文章编号：2096-1472(2023)-04-14-04
DOI:10.19644/j.cnki.issn2096-1472.2023.004.004

基于机器学习的剧本角色情感识别研究

蔡校育，邱美兰，李德旺

(惠州学院数学与统计学院，广东 惠州 516007)
✉910940859@qq.com; qiumeilan@hzu.edu.cn; ldwldw1976@126.com

摘　要：针对DataFountain平台举办竞赛所提供的剧本角色情感数据集，采用中文分词、去停用词和绘制词云图等工具对数据进行预处理，利用词频−逆向文档频率(TF−IDF)算法提取文本特征，分别建立了基于支持向量机和朴素贝叶斯算法的机器学习分类识别模型。将建立的新模型应用于剧本角色情感的识别和分析研究，结果表明，朴素贝叶斯分类模型的识别效果要优于支持向量机分类模型；并且，当拉普拉斯平滑系数$\alpha = 0.2$时，朴素贝叶斯算法的分类准确率接近于80%。

关键词：剧本角色；支持向量机；朴素贝叶斯；情感识别
中图分类号：TP181　**文献标识码**：A

Research of Screenplay Characters Emotion Recognition based on Machine Learning

CAI Xiaoyu, QIU Meilan, LI Dewang

(*School of Mathematics and Statistics*, *Huizhou University*, *Huizhou* 516007, *China*)
✉910940859@qq.com;qiumeilan@hzu.edu.cn; ldwldw1976@126.com

Abstract: Based on the emotion datasets of the script characters provided by the DataFountain platform competition, this paper proposes to preprocess the data by using tools such as Chinese word segmentation, removing stop words and drawing word clouds, and text features are extracted by using the Term Frequency–Inverse Document Frequency (TF-IDF) algorithm. Then, machine learning classification and recognition models based on Support Vector Machin (SVM) and Naive Bayes algorithm are established respectively. The two proposed models are applied to the recognition and analysis of script character emotion. The results show that the recognition effect of Naive Bayesian classification model is better than that of SVM classification model. In addition, when the Laplacian smoothing coefficient $\alpha = 0.2$, the classification accuracy of Naive Bayes algorithm is close to 80%.

Keywords: screenplay characters; SVM; Naive Bayes; emotion recognition

1 引言(Introduction)

对于影视制片人来说，剧本的好坏直接决定其商业价值和社会意义，因此，对剧本文本分析成为不可或缺的环节，其中剧本角色的情感识别是剧本分析中一个非常重要的任务。剧本角色情感识别是将剧本中涉及角色的对白和动作描述识别为某一种具体的情感倾向，属于情感分析[1]中句子级别的范畴，输入为剧本中的角色对白或动作描述的句子，输出其对应的情感倾向。

基于机器学习的情感分析是一种有监督的学习方法，属于文本机器学习[2]的范畴，目前常见的基于机器学习情感分析的算法有支持向量机(SVM)[3]、朴素贝叶斯[4]和逻辑回归[5]等，研究人员也开展了与此相关的大量研究工作[6]。本文将对非结构化的剧本数据使用情感分析技术进行处理，从而减少人工处理数据的工作量，利用机器学习算法快速挖掘非结构化数据中的价值，依据情感预测的结果为剧本分析提供有价值的参考，对影视作品的发展具有一定的指导意义。

2 剧本角色情感识别(Emotion recognition of screenplay characters)

2.1 数据集介绍

本文研究所需数据来源于DataFountain平台举办的剧

基金项目：广东省重点建设学科科研能力提升项目(2021ZDJS080)；惠州市哲学与社会科学基金项目(2022ZX046)；惠州学院教学质量工程项目(XJYJG2021045)；惠州学院“百名优秀青年教师”培养项目.

本角色情感识别竞赛所提供的数据集，该数据集的主要数据来源于一部电影剧本，通过人工的情感标注，同时对数据进行相应的处理，使之划分为三种情感(1：正向情感；0：中性；−1：负向情感)。该数据集共有36,612 条数据样本，而中性数据对于本文模型的研究用处不大，也易产生分歧，所以剔除中性数据，只保留正、负向情感，共10,143 条数据样本，部分数据内容如表1所示。

表1 数据集

Tab.1 Datasets

编号	内容	角色	情感值
1171_0001_A_7	c1开心地点了点头	c1	1
1171_0001_A_10	c1再次微笑着点头，然后举手敬礼，但是手的形状却是弯的	c1	1
1171_0001_A_11	o2笑了笑：军礼不是这么敬的。五指并拢，大臂带动小臂，举到齐眉	o2	1
1171_0002_A_21	c1四处张望，显然对这一切充满了好奇和胆怯	c1	−1
1171_0003_A_26	女演员跳舞的时候，分h3一直在旁边批评着，排练的时候，o2领着c1走了进来	h3	−1

2.2 数据预处理

因为中文语篇中词语不存在空隙，所以必须采用分词的方法进行识别，而在分词过程中，某些对分类不起作用的信息也要去掉，即删除停用词，最后将那些能传达重要信息的关键词从文本中抽取出来，并将文本表示为这些关键词的集合。数据预处理包括数据清洗、文本分词、删除停用词等。

2.2.1 文本分词

由于中文文本与英文不同，中文文本分词是预处理中不可缺少的关键步骤，因此在分类过程中使用词语表示文本时必须先进行分词处理。目前的分词技术已经逐步完善，其中jieba分词具有准确率高、性能优越及可扩展性等特点，是一款当下流行的中文分词技术。

jieba分词可以分为三种类型：精确模式、全模式和搜索引擎模式。其中，精确模式实现了对被分词文本的准确分割，并且不存在冗余词，本文将运用jieba分词工具中的精确模式进行分词操作，分词效果如表2所示。

表2 文本分词效果

Tab.2 The effect of text segmentation

分词前	分词后
c1开心地点了点头	c1/开心/地点/了/点头
c1再次微笑着点头，然后举手敬礼，但是手的形状却是弯的	c1/再次/微笑/着/点头/，/然后/举手敬礼/，/但是/手/的/形状/却是/弯/的
o2笑了笑：军礼不是这么敬的。五指并拢，大臂带动小臂，举到齐眉	o2/笑了笑/：/军礼/不是/这么/敬/的/。/五指/并拢/，/大臂/带动/小臂/，/举到/齐眉
c1四处张望，显然对这一切充满了好奇和胆怯	c1/四处张望/，/显然/对/这/一切/充满/了/好奇/和/胆怯
女演员跳舞的时候，分h3一直在旁边批评着，排练的时候，o2领着c1走了进来	女演员/跳舞/的/时候/，/分/h3/一直/在/旁边/批评/着/，/排练/的/时候/，/o2/领着/c1/走/了/进来

2.2.2 去停用词

对于文本分类而言，有些词语在文本中出现的次数并不能反映该词语在文本中的重要性。比如“一二三四”“你我他”“这个”“的”，这些没有特殊语义并且出现频繁的词语，即停用词。本文主要研究中文文本所体现的情感，这些停用词在很大程度上会对该研究产生影响，因此应该将这些停用词从文本中清除掉，避免它们对后续分类产生干扰。去停用词效果如表3所示。

表3 去停用词效果

Tab.3 The effect of removing stop word

去停用词前	去停用词后
c1/开心/地点/了/点头	开心/地点/点头
c1/再次/微笑/着/点头/，/然后/举手敬礼/，/但是/手/的/形状/却是/弯/的	微笑/点头/举手敬礼/手/形状/却是/弯
o2/笑了笑/：/军礼/不是/这么/敬/的/。/五指/并拢/，/大臂/带动/小臂/，/举到/齐眉	笑了笑/军礼/敬/五指/并拢/大臂/带动/小臂/举到/齐眉
c1/四处张望/，/显然/对/这/一切/充满/了/好奇/和/胆怯	四处张望/充满/好奇/胆怯
女演员/跳舞/的/时候/，/分/h3/一直/在/旁边/批评/着/，/排练/的/时候/，/o2/领着/c1/走/了/进来	女演员/跳舞/分/旁边/批评/排练/领着/走

通过对本文的数据集内容进行相应的预处理之后，可以绘制正、负向情感关键词词云图，如图1和图2所示。

图1 正向情感数据集词云图

Fig.1 Word cloud of positive sentiment datasets

图2 负向情感数据集词云图

Fig.2 Word cloud of negative sentiment datasets

从图1和图2两个词云图中可以看出，“看着”“坐

在”“我要”“画外音”“爸爸”等词语在两种情感中都是高频词，对本文的研究会产生相应的影响。因此，在停用词表中需添加这些词语，可以减少误差，提升模型预测的准确率。

2.3 特征抽取

经过分词和过滤掉停用词之后的文本数据集要进行特征抽取，即向量化之后输入分类模型中进行运算。TF−IDF算法是常用的文本向量化算法，它由词频(TF)和逆向文档频率(IDF)两个部分组成，即$TF-IDF=TF\times IDF$。TF−IDF的主要思想是当一个词语在一篇文章中出现的概率很高，而其他文章出现的概率很小，证明该词语有很好的区分效果，可以被用来进行分类[7]。

TF是指某个特定词语在文档中出现的次数，即

$$tf(i,j)=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

其中，$n_{i,j}$是词语i在文档j中出现的次数，$\sum_k n_{k,j}$表示文档j中所有词语的出现次数之和。

IDF是指衡量一个词的一般重要性的尺度。一个特定词语的IDF，可以用文档的总数量除以含有该词的文档数量，然后将其商取以10为底的对数，即

$$idf(i)=\lg\frac{|D|}{|\{j:t_i\in d_j\}|} \tag{2}$$

其中，$|D|$为一个语料库中的文档总数量，$|\{j:t_i\in d_j\}|$表示包含词语t_i的文档数目。如果在语料库中没有这个词语，就会导致分母为零，所以通常情况下使用$1+|\{j:t_i\in d_j\}|$。

2.4 模型建立

本文将使用Sklearn库(python中的机器学习库)中的支持向量机和朴素贝叶斯两种分类算法构建分类模型。因此，在完成数据预处理和特征工程相关工作后，接下来需对数据集进行划分、交叉检验、模型训练及分类预测等相关操作。

2.4.1 划分数据集

机器学习的分类方法需要大量的数据用于训练，特别是对神经网络的训练。在进行机器学习时，数据集被分为两类，一类是训练集，另一类是测试集。本次实验选取80%的数据作为训练集，20%的数据作为测试集。有时为了保证模型的精度，往往需要先进行k折交叉验证。k折交叉验证实质上是把一个数据集分成k份，每次选$k-1$份为训练集，剩余的1份作为验证集，然后取k个模型的平均测试结果作为最终的模型效果。本文将以10折交叉验证为基础，尝试探索两种分类模型的有效性。

2.4.2 交叉验证及结果

通过对朴素贝叶斯(Naive Bayes)和支持向量机(SVM)两种机器学习模型进行10折交叉验证，并将10次的交叉验证的准确率作为最终的结果。两种分类模型10次运行对应的准确率如表4所示，根据表4的结果绘制如图3所示的箱型图。

表4 两种分类模型准确率

Tab.4 Accuracy of two classification models

次数	SVM	Naive Bayes
1	0.7064	0.6985
2	0.7064	0.7015
3	0.7251	0.6847
4	0.6874	0.6736
5	0.6903	0.6903
6	0.6667	0.6637
7	0.7012	0.6736
8	0.6321	0.6430
9	0.6302	0.6716
10	0.6677	0.6479
平均值	0.6814	0.6748

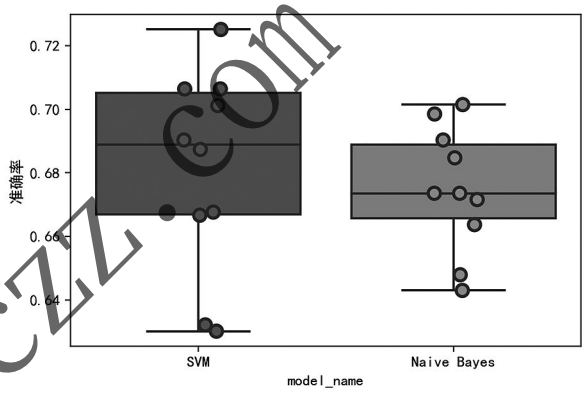

图3 分类模型准确率箱型图

Fig.3 The accuracy box chart of classification models

从图3中可以看出，两种模型相比，线性支持向量机的平均准确率要比朴素贝叶斯的准确率略高，但准确率较为分散，即存在不稳定性。因此，本文通过设置超参数的不同取值，进一步研究朴素贝叶斯算法的综合性能。

2.5 模型评估

本文利用混淆矩阵对朴素贝叶斯分类算法的性能进行评估，其中包括准确率、精确率、召回率、*F*1值和*AUC*指标[8−9]。根据朴素贝叶斯的拉普拉斯平滑法[10]选取不同的拉普拉斯平滑系数α，对朴素贝叶斯分类模型进行实验，得到实验结果如表5所示。从表5可以看出，最佳的拉普拉斯平滑系数介于0.1—0.5。通过调整超参数，可以使算法的性能变得更好。

表5 不同α值评价指标结果对比表

Tab.5 Comparison of evaluation index results for different α

α	准确率	精确率	召回率	*F*1值	*AUC*指标
0.05	0.764	0.79	0.86	0.82	0.729
0.1	0.769	0.79	0.87	0.83	0.733
0.2	0.775	0.79	0.88	0.83	0.736
0.5	0.757	0.75	0.92	0.83	0.698
1	0.78	0.73	0.95	0.83	0.674

通过前面模型分析及超参数的对比实验，运用朴素贝叶斯算法以及设置超参数拉普拉斯平滑系数$\alpha=0.2$进行学习，分别采用训练集和测试集进行预测[11]，得到如图4和图5所示的两种情况预测结果。

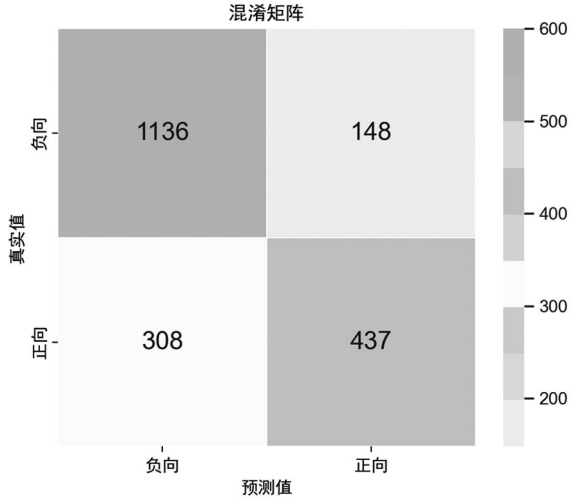

图4 测试集混淆矩阵

Fig.4 Confusion matrix of test datasets

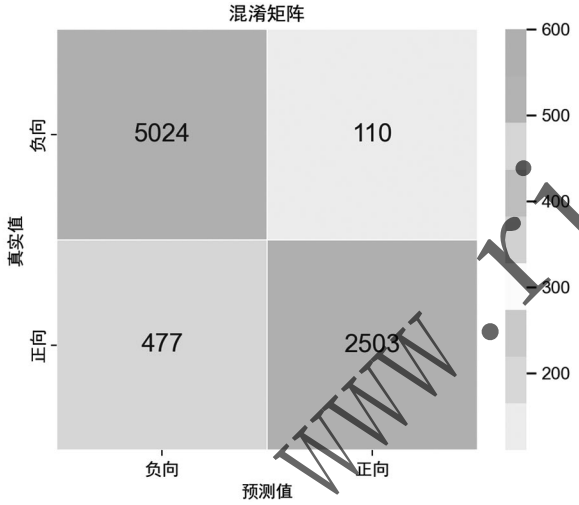

图5 训练集混淆矩阵

Fig.5 Confusion matrix of training datasets

从图4和图5两个混淆矩阵得出，朴素贝叶斯算法对测试集样本的预测结果准确度接近于80%，训练集样本的预测结果高达93%。

3 结论(Conclusion)

本文主要建立了基于支持向量机和朴素贝叶斯算法的两种情感分类与识别模型，对剧本中每句对白和动作描述中涉及的每个角色从多个维度进行分析并识别出情感。

首先，根据剧本角色情感文本的特点，对所获取的数据文本进行预处理，包括文本分词、去停用词、绘制词云图、特征抽取等，建立了基于支持向量机和朴素贝叶斯算法的两种情感分类与识别模型。其次，利用10 折交叉验证得出两种模型的预测准确率，分析了两种机器学习情感识别模型的预测效果，并通过不断调整模型中超参数的取值对模型进行优化。最后，根据研究结果得出朴素贝叶斯识别模型在剧本角色情感识别方面的效果要优于支持向量机的识别模型，并且，当超参数拉普拉斯平滑系数$\alpha=0.2$时，朴素贝叶斯识别模型的预测准确率接近于80%。

本研究的不足之处是尽管模型的训练有较好的拟合效果，但由于数据存在样本不均衡的现象，正向情感数据在总样本数据中所占的比重偏低，存在一定的过拟合现象。因此，在后续的研究中，应该增大正向情感的样本数据量，从而对本文的研究做进一步的改进和优化，使得预测结果更加准确、更具有可解释性。

参考文献(References)

[1] 李梦楠,汪明艳.基于机器学习的情感分析方法及应用研究综述[J].软件工程,2021,24(09):21-23,8.

[2] AGGARWAL C C. Machine learning for text[M]. Switzerland: Springer, 2018:5-26.

[3] 韩开旭.基于支持向量机的文本情感分析研究[D].大庆:东北石油大学,2014.

[4] 方小宇,罗补干,周铄洋,等.基于朴素贝叶斯算法的群众留言多标签分类的应用[J].科学技术创新,2021(09):100-102.

[5] 李佳儒,王玉珍,丁申宇.基于逻辑回归的在线评论情感分类方法研究[J].东莞理工学院学报,2020,27(05):50-54.

[6] 鲍继彬.基于机器学习的网购平台产品评论情感分析[D].哈尔滨:哈尔滨工业大学,2019.

[7] 王啸楠,尹辉平.基于自然语言处理的高校舆情情感倾向分析模型的研究[J].鞍山师范学院学报,2020,22(04):40-44.

[8] 王子.基于中文小说的对话角色和情感识别的研究与实现[D].北京:北京邮电大学,2021.

[9] 蒋帅.基于AUC的分类器性能评估问题研究[D].长春:吉林大学,2016.

[10] 陈翠娟.改进的多项朴素贝叶斯分类算法和Python实现[J].景德镇学院学报,2021,36(03):92-95.

[11] 李春林,武巾莉.基于机器学习的白酒板块股评情感分析[J].信息技术与信息化,2021(10):139-141.

作者简介：

蔡校育(1998-)，男，本科生.研究领域：机器学习，大数据分析.

邱美兰(1980-)，女，博士，讲师，人工智能高级工程师.研究领域：数据科学与计算，机器学习，深度学习.本文通信作者.

李德旺(1976-)，男，博士，讲师.研究领域：经济统计，大数据统计分析.