doi:10.3969/j.issn.1672-1292.2016.03.007

Twitter推文与情感词典SentiWordNet匹配算法研究

易顺明[1],周洪斌[1],周国栋[2]

(1.沙洲职业工学院电子信息工程系,江苏 苏州 215600)
(2.苏州大学计算机科学与技术学院,江苏 苏州 215006)

[摘要] 在Twitter情感分类研究中,经常会采用将推文中的单词匹配情感词典中的同义词条查找相应情感值的方法.但推文书写比较随意,包含许多俚语、缩写和特殊符号,导致许多词汇与情感词典中的词条无法匹配,匹配率不高直接影响推文的情感分类性能.针对Twitter的语言特征,提出了一套Twitter推文与情感词典SentiWordNet的匹配算法.该算法首先通过对推文内容进行数据清洗、替代处理、词性标注和词形还原等预处理,增加了命名实体识别、对hashtags内容的断词处理、基于Word Clusters的否定句处理和词组匹配等方法.实验结果表明,采用此方法的匹配率可达90%以上.

[关键词] 推文,情感分类,SentiWordNet,匹配算法
[中图分类号]TP391 [文献标志码]A [文章编号]1672-1292(2016)03-0041-07

A Matching Algorithm Between the Tweets in Twitter and SentiWordNet

Yi Shunming[1],Zhou Hongbin[1],Zhou Guodong[2]

(1.Department of Electronics and Information Engineering,Shazhou Professional Institute of Technology,Suzhou 215600,China)
(2.School of Computer Science and Technology,Soochow University,Suzhou 215006,China)

Abstract: In the research of the Twitter sentiment classification,a method is widely used to obtain sentiment values by mapping tweets' words with the synonym terms in the sentiment lexicon. However,tweets are usually written informally,which contain slangs,abbreviations and special symbols,many words in the tweets cannot be found in the terms of sentiment lexicon. Lower matching rate directly impacts the performance of sentiment classification. Based on the features of Twitter,a set of matching algorithm between tweets and sentiment lexicon SentiWordNet is proposed in the article. In this method,tweets are processed by data cleaning,alternative processing,POS tagging and word lemmatizing,along with some algorithms such as named entity recognition,hashtags word segmentation,negated context recognition with Word Clusters and phrase matching. Experimental results show that the matching rate reaches over 90%.

Key words: tweets,sentiment classification,SentiWordNet,matching algorithm

对社交媒体开展数据分析,是大数据分析和自然语言处理中的一个热门研究方向.作为被公众广泛接受的社交媒体平台,Twitter因其基于文本的特点和内容中包含较多的情感因素,成为社交媒体情感分析研究的主要对象.情感分析可以归类于人工智能中的分类问题,技术关键在特征抽取与分类模型选择.特征方面,研究者一开始将某类词性或某些词性的组合作为特征向量,如Hu和Liu[1]提出了将形容词作为情感特征,Wiebe等[2]将形容词、情态动词、副词的组合作为特征向量,而后Dave等[3]加入了n-gram语言模型.另一方面,也有研究从构建情感词典获取情感值来进行情感分类,Turney[4]最先尝试采用基于词典的无监督学习方法,Yu等[5]提出以句中每个情感词的平均值作为句子情感值,Esuli等[6]用情感词典中的注释信息来作极性判断.模型方面,Pang等[7]最早将机器学习中的朴素贝叶斯、最大熵、支持向量机

收稿日期:2016-07-17.
基金项目:国家自然科学基金(61003155、61273320).
通讯联系人:易顺明,副教授,研究方向:自然语言处理.E-mail:ysm2501@qq.com

等多种分类方法引入情感分类任务中,Mei 等[8]融合主题模型和情感模型来研究情感分类,近年来本文作者[9-11]将半监督学习作为研究情感分类的方向 . 在深度学习的情感分类研究中,Socher 等[12]提出了用递归张量神经网络处理情感分析的语义合成 . 国内有一些研究团队投入了 Twitter 情感分析的研究,其实验系统在 SemEval(国际语义评测大赛)中获得不错成绩,如华东师大的 ECNUCS[13]、哈工大的 Coooolll[14]等 .

在研究 Twitter 情感分析中,本文作者比较了基于词频特征向量与基于情感值特征向量的情感分类方法,实验表明后者的性能更优异[15]. 获取 Twitter 的情感特征向量的方法是到情感词典中去匹配相同单词/词性的词条,查询到推文相关单词的情感值 . 问题在于:一方面 Twitter 上使用语言非正式,口语化现象非常突出,与正式的文本有很大不同;另一方面,在 Twitter 推文与情感词典之间,存在单词词形、词性、词组及命名实体等的诸多不同,这些不同导致词汇匹配率降低,匹配不到的单词就没有相应的情感值,许多有情感色彩的单词或词组就会被误处理为中性词,加上推文中的否定上下文对情感值的极性变化,这些都导致情感分类的结果不很理想 .

针对上述 Twitter 推文与情感词典之间的匹配问题,本文提出了一套匹配算法 . 首先预处理推文数据,对推文中的网址(URLs)、@(at_mentions)、RT(retweet)进行数据清洗,对表情符号(emoticons)、缩写(abbreviations)、俚语(slangs)进行替代处理,对英文单词的多种形态、含有重复字母的单词(elongated word)作词形还原;围绕情感词典 SentiWordNet 的特点,设计出包括命名实体处理、对#(hashtags)内容作断词处理、找到 SentiWordNet 相应的词组以及否定词句(negation)处理等匹配算法 .

1　SentiWordNet的组成

目前比较流行的英文情感词典主要有 GI(General Inquirer)(http://www.wjh.harvard.edu/~inquirer/)、LIWC(Linguistic Inquiry and Word Count)(http://www.liwc.net)、MPQA(http://mpqa.cs.pitt.edu)、Opinion Lexicon[16]和 SentiWordNet(http://sentiwordnet.isti.cnr.it)等 .

本文使用的情感词典 SentiWordNet 由意大利信息科学研究所(ISTI)研制[17]. SentiWordNet 基于 WordNet,功能是给不同词性下的每个同义词条赋予不同的情感值,其 3.0.0 版包含了 117 659 条记录,每条记录由 POS(词性)、ID(词条编号)、PosScore(正向情感值)、NegScore(负向情感值)、SynsetTerms(同义词词条名)和 Gloss(注释)6 项组成 . SentiWordNet 将词性分为 4 类,即名词(n)、形容词(a)、动词(v)和副词(r),根据统计,这 4 类词性在 SentiWordNet 中分别有 82 115、18 156、13 767、3 621 条记录 . PosScore、NegScore 的值均在[0,1]之间,同义词条可以同时具有 PosScore、NegScore 值 . 在 SynsetTerms 中,可以有多个同义词,每个单词由“单词名#*n*”表示,其中 *n* 表示一词多义的个数,同一单词的每个不同词性、词义可对应不同的情感值,在 SentiWordNet 中,最多的单词“break”有多达 65 条记录 . SentiWordNet 共涵盖了 155 084 个英文单词或词组的 206 941 个同义词项,平均每个单词或词组对应有 1.33 条记录 .

同时,组成同义词名的可以是一个单词,也可能是一个词组,在 SentiWordNet 中,仅包含一个单词(一字词)的共有 85 567 个,而包含有两词或两词以上的词组共有 69 517 个,表 1 描述了具体的分布 .

表 1　SentiWordNet 中的单词与词组分布

Table 1　Distribution of words and phrases in senti wordnet

词组类别	名词数	形容词数	动词数	副词数	总数
一字词	55 387	18 093	8 433	3 654	85 567
二字词	51 687	3 039	2 674	447	57 847
三字词	8 574	262	306	285	9 427
四字词	1 530	52	89	80	1 751
五字词	312	3	14	12	341
六字词	82	0	5	2	89
七字词	24	2	1	0	27
八字词	24	0	0	0	24
九字词	11	0	0	0	11
合计	117 631	21 451	11 522	4 480	155 084

2 Twitter语言特征及预处理

Twitter 上每个推文(tweet)不超过 140 个字,可以是几个句子,或一个句子、一个短语 . 以下是两个实际的推文例子:

例 1 Netflix?? RT @Pink_Dagger:I'm watching the ORIGINAL X-Men cartoons I used to watch on Saturday Mornings….#throwback #iamhappyagain

例 1 推文共有 17 个单词,包含 Twitter 一些共有的语言特征,有 RT(retweet)、at_mentions(例如 @Pink_Dagger)、hashtags(例如#throwback、#iamhappyagain). 推文中还出现了部分常见的英文语言特征,如缩写(I'm)、专有名词(电影名 X-Men、日期名 Saturday)、动词时态(am watching). 而关键情感词 happy 隐于 hashtags(#iamhappyagain)中 .

例 2 Are you waiting for Nokia N9 getting released in US… well,that may not happen. http://www.wizard-journal.com/tech-news/nokia-n9-meego-smartphone-release-date.html:(

例 2 共有 15 个单词 . 除了例 1 中的一些特征外,还包含有 URLs(例如 http://www.wizardjoumal.com/…)、表情符号":("、否定词"not"、情感词"well"及一般疑问句式等 .

综合起来,推文中包含有 RT、@、#、URL、表情符号等 Twitter 特有的元素,也大量使用像 lol(laughing out loud)、fmi(for my information)这些约定俗成的俚语和像 u'r(you're)这样的缩写,单词拼写错误频频,也有用重复字母的单词(如:I looove it)随意书写的情形 .

针对推文中的上述特点,首先需要将 Twitter 进行数据预处理 . 通常预处理主要分为以下几步:数据清洗、替代处理、词性标注及词形还原等 .

2.1 数据清洗

数据清洗的工作是要删除推文中的 URLS 地址、@内容和 RT 标签,并对推文小写化 . 从情感分析的角度观察,推文中的 URLS、@内容和 RT 标签不涉及影响推文情感分类的相应内容,可以从推文中清除掉 . 而 hashtags 涵盖的话题信息中,可能包含一些情感词,如例 1 中 hashtags 中有情感词"happy",因此需要将 hashtags 保留下来 .

2.2 替代处理

替代处理是对表情符号、俚语和缩写词用其他单词替代 . 推文中有大量的表情符号,无法从 SentiWordNet 中直接匹配到,可将其按正向或负向情感的表情符号加以区分,如:)或 ☺ 表示正向,:(或 ☹ 表示负向,分别用单词 happy 或 sad 替代;俚语和缩写词甚至错写的单词在 SentiWordNet 中也匹配不到,要用完整的单词或词组作替代处理,如用"I will"替代"I'll",用"Tomorrow"来替代"2moro"、用"oh my God"替代"omg".

2.3 词性标注

在 SentiWordNet 词典中的单词或词组对应不同的词性有不同的情感值,因此在与 SentiWordNet 词典匹配前,需要对推文中的每个单词进行词性标注 . Python 的 NLTK 提供了基于宾大树库(Penn Treebank)词性标注符号的词性标注器 pos_tag,可完成大部分的词性标注 . 在完成词性标注后还需要将对应的词性转换成 SentiWordNet 词典的名形动副 4 种词性符号,如将 NN、NNS、NNP、NNPS 转为 n(名词),将 JJ、JJR、JJS 转为 a(形容词),而将不在名形动副之列的,标注为 o(other),作为其他词性 .

用 pos_tag 进行词性标注,也有一些差错,譬如动名词在句中不同位置出现时,可以是动词进行时态或名词,在 NLTK 中经常混淆 . 一个策略是用情感词典来增加词性标注的鲁棒性:对用 NLTK 的词性标注器标注可能错误的单词,将其放入 SentiWordNet 中进行匹配尝试,如果词典中有这个词条,就默认是正确的,否则选择别的词性加以尝试 . 此方法对于提高匹配率效果很好 .

2.4 词形还原

将英文单词中的多种形态、含有重复字母的单词(elongated word)还原成单词原形 . 动词的过去式、过去分词、进行时/动名词、第三人称单数形式,名词的复数形式,形容词的比较级、最高级形式等等,都是单词的变形 .

为了能让推文中的单词尽可能准确地匹配到 SentiWordNet 词典中的词条,而词典中的同义词条均是单词或词组,因此应采用词性还原而不是提取词干,其结果必须是一个完整的单词,而不是词的一部分,这是首先要注意的,因此 NLTK 提供的一些 stem 工具(包括 SnowballStemmer、PorterStemmer、LancasterStemmer、RegexpStemmer 等)都不能使用.这里使用的是 NLTK 中基于 WordNet 的词性还原工具 WordNetLemmatizer. 例如,单词"files"经过 WordNetLemmatizer 后还原成"file".

需要注意的是,WordNetLemmatizer 默认的是名词类型,当其他词性的单词使用这个工具还原时,必须指定这个单词的词性.

此外,并非所有的单词都需要还原单词原形.以"interest"为例,在 SentiWordNet 中既有"interest"词性为动词,又有其动词过去式与动名词形态"interested"和"interesting"词性均为形容词,若一味将后者还原成"interest",则在 SentiWordNet 中就会误匹配为动词词性的词条.例如,"an interesting show"中"interesting"是形容词,保留这个形态即可;"be interesting"是现在进行时态,这个"interesting"应该作词形还原,还原为动词 "interest". 由此可见,是否需要词形还原,需视单词的词性而定:若词性为 JJ(形容词),则不还原;词性为 VBG(动词、动名词、动词进行时)或 VBD(动词过去式)或 VBN(动词过去分词),则需要词形还原.

同时,在 Twitter 中有许多书写含有重复字母的单词,也可看成是单词的一种变形.重复字母的形式有两类,如"looove"中单个字母的重复和"hahahaha"中多个字母的重复.可以采用正则表达式匹配的方式,例如正则表达式为"(\w*)(\w)\2(\w*)",可以匹配到字符串中有一个重复字母的情形,"(\w*)(\w{2})\2(\w*)"匹配有两个字母被重复(如"haha")的情形,依次类推,迭代去掉重复,直至最后的单词包含在 WordNet 中.用这种算法,不仅可以将 looove 还原为 love,hahahaha 还原为 ha,还可保留 goose、retweet 这样的单词.但像"lololol"这样的字符串,因为最终"lol"不在 WordNet 中,就只能保留原样.

3　匹配算法

匹配算法包括命名实体处理、断词处理、NEG 标注、词组匹配等,其基本流程如图 1 所示.

3.1　命名实体处理

在 Twitter 中,人们习惯用多个连续的首字母大写的单词来表示一个命名实体.

例 3　Me and @hayleyrose booked in to see A Good Day To Die Hard on February 14 2013(Valentine's Day).

此例中,"A Good Day To Die Hard"表示电影名(中文名"虎胆龙威"),"Valentine's Day"是情人节,均属于一个整体(命名实体),而非各自独立的单词.从情感分析来看,"A Good Day To Die Hard"是个客观词,而不是因为有"good"存在就是个正向情感词.

NLTK 提供了一种基于 CONLL-2000 Chunk 语料库的 ne_chunk()方法来识别命名实体,它可以在上例的推文中找出 Valentine 这个命名实体,但不能将"A Good Day To Die Hard"标注为一个命名实体.

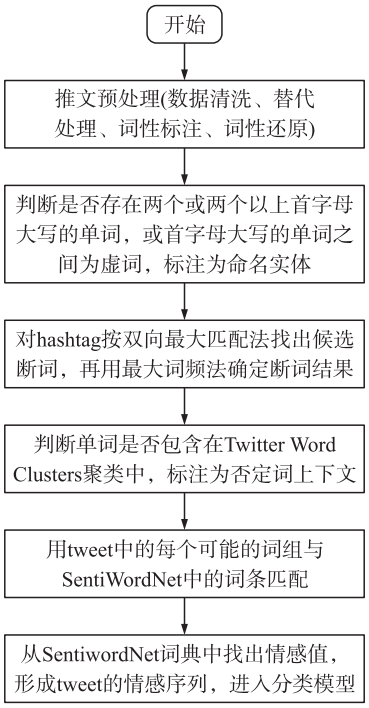

图1　匹配算法的基本流程

Fig.1　Basic flowchart of matching algorithm

由此需要标注出更宽泛的命名实体,在本文的系统中采用了一种基于英文命名实体书写规则的识别方法.根据英文命名实体书写规则,一般实词(名词、动词、代词、形容词、副词)首字母大写,虚词(介词、冠词、连词)小写,这些小写的单词包括冠词(a、an、the)、介词(in、on、to、of、by、as、for)和连词(and、but、with、not)等常用词,超过 5 个字母的虚词必须大写.因此,判定命名实体的策略是,首先判断推文中是否存在(1)两个或两个以上的连续单词,其首字母均为大写(不包含句首的单词);或者(2)在两个首字母大写的单词之间出现的均是字母长度小于 5 的英文虚词.若满足上述条件,则认定为命名实体,用"_"将其连接在一起.

一般地,这样处理得到的命名实体,在 SentiWordNet 中不存在,因此后续一律作为客观词 .

3.2 断词处理

hashtags 中的内容,如例 1 中的#iamhappyagain,英文单词之间没有空格划分,需要对其作断词处理 . 在断词处理算法中,使用了“双向最大匹配法”,通过与单词词频词典进行比对,并用“最大频率法”来评价两者匹配结果,最终获得分词 . 上例中,按照双向最大匹配法,去除“#”后分别得到的断词结果是“i am happy again”、“ia mh app ya gain”,通过找到 max_frequency 的相应句子,最后选定前者为断词结果 . 这种断词方法准确率比较高 .

3.3 NEG 标注

NEG 标注,即对推文中的否定句进行标注 . 根据 Pang 定义[7],一个否定上下文(negated context)为从否定词开始到标点符号结束的一段文字 .

在处理否定句式时,由于推文中的缩写、误写等导致否定词形态繁多,本文采用 CMU 提供的 Twitter 词聚类(Twitter Word Clusters)[18]资源来枚举否定词 . 这个资源通过无监督 Brown 聚类技术[19]从 56 345 753 个推文中提取出 1 000 个单词聚类 . 表 2 列出了 Twitter Word Clusters 中与否定词相关的 5 个聚类及单词样例 .

表 2 Twitter Word Clusters 中与否定词相关的聚类

Table 2 Word Clusters Associated with negative words in twitter

聚类编号	个数	单词样例
00111100	53	don't dont dnt don't dun idont don't …
00111110	76	can't cant couldn't cannot couldnt …
001111010	28	doesn't doesnt doesn't dosent dosnt …
0011110110	29	didn't didnt didn't dint ddnt didn …
0011110111	58	won't wouldn't wont shouldn't …

算法实现中,判断推文中的每个单词是否包含在上述聚类中,或者本身就是诸如 no,not,never,seldom,hardly,few,little,barely,scarcely,nothing,none,rarely,no one,nobody,neither,nor,without 等否定词,或以“n't”结尾,若是,则将从这个单词的后一个单词到标点符号前的一个单词,标注为否定上下文,其中的每个单词后面标“_NEG”,留待作情感极性与情感值计算处理 .

3.4 词组匹配

词组匹配,就是从推文中找到与 SentiWordNet 相应的词组,用“_”连接单词 . SentiWordNet 中除了大量的单个字词外,还有大量的二字词和多字词,若在推文中有相应的这种词,则将这个词组作为一个整体,而不是分割成单个词去匹配 .

例 4 Today for the first time I noticed how gorgeous it is when the sun comes up over the Washington Monument.

在这个推文中有两个词组:“for the first time”和“Washington Monument”,出现在 SentiWordNet 中,其中 “for the first time”为副词,PosScore=0、NegScore=0;“Washington Monument”为名词,PosScore=0.125、NegScore=0.125.

词组的匹配算法:用 tweet 中的每个可能的词组与 SentiWordNet 中的词条匹配,设匹配词组的最大词数为 m 个,以一个 tweet 的平均词数为 t 个,则每个 tweet 的计算复杂度为:

$$(m-1)\times(t-m+1)+\frac{(m-1)\times(m-2)}{2}. \tag{1}$$

依据式(1),在例 4 中,设 m=4,tweet 的词数 t 为 20 个,则每个 tweet 仅需要 57 次匹配 .

4 实验

4.1 实验方法

实验采用 SemEval2013 的数据集[20],其中用 Twitter API 收集到 TrainingB_output 推文共 5 357 条 .

采用两种方法进行比较:一般方法,指只对每一个推文进行常规的数据清洗和词性标注;另一种方法即采用本文的匹配算法,除数据清洗、词性标注、词形还原外,还进行替代处理、否定词处理、命名实体处理、断词处理、词组匹配等 .

4.2 实验结果分析

表 3 所示为以上两种方案经词性标注获得的词汇个数的统计表 .

从表3可以发现,采用本文的匹配算法后,推文数据集中总的单词数缩减了1576(116 578-115 002)个,原因是命名实体处理、词组匹配处理将单一的单词进行了合并,而替代处理、断词处理等又增加了部分单词,合并的数量多于增加的数量,总体上有所减少;同时,名形动副这些词性的单词数量反而增加了1 423(65 955-64 532)个,主要是原本被标注为其他词性的部分单词或词组被成功匹配,找到了相应的情感值.

表3　两种方案下的词汇数量统计

Table 3　Number of word in two schemes

处理方法	单词总数	名词数	形容词数	动词数	副词数	名形动副总数
一般方法	116 578	37 857	6 663	15 754	4 258	64 532
匹配算法	115 002	37 706	6 633	16 989	4 627	65 955

以表3中采用本文匹配算法获得的推文中的名形动副4种词性的单词或词组总数65 955个为基准,表4给出了两种不同方法与SentiWordNet词典中的词条相匹配结果.

从表4可以看出,采用本文的一系列匹配算法后,推文中单词与词典中词条的匹配率提高了20.8%;在匹配到的单词数中,有5%左右是词典中的词组词条.匹配率得以提高的原因,主要是通过对hashtags的内容断词处理、词形还原、词组匹配以及对NLTK的词性标注作柔性处理后,推文中更多的单词、词组与情感词典中的词条相一致.

此外,像命名实体的处理,将一些专用名词合并在一起,虽然在词典中也无法匹配,但却屏蔽了这些命名实体中有些单词可能的情感因素,客观上也提升了情感分类的准确度.

表4　匹配结果

Table 4　Results about match rate

处理方法	需匹配单词/词性个数	SentiWordNet匹配到单词/词性		其中:匹配到词组/词性	
		单词个数	匹配率	词组个数	占比
一般方法	65 955	45 980	69.7%	0	0%
匹配算法	65 955	59 702	90.5%	2961	5.0%

表5是与渥太华大学的uOttawa系统[21]的结果进行的比较,这些准确率数据是针对SemEval2013任务B,基于词袋(BOW)与SentiWordNet(SWM)词典匹配方法,分别使用SVM(支持向量机)和MNB(多项式朴素贝叶斯)模型,对训练集数据进行10折交叉验证后得到的,而本文系统增加了匹配算法(matching algorithm,MA)的环节.

表5　情感分析的准确率结果

Table 5　Accuracy results of sentiment analysis

系统	方法	SVM	MNB
uOttawa	BOW(tf-idf)	58.75%	59.56%
	BOW+SWM	69.43%	63.30%
our system	BOW(tf-idf)	58.12%	58.74%
	BOW+SWM	68.80%	62.96%
	BOW+MA+SWM	72.24%	65.13%

从表5的结果可以看出,在本文系统中,采用tf-idf的词袋方法和采用词袋+SentiWordNet的方法的结果略低于uOttawa的相关结果,而增加了Twitter与SentiWordNet匹配算法后,结果高于后者.uOttawa系统采用了扩展特征后表现较好.同时,在本文系统中,增加这一匹配算法后,在SVM、MNB模型中,准确率分别提高了3.44%、2.17%,可以看出匹配算法对于基于SentiWordNet方法的情感分析是有效果的.

4.3　问题分析

虽然本文采用的针对情感词典SentiWordNet匹配的推文数据预处理方法明显优于一般的推文预处理方法,但在实验过程中通过分析发现仍然存在以下问题:

(1)新词、俚语、表情符号等的处理:由于Twitter的书写随意性,无论是表情符号还是缩写、俚语、误写,都有许多新词、新的形态被不断创造出来,难于枚举.此外,对新词、俚语、缩写、表情符号等有时也会出现分割错误.

(2)否定句判断:推文中除了有单一的否定词出现,可以定义否定词与标点符号之间为否定上下文之外,实际上还有另外一些否定形态.一种表现在句子结构上,例如,有的没有否定词但有否定含义(如

“too…to…”句式),有的虽有否定词但没有否定含义(如“not only…but also…”句式),有的有多个否定词但不属于双重否定的情形(如“neither… nor…”),还有没有否定词但确是否定的隐式否定句.

5 结语

本文研究了基于情感特征向量的 Twitter 情感分类中推文与情感词典 SentiWordNet 之间的匹配问题.针对这些问题,提出了一套 Twitter 推文与情感词典 SentiWordNet 之间的匹配算法,改进了推文内容预处理策略(数据清洗、替代处理、词性标注和词形还原等),增加了针对 SentiWordNet 词条特点的匹配算法(命名实体识别、hashtags 内容断词处理、基于 Word Clusters 的否定句处理和词组匹配等),极大提高了推文与情感词典的匹配度,在本文的实验中采用此套方法的匹配率可达 90.5%.作为情感分类研究的基础,开展这样的工作还是非常有价值的.

[参考文献](References)

[1] HU M Q, LIU B. Mining and summarizing customer reviews[C]//Proceedings of KDD, USA, 2004: 168-177.

[2] WIEBE J, BRUCE R, O'HARA T. Development and use of a gold standard dataset for subjectivity classifications[C]// Proceedings of ACL, USA, 1999: 246-253.

[3] DAVE K, LAWRENCE S, PENNOCK D. Mining the peanut gallery: opinion extraction and semantic classification of product reviews[C]//Proceedings of WWW, Hungary, 2003: 519-528.

[4] TURNEY P. Thumps up or thumbs down? semantic orientation applied to unsupervised classification of reviews[C]// Proceedings of ACL, USA, 2002: 417-424.

[5] YU H, HATZIVASSILOGLOU V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences[C]//Proceedings of EMNLP, Japan, 2003: 129-136.

[6] ESULI A, SEBASTIANI F. Determining term subjectivity and term orientation for opinion mining[C]//Proceedings of EACL, Italy, 2006: 193-200.

[7] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? sentiment classification using machine learning techniques[C]// Proceedings of the EMNLP, USA, 2002: 79-86.

[8] MEI Q, LING X, WONDRA M, et al. Topic sentiment mixture modeling facets and opinions in weblogs[C]//Proceedings of WWW, Canada, 2007: 171-180.

[9] LI S, HUANG C, ZHOU G, et al. Employing personal/impersonal views in supervised and semi-supervised sentiment classification[C]//Proceedings of ACL, Sweden, 2010: 414-423.

[10] LI S, WANG Z, ZHOU G, et al. Semi-supervised learning for imbalanced sentiment classification[C]//Proceedings of IJCAI, Spain, 2011: 1 826-1 831.

[11] LI S, HUANG L, WANG R, et al. Sentence-level emotion classification with label and context dependence[C]//Proceedings of ACL, China, 2015: 1 045-1 053.

[12] SOCHER R, PERELYGIN A, WU J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]// Proceedings of EMNLP, USA, 2013: 1 631-1 642.

[13] ZHU T, ZHANG F, LAN M. ECNUCS: A surface information based system description of sentiment analysis in Twitter in the SemEval-2013(Task 2)[C]//Proceedings of SemEval2013, USA, 2013: 408-413.

[14] TANG D, WEI F, QIN B, et al. Coooolll: a deep learning system for Twitter sentiment classification[C]//Proceedings of SemEval2014, Ireland, 2014: 208-212.

[15] 易顺明,易昊,周国栋.基于情感特征向量的 Twitter 情感分类方法研究[C]//第 14 届全国计算语言学会议,广州,2015: 79.
YI S M, YI H, ZHOU G D. Twitter sentiment classification with sentimental feature vector[C]//Proceedings of CCL2015, Guangzhou, 2015: 79. (in Chinese)

[16] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A convolutional neural network for modeling sentences[C]// Proceedings of ACL, USA, 2014: 655-665.

[17] BACCIANELLA S, ESULI A, SEBASTIANI F. SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining[C]//Proceedings of LREC, Malta, 2010: 83-90.

(下转第 53 页)

虚拟现实、计算机视觉等诸多领域,基于特征点的图像拼接技术具有广泛的应用前景.

[参考文献](References)

[1] BROWN L G. A survey of image registration techniques[J]. ACM computing surveys,1999,24(4):325-376.

[2] ZITOVA B,FLUSSER J. Image registration methods:a survey[J]. Image and vision computing,2003,21(11):977-1 000.

[3] LOWE D G. Distinctive image features from scale in variant key points[J]. International journal of computer vision,2004,60(2):91-110.

[4] 阮芹,彭刚,李瑞. 基于特征点的图像配准与拼接技术研究[J]. 计算机与数字工程,2011,39(2):141-144.
RUAN Q,PENG G,LI R. Study on image registration and mosaic technology based on surf feature[J]. Computer & digital engineering,2011,39(2):141-144. (in Chinese)

[5] 王君本,卢选民,贺兆. 一种基于快速鲁棒特征的图像匹配算法[J]. 计算机工程与科学,2011,33(2):112-115.
WANG J B,LU X M,HE Z. An Improved algorithm of image registration ration based on fast robust feature[J]. Computer engineering & science,2011,33(2):112-115. (in Chinese)

[6] HERBERT B,ANDREAS E,TINNE T,et al. Speeded-up robust features(SURF)[J]. Computer vision and image understanding,2008,110(3):346-359.

[7] MARTIN A F,ROBERT C B. Random sample consensus:a paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM,1981,24(6):381-395.

[8] BROWN M,LOWE D G. Automatic panoramic image stitching using invariant features[J]. International journal of computer vision,2007,74(1):59-73.

[责任编辑:严海琳]

(上接第47页)

[18] OWOPUTI O,O'CONNOR B,DYER C,et al. Improved part-of-speech tagging for online conversational text with word clusters[C]//Proceedings of NAACL,USA,2013:380-390.

[19] BROWN P,DESOUZA P,MERCER R,et al. Classbased *n*-gram models of natural language[J]. Computational linguistics,1997,18(4):467-479.

[20] NAKOV P,KOZAREVA Z,RITTER A,et al. SemEval-2013 task 2:sentiment analysis in Twitter[C]//Proceedings of SemEval2013,USA,2013:312-320.

[21] POURSEPANJ H,WEISSBOCK J,INKPEN D. uOttawa:System description for SemEval 2013 task 2 sentiment analysis in Twitter[C]//Proceedings of SemEval2013,USA,2013:380-383.

[责任编辑:严海琳]