基于主题增强的递归自编码情感分类研究

朱 引 黄海燕

(华东理工大学信息科学与工程学院 上海 200237)

摘 要 中文文本情感分析旨在发现用户对事物、事件的情感倾向,然而现有研究往往忽视了文本之间的相互联系。提出一种基于主题增强的递归自编码情感分类模型,通过将文本的主题信息融入到递归自编码模型中,使得该模型可以更深层次地考虑文本的内容信息,提高其对文本情感的理解和泛化能力。在 COAE2014 数据集上的实验结果表明,将所提分类模型用于情感分类任务时可获得更优的分类效果,证实了其在实际问题中的适用性与可行性。

关键词 递归自编码,主题模型,情感分类,数据挖掘

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.12.022

Study on Recursive Auto-encoding Sentiment Classification Based on Topic Enhancement

ZHU Yin HUANG Hai-yan

(School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract The emotional analysis of Chinese text aims to discover the emotional tendencies of users to things and events, however, the existing studies often neglect the interrelationships between texts. In light of this, this paper proposed a recursive auto-encoding classification model based on topic enhancement. By incorporating the subject information of the text into the recursive auto-encoding model, this model can further consider the content information of the text and improve the capability to understand the text emotion and generaliza ability. The experimental results on the COAE2014 dataset show that the proposed classification model can achieve better classification performance when used for tasks of sentiment classification, thus verifying its applicability and feasibility in practical problems.

Keywords Recursive auto-encoder, Topic model, Sentiment classification, Data mining

1 引言

随着互联网技术的迅猛发展,网络中产生的文本信息也随之急剧膨胀,如何从海量的网络信息中进行有效的数据挖掘成为人们关注的热点。其中,分析互联网用户产生的网络评论文本,获取网络用户对特定事件的态度,是非常重要的研究方向。网络文本的情感分析技术可以提高企业对用户消费习惯的理解程度,加强对新闻热点捕捉的灵敏度,挖掘用户情感随事件发展的演化规律,为企业或政府的决策提供依据[1-2]。

情感倾向性问题首先由 Bo 等[3]提出,随即获得了各方面的关注,许多学者对此进行了深入研究。目前主流的研究工作通过两类方法实现:基于情感词典的方法和基于机器学习的方法[4]。前者主要采用预定义的情感词典作为辅助分析的策略,中文领域的情感词典包括台湾大学的 NTUSD 和知网发布的 HowNet,该类方法由于其简单易行,在实际生产中得到了广泛应用。Turney[5]使用了基于种子词的非监督学习方法,即选用种子词搜索网页中接近的未知词的互信息,来计算未知词的情感极性,以判断整个文本的情感极性;罗毅等[6]通过构建二级情感词典,对不同级别的情感词进行分层处理,采用 N-gram 获取文本特征,再进行情感倾向性分析。基于情感词的方法通过分析词语的情感极性来判断句子的情感倾向性,其缺陷在于无法考虑整个句子的语义,也无法对句子的结构进行分析,是一种相对浅显的理解方式。另一方面,情感词典的构建需要大量的人工标注,耗时耗力,导致其在跨领域应用时受到极大的限制。基于机器学习方法的原理是将词语样本编码成为词向量形式,继而对其进行语义合成并提取句子的特征,最后使用分类器判断其情感极性。而基于特征提取的方法分为浅层机器学习和深层机器学习。对于浅层机器学习的方法,主要先抽取词语的 n-gram 特征,再使用如支持向量机(Support Vector Machine, SVM)、朴素贝叶斯(Naïve Bayes, NB)等[7]分类器对文本进行情感倾向性判断。然而,浅层线性分类模型由于缺乏捕捉文本结构的能力,其性能难以得到提升。深层机器学习的方法主要是深度学习的方法,如 Socher 等[8]使用了递归自编码网络(Recursive Auto-Encoder, RAE), Tang 等[9]使用了循环神经网络(Recurrent

到稿日期:2017-11-07 返修日期:2018-02-02

朱 引(1993-),男,硕士生,主要研究方向为机器学习、文本挖掘,E-mail:yami.zhu@foxmail.com;黄海燕(1972-),女,博士,副教授,主要研究方向为控制与优化复杂工业过程建模,E-mail:huanghong@ecust.edu.cn(通信作者)。

Neural Network,RNN),Zhang 等[10] 使用了卷积神经网络(Convolutional Neural Network,CNN),他们都通过复杂的神经网络结构对向量化的文本信息进行抽取和整合,以获得文本深层次的特征表达。使用此类模型的优势在于可以在避免复杂的特征提取过程的同时考虑词与词之间的相对顺序,捕捉语意中复杂的语法特征,促使模型对文本的语义理解上升至新的层次。

本文在 RAE 的基础上构建情感分类模型,RAE 模型通过一个类似哈尔曼夫树的树形网络将多维矩阵向量信息编码为一个向量,以此进行语义信息的抽取与综合,相比于 RNN 网络和 CNN 网络,其在处理具有复杂语法规则的语句时具有很大优势,可以更好地理解句子的语法结构。梁军等[11] 提出了基于情感极性转移的递归自编码模型,对中文微博的情感倾向进行了理论分析和实验认证,通过递归神经网络发现与任务相关的特征,并将句子中每个词语的标签关联考虑在内。该模型虽然可以有效地识别数据的结构,但忽视了语句之间的相关性以及语句相对整个文本的内容信息。

对于短文本级别的情感分类任务,常规的处理方法是对单个句子进行语义分析和信息抽取,但对于特定领域的文本信息,文本的内容信息受其背景知识的影响,若信息量不足,将会导致转义的发生。如何充分利用上下文信息以及全局信息,建立句子与句子之间的联系,使得对文本情感倾向的判断不再局限于单个句子,从而提高模型对文本情感倾向的理解能力,是本文研究的重点。为解决此问题,本文提出了基于主题增强的递归自编码情感分类模型(Topic Enhancement Recursive Auto-Encoding,Tp-RAE),该方法利用递归自编码和主题模型,改进其递归自编码网络的重构误差计算方法,再将文本的主题信息融入到递归自编码模型的训练过程中,从而得到文本深层次的语义信息,以此加强模型对文本的理解和泛化能力。实验结果验证了所提算法的有效性。

2　基于主题增强的递归自编码模型

2.1　使用词向量表示句子

在自然语言任务中,词是最基本的单元,One-hot Representation 是最常用的一种表示方式。但这种采用稀疏编码表示文本信息的方式在实际应用中经常会遇到维数灾难的问题,且无法表征语义信息和挖掘数据之间的潜在联系[12]。采用词嵌入的方法将高维数据映射到低维空间中,不仅可以解决维数灾难的问题,而且可以获取各个词之间的关联属性,即意义相近的词在词嵌入的表示上更加接近。本文采用 word2vec[13] 的开源实现对文本数据进行编码表示。word2vec 实现了 CBOW(Continuous Bag-of-Words Model)和 SG(skip-gram)两种结构,并将其用于计算词语的向量表示,本文采用的是 CBOW 结构,即将一个词所在的上下文中的词作为输入,而该词本身作为输出进行模型训练,训练完成后,得到每个词到隐含层对应的每个维度的权重,即对应每个词的词向量。由此,基于该训练过程,将文本内容的处理转化为高维向量空间中的向量运算。

2.2　改进的递归自编码

递归自编码的目的是获取数据隐含的特定结构的表示,以实现对整个文档内容信息的高层次聚集。例如,“爱好”的向量表示为$[0.1,0.3,0.5,0.2]^{\mathrm{T}}$,“学习”的向量表示为$[0.3,0.6,0.2,0.3]^{\mathrm{T}}$,假设“爱好学习”的父节点为 p,“爱好”为第一个子节点 x_1,“学习”是第二个子节点 x_2,则 p 可以由 x_1,x_2 映射得到。

$$p=f(w^{(1)}[x_1;x_2]+b^{(1)}) \tag{1}$$

其中,$[c_1;c_2]\in R^{2n}$;$w^{(1)}\in R^{n\times 2n}$是一个参数矩阵;$b^{(1)}\in R^n$ 为偏置项;编码器函数 $f(\cdot)$选取双曲正切函数 $\tanh(\cdot)$。

由式(1)可以得到父节点的 n 维向量表示$[0.2,0.8,0.3,0.4]^{\mathrm{T}}$。

为验证父节点 p 对子节点 x_1,x_2 的表示准确度,在父节点 p 上建立一个重建层,重建两个子节点,即:

$$[x_1';x_2']=g(w^{(2)}p+b^{(2)}) \tag{2}$$

其中,$[x_1',x_2']\in R^n$ 是重建子节点;$w^{(2)}\in R^{n\times 2n}$为参数矩阵;$b^{(2)}\in R^n$ 是偏执项;解码器 $g(\cdot)$取恒等函数。

使用重构误差描述重建节点 x_1',x_2'与原始节点的相似程度,如式(3)所示:

$$E_{rec}=\|x_1-x_1'\|^2+\|x_2-x_2'\|^2 \tag{3}$$

对整个句子递归使用上述自编码结构,在迭代过程中每次采用贪心算法选取两个节点结构重构误差最小的节点组合,最终得到树形结构的信息网络(见图 1),根节点即为整个句子的向量表示。图 1 中,圆点表示原始节点,方点表示重建节点。

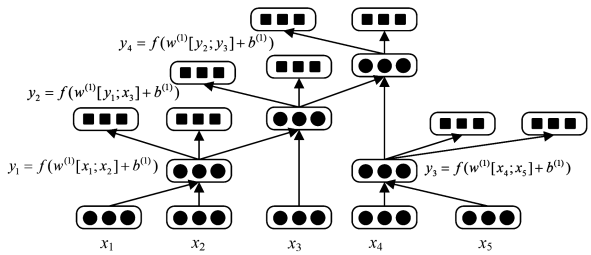

图 1　递归自编码结构图

Fig. 1　Recursive auto-encoding structure diagram

在获得整个句子的向量表示之后,便可以从更高的层次对句子的情感倾向进行分析。

通过分析可以发现,模型在生成树的过程中会出现根节点的左、右两个节点包含的叶子节点数目差距很大的问题,从而导致重建误差计算不平衡,因此需要在计算重建误差时重新考虑子节点的权重。

$$E_{rec}=\frac{n_1}{n_1+n_2}\|x_1-x_1'\|^2+\frac{n_2}{n_1+n_2}\|x_2-x_2'\|^2 \tag{4}$$

其中,n_1 和 n_2 分别表示左、右子节点 x_1,x_2 的节点数。通过节点权重的添加,可以对子孙节点较多的节点赋予更大的损失权重,从而减少误差。

构建递归自编码网络时,为减少网络的重构误差并优化网络性能,使用算法 1 构建该网络。

算法 1　*BuildTree*(*Nodes*)使用贪心算法构建递归自编码网络

输入:输入语句词向量的表示 $v_1,v_2,\cdots,v_n$

输出:网络 T,根节点 T. root

Setp1 输入节点 Nodes={v_1,v_2,…,v_n}

Setp2 While Nodes. size>1 do:

min Error=∞

j=−1

newNode=null

for i in1to Nodes. size−1 do:

p←s($W^{(1)}$[x_1;x_2]+$b^{(1)}$)

[x_1′;x_2′]←f($W^{(2)}$p+$b^{(2)}$)

error ←E_{rec}([x_1;x_2]$_p$,[x_1′;x_2′]$_p$)

if error< min Error then

min Error←error

j←i

newNode←p

End if

newNode. leftChild←Nodes[j]

newNode. rightChild←Nodes[j+1]

Nodes←Nodes−{Nodes[j],Nodes[j+1]}

End for

End while

Setp3 T. root←newNode

算法 1 对词向量集合 *Nodes* 进行遍历,将重构误差最小的元素两两合并,形成递归自编码网络 T。T 中的各叶子节点代表词向量,非叶子节点代表其子节点向量合并后的语义向量。

2.3 LDA 主题模型

LDA(Latent Dirichlet Allocation)主题模型由 Blei 等[14]提出,其中主题表示一个概念、一个方面,表现为一系列相关的词汇,一个主题下的词与该主题之间具有很强的相关性。主题模型的本质是一个 3 层的贝叶斯网络,词和文档通过潜在的主题相关联。本文使用 LDA 模型对文本进行主题划分,以获得单个文本相对于整个语料库的领域信息。LDA 模型假设文本为一定数量隐藏主题下随机产生的词语的集合,每个主题包含特定的词语及其概率分布。

LDA 模型给定语料库参数 α_L 和 β_L,任一文本 w 的概率分布为:

$$p(w|\alpha,\beta)=\prod_{m=1}^{M}\int p(\theta_m|\alpha_L)\left(\prod_{n=1}^{N}\sum_{z_{m,n}}p(z_{m,n}|\theta_{m,n})p(w_{m,n}|z_{m,n},\beta_L)\right)\mathrm{d}\theta_m \tag{5}$$

ThemesGet(w)LDA 模型的主题生成算法如算法 2 所示。

算法 2 *ThemesGet*(w)LDA 模型的主题生成算法

输入:训练语料 D,主题数 G,超参数 α_L 和 β_L

输出:文本属于各个主题的概率分布 S={r_1,r_2,r_3,…,r_g},r 表示各个主题对应的概率

Step1 生成主题参数 φ_k,φ_k 服从参数为 Dirichlet(β_L)的分布。

Step2 对每一个文本 w_m(1≤m≤M)进行采样:

1)采样 θ_m,θ_m 服从 Dirichlet(α_L)分布。

2)采样文档长度 N,N 服从 Poiss(ξ)分布。

3)对文档 w_m 中的第 n(1<n<N)个词进行采样:

①选择主题 $z_{m,n}$,$z_{m,n}$服从 Multinomial(θ_m)分布;

②生成 $w_{m,n}$,$w_{m,n}$服从 Multinomial($\phi_{z_{m,n}}$)分布。

4)语料库参数 α_L 和 β_L 由 Gibbs 采样[15]更新。

Step3 生成文本-主题概率分布以及主题-词概率分布。

式(5)及算法 2 中各符号的定义如表 1 所列。

表 1 符号定义说明

Table 1 Description of symbols and definitions

参数	定义
θ_m	第 m 篇文档的主题概率分布
φ_k	主题 k 的词概率分布
$w_{m,n}$	第 m 篇文档中的第 n 个单词
$z_{m,n}$	第 m 篇文档中的第 n 个主题
α_L	θ 的超参数
β_L	φ 的超参数
M	文本数
N	词数
G	主题数

2.4 融合主题信息的递归自编码分类模型

传统的递归自编码模型仅考虑对单个句子信息的融合,忽略了文本在特定语境下的词意转变,而采取将主题模型融合进递归自编码模型的方法,可提高模型在不同情境下对文本情感的分析效果和判断能力,在当前互联网时代文本内容类型繁多的背景下具有极大的实用价值。

通过递归自编码模型得到二叉树中每个节点的 n 维向量表示,根节点为整个句子的 n 维向量表示。为得到整条语句的情感倾向,在根节点处添加一层输出层:

$$s(p;\theta)=h(w^{(3)}p+b^{(3)}) \tag{6}$$

其中,$w^{(3)}\in R^{(u\times n)}$($h$ 是情感种类的数量,本文仅正负两类,故 $u=2$);$b^{(3)}\in R^k$ 为偏置项;激活函数 $h(\cdot)$选择 $softmax(\cdot)$函数。加入输出层的递归自编码模型如图 2 所示,其中三角形节点表示输出层。

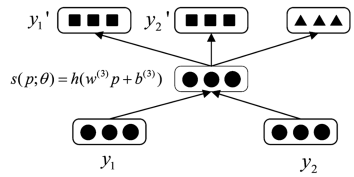

图 2 加入输出层的递归自编码模型

Fig. 2 Recursive auto-encoding model with output layer added

获取文本的主题分布 r^t 之后,为实现主题信息与递归自编码结构的融合,以主题信息为句子特征来作为分类依据,如图 3 所示,其中五角星节点为主题信息。

$$s(p;\theta)=h(w^{(3)}p+b^{(3)}+w^{(4)}r) \tag{7}$$

其中,$w^{(4)}\in R^{(i)}$(i 为主题的类别);r 为主题分布向量。

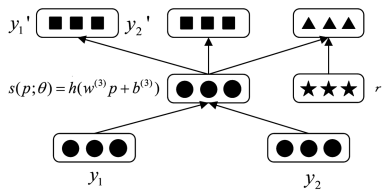

图 3 加入主题信息的递归自编码模型

Fig. 3 Recursive auto-encoding model with theme information

本文采用交叉熵作为代价函数(假定 t_i 表示句子情感倾向性为 T_i 的概率),交叉熵 E_{ce} 为:

$$E_{ce}=-\sum_{k=1}^{K}t_k\log s_k(p;\theta) \tag{8}$$

综合重建误差函数和标签误差函数,可得到模型的代价函数:

$$J=\frac{1}{N}\sum_{(c,t)\in s}E(c,t;\theta)+\frac{k}{2}\|\theta\|^2 \tag{9}$$

其中,$E(c,t;\theta)=\alpha E_{rec}+(1-\alpha)E_{ce}$($\alpha$ 为交叉熵权重比),S 为数据集,x 表示句子,t 表示该句子对应的情感标签,k 为惩罚项系数。

为求解上述优化问题,首先计算目标函数 J 的梯度,即关于参数 θ 的偏导数,然后采用随机梯度下降法[16]求解该函数的最优解。

算法 3　Tp-RAE 模型训练

输入:训练语料 S 及对应标签 t

输出:参数 $\theta=\{w^{(1)},w^{(2)},w^{(3)},w^{(4)},A,b^{(1)},b^{(2)},b^{(3)}\}$

Step1　使用 word2vec 初始化训练语料的词向量表示

Step2　初始化 LDA 模型

Step3　初始化参数 $\theta=\{w^{(1)},w^{(2)},w^{(3)},w^{(4)},A,b^{(1)},b^{(2)},b^{(3)}\}$

Step4　While 不收敛 do:

　　$\nabla J=0$

　　for all $\langle s,t\rangle\in S$ do

　　r←ThemesGet(x)

　　p←BuildTree(Nodes)

　　$\nabla J_i\leftarrow\partial J(s,t)/\partial\Theta$

　　$\nabla J\leftarrow\nabla J+\nabla J_i$

　　End for

　　$\Theta\leftarrow\frac{1}{N}\nabla J+\lambda\theta$

End while

3　实验分析

3.1　数据来源

本文采用的评测数据集为 COAE2014 微博数据集,共 40000 条数据,其中包含 20000 条已标注的数据(正负比例为 3∶2),用于情感分类任务。

表 2　COAE2014 数据集样例

Table 2　COAE2014 dataset samples

积极	消极
这个车真不错,我很喜欢,很享受!	什么东西啊,这样的垃圾也敢拿出来卖?
这家饭店的味道很地道,我吃了好多次了,强烈推荐	绝对不会再去第二次,服务差,味道一般,上菜速度慢
这个广告做得可以啊,有意思	烦死了,各种广告没完没了了还!
漫长的等待,终有收获的一天,今日正是此时	哎,这日子一天天的什么时候是个头哇

此外,选取 300 万篇网络收集的微信公众号文章(多领域,属于中文平衡语料)作为训练语料,总词数达到 420 亿,用于 word2vec 模型和 LDA 模型的预训练,以获取质量更优的词向量和精度更高的主题分布。

3.2　文本主题的获取与词向量的训练

本文中主题模型的参数设置请参考文献[14]中的选定方法,超参数的设置为 $\alpha_L=1.25$,$\beta_L=0.01$,$G=40$。

文本主题获取的核心目标是将每条语句依照其与整个文本库的关系划分其类别领域,为此先进行相应的文本预处理。首先,采用中科院计算技术研究所开发的中文分词工具 NLPIR[17]进行中文分词,然后去除停用词(主要包括常用词和标点符号),仅保留句子主干,减少无关词汇的影响,提高输入语料的准确度。

训练 LDA 模型可以获得主题-文本矩阵及每个文本的主题分布矩阵,表 3 列出了前 4 个主题-文本示例。

表 3　主题-文本示例

Table 3　Topic-text examples

主题	对应文本
Topic 1	朋友 男朋友 游戏 女生 回家 双排 凌晨 记录 开黑 冻伤 两点半 质问 好友 下路 网吧 不用 吐槽
Topic 2	孩子 父母 学习 习惯 最好 怪罪 教育 道理 老师 建议 知识 帮助 办法 伟人 家庭教师 成绩 理解
Topic 3	吃饭 下馆子 好吃 地道 再来 多次 服务 难吃 爆炸排队 等了 新鲜 慢 快 上菜 一般 品种 多 少
Topic 4	难过开心 伤心 郁闷 难受 欣喜 不错 垃圾 心烦 一般 地道 哈哈 呵呵 东西 哇 哎 什么 干嘛

第一个主题与交际相关,第二个主题与家庭教育相关,第三个主题与餐饮相关,第四个主题与个人情感相关。

在此基础上,获得测试集各条语句的主题概率分布示例,如表 4 所列。

表 4　主题概率分布示例

Table 4　Topic probability distribution examples

文本	对应主题概率分布
1	[0.45562,0.06584,0.03614,0.00467,0.06412,0.071…]
2	[0.55125,0.01322,0.01126,0.31245,0.04623,0.034…]
3	[0.02654,0.10546,0.05569,0.61247,0.01237,0.012…]
4	[0.00446,0.16574,0.34561,0.01235,0.01453,0.014…]
5	[0.06123,0.41322,0.15472,0.03356,0.04628,0.101…]

词向量训练结果质量的好坏对实验结果有着较大影响,为此,本文参照文献[18]中的方法进行无监督的词向量训练,设置词向量维度为 256。

3.3　可调参数选择

模型训练过程中的词向量长度、主题个数、迭代次数、惩罚项系数和交叉熵权重比等参数的选择对分类器的性能均可产生不同程度的影响。本实验涉及的词向量长度和主题个数参数已在前文进行了相应设定。我们考查在其他参数恒定的情况下,迭代次数对模型性能表现的影响,主要观测训练集准确率和测试集准确率随着迭代次数的增加而发生变化的情况,实验结果如图 4 所示。

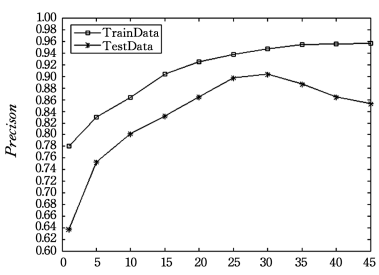

图 4　改变迭代次数对准确率的影响

Fig. 4　Impact of change of number of iterations on accuracy

由实验结果可知,随着迭代次数的增加,模型对训练集数

据的拟合效果越好，但经 30 次迭代后，测试集的准确率开始下降，出现过拟合情况，并导致模型的泛化能力减弱，因此综合考虑后设定迭代次数为 30。

以下测试实验针对惩罚项系数 α 及重建误差与交叉熵误差权重比 k 两个核心参数，研究其对模型性能的影响。设置词向量维度为 256，主题个数为 40，迭代次数为 30，改变 α 和 k 的参数值，观察其对模型分类性能的影响，实验结果如图 5 所示。实验数据采用 5000 句正负平衡的带有情感倾向的句子，由十折交叉验证得到准确率的平均值。

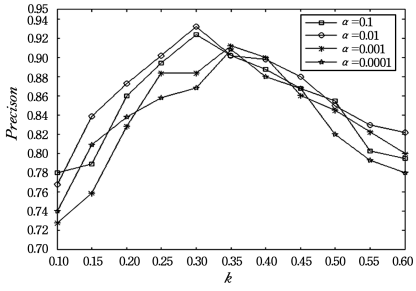

图 5 不同 α 和 k 对准确率的影响

Fig. 5 Impact of different α and k on accuracy

由惩罚系数对模型性能影响的折线图可看出，当惩罚系数为 0.01 时，模型的综合性能最好。另外，观测重建误差系数对模型影响的折线图可发现，当惩罚系数为 0.01，且重建误差与交叉熵误差的权重比为 0.3 时，算法的性能最佳，其后随着权重比的增加，模型性能呈下降的趋势。

综合多次实验的结果，模型各参数值的设置如表 5 所列。

表 5 可调参数的设置

Table 5 Settings of adjustable parameter

可调参数	参数值
词向量维度	256
迭代次数	30
正则惩罚系数	0.01
重建误差与交叉熵误差权重比	0.3
设定主题个数	40

3.4 对比实验与评价指标

为验证本文提出的模型的有效性和正确性，使用以下模型进行对比实验：1）基于词袋模型的朴素贝叶斯方法（NB）[7]；2）SVM 分类器[7]；3）平均词向量方法（VecAvg）；4）原始的递归自编码网络（RAE）[8]；5）带情感极性转移的递归自编码网络（MV-RAE）[11]；6）卷积神经网络（CNN）[10]；7）本文提出的主题增强递归自编码网络（Tp-RAE）。通过该实验对比本文提出的主题增强模型相对其他机器学习模型的性能优劣。

评价指标采用准确率（Precision）、召回率（Recall）和 F1 值（F1-measure）：

$$Precision=\frac{correct\_n}{predict\_n} \tag{10}$$

$$Recall=\frac{correct\_n}{labeled\_n} \tag{11}$$

$$F1\text{-}measure=\frac{2\times Precision\times Recall}{(Precision+Recall)} \tag{12}$$

其中，$predict\_n$ 表示模型预测属于某个类别的模型实例数目；$labeled\_n$ 表示实际属于某个类别的实例总数；$correct\_n$ 表示模型预测正确的实例数目。为了更加精细地分析模型的实验性能，对正例和负例分别进行分析，并使用“+”和“−”进行标记。

3.5 实验结果与分析

按照表 5 设定实验的各参数，在相同条件下对比各个模型的性能表现，结果如表 6 所列。

表 6 各模型实验结果的比较

Table 6 Comparison of experimental results of models

模型	*Precision*+ *Precision*−	*Recall*+ *Recall*−	*F1-measure*+ *F1-measure*−
NB	83.42 79.69	80.21 81.10	81.78 80.53
SVM	86.36 83.21	82.90 86.55	84.59 84.85
VecAvg	85.41 84.31	81.25 86.35	82.85 85.31
CNN	91.94 83.25	88.83 84.19	90.36 83.72
RAE	90.38 85.25	85.50 87.09	87.87 86.16
MV-RAE	90.56 86.92	88.90 88.21	89.73 87.56
Tp-RAE	**93.45** **88.21**	**90.16** **89.52**	**91.76** **88.86**

表 6 列出了不同模型在相同测试集和验证集上的实验结果，从实验结果可以看出：

1）VecAvg 模型采用平均词向量的方法，获得了与 SVM 模型相近的效果，且优于 NB 模型，表明词向量的特性对文本分析具有较大意义。而相比于 SVM，NB 和 VecAvg 模型，基于深度学习的方法在模型判别能力上取得了很大的提升，说明了在文本情感分析中依据词语的前后顺序对句子进行结构分析与识别的必要性。

2）以上模型的实验结果均出现了对正例（积极情感）的判别能力略优于负例（消极情感）的现象，说明消极情感相对于积极情感在语义表示上更加模糊，更容易出现判断错误。通过对发生判断错误的实例进一步分析可以发现，积极情感的语句更偏向于口语话的表达方式，而消极情感的语句更多地带有较为完整的语法结构。使用 CNN 模型对正例识别的 $F1$ 值高于 RAE 模型，但对负例识别的 $F1$ 值比 RAE 模型低，说明 RAE 模型对带有复杂语法结构的句子结构的识别能力略优于 CNN 模型，而对于口语化的句子的判别能力弱于 CNN 模型。MV-RAE 模型由于添加了极性转移机制，进一步强化了 RAE 模型对句子结构的理解能力，提升了模型对文本情感的判别能力，但其对积极情感的判别能力还是弱于 CNN 模型。

3）对于 Tp-RAE 模型，在融合了主题信息之后，相比于 CNN 模型，对正例识别的 $F1$ 值提升了 1.40%，对负例识别的 $F1$ 值提升了 5.14%，克服了 RAE 与 MV-RAE 模型在对正例判别能力上弱于 CNN 模型的缺陷，也进一步提高了模型对负例的判别能力。这说明通过以主题模型构建各单句之间的相互联系并将其融入到 RAE 模型的方法对提升模型的

判别能力具有显著效果。

结束语　本文提出一种将主题模型与递归自编码模型相融合的情感分类模型,该模型将文本的主题信息融入到改进后的递归自编码模型中,在不同维度上对文本信息进行更深层次的分析和理解,显著提高了模型在互联网应用环境中对复杂文本的情感分类精度。实验结果表明了该模型的有效性。该模型下一步的发展方向应为进一步细化主题划分,使其对应到具有实际意义的不同领域,以获得更好的解释性。此外,如何提高模型的训练速度和预测精度是今后需要研究的另一个内容。

参 考 文 献

[1] ZHANG Z Q, YE Q, LI Y J, et al. Literature review on sentiment analysis of online product reviews [J]. Journal of Management Sciences in China, 2010, 13(6): 84-96. (in Chinese)
张紫琼,叶强,李一军,等. 互联网商品评论情感分析研究综述[J]. 管理科学学报,2010,13(6):84-96.

[2] 赵军,许洪波,黄萱菁. 中文倾向性分析评测技术报告[EB/OL]. http://www.doc88.com/p-179806395884.html.

[3] BO P, LEE L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales[C]// Meeting on Association for Computational Linguistics. 2005: 115-124.

[4] ZHOU S C, QU W T, SHI Y Z, et al. Overview on sentiment analysis of Chinese microblogging[J]. Computer Applications and Software, 2013, 30(3): 161-164. (in Chinese)
周胜臣,瞿文婷,石英子,等. 中文微博情感分析研究综述[J]. 计算机应用与软件,2013,30(3):161-164.

[5] TURNEY P D. Thumbs up or thumbs down: semantic orientation applied to unsupervised classification of reviews[C]// Meeting on Association for Computational Linguistics. 2002: 417-424.

[6] LUO Y, LI L, TAN S B, et al. Sentiment analysis on Chinese Micro-blog corpus[J]. Journal of Shandong University (Natural Science), 2014, 49(11): 1-7. (in Chinese)
罗毅,李利,谭松波,等. 基于中文微博语料的情感倾向性分析[J]. 山东大学学报(理学版),2014,49(11):1-7.

[7] WANG S, MANNING C D. Baselines and bigrams: simple, good sentiment and topic classification[C]// Meeting of the Association for Computational Linguistics: Short Papers. Association for Computational Linguistics, 2012: 90-94.

[8] SOCHER R, PENNINGTON J, HUANG E H, et al. Semi-supervised recursive auto-encoders for predicting sentiment distributions[C]// Conference on Empirical Methods in Natural Language Processing. DBLP, 2011: 151-161.

[9] TANG D, QIN B, LIU T. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification[C]// Conference on Empirical Methods in Natural Language Processing. 2015: 1422-1432.

[10] ZHANG L, CHEN C. Sentiment Classification with Convolutional Neural Networks: An Experimental Study on a Large-Scale Chinese Conversation Corpus[C]// International Conference on Computational Intelligence and Security. IEEE, 2017: 165-169.

[11] LIANG J, CHAI Y M, YUAN H B, et al. Deep Learning for Chinese Micro-blog Sentiment Analysis[J]. Journal of Chinese Information Processing, 2014, 28(5): 155-161. (in Chinese)
梁军,柴玉梅,原慧斌,等. 基于深度学习的微博情感分析[J]. 中文信息学报,2014,28(5):155-161.

[12] TANG D, WEI F, YANG N, et al. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification[C]// Meeting of the Association for Computational Linguistics. 2014: 1555-1565.

[13] LEVY O, GOLDBERG Y. Neural word embedding as implicit matrix factorization[J]. Advances in Neural Information Processing Systems, 2014, 3(4): 2177-2185.

[14] BLEI D M, NG A Y, JORDAN M I. Latent dirichletallocation [J]. Journal of Machine Learning Research, 2003, 3(6): 993-1022.

[15] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 1011(1): 5228.

[16] ZHANG S, ZHANG C, YOU Z, et al. Asynchronous stochastic gradient descent for DNN training[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 6660-6663.

[17] XIAO H, MAO M Y. NLPIR Chinese character segmentation system based Chinese character segmentation tool: CN 106354714 A[P]. 2017. (in Chinese)
肖红,毛明扬. 一种基于 nlpir 中文分词系统的中文分词工具: CN 106354714 A[P]. 2017.

[18] ORDENTLICH E, YANG L, FENG A, et al. Network-Efficient Distributed Word2vec Training System for Large Vocabularies [J/OL]. https://arxiv.org.abs/606.08495.

(上接第 136 页)

[35] BAGGA A, BALDWIN B. Algorithms for scoring coreference chains[C]// The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference. 1998: 563-566.

[36] RECASENS M, HOVY E. BLANC: Implementing the Rand index for coreference evaluation[J]. Natural Language Engineering, 2011, 17(4): 485-510.

[37] LUO X. On coreference resolution performance metrics [C]// HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Vancouver, British Columbia, Canada. DBLP, 2005: 25-32.

[38] PRADHAN S, LUO X, RECASENS M, et al. Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation[C]// Meeting of the Association for Computational Linguistics. 2014: 30.