测情感标签，θ 为模型全部参数，λ 为 L2 正则化的惩罚系数，λ 越大，惩罚力度越大。

在此基础上，给出本文执行过程如算法 1 所示。

算法 1 基于多特征融合的评论文本个性化情感分类新方法

输入：用户评论集 $U^{\mathrm{P}}=\{U_i^{\mathrm{P}}\}(1\leqslant i\leqslant N)$，属性集 $F=\{F_i\}(1\leqslant i\leqslant N)$，用户的当前评论数据集 $U^{\mathrm{C}}=\{\{U_{ij}^{\mathrm{C}}\}\}(1\leqslant i\leqslant N,1\leqslant j\leqslant R_i,R_i$ 为用户 i 具有的当前评论数)，与 U_{ij}^{C} 对应的标签集 $\hat{y}_{ij}\in\{0,1\}$，用户社会关系 $G=(V,\boldsymbol{A})$，用户 i 的待分类当前评论 U_{ix}^{C}

输出：U_{ix}^{C} 的预测情感类别标签 y_{ix}

1. 按照 2.1 节的步骤对 U^{P} 和 F 进行处理，得到用户的特征向量集 $H_F=\{\boldsymbol{h}\}_i^{\mathrm{F}}(1\leqslant i\leqslant N)$。

2. 利用 node2vec 算法处理用户社会关系图 $G=(V,\boldsymbol{A})$，得到用户的社会关系向量集：$\boldsymbol{H}_R=\{\boldsymbol{h}_i^R\}(1\leqslant i\leqslant N)$。

3. for i=1:1:N

4. 利用 2.3 节的方法对用户 i 的当前评论数据集 U_i^{C} 进行处理，可得到用户 i 的当前评论向量集 $\boldsymbol{H}_i^{\mathrm{C}}=\{\boldsymbol{h}_{ij}^{\mathrm{C}}\}(1\leqslant j\leqslant R_i)$。

5. 利用式(11)将用户特征向量、关系向量、当前评论向量拼接融合，得到 $\boldsymbol{h}_{ij}$。

6. end for

7. 如式(15)所示，将 $\boldsymbol{h}_{ij}$ 输入全连接层并训练即可得到本文的情感分类模型 FC_model。

$$\mathrm{FC\_model}=\mathrm{FC\_train}(\mathrm{H},\hat{\mathrm{Y}},\Omega) \tag{15}$$

其中，FC_train 为全连接层训练函数，$H=\{\boldsymbol{h}_{ij}\}(1\leqslant i\leqslant N,1\leqslant j\leqslant R_i)$，$\hat{Y}=\{\hat{y}_{ij}\}$，$\Omega$ 为训练所需的参数，FC_model 为最终得到的训练模型。

8. 预测 U_{ix}^{C} 的情感类别时，先对 U_{ix}^{C} 进行步骤 4、步骤 5 的处理以得到多特征融合后的向量 $\boldsymbol{h}_{ix}$。在此基础上，将 $\boldsymbol{h}_{ix}$ 输入 FC_predict 模型中即可输出 U_{ix}^{C} 对应的情感类别 y_{ix}，如式(16)所示：

$$y_{ix}\leftarrow\mathrm{FC\_predict}(\mathrm{FC\_model},\boldsymbol{h}_{ix},\Omega) \tag{16}$$

其中，FC_predict 为情感类别预测函数。

这里进一步对本文算法的时间复杂度进行分析：生成用户特征向量和当前评论向量的时间复杂度均为 $O(N)$（N 为用户总数）；对于用户关系向量的生成所使用的 node2vec 算法，其随机游走部分的时间复杂度为 $O(N)+O(M)$（M 为用户关注网络的边数），节点向量生成部分的时间复杂度为 $O(N\lg N)$；全连接分类器的时间复杂度为 $O(3m+2n)$。可见，本文算法的整体时间复杂度为 $O(M+N\lg N)$。

3 实验结果与分析

3.1 数据集

本文针对东方财富网股吧平台的用户，使用八爪鱼专业版软件爬取其在 2021 年 11 月之前发布的评论（最大不超过 60 条）。每个用户选取最多 20 个关注者，最终得到了用户评论数据集，数据集中共含有 437 个用户，17391 条评论，479 条关注边数。对每个用户爬取包括粉丝数、吧龄、访问量、关注的人数、影响力值等属性信息。在所有的用户评论中，选取标题和正文总字数不超过 400 的评论进行积极性和消极性标注，以此得到 5793 条标注数据。

为了进一步提升数据集标注的质量，鉴于置信学习[13]可直接估计噪声标签与真实标签联合分布的优势，本文使用基于置信学习的 cleanlab 方法对数据进行除噪操作，得到本文的用户当前评论数据集（包含有 3321 条评论，288 个用户，其中积极评论数共 2414 条）。此外，本文使用准确率（Acc）、F_1 值和精确率（Pre）[14]来衡量模型整体分类的性能。实验训练集、验证集和测试集的划分比例为 8∶1∶1，每次实验的结果由 10 次随机实验的平均值得到。

3.2 对比方法及参数设置

选取以下几种基线方法进行对比。(1)传统机器学习方法：支持向量机（Support Vector Machines，SVM）[15]、随机森林（Random Forest，RF）[16]、朴素贝叶斯（Native Bayesian，NB）和逻辑回归（Logistics Regression，LR）[17]。(2)深度学习分类方法：全连接神经网络（Full Connection，FC）、Bi-LSTM[18]、TextCNN[19]、TextRCNN[20]、Bert[21]、MC-CNN-AttBiGRU[5]、BiGRU-CNN[22]、Att-CNN-BiGRU[23]。若无特殊说明，所有深度学习方法默认迭代轮次 $epochs=100$，批大

小 $batch\_size=64$，学习率 $learning\_rate=0.0001$，优化器为 Adam，词向量通过 skip-gram 模型获得，词嵌入维度 $m=100$，句子最大长度 $max\_len=100$，属性向量维数 $n=5$。

上述方法涉及的其他参数如表 2 所列。

表 2　不同方法的主要参数设置

Table 2　Main parameter settings of different methods

方法	实验设定
SVM	惩罚系数 $penalty\_factor=4.0$
LR	正则化系数 $c=1.6$
RF	决策树数目 $n\_estimators=200$
NB	平滑因子 $alpha=0.8$
Bi-LSTM	层数 $num\_layers=2$，学习率 $learning\_rate=0.00005$
TextCNN	卷积核尺寸 $kernel\_sizes$ 分别为 3,4,5，卷积核数量 $num\_kernel=100$
TextRCNN	lstm 层数 $num\_lstm=1$
Bert	利用 bert-base-chinese 模型得到预训练词向量，迭代轮次 $epochs=20$
MC-CNN-AttBiGRU	学习率 $learning\_rate=0.00003$，迭代轮次 $epochs=50$，词向量维数 $word2vec\_size=300$
BiGRU-CNN	学习率 $learning\_rate=0.00003$
Att-CNN-BiGRU	词向量维数 $word2vec\_size=200$
Our algorithm	全连接网络参数 $parameters$ 为 $d_1=256$，$d_2=128$，$d_3=2$

3.3　基于历史评论的用户个性特征有效性验证

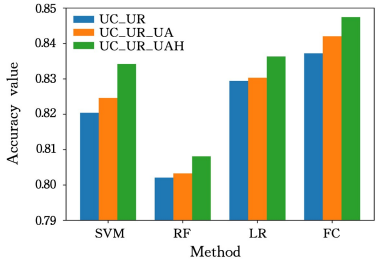
为了验证基于历史评论的用户个性特征对提升情感分类性能的有效性，以 SVM，RF，LR，FC 为分类器，通过改变以上分类器输入的特征向量以获得不同的情感分类方法，具体如下：

(1)UC_UR：输入的特征向量为用户当前评论向量拼接 node2vec 算法所得的用户社会关系向量。

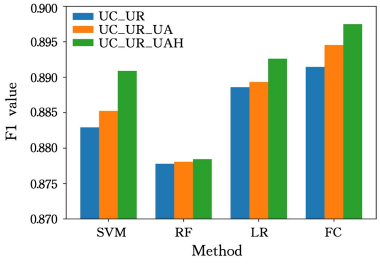
(2)UC_UR_UA：输入的特征向量为用户当前评论向量拼接用户社会关系向量、用户属性向量。

(3)UC_UR_UAH：输入的特征向量为用户当前评论向量拼接用户社会关系向量、基于历史评论和属性的用户特征向量。

图 5 给出了以上各分类器在不同输入特征下的情感分类结果。由图可知，UC_UR_UAH 的表现均优于 UC_UR_UA。当使用 SVM，RF，LR，FC 作为分类器时，UC_UR_UAH 相对于 UC_UR_UA 在 Accuracy 值上分别提高了 0.96%，0.48%，0.60%，0.54%，在 F_1 值上分别提高了 0.56%，0.04%，0.33%，0.30%，验证了从用户历史评论中提取个性特征信息并将其融入评论文本内容对情感分类性能的提升作用。



(a)Accuracy 值



(b)F_1 值

图 5　基于历史评论的用户个性特征有效性验证

Fig. 5　Validity verification of user personality characteristics based on historical reviews

进一步发现，UC_UR_UAH 的分类性能相比 UC_UR 有显著提升。当使用 SVM，RF，LR，FC 作为分类器时，UC_UR_UAH 相对于 UC_UR 在 Accuracy 值上分别提高了 1.38%，0.60%，0.69%，1.02%，在 F_1 值上分别提高了 0.80%，0.07%，0.40%，0.60%，说明融合基于用户历史评论和属性的用户特征向量对提升评论文本的情感分类性能的有效性。

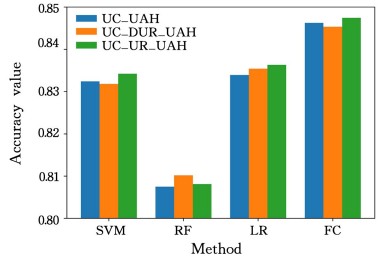
3.4　基于 node2vec 的用户社会关系特征有效性验证

为了验证基于 node2vec 的用户社会关系特征的有效性，仍以 SVM，RF，LR，FC 作为分类器，通过改变以上各分类器输入的特征向量获得不同的情感分类方法，具体如下：

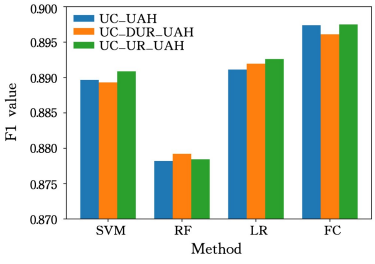
(1)UC_UAH：输入的特征向量为用户当前评论向量拼接基于历史评论和属性的用户特征向量。

(2)UC_DUR_UAH：输入的特征向量为用户当前评论向量拼接 DeepWalk 算法[24]所得的用户社会关系向量、基于历史评论和属性的用户特征向量。

图 6 中给出了以上各分类器在不同输入特征下的情感分类结果。



(a)Accuracy 值



(b)F_1 值

图 6　基于 node2vec 的用户社会关系特征有效性验证

Fig. 6　Validity verification of user social relationship features based on node2vec algorithm

由图 6 可知，UC_UR_UAH 表现均优于 UC_UAH。当

使用 SVM,RF,LR,FC 作为分类器时,UC_UR_UAH 相较于 UC_UAH 在 Accuracy 值上分别提高了 0.18%,0.06%,0.24%,0.12%,在 F_1 值上分别提高了 0.12%,0.02%,0.15%,0.01%,说明融合用户社会关系特征对情感分类性能具有较好的提升效果,同时也验证了在社交网络中用户情绪和行为易受到有关注关系的用户的影响。

进一步发现,除 RF 外,UC_UR_UAH 的表现均优于 UC_DUR_UAH。当使用 SVM,LR,FC 作为分类器时,UC_UR_UAH 相较于 UC_DUR_UAH 在 Accuracy 值上分别提高了 0.24%,0.09%,0.21%,在 F_1 值上分别提高了 0.16%,0.07%,0.14%。这说明相比 DeepWalk 算法,使用 node2vec 算法得到的用户社会关系向量对提升情感分类的性能更加有效,说明了使用 node2vec 算法可以得到更加符合用户社会关注网络结构、更具代表性的节点向量表达。

3.5 消融实验

为验证本文方法中不同特征向量对评论情感分类表现的影响,在图 5、图 6 基础上,以全连接网络为分类器,通过进一步改变输入的特征向量获得不同的情感分类方法。

(1)UC:输入的特征向量为用户当前评论向量。

(2)UC_UA:输入的特征向量为用户当前评论向量拼接用户属性向量。

(3)UC_UH:输入的特征向量为用户当前评论向量拼接用户历史评论向量。

(4)UC_UR_UH:输入的特征向量为用户当前评论向量拼接用户社会关系向量、用户历史评论向量。

表 3、表 4 分别列出了不同方法对应的 Accuracy 值和 F_1 值,由表知,在当前评论向量的基础上,单独添加用户社会关系向量、用户属性向量、用户历史评论向量均可以使 Accuracy 值和 F_1 值得到提升。相比只以当前评论向量为特征的分类方法,单独添加用户历史评论向量对提升情感分类性能更加有效,其在 Accuracy 和 F_1 值上分别提升了 0.6%,0.3%。进一步发现,UC_UR_UA,UC_UR_UH,UC_UR_UAH 在 Accuracy 和 F_1 值上均比在当前评论向量上加入相应单个特征的分类方法有所提升,验证了综合用户社会关系向量、用户属性向量及用户历史评论向量对情感分类性能的提升作用。由于本文综合了用户社会关系向量、基于历史评论和属性的用户特征向量与当前评论向量,因此获得了最好的情感分类效果,比只考虑当前评论向量的分类方法(UC)在 Accuracy 和 F_1 值上分别提升了 1.2%与 0.6%,可见,基于多特征融合的个性化情感分类方法能有效融合不同个性特征的影响,提高评论文本情感分类的准确性。

表 3 不同方法对应的 Accuracy 值

Table 3 Accuracy values of different methods

方法	*Acc*
UC	0.835
UC_UR	0.837
UC_UA	0.838
UC_UH	0.842
UC_UR_UA	0.842
UC_UR_UH	0.845
UC_ UAH	0.846
UC_UR_UAH	**0.847**

表 4 不同方法对应的 F_1 值

Table 4 F_1 values of different methods

方法	F_1
UC	0.891
UC_UR	0.891
UC_UA	0.892
UC_UH	0.894
UC_UR_UA	0.895
UC_UR_UH	0.896
UC_ UAH	0.897
UC_UR_UAH	**0.897**

3.6 综合比较

表 5 和表 6 分别列出了本文方法与传统机器学习方法在 Accuracy 值、F_1 值和 Pre 值上的对比结果。

表 5 与传统机器学习方法的 Accuracy、F_1 值、Pre 值比较

Table 5 Comparison of accuracy, F_1 values and Pre values of traditional machine learning

分类方法	*Acc*	F_1	*Pre*
SVM	0.818	0.881	0.838
RF	0.791	0.870	0.791
LR	0.819	0.882	0.840
NB	0.763	0.836	0.843
Our algorithm	**0.847**	**0.898**	**0.898**

表 6 与典型深度学习方法的 Accuracy、F_1 值、Pre 值比较

Table 6 Comparison of accuracy, F_1 values and Pre values of typical deep learning

分类方法	*Acc*	F_1	*Pre*
MC-CNN-AttBiGRU[5]	0.810	0.876	0.834
TextRCNN[20]	0.823	0.879	0.866
Bert[21]	0.824	0.882	0.853
Bi-LSTM[18]	0.827	0.883	0.857
Att-CNN-BiGRU[23]	0.829	0.885	0.859
BiGRU-CNN[22]	0.835	0.890	0.862
FC	0.835	0.891	0.864
TextCNN[19]	0.840	0.894	0.859
Our algorithm	**0.847**	**0.898**	**0.898**

由表 5 可知,本文方法的表现均优于传统机器学习方法(SVM,RF,LR 和 NB),比传统方法中表现最好的 SVM 方法在 Accuracy 值、F_1 值和 Pre 值上分别提升了 2.9%,1.7%和 6.0%。这是因为传统分类器仅将用户当前评论句向量作为特征,而忽略了用户的特征和社会关系信息,同时此类方法依赖输入的特征质量,影响了其分类性能。由表 6 可知,TextCNN 模型在本文数据集上的表现优于其他算法,本文方法相比 TextCNN 在 Accuracy 值、F_1 值和 Pre 值上分别提升了 0.7%,0.4%和 3.9%,说明了在用户当前评论文本特征的基础上综合考虑用户历史语言、属性和社交关系对提升情感分类性能的有效性。结合表 5 可知,表 6 中基于深度学习的情感分类方法表现虽普遍优于表 5 中的传统机器学习方法,但相较于本文方法而言仍具有一定劣势,这说明深度学习方法虽然能通过较强的特征学习能力改善模型分类表现,但是因其忽略了用户的个性特征以及社交关系信息,其在情感分类方面的表现受到限制。

结束语 本文提出了一种基于多特征融合的评论文本个性化情感分类的新方法,主要贡献包括:(1)充分利用用户大量无标注历史评论中蕴含的用户个性信息,为每个用户构建包含用户个性特征和属性特征的用户特征向量,并利用

node2vec算法融合用户之间的相互影响,提高社交网络用户表达的准确性。(2)综合用户属性特征、用户历史评论特征、用户社交关系特征和当前评论文本内容特征,提出基于多特征融合的情感分类方法,在保证个性化的同时考虑用户之间的相互影响。在从股吧爬取的真实数据集上的实验结果表明,与典型传统分类方法和深度学习方法相比,基于多特征融合的情感分类方法能有效提高情感分类表现。未来考虑将该方法与推荐系统相结合,通过获取投资者评论的情感倾向为投资者提出投资建议,以减少投资风险。

参 考 文 献

[1] MAQSOOD H,MEHMOOD I,MAQSOOD M,et al. A local and global event sentiment based efficient stock exchange forecasting using deep learning[J]. International Journal of Information Management,2020,50:432-451.

[2] CHEN K J,CHEN R H. Automatic Construction and Optimization of Stock Market Sentiment Dictionary[J]. Science Technology and Engineering,2020,20(21):8683-8689.

[3] ALKUBAISI G A A J,KAMARUDDIN S S,HUSNI H. Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naive Bayes Classifiers[J]. Comput. Inf. Sci.,2018,11(1):52-64.

[4] LIU Z,HUANG D,HUANG K,et al. Finbert:A pre-trained financial language representation model for financial text mining[C]// Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. 2021:4513-4519.

[5] CHEN Y,YAO L B,ZHANG G H,et al. Text Sentiment Orientation Analysis of Multi-Channels CNN and BiGRU Based on Attention Mechanism[J]. Journal of Computer Applications,2020,57(12):2583-2595.

[6] HU X,TANG L,TANG J,et al. Exploiting social relations for sentiment analysis in microblogging[C]// Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. 2013:537-546.

[7] LIU W,ZHANG M. Semi-supervised sentiment classification method based on weibo social relationship[C]// International Conference on Web Information Systems and Applications. Cham:Springer,2019:480-491.

[8] YANG J,ZOU X,ZHANG W,et al. Microblog sentiment analysis via embedding social contexts into an attentive LSTM[J]. Engineering Applications of Artificial Intelligence,2021,97:104048.

[9] JIANG Z L,ZHANG J. Multi-Head Attention Model with User and Product Information for Sentiment Classification[J]. Computer Systems & Applications,2020,29(7):131-138.

[10] WANG Q F,ZHOU M,WANG Z Q,et al. Graph Convolution Network for Sentiment Classification via User and Product Information[J]. Journal of ChineseInformation Processing,2021,35(3):134-142.

[11] ZOU X,YANG J,ZHANG J. Microblog sentiment analysis using social and topic context[J]. PloS One,2018,13(2):e0191163.

[12] GROVER A,LESKOVEC J. node2vec:Scalable feature learning for networks[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:855-864.

[13] NORTHCUTT C,JIANG L,CHUANG I. Confident learning:Estimating uncertainty in dataset labels[J]. Journal of Artificial Intelligence Research,2021,70:1373-1411.

[14] WANG Y W,ZHU C,ZHU J M,et al. User Interest Dictionary and LSTM Based Method for Personalized Emotion Classification[J]. Computer Science,2021,48(S2):251-257.

[15] WANG D,ZHAO Y. Using news to predict investor sentiment:Based on SVM model[J]. Procedia Computer Science,2020,174:191-199.

[16] KUMAR R,KAUR J. Random forest-based sarcastic tweet classification using multiple feature collection[M]// Multimedia Big Data Computing For IoT Applications. Springer,Singapore,2020:131-160.

[17] BIRJALI M,KASRI M,BENI-HSSANE A. A comprehensive survey on sentiment analysis:Approaches,challenges and trends[J]. Knowledge-Based Systems,2021,226:107134.

[18] DING F,SUN X. Negative-emotion Opinion Target Extraction Basedon Attention and BiLSTM-CRF[J]. Computer Science,2022,49(2):223-230.

[19] GUO B,ZHANG C,LIU J,et al. Improving text classification with weighted word embeddings via a multi-channel TextCNN model[J]. Neurocomputing,2019,363:366-374.

[20] GUO Z,ZHU L,HAN L. Research on Short Text Classification Based on RoBERTa-TextRCNN[C]// 2021 International Conference on Computer Information Science and Artificial Intelligence(CISAI). IEEE,2021:845-849.

[21] DEVLIN J,CHANG M W,LEE K,et al. Bert:Pre-training of deep bidirectional transformers for language under-standing[J]. arXiv:1810.04805,2018.

[22] SOARES L D,FRANCO E M C. BiGRU-CNN neural network applied to short-term electric load forecasting[J]. Production,2021,32,e20210087.

[23] WANG K,WANG M Y,LIU X,et al. Event detection by combining self-attention and CNN-BiGRU[J]. Journal of Xidian University,2022,49(5):181-188.

[24] PEROZZI B,AL-RFOU R,SKIENA S. Deepwalk:Online learning of social representations[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014:701-710.

WANG Youwei, born in 1987, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include machine learning, data mining and NLP.

LIU Ao, born in 1997, postgraduate. His main research interests include data mining and NLP.