弹幕消息情感分析

目录

1 研究背景

1.1 问题描述

直播网站的存在虽为网友们提供了自由表达意见和建议的平台，但由于其网站参与的主体大多数为年轻的受众，他们是一种自发性群体，深受“宅文化”的影响，较之现实生活中的群体更加松散、自由，以至于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、准确地检测并识别出其中的负面信息显得尤为重要。

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，所以迫切地需要一种能够自动检测的方法。

应用自然语言技术、机器学习方法进行情感分析可以在一定程度上分析出直播网站的情感趋势走向，进而实时检测出不良信息内容。

1.2 实现目标

检测并识别出争吵、辱骂、不良信息广告等内容，及时发出警告信息。

1.3 弹幕消息特点

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。

1.4 情感分析现状

情感分析主要分为两类，一是基于情感词典的文本情感分类，二是基于机器学习的文本情感分类，研究的分类对象主要是微博、商品评论、电影评论、博客、论坛文本等，鲜有针对弹幕消息进行情感分析的研究。

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、基于一定规则的情感得分计算，关键在于情感词典的构建、情感得分规则的制定，具体是通过搜索语句里面情感词、程度词、否定词的数量以及出现的位置，然后根据制定的规则来计算语句的情感得分。

基于词典的情感分析大体可以分为几个过程：预处理操作（情感词典、程度词词典、否定词词典的构建），计算语句情感得分（句子切割、中文分词、计算积极消极情感得分、计算积极可能性），绘制情感波动曲线（计算指定窗口大小的整体情感得分），显示异常弹幕信息。

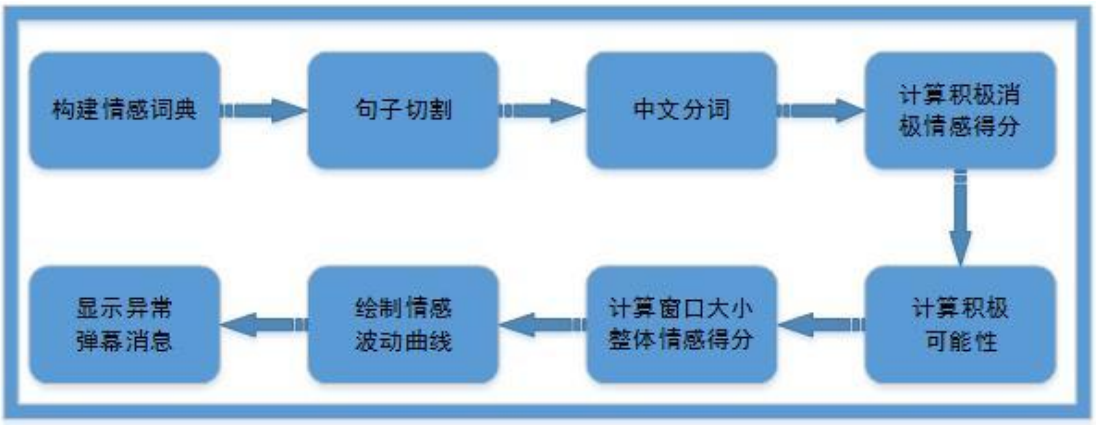

图 1 基于词典情感分析流程

2.2 过程描述

以“昨天天气不是很好，但今天天气非常不错”为例说明。

2.2.1 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。

[昨天天气不是很好，但今天天气非常不错]

2.2.2 中文分词

运用了 python 里面中文处理库 jieba，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。

以“我来到北京清华大学”为例说明：

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）

本次采用的是精确模式，简单讲就是句子里面最大概率分词模式，在实际应用过程中，需要导入用户词典，以此提高分词准确性，jieba 里面提供分词的函数接口是 jieba.cut()。

相应链接：http://www.cnblogs.com/eastmount/p/5055906.html

[[昨天，天气，不是，很，好],[但，今天，天气，非常，不错]]

2.2.3 计算句子积极消极得分

基本过程是从头到尾遍历一个句子，查看当前指向词语是否是情感词，如果是，再判断是积极还是消极情感词，并增加相应词性情感得分，然后：查看上一情感词后面与该情感词前面之间是否存在程度词，如果存在，按照程度词程度大小改变情感得分；查看上一情感词后面与该情感词前面之间是否存在否定词，如果存在，将情感得分取反。当处理完整个句子时，将所得到的积极情感得分、消极情感得分正向化，并计算积极消极得分总和。([-2,4]->[0,6],[4,-2]->[6,0],[-2,-4]->[4,2])

[[-1.5,0],[2,0]]-> [0.5,0]

2.2.4 计算句子积极可能性

目前我们得到的是句子的积极与消极得分[posScore,negScore]，但我们想知道的是句子属于积极的可能性有多大，这里分为四种情况考虑：

情况一，posScore 等于 negScore，则 posProbability 为 0.5

情况二，posScore、negScore 都不为 0，则 posProbability=(posScore)/(posScore+negScore)

情况三，posScore 为 0，posProbability 值在 1-(negScore+2)/ (negScore+3)与 1-(negScore+1)/ (negScore+2)之间，取这两个平均值处理。

情况四，negScore 为 0，posProbability 值在 (posScore+1)/ (posScore+2) 与 (posScore+2)/(posScore+3)之间，取这两个平均值处理。

0.657

2.2.5 计算指定窗口大小整体情感得分

类似于滑动窗口机制，根据指定窗口大小 windowSize，计算整体情感得分。

具体细节:

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

2）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)

2.2.6 绘制情感波动曲线

调用 python 绘图函数，绘制出情感波动曲线。

2.3 主要问题

1）情感词典不适用性

网上下载到的“情感词典”往往是根据某个领域总结出来的，并不适用于所有情况。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。

2）窗口大小设置

由于单独考虑一句话的情感倾向没有意义，当出现争吵等异常情况时，会持续一段时间，也即那段时间的语句消极成分会居多，所以可通过计算连续一定范围内语句的情感得分总和来评估情感趋势，关键问题是窗口设置多大比较合适？

如果窗口设置过小，可能导致返回的异常情况过多，警告作用不大；如果窗口设置过大，由于中和作用，可能会导致寻找不出异常情况。所以选择一个合适的窗口大小尤为重要，在实际使用中，可根据弹幕消息数量情况调节窗口大小，可设置在 50 至 150 之间。

3）情感得分阈值设置

情感得分阈值是判别异常情况的边界条件，阈值设置得好坏直接关系到输出异常语句的准确性。当情感得分低于阈值时，被认为是异常（争吵）情况，会返回相关的语句以及语句所在的位置；当情感得分高于阈值时，视为正常。

阈值设置的过高，会降低识别精度（不是异常的也会返回）
阈值设置得过低，会降低召回率（是异常的也不返回）
可根据绘制出的情感曲线来设置。

2.4 优化改进

基于词典的情感分析可优化改进的地方主要有三个：

用户词典的更改与完善，目前所使用的用户词典是根据一些书籍、网上文档等总结出来的，里面的用词比较规范，而弹幕消息里面的用词口语化，存在大量网络用语，这可能会导致分词的不准确性，所以可根据弹幕消息数据适当地完善下用户词典。

情感词典、程度词典的更改与完善，目前所使用的情感词典是根据社交媒体、电商评论总结出来的，应用于弹幕消息不大合适，可更改完善。

情感得分规则的优化，可尝试地提出一些新的方法来计算语句的情感得分。

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。

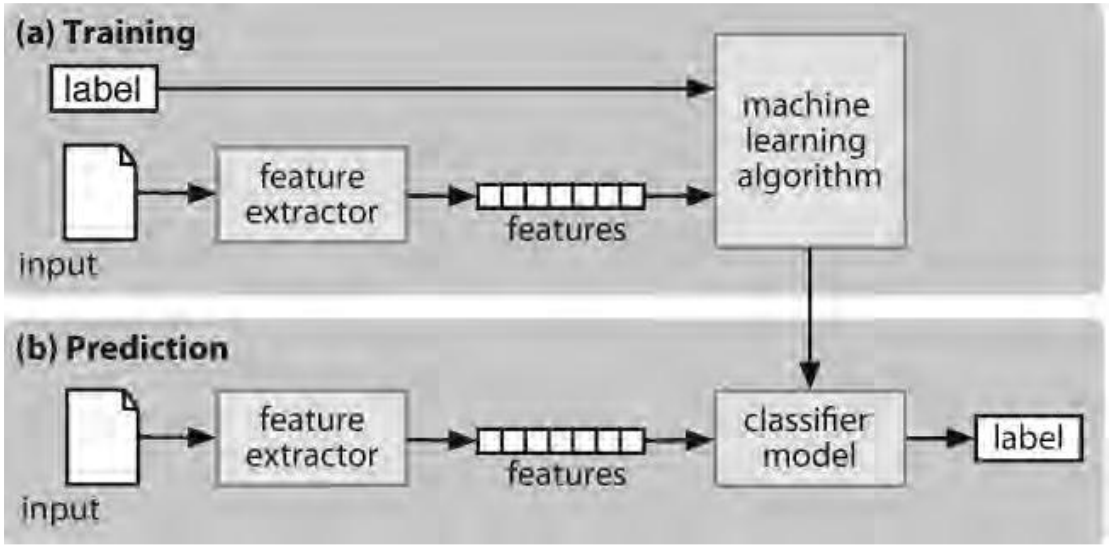

图 2 基于机器学习情感分析流程

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，标记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。

review_data	review_count	is_subjective	sentiment_tendency	is_erotic	key_words
好有型呐	3	1	1	0	
感谢拔刀、为紅颜赠送的1个666 送竹子抽Iphone 7P，猫币礼物抽华硕I7游戏本	1	1	1	0	
真的不尊重主播，别说了	1	1	0	0	
你有什么才艺 小马	1	0		0	
这盘贼刚	1	1	0	0	
一道美丽的风景线	1	1	1	0	

表 1 数据标注格式

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，然后将不符合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。

里面的一个基本思想是如果一个词在积极语句里的比例要远远高于在消极语句里的比例或在消极语句里的比例远远高于在积极语句里的比例，则可认为该词信息量丰富，反之，如果一个词在积极语句的比例大致等于在消极语句的比例，则可认为该词信息量不丰富。根据这个基本思想，计算每个词的信息量得分，再将得分逆序排列，挑选出排名靠前的 n 个词作为所选择的特征，这个 n 也就是所谓的特征维度。

4）特征表示

在训练分类器之前，首先要将所有的原始的标记数据表示成特征的形式。例如，“手机非常好用！”。如果选择所有词作为特征，则其可表示成[{“手机”: True, “非常”: True, “好用”: True, “！”: True}, positive]；如果选择双词搭配作为特征，则其可表示成[{“手机 非常”: True, “非常 好用”: True, “好用 ！”: True}, positive]；如果选择信息量丰富的词作为特征，则其可表示成[{“好用”: True}, positive]。

3.2.4 分类器构建

1）训练集与测试集

机器学习进行分类必须有数据给分类算法训练（训练简单讲就是构建一个能够拟合标记数据的过程），这样才能得到一个基于训练数据的分类器，有了分类器以后，自然就需要检测分类器的精度，所以要将标记数据集分为两部分，一部分作为训练集用于训练分类器，一部分作为测试集用于测试分类器分类效果。

2）分类算法的选择

影响分类效果的重要因素主要有两个，一个是分类特征，一个是分类算法，分类算法的好坏直接关系到预测结果的准确性。考虑到弹幕消息的特点，这里我们选择进行训练测试的分类算法有朴素贝叶斯（BernoulliNB、MultinomialNB），逻辑回归（LogisticRegression），支持向量机（NuSVC），K 最近邻（KNeighborsClassifier），神经网络（MLPClassifier）。

3）特征维度的确定

一个良好的特征维度不仅可以有效地减少计算量，而且可以较少噪音数据，提高分类精度，这里我们准备测试的特征维度在[500,3000]之间。

3.2.5 分类效果评价

考虑到分类中类别数目不平衡、相差很大的情况，所以引入了混淆矩阵的概念，混淆矩阵的横轴代表预测类别，纵轴代表实际类别。

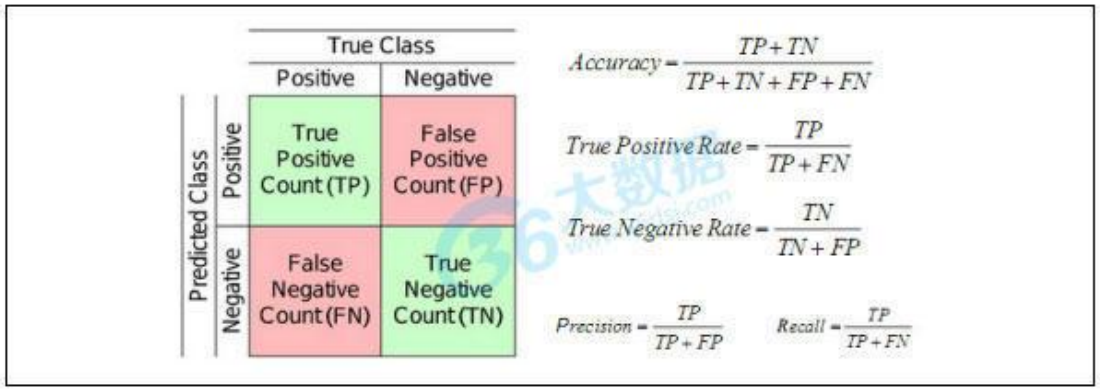

图 3 混淆矩阵

在评价分类器的分类效果时，有两个通用评价指标，一个是精度，一个是召回率。精度是被预测为正例的样本中实际也是正例的比例，召回率是正例样本中被正确预测的比例。通俗讲，精度是查的准，召回率是查的全。考虑到要分类的样本数据是一条条语句，所以我们选择以精度作为评价指标，由于正类与负类的数目大致相当，所以以正类精度与负类精度加权求和作为整体精度来评价分类效果。

3.3 主要问题

1）数据标注

目前标记的有效数据数目是 1154 条，数目不够多，覆盖的范围不够全面，可从不同种类的直播房间抓取一定数量的弹幕数据进行人工标注。

2）客观语句的过滤

客观语句的过滤采取的基于情感词典的方法，关键在于情感词典的质量。

3）分类算法选择

目前效果不错的分类算法有朴素贝叶斯，逻辑回归，可考虑使用一些集成算法如 adaboost 进行测试。

3.4 优化改进

基于机器学习的情感分析可优化改进的地方主要有几个：用户词典的更改与完善，情感词典的更改与完善，标记数据的增加，特征选择的优化，分类算法的改进等。

4 对比分析

4.1 实现效果

1）情感波动曲线图

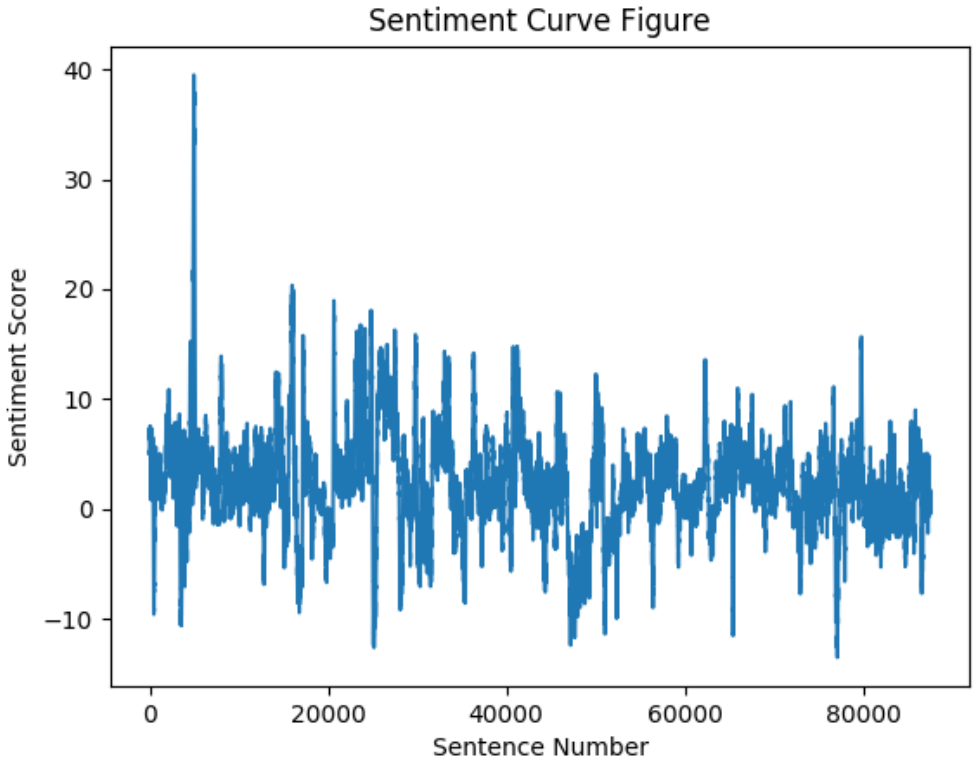

图 4 基于词典的情感波动曲线

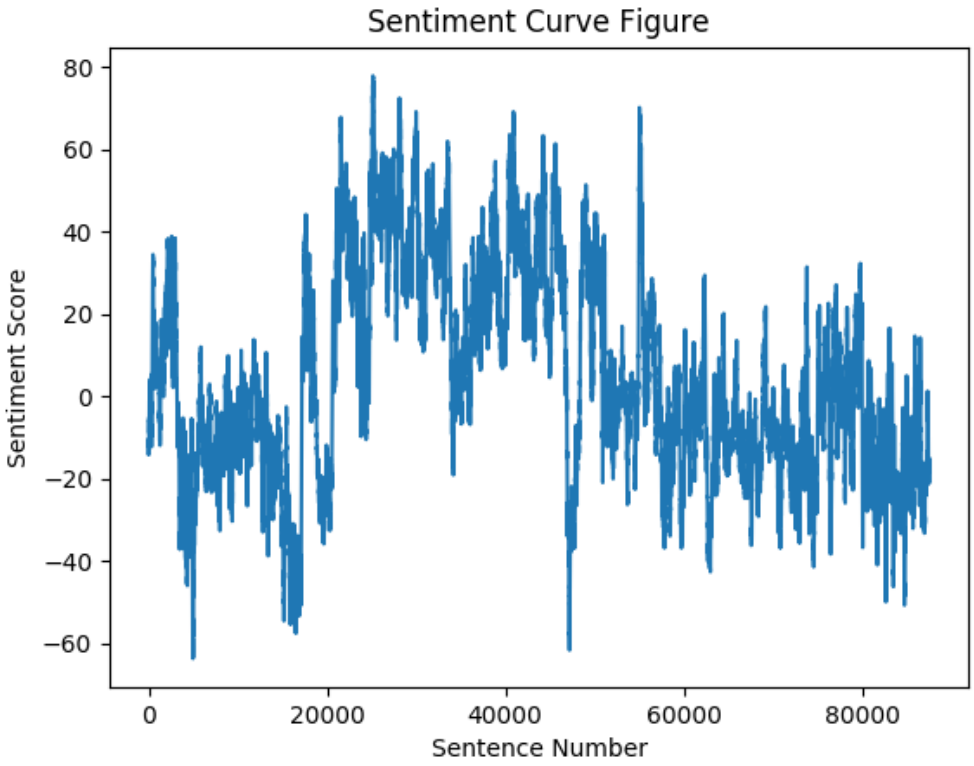

图 5 基于机器学习的情感波动曲线

2）类别成分占比图

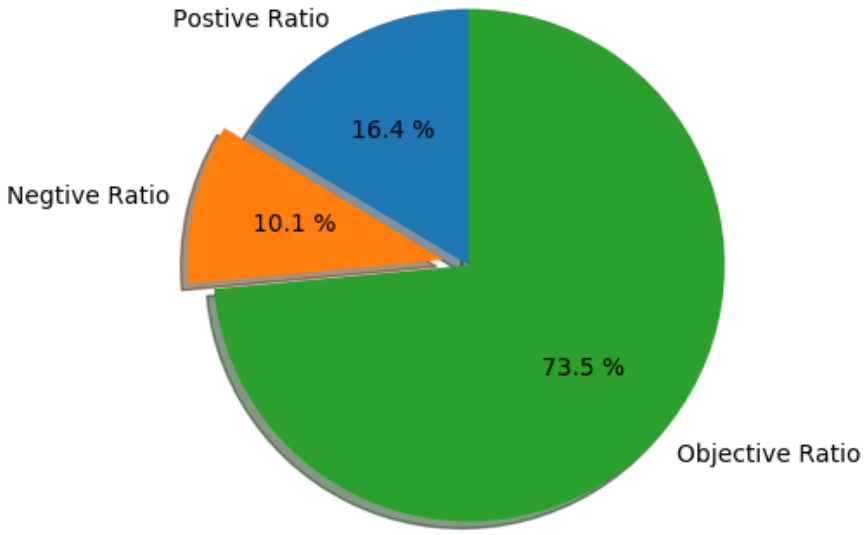

图 6 基于词典的类别成分占比

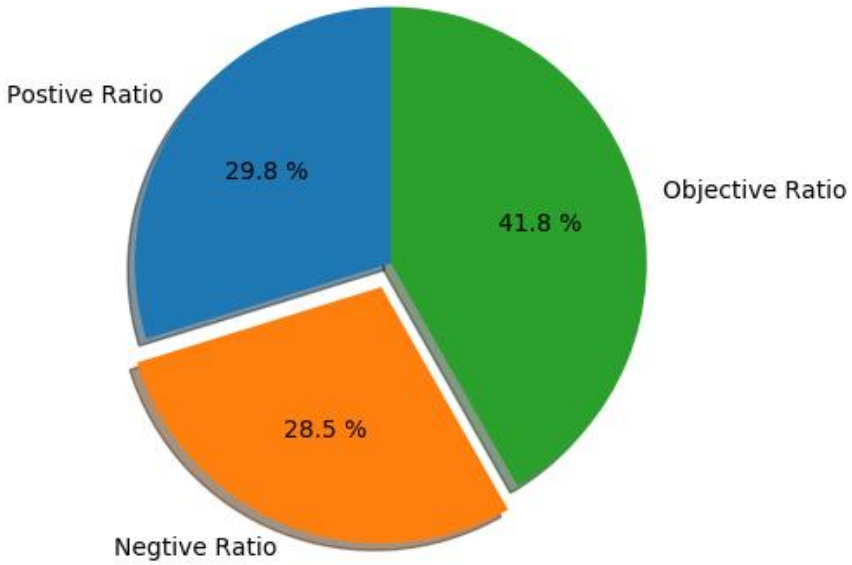

图 7 基于机器学习的类别成分占比

4.2 识别精度

测试样本数目	使用方法	精度
1154	基于字典情感分析	81.10%
	基于机器学习情感分析	88.10%

表 2 识别精度对比

4.3 时间性能

预测样本数目	使用方法	耗时/s
87642	基于字典情感分析	18.19144925
	基于机器学习情感分析	146.8819333

表 3 时间性能对比

4.4 可扩展性

1）基于词典的情感分析

基于词典的情感分析运用到的主要技术是中文分词、字符串匹配、基于一定规则的情感

得分计算，关键是情感词典、程度词典的构建，它可改变扩展的地方主要有三个：用户词典的更改，情感词典、程度词典的更改，情感得分规则的优化。

2）基于机器学习的情感分析

基于机器学习的情感分析运用的主要技术是中文分词、特征提取与选择、分类器训练、分类效果评价，关键在于标记数据质量与数目、特征选择、分类算法选择，它可改变扩展的地方主要有：用户词典的更改，情感词典的更改，标记数据数目的增加，特征选择的优化，分类算法的改进。

应用机器学习除了可以进行情感极性判别之外，还可进行主客观分析，以提取一些与弹幕消息相适应的关键词汇完善情感词典。

5 程序使用

5.1 环境配置

编译器：Python2.7.12
开发集成环境：Pycharm2017.1.5
相关库：nltk、sklearn、numpy、xlwt、wlrd、matplotlib

5.2 模块介绍

5.2.1 基于词典情感分析涉及模块

1）textProcessing（文本预处理）

完成数据读取、中文分词、词性标注、句子切割、停用词过滤、格式转换等功能。

用户词典（用于中文分词）保存在 D:/ReviewHelpfulnessPrediction/PreprocessingModule 目录下，名称为 userdict.txt ,可在里面增加一些新的词汇，每行的格式为：词语 词频。如果用户词典存储位置发生更改，则需要更改 jieba.load_userdict(userDictPath)。

停用词默认保存在 D:/ReviewHelpfulnessPrediction/PreprocessingModule 目录下，名称为 stopword.txt，用于 seg_fil_excel、seg_fil_txt 等函数中。

2）sentimentAnalyzeBasedDict

完成情感得分计算、情感曲线绘制等功能。

情感、程度词词典默认保存在 D:/ReviewHelpfulnessPrediction/SentimentDict 目录下，可以在里面增加相应的新的词汇，如果存储目录发生更改，需要修改 dictDir 的值。

sentiAnalyzeBaseDict()完成基于词典的情感分析工作，分析的文本数据默认保存在 D:/ReviewHelpfulnessPrediction/BulletData 目录下。输入参数包括：原始数据名称、原始数据文件格式、窗口大小、积极边界、消极边界、情感得分边界。该函数会将预测的原始数据所属类别结果保存在 D:/ReviewHelpfulnessPrediction/PredictClassRes 目录下，绘制出的情感波动曲线图、类别成分占比图保存在 D:/ReviewHelpfulnessPrediction/SentimentLineFig 目录下，并且会输出异常语句所在的位置以及相应的异常语句。

5.2.2 基于机器学习情感分析涉及模块

1）textProcessing（文本预处理）

完成数据读取、中文分词、词性标注、句子切割、停用词过滤、格式转换等功能。

用户词典（用于中文分词）保存在 D:/ReviewHelpfulnessPrediction/PreprocessingModule 目录下，名称为 userdict.txt ,可在里面增加一些新的词汇，每行的格式为：词语 词频。如果用户词典存储位置发生更改，则需要更改 jieba.load_userdict(userDictPath)。

停用词默认保存在 D:/ReviewHelpfulnessPrediction/PreprocessingModule 目录下，名称为 stopword.txt，用于 seg_fil_excel、seg_fil_txt 等函数中。

2）unlabelDataProcessToLabel

完成客观语句过滤、重复评论删除、已标记数据的检查与错误处理、标记数据合并等功能。

情感词典默认保存在 D:/ReviewHelpfulnessPrediction/SentimentDict 目录下，标记数据默认保存在 D:/ReviewHelpfulnessPrediction/LabelReviewData 目录下。

3）selectBestClassifier

完成最佳分类器的选择、最佳特征维度选择的功能。

已标记的训练数据保存在 D:/ReviewHelpfulnessPrediction/LabelReviewData 目录下，如果改变，需要修改 posNegDir、pos_review、neg_review 的值。情感停用词保存在 D:/ReviewHelpfulnessPrediction/PreprocessingModule/sentiment_stopword.txt 文件路径下。构建的分类器保存在 D: /ReviewHelpfulnessPrediction/BuildedClassifier 目录下，挑选出的最佳分类器名称以及最佳特征维度保存在 D: /ReviewHelpfulnessPrediction/BuildedClassifier/bestClassifierDimenAcc 文件路径下。

4）predictDataPosNegProbility

完成预测未知数据所属分类功能。

已标记的训练数据保存在 D:/ReviewHelpfulnessPrediction/LabelReviewData 目录下，如果改变，需要修改 create_word_bigram_scores()函数。

sentiAnalyzeBaseML()完成基于机器学习的情感分析工作，分析的文本数据默认保存在 D:/ReviewHelpfulnessPrediction/BulletData 目录下。输入参数包括：原始数据名称、原始数据文件格式、窗口大小、积极边界、消极边界、情感得分边界。该函数会将预测的原始数据所属类别结果保存在 D:/ReviewHelpfulnessPrediction/PredictClassRes 目录下，绘制出的情感波动曲线图、类别成分占比图保存在 D:/ReviewHelpfulnessPrediction/SentimentLineFig 目录下，并且会输出异常语句所在的位置以及相应的异常语句。

5.3 使用说明

1）基于词典情感分析

更改 sentimentAnalyzeBasedDict 模块里面 sentiAnalyzeBaseDict()函数的相应参数，运行。

2）基于机器学习情感分析

如果增加了标记数据，需要运行 unlabelDataProcessToLabel 模块里面的 save_label_data_to_spe_name()，将标记数据保存在特定位置，然后运行 unionFewLabelData(),将所有标记数据合并起来并保存在特定位置，再运行 selectBestClassifier 模块，训练出最佳的分类器，并将分类器保存在特定位置。

更改 predictDataPosNegProbility 模块里面 sentiAnalyzeBaseML()函数的相应参数，运行。

6 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