Journalism and Communications 新闻传播科学, 2024, 12(1), 23-28
Published Online February 2024 in Hans. https://www.hanspub.org/journal/jc
https://doi.org/10.12677/jc.2024.121004

基于文本挖掘的豆瓣电影评论的LDA主题模型分析

——以电影《让子弹飞》为例

唐　诗

重庆师范大学新闻与传媒学院，重庆

收稿日期：2023年12月8日；录用日期：2024年1月26日；发布日期：2024年2月2日

摘　要

网络评论的兴起，让普罗大众都能参与到对电影的解读和评价中来，撰写影评不再是专业影评人所独有的一项活动。电影网络评论作为对电影作品的评价，既是观众对电影好坏的直观反映，也是其观影体验的直接体现。对网络影评的分析有助于直接了解观众对电影的观感和评价，进而加深对经典电影大火基因的理解，并为电影制作提供相应借鉴。电影《让子弹飞》上映十三年后仍被奉为经典与神作，片中的经典桥段和台词一直被人们所津津乐道。本文通过对该部影片豆瓣影评的文本挖掘和LDA主题模型分析，探究观众评价中对影片的关注点和评论视角，基于高频主题词，挖掘深层主题，有助于客观整体评价电影和反映观众的真实感受，寻找该片持续大火的传播基因与密码。

关键词

豆瓣电影评论，文本挖掘，LDA主题模型，《让子弹飞》

Analysis of LDA Theme Model for Douban Movie Reviews Based on Text Mining

—Taking the Movie “Let the Bullets Fly” as an Example

Shi Tang

School of Journalism and Media, Chongqing Normal University, Chongqing

Received: Dec. 8th, 2023; accepted: Jan. 26th, 2024; published: Feb. 2nd, 2024

文章引用: 唐诗. 基于文本挖掘的豆瓣电影评论的 LDA 主题模型分析[J]. 新闻传播科学, 2024, 12(1): 23-28.
DOI: 10.12677/jc.2024.121004

Abstract

The rise of online reviews has enabled the general public to participate in the interpretation and evaluation of movies, and writing reviews is no longer a unique activity for professional film critics. Film online reviews, as an evaluation of film work, are not only an intuitive reflection of the quality of the movie by the audience, but also a direct reflection of their viewing experience. The analysis of online film reviews helps to directly understand the audience's perception and evaluation of movies, deepen their understanding of the genes of classic movies, and provide corresponding references for film production. The movie "Let the Bullets Fly" is still regarded as a classic and masterpiece thirteen years after its release, and the classic scenes and lines in the film have always been talked about by people. This article explores the focus and perspective of audience evaluation on the film through text mining and LDA theme model analysis of Douban film reviews. Based on high-frequency theme words, it explores deep themes, which helps to objectively evaluate the film as a whole and reflect the audience's true feelings, and to find the spreading genes and codes of the film's sustained popularity.

Keywords

Douban Movie Review, Text Mining, LDA Theme Model, "Let the Bullets Fly"

1. 引言

随着社会和技术的发展，互联网时代的电影网络评论也随之发展壮大。豆瓣电影、微博电影榜、猫眼、淘票票等各种电影打分、点评的平台的兴起，契合了用户在公众平台上表达自己的意见、态度、看法和情感的倾向，让不少观众养成了在观看电影的前后都去网络平台上查看相关电影的评分与评论，并留下自己的评分和评论的习惯。

豆瓣网是一个以书影音起家，提供关于书籍、电影、音乐等作品的 UGC 信息的社区网站[1]。其中的豆瓣电影，包含了最新的电影介绍、影讯查询、发布影评及购票服务。用户可以记录想看、在看和看过的电影电视剧，并打分、写影评。此外，还根据用户偏好，进行个性化的电影推荐。伴随着网络影评的异军突起，豆瓣网自 2005 年成立以来已经发展为电影口碑分享的重要平台[2]。在豆瓣电影的影评中，用户自发参与评价打分，豆瓣据此进行排序，从而给其他受众做参考，培养了受众在观看电影之前就来看豆瓣评分评价的一种行为习惯。豆瓣电影的评分、评论也时常成为媒体报道甚至专业影评的引用对象、参考对象。因此，选择豆瓣电影的评论文本，是因为这一平台更能“还原普通大众的平均看法”[3]，具有一定的代表性和参考价值。

所有用户都可以自己发布书评、影评，把自己的喜好展示给别人看。豆瓣的评价和评分能够直接影响用户对作品的感受和接受度，甚至可以直接决定观看意愿和购买决策。能直观反映其他用户总体认同度的打分评价，这是豆瓣的特色，也因此吸引了很多新用户的注册。

电影《让子弹飞》是一部 2010 年上映，由姜文执导，姜文、周润发、葛优等主演的剧情片。该片讲述了麻匪张牧之摇身一变，假扮新官“马邦德”上任鹅城县长，并与镇守鹅城的恶霸黄四郎展开激烈争

斗的故事。该片在极富娱乐性之余，也蕴含着丰富的细节和对白，使得观众对它的解读五花八门，对其评价也是褒贬不一。全片充满段子和隐喻，其中的“让子弹飞一会儿”“剖腹验粉”“站着挣钱”“翻译翻译，什么叫惊喜”“体面”“公平”等经典梗广为流传，直到十三年后的今天，这些段子和隐喻仍然具有强大的适应性和生命力。网友们常常在别的热点事件下借用《让子弹飞》里的经典桥段或台词进行评论，以至于网友直呼“让子弹飞怎么还不申遗？”“请全文背诵！”。尽管让该片申遗也是一种戏谑的说法，但网友们还是把对《让子弹飞》的细节和对白研究戏称为了“让学”，足可见观众对该片的高度肯定与赞扬。

网友们对于电影《让子弹飞》的评价究竟如何，这些评价大致可分为什么主题，体现了观众对该部影片何种关注视角？厘清观众对该部影片的不同评论主题，或许也能成为进一步深入研究经典的电影及其对白和隐喻，并探究其传播密码、为电影制作提供相应借鉴。

2. 研究设计

本研究通过 Python 程序语言编写爬虫程序，从豆瓣电影网站爬取电影《让子弹飞》的影评，再对爬取到的数据内容进行文本预处理——数据清洗、文本分词、去除停用词等，将有价值的数据进行 TF-IDF 分析，计算影评中的高频词汇，反映其评论热点，并使用 LDA 主题模型分析技术将主题词展现出来，试图找到有参考价值的评论信息。

2.1. 数据获取

由于豆瓣网的电影评论常被用户参考，且进行评论打分的用户在互联网中占比较大，该网站的评论数据具有一定代表意义。因此本文选取豆瓣电影网站(https://movie.douban.com/)中电影《让子弹飞》的影评作为研究对象，按“最受欢迎的”排布的评论数据进行了文本挖掘。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，因此本文只抓取了前十页的评论数据，包含了评论用户的名称、评级(分别是 1 到 5 星，对应的评级用词是很差、较差、还行、推荐、力荐)、发布时间、评论内容以及有用数和没用数等相关信息，共计 191 条数据。本文将主要选取这 191 条数据中评论内容作为研究的对象和文本数据。

2.2. 数据预处理

LDA 主题模型分析不会直接分析文本文档，而是分析基于这些文档形成文档词语矩阵(document-term matrix)，这个矩阵集合了每个词出现在文档的频率。因此文本预处理是文本挖掘必不可少的环节，目的是删除原始文本数据中的无用信息[4]。本文的数据预处理主要包括数据清洗、中文分词和去除停用词。首先是对爬取到的数据进行清洗，去除了表情、链接、无意义的符号或乱码以及一些空格和空行等。然后对文本进行中文分词，采用 jieba 中文分词，将评论语句分为单独词汇。随后对中文分词后的文本去除停用词，停用词主要是指那些在文本中出现的介词、代词、虚词等字符，以及嗯、啊、吧这之类的语气词，它们在评论中出现频率很高，但没有实际意义，因此需要对此类词语进行过滤。

3. 数据分析

3.1. 文本向量化

由于文本是人类语言而非计算机语言，因此还需要将非结构化的文本转变成结构化的数据。本文采用 TF-IDF 算法，来统计评论文本中的高频词，这里用到了用 TfidfVectorizer()函数完成向量化与 TF-IDF 预处理。在去除文档内出现几率过大或过小的词汇时，将参数 max_df 设置为了 0.99，将参数 min_df 设

置为了 0.01。

3.2. 构建 LDA 主题模型分析

本文采用基于 LDA (Latent Dirichlet Allocation，潜狄利克雷分布)算法的主题模型(topic model)方法[5]。主题模型是对文字隐含主题进行建模的方法，能够在海量互联网数据中自动寻找出文字间的语义主题。LDA 概率主题模型是基于贝叶斯网络模型的主题模型，挖掘文本间所隐含的主题信息，使用户快速了解文档的信息。主题对应的关键词是主题含义的折射，根据关键词可以概括出电影评论的主题内容。

3.3. 文本聚类结果

在做 LDA 主题模型分析前，先通过绘制主题数－困惑度曲线计算困惑度来确认主题个数。如图 1 所示。一般而言，困惑度越低，模型越好。根据图 1 的变化情况，将主题个数设置为 3。

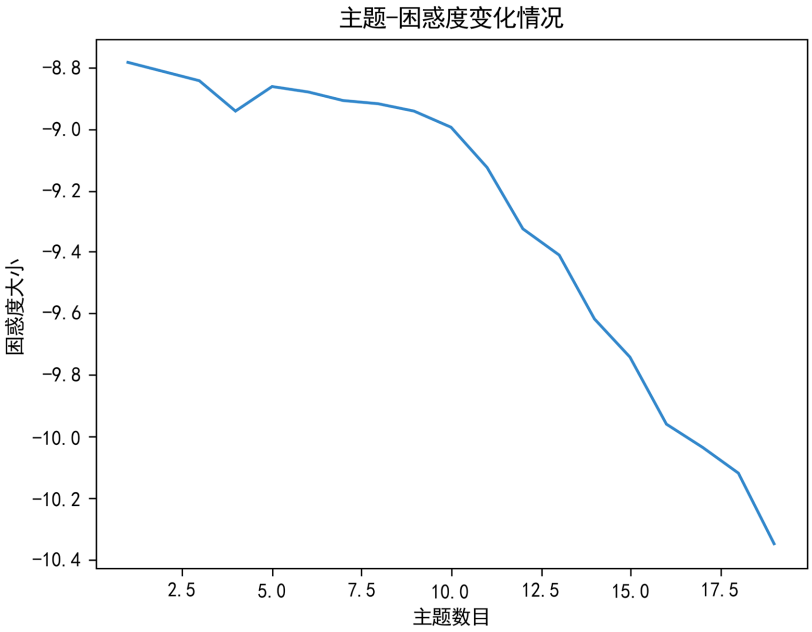

Figure 1. Number of topics-changes in confusion level
图 1. 主题数－困惑度变化情况

在 LDA 可视化部分，使用了 python 中的 LDAvis 模块对 LDA 模型结果进行可视化。本研究将主题数设置为 3，暂定每个主题输出前 20 个关键词，所得结果如表 1 所示。

Table 1. Topic feature words for different topics
表 1. 不同话题的主题特征词

主题 1	姜文 电影 子弹 觉得 火车 中国 重要 导演 意思 观众 已经 故事 大家 一点 葛优 片子 应该 东西 看到 影片
主题 2	权力 人民 张牧 蔡锷 日本 铁门 实际上 大哥 1920 现在 隐喻 问号 恐惧 发生 手枪 追随 权利 起义 影评人 1900
主题 3	麻子 黄四郎 师爷 县长 一起 最后 四爷 姜文 挣钱 兄弟 一会 革命 来说 辛亥革命 问题 后来 帽子 张牧 老三 葛优

从主题聚类效果表的关键词来看，豆瓣电影《让子弹飞》评论文本对该部影片的评论可以拆分为三

个主题，主题 1 中高频词，即姜文、电影、子弹、火车、中国、重要、导演、意思、观众、故事、葛优、片子、影片等，主要反映了对中国电影作品和行业评价的评论内容，以及对导演和主演的关注。主题 2 中高频词，即权力、人民、蔡锷、日本、铁门、大哥、1920、隐喻、问号、恐惧、手枪、追随、权利、起义、影评人等，这类评论更关注电影中的隐喻，反映出对电影的隐喻猜想和现实发散。主题 3 中高频词，即麻子、黄四郎、师爷、县长、四爷、姜文、挣钱、兄弟、革命、辛亥革命、问题、帽子、张牧、老三、葛优等，多数为电影角色和电影情节，这类评论更关注影片叙事和情节等，呈现出对电影本身的内容解读。

从主题聚类结果来看，针对不同主题的评论还是存在较为明显的差异。可视化分析结果显示，观众较多地在主题 1，即对电影作品评价、电影行业评价和对导演的关注这方面发表自己的看法，而关于主题 2，即对电影的隐喻猜想和现实发散，和主题 3，即对电影本身的内容解读的评论数量则相对要少些。如图 2 所示，左侧圆圈表示主题，右侧表示各个词语对主题的贡献度。

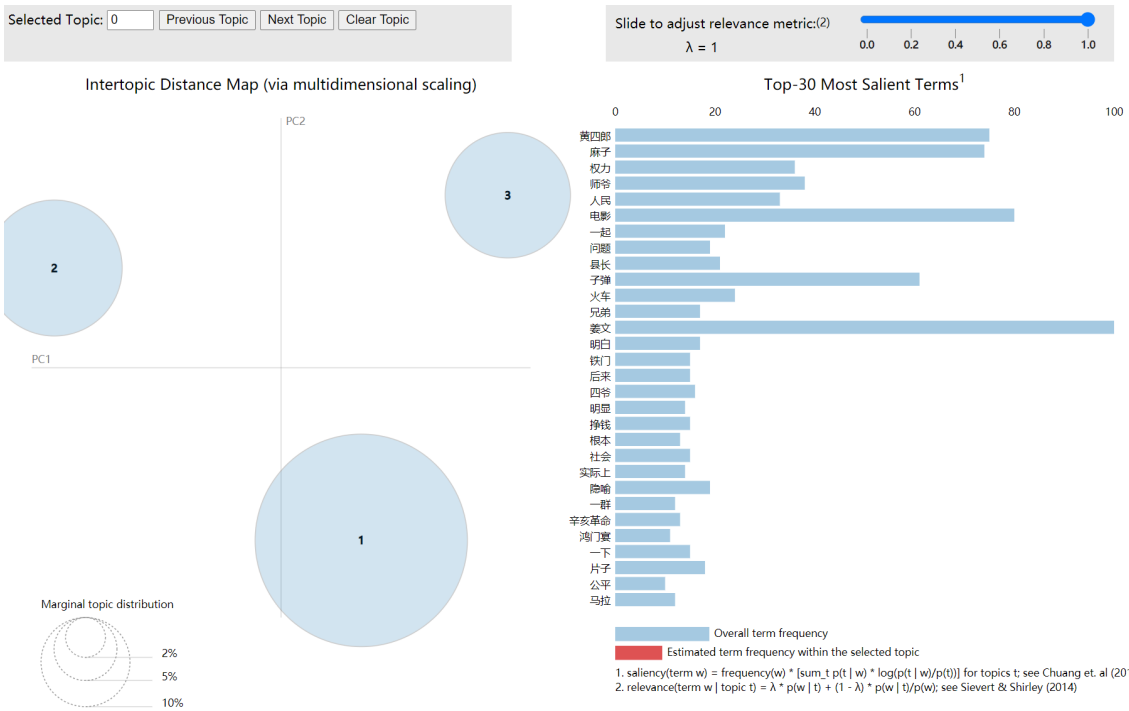

Figure 2. Visualization of LDA theme model
图 2. LDA 主题模型可视化

在主题 1，即对电影作品和行业评价以及对导演的关注中可以看出，不少观众是冲着导演和主演来的。姜文，作为一个备受争议又独树一帜的导演，受到了不少观众的关注和喜爱。在一部电影上映前，能否吸引到观众观看，除了作品本身的剧情内容，对导演的信任和支持也占有一席之地。例如评论者“溪流”的评论“姜文肯定给我们带来惊喜”，就导演姜文本身的才华和过往作品展开了评论，为此，在豆瓣电影评论中的热度高居第一位并获得了 36,440 次的“有用”和 1444 次的“没用”以及 3059 次的互动“回应”。又如评论者“鱼非鱼”的评论“三年以后，我们又等来了姜文”，这条影评联系了导演姜文之前的作品，找出《让子弹飞》和过往作品中的一些共同点和他一贯的拍摄偏好，以及几部作品之间的不同之处，在豆瓣电影评论中的热度高居第二位并获得了 11,302 次的“有用”和 515 次的“没用”以及 601 次的互动“回应”。

在主题 2，即对电影的隐喻猜想和现实发散中可以看出，对电影的暗线挖掘和隐喻猜想也是不少观众的一个讨论和评价热点。在他们看来，这部影片不是一部纯纯的幽默的商业片，而是用诙谐的方式反映社会现实，充满了魔幻的荒诞现实主义色彩。全片蕴含了大量隐喻和暗线，也有很多留白，正是这些隐喻和留白，将观众的想象力大大地调动了起来，给观众留下了足够的探讨空间，让他们对片中的细节和暗线深度挖掘，并作出自己的解读。如评论者“朽木”的评论“让子弹飞结局大揭秘”，评论中一共列举了 15 条对影片中的暗线解读，获得了 985 次的“有用”和 85 次的“没用”以及 159 次的“回应”互动。可见不论是赞同其解读还是另有观点，该部影片都有许多细节值得观众细细品味和讨论的空间。

在主题 3，即对电影本身的内容解读中，观众主要是针对影片中的人物、台词和剧情进行评价。略过对导演和主演的现实关注、对暗线和留白的隐喻猜想，这部分的评论关注到了电影作品本身。一部作品好不好，能不能成为经典，其中的人物塑造、台词打磨和剧情编排必不可少。不少中外经典电影作品之所以有着持久的生命力，正是因为其人物形象塑造得丰满，把一个个笔下、片中的人物演绎得有血有肉、有灵有性，例如《让子弹飞》剧中的张麻子、师爷、黄四郎、花姐等人物，每一个人都有自己的成长背景和人物立场，也有自己的故事发展，都不是片面单一的形象，而是丰满的、极具张力的。此外有不少评论都在复述评论者自己所认为的剧中经典台词和经典桥段，例如评论者“仰山雪”的评论“台词整理”，记录了该位用户认为剧中最经典的六句台词；评论者“小杨树”的评论“麻将、师爷、六子”，解读了麻将这个剧中元素以及师爷和六子这两位剧中人物角色。

4. 结语

技术带来的沟通便捷性和公共交流平台的扩大化促使观众和用户在网络平台上对电影作品进行网络评价时，可以畅所欲言，也可以和他人互动交流。本文基于文本挖掘，研究了豆瓣网中电影《让子弹飞》的网络影评。研究采用了计算机辅助内容分析的方法——LDA 主题模型来进行电影评论的主题聚类，从大量的评论文本中提炼出框架和主题，进而探寻观众评价中对影片的关注点和评论视角，试图找到有参考价值的评论信息。

计算传播学中使用的文本挖掘方法和 LDA 主题模型分析可以以数据驱动的形式实现海量文本分析，从大量的新闻文本中提炼出框架和主题，突破了以往面对海量影评内容“望洋兴叹”而只能对有限样本进行分析的局限。本研究基于计算传播学视角，借助文本挖掘和 LDA 主题模型分析，研究了电影《让子弹飞》的豆瓣影评，有助于对经典电影持续走红的传播密码进行分析。

参考文献

[1] 李峰媛. 从电视到网络我国悬疑剧发展新策略——以优酷网为例[J]. 艺术评鉴, 2019(1): 132-134.

[2] 马丽琳, 杨石华. 春节档电影网络评论的主流话语模式及社会心态——以近五年豆瓣电影评论为例[J]. 当代电影, 2022(4): 11-17.

[3] 刘颖琪. 国内外电影评分社区的现状及舆情探析[J]. 视听, 2018(1): 34-35.

[4] 唐铮, 徐子岳. 当他们谈论游戏时他们在谈论什么——基于微博评论的群体认知定量实践[J]. 新闻与写作, 2021(10): 53-61.

[5] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022.