王菁驿. 有声读物 APP 在线评论情感分析[J]. 智能计算机与应用,2024,14(4):180-183. DOI:10.20169/j.issn.2095-2163.240429

有声读物 APP 在线评论情感分析

王菁驿
(广西民族大学 管理学院, 南宁 530000)

摘 要: 本文利用朴素贝叶斯算法实现在线评论情感倾向分析。首先,利用爬虫获取数据,然后对数据进行预处理;其次,对在线评论文本采用 jieba 分词后,进行了文本分析与词频统计;使用朴素贝叶斯分类器构造一个基于朴素贝叶斯的情感分类模型,对模型进行训练后,使用有声阅读 app 的用户评价进行情感分析,得到在线评论情感倾向结果。

关键词: 朴素贝叶斯; 情感分析; 有声读物; 文本分析
中图分类号: G353.1 **文献标志码**: A **文章编号**: 2095-2163(2024)04-0180-04

Audiobook APP online review sentiment analysis

WANG Jingyi
(School of Management, Guangxi Minzu University, Nanning 530000, China)

Abstract: In this paper, we use the plain Bayesian algorithm to realize the analysis of online comment sentiment tendency. Firstly, the crawler was used to obtain the data, and then the data was preprocessed; secondly, the text analysis and word frequency statistics were carried out after jieba segmentation was applied to the text of the online reviews; a plain Bayesian classifier was used to construct a sentiment classification model based on plain Bayes, and after the model was trained, the user evaluations of audiobook reading apps were used for the sentiment analysis, and the results of the online reviews' sentiment tendency were obtained.

Key words: Naive Bayes; emotion analysis; audio books; text analysis

0 引 言

随着数字技术和人工智能的迅速发展,有声读物作为数字图书的重要形式之一,弥补了读者时间和空间的限制,提供了一种方便快捷的阅读方式[1]。根据 Grand View Research1 和 Acumen Research and Consulting 的报告,2022 年全球有声图书市场规模约为 53.6 亿美元,有声读物的普及度将不断提高[2]。未来有声读物 APP 将拥有更广阔的市场空间和发展潜力。

在线评论具有实时性、客观性、规模性、便捷性,挖掘获取的数据是用户主动发表的,更能反映用户真实的想法和需求。在线评论挖掘可以获取大量的用户评论数据,不受到样本量的限制[3]。通过分析在线评论,可为产品改进和优化提供有力的依据[4]。

本文爬取有声读物 APP 在线评论数据,建立分类器完成训练,并对分类器进行性能评估,最后利用训练好的情感分类器对有声读物 APP 在线评论进行词频统计和情感分析。通过数据分析了解用户需求、满足用户需求、改进应用程序质量,提升读者的阅读体验。

1 数据获取及数据预处理

1.1 数据获取

从移动应用市场 appstore 中的"图书专题"热门排行中筛选与有声读物的相关应用,本文选择了 10 个最受欢迎的有声书应用程序,即喜马拉雅、番茄听书、番茄小说、懒人听书、喜马拉雅儿童版、凯叔讲故事、荔枝 FM、蜻蜓 FM、得到、帆书(原樊登读书)。在评论区按照"最有帮助排序"的顺序爬取数据,爬取每条评论的评论时间、用户名、评分、评论内容,每个 APP 各爬取5 000 条评论,共爬取到50 000 条在线评论,共计1 964 016 字,具体情况见表 1。

作者简介: 王菁驿(1998-),女,硕士研究生,主要研究方向:数据挖掘,机器学习。Email:137361773@qq.com

收稿日期: 2023-04-10

表 1　爬取数据统计表

Table 1　Crawl data statistics

APP 名称	评论条数	评论字数
番茄小说	5 000	191 604
番茄畅听	5 000	173 381
喜马拉雅	5 000	209 767
懒人听书	5 000	194 646
荔枝 FM	5 000	168 987
蜻蜓 FM	5 000	305 500
帆书	5 000	268 634
得到	5 000	234 525
喜马拉雅儿童	5 000	82 528
凯叔讲故事	5 000	134 444
总共	50 000	1 964 016

1.2　数据清洗

收集的评论数据包含了部分对本文研究没有意义的文本数据。本文先对获得的评论数据进行相应的预处理，相关工作有数据清洗、过滤停用词、中文 jieba 分词等环节。

1）去除垃圾留言

一些评论中只有几个词、一个字符串或只有标点符号，对搜索没有实际的信息或意义，因此，本文将少于 3 个词的纯符号评论和数字评论删除。

2）消除重复留言

在评论数据集中，如果存在评论内容完全相同的数据，则删除多余的评论。在线评论中，这种现象被称为“刷评论”，用户会刷评论以获取奖励或满足其他目的，这会对数据分析和清洗产生影响。

3）消除不相关的评论

在线评论中，有些评论内容可能与有声专辑的内容或作者无关，这些评论可能是为了达到某种目的或者发泄情绪而发布的。此外，一些用户为了获得平台的奖励或者提高自己的影响力，可能会发表无关或者恶意评论。针对这些无关或者恶意评论，本文需要进行筛选和清洗，以确保最终的评论数据集是有价值的。

4）数据分词及去停用词

对数据进行清洗后，需要进行中文文本的分词和去停用词。分词是将连续的句子划分为单独的字词，而去停用词则是剔除那些对文本分析无实际意义的词。本文使用 jieba 词语分词工具来对中文文本进行分词。在使用 jieba 时，一些特定的词可能会在词典中缺失，自定义一个 stop words 词典，以便于得出的结果更加精准。本文的研究对象是用户对于软件的评论，这些评论中存在一些专业名词及一些流行词汇，加入一些例如“喜马拉雅”、“网络主播”、苹果手机”等词语。在分词处理完成后，借助了哈工大停用词表来剔除那些无实际意义的词，以便后续的数据处理和分析。

2　高频词分析

经过数据清洗后，统计出现频次较多的词语，即高频词，对评论内容进行分词，将评论词语按照出现频次进行统计，生成了一个包含前 25 个词频的列表见表 2。

表 2　高频词表

Table 2　High frequency word list

排名	高频词	频数
1	软件	6 389
2	广告	4 831
3	内容	3 405
4	故事	3 046
5	会员	2 797
6	小说	2 640
7	时间	2 067
8	垃圾	2 010
9	声音	1 745
10	孩子	1 722
11	功能	1 643
12	读书	1 576
13	平台	1 354
14	老师	1 251
15	免费	1 238
16	版本	1 229
17	音频	1 148
18	节目	1 053
19	收费	980
20	生活	925
21	界面	892
22	游戏	878
23	习惯	574
24	陪伴	555
25	恶心	415

表 2 可见高频词的分布情况，读者对有声读物的关注点主要体现在以下几个方面：“软件”、“广告”频数排名为第一和第二，说明读者在使用该类软件过程中普遍存在广告的困扰，而“时间”、“免费”、“收费”则说明多数 APP 采取免费+收费模式，

但是免费内容常伴随较长广告,令读者体验感大大降低。“内容”、“故事”及“小说”频数排名相对较高,读者对该 APP 提供的内容较为关注,这也是有声读物 APP 的核心竞争力,儿童更喜爱故事性的读物,成人更偏好多种类型的网络小说。“游戏”一词也有较高频次,说明读者同样较关注电竞、热门网游解说等音频。

词云图是一种数据可视化的形式,特别适合于展示文本数据信息。通过将高频关键词以视觉化的方式展现,可以更快速了解文本数据的主题和重点。本文将得到的词频数据制作成词云图,如图 1 所示。通过这些高频词汇可以更清楚地了解评论数据集的主要内容,从而更好地进行分析和挖掘。

图 1 词云图

Fig. 1 Word cloud map

3 基于朴素贝叶斯的情感分析

3.1 分类原理

朴素贝叶斯分类器是一种用于文本分类的分类模型,其核心思想是基于贝叶斯定理和特征之间独立性的假设来计算后验概率,从而实现分类[5]。在处理离散特征时,需要进行平滑处理以避免出现概率为 0 的情况,然后使用多项式模型进行分类[6]。

朴素贝叶斯算法的优点在于其收敛速度快,易于实现分类目标,具有较高的分类准确率和分类速度,不仅适用于大规模的多分类任务,也适用于增量训练,不易受数据缺失的影响[7]。

通过查全率、查准率、$F1$ 值评估构建好的朴素贝叶斯分类器。查准率表示预测正确的样本占总样本的比例,见式(1);查全率表示预测见正确样本的样本中真正正确的样本的比例,见式(2);$F1$ 值是查全率和查准率的调和平均数,查全率和查准率的加权平均值[8],见式(3)。

$$precision = TP/(TP + FP) \quad (1)$$

$$recall = TP/(TP + FN) \quad (2)$$

$$F1 = 2PR/(P + R) \quad (3)$$

目前针对文本情感分析的研究中,朴素贝叶斯法是最适合短评情感分析的机器学习方法之一。与长文本相比,短篇评论文本中的情绪词之间的关联性较小,这与简单贝叶斯算法中的属性独立假设相近。因此本文选择基于朴素贝叶斯的机器学习方法对在线评论进行情感分析。

3.2 情感标注

本文选择 APP 在线评论中的五分之一作为训练样本,然后进行人工标注以区分正面情感、负面情感和中性情感。主要表达了正向情感的评论文本标记为“+1”;如“读者体验不佳,费劲,广告泛滥,反正各种各样的问题。”等主要传达了负向情感的评论文本标记为“-1”;对于在线评论文本,如果其表达了中性情感或者与本事件无关,标记为“0”。本文一共标记了 10 000 个样本,标注结果如图 2 所示。

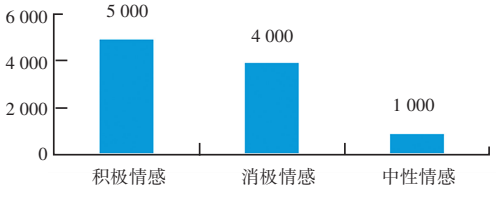

图 2 标注结果统计图

Fig. 2 Statistical chart of labeling results

本文采用基于朴素贝叶斯算法的情感分类模型。首先,将经过词汇向量化处理的训练集评论数据用于训练情感分类器;其次,通过测试,得出准确率为 84.6%和召回率为 82.3%。

3.3 实验结果与分析

将该分类器模型应用于所选取的 10 个 APP 评论数据进行预测,得到不同情感倾向的比例:正向情感 67%,负向情感 29%,中性情感 4%。可见大部分读者对于有声读物 APP 的评价是积极的,但是一些读者在使用过程中会产生负向情感,而中性情感的表达较少。总的来说,读者对于有声读物 APP 持有满意态度。那些持有负面情感的读者,大多数是因为广告太多,弹窗广告干扰读者体验,希望能够减少广告的数量。此外,读者还对功能有着各自的需求,比如需要开通“防沉迷模式”、“车载模式”等等。读者经常会因为使用过程中遇到的不愉快情况,比如频繁的广告、闪退的界面、迅速耗电的应用、缓慢的更新、高昂的听书费用、不感兴趣的阅读内容等等,而产生负面情感。

统计各个在线评论的不同情感倾向数量,得到不同情感倾向的比例,如图 3 所示。从图 3 可以看

出，积极情感比例最高的应用程序是懒人听书和喜马拉雅儿童，表明这些应用程序具有积极的读者评价和口碑；消极情感比例最高的应用程序是番茄小说和蜻蜓 FM，表明这些应用程序存在一些负面评价或读者体验问题；帆书和得到是知识付费类应用程序，读者对这些应用程序的内容和付费服务感到满意；荔枝 FM 的积极情感比例为 70%以上，比蜻蜓 FM 高出 20 个百分点以上，表明荔枝 FM 在读者中有更好的口碑和品牌形象。大多数应用程序的中性情感比例都比较低，表明读者在评价这些应用程序时倾向于使用情感化的词汇，而不是中性词汇。

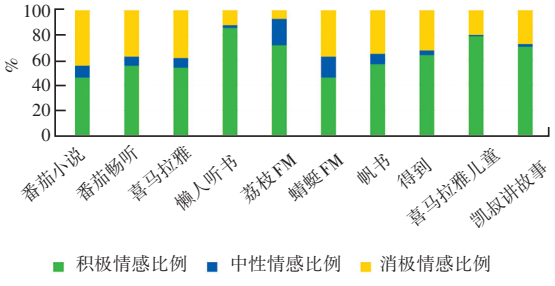

图 3　APP 情感分类比例

Fig. 3　APP sentiment classification proportion

4　结束语

本文通过词频统计和朴素贝叶斯分类器对有声读物 APP 进行数据分析，得出以下结论：

（1）读者对有声读物积极倾向较高，为 67%；中性情感少，比例为 4%；消极情感较少，比例为 29%。读者使用有声读物 APP 满意度较高；

（2）在不同的 APP 中，懒人听书和喜马拉雅儿童应用口碑好，使用感佳，读者的积极情感倾向分别达到了 86%和 79%。而番茄小说和蜻蜓 FM 消极情感比例高，需多加改进；

（3）从负面倾向评论中可以得出，内容质量、广告量、使用顺畅度、界面设计是读者关注的主要因素，运营商可以在这些方面进行提高和改进。

本文针对有声读物平台 APP 改进提出如下建议：

（1）有声读物平台需要关注用户的需求，提供高质量的内容，才能吸引更多的读者。选择或改编适合有声读物形式的高质量名著书籍，注意文本的结构、逻辑、语言等方面，让有声读物内容更具流畅性和生动性；注重播读质量，挑选优秀的播音员来录制、演播内容。

（2）打造可持续盈利模式。有声读物 APP 应减少广告量，改善广告模式。提供付费服务，让读者享受没有广告的阅读体验。运营商采取互动营销模式，与图书馆、出版社其他平台合作。利用名人效应，提升有声读物 APP 影响力。

（3）重视用户调研和反馈，以此优化应用程序的用户界面和交互设计，提升用户的使用体验，以此提升读者的满意度。通过不断提升技术手段，改善读者使用体验，根据用户需求和问题，提供更优质的服务，以提高读者的满意度和忠诚度。

参考文献

[1] 蔡翔，王睿. 从国民听书率看我国有声阅读产业发展趋势[J]. 现代出版，2018(1)：65-70.

[2] Audiobooks Market Size, Share and trends report, 2030 [OL]. San Francisco: Grand View Research, 2021. https://www.grandviewresearch.com/industry-analysis/audiobooks-market

[3] 曹树金，陈忆金，杨涛. 基于用户需求的图书馆用户满意实证研究[J]. 中国图书馆学报，2013，39(5)：60-75.

[4] 姜巍，张莉，戴翼，等. 面向用户需求获取的在线评论有用性分析[J]. 计算机学报，2013，36(1)：119-131.

[5] WEBB G I, KEOGH E, MIIKKULAINEN R. Naïve bayes[J]. Encyclopedia of Machine Learning, 2010, 15: 713-714.

[6] 辛梓铭，王芳. 基于改进朴素贝叶斯算法的文本分类研究[J]. 燕山大学学报，2023，47(1)：82-88.

[7] 叶光辉，李松烨，宋孝英. 基于多标签标注学习的城市画像文本分类方法研究[J]. 数据分析与知识发现，2023，7(5)：60-70.

[8] 许丽，焦博，赵章瑞. 基于 TF-IDF 的加权朴素贝叶斯新闻文本分类算法[J]. 网络安全技术与应用，2021(11)：31-33.