结合博客情感分析
词汇知识与文本分类

Prem Melville IBM
TJ Watson 研究中心。
邮政信箱 218
Yorktown Heights,NY 10598
pmelvil@us.ibm.com

Wojciech Gryc Richard D. Lawrence牛津大学计算实验室 IBM TJ Watson 研究中心。
Wolfson Bldg, Parks Rd PO Box 218 Oxford OX1 3QD,英国 Yorktown
Heights,NY 10598 wojciech.gryc@trinity.ox.ac.uk ricklawr@us.ibm.com

抽象的

网络上用户生成内容的激增为公司带来了新的机遇和重大挑战,公司越来越关注监控围绕其产品的讨论。跟踪网络日志上的此类讨论,可以提供有关如何改进产品或更有效地营销产品的有用见解。此类分析的一个重要组成部分是描述博客中表达的有关特定品牌和产品的情绪。情绪分析专注于自动识别一段文本是否表达了对该主题的正面或负面意见的任务。该领域的大多数先前工作都使用先前的词汇知识来判断词语的情绪极性。相比之下,一些最近的方法将该任务视为文本分类问题,他们仅根据标记的训练数据来学习对情绪进行分类。在本文中,我们提出了一个统一的框架,在该框架中,人们可以使用词类关联的背景词汇信息,并使用任何可用的训练示例针对特定领域细化此信息。

在不同领域的实证结果表明,我们的方法比单独使用背景知识或训练数据以及在文本分类中使用词汇知识的替代方法表现更好。

类别和主题描述I.2.6 [人工智能]:学习;I.5.1 [模式识别]:模型

一般条款

算法、经济学、实验

1. 引言

过去十年,互联网为各种规模的公司创造了巨大的机会,使它们能够接触客户、宣传产品和开展业务。在这种成熟的商业模式下,公司可以在很大程度上控制其通过网站上的内容,企业可以提高自己的网络声誉。Web 2.0 强调促进信息共享和用户协作的网站,正在迅速改变这一格局。越来越多的公司产品组合各方面的讨论焦点从单个公司网站转移到协作网站、博客和论坛,在这些网站上,任何人都可以发表评论,这些评论可能会影响大量潜在买家的看法和购买行为。这显然引起了营销组织的关注,不仅因为负面信息的传播难以控制,而且由于博客、论坛和其他 Web 2.0 现象的迅猛增长,甚至很难发现它。这种担忧催生了“buzz”一词,并引入了几个网站(例如 [2])和系统(例如 [27])来监控特定公司和产品品牌的博客声誉趋势。

对博客相关话题的自动分析提出了几点
从营销角度来看有趣的问题:

1. 给定大量(约 1 亿个)博客,我们如何才能找出不仅讨论特定产品而且讨论与该产品有某种相关的更高级概念的博客和论坛子集?

2. 确定了相关博客子集之后,我们如何确定这个领域中最权威或最有影响力的博客作者?

3. 我们如何检测和描述博客或论坛中提到的实体(例如产品)所表达的特定情绪?

这些营销问题中的每一个都为数据挖掘社区提出了有趣的技术问题,我们正在将其作为更广泛议程的一部分来解决。在本文中,我们仅关注情感分析问题。在博客分析的背景下,情感分析的目标是确定博客中包含的部分文本所表达的总体态度。文本可以是完整的博客文章,也可以是在提及特定实体(例如公司或产品)附近提取的片段。

KDD 09,2009年 6 月 28 日至 7 月 1 日,法国巴黎。

情感分析中的大多数先前工作都使用基于知识的方法,根据定义单词情感极性的词典和简单的语言模式对文本情感进行分类。最近,有一些研究采用机器学习方法 [20, 8],并构建基于已被训练的文档进行训练的文本分类器。

人类将其标记为正面或负面。基于知识的

方法不能很好地适应不同的领域,而
学习方法在人工注释文档方面需要付出很多努力。在本文中,我们提出了一种新的

机器学习方法可以克服这些缺点
通过有效地结合背景词汇知识和
监督学习。特别是,我们构建了一个基于充满情感的单词词典的生成模型,并且

第二个模型在标记文档上进行训练。然后自适应地汇集这两个模型的分布

创建复合多项式朴素贝叶斯分类器
捕获两个信息源。通过利用
我们大大减少了先验词汇知识的数量
所需的训练数据。此外,通过使用一些标记的
我们能够提炼背景知识的文档,
它基于通用词典,从而有效地适应
到新的域。我们在三个截然不同的领域展示了我们方法的通用性 博客讨论

企业软件产品、讨论美国的政治博客
总统候选人和在线电影评论。实证结果表明,我们的池化多项式方法比使用词汇

知识和标记数据,以及在半监督环境中使用背景知识的替代方法。

2. 基线方法

出于本文的目的,我们将情绪检测
作为将文档分为积极和消极情绪类别的二元极性分类。在本节中,我们描述了在此类文档分类中使用背景知识的两种基本方法。

2.1 词汇分类

在域中没有任何标记数据的情况下,可以
建立仅依赖背景知识的情绪分类模型,例如定义极性的词典

的话。给定积极和消极术语的词典,
使用此信息的一种直接方法是
测量这些术语在每个术语中出现的频率
文档。测试文档 D 属于的概率
正类可以计算为 P(+|D) = a+b ;其中a和b分别是文档中正面和负面术语的出现次数。 A

如果 P(+|D) > t,则将文档分类为正类,其中
t 是分类阈值;否则,该文档为
归类为负面。在缺乏任何关于术语相对积极性和消极性的先验信息的情况下,我们使用

t = 0.5,即,如果存在
文件中积极的术语多于消极的术语。我们将这种分类方法称为词汇分类器,

并将其用作我们的基线之一。

在本研究中,我们使用了 IBM 生成的词典
印度研究实验室是为其他文本挖掘应用程序开发的[22]。它包含 2,968 个单词

人类标记为表达积极或消极情绪。
总共有 1,267 个正面独特评论和 1,701 个负面独特评论
经过词干提取后的术语。需要注意的是,此列表
在构建时没有考虑特定的领域;
是使用训练示例进行学习的进一步动机
特定领域的内涵。

2.2 特征监督

纯词汇分类方案的替代方案
是在半监督学习环境中使用词典和未标记数据。目前很少有方法

将这些背景知识融入学习中,
我们将在第 5 节中列出。这些方法中的大多数都创建了
基于背景知识的伪示例,以及
然后使用现有的学习算法。在这里,我们描述了
刘等人的方法。 [15]因为它与我们的工作关系最密切;因为他们也使用朴素贝叶斯分类

并通过词类关联(即标记特征)利用背景知识。然而,与我们不同的是,他们

使用未标记的文档进行学习。给定每个类别的代表性单词集(即词典),他们创建一个

包含所有代表性单词的每个类别的代表性文档。然后他们计算余弦相似度

未标记集中的每个文档与代表性文档之间。他们分配每个未标记的文档

到相似度最高的类,然后训练
使用这些伪标记示例的朴素贝叶斯分类器。
刘等人。证明他们的方法表现更好
比仅使用词汇分类器。我们将他们的方法称为特征监督并将其作为基线

使用标记特征和监督学习
方法。

3. 池化多项式

庞等人。 [20]已经表明,仅使用词汇信息对于情感分类不如构建

来自训练示例的机器学习模型。然而,
完全忽略了提供的背景知识
用词典代替训练数据可能不是最佳的。作为一个
替代方案是,我们提出池化多项式分类器
它提供了一个框架,人们可以在其中构建一个复合朴素贝叶斯分类器,该分类器结合了背景
知识和训练示例。

下面我们介绍一下多项式的一些基本概念
朴素贝叶斯文本分类,描述所需
我们的框架。有关文本事件模型的更多详细信息
朴素贝叶斯分类器的分类和归纳可以是
见于[16]。

常用的多项式朴素贝叶斯分类器
文本分类依赖于三个假设[18]:(1)
文档由混合模型生成; (2) 有
每个混合成分与类别之间一一对应,并且 (3)每个混合成分都是单词的多项分布,即给定一个类别,该类别中的单词

一份文件是彼此独立制作的。

基于这个生成模型,文档的可能性 (D) 是所有混合的总概率之和

分量,即 P(D) = P(D|cj)P(cj);其中 P(cj) 是
类别cj的概率, P(D|cj) 是概率
给定类别的文档。我们指的是类的集合
为 C,对于二元情绪分类,其为 {+, −}。
计算文档的可能性,我们做了天真的
假设生成文档的单词 (wi)
独立地,因此文档 D 的概率是
在给定类cj中生成的概率为 P(D|cj) = 朴素贝叶斯分类规则使用 P(wi|cj)

贝叶斯定理
计算每个类别的类别成员概率,

$$P(cj|D) = \frac{P(cj) \cdot P(wi|cj)}{\text{压力(D)}} \quad (1)$$

并预测可能性最高的类别,即

$$\text{argmax}_{cj} P(cj) \prod_{i} P(wi|cj) \quad (2)$$

3.1 结合概率分布

池化分布是结合
来自多个来源或专家的信息;专家在哪里
通常以概率分布的形式表示。此类专家概率分布的组合一直是风险领域的研究热点

分析[3]。由于我们处理的是文本分类,
在我们的设定中,“专家”可以表示为多项式
概率分布与朴素贝叶斯分类中一样。我们考虑两个专家 一个基于标记训练数据学习,
另一个代表生成

解释词典的模型。我们将讨论后者,
下一节的背景知识模型;但对于
现在假设我们有这样的多项式参数
模型。

在本文中,我们只考虑两个来源的情况
信息,即目标域中的标记示例
和背景词汇知识。然而,池化多项式是一个通用框架,可用于组合多个多项式模型 - 这
些可以导出

来自不同相关领域或不同背景知识来源的训练数据。

概率数学组合文献
分布,包括 Winkler [30]、Gen-est 和 Zidek [11] 以及 French
[10] 的评论。在这里,我们比较了两个
公理方法,即线性意见池和
对数意见池。

线性意见池是一种聚合概率分布的有效方法,其历史可以追溯到拉普拉斯 [3]。

在这种方法中,总概率:

$$P(wi|cj) = \sum_{k=1}^{\text{钾}} \alpha k P k(wi|cj) \quad (3)$$

其中 K 是专家数量; $Pk(wi|cj)$ 表示
专家 k 为单词wi分配的出现在 a 中的概率
cj类文档;权重αk之和为 1。

另一种典型的组合方法是对数
意见池,使用乘法平均。在这个
逼近组合概率:

$$P(wi|cj) = Z \prod_{k=1}^{K} Pk(wi|cj)^{\alpha k} \quad (4)$$

其中 Z 是归一化常数,权重αk满足
确保$P(wi|cj)$ 是概率分布的限制。通常,权重之和被限制为一。

如果权重相等 (1/K),则 Log Pooling 等效
取各个分布的几何平均值。
这个方案之所以被称为日志池化,是因为
组合分布可以表示为各个分布的对数的线性组合。

对于这两种池化方案,我们根据专家解释训练数据时的错误来计算各个专家的权
重。具体来说,我们使用 S 型加权方案,

在哪里:

$$\alpha k = \text{对数} \frac{1 - \text{呃}}{\text{埃克}} \quad (5)$$

其中 errk是专家 k 在训练集上的错误;
αk被标准化为总和为 1。这是一样的
用于在 boosting 中组合加法模型的加权方案 [23]。

为了从训练数据中学习模型,我们根据观察到的词频计算条件$P(wi|cj)$,如下所示

标准朴素贝叶斯分类。此外,对于 Pool-ing Multinomials,我们需要构建一个
多项式模型
代表我们描述的背景知识
下一节将详细介绍。

3.2 生成背景知识模型

引入多项式朴素贝叶斯分类器涉及估计模型参数 $P(Cj)$ 和$P(wi|cj)$。在里面

由于缺乏关于类别分布的背景知识,我们从训练数据中估计类别先验 P(Cj)。因此对于
背景模型,我们只关注

给定类别的每个单词的条件概率。
我们假设
该词典是由人类专家隐式得出的
通过检查大量正面和负面情绪文档。因此,我们尝试选择

生成此类文档的多项分布。这些条件的确切值是

下面,基于一组属性,这些分布必须
满足。为了帮助我们推导,我们首先在下面建立一些重要的符号。

定义:
V – 词汇表,即我们领域内的单词集
P – 存在于 V 中的词典中的一组正项
N – 词典中存在于 V 中的一组否定词
U – 未知项集,即 V − (N ∪P)
m – 词汇量的大小,即 |V|
p – 正项的数量,即 |P|
n – 否定项的数量,即 |N |

性质 1:由于我们不知道相对极性
词典中的术语,我们假设所有正向术语都是
同样可能出现在积极的文件中,并且相同
对于负面文件来说是正确的,即

$$P(wi|+) = P(wj|+), \forall wi, wj \in P$$
$$\text{且} P(wi|-) = P(wj|-), \forall wi, wj \in N \quad (6)$$

我们指的是任何正项出现的概率
在正文档中,简单表示为P(w+|+)。同样地,我们
指的是任何否定项出现在a中的概率
负面文档为P(w−|−)。

此外,在没有任何文字知识的情况下
那些不在我们的词汇表中的,我们平等对待它们
类,即

$$P(wi|+) = P(wj|+), \forall wi, wj \in U$$
$$\text{且} P(wi|-) = P(wj|-), \forall wi, wj \in U \quad (7)$$

我们参考这些非词典的条件概率
项为P(wu|+)和P(wu|−)。

性质 2:如果文档Di有 α 个正项和 β
负项,文档Dj有 β 个正项,并且
α 为负项,我们希望Di被视为可能
是一份积极的文件,因为Dj很可能是一份消极的文件
文档。使用 (1) 和这个对称性给出

我们提出以下要求：

$$[P(w+|+)]\alpha[P(w-|+)]\beta = [P(w-|-)]\alpha[P(w+|-)]\beta$$

$$\Rightarrow \frac{P(w+|+)}{P(w-|-)}^{\alpha} = \frac{P(w+|-)}{P(w-|+)}^{\beta}$$

为了使所有 α 和 β 值都成立，我们需要

$$P(w+|+) = P(w-|-)$$

$$且P(w-|+) = P(w+|-) \quad (8)$$

也就是说，一个正项出现在
肯定性文件与否定性术语相同
出现在负面文件中；同样如此
用于具有相反极性的文档中出现的术语的条件概率。

性质 3：由于积极的文件更有可能
包含正项多于负项，反之亦然，
我们想：

$$P(w+|+) = r \times P(w-|+)$$

$$且P(w-|-) = r \times P(w+|-)$$

$$其中\ 0 < 1/r \leqslant 1 \quad (9)$$

我们将因子 r 称为极性水平 它衡量一个正项出现的可能性有多大。

与消极术语相比，出现在积极的文件中。

属性 4：由于我们的混合模型的每个组件都是
概率分布，我们有以下约束
每个类别的条件概率cj：

$$\sum_{i}^{*} P(wi|cj^{*}) = 1 \quad (10)$$

使用上述属性作为约束，我们现在可以得出
适合我们的背景知识的价值观
模型。

由(10)可知

$$pP(w+|+) + nP(w-|+) + (m-p-n)P(wu|+) = 1 \quad (11)$$

这给我们带来了以下不等式

$$pP(w+|+) + nP(w-|+) \leqslant 1$$

$$\Rightarrow pP(w+|+) + \leqslant 1 \quad \frac{nP(w+|+)}{r}$$

使用(9)。

由于 $0 < 1/r \leqslant 1$，因此，

$$P(w+|+) \leqslant p+n \quad \frac{1}{}$$

因为我们想分配最大概率质量
对于词典中的已知术语，我们将P(w+|+)设置为
可能的最大值，即

$$P(w+|+) = p + n \quad \frac{1}{} \quad (12)$$

现在，由（8)和（9)可知

$$P(w-|-) = p + n \quad \frac{1}{}$$

$$P(w+|-) = P(w-|+) = p + n \quad \frac{1}{} \times \frac{1}{r} \quad (13)$$

现在，求解（11)（类似地求解负
类），我们得到以下未知的条件
条款：

$$P(wu|+) = (p \quad \frac{n(1 - 1/r)}{+ n)(m - p - n)}$$

$$P(wu|-) = \frac{p(1 - 1/r)}{(p + n)(m - p - n)} \quad (14)$$

使用（12）、(13)和（14），我们现在有了所有必要的条件
参数来代表我们的背景知识。

4.实证评估

在本节中，我们展示了
池化多项式在情感分析中的应用
不同的域。

4.1 数据集

如前所述，我们的激励应用程序是自动分析与产品相关的博客文章

和/或品牌名称。为此，我们创建了一个
与 IBM Lo-tus 软件品牌相关的一组带标签的情绪示例。虽然我们主要感兴趣的是

科技博客的感悟，我们也应用了我们的
描述博客讨论中的情绪的方法
具体的政治候选人。

博客情感数据收集的一个重要部分
分析是从下载的网站中提取相关文本。博客本质上更加多样化，

布局和结构比电影或产品评论更合理。此外，许多博客都有大量评论

（通常来自博主以外的个人)以及
明确的引用。评论和引用常常表现出与主要内容完全相反的情绪。

然而，核心内容之间的自动区分，
引用和评论非常困难。我们使用 [9] 提供的算法，仅从部分文本中提取文本

HTML标签与文字比例在以上的网页
最低门槛。

Lotus 博客：我们的目标应用程序是检测情绪
围绕企业软件，特别是 IBM Lotus 协作软件。为了做到这一点，我们一直在监测 20,488
个技术博客，目前这些博客包含超过

170 万个帖子。我们创建的带标签的Lotus数据集，
由 14 个博客的帖子组成，其中 4 个积极发布针对该品牌的负面内容，其余

倾向于撰写更多积极或中立的帖子。在这些数据中
负面的博客文章经常抱怨用户界面挑战或软件错误。例如，一条评论

就像“有人可以告诉我为什么 Lotus Notes 需要
我的 CPU 使用率达到 99%？”可以被视为负面的，而
“达摩演示提供了一系列选择莲花的理由”是积极的。 Lotus数据是通过下载最新的

来自每个博主的 RSS 提要的帖子，或访问该博客的
档案，如果存在的话。然后阅读并标记每篇帖子
手动将其标记为正面的、负面的、中性的或不相关的。
对于我们的分析，仅保留正面和负面帖子，分别创建 34 个和 111 个示例的标记集。由
于一些博主倾向于始终如一地展示

正面或负面的情绪，只显示每篇帖子的正文
用于分析，从而避免博客标题和用户名等重复出现的信息，这些信息最终可能会

导致模型过度训练。

政治候选人博客:关注政治的帖子
取自不断更新的 16,741 条政治
博客,其中包含超过 200 万个帖子。我们专注于我们的
对随机选择的包含以下内容的帖子进行标记工作
在他们的网址中使用“奥巴马”或“克林顿”。与以 Lotus 为中心的帖子不同,政治帖
子全部来自不同的博客。
此外,从人类标注者的经验来看,政治情绪似乎更难标注

与软件评论相比,因为帖子往往更加情绪化,
讨论的问题只与候选人隐性相关(例如经济或外交政策),也可能使用文化参考来做
出判断。一篇帖子被标记为具有

对特定候选人的积极或消极情绪
(巴拉克·奥巴马或希拉里·克林顿)如果明确提到
候选人的正面或负面评价。客观陈述和报纸及其他来源的引文

被忽视。同样,如果博主对候选人做出含蓄的陈述(例如讨论种族主义或性别歧视),

选举中没有具体提到候选人),
帖子不会与情绪相关联。本质上,
只有对候选人有明确意见的帖子才会被标记并包含在该分析中。例如,“我认为

希拉里·克林顿在这场辩论中表现不佳。”将是一个
克林顿被标记为负面。另一方面,“奥巴马
列出了主要筹款人的名字,提供了比克林顿更详细的细节。”
会被视为中立,因为读者无法分配
不做个人价值判断,
陈述。在整个标签中,只有正面和负面
帖子被保留 那些被标记为中立和不相关的帖子被丢弃。同样,如果帖子被视为负面的

对一位候选人持积极态度,对另一位候选人持积极态度,那么
帖子也被废弃了。虽然在这样的情况下丢弃帖子
如果要部署分类器,这种方式就不是一种选择
制作环境,如此严格的后期流程
选择用于构建一个干净的测试集来评估我们的
方法。最终的政治数据集包含 49 个积极的
以及 58 条负面帖子。

电影评论:除了我们生成的博客数据外,我们还使用了电影的公开数据

Pang 等人提供的评论。 [20]。该数据包括
来自互联网的 1000 条正面评论和 1000 条负面评论
电影数据库。评论被贴上正面标签
评分高于 3.5 星且带有负面标签的
分配给其余的。

4.2 结果

我们将我们的方法,线性池化和日志池化与词汇分类器、特征监督进行了比较,如刘

等人[15],以及仅使用 Na¨ ve
贝叶斯分类器,如 Pang 等人 [19] 所述。对于朴素贝叶斯
模型中,我们使用 Lidstone 平滑 [14] 计算条件概率估计,其中 = 1 × 10−6,即

$$P(w_i|c_j) = \frac{t_{ij} + }{\sum_{i} t_{ij} + |V|} \qquad (15)$$

其中tij是单词出现的总次数
wi在属于cj类的所有文档中。在实践中,它是
通过使用 L2 对表示文档的每个词向量进行归一化,可以更好地控制不同的文档大小

规范。因此,我们使用这些标准化的词频
在我们的实验中。上述平滑考虑了 -

训练集中没有出现的高频词,但是
出现在测试集中。在 Lidstone 平滑中使用 << 1
因为单词概率往往比使用拉普拉斯平滑 (= 1) [1] 更适合文本分类。

对于池化多项式,我们设置极性级别,r =
100 即,词典中的肯定词被视为 100
出现在正面文件中的可能性比
一个否定词。 4.5 节介绍了对此参数的敏感性分析。我们比较了数据集上的不同情感
分类方法

第 4.1 节;我们通过过滤对数据进行预处理
使用 Porter 词干提取器 [21] 去除停用词和词干词。我们还删除了出现在

少于 3 份文件。
为了评估不同数量的训练数据的效果,我们生成多个学习曲线的平均值

运行 10 倍交叉验证。对于特征监督,我们
在学习的每个阶段使用可用的训练数据
曲线作为未标记数据池。对于 Lotus 和 Movies,
我们进行了 10 次交叉验证;对于更吵闹的
政治集我们使用了 20 次运行。由此产生的学习曲线
如图 1 所示,最后一点的准确率
总结在表 1 中。为清楚起见,我们仅列出
图中最相关的曲线。除了准确度之外,
我们还比较了 ROC 曲线下面积 (AUC)。
AUC 的结果没有在这里呈现,因为它们是相似的
到结果的准确性。

表 1:比较不同情绪分类方法的准确率。

模型词汇分	莲花政治电影		
类器 68.23 特征监督 57.93 朴素贝叶斯		55.20	63.40
88.40 线性池化 91.21 日志池化 88.42		46.19	57.59
		59.24	80.81
		63.61	81.42
		60.04	80.00

结果清楚地表明,对于所有数据集,通过

池化多项式框架比单独使用每个信息源表现更好。特别是,

线性池化表现最佳,其中所有改进
与其他相比,准确率具有统计学意义
根据配对 t 检验方法(p < 0.05)。一般来说,
日志池也有效,但改进不
与通过线性池化获取分布的凸组合一样重要。在本文的其余部分中,

除非另有说明,我们将使用池化多项式
参考线性池化。
学习曲线表明,正如预期的那样,很少有
训练样例最好只依赖背景
知识,而不是尝试估计模型参数
一组有限的标签。然而,随着数量的增加
标记数据后,就可以开始构建朴素贝叶斯模型
比词汇分类器更好。此外,
结合背景知识和训练数据
正如池化多项式始终是更好的选择。
多项式池化的影响最为显著
当有标记的示例很少时。然而,
从我们的数据分布中提取的标记示例的数量最终应该能够捕获所要提供的信息

通过背景知识来呈现。我们可以在电影中看到这种现象的开始,相对的

随着训练样例数量的增加,池化多项式相对于朴素贝叶斯的改进会减弱。

然而,在需要用
对于人类专家来说,提供数千个标签可能会很繁琐。
在这种情况下,能够通过少量训练示例实现高精度是一个显著的优势。例如,在电影中,
构建的朴素贝叶斯模型的精度

使用 800 个训练示例,可以通过 Pooling 实现
使用约 300 个示例的多项式。这个可以翻译
大大减少了人工注释的劳动密集型过程。使用与领域无关的资源,如

作为情感词典,也让我们能够快速适应
通过提供一些特定领域的训练示例来完善我们的背景知识,从而帮助
我们应对新领域的挑战。
仅通过词汇分类器使用词典不会
在我们的数据集上表现不佳。对于政治和
电影的准确率比基本率好一点
(多数阶层比率)分别为54%和50%;然而
对于 Lotus 来说,其准确率低于 77% 的基本准确率。
词汇分类器的基本假设是,如果词汇表中有更多积极的词汇,则该文档是积极的

比文档中的否定词多。除了
该词典并未涵盖所有可能出现的术语
在我们的词汇中,它也没有捕获特定领域的
术语的内涵。词汇分类器也无法
考虑积极和消极情绪的程度
与每个术语相关联。然而,很明显这些
纯词汇分类器的弱点可以通过以下方式克服
从一些训练示例中学习。培训对背景知识的影响更详细的分析为

第 4.4 节。
我们还观察到,特征监督方法
刘等人的表现相当糟糕,甚至比
词汇分类器。这与之前的结果相反
Liu 等人 [15]。然而,在他们的研究中,他们使用了非常小的
标记特征词典 – 少至 5 个,最多 30 个
每堂课。相比之下,在情感分类领域
包含数千个词类关联的大型词汇表
很容易获得,词法分类器分类器执行
更好。此外,正如之前指出的,我们
用途是通用的,需要针对每个领域进行改进。在特征监督中,从嘈杂的例子中学习

仅通过这个与领域无关的词典进行标记就可以清楚地
性能进一步恶化。
对于我们的博客域,我们注意到 Lotus 的情感分类准确率明显高于

对于政治。 Lotus 帖子源自一小部分
博客,博客和情感之间几乎是一一对应的,即每个博主要么喜欢,要么喜欢

讨厌 Lotus。这更符合
我们的生成模型与政治帖子相比,
源自许多不同的来源,具有积极的
以及对不同候选人的负面评论。因此,我们的方法在分类方面表现更好

Lotus 博客中的情绪与政治博客相比。
关于电影评论情绪分类的挑战,可以在[20]中找到一个很好的讨论 我们讨论

对情绪进行分类的一些挑战
对于下面的博客。

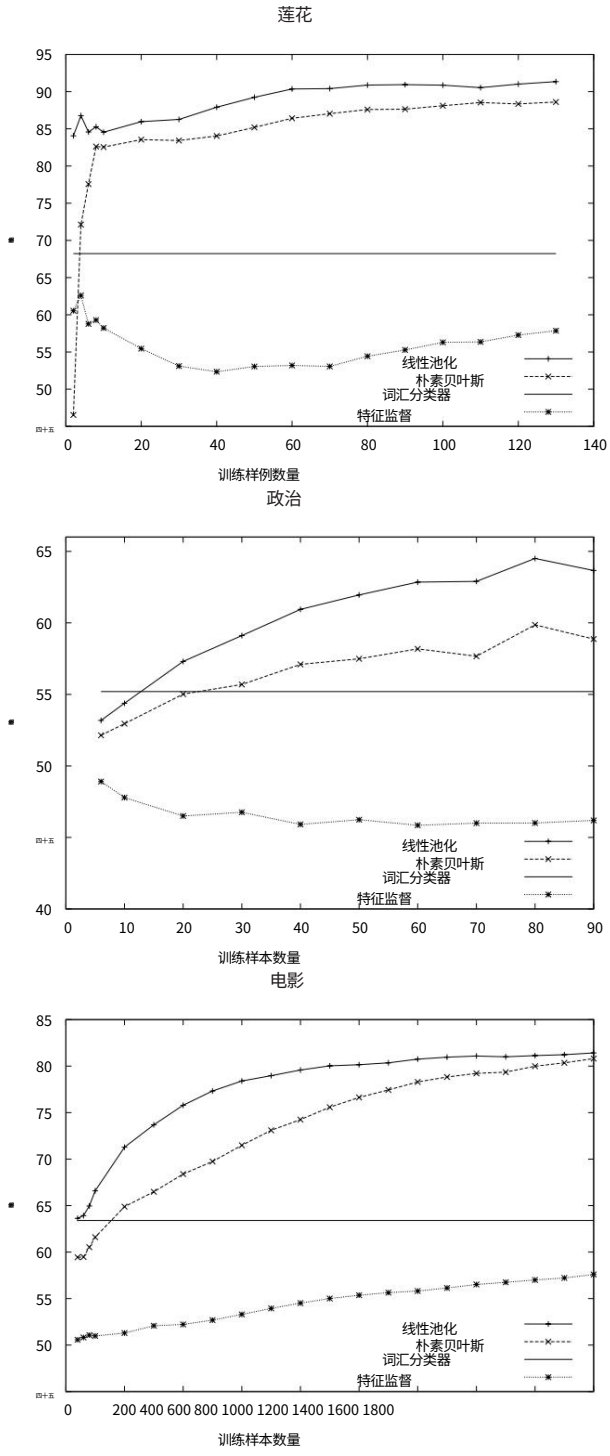

图 1:比较不同的情绪分类方法。

4.3 博客情绪分析面临的挑战

即使对于人类来说,将博客文章标记为正面或负面也非常复杂。在努力集中精力的
同时
分类帖子时仅根据帖子内容,仍然是
评论、引文和引用来自其他人
来源包含在帖子的主体中。当一个

博客文章页面成为某个讨论区域
主题,将整个页面标记为正面或负面可以
是相当困难的。对于政治博客来说尤其如此,
作家经常将多个候选人、政策或事件进行比较。情绪分析的问题

博主经常使用的事实使情况变得更加复杂
笑话、轶事和文化参考来说明他们的
观点,使得不熟悉相关事实或参考资料的人无法理解标记任务。这使得大多数算法的情感分类变得极其困难。

在相关工作中,人们观察到情绪分类器在技术性更强的话题上的表现往往比

社交或创意的。例如,Turney [28] 认为汽车和银行评论的准确度分别为 84% 和 80%,

分别,并且相同的方法提供了准确度
电影只有 66%,旅游目的地只有 71%。同样,在我们自己的结果中,政治帖子分类趋向于
比给以 Lotus 为中心的帖子贴上标签更糟糕。什么是
然而值得注意的是,我们提出的方法仍然
在两个领域都取得了更好的结果。

4.4 训练样本与背景知识

我们的结果表明,通用词典是一种非常有用的
不应忽视的先前知识来源,
代替训练数据。然而,这提出了一个问题
监督学习究竟如何影响我们的背景知识;以及一个单一的、全面的词典能否

解决我们所有的问题?
我们声称监督学习可以完善我们的知识
关于术语的情感极性;特别是,
帮助我们学习特定领域的内涵。我们可以通过检查我们的背景元素来支持这一说法

被训练样例改变的知识。例如
通过比较池化,可以轻松获得洞察力
多项式模型与背景知识模型
并确定哪些词典术语变化最大
情绪上戏剧性地。我们可以直接测量这一点,通过
计算给定类别的单词的条件概率的对数几率比的差异。对数几率

单词的log(P(wi|+)/P(wi|−))衡量了指示性
一个词属于正类文档。一个
可以在我们最终的池化多项式模型和背景知识中计算所有词典单词

模型。现在可以计算情绪偏差的变化
作为有和没有训练的对数赔率之间的差异。正值表示对应的词

积极情绪的权重增加;负值意味着它的权重被降低了

积极的情绪。
表 2 列出了排名前 10 的上调和下调权重术语
对于每个领域。这让我们对某些术语的领域特异性有了一些了解,这是不可能编码的

成一个通用词典。例如,单词
诸如战争、黑暗和复杂等元素在电影描述中可能与积极的体验相关,尽管它们可能

在其他情况下通常被认为是负面的。这
积极词汇的权重降低,例如天赋和
对电影的期望也与 Pang 等人[20] 在本研究中观察到的“期望受挫”的叙述相一致。

数据。
虽然表中的一些积极或消极的术语
2 可能看起来违反直觉,但加权和减权

表示模型在特定领域的学习讨论模式。例如,在没有特定语境的情况下,真理没有负面含义,但在特定语境中权重较低。

政治为中心的模型。通过探索包含
这个词,人们可以理解为什么 看似积极的术语是
用于讽刺和指责的帖子,或用于负面
语境。例如,“编造事实”、“转变谎言”
成真理”,甚至是特定的上下文,例如“有很多
赖特布道的真理”,这个话题通常包含
对巴拉克·奥巴马的负面情绪。

表 2:词典中最常出现的术语
根据训练示例增加或减少权重。

电影强调战争、缺陷、社会、黑暗、复杂、愤怒,	罐子, 急躁, 捕获, 外星人
天赋、理性、承诺、拯救、公平、本能、求爱、坚定、救赎	
政治加权反对,服务,边缘,辩论,难以置信,战争,服务,意味着,争夺,	戳
降低权重的真相、订阅、关联、连接、	团结、自由、明确、坚持、接受、公司
莲花加重酷,提问,错过,盗窃,奔跑,对不起,	诡计、社交、愚蠢、服务
贬低附加、建立、起源、包含、纠正、推理、选择、感知、给予、告知	

4.5 敏感性分析

池化多项式中的唯一参数是背景知识模型中使用的极性级别 (r);

在本节中我们将探讨其对该参数的敏感性。
这个极性水平是衡量它的可能性有多大
一个积极的术语出现在积极的语境中比较
为否定项,反之亦然。我们运行了所有测试
电影数据,因为它是最大的并且提供结果
方差最小。我们测量了模型的准确率
对于 2 到 1000 范围内的不同 r 值。结果,
如图 2 所示,表明池多项式相当
对该参数的变化具有鲁棒性 仅具有准确性
从 81.33 到 81.78 不等。实际上,我们所有实验都将此参数设置为 100。

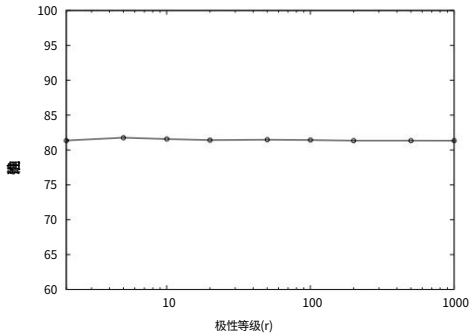

图 2:评估池多项式对极性水平参数的敏感性。

5. 相关工作

情绪分析的大部分工作都集中在识别在线文本段落中的积极或消极情绪。这些研究大致可分为两类:基于知识的方法和基于学习的方法。基于知识的方法主要使用语言模型或其他形式的背景知识来

对段落的情感进行分类。这方面的重点
领域是使用和生成词典来捕捉
言语的情感。这些方法包括手动
开发领域相关词典的方法[5]
半自动化方法 [12, 33, 13, 32],甚至几乎完全自动化的方法 [28]。正如 Ng 等人所观察到的
等人。 [17],大多数半自动化方法产生的词典并不令人满意,要么覆盖率高,精度低,要么
反之亦然。最近,Pang 等人。 [20]成功应用机器学习方法来分类情感

电影评论。他们将问题视为文本分类任务,使用每部电影的词袋表示

审查。他们证明了学习方法比简单地计算积极和消极的效果更好

使用手工编写的词典来识别情绪术语。然而,
正如我们在本文中提出的那样,他们没有考虑将此类背景词汇信息与监督学习相结合。他们的结果还表明,使用更复杂的

语言模型,结合词性和 n-gram
语言模型,不会比简单的单元模型有所改进
词袋表示。根据他们的发现,
我们还采用一元文本模型。庞等人。 [19]通过首先将句子分类为主观来扩展他们的工作
与客观句子相比,然后仅根据情绪极性对主观句子进行分类。他们证明

仅关注每条评论中的主观句子
他们能够提高整体情绪分类的准确性。这种两阶段方法也可以改善我们的结果;然而,在本文中,我们只关注

我们在极性预测阶段的进展。杜兰特和
Smith [8] 还将文本分类应用于分类
政治博客文章。虽然他们的数据和我们的差不多,
他们的任务是确定左翼与右翼的政治联盟
与我们识别积极和积极的目标有很大不同
负面情绪。使用朴素贝叶斯分类器
通过前向特征选择,他们能够超越
支持向量机。鉴于他们的成功,我们还使用朴素贝叶斯方法。威尔逊等人。 [29]还制定了情感检测
作为监督学习任务。然而,不只是使用
在文本分类方面,他们专注于语言特征的构建,并使用 Boostexter 训练分类器。结合语言规则方面的背景知识

这种分类器是未来工作的一个有趣的方向。
最近,人们对使用
监督学习中的背景、先验或领域知识 包括使用人类提供的特征与特定类别的关联的方法。这方面的大多数工作

专注于使用特征的此类先验类别偏差来生成
标记的示例然后用于标准监督
Schapire 等人 [24] 提出了一个这样的框架
用于增强逻辑回归,使用手工制定的规则
从相关特征列表生成来标记伪示例。他们将提升目标函数修改为

拟合训练数据以及基于这些数据的先验模型
伪例子。提供一些相关的功能
对于每个类别,Wu 和 Srihari [31] 为 unla- 分配标签

附带的文件,然后与
标记示例来构建加权边际支持向量机。与上述方法不同,我们使用特征类关联来直接构建生成模型。

Dayanik 等人 [6] 探索了将先验知识纳入 Logistic 回归的几种方法。在他们的研究中,
人类注释的相关特征被赋予了更多能力
通过分配更大的先验模式来影响分类,或者
方差比其他特征更大。他们的模式设置方法
与我们最接近。然而,他们报告说
这种方法不可靠,因为它偶尔会产生
最好的,但通常会产生最差的结果
其他方法。德鲁克等人。 [7]通过标记特征结合先验知识,这些特征用于直接

限制模型对未标记实例的预测。
他们的广义期望标准方法适用于任何判别概率模型,并且他们

展示其专门针对多项物流的实用性
回归。与他们的方法不同,该方法只使用未标记的
实例中,我们的方法受到监督并使用背景
带有标记实例的知识。然而,我们可以延长
我们的方法还利用未标记的数据,如中所述
第 6 节 Sindhwani 和 Melville [26] 提出了一种方法
合并标记特征和未标记文档
在标准正则化最小二乘法中。在以下设置中
标记数据非常有限,未标记数据丰富,
他们的方法比纯监督方法表现更好,
竞争性半监督技术。在文本分类之外,将背景知识纳入学习的研究也已开展。值得注意的是,Shavlik [25]

探索了基于知识的先验知识的使用
神经网络。

6. 未来的工作

在监督学习中使用背景知识是其中之一
减少在目标域中标记许多示例的负担的方法。另一个信息来源

可以利用的,在相关领域中被标记为示例。例如,软件评论的集合可能

捕获类似的依赖于领域的情感表达
也用于技术博客讨论。因此,
我们未来工作的重点是构建更好的分类器,
利用背景知识和标记数据
其他领域。最近在
迁移学习可以应用于扩展
基于背景知识的模型来整合数据
来自不同领域。这种
迁移学习负责训练和测试集
来自不同的发行版。戴等人。 [4] 提供一个
使用文本分类中的迁移学习解决方案
基于 EM 的朴素贝叶斯分类器。他们的解决方案首先估计分布Dl下的初始概率

标记数据,然后使用 EM 算法修改
使用未标记实例的测试分布Du模型
来自测试集。这个朴素贝叶斯转移分类器可以
很容易添加到 Pooling Multinomi-als 中的专家集合中,从而结合背景知识和迁移
在单一框架内学习。

七、结论

在本文中,我们做出了两个主要贡献。首先,
我们开发了一个有效的框架来整合词汇

监督学习中文本分类的知识。
其次,我们成功地将开发的方法应用于情绪分类任务 扩展了该领域的最新技术,该领域主要侧重于单独使用背景知识或监督学习。实证结果表明,即使只有几个训练示例,我们也可以在我们的框架中将背景词汇信息与监督学习相结合,从而产生比单独使用词典或训练数据更好的结果,以及在半监督设置中使用词典和未标记数据的方法。

尽管本文的主要重点是情感分析,但所开发的方法适用于任何可以获得一些相关背景信息的文本分类任务。在博客分析领域,此类信息可能存在于各种基于网络的社交和协作工具中,例如网络标签、大众分类法或网络目录。在分析博客时利用这种背景知识的替代来源为未来的工作提供了有趣的途径。

8.致谢

我们感谢 Yan Liu、Claudia Perlich、Vikas Sindhwani、Scott Spangler 和 Ying Chen 进行了深入的讨论。

9. 参考文献

[1] R. Agrawal、RJB Jr. 和 R. Srikant。 Athena:基于挖掘的文本数据库交互式管理。扩展数据库技术,2000 年。

[2] Blogpulse:尼尔森 buzzmetrics 的一项服务。http://www.blogpulse.com/。

[3] RT Clemen 和 RL Winkler。结合来自风险分析专家的概率分布。风险分析,19:187-203,1999。

[4] W. Dai、G.-R. Xue、Q. Yang 和 Y. Yu。迁移朴素贝叶斯分类器用于文本分类。AAAI,2007 年。

[5] S. Das 和 M. Chen。Yahoo! for Amazon:从股票留言板中提取市场情绪。亚太金融协会,2001 年。

[6] A. Dayanik、DD Lewis、D. Madigan、V. Menkov 和 A. Genkin。根据文本分类中的领域知识构建信息先验分布。SIGIR,2006 年。

[7] G. Druck、G. Mann 和 A. McCallum。使用广义期望标准从标记特征中学习。在 SIGIR,2008 年。

[8] KT 杜兰特和 MD 史密斯。网络的进步挖掘和网络使用分析,章节“使用监督机器学习技术结合特征选择预测网络日志帖子的政治情绪”。Springer,2007 年。

[9] 提取网页的主要内容。http://w-shadow.com/blog/2008/01/25/extracting-the-main-content-from-a-webpage/。

[10] S.弗伦奇。群体共识概率分布:一项批判性调查。贝叶斯统计 2,第 183-197 页。北荷兰省,1985 年。

[11] C. Genest 和 JV Zidek。结合概率分布:评论和带注释的参考书目。统计科学,1:114–135,1986。

[12] 胡明和刘斌。挖掘并总结客户评论。在 KDD,2004 年。

[13] S.-M. Kim 和 E. Hovy。确定意见。在科林,2004 年。

[14] B. Liu。Web 数据挖掘。Springer,2007 年。

[15] B. Liu, X. Li, WS Lee, 和 P. Yu. 通过标记词进行文本分类. In AAAI, 2004.

[16] A. McCallum 和 K. Nigam。朴素贝叶斯文本分类的事件模型比较。1998 年 AAAI 文本分类研讨会。

[17] V. Ng、S. Dasgupta 和 SMN Arifin. 研究语言知识源在评论自动识别和分类中的作用。在 ACL,2006 年。

[18] K.尼加姆。使用未标记的数据来改进文本分类。博士论文,卡内基梅隆大学,2001 年。

[19] B. Pang 和 L. Lee。情感教育:使用基于最小切割的主观性概括进行情感分析。在 ACL,2004 年。

[20] B. Pang、L. Lee 和 S. Vaithyanathan。竖起大拇指?使用机器学习技术进行情感分类。在 EMNLP,2002 年。

[21] MF 波特。后缀剥离算法,第 313-316 页。摩根考夫曼出版公司,1997。

[22] G. Ramakrishnan、A. Jadhav、A. Joshi, S. Chakrabarti 和 P. Bhattacharyya。通过词汇关系的贝叶斯推理来回答问题。在 ACL 多语言总结和问答研讨会上,2003 年。

[23] 雷·夏皮尔。学习性强弱弱。机器学习,5(2):197–227, 1990。

[24] RE Schapire、M. Rochery、MG Rahim 和 N. Gupta。将先验知识纳入提升中。在 ICML,2002 年。

[25] J.沙夫利克。结合符号学习和神经学习的框架。机器学习,1992 年。

[26] V. Sindhwani 和 P. Melville。用于半监督情感分析的文档-单词共同正则化。在 ICDM,2008 年。

[27] S. Spangler、Y. Chen、L. Proctor、A. Lelescu, A. Behal、B. He、T. Griffin、A. Liu、B. Wade 和 T. Davis。用于企业品牌和声誉分析的 COBRA-Mining Web。 IEEE 国际网络智能会议,2007 年。

[28] P. Turney。赞成还是反对?语义取向在评论无监督分类中的应用。ACL,2002 年。

[29] T. Wilson, J. Wiebe, 和 P. Hoffmann. 在短语级情绪分析中识别语境极性。在 EMNLP,2005 年。

[30] R L Winkler。主观概率分布的共识。管理科学,15:361–375,1968 年。

[31] X. Wu 和 R. Srihari。将先验知识与加权边缘支持向量机结合起来。KDD,2004 年。

[32] H. Yu 和 V. Hatzivassiloglou。走向回答观点问题:将事实与观点区分开来,并识别观点句的极性。在 EMNLP,2003 年。

[33] L. Zhuang, F. Jing, 和 X.-Y. Zhu. 电影评论挖掘与摘要. CIKM, 2006.