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）基于文档频率(TF-IDF)的特征选择

TF-IDF（词频-逆向文件频率）是一种基于统计的数学方法，其可以判断出一个词语或者短语再文档或者文档集中的重要程度。

TF-IDF的主要思想是：如果一个词在一篇文档中出现的频率越高，且出现的范围在整个文档中较为集中，则他在文档分类方面的额作用越明显。TF-IDF的计算公式：

其中，W(t,d)为特征项t的权重，tf(t,d)为特征项t的词频，N全部训练文档数量，nt为训练集中出现特征项t的文档的数量，分母为归一化因子。

若候选关键词t的TF-IDF特征值越高，则此候选关键词成为该文档的一个关键词的可能性越大，反之，此候选关键词成为该文档的关键词的可能性越小。

（2）信息增益（Information Gain）

信息增益（IG）对特征项ti判定的重要程度的主要评价指标是其对此分类所提供信息量。信息增益的具体表示如下：

其中，P(Cj)表示所属Cj类型文本出现的概率，P(ti)代表特征项ti出现概率，P(Cj|ti)表示包含特征项ti时属于Cj类的概率，M为类别数。信息增益表示的特征具有较好的可区分性，并且可以降低特征空间的维数。

(3)基于互信息(Mutual Information, MI)的特征选择

MI(互信息)是信息论这门学科中检测两个随机信号之间的依赖程度，特征项t和类别C的互信息就代表了他们的关联程度。具体gongshi

2.2.4.2 特征权重计算

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。

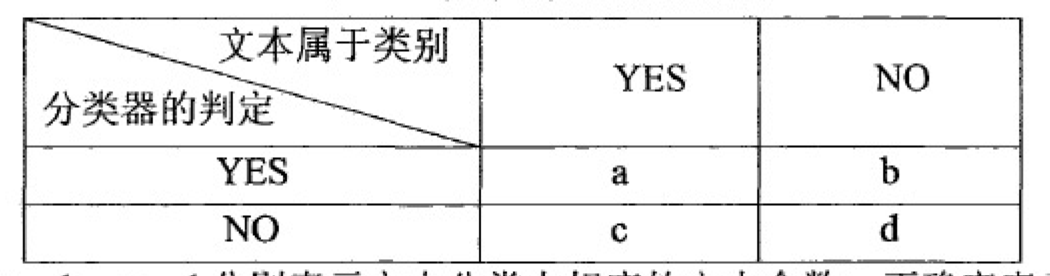
通过其数学公式推演将其转变为如下形式:

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。

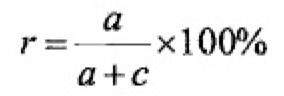
支持向量机和逻辑回归这两个算法在损失函数层面的区别是支持向量机采用的是hingeloss,逻辑回归采用的是logistica1 1oss。这西者都是0\_l损失函数的近似, 目的都是通过增大分类能力较大的数据点的权重, 减小分类能力较小的数据点的权重 。 支持向量机主要考虑支持向量, 即上述的那些区分能力最强的少数数据点 。 逻辑回归主要通过非线性映射函数来减小与分类平面距离较远的数据点的权重, 并增大分类能力较大的数据点的权重。两者的目的相同, 在很多实验中,这两种算法的结果也很接近的,所以本文进行了比较。逻辑回归相对来说实现简单,模型更好理解,而svM相比之下更加复杂 。 但是svM转化为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

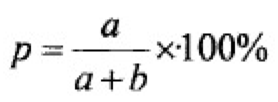
2.7 分类性能

2.7.1 正确率、召回率、F-测度值

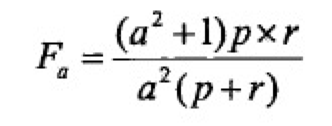
在自然语言处理领域中，针对文本分类问题的评价指标，本文主要介绍正确率、召回率、宏平均、微平均。

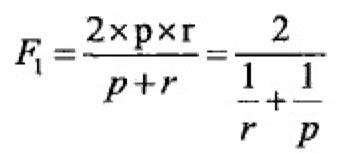
上表中a,b,c,d分别表示文本分类中相应的文本个数，正确率表示正确被检索到的数目与实际被检索到的数目的比值，召回率表示正确被检索到的数目与应该被检索到的数目。

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 一种基于情感词典的中文情感分析

3.1 问题描述

3.2 改进的情感词典分析方法

3.2.1 情感词典

3.2.2 ?

3.3 基于情感词典的中文情感分析算法

3.3.1 算法概述

3.3.2 算法模型

3.3.3 算法设计

3.4 实验结果与分析

3.3.1 实验性能指标

3.3.2 实验设计与结果分析

第四章 结合客观弹幕文本的机器学习情感分类模型

4.1 问题描述

4.2 基于弹幕文本的机器学习情感分类模型

4.3 结合客观弹幕语料的改进机器学习情感分类模型

4.4 实验结果及分析

4.5 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。