作品编号：

2025年（第十一届）全国大学生统计建模大赛

参 赛 作 品

参赛学校：西安电子科技大学

论文题目：

参赛队员：张佳昊 郭昊 高宇轩

指导老师：

# 摘要

随着中国电影市场规模持续扩大与社交媒体舆论影响力的深化，观众情感倾向与结构化市场特征的耦合作用日益显著。然而，传统票房预测模型对非结构化舆情数据的量化不足，且面临水军干扰、评分与内容不匹配等问题，亟需构建动态情感指数并融合多源数据以提升预测精度。

第二部分， 针对电影舆情数据碎片化与噪声干扰问题，本研究提出多维情感指数（SEI）构建框架。**数据整合层面**，融合豆瓣短评、微博超话实时数据、知乎长评及猫眼结构化评分，通过用户可信度加权（点赞数×用户等级）与时间衰减函数（衰减因子λ=0.95）抑制水军干扰，确保数据代表性。**情感量化模型设计**采用双路径策略：短文本基于TF-IDF扩展情感词典（覆盖89%网络新词）计算情感极性；长文本结合大数据大模型对情节、主题、视听效果等6个维度进行细粒度评分，解决“高评分低口碑”的数据矛盾。**动态融合阶段**，基于双模型方差倒数分配权重（短评模型权重0.62，长评模型权重0.38），结合逻辑函数与对数函数生成SEI。统计验证显示，SEI与票房走势的格兰杰因果检验显著（*p*<0.01），且方差贡献度达28.6%，证明其能有效表征舆情动态对市场的先导影响。

第三部分，为解决多源数据耦合与非线性预测难题，构建混合模型框架：时序建模层采用DLinear模型，将历史票房分解为趋势成分（移动平均窗口=7）与剩余成分（反映周期性波动），通过双线性网络分别建模，并引入外部变量（排片率、节假日标志、竞争影片数量）增强动态响应能力。特征交互层集成XGBoost，解析导演历史票房（衰减因子γ=0.8加权）、电影类型热度（猫眼搜索指数）与SEI的非线性关联。模型优化阶段，通过网格搜索确定超参数（输入长度=14，预测时程=7），采用MSE损失函数与早停策略（耐心值=10），防止过拟合。测试集上混合模型MSE为0.079，显著优于ARIMA（0.142）与单一DLinear（0.105）；SHAP值分析表明，SEI贡献度（31.2%）远超排片率（19.8%）与导演影响力（15.4%），验证舆情量化对预测的核心价值。案例分析显示，模型对《哪吒2》首周票房的预测误差仅为5.7%，且能识别“黑马”影片（如《你好，李焕英》）的异常舆情拐点，展现强鲁棒性。

本研究构建的DLinear-XGBoost混合模型可实现票房高精度预测（MAPE=7.3%），为片方优化排片策略、资本方量化投资风险提供数据支持；通过融合情感指数与结构化特征，突破传统模型的线性局限，推动预测技术向动态智能化转型。此外，模型输出的可解释性结论（如节假日效应权重=0.22）可辅助政策制定者平衡商业价值与文化导向，促进电影产业可持续发展。

**关键词：DLinear模型、情感指数、票房预测、时间序列分解、可解释性机器学习**

# 目录

# 表格和插图清单

# 引言

## 利用PEST模型分析研究背景

(1)政治环境（Political）

中国政府对文化产业的扶持政策为电影产业发展提供了坚实基础。近年来，国家电影局出台多项政策鼓励国产电影创作，例如《电影产业促进法》明确提出支持优质内容生产，并推动电影市场规范化。2021年发布的《“十四五”中国电影发展规划》进一步强调技术创新与内容升级，提出“提升电影产业链现代化水平”的目标。这些政策不仅推动了《哪吒》系列等高投入动画电影的制作，也为票房预测模型的应用提供了政策支持。此外，电影审查制度对内容导向的管控，使得观众情感与口碑成为影响票房的关键因素，进一步凸显了情感指数在预测中的重要性。

(2）经济环境（Economic）

中国电影市场规模持续扩大，2023年总票房已突破500亿元，稳居全球第二大市场。随着居民可支配收入增长，文化消费需求显著提升，电影票房的波动与经济周期、节假日消费习惯密切相关。例如，春节档和暑期档票房占比高达全年40%以上。与此同时，电影投资风险加剧，片方需通过精准预测优化排片和营销策略。结构化数据（如导演历史票房、排片率）与情感指数的结合，能够为资本方提供科学决策依据，降低市场不确定性。艾媒咨询数据显示，2022年超60%的观众选择观影前参考网络评价，表明情感分析对票房预测的经济价值。

(3)社会环境（Social）

社交媒体时代，观众参与电影口碑传播的主动性显著增强。以《哪吒之魔童降世》为例，其豆瓣评分8.4分，微博话题阅读量超50亿次，形成“现象级”社会讨论。新一代观众（Z世代）更依赖短评平台（如抖音、微博）获取信息，情感倾向直接影响观影决策。然而，水军刷分、评分与内容不匹配等问题（如“高评分低口碑”现象）增加了预测复杂性。通过构建多维度情感指数（情节、主题等），结合用户可信度加权，可更客观反映社会舆论的真实趋势，为预测模型提供可靠输入。

(4)技术环境（Technological）

自然语言处理（NLP）与机器学习技术的突破为情感分析提供了技术保障。传统TF-IDF加权情感词典可高效处理短文本（如微博评论），而大模型API（如GPT-4）能深度解析长文本的隐含情感，实现多维度评分。时间序列模型（如Informer）与集成学习（XGBoost）的融合，可同时捕捉结构化特征（导演影响力）和非结构化特征（情感指数）的交互效应。此外，SHAP值解释性技术能揭示模型决策逻辑，增强结果可信度。Gartner报告指出，2023年全球75%的企业已将情感分析纳入商业决策，印证了技术应用的成熟性。

## 研究意义及创新性

（1）研究目的及意义

通过构建动态情感指数与多模态票房预测模型，不仅能够在理论层面推动情感分析与市场预测的交叉融合，同时能为电影产业链的精细化运营提供切实可行的决策工具。理论层面，通过提出多源情感数据的动态加权融合方法，能够突破传统单一维度评分的局限，为情感分析领域提供兼顾效率与深度的技术路径；而通过探索非结构化情感指标（如情节、主题分项评分）与结构化市场特征（导演影响力、排片率）的协同作用机制，能够进一步丰富票房预测模型的理论框架。

实践层面，模型的高精度预测能力可助力片方优化排片策略、资本方量化投资风险、院线制定动态营销计划，从而提升行业资源配置效率；情感指数的实时监测与可视化分析，还可辅助内容创作者精准捕捉观众偏好，推动宣发策略从“经验驱动”向“数据驱动”转型。

社会层面，通过设计数据可信度加权与时间衰减机制，有效抑制水军评论与短期舆情噪声的干扰，促进电影评价体系的客观性与公平性；同时，模型输出的量化结论可为政策制定者平衡文化产品的内容质量与市场效益提供科学依据，助力国产电影在商业价值与文化价值间实现良性互动。

（2）创新性

## （三）文献综述

近年来，随着社交媒体数据的爆炸式增长和自然语言处理技术的突破，基于文本情感分析的预测模型在金融与影视领域受到广泛关注。现有研究主要围绕社交媒体数据的情感挖掘及其对市场表现的预测能力展开，但在数据源选择、模型架构设计以及多模态数据融合等方面仍存在优化空间。

在影视领域，Bogaert等（2021）通过跨平台比较揭示了社交媒体数据对票房预测的显著价值。研究发现，相较于Twitter，Facebook的用户生成内容（UGC）和营销生成内容（MGC）具有更强的预测能力，尤其是评论数量和情感极性组合变量对票房解释力最高。该研究通过信息融合敏感性分析验证了多算法集成模型的鲁棒性，但其情感分析仍依赖于传统词典法和浅层机器学习模型，未深入挖掘长文本的语义特征。类似地，张宸瑞（2022）在金融领域的研究中，采用预训练语言模型BERT对股吧评论进行情感分析，发现情绪指数与股价波动存在显著相关性（MIC=0.38）。该研究创新性地利用深度学习解决一词多义问题，但其情绪维度单一（仅正向、中性、负向三分类），且未结合结构化市场数据构建复合预测模型，限制了预测效用的实际落地。

现有研究在方法论上呈现两大趋势：一是数据类型的多元化整合，如Bogaert等（2021）将电影特征数据与社交媒体UGC/MGC相结合，Houston等（2018）提出预发布社交声量可作为票房领先指标；二是模型架构的深度优化，如张宸瑞（2022）通过领域适配的BERT微调提升金融文本情感识别准确率至75.53%。然而，这些研究仍存在三方面局限：其一，文本处理粒度不足，未区分短评与长评的语义差异，短评依赖TF-IDF加权词典法易受噪声干扰（石善冲等，2018），长评采用单一情感标签则忽略多维评价属性；其二，多模态数据协同机制缺失，结构化特征（如导演历史票房）与非结构化情感指数的动态交互尚未充分建模；其三，时间序列特性利用不足，现有研究多采用静态特征工程，未有效捕捉情感传播的滞后效应与衰减规律。

针对上述问题，本研究提出融合多粒度情感分析与时空特征建模的创新框架。首先，构建分层情感计算模型,以此来为评论文本赋予一个合理的情感指数；其次，引入混合专家系统（MoE），将情感指数与电影属性（导演影响力、档期竞争等）、市场动态（排片率、预售数据）等结构化特征进行动态权重分配，通过XGBoost与DLinear模型的集成捕捉线性和非线性关系。最后，采用格兰杰因果检验验证情感指数的领先性，并利用SHAP值解析多特征交互效应，弥补现有研究归因机制不透明的缺陷。该框架既继承了Bogaert等（2021）的多平台数据融合优势，又通过深度学习细粒度分析拓展了张宸瑞（2022）的文本挖掘深度，为跨领域情感经济研究提供新的方法论范式。

## （四）研究方法与技术路线

# 二、情感指数的获取（我还没想好一个合适的名字）

## （一）问题描述

随着互联网与社交媒体的快速发展，在线评论与评分已成为公众对电影情感表达的重要载体。然而，现有评分体系（如豆瓣五星制）存在显著局限性：其一，单一评分难以反映用户对电影情节、主题等多维度的细分情感；其二，短文本评论与长文本评论的信息挖掘方式差异较大，传统情感分析方法难以兼顾；其三，数据噪声（如水军评论、评分与文本情感不匹配）可能降低情感分析的可靠性。此外，情感动态随时间衰减的特征以及评论影响力（如用户可信度与互动量）的差异化未被有效量化，导致现有情感指标与票房预测模型的关联性不足。

针对上述问题，该部分旨在构建一个动态加权的情感指数（Sentiment Emotion Index, SEI），以综合多维评论数据并提升预测精度。

## （二）数据介绍

1.数据来源

本研究的数据采集基于多源异构数据融合框架，涵盖短文本评论、长文本评论及结构化评分三类数据，分别从主流社交媒体平台与专业票务平台获取。

短评方面，通过定向网络爬虫技术，从豆瓣电影与微博超话平台抓取用户实时评论数据，提取字段包括评论文本内容、发布时间、点赞数及回复数。其中，豆瓣电影提供高密度的电影垂直领域讨论，微博超话则聚焦热点影片的即时舆论反馈。

长评方面，知乎影评与猫眼专业影评作为长文本数据源，通过API接口授权获取结构化数据。采集字段涵盖完整评论文本、用户评分（1-5星）及用户等级信息。知乎影评侧重深度分析与多维度观点表达，猫眼专业影评则提供大众用户的综合反馈。

2.数据处理

为确保数据质量与后续建模的可靠性，本研究基于多层级清洗框架对原始数据进行系统化处理。清洗流程涵盖冗余信息剔除、语义噪声过滤及文本规范化三大核心模块。

针对数据重复性问题，采用哈希算法识别完全一致的评论，并结合Jaccard相似度（阈值>0.8）与余弦相似度（阈值>0.75）双指标检测语义高度重叠的评论，有效消除水军刷评及用户跨平台重复发布的影响。其次，通过构建正则表达式规则库（如“高清资源|私信领取”）与随机森林分类器，精准识别广告类垃圾内容，同步筛除非电影相关评论。在此过程中，LDA主题建模技术被用于量化评论文本与电影主题的关联性，设定相似度阈值（<0.3）以移除讨论演员私生活或无关话题的噪声数据。此外，利用正则匹配清除URL链接、联系方式及非中文字符，确保数据纯净度。

针对文本结构杂乱问题，对原始评论进行深度格式化处理：清除HTML标签、特殊符号及非标准标点，统一全角与半角格式；过滤无意义短评（如“好”“嗯”）及重复字符噪声（如“好好好”→“好”）。同时，对多模态内容进行语义映射，例如将Emoji表情转换为等效文本（如“😂”→“笑哭”），并基于同义词词林完成同义表达归一化（如“棒极了”→“优秀”）及繁简转换（如“電影”→“电影”），以提升文本分析的标准化程度。

针对噪声评论识别与质量控制，进一步采用混合策略处理非自然语言与极端评论：通过规则引擎识别过度使用表情符号或重复字母的噪声内容（如“hahahaha”→“哈哈”），并结合BERT情感分析模型检测极端情绪表达（如“史上最烂电影”），此类评论被标记而非直接删除，以保留数据多样性。此外，统计用户评论频率与评分一致性，构建水军识别模型筛除异常高频或低可信度用户。最后，引入时间衰减因子降低历史数据的权重，并通过TF-IDF高频词分析与人工抽样复检，确保清洗结果的完备性与准确性。

## 短评情感指数的获取

1.评论分词处理

在得到经清洗过的大量短评后，利用python中jieba库对短评进行分词。Jieba库作为优秀的中文分词第三方库，依靠中文词库，确定汉字之间的关联概率，汉字间概率大的组成词组，形成分词结果。除了分词，我们还可以添加自定义的一些常用词组，让其帮助我们划分，将每个短评划分为单个的词语合集。

2.模型建立

①关键词TF-IDF算法

TF-IDF(Term Frequency-Inverse Document Frequency)是一种用于资讯检索与文本挖掘的常用加权技术。TF-IDF是一种统计方法，用以评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF加权的各种形式常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。利用该方法，可以通过加权解决有些词在文本中尽管词频高，但是并不重要的问题。

TF（Term Frequency）,即词频。它表示一个词在所有短评中出现的次数，为了消除数据量大小的影响，将其定义为：

|  |  |
| --- | --- |
|  | (1) |

TF（,d）值越大，说明词越重要。

IDF（Inverse Document Frequency）,即逆文档频率，它是表达词语在语料库中重要性的一个重要指标，通常被定义为：

|  |  |
| --- | --- |
|  | (2) |

其中N为语料库中文档的总数，为包含词的文档数量，log通常为自然对数（底为e）或以10为底的对数，分母＋1是为了避免分母为0。如果所有文章都包含t这个词，那么t的IDF（,D）=log(1)=0，即重要性为0，停用词的IDF约等于0。如果某个词只在很少的文章中出现，则IDF很大，其重要性也就越高。

将TF与IDF结合，得到TF-IDF值：

|  |  |
| --- | --- |
|  | (3) |

②基于大连理工情感词典的情感得分计算

本研究选择的基础词典为大连理工大学情感词汇本体库，在该词典中，情感词的初始情感强度被设置为1、3、5、7、9五个等级，较其他词典而言，强度划分的更为细致，9表示强度最大，1为强度最小。情感词的情感极性有中性、褒义、贬义三类，分别对应值0、1、2。为便于计算机作情感计算，我们将代表贬义的极性2修改为-1。

情感词汇本体中的词性种类一共分为7类，分别是名词（noun）、动词（verb）、形容词（adj）、副词（adv）、网络词语（nw）、成语（idiom）、介词短语（prep）。同时每个词在每一类情感下都对应了一个极性，分别为中性、褒义和贬义。

表1 情感词汇本体格式样例

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 词语 | 词性种类 | 词义数 | 词义序号 | 情感分类 | 强度 | 极性 |
| 无所畏惧 | idiom | 1 | 1 | PH | 7 | 1 |
| 手头紧 | idiom | 1 | 1 | NE | 7 | 0 |
| 周到 | adj | 1 | 1 | PH | 5 | 1 |
| 言过其实 | idiom | 1 | 1 | NN | 5 | 2 |

最后检查是否有否定词与程度副词，给定否定词判定系数，初始值为1，若有否定词，则置为-1，即否定词权重。程度副词根据搜集到的文本赋予其权重。

表2 否定词表

|  |  |
| --- | --- |
| 权重 | 否定词 |
| -1 | 不、不是、不能、不大、不必、不可、没、没有、不要、无、非、并非、莫、弗、毋、勿、未、未有、枉、否、别、無、休、甭、白、空、徒、徒然、毫不、s绝不…… |

表3 程度副词

|  |  |  |  |
| --- | --- | --- | --- |
| 等级 | 权重值 | 程度副词（列举） | 数量 |
| “极其/最“ | 3 | 非常、极、极其、极度、充分、绝对、十分 | 69 |
| “超“ | 2.1 | 超、浮、过、过分、何止、偏、忒、开外 | 30 |
| “很“ | 1.5 | 不少、甚、很是、颇为、尤其、分外、大为 | 42 |
| “较“ | 1.06 | 更为、愈加、越发、还要、较、较为、进一步 | 37 |
| “稍“ | 0.75 | 略、略微、一点儿、有些、稍微、稍许、稍稍 | 29 |
| “欠“ | 0.53 | 半点、丝毫、不丁点儿、不甚、不大、相对 | 12 |

综上，基于情感词典的情感得分为：

|  |  |
| --- | --- |
|  | (4) |

式中，表示词汇的情感值，表示词汇的情感强度，表示词汇的情感极性，为否定词判断系数，表示程度副词的权重。

③二者的加权融合

到此，我们已经得到了由TF-IDF计算出的词语重要程度，以及基于情感词典得到的词语情感分数,现在我们将二者加权融合得到最终第i条短评的得分：

|  |  |
| --- | --- |
|  | (5) |

④最终每日短评得分计算

设得到的每日短评数量为m，则最终每日短评得分为：

|  |  |
| --- | --- |
|  | (6) |

## （四）长评情感指数的获取

1.模型介绍

该部分基于预训练语言模型的架构范式，采用模型即服务(MaaS)的设计理念，在本地部署了Deepseek-R1-Distill-Qwen-14B，针对评论理解这一任务进行了模型微调，最后标准化接口实现了评论推理模块的集成，有效平衡了推理速度与推理效。

首先筛选出有效评论（有效评论有效值为1，无效评论有效值为0），其次分别对有效长评的情感得分、故事讲述、角色表演、制作水平四个维度进行评分，随后对每日所有有效长评的每一维度的评分分别求和，则得到每日评论的四个维度的初始得分RawScorei,t,其中i为维度，t为日期（1月29日-3月24日）。

2.数据的Z-score标准化

①Z-score标准化分数计算

Z-score标准化是数据处理的一种常用方法，能够将不同量级的数据转化为统一量度的Z-score分值进行比较，提高了数据可比性，削弱了数据解释性，将其定义为：

|  |  |
| --- | --- |
|  | (7) |

设N为总天数，其中，

|  |  |
| --- | --- |
|  | (8) |

|  |  |
| --- | --- |
|  | (9) |

②异常值检验

根据3原则,对于不在范围内的值进行截断，低于下限的值截断取-3，高于上限的值截断取3。

该研究中，2月14日受节日影响，评论数据量较大，因而其标准化分数也较高，造成了分数值的较大波动，该天在情感得分、故事讲述及制作水平三个维度均检验出异常值，分别为3.1894、3.9676和3.9524，将他们截断为3。

由此，我们对初始得分，进行了标准化，更新了每日的各维度情感得分。

3.基于熵权法的维度权重计算

①模型介绍

熵权法是一种基于信息熵的客观赋权方法，通过计算指标的变异程度（信息量）确定权重。信息熵越小，指标的变异程度越大，提供的信息越多，权重越高；反之，熵值越大，权重越低。

包含的参数：

信息熵ej：衡量指标数据分布的混乱程度。熵值越小，数据分布越不均衡，信息量越大。

差异系数dj：将熵值转化为正向指标，直接反映信息量大小。

权重wj：通过归一化差异系数，实现权重的客观分配。

②构建判断矩阵：

评价天数m天，（m=55），每天具有维度n个（n=4），构建判断矩阵：

|  |  |
| --- | --- |
|  | (10) |

③数据标准化（非负化处理）

为消除原始矩阵存在的负数和量纲差异，需要先将判断矩阵正向化标准化，使用最小-最大标准化方法：

|  |  |
| --- | --- |
|  | (11) |

其中，min(z.j)与max(z.j)为同一维度j下的最大值与最小值，确保元素非负，得到标准化矩阵。

④计算概率矩阵P

对每个维度j，计算样本i的占比:

|  |  |
| --- | --- |
|  | (12) |

满足,即每列概率和为1。

⑤计算信息熵

第j个指标的信息熵为：

|  |  |
| --- | --- |
|  | (13) |

其中(n为样本数)，若=0,定义。

⑥计算差异系数

差异系数反映指标的信息量：

|  |  |
| --- | --- |
|  | (14) |

越大，指标权重越高。

⑦确定权重

权重为差异系数归一化结果：

|  |  |
| --- | --- |
|  | (15) |

最终权重向量为。

4.引入时间衰减

## （五）结论分析与讨论

# 基于深度学习…预测票房（依旧没想好高级名字）

## 模型介绍与建立（模型假设）

## 数据介绍与处理

## 曲线拟合及结论分析

# 模型检验与效果分析

## 情感指数检验

## （二）深度学习那部分检验

# 五、结论与建议

## 结论

## 不足和建议

# 参考文献

# 附录

# 致谢