学校代码：10270　　分类号：TP391.1　　学号：182502742

上海师范大学

硕士专业学位论文

基于 FV-SA-SVM 的电影评论情感分析

学　　院：数 理 学 院

专业学位类别：应 用 统 计 硕 士

专 业 领 域：应 用 统 计

研 究 生 姓 名：张 成 博

指 导 教 师：崔 百 胜

完 成 日 期：2020 年

论文独创性声明

本论文是我个人在导师指导下进行的研究工作及取得的研究成果。论文中除了特别加以标注和致谢的地方外，不包含其他人或机构已经发表或撰写过的研究成果。其他同志对本研究的启发和所做的贡献均已在论文中做了明确的声明并表示了谢意。

作者签名：张成博　　日期：2020-06

论文使用授权声明

本人完全了解上海师范大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它手段保存论文。保密的论文在解密后遵守此规定。

作者签名：张成博　　导师签名：崔百胜　　日期：2020-06

论文题目：基于 FV-SA-SVM 的电影评论情感分析

学科专业：应用统计

学位申请人： 张成博

指导老师：崔百胜

摘 要

近年来，中国经济飞速发展，早已成为全球第二大经济体，人民生活水平不断提高，人民享受生活的方式越来越多样化，看电影则是主要的形式之一。随着观众人数的激增，电影市场规模也在逐渐扩大。根据国家电影局数据显示，2018 年全国电影总票房为 609.76 亿元，我国已坐稳全球第二大电影市场。

21 世纪，互联网技术日新月异，随着手机的普及以及众多观影 APP 如雨后春岁般的涌出，人们可以随时随地的在网上购买电影票以及发表观影评论。根据观众的影评，可以得知他们的情感倾向，深度剖析观众对于电影的看法，从中总结出优点以及不足，指引电影业朝着更好的方向发展，使得观众影评的价值最大化。

本文从猫眼 APP 爬取了动作、喜剧、青春以及悬疑四个类型共八部电影的影评，首先对影评进行预处理，然后使用 FV-SA-SVM 将影评划分为积极评论和消极评论两类，结果显示 FV-SA-SVM 算法的准确率分别达到了 97.8%、95.3%、96.1%以及 97.4%。接着将这种分类算法与 SA-SVM 算法、传统分类算法进行比较，发现 FV-SA-SVM 算法的准确率、精确率、召回率以及 F1-Score 这四个指标均优于 SA-SVM 算法和传统分类算法，从而验证了 FV-SA-SVM 算法在影评情感分类上的优越性能。而后将此模型应用于《上海堡垒》、《战狼 2》以及《红海行动》这三部动作电影的情感分类，分类准确率分别是 94.7%、93.4%以及 92.9%，均优于 SA-SVM 以及传统分类算法。

接下来对分类后的数据使用情感字典打分的方法获取影评的情感值进行统计分析；而后对影评进行语义网络分析，构建语义网络，实现数据可视化；再对影评进行主题挖掘，提取出现次数最多的前 4 个主题，挖掘电影特

征；最后使用 k-Means 算法进行聚类分析，通过对比聚类结果发现动作片中观众最在意的是电影特效、演技动作这两个方面，低分动作片是因为这两个方面做的并不好，最后本文根据这两个方面提出了相关的建议。

本文研究的结论，即可以帮助电影制片商了解观众需求、从而改进自身电影作品，同时也可以帮助其他观众选择是否观看此电影。

关键词：FV-SA-SVM；情感分析；k-Means；LDA；自然语言处理；

Abstract

In recent years, with the rapid development of China's economy, China has already become the second largest economy in the world. People's living standards are constantly improving, and people's ways of enjoying life are becoming more and more diversified. Watching movies is one of the main forms. As audiences have soared, so has the market. According to the State Film Administration, China's box office totaled 60.976 billion yuan in 2018, making it the world's second-largest movie market.

In the 21st century, Internet technology is changing with each passing day. With the popularity of mobile phones and numerous movie-watching apps, people can buy movie tickets and post comments online anytime and anywhere. Through the user's comment text data, we can know their emotional tendency, further analyze the audience's views on movies, and summarize the advantages and disadvantages, so as to guide the film and TV industry to develop in a better direction and maximize the value of the user's comment text data.

This article crawls the film reviews of eight films in four categories: action, comedy, youth and suspense from the Cat's Eye APP. First, the film reviews are pre-processed, and then FV-SA-SVM is used to divide the film reviews into positive and negative reviews The results show that the accuracy of the FV-SA-SVM algorithm reaches 97.8%, 95.3%, 96.1%, and 97.4%, respectively. Then compare this classification algorithm with SA-SVM algorithm and traditional classification algorithm, and find that the accuracy, precision, recall and F1-Score of FV-SA-SVM algorithm are better than SA-SVM algorithm and Traditional classification algorithm, thus verifying the superior performance of FV-SA-SVM algorithm in the sentiment classification of movie reviews. Then applied this model to the emotion classification of the three action movies of "Shanghai Fortress", "War Wolf 2" and "Red Sea Action". The classification accuracy rates were 94.7%, 93.4% and 92.9%, which were better than SA-SVM and traditional classification algorithms.

Next, the sentiment dictionary score method is used to obtain the sentiment

value of the movie review for statistical analysis. Then, the semantic network analysis of the movie review is carried out to construct a semantic network to realize data visualization. Then the subject of the movie review is mined to extract the most frequent occurrences. The first four themes, mining movie features; Finally, k-Means algorithm is used for cluster analysis. By comparing the clustering results, it is found that the audiences most concerned about the movie special effects and acting actions are the two aspects of the action movie. The low score action movie is because These two aspects are not good, and finally this article puts forward relevant suggestions based on these two aspects.

The conclusion of this paper is that it can help film producers understand the audience's needs and improve their own film works, and also help other viewers choose whether to watch this movie.

KeyWords： SA-SVM; sentiment analysis; k-Means; LDA; natural language processing;

目 录

第 1 章 绪论

1.1 研究背景

电影一直是我们平时的生活娱乐中不可或缺的一部分，最近这几年，凭借着国民经济的连续高速增长以及社会和国家对于文化产业的大力支持，整体上中国电影产业的发展环境不断优化和提升，中国电影产业在 2009-2018 年这十年里飞速发展，票房以年均 35%的速度节节攀升，行业内称“黄金十年”。

从图 1-1 中可以看出我国电影票房增长迅猛，根据文化部相关数据表示，国内电影票房市场近年来维持快速增长的发展趋势，观影人次从 2012 年的 4.4 亿人次，增长至 2018 年的 17.16 亿人次，复合年均增长率达到 25.5%。国内电影票房从 2012 年的 170.7 亿元增长至 2018 年 609.76 亿元，复合年均增长率达到 23.6%。我国电影产业在国民经济新的发展形势下实现了快速增长。只凭借对电影票房的收入进行衡量，中国电影市场已经发展成为继美国后的全球第二大电影市场。

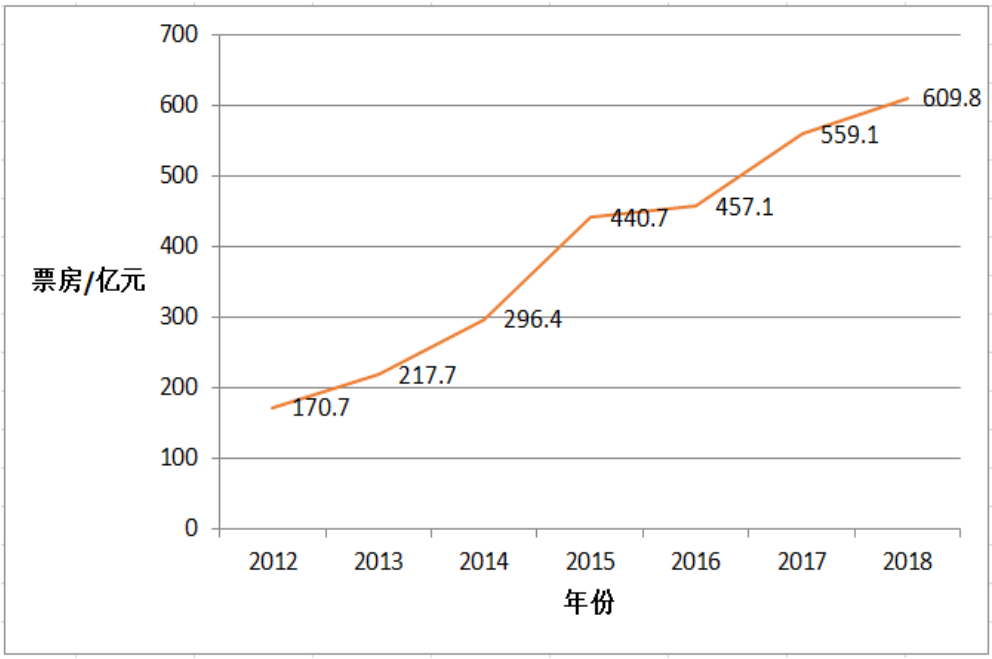


图 1-1 2012-2018 年全国票房统计情况

2018 年一大批国产电影通过口碑实现逆袭，取得了优异成绩，有六部电影票房超 20 亿元。2018 年度电影票房排行榜前十名有六部国产电影，四部引进片皆来自美国。其中《红海行动》、《唐人街探案 2》、《我不是药神》、《西虹市首富》四部国产影片占据 2018 年电影票房前四。其中《红海行动》以 36.5 亿票房成绩问鼎年度冠军。

依照如此的发展趋势，2020 年将很有可能实现中国票房超过北美的黄金交叉，中国电影市场也将从 2020 年开始长期维持全球票房第一，而中国电影的巨大市场规模也将是中国电影产业发展的核心动力。

1.2 研究的目的和意义

1.2.1 研究目的

这几年优秀的国产电影层出不穷、大放异彩，尤其是 17 年上映的《战狼 2》以及 18 年上映的《红海行动》分别凭借 56.8 亿元和 36.5 亿元人民币票房夺得 17 和 18 年的年度冠军。虽然好电影层出不穷，但是期初人们寄予厚望结果票房惨淡的电影也不是没有，例如 2019 年上映的同为动作片的《上海堡垒》，它的总票房只有 1.2 亿人民币，可以说是令人大跌眼镜。

所以本文有两个目的：首先，提出一种新的分类模型，即 FV-SA-SVM，之后对不同类型的电影影评进行情感分类研究，并和传统模型进行对比，来验证本文所用模型是否优于传统模型。

其次通过对《上海堡垒》、《战狼 2》以及《红海行动》影评的统计分析并且对比，了解观众对于动作片的情感倾向，观众评赏动作片时主要关注电影哪些元素，还研究了好评、差评的形成因素。以及基于 FV-SA-SVM 模型进行影评分类，并且与情感字典打分相结合，获得更好的情感分类效果。

本文还将通过对影评进行聚类分析以及 LDA 主题建模，获得影评中观众关注的电影重要方面，从观众的不同角度来获取他们对电影的不同看法，可以充分的了解电影的优点以及不足，总结得出有用结论，为以后的同类型电影的发展提供有价值的建议。

1.2.2 研究意义

1. 理论意义

近几年，数据挖掘情感分析已经成为一个关注的热点和研究的重要课题，其中的电影评论文本情感分析更是成为热中之热，越来越多的专家和学者通过挖掘电影评论文本进行情感分析。在互联网影评情感分析领域里，从目前的数据挖掘研究成果总体来看，仍然以机器学习的方法为主，研究结果还有很多需要改善的地方和空间，SA-SVM 是针对 SVM 的一个改进方法，因为 RBF 函数具有收敛域宽、参数少、通用性好等诸多优点，是一个很好的文本分类依据函数，所以采用 RBF 函数建立 SVM，但是 SVM 的性能同时受到惩罚系数 C 和核函数系数 σ 的影响。

通过模拟退火算法(SA)寻找 SVM 的最优参数，跳出局部最优解从而获得全局最优解，传统 TFIDF 方法只是考虑了特征词的总频率而忽视了其在不同类别中的分布。为了同时考虑到特征词的总频率以及在类别中的分布，本文使用引入类频方差（FV）的 TF-IDF（FV-TFIDF）权值计算方法，再将 SA-SVM 与 FV-TFIDF 结合构建一种新的分类器−FV-SA-SVM, 使得影评数据的分类效果最大化。

2. 现实意义

本篇所研究的结果，对观众、电影片商都具有很重要的应用价值。第一，对于观众来说，本文的研究相关结论可以辅助他们了解电影的优缺点，从其他人的影评中找到电影吸引自己的地方，从而快速选择是否观看此电影；第二，对于电影片商来说，本文的研究结果不仅可以帮助他们更详细的了解自身电影作品的观众口碑，从而加强自身电影作品的更新改造，而且还能帮助他们快速、准确地洞察市场竞争的相关现状，从而准备更充分的应对方法去占领市场空间。综上所述，本文的研究还是非常有现实意义的。

1.3 研究内容、方法和技术路线

1.3.1 研究内容

1.通过 python 爬虫技术从猫眼 APP 爬取中文数据。

2.文本预处理。

第一步分词，在这里我们运用 Python 里的 jieba 分词。为了提升分词的准确率，需要在分词前导入自定义词典。本研究选择导入 NTUSD 情感词典，除此之外还要将电影中出现的专有名词导入其中。

第二步，去停用词。使用将哈工大停用词处理词库、百度停用词等诸多停用词表进行合并去重操作，设计出的含近 1800 个停用词的词典，将出现的没有任何意义的词语删去。

第三步，计算权重。在大多数实验过程中使用 TFIDF 计算权重。TF(term frequency,词频)：指在一个给定文件中，某一特定词在该文件中的次数。IDF(inverse document frequency，逆文档频率)某一特定词语的 IDF，可由文件库中的文件总数除以该特定词出现的文件数，之后将计算得到的值取个对数。公式如下：

$$tf\text{-}idf_{i,j}=\frac{n_{i,j}}{\sum_k n_{i,j}}\times\log\left(\frac{n_d}{df(d,w_i)+1}\right) \tag{1.1}$$

其中，$n_{i,j}$ 是指特定词 i 在文件 j 中的词数，而第一个分母是指文件 j 中的总词数。n_d 是所有文档总数，$\mathrm{df}(d,w_i)$ 是文档库 d 中所包含的词语 w_i 的文档个数，加 1 是为了防止 $\mathrm{df}(d,w_i)$ 出现为零的情况。

但它也有其缺点，单纯以“词频”衡量一个词的重要性，不够全面，有时重要的词可能出现次数并不多。本文使用一种引入类频方差（Frequency variance）的 TFIDF 权重计算方法（FV-TFIDF），并与 SA-SVM 分类算法相结合，本文称之为 FV-SA-SVM。且与 SA-SVM 以及传统的分类算法进行对比。

3.将预处理后的数据分类成积极和消极两类，再对已分类数据通过情感字典打分，并进行语义网络图谱的构建。

4.使用 LDA 主题模型分别研究电影的积极、消极影评，获取积极、消极影评中主题的分布情况，不同的情感，对各个主题的出现次数按照由大到小的顺序进行排序，将位于前列的潜在主题作为影评的热点，并按照潜在主题上的电影特征词概率分布情况得到影评集合相应的影评词。

5.对分类数据再进行聚类分析，将影评划分为某些类别，找出三部电影影评包含的相同的类别，通过好评率进行比较分析。

6.总结研究过程得出结论。

1.3.2 研究方法

1.数据收集

通过 python 爬虫技术从猫眼 APP 爬取中文数据。

2.文本预处理

预处理步骤包括分词，去停用词。分词工具有很多，例如：哈工大的 LTP、盘古分词等，在这里我们运用 Python 里的 jieba 分词。为了提升分词的准确率，还需在分词前导入自定义词典，本文将电影中出现的词语导入其中。然后进行去停用词，本研究使用将哈工大停用词处理词库、百度停用词等诸多停用词表进行合并去重操作，设计出的含近 1800 个停用词的词典，然后将出现在评论数据中无用词语删去。

3.FV-SA-SVM 分类算法

在中文文本中人工标注多条数据，积极消极各占一半，分别用 0,1 表示，将其称之为标记数据。将标记数据划分为训练集和测试集，测试集占比 0.2。因为计算机只能识别数字，所以我们要将预处理后的文本转化为向量，通常采用 TFIDF 进行权值计算，使用的是朴素贝叶斯，K 最近邻法以及支持向量机（SVM）等分类算法。

但是，在文本分类研究中，文本库中的文本通常被标记成几个不同类别，而 TFIDF 算法只考虑特征词在整个文本库中出现的总频率，忽略了在类别中的分布，例如某个词语 w_i 在文本库中的几个类别的文本中出现频率较高，而在其他几个类别的文本中出现频率较低，说明该 w_i 对文本的判别具有一定贡献，而传统的 TFIDF 算法没有考虑这种不同类别间的分布情况，导致某些对类别判断具有贡献的词丢失[1]。

因此本文提出引入类频方差（Frequency variance）的 TF-IDF 算法，称之为 FV-TFIDF，类频方差衡量的是词语在不同类别的分布情况，计算公式如下所示:

$$T_i = \frac{\sqrt{(\sum_{j=1}^{N}(\frac{df(d,w_i)}{N} - df(d_{cj},w_i))^2)}}{N} \tag{1.2}$$

T_i 为词语 w_i 的类频方差，N 为文本类别数，$df(d,w_i)$ 为整个文本库 d 中包含词语 w_i 的文档个数，$df(d_{cj},w_i)$ 为在类别 c_j 中包含词语 w_i 的文档个数。T_i 越大说明词语 w_i 在类别中波动越大，分布越不均匀，对类别的判断作用越大，所以基于类频方差（Frequency variance）的 TF-IDF 计算公式如下所示:

$$FV-TFIDF = tf - idf \times T_i \tag{1.3}$$

大量研究表明在中文文本分类上 SVM 具有良好的泛化能力，且 RBF（径向基核函数）具有收敛域宽、参数少、通用性好等优点，所以采用 RBF 核函数建立 SVM。

基于 SVM 的文本分类性能与其惩罚因子 C 和核函数 RBF（径向基核函数）参数 σ 等密切相关，直接影响文本分类精度[2]。所以本文引用 FV-SA-SVM 文本分类方法，其原理是将引入类频方差（Frequency variance）的 TFIDF 权重算法与利用 SA（模拟退火）对于 SVM 良好的寻参能力相结合构建 FV-SA-SVM 分类器。

最后使用 FV-TFIDF 权重算法与模拟退火算法（SA）构建的 FV-SA-SVM 分类器进行模型训练，并通过准确率(A)、召回率(R)、精确率(P)以及 F1-score 这四个指标来和传统分类算法进行对比，验证其高效性、有用性。

改进后的分类算法如下图：

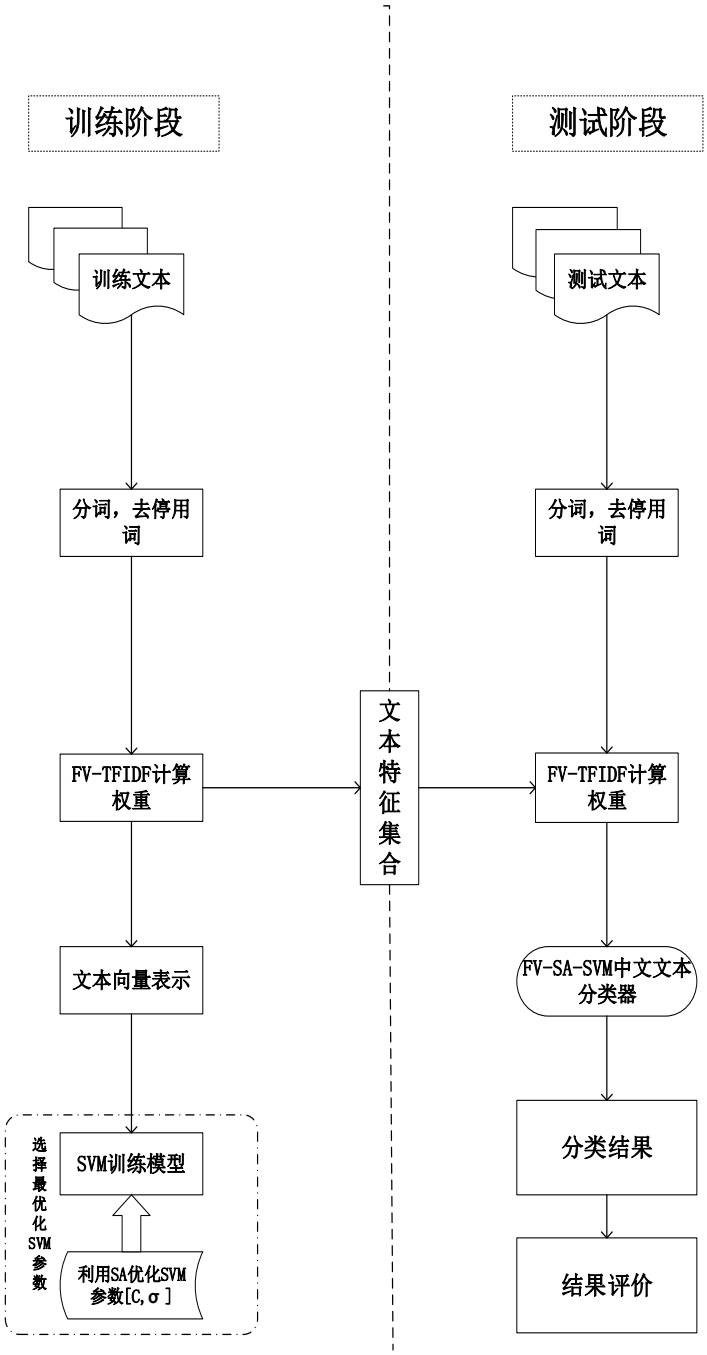


图 1-2 改进后的分类过程

准确率(Accuracy)：预测结果和真实结果占整个样本总数的百分比即为准确率。公式如下：

$$A=\frac{TP+TN}{TP+TN+FP+FN} \tag{1.4}$$

召回率(R)：是针对我们原来样本而言，表示有多少样本中的正例（一种是把正类预测为正类即 TP，一种是把正类预测为负类即 FN）被预测正确了。公式如下：

$$R=\frac{TP}{TP+FN} \tag{1.5}$$

精确率（P）：是针对我们的预测结果而言，表示的是预测为正的样本中，有多少个是真正的正样本。公式如下：

$$P=\frac{TP}{TP+FP} \tag{1.6}$$

F1-score:作用就是调和 P 和 R 的矛盾。公式如下：

$$F=\frac{2PR}{P+R}=\frac{2TP}{2TP+FP+FN} \tag{1.7}$$

4.语义网络的构建

语义网络主要用于更好的理解自然语言和加深对科学领域的认知，是一种语言的概念关系的表达。语义网络是一幅由节点和弧构成的语义网络描述图，节点表示概念、事件，是各种可以用文字表达的事物，有向弧表示这些事件存在的语言意义上的关系，弧方向是语言关系的因果指向，例如：A→B，A 是主动方，B 是被动方，A 与 B 有语义联系。语义网络的每一条弧代表了两个概念之间的语义关系，随着词汇概念的增多，就会成为复杂的语义网络[3]。

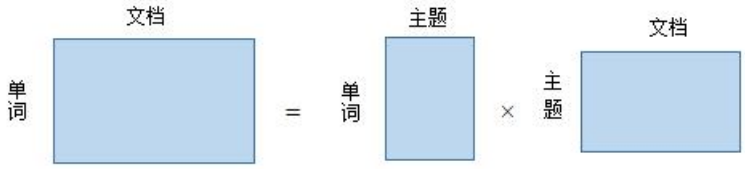
我们使用 ROSTCM6 工具进行分析时，首先分别对积极和消极评论重新进行预处理，并提取出高频词，这些高频词之间的语义联系才是我们真正需要去采集的，其余的个性化词汇之间的关系不具有代表性，然后把显著的无意义的成分过滤掉，最后抽取行特征，构建语义网络[3]。

5.LDA 模型进行主题抓取。

主题模型通过对影评中隐含的主题进行挖掘，能够把两个通过词特征被认定为没有相似性的词汇以一定概率放在同一主题下，从而提取影评中主题相关度的方法。生成模型，就是认为每一篇文章的每一个词都是通过以“一定的概率选择了某个主题，并从这个主题中以一定的概率选择了某个词语”的过程得到的。由此，如果要产生一篇文章，每个词语出现的概率可表示为如下公式：

$$P(\text{词语}|\text{文档}) = \sum_{\text{主题}} P(\text{词语}|\text{主题}) \times P(\text{主题}|\text{文档}) \qquad (1.8)$$

上式的概率公式可以用矩阵图表示为：



其中“文档-单词”矩阵表示每个文档中每个单词的词频，即出现的概率；“主题-单词”矩阵表示每个主题中每个单词的出现概率；“文档-主题”矩阵表示每个文档中每个主题出现的概率。我们首先对文本进行了分词等工作，然后统计计算出逐个单词词频， 就可以得到第一个矩阵，LDA 模型就是通过第一个矩阵得到第二、三个矩阵得以实现的。 文档生成过程如下图所示：

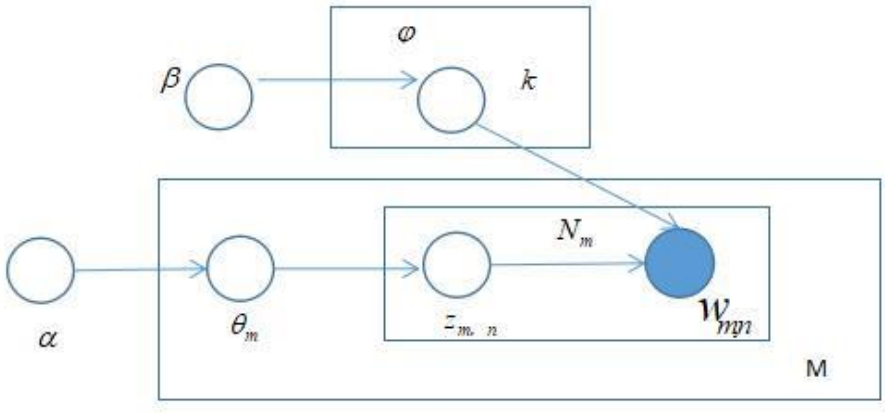


图 1-3 LDA 生成文档示意图

图 1-3 中，主题总数为k。文档个数是m，m个文档的总单词数N_m。α是每个文档下主题（Topic）的多项分布的 Dirichlet 先验参数，β是每个主题下词的多项分布的 Dirichlet 先验参数。$Z_{m,n}$是第m个文档中第n个词的主题，$W_{m,n}$是m个文档中的第n个词。剩下来的两个隐含变量θ_m和φ分别表示第m个文档下的主题（Topic）分布和第k个主题下词的分布。LDA 目的是根据已有数据来构建θ和φ，即估计文档-语义以及语义-单词层级的概率分布。图 1-3 中的 z 为未知变量，而θ和φ分别有一个带有超参数的α和β的 Dirichlet 先验分布，所以 LDA 的目的就是估算α和β，LDA 模型的参数含义和参数估计对于提取表示文档集的特征非常重要。

6.文本聚类

使用划分聚类中最经典的 k-Means 聚类算法，分别得到评论关注的几个方面，比如：人物、剧情、演技、特效。

步骤如下：

输入：需要输入分类簇的数目 K 以及包含 n 个数据的集合。

输出：K 个聚类完成的簇

步骤 1：从数据中选择 K 个数据作为初始聚类中心；

步骤 2：计算给定的数据集分别到初始化聚类中心的几何距离

步骤 3：按照距离最短原则将每个数据分配到最邻近的簇中

步骤 4：使用每个簇中样本数据的几何中心作为新聚类中心；

步骤 5：反复迭代算法中步骤 2、步骤 3 和步骤 4 直到算法收敛为止

步骤 6：算法结束，得到输出结果。

1.3.3 技术路线

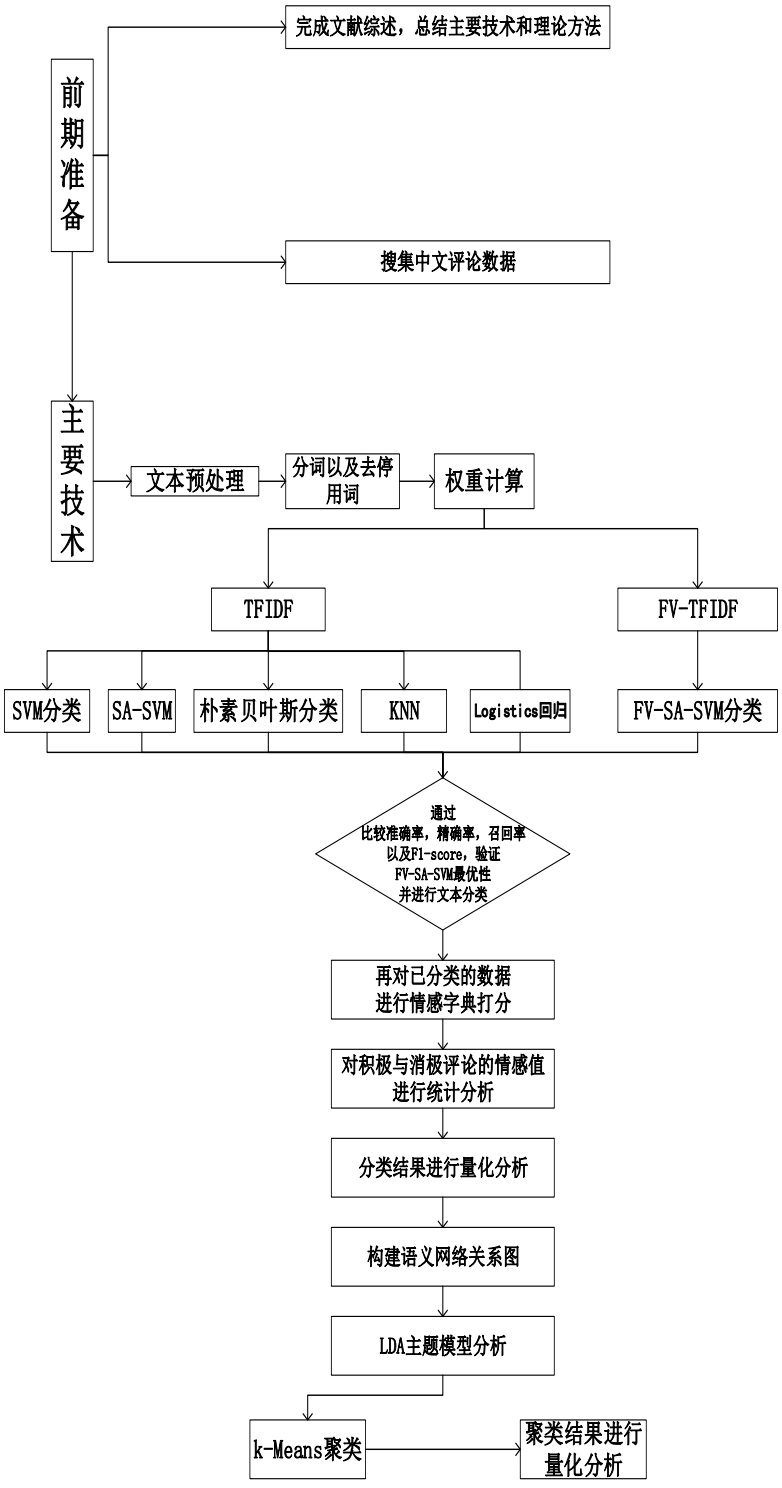

前期准备
完成文献综述，总结主要技术和理论方法
搜集中文评论数据
主要技术
文本预处理
分词以及去停用词
权重计算
TFIDF
FV-TFIDF
SVM分类
SA-SVM
朴素贝叶斯分类
KNN
Logistics回归
FV-SA-SVM分类
通过比较准确率，精确率，召回率以及F1-score，验证FV-SA-SVM最优性并进行文本分类
再对已分类的数据进行情感字典打分
对积极与消极评论的情感值进行统计分析
分类结果进行量化分析
构建语义网络关系图
LDA主题模型分析
k-Means聚类
聚类结果进行量化分析

图 1-4 技术路线图

1.4 本文创新点

1.首先，本文聚焦于电影领域，将文本情感分析以及主题挖掘的相关理论用于分析近两年票房冠军电影（《战狼 2》、《红海行动》）影评中的观众情感态度，相比较于文本情感分析方法在其他领域（如手机、新闻评论等）中的更多运用，本文的研究在应用层面比较创新。

2.其次因为传统 SVM 分类算法性能会受到惩罚系数 C 以及核函数 RBF（径向基核函数）参数 σ 等密切相关，直接影响文本分类精度。所以本文使用了一种基于模拟退火算法(SA)寻找 SVM 最优参数的分类算法，并将另外一种基于类频方差改进的 TFIDF 特征提取方法（FV-TFIDF）与其相结合，构建一种全新的 FV-SA-SVM 分类算法进行文本数据的情感分析，并与参考文献中的 SA-SVM 分类算法以及传统的 KNN、朴素贝叶斯以、Logisitic 回归、SVM 等分类器通过比较准确率、精确率、召回率以及 F1-score 来验证新算法性能更好。

3.通过 FV-SA-SVM 算法将影评进行分类之后再使用情感字典打分的方法对已经分类的评论进行打分，将打分结果与之前分类结果相异常的影评筛选出来重新进行人工分类，尽可能地保证数据分类的正确性，极大地降低了错分的概率。

4.本文研究的是近两年票房冠军电影，并且还是同一类型的，具有时效性与新颖性，之前并没有人做过。

第 2 章 文献综述与相关理论

2.1 文献综述

2.1.1 文本情感分析定义

文本情感分析，又被称为观点识别，情感挖掘等。是指对带有情感色彩的主观性文本进行采集、处理、分析、归纳和推理的过程，涉及到人工智能、机器学习、数据挖掘、自然语言处理等多个研究领域。情感分类是其核心研究问题，主要包括两类研究任务，一是对主客观文本进行区分，降低对情感文本的噪音影响，提高情感分析性能；另一个是对主观性文本进行情感分类[4]。根据情感划分的粒度，包括：①二元分类，主要指积极情感/消极情感；②多元分类，根据人们情绪表达进行细分，将情感分为“快乐、悲哀、褒扬、贬斥、信心和意外”等十大类[5]；根据文本划分的层次粒度，包括面向深层次语义对象的细粒度情感分析和面向句子和篇章的粗粒度情感分析。

互联网的快速发展推动了电子商务以及各大网络平台的普及，互联网用户由信息接受者向信息发布者转变，不仅关注其他人对某件事或某件商品的评论和看法，也越来越乐于分享自己的意见看法或表明态度。因此，有必要对各大平台上的带有情感色彩或情感倾向的言论进行整理分析，这可能带来多方面的效益，例如分析购物网站上的商品评论便于消费者详细了解商品信息，从而优化消费决策，也便于企业根据消费者的反馈信息掌握自己的优劣势，从而优化战略决策；针对某一舆情，通过对相关微博的评论进行整理分析，政府可以将其作为颁布相关政策的有力依据；针对电影评论而言，通过进行情感分析不仅可以引导观众的观影决策， 而且可以使制片商调整他们的营销策略。

所以，对含有情感色彩的文本进行情感极性判断具有巨大的商业价值和社会价值。

2.1.2 情感分析方法

主流的情感分析研究方法分为两种类型：无监督的文本情感分析和有监督的文本情感分析。无监督的文本情感分析主要是运用情感词的相关信息进行文本情感倾向判别。有监督的文本情感分析主要是运用朴素贝叶斯、支持向量机等有监督学习算法进行情感分类。

基于情感知识构建情感词典并将其作为工具是判断主观性文本情感极性的传统方法，大部分的情感词典是人工构建的，基本原理是根据经验将广泛使用的情感词进行归纳整理，当文本输入后就与词典内容进行匹配，寻找文本中与情感词典中重合的情感词，从而判断文本的情感极性。

Hatzivassiloglou 等(1997)[6]的研究表明形容词能很好的表示句子的主观性和情感倾向，他们提出了通过计算单个形容词的语义倾向值的方法进行文本情感倾向的识别。Turney 等(2002) [7]则认为词组比单词能更准确的表达情感，他们的实验抽取了语料库中特定模式的情感词组，通过计算抽取词组的语义倾向平均值识别文本整体的情感倾向。

Theresa Wilson 等(2010)[8]首先提出了一种新的短语级情感分析方法，确定了一句表达是中性的还是极性的，然后消除极性表达式的极性的相关歧义，通过采用这种方法，系统能够自动识别情感表达的大部分子集的情境极性，实现显著超过基线的结果。

Taboada 等(2011) [9]利用情感词典中情感词的关联度和情感倾向强度等信息计算文档的情感得分，以此进行文本情感倾向识别。

而在近几年，Erik Cambria 等(2013)[10]运用大量有价值的非结构化信息进行公众观点的挖掘和情感分析。Tajinder Singh 和 Madhu Kumari(2016)[11]使用 n-gram 和条件随机字段来检查俚语的重要性，所提出的预处理方法依赖于俚语词汇与其他共存词语的结合，来检查俚语词汇的意义和情感翻译，表明了预处理对 Twitter 数据情感分析的影响。

因为英文词典资源丰富且具有优势，于是李寿山等(2013) [12]利用英文种子词典，借助机器翻译系统，构建了中文情感词典。上述的情感词典是最基础的情感词典，情感词覆盖率较低，无法结合语境、语义，也无法识别同义词、近义词等，所以在后来的研究中学者们对情感词典进行了进一步完善。

考虑到语境迁移的影响，现有的情感词典应用到旧词新用的语料中分类效果较差，阳爱民等(2013) [13]用若干个情感种子词计算基础情感词的情感倾向值，利用搜索引擎返回的共现数构建了情感词典。王志涛等(2015) [14]利用 40 万条微博数据构建新词词典，对已有情感资源进行拓展，并对不同语言层次定义不同的规则，还以表情符号作为附加信息提供辅助作用。

周杰(2016)[15]通过比较分析基于情感词典的方法和基于机器学习的方法，针对两者的不足提出一种新的基于情感词典和句型分类的中文情感分析方法，并利用拉普拉斯平滑的情感倾向点互信息 SO-PMI 算法对微博情感词典进行扩展，深入分析不同句型对句子情感倾向的影响。

张克亮(2016)等[16]融合了情感词典资源和概念层次网络语境框 架的优势，将文本的情感分析分为两个阶段：特征词、句子和句群判定阶段；基于 HNC 语境框架的句与句群情感分析阶段。上述学者改进的情感词典主要是用于中文文本的情感分析，但随着英文的国际化，网友们在网络平台上发布的言论中除了中文还夹杂着少许英文甚至混合着多种语言，可见上述研究存在一定的局限性。

栗雨晴等(2016) [17]考虑到目前文本情感分析工作多针对单一语种的状况，提出了一种双语的情感词典，然后利用半监督高斯混合模型分类算法和基于对称相对熵的 K 近邻算法对微博文本进行情感分类，双语词典的提出和成功应用是一种包含时尚性的进步。Vileras(2017)等[18]为解决 Twitter 上多语言极性分类问题引入了带有情感标签的代码转换 Twitter 语料库。

学习是人类具有的一种持续性智能行为，目前计算机也已经初步具备了这种能力，即机器学习。基于机器学习对文本进行情感分析的原理是人工提取文本特征后由计算机根据某种特定的算法对文本进行处理然后输出情感分类。相较于完全依赖人工构建情感词典的方法，机器学习具有明显的优势，一方面能有效地缓解劳动力的负担且减少非理性判断，另一方面能构建庞大的数据库且能根据时代发展及时对词库进行更新。

在机器学习方法中朴素贝叶斯 NB 和支持向量机 SVM 是常用的监督学习算法，但是有研究指出，NB 和 SVM 单独使用时分别会面临独立条件假设和核函数选择方面的问题，所以 Sharma 和 Dey(2013)[19]通过使用 Boosting 技术整合“弱”支持向量机分类器，利用了 Boosting 的分类性能，同时使用 SVM 作为基础分类器，研究结果表明集成分类器在准确率上明显优于单

纯的 SVM 分类器。Manek 等(2017)[20]提出了基于基尼指数的支持向量机分类器的特征选择方法。

评论类的文本缺乏逻辑性，文本多呈无序性，一般的监督学习算法处理无序文本时准确率较低，Perikos 和 Hatzilygeroudis (2016)[21]设计了一种集成分类器，它是基于 3 个分类器：第 1 个和第 2 个是统计学（朴素贝叶斯和最大熵），第 3 个是基于知识的工具，对自然语言句子进行深入分析。类似地，Tripathy 等(2016)[22]将文本分别以 1 个词、2 个词、3 个词以及其组合的方式划分，然后分别用朴素贝叶斯、最大熵、随机梯度下降和支持向量机的方法进行评论情感分析。

传统的文本情感分析方法主要有人工构建情感词典的方法或基于监督的机器学习模型，但是这 2 种方法不仅耗费大量的人力，而且在大数据时代任务完成效率和任务完成质量较低。深度学习可以通过构建网络模型模拟人脑神经系统对文本进行逐步分析、特征抽取且自动学习优化模型输出，以提高文本分类的正确性。

神经网络模型的使用不可避免地要涉及词向量嵌入技术，即将人类语言转换成机器语言，例如 Word2Vec，Giatsoglou 等(2017)[23]将 Word2Vec 提供的上下文敏感编码与词典提供的情感信息相结合。 虽然词向量嵌入技术考虑了单词的上下文，但是忽略了整体文本的情感，Tang 等(2016)[24]提出通过在情感嵌入中将文本的情感信息连同词的情境一起编码来解决这个问题，并开发了一个具有裁剪损失功能的神经网络自动收集情感信号。Fernandez－Gavil－anes 等(2016)[25]提出新的无监督情感分析算法，该算法使用了依存句法来判断情感的极性。在进行文本情感分析任务时，经常出现同一句子中有情感极性不一致的多个情感词，梁斌等(2017)[26]认为注意力机制能有效解决上述问题，于是将词向量注意力机制、词性注意力机制和位置注意力机制相结合构造了多注意力卷积神经网络。

2.1.3 情感分析方法改进

目前大多数文献研究的是基于改进机器学习算法的情感分类方法。因为它相比于情感词典的方法具有明显的优势。众多学者在研究情感分类的过程中，提供了许多优秀的分类算法，Shathi 等(2017)[27]将贝叶斯算法应用于文

本分类中；Bahassine 等(2016) [28]使用决策树算法对文本进行分类；Goudjil 等(2018) [29]采用 SVM 算法对文本分类进行技术研究。经过大量实验表明，在中文文本分类上，SVM 具有较强的泛化能力。基于 SVM 的文本分类性能与其惩罚因子 C 和核函数参数 σ 等密切相关，直接影响文本分类精度[30-31]。庄严等(2011) [32]提出了基于蚁群优化算法(ACO)的支持向量机选取参数算法；陈晋音等(2018)[33]提出了基于粒子群算法(PSO)的支持向量机的参数优化。

2.1.4 文献评述

通过对以上文献的梳理，可以发现学术界对情感分析的研究在不断的推进，但是仍然存在以下几个方面的问题：

1. 以上学者在研究文本情感方面虽然已经意识到了传统的情感词典的局限性并做出了改进，但是仍有一定的局限性，他们并没有突破情感词典的限制。情感词典无论怎样拓展完善都存在“词典”这一边界，它无法涵盖所有情感表达形式且随着时代发展出现的新词无法及时涵盖进去，这使得文本情感判断准确率较低。

2. ACO 算法的收敛速度较慢易陷入局部最优，PSO 算法易早熟收敛且局部寻优能力较差。

模拟退火算法(SA)也是一种启发式算法[34]，能较强地跳出局部最优，提高全局寻优能力。所以本文使用一种基于模拟退火算法优化 SVM 参数的方法，并应用于中文情感分类中。利用 SA 良好的寻优性能构建的 SVM 中文文本分类器。

2.2 相关理论

2.2.1 支持向量机

支持向量机（support vector machines）是一种二分类模型，它的目的是寻找一个超平面来对样本进行分割，分割的原则是间隔最大化，最终转化为一个凸二次规划问题来求解。由简至繁的模型包括：

1.当训练样本线性可分时，通过硬间隔最大化，学习一个线性可分支持向量机；

2.当训练样本近似线性可分时，通过软间隔最大化，学习一个线性支持向量机；

3.当训练样本线性不可分时，通过核技巧和软间隔最大化，学习一个非线性支持向量机；

给定训练样本集 $D=(x_1,y_1),\cdots,(x_m,y_m)$，其中 $y_i\in[-1,+1]$，分类学习最基本的想法就是基于训练集 D 在样本空间中找到一个划分超平面，将不同类别的样本分开。

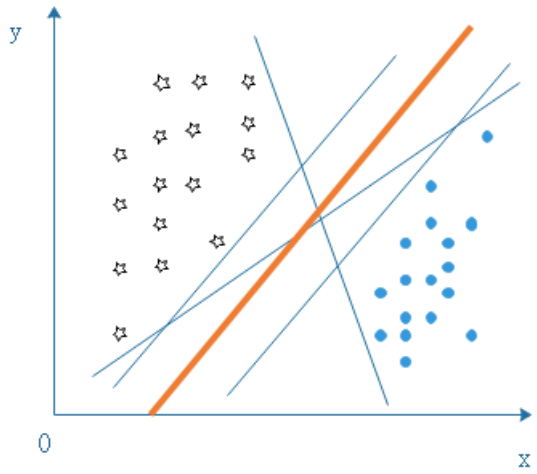


图 2-1 存在多个划分超平面将两类样本分开

直观看上去，能将训练样本分开的划分超平面有很多，但应该去找位于两类训练样本“正中间”的划分超平面，即图 4 中红色的那条，因为该划分

超平面对训练样本局部扰动的“容忍”性最好，例如，由于训练集的局限性或者噪声的因素，训练集外的样本可能比图 4 中的训练样本更接近两个类的分隔界，这将使许多划分超平面出现错误。而红色超平面的影响最小，简言之，这个划分超平面所产生的结果是鲁棒性的。

如果一个线性函数能够将样本分开，称这些数据样本是线性可分的。那么什么是线性函数呢？其实很简单，在二维空间中就是一条直线，在三维空间中就是一个平面，以此类推，如果不考虑空间维数，这样的线性函数统称为超平面。我们看一个简单的二维空间的例子，O 代表正类，X 代表负类，样本是线性可分的，但是很显然不只有这一条直线可以将样本分开，而是有无数条，我们所说的线性可分支持向量机就对应着能将数据正确划分并且间隔最大的直线。

对于非线性问题，线性可分支持向量机并不能有效解决，要使用非线性模型才能很好地分类。非线性问题往往不好求解，所以希望能用解线性分类问题的方法求解，因此可以采用非线性变换，将非线性问题变换成线性问题。对于这样的问题，可以将训练样本从原始空间映射到一个更高维的空间，使得样本在这个空间中线性可分，如果原始空间维数是有限的，即属性是有限的，那么一定存在一个高维特征空间使样本可分。令 $\varphi(x)$ 表示将 x 映射后的特征向量，于是在特征空间中，划分超平面所对应的模型可表示为：

$$y = \omega^T \phi(\chi) + b \tag{2.1}$$

式中：$\phi(x)$ 为标准正态分布函数，ω 表示权值向量，b 表示偏移向量。

要将数据从低维映射到高维需要使用核函数，用的最多的核函数有线性核函数、高斯核函数以及径向基核函数(RBF)。

2.2.2 朴素贝叶斯分类算法

朴素贝叶斯（NB）是所有贝叶斯模型中最为简单有效的一种，并且在实际应用中也较为成功。它假定一个特征对于给定类别的影响独立于其他特征。对于文本分类来说，假设特征词彼此之间相互独立。NB 分类模型主要建立在贝叶斯定理的基础上，利用概率统计进行学习分类，预测一个文档属于各个类别的可能性，最终将文档归到可能性最大的一类中。

根据贝叶斯定理，一个未知样本 d_i 属于类别 c_j 的后验概率为：

$$P(c_j \mid d_i) = \frac{p(d \mid c_j) \times P(c_i)}{P(d_i)} \quad (2.2)$$

对于所有类别来说，$P(d_i)$均为常数。其中，$P(d_i \mid c_j)$为样本d_i属于类别c_j的条件概率。$P(c_j)$为类别c_j的条件概率，其值为c_j类样本数除以训练样本集总数。样本d_i由其包含的特征词表示，$d_i = (t_{i1}, t_{i2}, \cdots, t_{im})$，$m$为样本$d_i$中特征词的总个数，$t_k$是第$k$个特征词。基于特征独立假设，样本的类条件概率$P(d_i \mid c_j)$可以由样本中出现的特征词的类条件概率求得：

$$P(d_i \mid c_j) = P((t_1, \cdots, t_k, \cdots, t_m \mid c_j)) = \prod_{k=1}^{m} P(t_k \mid c_j) \quad (2.3)$$

其中，$P(t_k, c_j)$表示单词t_k在类c_j中出现的概率。

朴素贝叶斯分类根据公式 2.2，将未知样本归于后验概率最大的类。

$$P(c_j \mid d_i) = \arg\max\{P(d \mid c_j)P(c_j)\} \quad (2.4)$$

朴素贝叶斯分类的优点在于容易实现，且多数情况下取得的效果都比较不错。不过，该模型有效的前提是假设各特征之间相互独立。当满足条件时，分类速度快、准确率高。但实际上特征项之间往往具有一定依赖关系，很难保证相互之间条件独立。

2.2.3 KNN 分类

KNN 分类算法（K-Nearest-Neighbors Classification），又叫 K 近邻算法，是一个概念极其简单，而分类效果又很优秀的分类算法。他的核心思想就是，要确定测试样本属于哪一类，就寻找所有训练样本中与该测试样本“距离”最近的前 K 个样本，然后看这 K 个样本大部分属于哪一类，那么就认为这个测试样本也属于哪一类。简单的说就是让最相似的 K 个样本来投票决定。KNN 算法的结果很大程度上取决于 K 的选择，K 通常是小于等于 20 的整数。

该方法中，通过计算对象间距离来作为各个对象之间的非相似性指标，避免了对象之间的匹配问题，对象间的距离通常用欧氏距离或者曼哈顿距离来衡量：

欧氏距离：$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \quad (2.5)$$

曼哈顿距离： $$d(x,y)=\sqrt{\sum_{k=1}^{n}|x_k-y_k|}$$ (2.6)

此外，KNN 通过依据 K 个对象中占优的类别进行决策，而不是单一的对象类别决策。就是在训练集中数据和标签已知的情况下，输入测试数据，将测试数据的特征与训练集中对应的特征进行相互比较，找到训练集中与之最为相似的前 K 个数据，则该测试数据对应的类别就是 K 个数据中出现次数最多的那个分类。

其算法的描述为：

（1）计算测试数据与各个训练数据之间的距离；

（2）按照距离的递增关系进行排序；

（3）选取距离最小的 K 个点；

（4）确定前 K 个点所在类别的出现频率；

（5） 返回前 K 个点中出现频率最高的类别作为测试数据的预测分类。

该算法只计算“最近的”邻居样本，某一类的样本数量很大，那么或者这类样本并不接近目标样本，或者这类样本很靠近目标样本。无论怎样，数量并不能影响运行结果。可以采用权值的方法（和该样本距离小的邻居权值大）来改进。

该方法的另一个不足之处是计算量较大，因为对每一个待分类的文本都要计算它到全体已知样本的距离，才能求得它的 K 个最近邻点。目前常用的解决方法是事先对已知样本点进行剪辑，事先去除对分类作用不大的样本。该算法比较适用于样本容量比较大的类域的自动分类，而那些样本容量较小的类域采用这种算法比较容易产生误分。

实现 K 近邻算法时，主要考虑的问题是如何对训练数据进行快速 K 近邻搜索，这在特征空间维数大及训练数据容量大时非常必要。

KNN 算法的优点是运行不繁琐，没有参数进行判断；但是局限性在于 K 值的选定耗费时间长，但是相比于其他算法，KNN 算法操作简单，得到了广泛使用。

2.2.4 逻辑回归分类(简称 LR)

LR 的主要思路是将数据拟合到一个 logistic 函数中，然后对事件发生的几率进行预测。首先，对于空间中点的分布和轨迹，是通过线性回归（也

即特征的线性组合）去拟合。但在现实生活中，很多样本用特征的线性回归来构建，已经远远不能满足需求。所以，在建好一个线性回归模型之后，我们可以设置一个阈值 0.5，将 $h_\theta(x) > 0.5$的这些点预测为正例，而将 $h_\theta(x) < 0.5$的样本预测为负例。这里其实就是用到逻辑回归的思想，来解决生活中这样的二分类问题。其中所用到的函数就是上面列出的 sigmoid 函数。sigmoid 函数公式如下：

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2.7}$$

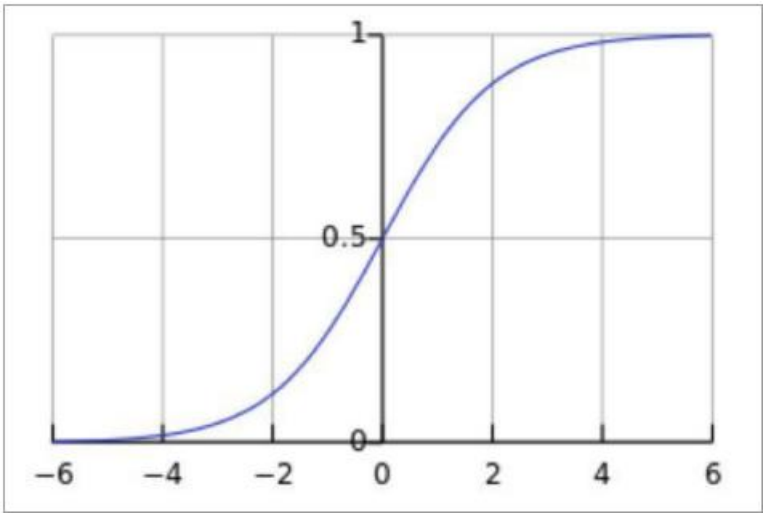


图 2-2 sigmoid 函数

从图 2-2 可以看到 sigmoid 函数的图像是一个 s 形的形状。它其实只在[0,1]之间取值。sigmoid 函数值是随着 x 的值的变小而逐渐变小，最终趋于 0；反之，x的值逐渐变大的同时，会使得函数值逐渐趋于 1。而这个范围正好符合一个概率的取值范围。正是由于 sigmoid 函数这些优良性质，所以很多领域的研究都需要它。

对于任意样本，$X = (x_1, x_2, \cdots, x_n)^T$，可将两分类问题的假设函数写成：

$$h_\theta(x) = \sigma(\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n) = \sigma(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \tag{2.8}$$

其中，$\theta = (\theta_0, \theta_1, \cdots, \theta_n)^T$为未知参数，二分类的判别公式为：

$$y(x)=\begin{cases}1, h_\theta(x)\geq 0.5\\ 0, h_\theta(x)\leq 0.5\end{cases} \tag{2.9}$$

定义整体损失函数：

$$L(\theta)=\frac{1}{m}\sum_{i=1}^{m}\cos t(x_i,y_i) \tag{2.10}$$

对于单个样本的损失函数，

$$\cos t(x_i,y_i)=\begin{cases}-\log(h_\theta(x)), y_i=1\\ -\log(1-h_\theta(x)), y_i=0\end{cases} \tag{2.11}$$

可以写成整体为：

$$\cos t(x_i,y_i)=-y_i\log(h_\theta(x))-(1-y_i)\log(1-h_\theta(x)) \tag{2.12}$$

logistic 回归可以解决分类或回归问题，它的流程就是：第一步，建立代价函数；第二步，通过方法的不断优化和迭代、求解出最优的模型参数组合；最后一步，通过一定的指标来测试和验证模型的好坏。

对 logistic Regression 来说，梯度下降算法如下：

$$\theta^{t+1}=\theta^t-\alpha\frac{\partial L(\theta)}{\partial\theta}=\theta^t-\alpha\sum_{i=1}^{n}(y_i-\sigma(\theta^T x_i))x_i \tag{2.13}$$

其中，参数 α 叫学习率，就是每一步会走多远。但是需要注意的是，这个值的设定也必须在适当的范围，如果设置的太多，那么很容易就造成局部最优问题，即一直只能在最优值附近徘徊。

2.2.5 K-Means 聚类

K-means 算法是最常用的聚类算法，它属于划分聚类的一种，主要思想是:在给定 K 值和 K 个初始类簇中心点的情况下，把每个点(亦即数据记录)分到离其最近的类簇中心点所代表的类簇中，所有点分配完毕之后，根据一个类簇内的所有点重新计算该类簇的中心点(取平均值)，然后再迭代的进行分配点和更新类簇中心点的步骤，直至类簇中心点的变化很小，或者达到指定的迭代次数。

它的优点是原理比较简单，实现也是很容易，收敛速度快。算法的可解释度比较强。主要需要调参的参数仅仅是簇数 k。

它的缺点是需要事先确定分类的簇数，即 k 值；采用迭代方法，得到的结果只是局部最优。对初始值的选取比较敏感。当数据量非常大时，算法的时间开销是非常大的；若簇中含有异常点，将导致均值偏离严重，对噪声和孤立点数据敏感。

2.2.6 LDA 主题模型

在文本挖掘领域，大量的数据都是非结构化的，很难从信息中直接获取相关和期望的信息，一种文本挖掘的方法：主题模型（Topic Model）能够识别在文档里的主题，并且挖掘语料里隐藏信息，并且在主题聚合、从非结构化文本中提取信息、特征选择等场景有广泛的用途。主题可以被定义为“语料库中具有相同词境的词的集合模式”，比如说，主题模型可以将“健康”，“医生”，“病人”，“医院” 集合成 “医疗保健” 主题； “农场”，“玉米”，“小麦” 集合成 “农业”主题。

LDA 模式是生成式模型，在这里，假设需要建模的数据为 X，标签信息为 Y。

判别式模型：对 Y 的产生过程进行描述，对特征信息本身不建模。判别式模型有利于构建分类器或者回归分析生成式模型需要对 X 和 Y 同时建模，更适合做无监督学习分析。

生成式模型：描述一个联合概率分布 P(X,Y)的分解过程，这个分解过程是虚拟的过程，真实的数据不是这么产生的，但是任何一个数据的产生过程可以在数学上等价为一个联合概率分布。

LDA 是一种矩阵分解技术，在向量空间中，任何语料（文档的集合）可以表示为文档（Document - Term，DT）矩阵。下面矩阵表达了一个语料库的组成：

表 2-1 文档−词语矩阵

.	W1	W2	…	Wm
D1	0	2	…	3
D2	1	4	…	0
…	…	…	…	…
Dn	1	1	…	0

其中， N 个文档 $D1, D2, \cdots, Dn$ 的组成语料库， M 个词 $W1, W2, \cdots, Wm$ 组成词汇表。矩阵中的值表示了词 W_j 在文档 D_i 中出现的频率，同时，LDA 将这个矩阵转换为两个低维度的矩阵， $M1$ 和 $M2$ 。

表 2-2 文档−主题矩阵

.	Z1	Z2	…	Zk
θ1	0	2	…	3
θ2	1	4	…	0
…	…	…	…	…
θn	1	1	…	0

上面显示了 $M1$ 矩阵的情况，它是一个 $N*K$ 大小的 document - topic 矩阵， N 指文档的数量， K 指主题的数量， $M1$ 中， θ_i 是一个长度为 K 的向量，用于描述当前文档 θ_i 在 K 个主题上的分布情况， Z 表示具体的主题。

表 3 词语-主题矩阵

.	W1	W2	…	Wm
Φ1	0	2	…	3
Φ2	1	4	…	0
…	…	…	…	…
Φn	1	1	…	0

上面显示了 $M2$ 矩阵的情况，它是一个 $K*V$ 维的 topic - term 矩阵，K 指主题的数量，V 指词汇表的大小。$M2$ 中每一行都是一个 φ 分布，也就是主题 φ_k 在 m 个词上的多项式分布情况，可以通过学习得到。

LDA 文档生成流程：LDA 假设文档是由多个主题的混合来产生的，每个文档的生成过程如下：从全局的泊松分布参数为 α 的分布中生成一个文档的长 N；从全局的狄利克雷参数为 *alpha* 的分布中生成一个当前文档的 θ。对当前文档长度 N 的每一个字都有从 θ 为参数的多项式分布生成一个主题的下标 z_n，从 θ 和 Z 共同为参数的多项式分布中，产生一个字 w_n。

这些主题基于词的概率分布来产生词，给定文档数据集，LDA 可以学习出，是哪些主题产生了这些文档。对于文档生成过程，则有，首先对于文档 n 中的每一个字，都先从文档矩阵 $M1$ 中的 θ_i 中产生一个下标，告诉我们现在要从主题矩阵 $M2$ 中的哪一行 ϕ_m 生成当前的字。

第 3 章 影评数据获取与预处理

3.1 数据的来源

从猫眼 APP 上爬取《上海堡垒》、《战狼 2》以及《红海行动》的影评数据。

3.2 获取数据的技术实现

我们使用谷歌浏览器在 PC 端打开网页发现评论只有区区十几条，这显然是不够的，因为手机 APP 上显示有十几万条评论。所以我们使用手机(m)端来获取数据。

通过 F12 召唤开发者工具，然后点击 切换设备工具栏（toggle device toolbar），如下图所示，然后刷新页面，即可切换成移动端的界面，评论数据也都可以正常显示出来了。

图 3-1 猫眼 APP 移动端界面

评论区的数据是通过 Ajax 动态加载出来的，也就是说，向下滑动到底之后，页面再向服务器发送请求，加载后面的评论数据，请求的 URL 结构如下所示，关键的参数是 offset 和 startTime。（经测试，两个参数均可以实现 “翻页” 的效果，但是并不需要两个都做改变，固定一个的值，然后循环改变另一个的值即可），由于此处我是希望爬取电影上映之后的评论数据(毕竟是看过电影之后再评论的更加可靠一些),所以改变 startTime 可能更加合适一些。

爬虫部分比较简单，都是一些常规操作，这里简单介绍一下流程，细节就不做过多解释了。本文爬虫包含了五个函数，如下：

get_data ：其参数是目标网页 url，这个函数可以模拟浏览器访问 url，获取并将网页的内容返回。

parse_data ：其参数是网页的内容，这个函数主要是用来解析网页内容，筛选提取出关键的信息，并打包成列表返回。

save_data ：其参数是数据的列表，这个函数用来将列表中的数据写入本地的文件中。

main ：这个函数是爬虫程序的调度器，可以根据事先分析好的 url 的规则，不断的构造新的请求 url，并调用其他三个函数，获取数据并保存到本地，直到结束。

if __name__ == '__main__' ：这是主程序的入口，在这里调用 main 函数，启动爬虫调度器即可。

除此之外我们需要注意以下几点：

1. 安装必要的 python 库，就是代码一开头 import 的那些，安装的方式也很简单，在 python 终端输入命令行：

pip install 库名

2. 修改请求 URL，请将函数 main 里的 url 中的数字改成你要爬取的电影的 ID（78304 是电影《战狼》的 id ）。

3. 修改 end_time 的值，请将这里的值改为你要爬取的电影的上映时间（或者你希望爬取截止的日期）

end_time = '2018-11-16 00:00:00' # 电影上映时间

4. 修改 save_data 中 ，请将这里的 filename 修改为你希望的存储路径及文件名。

filename = 'Data/comments.csv'

爬去的数据如下图：

comment
大陆电影，还有什么值得可以说的呢？
总之就是喜欢，国产片拍成这样很不错了。
好看！打的挺精彩的！！
在我的眼里，这是部不错的电影
和初恋一起看的可是如今，哎
觉得一般般的感觉
电影很好看，最近才看到还需要评价，很方便
少有的中国式武装英雄片
很好不错，很有震撼力。
国产的电影质量越来越好了
给朋友买的，都过去很久了才来评价，不好意思啊！朋友跟他好基友一起去看的，咱可不去看这一类硬汉类的电影。哈哈电影还可以
不错！挺好看的！个人还是蛮喜欢的
犯我中华者，虽远必诛！军人！血性！
很好看的一部，激动又搞笑，赞一个，非常期待再出续集 ㄣ(ˆ0ˆ)ㄥ
特别喜欢他的这种特战片。演的淋漓尽致的。看的我这个开心啊 挺好 挺好
好看，展现了中国军人的血性和本色
好看，真的好看，虽远必诛
挺不错的，一直很期待没有失望
还不错，跟想象中的差不多，推荐大家一起来看哦
一般吧，没有想想的好看
可以哈，就是感觉太短了，意犹未尽啊。
好看，不过去的时候已经完成一半了
超喜欢，去电影院看了两边！！！
整体来说，还可以
我觉得超级不错，具体讲什么其实我也忘记了，反正都不错吧
十点多了，我的人家都市全部的话。我
电影一般般，不是太好看
不错，孩子选的，很有意义

图 3-2 爬取的影评

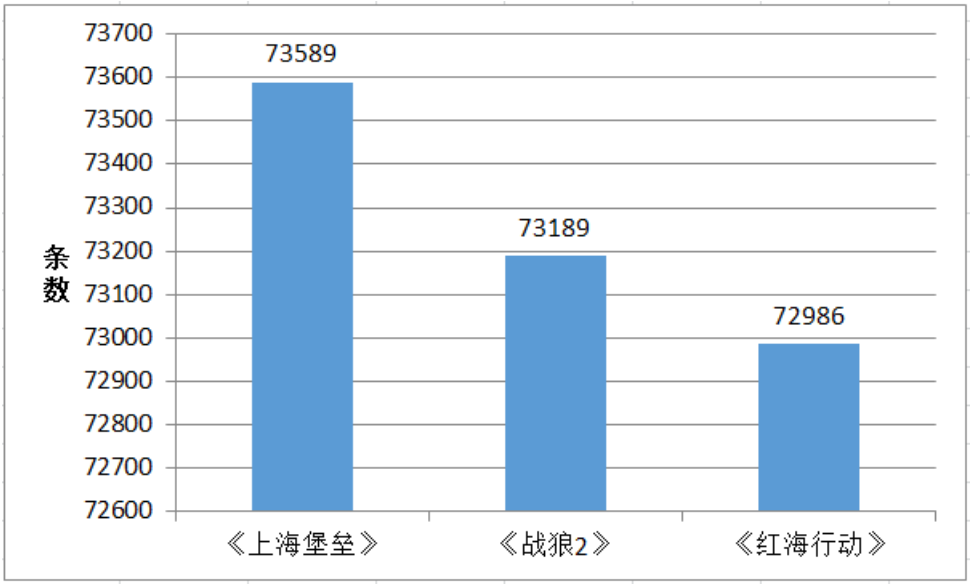


图 3-3 初始影评数量

3.3 数据的预处理

在爬取到影评之后，首先要将影评进行数据清洗：影评中掺杂着些毫无作用的内容，要将它们从影评集合里清除。

3.3.1 评论去重

第一步，删除影评中重复的数据。重复的数据包括：（1）评论没有被观众按时填写时，后台自动评论。（2）观众直接复制他人评论。本文采用删除重复影评的方法处理上述情况。

3.3.2 短文本删除

在影评中，一些影评字数只有区区一、两字，例如：可以，还行等。尽管体现了观众对于电影的情感，然而却挖掘不出更有价值的信息。本文采取规定最低影评字数的方法处理这种情况，将字数低于 6 的影评去除。

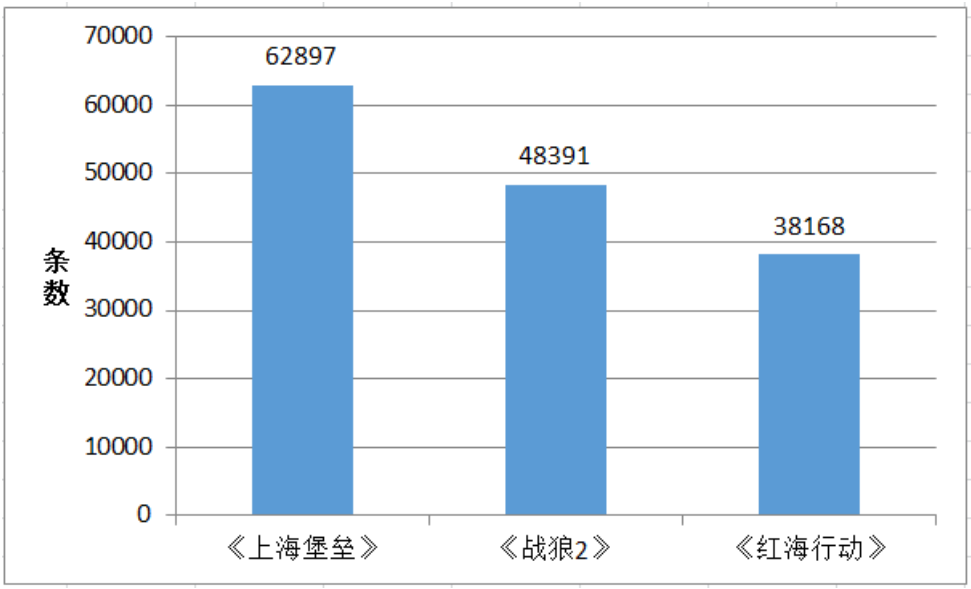


图 3-4 去重及短文本删除后的剩余影评

3.3.3 删除数字和英文

在影评中通常夹杂着英文字母和数字，这类符号对本研究毫无价值，所以本文将去除影评中的 26 个英文字符以及 0-9 的数字。

3.3.4 中文分词

中文分词是指将一句话切分成单独的几个词，每个词之间都由一个空格隔开。现有的分词方法可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。

主要的统计模型有：N 元文法模型（N-gram），隐马尔可夫模型（Hidden Markov Model ，HMM），最大熵模型（ME），条件随机场模型（Conditional Random Fields，CRF）[35]。

结巴(jieba)分词是目前国内最流行的中文分词工具。结巴分词支持三种模式：

①精准模式：能够将语句最精确地分开，适合文本分析；

②全模式：把语句中皆能够成词的词语都扫描出来，非常迅速，然而不能解决歧义问题；

③搜索引擎模式：在精准模式的基础上，再次将长词切分，提高召回率，适用于搜索引擎分词[35]。

除此之外还有 THULAC、NLPIR 等一系列分词工具，在此不做赘述。

本研究使用 jieba 的精准模式对影评数据进行分词处理。并且为了提升分词的准确率，需要在分词前导入自定义词典，词典中包含了电影中的特有名词，还要添加 NTUSD 情感词典。

3.3.5 去除停用词

影评中通常出现一些使用次数很高但却毫无含义的标点符号、词语以及表情符号，例如：好、嗯、挺不错的等，我们称呼这些毫无价值的部分为停用词。在进行分析研究之前可以把此类词语删除来提升效率。

本文决定使用将哈工大停用词处理词库、百度停用词等诸多停用词表进行合并去重操作，设计出的含近 1800 个停用词的词典，并在其中添加二十

多个表情符号，最后将影评数据中的这类词删掉。整理的相关停用词表如下：

表 3-1 部分停用词表

停用词表
一下 一个 一些 一何 一切 一则 一则通过 一天 一定 万一 三天两头 三番两次 三番五次 上 阿 哎 哎呀 哎哟 唉 俺 啊 俺们 按 按照 吧 吧哒 把 罢了 末##末 被 本 本着 比 比方 比如 人民 ,?、。 “” 《》！，：；？ ……

分词以及去停用词后的数据如下图：

1	观看 真的 赞 吴京 电影 满腔热血
2	扬国威 影片 完后 热血 澎涨 希望 吴京 拍出 好片 中国 点赞
3	太爽了 完美 吴京 OK
4	比战 狼 提前 上映 估计 吴京 事
5	喜欢 吴京演 高度 气魄
6	一部 棒 电影 吴京 电影 失望
7	看过 久 评价 两遍 第二次 亲友团 享受 猫眼 9.9 元 优惠价 猫眼 买 电影票 此片 大热 评价 褒贬不一 力挺 欣赏 电影
8	中国 军人 气势 喜欢 吴京
9	第三部 想 吴京 希望 上影
10	吴京 武术 一部 电影 电视剧
11	电影 表达 中国 军人 硬汉 形象 吴京 所说 中国 护照 世界 地方 中国 护照 世界 地方 回来 点赞 中国 骄傲 加油 祖国
12	喜欢 吴京 铁血男儿 电影 长大
13	相较 战狼 喜欢 战狼 喜欢 吴京 战狼 先播 先入为主
14	想 模仿 吴京 电影
15	红海 热血沸腾 一部 优秀的 爱国 影评 吴京 带给 强大的 祖国 中国 公民 在世界上 国家 受 保护 先进 装备 丢下 人质
16	一部 电影 吴京点 赞
17	激动人心 喜欢 吴京
18	太棒了 喜欢 吴京 这部 电影 淋淋 尽致 展现 扬 我国 威 四个 字
19	老公 喜欢 吴京 新意
20	太棒了 希望 吴京 新 作品
21	这部 赞 集体主义 吴京 单打独斗 更上 层楼
22	吴京 感人 震撼
23	影片 体现 爱国 情节 吴京
24	评论 真 不好 写 吴京能 拍出来 票子
25	震撼人心 演艺圈 更上一层楼 多一些 吴京 男子汉 少 陆晗 恶心
26	超级 喜欢 吴京 超级 喜欢 演出 气

图 3-5 分词及去停用词之后的影评

3.4 文本特征抽取

因为计算机只能处理结构化的数据，而文本属于非结构化数据，计算机不能直接处理，所以需要将文本转化为结构化数据才可以处理。文本特征抽

取是关键，文本特征抽取就是抽取出文本里的特征词并且量化的过程。经过特征抽取处理之后文本变成了计算机可以处理的结构化数据[36]。

文本特征抽取的方式多种多样，最主要的不同之处在于构建的估计函数不尽相同。文本特征抽取方式主要通过文本特征向量表示，进而基于统计的特征抽取的不同方法进行计算。

(1)文本向量表示

文本特征向量表示也就是向量空间模型，主要思想就是将影评用向量来表示。向量空间模型将影评数据表示为 $(t_1,t_2,\cdots,t_i)$ 和 $(w_1,w_2,\cdots,w_i)$ 构成，其中 $(t_1,t_2,\cdots,t_i)$ 是特征词表示，$(w_1,w_2,\cdots,w_i)$ 是特征全重表示，用向量 $d=(t_1,w_1,\cdots t_i,w_i)$ 表示影评数据。

(2)TF 方法

TF 方法即词频方法，根据该特征项在影评中出现的次数多少来决定该词语的权值，它比较容易操作，但也会遗漏一些出现次数少但作用大的。公式如下：

$$TF_{w,D_i}=\frac{count(w)}{|D_i|} \tag{3.1}$$

其中，$count(w)$ 为关键词 w 的出现次数，$|D_i|$ 为文档 D_i 中所有词的数量。

(3)IDF 方法

IDF 方法即反文档频率方法，概括来讲，特征项普遍程度的表达是它的主要目的，如果许多评论中出现同一个特征项，则它的值应该很低；而反过来如果一个特征项出现在比较少的评论中，则它的值应该很高；还有一种特殊的情形，若所有评论中都出现过这个特征项，那么它的值为 0。

一个词语的 IDF 可通过评论集中的评论总数除以包含该词的文件数，再将结果取对数得到。公式如下：

$$IDF_w=\log\frac{N}{\sum_{i=1}^{N}I(w,D_i)} \tag{3.2}$$

其中，N 表示评论总数，$I(w,D_i)$ 表示评论 D_i 是否包含关键词，如果有则是 1，没有则是 0。倘若特征词 w 在所有评论中均没有出现，则 IDF 公式中的分母是 0；所以需要对 IDF 做平滑（smooth）：

$$IDF_w=\log\frac{N}{1+\sum_{i=1}^{N}I(w,D_i)} \tag{3.3}$$

(4)TF-IDF 方法

TF-IDF 即词频-逆文本频率，它综合考虑了特征词的词频与反文档频率，是最主要的一种方法，计算公式如下：

$$TF\text{-}IDF_w = \frac{count(w)}{|D_i|} \times \log \frac{N}{1+\sum_{i=1}^{N} I(w, D_i)} \tag{3.4}$$

其中，第一个式子是 TF 的计算公式，第二个则是 IDF 的计算公式。

3.5 本章小结

本章内容首先详细讲述了如何使用爬虫技术爬取影评数据，并对所爬取的数据进行了一个展示，其次讲解了文本数据的预处理，包括文本去重、短句删除、分词以及去停用词，还展示了分词以及去停用词之后的数据，并且概括讲解特征词抽取的概念以及特征抽取的几种方法(TF、IDF、TF-IDF 算法)，使人们能够一目了然，对自然语言处理有了一个初步的认识。

第 4 章 情感分类以及算法改进

4.1 影评情感分类

4.1.1 情感分类相关理论

情感分类主要是对含有主观倾向的词语的判断来确定影评中包含观众的观点及情感倾向。通过研究影评中观众的情感方向，可以让电影制作方以及投资方把握到观众对于电影的看法，从而加以改进影片制作。现如今情感分类主要有两种方法：第一种是以情感词典为基础进行情感分类；第二章是基于机器学习的方法来进行情感分类。第一种是依据设置好的词典来抽取评论中的情感词语，然后得到该评论所表达的情感倾向。因此，情感词典是否设立完善将影响其最终的分类效果。

而第二种方法将作为特征词的情感词转化为矩阵后，再利用例如 SVM 等分类模型，对评论进行分类。所以这种方法的最终分类效果主要依赖于训练集合的筛选以及是否正确标注情感。

对于电影评论数据，本研究采用基于机器学习的方法进行情感分类，以猫眼 APP 观众评论举个例子：

“借着观后情绪发表一下评论，实在是太精彩了，通过电影的形式展示了中国的各种强大，各种武器装备，特种兵的强大以前只能看到美国电影里有，更加融入了中国式的军人团队情感，太感人了热泪盈眶，在不久的将来中国会更加强大的这是必然趋势”

从评论中可以看到“精彩”、“强大”、“感人”、“热泪盈眶”等正面词语，由此判断此句为正面评论，将其标记为 1。

“一般般就那回事了不怎么好～～”

从评论中可以看到“一般般”、“不怎么好”等描述，可以判断此句为负面评论，将其标记为 0。

4.1.2 FV-SA-SVM 分类算法

一般的基于机器学习的分类算法包括朴素贝叶斯、SVM、KNN 以及逻辑回归等，在第一章已经详细介绍了，在此不做赘述。通过大量实验表明，支持向量机在文本分类中的泛化能力强劲，相比于其他分类器更适合进行影评的情感分类。

但是支持向量机的分类性能却受到了惩罚因子 C 与核函数参数 sigma 的影响。那么如何选择合适的参数来使支持向量机性能更强就属于一个寻找最优参数的过程。近年来，有许多国内外学者提出过很多优化支持向量机参数的方法，例如：基于蚁群优化算法选取参数，基于粒子群算法的最优参数寻找等。

但是以上优化方法有其各自的缺点，ACO 算法的收敛速度较慢易陷入局部最优，PSO 算法易早熟收敛且局部寻优能力较差。所以本实验使用一种通过模拟退火算法(SA)寻找 SVM 最优参数的方法，因为模拟退火算法可以较强的跳出局部最优值，提高全局寻优的能力。

以下图为例，假设初始解为左边蓝色点 A，SA 会快速搜寻到局部最优解 B，但在搜索到 B 点后，不是就此结束，而是会以一定的概率向右边移动。或许经过几次这样的移动后会到达全局最优点 D，于是就跳出了局部最小值[37]。

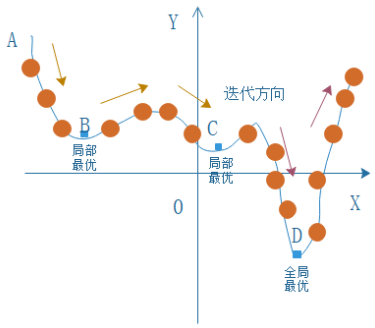


图 4-1 模拟退火示例

根据热力学原理，在温度为 T 时，出现能量差为 dE 的降温的概率为 p(dE)，表示为:

$$P(dE)=\exp(\frac{dE}{KT}) \tag{4.1}$$

其中 k 是波尔兹曼常数，k=1.3806488(13)×10^(−23)，dE<0。因此 dE/kT<0，则 p(dE)取值范围是(0,1)。满足概率密度函数的定义。其实这条公式更直观意思就是：温度越高，出现一次能量差为 p(dE)的降温的概率就越大；温度越低，则出现降温的概率就越小[37]。

在实际问题中，这里的“一定的概率”的计算参考了金属冶炼的退火过程。假定当前可行解为 x，迭代更新后的解为 x_new，那么对应的“能量差”定义为：

$$\Delta f = f(x\_new) - f(x) \tag{4.2}$$

其对应的“一定概率”为：

$$P(\Delta f)=\begin{cases}\exp(-\frac{\Delta f}{KT})\text{最小值优化问题}\\ \exp(\frac{\Delta f}{KT})\text{最大值优化问题}\end{cases} \tag{4.3}$$

模拟退火算法步骤如下图：

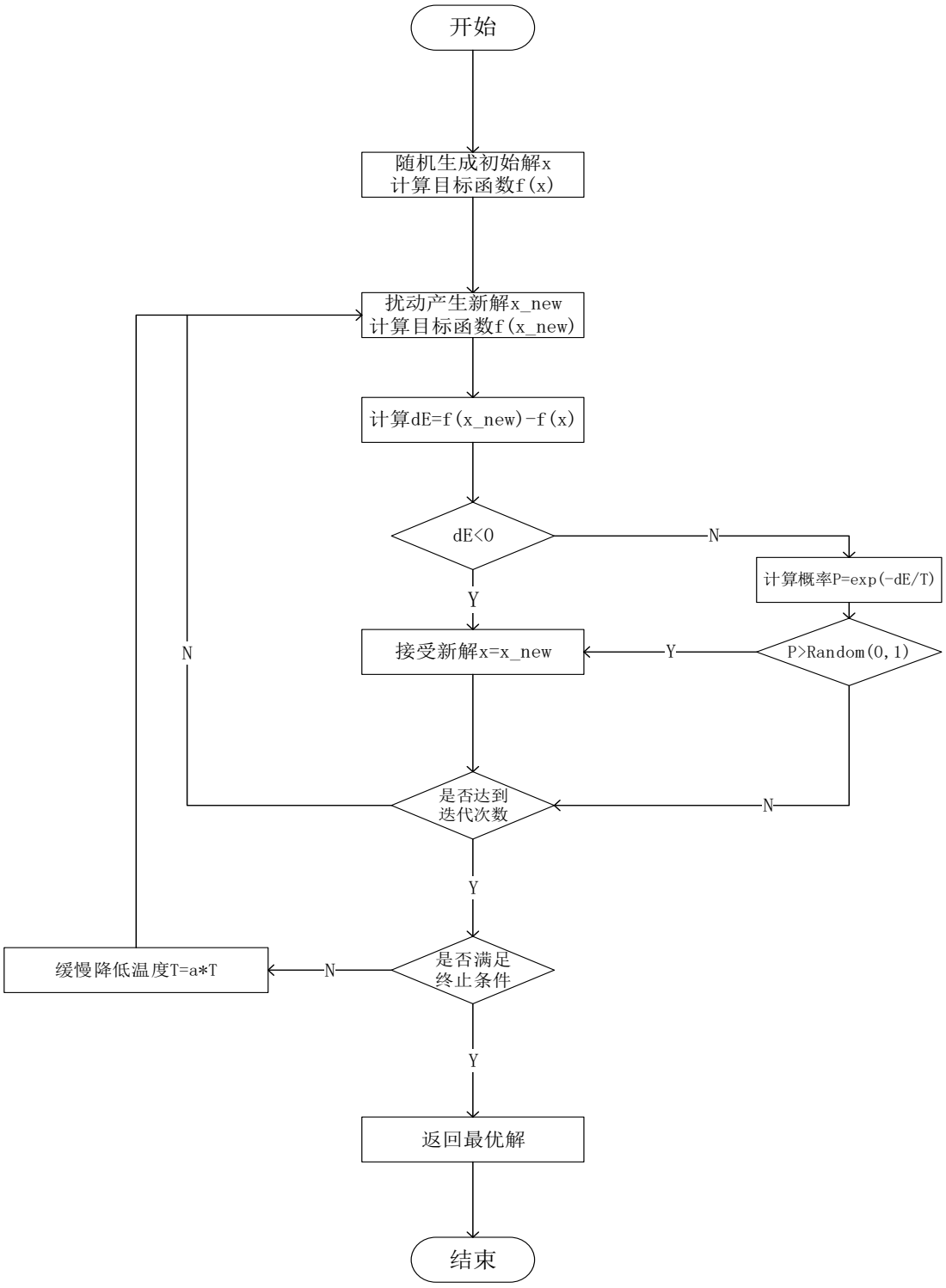


图 4-2 模拟退火算法步骤

在本文的研究中我们要使FV-SA-SVM分类性能最优即就是要寻找最优参数使得 SVM 的决策函数取得最大值。所以其本质是寻找全局最大值的过程。

4.1.3 FV-SA-SVM 实证

FV-SA-SVM 的原理以及步骤已经介绍过了，接下来就是使用 FV-SA-SVM 进行情感分类实验，实验硬件：操作系统为 Windows10 企业版；处理器为英特尔 Pentium(奔腾)G4560 @ 3.50GHz 双核；内存为镁光 DDR4 2400MHz 8GB 与英睿达 DDR4 2400MHz 8GB，一共 16GB；硬盘为日立 HGST HTS721010A9E630 1TB；软件平台：Python 3.6。

首先，我们要验证 FV-SA-SVM 模型对于不同类型电影的影评是否具有情感分类上的普适性、优越性，并且和传统的分类模型以及文献中的 SA-SVM 模型进行对比。本文从动作、喜剧、青春以及悬疑四个类型中选取八部具有代表性（高评分、影评多）的电影来进行研究。具体影片如下表：

表 4-1 所选不同类型的影片

动作	喜剧	青春	悬疑
《金刚：骷髅岛》	《唐人街探案 2》	《少年的你》	《犯罪现场》
《阿丽塔》	《一出好戏》	《夏洛特烦恼》	《窃听风云》

将爬取到的数据首先进行短文本删除以及文本去重操作之后再进行语料标注，标注时为了降低标注的主观性，我采取的办法是找四位同学一起标注，最后将有异议的影评进行讨论后再标注。最终标注情况如下表：

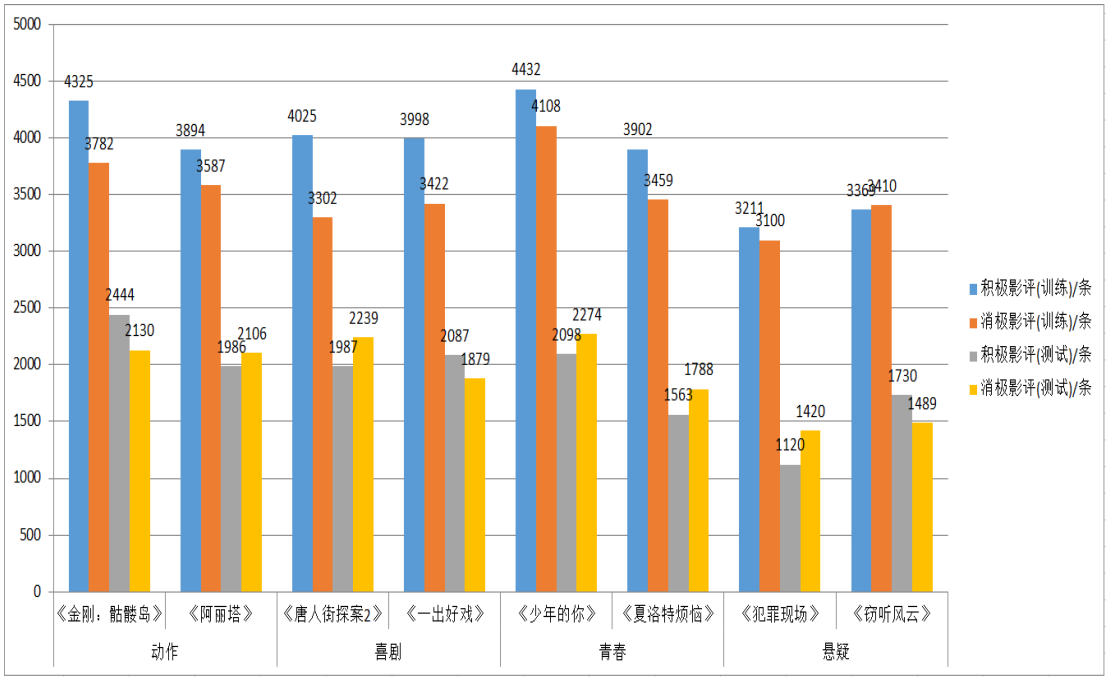


图 4-3 不同类型电影语料标注情况

本文中为了方便后续情感分类我们将同类型电影的影评数据进行汇总，结果如下图：

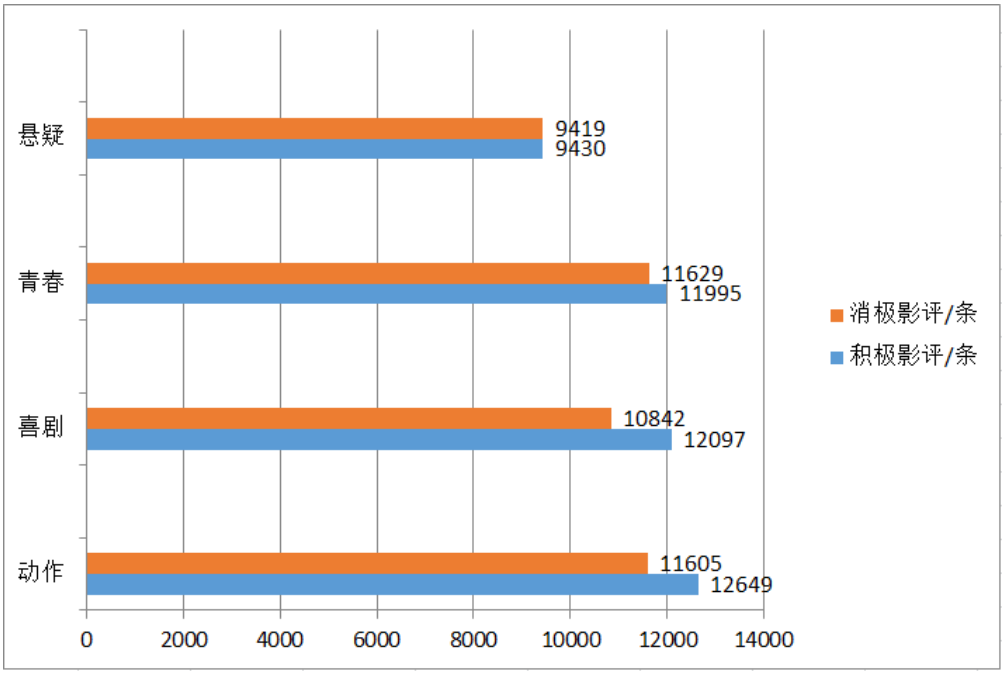


图 4-4 不同类型电影预料汇总

接下来使用 FV-SA-SVM 模型进行分类实验，实验结果所得准确率、精确率、召回率、F1-score 以及[C,σ]最优值如下图：

表 4-2 FV-SA-SVM 分类结果

	准确率(A)	精确率(P)	召回率(R)	F1-score	[C,σ]
动作	0.978	0.97	0.95	0.96	[109.14,0.95]
喜剧	0.953	0.95	0.95	0.95	[21.47,1.54]
青春	0.961	0.97	0.93	0.95	[11.89,0.58]
悬疑	0.974	0.96	0.95	0.95	[38.54,1.89]

然后使用文献中的 SA-SVM 模型以及 SVM 、Naive Bayes、Logistic 回归、KNN 等传统模型再次实验，并与 FV-SA-SVM 进行对比，所得结果如下图：

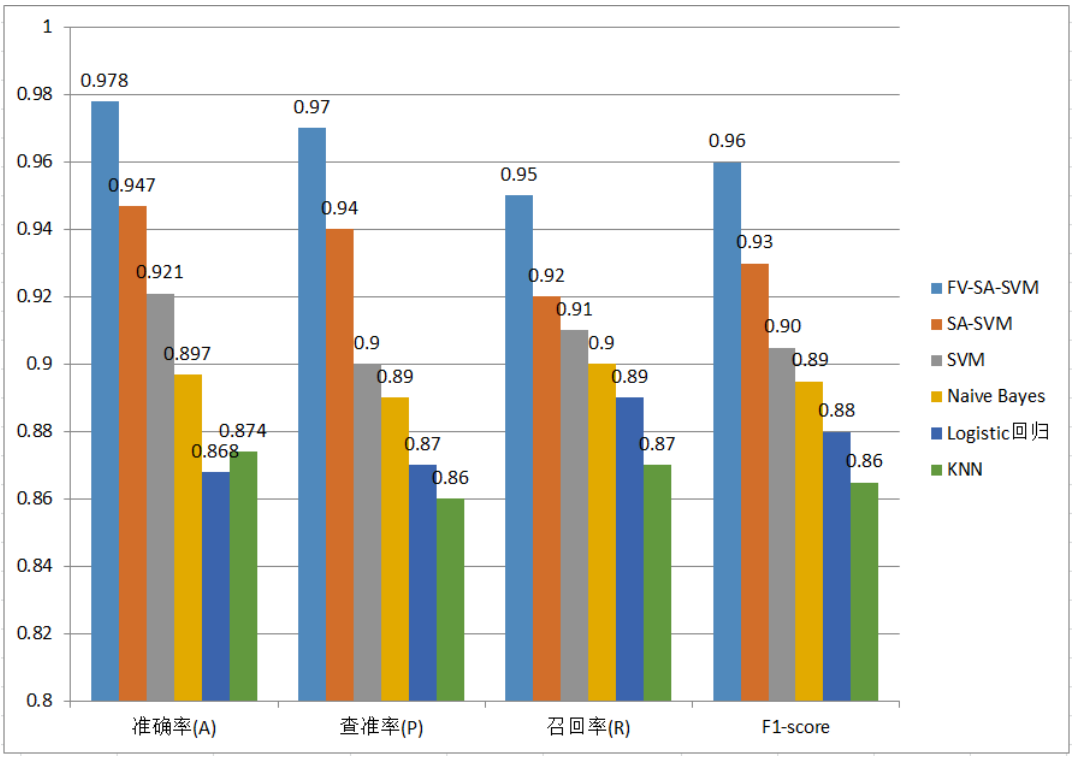


图 4-5 动作片分类结果比较

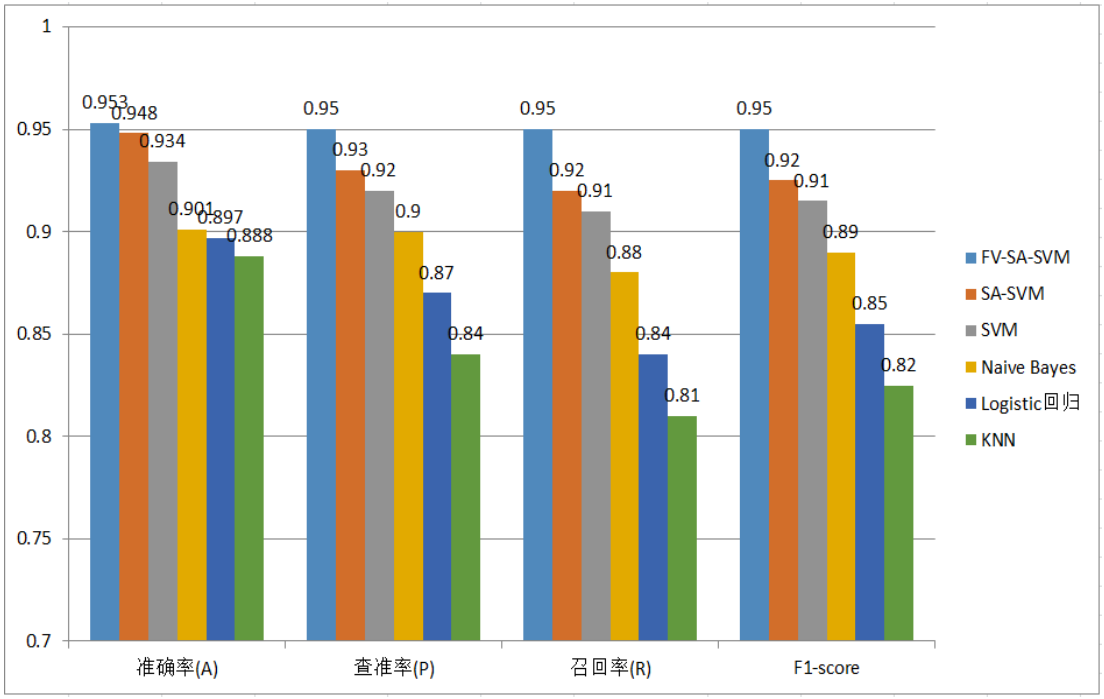


图 4-6 喜剧片分类结果比较

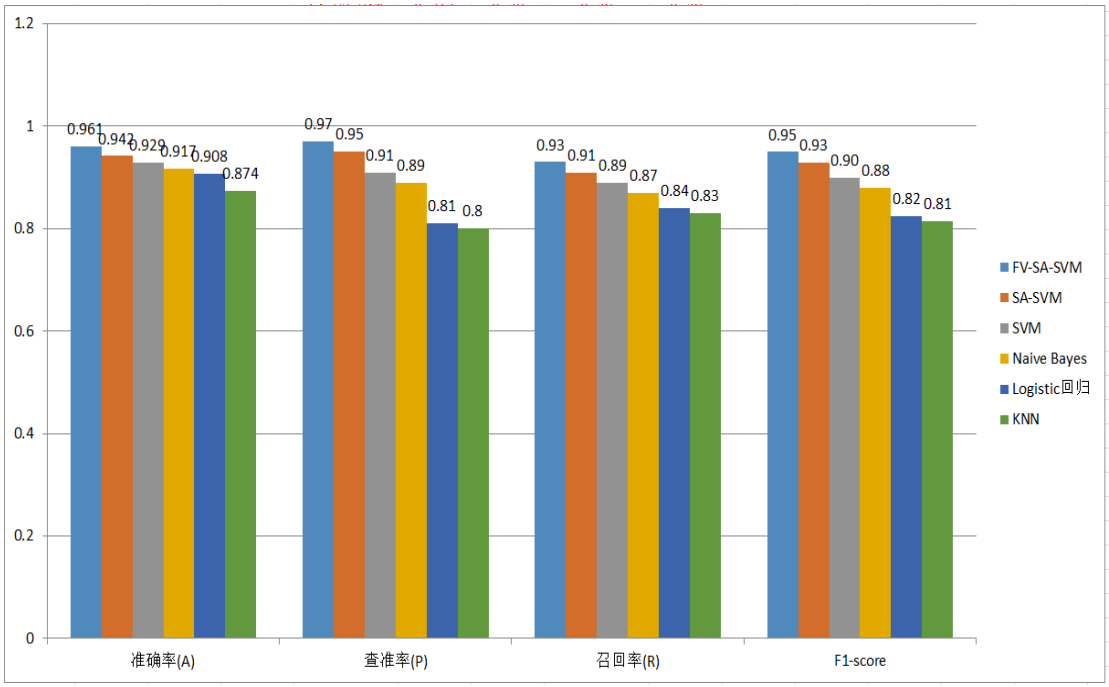


图 4-7 青春片分类结果比较

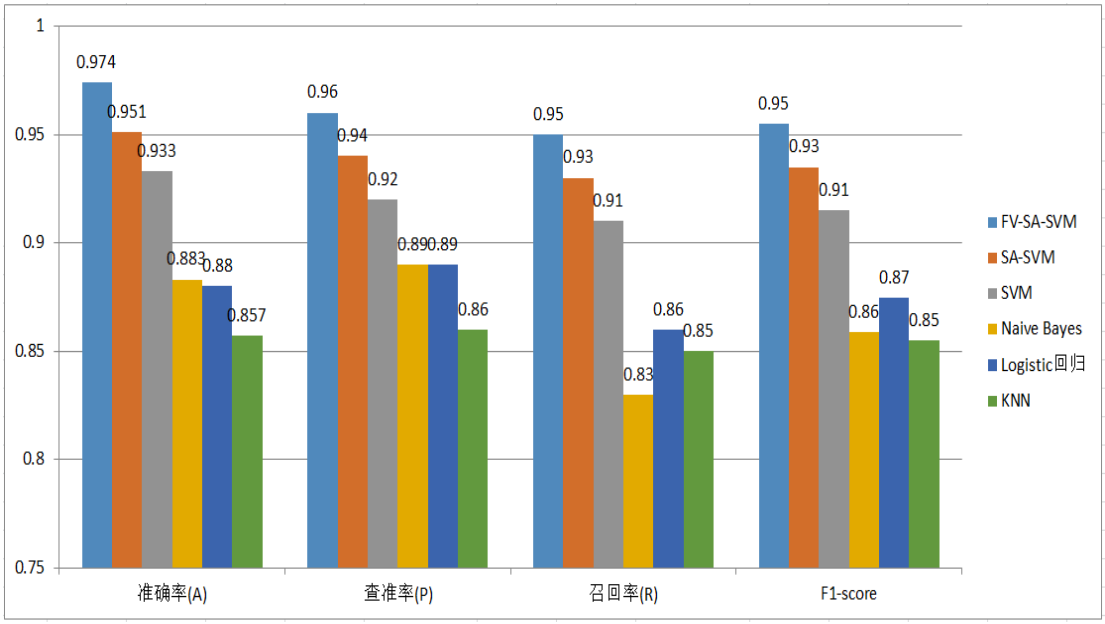


图 4-8 悬疑片分类结果比较

由图 4-5 至图 4-8 可知，综合比较不同算法的准确率、精确率、召回率以及 F1 值这四个指标，我们认为参考文献中的 SA-SVM 算法要比传统分类算法效果好，而本文所用的 FV-SA-SVM 分类模型则优于 SA-SVM 以及传统模型，这证明了 FV-SA-SVM 在不同类型电影的影评情感分析上具有很强的适用性、普适性以及优越性，为接下来的实验奠定了坚实的基础。

下一步，需要实验的语料为《上海堡垒》、《战狼 2》以及《红海行动》这三部电影的影评，首先选取部分影评进行人工标注，具体情况如下表：

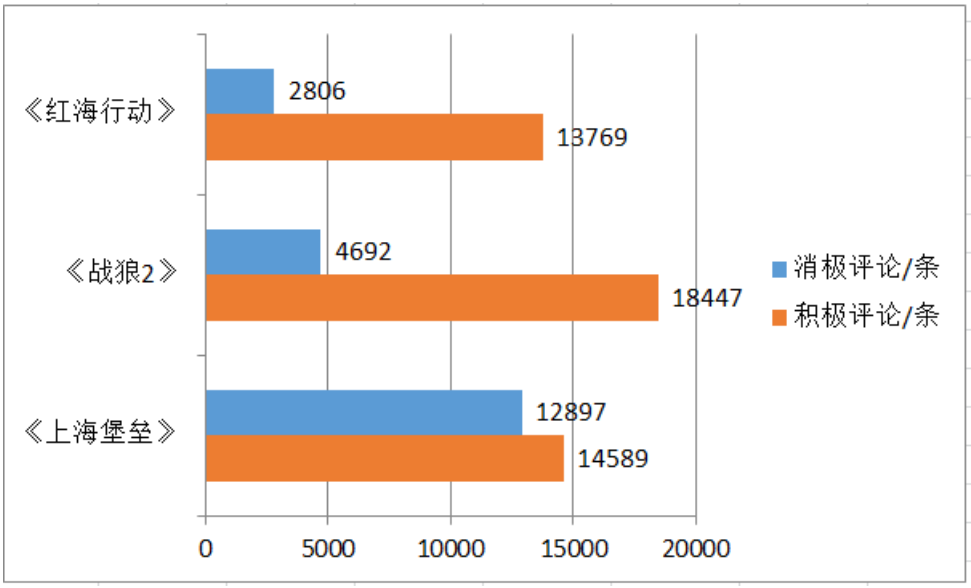


图 4-9 初始语料标注情况

经过 FV-SA-SVM 训练之后所得准确率、精确率、召回率、F1-score 以及[C,σ]最优值如下表：

表 4-3 FV-SA-SVM 分类结果

	准确率(A)	精确率(P)	召回率(R)	F1-score	[C,σ]
《上海堡垒》	0.947	0.93	0.95	0.94	[118.19,0.53]
《战狼 2》	0.934	0.91	0.94	0.92	[18.38,1.44]
《红海行动》	0.929	0.92	0.93	0.92	[17.51,0.16]

由表 4-3 可以看出 FV-SA-SVM 分类算法对于《上海堡垒》、《战狼 2》以及《红海行动》影评的准确率、精确率、召回率以及 F1 值皆达到了 0.90 以上，说明此方法切实有用、高效。

而后分别使用 Naive Bayes、Logistic 回归、KNN、SVM、SA-SVM 分类算法进行分类，并与 FV-SA-SVM 分类效果进行比较，结果如下表：

表 4-4 《上海堡垒》分类结果比较

	准确率(A)	精确率(P)	召回率(R)	F1-score
FV-SA-SVM	0.947	0.93	0.95	0.94
SA-SVM	0.928	0.92	0.93	0.92
SVM	0.888	0.87	0.90	0.88
Naive Bayes	0.876	0.84	0.84	0.81
Logistic 回归	0.838	0.82	0.81	0.81
KNN	0.825	0.80	0.82	0.81

由表 4-4《上海堡垒》分类结果比较可以看出各个分类算法的准确率、精确率、召回率以及 F1-score 皆在 0.80 以上，参考文献中的 SA-SVM 要比传统分类算法效果好,而本文中 FV-SA-SVM 的四项指标皆达到了 0.9 之上，实验证明 FV-SA-SVM 要比参考文献中 SA-SVM 以及 SVM、Naive Bayes、KNN 等传统算法分类性能要好，能最有效地将未分类数据进行分类。

表 4-5 《战狼 2》分类结果比较

	准确率(A)	精确率(P)	召回率(R)	F1-score
FV-SA-SVM	0.934	0.91	0.94	0.92
SA-SVM	0.920	0.92	0.91	0.91
SVM	0.889	0.89	0.87	0.88
Naive Bayes	0.874	0.88	0.89	0.88
Logistic 回归	0.859	0.90	0.87	0.88
KNN	0.843	0.86	0.83	0.85

由表 4-5《战狼 2》分类结果比较可以看出各个分类算法的准确率、精确率、召回率以及 F1-score 皆在 0.83 以上，参考文献中的 SA-SVM 要比传统分类算法效果好，而本文 FV-SA-SVM 的四项指标皆达到了 0.9 之上，实验证明 FV-SA-SVM 要比参考文献中 SA-SVM 以及 SVM、Naive Bayes、KNN 等传统算法分类性能要好，能最有效化地将未分类数据进行分类。

表 4-6 《红海行动》分类结果比较

	准确率(A)	精确率(P)	召回率(R)	F1-score
FV-SA-SVM	0.929	0.92	0.93	0.92
SA-SVM	0.894	0.89	0.90	0.89
SVM	0.882	0.87	0.89	0.88
Naive Bayes	0.866	0.88	0.87	0.87
Logistic 回归	0.869	0.87	0.85	0.86
KNN	0.870	0.86	0.87	0.85

由表 4-6《红海行动》分类结果比较可以看出各个分类算法的准确率、精确率、召回率以及 F1-score 皆在 0.85 以上，参考文献中的 SA-SVM 要比传统分类算法效果好，而本文 FV-SA-SVM 的四项指标皆达到了 0.9 之上，实验证明 FV-SA-SVM 要比参考文献中 SA-SVM 以及 SVM、Naive Bayes、KNN 等传统算法分类性能要好，能最有效化地将未分类数据进行分类。

根据以上结果我们可以发现 FV-SA-SVM 相比于 SA-SVM、SVM、Naive Bayes、Logistic 回归以及 KNN，它的分类性能无疑是最好的，所以接下来我们使用 FV-SA-SVM 分类算法对待分类影评数据进行初次分类，结果如下图：

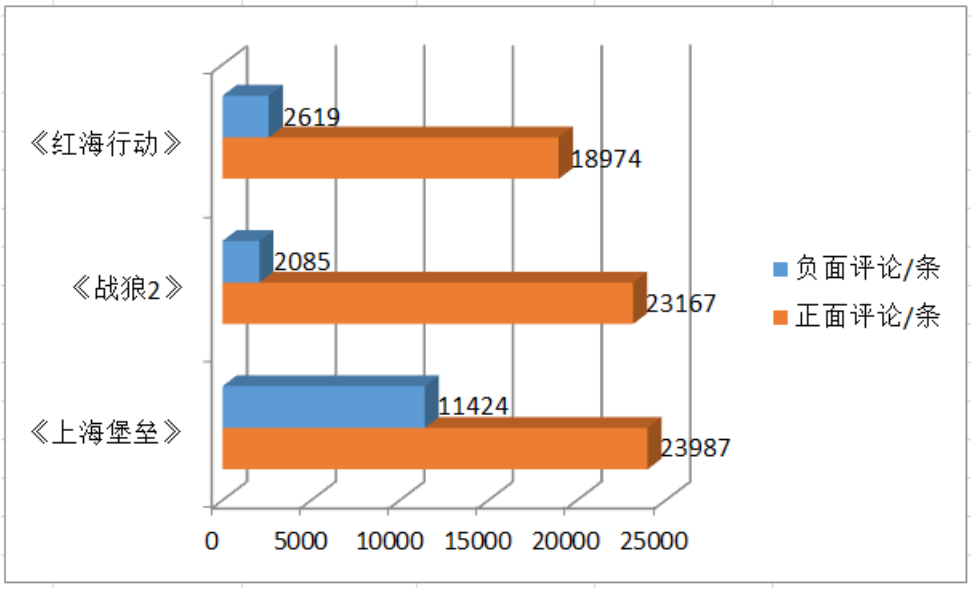


图 4-10 初始分类结果

然后我们再对初始分类的影评通过情感词典进行打分，筛选出错分的数据进行重新划分，并且为了更深入的了解这三部电影不同情感的分布情况，我们根据打分情况认为积极影评中得分属于(0,10]之间的是一般程度，(10,20]是中等程度，大于 20 属于高等。则最终分类结果如下图：

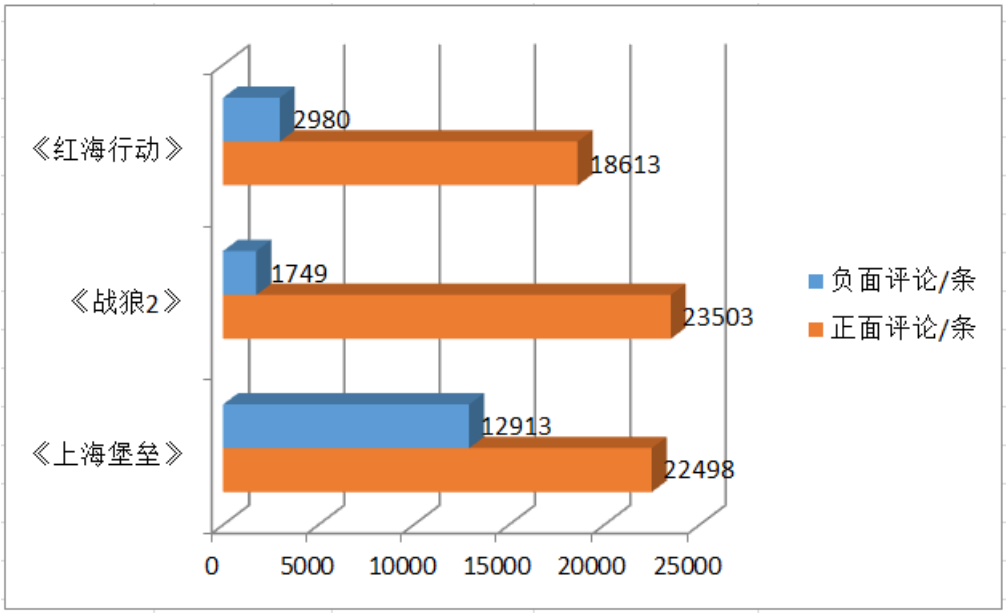


图 4-11 最终分类结果

按照评分情况我们研究三部电影情感倾向的分布情况，得出下图：

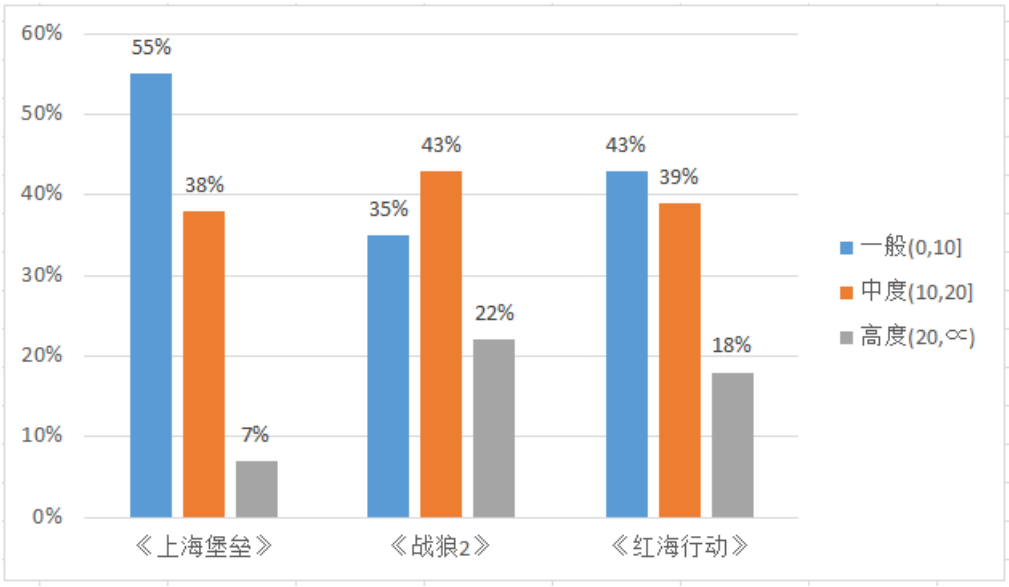


图 4-12 三部电影积极情绪分段统计图

由图 4-12 可以看出三部电影的积极评价中，一般积极情绪和中度积极情绪评价占比最多，尤其是《上海堡垒》，它的一般积极情绪和中度积极情绪评价占比总和达到了 93%。相比较而言《战狼 2》的高度积极情绪所占比例是三部电影之最，达到了 22%，其次是《红海行动》的 18%，而《上海堡垒》的高度积极情绪占比是最低的，只有 7%。以上数据说明观众对于《战狼 2》、《红海行动》的观影体验相比于《上海堡垒》还是很好的。

相反，消极影评中情感打分位于[-10,0)之间属于一般程度，位于[-20,-10)之间的属于中等程度，而位于-20 以下的属于高等程度。按照情感打分结果研究三部电影情感倾向的分段情况。

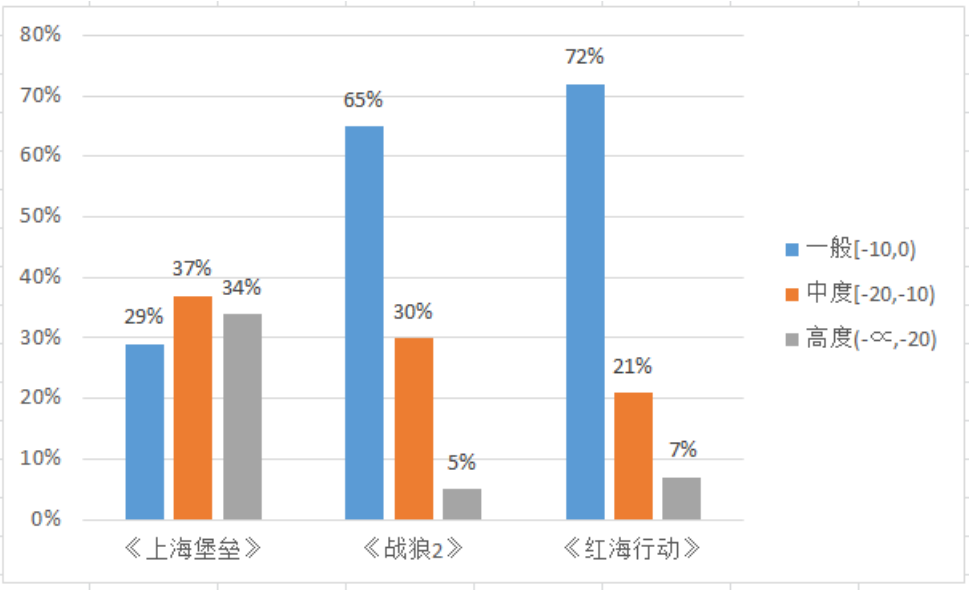


图 4-13 三部电影消极情绪分段统计图

由图 4-13 可知，从这三部电影的消极影评情感分段情况来看，《战狼 2》与《红海行动》中的一般消极情绪与中等消极情绪所占比重更大，而高度消极情绪所占比重较少，但是《上海堡垒》中高度消极情绪所占比重是最大的，达到了 34%，远超其余两部电影的 5%和 7%，说明观众对于《上海堡垒》的观影体验是非常不满意的。

目前已经将所有影评数据完成了情感倾向分析，将三部电影所有数据进行汇总，得出每部电影的情感分布如下表：

表 4-7 好评率与差评率

	《上海堡垒》	《战狼 2》	《红海行动》
积极情绪	0.59	0.87	0.85
消极情绪	0.41	0.13	0.15

为了更好的观察我们做下图：

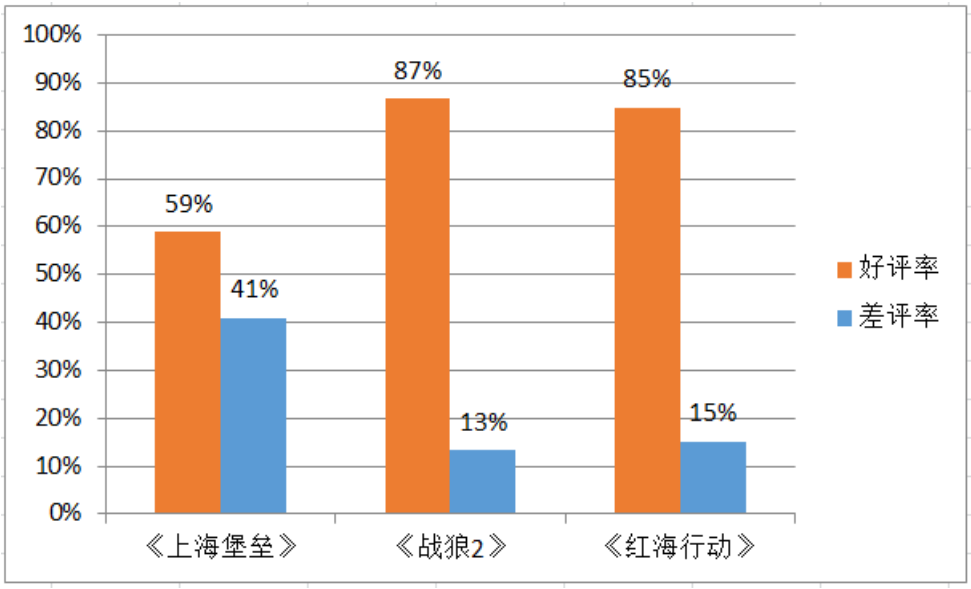


图 4-14 好评率与差评率

由图 4-14 可以看出三部电影不同情感的分布比例，其中《战狼 2》与《红海行动》的好评率都在 80%以上，说明观众对于这两部电影的观影体验综合来说还是不错的，尤其是《战狼 2》，观众观影体验最好。而《上海堡垒》好评率只有区区 59%，说明观众对于它的观影体验是糟糕的，这部电影整体来说是失败的，它 1.2 亿的总票房恰恰也证明了这一点。

4.2 基于语义网络的影评分析

我们使用 ROSTCM6 来绘制语义网络，步骤包括分词、提取高频词、过滤无用词、提取行特征词表、提取共现矩阵词、绘制语义网络。

对《上海堡垒》的积极和消极影评进行语义网络分析，可视化图像如下：

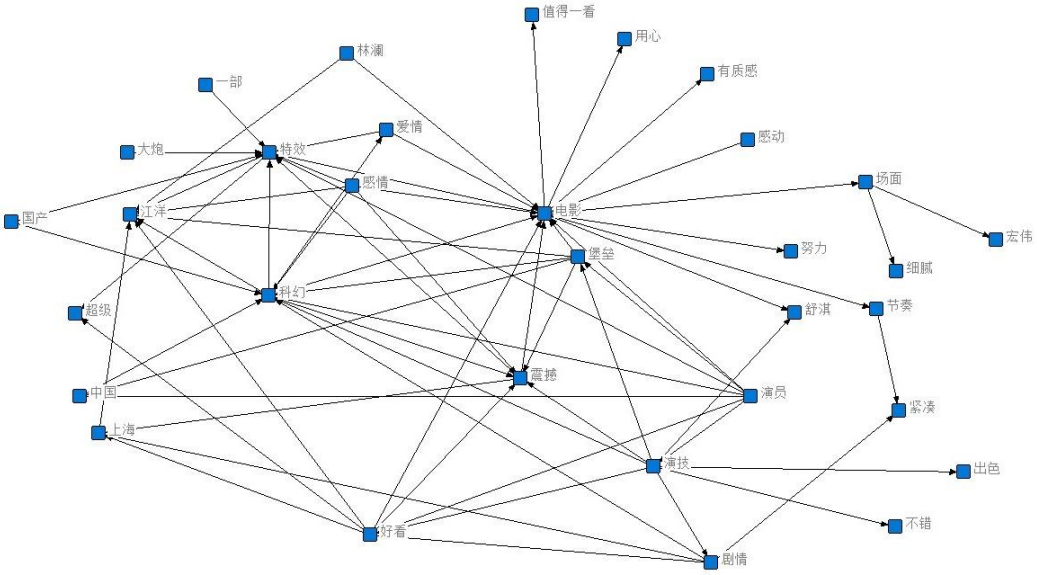


图 4-15 《上海堡垒》积极影评语义网络图

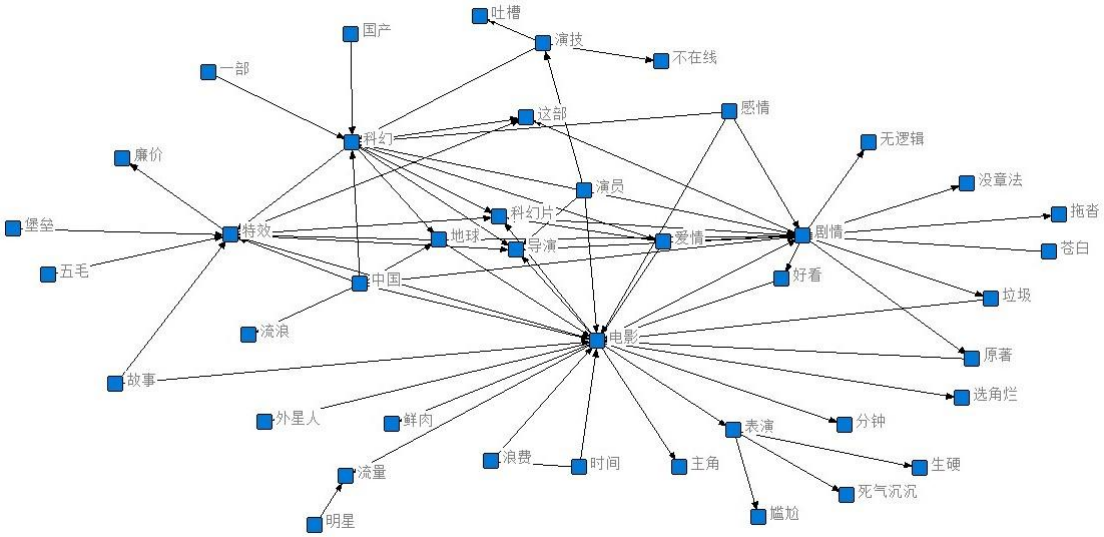


图 4-16 《上海堡垒》消极影评语义网络图

从语义网络图中可以看出《上海堡垒》的积极评论内容有：电影有质感、用心，场面细腻，特效逼真、震撼，节奏紧凑，演员演技出色。

《上海堡垒》消极评论内容有：剧情无逻辑、没章法，表演生硬尴尬、

死气沉沉，电影选角烂，节奏拖沓。

对《战狼 2》的积极和消极影评进行语义网络分析，可视化图像如下：

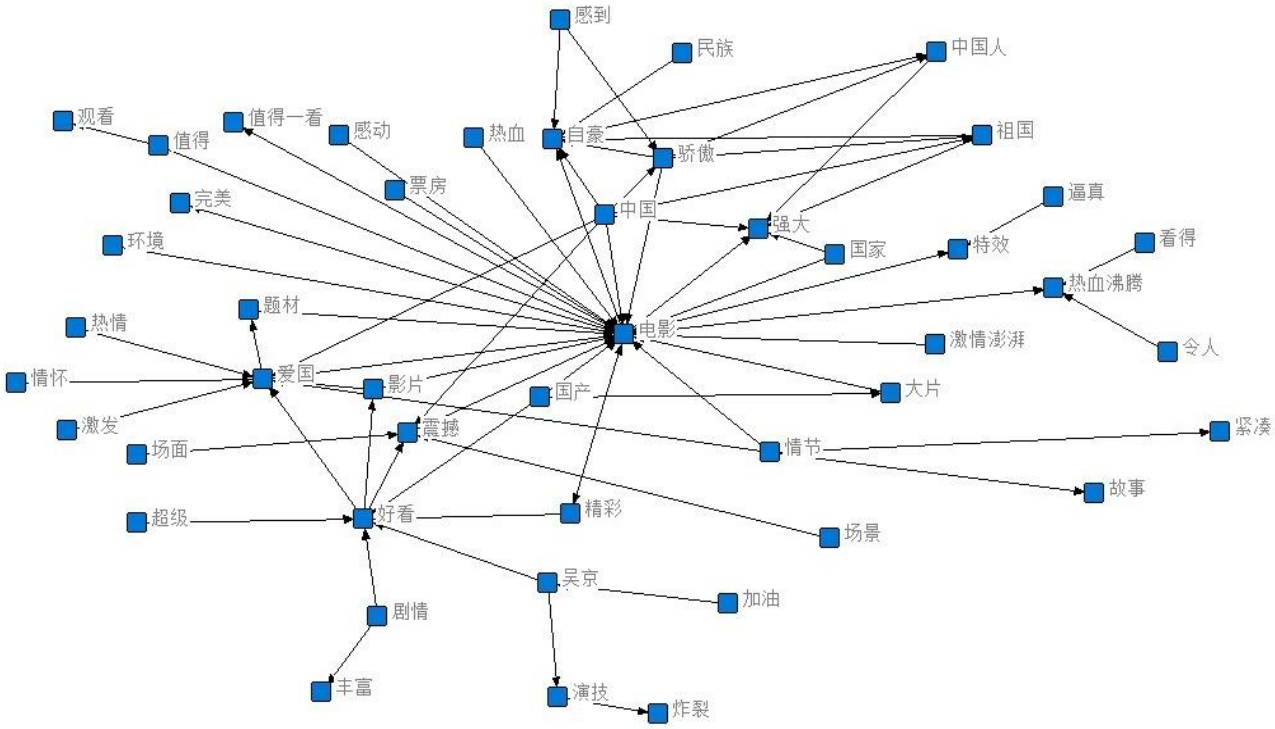


图 4-17 《战狼 2》积极影评语义网络图

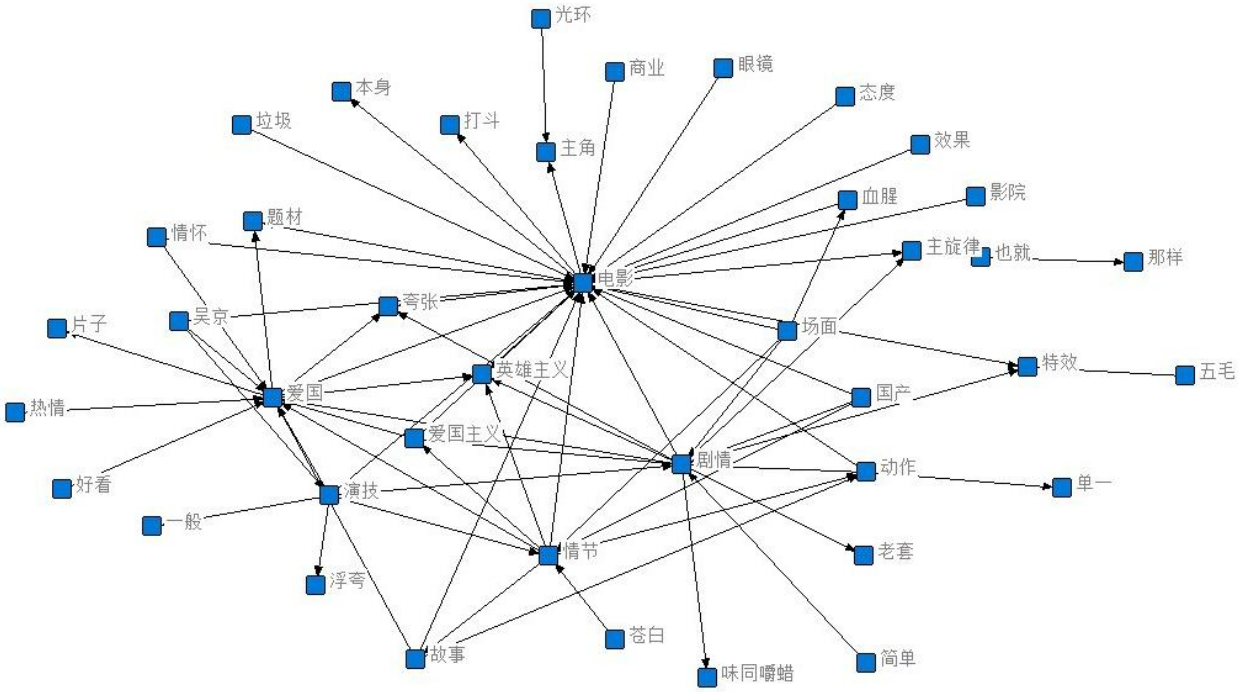


图 4-18 《战狼 2》消极影评语义网络图

从语义网络图中可以看出《战狼 2》的积极评论内容有：电影精彩热血，场景震撼、特效逼真，国家强大令人骄傲自豪、热血澎湃，剧情丰富、紧凑，吴京演技炸裂。

《战狼 2》的消极评论内容包括：电影五毛特效、画面血腥、演技浮夸、动作单一、剧情老套、味同嚼蜡、情节苍白。

对《红海行动》的积极和消极影评进行语义网络分析，可视化图像如下：

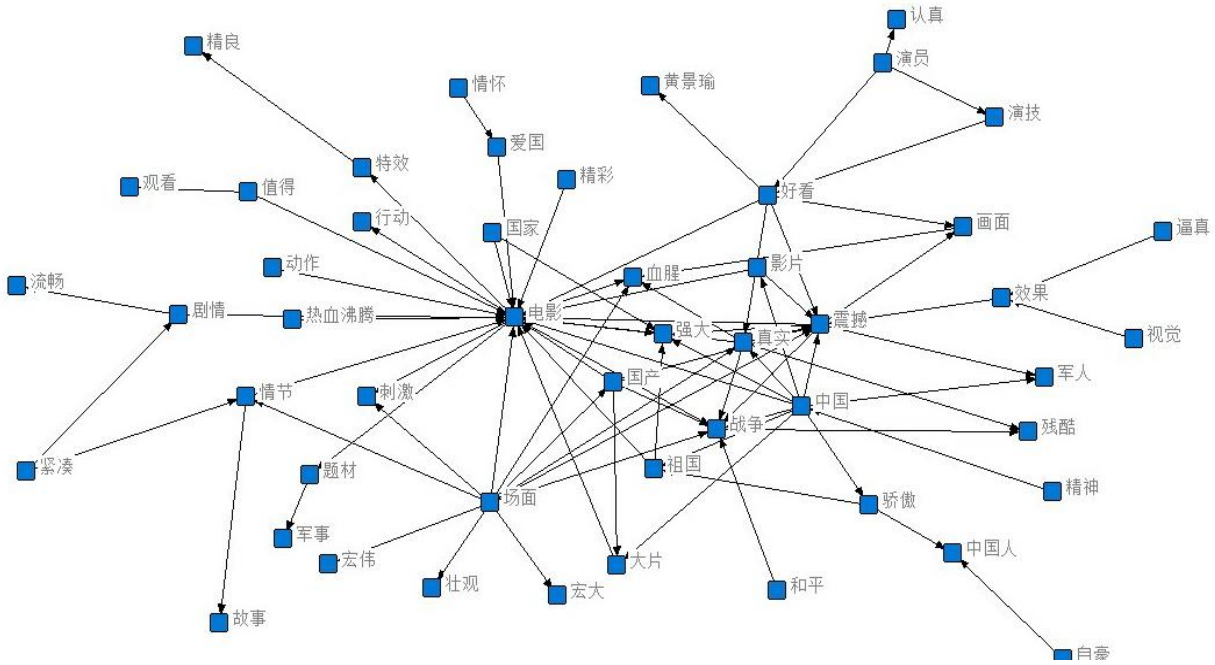


图 4-19 《红海行动》积极评论语义网络图

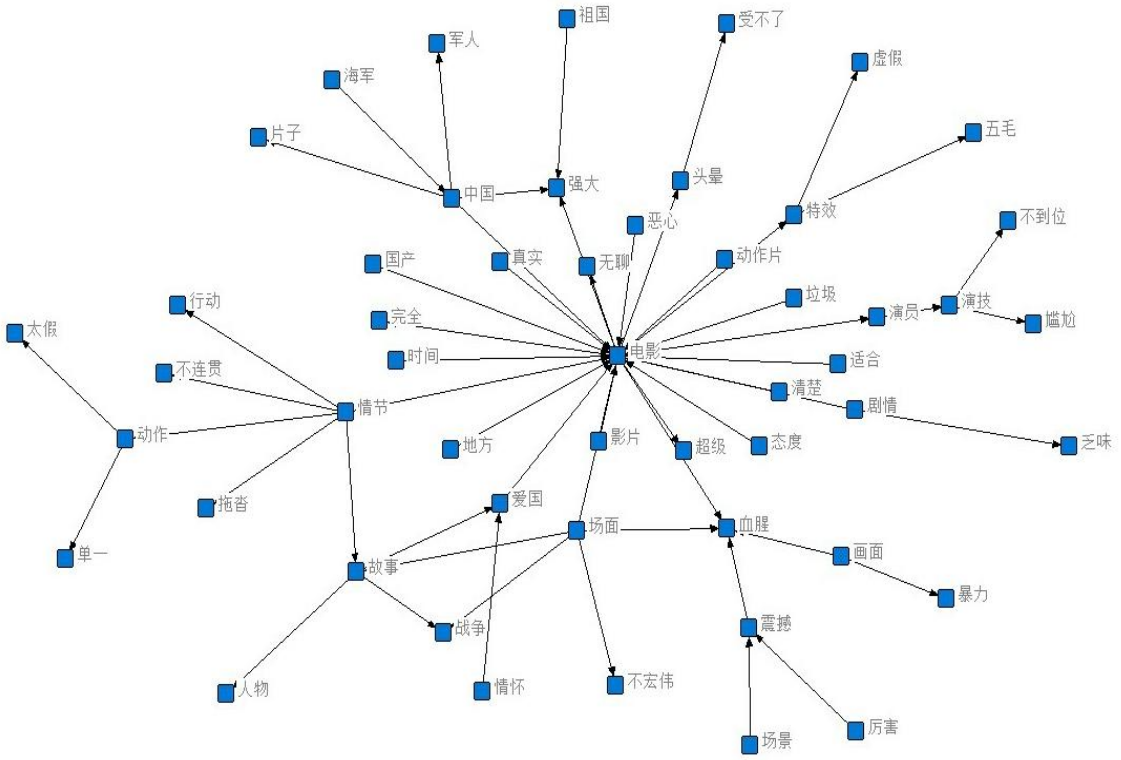


图 4-20 《红海行动》消极评论语义网络图

从语义网络图中可以看出《红海行动》的积极评论内容有：场面壮观、宏伟、刺激，情节紧凑、剧情流畅，令人热血沸腾，黄景瑜导演的好，特效精良，视觉效果逼真，演员认真且演技好。

《红海行动》的消极评论内容包括：电影无聊，五毛特效、虚假，动作单一，场面不宏伟，情节拖沓、不连贯，画面场景血腥暴力，剧情乏味，演员演技不到位、尴尬。

综合比较上述语义网络分析图可以发现，三部电影积极评价包含：

(1)电影精彩，场面壮观、宏伟；

(2)特效震撼、精良，场面刺激

(3)剧情紧凑、情节连贯；

(4)演技出色；

三部电影消极评价包含：

(1)画面血腥暴力令人不适；

(2)剧情乏味、老套，节奏拖沓

(3)特效差；

(4)演技拙劣；

我们能够发现语义网络将所有节点连接在一起，可以看到不同概念、不同事物间的联系，但是语义网络的概念很多，尽管可以看到很多概念间的联系，但是为了更好地提取主题，我们需要进行 LDA 主题挖掘以及聚类分析来发现电影的优点以及不足，从而提出改善方法。

第 5 章 基于 LDA 主题分析与聚类分析

5.1 影评数据的 LDA 主题分析

对这三部电影进行情感分析时，需要去挖掘影评中的潜在主题，为的是研究电影的评论关注点。主题是每条影评的中心主旨，如果某个潜在主题同时是多条影评中的主题，则这个潜在主题很有可能是整个影评中的核心关注点。我认为影评中的特征词项是 LDA 模型中的可观测变量，在特征词中、潜在主题中越是高频词就越来越可能成为核心关注点中的评论词。

因为每个影评都可能有多个主题，而且我们篇幅有限，所以我们将主题出现的次数从大到小进行排序，选取出现次数位居前四的主题作为影评中的核心关注点。

我们要使用 LDA 进行主题提取之前，首先要确定主题的个数，主题个数的提取可以使用最大似然方法来进行优化，首先抽取每个主题的对数似然估计值，而后计算各个主题的调和平均数，将其最为模型的最大似然估计，最后画出主题数-似然估计曲线图，选取最大值的那个点作为最佳主题数。

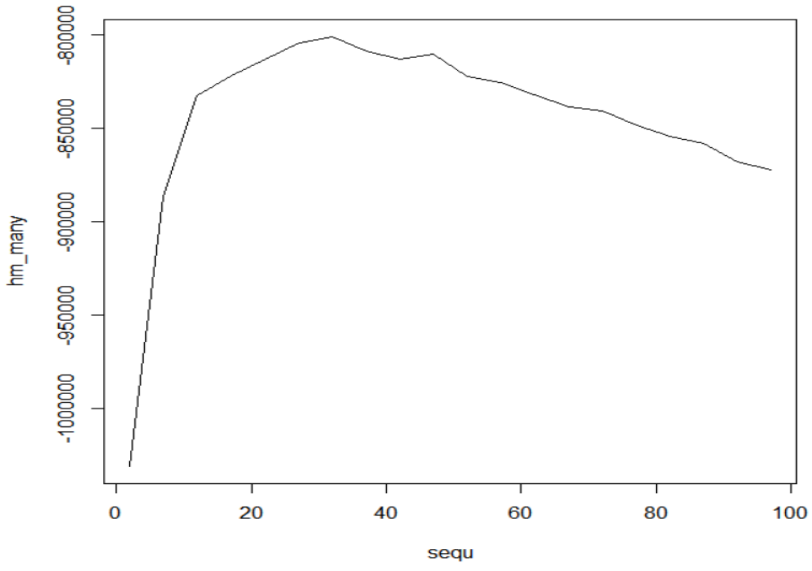


图 5-1 《上海堡垒》积极影评主题数-似然估计曲线图

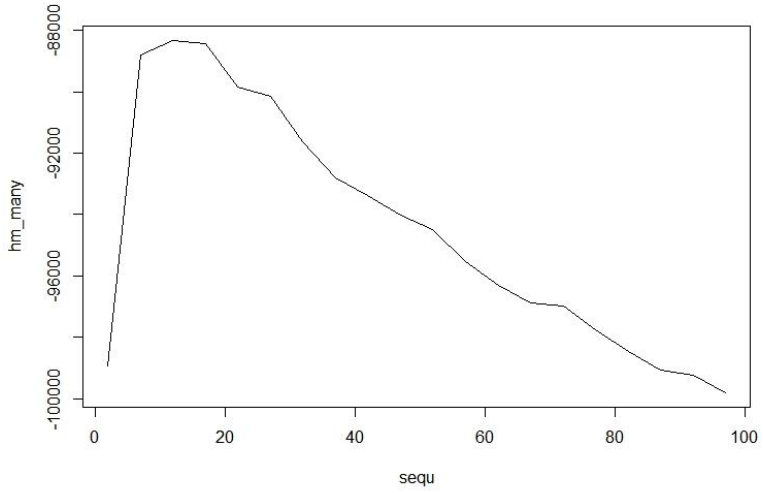


图 5-2 《上海堡垒》消极影评主题数-似然估计曲线图

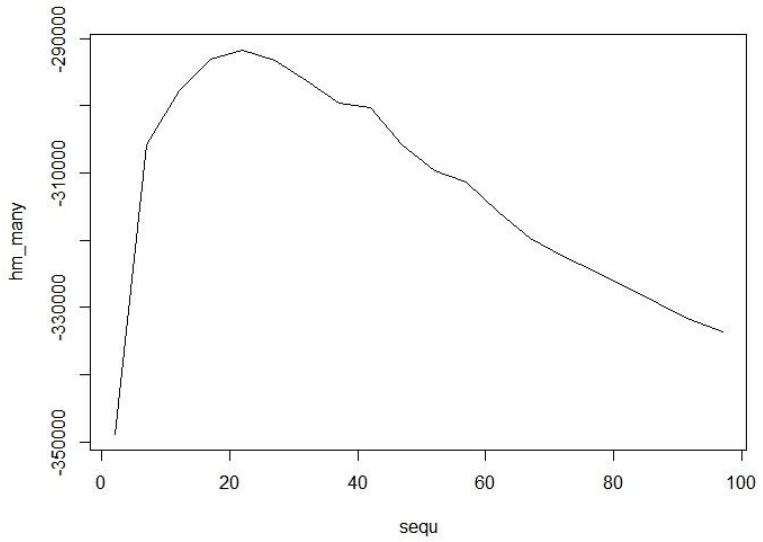


图 5-3 《战狼 2》积极影评主题数-似然估计曲线图

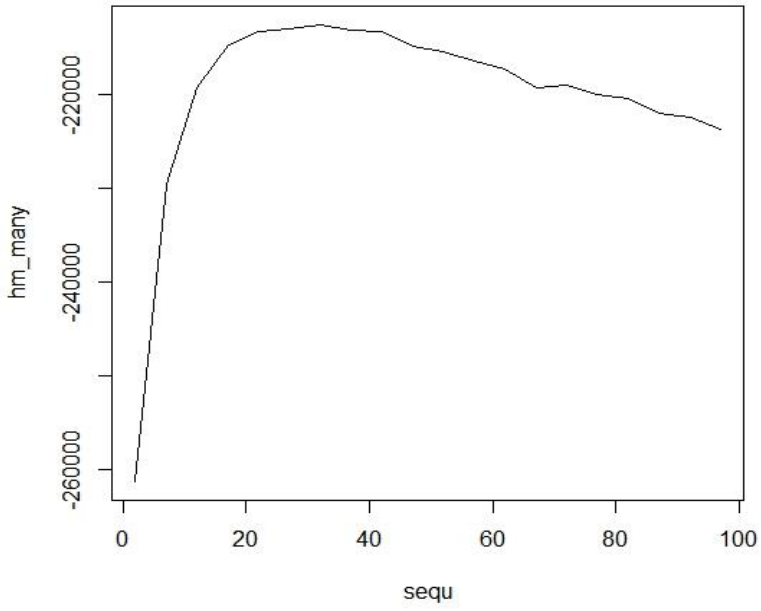


图 5-4 《战狼 2》消极影评主题数-似然估计曲线图

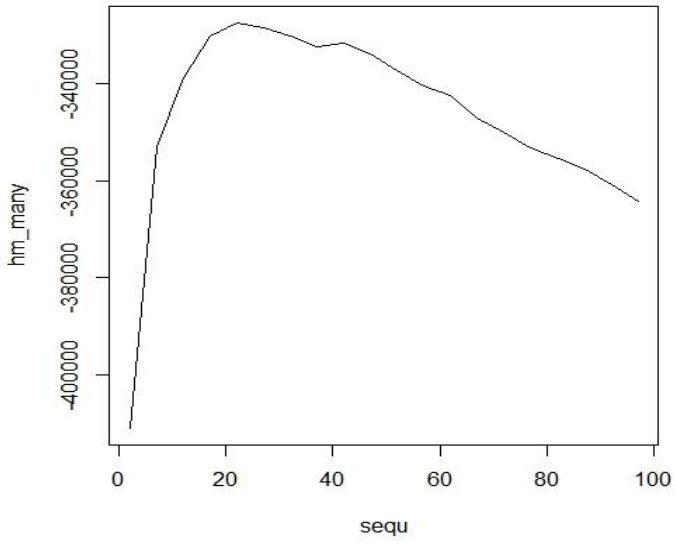


图 5-5 《红海行动》积极影评主题数-似然估计曲线图

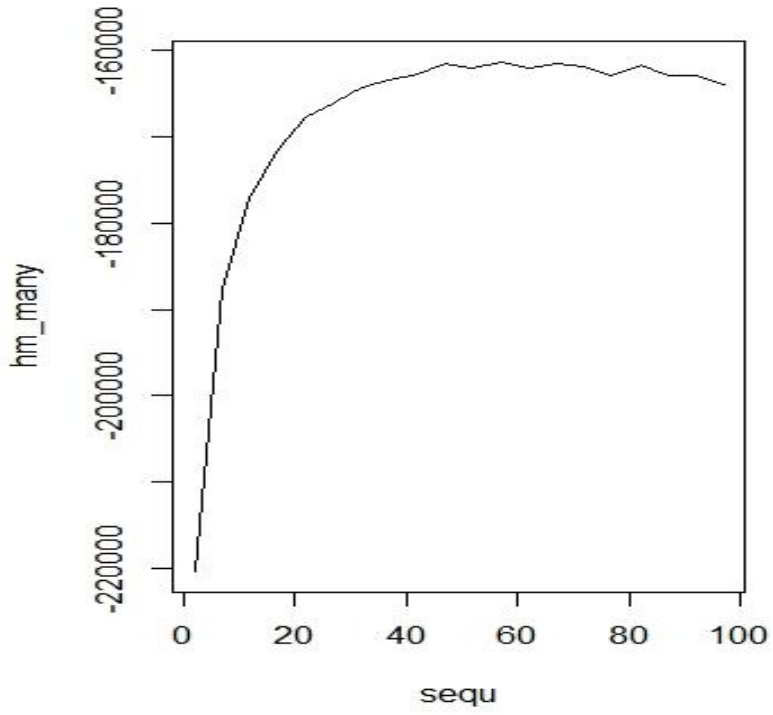


图 5-6 《红海行动》消极影评主题数-似然估计曲线图

由上述图像可知，《上海堡垒》正面影评最佳主题数为 32，负面影评最佳主题数为 12；《战狼 2》正面影评最佳主题数为 22，负面影评最佳主题数为 32；《红海行动》正面影评最佳主题数为 21，负面影评最佳主题数为 57。

分别进行 LDA 主题建模并列举出出现次数位居前四的主题，建模结果如下：

表 5-1 《上海堡垒》正面影评潜在主题

主题 1	主题 2	主题 3	主题 4
紧凑	视觉	身手	舒淇
内容	特效	演技	敬业
精彩	场景	到位	鹿晗

剧情	国产	精彩	认真
情节	场面	惊险	高以翔
狗血	震撼	刺激	做作
跌宕起伏	宏大	酷炫	认真
太快	逼真	过瘾	实力
演绎	炸裂	触目惊心	叹为观止
节奏	国产	犀利	无人能及

根据《上海堡垒》正面影评的 4 个潜在主题的挖掘情况，我们可以看到在主题 1 中的高频特征词语有剧情、情节、紧凑、跌宕起伏、扣人心弦等词语，这一主题主要反映《战狼》剧情精彩，节奏紧凑，内容扣人心弦；在主题 2 中的特征词语有视觉、场景、特效、逼真、身临其境等词语，这一主题主要用于反映《上海堡垒》特效逼真，场面宏大令观众感到震撼，给观众一种身临其境的感觉；在主题 3 中出现的高频特征词有演技、身手、到位、酷炫等词语，这一主题主要反映的是演员演技到位、犀利，动作酷炫、精彩；主题 4 中出现的词语是舒淇、鹿晗、敬业、认真等词语，这一主题主要反应演员态度认真、敬业、是实力派；

表 5-2 《上海堡垒》负面影评潜在主题

主题 1	主题 2	主题 3	主题 4
简单	内容	单一	鹿晗
画面	剧情	太烂	失望
随意	没章法	身手	不配
场面	无逻辑	生硬	脱节
缺陷	拖沓	演技	差劲
特效	味同嚼蜡	尴尬	舒淇
廉价	苍白	浮夸	幽默
粗糙	老套	动作	凑合
震撼	好看	死气沉沉	无语
太假	硬伤	蹩脚	模仿

根据《上海堡垒》负面影评的 4 个潜在主题的挖掘情况，我们可以看到

在主题 1 中的特征词语有特效、画面、缺陷等词语，这一主题主要反映特效廉价、有缺陷，制作粗糙；主题 2 中的高频特征词语有内容、拖沓、剧情、老套等词语，这一主题主要反映剧情老套、节奏拖沓，内容浮夸；在主题 3 中的特征词语有演技、太烂等词语，这一主题反映了演员演技太烂，动作生硬、蹩脚；在主题 4 中的特征词语有舒淇、鹿晗、模仿、不配等词语，这一主题主要反映观众对于演员的失望。

表 5-3 《战狼 2》正面影评潜在主题

主题 1	主题 2	主题 3	主题 4
节奏	震撼	张翰	超燃
澎湃	一流	吴京	到位
紧凑	特效	导演	演技
跌宕起伏	音效	正能量	值得一看
情节	效果	认真	实力
剧情	身临其境	硬汉	表演
丰富	逼真	威武	能量
夸张	良心	游刃有余	爆表
回味无穷	不太	淋漓尽致	在线
不太	好莱坞	失望	还行

根据《战狼 2》正面影评的 4 个潜在主题的挖掘情况，我们可以看到在主题 1 中的高频特征词语有剧情、节奏、澎湃、回味无穷等词语，这一主题主要反映了电影剧情丰富，情节紧凑，内容感人；在主题 2 中的高频特征词语有特效、一流、音效、震撼等词语，这一主题主要反映了电影特效一流，制作良心，能和好莱坞有的一拼；在主题 3 中的高频特征词语有张翰、吴京、导演、正能量等词语，这一主题主要反映了观众对于演员的认可，觉得他们把角色扮演的淋漓尽致，态度认真、负责；在主题 4 中的高频特征正词语有表演、到位、演技、爆表等词语，这一主题主要反映了演员表演到位、演技爆表、值得一看。

表 5-4 《战狼 2》负面影评潜在主题

主题 1	主题 2	主题 3	主题 4
老套	尴尬	暴力	张翰
狗血	浮夸	打斗	导演
内容	抄袭	太假	吴京
剧情	精彩	一般	别扭
无脑	不伦不类	画面	傻不拉几
牵强	评价	特效	做作
真心	没特色	血腥	没实力
漏洞百出	演技	夸张	个人英雄
还好	生硬	鸡血	努力
不接地气	一般	民族	一无是处

根据《战狼 2》负面影评的 4 个潜在主题的挖掘情况，我们可以看到在主题 1 中的高频特征词语有剧情、漏洞百出、牵强等词语，这一主题主要反应电影剧情老套、牵强，内容漏洞百出；在主题 2 中的高频特征词语有演技、浮夸、抄袭等词语，这一主题主要反应演员演技尴尬、浮夸、不在线，动作生硬、没特色；在主题 3 中的高频特征词语有特效、太假、画面、暴力等词语，这一主题主要反应电影特效一般，令人感觉太假、不真实，而且画面暴力；在主题 4 中的高频特征词语有张翰、吴京、别扭等词语，这一主题主要反应观众认为吴京、张翰没有实力而且做作，将角色演得很别扭、一无是处。

表 5-5 《红海行动》正面影评潜在主题

主题 1	主题 2	主题 3	主题 4
情节	场面	演技	黄景瑜
题材	逼真	炸裂	张译
紧凑	震撼	在线	辛苦
跌宕起伏	感同身受	训练	颜值
荡气回肠	刺激	很棒	生动
剧情	特效	超级棒	实力派

尿点	宏伟	无可挑剔	敬业
热血	血腥	动作	铁血
扣人心弦	强大	不突兀	刻画
新颖	搞笑	超赞	给力

根据《红海行动》正面影评的 4 个潜在主题的挖掘情况，我们可以看到在主题 1 中的高频特征词语有剧情、题材、新颖、扣人心弦等词语，这一主题主要反应电影题材新颖，过程毫无尿点，情节紧凑；在主题 2 中的高频特征词语有特效、逼真、感同身受等词语，这一主题主要反应电影特效投入了大手笔，场面宏伟刺激；在主题 3 中的高频特征词语有演技、在线、无可挑剔、炸裂等词语，这一主题主要反应演员演技炸裂，动作都是提前训练，令人感觉超赞；在主题 4 中的高频特征词语有黄景瑜、张译、颜值、敬业等词语，这一主题主要反应观众对于演员本身的认可。

表 5-6 《红海行动》负面影评潜在主题

主题 1	主题 2	主题 3	主题 4
情节	画面	演技	演员
空洞	血腥	浮夸	张译
单调	少儿不宜	尴尬	海清
拖沓	差劲	一般	导演
脱节	眩晕	不到位	形象
雷同	夸大	单调	空洞
剧情	特效	动作	不符合
老套	尺度	粗糙	没选好
偏少	太假	生硬	模仿
抄袭	五毛	蹩脚	人物性格

根据《红海行动》负面影评的 4 个潜在主题的挖掘情况，我们可以看到在主题 1 中的高频特征词语有情节、空洞、拖沓、雷同等词语，这一主题主要反应电影情节空洞，内容拖沓，与前人电影雷同；在主题 2 中的高频特征词语有画面、血腥、特效、眩晕等词语，这一主题主要反应电影特效制作太

假，画面过于血腥，令人眩晕；在主题 3 中的高频特征词语有演技、浮夸、尴尬等词语，这一主题主要反应演员演技尴尬不到位，动作单调生硬；在主题 4 中的高频特征词语有张译、海清、形象等词语，这一主题反映了观众认为演员形象不符合角色，没有把角色演活。

综上，由《上海堡垒》、《战狼 2》与《红海行动》这三部电影挖掘的潜在主题中，观众对于这三部电影的积极情绪体现在以下几个方面：

1.剧情精彩、情节紧凑、内容跌宕起伏、荡气回肠。

2.视觉特效令人震撼、场面宏伟。

3.演员演技炸裂，打斗触目惊心、过瘾，动作犀利、热血沸腾。

4.演员敬业、给力，态度认真、负责，将角色演的绘声绘色。

观众对于这三部电影的消极情绪体现在以下几个方面：

1.剧情老套、内容单调、拖沓。

2.画面太过血腥、特效粗制滥造。

3.演员演技浮夸，动作生硬、蹩脚。

4.演员演的角色不生动、中规中矩、毫无特色。

5.2 影评聚类分析

5.2.1 聚类分析的相关理论

聚类算法即使得簇内数据相较于簇外数据之间的相似度要高的分类方法。本文针对情感分类之后的影评分别采用聚类方法，来获得不同情感的观众对于电影所持有的不同看法，获取评论中观众关注电影的哪些方面。把握观众对于电影的不同观点。

5.2.2 观众评论聚类分析

对经过预处理后的正负面评论分别使用空间向量模型，建立文档-词项矩阵(TF-IDF)，矩阵中的每个值的含义为相应行上的词项乘上相应列上的

文档的权重值，将文本信息转变为利于分析的定量信息。

本文使用 k-Means 聚类算法分别来对正负面影评进行聚类，因为首先要确定初始质心个数即 k 值，所以我们使用轮廓系数来确定，轮廓系数计算公式如下：

$$SC_i = \frac{b_i - a_i}{\max(b_i, a_i)} \tag{5.1}$$

其中 b_i 是样本点到所有非本身所在簇的点的平均距离，a_i 是样本点到所有它属于的簇中其它点的平均距离。可见轮廓系数的值是介于 $[-1,1]$，越趋近于 1 代表内聚度和分离度都相对较优。

本文确定 k 值的取值范围 $[2,10]$，通过电脑计算当 k 在此范围内的轮廓系数的值，选轮廓系数最大的那个数字为 k 值，最终结果如下表：

表 5-7 选取的 k 值

	《上海堡垒》	《战狼 2》	《红海行动》
积极评论 k 值	7	6	6
消极评论 k 值	6	5	6

因为聚类结果可能出现相似的类，所以我们将相似的类簇划分为一个类别，则处理过后的三部电影的积极和消极影评的最终聚类结果如下表所示：

表 5-8《上海堡垒》影评聚类结果

《上海堡垒》					
正面评论	演员导演	剧情	演技动作	特效	其他(热血沸腾、激情)
负面评论	演员导演	剧情	演技动作	特效	其他(不值得、一般般、开挂)

表 5-9《战狼 2》影评聚类结果

《战狼 2》					
正面评论	演员导演	剧情	演技动作	特效	其他(正能量、爱国主义)
负面评论	演员导演	剧情	演技动作	特效	其他(政治宣传、少儿不宜)

表 5-10 《红海行动》影评聚类结果

《红海行动》					
正面评论	演员导演	剧情	演技动作	特效	其他(震撼、强大、大无畏精神)
负面评论	演员导演	剧情	演技动作	特效	其他(费劲、无望后悔、抗日神剧)

从以上聚类结果来看，《上海堡垒》、《战狼 2》与《红海行动》的正负面评论共同关注点都有人物、剧情、特效、演技这四个方面，但是上表并不直观，所以通过做可视化图像使聚类结果更加直观具体，图像如下：

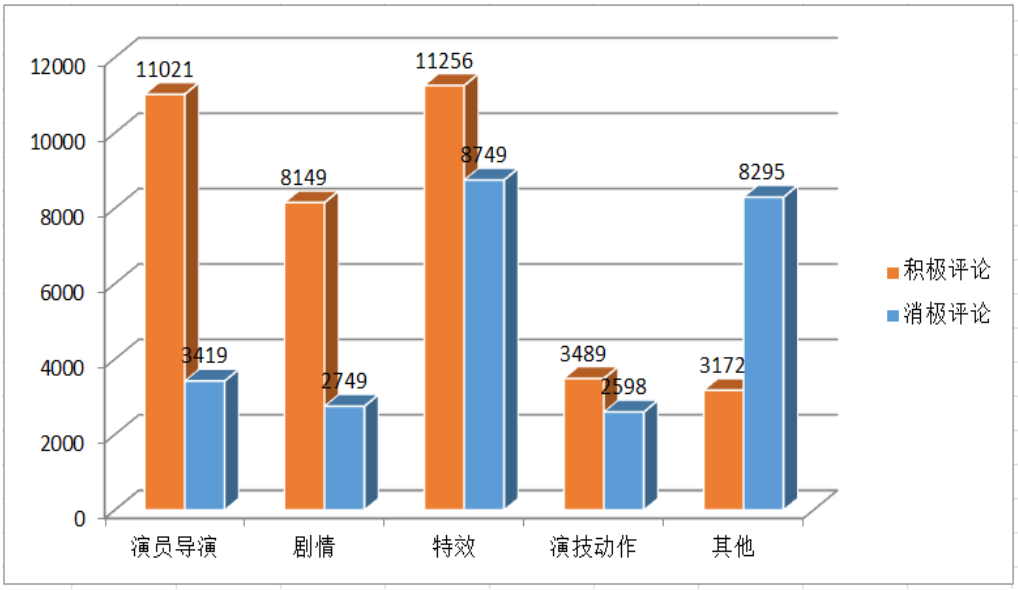


图 5-7 《上海堡垒》影评聚类结果

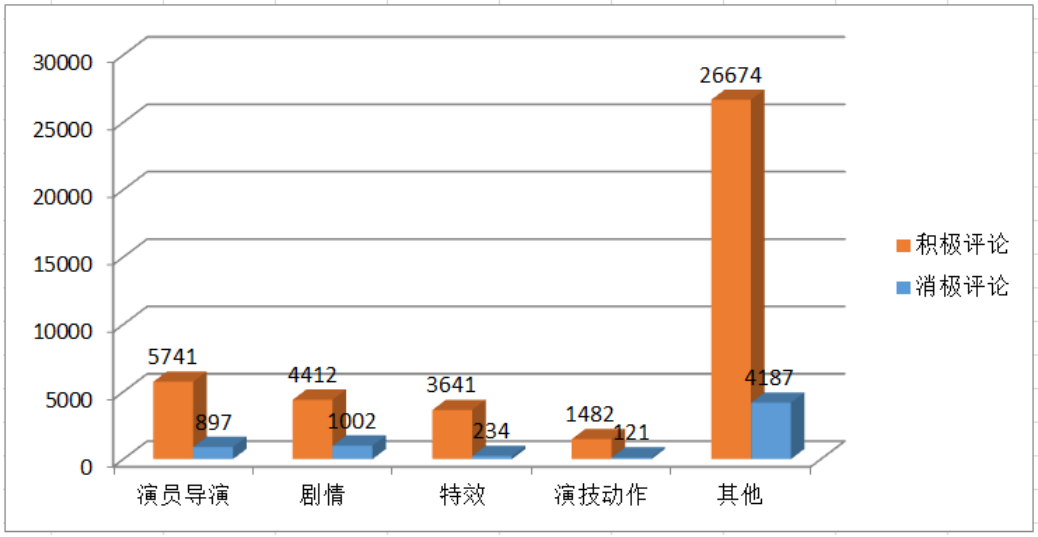


图 5-8 《战狼 2》影评聚类结果

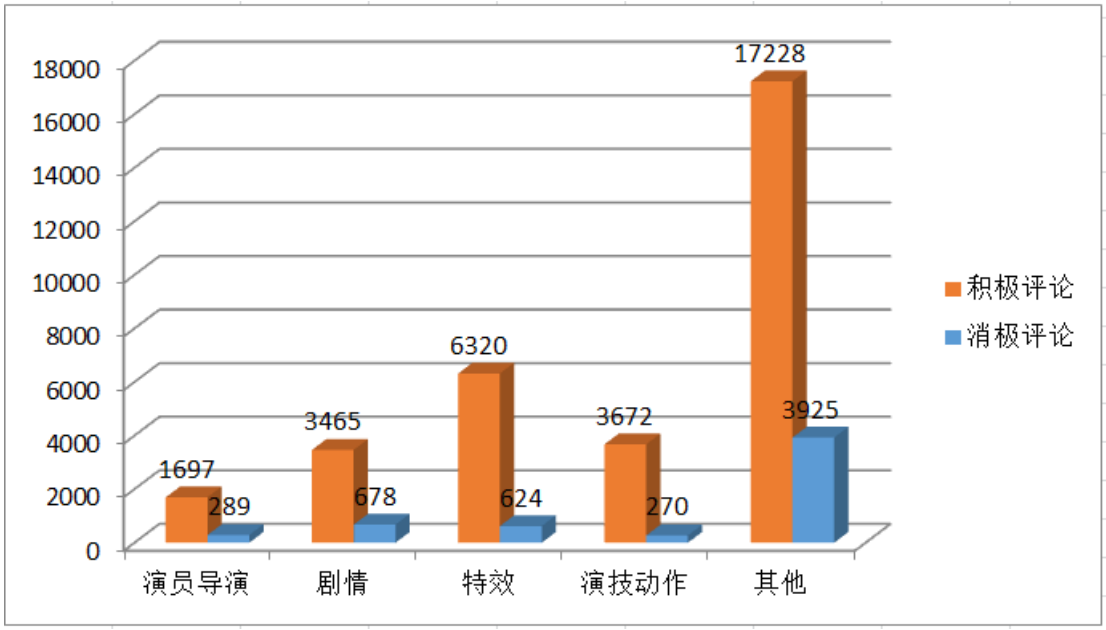


图 5-9《红海行动》影评聚类结果

我们通过做可视化图将这三部电影中观众关注的相同方面进行比较，结果如下图：

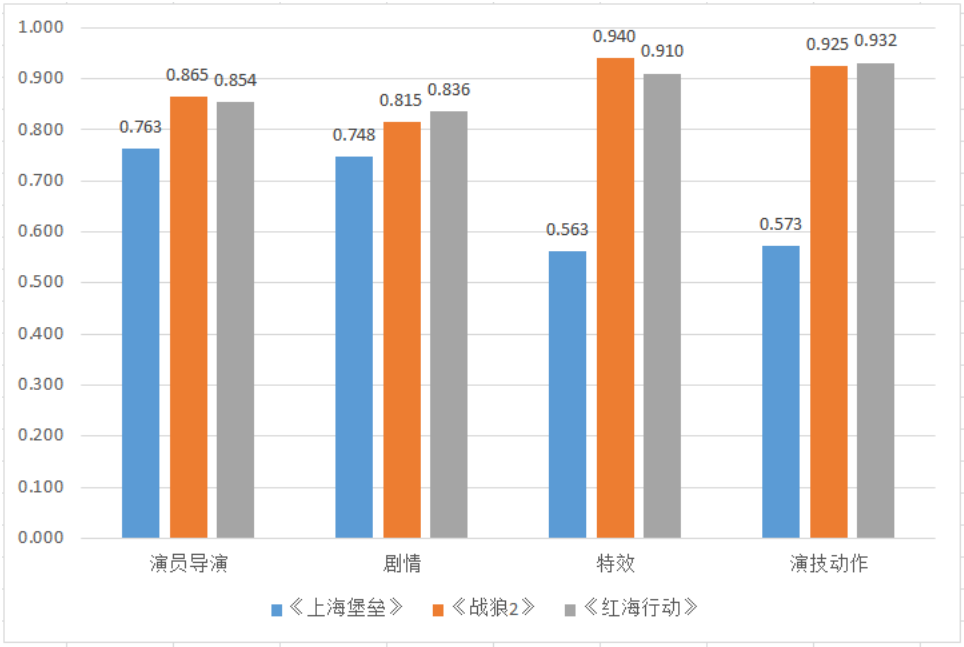


图 5-10 影评好评率比较图

从图 5-10 可以发现关于《上海堡垒》影评类簇好评率从高到低依次为演员导演、剧情、演技动作、特效；《战狼 2》影评类簇好评率从高到低依次为特效、演技动作、演员导演、剧情；《红海行动》影评类簇好评率从高到低依次为演技动作、特效、演员导演、剧情。

首先我们从整体上看，《战狼 2》与《红海行动》均优于《上海堡垒》，说明观众对于前两部电影的观影体验要好于后者，这也符合各大电影评分网站上的评分结果。

其次，从演员导演和剧情这两类的对比结果来看，我们可以发现《战狼 2》与《红海行动》的好评率皆在 0.8 以上，略好于《上海堡垒》，《上海堡垒》的好评率离 0.8 还有些差距，但差距并不是很大。这说明观众觉得《上海堡垒》不好看的主要原因并非是因为演员导演和剧情这两类。

然后再从特效和演技动作这两类的对比结果来看，《战狼 2》和《红海行动》皆要明显好于《上海堡垒》，前两部电影的好评率皆超过了 0.9，而《上海堡垒》得好评率则连 0.6 都没有达到，差距极大。

综合以上对于《上海堡垒》、《战狼 2》以及《红海行动》这三部电影评论的情感分析结果来看，我们可以得出结论即对于动作片电影来说，观众更在意的是整部电影的特效和演技动作这两个方面，逼真精致的特效、专业

的演技动作是最能吸引观众的两点。

所以对于后续的国产动作片电影提出以下两点建议：

1.特效是重头戏，观众很看重的一点就是特效，所以在特效、场景、画面要下大功夫，不能粗制滥造，画面不要过于血腥，要找专业的、口碑好的、有经验的特效公司进行特效制作。

2.其次是演技动作方面。需要专业的动作指导，动作要丰富，如行云流水一般流畅，避免生硬、蹩脚。

第 6 章 总结与展望

6.1 总结

本文主要挖掘了猫眼 APP 上《上海堡垒》、《战狼 2》与《红海行动》这三部动作电影的评论数据，首先使用 python 爬虫技术抓取电影的文本评论数据；其次对电影评论文本数据进行情感倾向分析，其次使用 FV-SA-SVM 分类算法对数据进行初步分类，把每部电影的影评分成正面影评、负面影评两个部分，再使用情感字典打分的方法对已分类的每条影评进行打分，并挑选出异常影评数据再进行分类，这样最大程度降低影评错分；紧接着对已经分类之后的影评数据进行社会语义网络分析，并且绘制了语义网络图，通过概念之间的语义关系发现电影的特征，对比电影的不同；然后分别对电影的积极评论、消极评论进行 LDA 主题模型分析，挖掘潜在主题，从不同的潜在主题中，发现电影的可取之处以及不足；最后对电影的积极评论、消极评论进行聚类分析，得到影评中观众关注的有剧情、特效、演技、人物四大方面，并把结果进行对比分析总结。

所以本文主要结论有以下两点：

第一，本文 FV-SA-SVM 分类算法的准确率、精确率、召回率以及 F1 值这四个分类指标均优于文献中 SA-SVM 算法以及传统分类算法，证明了 FV-SA-SVM 分类算法在影评情感分类领域具有很好的优越性能。

第二，通过对《上海堡垒》、《战狼 2》以及《红海行动》这三部电影影评挖掘分析，我们发现优秀的动作电影中最吸引观众的是电影的特效和演技动作两个方面。而《上海堡垒》之所以评分和票房低无疑是因为特效和演技动作这两点表现太差所致。

接下来，对于后续的国产动作片电影提出以下建议：

1. 特效是重头戏，观众很看重的一点就是特效，所以在特效、场景、画面要下大功夫，不能粗制滥造，画面不要过于血腥，要找专业的、口碑好的、有经验的特效公司进行特效制作。

2. 其次是演技动作方面。需要专业的动作指导团队，动作要丰富，如行云流水一般，避免生硬、蹩脚。

6.2 展望

本文数据来源有些单一，只爬取了猫眼电影 APP 上的影评，也应该从其他影评网站爬取影评，例如淘票票、时光网、豆瓣电影等，可以以此来保证数据来源的多样性。

还有本文使用的 FV-TFIDF 权重计算方法虽然是依据 TFIDF 改进而来的，但它本质上并不是一种全新的算法，接下来的实验中可以使用一些新的方法，例如 Word2Vec、Doc2Vec 等，其中 Word2Vec 可以训练词向量，可以把特征映射到 K 维向量空间，可以为文本数据寻求更加深层次的特征表示，而 Doc2vec 模型其实是在 Word2vec 模型的基础上做出的改进，考虑了单词之间的排列顺序对句子或文本信息的影响。

还有通常在大样本数据的分类研究中，深度学习算法比传统机器学习分类算法效果要好，这些都是本文的不足，在接下来的研究中，可以使用深度学习算法，希望可以获得进一步的结果。

参考文献

[1] 王根生，黄学坚，基于 Word2vec 和改进型 TF-IDF 的卷积神经网络文本分类模型. 小型微型计算机系统，2019.40(05)：第 1120-1126 页.

[2] 郭超磊，基于 SA_SVM 的中文文本分类研究_郭超磊. 计算机应用与软件，2019. 36(3):第 277-281 页.

[3] 杨瑞欣，电商空调产品的评论数据情感分析，2017，山西大学. 第 56 页.

[4] 杨立公，朱俭，汤世平. 文本情感分析综述[J]. 计算机应用，2013，33(6)：1574-1607.

[5] 杨小平, 张中夏, 王良, 等. 基于 Word2Vec 的情感词典自动构建与优化[J]. 计算机科学，2017，44(1)：42-47.

[6] Hatzivassiloglou V,McKeown KR., Predicting the Semantic Orientation of Adject ives. Proceedings of the Acl, 1997: p. 174--181.

[7] Turney, P.D. Thumbs Up or Thumbs Down: Semantic Orientation Applied to Un supervised Classification of Reviews. in 40th Annual Meeting of the Associatio n for Computational Linguistics:(CD:CD-CNF-0517). 2002. Philadelphia,PA.

[8] Wilson, Theresa, Wiebe, et al. Recognizing contextual polarity in phrase-level se ntiment analysis[J].2010,7(5)P.12

[9] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis[J]. Computational linguistics,2011,37(2): 267-307.

[10] Cambria E, Schuller B, Xia Y, et al. New avenues in opinion mining and senti ment analysis[J].IEEE Intelligent Systems, 2013, 28(2): 15-21.

[11] Tajinder Singh, Madhu Kumari. Role of Text Pre-processing in Twitter Sentime nt Analysis [J]. Procedia Computer Science, 2016, 89:549-554.

[12] 李寿山，李逸薇，黄居仁等. 基于双语信息和标签传播算法的中文情感词典构建方法[J]. 中文信息学报，2013，27(6)：75-81.

[13] 阳爱民，林江豪，周咏梅. 中文文本情感词典构建方法［J］. 计算机科学与探索，2013，7(11)：1033-1039.

[14] 王志涛，於志文，郭 斌，等. 基于词典和规则集的中文微博情感分析［J］. 计算机工程与应用，2015，51(8)：218-225.

[15] 周杰. 基于情感词典与句型分类的中文微博情感分析研究［D］. 银川：宁夏大

学，2016.

[16]张克亮，黄金柱，曹蓉，等．基于 NHC 语境框架和情感词典的文本情感倾向分析［Ｊ］．山东大学学报（理学版），2016，51(7)：51-58.

[17]栗雨晴，礼欣，韩煦，等．基于双语词典的微博多类情感分析方法［Ｊ］．电子学报，2016，44(9)：2068-2073.

[18] Vilares, D., M.A. Alonso and C. Gomez-Rodriguez, Supervised sentiment analysis in multilingual environments. Information Processing & Manag ement: Libraries and Information Retrieval Systems and Communication Networks: An International Journal, 2017. 53(3): p. 595-607.

[19] Sharma, A. and S. Dey, A boosted SVM based ensemble classifier for sentiment analysis of online reviews. ACM SIGAPP Applied Computing Review, 2013. 13: p. 43-52.

[20] Manek A S, Shenoy P D, Mohan M C, et all.Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier[J].World Wide Web, 2017, Vol.20 (2), pp.135-154

[21] Isidoros Perikos,Ioannis Hatzilygeroudis. Recognizing emotions in text using ensemble of classifiers[J]. Engineering Applications of Artificial Intelligence,2016,51.

[22] Abinash Tripathy,Ankit Agrawal,Santanu Kumar Rath. Classification of sentiment reviews using n-gram machine learning approach[J]. Expert Systems With Applications,2016,57.

[23] Maria Giatsoglou,ManolisG. Vozalis,Konstantinos Diamantaras,Athena Vakali,George Sarigiannidis,Konstantinos Ch. Chatzisavvas. Sentiment analysis leveraging emotions and word embeddings[J]. Expert Systems With Applications,2017,69.

[24] Information Technology; Recent Findings from Microsoft Research Has Provided New Information about Information Technology (Sentiment Embeddings with Applications to Sentiment Analysis)[J]. Computers, Networks & Communications,2016.

[25] Fernandez-Gavilanes, M., et al., Unsupervised method for sentiment analysis in online texts. Expert Systems with Application, 2016. 58(Oct.): p. 57-75.

[26] LiangBin,LiuQuan,etal. Aspect-based sentiment analysis based on multiattention

CNN[J].Computer research and development,2017,54(08):1724-1735.

[27] Shathi, S.P., et al. Enhancing Performance of naïve bayes in text classification by introducing an extra weight using less number of training examples. 2016 International Workshop on Computational Intelligence (IWCI).IEEE， 2017: 142-147.

[28] Bahassine, S., A. Madani and M. Kissi. An improved Chi-sqaure feature selection for Arabic text classification using decision tree. in International Conference on Intelligent Systems: Theories & Applications. 2016.

[29] Goudjil M，Koudil M，Bedda M，et al．A novel active learning method using SVM for text classification［J］．International Journal of Automation and Computing，2018(3): 1-9．

[30] 陈健飞，蒋刚，杨剑锋．改进 ABC-SVM 的参数优化及应用［J］．机械设计与制造，2016(1) ：24-28．

[31] 张进，丁胜，李波．改进的基于粒子群优化的支持向量机特征选择和参数联合优化算法［J］．计算机应用，2016，36(5) ：1330-1335．

[32] 庄严，白振林，许云峰．基于蚁群算法的支持向量机参数选择方法研究［J］．计算机仿真，2011，28(5) ：216-219．

[33] 陈晋音，熊晖，郑海斌. 基于粒子群算法的支持向量机的参数优化［J］．计算机科学，2018，45(6) ：197-203．

[34] 王万良，陈超，李笠，等．基于模拟退火的自适应水波优化算法［J］．计算机科学，2017，44(10) ：216-221．

[35] 戴立武． 基于深度神经网络的中文情感分析研究[D]. 华南理工大学, 2019.

[36] 王任远． 网购评语情感挖掘研究[D]. 大连海事大学, 2014.

[37] 段敬民, 常跃军, 李赞祥, 崔建明. 基于退火算法的物流配送网的求优研究[J]. 中国工程科学, 2012,14(07):109-112.

[38] Punniyamoorthy, M. and S. Manochandar, Scaling feature selection method for enhancing the classification performance of Support Vector Machines in text mining. Computers & Industrial Engineering, 2018. 124: p. 139-156.

[39] Redhu, S., Sentiment Analysis Using Text Mining: A Review. International Jour nal on Data Science and Technology, 2018. 4: p. 49.

[40] 范佳健. 微博评论信息的聚类分析[D]. 安徽大学, 2019.

[41] 王伟等, 一种基于 LDA 主题模型的评论文本情感分类方法. 数据采集与处理, 2017.

32(3): 第 629-635 页.

[42] Abu Taher, S.M., K. Afsana Akhter and K.M. Azharul Hasan. N-Gram Based Sentiment Mining for Bangla Text Using Support Vector Machine. 2018: IEEE.

[43] Ahn, J., M. Kang and K. Lee, Opinion mining using ensemble text hidden Markov models for text classification. Expert Systems with Application, 2018. 94(Mar.): p. 218-227.

[44] ZHAI Dong-hai, YU Jiang, GAO Fei, 等. 最大距离法选取初始簇中心的 K-means 文本聚类算法的研究[J]. 计算机应用研究, 2014, 31(3):713-715.

[45] 朱连江, 马炳先, 赵学泉. 基于轮廓系数的聚类有效性分析[J]. 计算机应用, 2010(s2):139-141.

[46] Arif, M.H., et al., Sentiment analysis and spam detection in short informal text using learning classifier systems. Soft computing: A fusion of foundations, methodologies and applications, 2018. 22(21): p. 7281-7291.

[47] Cheng, L. and C. Huang, Exploring contextual factors from consumer reviews affecting movie sales: an opinion mining approach. Electronic Commerce Research, 2019.

[48] 周庆平等, 基于聚类改进的 KNN 文本分类算法. 计算机应用研究, 2016. 33(11): 第 3374-3377,3382 页.

[49] 刘述昌，张忠林. 基于中心向量的多级分类 KNN 算法研究［J］. 计算机工程与科学，2017，39(9) : 1758 － 1764.

[50] 甄志龙. 文本分类中的特征选择方法研究［M］. 长春: 吉林大学出版社，2016.

[51] Butnaru A M, Ionescu R T. From image to text classification: a novel approach based on clustering word embeddings [J]. Procedia Computer Science, 2017：112-120.

[52] Sabbah T, Selamat A, Selamat M H,et al. Modified frequency-based term weighting schemes for text classification [J]. Applied Soft Computing, 2017, 58(9): 193-206.

[53] 刘志康. 一种改进的混合核函数支持向量机文本分类方法[J]. 工业控制计算机, 2016,29 (6): 113-114.

[54] 钱慎一, 杨铁松. 基于微博电影评论的情感分析研究[J]. 现代计算机(专业版), 2017(05):48-51.

[55] 牟兴. 基于中文微博的电影评论情感极性分类及舆论演化分析[D]. 西华大

学, 2017.

[56] 李明. 面向微博电影评论的情感分类研究[D]. 云南财经大学, 2014.

[57] Siddhartha Kumar Arjaria, J.M.M., Polarity Based SVM Techniques for Analyzing Movie Review. International conference on advanced computing and software engineering (ICACSE-2019), 2019.

[58] Yoosin Kim, M.K.S.R., Text Mining and Sentiment Analysis for Predicting Box Office Success. KSII Transactions on Internet and Information Systems, 2018. 12(8).

[59] 唐利. 网络电影评论的情感倾向性分类研究[J]. 遵义师范学院学报, 2018,20(06):160-164.

[60] 胡晓康. 基于 SOW-BTM 的网络电影评论情感分类研究[D]. 山西财经大学, 2018.

[61] 甘雨涵. 基于 Stacking 方法的电影票房预测[D]. 上海师范大学, 2018.

[62] 殷复莲, 潘幸艺, 柴剑平. 基于词向量的电影评论情感分析方法[J]. 现代电影技术, 2017(08):4-9.

[63] 冯莉. 面向英文电影评论的文本情感倾向性分类研究[D]. 大连海事大学, 2013.

[64] 郭伟. 网络电影评论的情感挖掘分析[D]. 吉林大学, 2010.

[65] 殷复莲, 潘幸艺, 柴剑平. 基于词向量的电影评论情感分析方法[J]. 现代电影技术, 2017(08):4-9.

致 谢

时光荏苒，为期两年的研究生生活即将结束，在上海师范大学攻读硕士的两年，是自己收获最为丰硕的两年。在这两年里，我进一步学习到了更多的专业知识，培养了独立学习与思考的能力。同时，在学校老师的指导下，我的社会实践能力有了很大的提升，对自己未来的职业规划有了清晰明确的认识。在读研的过程中，我也由入学前不严谨、急躁、科研能力匮乏的女孩成长为一名乐观、冷静、踏实、热爱科研的女生。这些成长变化，离不开学校老师、同学、家人们的关心和帮助。在即将为我的研究生生活画上句号的时刻，我应该向你们致以我最诚挚的谢意，在如此优越的环境下，我们都将成长得越来越美好！

首先，要对我的导师崔百胜教授表示衷心的感谢。在两年的硕士研究生学习过程中，崔百胜老师在学习上、生活上和工作上给予我极大的关怀和帮助。在指导我们的学习和研究时，他始终耐心地给与指导和帮助，给我提出尘贵的意见。崔百胜老师精湛的专业知识、严谨的治学态度、诚信的为人风尚、踏实的工作作风，都给我留下深刻的印象，使我受益匪浅。他大胆开拓的进取精神，循循善诱的启发方式，和蔼可亲的待人态度，都是我今后治学为人的楷模，激励着我在今后的人生道路上不断进取。

感谢所有上师大应用统计专业的老师和同学们，他们提出的宝贵建议和学习上的帮助使我获益良多，有了他们，两年的研究生生活才是那么的充实和精彩。

最后，感谢含辛茹苦育我成长的亲人和不断给我帮助的好友，感谢他们在生活上给予我无微不至的关怀，用无私的爱鼓舞我专心学习和研究。