**基于\*\*词向量的图书评论中英情感词典构建研究**

**2基于词向量的情感词典特征表示方法**

**2.1基于多语料+TF-IDF的情感词提取**

领域情感词典中如何抽取出领域相关性强的情感词是十分重要的。对于对单领域语料来说TF-IDF值计算并不能很好的体现领域性，因为实质上TF-IDF是表示单词对于语料库中的某个文件的重要程度的指标，如果只有单一领域语料的话我们必须将句子作为文件，整体语料作为文件集。以这样的方法得到的TF-IDF权值受两个因素影响，一个是常用性，另一个则是领域性。很显然领域性低的常用单词并不是我们期望的排名靠前的结果，例如在餐厅评论中，“美味可口”的优先级是应当高于“好”的。所以此处需要引入外部的非领域语料来提高领域独有性对权重的影响，降低常用性对权重的影响，这里只是一个相对的增强与削减，对于领域性强且常用的词仍然会有较高的权重，并不会产生矫枉过正的问题。

利用非领域语料的计算方法是：计算所有在领域语料中出现的单词在非领域语料中词频，以一定的值为基准（例如平均词频），词频和权重呈反比关系，也就是一个反函数，对在非领域语料中未出现过的词，即反函数分母可能为0的情况下，给予该单词一个足够小的词频或者直接给定一个足够大且合理的权值。公式表达为：

在选择种子词，以及筛选情感词时，常用的做法时利用TF-IDF算法计算单词权重，而TF-IDF值有可能会出现数值非常大的情况，过大的数值不利于进行计算，所以本研究对每个句子序列的TF-IDF值进行一次softmax处理以降低数值同时增强TF-IDF值的影响。

在TF-IDF的权重模式有多种计算方式，相较于常规的TF-IDF计算方式来说，根据文档情况选择其它方法能够提升权值的合理性。本研究中TF权重使用了log标准化（log normalization）的方法而IDF权重则采用了逆向文档频率平滑（inverse document frequency smooth）的方法，如下两图所示：

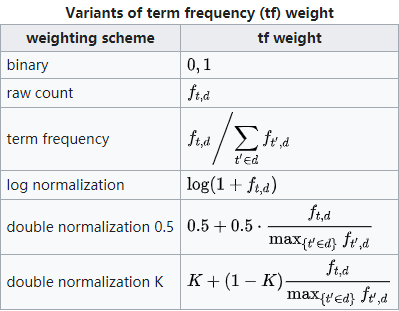


图1词语频率权重的变体

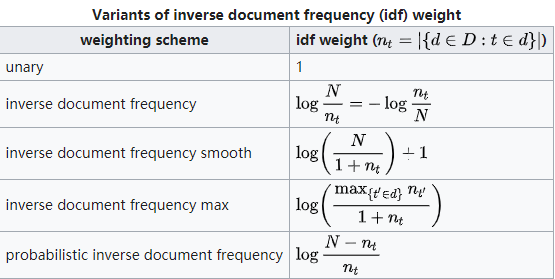


图2逆向文档频率权重的变体

**2.2基于fasttext的数据表示**

在利用传统的情感词典在处理分类任务时，利用向量来表示文本输入模型的通常做是对文本采用one-hot编码进行embedding操作，这样操作往往会造成向量过于稀疏不方便计算，同时也无法体现出词语之间的关联性。而使用词向量进行word embedding（词嵌入）操作，能够将词语向量控制在一个合理的大小之内，并且可以很好的反应词语之间的关联。但是词向量表现的是词语之间的语意关联，并不能很好的解释词语之间的情感关联，例如“不错”与“不差”在词向量中经常会有较大的关联性，这两个词的情感极性完全相反，在我们进行分类任务时会期望这两个词有较大的差异而非强关联性。当我们既需要强化词语之间的情感关系而使用情感词典时，同时又希望能够使其合理地进行文本表示，给分类模型提供一个合理的输入。

**2.3基于词向量的情感词典构建方法**

基于词向量的领域情感词典构建主要分为两个大部分，一部分根据相应算法生成低维度、可读的领域情感词典，这与常见的情感词典类似。另一部分则是通过融入预训练词向量，将情感词典的维度提高，将高维度的词向量作为词嵌入的基础，为机器学习模型提供合理的输入数据。

**算法：**

**输入：**领域语料re\_c、非领域语料irre\_c、词向量辅助训练语料ft\_c、融合情感词典sl。

**输出：**高维表示情感词典MCBSL、低维情感词典表示MCBSL\_Low

**步骤1：**将re\_c与ft\_c合并，进行分词以及去停用词处理，使用fasttext训练得到词向量。

**步骤2：**对re\_c进行分词以及去除停用词以及词性筛选处理，并进行序列化处理，生成格式化文本形如：，其中i为re\_c的句子数，n为语料中的最长句子长度。

**步骤3：**计算各词的TF-IDF值，TF-IDF的计算公式为，其中t为词符(token)，d为句子序号，N为re\_c的句子数量，nt为包含t的句子数量。并对每个句子序列进行softmax处理。得到TFIDF矩阵： 。然后根据TFIDF\_Seq求出每个词的TF-IDF权重，得到TF-IDF值词表：，其中m为语料中的词形(type)数量。

**步骤4：**对非领域语料irre\_c进行分词以及去停用词操作，利用公式

计算得到每个词的权重MCW，其中代表的词频，参数k为公式的权重值可进行调整，默认情况下取所有情感词在非领域语料中的均值，为当不存在于irre\_c中时词频的缺省值，默认情况取值为0.5。最终生成多语料权重词表：

**步骤5：**将TFIDF\_L与MCW\_L中相同词的权重相乘，构成情感词的融合权重词表，这其中权重weight的计算公式为。

**步骤6：**将sl中的情感极性与Weight\_L相结合生成维度为d的种子词表Seeds。首先求得各情感词的情感权重值，其中p代表了sl中的情感词极值，计算方法为各情感词典的极值加权之后求和。然后将词表按照倒序排序，选取权重值为正的前d/2个词，权重值为负的前d/2个词，最终得到种子词表。

**步骤7：**通过设置阈值α进行筛选的去除权值过低的情感词，生成低维表示的情感词典，其中wp值为sl中情感词的极值与权重值wight的乘积，即。

**步骤8：**生成高维情感词典表示MCBSL，其中每个情感词的维度为d，表达式为，参数e的计算方法为，SIM函数的具体表示为，利用训练好的fasttext词向量分别得到情感词与种子词的向量表示与，然后利用词向量计算情感词与种子词之间的相似度，即。

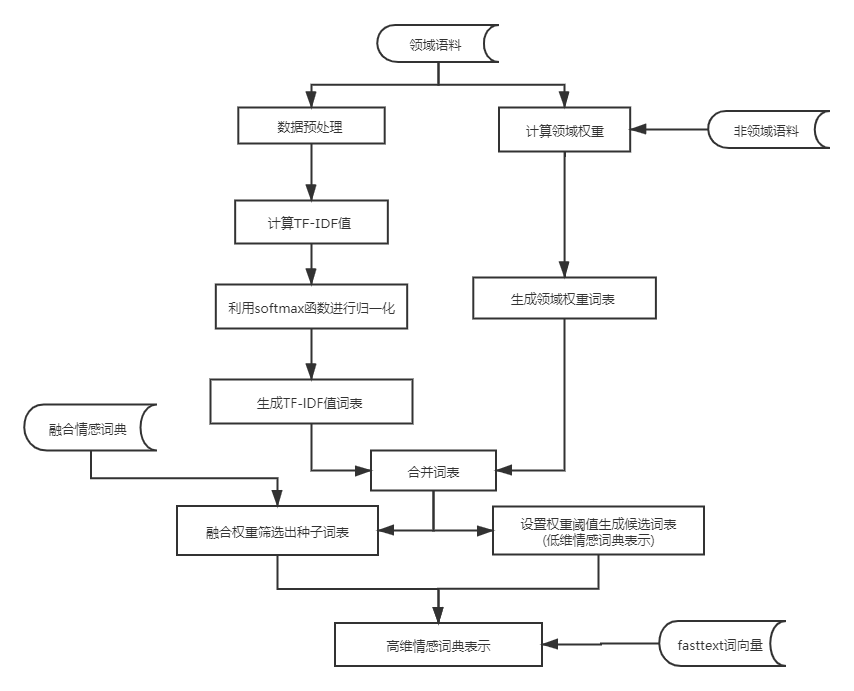


图3 算法流程图

**3实验及结果分析**

**3.1 实验设计**

为了检验本研究得到的情感词典在情感分析领域的效果，本实验采用对照实验的方法，利用情感本体、词向量以及情感词典进行文本情感分类来验证情感词典的效果。同时将情感词典中使用的领域强度的判断方法进行对照实验，以检验本研究中所使用的方法的效果。分类方法统一采用SVM。将正负向语料的精度、召回率、f1值，以及总体的准确率作为评价指标，来检验情感词典对文本情感分类任务的效果。

**3.2 实验数据**

中文语料的实验数据来源于谭松波老师的酒店评论公开数据集共4000条，其中正向与负向(积极与消极)领域语料各2000条。实验选取正负向各1000条数据作为训练数据，剩余的正负向各1000条作为测试数据。英文语料采用有Blitzer收集整理的Amazon评论，共有图书、dvd、电子产品、厨房用品、影像5个领域 。每个领域标注语料6000条，以及若干无标注数据。实验数据选取各领域正负向各1000条数据，选取其中10%即200条数据作为训练集。

中文多语料计算领域强度使用的非本领域语料，来源于SMP-EWCT2020的微博评论共46421条。英文的外部语料来自于纽约时报新闻评论的公开数据，共49868条。这部分语料通用性较强，没有特别的领域性所以选择其作为非领域语料使用，进行分词处理后作为外部语料进行计算。用于进行分类实验的语料来源及结构如下表：

表1 分类语料统计表

|  |  |  |  |
| --- | --- | --- | --- |
| 数据名称 | 正向语料数量 | 负向语料数量 | 无标注数据 |
| ChnSentiCorp-Htl-ba-4000 | 2000 | 2000 | 0 |
| ChnSentiCorp-Htl-uba-10000 | 7000 | 3000 | 0 |
| Amazon reviews（books） | 3000 | 3000 | 9750 |
| Amazon reviews（dvd） | 3000 | 3000 | 11843 |
| Amazon reviews（electronics） | 3000 | 3000 | 17009 |
| Amazon reviews（kitchen） | 3000 | 3000 | 13856 |
| Amazon reviews（video） | 3000 | 3000 | 28363 |

**3.3 实验参数**

中文语料的预处理工作采用jieba进行分词处理，并且使用paddle模式，去除助词、标点符号、非语素字、介词、量词、数词、叹词。中文停用词表采用哈工大的中文停用词表。英文语料通过spacy筛选出与中文一样的词性，停用词采用spacy的英文停用词表。

实验选取情感本体作为基础情感词典的对照组，采用One-hot编码的形式，句子向量的维度为情感本体中所有词语的个数，编码值为情感强度。使用word2vec以及fasttext方法对训练数据进行词向量的训练，设定维度为100，迭代次数为30，词的最小出现次数为2，其余为默认设置，句子向量的计算采用词向量求和取平均的方式生成，句子向量的维度等同于词向量的维度。情感词典中领域强度的计算采用tf-idf的多语料融合改进方法，维度同样设定为100，句子向量的生成方式与词向量方式相同。

关于本文提出的情感词典构建过程中各项参数的选取，通过外部计算情感词权重的公式中，参数k取所有情感词在非领域语料中的均值，fm值取0.5。选取情感词的阈值α取值为正负向各自排名在70%的情感词的wp值。融合情感词典采用大连理工情感本体（DUTIR）与SenticNet6.0两个情感词典进行融合。

情感分类采用SVM（支持向量机）方法构建分类器，利用sklearn进行SVM的具体实现，其中核函数采用liner，并且采用概率估计，其余参数为默认。

**3.4 实验结果与分析**

（1）情感词对照实验

引入了非领域语料之后，会抽取出更多的非常用但领域性强的情感词。我们利用算法中生成低维情感词典，对不同方法所选择的情感词进行一个直观地对比，来观察多预料计算权重的方法是否更为合理。

表2引入非领域语料之后的中文词表对照

|  |  |  |  |
| --- | --- | --- | --- |
|  | 极性 | 数量 | 示例 |
| 共有词条 | 正向 | 479 | …整洁，豪华，宽敞，价廉物美，诚挚… |
| 负向 | 427 | …最差，不足之处，简陋，大失所望，美中不足… |
| 引入非领域语料后追加的情感词 | 正向 | 62 | …宾至如归，方便，便利，便宜，没得说… |
| 负向 | 69 | …轰鸣声，坑坑洼洼，偏僻，形同虚设，置若罔闻… |

表3引入非领域语料之后的英文图书评论词表对照

|  |  |  |  |
| --- | --- | --- | --- |
|  | 极性 | 数量 | 示例 |
| 共有词条 | 正向 | 1804 | …great，excellent，wonderful，new，easy… |
| 负向 | 1842 | …bad，boring，worst，disappointing，confusing… |
| 引入非领域语料后追加的情感词 | 正向 | 75 | …gastronomic，decorative，suggestive，impressively，unforgettably… |
| 负向 | 80 | …machiavellian，sissy，unitarian，regretful，musty… |

从上表可以看出，引入非领域语料能够抽取出一些非常用的领域情感词汇，这些情感词的领域性较强，但是并不常用所以在单一的领域语料中TF-IDF值并不会很高，需要通过引入非领域语料才能够提高这些词的权重，所以为了使领域情感词的权重更为合理，引入非领域语料是有效的操作。

（2）TF-IDF算法改进对照实验

为了验证本研究采取的改进TF-IDF方法相较于常规TF-IDF方法有更好的效果。对于传统TF-IDF方法，以及采用上文提到的改进后的TF-IDF方法（进行了对照实验，利用ChnSentiCorp-Htl-ba-4000的4000条酒店进行分类计算，进行10次简单交叉验证，通过精度、召回率、f1值以及总体准确率对两种生成情感词典进行对照。

表4情感词抽取方法对照实验

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 情感词抽取方法 | macro precision | macro recall | macro f1 | accuracy |
| 常规TF-IDF | 82.83% | 82.76% | 82.76% | 82.78% |
| 改进TF-IDF | 83.78% | 83.73% | 83.73% | 83.74% |

通过上表可以看出，本研究所采用的基于改进TF-IDF方法的情感词领域性计算方式在总体的准确率上相较于常规方法表现更好，对于正负向的语料的情感判断精度较为平均，f1值较高可以证明该方法相较于传统的TF-IDF值在分类任务中有更好的效果。

（3）文本情感分类对照实验

为了验证本文所提出的情感词典构建方法具有更好的效果，所以将生成的高维情感词典MCBSL与大连理工大学的情感本体、哈工大的Chinese-BERT-wwm（词向量生成使用bert-as-service工具）、基于nlpir微博内容语料库的23万条微博数据以及ChnSentiCorp-Htl-uba-10000的1万条酒店评论数据训练的word2vec词向量以及fasttext词向量、单独使用TF-IDF算法计算权重生成的情感词典进行对照实验。对ChnSentiCorp-Htl-ba-4000的4000条酒店评论数据进行10次简单交叉验证，取平均值，设定领域情感词典MCBSL的维度为100维，取10次平均得到下表结果：

表5中文情感分类对照实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 文本表示方法 | macro precision | macro recall | macro f1 | accuracy |
| 情感本体 | 76.46% | 75.21% | 74.94% | 75.25% |
| BERT | 63.36% | 63.33% | 63.31% | 63.33% |
| word2vec | 83.24% | 83.21% | 83.22% | 83.24% |
| fasttext | 83.57% | 83.55% | 83.54% | 83.55% |
| TF-IDF | 83.18% | 83.10% | 83.09% | 83.10% |
| **MCBSL** | **84.05%** | **83.99%** | **83.98%** | **83.99%** |

同时为了验证方法在英文语料中的效果，分别对5个领域的商品进行了分类实验。其中各个领域的词向量训练数据采用各个领域的标注数据以及非标注数据，详细结构见表1。情感词典采用SenticNet6.0，BERT方法的模型采用google的cased\_L-12\_H-768\_A-12英文预训练模型，词向量的训练数据为各领域的标注数据加上无标注数据。对训练数据进行10词简单交叉验证，取分类准确率的评价值作为评价指标，设定领域情感词典MCBSL的维度为100维。

表6英文情感分类对照实验结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | book | dvd | electronics | kitchen | video |
| SenticNet | 61.55% | 64.80% | 66.19% | 68.11% | 66.13% |
| BERT | 49.74% | 50.16% | 49.72% | 49.91% | 49.74% |
| word2vec | 68.48% | 68.69% | 68.32% | 68.93% | 69.41% |
| fasttext | 67.42% | 65.86% | 67.74% | 69.26% | 69.53% |
| TF-IDF | 69.14% | 70.96% | 67.35% | 67.68% | 71.39% |
| MCBSL | **70.66%** | **71.43%** | **71.73%** | **71.22%** | **72.31%** |

通过表5可以看出，在准确率方面本研究中的领域情感词典取得了比情感本体以及两种词向量更好的效果，并且在正负向的情感分类中表现更为均衡。并没有出现不平衡的现象。表6也可以体现出在英文语料中本方法依然具有较好的效果。

（4）领域情感词典维度对照实验

领域情感词典的维度也会对结果造成较大的影响，所以通过选取不同的种子词数量来控制情感词典的维度，通过不同维度下的分类结果来验证本方法相较于使用TF-IDF计算权重的方法在分类问题中表现更佳。修改实验3中的维度参数，对数据进行5次简单交叉验证，对照了维度为20，50，100，120的情况。各维度下的结果如下：

表7 维度对照实验

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | macro precision | macro recall | | | macro f1 | | accuracy |
| 生成方法 | 维度20 | | | | | | |
| TF-IDF | 81.16% | | 81.07% | 81.07% | | 81.08% | |
| MCBSL | 81.94% | | 81.90% | 81.90% | | 81.91% | |
|  | 维度50 | | | | | | |
| TF-IDF | 82.68% | | 82.60% | 82.60% | | 82.61% | |
| MCBSL | 83.95% | | 83.90% | 83.90% | | 83.91% | |
|  | 维度100 | | | | | | |
| TF-IDF | 83.78% | | 83.73% | 83.72% | | 83.74% | |
| MCBSL | 84.10% | | 84.05% | 84.05% | | 84.06% | |
|  | 维度120 | | | | | | |
| TF-IDF | 83.86% | | 83.80% | 83.80% | | 83.82% | |
| MCBSL | 84.11% | | 84.06% | 84.06% | | 84.07% | |
|  | 维度150 | | | | | | |
| TF-IDF | 84.09% | | 84.03% | 84.04% | | 84.05% | |
| MCBSL | 84.18% | | 84.13% | 84.13% | | 84.14% | |
|  | 维度200 | | | | | | |
| TF-IDF | 84.13% | | 84.07% | 84.08% | | 84.09% | |
| MCBSL | 84.20% | | 84.15% | 84.15% | | 84.16% | |
|  | 维度300 | | | | | | |
| TF-IDF | 84.12% | | 84.07% | 84.07% | | 84.08% | |
| MCBSL | 84.22% | | 84.17% | 84.17% | | 84.18% | |

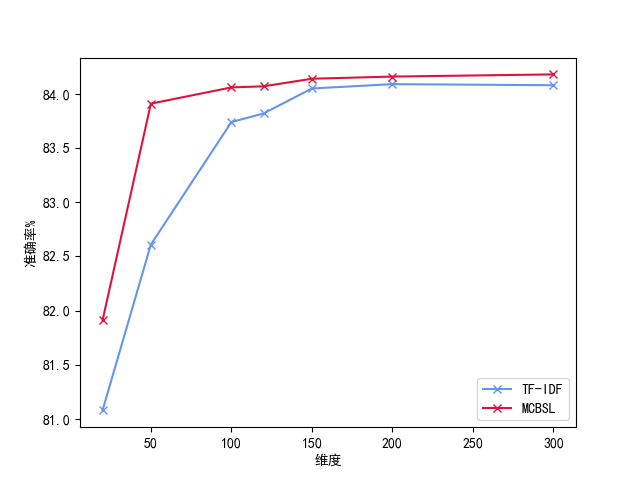


图4 维度-准确率变化关系图

通过图4可以看出，当维度大于150维时准确率趋于平稳，不再有太大的波动。同时在合理的维度范围内，MCBSL的效果总是优于TF-IDF算法的。