基于情感标签的极性分类

周 孟,朱福喜
(武汉大学计算机学院,湖北武汉 430072)

摘 要: 情感极性分析是文本挖掘中一种非常重要的技术.然而在不同领域中,很多情感极性分类系统存在分类精度低和缺少大量标注数据的缺陷.针对这些问题,提出了一种基于情感标签的极性分类方法.首先通过所有文本建立 Sentiment-Topic 模型,抽取出文本的情感标签;然后利用情感标签将文本划分为两个子文本,并通过 Co-training 算法对子文本进行分类;最后合并两个子文本的分类结果,并确定文本的情感极性.实验结果表明该方法具有较高的分类精度,而且不需要大量的分类样本.

关键词: 极性分类;情感标签;半监督学习;co-training 学习
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2017)04-1018-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.04.034

Polarity Classification Based on Sentiment Tags

ZHOU Meng,ZHU Fu-xi
(*Computer School*,*Wuhan University*,*Wuhan*,*Hubei* 430072,*China*)

Abstract: Sentiment analysis is a very important technology in text mining. However,a number of systems require amounts of annotated training data in different fields. In order to solve these problems,an approach to polarity classification based on sentiment tags is proposed. Firstly,on the basis of all the documents,the sentiment-topic model is developed and the sentiment tags for each review are extracted. Then each review is divided into two sub-texts by these sentiment tags,and each sub-text is classified by exploiting the co-training algorithm. Finally,the category results of two sub-texts are combined to determine document-level polarity of each review. Experimental results show that compared with other algorithms,the method improves the classification precision without a large number of annotated samples.

Key words: polarity classification;sentiment tag;semi-supervised learning;co-training learning

1 引言

近些年来,文本情感极性分析受到越来越多的关注[1~5],这是因为许多用户在互联网上分享自己的观点或体验来表达自己的各种情感色彩和情感极性,如喜、怒,哀、乐和批评、褒扬等[6].此外,生产厂商可以根据大众的口碑有针对性地提高产品的服务质量,赢得更多的经济效益.

目前,由于文本中含有复杂情感表达和大量的情感歧义,导致分类的准确率很低.例如,“交通方便服务周到,很不错的酒店.唯一不满的是价格没有竞争力,比汉庭还贵.而且没有早餐送,比较遗憾.”作者首先给予了正面评价,之后又提出了不足之处.而评论的情感极性是正面的,正是由于情感词“方便”、“周到”和“不错”的重要性远远要高于其他情感词.并且“方便”与“周到”、“不错”之间存在着潜在的联系.这是因为在对酒店正面评价时,自然而然地,文本中还有一些其他对酒店正面评价的情感词.因此,文本的情感极性分类是一项非常复杂的任务[7],如何从文本中抽取出关键的情感词,将对文本的情感极性分析具有极其重要的作用.

针对上述问题,本文研究通过情感标签对文本进行情感极性分析.所谓情感标签,是指文本中那些关键的情感词.在情感标签抽取时,考虑了情感词的3个特征,分别是情感词的关键度、情感词间的相关度和情感词的位置特征.关键度是用来衡量该词在文本中的重要程度,相关度是衡量其他词对该词的影响程度,位置特征是该词所属句子在文本中的具体位置.

收稿日期:2015-10-26;修回日期:2016-03-07;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61272277)

本文首先抽取情感标签,然后通过情感标签将文本划分为包含情感标签的文本和不包含情感标签的文本. 然后对两个子文本进行 Co-training 学习,并进行极性分类. 最后合并两个子文本的分类结果,进而得到文本的整体极性分类.

2　相关工作

文本情感极性分类的方法很多,按照语言学的粒度,大致可以分为词语级[2]、短语级[8]、句子级[9~11]和篇章级[1]根据学习方法的不同,文本的情感极性分类大致可以分为三种,即无监督学习、半监督学习和有监督学习. 由于本文的方法属于半监督学习,因此本节着重研究三种学习方法的情感极性分类.

在无监督学习中,Turney[2]使用了词对进行情感极性分类. 该方法首先对文本中的词进行词性标注;然后通过一些预定义的规则抽取一些词对搭配,使用点互信息(PMI)的方法进行词对的情感极性数值计算;最后对文档中的所有词对的情感极性数值综合计算,根据计算结果判断整个文档的情感极性. 冯时等人[4]提出了一种基于依存句法分析技术的情感极性分类:首先使用句法分析工具分析文本;然后抽取预定义的依存关系对,通过语法距离计算修饰词对情感词的修饰强度;最后通过修饰强度和情感词的初始分值计算文本的情感极性分值,根据结果的正负号判断该文本所属的类别. Zagibalov 等人[12]提出了一种基于种子词的情感极性分类技术,并将其用于情感极性分类系统中.

在半监督学习中,Dasgupta 等人[13]引入了谱分析技术,用于发现无歧义的评论,然后在这些评论的基础上,通过主动学习、直推式学习和集成学习相结合的方法对评论进行情感极性分析. Goldberg 等人[14]使用了基于图的半指导分类算法,对用户评论进行褒贬分类. 此外,Wan 等人[15]提出了基于 Co-training 的方法,并通过有标注的英语语料和无标注的中文语料训练学习,从而得到中文情感极性分类器. Li Tao 等人[16]提出了基于词语的先验知识和矩阵分解的方法来解决情感极性分类问题. OUYANG 等人[17]同时考虑篇章级和词语级两个粒度上的情感/主题分布,并提出多粒度的主题情感混合模型,通过实验证明该算法分类效果优于主题情感混合模型的分类效果.

在有监督学习中,Pang B 等人[1]将文本的情感极性分类看作为二值分类问题,即只有正面的和负面的,并使用了机器学习的方法对电影评论进行了情感极性分析,发现使用 unigram 特征的方法分析的效果最好. 然而,Cui 等人[9]通过实验证明,unigram 特征适用于训练语料比较少的情况下,随着训练语料的增多,n-gram 特征分析的效果更优. Kim 等人[10]除了考虑 n-gram 模型之外,还引入了位置特征来对句子进行情感极性分类. Zhao 等人[11]将句子情感极性分类划分为三层分类,并使用 CRF 模型对其分类结果进行融合. Melville 等人[18]将词典知识应用于博客的情感极性分类. Li 等人[19]从主客观的角度先对文本进行分类,然后分别训练分类器,最后对分类器的结果进行融合,也取得了一定的效果. 此外,卜湛等[20]人提出了一种高效的情感计算框架去获得短文本的情感倾向,并且将情感计算和博弈论相结合提出情感演化预测模型算法,进而对交互行为进行预测,并验证了情感计算框架的有效性和情感演化预测算法的准确性.

在无监督的情感分类中,对其分类的结果需要进行大量的分析和后处理,才能得到可靠的分类结果. 在有监督的情感分类中,人为主观因素较强,并且训练样本的选取和评估需要较多的人力和时间. 因此,结合无监督学习和有监督学习的优缺点,本文在情感极性分析时采取了半监督的学习方法. 此外,为了区分那些情感复杂的文本,本文通过情感标签将其划分为两个子文本,然后对子文本进行情感分类,最后对两个子文本的分类进行融合,进而得到文本的情感极性分类.

3　基于情感标签的 Co-training 分类

为了划分复杂情感的文本,本文提出了一种基于情感标签的 Co-training 分类方法. 该分类方法分为抽取情感标签和极性分类两个过程. 在抽取标签过程中,首先通过大量文本,建立情感主题模型(Sentiment-Topic model),然后根据 Sentiment-Topic 模型,计算文档 d 中的情感词的关键度、相关度和情感词的位置权重,最后通过情感词的三个特征的权重,计算情感词的总权重. 对文档 d 中的所有情感词按总权重进行降序排序,选择出靠前的若干情感词作为文档 d 的情感标签. 由于情感标签是反映文档的具体情感的重要标志,因此这些标签所在的那些句子也显得尤为重要. 于是本文在情感分类过程中,首先根据文档 d 是否含有情感标签,将其划分为两个子文本 d_1 和 d_2,再对两个子文本 d_1 和 d_2 进行 Co-training 分类学习,并对学习的结果进行融合,最后由融合的结果来确定文档 d 的情感类别.

3.1　Sentiment-Topic 模型的建立

文本中情感极性往往通过文本中的情感词来判断,而情感词大部分是一些词性为形容词、动词或修饰副词等这些词性的词语. 于是本文在所有文档上训练 Sentiment-Topic 模型,并且在训练过程中只使用了形容词、副词和动词等不同词性的词语,没有考虑文本中其他词性的词语.

Sentiment-Topic 模型以 LDA 模型[8,21,22]为基础,通

过概率模型来抽取文本中的情感标签. 虽然 Sentiment-Topic 模型与 LDA 模型有些类似,但 Sentiment-Topic 模型主要是针对情感词的模型. 如图 1 所示,在模型中引入了情感词检查过程,通过此过程可以过滤非情感词,从而仅根据情感词建立文档-情感主题分布 θ 和情感主题-词语分布 φ.

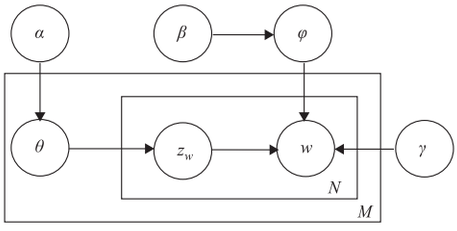

图1 Sentiment-Topic模型

在图 1 中,w 表示一个情感词语;γ 用来检查词语 w 是否为情感词,其取值为 1 或 0,如果值为 1,则表示词语 w 是情感词,否则不是情感词,并过滤掉词语 w;z_w 表示词语 w 的情感主题;α 和 β 是模型的先验分布参数,由于没有理由假设某个情感主题的强度有高或低,因此在该模型中所有隐含的情感主题都是平等的,于是本文中的 α 的所有分量值取值相同,同时 β 的所有分量值取值也相同. 对于语料库中的每篇文档,Sentiment-Topic 模型的具体生成过程如下:

(1)对于每篇文档,从情感主题分布中抽取一个主题;

(2)从上述被抽取到的主题所对应的词语分布中抽取一个词语;

(3)检验抽取到的词语是否为情感词;

(4)重复上述过程直至遍历文档中的所有情感词.

3.2 情感标签的抽取

情感标签是文本中那些具有较高权重的形容词、副词或动词等. 在计算情感词语的权重时,综合考虑了情感词的三个特征值,即关键度、相关度和位置权重.

3.2.1 情感词的关键度计算

情感词的关键度是体现情感词在文本的重要程度,关键度越大,说明该词越重要,反之亦然. 关键度可以通过很多方法来计算,比如词频、概率等等. 本文采用了概率的方法,即词语 w 在文本 d 中出现的概率,其计算公式如下:

$$C(w) = \sum_{k=1}^{K} p(w|z_k)p(z_k|d) \tag{1}$$

其中 K 为情感主题的数目,$p(w|z_k)$ 为情感词 w 在情感主题 z_k 下的概率,$p(z_k|d)$ 为文本 d 在情感主题 z_k 的概率,$p(w|z_k)$ 和 $p(z_k|d)$ 的概率分别通过文档-情感主题概率分布 θ 和情感主题-词语概率分布 φ 来获得.

3.2.2 情感词的相关度计算

在实际评论中,用户对评价对象的不同维度或属性之间的观点评价明显存在一定的潜在关联性. 例如,如果用户认为酒店的服务很差,则会很自然而然地对酒店的其他方面提供贬义性的评价. 此时抽取的情感标签之间就存在一种潜在关联性. 于是本节将深入讨论如何有效地考虑情感词之间的关联信息来改善情感词在文本中的权重.

为了解决这个问题,本文提出了基于情感主题的协同过滤方法,其主要基本思想是在情感主题上挖掘不同词语之间的相关性,即通过皮尔逊相关系数 PCC 的方法. 由于不同词语在情感主题上的分布不同,于是两个情感词间的相关度可以通过所有情感主题上的皮尔逊相关系数来计算. 其计算公式如下:

$$\mathrm{PCC}(w_i,w_j) = \frac{\sum_{k=1}^{K}[p(w_i|z_k) - p'(w_i)][p(w_j|z_k) - p'(w_j)]}{\sqrt{\sum_{k=1}^{K}[p(w_i|z_k) - p'(w_i)]^2}\sqrt{\sum_{k=1}^{K}[p(w_j|z_k) - p'(w_j)]^2}} \tag{2}$$

其中 $p'(w_j)$ 表示词语 w_i 在所有情感主题上的平均分布.

由于在对某一事物进行正面(或负面)评价时,自然而然地,其周围可能还有其他正面情感词对该事物进行正面(或负面)评价,所以情感词的相关性是其他情感词(与其具有相同极性的词语)对该词的影响程度. 本文中采用了加权平均的方法来计算情感词的相关度,其计算公式如下:

$$R(w) = \frac{\sum_{y=1\wedge w\neq w_y}^{V}[p(w_y|d) \times |\mathrm{PCC}(w,w_y)|]}{\sum_{y=1\wedge w\neq w_y}^{V}|\mathrm{PCC}(w,w_y)|} \tag{3}$$

其中 $p(w_y|d)$ 表示词语 w_y 在文本中的出现的概率,V 表示文本中与词语 w 极性相同的数目.

3.2.3 情感词的位置权重计算

在整篇文档中,文本的开头和结尾通常是总结性的关键句子,对于整篇文档特别重要,而文本中的中间句子往往是细节信息,其重要性要低于文本的开头或结尾的句子. 因此开头和结尾句子中的情感词更重要,而中间句子中的情感词的重要性较低. 于是本文采用了以下方法来计算情感词的位置权重.

$$L(w) = a \times \mathrm{pos}(w)^2 + b \times \mathrm{pos}(w) + c \tag{4}$$

并且

$$-\frac{b}{2a} = \frac{n}{2};a>0;b<0;c = 1-a-b;\frac{4ac-b^2}{4a} = 0.001.$$

其中 n 为文档中的句子数目,$\mathrm{pos}(w)$ 表示词语 w 所在

句子的位置,它是一个整数. 实际上,$L(w)$是一条开口向上的抛物线,横坐标表示情感词所在句子的位置,取值在1到n之间,纵坐标表示情感词的位置权重. 此外,只要设置的a、b和c满足上述条件,那么计算的位置权重不仅位于(0,1]区间内,而且文本中正中间句子的情感词的位置权重也是最低的(默认最低值为0.001),开头和结尾句子中的情感词的权重较高.

3.2.4　情感标签的抽取

在抽取情感标签时,本文着重考虑了情感词的3个特征因素,通过这3个因素来计算情感词的综合权重,然后选择权重较高的情感词作为文本的情感标签. 由于关键度、相关度和位置权重是从情感词的3个维度来量化情感词的,属于不同的特征空间,因此本文没有直接采用算术平均的方法计算情感词的总权重,而是采用了调和平均的方法来计算情感词的总权重,即

$$\text{weight}(w)=\frac{3\cdot C(w)R(w)L(w)}{C(w)R(w)+C(w)L(w)+R(w)L(w)} \tag{5}$$

其中$C(w)$为情感标签的关键度,$R(w)$为情感标签的相关性权重,$L(w)$为情感标签的位置权重. 由于概率的性质,可知$C(w)$的值域范围为[0,1];在计算位置权重时,如果a、b和c三个参数满足设置的条件,那么$L(w)$取值在(0,1]区间内;由于相关系数的值域范围为[0,1],很容易证明$R(w)$的值域范围也为[0,1].

最后,根据总权重的大小,对所有情感词进行降序排列,选择出比较靠前的情感词作为文本的情感标签.

3.3　基于情感标签的 Co-training 学习算法

Co-training 算法最早是由 Blum 和 Mitchell 提出的[23],最初的 Co-training 算法只是对抓取的网页进行分类,并没有对文本的极性进行分析. 于是采用了 Co-training 算法的思路,并通过情感标签将训练文本集划分为两个子文本集:即包含标签的文本集合和不包含标签的文本集合. 然后训练两个基础分类器,用包含标签的文本集训练得到分类器f_1,用不包含标签的文本集训练得到分类器f_2. 在这两个分类器中进行测试时,不仅要输出文本的类别,还要输出所属每个类别的概率,最后对分类的结果融合. 融合的方法很多,本文采用了以下方法对分类结果进行融合,即

$$i=\arg\max p_1(c_i|d)p_2(c_i|d) \tag{6}$$

其中$p_1(c_i|d)$表示在分类器f_1中,文本d为类别c_i的概率,$p_2(c_i|d)$表示在分类器f_2中,文本d为类别c_i的概率.

基于情感标签的极性分类算法(Polarity Classification of the Tags based on Criticality, Location and Relevance, PCTCLR),如算法1所示.

算法1

输入:包含情感标签的文本集合S_1,不含情感标签的文本集合S_2,有标注的文本集合L,无标注的文本集合U,Sentiment-Topic 模型,标签个数k,参数m和n

输出:文本集合U中每个文本的情感类别

步骤:

While U 不为空
　for $d\in U$ do
　　抽取文档d中的所有情感词,并放入集合W中
　　for $w\in W$ do
　　　计算总权重 weight(w)
　　end for
　　对d中所有情感词按照总权重进行降序排列,并选择前k个词作为文本的情感标签
　　通过文本情感标签,将文本d划分为d_1和d_2两个子文本,分别加入S_1和S_2
　　通过文本集合S_1学习分类器f_1,通过文本集合S_2学习f_2
　　通过分类器f_1和子文本d_1得到文本d的正、负面概率,通过分类器f_2和d_2得到d的正、负面概率
　　通过式(6)合并两个分类器的结果,得到文本d的类别
　end for
　从U的最可信的文本中选择m个正面和n个负面的文本添加到L中,然后将其对应的子文本分别添加到S_1和S_2中,并将选择的文本从U删除
end while

4　实验方法与分析

4.1　实验设置

为了验证新方法的有效性,我们采用了中科院计算所谭松波博士精心整理的一个较大规模的酒店(hotel)评论语料. 语料规模为6000篇,正面为3000篇,负面为3000篇. 另外,还从数据堂网站上下载了 monitor(683篇),digital(1705篇),network(350篇)和 mobile(2317篇)等领域的语料.

在实验中,本文的超参数设置参考了文献[24]中的方法,超参数的设置为$\alpha=1$,$\beta=0.01$,情感主题个数$T=10$. 此外,为了验证新方法的有效性,本文使用了 Self-Training 和 Transductive SVM(T-SVM)方法进行实验,并与 PCTCLR 方法进行了对比.

4.2　实验流程

本文的总体实验流程如下:

(1)对要分析的评论进行预处理操作. 在该过程中,使用了中科院的分词系统 ICTCLAS2013 对评论进行分词.

(2)根据预处理的评论,训练 Sentiment-Topic 模型,估计模型参数θ和φ.

(3)通过 Co-training 学习算法对未标注文本进行

分类.

4.3 实验结果与分析

为了研究新方法的有效性,本文在不同的语料下进行了不同情况下的实验.

4.3.1 极性分析效果对比

为验证新方法在不同领域中的有效性,采用了5个领域中的语料作为数据集实验,并与其他两种方法进行了对比. 实验结果如表1所示.

表1 不同方法之间情感极性分析结果对比

		Self-training	T-SVM	PCTCLR
monitor	precision	0.717	0.778	0.793
	recall	0.701	0.763	0.782
	F-measure	0.709	0.770	0.787
digital	precision	0.747	0.801	0.815
	recall	0.728	0.793	0.801
	F-measure	0.738	0.797	0.808
network	precision	0.735	0.799	0.803
	recall	0.715	0.784	0.791
	F-measure	0.725	0.791	0.797
mobile	precision	0.771	0.810	0.827
	recall	0.726	0.785	0.799
	F-measure	0.748	0.797	0.813
hotel	precision	0.734	0.798	0.818
	recall	0.721	0.779	0.802
	F-measure	0.727	0.788	0.810

从表1可以看出,在三种方法中,T-SVM方法在准确率上比Self-Training方法平均提高了4% ~ 6%,而PCTCLR方法比T-SVM方法平均提高了将近3%;在召回率上,T-SVM方法比Self-Training方法也提高了将近5%,而PCTCLR方法又提高了将近3%. 显然,Self-Training方法分析的效果是最差的,这是由于Self-Training方法在进行分类时,使用了朴素贝叶斯分类器,而T-SVM方法主要是使用了SVM方法对其进行分类,并且在一般情况下,SVM方法比朴素贝叶斯分类方法都具有较好的性能,因此,T-SVM方法的性能比Self-Training方法的性能较好. 与上述两种方法相比,PCTCLR方法考虑了文本中的复杂情感的表达,因此,PCTCLR方法的性能优于Self-Training和T-SVM的性能.

4.3.2 参数的影响

在PCTCLR算法中,设置了三个参数k、m和n,三个参数的选择也会影响最终的极性分类效果. 为了验证参数对分类的效果,本文选择了digital、mobile和hotel三个领域中的语料,并采用了不同的参数k、m和n,进行了10折交叉验证,具体实验结果如表2所示.

表2 不同参数的实验结果对比

选择的参数	precision	recall	F-measure
$k=2,m=2,n=4$	0.783	0.768	0.775
$k=2,m=4,n=2$	0.797	0.788	0.792
$k=2,m=4,n=4$	0.804	0.786	0.795
$k=4,m=2,n=4$	0.788	0.779	0.783
$k=4,m=4,n=2$	0.8	0.791	0.795
$k=4,m=4,n=4$	0.82	0.801	0.81
$k=10,m=2,n=4$	0.776	0.766	0.771
$k=10,m=4,n=2$	0.787	0.78	0.783
$k=10,m=4,n=4$	0.796	0.779	0.787

从表2中可以看出,当m和n的确定时,k为4时的极性分类效果是最佳的,其次是k为2,分类结果最差的是k为10. 并且$k=4$时的准确率比$k=10$时提高了2.4%(从0.796提高到0.82). 这是因为标签的数目越多,越不容易区分复杂的情感表达,因此会导致极性分类的效果会很差. 当k的个数较少时,对于那些语言较长,情感较复杂的文本,划分后的某个子文本可能还会存在较复杂的情感,仍会减弱系统分类的性能. 因此对文本进行划分时,选择合适的k,对极性分类是至关重要的.

另外,参数k确定后,$m=4$和$n=4$时的分类效果最佳的,其次是在$m=4$和$n=2$,分析结果最差的是$m=2$和$n=4$. 这是因为有很多文本较倾向于正面,将正面的文本分类到负面的文本集合后,也会影响分类效果. 于是,在每次迭代中,适当选择合适的m和n也会影响系统分类的整体性能.

4.3.3 情感词三个特征的影响

不同的情感标签都会影响极性分类的性能,而标签的选择需要根据情感词的三个特征来确定. 为验证情感标签的三个特征是否影响极性分类,本文采用了digital、mobile和hotel三个领域中的语料,分别进行PCTCL(Polarity Classification of the Tags based on Criticality and Location)、PCTCR(Polarity Classification of the Tags based on Criticality and Relevance)和PCTLR(Polarity Classification of the Tags based on Location and Relevance)三种方法的实验,具体实验结果如图2所示.

无论是mobile数据集,还是digital和hotel数据集,PCTCLR方法比PCTCL方法在各性能方面平均提高了1% ~ 2%. 这是因为在抽取标签时,只考虑了关键度和位置权重两个因素. 显然,情感词之间的相关度具有一定的影响作用;与其类似,PCTCLR方法比PCTCR方法、PCTLR方法在准确率、召回率和F-measure上仍具有较好的分类效果,这是因为PCTCR只考虑了关键度和相关度这两个因素,并没有考虑情感词在文本中的位置关系. 而PCTLR方法只考虑了位置权重和相关度,并没有考虑情感词在文本中的关键度. 由此可知,在抽取标签时,情感词的关键度、相关度和位置关系三个因

素的影响是不容忽视的.

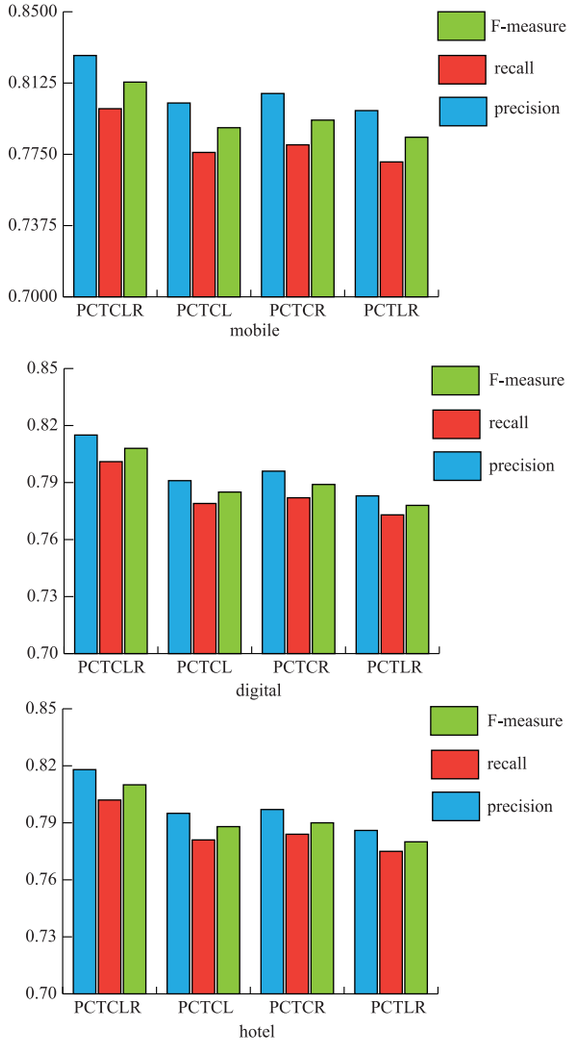

图2　基于不同特征的分类方法的性能对比

5　总结与展望

本文提出了一种基于情感标签的情感极性分类方法,并建立了 Sentiment-Topic 模型,用于抽取文本的情感标签,并通过情感标签对文本进行极性分析. 本文将该方法用于不同领域的文本情感分类中,分类的效果都得到了显著提升.

虽然文本的情感极性分类取得了一定的效果,但还有许多工作要做. 例如,情感词在语境中的影响. 一些情感词具有两面性,在不同的语境中表现出不同的情感极性,这就需要把情感词的极性分析的任务做细做深. 此外,文本情感极性分类的大部分工作都集中在篇章级的褒贬分类,然而更有意义的情感分类是针对评价对象的情感分类,即在情感篇章中挖掘出某评价对象的情感极性.

参考文献

[1] PANG B,LEE L,VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniques[A]. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10[C]. USA: Association for Computational Linguistics,2002. 79 – 86.

[2] TURNEY P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews [A]. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics[C]. USA: Association for Computational Linguistics,2002. 417 – 424.

[3] PANG B,LEE L. Opinion mining and sentiment analysis [J]. Foundations and Trends in Information Retrieval, 2008,2(1/2):1 – 135.

[4] 冯时,付永陈,阳锋,等. 基于依存句法的博文情感倾向分析研究[J]. 计算机研究与发展,2012,49(11):2395 – 2406.
FENG Shi,FU Yongchen,YANG Feng,et al. Blog sentiment orientation analysis based on dependency parsing[J]. Journal of Computer Research and Development,2012,49 (11):2395 – 2406. (in Chinese)

[5] TAN C,LEE L,TANG J,et al. User-level sentiment analysis incorporating social networks[A]. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. USA: ACM,2011. 1397 – 1405.

[6] 赵妍妍,秦兵,刘挺. 文本情感分析[J]. 软件学报,2010, 21(8):1834 – 1848.
ZHAO Yanyan,QIN Bing,LIU Ting. Sentiment analysis [J]. Journal of Software,2010,21(8):1834 – 1848. (in Chinese)

[7] YESSENALINA A,YUE Y,CARDIE C. Multi-level structured models for document-level sentiment classification [A]. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing[C]. USA: Association for Computational Linguistics,2010. 1046 – 1056.

[8] 王李东,魏宝刚,袁杰. 基于概率主题模型的文档聚类 [J]. 电子学报,2012,40(11):2346 – 2350.
WANG Li-dong,WEI Bao-gang,YUAN Jie. Document clustering based on probabilistic topic model[J]. Acta Electronica Sinica,2012,40(11):2346 – 2350. (in Chinese)

[9] CUI H,MITTAL V,DATAR M. Comparative experiments on sentiment classification for online product reviews[A]. Proceedings of AAAI the 21st National Conference on Artificial Intelligence[C]. USA: AAAI,2006. 1265 – 1270.

[10] KIM S M,HOVY E. Automatic identification of pro and con reasons in online reviews[A]. Proceedings of the COLING/

ACL on Main Conference Poster Sessions[C]. USA:Association for Computational Linguistics,2006. 483 – 490.
[11] ZHAO J, LIU K, WANG G. Adding redundant features for CRFs-based sentence sentiment classification[A]. Proceedings of the Conference on Empirical Methods in Natural Language Processing[C]. USA: Association for Computational Linguistics,2008. 117 – 126.
[12] ZAGIBALOV T, CARROLL J. Unsupervised classification of sentiment and objectivity in Chinese text[A]. International Joint Conference on Natural Language Processing[C]. USA: Association for Computational Linguistics,2008. 304 – 311.
[13] DASGUPTA S, NG V. Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification[A]. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2[C]. USA: Association for Computational Linguistics,2009. 701 – 709.
[14] GOLDBERG A B, ZHU X. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization[A]. Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing[C]. USA: Association for Computational Linguistics,2006. 45 – 52.
[15] WAN X. Co-training for cross-lingual sentiment classification[A]. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1[C]. USA: Association for Computational Linguistics,2009. 235 – 243.
[16] LI T, ZHANG Y, SINDHWANI V. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge[A]. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1[C]. USA: Association for Computational Linguistics,2009:244 – 252.
[17] 欧阳继红,刘燕辉,李熙铭,周晓堂. 基于 LDA 的多粒度主题情感混合模型[J]. 电子学报,2015,43(9):1875 – 1880.
OUYANG Ji-hong, LIU Yan-hui, LI Xi-ming, ZHOU Xiao-tang. Multi-grain sentiment / topic model based on LDA[J]. Acta Electronica Sinica, 2015, 43(9): 1875 – 1880. (in Chinese)
[18] MELVILLE P, GRYC W, LAWRENCE R D. Sentiment analysis of blogs by combining lexical knowledge with text classification[A]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. USA: ACM, 2009. 1275 – 1284.
[19] LI S, HUANG C R, ZHOU G, et al. Employing personal / impersonal views in supervised and semi-supervised sentiment classification[A]. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics[C]. USA: Association for Computational Linguistics, 2010. 414 – 423.
[20] 卜湛,伍之昂,曹杰,朱贵祥. 在线评论情感计算与博弈预测[J]. 电子学报,2015,43(12):2530 – 2535.
BU Zhan, WU Zhi-wang, CAO Jie, ZHU Gui-xiang. Affective computing and game theory based prediction for online reviews[J]. Acta Electronica Sinica, 2015, 43(12): 2530 – 2535. (in Chinese)
[21] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993 – 1022.
[22] 江雨燕,李平,王清. 基于共享背景主题的 Labeled LDA 模型[J]. 电子学报,2013,41(9):1794 – 1799.
JIANG Yu-yan, LI Ping, WANG Qing. Labeled LDA model based on shared background topic[J]. Acta Electronica Sinica, 2013, 41(9): 1794 – 1799. (in Chinese)
[23] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[A]. Proceedings of the 11th Annual Conference on Computational Learning Theory[C]. USA: ACM, 1998. 92 – 100.
[24] STEYVERS M, GRIFFITHS T. Probabilistic topic models[A]. Latent Semantic Analysis: A Road to Meaning[M]. USA: Lawrence Erlbaum, 2006.

作者简介

周 孟 男,1986 年生于河南濮阳,博士生、CCF 学生会员,主要从事数据挖掘及自然语言处理方面的研究工作.
E-mail:angel19851229@163.com

朱福喜(通讯作者) 男,1957 年生于湖北新洲,教授、博士生导师. 主要从事人工智能、知识挖掘和分布式计算等方面的研究工作.
E-mail:fxzhu@whu.edu.cn