浙江工商大學

《文本数据挖掘》期末大作业

题目：流浪地球

学　　院：统计与数学学院
专　　业：数据科学与大数据技术
班　　级：大数据 1902
学生姓名：胡夏冰　　1902110227
指导教师：崔峰

2021 年 12 月 26 日

《流浪地球》影评文本数据分析

摘要

2019 年 2 月 5 日——大年初一，《流浪地球》全国上映。在豆瓣评分上，首日开分站稳 8 分以上，延续了之前点映的高口碑。微博上也跟着出现吴京客串 31 天与投资 6000 万的热搜。本文爬取了豆瓣的部分影评数据，时间跨度自 2019 年 1 月 11 日至 2019 年 3 月 13 日。

在互联网时代，信息技术飞速发展，人们越来越多地在网络平台上发表自己的观点和意见。随着这些评论数据量的爆炸式增长，如何提取利用其中的情感信息也成为人们的关注热点，文本情感分析技术随之兴起。情感分析是对含有情感色彩的主观性文本进行分析，挖掘出其蕴含的情感倾向的过程础。作为自然语言处理领域的一个重要分支，情感分析在理论方面有着较高的研究意义。本文针对《流浪地球》数据集，分别从舆情分析，情感分析以及情感极性分析等方面提取有用的信息。

针对舆情分析，本文通过影评内容词云图展示、影评数与日期之间的关系、影评数与时刻的关系、网友评分高低与时间的关系、网友评分高低与入会时间的关系以及影评情况地理位置分布图等方面进行分析。

针对情感分析，本文通过 BosonNLP 情感词典，进行影评数据的情感分析，对每一条评论进行情感打分。实验发现，受影评文本长度的影响，部分影评内容的情感得分绝对值过大。传统的基于情感词典的情感分析局限性较大。

针对情感极性分析，本文比较了基于卷积神经网络的预测结果和基于文本类的预测结果，发现基于卷积神经网络的预测结果略高于基于文本类的预测结果。本文猜测可能导致的原因有：数据集样本量太少，深度学习所使用的神经网络模型，需要样本量只有在一定大的前提下预测效果才会明显高于传统机器学习模型所预测的结果；对于文本的情感极性分析，深度学习中的循环神经网络预测效果会更好。文本类数据在一定程度上涉及到了时序相关的特征，而当 CNN 在提取特征时，未考虑到时许特征，这也是 RNN 模型优于 CNN 之处。

在后续的工作中，本文将完善基于深度学习的情感极性分析预测方法。试图利用循环神经网络，长短期记忆网络，以及注意力机制的更为复杂的模型进行试验。

关键字： 流浪地球　舆情分析　情感分析　情感极性分析

目录

一、 引言

2019 年 2 月 5 日——大年初一，《流浪地球》全国上映。在豆瓣评分上，首日开分站稳 8 分以上，延续了之前点映的高口碑。微博上也跟着出现吴京客串 31 天与投资 6000 万的热搜。《流浪地球》根据刘慈欣的同名小说改编，故事背景设定在 2075 年，讲述了太阳即将毁灭，已经不适合人类生存，而面对绝境，人类将开启“流浪地球”计划，试图带着地球一起逃离太阳系，寻找人类新家园的故事。该片由屈楚萧、赵今麦、李光洁、吴孟达等主演，吴京特别出演，于 2019 年 2 月 5 日在中国内地上映。

《流浪地球》2019 年 2 月 4 日（农历大年三十）午夜场票房报收 1280 万元人民币，夺得当日的票房冠军。据猫眼电影统计，截至 2019 年 2 月 24 日，《流浪地球》在中国大陆票房超过 43.57 亿元人民币。因春节档档期竞争激烈，春节首日该片只有 11.4% 的排片。但因口碑良好于上映第 3 天起登顶日冠。2 月 8 日，上映第 4 天该片票房突破 10 亿。2 月 10 日，即中国大陆地区春节长假的最后一天（上映第 6 天），该片总票房突破 20 亿，成为中国大陆影史第 14 部破 20 亿的电影，截至 2019 年 2 月 12 日，该片投资方中国电影披露收益超 1 亿元人民币，北京文化收益也在 8000 万左右。2 月 14 日（电影上映第 10 天）下午，片方宣布该片票房突破 30 亿，总额暂跻身中国电影历史第六位，打破了《战狼 2》获得 30 亿元人民币票房的最快速度记录，成为第六部破 30 亿的电影，也是最快获得 30 亿元票房的电影。

本文爬取的评论数据时间区间为 1 月 11 日至 3 月 13 日，进行分析之前，了解到豆瓣评分有以下特点：豆瓣评分在电影没上映前即可评分，但不会显示出来，需在上映后评分人数达到指定数量，才会显示出来，因此将 2 月 5 日前的评论当做有效评论。利用网上网友观看过影片后对影片的评论（即影评）和评价（即推荐等级）展开分析。本文对推荐等级进行数据可视化，研究推荐等级与评分、评论间的关系。

电影评论的目的在于分析、鉴定和评价蕴含在银幕中的审美价值、认识价值、社会意义、镜头语言等方面，达到拍摄影片的目的，解释影片中所表达的主题，既能通过分析影片的成败得失，帮助导演开阔视野，提高创作水平，以促进电影艺术的繁荣和发展。随着影视业的不断发展，看电影已经成为家常便饭，看电影的方式也五花八门，因此影评数量众多，影评的质量层次不齐。如何从海量影评中获取关键信息，并保证信息较为准确成为一大难点。

本文针对《流浪地球》影评数据集，拟通过舆情分析、情感分析、情感极性分析（文本分类）等方面进行分析。

二、 数据预处理

2.1 数据集描述

本文先对流浪地球评论数据集进行简略的描述。该数据集一共有九个属性，分别为：citys, content, evaluate, labs, nams, scores, times,user_info 和 votes, 其具体含义如表 1 所示：

表 **1** 特征统计

属性 (特征)	含义	是否有缺失值	数据类型
citys	评论者所在的城市	否	object
content	评论内容	否	object
evaluate	评论者对电影的推荐程度	否	object
labs	评论者是否看过电影	否	object
nams	评论者用户名	否	object
scores	评论者对电影的打分	否	object
times	评论时间	否	object
user_info	评论者个人信息	否	object
votes	对评论的点赞数	否	int64

考虑到重复值会占用多余的内存空间，并且在数据分析时也会增加数据的相关性，影响数据分析的结果。数据集通过去重处理后，去除重复样本记录 20 条。

2.2 特征处理

2.2.1 nams、labs 特征的处理

对于 nams 特征，用户名可能涉及到用户的隐私，且对本文情感分析的作用不大。因此，本文直接去除该特征。

对于 labs 特征，本文统计了该特征的特征值分布情况，发现特征值都为“看过”。因此，labs 特征对本文情感分析的作用也不大，本文直接去除该特征。

2.2.2 votes、content 特征的处理

对于 votes 特征和 labs 特征，在数据预处理过程中先保持不动。

2.2.3 citys 特征的处理

对于 citys 特征，本文查看了该特征的具体特征值，发现所有特征值两侧都存在“['*']”这一无用字符。因此，本文对 citys 特征进行去“['*']”字符处理。

2.2.4 times 特征的处理

对于 times 特征，该特征值主要由日期“年-月-日”及具体时间“时-分-秒”构成。因此，本文将重新构造 date 特征代表日期“年-月-日”，time 特征代表具体时间“时-分-秒”，同时去除原特征 times。

2.2.5 user_info 特征的处理

对于 user_info 特征，该特征值主要由用户名和入会时间构成。因此，本文仅提取其中的入会时间部分，并将缺失值用评论时的日期 date 代替，以此构建新特征 join_time。

2.2.6 evaluate、scores 特征的处理

对于 evaluate 特征，本文先查看了该特征的特征值分布情况。发现存在“2019-02-10 18:09:43”这类的脏数据共计 14 条。

对于 scores 特征，本文同样查看了该特征的特征值分布情况。发现存在缺失值同样共计 14 条。初步看来，客户评分 scores 与 evaluate 有密切关联。因此，本文对于 scores 的缺失值，拟通过 evaluate 进行填补。对于 evaluate 的脏数据，拟用 scores 进行替换。

处理过程中发现，scores 的缺失值和 evaluate 的脏数据为同一样本数据。考虑到这些样本数据量不大，缺失值与脏数据失去了分析的意义，因此本文直接剔除这些样本。

对于 scores 与 evaluate 特征，进一步的，本文做了相关性分析。首先，粗略地将 evaluate 特征值划分为五个等级，并分别赋分：力荐 =50，推荐 =40，还行 =30，较差 =20，很差 =10。

将 scores 数据类型进行转换后，可以通过计算皮尔逊相关系数来观测两者相关性。皮尔逊相关系数计算公式如下：

$$r=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}} \tag{1}$$

其中，$\bar{X}$ 表示 scores 特征的平均值，X_i 表示第 i 个 scores 特征的特征值，$\bar{Y}$ 表示 evaluate 特征的平均值，Y_i 表示第 i 个 evaluate 特征的特征值。

接着，画出特征值散点矩阵分布图和相关系数热力图为：

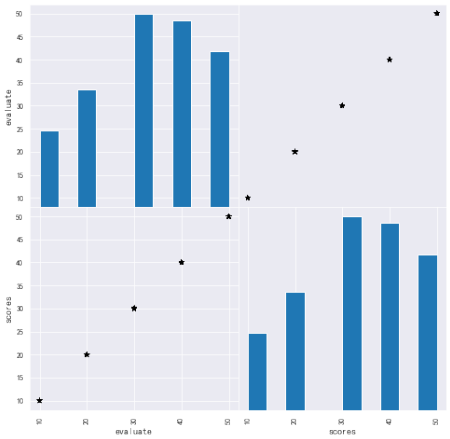

图 **1**　散点矩阵分布图

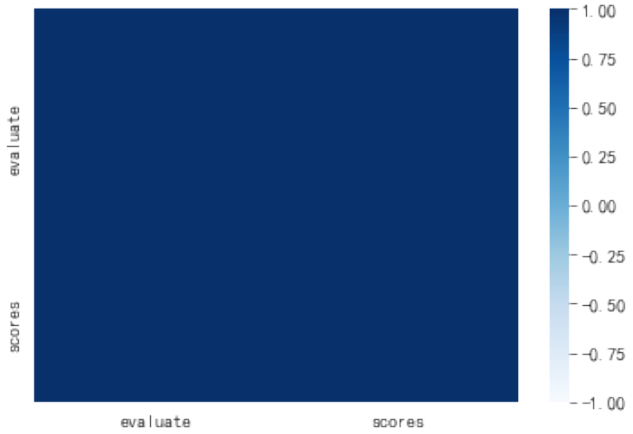

图 **2**　相关系数热力图

通过特征值散点矩阵分布图和相关系数热力图可以发现，scores 特征与 evaluate 特征的相关性为 1。表明了 scores 特征与 evaluate 特征线性相关，当自变量之间存在共线性时，模型的参数会变得极其不稳定，模型得预测能力会下降。很难确切区分每个自变量对因变量的影响，因此增加了对于模型结果得解释成本。对此，本文删去 evaluate 特征。

2.3 处理后的数据集

最后，将预处理得到的数据做个初步展示 (取前五条数据)，结果如表 2 所示：

表 **2**　预处理后的数据展示（部分）

	citys	content	scores	votes	date	time	join_time
0	北京	一个悲伤的故事：.....	40	35161	2019-02-05	00:24:35	2018-10-07
1	北京	电影比预期要更恢...	40	68629	2019-02-04	15:56:16	2005-07-18
2	北京	还能更土更儿戏一点吗...	10	69686	2019-01-28	22:06:27	2008-01-28
3	北京	1. 终于，轮到我...	50	59980	2019-01-29	20:10:48	2008-08-30
4		真为吴京的...	10	38488	2019-02-05	01:55:20	2019-02-05

经过预处理的数据集删去了一些对分析无关的信息，如用户昵称等。添加了一些与分析相关的数据，如评论时间、评论时刻、入会时间等。同时也对数据集进行了进一步的规范化处理。

三、 探索性数据分析及可视化

探索性数据分析是指针对已有数据通过汇总统计、可视化、方程拟合、计算特征量等方式探索数据本身的结构和规律的一种数据分析方法。在对数据进行操作前，首先需要运用数据框的属性和方法查看数据的形状和分布状况。只有对数据有一定的了解，才能采取相应的操作对该数据进行分析。

3.1 影评内容词云图展示

在绘制词云图之前，需要对影评进行分词、去停用词处理。本文采用 jieba 库对影评内容进行分词，并选取课堂上提供的停用词库进行去停用词处理。一般的，还有哈工大停用词表，中文停用词表，百度停用词表，四川大学机器智能实验室停用词库等公开的停用词库，可供大家自行选择。

词云图直观地显示出了一些备受关注的高频热议词汇。为提高词云图的美观性，本文通过 stylecloud 库进行词云图展示，效果如下：

图 **3**　影评词云图

由图可知，影评主要围绕“中国”，“地球”，“科幻”，“流浪”等话题展开，影评不少有对中国科幻电影元年，刘慈欣导演，吴京等的评价。也可以从侧面反映出人们对电影的关注点所在。

其中，“特效”的提及频率也很高，特效是电影前期主力宣传点之一，相较之前的国产电影确实有了很大的进步，查阅数据集发现，目前多数评价为正面评价。制片方的

确花了较多的经费用于电影的后期特效处理，并且请来了国际知名的 VFX 团队。

“国产”、“好莱坞”、“元年”。可以看出，《流浪地球》作为被称为“中国科幻电影希望”的国产科幻电影，观众自然而然地将它与好莱坞产品做比较，并提出了很多具有参考价值的评论。也存在部分评分由于国产而降低评分标准或恶意给打低分的情况。

为进一步探索不同影评内容所涉及的关键词所在，本文以 30 分为界限，将影评分为两类：正向影评（大于等于 30 分）和负向影评（小于 30 分）。并分别绘制词云图如图 4、图 5 所示：

图 **4**　正向评论词云图

图 **5**　负向评论词云图

3.2 影评数与日期的关系

随着电影热度的不断提升，当日影评数也可以从侧面反应网友对该电影的关注度。因此，本文做出影评数随日期变化的关系图，具体如图 6 所示：

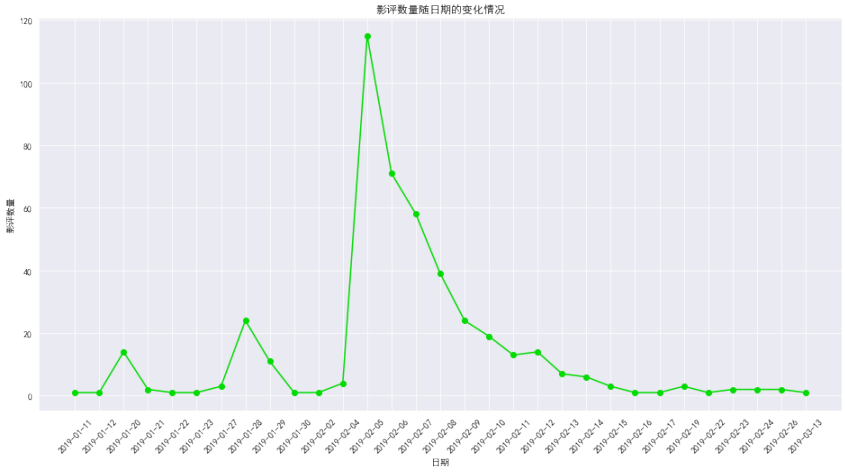

图 **6**　影评数随日期变化关系图

由于允许爬取的评论数量和时间问题，部分数据不是很明显。但依然可以得出一些发现。从 2019 年 1 月 11 日至 2019 年 2 月 5 日，影评数量总体呈现上升趋势。在影片上映开始的一周内，为评论高峰，尤其是上映 3 天内，这符合常识，但是也可能有偏差，因为爬虫获取的数据是经过豆瓣电影排序的（按投票数），倘若数据量足够大得出的趋势可能更接近真实情况。

另外发现，影片在上映前也有部分评论，分析可能是影院公映前的小规模试映，且这些提前批的用户的评分均值，差不多接近影评上映后的大规模评论的最终评分，从这些细节中，或许可以猜测，这些能提前观看影片的，可能是资深影迷或者影视从业人员，他们的评论有着十分不错的参考价值。

3.3 影评数与时刻的关系

影评数与时刻的关系可以粗略反映网友在观影时间上的选择，本文分别对所有时期和 2019 年 2 月 5 日统计其影评数，结果如图 7、图 8 所示：

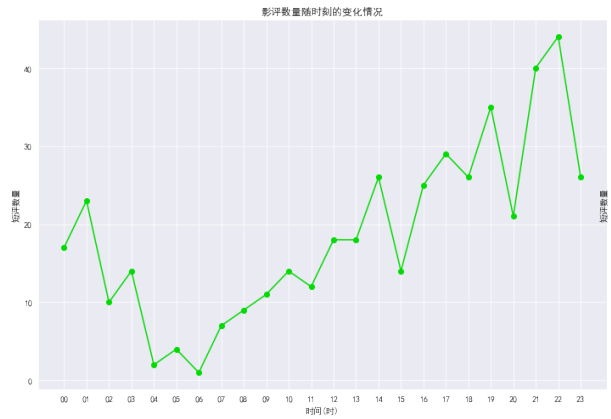

图 7　所有时期

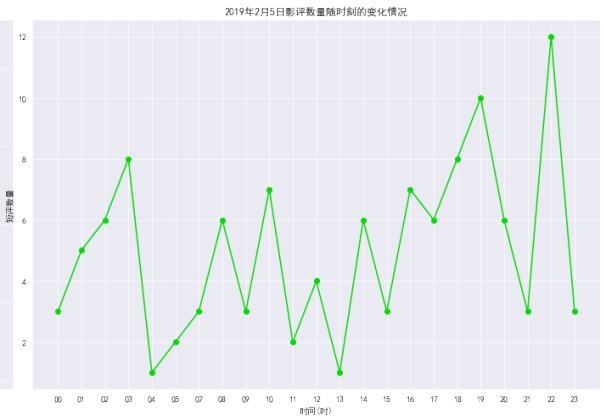

图 **8**　**2019** 年 **2** 月 **5** 日

我们可以看到，在一天的 24 小时中，每个时间段都有网友在发表短评，但更多的分布在晚上 6:00 之后，这个时间段处于广大网友的休息时间，网友有充足的时间去观看影片，并发表评论、评分。

3.4 网友评分高低与时间的关系

一个影片的评分在一定程度上反映了该电影的口碑，表现该电影的需求热度。有相当大比例的观众会因为购票网站评分较高而去选择观看一部影片，所以在电影行业中，一般情况下，映后的口碑与最终的票房动力成正相关。本文通过时间序列的关系，以时间轴为基础，分析网友用户评分随日期变化的关系，如下图 9 所示。

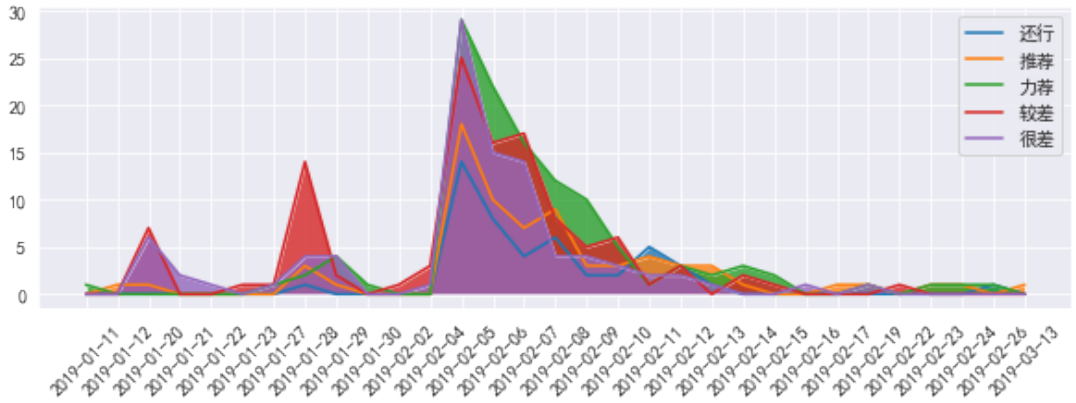

图 **9**　评分随日期变化关系图

在影片上映之前，较高分数偏多，如 1 月 28 日推荐及以上的网友数远超过其余推荐程度。在影片上映当天，评分网友数量最多，且以还行的网友较多。可以看出网友对影片期待值较大，这一方面是刘慈欣小说本身所带有的 IP 效应，另一方面是自《战狼》系列后大家对吴京抱有极大的期待，但是影片整体表现一般，评分随影片上映时间降低。

3.5 网友评分高低与入会时间的关系

不同入会时间的网友评分也会不同。一般而言，入会时间越长，间接性的说明了该网友对电影有更独到的见解。对此，本文通过时间序列的关系，以入会时间为自变量，分析网友评分与入会时间的关系，如图 10 所示。

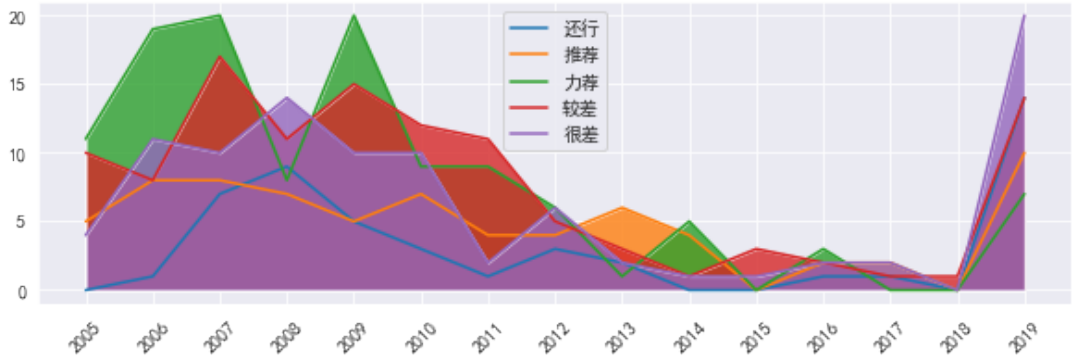

图 **10**　评分随入会时间变化关系图

总体而言，入会时间越长的网友对这部电影的评价越好。最早入会的 2005 年豆瓣用户对电影的评价两极分化较严重。明显可见的几点是：1. 在 2005-2011 年间的各个评分段呢评分数量都特别高，而在 2012-2018 年间各个评分段的评分数量都特别少，之后在 2019 年各个评分段评价数量都有不同程度的增长。2. 评价很差的人数在 2019 年增长最多。3. 在 2005-2011 年间，评价力荐、很差和较差的评价数量占总评论的很大一部分。2012-2018 年间各个评价占比都差不多，很少。

3.6 影评情况的地理位置分布图

影评的地理位置分布情况，可以间接地反映当地居民消费水平。只有当人们不再受物质需求的压迫时，才会进行精神需求。本文给出了影评情况的地理位置分布图，如图 11 所示：

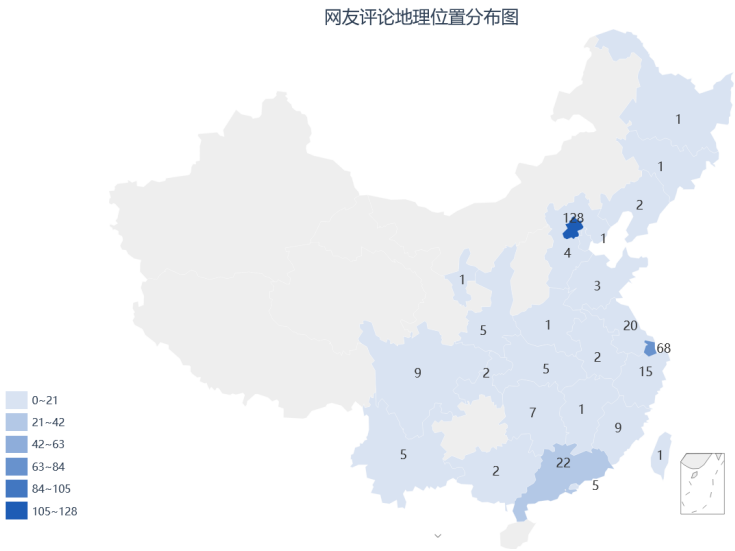

图 **11**　影评情况的地理位置分布图

可以看到，北京上海及沿海城市的评论用户较多，原因可能是因为这些地区的消费水平比较高，人们的生活也相对来说较为丰富，可以更好的追求精神上的满足。

四、 影评情感分析

文本的情感分析是一个涉及多领域的综合性研究学科，包括语言学、心理学、机器学习和统计分析等。对文本情感分析任务而言，数据集的组成部分主要有评价内容、评价对象、情感类别等。根据分类的目的不同，情感分析工作的子任务有主客观的分析和情感态度的分类两种。对于不同的文本情感分析任务，也可以采用不同的情感类别体系。

传统的情感分析技术主要可以分为两种，一种是基于情感词典的分析，一种是基于机器学习算法的情感分析。如图 12 所示：

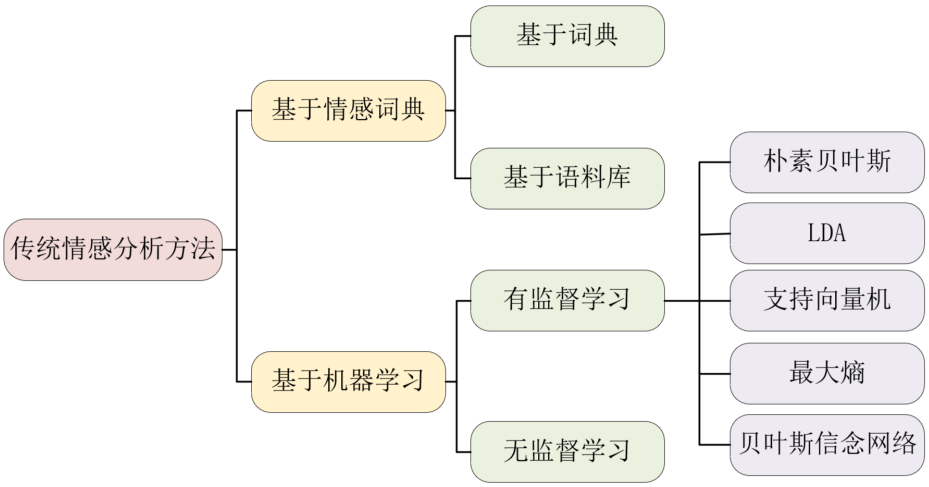

图 **12**　传统的情感分析组成框架

在使用基于情感词典进行情感分析时，依据已经构建好的情感词典，对文本中出现的情感词汇进行匹配，再根据词典中查询得到的结果，对这些情感词进行组合分析，进而判断文本的情感态度倾向。因此，情感词典的质量直接关系到对情感态度进行分析的系统的性能，情感词典的构建工作就变得尤为关键。

构建情感词典的方法可以分为三类: 基于手工方法、基于词典和基于语料的方法。但随着互联网的发展，信息量剧增，也出现了很多新兴词汇，对情感词典的构建维护工作造成了巨大的挑战, 单纯使用情感词典的方法完成情感分析工作也就存在着很大的局限性。

基于机器学习的文本情感分析方法，在近年来得到了广泛的关注。这种使用统计学习分析的方法不再依赖情感词典，而是根据各种机器学习算法对文本进行统计分析。机器学习有两种主要的方式: 有监督学习和无监督学习。2002 年，Pang 等人最早使用机器学习算法进行情感分析的研究, 在论文中比较分析了支持向量机、朴素贝叶斯、最大熵等算法，取得了良好的分类效果。

基于深度学习的文本情感分析方法，最早于 2006 年被 Hinton 等人提出。随着词向量技术的提出和深度学习理论的发展，神经网络模型逐渐被应用到情感分析等自然语言处理相关的研究任务中。

4.1 基于词典的情感分析

基于情感词典的分析方法是情感挖掘分析方法中的一种，其普遍做法是：首先对文本进行情感词匹配，然后汇总情感词进行评分，最后得到文本的情感倾向。目前使用较多的情感词典主要有：BosonNLP 情感词典，清华大学李军中文褒贬义词典，台湾大学中文情感极性词典，知网情感词典，知网程度副词词典等。

本文选用 BosonNLP 情感词典进行分析。BosonNLP 情感词典是由波森自然语言处理公司推出的一款已经做好标注的情感词典。词典中对每个情感词进行情感值评分，BosonNLP 情感词典大概如下表 3 所示：

表 **3** **BosonNLP** 情感词典

正向词	情感得分	中性词	情感得分	负向词	情感得分
借东风	4.523561956	生鱼片	0.001622123	公主岭市	-2.910417137
傻	4.523561956	破旧	0.001622123	片假名	-2.910417137
共建	4.523561956	雷霆	0.002364015	硝酸	-2.910417137
分离式	4.523561956	鞋子	0.00244663	罪魁	-2.910417137
力耶	4.523561956	庸俗	0.002475372	量刑	-2.910417137
包玫	4.523561956	退休	0.002584329	前兆	-2.91022687
包申通	4.523561956	本能	0.002666772	失算	-2.909219602
南帕亚	4.523561956	间谍	0.002683542	疼	-2.907328155
吉虹颖	4.523561956	琐事	0.002942021	候诊	-2.905388626
吴锦泉	4.523561956	嘉定	0.003068501	孤苦	-2.905388626
哈密地区	4.523561956	胃肠	0.00310488	怒不可遏	-2.905388626

通过 BosonNLP 情感词典，可以得到各评论的情感得分，如表 4 所示：

表 **4** 各评论的情感得分

	content	sentiment
0	一个悲伤的故事：太阳都要毁灭，地球都要流浪了，我国的校服.....	-6.087825e-01
1	电影比预期要更恢弘磅礴，晨昏线过后的永夜、火种计划、让地...	2.168741e+10
2	还能更土更儿戏一点吗？毫无思考仅靠煽动，毫无敬畏仅余妄...	9.920909e-02
3	1. 终于，轮到我们仰望星空。2. 后启示录死亡废墟，赛博朋克...	2.330640e+16
4	真为吴京的演技尴尬，总是摆出一副大义凌然的样子，好奇为什...	9.804465e-01

可以发现，第二条评论、第四条评论的情感得分明显过大。这是由于进行情感得分

计算的过程中，情感程度副词的应用极大的改变了情感得分。一般而言，评论的内容越长，情感得分的绝对值越大。从中也可以看出传统的基于情感词典的情感分析局限性。

4.2 基于深度学习的情感分析

深度学习是人工神经网络在使用多层网络进行任务学习中的应用，随着深度学习在图像和语音处理方面取得重大进展，它在情感分析领域也开始被广泛应用，目前深度学习模型包括卷积神经网络（Convolutional Neural Networks，CNN）、循环神经网络（Recurrent Neural Networks，RNN）、LSTM、BiLSTM（Bi-directional Long Short-Term Memory）、门控循环单元（Gated Recurrent Unit，GRU）和注意力机制等。

与基于情感词典与机器学习的方法相比，深度学习有更强的表达能力和模型泛化能力，但是缺乏大规模的训练数据也是深度学习在情感分类中遇到的问题，此外，梯度消失与爆炸、模型参数的设置与模型的复杂度也是需要解决的问题。

受硬件、时间以及个人能力等因素的限制。本文仅完成了通过 CNN 进行情感极限分析的任务。在分析前，先简要介绍一下 CNN 的基本概念。

卷积神经网络 (Convolutional Neural Networks, CNN) 是一种前馈神经网络。在传统的神经网络中，通常是特征向量作为输入的数据，而要得到特征向量，需要经过人工去对特征进行设计，然后进行组合计算得到特征向量。而人工进行的特征选择有其局限性，可能并不一定准确，卷积神经网络在特征提取上则有其突出的优势。近些年来，卷积神经网络不仅在图像相关领域取得了突出的表现，在 NLP 领域也取得了突破性进展，是深度学习相关研究中的热点理论模型。

CNN 采用局部感受野、权值共享和时空亚采样的思想，显著地减少了网络中自由参数的个数，是深度学习的代表算法之一，它的提出是受到了人们对生物的视觉感知机制相关研究的启发。CNN 是多层网络的结构，如图 13 所示，主要由卷积层、池化层、全连接层等构成，本文将主要介绍一下卷积层和池化层。

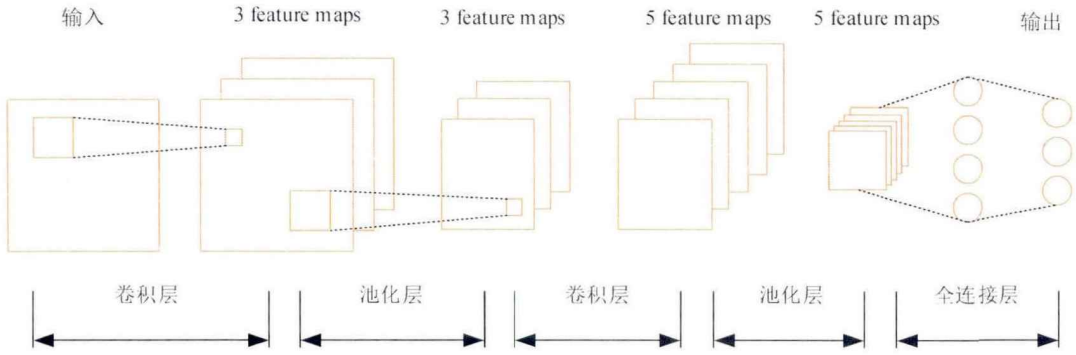

图 **13**　多层 **CNN** 结构

卷积是一个数学概念，一维卷积是对信号进行移动平均，而二维卷积则用于对图像

和视频信息进行处理。卷积层，是卷积神经网络中一个重要的层次，主要是对输入数据的特征信息进行学习，可以使用多个卷积核来计算不同的特征输出图（Feature Map)。局部感知和参数共享是卷积层的两个主要特征。

局部感知: 认为在图像中，局部的像素间联系比较紧密，在学习过程中，没必要学习全部的信息，只需要对局部进行感知，然后将多个局部进行综合就能够得到全局的信息。

参数共享: 使用卷积核进行局部特征提取的过程中，卷积核是不变的，整个图像的卷积是共享一个卷积核的，即每个神经元的参数是一致的。权值参数的共享能大量减少需要训练的参数数量，提升训练的速度。

使用单独的卷积核因参数共享可能使得特征的提取不够充分，这种情况可以通过采用多个卷积核进行卷积来学习更多的特征。卷积层中的激活函数（Activation Function）是对卷积层的输出做非线性变换。

池化层 (Pooling Layer）能够将经过卷积层输出的特征向量降维，在保留了有价值的信息的同时减少了数据量。如今，常用的池化方法有最大池化（MaxPooling）和平均池化 (Average Pooling)，最大池化的过程如图 14 所示。顾名思义, 最大池化就是在池化过程中在每个 filter 覆盖范围内保留最大值。如果 FeatureMap 是有深度的, 那么进行池化的过程中, 会对每一层 Feature Map 分别做池化，经过池化后的深度不改变。

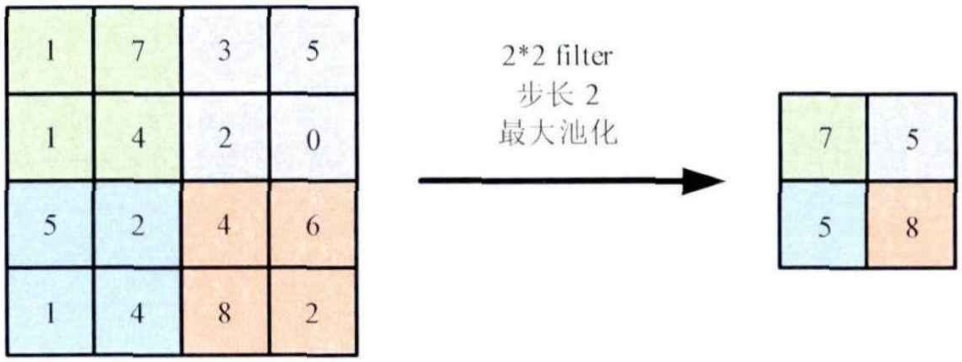

图 **14　CNN** 中的池化层

通过卷积层和池化层，可以学习到深层次的抽象特征表示，而全连接层的主要作用是分类, 将通过卷积、池化操作后学习到的特征表示经过非线性操作，输出分类的类别标签。

最早 CNN 是被应用于图像相关领域, 为了能够利用其突出的特征提取能力，研究人员也尝试着将卷积神经网络应用到自然语言处理的任务上, 取得了一定的成果。在自然语言处理领域，卷积神经网络通过卷积核的运算对输入的词向量进行特征的提取，提取文本信息建立特征表示，进而完成文本建模分析等任务。

本文通过基于 CNN 的情感极性分析与基于文本类的情感极性分析结果比较，来评价传统机器学习与深度学习在进行情感极性分析时的好坏。对于二分类问题，本文通过

混淆矩阵，F1 值，召回率、查全率等指标进行评价。两者的混淆矩阵如图 15、图 16 所示。

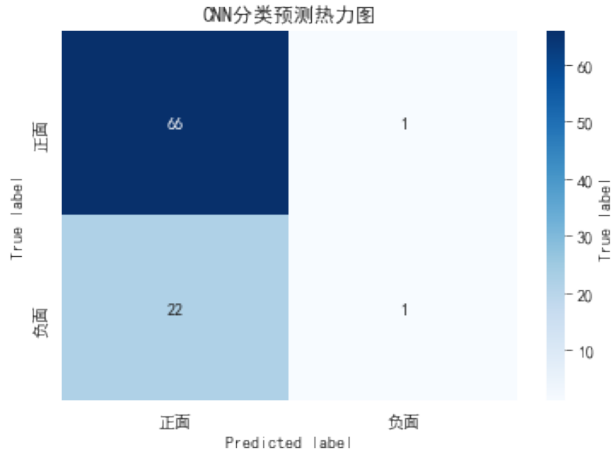

图 **15**　卷积神经网络极性分析

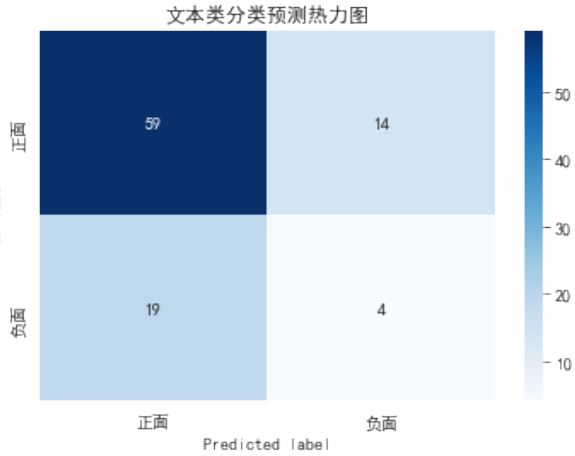

图 **16**　文本类极性分析

通过对比混淆矩阵，本文发现基于卷积神经网络的情感极性分析与基于文本类的情感极性分析预测结果相差不大。本文猜测可能导致的原因有：数据集样本量太少，深度学习所使用的神经网络模型，需要样本量只有在一定大的前提下预测效果才会明显高于传统机器学习模型所预测的结果；对于文本的情感极性分析，深度学习中的循环神经网络预测效果会更好。文本类数据在一定程度上涉及到了时序相关的特征，而当 CNN 在提取特征时，未考虑到时许特征，这也是 RNN 模型优于 CNN 之处。

五、 总结

本文针对《流浪地球》影评数据集，分别通过舆情分析、情感分析、情感极性分析（文本分类）等方面进行分析。文本情感分析作为自然语言处理领域的一项基本任务，已经成为一个重要的研究领域, 研究工作有着很高的理论意义和实际应用价值。

同时，在后续的工作里可以对以下几个方面进行进一步完善。

第一，本文的情感分析工作是对文本整体进行的情感分析分析，但实际上部分文本中对事物的不同方面所表达的情感倾向是不同的，如“这部电影特效做的很棒，但剧情没意思”。后续会对不同的方面的态度倾向进行细分，结合具体对象的抽取完成情感分析工作。

第二，文本中同一个词语在不同领域蕴含的信息是有差异的，特定领域的数据训练出的模型并不能有效地迁移到其他任务上。针对情感极性分析问题，后续工作中会结合 RNN、Transformer 等相关理论进行研究。

参考文献

[1] 李春林, 武巾莉. 基于机器学习的白酒板块股评情感分析 [J]. 信息技术与信息化,2021(10):139-141.

[2] 沈克琳, 吉秉彧, 李然. 基于卷积神经网络的英文篇章情感量化方法 [J]. 信阳师范学院学报 (自然科学版),2021,34(01):130-137.

[3] 胡晓菁, 曲春歌, 冯媛. 基于人工智能技术的邮政舆情监测分析研究 [J]. 邮政研究,2021,37(06):35-40.DOI:10.13955/j.yzyj.2021.06.07.06.

[4] 常城扬, 王晓东, 张胜磊. 基于深度学习方法对特定群体推特的动态政治情感极性分析 [J]. 数据分析与知识发现,2021,5(03):121-131.

[5] 方博平, 郭佳怡, 陆欣怡, 王梦怡, 宋涛. 基于文本挖掘技术的智慧政务舆情分析研究 [J]. 科技风,2021(34):86-88.DOI:10.19392/j.cnki.1671-7341.202134029.

[6] 王婷, 杨文忠. 文本情感分析方法研究综述 [J]. 计算机工程与应用,2021,57(12):11-24.

[7] 邱祥庆, 刘德喜, 万常选, 李静, 刘喜平, 廖国琼. 文本情感原因自动识别综述 [J/OL]. 计算机研究与发展:1-30[2021-12-24].

[8] 周晓兰, 戴香平, 陈洪龙. 基于朴素贝叶斯模型的评论文本情感分析 [J]. 科学技术创新,2021(33):88-90.

[9] 杨青, 张亚文, 朱丽, 吴涛. 基于注意力机制和 BiGRU 融合的文本情感分析 [J]. 计算机科学,2021,48(11):307-311.

[10] Li B , Zhou H , He J , et al. On the Sentence Embeddings from Pre-trained Language Models[J]. 2020.

[11] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. arXiv, 2017.

[12] Nguyen, Tuan-Linh, Kavuri, Swathi, and Lee, Minho. A Fuzzy Convolutional Neural Network for Text Sentiment Analysis. 1 Jan. 2018 : 6025 ‑6034.