Modern Linguistics 现代语言学, 2025, 13(6), 60-65
Published Online June 2025 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2025.136564

Hans汉斯

基于Python情感分析技术的字幕翻译研究

——以豆瓣电影网上《地球上的星星》影评为例

许知行

上海海事大学外国语学院，上海

收稿日期：2025年4月10日；录用日期：2025年5月27日；发布日期：2025年6月10日

摘　要

本研究聚焦宝莱坞电影《地球上的星星》在豆瓣电影网的观众评论，旨在探究其字幕翻译对电影文化传播的影响，来为电影字幕翻译提供参考。研究使用Python技术爬取影评数据并进行处理，再从语义和情感层面对数据进行分析，研究发现观众评论主要围绕电影本身、演员及现实影响，其高频词汇表明字幕翻译符合观众预期，准确阐明电影主旨，以目标语读者接受的方式保留印度特色且引发观众共鸣，使其反思现实和本国国情。情感分析显示正向情感评论占比较高，观众从多个方面对电影本身及字幕翻译给予肯定，这再次证明该电影字幕翻译的成功，其翻译传达电影的核心主题和关键情节，对外传播印度文化，彰显演员本身魅力，并引起观众对本国国情的反思。此外，尽管负面情感评论占比较少，仍提醒译者字幕翻译应充分考虑目标语观众接受程度，避免生硬照搬引起观众反感。

关键词

Python，情感分析，《地球上的星星》，字幕翻译

Research on Subtitle Translation Based on Python-Powered Sentiment Analysis Technology

—A Case Study of the Movie Reviews of “Like Stars on Earth” on Douban

Zhixing Xu

College of Foreign Languages, Shanghai Maritime University, Shanghai

Received: Apr. 10th, 2025; accepted: May 27th, 2025; published: Jun. 10th, 2025

Abstract

This research investigates movie reviews of the Bollywood film Taare Zameen Par (“Like Stars on

文章引用: 许知行. 基于 Python 情感分析技术的字幕翻译研究[J]. 现代语言学, 2025, 13(6): 60-65.
DOI: 10.12677/ml.2025.136564

Earth") on Douban to explore the impact of subtitle translation on film cultural dissemination, providing references for movie subtitle translation. The research uses Python to crawl and process the movie reviews, and then analyzes the data from the semantic and sentiment levels. The research finds that reviews mainly focus on the movie itself, the actors and the real-world impact. The high-frequency words indicate that the subtitle translation meets the audiences' expectations, accurately clarifies the main idea of the movie, preserves the characteristics of India in a way that is acceptable to the target language readers, and triggers empathy among the audience, causing them to reflect on the reality and the situation of their own country. Sentiment analysis shows a high proportion of positive sentiments, with viewers praising the movie itself and the subtitle translation in many ways, which proves once again that the subtitle translation of the movie is a success. The translation conveys the core themes and key plots of the movie, spreading Indian culture abroad, highlighting the charm of the actors themselves, and prompting viewers to reflect on the situation of their own country. In addition, despite the relatively small proportion of negative comments, translators are reminded that subtitle translation should fully consider the acceptance of the target language audience, and avoid causing aversion due to rigid imitation.

Keywords

Python, Sentiment Analysis, "Like Stars on Earth", Subtitle Translation

1. 引言

电影作为一种举足轻重的文化载体，自诞生便以独特的表现形式、强大的叙事能力，在全球范围内产生了深远影响，其不仅为观众提供感官盛宴，更通过鲜活画面与深刻情节，传递丰富的文化内涵与价值观。在全球化、信息化的浪潮下，电影的传播速度及影响力与日俱增，是当今对外文化交流的重要桥梁[1]。而电影字幕的翻译能直接影响观众对电影内容的理解和接受程度，进而影响电影在不同文化背景下的传播效果。

宝莱坞是印度电影业代称，其电影以音乐、戏剧和舞蹈的融合而闻名于世，其中情感更是作为"通用语言"跨越语言障碍，与全球观众产生共鸣[2]。而宝莱坞巨星阿米尔·汗更是被誉为"印度良心"，其电影深刻批判印度的社会问题，国际影响力深远。阿米尔·汗导演的《地球上的星星》讲述一个患有读写障碍症的儿童与老师之间发生的故事，其揭示了印度教育的缺陷，抨击了应试教育体制，在我国引起了广泛讨论。

因此，本文将研究目标定为豆瓣电影网上观众对《地球上的星星》的评论，运用 Python 情感分析技术获取并分析相关数据，挖掘豆瓣上观众对其的情感态度和评价，以期探寻宝莱坞电影对外输出的成功诀窍，以促进电影字幕翻译的发展，为今后的电影字幕翻译工作提供参考。

2. 文献综述

情感分析，又名意见挖掘，是自然语言处理的分支，涉及使用计算机算法识别、提取、分析和解释文本数据中的情感倾向和主观性，旨在确定作者对特定主题或对象持有的是积极还是消极(或中性)的态度[3]。而随着互联网的发展，情感分析在商业智能、市场研究等多个领域发挥着重要作用。

情感分析的方式主要分为利用情感词典的无监督机器学习方式和利用机器学习的有监督学习方式。

前者使用含情感色彩词语的情感词典，根据词典和自定规则计算研究文本得分，以判断其情感色彩[4]；后者由监督机器学习算法来进行情感极性划分与分类。两者区别在于后者需准备两份数据集，分别用于训练和测试。在对文本进行分类前，应先用标注好的训练数据来训练机器学习模型，让其学会提取关键信息以进行分类，再用同一标注文本测试模型效果。流行算法包括朴素贝叶斯和 SVM，其能通过学习训练集的特征进行情感分类，在处理大规模数据时更为有效[5]。

此外，近年来，深度学习方法在情感分析中也显示出巨大潜力，其可模拟人类的思考分析和神经网络的工作来判断情感倾向，可分为有监督和无监督的深度学习。常见模型有卷积神经网络(CNN)和基于循环神经网络(RNN) [6]。而注意力机制的引入进一步提升了深度学习模型的性能，尤其是在处理具有复杂结构的文本数据时[7]。

在电影评论情感分析领域，研究者多利用情感分析技术挖掘观众的观影偏好和市场走向。比如，吴小兰和常静宜[8]曾对豆瓣上《流浪地球》的影评进行情感分析以研究电影成功的要素。由此可见，情感分析可以运用在电影制作领域，还能为观众选择电影提供支持。

3. 研究方法

研究方法主要分为数据来源、研究步骤两大板块，其中研究步骤分为数据爬取和数据处理两个步骤。

3.1. 数据来源

豆瓣电影网是中国最大的电影分享与评论社区，为电影爱好者提供一个平台来发表他们对电影的看法，在国内的权威性和影响力都较高，拥有庞大的影迷社区和电影数据库。豆瓣电影网的用户群体集中在 20 到 39 岁，通常学历较高(本科及以上)、收入较高，且覆盖面极广，可见其是反映社会热点，以及青年观众观点、态度和立场的不二之选[9]。综上所述，考虑到豆瓣电影网的影响力，其评论的广泛性、真实性和深度，本文选择《地球上的星星》在其上的用户评论数据来进行分析，从受众评论角度研究国内观众对《地球上的星星》及其字幕翻译的情感态度和评价。

3.2. 研究步骤

3.2.1. 数据爬取

本研究访问了豆瓣电影网上《地球上的星星》的短评页面，从页面源代码中获取 cookie 值，以模拟登录，并使用 parsel 第三方库解析 html 内容以获取评价内容和星级等数据。受限于豆瓣的反爬虫机制，本次共爬取到 430 条数据。

3.2.2. 数据处理

首先，本研究初步清洗所得数据，将繁体字转化为简体汉字，并筛除内容为无关文字、乱码等 10 条数据，剩余 420 条数据。其次，对文本进行降噪。本研究使用 Jieba 分词软件对文本进行分词，借用哈工大停用词表去除高频出现但信息量低的词，并基于朴素贝叶斯分类器得出高频词汇云图、词频分析和情感分析等结果，以挖掘并分析文本内容。

4. 结果与分析

4.1. 语义分析

语义分析是自然语言的分支，其通过量化分析文本中特征词的结构、指向和特征来识别文本中的实体、关系、情感等语义信息[10]。而文本中词汇使用的频次也能在一定程度上反映出受众对特定对象的关注焦点和情感态度。由此，本研究对《地球上的星星》影评文本中关键词或主题词进行词频统计分析，

以了解观众对电影的关注焦点和对其翻译的评价，并借此为其他电影的翻译提供参考。

本研究先借助 jieba 分词脚本，对处理过的影评数据进行分词、去除停用词等操作，并根据统计出的高频词绘制出高频词汇云图，见图 1：

Figure 1. Word cloud of high-frequency words
图 1. 高频词汇云图

根据《地球上的星星》影评的词频分析可知，观众评论主要围绕电影本身(如“伊夏”“剧情”)、演员个人(如“阿米尔”)和现实影响(如“教育”“应试教育”)，这些词汇表明了翻译准确传达出电影的主旨，也表现出字幕翻译未影响到观众观感，其对电影反响热烈，特别是“感动”“励志”等词，可见字幕翻译符合观众预期，促进其对电影产生共鸣。

此外，“印度”和“中国”等词汇的高频出现，表明电影字幕翻译包含印度文化特定元素，保留了印度特色，却也不让国人反感，反而引起了国人对两国国情的思考。可见字幕翻译应在最大程度保留源语言特色的同时，又考虑到目标语观众的接受程度，以温和、能被接受的方式输出源语言国家的文化。而“教育”“孩子”等词的频繁提及，说明字幕翻译准确阐明了电影的核心主题和关键情节，得到了观众的认可，并引发其对现实的反思。因此，优秀的字幕翻译可以让观众忽视国家间国情的不同，从他国现实反思己国不足，这是生硬照搬地翻译无法做到，需要一定程度的同化，并找到两国国情的相接点，从而更好引起观众共鸣。此外，“音乐”等词的高频使用，也体现电影中音乐的翻译表达出乐曲的神韵，其深受观众的喜爱。

高频词汇云图表明大部分观众对《地球上的星星》持肯定态度，赞扬其教育意义、对特殊群体的呈现和电影本身的塑造，这从侧面反映出电影的字幕翻译是成功的，其阐明电影的含义精髓，保留印度特色风采，又符合国人观影习惯。然后，也有“煽情”“拖沓”等词在质疑电影剧情及其翻译过于矫情，不符合中国人说话习惯。总而言之，高频词分析真实反映出观众的态度和关注点，为此类电影的制作及字幕翻译提供了宝贵的反馈。

因此，本研究还提取出《地球上的星星》电影评论中排名前 10 的高频词汇(见表 1)，以更直观地展示出观众对电影及字幕翻译的真实评价和关注热点。

Table 1. Frequency ranking of top 10 words
表 1. Top 10 词频统计

排名	词条	频率
1	电影	203
2	孩子	174
3	老师	133
4	印度	104
5	教育	85
6	阿米尔	81
7	世界	66
8	小时	36
9	故事	34
10	影片	34

4.2. 情感分析

情感分析，也称意见挖掘，是自然语言处理的分支，还涉及信息检索、数据挖掘等领域，其主要用于定性主观性文本的态度、观点和情感倾向，广泛应用于舆情分析等多个领域。进行文本情感分析的过程通常包括：情感信息检索、情感信息抽取和情感信息分类[11]。

本研究在对提取到的文本进行分词、去停用词处理后，使用 Sklearn 机器学习库提供的朴素贝叶斯分类器对文本进行了分类。朴素贝叶斯分类器需要预先抽取训练样本进行训练，来建立分类模型。因此，本研究抽取了四分之一的文本(110 条)，人工将其标注为正向和负向情感，以充当训练集。若影评文本传达积极的正向情感，如“太感人了，泪水哗哗流，歌也好听，翻译也很融洽”等，则标注为“+1”；而若影评文本传达消极的负向情感，如“情节很俗套，小朋友的门牙很囧”等，则标注为“−1”。经统计，人工标注的训练集文本共计 110 条，其中正向情感有 85 条，占比 77%；负向情感有 25 条，占比 23%。

训练过分类器之后，再通过十折交叉验证的结果验证其性能指标，得到查准率(precision)和召回率(recall)的加权平均值 F1 为 0.7821，说明分类效果较好(取值为 0~1，数值越大则表示查准率或召回率越高)，并对标记文本进行分类，得到正向情感评论共 267 条，占比 86%；负向情感评论共 43 条，占比 14%。

由此可知，正向情感评论的占比要更高，而正向情感评论词集中在赞扬电影本身(如励志、喜剧)、演员个人魅力(如阿米尔汗)和电影的现实影响(如教育、应试教育)，可见观众对于《地球上的星星》的认同，而这也从侧面映证了该电影字幕翻译的成功，其翻译以目标语观众认可的方式准确传递出原片信息，保留一定印度特色，彰显演员本身魅力，并引起观众对本国国情的反思。

而负面情感评论占比虽较少，也不容忽视，其主要集中于吐槽电影时长、节奏(如小时、拖沓)和浓厚的印度风格(如歌舞)，这两点虽主要吐槽电影本身，但也为电影字幕翻译敲响警钟，电影字幕翻译应考虑到目标语言国家观众的接受程度，在一定程度上采用同化翻译，而非生硬照搬，引起目标语言国家观众不满。

综上，根据《地球上的星星》影评的情感表达和词汇特征可知中国观众对其电影及翻译较为认同和欢迎，这为日后电影字幕翻译提供了借鉴。

5. 结语与展望

本研究通过对《地球上的星星》在豆瓣电影网上的电影评论进行数据爬取、处理以及语义和情感分析，深入探究了字幕翻译在电影对外文化输出中的作用，研究发现如下：

从语义特征方面看，高频词汇云图和词频统计清晰地展现了观众的关注焦点，主要围绕电影本身、演员、现实影响等方面，可见字幕翻译符合观众预期，准确传达电影主旨，在最大程度保留印度特色的同时，又考虑到目标语观众的接受程度，温和地输出源语言国家文化，以一定程度的同化使得观众共鸣，引发其反思现实和本国国情。

从情感特征方面来看，正向情感评论占比较高，观众从多个方面对电影及字幕翻译给予肯定，再次证明该电影字幕翻译的成功，其翻译以目标语观众认可的方式传递出原片信息，保留一定印度特色，彰显演员本身魅力，并引起观众对本国国情的反思。尽管负面情感评论占比较少，仍提醒译者字幕翻译应充分考虑目标语观众接受程度，避免生硬照搬。

然而，本研究仍存在一些局限，一是调查对象仅为豆瓣网的影评用户，未能覆盖所有观影者，且受限于豆瓣网的反爬虫机制，研究数据较为有限。此外，尽管本研究已经剔除无法反映实际语义和偏离主题的评论文本，却无法分辨过滤水军操控的评论内容。因此，本研究可能存在一定的样本偏差，后续研究可以进一步改进代码或使用访谈法等来提升研究的信度与效度；二是本研究未对文本情感词汇的极性进行分类研究，因此情感极性对受众态度、立场以及传播效果的影响在后续研究中也可作进一步探讨。

参考文献

[1] 肖成笑. 电影艺术中的文化价值观塑造与传播机制[J]. 电影文学, 2024(18): 77-81.

[2] 赵琼. 新时期宝莱坞电影的跨文化传播[J]. 电影文学, 2020(13): 50-52.

[3] 林煌, 李弼程. 基于 BERT 模型和图注意力网络的方面级情感分析[J]. 计算机科学, 2024, 51(S2): 118-124.

[4] 董昱灿, 赵奎. 基于注意力机制多特征融合与文本情感分析的日志异常检测方法[J]. 四川大学学报(自然科学版), 2024, 61(2): 76-86.

[5] 杨程, 车文刚. 基于多门混合专家网络的情感分析与文本摘要多任务模型[J]. 现代电子技术, 2024, 47(1): 94-99.

[6] 林伟, 陈雁. 融合 BERT-BiGRU 和多尺度 CNN 的中文微博情感分析[J]. 中国电子科学研究院学报, 2023, 18(10): 939-945.

[7] 刘忠宝, 雷宇飞. 融合多尺度特征的多模态情感分析模型设计与实验[J]. 实验室研究与探索, 2024, 43(9): 78-83+96.

[8] 吴小兰, 常静宜. 基于情感计算的电影成功要素分析——以《流浪地球》为例[J]. 传媒观察, 2020(9): 59-67.

[9] 周彦杉. 大众接受效果与科幻文艺创作——以《三体》小说和电视剧的豆瓣评价为例[J]. 当代文坛, 2023(5): 109-115.

[10] 张换香, 彭俊杰. 基于方面级情感分析的深度语义挖掘模型[J]. 电子学报, 2024, 52(7): 2307-2319.

[11] 王玮, 温世阳. 情感分析在社会化媒体效果研究中的应用——基于分类序列规则的微博文本情绪分析[J]. 国际新闻界, 2017, 39(4): 63-75.