

基于情感分类的《热辣滚烫》LDA 主题挖掘

关瑞雪¹, 关瑞勇²

(1. 石家庄铁道大学管理学院, 石家庄 050000; 2. 河北农业大学文管学院, 沧州 061000)

摘要: 以豆瓣电影网关于电影《热辣滚烫》的用户评论为数据源,采用情感分类和 LDA(Latent Dirichlet allocation)主题模型对影评进行深入分析。首先对影评数据进行预处理,包括去重、分词和构建停用词表,匹配情感词表和修正情感倾向,将评论分为正面和负面两类。绘制正面和负面情感词云,揭示了观众对电影中女性力量的展现、励志元素的融入及幽默感表达等的正面评价,以及对剧情创新不足、营销宣传过度等负面反馈。LDA 主题分析进一步识别“励志人生”和“重塑自我”作为正面主题,以及“过度营销”和“抄袭翻拍”作为负面主题。研究结果为电影制作提供了宝贵的观众洞察,建议制作团队在未来作品中注重剧情创新、真实营销、角色深度塑造和文化传播等,从而提升电影的艺术价值和市场竞争力。

关键词: 情感分类; 主题挖掘; 重塑自我

中图分类号: G271.4 **文献标志码:** A **文章编号:** 1671-1807(2024)16-0219-05

在当今数字化时代,电影的评论数据成为理解市场反馈和观众情感倾向的重要资源。黄剑波等^[1]运用自然语言处理技术,挖掘用户影评情感倾向,将影评情感值与用户评分结合,预估模型预测点击率,为用户提供电影推荐服务。姜霖和张麒麟^[2]利用评论挖掘技术发掘用户的偏好,构建了电影的推荐策略;谢治海等^[3]、崔为之^[4]基于影评情感类型结合票房预测模型,进行电影票房的预测。旷开金等^[5]运用语义网络图、LDA 模型对《我和我的祖国》进行情感分析,结果表明观众对电影持力荐态度,整体评论积极且正面;杨秀璋等^[6]基于 LDA-BiLSTM 模型和知识图谱融合的影评文本挖掘方法精准预测影评的情感类别。贾晨^[7]则聚焦电影《夺冠》从语义、情感及内容主题 3 个维度对网络影评进行实证分析,探寻受众对体育电影的关注焦点与态度取向。当前研究学者主要集中在基于影评情感倾向的电影推荐、面相情感的票房预测、影评情感分类与分析等方面,针对电影评论的情感主题探究较少。电影《热辣滚烫》自上映以来,便引发观众的广泛讨论。其评论涵盖从剧情到演技、主题升华再到营销策略等多个方面,反映观众对电影的全面评价。故本文选取豆瓣网电影《热辣滚烫》的用户评论作为分析对象,旨在通过情感分类和主题模型深

入探讨观众对于该电影的看法和情感反应。

1 研究设计

1.1 数据采集与处理

爬取豆瓣电影网关于《热辣滚烫》的电影短评,由于网站限制,只能爬取到 660 条评论,因此将 660 条影评作为本文的研究对象。在获取影评数据的基础上,进行去重、分词、去除停用词,提取名词、动词、动名词等预处理工作。考虑到影评采集的时候具有重复性的评论以及标点符号,故对其进行去重及标点符号的删除等操作,采用结巴中文分词工具包对影评数据进行分词处理,将影评分成单个的词语,形成新的数据文档^[8]。为了消除部分无关词汇对研究的影响,结合哈尔滨工业大学停用词表和研究内容,构建停用词表,剔除如“为了”“如果说”等与研究主题无关的词语,提取分词中的名词、动名词及动词作为研究对象,对主题进行分析。

1.2 情感分类

1.2.1 匹配情感词

情感倾向反映观众对电影的喜好程度,表现为正面情感、负面情感或中性情感。进行评论情感倾向的分析需要识别和匹配情感相关的词汇,以知网提供的“情感分析用词语集”为情感词参考^[9]。将“中文正面评价”与“中文正面情感”两个词表合并,

收稿日期: 2024-04-08

作者简介: 关瑞雪(1998—),女,河北邯郸人,硕士研究生,研究方向为管理科学与工程、文本挖掘;关瑞勇(2005—),男,河北邯郸人,研究方向为物流管理、文本挖掘。

将每个词汇的初始权重值设定为1,构建正面评论情感词表。同样合并“中文负面评价”与“中文负面情感”词表,将每个词汇的初始权重值设定为-1,形成负面评论情感词表^[10]。对评论文本中的情感倾向进行量化分析,准确地理解和评估观众对于电影的情感反应。

在基于词表的情感分析方法中,分析的准确性取决于情感词表的质量和适用性。对于特定研究场景的适应性,应对词表进行适当的调整和扩充。针对本文场景,在知网提供的词表基础上进行优化,例如“好评”“差评”“五星”等词只有在电影评论中出现,根据词语的情感倾向添加至对应的情感词表内。将“力量感”“真善美”“意志力”“重塑”等词添加进正面情感词表。将“差评”“难看”“翻拍”“抄袭”等词加入负面情感词表。

1.2.2 修正情感倾向

情感倾向的判定依赖于对否定词的识别和处理。在汉语中,否定词的使用会影响情感词的极性,即情感的正负面。在存在多重否定的情况下,情感极性的判断需要更加细致和复杂,为了准确地判断情感值,本文载入否定词表,处理否定修饰词,对情感值的方向进行修正^[11]。计算每条评论的情感得分,根据修正后的情感值对评论进行分类,将其归为正面评论或负面评论,从而得到更加准确的情感分析结果。否定词表包括但不限于“不”“没”“无”“非”“莫”“弗”“毋”“未”“否”“别”“無”“休”“不是”“不能”“不可”“没有”“不用”“不要”“从没”“不太”等。

1.3 LDA 主题模型

1.3.1 LDA 主题模型

隐含狄利克雷分布主题模型简称 LDA(latent Dirichlet allocation),是一种广泛使用的统计方法,它通过算法对文本数据进行深入的语义分析,并在此基础上进行文本的自动聚类,以识别和分类潜在的文本主题。LDA 与潜在语义分析模型(latent semantic analysis, LSA)和概率潜在语义分析模型(probabilistic latent semantic analysis, PLSA)等其他主题模型方法相比,具有其独特的优势和应用场景。LDA 模型构建了一个包含文档层、主题层和词汇层三个层次的结构^[12]。在这个模型中,每个主题都与一个关键词的概率分布相关联,而每篇文档则与一个主题的概率分布相对应。LDA 的核心在于通过这种层次结构揭示文本中的主题分布和词汇分布。LDA 模型的建立包括文本数据分词和清洗,确定最合适的

主题数量,提取与各个主题相关的核心词汇,实现对文本数据的降维处理,并利用可视化工具展示聚类结果。结合文档中主题的分布情况和词汇的语义信息,实现对文本的主题的有效分类。

1.3.2 寻找最优主题数

采用基于相似度的自适应方法来选择最优的 LDA 模型,可以在无须人工干预的情况下自动确定最合适的主题数量,并进行深入的主题分析^[13]。这种方法通过迭代过程,有效地寻找到最佳的模型结构,具体实施步骤如下:①首先设定一个初始的主题数目 k ,利用此值 k 训练出一个初始的 LDA 模型;②对于得到的各个主题,计算它们之间的相似度,通常采用平均余弦距离作为度量;③根据计算出的相似度,调整主题数目 k ,可能是增加或减少,然后重新训练 LDA 模型;④使用新的 k ,再次计算各主题间的相似度;⑤重复步骤②~步骤④,直到找到一个使得主题间相似度最优化的 k 。

在这个过程中,从词频出发,通过计算不同主题间词汇的相似度来评估主题内容的接近程度。词汇越相似,表明主题内容越相关。通过 LDA 模型识别出在不同主题数量下的关键主题词,并将这些主题词汇集成一个统一的集合。构建任意两个主题间的词频向量,并计算这两个向量的余弦相似度。余弦相似度的值越高,表示这两个主题之间的内容越相似。通过对所有主题数的余弦相似度取平均评估不同主题数量下模型的优劣,并据此确定最优的主题数目^[14]。它通过自动化的迭代过程,减少人工调整主题数量的需要,同时确保主题分析的质量和效率。通过这种方式能够更加精确地捕捉文本数据中的潜在主题结构,为后续的文本挖掘和分析工作提供坚实的基础。

2 正负情感云

通过匹配情感词、修正情感倾向的代码操作之后,提取正负面评论信息,利用运用 Text Blob 中的 Word Cloud 子包对预处理后的数据进行词云绘制,出现次数较高的词汇以可视化的方式呈现。借助词云图能高效地排除低频低质的信息,弱化非关键区域,突出呈现关键信息,便于更好地进行情感主题分析。

2.1 正面情感词云

正面情感词云如图1所示,深刻地揭示了观众对这部电影的积极反响,其中几个核心元素尤为突出,如女性力量的展现、励志元素的融入以及幽默感的表达等。

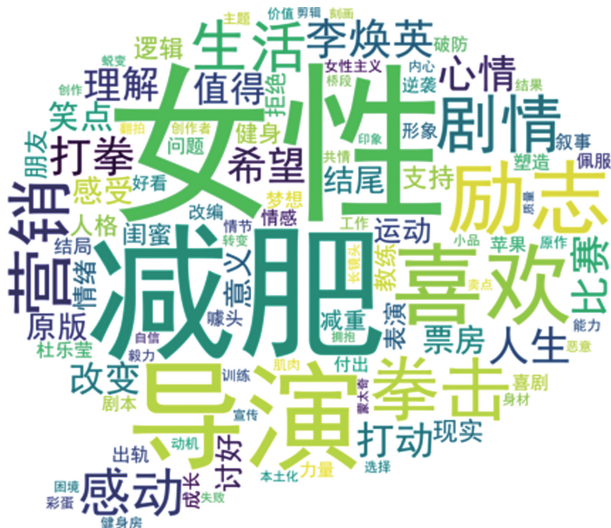


图 1 正面情感词云

女性力量的展现是这部电影获得正面评价的关键因素之一。目前女性角色的独立和力量越来越受到重视,而《热辣滚烫》恰恰捕捉到了这一社会趋势。电影中的女性角色不仅是情感的载体,更是行动的推动者,展现了女性在面对生活挑战时的坚韧和智慧。其次,励志元素的融入为电影增添了深度和感染力。电影通过角色的奋斗和成长,传递了积极向上的信息,激发了观众的希望和梦想。这种励志的表达不仅限于表面的成功和成就,更深入地探讨了个人价值的实现和自我超越的可能性。观众在角色的逆袭和付出中看到了自己的影子,感受到了电影传递的正能量,这种情感上的共鸣极大地增强了电影的吸引力。再者,幽默感的表达是电影吸引观众的另一大法宝。《热辣滚烫》巧妙地将幽默元素融入剧情之中,既为观众提供了欢笑,又增加了剧情的趣味性。电影中的笑点等关键词,反映了观众对这种幽默表达方式的喜爱和认可。

此外,电影在其他方面也获得了观众的好评。剧情的逻辑性和叙事结构的清晰,使得观众能够顺畅地跟随故事的发展,享受观影过程。角色的形象和演员的表演也得到了观众的认可,这些角色的塑造和演绎增添了电影的艺术魅力。这部电影不仅是一部艺术作品,更是一面镜子,映照出当代社会的价值观念和观众的情感需求。

2.2 负面情感词云

负面情感词云如图 2 所示,针对电影《热辣滚烫》的负面反馈主要集中在减肥营销的策略、翻拍作品的接受度,以及纪录片元素的运用。词云中的“营销”“减肥”两词凸显了观众对于电影宣传手法



图 2 负面情感词云

的不满。电影营销团队采用贾玲为角色减重的事迹作为宣传适得其反,营销活动过度夸大了电影内容与质量。减肥作为营销点被视为一种噱头,而非电影艺术价值的真实体现。翻拍作品的接受度问题在词云中的“抄袭”等词汇中得到体现。电影的改编缺乏创新、未能忠实于原作,这导致观众对电影的整体评价降低。最后,纪录片元素的运用在词云中通过“纪录片”一词显现。纪录片风格与真实性和深度探索相关联,电影试图融入纪录片元素,没有达到预期的深度,在艺术表达上存在一定的缺陷。这些方面的不足导致观众的不满和批评,为电影制作方提供了改进的方向,以便在未来的作品中更好地满足观众的期待。

3 基于情感分类的 LDA 主题分析

3.1 确定最优主题数

在进行 LDA 主题挖掘时,最优的主题数直接影响到模型的解释性和预测能力。分析正面评论和负面评论的平均余弦相似度图形,评估每个主题的内聚性和差异性,确定最优主题数。余弦相似度表示主题间的距离,余弦相似度越高则表明主题间重叠较多,反之,余弦相似度越低,模型的分类效果就越好^[15]。

如图 3(a)所示,正面评论的平均余弦相似度图形随着主题数的增加,相似度呈现先上升后稳定的趋势。在主题数较少时,相似度较低,这可能是因为较少的主题数无法充分捕捉评论中的多样性,导致同一主题内的评论内容差异较大。随着主题数的增加,相似度逐渐提高,表明主题内的评论内容更加一致。当主题数达到一定数量后,相似度增长趋缓,这意味着增加的主题数并没有带来额外的信

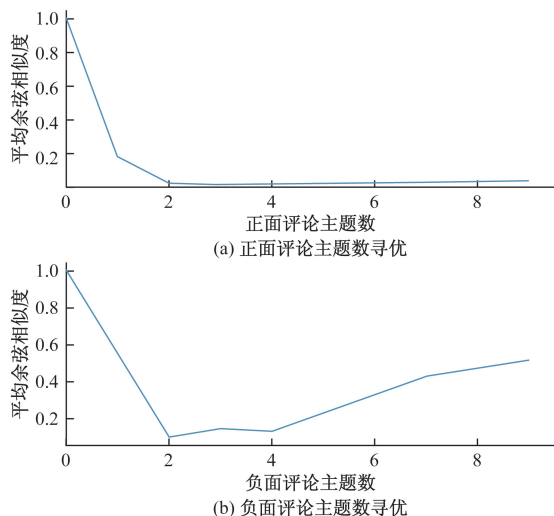


图3 正负面评论余弦相似度

息,而是增加了模型的复杂性。在主题数为2时,正面评论的余弦相似度最小,故确定正面评论的主题数为2。如图3(b)所示,负面评论的平均余弦相似度图形呈现先下降后缓慢上升的趋势。负面评论的相似度在较少主题数时下降得更快,这表明负面评论的内容更加集中和一致,需要较少的主题就能达到较高的内聚性。主题数为2时,主题之间的相似性最低,故确定负面评论的主题数也为2。

3.2 正负主题情感分析

通过LDA模型提取出两个正面主题和两个负面主题,如表1所示,正面主题“励志人生”中,关键词如“减肥”“励志”“拳击”“笑点”“理解”“现实”“希望”等,揭示了电影在传递积极向上信息方面的成功。观众在电影中找到了激励人心的元素,特别是主角面对生活挑战时的奋斗和努力,这种正能量的传递与观众产生了共鸣。电影通过展现主角的励志故事,激发了观众对更好生活的向往和追求,这种情感上的共鸣是电影获得正面评价的重要因素。“重塑自我”主题下的词汇,如“女性”“导演”“减肥”“感动”“改变”“闺蜜”“意义”等,指向了电影中关于个人成长和自我提升的探讨。通过主角的故事,深入挖掘了女性自我实现和自我价值的主题,观众对于电影中展现的女性力量和独立精神表示赞赏,这反映了现代社会对性别平等和女性自主权的重视。同时,电影中对个人成长和变化的描绘,也触动了观众对自我发展和生活改善的渴望。

然而,负面主题“过度营销”揭示了观众对电影宣传策略的不满。关键词“营销”“减肥”“恶心”“失败”“讨好”“噱头”“翻拍”“纪录片”“宣传”等,反映

表1 正负情感主题抽取结果

主题倾向	主题概括	主题词
正面主题	励志人生	减肥、励志、导演、女性、剧情、营销、拳击、笑点、理解、现实、李焕英、生活、逻辑、票房、希望、教练、叙事、减重、人生、表演、情绪、比赛、情感、健身、健身房、喜剧、工作、情节、卖点、彩蛋
	重塑自我	女性、导演、减肥、喜欢、感动、打拳、原版、运动、值得、打动、改变、闺蜜、意义、问题、塑造、结尾、剧本、改编、讨好、支持、朋友、人格、佩服、动机、逆袭、结局、拒绝、共情、心情、出轨
负面主题	过度营销	减肥、营销、拳击、恶心、剧情、小品、破防、失败、讨好、理解、恶意、自恋、问题、噱头、翻拍、纪录片、导演、工作、烂片、花钱、宣传、日本、改编、身材、对手、拥抱、摆烂、浪费时间、环境、长镜头
	抄袭翻拍	减肥、营销、难看、纪录片、改编、原版、恶心、宣传、拳击、小品、赚钱、剧情、噱头、反感、煽情、动机、理解、抄袭、花钱、套路、打拳、翻拍、刺激、情节、生活、对手、低分、逻辑、选择、剧本

观众认为电影过分依赖外在的宣传手段,忽视了电影本身的质量和內容。特别是将减肥作为营销点,被观众视为一种噱头,引起了观众的反感,忽视了电影艺术价值。此外,“抄袭翻拍”主题下的关键词,如“难看”“纪录片”“改编”“原版”“恶心”“宣传”“抄袭”“套路”等,指出了观众对电影原创性缺失的批评。观众认为电影在剧情、风格、表现手法上过于模仿《百元之恋》,导致对电影的整体评价降低。综合来看,观众对《热辣滚烫》的评价呈现出明显的两极分化。正面主题反映了电影在励志和个人成长方面的成功,而负面主题则揭示了观众对于电影营销手段和原创性缺失的担忧。

4 结论与建议

4.1 结论

(1)本文爬取电影《热辣滚烫》的影评,对文本数据进行预处理、情感分类绘制词云,通过词云呈现,发现情感分类的效果很好,从正面情感和负面情感两个方面进行了解析。通过正负情感词云分析发现电影《热辣滚烫》在女性力量的展现、励志元素的融入、幽默感的表达以及叙事逻辑、演员表演等方面都得到了观众的积极评价,而在剧情创新、营销宣传、制作质量等方面未能满足观众的期待。

(2)通过LDA主题分析进一步确定电影热辣滚烫的正面主题和负面主题,正面主题突显了影片在激励人心和重塑自我方面的积极作用,而负面主

题则暴露出观众对电影的营销策略和创新性不足的不满。

4.2 建议

通过对电影《热辣滚烫》的影评数据分析,制作团队在翻拍作品时需要在尊重原作的基础上,提供新颖的元素和深入的解读,避免陷入陈规旧套,同时在营销上保持真实性,以免过高的期望导致观众失望。在角色塑造方面,应注重多维性和背景深度,选择合适的演员并给予他们表现的空间。此外,提升电影的整体制作质量,注重视觉效果、音效设计和剪辑节奏等细节,以及深化主题探讨和文化元素的融合,将电影打造成文化传播的载体。最后,制作团队应密切关注并重视观众反馈,持续改进电影的各个方面,以满足观众期待,提升电影的艺术价值和市场竞争力,赢得观众的广泛认可。

参考文献

- [1] 黄剑波,陈方灵,丁友东,等. 基于情感分析的个性化电影推荐[J]. 计算机技术与发展, 2020, 30(9): 132-136.
- [2] 姜霖,张麒麟. 基于评论情感分析的个性化推荐策略研究——以豆瓣影评为例[J]. 情报理论与实践, 2017, 40(8): 99-104.
- [3] 谢治海,朱敏,牛红宇,等. 面向票房预测的影评情感可视分析[J]. 计算机应用研究, 2020, 37(10): 2945-2950.
- [4] 崔为之. 基于消费意图和情感分析的电影票房预测[D]. 湘潭:湘潭大学, 2020.
- [5] 旷开金,廖海琳,裴文庆. 基于文本挖掘的影评数据情感分析——以《我和我的祖国》为例[J]. 武夷学院学报, 2022, 41(5): 68-74.
- [6] 杨秀璋,武帅,廖文婧,等. 基于 LDA-BiLSTM 模型和知识图谱的电影影评文本挖掘研究[J]. 现代计算机, 2023, 29(8): 12-19.
- [7] 贾晨. 基于电影《夺冠》网络影评的情感分析与数据挖掘研究[J]. 南京体育学院学报, 2023, 22(12): 16-22.
- [8] 蔡文乐,周晴晴,刘玉婷,等. 基于 Python 爬虫的豆瓣电影影评数据可视化分析[J]. 现代信息科技, 2021, 5(18): 86-89.
- [9] DANYAL M M, KHAN S S, KHAN M, et al. Sentiment analysis of movie reviews based on NB approaches using TF-IDF and count vectorizer[J]. Social Network Analysis and Mining, 2024, 14(1): 1-15.
- [10] ALAMIN Z, LOROSAE T A, RAMADHAN S, et al. Improving performance sentiment movie review classification using hybrid feature TFIDF, N-gram, information gain and support vector machine[J]. Mathematical Modelling of Engineering Problems, 2024, 11(2): 36-51.
- [11] 尤众喜,华薇娜,潘雪莲. 中文分词器对图书评论和情感词典匹配程度的影响[J]. 数据分析与知识发现, 2019, 3(7): 23-33.
- [12] 林轶,曹清芳. 基于 LDA 模型及情感分析的电影主题公园评论文本研究[J]. 科技和产业, 2023, 23(22): 93-98.
- [13] 李庆真,唐焱. 基于主题模型的社交媒体倦怠行为研究[J]. 杭州电子科技大学学报(社会科学版), 2022, 18(5): 54-59.
- [14] 余亚烽,刘兴红,陶胜阳,等. 基于情感分类与主题挖掘的 MOOC 课程评论研究[J]. 考试研究, 2024, 20(1): 86-100.
- [15] 郭佳怡,方博平,陆欣怡,等. 基于文本挖掘和情感分析方法的“智慧旅游”服务质量感知研究[J]. 现代信息科技, 2023, 7(6): 1-5.

“Hot and Hot” LDA Theme Mining Based on Emotional Classification

GUAN Ruixue¹, GUAN Ruiyong²

(1. School of Management Shijiazhuang Tiedao University, Shijiazhuang 050000, China;

2. School of Culture and Management, Hebei Agricultural University, Cangzhou 061000, Hebei, China)

Abstract: Taking user reviews from Douban Movie website regarding the film “Hot and Spicy” as the data source, sentiment classification and LDA(Latent Dirichlet Allocation) topic model was employed for an in-depth analysis of the movie reviews. Firstly, data was preprocessed, which included deduplication, word segmentation, construction of a stop word list, matching with a sentiment lexicon and correction of sentiment tendencies, categorizing the reviews into positive and negative sentiments. Mapping the positive and negative emotional word clouds reveals the audience’s positive comments on the display of female power, the integration of inspirational elements, and the expression of humor in the film, as well as negative feedback on the lack of plot innovation and excessive marketing publicity. Further LDA topic analysis identifies “Inspirational Life” and “Self-Reconstruction” as positive themes, and “Over-Marketing” and “Plagiarism and Remakes” as negative themes. The results provide valuable audience insights for film production, suggestions are proposed that production teams should focus on plot innovation, authentic marketing, in-depth character portrayal, and cultural dissemination in future works to enhance the artistic value and market competitiveness of the film.

Keywords: sentiment classification; topic mining; self-reconstruction