网络文本情感计算课程结业报告

姓　　名：胡明明

学　　号：1133010101

所学专业：数字媒体艺术

报告题目：基于词典匹配和机器学习的文本情感

提交日期：2015 年 12 月 4 日

基于词典匹配和机器学习的文本情感分析

摘要：随着互联网的发展，web 文本迅速增多，对于这些文本，特别是用户主动发布的评论数据进行挖掘和分析，识别出其情感趋向和演变规律，可以更好地理解用户的消费习惯，分析热点舆情，给企业、政府等机构提供重要的决策依据。本文主要通过一组对照试验，分别使用了词典匹配的方法和机器学习的方法进行文本情感分析，来比较两种方法的异同之处。其中机器学习的算法选用的是朴素贝叶斯算法。

关键词：文本情感分析、朴素贝叶斯、词典匹配

引言：博客、微博、SNS 社交网站的出现，增强了网络的人际交互性与及时性，使互联网逐渐成为普通大众交流观点、抒发情感的平台，同时也积累下关于人类心理和行为的海量文本信息。这些主观性文本每天以指数级的速度增长，仅靠人工进行分析需要消耗大量的人力和时间。因此采用计算机来自动地分析这些主观性文本表达的情感成为一种新的研究方向，即文本情感分析或称为意见挖。

文本情感分析是指对包含用户表示的观点、喜好、情感等的主观性文本进行检测、分析以及挖掘。文本情感倾向分析作为一个多学科交叉的研究领域，涉及包括自然语言处理、计算语言学、信息检索、机器学习、人工智能等多个领域。

文本情感分析的一个重要应用领域是对互联网上出现的大量产品评论进行挖掘与分析，主要目的是能够比较精确地发现产品的优缺点。此外还有新闻评论的情感分析，可以用于网民舆情监控。可以根据消费者意图做推荐，可以做股票预测，另外在搜索引擎中也有重要应用。

情感分析是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。按照处理文本的粒度不同可以分为词语级、短语级、句子级和篇章级等。情感分类的方法主要有两种：无监督的分类算法和有监督的分类算法。其中无监督的分类算法是基于情感词典及规则的，它的优点是无需标注数据，缺点是构建词典和规则耗时耗力，准确率不高。有监督的分类算法是基于机器学习的，它的优点是分类效果提升，缺点是依赖标注语料和特征选择。

一、基于词典规则的无监督分类算法

情感词典，是按照不同的情感倾向对情感词分类后的词典。基于情感词典及规则方法，顾名思义，即按照人工构建的情感词典和指定的相关规则来进行情感倾向性判定的方法。简单来说，就是找出句子里的积极情感词和消极情感词，如果积极词的个数大于消极词的个数，就判定这个句子是积极的，反之就判定这个句子是消极的。评价：这种方法往往准确率非常高召回率很低；规则集需要人工精心撰写；建立和维护规则集的过程比较费事费力。

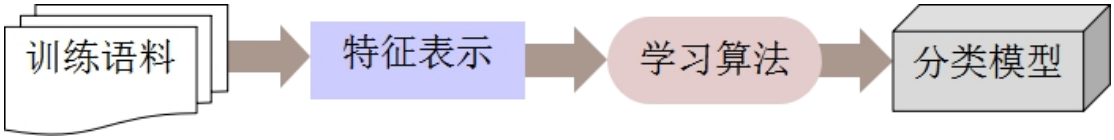
对于结构简单的句子，此方法会取得较好的分类效果。但是在句式复杂的某

些中文微博面前，基于情感词典及规则的方法显得很乏力；与此同时，人工构建情感词典在实际操作中，受到成本和规模的限制，不适于广泛推广。如果情感词典的规模过小，则会遗漏很多情感词，无法识别文本的情感倾向，特别是对于微博这类短文本，更不易命中情感词；如果情感词典的质量不高，也会造成情感分析结果的错误。

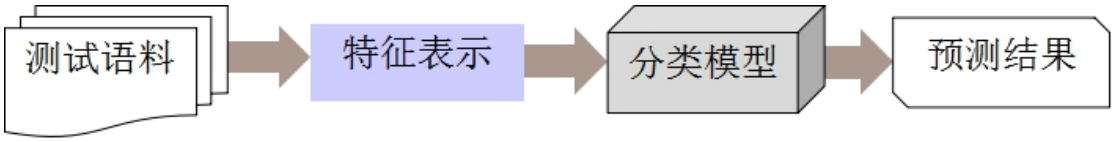
二、基于机器学习的有监督分类算法

机器学习算法分为做分类主要分为两步，一是训练，二是测试。

训练过程：



预测过程：



两个过程都需要先提取特征，训练过程是学习算法通过处理训练预料得出一个分类模型，测试过程是使用这个分类模型应用处理测试预料。

特征表示：对文本进行特征的抽取，转化为机器可理解的向量的表达形式。学习算法包括：朴素贝叶斯、最大熵、支持向量机等。特征抽取有很多种方式：词袋模型、否定特征、情感词频率特征、词性特征、N-gram 特征、强度词词典特征、句法依存特征、词向量特征。

三、评价指标

1. 正确率 = 正确识别的个体总数 / 识别出的个体总数

2. 召回率 = 正确识别的个体总数 / 测试集中存在的个体总数

3. F 值 = 正确率 * 召回率 * 2 / (正确率 + 召回率) （F 值即为正确率和召回率的调和平均值）

正确率是评估捕获的成果中目标成果所占得比例；

召回率，顾名思义，就是从关注领域中，召回目标类别的比例；

而 F 值，则是综合这二者指标的评估指标，用于综合反映整体的指标。

四、实验

实验步骤

1、下载数据

3000 句褒贬中的数据（已标注），积极消极中性各 1000 句、褒义词词典、

贬义词词典。

2、数据预处理

数据清洗，把一些 html 标签、@user、url 等去掉（已完成）；文本分词，依托 LTP 分词工具进行分词。

3、文本情感分类

一个是基于词典规则的情感分类，一个是基于机器学习的情感分类。

4、分类效果评估

使用五折交叉验证的方法，4/5 的数据作为 train，1/5 的数据作为 test，分类效果取平均；计算其准确率、召回率、F 值。

基于情感词典的分类实验

思路：首先载于积极词典和消极词典，把词性和词以键值对的方式存在字典里。然后读取语料文本，利用离线 LTP（Python 版）分词，遍历每句话的每一个词，在字典中查找对应的词性，如果没有找到就查找下一个词，如果查到一个词为积极词性，则使积极词的计数器增加 1，其他类似。当一句话的所有词分析介绍后，根据积极词个数和消极词个数进行情感分类，分类方法如下：

```
If count[pos] == count[nega] == 0:
   Classify =  'Neutral'
Elif count[pos] == count[nega] != 0:
   Classify =  'Positive'
Elif count[pos] > count[nega]:
   Classify =  'Positive'
Elif count[pos] < count[nega]:
   Classify =  'Negative'
```

如果积极词的个数等于消极词的个数且都不等于零，就把句子判为积极。

	中性	积极	消极	总计
标注词性	1000	1000	1000	3000
判断结果	535	1599	866	3000
判断正确	371	897	672	1940
正确率	0.693	0.561	0.776	0.646
召回率	0.371	0.897	0.672	
F 值	0.483	0.691	0.721	

修改判断条件，如果积极词和消极词的个数相等且不为零，就把句子判为消极。

	中性	积极	消极	总计
标注词性	1000	1000	1000	3000
判断结果	535	1369	1096	3000
判断正确	371	897	790	2058
正确率	0.693	0.655	0.721	0.686
召回率	0.371	0.897	0.79	
F 值	0.483	0.757	0.754	

修改判断条件，如果积极词和消极词的个数相等且不为零，就把句子判为中性。

	中性	积极	消极	总计
标注词性	1000	1000	1000	3000
判断结果	765	1369	866	3000
判断正确	418	897	672	1987
正确率	0.546	0.655	0.776	0.662
召回率	0.418	0.897	0.672	
F 值	0.474	0.757	0.721	

结论：

1. 基于词典规则的分类方法，情感分类的平均正确率为 66.47%。
2. 当积极词和消极词个数相等且不为零时，如果把它分为积极或中性，都会提高该类的召回率，降低正确率，F 值降低。但如果把它分为消极，则召回率提高，正确率降低，F 值却增大了。
3. 当积极词和消极词个数相等且不为零时，把句子判为消极情感，整体的正确率大于判为中性，更大于判为积极。

基于贝叶斯的机器学习方法分类实验

首先使用训练预料，根据朴素贝叶斯公式得出一个模型，然后把这个模型应用到测试预料进行分析。

已知某条件概率，如何得到两个事件交换后的概率，也就是在已知 P(A|B)的情况下如何求得 P(B|A)。这里先解释什么是条件概率：

$P(A|B)$表示事件 B 已经发生的前提下，事件 A 发生的概率，叫做事件 B 发生下事件 A 的条件概率。其基本求解公式为 $$P(A|B) = \frac{P(AB)}{P(B)}$$

贝叶斯定理之所以有用，是因为我们在生活中经常遇到这种情况：我们可以很容易直接得出 P(A|B)，P(B|A)则很难直接得出，但我们更关心 P(B|A)，贝叶斯定理就为我们打通从 P(A|B)获得 P(B|A)的道路。

下面不加证明地直接给出贝叶斯定理： $$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

根据贝叶斯定理我们进行以下实验。

A 组：3000 句预料。选其中 500 做训练预料，其余 2500 做测试预料，共计 6 组。

	A 组	B 组	C 组	D 组	E 组	F 组	平均值
标注中性	834	833	834	833	833	834	833.5
标注积极	833	834	833	834	833	833	833.34
标注消极	833	833	833	833	834	833	833.16
判断中性	753	774	750	771	767	806	770.17
判断积极	940	885	984	1001	876	837	920.5
判断消极	807	841	766	728	857	857	809.34
中性正确	751	765	744	759	760	791	761.67
积极正确	719	710	731	741	704	692	716.17
消极正确	685	710	662	639	722	713	688.5
中性 准确率	0.9973	0.9883	0.992	0.9844	0.9908	0.9813	0.9890
中性 召回率	0.9004	0.9183	0.892	0.9111	0.9123	0.9484	0.9137
中性 F 值	0.9464	0.952	0.939	0.9463	0.9499	0.9645	0.9496
积极 准确率	0.7649	0.8022	0.7428	0.7402	0.8036	0.8267	0.78
积极 召回率	0.8631	0.8513	0.8764	0.8884	0.8451	0.8307	0.8591
积极 F 值	0.8111	0.826	0.804	0.8075	0.8238	0.8286	0.8168
消极 准确率	0.8488	0.8442	0.8642	0.8777	0.8424	0.8319	0.8515
消极 召回率	0.8223	0.8523	0.7947	0.7671	0.8657	0.8559	0.8263
消极 F 值	0.8353	0.8482	0.8279	0.8186	0.8538	0.8437	0.8379
准确数	2155	2185	2137	2139	2186	2196	2166.33
错误数	345	315	363	361	314	304	333.67
准确率	0.862	0.874	0.8548	0.8556	0.8744	0.8784	0.866

描述：

1. 判断为中性的准确率、召回率、F 值都在 90%以上，均高于其他两种情况。
2. 判断为积极的准确率低于 80%，召回率和 F 值在 80%以上。
3. 判断为消极的准确率、召回率、F 值都在 80%以上。
4. 所有情感分析的准确率为 86.6%。

B 组：现在把训练预料和测试预料对调，即用 2500 句做训练，500 句做测试。

	A 组	B 组	C 组	D 组	E 组	F 组	平均值
中性	166	167	166	167	167	166	166.5
积极	167	166	167	166	167	167	166.67
消极	167	167	167	167	166	167	166.83
判断中性	163	160	159	155	165	157	159.83
判断积极	186	183	184	188	182	184	184.5
判断消极	151	157	157	157	153	159	155.67
中性正确	157	160	159	155	165	157	158.83
积极正确	151	153	156	157	155	157	154.83
消极正确	140	141	145	144	140	149	143.17
中性 准确率	0.9631	1	1	1	1	1	0.9938
中性 召回率	0.9457	0.958	0.9578	0.9281	0.988	0.9457	0.9538
中性 F 值	0.9544	0.9785	0.9784	0.9627	0.9939	0.9721	0.9733
积极 准确率	0.8118	0.836	0.8478	0.8351	0.8516	0.8532	0.8392
积极 召回率	0.9041	0.9216	0.9341	0.9457	0.9281	0.9401	0.9289
积极 F 值	0.8555	0.8767	0.8888	0.887	0.8882	0.8945	0.8817
消极 准确率	0.9271	0.898	0.9235	0.9171	0.9150	0.9371	0.9196
消极 召回率	0.8383	0.844	0.8682	0.8622	0.8433	0.8922	0.858
消极 F 值	0.8805	0.8703	0.8950	0.8888	0.8777	0.9141	0.8877
准确数	448	454	460	456	460	463	456.83
错误数	52	46	40	44	40	37	43.17
准确率	0.896	0.908	0.92	0.912	0.92	0.926	0.9136

数据描述：

1. 判断为中性的准确率、召回率、F 值都在 95%以上。
2. 判断为积极的准确率、召回率、F 值在 83% - 92%之间。
3. 判断为消极的准确率、召回率、F 值在 85% - 92%之间。
4. 所有情感分析的准确率为 91.36%。

结论：

1. 对比基于词典规则的实验和机器学习的实验，我们得到：

① 基于机器学习的方法在准确率、召回率、F 值三方面远高于基于词典规则的方法。

② 基于词典规则的方法平均正确率为 68。6%。

③ 基于机器学习的方法平均正确率为 91.36%。

2. 对比基于机器学习方法的 A 实验和 B 实验，可以看出，随着训练语料的增大，准确率、召回率、F 值有明显提高。

五、结语

互联网在发展过程中，不断尝试理解它的用户。起初，要求用户有意识的输入，由用户自己选择自己的所需；到后期的推荐、点评、大数据分析，情感分析，对用户需求的理解精度进一步提升。随着机器学习方法的愈加成熟，情感分析的准确率会越来越高，从而使产品的用户体验更加人性化。