文章编号：2095-2163(2023)02-0210-04 中图分类号：TP393.08 文献标志码：A

基于朴素贝叶斯的影评情感分析研究

邓慈云，余国清
（湖南信息职业技术学院，长沙 410200）

摘 要：本文基于 python 技术，提出了利用朴素贝叶斯算法实现的影评情感分析系统。首先利用 Scrapy 爬虫框架获取数据集，然后使用 pandas 库和正则表达式等技术完成数据清洗；对影评文本采用 jieba 分词后，使用多项式贝叶斯分类器，构造出一个基于朴素贝叶斯的情感分类模型。通过对模型进行训练，并使用豆瓣网站采集的影评数据进行分类预测。实验结果表明，该模型具有良好的分类效果。

关键词：朴素贝叶斯；情感分析；python；数据采集

Research on emotion analysis for film reviews based on naive Bayes

DENG Ciyun, YU Guoqing
(Hunan College of Information, Changsha 410200, China)

【Abstract】 Based on Python technology, this paper proposes a film review emotion analysis system using naive Bayesian algorithm. Firstly, the data set is obtained by using the Scrapy crawler framework. Then the data cleaning is completed by using Pandas and regular expression technology. After Jieba word segmentation for the film review text, a sentiment classification model based on Naive Bayes is constructed by using polynomial Bayesian classifier. After training and testing with the film review data collected on Douban website, the experimental results show that the model has a good classification effect.

【Key words】 naive Bayes; emotion analysis; python; data acquisition

0 引 言

随着社会和经济的高速发展，人们在精神生活、娱乐等方面的需求越来越高，电影已经成为大众精神生活中不可分割的一部分。2021 年中国电影市场累计票房达 472.58 亿，较 2020 年增长 131.5%。2021 年国产电影总产量 740 部，同比 2020 年增长 13.8%。面对日益壮大的电影市场以及不同的题材，经常会发现同一部影片在不同平台上的评分存在较大的差异。

情感分析是指用机器学习的方法解析出文本中情感极性信息，归纳出用户的情绪、态度、倾向等情感意向的过程，其是自然语言处理（Natural Language Processing，NLP）的一个分支内容[1]。文本情感分析是指提取文本中主观信息的一种 NLP 任务，其具体目标通常是找出文本所对应的正负情感态度。情感分析可以在实体、句子、段落乃至文档上进行。对于情感分析，只需要准备标注了正负情感的大量文档，就能将其视作普通的文本分类任务来解决[2]。

近年来，诸多学者对影评的文本情感分析以及如何提高结果的准确率进行了研究，并取得了一定的研究成果。如：文献[3]中提出了在影评的文本情感分析中，将机器学习方法与分层技术结合，针对具有异质结构的文本数据的算法。文献[4]提出了使用 Keras 内置的 Tokenizer 模块建立字典，利用字典将影评文字进行预处理后，通过 Keras 框架构建 MLP 模型并训练。文献[5]提出了一种加入注意力机制的联合神经网络模型，用来对影评进行情感分析。文献[6]提出了一种基于 Keras 平台实现的双向 LSTM（BiLSTM）的影评情感分析算法。综上研究分析，影评的文本情感分析的准确率依然不高，亟待进一步探索和研究更具实用性、通用性的算法和模型。

为了能够客观全面的了解观众对影片的真实感受，本文利用 python 作为编程语言，使用 Scrapy 框

基金项目：湖南省哲学社会科学成果评审委员会课题（XSP22YBC417）。

作者简介：邓慈云（1983-），女，硕士，讲师，主要研究方向：人工智能和可视化技术；余国清（1972-），男，硕士，副教授，主要研究方向：软件工程和职业教育。

收稿日期：2022-04-15

架爬取豆瓣电影网站影评数据，构建分类模型完成训练，并评估训练器的分类效果；最后利用训练后的分类器，对中文影评文本进行情感分析和文本分类，让观影者能够快速地从大量影评中得到有价值的信息，也让影视工作人员了解观影者的喜好以及主观情感倾向。

1　文本情感分析

1.1　数据来源

本文使用豆瓣网电影（https://movie.douban.com/chart）影评数据信息，其中数据字段包含：电影详情信息（电影类型、上映时间、演员列表等）；电影短评内容（用户、是否观看、五星评分、评论时间、有用数、评论内容等），将其作为分析的目标数据。

1.2　数据采集

Scrapy 是用 Python 语言开发的一个快速、高层次的屏幕/Web 抓取框架，用于抓取 Web 站点并从页面中提取结构化数据。Scrapy 使用 Twisted 异步网络请求框架来处理网络通信，不需要额外实现异步框架，而且包含各种中间件接口，能够灵活地实现各种需求[7]。

使用 Scrapy 框架爬取豆瓣电影影评数据的过程为：首先利用 selenium 实现模拟自动登录，然后从 Top250 电影排行榜里爬取电影信息和链接地址，接下来根据链接地址爬取相关影片的具体信息和影评信息并保存到 csv 文件中。

1.3　数据清洗

数据清洗是指在获取文本数据之后，对数据进行重新审查和效验的工作。主要包括：缺失值清洗、重复值清洗和错误值清洗。通过对采集的数据进行查看和分词后，影评文本中存在以下情况及相应处理方法：

（1）对于英文、长度过短、重复及无实际意义的灌水文本，可通过正则表达式进行英文识别，通过长度过滤内容过少的评论。

（2）对于有缺失值的文本，通过查找确认存在有缺失值的记录，然后使用 Pandas 库实现删除相应记录。

（3）对于存在简体、繁体混杂的文本，通过使用 opencc 库，实现将繁体中文转换成简体中文。

（4）去除停用词。汉语中有一类没有多少意义的词语，比如助词“的”、连词“以及”、副词“甚至”、语气词“吧”，称为停用词。借助预先准备好的停用词字典，通过查询字典的方式，剔除停用词。

1.4　基于朴素贝叶斯的情感分析算法

1.4.1　算法流程

基于朴素贝叶斯的情感分析的实现过程如图 1 所示。首先对影评使用 jieba 库进行分词，去停用词等预处理，然后构建分类模型并用训练集进行训练，同时利用测试集评估训练器的分类效果，最后利用训练后的分类器对分类文本进行情感分类。

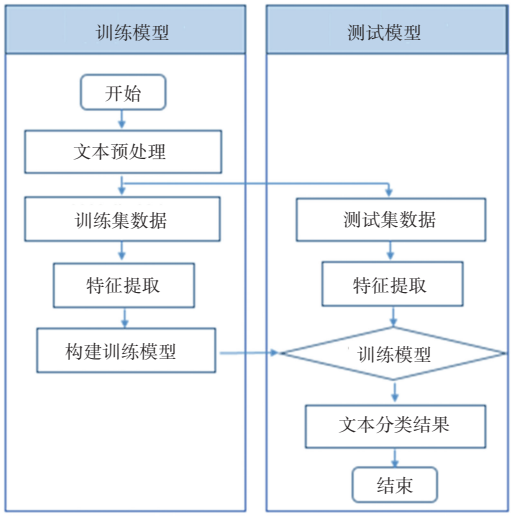

图 1　训练过程和分类过程

Fig. 1　Training process and classification process

1.4.2　朴素贝叶斯

朴素贝叶斯是分类器中最常用的一种生成式模型，其基于贝叶斯定理将联合概率转化为条件概率，利用特征条件及独立假设简化条件的概率进行计算。朴素贝叶斯法的目标是通过训练集学习联合概率分布 $p(X,Y)$，由贝叶斯定理可以将联合概率转化为先验概率分布和条件概率分布之积[8]，表达形式如下：

$$p(X=x,Y=c_k)=p(Y=c_k)P(X=x\mid Y=c_k)$$

其中，类别的先验概率分布 $(p(Y=c_k))$，可以通过统计每个类别下的样本多少（极大似然）来估计。即：

$$p(Y=c_k)=\frac{count(Y=c_k)}{N}$$

假设第 i 维 x_i 有 m_i 种取值，则组合起来 x 一共有 $\prod_{i=1}^{n} m_i$ 种，该条件概率分布的参数数量呈指数级的。当特征数量达到十万量级时，参数估计实际上是不可行的。为此，朴素贝叶斯法“朴素”地假设所有特征是条件独立的。该条件独立性假设为：

$$p(X=x,Y=c_k)=p(X_1=x_1,\cdots,X_n=x_n\mid Y=c_k)=\prod_{i=1}^{n}p(X_i=x_i\mid Y=c_k)$$

在预测时,朴素贝叶斯法最终的分类预测函数为

$$y = \arg\max_{c_k} p(Y = c_k) \prod_{i=0}^{n} p(X_i = x_i \mid Y = c_k)$$

1.4.3 朴素贝叶斯分类器

朴素贝叶斯分类器通过计算一个样本属于某一类的概率(后验概率),进而比较概率大小,来决定样本的分类结果。分类器需要数据集作为已知样本集,并且需要这些样本的分类结果,最后对新给出的样本集进行分类。具体来说,假设已经得到样本集 $D = \{x_1, \cdots, x_n\}$,每一个 x_i 都有 k 个特征,分别记为 a_i,可能类别为 $Y = \{y_1, \cdots, y_m\}$,根据每个 x_i 的特征,其会被分类到某一个 y_j 类中[9]。朴素贝叶斯分类器的特点是实现模型简单,且分类快速而精确。

本文从豆瓣网站爬取 Top250 排行榜中约 40 000条影评作为语料库。对影评分析情感倾向时,将评分中的推荐、力荐、还行视为积极评论,用数字 1 表示;将较差和很差视为消极评论,用数字 0 表示,积极评论和消极评论各占 50%。图 2 给出了样本实例,其中第一列数字为情感标签,第二列文字为影片评论内容。

1,没看会后悔 三个小时一点也不长 群像很好人物立体 很震撼
1,真的全程无尿点!规模之大,场面之震撼,到处都能激起属于中国人的感情共鸣。我们的志愿军在敌
1,非常好看的影片,里面的人物生动感人,70年前的战争更会比演绎的更惨烈,片子里时时说为了后辈
0,去看了点映,值得票价,三个小时看下来还好,一直战争戏容易麻痹双眼,但是也刺激。我不喜欢红
0,这是中国人民志愿军,不是普通意义的战争片,不是搞一堆坦克飞机和冲锋时人山人海就能够拍好的
1,有多久没看到当惊龙的编剧作品就有多期待《长津湖》,真的完全出乎意料的满意!整体故事相当完
1,一如既往的套路…民族情绪煽动真的ok…fine…观影的时候一直在思考:第一,无谓牺牲是否存在?
0,想的太多,拍得太散。人工雕刻痕迹过重。
1,外公参加过长津湖战役,看到了外公的影子,那场战役能够活下来已经很难了。战争镜头真好看
0,这拍的是动作片…美军除了空军,陆军就是屑。无语,真当抗美神剧在拍。不少画面甚至没制作完成
1,让我感动的不是演员,而是我联想到在几十年前有这么一群平凡而伟大的人为了我们的国家和下一代

图 2 标注情感标注的影评评论示例

Fig. 2 Sample set of film comments tagged with sentiment tags

(1)训练集与测试集的分割比率

如何设定训练集和测试集的分割比率,对朴素贝叶斯分类器的性能影响十分明显。本文使用 Python 的可视化工具 pyecharts,绘制精确度和分割比例折线图,找到训练集和测试集的最佳比例为6∶4。其中,影评数据的 60%作为训练集,40%作为测试集。

(2)分类器的选择

sklearn 的 naive_bayes 模块提供了高斯朴素贝叶斯、多项式朴素贝叶斯和伯努利朴素贝叶斯等 3 种用于构建朴素贝叶斯模型的类,其分别对应 3 种不同的数据分布类型。本文实验选择的是多项式贝叶斯分类器。

2 实验结果分析

2.1 情感分类结果分析

通过模型训练,获得适合影评情感分类的新模型。为了验证训练模型的效果,采集了 300 多条来自豆瓣网站的影评记录,并事先进行人工情感标签标注(数字 1、0 分别对应积极评论和消极评论)。测试中,将分数大于或等于 0.5 的评论判断为积极评论,否则判断为消极评论。将预测结果与人工判定结果进行对比,准确率达到了 92%,证明该模型训练过程是有效的。测试结果如图 3 所示。

label	comment	差评	好评	label	comment	差评	好评
1	沂蒙山小调一出,	0.324	0.676	0	这是中国人民	0.666	0.334
1	先看1950他们正	0.194	0.806	1	有多久没看到	0.422	0.578
1	看了点映,有笑	0.376	0.624	1	一如既往的套	0.68	0.32
1	小村庄夜战那段	0.325	0.675	0	想的太多,拍	0.723	0.277
1	场面拍得挺好,(	0.429	0.571	1	外公参加过长	0.373	0.627
1	(写在前面:我	0.633	0.367	0	这拍的是动作	0.476	0.524
1	我看见那句为抗	0.22	0.78	1	让我感动的不	0.257	0.743
1	点映观影,场面	0.425	0.575	1	好片子,剪坏	0.637	0.363
1	万里一个新兵在	0.548	0.452	1	事无巨细的开	0.297	0.703
1	该夸的也夸了,这	0.38	0.62	1	令人敬仰的是	0.396	0.604
1	CG做得实在一言	0.78	0.22	1	今天看了长津	0.3	0.7
1	可能因为知道结	0.366	0.634	0	好看就完事了	0.61	0.39
1	战斗场面篇幅之	0.525	0.475	1	为什么到现在	0.585	0.415
1	篇幅可以删减一	0.708	0.292	1	虽然有些不如	0.499	0.501
1	长津湖一战我军	0.54	0.46	1	认真安利《长	0.278	0.722
1	以前一直不理解	0.129	0.871	1	兰晓龙没让人	0.851	0.149
1	非常合格的献礼	0.786	0.214	1	刚入朝在干河	0.139	0.861
1	人物情感突如其	0.366	0.634	1	好几个画面还	0.496	0.504
1	国内战争题材电	0.49	0.51	1	个十百千万,	0.492	0.508
1	人物,从正派到	0.753	0.247	0	时长太长了,	0.699	0.301
1	“《长津湖》最	0.104	0.896	1	没有金刚川那	0.377	0.623
1	我觉得下次别设	0.662	0.338	0	拉满期待值去	0.777	0.223
1	不知道为什么有	0.561	0.439	1	中国第一部史	0.303	0.697
1	你说它没剧情吧,	0.658	0.342	1	太震撼了,太	0.422	0.578
1	只感受到了林超	0.385	0.615	1	从火车出逃的	0.222	0.778
1	看了点映,第七	0.19	0.81	1	冗长而又充斥	0.626	0.374
1	感人的是历史,	0.514	0.486	1	万里这个角色	0.82	0.18
1	中国所有的主旋	0.772	0.228	1	电影太震撼了	0.106	0.894
1	我就知道,肯定	0.856	0.144	1	直接比喻成:	0.646	0.354
1	卧槽尼玛的美国	0.727	0.273	1	香港导演们,	0.441	0.559

图 3 测试结果

Fig. 3 Test result

2.2 情感值分析

通过对影评数据的情感分析得到情感值(取值范围 0~1)。通过使用 matplotlib 可视化工具,绘制出情感建议值的直方图,如图 4 所示。由图中可以发现,观众对该电影整体的情感倾向是积极的。其中情感值分布在 0.1 和 1.0 左右的数量占比约为3.2%,情感值分布在 0.5 以上的数量占比约为84.1%。

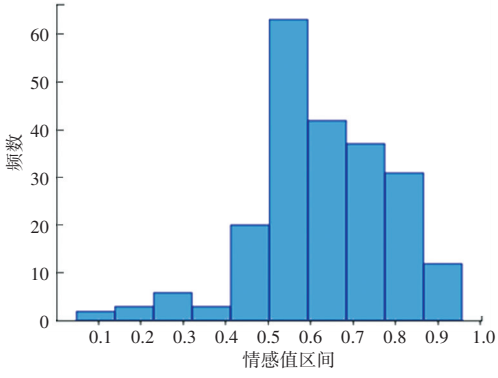

图 4 情感值区间统计

Fig. 4 Affective value interval statistics

3 结束语

在各种各样的分类器中,朴素贝叶斯法可算是

(下转封三)