**自动构建消除语义歧义的领域情绪词典**

**摘要**

现有的情绪词典忽视单词在不同语境下的情感歧义问题，并且仅仅给单词赋予一个积极或者消极的极性。在本篇论文中，我们提出了一个自动化的方法来构建特定领域的情绪词典（SDS-lex）来避免情绪歧义，该方法不仅结合了来自现有词典的情绪信息，还通过使用我们改进的TF-IDF算法结合了来自语料库的情绪信息。ITF-IDF 算法通过考虑单词的重要性和不同词性 （POS） 在标有不同情感倾向的语料库中的分布来计算单词的情感。

**介绍**

现有的许多词典如MPQA Subjectivity Lexicon 给出了单词对应的情绪极性或者值和词形，可以在本篇论文中作为先验情绪知识使用。然而，它们中的大部分都是通用的，灵活性有限，不能直接用于某些特定领域。因此，仍存在需要值得考虑的因素。

构建一个情绪词典需要特定领域的文集和现存的情绪词典来形成基础。现存的工作都只根据不同的文集来区分单词的情绪。

然而，一个单词即使是在同一个文集中也可能有多种含义，例如patient。因此，如何区分单词的情绪表达是一个值得考虑问题。单词的词性可以作为一个指示器来区分不同的情绪表达，例如patient作为名词和作为形容词。受到这个启发，我们使用ITF-IDF算法，同时考虑在不同极性的文集中的情绪单词的差异。

再次重申，在本篇论文中，我们提出了一个自动构建特定领域情绪词典的方法，它集合了两部分的情绪信息，一是通过基于文集的ITF-IDF算法计算得到的情绪值，二是从现有情绪词典中获得的先验情绪知识。我们的ITF-IDF算法使用不同的极性标签和POS因子来获得特定领域内单词的更精细的情感值，而传统的TF-IDF只关注单词在不同文档中的重要性。然后，我们会结合先前的情感知识，克服单词情感对语料库的严重依赖性，并补充训练文集中未包含的单词情绪。

**相关工作**

**现有情绪词典资源**

列举了很多词典。现有的情绪词典主要包括通用词典和特定领域词典。大部分通用词典都是人工标注的，分时费力。大部分特定领域的词典是使用特定领域文集中的单词共现构建的，例如产品评审。尽管这种词典考虑了领域和话题，但是它们忽略了单词在不同情感极性的语料库中的分布，以及同一主题上的某些单词可能具有不同情感极性的事实。

**情绪词典的构建方法**

主要有三种构建情绪词典的方法。第一种基于现有的情绪词典或知识基础。第二种通过分析特定领域的文集中单词的关系来自动构建情绪词典。第三种结合文集和现有的情绪词典来共同学习情绪词典。本篇论文中的方法使用了第三种。并且，我们的模型中采用的改进的TF-IDF方法考虑了同一情感类别中每个单词的POS的不同分布，提供了更详细和稳定的单词情绪分析。

另外，由于一个词在不同的语境和主题下可能具有不同的极性和强度表达，因此仅根据上下文或主题进行情感分析是不够的。

**基于改进的TF-IDF算法构建一个特定领域的情绪词典**

对于每个单词，我们使用POS方法标记细粒度的单词含义，此步骤可以减少语义消歧。然后，通过ITF-IDF算法计算具有不同POS的单词的不同情绪值。由于特定领域语料库的词汇限制，我们还纳入了现有的通用情感词典，旨在增加单词的覆盖率并微调单词的情感值。

**基于文集的情绪计算**

本文提出改进的TF-IDF（ITF-IDF）算法，将POS信息与不同情感语料库中的词分布相结合，计算特征词的重要性，降低训练过程中的不确定性。

计算每个特定词性的单词的情绪，设置了四条前提。

H1: 与在某种情绪倾向的文集中，单词该种词性的频率与单词的频率的比值正相关，对应一个值。

H2: 与在某种情绪倾向的文集中，单词该种词性的频率与该种情绪文集的个数的比值正相关，对应一个值。

H3: 与单词该种词性在积极文集和消极文集中频率的差值的绝对值正相关，对应一个值。

H4: 与单词该种词性在积极文集和消极文集中的总频率负相关，对应一个值。

特定词性的某个单词在在某种情绪倾向的文集中的情绪值就是上面四个值的和。特定词性的某个单词的情绪值就是积极文集中的值减去消极情绪文集中的值的差值，记为。

大家看原文3.1 Corpus-based sentiment calculations吧，有公式，更加明了详细，我不好用文字表述。

**基于多种词性构建特定领域情绪词典**

我们引入了ITF-IDF算法来解决词义歧义并发现新词，但是当使用小型训练语料库时，它对一般情感词的覆盖率很差。为解决这个问题，我们结合了先验情绪知识。

使用了MPQA和SentiWordNet3.0这两个现有词典。对于MPQA，由于只有极性和强度标签，没有值，每个特定词性的单词的先验值定义规则为：如果是强度标签为strong，则绝对值为1；若为weak，则为0.5；如果为积极则是整数；消极则为负数。对于SentiWordNet3.0，包含情绪值和一个相似单词集，但是忽略了某些但可能有多个情绪值，尽管有相同的词性。为解决这个问题，先验值定义为多个情绪值的平均值，即包含在类似概念的词中的特定词性的某个单词的情绪值的和，除以单词该词性在整个词典中出现的次数。

那么，对于每个特定词性的单词，先验情绪值为以上两个先验值的平均值，记为。

最终，一个特定词性的单词的情绪值为(1-) , β是融合贡献率，在0到1之间变化，用于调整先验知识和基于语料库的情感知识之间的比例。