DOI:10.13718/j.cnki.xsxb.2021.01.006

基于 NLP 和机器学习的短文本作者识别算法①

吴桂玲

信阳农林学院 信息工程学院,河南 信阳 464007

摘要:针对当前垃圾邮件账户撰写虚假在线评论,降低评论网站可信度的问题,提出一种基于自然语言处理和机器学习的短文本作者识别算法,该算法将自然语言处理技术(Natural Language Processing,NLP)与不同的机器分类器相结合,根据多个不同的语言特征解决了简短嘈杂的评论文本的作者识别问题.实验结果表明,相对于基线模型而言,本文算法在引入 NLP 技术后,仅采用一元语法和一元与二元语法相结合的两个 N-gram 模型的分类精度均有明显提高,充分说明本文算法的有效性.

关 键 词:自然语言处理;机器学习;作者识别;N-gram 模型

中图分类号:TP391 **文献标志码**:A **文章编号**:1000-5471(2021)01-0032-06

随着社交媒体的兴起,许多评论网站在收集在线产品和服务意见方面变得有影响力,同时评论网站需要客户提供准确而真实的评论[1].但是为了在评论网站上提高、损害产品或服务的声誉,个人和公司可能通过制造虚假评论来实施欺诈.这可以由在网站上拥有多个垃圾邮件账户的用户完成,同一作者的多个虚假评论将在多个账户中发布.因此,能够准确有效地识别欺诈性评论,对于网站维护其提供的评论信誉非常重要[2-3].作者识别任务可用于此域中,以确定两个或多个垃圾邮件账户是否属于同一个作者,并帮助删除这些账户及其相关联账户的评论[4]

作者识别(Author identification,AI)是指在一组封闭已知作者身份的文档集合中识别给定文档的作者的任务[5].随着社会化媒体服务的发展,短文本作者归属变得非常必要.当前已有一些方法用于作者识别.文献[6]提出了一个新的作者-文档-主题模型,该模型在作者和文档两个层次上为语料库建立模型,以解决短文本的作者归属问题,同时设计了新的分类算法来计算文本之间的相似度,寻找匿名文本的作者.文献[7]针对 Twitter 数据上的作者验证问题,提出了一种基于多层感知器的深度学习方法,获得了很高的分类成功率.文献[8]针对古希腊文字著作权归属问题,提出了一种仅使用形态-句法数据来识别作者身份的方法,不依赖特定的词汇项目,提高了小文本分类的精度.文献[9]提出了一种主题漂移模型,该模型可以监视作者写作风格的动态并同时了解作者兴趣,通过对时间信息和单词顺序的敏感分析,识别匿名文本中的作者身份.文献[10]采用了一组文本失真方法来掩盖主题相关信息,该方法将输入文本转换成一种更为主题中立的形式,同时保持与作者个人风格相关的文档结构,并控制语料库以提高其在具有挑战性任务中的表现.文献[11]通过情感定位和帖子长度来表征每个用户的发帖风格,然后结合字符 N-gram 和单词的 N-gram 模型来训练支持向量机,识别评论区短文本作者.

由于在线评论的文本长度很短,语言特征有限,使得上述方法在识别多个帐户的评论是否属于同一作者较为困难.本文提出一种基于自然语言处理和机器学习的短文本作者识别算法,该算法将自然语言处理

① 收稿日期:2020-02-28
基金项目:河南省科技攻关计划项目(182102210533).
作者简介:吴桂玲,讲师,主要从事模式识别与图像处理、物联网服务计算研究.

技术与不同的机器学习方法相结合，根据多个不同的语言特征，以期解决简短嘈杂的评论文本的作者识别问题.

1　自然语言处理

自然语言处理(Natural Language Processing，NLP)是计算机科学领域和人工智能领域中的一个重要分支，它关注计算机与人类之间使用自然语言的互动交流，最终实现计算机能够以一种智能与高效的方式理解和生成语言的目标. NLP 技术应用于多个领域，在信息检索、情感分析、文章作者识别、机器翻译等方面均取得了较好的效果，研究方向也从词汇语义成分的分析向叙事理解扩展.

目前，自然语言处理技术是基于统计机器学习，这些算法的输入是一大组从输入数据生成的特征. 根据自然语言的上下文关系特性建立数学模型，此类模型具有能够表达不同可能的答案，可以产生更可靠的结果.

2　短文本作者识别算法

本文提出的识别方法分为 4 个部分：①数据预处理，预处理评论；②特征表示，从预处理评论中提取特征；③自然语言处理，为评论添加上下文；④机器学习分类，对评论作者进行分类.

评论的分类包括训练阶段和预测阶段. 在训练阶段，对包含多个评论的训练数据进行预处理，并将其传递至特征提取器. 在特征提取器中，将多个 NLP 技术应用于评论并生成特征，然后利用这些特征训练机器学习的分类器. 最后在预测阶段使用这个分类器来预测新评论(输入)的作者(标签).

2.1　数据预处理

进行数据预处理的目的是标准化数据集中的数据并减少特征集中的数量，这样的处理也适用于将数据拟合到所有类型的分类模型上，因为该处理方式能够有效提高分类模型的计算效率. 使用的预处理步骤：①小写字母归一化：为防止同一单词的多个版本，所有单词均被归一化为小写形式，以便将它们全部计数；②标点符号的分离：大多数标点符号与句子中的单词结合在一起，使得同一单词形式的多种变体重复出现，但是标点符号可能是识别算法中的重要特征，因此预处理中将标点符号与单词分开，而不是删除；③删除出现次数为 1 的单词：如果一个单词在整个数据集中只出现一次，那么这个单词很可能不会对分类模型产生影响，因此这些单词被视为不相关的，可以删除.

2.2　特征表示

作者通常比其他人更喜欢某些单词或者短语. 为了捕捉作者的写作风格并帮助区分不同的作者，本文采用了 N-gram 模型[12]. N-gram 是一种基于概率判别的语言模型，基本原理是基于马尔可夫假设：第 n 个词的出现只与前 $n-1$ 个词相关，与其他任何词都不相关；整个语句出现的概率等于每个词出现的概率乘积，每个词出现的概率可以通过语料中统计的次数得到.

假设一条长为 n 的语句词序列为 $S=(w_1, w_2, \cdots, w_n)$，其中 $w_i(1\leqslant i\leqslant n)$ 是 S 的单词，S 出现在语料库中的概率为

$$P(S)=P(w_1, w_2, \cdots, w_n)=\prod_{i=1}^{n} P(w_i \mid w_1, w_2, \cdots, w_{i-1}) \tag{1}$$

从而将序列的联合概率转化为一系列条件概率的乘积. 假若当前词的出现仅依赖于其前面出现的一个词时，上式可以变为一个二元模型.

$$P(S)=\prod_{i=1}^{n} P(w_i \mid w_{i-1}) \tag{2}$$

在训练 N-gram 模型时，一个很重要的过程就是预测模型的条件概率为

$$P(w_i \mid w_{i-1})=\frac{P(w_{i-1}, w_i)}{P(w_i)}=\frac{c(w_{i-1}, w_i)}{c(w_i)} \tag{3}$$

其中，$c(w_i)$ 和 $c(w_{i-1}, w_i)$ 分别表示为词 w_i 与二元词组 (w_{i-1}, w_i) 在训练语料库中出现的次数.

随着N-gram数量的增加,用于训练分类器不同特征的数量增加,训练分类模型所花费的时间随之呈指数增加.因此,在本文算法中,仅对一元语法、一元语法与二元语法两者的组合进行评估.

2.3 自然语言处理技术

NLP技术[13]通过为分类器提供每次评论的上下文来帮助分类器.因此,它们有助于更好地理解每个作者的文体风格和语言特征.本文采用3种不同的NLP技术进行处理.

2.3.1 词干提取

词干提取是去除单词前后缀得到词根的过程,常见的前后词缀有名词复数、进行式、过去分词等形式,如单词"likes""liking""likely""liked"将全部简化为它们的词根形式"like".将具有相同词根形式的单词所有变体合并在一起,可以确保共同计算词根形式的频率,因此这会增加词根形式出现的可能性,从而可以对分类模型产生影响.在本文方法中用于分类的词干算法是Snowball词干算法[14].

2.3.2 词形还原

词形还原是基于词典,将单词的复杂形态转变成最基础的形态.不同于词干提取采用的缩减方式,词形还原主要采用转变的方法,通过对词形进行分析识别后将词转变为其原形.将形式不同但含义相同的词统一起来,可以方便后续的处理和分析.在本文方法中用于分类的词形还原方法是WordNet Lemmatizer算法[14].

2.3.3 停用词移除

停用词是指在处理自然语言数据前后会自动过滤掉某些词,这些词对文本分类没有影响.停止词是最常见词的集合,通过将这些词作为特征可能会导致文本的错误分类.因此,需要对停用词作出标记进行删除.本文根据Python自然语言工具包NLTK(Natural Language Toolkit)提供的128个停止词列表在评论文本中有依据地删除[14].

2.4 机器学习分类器

在机器学习中,分类器的作用是在标记好类别的训练数据基础上判断一个新的观察样本所属的类别.本文方法采用3种不同的机器学习分类器用于文本分类任务,分类的内容是判断评论文本是否属于同一作者.

2.4.1 支持向量机(SVM)

SVM是大幅度线性分类算法.该算法训练模型的目的是找到两个不同类别之间的分离超平面向量 $\boldsymbol{w}$,以使分离距离最大化.超平面向量 $\boldsymbol{w}$ 的推导可以定义为

$$\boldsymbol{w}:=\sum_i \gamma_i a_i \boldsymbol{r}_i,\ \gamma_i \geqslant 0 \tag{4}$$

其中,γ_i 是通过求解对偶优化问题得到,$a_i \in \{-1,1\}$,1和-1表示两个不同的类.当 $\gamma_j > 0$ 时,评论向量 $\boldsymbol{r}_j$ 由于对超平面向量 $\boldsymbol{w}$ 起到重要影响,称为支持向量.为了确定分类,可以简单地测试每个评论 $\boldsymbol{r}_j$ 位于超平面向量 $\boldsymbol{w}$ 的位置.

本文所使用的支持向量机是一个二值分类器,它训练支持向量机在每对候选作者之间进行分类,以便执行多类分类.这种训练支持向量机的方法称为一对多,每个新文本评论可以通过每个支持向量机,并选择最有可能的类.在本文方法中由于培训的类别很多,导致多项式核的计算时间太长而不可行,所以采用线性核而不是复杂的多项式核训练支持向量机.

2.4.2 多项式朴素贝叶斯MNB

MNB是一个概率分类器,假定每个特征在给定特征类别的情况下有条件地独立于其他特征.因此,分类模型不考虑单词的顺序,只考虑单词在文本中的出现频率.该模型基于贝叶斯规则,具有最高可能性的作者类 a 被视为评论作者,其定义为

$$P(a \mid \boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_n) = \frac{P(\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_n \mid a) \cdot P(a)}{P(\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_n)} \tag{5}$$

其中,$\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_n$ 为特征向量,$P(\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_n)$ 是通过将 $\boldsymbol{r}$ 评论中出现的每个特征 $\boldsymbol{f}$ 的计数除以评论总数来计算的.每个作者的可能性计算为

$$P(a)=\frac{Na}{N_S} \tag{6}$$

其中，Na 表示作者 a 的评论数量，N_S 表示总的评论数量.

为了训练 MNB 分类器，使用了拉普拉斯平滑度和用于学习先前类别概率的选项，这是为了防止在训练数据集中不存在评论特征时生成零可能性现象的发生.

2.4.3　最大熵 ME

ME 是另一个概率分类器，该分类器与 MNB 的不同之处在于，假设中不认为所有特征有条件地相互独立. 因此，当使用基于特征的模型时可以迭代添加特征，而不会出现重叠特征. 为了计算一个类(作者)的可能性，使用最大熵原理，选择最接近均匀分布的分布. 因此，该分类器除了在训练数据时给出的假设外，没有其他假设.

假定包含 N 个样本对$\{(r_1, a_1), \cdots, (r_N, a_N)\}$的训练数据被表示为单词出现的向量，则最大熵模型的可能性定义为

$$P(a \mid r, \mu)=\frac{\exp(\sum_i \mu_i \boldsymbol{f}_i(a, r))}{\sum_a \exp(\sum_i \mu_i \boldsymbol{f}_i(a, r))} \tag{7}$$

其中，$a \in A$ 表示作者类的集合；$r \in R$表示评论的集合；μ表示权重向量，是通过最大化每个评论 r 的每个特征 $\boldsymbol{f}_i(a, r)$ 的条件概率来计算的.

本文使用机器学习分类器和自然语言处理技术的不同组合进行作者识别，为了实现每个分类模型与 NLP 技术的最佳组合，使用了前向选择. 前向选择是从分类模型中不使用任何 NLP 技术开始，然后选择性地向模型中添加具有最显著准确性改进的单个 NLP 技术的过程.

3　实验结果与分析

为了验证本文算法的性能进行了 4 个不同的实验，前 3 个实验采用不同的机器学习分类器和自然语言处理技术的组合. 第一个实验作为基线分类模型，在实验中每一个机器学习分类器与一个一元语法向量模型或者二元语法向量模型相结合；第二个实验将一元语法基线模型与本文提出的多种 NLP 技术相结合，以提高基线模型的精度；第三个实验将联合的一元语法和二元语法基线模型与众多 NLP 技术相结合. 第 4 个实验将机器学习分类器和自然语言处理技术的最优组合与最新的识别算法进行对比.

交叉验证是一种评估分类模型准确性的方法. 为了评估不同的作者识别模型，使用了 k 折交叉验证. 即将数据集拆分为 k 个大小相等的分区，$k-1$ 个分区用于训练集，剩余 1 个分区用于测试集. 使用训练集对分类模型进行训练，然后在测试集上进行测试，以计算模型的准确性. 此过程重复 k 次，将每个不同的分区用作测试集，将剩余分区用作训练集. 然后，将这些迭代中计算出的所有精度的平均值推广为单个精度估计. 通过使用这种验证方法，可以使用更多数据来训练模型. 此外，交叉验证确保数据集中的所有数据都用于训练和测试，而不会对准确性结果造成任何偏差. 在本文实验中，使用了 10 倍交叉验证.

3.1　数据集

为了保证在在线评论上执行本文算法，使用了由餐厅和酒店评论组成的 Yelp 评论数据集. 该数据集包含了来自 192 609 家企业的 1 637 138 位不同作者的 6 685 900 篇独特评论，其中数据集内的一些作者只发表了少量的评论. 由于没有足够的语言特征来获取和学习，很难对那些作者使用本文算法. 出于这个原因，在实验中使用的数据集只包括每个作者至少发表了 100 篇评论的作者，这与假评论者倾向于发表许多评论的事实是一致的.

3.2　实验结果

表 1 给出了仅使用一元语法或采用了一元语法和二元语法组合的基线模型获得的结果 . 从表 1 中可以看出，这种组合为 SVM 和最大熵提供了很高的分类精度. 但是，MNB 的分类准确性较低. 这些结果在所测试的两种类型的 N-gram 模型中都是一致的. 对于 SVM 和 ME 分类器，当将使用的 N-gram 模型从一元语法增加到一元语法与二元语法组合时，准确性都会提高. 表现最佳的基准模型是采用一元语法和二元语法组合的 SVM.

表1 基线分类器模型的分类精度 %

机器学习分类器	一元语法分类精度	一元和二元语法组合的分类精度
MNB	47.1	46.7
SVM	84.2	89.0
ME	80.7	81.1

表2和表3分别给出了针对仅采用一元语法、一元语法与二元语法同时采用时，机器学习分类器与NLP技术最佳组合的测试结果．表2、表3的结果表明，NLP技术在两个N-gram模型中具有相同的效果．这是由于NLP技术的相同组合在两个N-gram模型中均提供了最佳结果．与基线模型相比，结合NLP技术的作者识别算法的精度有明显的提高，充分说明本文算法的有效性．

表2 一元语法时分类器与NLP技术最佳组合的分类精度

机器学习分类器	NLP技术	分类精度/%
MNB	停用词+词干提取	48.2
SVM	词形还原	86.1
ME	词干提取	81.8

表3 一元语法与二元语法结合时最佳组合的分类精度

机器学习分类器	NLP技术	分类精度/%
MNB	停用词+词干提取	49.9
SVM	词形还原	93.5
ME	词干提取	84.3

4 结 语

本文提出一种在短文本上执行作者识别的方法，该算法结合多种机器学习分类器和自然语言处理技术，根据不同的语言特征，解决了评论网站上在线评论文本的作者识别问题．为了验证本文算法的有效性，采用3组不同机器学习分类器和自然语言处理技术最佳组合的实验．实验结果表明，相对于基线模型而言，引入NLP技术的分类精度有明显提高，而且在一元语法和二元语法相结合的N-gram模型中，SVM分类器与词形还原NLP技术组合的分类精度达到93.5%，从而证实了本文算法对作者人数众多，且评论规模很小的数据集能够产生良好的准确性．

参考文献：

[1] ARAUJO L, MARTINEZ-ROMO J, DUQUE A. Discovering Taxonomies in Wikipedia by Means of Grammatical Evolution [J]. Soft Computing, 2018, 22(9): 2907-2919.

[2] BARTOLI A, DE LORENZO A, MEDVET E, et al. Active Learning of Regular Expressions for Entity Extraction [J]. IEEE Transactions on Cybernetics, 2018, 48(3): 1067-1080.

[3] RUANO-ORDÁS D, FDEZ-RIVEROLA F, R MÉNDEZ J. Using Evolutionary Computation for Discovering Spam Patterns from E-mail Samples [J]. Information Processing & Management, 2018, 54(2): 303-317.

[4] YOUNG T, HAZARIKA D, PORIA S, et al. Recent Trends in Deep Learning Based Natural Language Processing [Review Article [J]. IEEE Computational Intelligence Magazine, 2018, 13(3): 55-75.

[5] GÓMEZ-ADORNO H, POSADAS-DURÁN J P, SIDOROV G, et al. Document Embeddings Learned on Various Types of N-grams for Cross-topic Authorship Attribution [J]. Computing, 2018, 100(7): 741-756.

[6] ZHANG H W, NIE P, WEN Y L, et al. Authorship Attribution for Short Texts with Author-Document Topic Model [M]// Knowledge Science, Engineering and Management. Cham: Springer International Publishing, 2018.

[7] Yilmaz M, Mutlu B, Utku A, et al. Deep learning approach for author verification problem on twitter [C]//2018 26th Signal Processing and Communications Applications Conference (SIU). Cesme-Izmir: IEEE, 2018.

[8] GORMAN R. Author Identification of Short Texts Using Dependency Treebanks without Vocabulary [J]. Digital Scholarship in the Humanities, 2020, 35(4): 812-825.

[9] YANG M, CHEN X J, TU W T, et al. A Topic Drift Model for Authorship Attribution [J]. Neurocomputing, 2018,

273：133-140.
[10] STAMATATOS E. Masking Topic-related Information to Enhance Authorship Attribution [J]. Journal of the Association for Information Science and Technology, 2018, 69(3): 461-473.
[11] Leepaisomboon P, Iwaihara M. Utilizing Latent Posting Style for Authorship Attribution on Short Texts [C]//2019 IEEE International Conference on Dependable Autonomic and Secure Computing (DASC). Fukuoka: IEEE, 2019.
[12] AHMED H, TRAORE I, SAAD S. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques [M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017.
[13] K V, GUPTA D. Unmasking Text Plagiarism Using Syntactic-semantic Based Natural Language Processing Techniques: Comparisons, Analysis and Challenges [J]. Information Processing & Management, 2018, 54(3): 408-432.
[14] VIJAYAKUMAR B, FUAD M M M. A New Method to Identify Short-Text Authors Using Combinations of Machine Learning and Natural Language Processing Techniques [J]. Procedia Computer Science, 2019, 159: 428-436.

Author Identification Algorithm of Short Text Based on Natural Language Processing and Machine Learning

WU Gui-ling
College of Information Engineering, Xinyang Agriculture and Forestry University, Xinyang Henan 464007, China

Abstract: In order to reduce the credibility of comment websites, an author identification algorithm of short text based on natural language processing and machine learning has been proposed. This algorithm combines natural language processing (NLP) with different machine classifiers, and solves the author recognition problem of short and noisy comment text according to different language features. The experimental results show that, compared with the baseline model, the proposed algorithm combines with either unigram only or both unigram and bigram by introducing NLP technology, and the classification accuracy is significantly improved, which fully shows the effectiveness of the algorithm.

Key words: natural language processing; machine learning; author identification; N-gram model

责任编辑　夏　娟