Journal of Computer Applications
计算机应用, 2020, 40(2): 473 - 478
ISSN 1001-9081
CODEN JYIIDU
2020-02-10
http://www.joca.cn

文章编号:1001-9081(2020)02-0473-06
DOI:10.11772/j.issn.1001-9081.2019101768

目标依赖的作者身份识别方法

李　扬[1],张　伟[1*],彭　晨[2]
(1. 华东师范大学 计算机科学与技术学院,上海 200062; 2. 中国科学院电子学研究所苏州研究院,江苏 苏州 215123)
(*通信作者电子邮箱 zhangwei.thu2011@gmail.com)

摘　要:作者身份识别任务旨在判断一篇文档的作者,但目前已有的作者身份识别方法都是目标独立的,意味着这些方法在预测作者身份时假设没有任何限定条件,这与实际情况不相符合。为了解决限定条件下的作者身份识别问题,提出了一种目标依赖的作者身份识别方法TDAA。首先,使用用户评论对应的商品ID作为限定信息;其次,为了使文本建模过程更加具有普适性,使用BERT提取预训练的评论文本特征;然后,使用卷积神经网络(CNN)进行深层次的文本特征提取;最后,为了将两种不同的信息融合起来,讨论了两种不同的融合方式。在亚马逊电影评论(Amazon Movie_and_TV)和CD评论(CDs_and_Vinyl_5)两个数据集上的实验结果表明,所提出的方法在精确率评价指标上较对比方法提高了4%~5%。

关键词:作者身份识别;目标依赖;卷积神经网络;信息融合;预训练语言模型
中图分类号:TP391.1　　**文献标志码**:A

Target-dependent method for authorship attribution

LI Yang[1], ZHANG Wei[1*], PENG Chen[2]
(1. *School of Computer Science and Technology, East China Normal University, Shanghai* 200062, *China*;
2. *Institute of Electronics, Chinese Academy of Sciences, Suzhou Jiangsu* 215123, *China*)

Abstract: Authorship attribution is the task of deciding who is the author of a particular document, however, the traditional methods for authorship attribution are target-independent without considering any constraint during the prediction of authorship, which is inconsistent with the actual problems. To address the above issue, a Target-Dependent method for Authorship Attribution (TDAA) was proposed. Firstly, the product ID corresponding to the user review was chosen to be the constraint information. Secondly, Bidirectional Encoder Representation from Transformer (BERT) was used to extract the pre-trained review text feature to make the text modeling process more universal. Thirdly, the Convolutional Neural Network (CNN) was used to extract the deep features of the text. Finally, two fusion methods were proposed to fuse the two different information. Experimental results on Amazon Movie_and_TV dataset and CDs_and_Vinyl_5 dataset show that the proposed method can increase the accuracy by 4%-5% compared with the comparison methods.

Key words: authorship attribution; target-dependent; Convolutional Neural Network (CNN); information fusion; pre-trained language model

0　引言

作者身份识别(Authorship Attribution)的主要思路是将文档中隐含的作者无意识的写作习惯通过某些特征表示出来,凸显作品的文学特征及写作风格,以确定匿名文本的作者。作者身份识别可以在许多实际问题中发挥作用,比如,可以帮助历史学家从一些候选的作者中推测出文献中的一段话的作者,可以帮助网络执法者鉴别出发布不良信息的用户等。

目前现有的许多方法都是根据不同的文本特点来设计文本特征,这些特征包括文本的单词级别的n元语言模型(word n-gram)和字符级别的n元语言模型(character n-gram)、文本的主题分布、文本的语法和语义特征等[1]。根据不同的文本特征,可以设计使用不同的方法。支持向量机(Support Vector Machine, SVM)、随机森林(Random Forests, RF)和隐式狄利克雷(Latent Dirichlet allocation, LDA)主题模型等都是解决作者身份识别问题常用的方法[2-3]。近年来,深度学习技术在文本表示方面取得了很好的效果,因此卷积神经网络(Convolutional Neural Network, CNN)和长短期记忆(Long Short-Term Memory, LSTM)神经网络也被用来解决作者身份识别问题[4-7]。

但是,在实际问题中,作者身份识别问题通常被限定在某一范围内。比如,法官想要通过一篇文档来确定犯罪嫌疑人,而且证据表明犯罪嫌疑人是一个年龄在40和50岁之间的女性。如果利用现有的方法,即仅利用文档进行作者身份识别,判断的结果可能是一个年龄在20和30岁之间的男性,显然这个结论是错误的。原因在于现有的方法没有利用作者依赖的信息,即在开放域上利用文档预测作者,忽略了依赖信息的重要性。

收稿日期:2019-09-18;**修回日期**:2019-10-18;**录用日期**:2019-10-24。　**基金项目**:国家自然科学基金青年基金资助项目(61702190)。
作者简介:李扬(1994—),男,山西运城人,硕士研究生,主要研究方向:数据挖掘; 张伟(1988—),男,重庆人,副教授,博士,主要研究方向:用户数据挖掘、自然语言处理; 彭晨(1986—),男,江苏常州人,副研究员,博士,主要研究方向:空间信息处理。

通常来讲，作者身份的限定条件可能是某种离散的属性，比如性别、年龄和婚姻状况等，但由于上述信息包含用户隐私，获取成本比较高，因此本文使用易于获得的亚马逊商品评论作为实验数据集。在商品评论数据集中，用户选择购买某种商品与其年龄、性别、收入情况和兴趣爱好等方面息息相关，可以在某种程度上反映用户的属性信息。因此，选择使用商品ID作为限定条件，提出了一种目标依赖的作者身份识别算法，可以避免复杂的文本特征设计，同时有效利用依赖信息进行作者身份预测。本文的主要工作如下：1）提出了一种使用目标依赖信息解决作者身份识别问题的算法，探索了两种不同的信息融合方式，并在亚马逊电影评论和CD评论数据集上证明了这两种融合方式可以用于解决限定条件下的作者身份识别问题。2）提出了一种使用BERT（Bidirectional Encoder Representation from Transformer）提取文本特征的方法，避免了针对不同类型的数据集设计不同文本特征的复杂性，并用实验验证了这种方法优于目前已有的文本特征提取方法。

1 相关研究

1.1 文本建模与作者身份识别问题

针对文本建模，许多工作研究了不同的文本特征提取方法，大致可以分为以下几种特征：1）词汇级别的特征，包括单词长度、文档长度、文档中词汇的丰富程度和错误词汇数量等；2）字符级别的特征，包括字符的类别（字母或者数字）和字符n元模型等；3）语法特征，包括词性和句子结构等；4）语义特征，包括语义依赖分析和功能分析等。考虑到评论数据的生成受到用户和商品的同时影响，Zhang等[8-9]利用主题模型和矩阵分解模型同时对用户评论和商品进行建模。

之前的研究中，作者身份识别大致分为两种思路[9]。第一种思路是基于相似度的方法。这种方法的做法是将作者的所有文本信息拼接为单个文档，将单个文档的特征作为该作者的特征。对于一条新的文本，通过比较该文本与已知文本的相似度，将相似度最高的已知文本的作者作为未知文本的预测结果。基于这个思路，Seroussi等[10]将一个作者发布的所有文本组成一条文本，然后使用主题模型的方法从文本中提取出文本的主题分布作为作者的特征，对于一条新的文本，计算其主题分布与作者特征之间的Hellinger距离，距离最小的作者作为预测结果。另一个思路是基于分类的思想，大多数的研究都基于此方法。Schwartz等[1]使用单词级别n元模型和字符级别的n元模型作为文本的特征，这样可以将一段文本信息映射为二值的特征向量，然后使用支持向量机（SVM）进行分类。Zhang等[11]除了使用n元模型，还加入了语义特征，具体操作是：对文本进行语法分析得到文本的语法树，并对树中的每一个节点进行编码，将这些节点的编码作为文本的语义特征；接下来，将文本的语义信息和内容信息作为两个不同的通道使用CNN进行分类。

以上方法与本文提出的方法最大不同之处有两点：一是需要复杂的特征设计与处理；二是忽视了在实际问题中作者身份存在限定条件这一特征，仅利用文档信息进行作者身份判定。这些方法可能得出与限定条件相悖的结论。

1.2 预训练语言模型

在自然语言处理领域，词向量被广泛应用在多种任务中，比如，文本分类、问答系统以及文本检索等。Word2Vec[12]是目前最常用的词嵌入模型之一，它实际上是一种浅层的神经网络模。常用的模型包括根据上下文出现的词语来预测当前词语生成概率的连续词袋（Continues Bag of Words，CBOW）模型和根据上下文的词语预测当前词语生成概率的跳字（Skip-gram）模型。但是由于Word2Vec输入的上下文有限，使得其无法解决多义词的情况。BERT[13]是一种新的语言表示模型。不同于Word2Vec，BERT使用文本内容的左、右语境进行预训练得到文本的深度双向表征，因此，BERT通过添加额外的一层神经网络进行微调，就可以在多种任务上达到最优的效果。本文将通过BERT模型得到的文档向量作为文本特征。

1.3 多模态学习

多模态学习是一种利用多种数据类型进行学习的方式。多模态学习需要利用好各种数据类型的内在关系，使得不同的数据类型可以提供有效且互补的信息。信息融合首要的问题是解决融合发生的位置，一般可以分为三种，分别是特征多模态融合（feature multimodal fusion）、决策多模态融合（decision multimodal fusion）和混合多模态融合（hybrid multimodal fusion）[14]。特征多模态融合是对不同的特征在进入模型之前进行融合；决策多模态融合方式需要在特征输入模型之前保持相互独立，而在各自通过模型之后进行融合；混合多模态融合既在输入之前进行融合又需在通过模型之后进行融合。其次需要解决融合内容这一问题，不同的数据类型有不同的表示方式，如何选择数据的表示类型是解决这一问题的关键。

本文探索了两种信息融合方式在作者身份识别问题中的应用，分别代表了前期融合和后期融合：前期融合和后期融合的区别在于融合发生的位置不同，前期融合在输入模型之前对不同的数据类型进行融合，后期融合是对不同的数据类型在输入模型之后进行融合。

2 目标依赖作者身份识别算法

2.1 问题描述与符号定义

设数据集中包含的用户集合为$U=\{u_1,u_2,\cdots,u_n\}$，评论集合为$R=\{r_1,r_2,\cdots,r_m\}$，商品集合为$D=\{d_1,d_2,\cdots,d_q\}$，其中，n、m和q分别为用户数量、评论数量和商品数量。目标依赖的作者身份识别算法是根据作者产生的评论和对应评论的商品(r_i,d_i)从候选集U中找到对应的评论的作者u_i。本文使用的符号表述见表1。

表1 符号定义

Tab. 1 Symbol definition

符号	描述	符号	描述
L	文本最大长度	$\boldsymbol{B}\in\mathbb{R}^{\lvert C\rvert\times d}$	词向量表
m	卷积核的数目	$\boldsymbol{E}\in\mathbb{R}^{s\times L\times d}$	文档向量
σ	激活函数	$\boldsymbol{P}\in\mathbb{R}^{b\times d}$	商品ID向量表
$Conv2D$	二维卷积操作	$\boldsymbol{p}$	商品ID向量
d	词向量维度		

2.2 预训练文档特征提取

为了避免复杂的特征设计，本文采用BERT提取预训练的词向量，如图1所示。具体地，对于用户的评论文本，首先将其分词后得到$Tok_1,Tok_2,\cdots,Tok_n$，通过BERT预训练模型，

可以得到词向量表 $\boldsymbol{B}$。查询词向量表 $\boldsymbol{B}$ 后得到 $\boldsymbol{Vec}_1, \boldsymbol{Vec}_2, \cdots, \boldsymbol{Vec}_n$ 分别对应 $Tok_1, Tok_2, \cdots, Tok_n$ 的向量表示。文档向量表示是将 $\boldsymbol{Vec}_1, \boldsymbol{Vec}_2, \cdots, \boldsymbol{Vec}_n$ 拼接。如图1所示。

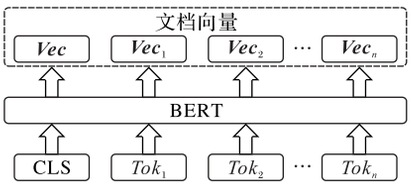


图1　预训练的文档特征提取

Fig. 1　Pre-trained document feature extraction

2.3　基于文档向量的卷积神经网络

卷积神经网络的输入为用户评论文本。首先将用户评论文本通过BERT得到对应的文档向量。设定文本的最大长度为 L,对于不足最大长度的作填充处理,超过最大长度的作截断处理。因此,输入即为文档向量 $\boldsymbol{E}$。

在得到输入文档向量后,需要对文档向量进行二维卷积操作。首先是一个卷积核 $\boldsymbol{H} \in \mathbb{R}^{d \times w}$ 作用于输入的文档向量,其中 w 为卷积核的宽度。由此产生的特征矩阵 $\boldsymbol{O}$,经过激活函数 σ 后加上偏置项 b 可以得到文档向量经过卷积处理的特征:

$$\boldsymbol{O} = \boldsymbol{H} \cdot \boldsymbol{C}[i; i + w - 1] \tag{1}$$

$$g_{\text{text}} = \sigma(\boldsymbol{H} \cdot \boldsymbol{C}[i; i + w - 1] + b) \tag{2}$$

式(1)和式(2)定义为 $Conv2D$。

最大池化作用于 g:

$$y_k = \max_i g_{\text{text}}[i];\ k = 1, 2, \cdots, m \tag{3}$$

其中,m 是特征层的个数。最大池化保证了 y_k 中含有每个特征层中最重要的信息。将所有的 y_k 拼接起来即可得到文本特征:

$$\boldsymbol{f}_{\text{text}} = [y_1, y_2, \cdots, y_m] \tag{4}$$

在得到文本特征之后,需要使用Softmax层进行分类。Softmax层的输入为上述文本特征,为了得到模型对每个用户的预测分数,需要将 $\boldsymbol{f}_{\text{text}}$ 与权重矩阵 $\boldsymbol{W} \in \mathbb{R}^{n \times m}$ 相乘:

$$c = \boldsymbol{W} \times \boldsymbol{f}_{\text{text}} \tag{5}$$

经过Softmax函数归一化后可得该文档属于第 i 个作者的概率:

$$p(i|x) = \mathrm{e}^{o_i} \Big/ \sum_{j=1}^{n} \mathrm{e}^{o_j} \tag{6}$$

因此,Softmax函数的输出为 $\boldsymbol{S} = [s_1, s_2, \cdots, s_n]$,其中 $s_i = p(i|x)$,$\boldsymbol{S} \in \mathbb{R}^{1 \times n}$,$\sum_{i=1}^{n} s_i = 1$。模型结构如图2所示。

2.4　加入目标依赖信息的卷积神经网络

本文选择的依赖信息为作者评论对应的商品ID。商品ID对于预测评论的作者的作用是可解释的:一个用户更倾向于购买自己喜欢的商品,这包括商品的类别、价格和美观程度等。商品ID是一个离散的数据,本文的目的是将此ID转化成为一个稠密的向量,使得这个向量可以从某种意义上表示该商品的各种特征。

接下来,对前期融合和后期模态两种融合方式进行详细的介绍。

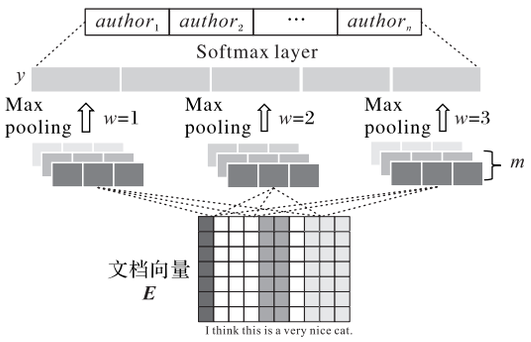


图2　基于文档向量的卷积神经网络

Fig. 2　CNN based on document vector

2.4.1　前期融合

前期融合将商品ID向量与文档向量在输入到卷积神经网络之前进行融合。具体地,将商品ID通过查商品ID向量表 $\boldsymbol{P}$ 得到对应的向量表示 $\boldsymbol{p}$。然后将该向量与文档向量进行拼接,将拼接后的向量输入到卷积神经网络中:

$$g = Conv2D([\boldsymbol{E}, \boldsymbol{p}]) \tag{7}$$

最大池化作用于 g:

$$y_k = \max_i g[i];\ k = 1, 2, \cdots, m \tag{8}$$

拼接 y_k:

$$\boldsymbol{f} = [y_1, y_2, \cdots, y_m] \tag{9}$$

最后使用Softmax层进行分类:

$$\boldsymbol{s} = \text{softmax}(\boldsymbol{f}) \tag{10}$$

模型的结构如图3所示。

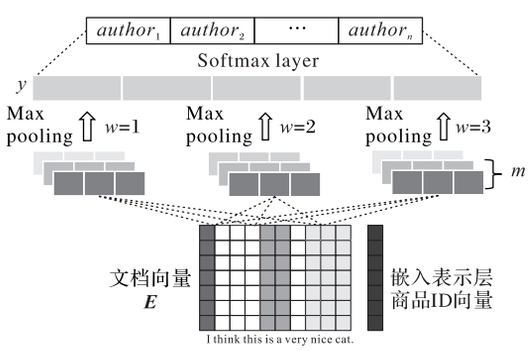


图3　前期融合模型

Fig. 3　Earlier-stage fusion model

2.4.2　后期融合

后期融合将商品ID向量与经过卷积操作之后的文档向量进行拼接。具体地,将文档向量经过卷积操作后与商品ID向量进行拼接,商品ID向量并没有参与卷积与最大池化操作:

$$g = Conv2D(\boldsymbol{E}) \tag{11}$$

最大池化作用于 g:

$$y_k = \max_i g[i];\ k = 1, 2, \cdots, m \tag{12}$$

拼接 y_k:

$$\boldsymbol{f} = [y_1, y_2, \cdots, y_m] \tag{13}$$

使用Softmax函数分类:

$$\boldsymbol{s} = \text{softmax}([\boldsymbol{f}, \boldsymbol{p}]) \tag{14}$$

模型结构如图4所示。

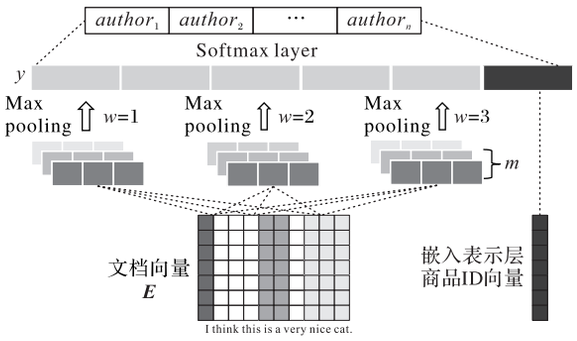


图 4 后期融合模型

Fig. 4 Later-stage fusion model

3 实验结果及分析

3.1 指标定义与实验数据集

本文使用准确率 Acc(Accuracy)、宏召回率 R_{macro}(macro-Recall)和宏 F1($F1_{\text{macro}}$)来评价算法的性能。其中准确率是所有类别整体性能的平均,而宏召回率和宏 F1 衡量了模型对不同类别的性能,各项指标的定义如下:

$$Acc = \frac{1}{N}\sum_{n=1}^{N} I(y^{(n)} = y'^{(n)}) \tag{15}$$

其中:N 为样本数量,$I(\cdot)$ 为指示函数,$y^{(n)}$ 为样本的真实标记,$y'^{(n)}$ 为预测结果。

类别 c 的召回率定义为:

$$R_c = \frac{TP_c}{TP_c + FN_c} \tag{16}$$

类别 c 的精确率定义为:

$$P_c = \frac{TP_c}{TP_c + FP_c} \tag{17}$$

其中:TP_c 表示真正例数(True Positive, TP),FN_c 表示假负例数(False Negative, FN),FP_c 表示假正例数(False Positive, FP),TN_C 表示真负例数(True Negative, TN)。

宏召回率定义为:

$$R_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C} R_c \tag{18}$$

宏精确率定义为:

$$P_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C} P_c \tag{19}$$

宏 F1 定义为:

$$F1_{\text{macro}} = \frac{2 \times P_{\text{macro}} \times R_{\text{macro}}}{P_{\text{macro}} + R_{\text{macro}}} \tag{20}$$

实验将原数据集划分为训练集:验证集:测试集=6:2:2,其中验证集用于调整参数,测试集用作最终测试。为了验证模型在不同领域的有效性,实验采用亚马逊电影评论(Amazon Movie_and_TV)和 CD 评论(CDs_and_Vinyl_5)两个数据集。由于上述两个原始数据集比较稀疏,而本文的实验数据既要求同一作者包含一定数量的评论信息,同时也要求同一商品包含一定数量的评论信息。因此,需要从原始的数据集中检索出一些满足上述要求的用户和商品。两个数据集的统计信息如表 2 所示。

表 2 数据集统计信息

Tab. 2 Dataset statistics

数据集	商品数量	用户数量	评论数/用户	评论数/商品	总评论数
电影评论	250	610	37.37	91.17	22 793
CD 评论	600	800	51.27	38.45	30 763

3.2 实验方法

表 3 中包含了神经网络的具体结构及参数。为了减轻过拟合,在每个卷积层之后加入 50% Dropout,使用 ReLU 作为激活函数,使用 Adam[16] 作为优化器,学习率为 10^{-4} 来训练网络。

表 3 神经网络结构及超参数

Tab. 3 Neural network architecture and hyperparameters

名称	层数	数值
最大长度 L	—	1 000
向量维度 d	—	300
卷积	3	m = 300, w = [1, 2, 3]
全连接	1	# of classes

3.3 不同实验方法对比

将所提出的模型与以下模型进行对比:

1) CNN-2:Shrestha 等[15]使用 Character n-gram 作为输入,使用一个 Embedding 层将输入映射为稠密的矩阵,然后依次通过卷积层和全连接层后使用 Softmax 函数得到输出。实验发现,当选择 2-gram 时,在验证集上的效果最好,将此模型记为 CNN-2。

2) LSTM-1:LSTM 已经被成功用于文本分类的任务中[4-5]。使用 Character n-gram 作为输入特征,将所有单向 LSTM 单元的输出进行求和作为文本特征,最后用 Softmax 函数进行分类。

3) SVM:Schwartz 等[1]选择 word n-gram 和 character n-gram 作为输入,使用 SVM 进行分类。实验证明,使用 word n-gram 和 character n-gram 作为输入与仅仅使用 character n-gram 作为输入的结果基本相同,在实验中将不考虑 word n-gram。实验中使用 character n-gram 长度为 4 作为特征,线性核 SVM 作为分类器。

4) RF:RF 是机器学习中经典的多分类方法,它包含多个决策树,在分类问题中往往具有很好的效果。实验采用 character 3-gram 作为输入,使用 sklearn 实现的随机森林分类器在验证集上调参,效果最好的分类器用于测试。

5) Syntax-CNN:Zhang 等[11]除了使用 n 元模型,还加入了语义特征,这样可以将文本的语义信息和风格信息融合起来得到比较好的效果,但同时增加了模型复杂度。

6) LDA-S:Seroussi 等[10]将作者的所有文本拼接后将单词频率作为 LDA 的输入,得到每个作者的主题分布,使用 Hellinger 距离度量新文本的主题分布与作者主题分布的距离,距离最近的为预测结果。

7) CNN-product:CNN-product 是一种仅利用商品 ID 向量预测用户的方法。首先将商品 ID 通过嵌入层将其表示为向量,然后通过一层卷积神经网络和最大池化得到卷积特征,最后使用 Softmax 函数进行分类。

不同方法的实验结果如表 4 所示。从表 4 中可以看出,后期融合在两个数据集上都取得了相比其他方法最好的结果,

在以后的实验中，将后期融合记为TDAA（Target-Dependent method for Authorship Attribution）。TDAA的效果优于前期融合，其原因可能是：前期融合将商品向量与文本向量视为相同的输入，忽视了两者所含的不同信息，使得模型无法学得互补的信息。

表4　两个数据集上不同方法的评价指标结果对比

Tab. 4　Comparison of evaluation results of different methods on two datasets

方法	电影评论数据集			CD评论数据集		
	Acc	R_{macro}	$F1_{\text{macro}}$	*Acc*	R_{macro}	$F1_{\text{macro}}$
CNN-2	0.519	0.411	0.415	0.683	0.581	0.579
LSTM-1	0.363	0.262	0.259	0.464	0.362	0.363
SVM	0.452	0.354	0.351	0.619	0.523	0.521
RF	0.307	0.209	0.205	0.492	0.401	0.399
Syntax-CNN	0.505	0.401	0.405	0.656	0.566	0.565
LDA-S	0.285	0.188	0.186	0.349	0.251	0.252
CNN product	0.018	0.006	0.003	0.012	0.003	0.004
前期融合	0.556	0.449	0.443	0.708	0.612	0.608
后期融合	0.569	0.467	0.465	0.725	0.621	0.622

与仅利用商品信息的对比：仅利用商品信息的方法在两个数据集上的准确率均不足0.1，本文方法远高过它。

与仅利用文本信息对比：在对比方法中，本文方法比其他方法中最优的结果仍高出4%~5%。在传统的机器学习分类方法中，效果最好的是SVM，它也是被广泛应用在作者身份识别问题中的一种方法；LDA-S效果不如其他机器学习方法的原因可能是商品评论数据集的主题分布比较集中，作者之间的主题分布差异不大，造成分类的难度增加；LSTM-1捕获的信息可能更多是语义上的，与作者的写作风格无关，因此与其他深度学习方法差异较大。

3.4　目标依赖信息对作者身份识别效果的影响

为了比较在相同文本特征下加入目标信息与不加目标信息的结果，设计了如下实验：采用character *n*-gram作为文本特征，使用CNN-2作为实验方法，融合方式采用后期融合与前期融合，对比有无依赖信息对效果的影响。在电影评论数据集上加入商品ID与不加商品ID的对比结果如表5所示。

表5　*n*-gram特征下目标依赖信息对*Acc*的影响

Tab. 5　Impact of target-dependence information on *Acc* based on *n*-gram feature

方法	电影评论	CD评论
CNN-2	0.519	0.682
前期融合	0.522	0.686
后期融合	0.540	0.706

采用BERT提取的文本特征，使用CNN作为实验方法，融合方式采用后期融合，对比有无依赖信息对实验结果的影响，实验结果如表6所示。

表6　预训练特征下目标依赖信息对*Acc*的影响

Tab. 6　Impact of target-dependence information on *Acc* based on pre-trained feature

方法	电影评论	CD评论
CNN-2	0.548	0.703
前期融合	0.554	0.710
后期融合	0.568	0.725

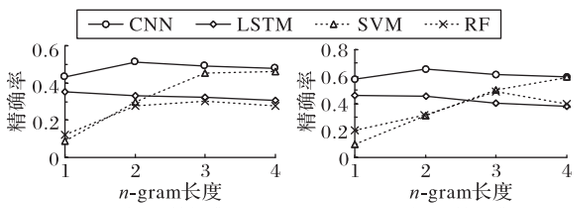
对比表5和表6，可以得出如下结论：

1）使用相同的分类模型，通过BERT提取的文本向量可以比使用character *n*-gram作为文本特征的分类精确率高出2%~3%，说明使用BERT提取文本特征的方法是一种普适且有效的方法。

2）后期融合是一种有效的融合方式。在使用相同文本特征的条件下，对目标依赖信息的后期融合可以比不加依赖信息的方法分类精确率高出2%左右。

3.5　不同长度*n*-gram对于实验结果的影响

为了探究不同长度的character *n*-gram对于实验结果的影响，设计了*n*-gram长度分别为1、2、3、4的实验，所采用的方法为CNN、SVM、RF与LSTM，实验结果如图5所示。



(a) 电影评论数据集　　(b) CD评论数据集

图5　两个数据集上*n*-gram长度对*Acc*的影响

Fig. 5　Impact of different *n*-gram length on *Acc* on two datasets

从图5可以看出，不同长度的*n*-gram对不同的方法影响不同。*n*-gram长度的增加会造成LSTM效果下降；而对于SVM，长度增加会使其效果变好；对于CNN和RF而言，存在一个最合适的长度使其性能最佳。因此，对于不同的方法首先要通过实验找出最佳的*n*-gram长度。TDAA由于没有使用*n*-gram特征，因此效果不受影响。

4　结语

本文提出了一种目标依赖的作者身份识别算法，解决了在限定条件下的作者身份识别问题。本文方法免去了复杂的特征设计，利用BERT提取文本信息，使得该方法更加具有普适性。利用商品ID作为对作者身份的限制条件，这种方法可以很好地推广到其他对作者身份有限制条件的应用场景中。

提高作者身份识别问题的效果的另一个思路是提高文本分类的效果。目前许多先进的模型被用于文本分类，如Li等[17]提出了一种对抗学习网络（Adversarial Network）来提高文本分类的效果；胶囊网络[18]最初被用在图像分类任务上，其动态路由机制实现了输出对输入的某种聚类；Zhao等[19]首先尝试了使用胶囊网络实现文本分类，并取得了很好的效果。这些方法都可以被用来解决作者身份识别问题。

参考文献（References）

[1] SCHWARTZ R, TSUR O, RAPPOPORT A, et al. Authorship attribution of micro-messages[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2013: 1880-1891.

[2] LAYTON R, WATTERS P, DAZELEY R. Authorship attribution for twitter in 140 characters or less[C]// Proceedings of the 2nd Cybercrime and Trustworthy Computing Workshop. Piscataway: IEEE, 2010: 1-8.

[3] KOPPEL M, SCHLER J. Authorship verification as a one-class classification problem[C]// Proceedings of the 21st International Conference on Machine Learning. New York: ACM, 2004: 1-7.

[4] TANG D, QIN B, LIU T. Document modeling with gated recurrent neural network for sentiment classification[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Pro-

cessing. Stroudsburg, PA: Association for Computational Linguistics, 2015: 1422-1432.

[5] TAI K S, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks [EB/OL]. [2019-02-20]. https://arxiv.org/pdf/1503.00075.pdf.

[6] KIM Y. Convolutional neural networks for sentence classification [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2014: 1746-1751.

[7] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification [C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2015: 649-657.

[8] ZHANG W, YUAN Q, HAN J, et al. Collaborative multi-Level embedding learning from reviews for rating prediction [C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2016: 2986-2992.

[9] ZHANG W, WANG J. Integrating topic and latent factors for scalable personalized review-based rating prediction [J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(11): 3013-3027.

[10] SEROUSSI Y, ZUKERMAN I, BOHNERT F. Authorship attribution with latent Dirichlet allocation [C]// Proceedings of the 15th Conference on Computational Natural Language Learning. Stroudsburg, PA: Association for Computational Linguistics, 2011: 181-189.

[11] ZHANG R, HU Z, GUO H, et al. Syntax encoding with application in authorship attribution [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2018: 2742-2753.

[12] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2013: 3111-3119.

[13] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2019-02-20]. https://arxiv.org/pdf/1810.04805.pdf.

[14] ATREY P K, HOSSAIN M A, EL SADDIK A, et al. Multimodal fusion for multimedia analysis: a survey [J]. Multimedia Systems, 2010, 16(6): 345-379.

[15] SHRESTHA P, SIERRA S, GONZÁLEZ F, et al. Convolutional neural networks for authorship attribution of short texts [C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2017: 669-674.

[16] KINGMA D P, BA J L. Adam: a method for stochastic optimization [EB/OL]. [2019-02-20]. https://arxiv.org/pdf/1412.6980.pdf.

[17] LI Y, YE J. Learning adversarial networks for semi-supervised text classification via policy gradient [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2018: 1715-1723.

[18] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules [C]// Proceedings of the 2017 Conference on Neural Information Processing Systems. [S. l.]: CUED Publications database, 2017: 3856-3866.

[19] ZHAO W, YE J, YANG M, et al. Investigating capsule networks with dynamic routing for text classification [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2018:3110-3119.

This work is partially supported by the Young Scientists Fund of the National Natural Science Foundation of China (61702190).

LI Yang, born in 1994, M. S. candidate. His research interests include data mining.

ZHANG Wei, born in 1988, Ph. D., associate professor. His research interests include user data mining, natural language processing.

PENG Chen, born in 1986, Ph. D., associate research fellow. His research interests include geospatial information processing.