

自然语言处理大作业-情感分类

陈立国*

中国科学院计算机网络信息中心

2016.5.31

*学 号 : 2015E8016061028, 电 子 邮 件 : sunshinenum@foxmail.com, 电
话: 18101387977

1 引言

随着web2.0时代的加速发展，人们越来越倾向于在网络上表达自己的情感倾向，由于人工处理这些大众舆论费时费力，因此诞生了情感分析的研究方向。

对于新闻评论的情感分析可以及时的做到舆情监测，对于产品评论的舆情分析有助于公司及时了解产品的反馈[1]。

2 国内外相关研究

情感分析又称为意见挖掘，对带有情感倾向的文本进行分析、处理、归纳和推理的过程。此方面的研究根据处理文本的粒度不同可以分为：词语级别的、短语级别的、句子级别的、篇章级别的等几个研究层次。[4][2]根据处理文本的类别分为基于新闻评论的情感分析和基于产品评论的情感分析，基于新闻评论的情感分析多用于舆情的分析和信息预测，基于产品评论的情感分析多用于分析消费者对产品的喜好和倾向[6]。

情感分析中，抽取评价词语是一个重要的任务。评价词语的抽取和判别一般采用两种方法，分别是基于词典的方法和基于语料库的方法。基于词典的方法一般是使用WordNet等词典中的极性信息和同义反义信息来识别情感倾向和强烈程度，但是这种方法依赖情感词典，词典构建成本大。基于语料库的方法简单易行，但是需要标注数据，而且针对一种评论的模型应用到另一种评论后效果会大打折扣。[5][3]

3 我们的思路

本文对多种产品评论进行情感分类，通过自己设定的一种评分指标对评论抽取特征进行分类，取得了非常好的效果。为了区分正样本和负样本，本文的目标是找出这样一些词，这些词的在正样本中出现的次数比负样本的次数多很多（赞美的词语），或者负样本中的出现次数比正样本多很多（负面的词语），使用这些词语作为特征。为了区分这些词语和普通的词语，本文利用调和平均数的性质，当一个词语在正负样本中词频差距大时，两个样本中该词出现次数的调和平均数会变小，当差不多时，调和平均数会变大，根据这个特性。提出了一种基于调和平均数的特征提取方法。

本文采用词袋模型，首先使用分词软件对评论进行分词，然后统计词语 W_i 在正向和负向评论中出现的次数 C_{ij} ，其中 j 对应分类的数目，在本文中 $j \in 0, 1$ ，然后计算词语得分：

$$Score(W_i) = \frac{\sum_{j=0,1}(C_{ij})}{\prod_{j=0,1}(C_{ij}) + 1} * \log(\sum_{j=0,1}(C_{ij})) \quad (1)$$

将词语按照 $Score$ 排序，取排名前 n 个的词语作为特征词，对于每个评论生成特征向量，最后输入分类器分类。

在实验中特征的表示非常重要，为了选出那些在正面和负面分类中词频相差较大的词语，对词在正负样本中的词频取调和平均的倒数，为了防止分母为0，分母上加1，为了考虑词频对 $Score$ 的影响，最后乘以 $\log(\sum_{j=0,1}(W_{ij}))$ 。在训练样本数目达到4000时，使用本特征4折交叉验证得到的准确率为0.97175，远高于使用tf-idf的0.909。

4 实验结果及其分析

为了验证此特征的效果，实验中使用了tf(table1)、tf-idf(table2)进行对比实验，实验使用朴素贝叶斯分类器，实验所使用的数据为产品评论数据，包括Movie reviews of sentences (Pang and Lee)，豆瓣网的书评、京东电脑评论、携程网房间评论，每种评论分积极和消极，各有2000各样本，第一个数据集各1000，为英文数据，无需分词，后三个中文数据集需要进行分词处理，分词使用分词工具jieba。

本特征所选择的词汇有一定特点，首先正负两类中该词汇出现次数有较大差异，其次，该词汇出现次数相对较多。所以对于某种评论正负样本中的词汇有较大差异时效果较好。

通过以上实验数据发现，在处理书评时，本文所提出的特征有极好的效果，得分明显高于tf-idf特征。仔细分析，原因如下：对于评论的情感词，在不同的领域是不同的，对于电脑评论可以用“好”、“不好”来评价，这种词前加否定的特征词在tf-idf中是会被选出的，tf-idf选中词频达标、文档频率不是很高的特征，而在书籍的评论中，特征词会多于在电脑的评论，电脑的评价方面相对单一，因此评价词也相对较少，评论中往往围绕配置、使用体验展开，因此总体来说特征词数目较少，但是每个特征词的词频很高。

数据集	准确率	召回率	F值
pang_lee_movie_reviews	0.694	0.417	0.575
当当网书评	0.887	0.925	0.891
京东网电脑评论	0.893	0.890	0.892
携程网旅店评论	0.884	0.0.874	0.882

Table 1: tf特征实验结果

数据集	准确率	召回率	F值
pang_lee_movie_reviews	0.793	0.808	0.737
当当网书评	0.909	0.908	0.909
京东网电脑评论	0.890	0.894	0.890
携程网旅店评论	0.891	0.886	0.890

Table 2: tf-idf特征实验结果

数据集	准确率	召回率	F值
pang_lee_movie_reviews	0.962	0.963	0.926
当当网书评	0.972	0.953	0.971
京东网电脑评论	0.895	0.845	0.889
携程网旅店评论	0.904	0.962	0.909

Table 3: 调和平均数-词频特征实验结果

而对于书籍和影评，仁者见仁，智者见智，评价的方面很多，评论特征词也相对较多，但是每个特征词的数目不会太多。而积极词汇一般有别于消极词汇，比如“金玉良言”是褒义词，在消极词汇中一般不会出现这个词，再比如“启迪”、“跌宕起伏”、“毒害”、“肤浅”等等。综上所述，调和平均数-词频特征对书籍这类正负样本特征词不同的分类效果较好，在其他类别的评论中表现一般。

5 结束语

本文主要是探讨了调和平均数-词频特征对于不同商品评论的情感分类效果，通过对情感词的分析发现，书籍类评论的情感词正负样本的差距较大，使用调和平均数-词频特征效果很好，而对于正负样本特征词差距较小的评论，表现和tf-idf等相差不大。每个方法都有适合使用的领域数据集，对于书评数据和影评数据，使用本文提出的方法效果非常好，说明这种模型是适用于这类数据的。

References

- [1] 刘楠. 面向微博短文本的情感分析研究. 博士, 2013.
- [2] 张紫琼, 叶强, and 李一军. 互联网商品评论情感分析研究综述. 管理科学学报, (06):84-96, 2010.
- [3] 徐旭. 基于特征极性的跨领域情感分类方法研究. 硕士, 2015.
- [4] 朱力. 中文词语情感倾向研究. 硕士, 2009.
- [5] 欧倩倩. 基于特征关联的跨领域情感分类方法研究. 硕士, 2014.
- [6] 魏慧玲. 文本情感分析在产品评论中的应用研究. 硕士, 2014.