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

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。

2.2.1.1 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，如图所示。

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。

2.2.1.2 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 文本特征与权重计算

2.2.4.1 文本特征选择

良好的特征应该具有四大特征：可区分性、可靠性、独立性以及特征词的数量精简。通常文本的初始特征是由文本分词后所有的词语构成，这会造成文本特征空间的爱高而使文本向量稀疏，对计算速度和分类准确度造成很大影响，因此给文本特征空间进行将为是必要的。目前常见的特征选择方法有文档频率、信息增益、互信息、CHI统计等。

1. 情感词

在对文本情感分类时，往往文本中含有的少数带有情感倾向的词汇最直接表现文本情感的倾向性。如正向词汇“高兴”和负向词汇“难过”。由于中文词汇的复杂，情感词汇非常丰富，多位形容词、副词等。本文选择四个情感词典进行情感特征选择。其中包含整理好的Hownet、NTUSD、大连理工大学的本体词汇。其中由于前两者Hownet、NTUSD词典并没有标注情感词的情感极性。所以我们将正向词汇的得分设定为1.0，负向词汇的得分为-1.0。在情感词典特征上我们采用下面四个规则进行情感分数的计算：

规则1 分别计算情感文本中的正向词、负向词的数量；

规则2 分别计算情感文本中的正向词、负向词的得分总数；

规则3 分别计算情感文本中的得分最大正向词、负向词的分值。

规则4 分别计算情感文本中的最后一个情感词的分值。

1. 否定词特征

含有主观倾向的语句往往有很明显的否定词。与传统文本情感分类不同，“不”“没”等否定词不再作为停顿词被删除。在句子里“不”或“没”的否定范围时“不”或“没”的全部词。一个词在不在否定范围内对正确情感分类产生很大影响。例如：他一直没上班/他没一直上班；你没天天学习/你天天没学习。本文采用否定特征是以句子出现否定词为否定特征的开始直至句子结束都加上否定标记。并且记录否定词的个数也作为否定特征的一部分。

1. 程度词特征

含有主观倾向的语句不仅含有情感词汇，而且情感倾向的强烈程度往往对情感分类有很重要的贡献。针对这个特点，文本中的程度副词

1. 标点符号特征

？？？？？。

1. n-gram特征

对于给定的文本都可以看作是长度不同序列的集合。在这些序列中相邻的N个字或词成为n-gram，n-gram算法的基本思想是通过一个大小为N的滑动窗口将文本内容进行切分，形成长度为N的片段序列，每个片段序列成为gram。使用n-gram特征，尽可能获取有限长度短文本的未登录情感词和情感信息。

传统分词技术对于短文本的分词存在的明显缺陷甚至可能会改变原有评价对象。本文鉴于此类情况增加n-gram特征：对于1-gram是单个的字或词对于特征的选择并没有多大的意义，所以本文选择从2-gram开始，但超过4-gram同样没什么意义。

1. 词聚类特征

词向量具有良好的语义特性，是表示词语特征的常用方式。考虑到很多字、短语表现形式不同，但是表达的意义相同。采用词聚类的方法能够有效的避免处理相近的词语多余的计算量。Word2vec是Mikolov等所提出模型的一个实现，可以用来快速有效的训练词向量。可以吧对文本内容的处理简化为向量空间中的向量运算，计算出向量空间上的抗四度，来表示文本语义上的相似度。

word2vec包含了cbow和skip-gram两种训练模型。本文在skip-gram模型完成了词聚类构建了词聚类词典。在这个模型中，给出了语料库的词w及其上下文c，在给定条件概率和语料库Text，去求参数使得属于语料库的概率最大化：

在这里C(w)代表一组单词w的上下文。本文使用该工具运用在收集到的语料库上聚100类后得到1533个基元。

提取完如上特征之后，还需要对特征依次赋权值。

2.2.4.2 特征权重计算

在向量空间模型表示文本中，我们首先会根据各个特征项对表示文本的贡献度以及对文本分类能力进行特征选择，而我们会使用特征权重来量化这种分类能力。特征甲醛就是赋予各个被选择出的特横的权值的过程，其常用的方法有以下三种：

1. 布尔函数

布尔函数的数学表达式如公式：

为文本中的第i个特征项的权值，表示文本中的第i个特征项在该文本中出现的次数，即词频。从公式可以看出，该方法不考虑特征项出现的次数，仅仅简单笼统的根据出现与否特征性赋予1或0的权重，因此其忽视了词频对文本分类的影响，未能很好的体现各个特征项的区分性。

1. 开根号函数

开根号函数方法就是将特征项的词频的算术平方根作为该特征词的权重，其公式如：

该公式各项的表示同前面的公式，虽然该方法较布尔函数的方法有了改进，不是笼统的0，1权重，但是该方法得到的权值是随着特征项的词频增加呈单调增加趋势的，这违背了特征选择的原则。开根号函数方法在表示特征对文本的贡献度的效果方面也不是很好。

1. TF-IDF方法

TF-IDF方法实现技术简单且高效，因而其得到了广泛的应用，计算公式如下所示：

TF称为词频，是特征项t在文本中出现的频率，IDF称为逆向文档频率，表示含有特征项t的文本在整个文本集中的分布情况。这种方法既考虑了特征项的词频，又考虑了该特征项的反文档频率，TF表示的是该特征项在单个文本中出现的频率，IDF则表示的是该特征项在整个文档集中的出现频率。TF说明了特征项对该文本内容的表示能力，IDF则说明特征项对文本分类的贡献能力。

本文根据特征不同选取不同的特征权值，如否定词特征，程度词特征选取布尔函数赋值，情感词特征、词聚类特征等选取TF-IDF函数赋值。

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

2.6.1 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 。

2.6.2 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。

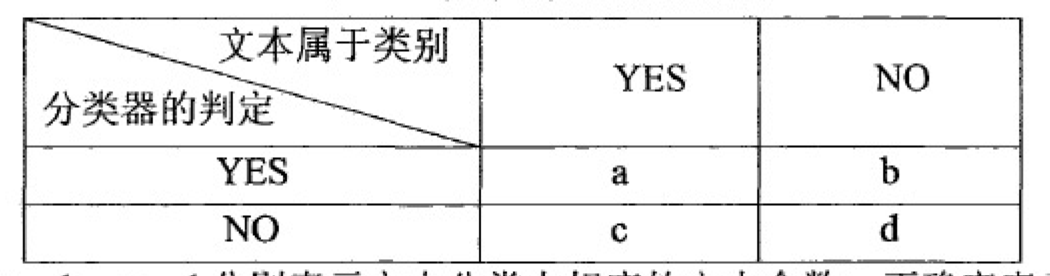
通过其数学公式推演将其转变为如下形式:

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。

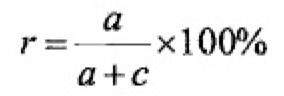
支持向量机和逻辑回归这两个算法在损失函数层面的区别是支持向量机采用的是hingeloss,逻辑回归采用的是logistica1 1oss。这西者都是0\_l损失函数的近似, 目的都是通过增大分类能力较大的数据点的权重, 减小分类能力较小的数据点的权重 。 支持向量机主要考虑支持向量, 即上述的那些区分能力最强的少数数据点 。 逻辑回归主要通过非线性映射函数来减小与分类平面距离较远的数据点的权重, 并增大分类能力较大的数据点的权重。两者的目的相同, 在很多实验中,这两种算法的结果也很接近的,所以本文进行了比较。逻辑回归相对来说实现简单,模型更好理解,而svM相比之下更加复杂 。 但是svM转化为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

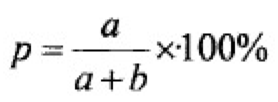
2.7 分类性能

2.7.1 正确率、召回率、F-测度值

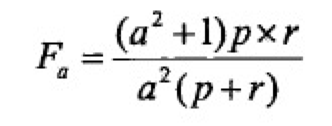
在自然语言处理领域中，针对文本分类问题的评价指标，本文主要介绍正确率、召回率、宏平均、微平均。

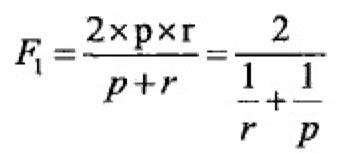
上表中a,b,c,d分别表示文本分类中相应的文本个数，正确率表示正确被检索到的数目与实际被检索到的数目的比值，召回率表示正确被检索到的数目与应该被检索到的数目。

正确率、召回率具体公式如下表示：

召回率：

正确率：

准确度和召回率可以合并成另一个度量F1度量。原则上，F1度量表示召回率和准确度的调和均值，F1度量认为正确率和召回率的重要程度一样，今儿同意这两个量得到一个衡量标准。而根据正确率和召回率的权重不同，可得到不同Fa-measure来评价性能，计算公式如下：

当a为1时，得F1度量的计算公式：

2.7.2 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嘞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。

2.8 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SVM的高维混合特征弹幕文本情感分析算法**

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。

3.1 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。

1. 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”。

1. 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。

1. 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。

3.2 基于SVM的高维混合特征的弹幕情感分析算法

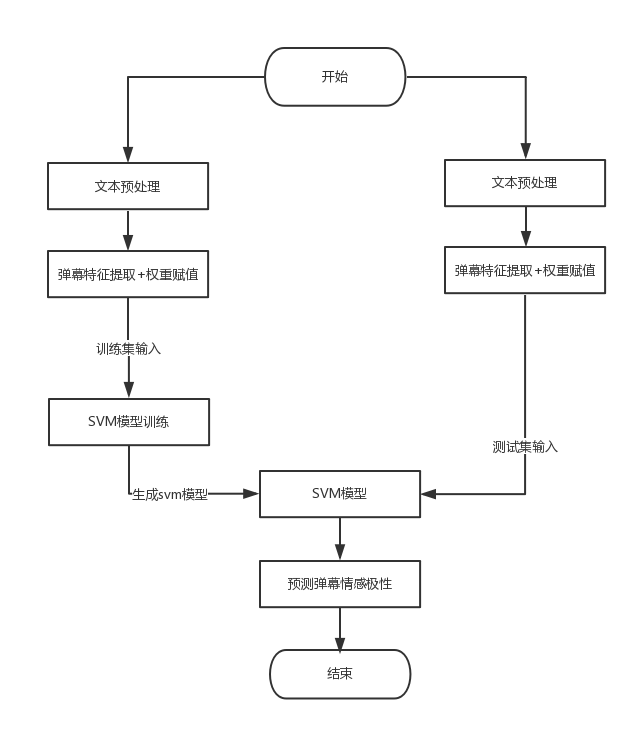
采用传统的文本特征作为支持训练机的特征输入，生成的支持向量模型可以用来对普通web文本进行倾向性分类。但是针对弹幕文本有诸如文本较短，表达方式多样等特有的问题，传统的特征工程在这一场景下性能会大大下降。在这种情况下，我们提出一种基于SVM的高维混合特征的弹幕文本情感倾向性分析算法，可以针对弹幕短文本进行有效的倾向性预测。

3.2.1 算法概述

由于该算法是基于SVM的高维混合特征的情感极性分析方法，所以对于文本的情感分析将会采用使用机器学习的文本分类方法。采用支持向量机分类模型，使用训练文集进行向量空间模型表示、特征提取得到特征向量集，然后对SVM分类器进行训练。最后通过得到的SVM训练器对文本进行分类，得到整个文本情感分类方法的结果。

3.2.2 算法模型

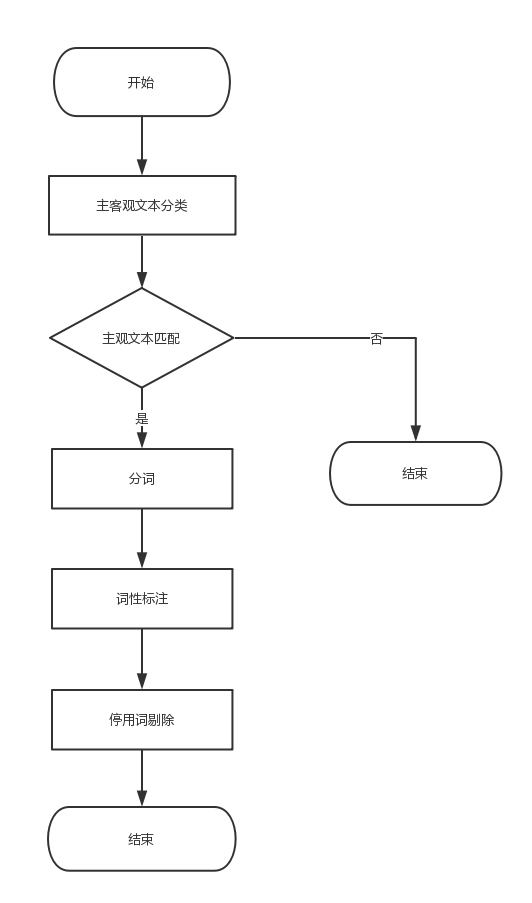
基于SVM的高维混合特征的分类算法主要是先对待分析文本语料进行文本特征提取，通过这种方法生成一个混合特征的向量，然后将该向量作为SVM的输入并进行学习，在通过训练好的SVM分类器进行文本分类。该模型又文本预处理、词典语义分析、机器学习分析三个大的部分组成。第一部分主要是进文本预处理，使用Jieba汉语词法分析系统对所有文本进行中文分词、词性标注和去停用词操作。第二部分是提取弹幕文本中的多个特征向量，得到特征向量数组。第三部分是通过使用支持向量机的分类方法对文本情感极性进行分类。算法模型图如下图所示：



3.2.3 算法设计

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。具体特征项描述如下：

？？？

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。

3.3 实验结果与分析

由于目前针对网络弹幕文本的语料还没有一个公开标注的数据集可以直接使用，因此本文选用了来自bilibili.com,acfun.com,iqiyi.com,youku.com等热门弹幕视频网站的弹幕文件，作为本次试验的语料。为了保证实验的数据集一致，本文将主要就提出的改进算法与改进钱的算法进行对比。

3.3.1 实验性能指标

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。嘉定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：

3.3.2 实验设计与结果分析

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。

1. 几种特征组合的弹幕情感分类实验

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特征这五种基础特征，以情感词特征作为基本特征，分别加入其他四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 | 77.80% | 81.20% | 80.10% |
| 情感词+否定词+  程度词+标点符号 | 82.10% | 84.30% | 80.20% |
| 情感词+否定词+  程度词+标点符号  +n-gram | 85.30% | 87.00% | 85.90% |

将实验数据永柱状图表示如下图 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。

1. 几种特征组合的弹幕情感分类实验

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特征这五种基础特征，额外加入针对短文本添加的词聚类特征进行对比试验。

其实验结果如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 特征组合 | 查准率 | 查全率 | F值 |
| 五种基本特征组合 | 85.30% | 87.00% | 85.90% |
| 五种基本特征组合  +词聚类特征 | 88.00% | 87.50% | 86.10% |

将实验数据永柱状图表示如下图 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。

3.4 本章小结

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 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。

4.1 问题描述

4.2**结合客观弹幕文本的机器学习情感分类模型**

4.2.1 ？？

4.2.2结合客观弹幕文本的机器学习情感分类模型

4.2.3 参比模型

4.2.4 模型评价标准

4.3 实验结果及分析

4.3.1 数据来源标准化

4.3.2 弹幕文本分析

4.3.3 样本选取对模型预测结果的影响分析

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

1.1 情感分析在网络评论领域的应用现状

随着网络视频资源的日渐丰富，网络视频用户的数量也在不断增加。用户观看视频后在视频网站、影视论坛等留下的影评能够一定成熟上影响其他用户的观看选择。一般的评论界面大多位于视频界面的下方，用户一般是在观看前或者观看后浏览评论区。与传统评论模式不同，弹幕评论的文本可以直接显示在视频界面内，即用户可以在观看视频的过程中即使分享自己的感受、观点。包含大量个人观点的弹幕评论作为弹幕视频的重要组成部分，也成为网络视频用户对视频进行二次编辑的一种方式。

目前针对网络视频弹幕的研究多集中在文化传播领域。弹幕视频的流行是“吐槽”文化发展的结果和动力。社交平台是的人与人之间都能形成交流、互动，而各种影视作品、原创视频更是经常成为讨论的热点。为了满足观众的吐槽欲望，弹幕视频营运热盛，他是观看视频和评论视频同时进行，观众可以最大限度的发表对视频的吐槽和批评，形成现场直播式的互动。随着弹幕功能的流行，弹幕数据中的情感倾向和观点会更具普遍性和参考性，这些情感信息可以为用户在选择视频时提供参考，满足更多元、个性化的视频检索需求。

尽管已有研究针对弹幕进行情感分析，但是发现参照的研究方法是传统的情感分析手段，刻意剔除了客观语料，仅提取了主观语料进行情感值计算。原因是认为客观语料不包含有带有情感成分的语料。

但是在当今web社交媒介中，每天都会有很多的网络新词出现，网友们尤其喜欢用这种客观语料性质的网络新词或短语进行交流，通过“隐喻”的方式表达自己的观点和抒发情感。这种新型表达方式目前非常流行，尤其在弹幕这种媒介中特别流行。那这里所谓的客观语料其实可能包含大量的情感成分，不应再被剔除出去，反而应该被做单独的情感分析，整合进入完整的弹幕情感分析模型中。

综上所述，基于结合客观语料的弹幕情感分析具备一定的应用需求，弹幕数据即使很有价值的情感分析数据来源。

1.2 情感分析在视频评论领域的应用需求

1.3 结合客观语料的弹幕情感分析的技术思路

1.4 情感语料库

本文语料库使用的是大连理工大学信息检索研究室开源发布的情感词汇本体库，该语料已经广泛用于中文情感的研究。

该语料库的情感分类体系是在国外比较有影响的Ekman的6大类情感分类体系的基础上构建的，加入情感类别“好”对褒义情感进行了更细致的划分。最终词汇本体中的情感共分为7大类21小类。共含有情感词共计27466个，情感强度分为5个等级。

中文情感词汇本体比较适用于解决多类别情感分类的问题。