来自评级的上下文感知情感检测

听涛路瑞海东，巴里史密斯

摘要用户生成内容的爆炸式增长，特别是推文，客户评论，可以通过明确或隐含地利用内容及其伴随的情感信号之间的一致性来自动构建情感词典。在这项工作中，我们描述了自动生成特定于领域的情感词典的新技术，这些词典针对给定域的语言模式和习语进行了优化。我们描述了我们如何使用评论评级作为情绪信号。我们还描述了一种识别情绪中的情境变化的方法，并展示了如何在实践中利用这些变化。我们在许多不同的产品领域评估这些想法。

1 介绍

情感分析和意见挖掘技术旨在识别，分析和理解用户提供的文本材料中的主观信息。在利用用户生成的内容从短篇推文和状态更新到更详细的客户评论时，它们变得越来越受欢迎。例如，[7]从关于股票的新闻文章中挖掘正面和负面情绪，并用它来形象化用户的市场情绪和定价信息。同样Tumasjan等人。[19]证明了通过提取推文中包含的政治情绪来预测选举结果的潜力。别处
[15]描述了如何通过总结来自客户评论的不同产品特征的情绪信息来帮助用户比较不同的产品。和董等人。提出一种自以为是的产品推荐技术

路.
复旦大学电子邮箱：luyc13@fudan.edu.cn

瑞海东，巴里史密斯
都柏林大学数据分析洞察中心电子邮件：[Ruihai.Dong,
Barry.Smyth] @ insight-centre.org

挖掘相似性和情绪来估计评论评级并产生更有效的产品推荐[5,4]。

处理大量非正式的，固定的，多域的，用户生成的内容的要求已被证明是对传统情感分析技术的挑战，传统的情感分析技术通常基于静态词典[10]，固定规则[9]或标记数据[14,16]。特别是，这种静态方法很少涉及跨越不同领域出现的常见习语。即使在单个领域内，根据具体情况，不同的术语也可以表示非常不同的情绪极端情况。例如，在平板电脑评论中，术语“高”在修改外观质量时具有正极性，但在修改外观价格时具有负极性。

这些挑战表明需要一种更灵活，更动态的情感分析方法，也许可以基于与目标领域相关的共同语言模式。前进的方法是利用通常与用户生成的内容相关联的情绪信号（隐式或显式）。例如，用户倾向于使用表情符号，如:)，:(或标签（#cool，＃fail等）来表达情感意图。这些信号可用于增强情感学习算法;参见[2,16]例如，Davidov等人[2]提出了一种监督分类技术，通过使用50个Twitter标签和15个“表情符号”作为情感标签来避免强化手动注释.Liu等人[16]提出了一个模型（ESLAM）利用手动标记的数据和表情符号来训练和平滑Twitter情绪分析。

对于较短形式的用户生成内容，指导性假设是积极情绪词通常与积极情绪信号的关系比与负情绪信号关系更密切，反之亦然;见例如[11]。Kiritchenko等。介绍了一种在Twitter上评估短期非正式短信情绪的微弱策略方法[13]。Hashtags用作消息的情感极性的指示符，并且它们使用逐点互信息（PMI）技术来测量候选情感词与推文的极性之间的关系。

较长形式的用户生成内容（如博客或客户评论）不太适合使用本地情感信号：它们比其他情感更长，更多样化，并且通常包含多个主题并表达复杂多变的情绪。例如，“泰国红咖喱味道鲜美，.....但售价12美元，相当昂贵”表达了对泰国红咖喱的积极情绪，但对价格的负面情绪。然而，可能存在隐含的意图指示。例如，Bross和Ehrig [1]利用利弊摘要中包含的信息作为情绪信号;假设是作者在描述专业人士的产品方面时选择正面表达，而负面表达在缺点中是常见的。Wu和Wen [22]通过计算他们与正负搜索引擎命中的统计关联来预测名词的情绪，这是通过使用具有正面和负面查询模式的搜索引擎百度获得的。

明确的评级也是评论的重要信号。评论的评分反映了评论的整体情绪，但不一定反映每个评论主题的情绪。例如，Kiritchenko等[12]创建了一个域 -

通过利用一星，二星评级作为负信号和四星，五星评级作为正面信号，餐馆的特定情感词典。但更有趣的挑战是，如果评级可以成为在这些较长形式的用户生成内容中构建上下文感知情感词典的有用信号。

在本文中，我们描述了一组用于自动构建上下文敏感的特定于域的情感词典的新技术。我们还使用评论评级作为情感信号，但描述了如何利用上下文。我们在许多不同的产品领域评估这些想法。我们使用来自TripAdvisor和亚马逊的数据将我们的词典与传统的“一刀切”词典进行比较，我们通过将其性能与多个基线和最先进的情感极性进行比较来评估词汇的质量。预测任务。

2 创建特定于域的情感词典

我们描述了一种自动构建特定于域的情感词典的方法，给出了一组用户生成的评论，这些评论可以可靠地分为正分组和负分组。出于本文的目的，我们将重点关注来自TripAdvisor的酒店评论领域 - 从而创建与旅行相关的词典 - 但许多其他领域也适用于所描述的方法。为了将我们的评论分为正面和负面组，我们使用相应的评论评分，重点关注此例中的五星（正面）和一星（负面）评论。正如引言中所提到的，基本原则是在正面评论中经常出现但在负面评论中不常见的单词更可能是积极的，反之亦然;通过比较单词的正面和负面评论频率，我们可以计算出细粒度的情绪评分。

2.1 词级情绪评分

我们从n个正（5星）和负（1星）评论的集合开始，对于每个单词，我们计算它在正（pos（w））和负（neg（w））中出现的次数评论。从表面上看，pos（w）和neg（w）之间的相对差异表示w的极性。但是，我们观察到，在许多在线评论设置中，正面评价和负面评论的长度之间可能存在显着差异，从而导致需要某种规范化方法。

例如，TripAdvisor上的负面评论的平均长度是191个单词，显着高于平均正面评价的120个单词。鉴于我们在训练集中确保了相同数量的正面和负面评论，我们可以通过将字数除以总评论长度（分别为正面和负面评论的POS和NEG）来应用简单的标准化程序，如公式1和2所示。

$$要求^{+}F\ (W) = \frac{pos(w)}{POS机} \tag{1}$$

$$REQ的\ (双膦酸^{-})\ f\ (w) = \frac{(W)\ 。}{负} \tag{2}$$

接下来，我们计算一个单词的极性差异（或PD）作为其正负频率之间的相对差异;因此，在负面评论中更有可能出现在正面评论中的单词将具有正PD值，而更可能在负面评论中出现的单词将具有负PD值。接近0的PD值表示中性极性或情绪，而极值-1或+1的PD值分别表示强的负极性或正极性。

$$PD_{。}\ (\ ) = \frac{f\ req^{+}\ (w)\ -\ f\ req^{-}\ (w)}{f\ (w)\ 的要求tf41)\ +\ f\ (w)\ 的要求tf42)} \tag{3}$$

最后，我们计算一个单词（发送（w））的情绪分数作为其PD值的有符号平方;参见等式4，其中如果x <0则sign（x）返回-1，否则返回+1。这种转变的效果是相对放大非中性的PD值之间的差异。

$$发送\ (w) = PD\ (w)^{2} * 符号\ (PD\ (w)\ ) \tag{4}$$

2.2 处理否定

到目前为止讨论的方法在评估情绪方面没有考虑否定问题。例如，当在负面背景中使用时，单词情绪的极性可以完全改变;例如，当在负面背景中使用时，像“极好”这样的高度积极词语的情绪几乎完全逆转。例如，“签到经历远非太棒了”。然而，处理否定并非直截了当，并且对情感分析的这一方面给予了相当多的关注;见[20]。

在这项工作中，我们通过使用常见否定词（例如“not”）和短语（例如“远离”）的词典来处理否定，并在给定情感词w的k字距离内寻找这些词，在同一词内句子。当负面术语以这种方式与情感词w相关联时，我们假设w处于否定的上下文中。为了应对这种负面背景，我们维持两种形式的w。当在常规（非否定）上下文中找到w时，以如上所述的正常方式处理w。然而，当在否定的上下文中找到w时，我们为其否定形式保留单独的计数，我们将其标记并记录为不是w。例如，我们可能会发现许多正面评论中出现“完美”这个词，但负面评论较少，因此根据公式4，它会被赋予正发送（w）值。但是，我们也可能会发现提及特征的评论

那些“不完美”或“远非完美”。在这些评论中，情感词（“完美”）出现在否定的上下文中，但是这些出现的计数是参考不存在的情况存储的，允许对被发送的（不是每个）的否定形式进行独立的情绪计算。

为了滤除噪声，我们设置最小值的阈值t。在训练集中发现少于t次的单词将从词典中删除。阈值的设置保证了词典中包含的意见词确实是域中的高频词，并且还弥补了我们的情感得分公式在处理仅在训练集中出现几次的词时所具有的不足。

3 创建一个上下文敏感的情感词典

到目前为止，我们已经描述了构建特定于域的情感词典的方法;一个可以捕捉不同域之间的惯用差异，从而提供更准确的情绪记录。但是，这仍然不能解决在域内更精细的粒度级别可能发生的情绪转变。例如，将服务水平描述为“低”显然是消极的，但“低”价格通常是正的。为了解决这个问题，我们需要一个上下文敏感的词典来记录单词在不同的上下文中如何呈现不同的情感，例如，当应用于不同的功能时。

为了构建一个上下文敏感的情感词典，我们采用类似的整体方法来构建特定领域的情感词典。我们从两组正面和负面评论开始，计算每组中的单词出现次数，并处理否定效应。但是这次我们记录关于个别特征词的单词出现，并记录与正面和负面评论中的特定特征相关联的情感词。

假设特征词是以高于平均频率出现的名词;因此，它们是用户倾向于在给定评论域中讨论的共同方面（例如，酒店评论中的“餐馆”或膝上电脑评论中的“显示”）。为了识别特征修改意见词（情感词），我们首先识别句内词性模式;也就是说，POS标签序列的一端是特征词，另一端是情感词。当情绪词用于修改特征时，我们假设存在一定数量的固定搭配。这种搭配可以由相应的POS标签序列捕获，使得可以采用频繁的词性模式来识别特征和意见词之间的关系。在寻找模式时，我们关注那些表明情感 – 特征关系的模式。

但是在开始时，对于句子中的给定特征，我们既没有相应的情感词也没有POS模式。因此，域特定词典用于提供一组候选情感词，其极性差的绝对值不小于0.3。我们假设这些具有强极性差异的附近单词用于修改给定的特征。然后我们扫描并计算

训练集中的情感特征模式，并将具有高于平均值的出现次数的那些模式标记为有效模式。之后，我们可以通过匹配有效模式来搜索特征修改情感词。对于每个情感 – 特征词配对，我们将在正和负评论中共同出现的次数（w，f）分别计为pos（w，f）和neg（w，f）。然后我们计算每个（w，f）配对的相对频率，如等式5和6，其中pos（f）和neg（f）表示特征f在正和负评论中出现的次数。

$$要求^{+}f\ (w,\ f)\ =\ \frac{pos(w,\ f\ )}{pos(f)} \tag{5}$$

$$要求^{-}f\ (w,\ f)\ =\ \frac{(W,\ f)\ 。}{(f)\ 。} \tag{6}$$

接下来，我们可以分别如等式7和8中那样计算基于特征的极性差异和相应的情绪分数。

$$PD的WF\quad (,) = \frac{f\ req^{+}\ (w,\ f)\ -\ f\ req^{-}\ (w,\ f)}{的要求tf76\ (F)\ +\ F,\ W,\ f)\quad 的要求tf77)\ \ (W,\ f)} \tag{7}$$

$$发送\ (w,\ f)\ =\ PD\ (w,\ f)^{2}\ *符号\ (PD\ (w,\ f)\ ) \tag{8}$$

（8）通过这种方式，在给定的域内，我们创建一个情感词典，其中每个情感词w与共同出现的特征词f相关联，并与反映（w，f）的相对可能性的情感分数相关联。）在正面或负面评论中发生配对。再一次，（发送（w，f）的正值表明w在正面评价中更有可能与f共存，而负值则对于发送（w，f）意味着w倾向于与负面评论中的f共同出现。

我们可以以与生成特定于域的词典时采用的方法完全类似的方式处理特定于上下文的词典生成中的否定。简而言之，出现在否定上下文中的情感词用“非”前缀记录，并且它们（基于特征的）情绪分数独立于非否定版本计算。最小出现的阈值设置为与域独立词典的情况相同。

4 评估

在本文的开头，我们强调了传统情感词典的两个问题 – 领域问题和背景问题 – 它们反对一种通用的情绪标签方法。然后，我们通过自动生成来自用户生成的评论的特定于域和上下文的词典来描述解决这些问题的方法。在本节中，我们将描述最近一项研究的结果，该研究旨在评估这些方法对来自不同内容领域的实际数据集的有效性。

4.1 数据集

作为我们评估的基础，我们收集了来自TripAdvisor和亚马逊的大量评论。表1列出了这些数据的摘要描述。例如，TripAdvisor数据集包含的酒店总数为867,644条。亚马逊数据集包括6种类别（数码相机，GPS，笔记本电脑，手机，打印机和平板电脑）的电子产品的90,138条评论。两个数据集中的评论按1到5的等级评定。

表1 TripAdvisor和Amazon审核数据集的统计信息。

	到到网		亚马逊	
评分	#点评	平均。莱恩	#点评	平均。莱恩
1	33,646	191	14,708	143
2	43,372	186	6,703	190
3	112,235	161	18.962	204
4	272,077	133	18,962	208
5	406,314	120	42,100	163

我们需要用户生成的评论来源作为培训数据来生成我们的词典，并作为测试数据来评估结果。对于每个数据集，我们从评级-1（负面）和评级-5（正面）评论的随机抽样构建我们的训练集;在TripAdvisor的情况下，我们选择30,000个正面和负面评论，对于亚马逊，我们选择12,000个。剩余的评论（无论评级如何）被用作测试数据的来源，我们从测试集的五个评级级别中的每一个中随机选择2,000。

作为评估我们的算法生成词典的基线，我们选择流行的词典Bing Liu意见词典[10];我们将在下文中将其称为BL。这是情感分析文献中使用最广泛的一种，并且是一种通用情感词典，其中每个意见词被标记为否定的正面。

作为第二个基线，我们也使用SO-CAL词典[18]。虽然Bing Liu的意见词典只包含没有情感分数的正面和负面词汇列表，但SO-CAL词典在三个方面与它不同：（1）词汇被POS标签分开；（2）词典为每个意见词指定情感分数（而不是二元标签），以及（3）词典以与我们类似的方式处理否定;即使用“not”前缀。换句话说，SO-CAL词典在结构上更类似于我们的域特定词典。我们在下面将SO-CAL词典称为SOCAL。

4.2 特定领域的词典覆盖范围

首先，我们使用第2节中描述的技术为TripAdvisor和Amazon域生成特定于域的词典 - 我们将以这种方式生成的词典称为DS - 并将这些词典与BL进行比较。我们通过比较DS和BL词汇的覆盖范围（它们的术语重叠和差异）来做到这一点。请注意，出于评估的目的，我们仅考虑BL和DS中也存在于测试数据集中的术语。

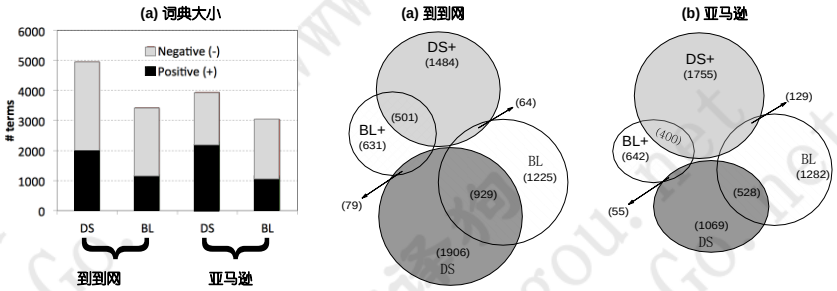

图1 TripAdvisor和Amazon数据集的DS和BL词典的覆盖率分析。

结果在图1中显示为TripAdvisor和亚马逊数据的词典大小和维恩图的条形图。在每种情况下，我们提出如下4个区域：

1. *DS* +：特定于域的词典中的正项数；
2. *DS*-：特定于域的词典中的否定词的数量；
3. *BL* +：Bing Liu词汇中的正词数；
4. *BL*-：Bing Liu词汇中的否定词数。

在这个阶段可以做出许多观察。首先，很清楚DS词典往往比BL词典更大（总体而言，相对于正面和负面术语）。图1显示了TripAdvisor和Amazon数据集的DS和BL词典中的正项和负项的数量。例如，与DS词典中的3,429个术语（1,211个正面和2,218个底片）相比，TripAdvisor DS词典总共包括4,963个术语（2,049个正词和2,914个底片）；亚马逊数据的比例相似。显然，在TripAdvisor和亚马逊的评论中有很多单词，可以与正面或负面的观点相关联，但更常规的BL词典中没有这些观点。

第二个观察结果是DS和BL词典之间存在相对较小程度的重叠。对于TripAdvisor，总重叠仅为1,567个术语，而亚马逊仅为1,112个术语。也许更重要的是，每个词典中也有大量的术语不存在于另一个词典中。例如，TripAdvisor DS词典包括1906个正面和1484个负面术语

它是独一无二的，BL词典中没有。这些术语通常与正面或负面的酒店评论相关，但通常不包括在传统词典中。这些术语的一个很好的例子包括酒店领域中的经理或地毯等字样。在每种情况下，该术语通常与否定评论相关联;旅行者很少在评论中谈论酒店经理，除非它是负面的，当他们提到酒店的地毯时，几乎总是抱怨他们缺乏清洁。另一方面，BL中唯一的631个正面和1,225个负面术语（并且从DS中缺失）根本不会作为酒店评论中的常见情感词出现。我们注意到亚马逊数据的模式非常相似，尽管很明显很多条款都不同。

第三，虽然两个词汇都同意大多数重叠术语的极性，但也有明显不同意见的重叠术语。例如，在TripAdvisor中，我们看到79个被DS分类为负但在BL词典中被列为正数的术语;像“便携式”这样的词通常在酒店评论中具有负面含义（例如，客人经常抱怨便携式电视）但BL将其归类为正面词。相反，有64个术语被DS归类为阳性，但被BL认为是负面的。例如，诸如“奢侈”或“沉没”之类的词被BL认为是否定的，但在酒店评论中总是积极的。我们再一次在Amazon结果中看到了类似的模式。

到目前为止，我们已经证明，我们生成特定领域词典的方法有可能在正面和负面的用户生成的评论中识别出更多种类的情感词;值得记住的是，这些词汇包括与我们的上下文敏感词典相同的术语。虽然这些词中的一些与常规情感词典中的词匹配，但大多数词却没有。对于那些在传统情感词典中发现的词，根据我们的特定领域方法，一些被发现具有相互矛盾的情感极性。当然，这些都没有说明DS词典的质量，也没有考虑过添加上下文敏感性的影响。因此，下一节我们将考虑对词汇质量进行更客观的测试，并更充分地考虑这些问题。

4.3 情绪极性预测

作为对词典质量的独立测试，我们建议使用评论内容和情绪数据来自动预测评论的整体极性。为此，我们将每个评论表示为情绪向量并训练分类器以预测整体评论极性。我们将通过改变词典和分类器来评估许多不同的条件。我们还将分别评估处理否定的影响，然后评估我们的上下文敏感方法的好处。

4.3.1 建立

首先，来自先前分析的测试数据将用作此评估的总体数据集。这是必要的，因为我们以前的训练数据用于构建词典，我们需要确保它不参与本实验的分类器训练或测试。出于此分类任务的目的，我们将1星级和2星级评价（每个领域中的4,000个评论）标记为阴性，将4星级和5星级评价标记为阳性（再次评估4,000个）。

在词典方面，我们将把我们的DS方法与BL和SOCAL进行比较，作为传统的“一刀切”词典的例子。此外，我们还将包括另一个我们将根据[13]的工作指定为PMI的替代方案。这个PMI词典的重要性在于它以与我们的DS方法类似的方式生成，但使用互信息分数来估计单词的情感极性。因此，它提供基线域特定词典。

在分类方法方面，我们将考虑贝叶斯网（BN）[21]，基于规则的方法，以JRip [21]和随机森林（RF）[21]的形式。我们还将根据第2节比较分类器性能是否进行任何否定测试。对BL的否定的实现是通过简单地根据否定来切换情绪的原始极性来实现的。最后，我们将通过将其应用于DS和PMI词典来考虑我们的上下文敏感方法（第3节）的影响。将我们的方法应用于PMI类似于我们在第3节中描述上下文敏感性的方式，因为可以针对单个范围内的特征词单独计算互信息分数。然而，传统的二元情感词典BL不适用于这种情境敏感性方法。

4.3.2 方法

我们的基本分类任务首先将每个评论表示为一个简单的特征向量。为此，我们通过词性标签（形容词，副词，名词，动词）对评论中的每个单词进行分类，并对情感分数求和（BL为-1或+1与其他词典的实值分数）。因此，每个评论由4元素情感向量表示，并与二元类标签（正或负）相关联。我们执行10次交叉验证，将数据拆分为训练和测试集，使用训练数据构建适当的分类器，并根据测试数据对其进行评估。在本实验中，我们使用WEKA [1]来运行分类。接下来我们将重点介绍每种情况的f-measure分数。

[1] http://www.cs.waikato.ac.nz/ml/weka/

表2 TripAdvisor和亚马逊数据集的精确度，召回率，F度量;域特异性，上下文敏感结果之间的差异在0.05水平上是显着的。

		没有否定			否定			否定+背景		
	方法	预。	召回	F-措施	预。	召回	F-措施	预。	召回	F-措施
	的DS - BN	0.935	0.935	0.934	0.946	0.946	0.946	**0.954**	**0.954**	**0.954**
	PMI的BN	0.933	0.933	0.933	0.941	0.941	0.941	0.949	0.949	0.949
	B1 - BN	0.857	0.856	0.856	0.882	0.882	0.881	-	-	-
	南加州-BN	0.837	0.837	0.837	0.858	0.858	0.858	-	-	-
	jrip DS	0.928	0.928	0.928	0.942	0.942	0.942	0.948	0.948	0.948
	jrip PMI	0.93	0.93	0.93	0.938	0.938	0.938	0.945	0.945	0.945
	jrip bl	0.853	0.853	0.853	0.882	0.882	0.881	-	-	-
	社会- jrip	0.842	0.842	0.842	0.862	0.862	0.862	-	-	-
	的DS -射频	0.932	0.932	0.932	0.943	0.943	0.943	0.95	0.95	0.95
	MIP -射频	0.927	0.927	0.927	0.941	0.941	0.941	0.949	0.949	0.949
	的BL -射频	0.839	0.839	0.838	0.868	0.867	0.867	-	-	-
	南加州-RF	0.814	0.814	0.814	0.845	0.845	0.845	-	-	-
	的DS - BN	0.82	0.82	0.82	0.853	0.853	0.853	**0.861**	**0.861**	**0.861**
	PMI的BN	0.819	0.818	0.817	0.846	0.846	0.845	0.857	0.857	0.856
	B1 - BN	0.726	0.726	0.726	0.762	0.762	0.762	-	-	-
	南加州-BN	0.703	0.703	0.703	0.73	0.73	0.73	-	-	-
	jrip DS	0.801	0.801	0.801	0.847	0.847	0.847	0.859	0.859	0.859
	jrip PMI	0.802	0.802	0.801	0.839	0.839	0.839	0.854	0.854	0.854
	jrip bl	0.738	0.738	0.738	0.76	0.76	0.76	-	-	-
	社会- jrip	0.715	0.715	0.715	0.731	0.731	0.731	-	-	-
	的DS -射频	0.802	0.799	0.799	0.848	0.848	0.848	0.859	0.859	0.859
	MIP -射频	0.797	0.793	0.793	0.84	0.84	0.84	0.851	0.85	0.85
	的BL -射频	0.709	0.709	0.708	0.743	0.743	0.743	-	-	-
	南加州-RF	0.678	0.678	0.678	0.707	0.707	0.707	-	-	-

4.3.3 结果

对于TripAdvisor数据和亚马逊数据，结果显示在表2中。在第一列中，我们列出了基于潜在情感词典和分类算法使用的基本分类技术;例如，DS BN是指具有贝叶斯网络分类器的特定于域的词典。接下来的三列显示了使用无否定，否定和否定+上下文构建的精度（召回，f-度量）分数分类器。

4.3.4 讨论

我们可以看到特定领域的词典（DS和PMI）的表现非常相似，f-measure得分大约为0.93，明显高于一刀切的BL和SOCAL词典的平均f-测量范围为0.81至0.85。我们还可以看到，通过在词典生成过程中明确处理否定，我们可以提高整体的分类性能。此外，通过将背景纳入DS和PMI词典 - 我们不能轻易 -

我将这种方法应用于BL和SOCAL词典 – 我们可以进一步提高分类性能，尽管在TripAdvisor数据的情况下更为轻微。

亚马逊域的结果大致相似，但这次我们发现DS在各种条件下的表现略好于PMI。否定和背景的引入也推动了更大的f-measure改进（与TripAdvisor相比），这表明否定和背景在亚马逊的评论中扮演着更重要的角色。话虽如此，否则在应用于BL词典时几乎没有效果。

在这个分析中，我们专注于f测量结果，但也注意到精确度和召回率的类似效果。总的来说，当我们将特定领域，上下文敏感词典与更传统的“一刀切”词典进行比较时，显然存在显着的分类性能优势。例如，根据表2，表现最佳的TripAdvisor条件是DS使用具有否定和上下文的贝叶斯网络分类器来产生大约0.954的f度量分数。相比之下，BL的相应得分仅为0.881。同样，在亚马逊域中，我们看到DS使用贝叶斯网络分类器的最佳性能，其f-measure得分约为0.861，而BL只有0.762。这些差异 – 在特定领域，上下文敏感结果和BL和SOCAL结果之间 – 使用配对t检验发现在0.05水平显着[3]。

(a) 正。 (b) 否定。

图2来自TripAdvisor数据集的特定领域形容词和副词。

为了更具体，我们还为TripAdvisor和亚马逊数据集的DS词典生成单词云，以便直观地评估词典的质量。出于空间原因，我们仅显示来自图2中的TripAdvisor数据集的特定领域的形容词和副词，以及来自图3中的亚马逊数据集的屏幕上下文中的上下文感知情绪词。在这些图中，大小表示情绪强度和亮度代表频率。从图2中我们可以看出，自制，铺砌，梦幻，光线充足，尺寸合适等具有强烈的正极性，但最差，劣质，无应答等具有强烈的负极性。从图3中，我们可以看到锐利通常用于描述屏幕，并且在屏幕环境下具有最强的正极性，但是，它具有强烈的负面

(a) 积极的。 (b) 否定。

图3来自亚马逊数据集的屏幕上下文中的上下文感知情绪词。

在酒店评论中描述桌子角落时的极性。我们也看到了*没*有反应，空白，坏等对屏幕有强烈的负极性。

5 结论

在处理用户生成内容中常见的惯用语和上下文词使用差异时，一刀切的情绪词典已被证明是有问题的。在一种情况下可靠积极的词语可能在其他地方产生负面情绪，反之亦然。在本文中，我们描述了一种新技术，通过将评论评级作为信号，自动构建上下文敏感的，特定于领域的情感词典。评估结果基于Don和亚马逊的评论数据证明了这些词典与更传统的基线相比的有效性。与传统词典相比，我们的特定领域词典捕获了许多新的情感术语，并且通过注意上下文，他们能够更好地处理域内可能发生的偶然情绪变化。

致谢这项工作得到爱尔兰科学基金会的资助，编号为SFI / 12 / RC / 2289。

参考

1. Bross，J.，Ehrig，H。：为客户审查挖掘自动构建领域和特定方面的情感词典。见：第22届ACM国际信息与知识管理会议论文集，第1077-1086页。ACM（2013年）
2. Davidov，D.，Tsur，O.，Rappoport，A。：使用twitter hashtags和smileys增强情感学习。参见：第23届国际计算语言学会议论文集：海报，COLING '10，第241-249页。计算语言学协会，美国宾夕法尼亚州斯特劳兹堡（2010年）。网址http://dl.acm.org/citation.cfm?id=1944566.1944594

3. Dietterich, TG: 用于比较监督分类学习算法的近似统计检验。神经计算10（7），1895-1923（1998）
4. Dong, R., OMahony, MP, Smyth, B。：在固定产品推荐中的进一步实验。见：第22届国际案例推理会议论文集，ICCBR '14，第110-124页。施普林格（2014年）
5. Dong, R., Schaal, M., O'Mahony, MP, Smyth, B。：从在线评论中分类和推荐的主题提取。在：第23届国际人工智能联合会议论文集，IJCAI '13。加利福尼亚州门洛帕克AAAI出版社（2013年）
6. Esuli, A., Sebastiani, F。：Sentiwordnet：一个公开可用的意见挖掘词汇资源。在：LREC会议记录，第一卷。6，pp.417-422。CiteSeer（2006年）
7. Feldman, R., Rosenfeld, B., Bar-Haim, R., Fresko, M。：基于混合方法的股票的股票声纳分析。在：第二十三届IAAI会议（2011年）
8. Go, A., Bhayani, R., Huang, L。：使用远程监督的Twitter情绪分类。CS224N项目报告，斯坦福1,12（2009）
9. Hatzivassiloglou, V., McKeown, KR：预测形容词的语义方向。参见：计算语言学协会第35届年会论文集和计算语言学协会欧洲分会第8次会议，第174-181页。计算语言学协会（1997年）
10. Hu, M., Liu, B。：在客户评论中挖掘意见特征。参见：第19届全国人工智能会议论文集，AAAI'04，第755-760页。AAAI出版社（2004年）。网址http://dl.acm.org/citation.cfm?id=1597148.1597269
11. Hu, X., Tang, J., Gao, H., Liu, H。：带有情绪信号的无监督情绪分析。见：第22届万维网国际会议论文集，第607-618页。国际万维网会议指导委员会（2013年）
12. Kiritchenko, S., Zhu, X., Cherry, C., Mohammad, SM：Nrc-canada-2014：检测客户评论中的方面和情绪。参见：第八届国际语义评估研讨会论文集（SemEval 2014），第437-442页（2014）
13. Kiritchenko, S., Zhu, X., Mohammad, SM：短篇非正式文本的情感分析。人工智能研究杂志，第723-762页（2014）
14. 刘，B：情绪分析和意见挖掘。人类语言技术综合讲座5（1），1-167（2012）
15. Liu, B., Hu, M., Cheng, J。：意见观察员：分析和比较网络上的意见。参见：第14届万维网国际会议论文集，WWW '05，第342-351页。ACM，纽约，纽约，美国（2005年）。DOI 10.1145 / 1060745.1060797。网址http://doi.acm.org/10.1145/1060745.1060797
16. Liu, KL, Li, WJ, Guo, M。：表情符号平滑了推特情绪分析的语言模型。在：AAAI（2012）
17. Lu, Y., Castellanos, M., Dayal, U., Zhai, C。：自动构建情境感知情感词典：优化方法。参见：第20届万维网国际会议论文集，WWW '11，第347-356页。ACM，纽约，纽约，美国（2011年）。DOI 10.1145 / 1963405.1963456。网址http://doi.acm.org/10.1145/1963405.1963456
18. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M。：基于Lexicon的情感分析方法。计算语言学37（2），267-307（2011）
19. Tumasjan, A., Sprenger, TO, Sandner, PG, Welpe, IM：用推特预测选举：140个字符揭示了政治情绪。ICWSM 10,178-185（2010）
20. Wiegand, M., Balahur, A., Roth, B., Klakow, D., Montoyo, A。：关于否定在情绪分析中的作用的调查。见：关于自然语言处理中否定和推测的研讨会论文集，第60-68页。计算语言学协会（2010年）
21. Witten, I., Frank, E。：数据挖掘：实用的机器学习工具和技术。摩根考夫曼（2005）
22. Wu, Y., Wen, M。：消除模糊形容词的动态情绪。在：第23届国际计算语言学会议论文集，COLING '10，第1191-1199页。计算语言学协会，美国宾夕法尼亚州斯特劳兹堡（2010年）。网址http://dl.acm.org/citation.cfm?id=1873781.1873915