**弹幕消息情感分析**

目录

[1 研究背景 2](#_Toc489869003)

[1.1 问题描述 2](#_Toc489869004)

[1.2 实现目标 2](#_Toc489869005)

[1.3 弹幕消息特点 2](#_Toc489869006)

[1.4 情感分析现状 2](#_Toc489869007)

[2 基于词典情感分析 2](#_Toc489869008)

[2.1 基本思路 2](#_Toc489869009)

[2.1.1 句子切割 3](#_Toc489869010)

[2.1.2 中文分词 3](#_Toc489869011)

[2.1.3 计算句子积极消极得分 4](#_Toc489869012)

[2.1.4 计算句子积极可能性 4](#_Toc489869013)

[2.1.5 计算窗口大小话语整体情感得分 4](#_Toc489869014)

[2.1.6 绘制情感波动曲线 4](#_Toc489869015)

[2.2 主要问题 4](#_Toc489869016)

[2.3 算法介绍 5](#_Toc489869017)

[2.4 实现效果 5](#_Toc489869018)

[2.5 优势、劣势分析 5](#_Toc489869019)

[3 基于机器学习情感分析 5](#_Toc489869020)

[3.1 基本思路 5](#_Toc489869021)

[3.2 主要问题 5](#_Toc489869022)

[3.3 过程描述 5](#_Toc489869023)

[3.4 算法介绍 5](#_Toc489869024)

[3.5 实现效果 5](#_Toc489869025)

[4 对比分析 5](#_Toc489869026)

[4.1 识别精度 5](#_Toc489869027)

[4.2 时间性能 5](#_Toc489869028)

[4.3 可扩展性 5](#_Toc489869029)

[4.4 适用情况 5](#_Toc489869030)

# 1 研究背景

## 问题描述

为了满足观众的吐槽欲望, 弹幕视频应运而生, 它使观看视频和评论视频同时进行, 观众可以最大限度地发表对视频的赞叹或批评, 形成现场直播式的互动。但由于网站对弹幕内容的监管存在许多疏漏，网民素质参差不齐，弹幕视频里时常会出现一些争吵、辱骂或者一些未经许可的广告等信息，而这些负面信息往往会使得直播网站和主播深受困扰，也一定程度上影响了双方的收入以及观众的观看体验，所以如何快速、准确地识别并检测出其中的负面、消极的信息显得尤为重要。目前由于主播人数众多，直播网站采取的是人工管理方式，需要耗费大量的人力物力资源，迫切地需要一种新的省时费力的方法。

## 实现目标

实时识别出争吵、辱骂、不良信息广告等内容，及时发出警告信息。

## 弹幕消息特点

从消息内容出发，弹幕消息既是对直播视频内容的实时评论，也是观众与观众之间、观众与直播者之间的一种交流，所以它既具备评论带有一定情感倾向的特点，也同时含有聊天信息简短、用词不规范、存在大量网络用语等特征。

从消息的发送者出发，也就是直播观众，年轻人

## 情感分析现状

针对商品评论

# 2 基于词典情感分析

## 2.1 基本思路

基于词典的情感分析本质上是字符串匹配、基于一定规则的字符串处理，关键是情感词典的构建、情感得分规则的制定，具体是通过搜索一句话语里面情感词、程度词、否定词的数量以及出现的位置来计算话语的情感得分。

预处理操作（情感词典（积极、消极）、程度词词典、否定词词典 构建（借助语料库（用词规范特点）（ 弹幕消息特点）（可以加强、优化地方）））->计算一句话语的情感得分（句子切割（标点符号），中文分词（难点 概率模型），搜索一句话语里面情感词、程度词出现的位置，计算情感得分）->计算窗口大小话语的整体情感得分->绘制情感波动曲线

画一个流程图说明下

图1 基于字典情感分析流程图

### 2.1.1 句子切割

根据停顿分割标点符号[，。；！？…… ]将语句切割成若干个句子

### 2.1.2 中文分词

运用了jieba，python里面中文处理库

结巴中文分词涉及到的算法包括：  
(1) 基于Trie树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG)；

(2) 采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合；  
(3) 对于未登录词，采用了基于汉字成词能力的HMM模型，使用了Viterbi算法。  
结巴中文分词支持的三种分词模式包括：  
(1) 精确模式：试图将句子最精确地切开，适合文本分析；  
(2) 全模式：把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义问题；  
(3) 搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

采用的是精确模式（一个句子里面最大概率分词模式）

实际操作过程中：需要导入用户词典（词语 词频），提高分词准确性（这也是可以优化的地方，我们可以将弹幕消息里面常出现的网络用语等添加进去）jieba.load\_userdict(dict\_path) 导入用户词典

jieba.cut(sentence) 分词

举例说明：

我来到北京清华大学

精确模式：我/来到/北京/清华大学

全模式: 我/来到/北京/清华/清华大学/华大/大学

搜索引擎模式: 我/来到/北京/清华/华大/大学/清华大学 （精确性、召回率）

http://www.cnblogs.com/eastmount/p/5055906.html

### 2.1.3 计算句子积极消极得分（重要）

搜索情感词（积极或消极）（情感词个数越多，情感基础分值越大）-查看情感词前面是否有程度词（相当于情感得分权重，不同程度词（比较、非常（好）），权重大小不同）-查看情感词前面是否有否定词（情感得分取反）-情感得分正向化([-2,4]->[0,6])

举例说明：今天天气真好。 [1.2,0]

### 2.1.4 计算句子积极可能性

[posScore,negScore]->[posProbability]

posProbability=(posScore)/(posScore+negScore)

当两者相等或其中一个为零时，特殊计算

以0.5为积极消极分界点

### 2.1.5 计算窗口大小整体情感得分

设置窗口windowSize 也即每windowSize大小的消息计算一次整体情感得分

具体处理:

积极消极得分边界(考虑到有些话语不带有情感倾向，属于客观语句，所以得设置积极消极得分边界posBounder，negBounder)(具体实现 posProbability>=0.6 评定为积极 posProbability <=0.4 评定为消极)

根据可能性大小加权处理（如果一条语句积极可能性很大，则其权值也会很大）：

对于积极语句(posProbability-posBounder)/(1-posBounder)

对于消极语句(negBounder-posProbability)/(negBounder)

设置原因：

单独考虑一句话的情感倾向没有意义

当出现争吵、辱骂等不正当情况时，会持续一段时间，也即那段时间的话语消极会居多

选择多大合适：

不能太小（太小导致返回的异常情况太多）

不能太大（太大导致找不出异常情况 中和）

需要根据具体情况（消息整体数目）逐渐调节窗口大小，目前设定为100

设置异常情况情感得分阈值

低于阈值，认为是异常情况（争吵情况），会返回相关的语句以及语句所在的位置

高于阈值，正常

阈值设置多大合适：

过高，会降低精度（不是异常的也会返回）

过低，会降低召回率（是异常的也不返回）

想要达到的目标是把所有可能出现异常的情况返回，所以需要设置得偏高一点

目前处理是设置成-7 可以根据情感曲线再去设置

### 2.1.6 绘制情感波动曲线

调用python绘图函数，绘制出情感波动曲线

## 2.2 主要问题

基于“情感词典”的方式，没有什么学习过程，可以快速实现，毕竟最困难的部分“总结情感词典”已经实现完了。但是缺点在于：**准确率不高**。主要存在如下两类问题：

1）**如何处理否定词**。一般来说，应该制订规则，即，如果一个情感词出现在否定词后面，那么该情感词的“情感极性”就应该取反。但是，自然语言千变万化，没有那么简单。比如英语中的cannot recommend any more，中文中的“不要太便宜了吧”，上述规则就不起作用。

2）**网上下载到的“情感词典”往往是根据某个领域总结出来的**，比如社交媒体或是电商的评论。如果使用一个根据社交媒体得到的“情感词典”来判断电商评论的情感，可能电商评论中大量表达情感的词在“情感词典”中根本就找不到。

3）窗口大小设置

4）情感得分阈值设置

## 2.3 实现效果

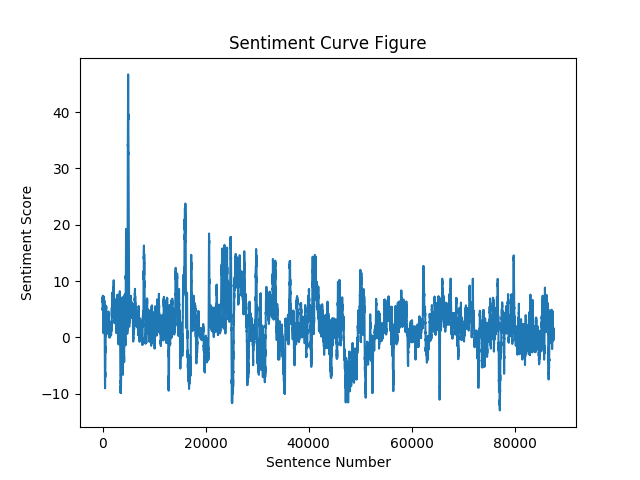


图2 情感波动曲线图

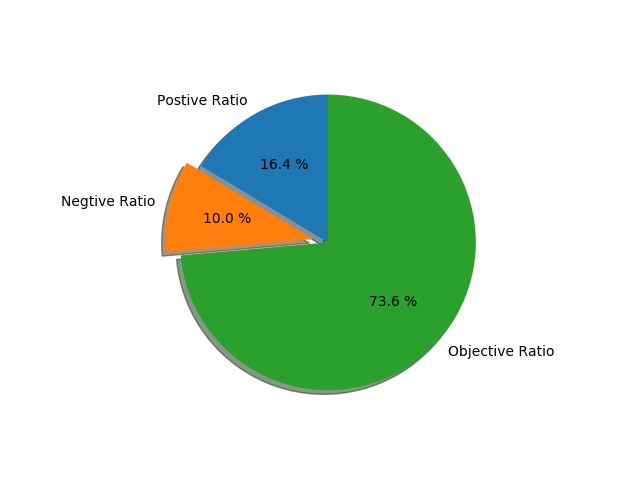


图3 积极消极比例图

## 2.4 结果评价

时间性能：sentiment Analyze based dict running time: 133.935403418 handle review num: 87642

识别精度：81%左右 根据标记数据计算

原因分析：

分词词典、情感词典、程度词词典里面收集的词汇与弹幕消息里面的用词不大匹配

基于词典的情感分析依赖于字符串匹配规则，不够灵活，无法解决歧义以及否定词问题

计算情感得分算法不够完善

## 2.5 优化改进

词典更新

# 3 基于机器学习情感分析

## 3.1 基本思路

机器学习以统计理论为基础，利用[算法](http://lib.csdn.net/base/datastructure)让机器具有类似人类般的自动“学习”能力，它对已知的训练数据做统计分析从而获得规律，再运用规律对未知数据做预测分析。

应用机器学习的方法进行情感分析需要进行数据标注、特征提取、特征选择、分类器构建、分类器训练、数据预测等工作。



图4 基于机器学习情感分析流程图

## 3.2 过程描述

### 3.2.1 数据标注

数据标注是对原始数据的人工分类，是给原始数据添加一个类标签，标记后的数据作为机器学习过程中所需要的训练集来训练分类器，训练数据质量的好坏直接关系到对未知数据结果预测的准确性。

情感分析可分为主客观分析、情感倾向分析、情感程度分析，考虑到弹幕消息的特点以及需要实现的目标，这次工作主要是进行主客观分析以及情感倾向分析，所以需要将数据进行主客观以及情感倾向的标注工作。



表1 数据标注格式表

格式说明：

review\_data：原始数据

review\_count：原始数据出现次数

is\_subjective：主客观标记，1表示主观，0表示客观

sentiment\_tendency：情感倾向标记，1表示积极，0表示消极

is\_erotic：不良内容标记，1表示是不良内容，0表示不是不良内容

key\_words：原始数据的关键词

### 3.2.2 数据预处理

为了减少标注工作，为了提高标记数据的质量，为了构建效果更好的分类器，往往需要进行数据预处理工作。

数据预处理主要包括读取与写入数据、数据格式转换、客观语句的过滤、重复语句的删除、标记数据的检查与错误处理、标记数据的合并、句子切割、中文分词、词性标注、去除停用词等工作。

标记数据的检查与错误处理主要检查的是已标记数据是否符号规定格式，所填的数值是否位于指定的区间范围之内，然后将不符号规范的数据所在的位置显示出来，提醒用户。

客观语句的过滤依据的是已导入的情感词典，判断一句话是否是客观句的关键因素是这句话是否含有情感倾向的词语，如果不含有，就判定为客观句，所以客观语句的过滤效果取决于情感词典的质量。

### 3.2.3 特征提取

1）特征定义

特征是分类对象所展现的部分特点，是实现分类的重要依据。我们经常会做出分类的行为，那我们依据些什么进行分类呢？

举个例子，如果我看到一个年轻人，穿着新的正装，提着崭新的公文包，快步行走，那我就会觉得他是一个刚入职的职场新人。在这里面，“崭新”，“正装”，“公文包”，“快步行走”都是这个人所展现出的特点，也是我用来判断这个人属于哪一类的依据。这些特点和依据就是特征。可能有些特征对我判断更有用，有些对我判断没什么用，有些可能会让我判断错误，但这些都是我分类的依据。

我们没办法发现一个人的所有特点，所以我们没办法客观的选择所有特征，我们只能主观的选择一部分特征来作为我分类的依据，这个选择也就是机器学习里面所说的特征选择，特征选择的目的是挑选出排名靠前的、有助于判别事物所属分类的特征。

2）特征提取

在情感分析中，一般从“词”这个层次来提取特征。

比如这句话“手机非常好用！”，它的类标签是“Positive”。里面有四个词（把感叹号也算上），“手机”，“非常”，“好用”，“！”。可以认为这4个词都对分类产生了影响，都是分类的依据。

同样的，对这句话，我也可以选择它的双词搭配（Bigrams）作为特征。比如“手机 非常”，“非常 好用”，“好用 ！”这三个搭配作为分类的特征。以此类推，三词搭配（Trigrams），四词搭配都是可以被作为特征的。

3）特征选择

特征选择也就是所谓的特征降维，特征降维本质上就是减少特征的数量。特征降维的好处主要有两个，第一能够提高计算速度，第二如果用一定方法选择信息量丰富的特征，可以减少噪音，有效提高分类精度。

那么如何选择信息量丰富的特征，答案是运用统计方法。统计方法包括词频（Term Frequency）、文档频率（Document Frequency）、互信息（Pointwise Mutual Information）、信息熵（Information Entropy）、卡方统计（Chi-Square）等等，这里选择使用卡方统计。

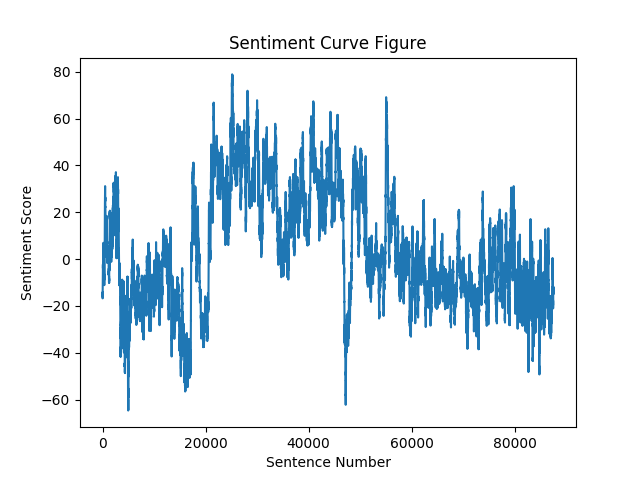
### 3.2.4 分类器构建

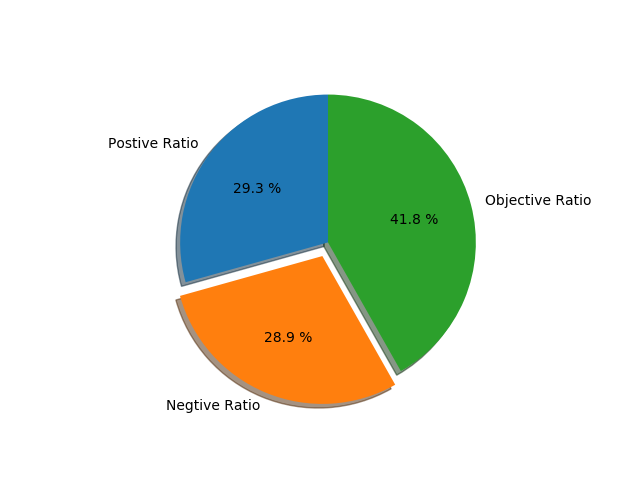
### 3.2.5 未知数据预测

## 3.3 主要问题

## 3.4 算法介绍

## 3.5 实现效果





## 3.6 结果评价

时间性能：sentiment Analyze based machine learning running time: 18.0294308669 handle review num: 87642

识别精度：

# 4 对比分析

## 4.1 识别精度

## 4.2 时间性能

## 4.3 可扩展性

## 4.4 适用情况

# 5 参考资料

1. 情感分析算法设计：

<http://site.douban.com/146782/widget/notes/15462869/note/355625387/>

2）文本分类技术：

<http://www.blogjava.net/zhenandaci/category/31868.html>

3）对推特数据进行文本情感语义分析：

<https://zhuanlan.zhihu.com/p/25873307>

4）游戏口碑的风向标——短文本聚类和维度口碑分析技术分享：

<http://wetest.qq.com/lab/view/30.html>

5）自然语言处理（NLP）与自然语言理解（NLU）的区别：

<http://blog.csdn.net/riverflowrand/article/details/51355238>

6）用“一袋子词”进行情感分析：

<http://www.infoq.com/cn/articles/Sentiment-analysis-using-bag-of-words?utm_source=articles_about_qinggandwe&utm_medium=link&utm_campaign=qinggandwe>

7）郑飓飓, 徐健, 肖卓. 情感分析及可视化方法在网络视频弹幕数据分析中的应用[J]. 现代图书情报技术, 2015, 31(11):82-90.

8）使用python+机器学习方法进行情感分析：

<http://www.10tiao.com/html/284/201607/2652389939/1.html>

9）鲁岳. 视频弹幕为对象的短文本情感分析研究[D]. 华中师范大学, 2016.

10）徐琳宏, 林鸿飞, 赵晶. 情感语料库的构建和分析[J]. 中文信息学报, 2008, 22(1):116-122.

邓扬, 张晨曦, 李江峰. 基于弹幕情感分析的视频片段推荐模型[J]. 计算机应用, 2017, 37(4):1065-1070.

11）中文分词：

Wang K, Zong C, Su K Y. Which is more suitable for Chinese word segmentation, the generative model or the discriminative one[J]. Paclic, 2013.

12）参考书籍：

Natural Language Processing with Python