

话题型微博语言特点及其情感分析策略研究

侯敏 滕永林 李雪燕 陈毓麒 郑双美 侯明午 周红照

中国传媒大学 国家语言资源监测与研究有声媒体中心

摘要 微博已经成为现代人们传递态度、意见和评价的重要形式。本文以语言主观性为理论依据,分析了话题型微博语言句子简短、负面倾向居多、情感表达强烈而理性评价淡化、口语色彩浓重、观点表达的隐晦和非直接、评价对象省略、非规范性等语言特点,并就此提出基于短语情感词典及语义规则的观点句识别及评价对象提取的策略。实验和评测结果证明,这些策略和方法取得了较好的效果。

关键词 话题型微博;语言特点;情感分析;评价对象;

随着网络技术的不断发展,微博已经成为现代人们传递态度、意见和评价的重要形式。传递态度、意见和评价的句子称为观点句。观点句是主观句的一种类型,主观句是语言主观性的一级表达形式,主观性是语言的一种重要属性,也是本质属性之一。语言的主观性是一个复杂的理论问题,从不同角度、不同层面,可以对它有不同认识和解释。本文从话语分析角度出发,主观性指的是“说话人的立场、态度和情感。”(沈家煊,2001)那么,表达说话人立场、态度和感情的句子就是主观句。根据表达侧重点不同,主观句至少可以分为推测句(如“他可能不来了。”)、意愿句(如“我希望你好运。”)、感情句(如“我很高兴。”)、观点句(如“他很优秀。”)。语言信息处理中的“情感分析”或称“倾向性分析”主要分析的是观点句。一般来说,观点句的主要表达手段是使用带有明显评价倾向的词语,主要是形容词,以及一些副词、动词和名词,人们称其为“情感词”。但在不同的语域中,人们表达观点使用的手段也有所不同。微博作为一种新兴媒体,有它独特的文本结构形式。话题型微博指的是围绕某一话题即标签(hashtag)阐发意见、进行讨论的微博形式,因此在观点句的使用、表达观点使用的语言手段以及评价对象的隐现上也有与众不同的特点。本文重点讨论的是话题型微博的语言特点以及对其进行情感分析时可以采取的策略。

以下文章分为四部分:第一部分讨论话题型微博的语言特点,第二部分是话题型微博情感分析所采取的策略;第三部分是实验及其结果分析;最后是结语。

1 话题型微博的语言特点

除了一些名人,微博的主人用的大都是网名,发表意见可以无所顾忌;微博的传播往往是蒲公英式,一个微博主发出一条引人瞩目的微博,就会形成一个主题,体现为标签,其他的微博会围绕这个主题发表意见,下一轮微博还可能会针对上一轮微博发表意见,以此类推,于是就形成了一个带有鲜明主题的话题型微博群。话题型微博群规模的大小取决于人们对该主题的关注程度。与一般的微博^[7]不同,话题型微博这样的文本结构形式就决定了它的语言具有句子简短、负面倾向多、语句口语化程度强、表达情感强烈而理性评价淡化、评价对象在句中不直接出现、语言不够规范等特点。下面分别论述。

1.1 句子简短单句多

微博有字数限制,因而往往短小。话题型微博又因为有一个明确的话题,人们是围绕这个话题发表意见,写微博,只要表达了心情、态度,三言两语也可,只字片语也行,所以与一般文体相比,话题型微博

中句子相对简短，单句居多。表 1 显示了对 400 篇语言舆情评论文和 20 篇话题型微博平均句长的统计结果。

表 1. 评论文与话题型微博平均句长比较表

文体形式	文本数	汉字数	句子数	平均句长（字）
语言舆情评论文	400	356511	9366	38.06
话题型微博	20	68726	3416	20.11

从表 1 可以看出，评论文平均句长是 38 个字，而话题型微博平均句长只有 20 个字，几乎只有评论文的一半长。20 个字，大约有 10 个词，很难形成复句。

1.2 观点负面倾向多

话题型微博的形成是基于一定的社会话题和社会事件，而当今社会引起关注的话题负面性较多，这就使得话题型微博在表达观点时，以否定倾向居多。表 2 是对 400 篇评论文和 20 篇话题型微博中观点句的统计结果。

表 2. 评论文与话题型微博观点句及其极性比较表

文体形式	句子数	观点句数	负面倾向句子数	负面倾向句比例（%）
语言舆情评论文	9366	6190	3083	49.81
话题型微博	3416	2207	1766	80.02

表中数据显示，语言舆情评论文观点句中，负面倾向的只占 49.81%，将近一半，而话题型微博中负面倾向句却占到了观点句的 80.02%，可以看出，话题型微博的负面倾向更加明显。

1.3 表达情感强烈而理性评价淡化

由于传播空间的相对自由和匿名评论的相对隐私，因而网友在表达观点时会采用一些比较极端和激烈的形式，又因微博字数的限制，无法以逻辑鲜明的论述方式进行理性表达，所以在话题型微博中，观点句的表达往往感情色彩强烈，而理性评价淡化，脏话、粗话等表现力强的不雅语汇大量出现，如“傻逼、吃软饭、狗东西、狗日的、狗逼、操蛋、猪狗不如、不得好死、有病、吃屎”等等，这也成为话题型微博观点句在表达情感和态度时一种较广泛的表达方式。

1.4 口语色彩浓重，情感因子颗粒度加大

微博虽然用的是书面形式，但交际的实时性、互动性，使得它具有浓重的口语色彩，人们往往会使用一些口语化词语把自己内心的真实感受直接表达出来，体现句子观点的情感因子颗粒度加大，往往不再是词，而是短语。例如：

- (1) #假和尚搂女子#得瑟，使劲得瑟，被抓了吧
- (2) #90 后暴打老人#我滴妈~！这还真特么争气 一群欠收拾的
- (3) #90 后暴打老人#什么玩意了。
- (4) #官员财产公示#有个鬼用，公示出来的也要有人信呀？

其中的“得瑟、我滴妈、欠收拾、什么玩意、有个鬼用”都口语色彩浓重，明确表达了说话人的观点，

其中除了“得瑟”是词，其余都是短语。

1.5 观点表达的隐晦、非直接性

在话题型微博中，除了用一些很“给力”的词语明确表达观点外，人们还会采用一种隐晦的、非直接的方式，以言外之意表达观点。如：

(1) #食用油涨价#**我可以说脏话吗？**

这是一个疑问句，字面上看没有表达任何观点，然而在“食用油涨价”这一语境下，可以理解为说话者表达的是一种想说脏话来发泄的冲动，因而也就间接地表达了对话题不满的情绪和贬义的态度。

(2) #三亚春节宰客#当地的政府部门这么做的用意是什么？掩耳盗铃？越抹越黑？还是让游客永远不去三亚？**应该去测测智商了！**

“应该去测测智商”则是说话者以此表达对“当地政府部门”的不满。

(3) #假和尚搂女子#**狐狸尾巴总会露出来的**

“狐狸尾巴总会露出来的”是对假和尚做坏事最终会被曝光的一种讽刺。

(4) #韩寒方舟子之争#我真叫您教主了，正话反话都让您说了，**您这逻辑是美术老师教的吧**

“您这逻辑是美术老师教的吧”，这个句子内也没有情感词，但根据标签的语境可以看出，是讽刺方舟子的逻辑性缺失。

1.6 评价对象省略

由于话题型微博的话题标签与文本存在着密切的关联，网友在进行微博发布时，往往可以直接对整个话题或话题的某一部分进行评价和表达态度，因而标签或者标签的一部分甚至标签外事物也就成为了评价对象。所以，话题型微博的评价对象常在文本中省略，观点句更加短小精悍，甚至短小到一个短语、一个词，也可以被人理解，不会造成传播上的歧义。如：

(1) #菲军舰恶意撞击#**抗议！**

(2) #菲军舰恶意撞击#**真可恶**

(3) #菲军舰恶意撞击#**…欺人太甚**

(4) #菲军舰恶意撞击#**可悲啊……**

上面这些观点句都缺省了评价对象，在表达评价时只用了极简短的话语，但读者仍然能够理解它们评价的对象：(1)的评价对象是整个标签“菲军舰恶意撞击”这件事，(2)是标签中的“菲军舰”，(3)是标签中的“菲（律宾）”，(4)是标签之外的“中国”。

1.7 语言不够规范

含有大量非规范性的语言文字，也是话题型微博语言的一个特点。这种非规范性可能是由于几方面原因造成：一方面，是非人为的原因造成的输入失误或常识上的汉字书写错误；另一方面，微博表达自由，加之敏感话题在法律上的限制，会人为地加入一些噪声、非规范词、非规范符号和非规范语言格式。例如：

(1) #疯狂的大葱#找点空闲，找点时间！带着钱包常去 <http://url.cn/1T8Gpq> 淘宝小店去看看！（网站链接等噪声）

(2) #假和尚搂女子#真 **TM** 不要 **Face**。（中英文混用，TM 是“他妈”的拼音缩写，不要 Face 是“不要脸”）

(3) #假和尚搂女子#不过这二和尚佛景不逊达摩释迦，他们真的能做到‘色即是空，空即是色’？？？。。。”(标点符号和中文内容混用的情况)

(4) #彭宇承认撞了南京老太#我很愿意相信这是真的，但是！(“但是”作为一个连词，功能一般是连接句子，往往不会作为句尾结束的标志，而这种微博观点句的表达，则是一种不完整的表达方式)

上述话题型微博的语言特点就决定了我们在对其进行情感分析时所采取的对应策略。

2. 话题型微博情感分析所采取的策略

情感分析，目前主要有两种基本方法。一种是基于语义的词典及规则方法^{[6][8][10]}，一种是基于标注语料的机器学习方法^{[1][5][8][11]}。我们认为，情感分析，分析的是语言的意义，要想在对语言进行深入分析的基础上完成情感分析，规则的方法更有优势。所以我们主要采取的是基于语义的词典加规则的方法，统计方法只在建立词典确立词条属性、建立规则库确定规则优先级时发挥作用。针对话题型微博，我们主要采用了加大情感因子颗粒度，构建基于短语的情感词典、通过短语规则确定句子极性、重点研究否定形式、建立基于话题的OBJ表单等策略。

2.1 构建基于短语的情感词典

识别观点句，学界的通常做法是以情感词作为依据，带有情感词的是观点句，否则就是非观点句(这种以词定句的做法显然有问题，这点我们拟另文讨论)。也有一些学者注意到了短语在情感分析中的作用^{[2][12]}。话题型微博中的观点句由于情感表达强烈而理性评价淡化，各种脏话、粗话、口语词语大量使用，加之很多非直接性的表达，使得影响句子的情感因子的颗粒度加大，因而仅仅依靠传统的情感词典，依靠句子中是否有情感词来进行观点句的判断，会造成很多错判，因此我们构建了一部基于短语的情感词典。

短语情感词典突破了以词为主要收录对象的局限，将明确表达观点的情感短语甚至短句都收录其中，如“不得好死、不咋的、扯犊子、就好了、瞎折腾个屁、不是找抽吗、什么玩意儿、没一个好东西、好不到哪儿去、以为自己是谁啊、可以说脏话吗、这是人的行为吗、良心让狗吃了”等等。当然，一些明确表达观点的口语词、网络用语和脏话也必须收录到情感词典中，如“得瑟、傻逼、狗逼、操蛋、犯贱、傻B、欠揍、找死”等，这样包含这些情感短语和情感词的观点句就能够被准确地识别出来，从而在不降低准确率的前提下，提高了系统对观点句的召回率。

2.2 通过短语规则确定观点句及其极性

话题型微博语言简单，短句多，长句少；单句多，复句少。针对这种特点，我们建立了短语规则库，试图解决情感短语词典不能处理的那部分句子是否观点句以及句子的极性问题，主要是长距离搭配形成的情感短语或者是有歧义的短语搭配。如：

(1) #六六叫板小三#这种事是 3 方的责任;不能单独怪哪一方;

(2) #90 后暴打老人#我们指责的不应该是 90 后，人之初性本善，这是国家、学校、家长教育的责任。

其中“是……的责任”是一个表达负面评价的远距离搭配，中间可以加进各种不同成分，基本都是评价对象。

(3) #苹果封杀 360#这不是腾讯和 360 的问题，这是**中国跟美国的问题**。

(4) #国旗下讨伐教育制度#绝对不是教育制度的问题 **是教师的问题**啊

这两个句子中都有“是……的问题”这样的搭配，但句(3)中的“是……问题”不表示评价，不是情感短语，句(4)中的“是……问题”就表示对“教师”的评价，是情感短语。这样的长距离、有选择搭配，也只能靠短语规则来解决其观点句的判定问题。

2.3 建立否定形式规则库

话题型微博观点句具有负面倾向居多的特点。负面观点句的表达主要靠两种手段，一种是具有负面倾向的情感词或情感短语，一种是运用否定形式表达。利用情感短语以及短语规则识别出来的观点句，其倾向极性都在词典或规则上明确标注着，所以观点句的倾向极性识别不成问题。问题较大的是有否定形式的观点句，一般的否定会改变句子的倾向极性。如：

(1) #韩寒方舟子之争#急急忙忙起诉，**绝非上策**。

(2) #菲军舰恶意撞击#我们的政府**不够强硬**，每次都是强烈抗议，别人鸟你么？

(3) #90 后当教授#现在的教授 真是越来越**不值钱**了！！

(4) #官员财产公示#**没有公平，没有正义**，处处充斥着欺诈。

遇到这种情况，只需在规则的操作部分在情感词的极性值上乘以-1，就可以解决句子的极性判断问题。但否定形式的情况极其复杂，并不都是一个简单的“乘-1”就能解决的问题。如：

(5) #官员财产公示#**太不和谐**了吧

(6) #洗碗工剩菜被开除#有些事情合情并**不一定合理**

(7) #假和尚搂女子# **没有不沾腥**的和尚。。。。

(8) #官员财产公示# **所谓的民主**，就是你是民，我是主。

可以看到，在(5)中，因为程度副词“太”的存在，在改变倾向极性的同时，加强了情感强度；(6)中，因为有“一定”，否定程度弱化了；(7)中，否定词语连用，就变成了对“沾腥”的肯定；(8)中的“所谓”一般不会看作是否定词，但它确实确实改变了“民主”的极性。

针对这些情况，我们在语料中尽可能全面地搜集了可以改变句子极性的词语，形成了一个否定形式词集，并就各种否定类型建立规则库，以保证这类观点句倾向极性的准确判断。

2.4 建立基于话题的 OBJ 表单

话题型微博语言话题集中，评价对象往往省略，它的评价对象或者是标签，或者是标签的一部分，或者在句子的上文或下文，因此在句内根本无法提取到评价对象。针对这种情况，我们在评价对象提取模块中建立了一个基于话题的 OBJ 表单。

例如下面 4 个观点句，标签话题都是“六六叫板小三”，句中都没有评价对象。但仔细分析，这 4 个观点句的评价对象依次为整个标签、六六和小三、六六、小三。这说明，对于话题型微博来说，评价对象缺省的观点句，可以通过回溯话题标签的方式，在标签中提取评价对象。

(1) #六六叫板小三#**无聊**…

(2) #六六叫板小三#**都贱货**！

(3) #六六叫板小三#维护自己的爱情，**做的对，给力**

(4) #六六叫板小三#破坏别人家庭，变态。

OBJ 表单必须建立在话题的基础上。“六六叫板小三”这一话题下的“无聊”是评价“六六叫板小三”这件事的，“官员调研”这一话题下也有“无聊”，则是评价“官员调研”这一现象的，“苹果封杀 360”这一话题下的“无聊”则是评价“苹果封杀 360”这一事件的。应说明的是，建立 OBJ 表单比较适合话题型微博，换一种文体，也许不会有效。

3 实验结果及分析

运用上述策略，我们构建了一个基于情感短语词典以及语义规则的情感分析系统 CUCsas，并以中国计算机学会（CCF）2012 年主办的第一届中文微博情感分析评测提供的评测语料为对象，对 20 篇话题型微博语料进行了观点句识别、观点句极性识别和观点句评价对象提取的实验。表 3 是实验结果（也是评测结果）。

表 3 情感分析系统 CUCsas 对话题型微博语料的分析结果

实验（评测）项目	微平均			宏平均		
	正确率	召回率	F值	正确率	召回率	F值
观点句识别	0.756	0.812	0.783	0.757	0.797	0.773
观点句极性识别	0.844	0.685	0.756	0.842	0.672	0.745
评价对象识别（严格）	0.303	0.275	0.288	0.305	0.265	0.278
评价对象识别（宽松）	0.387	0.356	0.371	0.391	0.344	0.359

上述成绩在本次参加评测的 34 支队伍提交的 53 组结果中，任务 1 观点句识别获得第二名，任务 2 观点句极性识别获得第三名，任务 3 观点句评价对象识别（包括宽松和严格）获得第一名。虽然相对成绩不错，但我们深知，这三项任务的绝对成绩即正确率还不高，任务 1 和任务 2 的最好成绩还不到 80%，任务 3，我们的成绩最好，但即使宽松评测，F 值还不足 40%。看来，情感分析，离实用化的要求还有相当一段距离，任重道远。

我们对系统分析结果中识别错误的句子进行了分析和归纳，产生错误主要有以下几个原因。

1. 在观点句及其极性的识别上，由于规则方法本身的限制，使得我们建立的情感词典和规则库很难覆盖全部语言现象和微博全部的语料，例如“#彭宇承认撞了南京老太#这事儿还有人信啊...”，这条微博没有被系统识别为观点句，因而极性识别和评价对象提取也都没有再继续进行。但是从这条微博内容来看，明显带有说话人对彭宇这件事情的强烈不满和质疑，应该是观点句。再回溯到我们建立的情感词典，其中“谁信呢”作为一条情感短语已收录其中，但是这条微博变换了语言形式。如果能在情感词典中建立一个“还有人信啊”的短语条目，那么这条微博也能在系统中进行正确分析。
2. 对微博语料进行预处理时，由于分词错误使得一些词不能与情感词典中的词条匹配，因而造成了观点句识别上的流失，例如“#皮鞋果冻#以前敢吃，现在不敢吃了，想想都好恶心”，在分词的时候，因为“好恶”被分在一起，“恶心”两个字被切开，因而无法到情感词典进行匹配。

3. 对观点句极性的认识不同, 也造成系统分析结果与评测答案不一致, 如“#菲军舰恶意撞击#转向舵是不会失效的, 更不会这么巧。”这句微博, 系统认为是观点句, 因为这句微博体现了对菲军舰撞击原因的质疑, 但是评测答案认为这句微博是非观点句。
4. 微博语言表达上的隐晦、非直接性, 使得系统在处理深层语用时也会出错, 例如“#皮鞋果冻#JS 有点良心吧”这一观点句, 系统分析结果显示为肯定倾向, 但是从这句话的深层语义分析, 是表示对“JS (奸商)”的一种讽刺, 是劝诫他们应该有点良心, 应该是否定倾向。
5. 由于微博语料的非规范性, 错别字和标点符号的误用现象比较普遍, 这也在很大程度上影响了规则的匹配, 造成一些观点句评价对象的提取错误。例如, “#彭宇承认撞了南京老太#---老百姓太不容易了, 希望我们的党能领导人民好好走下去。”这一观点句, 由于“老百姓”前面有一个标点符号, 因而在进行规则匹配时候, 这个标点符号也被算到评价对象中来, 评价对象就成为了“---老百姓(NEG)”, 在严格评测中就是个错误。

由于时间的限制, 我们所做的工作还不全面, 研究还不够深入, 对于语料还没有充分认识, 因此出现一些错误也是不可避免的。

4 结语

我们针对话题型微博的语言特点, 采用基于短语情感词典及语义规则的方法构建了一个情感分析系统, 在中国计算机学会组织的第一届中文微博语言情感分析评测中取得了较好的成绩。但是限于时间和我们的能力以及微博语言本身的复杂和不规范, 系统分析结果还存在很多不足。汉语的表达方式丰富多样, 针对不同的语域类型, 应采取不同的分析策略, 要真正建立一个面向实用的情感分析系统, 还有很长的路要走。下一步工作的重点是对汉语的情感表达手段进行更深入、更为精密化的研究, 逐步通过规则和统计相结合的方法, 争取建立一个科学、完善, 真正可以在语言研究和社会实践中实际应用的情感分析系统。

参考文献

- [1]樊娜, 蔡皖东, 赵煜. 基于最大熵模型的观点句主观关系提取, 计算机工程, 2010, 36 (2)
- [2]李钝, 曹付元, 曹元大, 万月亮. 基于短语模式的文本情感分类研究, 计算机科学, 2008, 1135(14)
- [3]刘鸿宇, 赵妍妍, 秦兵, 刘挺. 评价对象抽取及其倾向性分析, 中文信息学报, 2010, 24(1):84
- [4]沈家煊. 汉语的主观性和汉语语法教学. 汉语学习, 2009
- [5]王根, 赵军. 基于多重冗余标记CRFs的句子情感分析研究* 2007
- [6]王素格, 杨安娜, 李德玉. 基于汉语情感词表的句子情感倾向分类研究, 计算机工程与应用, 2009, 45 (24), 153
- [7]谢丽星, 周明, 孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取. 中文信息学报, 2012, 26(1):74
- [8]杨江. 语言文字舆情自动监测的基本方法和关键技术研究[D]. 中国传媒大学, 2011
- [9]张博. 基于SVM的中文观点句抽取[D]. 北京邮电大学, 2011
- [10]朱嫣岚, 闵锦, 周雅倩, 黄萱菁, 吴立德. 基于HowNet的词汇语义倾向计算. 中文信息学报2006 Vol20 No.1
- [11]Pang. B, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment. Classification using Machine Learning Techniques. In: Proceedings of EMNLP 2002
- [12]Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, HLT-EMNLP-2005