北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2014.013

面向微博短文本的细粒度情感特征抽取方法

贺飞艳[1]　何炎祥[1]　刘楠[1,2]　刘健博[1]　彭敏[1,†]

1. 武汉大学计算机学院，武汉　430072; 2. 军事经济学院军需系，武汉　430035; † 通信作者, E-mail: pengm@whu.edu.cn

摘要　结合 TF-IDF 方法与方差统计方法，提出一种实现多分类特征抽取的计算方法。采用先极性判断，后细粒度情感判断的处理方法，构建细粒度情感分析与判断流程，并将其应用于微博短文本的细粒度情感判断。通过 NLPCC2013 评测所提供的训练语料对该方法有效性进行验证，并在 NLPCC2013 评测任务中运用该方法，证实该方法具有较好的抽取效果。

关键词　自然语言处理；文本情感分析；细粒度情感；多分类特征抽取

中图分类号　TN914

A Micro-blogging Short Text Oriented Multi-class Feature Extraction Method of Fine-grained Sentiment Analysis

HE Feiyan[1], HE Yanxiang[1], LIU Nan, LIU Jianbo[1], PENG Min [1, †]

1. School of Computer,Wuhan University, Wuhan 430072; 2. Department of Quartermaster,Military Economic Academy,Wuhan 430035;
† Corresponding author, E-mail: pengm@whu.edu.cn

Abstract　Combined with TF-IDF method and variance statistical forumla, a new method for the extraction of multi-class feature is presented. This micro-blogging short text oriented extraction method is used to determine the fine-grained sentiment type. Then the processes of fine-grained sentiment analysis is bulit. This method is verified by the train data of NLPCC 2013 sentiment analysis task. This method is used to praticipate the NLPCC 2013 evaluation, and the effectiveness of this method is proved by the good ranking of the subimitted data.

Key words　natural language processing; text sentiment analysis; fine-grained sentiment analysis; multi-class feature extraction

情感特征抽取是文本情感倾向分析的基础工作。由于中文词汇量巨大，如果为所有词汇人工标记其情感倾向，用以判断文本内容的情感，虽然理论上可行，但需要巨大的人力资源。而且，随着网络语言的不断丰富以及词汇在不同领域中所表达含义的不同，加大了建立情感词典的难度。此外仅依靠人工整理的情感词典，识别效果受其规模和更新速度的影响。本文尝试以人工标记情感倾向的短文本语料和爬取的带评分的产品评价作为粗糙的标注训练语料，从中抽取可用的情感特征，扩充现有的情感词典来提高文本情感倾识别的准确率。

由于人类情感丰富多样，对其情感划分并没有统一的标准。常见的划分方法如赵妍妍等[1]把情感信息的分类任务分为两种：1) 主、客观信息的二元分类；2) 是主观信息的情感分类，包括最常见的褒贬二元分类以及更细致的多元分类。对于多元分类，徐琳宏等[2]在 Ekman[3]提出的 6 类情感基础上，将褒义细分为“好”和“乐”，形成 7 类情感。

对于特征抽取，Yang 等[4]针对文本分类问题，分析和比较信息增益 IG(Information Gain)、DF (Document Frequency)、互信息 MI(Mutual Information)和卡方统计 CHI(Chi-Square statistic)等方法后，认为 CHI 和 IG 方法效果较好。熊忠阳等[5]分析 CHI 方法仅考虑文档频率的不足，提出通过增加词频等

国家自然科学基金(61070083)和国家科技支撑计划子课题(2011BAK08B03-01)资助
收稿日期: 2013–07–04; 修回日期: 2013–09–17; 网络出版时间: 2013-11-11 10:25

信息改进其识别效果。Zagibalov 等[6]提出采用非监督训练方法抽取情感种子词，用以判断中文文本的极性。Barbosa 等[7]利用 Tweets 标注数据作为训练数据，通过提取特征，采用二步分类法对 Tweets 的情感倾向进行分类，即先进行主客观分类，然后再对分为主观的 Tweets 进行正负向情感分类。Li 等[8]用图模型方法来判断文本的极性。

1 细粒度特征抽取方法

如果把汉语词汇直接用作文本特征进行情感分类，其数据规模巨大，计算复杂度高。挑选具有强烈情感信息的特征词汇作为分类的依据，则可以大幅度降低特征的维数并提高分类的准确率。我们采用分级处理的方法，先进行粗粒度处理，找出具有情感情感倾向的词汇，判断其情感极性，再对这些词汇进行细粒度处理，计算其对各类情绪的影响。尝试考虑词频、反文档频率和各分类中文档的先验分布以及特征项在各类情绪中的分布情况，获得一种具有较高准确率的细粒度情感特征抽取方法。

1.1 权重计算

TF-IDF 是一种常用的权重计算方法，由于考虑了词频和反文档频率的影响，使得在少量文本中有较高的出现频率的词汇有较高的权重。我们利用分类中词频和文档频率之间关系，使用 TF-IDF 方法计算特征词在各分类中的权重。

对于由 K 个分类组成的文档集合 D,有 $D=\{D_1,\cdots,D_k\},k\in K$，$D_k$ 表示第 k 个分类中文档的总数量，对于 n 个特征词组成的特征词集:

$$T=\{t_1,\cdots,t_i\},i\in N。\tag{1}$$

词频(trem frequency): tf_{ik} 表示特征词 t_i 在文档 D_k 中出现的总次数。

反文档频(inverse document frequency): idf_{ik} 表示特征词 t_i 的反文档频率，d_{ik} 表示 t_i 文档集 D_k 中所出现的文档数量。idf_{ik} 的计算一般会设置一个常数以保证平滑。考虑到某一特征词在某类文档中可能并没出现，我们以式(2)进行计算:

$$\mathrm{idf}_{ik}=\log\left(\frac{D_k}{d_{ik}+0.5}+0.5\right),\tag{2}$$

则特征词 t_i 在文档集 D_k 中的权重为

$$W_{ik}=\mathrm{tf}_{ik}\times\mathrm{idf}_{ik}=\mathrm{tf}_{ik}\times\log\left(\frac{D_k}{d_{ik}+0.5}+0.5\right)。\tag{3}$$

通过权重计算，比单纯利用词频更能体现重要特征词对文档集的影响。但由于该方法本身不包含分类的相关信息，只能用于计算特征词在某一分类文档集中的权重，无法对比该特征词在不同类别之间的权重差异。实际上，人们更倾向于找出在某一类中有着突出表现的特征词，并将其归为该类中。

基于以上考虑，在计算特征词在各类文档集中 TF-IDF 权重的基础上，进一步考虑通过统计学上的方差方法，优选偏移量大的部分。在原本含义中，方差越大，意味着数据的波动越大，也就越不稳定。用在特征项与各类的关系中则可以认为，波动越大，词汇越有可能在某类或某几类中有突出值:

$$D(x)=E(x^2)-E(x)^2=\frac{1}{N}(\sum_{i=1}^{N}x_{\mathrm{i}}^2-N_x^{-2})。\tag{4}$$

如直接对 K 类中的某特征词 x_i 的词频 tf_i 计算其方差值，可得

$$D(\mathrm{tf}(x_i))=\frac{1}{K}(\sum_{k=1}^{K}\mathrm{td}_{ik}^2-K\overline{\mathrm{tf}_i}^2),\tag{5}$$

其中 $\overline{\mathrm{tf}_i}$ 为特征词 x_i 在各类中的平均词频，有 $\overline{\mathrm{tf}_i}=\frac{\sum_{k=1}^{K}\mathrm{tf}_{ik}}{K}$。

结合前面的 TF-IDF 方法计算出特征词在各类中的权重 W_{ik}，同样计算出其方差:

$$D(\mathrm{tf}_i\mathrm{df}(x_i))=\frac{1}{K}(\sum_{k=1}^{K}W_{ik}^2-K\overline{W}_i^2),\tag{6}$$

其中 $\overline{W}_i$ 为特征词 x_i 在各类中的平均权重，有 $\overline{W}_i=\frac{\sum_{k=1}^{K}W_{ik}}{K}$。

这样，就可根据方差值的大小进行排序，从而将偏差度大、词频较高且集中的特征词抽取出来。该词汇的权重不受分类中其他词汇的影响，影响词汇权重的为其词频、在分类中出现的文档数以及文档集本身的先验分布。

1.2 情感倾向判断

最简单有效的情感倾向判断方法是通过词频(TF)判断，特征词在哪个类别出现的频率最高，就判断为哪一类情感。但由于训练语料往往不具备完备性和平衡性的特点，单纯利用词频效果并不理想。通过权重计算可获取某一特征词在各个类别中的权重，如果一个特征词在多个类别中的权重相同，则无法单纯以权重值大小来判断其所属的情感类别，但如果一个特征在某一类别出现的权重值远高于其他类别，则可认为其有较大概率属于该类别。我们用 TF-IDF 方法计算出特征词在多个类别中的

权重大小, 选择权重最大的类别为该特征词的情感倾向。从结果可以明显看出, 选择 TF-IDF 方法判断情感倾向其准确率要高于 TF 方法。

计算出特征词 x_i 分别在 K 类中的 TF-IDF 权重 W_{ij} 后, 我们可考虑选择权重最大的类别为该特征词的情感倾向。

TF 判断情感倾向:

$$\text{Typeof}_{\text{tf}}(x_i) = \text{argmax}_k \text{tf}_i K, \tag{7}$$

TF-IDF 判断情感倾向:

$$\text{Typeof}_{\text{tfidf}}(x_i) = \text{argmax}_k W_i K \text{ 。} \tag{8}$$

1.3 归一化处理

常用的归一化方法是通过将特征项在某分类的值除以该特征项在各类的总值, 从而将其转换到[0,1]之间:

$$f(x_{ik}) = \frac{x_i}{\sum_{k=1}^{K} x_{ik}}。 \tag{9}$$

为了避免低频词可能出现的噪音, 我们先计算特征项在各类的算术平均值, 然后将小于算术平均值的数值去除, 仅保留高于平均值的部分, 然后再除以保留的总值, 作为归一化处理方法。

$$\bar{x}_i = \frac{\sum_{k=1}^{K} x_{ik}}{K}, \tag{10}$$

$$f(x) = \begin{cases} \dfrac{x_i}{\sum_{k=1}^{K}(x_{ik} \mid x_{ik} \geqslant \bar{x}_i)}, & x_{ik} \geqslant \bar{x}_i, \\ 0, & x_{ik} < \bar{x}。 \end{cases} \tag{11}$$

2 抽取流程

我们通过参加 NLPCC2013 中文微博情绪识别任务, 对细粒度情感特征的抽取方法进行验证。采用任务指定的情感本体库作为情感词典, 并通过本文提出的抽取方法从训练语料中抽取辅助识别的情感特征, 加以补充情感词典, 实现微博情感句的识别和微博的情感识别。

2.1 情感本体库

根据参与 NLPCC2013 评测任务的要求, 采用大连理工大学提供的情感本体库作为细粒度情感划分的依据, 其情感分为 7 大类 21 小类, 情感强度分为 1, 3, 5, 7, 9 共 5 档, 共含有情感词共计 27466 个。语料中 7 个大类的划分和训练语料的细粒度情感分类方式一致。

根据其分类特点, 我们将 7 个大类又归结为积极、消极和中立 3 大类, 形成如表 1 的树形关系, 用于指导我们的识别流程。

2.2 情感特征库

根据前述特征抽取方法, 我们从训练语料中抽取具有较强情感特征的表情、高倾向词汇、强调标点来构建情感特征库。通过将情感本体库与情感特征库结合在一起共同识别, 提高了识别效果。

2.2.1 表情符号

用户在网络上生成文本信息时, 受到输入方式的限制, 往往会选择“^ _^”笑脸等表情符号, 用来

表 1 细粒度情感类别从属关系

Table 1 Level relationship of fine-grained sentiment types

客观 Objective	主观(subjective)						
	积极(Postive)		消极(negative)				中立(erutal)
	乐(happy)	好(like)	怒(anger)	哀(sad)	惧(fear)	恶(disgust)	惊(surprise)

表 2 抽取出的部分特征项

Table 2 Part of extracted sentiment features

特征项	keguang	happy	like	anger	sad	fear	surprise	disgust
！/!	0.38	0	0.22	0.4	0	0	0	0
~/~	0.36	0.26	0.38	0	0	0	0	0
好/烦	0	0	0	0	0	0	0	1
啊啊/啊啊	0	0	0.16	0.56	0	0	0	0.28
给/力	0.33	0.19	0.47	0	0	0	0	0
尼/玛	0	0	0	1	0	0	0	0
节日/快乐	0.19	0.39	0.42	0	0	0	0	0
坑/爹	0	0	0	0.59	0	0	0	0.41
你/妹	0	0	0	1	0	0	0	0
他/妈	0.4	0	0	0.6	0	0	0	0
心/痛	0.14	0	0	0	0.86	0	0	0
大/哭	0	0	0	0	1	0	0	0

表达文字难以表述的情感。但在文本处理过程中，这种非正文的字符往往被停止词过滤掉。这些表情符号难以通过现有分词系统获得，因此我们对训练语料采用逐字符滑动的方法，提取窗口长度为 2~5 的高频字符组合，通过特征抽取方法获取，并加以人工整理。此外，对于部分只在测试语料中出现的高频表情符号，我们同样以字符滑动方式抽取出来，并对其人工打分，总共获得 87 个字符表情。

2.2.2 情感特征词

我们利用前述特征抽取方法，抽取训练语料中的高频词汇组合，并根据权重判断其情感倾向。首先通过分词系统进行预处理，然后根据分词结果选择单一词汇，以及对其后窗口长度为 2~5 的词汇组合进行逐词滑动，和表情同样的方法抽取出权重较高的词汇组合作为情感特征词用于辅助识别，该方法共获得约 200 个词汇组合。其中多为网络口头用语，未在情感本体库中收录。

2.2.3 强调标点

通过观察发现，微博用户往往会以连续的标点符号，或者连续的拟音词来表情其强烈的情感，如“??”“!!”“啊啊”之类。出现这样连续标点的时候，一般并非笔误，而是在用其表达比句子本身更强烈的情感倾向。对于此类强烈标点和拟音词，我们收集在句中连续出现 2 个以上的相同符号。通过特征抽取方法获得少量特征项，用以辅助识别。

3 处理流程

3.1 预处理

我们选用中科院 ICTCLAS 分词系统进行预处理，并通过正则匹配的方式去掉微博文本中的话题标签、@用户、网址等信息，以免影响判断结果。此外，训练样本中存在不少重复微博，我们人工删除个别重复率较高的微博。

3.2 权重计算

情感本体库对词汇具有情感强弱的划分，我们沿用其权重。对于抽取出来的情感特征库，对其归一化后，将归一化的值作为其权重，由于其权重整体上弱于情感本体库，可以认为是在其基础上加以修正。

对于微博中的某个句子，通过判断句中是否包含情感特征词，将该词分别所属的各个情感类别的分值累加，获得该句子各个细粒度情感的分布。我们针对 7 个细粒度情感分别设定其分数，anger 粒度分数 AngerValue、disgust 粒度分数 DisgustValue、fear 粒度分数 FearValue、happiness 粒度分数 HappyValue、like 粒度分数 LikeValue、sadness 粒度分数 SadValue、surprise 分数 SurpriseValue。此外，还设定主观分数 SubjectValue，其分数为 7 个细粒度情感的总和，积极情感分数 PosValue 为 happiness 和 like 分数之和，消极情感分数 NegValue 为 anger、disgust, fear 和 sadness 分数之和，把 surprise 分数作为中立情感分数，其分值不计入积极或消极中。

$$\begin{aligned}\text{SubjectValue} &= \text{PosValue}+\text{NegValue}+\text{SurpriseValue}\\ &= (\text{HappyValue}+\text{LikeValue})+\\ &\quad (\text{AngerValue}+\text{DisgustValue}+\\ &\quad \text{FearValue}+\text{SadValue})+\\ &\quad \text{SurpriseValue}。\end{aligned}\tag{12}$$

考虑到句子中可能出现的否定反转现象，我们利用自己收集整理的否定词典，当情感词出现在否定词周边窗口为 2 的距离之内时，则对其进行情感倾向进行反转。被否定反转的特征词的细粒度情感不加分，如果属于积极情感，则对消极情感 negValue 加其权重的分数，如果属于消极极情感，则对积极情感 PosValue 加其权重的分数。

3.3 情感判断

通过计算的权重分数，我们采用分级处理的方法，先判断主观、客观，再判断积极、消极和中立(这里的中立指 surprise 类别)，最后辨别其他细粒度情感。通过计算出微博句中的第一和第二情感，再汇总得出整个微博的情感倾向。

具体流程如下(见图 1)。

1) 主客观的判断。如果 $\text{SubjectValue} > 0$，则为主观句，转 2，否则判断为客观句。

2) 积极消极的判断。根据 posValue、negValue、SurpriseValue 值的大小判断。如果 posValue 值最大则为积极句；如果 negValue 值最大则为消极句；如果 SurpriseValue 值最大则为中立句，判断最大情感为 Surprise。

3) 细粒度的判断。如判断为积极或消极句，再根据细粒度情感中值最大者，判断其情感倾向。对积极句进行 happiness 和 like 的判断，对消极句进行 Anger, Disgust, Fear 和 Sadness 的判断。

4) 第二情感判断。如需判断第二情感，则去除已选择的最大情感值，为 7 个细粒度情感中值排名第一的作为第二情感。如一个句子中，其值为 $\text{disgustValue}=3$，$\text{happyValue}=4$，$\text{angerValue}=2$，则

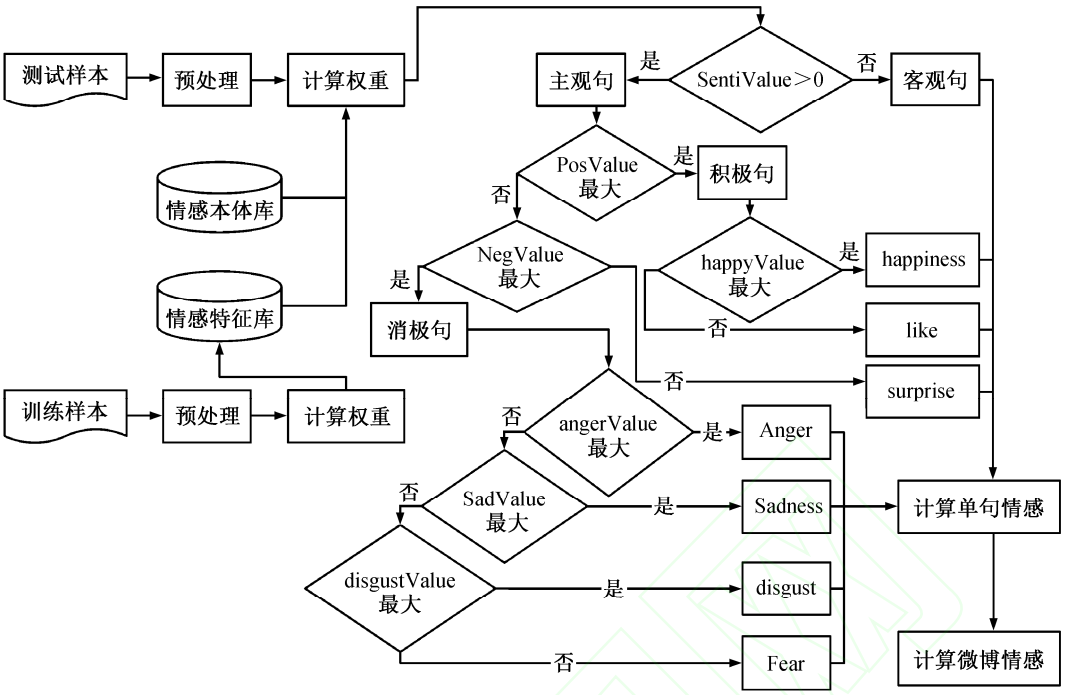

图 1　细粒度情感分析流程

Fig. 1　Flowchart of fine-gtained sentiment analysis

首先判断其主要情感为消极句, 然后判断最大细粒度情感为 disgust, 去除该情感后, 选择排名第一的细粒度情感值 happyValue, 判断 happiness 为第二情感。

5) 微博的情感判断。微博情感判断是在判别出微博句子的粒度情感基础上进行的, 采用句子的第一情感作为判别基础, 微博句子中判断为某一个粒度情感最多的句子的情感作为整个微博情感, 如果各细粒度情感判断的句子数量相等, 则把细粒度情感值最大的那个类别作为微博情感。如同时有多个类别的情感值相等且都为最大, 就再统计微博句中积极、消极和中立句出现的数目, 选择数目最多的类别, 再辨别为其中情感值最大的细粒度情感类别。

4 实验与分析

4.1 实验设计

4.1.1 实验数据

我们选用 NLPCC2013 评测任务所提供的训练语料对该特征项抽取方法进行实验, 其中包括 4000 条微博, 人工去重后, 共整理 13252 句。其中标记了客观句与 happiness, like, anger, disgust, fear, sadness 和 surprise 共 7 种细粒度情感。统计情感本体库中词汇在该语料中的出现情况, 共有 3533 个词汇出现。抽取其判别结果为 7 个细粒度情感的部分, 与情感本体库中标记的情感类别进行对比。对于客观值最高的, 则认为该特征并非情感特征, 不予抽取。

4.1.2 实验方法

作为对比, 我们选择 TF 方法、CHIMAX 方法、总 TF-IDF 方法、D(TF)方法、D(TF-IDF)方法来计算权重, 并对各个方法分别采用 TF 和 TF-IDF 判断情感倾向。在方法后加上“-TF”和“-TFIDF”后缀予以标识。

1) TF 方法。TF 方法为最简单直观的方法, 仅统计特征项在各细粒度情感类别中出现的词频数。我们以此为 baseline。

2) CHIMAX 方法。CHI 方法(χ^2 检验)可以衡量特征项 t 与类别 c 的相关程度, 为常见的特征抽取方法, 因此拿来作为对比。该方法考虑包含特征项的文档频率, 在微博中可将一条博文作为一个文档来统计。如果 A 表示包含词条 t 且属于类别 c 的文档频数, B 为包含 t 但是不属于类别 c 的文档频数, C 表示属于类别 c 但是不包含 t 的文档频数, N 表示语料中文档总数, D 表示既不属于 c 也不包含 t 的文档频数。则 t 对于 c 的 χ^2 值为

$$\chi^2(t,c)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)}。\qquad(13)$$

考虑到 N, $A+C$, $B+D$ 均是常数, 上式可以简

化为

$$\chi^2(t,c)=\frac{(AD-BC)^2}{(A+B)(C+D)}。\tag{14}$$

特征 t 与类别 c 的统计相关性越强, $\chi^2(t,c)$ 的值就越大, 此时特征 t 包含的与类别 c 有关的鉴别信息就越多。由于 CHI 本身仅考虑类别是否包括含有特征项的文档。因此适用于两类中计算相关性, 对于多类划分, 则选其在各类中的最大值为其权重:

$$\chi^2_{\max}(t)=\max_{1\leqslant k\leqslant K}\{t,c_k\}。\tag{15}$$

我们选用此方法作为对多类情感权重计算的方法, 称为 CHIMAX 方法。此外, 还有采用平均值的方法, 经过实验发现其结果相差不大, 因此未将其记录。

3) TF-IDF 方法。TF-IDF 方法无法用来比较特征项在各个子类别中的权重。我们选用该方法, 用其来计算特征项在整个文本集中的权重, 并根据权重大小进行排序。其中 TF 指特征项在训练样本中所有分类中的词频总和, 反文档频率 IDF 统计在训练样本中该特征项所出现的文档总数。

4) D(TF)方法。我们将方差统计方法和最基本的词频数相结合, 用于观察该方法本身对词频所产生的变化, 对特征项在各类别中的词频数计算方差, 并根据其结果大小排序, 如式(5)。

5) D(TFIDF)方法。该方法是本文所提出的方法, 通过结合方差统计方法和各分类 TF-IDF 权重计算方法, 首先计算出各特征项在子分类中的 TF-IDF 值, 然后在对 TF-IDF 值计算方差, 并根据结果大小排序, 如式(6)。

4.2 结果与分析

4.2.1 实验 1

由于不同方法其权重设定不同, 我们将各个方法的结果按照大小进行排序, 分别抽取排序靠前的 50, 100, 150, 200 个结果, 计算其准确率。考虑到低频词噪音信息较大以及低频词容易受到排序方式等外部影响, 我们过滤词频为 1 的词汇, 并分别按拼音正序、倒序进行排列统计结果。

从表 3 可见, 对于不同方法, 采用 TF-IDF 判断情感, 其效果都好于采用 TF 判断情感, 说明选用 TF-IDF 判断情感的方法具有更好的效果。在不同方法对比中, 当抽取范围为 50 时, 各种方法其准确率相差不大, 甚至单纯统计词频具有更高的准确率, 这时词频高低对结果影响更为明显, 而随着抽取范围扩大, 采用方差统计方法的准确率高于其他方法。最终将所有抽取的结果予以平均, D(TFIDF)-TFIDF 方法具有最高的准确率, CHIMAX-TFIDF 方法的准确率与其较为接近。CHIMAX 方法由于仅考虑特征词在文档中出现的频率, 倾向于选择出现文档频率较高的特征词, 但对于低频词的准确率较低而且波动较大。

4.2.2 实验 2

此外, 我们对不同词频下 TF 和 TFIDF 辨别的准确率进行对比, 实验结果见表 4。随着去除低频词汇数量的增多, TFIDF 的召回率下降较大, 但其准确率明显高于 TF 方法。通过该方法从训练语料中抽取出的特征词虽然规模较小, 但具有更高的置信度。进一步证明了 TFIDF 判断情感方法的有效。

表 3 实验 1 结果

Table 3 Result of Experiment 1

抽取方法	升序排列结果				降序排列结果				汇总准确率/%			
	TOP 50	TOP 100	TOP 150	TOP 200	TOP 50	TOP 100	TOP 150	TOP 200	TOP 50	TOP 100	TOP 150	TOP 200
CHIMAX-TF	35	65	85	98	35	65	85	100	70.00	65.00	56.67	49.50
CHIMAX-TFIDF	32	65	84	111	32	65	84	106	64.00	65.00	56.00	54.25
TF-TF	35	58	78	95	35	58	77	95	70.00	58.00	51.67	47.50
TF-TFIDF	32	56	79	109	31	55	80	102	63.00	55.50	53.00	52.75
TFIDF-TF	35	58	78	93	36	58	77	95	71.00	58.00	51.67	47.00
TFIDF-TFIDF	31	56	77	106	31	54	74	102	62.00	55.00	50.33	52.00
D(TF)-TF	36	62	81	103	36	62	80	103	72.00	62.00	53.67	51.50
D(TF)-TFIDF	34	58	80	107	33	58	80	106	67.00	58.00	53.33	53.25
D(TFIDF)-TF	31	59	83	102	36	63	83	104	67.00	61.00	55.33	51.50
D(TFIDF)-TFIDF	34	60	83	110	34	60	83	109	68.00	60.00	55.33	54.75

4.3 评测结果

我们采用最终选取的 D(TFIDF)-TFIDF 抽取方法参加 NLPCC2013 评测的中文微博情绪分析评测任务。该任务包括微博情绪识别与分类(任务 2.1)和情绪句识别和分类(任务 2.2)两个子任务。任务 2.1 共有 19 组队伍提交 58 组结果, 由于通过训练语料扩展了情感特征库, 从而提高了召回率, 拉高了 *F* 值。在评测中我们获得中等靠前的成绩(表 5)。在任务 2.2 共有 8 组队伍提交 12 组结果, 我们具有领先的平均精度(表 6)。从而证实该抽取方法是有效的。

由于该方法仅考虑句中的情感倾向, 未进一步考虑上下文关系, 因此在评测中部分情感倾向未能识别出来, 此外, 由于 TF-IDF 方法对词频与句频关系较为敏感, 如果一句中由于表达强烈情感而出现连续重复的词汇, 或文档集中出现重复的文本, 则会影响权重的计算, 对文档集的预处理会直接影响

表 4　实验 2 结果

Table 4　Result of experiment 2

去除低频词	TF				TFIDF			
	正确数	召回数	准确率/%	召回率/%	正确数	召回数	准确率/%	召回率/%
0	413	987	42	12	426	967	44	12
1	123	276	45	7	136	256	53	8
2	79	157	50	7	71	137	52	6
3	60	105	57	7	50	85	59	6
4	49	81	60	7	38	63	60	6
5	40	64	63	7	29	47	62	5
6	35	49	71	8	23	31	74	5
7	30	39	77	8	20	24	83	5
8	25	33	76	8	15	18	83	5
9	21	26	81	7	14	15	93	5
10	19	22	86	7	11	12	92	4

表 5　NLPCC2013 评测任务 2.1 结果

Table 5　Result of NLPCC2013 Task 2.1

结果序号	情绪句判别任务/%						情绪识别任务 Close/%						情绪识别任务 Open/%					
	Close			Open			宏平均			微平均			宏平均			微平均		
	正确率	召回率	*F* 值	正确率	召回率	*F* 值	正确率	召回率	*F* 值	正确率	召回率	*F* 值	正确率	召回率	*F* 值	正确率	召回率	*F* 值
15-1	52.63	94.36	67.57	52.13	97.10	67.84	21.55	28.03	24.36	21.64	38.80	27.78	20.93	28.36	24.09	21.25	39.59	27.66
15-2	52.63	94.36	67.57	52.13	97.10	67.84	21.60	28.10	24.42	21.48	38.52	27.58	21.22	28.53	24.34	21.08	39.27	27.44
avg	58.92	64.82	58.03	59.68	79.55	66.71	21.45	19.33	19.54	24.81	26.84	24.14	23.99	26.60	24.71	25.80	34.32	28.85
max	74.94	95.17	72.71	70.53	97.10	72.86	28.44	30.64	28.73	38.34	39.76	34.12	28.42	34.80	31.29	32.32	**39.59**	35.21

表 6　NLPCC2013 评测任务 2.2 结果

Table 6　Result of NLPCC2013 Task 2.2

结果序号	情绪句识别和分类/%			
	Close		Open	
	平均精度(宽松指标)	平均精度(严格指标)	平均精度(宽松指标)	平均精度(严格指标)
15-1	33.25	32.08	**36.50**	**34.84**
15-2	32.43	31.87	–	–
avg	24.87	24.11	26.83	25.98
max	34.39	33.05	**36.50**	**34.84**

到最终结果。

5 结语

本文通过将 TF-IDF 方法和方差方法相结合，使其可在多分类中的权重具有可比性，从而提出一种从多分类中抽取情感特征的方法，并将其应用于微博短文本的细粒度情感特征抽取中，构建了细粒度情感分析与判断流程。并运用该方法对 NLPCC2013 的训练样本集做实验以及参加 NLPCC2013 评测任务，证实该方法具有较好的抽取效果。

参考文献

[1] 赵妍妍，秦兵，刘挺. 文本情感分析. 软件学报，2010, 21(8): 1834−1848

[2] 徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造. 情报学报, 2008, 27(2): 180−185

[3] Ekman P. Facial expression and emotion. American Psychologist, 1993, 48(4): 384

[4] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization. ICML, 1997, 97: 412−420

[5] 熊忠阳，张鹏招，张玉芳. 基于χ^2 统计的文本分类特征选择方法的研究. 计算机应用，2008, 28(2): 513−514

[6] Zagibalov T, Carroll J. Automatic seed word selection for unsupervised sentiment classification of Chinese text // Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, 2008: 1073−1080

[7] Barbosa L, Feng J. Robust sentiment detection on twitter from biased and noisy data // Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), 2010: 36−44

[8] Li Binyang, Zhou Lanjun, Feng Shi, et al. A unified graph model for sentence-based opinion retrieval // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 1367−1375