论文题目：基于视频弹幕分析的一种结合客观语料分析的情感分析模型

1 绪论 4

1.1 研究背景与意义 4

1.1.1 研究背景 4

1.1.2 研究意义 4

1.2 国内外研究现状 4

1.2.1 文本情感分析研究 4

1.2.2 文本情感分析在弹幕领域的应用 5

1.3 论文的主要研究内容 5

1.4 论文的组织结构 6

第二章 相关理论及技术 7

2.1 情感分析流程 7

2.2 文本情感分析常用技术 7

2.2.1 文本预处理 7

2.2.3 文本表示模型 9

2.2.5 情感词典 10

2.3 机器学习分类算法与性能 10

2.3.1 贝叶斯 10

2.3.2 支持向量机 11

2.4 分类性能 12

2.4.1 准确率、召回率、F-测度值 12

2.4.2 微平均和宏平均 13

2.5 本章小结 13

第三章 基于情感词典和LDA的情感特征提取 15

3.1 情感词典的构建 15

3.1.1通用基础词典 15

3.1.2扩充词典 15

3.2 基于LDA的文本主题模型的构造与改进 16

3.2.1潜在狄利克雷分配主题模型构造 16

3.2.2改进LDA模型的文本主题分类 17

3.3 基于词典和LDA的文本情感值计算 18

3.4实验结果及分析 19

3.4.1基于词典与LDA结合的情感分类实验 19

3.5本章小结 21

第四章 基于弹幕文本的SVM情感分析算法研究 22

4.1 问题描述 22

4.2 基于弹幕文本的SVM情感分析算法研究 23

4.2.1一种基于SVM的高维混合特征的弹幕文本情感倾向性分析算法 23

4.2.2结合客观弹幕文本的机器学习情感分类模型 26

4.3 实验结果与分析 31

4.3.1基于SVM的高维混合特征弹幕情感分类实验 31

4.3.2结合客观语料的弹幕情感分类对比实验 32

4.4 本章小结 33

第五章 基于弹幕文本的情感分析系统实现与测试 34

5.1 基本流程 34

5.2 算法设计 34

第六章 结论与展望 35

6.1 结论 35

6.2 展望 35

# 1 绪论

## 1.1 研究背景与意义

### 1.1.1 研究背景

随着互联网的迅速普及，多媒体处理技术，网络传输、视频数据处理等相关技术快速发展，来自不同领域的视频数据正在以惊人的速度增长。近几年随着新媒体技术的发展，一种新类型的视频—弹幕视频再世界范围内变得愈来愈流行，其评论内容直接显示在视频界面上，当某个视频有很多评论是，视频界面上就会产生如同无数子弹飞过的效果。目前国内最著名的量大弹幕视频网站分别是AcFun和bilibili，网友简称A站和B站。

### 1.1.2 研究意义

用户可以在观看网络视频中发表自己的实时评论，并与其他用户产生共鸣或者互动。在用户所发的弹幕中，往往包含着于视频之于当前片段的情感相吻合的文字，这对展开进一步的研究起到很好的帮助作用。作为一种新型的消息传递方式，弹幕评论与传统的短文本有着很大区别，主要表现在两个方面：1）弹幕内容与视频时间节点有着高度相关性，可以反映一定的剧情情感波动；2）内容口语化，符号化。用户往往通过当下流行的热门词语，通过隐喻的方式表达自己当前的情感。

## 1.2 国内外研究现状

### 1.2.1 文本情感分析研究

目前针对网络评论的情感分析应用研究多集中再网络舆论和商品评论两个领域。例如，利用情感分析技术对微博舆论中的热点话题、公众观点等进行提取，或者面对商品的用户在线评论进行褒贬识别。

根据情感分类的方法，情感分析在微博、网络商品评论领域的应用研究可分为两大类。

基于情感词的方法通常是现构建一个情感词典，词典中包含情感词所属的类别及对应的情感值，如正整数值表示正向情感，负整数值表示负向情感。Hu等对每一条用户使用反馈中各类别情感词的数量进行统计，在此基础上判断用户对产品的情感倾向。为了进一步量化用户评论的情感，Yu等把一条评论中所有的情感词的情感之平均数作为该条评论的情感值。Kim等考虑到否定词和双重否定对情感之的影响，采用成绩发判断某一条评论的极性。Yang等通过定义情感词权重词典、否定词词典、程度词词典及感叹词词典，综合考虑计算每条微博的情感指数。

机器学习方法主要是用用机器学习模型，对训练集的情感特征进行学习，提取系统输入输出之间依赖关系，从而应用于对测试集的分类判断。Pang等利用支持向量机，朴素贝叶斯、最大熵三种方法对电影评论进行分类，发现支持向量机的分类效果最好，而最大熵和朴素贝叶斯的分类效果相当。刘志明等研究发现，采用支持向量机的机器学习算法、信息增益的特征选取算法和TF-IDF的特征项权重计算方法，三者的结合对微博段文本的情感分类效果最好。Yu等则认为通过机器学习方法对新闻进行自动情感分类时，同时选取形容词、副词和名次作为情感特征项比选取单一词性的效果要好。

### 1.2.2 文本情感分析在弹幕领域的应用

目前，基于视频弹幕数据的弹幕分析研究工作相对较少。其中，有人采用数据可视化的形式将弹幕包含的情感值进行了分析和展示；其次，有人通过识别弹幕密集区域的情感波动，进行了精彩视频片段的推荐；以上两个研究工作都参照了传统的短文本情感分析手段进行研究，但是忽略了一个重要问题：弹幕视频内容口语化符号化，大量的通过单词或短语，借用隐喻的表达方式表达情感；而这些不包含主观情感词汇的客观词在传统的情感分析手段下，会被忽略，从而造成情感分析的不全面。

## 1.3 论文的主要研究内容

针对上述提出的问题，本文提出了一种基于情感分析的改进模型。利用弹幕的时间相关特性，计算出弹幕中的主观弹幕的情感值，映射给相关的客观词。通过海量的样本数据反复调整客观词的情感向量。在后续的情感分析工作中，进行过情感向量标注的客观词被加入到情感分析模型中，完善情感分析模型的完备性，提高情感分析的准确度。

结合客观语料的情感分析的目的是能够将提取出客观语料的情感倾向，并将这种倾向用向量的形式添加到情感分析模型当中。提取客观语料的情感倾向是本文的研究重点。针对这一目的，我们的工作任务如下。

1）获取有代表性的弹幕数据资源，并做基本的数据清洗工作。

2）识别出视频中的情节高潮，即弹幕数量异常密集的区域；并选取其中一段作为研究的语料来源。

3）针对弹幕数据进行主客观分类，其中主观句识别的关键是情感特征的识别。这一步的工作采用借助情感词典，对主观弹幕进行识别。

4）针对主观弹幕的情感向量评估方法，本文采用基于支持向量机的机器学习算法，借助开源分词工具进行分词和词性标注；利用多维情感分类的中文语料库训练模型。生成多维情感标注的情感向量。

5）客观弹幕中高频词提取，主要通过在客观弹幕中，分词统计出现频率；认为出现频率较高的单词，和主观弹幕的情感倾向相关度更高。

6）最后将主观弹幕提取出的情感向量和客观弹幕中的高频单词进行映射。

7）通过其他视频弹幕高潮片段的弹幕数据，重复如上的分析，并对客观词的情感向量进行调整。随后我们可以将这些高频客观词加入情感语料库中，提供给下一步的情感分析工作。

## 1.4 论文的组织结构

# 第二章 相关理论及技术

本章主要就少了本情感分析的基本流程，以及所设计的相关理论知识体系与相关技术。如文本与处理、文本表示、特征提取与选择、特征权重计算、情感词典、分类算法、分类起性能评估等理论基础。

## 2.1 情感分析流程

根据所使用技术的不同，情感分析的技术通常可以分为两类：一类是基于情感词典的额方法，其主要步骤是对文本进行文本与处理后，与以构建的情感词典进行对比，获取到相应的情感词，今儿计算该文本的额情感倾向。另一种是基于机器学习方法，即首先对语料库进行训练数据集和测试数据集的划分。经过文本预处理后，得到文本特征的矩阵化表示，用文本分类算法完成对文本情感的倾向判别。基于机器学习的情感分类的基本过程如图所示。基于机器学习的情感分析可以分为两部分：训练分类起和训练模型的测试。在训练分类起部分，首先进行文本预处理和特征选择以及训练集文本的表示，并且通过分类算法训练特征向量后，获得分类模型。在测试训练模型部分，通过测试集文本，特征选择，特征表示，然后在训练模型的分类上利用测试集对训练模型进行测试。因此，文本预处理，文本特征选择和表示，机器学习分类算法是文本情感分类研究中必不可少的部分，下面将对以上部分的理论和技术方法进行更详细的介绍。

## 2.2 文本情感分析常用技术

### 2.2.1 文本预处理

预处理，就是过滤文本中的不具备实际意义的词汇，得到人物模型所需的文本格式，对于机器来说，现在他只是一个不透明的字符串值。为了能够让机器识别，就需要经过句末检测、切词、词性标注等过程。接下来主要介绍情感分析任务中关于文本预处理的相关工作。

1）中文分词

分词是对文本处理和文本分类任务的第一步，中文分词与英文分词在分词过程中会有很大的不同，英文文本中英文单词间有空格对单词进行切分，所以在英文文本中空格起着切分英文词汇的作用。但是在中文分词中这个简单易行的优势将不会得到体现。因此在没有对中文文本进行相关处理时就应先对文本进行分词处理。

目前主要有三类基于分词的研究方法。

其中基于词典的字符串匹配的分词方法根据词典、顺序扫描以及匹配等展开研究，其优势在于实现相对容易，切通过匹配的方法能准确辨认出文本出现的词语，但是未登陆词的处理工作没有进行研究，并且词典之间犹豫标准的不同意会出现词汇歧义的问题。由于基于顺序扫描与匹配的过程，其匹配计算量较大。基于统计的分词算法主要根据相邻的字的共现频率，通过其共现次数确定其关联关系。崇勇的统计模型有互信息、N元文法模型等。基于统计的分词方法的优势主要体现在消除歧义和未登录词的识别两个方面。

在对中文分词的研究过程中，基于统计的中文分词其主要思路时词汇在上下文中，统计字与字之间的词语组合，两个字在同一文本中的共现频率高，其构成一个词的可能性就越大。基于统计的中文分词其优势主要体现在不受限于词典与文本的相关规则，但由于要用语料库对其进行模型构建与测试，其构建过程相对复杂，切分词的京都会收到训练文本等相关因素的影响。进行基于统计的中文分词方法常用的统计模型有：互信息、N元文法、HMM(隐马尔可夫链)、最大熵等。

本文将选用NodeJs中文分词中的Jieba分词开源库对文本进行中文分词。Jieba分词的基本原理主要是基于Trie树结构对文本进行扫描，并输出分局中词组成像一个的有向无环图结构，根据动态规划算法查找最大概率的路径。Jieba分词采用基于HMM模型使用Viterbi算法解决未登录词的识别问题。

2）词性标注

词性标注简称标注，是指对文本进行分词处理之后，对其相应的词汇进行词性的标注的过程，即标注分词后的词汇是动词、名次、形容词还是其他词性，然后将其词性（Part-of-speech，POS）信息分配给相应单词，并用标签进行标识。目前比较常用的词性标注算法如图。

基于规则的方法主要原理是在规范的规则下，对语料进行词性标注。其优势是标注过程相对简单且易懂，当爱方法需要人工执行消除器一规则，随着语言形式与新兴词的出现，人工制定的歧义消除规则不但任务繁重且难以保证规则之间是否存在歧义。所以方法在处理新兴词汇和网络词汇是较为困难。

基于统计的方法其主要原理是通过分类算法的方式来自行推衍文本中的信息与规律。其优势在于该方法主要根据对数据的统计，通过算法自行归纳规则，无需人工制定标注规则。但是由于是基于数据的统计，其总结推算出来的是一种基本似然规律，无法对其进行定性的总结归纳。由于语料库的不全面性会导致词性判断的片面性，因此要考虑由于稀疏数据造成过度拟合等问题。

根据规则与统计的方法的利弊，结合规则与语料库统计的方法形成基于规则与统计相结合的方法。其主要原理是对词性进行标注时，使用规则的方法对简单语句中的词汇进行词性标注，节省资源。若使用规则的方法处理较为困难时，再使用基于统计的方法来处理相应词性标注的问题。

在中文情感分析中，中文语句会出现很多不含邮情感色彩的虚词，为了使情感词典的构成尽量简化与计算量的减少，我们着重对语言特征与属性特征的词汇进行词性标注和收录。在词性标注中，标注的词性只具有参考作用，其并不具有绝对的准确性。对于上下文中标注的词性存在歧义问题，也有相关的学习方法对其研究提供了较为夯实的基础与清晰的思路，例如隐形马尔科夫模型、基于转换的学习方法。虽其取得了相对较好的研究效果，仍然还需要广大学者对其继续研究。

### 2.2.3 文本表示模型

文本主要是由字、词、标点符号以及相应的句式构成。为了确保文本分类的高效性与准确性，文本形式话表示就显得格外重要，因此要找到技能标识不同种类的文本并能准确的反映文本内容的方法是进行文本分析的必要步骤。

文本的形式化表现就是从文本中抽取能表示文本特征的信息，并通过规则对其加以形式化表现。目前通常的文本表示方法有向量空间模型（VSM）、语言模型、后缀树表示法以及本体表示法等。本文将采用向量空间模型作为文本的形式化表现方法。向量空间模型（Vector Space Model，VSM）的主要原理是将文本应设成一个向量的点，文本的每个特征由向量的每一维来表示。通过对文本进行分词和特征提取处理再进行计算。

空间向量模型的主要原理如下所述：

1）文档（Document）：文档是VSM中处理的对象单元。

2）特征项（features term）：在VSM中，最小不可分的单元就是特征项，特征项包括字、词、词组和短语等。

3）特征项的权重（feature weight）：每个文档中的特征项ti标识在文档中重要程度的值，即权重wi。通过VSM，就可以将一个吻昂表示为n维空间的一个空间向量。

4）空间向量模型：给定一个文档D(t1,w1;t2,w2;…;tn,wn)，D满足以下两条约定：其一，不考虑文档的内部结构；其二，文档中各个特征项没有重复。

特征项t1,t2,…,tn可以看作n维坐标系，对用的权重表示为：w1,w2,…,wn为坐标值，则文档表示为向量D(w1,w2,…,wn)

2.2.4 特征选择

？？？？

### 2.2.5 情感词典

情感词库的构建是内容情感分析任务的一个重要环节。Subisie等人(2001)利用手工建立了一个情感词典，并标明了情感词典的强度和向心度，即情感词表示情感的力度及所定义的情感类别之间的相关程度。而知网(Hownet)所提供的情感词典将情感词分为：“正面评价词语(Plus Feeling)”、“负面评价词语(Minus Feeling)”、“正面情感词语(Plus Sentiment)”、“负面情感词语(Minus Sentiment)”、“主张词语(Opinion)”、“程度级别词语(Degree)”六类不同的词来构建一个中国那你问的情感词典。台湾大学的研究组进一步针对中文的情感子或者词进行研究，并在此基础上提出了一个积极和消极两元分类的NTUSD情感词典。因此情感词典的构建本质就是对文本中存在的情感词进行识别、分类和标注的过程，在此基础上为更多情感词的识别与内容的情感分析奠定了基础。

而在实际内容情感分析的过程中，常将所构建的情感词典分为：基础情感词典、拓展情感词典及领域情感词典三种类型。其中，基础情感词典是利用已有的情感词典，根据不同情感类型进行消重合并后所形成的情感极性词典。

拓展情感词典则是在基础词典构建的基础上，根据不同的实际需要来对一构建的情感词典进行扩展。例如：哈工大自然语言研究小组利用中文同义词林中的同义词来扩展了基础情感词典。另外，WordNet-Affect(2004)情感词典也是针对WordNet词典中的同义词集合进行了情感的标注。

另外，由于不同应用领域的差异，相同的情感词在不同领域中表现出的情感倾向与急性也存在着较大差异，为了提供情感分析的准确率与质量，构建面向领域的情感词典则成为了实际应用分析的关键。例如：Turney(2002,2003)采用了在特定的语料库中，将新词与带有倾向的极性词进行相关性计算并获得新词的情感倾向评分。

## 2.3 机器学习分类算法与性能

### 2.3.1 贝叶斯

贝叶斯分类算法(B1,yes)是一种基于贝叶斯原理分类策略的监督学习算法,贝叶斯分类器定义如下:

设待分类项为x,预先设定类别集合C={y1,y2,…yn},计算p(y,lx), p(y2lx),p(ynlx)。当p(yklx)=max{p(y1lx),p(y2l,) ,pCy1,lx)}, x属于类别yk。

设x={x,, x2, \_,xm]l, xk表示x的特征属性。如果各预先设定的类别下各个特征属性的条件独立,.则:

式中分子根据条件概率估计计算:

代入公式(2.13),即可得x的类别。

贝叶斯分类算法有两种常用的模型, 分别为多项式模型(multinomial model)和伯努利模型(Bemou11i model)。

一篇文档是由许多单词组成,那么文档可以表示为: d=(t,,t2,…,tk),其中tk是在文档中-出现并且可以重复出现的单词,文档d由k个单词组成。在多项式模型中,先验概率P(c)的值可以表示为类别 c下单词总数与训练样本中所有单词个数的比值。 在类别c下单词t,,在所有文档中出现次数的总和与1的和与类别c下单词总数与lvl的和,两和的比值即为条件概率P(tklc)的值:

多项式模型和伯努利模型的区别:第一,两个模型的计算粒度不同,多项式模型的计算粒度是单词, 而伯努利模型的计算粒度是文件 。 第二, 后验概率的计算涉及的单词范围不同, 在后验概率计算过程时, 多项式模型中后验概率的计算只考虑文档中出现的单词, 伯努利模型中后验概率的计算考虑了所有的单词 。

### 2.3.2 支持向量机

用于分类的算法有很多,如:逻辑回归分类法、贝叶斯分类法等。它们的共同特征是分类误差最小化。 但该特征存在两个问题: 一是不能判别划分数据的可信程度; 二是可能得到的解有多个甚至无穷个, 不能判定所有解的优.劣 。 为解决上述问题就出现了支持向量机(SupportVlectorMachine, SVM) 。 SVM的核心思想是:使各点的间隔最大化。

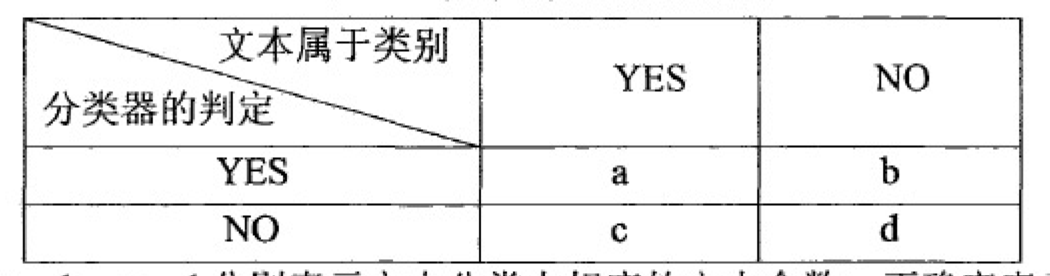
通过其数学公式推演将其转变为如下形式:

支持向量即只在虚线上的点, 支持向量机就是在寻找这些点到超平面的距离相等且最短的过程。

支持向量机和逻辑回归这两个算法在损失函数层面的区别是支持向量机采用的是hingeloss,逻辑回归采用的是logistica1 1oss。这西者都是0\_l损失函数的近似, 目的都是通过增大分类能力较大的数据点的权重, 减小分类能力较小的数据点的权重 。 支持向量机主要考虑支持向量, 即上述的那些区分能力最强的少数数据点 。 逻辑回归主要通过非线性映射函数来减小与分类平面距离较远的数据点的权重, 并增大分类能力较大的数据点的权重。两者的目的相同, 在很多实验中,这两种算法的结果也很接近的,所以本文进行了比较。逻辑回归相对来说实现简单,模型更好理解,而svM相比之下更加复杂 。 但是svM转化为对偶问题后, 进行分类时只需要计算与分类平面较近的少数几个支持向量, 在处理非线性可分问题时引入复杂核函数计算时, 能够简化模型和提高计算速度。

## 2.4 分类性能

### 2.4.1 准确率、召回率、F-测度值

在自然语言处理领域中，针对文本分类问题的评价指标，本文主要介绍准确率、召回率、宏平均、微平均。

上表中a,b,c,d分别表示文本分类中相应的文本个数，正确率表示正确被检索到的数目与实际被检索到的数目的比值，召回率表示正确被检索到的数目与应该被检索到的数目。

实验结果的评价指标应该考虑到多方面因素的影响，保证客观、公正的评价。一般的情感极性分类性能判定又准确率(Precision)、召回率(Recall)、F值三种评价指标。假定A代表分类算法能够将带测试文本准确的分道该极性类别的文本语料数量；B代表分类算法对文本错误的分道该极性类别的文本语料数量；C代表分类算法未能将文本准确的分道该极性类别的文本语料数量。

准确率，也叫查全率，记为P,顾名思义，该指标是指判定分类的准确性。准确率是判定该类别极性书与所有判定为该类别极性书的比值。其计算公式如下：

召回率，也叫查准率，记为R，该指标主要是为了判定分类模型的完备性，是正确判定为该类别极性书与本应该判定为该类别的比值。其计算公式如下：

F值记为F，由于实际测试中，当准确率高时召回率低，当召回率高时，准确率低。两者一定程度上呈相互制约的关系。为了寻求两者的平衡点引入了综合度量指标F值作为两者的调和平均数。其计算公式如下：

### 2.4.2 微平均和宏平均

文本分类后的各类文档都有相应的召回率和准确度。我们可以通过对每个类的分析来引申到对整个分类起的分析。将采用微观平均和宏观平均作为其相应的评价标准。宏平均评价指标是通过对每一类别求和后取平均的方式，得到其相应的平均查全率与正确率，进而得出整体的正确率和召回率。同理可知，微平均是对整个实例文档进行其相应的求和再平均方式。因此，宏平均是类的算术平均指标，而为平均是整个文档的算术平均指标。

## 2.5 本章小结

本章节首先介绍了文本情感分析的一般流程：文本预处理、特征表示、机器学习、文本分类等步骤。接着详细介绍了文本预处理的步骤中，中文分词和词性标注等。文本预处理后的文本表示模型将采用向量空间模型。2.4介绍了几种传统的特征提取方式，如文档频率、信息增益法等并分析了方法的利弊。2.5介绍了情感词典的现状及成果。2.6介绍了常用的分类方法同时引入了分类起性能的评估指标：正确率、召回率、F1-measure、微平均和宏平均。同时，正确率、召回率、和F1-measure也是本文所采用的分类评价标准。

第三章 基于情感词典和LDA的情感特征提取

基于词典的情感分析方法是当前处理文本情感分析比较常用且较成熟的方法,本章将视频的弹幕文本作为研究对象,将文本中的情感特征量化出来。人类的行为和情感是相辅相成的,通过情感特征量化可以探索用户群体行为具体是受哪些情感特征的影响, 其有助于对狂交网络用户行为分析的进一步工作的展开。

## 3.1 情感词典的构建

### 3.1.1通用基础词典

本文知网的"情感分析用语词集"和台湾大学的NTUSD词典以及正面词汇、负面词汇和程度副词组合为通用情感词典。知网发布的"情感分析用语词集"中有中文情感分析用词语集,并且台湾大学发布的NTUSD词典中包含2810个正极性词语和8276个负极性词语。

首先筛选并去除知网的"情感分析用语词集"和台湾大学NTUSD词典中存在许多情感极性相反的,或者情感意义比较多的词汇,然后将知网(Hownet)里面的正面评价词语、正面情感词语和NTUSD词典的positive词典进行词语去重,然后对其进行组合,成为通用基础词典中的积极情感词典。把知网的负面评价词语、负面情感词语和NTUSD词典的negative词典进行词语去重,然后对其进行组合,成为通用基础词典中的消极情感词典。

通过使用这种扩充策略,对HowNet和NTUSD词典进行相应的关系映射,将所得词汇加入扩充词典中。通过使用构建策略,能有效提高工作效率,并且能増加许多有效的词汇对后续部分的词典扩展也有很好的借鉴意义。最后再把通用情感词典中重复的词语剔除,成为最终的通用基础情感词典。

### 3.1.2扩充词典

由于网络新兴词的出现,所以在当下社交网络中进行情感分析时,我们要对网络口语化用语进行正确的评断。网络新词汇不同于传统用语等,比如"然并卵","香菇蓝瘦"等都带有浓烈的主观情感。

这些网络新兴词汇在词典中是没有的,但是对情感分析中情感极性的判断影响确非常大。因此,扩充网络特殊情感词典在基于词典的情感分析中甚至起到决定性的作用。本文所扩充的网络新兴情感词汇,主要来自于搜狗拼音词库网络流巧新词,几乎涵盖了在各大社交网站中各种网络流行词汇,共计23261个。为了提高其分析准确率,利用哈工大的《同义词词林扩展版》找到情感词的同义词,这样就拓展了基础情感词典。

## 3.2 基于LDA的文本主题模型的构造与改进

3.2.1潜在狄利克雷分配主题模型构造

潜在狄利克雷分配模型(LatentDiriletAllocation,LDA)是JordanMichael、吴恩达、David等人提出的一种主题模型,其本质是一个由文档、主题、词汇层组成的贝叶斯模型。2004年Griffi等人在Blei的基础模型上对参数施加Dirichlet先验分布,使得LDA成为一个完整的概率生成模型。LDA模型中在选定主题时引入了先验概率,为每一篇文档都设定了两个先验狄利克雷先验参数:和。这两个参数能为任一文档随机分配主题分布和词语分布,文档的主题分布可以产生新的文档：p()=Dirichlet() ;主题下词语的分布遵循狄利克雷先验分布：p()=Dirichlet()。 利用上述的参数,LDA模型就可以预测一篇新文档的出现概率。

在LDA模型下可通过生成模型来处理文档和主题。生成模型指的是文档中的每一个词语都是某一概率选中某个主题,又在该主题下—定概率选中了该词语"这样的方式生成的。文挡中的每一个词语会下的概率出现:



LDA模型生成文档的步骤如下:

1)使用LDA模型的参数进行取样得到第i篇文档的主题分布

2)从得到的多项主题分布取样得到组成文档i的第j个词所属的主题

3)使用LDA模型的参数进行取样得到主题的词语分布:

4)从得到的多项词语分布进行采样得到词语

LDA模型是一个三层的贝叶斯模型,该模型的二项分布的共轭先验概率分布是Beta分布;而多项分布的共轭先验概率分布则是狄利克雷(Dirichlet)分布。

本文所采用的采样方法-吉布斯采样法是马尔科夫链蒙特卡罗的一种简单实现方式。在使用吉布斯采样法对联合分布进行采样时,由于w是可见参数,只有z是潜在变量,那么最终要获取的采样分布就是。因此,LDA模型的吉布斯采样公式如下:



上式的右半部分实质上就是P (主题|文档)\*p (词语|文档),两个概率相乘所得的概率即为"文档->主题->词语"的路径概率。在整个LDA模型中,主题的总数是K个所以吉布斯采样公式(公式3.5)的实质就是在由K个主题构成的K条路径上实现采样。吉布斯采样法的采样路径如下图所示。



通过使用吉布斯采样法对LDA模型实现训练,目的就是获取给定预料的"主题-词语"样本,在获取样本之后,就可以对LDA模型中的词语分布,…。以及主题分布,…进行估计。对于一篇新的文档,通过训练得到的参数,我们就可以获得这篇文档主题分布。

训练步骤如下：

1)对语料进行初始化操作，为语料文档中的每一个词语随机分配一个主题编号z

2)对语料库进行扫描操作，按照吉布斯采样共识对每个词w进行采样获取其所属主题，讲获取的结果在语料中进行功更新

3)重复进行第二步的操作，当吉布斯采样收敛时停止

4)对语料库的主题-单词频率矩阵和文档-主题矩阵进行统计，得到这两个矩阵就是要获取的LDA模型

### 3.2.2改进LDA模型的文本主题分类

由于LDA主题模型在表示文本时使用是词袋模型,在处理词语的时候采用的是公平对待的方式而忽略了词语的权重。因此,LDA模型经过训练得到的主题分布一般会偏向高频词汇,针对词汇的幂率分布问题,本文采用高斯函数来实现对文本中词语的加权操作,目的是提高中频词语的权重w及降低高频词语的权重。实现加权操作的前提是:待处理文档中的所有词语都具有相同的权重;同时也要保证在同一主题下词语出现的频率越高(分布比例越大),其重要性也就越高。

在使用TF-IDF进行加权操作时,假设条件是:在一个文档中出现频率很高的词语,在同类别其它文档中出现的频率同样也很高,反么亦然。如果无法满足上述假设,加权效果将会大打折扣。诊断数据为例,由于每个词语在每个文档中的出现次数都是1,即TF=1,如果在这种情况下进行使用TF-圧)F方法进行加权,会产生词频和权重呈反比的现象,是不符合常识的。使用高斯公式对文档中的任一词语m进行加权的公式如公式3.6所示:上式中的方差公式如下：



其中的表示词语在文档中的词频数;表示在整个词频数集合当中词频数居中的词语i的词频数。为了保证数据集在加权前后总词数不发生变化，需采取以下方式:



经过加权进行吉布斯采样时,将文档d的一个词语w归为某一个主题j时,词语分布的值不是累加1,而是累加一个;同理主题分布累加的也是一个。

## 3.3 基于词典和LDA的文本情感值计算

基于词典和LDA文本分析方法的工作框架如图所示 。



基于词典和LDA的情感分析算法的描述如下所示:

输入:情感词典DIC,评论文本集D={d...,}

算法描述:

1)预处理D；

2)遍历情感词,记录积极情感词posWord和消极情感词negWord,以及各自相应的位置posWordPos、negWordPos；

3)在遍历到的情感词后向前查找程度副词,找到一个程度副词就停止搜寻。为程度副词设定情感权重乘情感权重；

4)在情感词前查找否定词notWord,遍历全部否定词notWord,若数量为奇数,乘-1,若为偶数,乘1；

5)判断分句结尾是否有感叹号,有感叹号则往前寻找情感词,有则相应的情感权值+2；

6)根据权重计算后得到的向量空间矩阵与改进后的LDA模型进行加权情感特征权值,计算出积极情感权值Spos\_i和消极情感权值Sneg\_i,用数组记录起来；

7)计算所有评论的情感值posW、negW；

8)计算每条评论的积极情感均值posAvg,消极情感均值negAvg,积极情感标准差posStd,消极情感标准差negStd；

9)将posW、negW、posAvg、negAvg、posStd、negStd添加到特征集合T；

输出:特征集合T

词典和LDA的弹幕文本情感值计算如式:



其中,为基于词典特征集合,为基于LDA主题模型的特征集合,表示两类特征的权重。

## 3.4实验结果及分析

### 3.4.1基于词典与LDA结合的情感分类实验

对弹幕文本进行情感倾向性分析过程中,在其对情感特征项进行提取时,采用专业领域情感词典情感词抽取与LDA主题模型联合表征的特征计算与表示。通过该方法进行情感极性计算,得出其情感权值、情感均值和情感标准差。基于词典和LDA组合分类实验评价指标数据,如表所示。

|  |  |  |  |
| --- | --- | --- | --- |
|  | 精度 | 召回率 | F1值 |
| 基于词典的  情感分析 | 0.8761 | 0.8791 | 0.8767 |
| 基于词典与LDA组合的情感分析 | 0.8925 | 0.8982 | 0.8910 |

由表得,基于词典与LDA组合方法的精度比基于词典的方法的精度高,基于词典方法的情感特征的精度达87.%%,基于词典与LDA组合方法的情感特征的精度达89.25%。

由图可知,基于词典与LDA组合方法的情感特征的分类效果的准确率要比基于词典方法的情感特征的准确率高。且其各项情感特征项的分类趋势与上述实验呈现正相关关系,证明了词典与LDA结合进行情感分析的有效性。文本的预处理工作主要是对初始弹幕文本进行分词和停用词处理,从弹幕文本集中去除噪声词汇造成的干扰。在文本预处理后,根据情感词典从文本中筛选出特征词,完成情感词抽取及特征选取,对所选特征进行权重计算得到文本向量空间矩阵,通过LDA主题模型的持征扩展得到经过扩展以后的向量空间矩阵。每一行是一个扩展后的特征向量的原始权重,通过公式情感值计算方式后,最终得到其积极消极情感权值,积极消极情感均值、标准差。

## 3.5本章小结

本章介绍了基于词典的网络新型文本的情感分析。首先,对于被清洗掉的短小的评论文本采用基于词典的分析方法。本文构建由基础词典、扩展词典的情感词典。通过对比实验可得情感词典与LDA主题模型进行结合后得到的情感特征的准确度更为精确。将得到的情感词典与LDA主题模型进行组合选取文本特征,并按照其相应的情感倾向性计算规则,得到情感权值、情感均值、情感标准差这兰种情感特征项。

# 第四章 基于弹幕文本的SVM情感分析算法研究

上一章主要介绍了本文涉及到的基本概念与关键技术。从本章开始，将介绍一种针对弹幕文本的情感分析方法。本章首先对研究的主要问题进行描述，指出现有web短文本情感分析方法的问题。然后描述了弹幕文本的挖掘过程，并就这些问题提出了一种基于SVM的高维混合特征弹幕文本情感分析算法，最后将本章提出的算法与常用SVM分析算法进行实验对比。

## 4.1 问题描述

目前主流的文本情感分析算法在对弹幕短文本进行分析时效果往往不够理想。这主要是由于弹幕这种形式的短文本和普通的网络评论文本有着显著的特征差异。这种差异主要体现在以下几个方面。

1)内容形式较短。

在弹幕这种基于web视频流媒体之上存在的新型评论文本形式中，单个文本的长度相较于传统的网络评论文本有显著的差异。弹幕的长度整体偏短，绝大部分弹幕长度在一句话以内。弹幕长度短会导致可提取有效信息减少，影响情感分析结果不够准确。

2)表达形式多样。

核心的特点就是表达形式多样，相对传统评论更加不受约束，这一点深受弹幕的主要受众们（主要涵盖了青少年和年轻人群）的喜爱。为了表述强烈的情感寓意，弹幕中会高频的出现字符表情，这种形式的目的是在表达中凸显俏皮和幽默的气氛，用来达到加强情绪表达和观点阐述的效果。还有很多弹幕为了简洁明快的表述观点，会用一些缩略用语来代替标准但冗长的原语句。这种表达方式在视频情节的一些高潮区间会大量出现，目的是表达出观众们对特定剧情的强烈态度和情绪。还有一种很重要的弹幕表达形式是“隐喻”，弹幕深度用户们喜欢用一些历史上的典故来来用讽刺的表达方式去帮助表达自己的观点和情绪，在社区中这种表达方式也叫“梗”。

3)术语的特定性。

弹幕在产生初期使用者的圈子局限于动漫爱好者，由此造成了会在弹幕中出现很多动漫圈特有的术语，如“卡哇伊”，“宅”，“前方高能”。之后在弹幕发展的过程里，也不断的产生着很多特有的术语。如“蓝瘦，香菇”，“膝盖中箭”等等。在分析过程中对专业术语的缺乏显然会导致情感分析结果不够准确。

4)倾向性特征词中包含大量的动词。

在弹幕文本中，用户往往会使用动词描述自己对当前视频情节的看法。按照传统的弹幕文本情感分析方法，评论文本的动词往往可以用来表达用户的情感极性，如“喜爱”、“加油”、“爆发”等词通常具有积极的含义；而诸如“丧气”、“抱怨”等词通常具有消积的含义。但是需要特别注意的是，弹幕文本的表达中，很多动词表达的情感倾向和传统含义截然相反。如“智商感人”等表述的非但不是表示积极倾向，反而是描述一种消极，失望的含义。传统的情感分析算法应用于弹幕文本情感倾向性分析，效果不好。通过结合弹幕术语的特定性，以及弹幕情感倾向的动词词性，建立弹幕倾向性的模式，可以比较准确的进行弹幕文本情感倾向性判定。

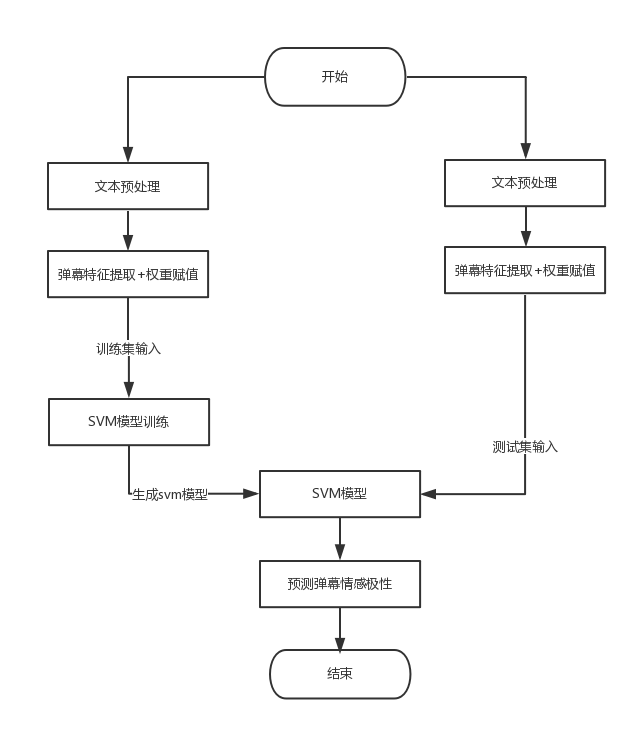
基于以上的弹幕文本的特点，本文提出了一种基于SVM的高维混合特征的弹幕情感分析算法。并随后以此模型为基础，构造了一个结合客观弹幕语料的SVM情感分析模型。

## 4.2 基于弹幕文本的SVM情感分析算法研究

采用传统的文本特征作为支持训练机的特征输入，生成的支持向量模型可以用来对普通web文本进行倾向性分类。但是针对弹幕文本有诸如文本较短，表达方式多样等特有的问题，传统的特征工程在这一场景下性能会大大下降。针对如上两种问题，我们在此章提出了两种改进基于弹幕的情感分析的算法。1）一种基于SVM的高维混合特征的弹幕文本情感倾向性分析算法；2）结合客观弹幕文本的机器学习情感分类模型。

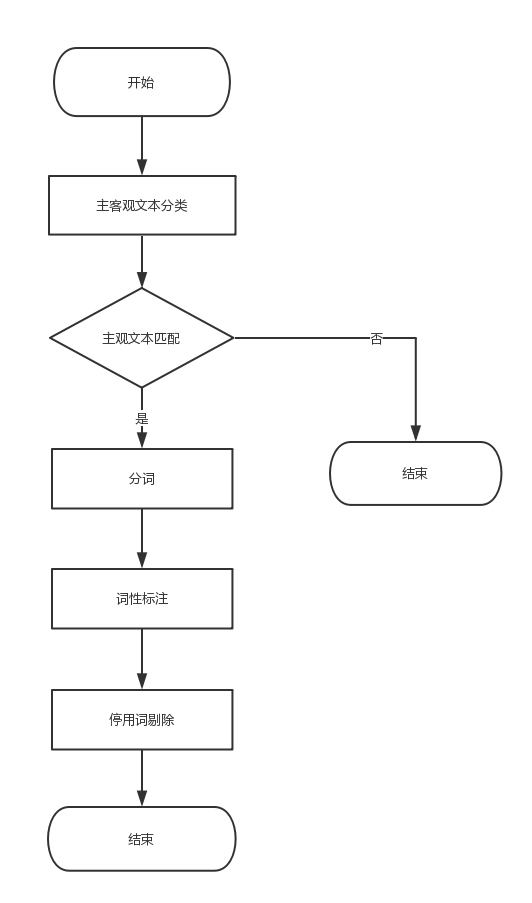
### 4.2.1一种基于SVM的高维混合特征的弹幕文本情感倾向性分析算法

基于SVM的高维混合特征的分类算法主要是先对待分析文本语料进行文本特征提取，通过这种方法生成一个混合特征的向量，然后将该向量作为SVM的输入并进行学习，在通过训练好的SVM分类器进行文本分类。该模型又文本预处理、词典语义分析、机器学习分析三个大的部分组成。第一部分主要是进文本预处理，使用Jieba汉语词法分析系统对所有文本进行中文分词、词性标注和去停用词操作。第二部分是提取弹幕文本中的多个特征向量，得到特征向量数组。第三部分是通过使用支持向量机的分类方法对文本情感极性进行分类。算法模型图如下图所示：



基于SVM的高维混合特征的情感倾向分析算法的设计思路分为如下三个步骤：

首先需要完成弹幕文本预处理。预处理即过滤文本中不具备实际意义的词汇，得到任务模型所需的文本格式。其中涉及到的具体任务包含中文分词、词性标注、停用词去除等。具体算法流程图如下：



第二步，需要对预处理完成的弹幕文本进行情感特征提取。本算法选定了经典的文本情感特征项，诸如情感词特征、否定词特征，词性标注特征，ngram特征，还新加入了词聚类特征，希望针对弹幕这种短文本特征稀疏的问题加以改进。

第三步，进行SVM高维混合特征情感分类器的训练工作。通过将上一步中生成的弹幕文本特征向量作为svm的输入，训练并生成支持向量模型。接着使用第一步中生成的测试数据输入模型，进行验证测试。通过调整特征维数对情感分类问题的影响，以期得出最佳的特征维数。由于位数太低会造成区分能力差，不能够准确分类等问题；而维数太高又会出现数据系数、过拟合等问题，继而降低分类效率。因此，关于特征位数的选择对情感分析的分类问题也同样起着至关重要的作用。

与此同时，针对弹幕短文本提出的词聚类特征，是否选择此特征对模型预测准确度的影响需要通过实验得到验证。

利用训练集和测试集得到最佳的分类算法和特征维数后，还需要对语料库中的测试剂进行实证测试。将测试后的得到的分类预测结果和人工标注的标签进行对比，得到基于整个语料范围中，在最佳特征位数的条件下分类准确度最高的分类算法。也需要证明本文提出的词聚类特征针对弹幕短文本进行情感预测是否有改进效果。

以下针对基于SVM的高维混合特征的弹幕情感分析算法做以小结。

本章算法充分考虑弹幕短文本的特点，从多位混合特征的角度进行文本的特征抽取，做到尽可能兼顾语义和情感，并且取得了较好的实验效果，用实验的方法验证了本章方法的有效性和鲁棒性。

本章算法主要的贡献有：提出了基于SVM的高维混合特征框架，采用了正则化的手段。考虑到语义对短文本情感分类的正确率影响，将词聚类加入到情感分析的特征，提高了一定的准确率。

虽然研究取得一定的改进，但也存在不足之处：1.弹幕标注数据匮乏，对分类模型的准确度有一定的影响。2.弹幕文本的客观语料非但不是噪声数据，还包含了重要的情感特征，在该算法中被剔除未加以利用。接下来的工作将着手解决上述问题的不足。

### 4.2.2结合客观弹幕文本的机器学习情感分类模型

本章在前一章基于svm的弹幕文本情感分类模型的基础上，建立了一个结合客观语料进行情感分析的模型。在上一章里，我们首先对弹幕语料做文本预处理，将一般视为文本噪声的客观语料剔除。只利用主观语料作为随后训练模型的输入数据。针对弹幕这种特殊的web短文本，它的特点决定了它有一大部分信息是基于客观语料，通过“隐喻”的表达手法传送的。本章针对这一实际问题提出了结合客观语料的弹幕情感分析模型，。最后，通过实证分析，给出改进模型的评价指标。分别从平均绝对误差，平均相对误差以及趋势正确率三个方面评价一个模型的优劣程度。

以下我们针对弹幕这种特殊文本的一些典型特质分析归纳如下。

本文已经在第一章对弹幕这种特殊的短文本形式的特点做了详细的分析。弹幕中超过一半的语料都是客观语料，用户善于用“隐喻”的方式去基于客观语料去表达自己的情感和观点。

这种表达方式可以分为两种，详细介绍如下。

1）象形文字/符号。

例子1: “orz” ，形状像一个面朝左跪在地上的小人，表示强烈的悲伤的氛围。

例子2: “23333”，意思是“啊哈哈哈哈”，模拟大笑的声音，表示开心的心情。

2）借用一些典故的用语反讽，语义上的情感倾向和实际倾向恰恰相反。

例子1: “xx感人”，语义上是说xx让人感动，而实际的用意则是负面态度，暗指xx这件事情让人不够满意。

例子2: “我从未见过如此厚颜无耻之人”，原文是来自三国演义中王朗的台词，原为贬义；后在弹幕中则是褒义，形容某人做某件事情的行为方式很巧妙，这是一种亲昵的表达方式。

3）还有一些出现在弹幕中的短语模式，表达出的情感倾向需要根据具体上下文才能确定。

例子1: “膝盖”，当弹幕中出现“跪了，把我的膝盖送给你”时表示敬佩的意思，表示褒义；当弹幕出现“膝盖中箭”时，表示的是很受伤的负面情绪。这种一词双意的状况也是弹幕文本情感一种很特别的表达形式。

4）一部分弹幕是一些常用弹幕文本的缩略语。

例子1:“颜表立”，全文是“用颜色表达立场”，是动漫视频中常见的剧情内容；

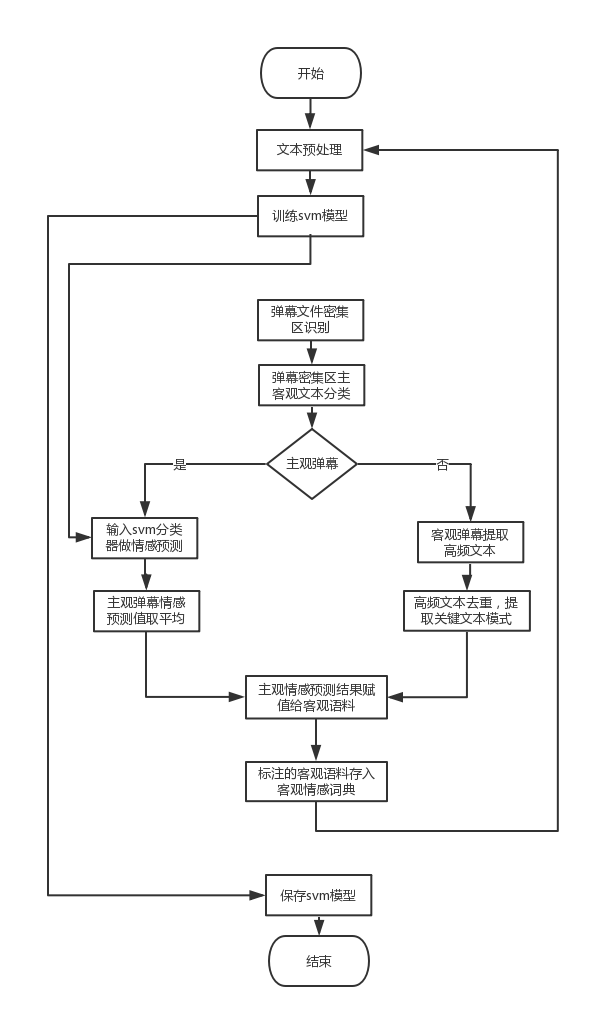
例子2:“火钳刘明”，全文是“这个视频火之前，我在这里留一个名字”；这里不仅缩略用语，而且还有使用同音词代替。

如上可以发现，各种各样的客观语料组成了弹幕这种新兴的网络文本载体。这些文本包含了及其大量且重要的情感倾向信息，我们在对弹幕的情感分析过程之中非但不能将其视为噪声忽略，还应该着重加以利用，来改进我们的弹幕情感分析模型的预测准确率。

以下我们介绍结合客观弹幕文本的机器学习情感分类模型的基本架构，具体描述如下。

针对弹幕文本的情感分析所基于的情感词典，目前使用的是研究领域内较为认可的Hownet和NTUSD两个情感词典的整合生成的情感词典。它的优点是较为丰富的包含了主要的情感词汇，能有效帮助进行情感语义信息提取；不足是针对客观语料表达为主的弹幕文本，这样的情感词典便无法再帮助进行情感语义的抽取。我们需要建立一个基于客观语料生成的情感语料库，将它和我们的情感词典相结合去进行情感分类模型的后续工作。

关于这个基于客观语料库如何去生成，我们进行了建模，模型流程图如下图：



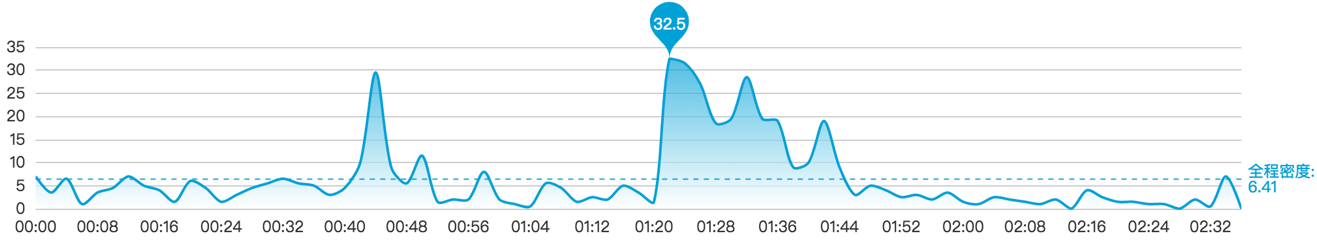
如图可见，我们的模型建立可以分为如下这几个模块。

1）初始svm模型生成

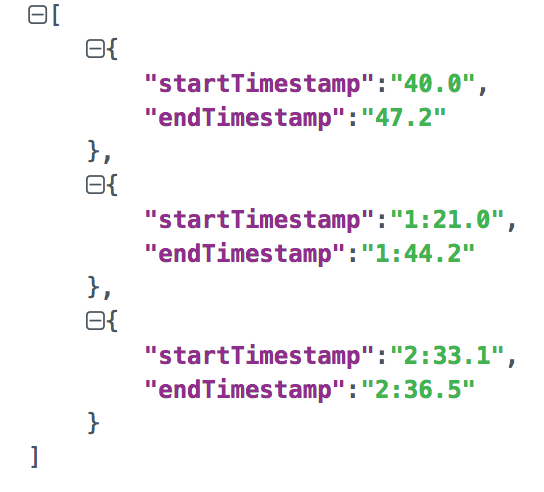
这一模块是作为整个模型的前置模块，基本流程是基于上一章中我们建立的针对弹幕短文本的svm情感分析模型的训练过程去执行的。目的是根据弹幕主观语料先初始化一个基础版本的svm情感分析模型。

2）弹幕文件密集区域识别

这一部分是整个模型核心的一个模块。我们希望通过分析弹幕文件中，每一个时间点的弹幕密度，识别在当前弹幕文件中弹幕密集区域的区间集合。



上图是一段23分钟的视频弹幕在时间轴上的密度分布折线图；通过密集区域识别，我们获取到的密集区域数组如下：



并依次对每一个密集区间的弹幕进行主客观分类。具体概念如图：



3）对主观语料情感倾向分类

基于上一章训练好的初始svm弹幕情感分类器之上，我们将当前密集弹幕时间区间的主观弹幕部分输入分类器，并获取主观弹幕的平均情感倾向。输出的结果和对应密集弹幕时间区间相关联，并存入文件中。

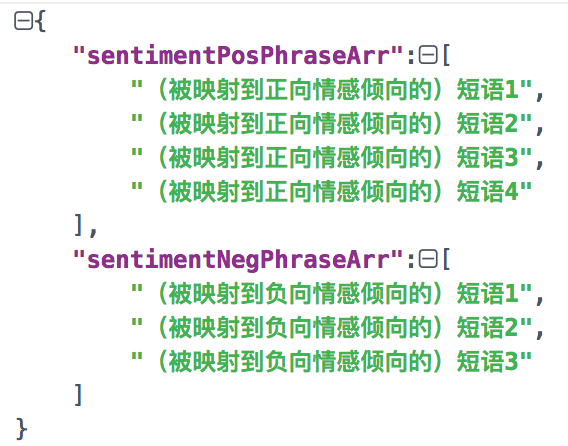
4）对客观语料提取高频短语模式

当前密集弹幕时间区间中的客观语料部分，我们希望提取其中的高频出现的文本模式，并对其进行情感赋值。我们首先对客观语料按照出现频率由高到低进行排序；针对相似的弹幕语料进行关键文本模式的提取；去除重复的语料文本；最后提取排序靠前的短语作为结果输出，存入文件。

5）对前两步产生的情感分类结果和短语模式映射

按照对应关系，将同属一一个密集弹幕区间的主观语料情感值和客观语料高频短语模式进行映射，将映射结果存入新建的客观情感语料库。

具体数据格式如下图：



6）结合客观情感语料库重新训练svm模型

如题所述，我们结合新生成的客观情感语料库重新训练针对弹幕文本的svm情感分类器。与初始化svm分类器时不同的是，我们在文本预处理阶段，不再将客观弹幕作为噪声数据剔除；而是将客观弹幕结合客观情感语料库进行情感分析，作为训练svm模型时一个重要的特征向量。重新训练svm模型，供随后对弹幕进行情感分类使用。

我们这样做的初衷是，弹幕文本密集的区域往往是由于对应时间点的视频存在一些剧情的高潮，引发人们通过发弹幕的方式去进行情感共鸣的回馈。所以我们有理由认为，在这些时间区间中的弹幕大概率是针对相同的目标事件才产生的。所以我们可以尝试对该区间内的含有情感的主观语料进行情感预测，并将情感预测值赋给遇见内高频出现的客观短语（长短不定的文本模式），并保存在我们新建立的客观情感语料库。

关于该模型的实验参比模型，为了验证结合客观语料的svm情感分类器模型的预测效果是否有一定的提升，我们选用上一章中基于六种情感特征的不包含客观语料的svm情感分类器模型作为参比模型。

## 4.3 实验结果与分析

### 4.3.1基于SVM的高维混合特征弹幕情感分类实验

本章的实验主要针对的问题是，基于SVM的高维混合特征的弹幕文本情感倾向性分析算法与基于传统特征提取的SVM算法的优劣比较。

此次试验准备基于情感词特征、程度词特征、否定词特征，标点符号特征，n-gram特征这六种基础特征，以情感词特征作为基本特征，分别加入其他五种特征进行组合和试验。

为了研究基于SVM的高维混合特征的弹幕文本倾向性分析算法在弹幕领域内是否对最终的分析结果有所提升，本文选取了1500条弹幕文本作为样本进行测试。其中标注情感正向450条，情感负向375条，情感客观675条。

其实验结果如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 特征组合 | 查准率 | 查全率 | F值 |
| 情感词 | 82.1% | 69.2% | 73.1% |
| 情感词+否定词+  程度词+标点符号 | 88.1% | 69.9% | 77.2% |
| 情感词+否定词+  程度词+标点符号  +n-gram | 91.2% | 71.5% | 80.2% |
| 情感词+否定词+  程度词+标点符号  +n-gram+词聚类特征 | 91.9% | 75.2% | 83.2% |

将实验数据用柱状图表示如下图 所示：

从柱状图可以发现，本文改进的算法当特征组合更丰富的时候，查准率和查全率上都有着显著的提升。

### 4.3.2结合客观语料的弹幕情感分类对比实验

此次试验准备基于上一章训练好的svm弹幕情感分类器模型为基础，为了研究模型预处理阶段是否考虑对客观语料做情感分析这一问题上，是否对最终的分析结果有所提升。

本文选取了1500条弹幕文本作为样本进行测试。其中标注情感正向450条，情感负向375条，情感客观675条。

其实验结果如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 模型选择 | 查准率 | 查全率 | F值 |
| 基础svm情感分类模型 | 91.9% | 75.2% | 83.2% |
| 结合客观语料的  基础svm情感分类模型 | 93.1% | 81.0% | 88.7% |

将实验数据用柱状图表示如下图 所示：

## 4.4 本章小结

# 第五章 基于弹幕文本的情感分析系统实现与测试

## 5.1 基本流程

## 5.2 算法设计

# 第六章 结论与展望

## 6.1 结论

## 6.2 展望