Statistics and Application 统计学与应用, 2024, 13(4), 1128-1139
Published Online August 2024 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2024.134115

Hans汉斯

基于文本挖掘的电影消费者满意度影响因素研究

屈文鑫，符俊雄

广东金融学院金融数学与统计学院，广东 广州

收稿日期：2024年7月9日；录用日期：2024年7月30日；发布日期：2024年8月12日

摘　要

本文以豆瓣电影评论为研究对象，通过文本挖掘的相关方法，对电影消费者满意度的因素构成进行探索。首先，爬取豆瓣电影评论数据2万条，然后对预处理后的评论数据进行词频统计和词云图分析，归纳出影响观众对电影满意度的主要因素。接着用向量化处理后的文本数据，通过k-means算法的fit函数总结出电影消费者关注的主要特征因素，建立消费者满意度指标体系。然后，构建情感词典，计算各特征因素的情感得分，输出满意度得分表，并分析电影消费者的总体满意情况。最后通过建立贝叶斯网络模型，得出各满意度影响因素之间的关系和影响程度。结果表明，电影消费者满意度影响因素及其影响系数排序为：角色(0.1945)、导演(0.1693)、剧情(0.1618)、演员(0.15)、题材(0.1407)、表演(0.1254)、视听(0.0585)，并从电影配置、电影设计、电影表现三个方面提出优化建议。

关键词

电影评论，消费者满意度，文本挖掘，贝叶斯网络模型，情感分析

A Study on the Influencing Factors of Movie Consumer Satisfaction Based on Text Mining

Wenxin Qu, Junxiong Fu

School of Financial Mathematics and Statistics, Guangdong University of Finance, Guangzhou Guangdong

Received: Jul. 9th, 2024; accepted: Jul. 30th, 2024; published: Aug. 12th, 2024

Abstract

This article takes Douban movie reviews as the research object and explores the factors influencing consumer satisfaction in movies through text mining methods. Firstly, 20,000 pieces of Douban

文章引用: 屈文鑫, 符俊雄. 基于文本挖掘的电影消费者满意度影响因素研究[J]. 统计学与应用, 2024, 13(4): 1128-1139.
DOI: 10.12677/sa.2024.134115

movie review data are crawled, and then the preprocessed review data are analyzed by word frequency statistics and word cloud map to summarize the main factors that affect audience satisfaction with the movie. Then, the vectorized text data are used to summarize the main feature factors that movie consumers are concerned about through the fit function of the k-means algorithm, and establish a consumer satisfaction index system. Then, the emotion dictionary is constructed to calculate the emotional scores of each feature factor, output a satisfaction score table, and analyze the overall satisfaction of movie consumers. Finally, a Bayesian network model was established to determine the relationship and degree of influence among various satisfaction influencing factors. The results show that the factors influencing consumer satisfaction in movies and their influencing coefficients are ranked as follows: character (0.1945), director (0.1693), plot (0.1618), actor (0.15), theme (0.1407), performance (0.1254), and audio-visual (0.0585), and optimization suggestions are put forward from three aspects of movie configuration, design, and performance.

Keywords

Movie Reviews, Consumer Satisfaction, Text Mining, Bayesian Network Model, Sentiment Analysis

1. 引言

如今互联网技术惠及我们的生活，使得人们越来越多地在网络上进行商品的挑选和购买。2019 年，中国成为全球电商市场最大国家，以 8630 亿美元的电商收入总额位列世界第一，占据全球约 45.42% [1]。互联网技术蓬勃发展下中国网民数量不断增多，根据中国互联网中心发布的第 52 次《中国互联网络发展状况统计报告》显示，到 2023 年 6 月底，我国网站数量达到 383 万个，我国网民数量达到 10.79 亿，互联网普及率高达 76.4% [2]。如今网络成为了人们认识世界和向世界发声的重要渠道。

近年来，在习总书记的带领下，我国经济持续稳步增长，经济和文化的发展齐头并进。我国电影市场规模逐年扩大，且票房总额连续多年快速增长，如今电影已成为我们日常生活娱乐的一种主要途径。2018 年全球电影票房达 417 亿美元，中国电影票房达 89 亿美元，中国承担了推动全球电影票房增长的重要帮手[3]。根据 1905 电影网的《2023 中国电影——年度调查报告》显示，截止 2023 年 12 月 27 日，北美票房累计 87.05 亿美元，中国以 75.71 亿美元累计票房位列全球第二，荧幕总数多达 86,310 块。中国电影呈现出欣欣向荣的发展状况，票房稳定增长，稳居世界第二大电影市场的位置。《“十四五”中国电影发展规划》提出，2035 年，我国将建成电影强国，中国电影实现高质量发展，电影创作生产能力明显进步，彰显中国精神、价值、力量、美学的精品作品不断涌现，以国产电影为主导的电影市场规模在全球领先[4]。

本文通过文本挖掘技术对电影消费者的满意度进行研究，旨在得出电影消费者满意度的主要影响因素及其对满意度的影响机制，为电影出品方改进电影、提高竞争力指明方向，也帮助消费者更好地享受到优质的影片，最终实现消费者和出品方双赢。

2. 国内外研究现状

近年来，国内外学者们在借助文本挖掘研究消费者满意时通常喜欢集中在文本数据中的数值型信息方面和文本本身内容含义方面进行研究。一是评论文本数据中对数值型信息的探究，如 Evanschitzky 等

[5]研究了德国居民的网络购物行为。徐小琳(2010) [6]以 Charles 的医疗决策过程理论及 Makoul 的理想医疗决策过程的构成要素为理论依据，通过对 209 名外科住院手术患者进行测试，编制了患者对医疗决策参与的满意度量表。李燕飞(2016) [7]基于淘宝评论，采用多元回归分析探讨了影响商品销量的因素。张璇(2023) [8]基于携程蒙古旅游景区在线评论，通过 IPA 分析法构建了“重要性－满意度”二维四项方格图，分析得到蒙古旅游景区多个方面的改进方案。其二是通过对评论文本中的信息进行分析，通常是对消费者评论进行主题提取和计算情感倾向。如安翔等(2018) [9]基于京东和天猫平台下北大荒米业竞争对手的产品评论数据，通过情感词典和 LDA 主题模型挖掘出消费者的满意度，得出不同品牌的差异，为北大荒米业改善商品带来了决策依据。严军超等(2019) [10]采用朴素贝叶斯算法、SVM 和 RNN 算法作为社交媒体数据文本的文本分类模型，对比不同算法下的分类效果，选择最优的模型作为情感分析的研究。赵志杰(2020) [11]采用 TF-IDF 算法和贝叶斯算法与情感分析技术获得酒店顾客情感得分，构建了酒店顾客满意度评估模型。辛雨璇和王晓东(2021) [12]对电影网站的评论文本进行挖掘，通过构建情感词典及 TF-IDF 的方法，实现评论文本的高频词可视化分析并挖掘出评论文本中的情感倾向，然后使用贝叶斯分类器挖掘出评论信息的深层含义。

根据对文献的查找和研究得知文本挖掘作为多项技术的交叉研究领域，很多学者借用文本挖掘进行意见挖掘，在消费者满意度上已经取得了很多成果，尽管文本挖掘的研究方向有很多，不过对电影行业的研究较少。目前关于电影行业的研究主要集中在票房收入、影院建设和文化发展等方面，而在微观层面研究不足。因此，本文从消费者的视角出发，以豆瓣电影评论为研究对象，通过文本挖掘的相关方法，对电影消费者满意度的因素构成进行探索，为电影行业的研究方向进行补充。

本文的设计思路为：首先，借助 Python 爬取豆瓣电影网站的电影评论，通过清洗并处理获得在线评论的相关高频词和词云图，进而探究电影消费者满意度的影响因素。接下来，对预处理后的数据构建相似词词表、计算情感得分、消费者满意度得分和构建贝叶斯模型等，挖掘各影响因素之间的联系和对满意度的影响机制。最后，根据研究结果总结为电影制作提出建议。

3. 数据选取与预处理

3.1. 数据来源

豆瓣电影网站是中国最受欢迎的电影评分网站之一，有着用户群体广泛、用户评分和评论数量大、覆盖范围广等优点，因此该网站的评论数据具有一定的代表意义。所以，本文以豆瓣电影的电影短评为研究数据，选取了其 2020 年到 2023 年间中国大陆出品的 100 部电影的评论数据。但由于豆瓣电影对数据爬虫的 IP 限制，无法直接抓取电影的全部评论数据，并且在抓取数据过多时极易出现大量数据重复的现象。因此，在尽量减少数据重复的情况下，对选取的电影每部随机选取 200 条评论数据，包含了用户的名称、评级(分别是很差、较差、还行、推荐、力荐)、评论内容、发布时间等相关信息，一共两万条电影评论数据为样例。爬取评论的时间为 2023 年 12 月 20 到 30 号，评论截止日期为 12 月 30 号。

3.2. 数据预处理

本研究收集了共计 2 万条文本评论数据。然而，在爬取数据的过程中，出现了采集样本数目大、爬取次数频繁、网站反爬机制等情况，因此导致所采集的数据容易地出现了一些噪声。为了确保后续分析结果的准确有效，对爬取到的数据进行清洗，主要内容包括：剔去空值、删除重复评论、删除无效评论等。数据清洗后，共保留了 16,689 条有效评论数据。再利用 Python 软件进行中文分词、去除停用词、词性标注等。

3.3. 基于词频统计的特征分析与可视化

在电影影评中出现的高频词，在一定程度上体现出观众所关注的影片的关键因素。对采集的评论数据进行预处理后，利用 Python 运行 Counter 函数输出词频最高的前 100 个关键词及其词频数，并使用 WordCloud 模块绘制电影评论数据的词云图。词语大小反映了其在文本中出现的次数多少和重要程度，揭示了电影观众对电影关注的主要信息。词云图如图 1 所示。

Figure 1. Word cloud map of keywords of movie reviews
图 1. 电影影评关键词词云图

根据词频统计词云图可知，“故事”、“演员”、“导演”、“剧情”、“人物”、“角色”、“喜欢”、“中国”、“女性”、“演技”、“表演”、“不错”、“生活”等词语出现的频率比较高，说明了消费者对于近年电影的好感度比较高，同时也反映了观众的关注主要集中在电影的剧情、导演的能力、人物角色的表演和与中国、女性相关的题材等方面。

4. 电影消费者满意度影响因素的分析

4.1. 消费者满意度影响因素的提取

影评数据里有许多影响观众满意度的属性特征，本文使用 Word2vec 的 CBOW 模型将预处理后的 16,689 条有效评论数据向量化。根据模型的对数据的最终训练得到词向量空间，然后计算词汇的余弦相似度，得出评论数据中的词向量列表。接着本文借助 k-means 聚类算法的 fit 函数将词向量列表中相似度较高或指向同一类对象的词语进行聚类。通过对评论数据的向量化和聚类处理，一共确定了 7 类属性特征，得到电影评论数据的特征词及其对应的相似词词典。将特征词的部分同义词展示如下表 1 所示。

为了比较不同影响因素的相关关系，依据文本聚类的结果总结出了“电影配置”、“电影设计”和“电影效果”三类影响因素，从而构建消费者满意度概念模型。其中，影响因素“电影配置”类包括题材和演员两个特征词，“电影设计”类包括导演、剧情和角色三个特征词，“电影展示”类包括视听和表演两个特征词。

Table 1. Partial feature similarity word list
表 1. 部分特征相似词表

特征词	相似词
剧情	剧情、故事、剧本、情感、发展、扣人心弦、情感表达、情绪、情感共鸣、推进、转折、紧凑、跌宕起伏、引人入胜、悬念、眼眶里、动人
演员	演员、男演员、女演员、阵容、选角、素质、形象、评价、扮演、效应、光环、敬业、才华、魅力、颜值，明星

续表

表演	表演、演技、演出、表现、表演力、精湛、出众、展现、大师、炉火纯青、高超、扎实、到位、不凡、精彩、出色、一流、非凡、华丽、扮演
角色	人物、角色、性格、发展、关系、设定、描述、转变、特点、形态、关注、真真切切、少好、俗人、泪花
导演	编剧、制片人、导演风格、导演风格、拍摄手法、拍摄法、节奏、剪辑、导演执导、执导、影片评价、朝外、表达
题材	中国、女性、时代、动作、爱情、悬疑、科幻、喜剧、战争、惊悚、犯罪、动画、恐怖、纪录片、奇幻、冒险、家庭、历史、儿童、情色、古装、古风、爆笑、生活、主义、旋律、旧时、应景
视听	画面、声音、视觉、听觉、音效、画面、音乐、配乐、影像、视角、音频、感官、观感、色彩、画质、音响、音调、翻新、流水线

4.2. 基于情感分析的消费者满意度文本数据量化

本文通过构建情感词库，基于语义规则对提取的 7 个影响因素进行情感分析，输出电影评论的量化数据。情感词典是一种用于分析文本情感倾向的工具，包含情感词语及其情感倾向。根据情感词语，可以辨认出消费者正面情感倾向(如高兴、表扬)和负面情感倾向(如生气、沮丧)。本研究构建了一个适用于豆瓣电影评论的情感词库，其中包含情感词、副词、否定词的词表。基于构建的情感词库，并结合情感倾向性计算规则，对豆瓣电影评论计算情感得分。

4.2.1. 情感词库的建立

情感分析是指对文本的情感倾向进行识别和判断的过程。首先需要建立情感词库，本研究选取了知网(HowNet)情感词典、台湾大学(NTUSD)情感词典和大连理工大学情感词典为基础的情感词典来构建情感词库。本文构建的情感词库包括以下几个方面。

1) 情感词词表包含正面情感词语和负面情感词语。本研究将所选取的基础情感词典中的情感词进行合并，并根据研究的实际情况添加了一些网络用语作为补充的情感词语，构建了情感词词典。然后对情感词词典中正向的词设权重为 1，负向的词设权重为−1。

2) 副词词表包含情感褒贬强度的词语。例如，“感觉可以了”和“感觉格外可以”、“表演差”和“表演太差”它们的情感程度并不一样。本研究的程度副词词表基于所选取的基础情感词典，根据以往学者的经验，将基础情感词典中五种情感倾向强度的副词赋予了合适的权重。

3) 否定词表用于判断情感倾向。否定词的出现往往会反转一句话的情感倾向例如，“演技不出彩”中的“出彩”是褒义词，但前面加了否定词“不”之后，句子的情感倾向就变成了负面的。此外，有的句子会出现双重否定的情况，需要另外补充到情感倾向判断中。在数据预处理的基础上，本研究基于所选取的基础情感词典中的否定词表，进行合并和整理后构建了本文所需的否定词表。

4.2.2. 情感得分的计算

在完成情感词词表、副词词表、否定词词表的构建后，本文基于词性模板，用正则表达式将评论数据抽取为短句。然后通过构建好的情感词库来计算短句情感值，以及评论数据的情感值。本研究的情感值计算规则参考了郭立秀(2018) [13]在生鲜电商顾客满意度研究中的情感计算规则。以下是几种短句情感值的计算规则：

1) 当识别出短句有特征词但不含有情感词时，将该短句视为中性评价，赋予情感得分为 0。

2) 当识别出短句中只有特征词和有情感词时，视情感词权重为短句情感得分，情感值的得分公式见(1)，其中 F 表示短句的情感值，f 反映了情感词的权重。

$$F = f(x) \tag{1}$$

3) 第三种情况，当识别出短句是特征词 + 副词 + 情感词的组合形式，即包含程度副词句子时，此时计算公式见(2)，其中 n 代表句子中副词的个数，f 代表副词所对应的权值。

$$F = f(x)\prod_{i=1}^{n} A_i \tag{2}$$

4) 当识别出短句含有否定词时，一般出现两种情况，首先是特征词 + 否定词 + 情感词的组合，另外是特征词 + 否定词 + 副词 + 情感词的组合，此时情感得分为各权重相乘，计算公式(3)如下：

$$F = f(x)\prod_{i=1}^{n} A_i \prod_{j=1}^{m} g(y_j) \tag{3}$$

其中，n 代表副词的个数，m 代表否定词的个数，表示否定词权值函数。当 m 的值为奇数时，权值为−1，若为偶数，则权值为 1。依次将对应的权值积累起来就可以计算出短句的情感得分。总结出来的短句可能出现类型如表 2 所示。

Table 2. Short sentence types and examples
表 2. 短句类型及示例

短句可能出现类型	评论观点短句示例
特征词 + 情感词	画面震撼
特征词 + 副词 + 情感词	演技很好
特征词 + 否定词 + 情感词	价格不贵
特征词 + 否定词 + 副词 + 情感词	剧情不是很感人

本文采用 Python 对特征情感词对进行情感分析，样本数据为豆瓣网站的评论文本，程序在之前数据预处理的基础上，利用正则表达式抽取短句，导入所构建的各类词典，将短句依次与特征词表、情感词词表、副词词表、否定词表进行匹配，得出各个匹配成功词表得分，最后综合计算评论数据的情感值。本文将每条评论数据的最终情感值作为该评论数据的总体满意度。部分计算结果如表 3 所示。

Table 3. Partial emotional score calculation results
表 3. 部分情感得分计算结果

序号	评论内容	影响因素 1	得分	影响因素 2	得分	总体满意度
1	邱礼涛今年四涛里最好一涛，动作场面很顶。	演员	3.5	视听	3.5	7
2	文牧野不愧是青年导演的佼佼者，整部电影的画面细腻镜头语言近乎完美！	导演	5	视听	5	10
3	全员演技在线，但导演想表达的东西有点多。	表演	2	导演	-1	1
4	画面很有冲击，作为类型片已无可指摘。	视听	3	题材	1	4
5	乔杉和洪班长看着就很讨喜，演技简直浑然天成。	演员	3	表演	1	4

4.3. 消费者满意度的贝叶斯网络模型

通过量化的评论数据，能了解到消费者在各种指标下对电影的满意程度。本文研究在量化的电影评

论数据的基础上进一步分析，求得各特征因素的重要程度。将处理后的满意度数据导入 Clementine 软件中，借助该软件建立电影消费者满意度的贝叶斯网络模型。通过模型输出结果总结各特征因素在电影消费者满意度中的相对重要程度。

4.3.1. 消费者满意度贝叶斯网络模型的构建

本文借助 Clementine 软件，将整理过的电影消费者满意度数据导入到软件中，对数据进行筛选过滤和类型选择操作，以消费者满意度为输出目标，其余 7 个变量为输入变量，其中输出目标的类型为 set，输入变量的类型为 range。将模型的结构类型选择为 TAN 即增强型朴素贝叶斯树，学习方法选择最大似然法。构建的贝叶斯网络模型如图 2 所示。

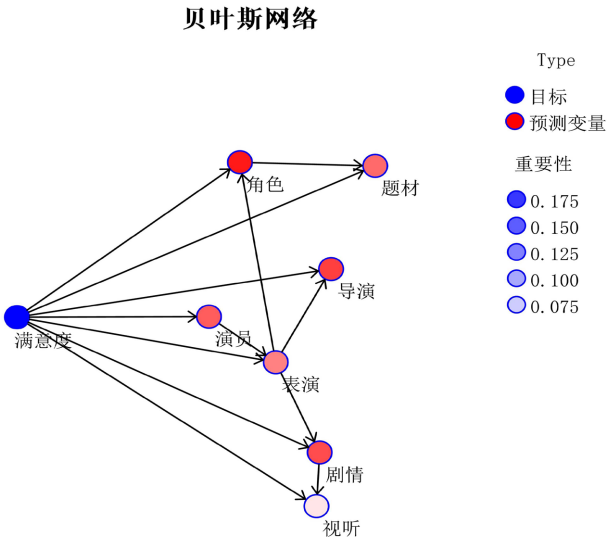

Figure 2. Bayesian network model diagram
图 2. 贝叶斯网络模型图

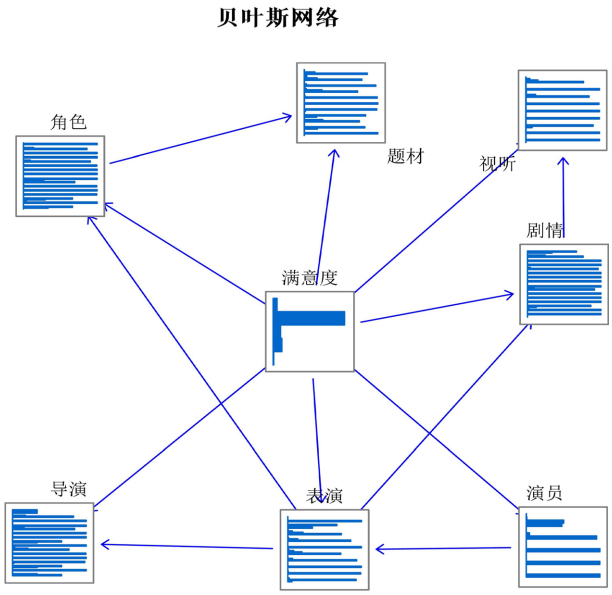

Figure 3. Bayesian network conditional probability distribution diagram
图 3. 贝叶斯网络条件概率分布图

如图 2 所示，最左侧的节点表示输出目标变量，右侧剩余的节点为输入节点的预测变量，预测变量的颜色越红，表明其重要性越高，反之则越低。根据图片可知，预测变量“角色”和“导演”的相对较红，说明其重要性相对其它预测变量更高。通过输出的模型结构图可知，目标变量即满意度通过方向箭头指向 7 个预测变量，说明满意度与所有预测变量都有存在一定程度上的关联，同时预测变量之间也包含某种程度上的联系。

4.3.2. 消费者满意度贝叶斯网络模型的分析

贝叶斯网络条件概率分布图能够展现其中各节点的相关性，更直观地体现各特征因素之间的相互关系。在图 3 中，可以清楚地看到中心的满意度作为目标变量是其余变量的父节点，同时可以发现变量“演员”是变量“表演”的父节点，变量“表演”是变量“角色”、“导演”、“剧情”的父节点，变量“角色”是变量“题材”的父节点，变量“剧情”又是“视听”的父节点，而变量“视听”、“题材”、“导演”无子节点。

根据所得贝叶斯模型得出部分预测变量概率分布如下表 4 所示。从满意度的条件概率可以看出，满意度节点中等级为 4 的概率最大，为 0.777，说明电影消费者对电影的总体满意度较高。

Table 4. Probability distribution table of satisfaction conditions
表 4. 满意度条件概率分布表

满意度	1	2	3	4	5
概率分布	0.005	0.094	0.08	0.777	0.042

接下来分别对“角色”、“导演”和“剧情”三个方面的条件概率进行分析。从“角色”的条件概率(表 5)可以看出，满意度和“表演”都是“角色”的父节点。在满意度节点为 4 的情况下，“角色”节点得分在−18.3~22.05 的概率最高，最高值为 22.05，此时概率值为 0.995，说明电影消费者对电影的角色塑造比较满意。

Table 5. Role conditional probability table
表 5. 角色条件概率表

父级节点		概率分布				
表演	满意度	≤−58.65	−58.65~−18.3	−18.3~22.05	22.05~62.4	>62.4
≤−11.9	1	0	0.333	0.666	0	0
≤−11.9	2	0	0	1	0	0
−11.9~2.2	1	0.039	0.294	0.666	0	0
−11.9~2.2	2	0	0.004	0.995	0	0
−11.9~2.2	3	0	0	1	0	0
−11.9~2.2	4	0	0	0.995	0.004	0
−11.9~2.2	5	0	0	0.697	0.227	0.024
2.2~16.3	1	0	1	0	0	0
2.2~16.3	2	0	0	1	0	0
2.2~16.3	3	0	0	1	0	0
2.2~16.3	4	0	0	0.995	0.004	0
2.2~16.3	5	0	0	0.909	0.09	0

续表

16.3~30.4	2	0	0	1	0	0
16.3~30.4	4	0	0	1	0	0
16.3~30.4	5	0	0	0.862	0.137	0
>30.4	5	0	0	0	1	0

从“导演”的条件概率表(表 6)可以看出，满意度和“表演”都是“导演”的父节点。由表可知，“导演”节点的主要得分在 22.05 以下。当父级节点满意度为 2 到 5 时，“导演”节点得分在−58.65~−18.3 的概率最高，最高值为−18.3，此时子节点“导演”各个概率值都接近于 1，说明电影消费者对电影导演水平的整体满意度比较适中。

Table 6. Director conditional probability table
表 6. 导演条件概率表

表演	满意度	≤−58.65	−58.65~−18.3	−18.3~22.05	22.05~62.4	>62.4
≤−11.9	1	0.666	0.333	0	0	0
≤−11.9	2	0	1	0	0	0
−11.9~2.2	1	0.333	0.666	0	0	0
−11.9~2.2	2	0.017	0.982	0	0	0
−11.9~2.2	3	0	1	0	0	0
−11.9~2.2	4	0	0.996	0.003	0	0
−11.9~2.2	5	0	0.765	0.206	0.028	0
2.2~16.3	1	0	1	0	0	0
2.2~16.3	2	0.333	0.666	0	0	0
2.2~16.3	3	0	1	0	0	0
2.2~16.3	4	0	0.997	0.002	0	0
2.2~16.3	5	0	0.844	0.155	0	0
16.3~30.4	2	1	0	0	0	0
16.3~30.4	4	0	1	0	0	0
16.3~30.4	5	0	0.827	0.172	0	0
>30.4	5	0	0.333	0.333	0	0.333

Table 7. Plot conditional probability table
表 7. 剧情条件概率表

父级节点		概率分布			
表演	满意度	≤−110.25	−110.25~−22.5	−22.5~36	>36
≤−11.9	1	0	0	1	0
≤−11.9	2	0	0	1	0
−11.9~2.2	1	0.019	0.117	0.862	0

续表

−11.9~2.2	2	0	0	1	0
−11.9~2.2	3	0	0	1	0
−11.9~2.2	4	0	0	0.999	0
−11.9~2.2	5	0	0	0.991	0.088
2.2~16.3	1	0	1	0	0
2.2~16.3	2	0	0	1	0
2.2~16.3	3	0	0	1	0
2.2~16.3	4	0	0	1	0
2.2~16.3	5	0	0	0.961	0.038
16.3~30.4	2	0	0	1	0
16.3~30.4	4	0	0	1	0
16.3~30.4	5	0	0	0.758	0.241
>30.4	5	0	0	0.333	0.666

从“剧情”的条件概率表(表 7)可以看出，满意度和“表演”都是“剧情”的父节点。在满意度节点为 2~4 时，“剧情”节点得分在−22.5~36 的得分的概率最高，最高值为 36，此时各个概率值都接近 1，说明电影消费者对电影剧情的满意度比较高，极少消费者表示不满意。

进一步对贝叶斯网络模型的预测变量重要性进行排名。从图 4 可以看出，电影消费者最为在意的前 3 个因素依次是“角色”、“导演”、“剧情”，接下来是“演员”、“题材”、“表演”，最后是“视听”，所以可知电影的角色塑造对观众满意度的影响最大，而电影视听表现对满意度的影响最小。

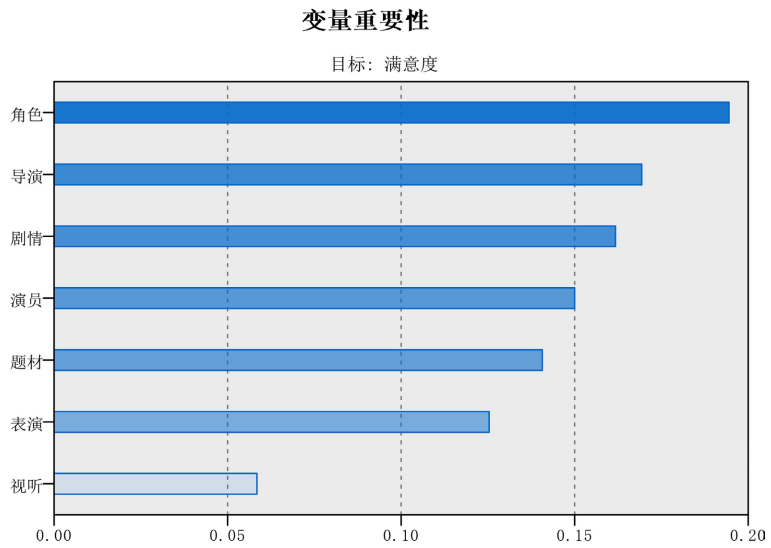

Figure 4. Predictive variable importance ranking
图 4. 预测变量重要性排名

为了更详细地了解各特征因素对满意度的作用程度，得出各个子节点重要性得分表(表 8)，预测变量重要性得分表中的每个得分都代表一个变量的重要性程度，该得分越高，代表该变量对目标变量的影响

越大，在该表中各个预测变量的得分之和通常为 1。由表 8 可知，在此次构建的贝叶斯网络模型中，最重要的预测变量为“角色”，其作用程度为 0.1945。其次是“导演”、“剧情”、“演员”、“题材”、“表演”，重要性分别为 0.1693、0.1618、0.15、0.1407 和 0.1254，说明消费者评判电影的好坏时主要看重角色展现、导演的技术水平、剧情好坏、演员阵容、电影题材和演员的表演水平，这些因素对电影消费者满意度的影响程度较高。最后是“视听”，电影的视听效果重要性为 0.0585，得分最低，说明该因素对电影消费者满意度的影响程度较低。在电影消费者满意度分析中，角色展现的影响程度最高，可能是因为角色展现是影片中最基础的要素之一，电影制作的目的之一就是通过角色的塑造来吸引观众。如果电影中的角色形象鲜明、性格突出、情感真挚，观众会更容易对这些角色产生共鸣，从而更容易产生情感投入感和情感参与感，提高对电影的喜爱度和满意度。除此之外，导演、剧情、演员、题材和表演这几个因素也对消费者评价产生影响，这几个因素的重要性差异程度较小，这也说明消费者对于电影的偏好视角越来越多元，不再只因为某些特殊原因，比如粉丝为了追捧喜爱的演员明星而无差别地选择观影，观众更愿意从多个角度全方面地对电影质量进行评价。近几年我国电影行业不断涌现出一些新主题的电影，如女性主义题材、爱国主义题材、国漫等电影作品，影片质量也有了极大的提高，这有助于促进我国电影事业的繁荣与发展，影响因素分析结果符合当前电影行业发展趋势。而视听效果对电影消费者满意度的影响程度相对较低，可能是因为现代电影技术日益发达，视听效果已经成为了电影制作中的基本要素，一般来说，观众已经对视听效果有了较高的期待和预期，这样就导致了视听效果提升对于观众满意度的提高并不是很明显。当然，如果某部电影的视听效果偏差较大或者远低于预期，则会对观众的满意度产生较大的负面影响。总体而言，在各特征因素影响程度中，角色展现 > 导演水平 > 剧情好坏 > 演员阵容 > 电影题材 > 演员的演出水平 > 电影的视听效果。

Table 8. Predictive variable importance score
表 8. 预测变量重要性得分

节点	重要性
角色	0.1945
导演	0.1693
剧情	0.1618
演员	0.15
题材	0.1407
表演	0.1254
视听	0.0585

5. 结论与建议

本文以豆瓣电影评论为研究对象，通过文本挖掘的相关方法，对电影消费者满意度的因素构成进行探索。通过对评论高频词的分析，发现消费者的关注主要集中在电影的剧情、导演的能力、人物角色的表演和与中国、女性相关的题材等方面。对评论数据进行向量化和聚类，将特征因素归类成“电影配置”、“电影设计”、“电影表现”三大方面，并据此构建电影消费者满意度概念模型。通过构建情感词库，运用语义分析规则进行情感分析，计算各特征因素的得分和满意度评价，基于情感分析的结果构建消费者满意度影响因素的贝叶斯网络模型，得到贝叶斯网络模型结构图、条件概率分布图和预测变量影响程度的排名。结果表明，预测变量影响程度的排名结果中各个特征因素和其对应系数的排名为：角色(0.1945)、

导演(0.1693)、剧情(0.1618)、演员(0.15)、题材(0.1407)、表演(0.1254)、视听(0.0585)。基于上述结论，本研究从电影配置、电影设计、电影表现三个方面提出优化建议。

首先，为了实现更加长远的发展，电影出品方必须从大局入手，从消费者的关注点入手。根据贝叶网络模型的分析结果可知，消费者在评判电影时注重电影的题材类型、注重电影的演员阵容选择。对此，电影出品方应当充分调研市场需求，了解目标受众的喜好和兴趣点的电影题材，明确电影的类型定位。在演员选取时，可以研究演员过往作品的票房表现和影评口碑，了解其受众群体，可以考虑演员之间的化学反应和配合度，营造和谐的表演氛围。其次，作为电影剧本设计者，导演需要对艺术有敏锐的感知力，能够准确地把握剧本的情感和主题，并通过镜头语言和表演来传达给观众，同时需要与演员、摄影师、剪辑师等各个创作团队成员进行有效的沟通和协调，合理安排拍摄进度，有效利用时间资源，对细节有敏锐的观察力，能够捕捉到人物情感和场景细节，并通过镜头语言来表达。最后，演员应该从各方面提升情感表达、身体语言、语音表达等方面的能力。而电影团队也应该加强视觉表现输出，这包括摄影、美术设计、服装和化妆等方面的精细工艺，能够使观众在视觉上感到享受。通过合理的构图和精心设计的镜头运动增强观众的观影体验，结合音效设计和配乐，增强电影的氛围和情感表达，让观众更加沉浸在电影的世界中。

参考文献

[1] Statista Digital Market Outlook (2020) eCommerce Report 2020. https://www.statista.com/markets/

[2] 国家图书馆研究院. 中国互联网络信息中心发布第 52 次《中国互联网络发展状况统计报告》[J]. 国家图书馆学刊, 2023, 32(5): 13.

[3] 陆佳佳, 刘汉文. 2018 年中国电影产业发展分析报告[J]. 当代电影, 2019(3): 13-20.

[4] “十四五”中国电影发展规划[N]. 中国电影报, 2021-11-17(002).

[5] Evanschitzky, H., Iyer, G., Hesse, J. and Ahlert, D. (2004) E-Satisfaction: A Re-Examination. *Journal of Retailing*, **80**, 239-247. https://doi.org/10.1016/j.jretai.2004.08.002

[6] 徐小琳. 患者对医疗决策参与的满意度量表的编制及信效度考评[D]: [硕士学位论文]. 长沙: 中南大学, 2010.

[7] 李燕飞. 在线评论对消费者满意度及商品销量的影响研究[D]: [硕士学位论文]. 广州: 广东工业大学, 2016.

[8] 张璇. 基于网络文本挖掘的游客满意度研究[D]: [硕士学位论文]. 济南: 山东大学, 2023.

[9] 安翔, 李世鑫, 白雪, 杜禹墨. 北大荒米业竞争对手产品评论数据挖掘[J]. 北方经贸, 2018(8): 44-47.

[10] 严军超, 赵志豪, 赵瑞. 基于机器学习的社交媒体文本情感分析研究[J]. 信息与电脑(理论版), 2019, 31(20): 44-47.

[11] 赵志杰, 刘岩, 张艳荣, 周婉婷, 孟令跃. 基于 Lasso-LDA 的酒店用户偏好模型[J]. 计算机应用与软件, 2021, 38(2): 19-26.

[12] 辛雨璇, 王晓东. 基于文本挖掘的电影评论情感分析研究[J]. 牡丹江师范学院学报(自然科学版), 2021(1): 25-28.

[13] 郭立秀. 基于文本挖掘的生鲜电商顾客满意度研究[D]: [硕士学位论文]. 成都: 西南交通大学, 2018.