贝叶斯理论在情感分析中的应用

——以“Bilibili”网站网友对于“回形针 Paperclip”的态度为例

某

(某学院 某专业 某学号)

摘　要：使用 Bilibili 网站对于“回形针 Paperclip”视频的评论，利用贝叶斯理论构建朴素贝叶斯模型，并利用模型研究近期网民对于“回形针 Paperclip”的情感态度变化。

关键词：贝叶斯公式；朴素贝叶斯模型；情感分析；最大似然估计

一、引言

2021 年 6 月 18 日，一个名为“回形针 Paperclip”(以下简称“回形针”)的科普自媒体进入舆论的风口浪尖，起因是另一自媒体“赛雷话金”发视频抨击回形针，揭露回形针的一位前编辑现供职于美军陆军研究实验室，且另一位前编辑在国外社交媒体肆意散布反华言论。

这并不是回形针第一次进入舆论的风口浪尖，早在 2020 年 3 月 21 日，回形针在其视频《如何快速消灭全世界的森林》中提出的“我们的肉蛋奶消费可以实实在在影响巴西雨林的存亡”的论点即引发争议(以下简称“肉蛋奶”事件)。同时期，网友还揭露了回形针存在的其他问题，如其视频的 YouTube 版本，中国地图没有台湾省以及断章取义剪辑罗翔的普法视频等。

二、提出问题

以 Bilibili 网站(以下简称“B 站”)的视频评论为例，使用机器学习的方法，利用贝叶斯理论构建朴素贝叶斯模型，分析评论中的情感内容，研究 B 站的网友对于回形针的态度随时间的变化。

三、数据获取与预处理

此部分与概率论及数理统计无关，故只进行简要介绍。

B 站给予网友两种发表意见的方式，一种为评论，另一种被称为“弹幕”，

由于弹幕无法进行点赞互动等操作，所以很难代表主流意见，且受水军等因素影响较为严重，所以选择使用评论获知网友意见。从 2020 年 3 月至今，随机选取 10 个视频，利用爬虫 1获取这些视频的热门评论 2。由于 B 站反爬机制较为严格，所以采用了伪造 cookie 等技术，这里不进行详细介绍。

在获取到评论之后，对评论进行处理，删除无意义的评论，并为所有评论手动添加专家标注，“0”代表正常交流，“1”代表不满言论。而后进行分词，将评论切分为很多词汇，即将每条评论转换为一个词序列，便于之后模型的训练。中文分词也是利用贝叶斯模型进行的，这里直接使用了结巴分词 3，并根据 B 站的用户习惯添加了自定义词典以保证正确分词，如 DOGE(含有开玩笑的含义)，三连(表示对视频上传者的喜爱)等。

四、训练贝叶斯模型

训练贝叶斯模型，其核心就是贝叶斯公式，即：

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)}$$

在该公式中，x代表测试样本，c代表类别标签，$P(c|x)$表示某个测试样例x属于类别c的概率，$P(c)$为c类别的样例在总体中出现的概率，$P(x|c)$为x出现在c类别样本中的概率。

贝叶斯模型的核心思想如下：通过训练样本估计$P(c)$、$P(x|c)$，生成贝叶斯模型。在预测时，通过模型计算出$P(c|x)$，选取概率最大的类别作为预测结果，即：$h(x) = \arg \max_{c\ in\ y} P(c|x)$。

在文本分析模型中，对于$P(x|c)$，如果直接求取，意味着直接求取特定词序列出现概率，以以下两个词序列为例：

list1={“加油”,“加油”,“加油”}

list2={“加油”}

如果直接求取$P(x|c)$，意味着上述两个词序列无任何关联，即$P(list1|c)$和$P(list2|c)$是分别计算，完全无关的，这显然是不合理的，且某个特定的词序列

1 爬虫部分代码在 spider.py 中
2 共约 1000 条，在 trainset.csv 中
3 https://github.com/fxsjy/jieba

出现的概率可能过低，甚至不会出现，进而导致模型构建无法继续进行，所以这里按照朴素贝叶斯模型理论将$P(x|c)$进行拆分：

$$P(x|c) = \prod_{i=1}^{m} P(x_i|c)$$

其中，x_i为每个词汇的取值，这样，就把整个句子出现的概率，拆分为各个词汇出现概率的乘积。这里假设各个词汇的出现概率是相互独立的（这是不符合实际情况的，由于这里训练集只有 1000 条左右，而且评论的长度一般较短，所以没有使用复杂的贝叶斯网）。

某些词汇可能不会出现，导致某些概率值为 0，影响结果，所以采用了拉普拉斯修正，保证概率不为 0，公式如下：

$$P\left(x_i|c\right) = \frac{\left|D_{c,x_i}\right| + 1}{|D_c| + N_i}$$

其中 $\frac{\left|D_{c,x_i}\right|}{|D_c|}$即为原有的$P\left(x_i|c\right)$，$N_i$为第$i$维上，可能出现的所有种类数，在这里为所有可能的词汇种类，在这里经过实际测试后，发现在分母不包括N_i的情况下，仍能保证概率不为 0，且分类的准确率更高，所以进行预测时去掉了N_i。

另外，为防止过多的概率相乘导致下溢，以正确处理词数较多的文本。在这里使用了$lnP(x|c) = \sum_{i=1}^{m} lnP(x_i|c)$来取代$P(x|c)$，将乘法改为加法。

在本模型当中，由于分类的类别只有两类，所以在具体判断过程中，可以直接使用不同类别概率的比值进行判别，即：

$$\frac{P\left(c_{positive}|x\right)}{P\left(c_{negative}|x\right)} = \frac{P\left(c_{positive}\right)P\left(x|c_{positive}\right)}{P\left(c_{negative}\right)P\left(x|c_{negative}\right)}$$

这种方式可以将$P(x)$消去，在建立模型时只需要计算两类样本在训练集中出现的概率与各类样本中各词汇出现的概率。

在实践过程中发现，在上述策略构建的朴素贝叶斯模型下，1000 次测试的平均准确率约为 75%，准确率较低，所以采取了以下两个措施提高准确率：

1.拟合正态分布模型

通过观察发现，两类评论的长度略有差异。因为正常交流往往涉及到对视频科普内容的讨论，相对而言长度更长。所以使用最大似然估计法，假定评论长度的分布符合正态分布，拟合一个正态分布模型，将某长度出现的概率也用

于模型的训练。即在计算$P(x|c)$时，连乘积中加入$P(length|x)$。

2. 加入停止词机制 4(stopwords.txt)

对错误分类的训练样本进行进一步研究，发现某些词序列含有无意义的词汇或标点，例如“啊”、“而且”、“。”等。于是加入了停止词机制，对于拆分出的某些无意义的词汇，将直接删除，不进入贝叶斯模型的统计范围。

在加入上述两种提高准确率的机制后，准确率提高到 85%，猜测无法继续提高的原因如下：

1. B 站评论内容的复杂性

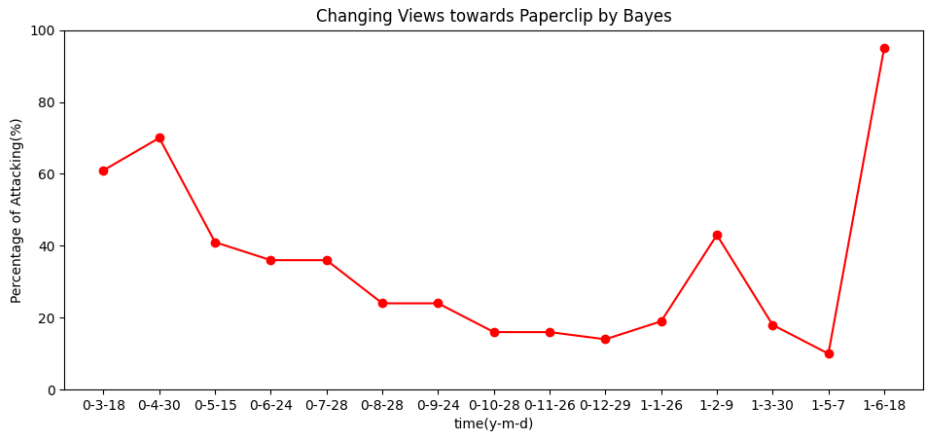
由于 B 站网友平均年龄较低，产生了很多隐藏含义，以及某些不符合常规的语言规范的转折，导致判断语义相对困难，即使是人类来判断，准确率也很难做到很高。使用复杂的贝叶斯网可能会取得更好的效果，但是 85%的正确率总体来看已经足够。

2. 训练样本数目的匮乏

由于每个评论样本都需要手动为其添加专家标注，工作量很大且时间有限，所以只选取了 1000 余条评论作为测试集中的样本，样本数目有所不足。

五、利用训练出的模型对网友情感进行分析

自 2020 年 3 月开始，每月选取一个视频获取评论 5，通过模型测试不满言论在视频评论中的占比，得到测试结果：



4 以 https://github.com/goto456/stopwords 为基础

5 在 comment_predict.csv 中

根据时间线来看，2020 年 3 月 20 日发生了“肉蛋奶”事件，所以部分人回到上一个视频(3 月 18 日)表达不满，而在 4 月 30 日，回形针发布了不诚恳的道歉视频，导致网友的不满达到一波高潮，但是仍有小部分人选择继续支持。随着事件热度下降，网友对于其不满态度也逐步下降。

2021 年 1 月，回形针因自媒体“赛雷话金”的抨击而再次登上热搜。30 日，回形针发布了视频《直面问题 ： 我们为什么又上热搜了？》，对此做出回应，此时对其不满达到新的高潮，但此次由于问题仅涉及科普领域，所以不满情绪比之前少很多。风波过后，不满情绪很快回落。

6 月 18 日，自媒体“赛雷话金”再次抨击回形针，揭露回形针编辑的身份。此时回形针进行资本主义文化渗透的危险面目已经昭然若揭，几乎已经没有网友支持回形针，对其不满态度也达到顶峰。

六、总结

本次拟合的朴素贝叶斯模型虽然只有 85%的正确率，但是其预测结果已经能够很好说明网友对于回形针的态度变化。此模型不仅使用了朴素贝叶斯模型完成了文本的情感分析基本内容，还加入了正态分布的拟合与停止词机制，进一步提高了分类的准确率。总体来说，朴素贝叶斯模型是对于文本情感分析的一种很好的预测模型，相较于支持向量机(SVM)更加容易理解与训练。

源代码：（代码用途详见 README.md）

https://github.com/NicerWang/bilibili-papercilp-viewlearning

参考文献：

[1] 维基百科：回形针 Paperclip.
[2] 维基百科：朴素贝叶斯分类器.
[3] 周志华，《机器学习》，清华大学出版社，2016.
[4] 盛骤, 谢世千, 潘承毅 等编，《概率论与数理统计(第四版)》，高等教育出版社，2008.