
Date：2018.4.8
Author：baby_qian

初始版本：

具体流程

0. 构建正负情感词词典、否定词词典、程度词词典，并将这些字典里面的词加入到分词器自定义词表中；

1. 对给定需要进行情感分析的句子进行分词；

2. 对分词结果进行遍历，查找里面包含的情感词，并根据情感词的前位词给与该情感词相应的分数；情感词有以下三种情况的前位词：(1). 前位词不为任何否定词或程度词，则返回该情感词的分数，(2). 前位词为否定词，则对情感词分数取反，(3). 前位词为程度词，则返回情感词分数*程度词分数；

3. 累加句子中所有情感词的分数得到句子的最终情感分数。

一、情感词典组成框架

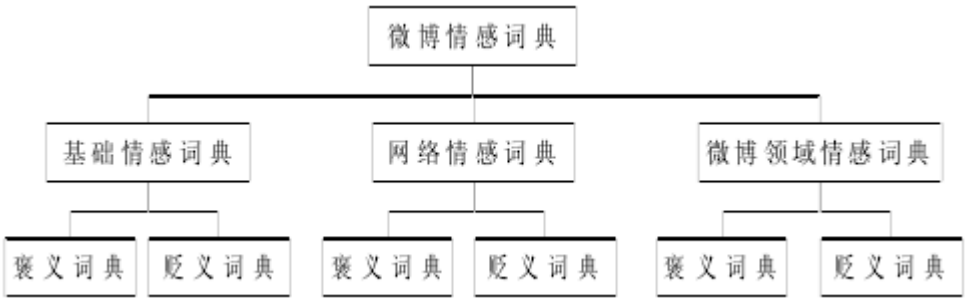

图 1 微博情感词典组成

二、具体分阶段工作

2.1 基于人工构建情感词典

1、将多个较知名的情感词表进行综合，再去重；

潜在问题：不同的情感词表采用的评估情感词方法不一样，会导致后期在给情感词计算权重时，难以选取一个都适用的计算方法。

2、前期实验阶段，为了便于处理，将情感词典中的褒义词权重统一设置为 1，贬义词权值统一设置为-1；

潜在问题：在情感倾向分析研究中，为了区分两者的程度差别，通常是采取给每个情感词赋予不同的权值来体现，后期还需要修改这一步。

3、构建否定词表，否定词权值统一设置为-1；

由于汉语中存在多重否定现象，即当否定词出现奇数次时，表示否定意思；当否定词出现偶数次时，表示肯定意思；

因此，可以引入一个“滑动窗口”的概念，窗口的大小以一个句子为标准，可根据文本中的标点符号作为判断依据，假设情感词所在位置的前面窗口范围内出现奇数个否定词，须将原情感词的情感极性取反（实际计算时乘-1）。

4、构建程度副词表，分为 4 个等级

量级	权重	程度副词示例	个数
极量	2	极其、最、非常、极度、绝对、无可估量、至极无以伦比、卓绝、过度……	99
高量	1.75	不过、不少、出奇、大为、分外、格外、何等颇为、太、特别、着实……	42
中量	1.5	大不了、更、更加、还、还要、较、较为、愈加进一步、足足、越发……	37
低量	0.5	半点、不怎么、轻度、弱、丝毫、稍微、略略不丁点儿、略微……	41

引入一个“滑动窗口”的概念，在实际查找时，以一个句子为滑动窗口的单位。

5、构建停用词表，实际中使用时，对句子先进性分词，然后依照停用词表去除停用词。

6、感叹句的分析，将感叹句权值设置为增强 2 倍关系。

通常感叹句是依附于它所在情感句的情感极性，可以是对正面情感或负面情感的程度加深。

可以做简单处理，将“！”的权值设为 2，在对感叹句具体处理时，首先读入文本预处理后字符串 S 中的特征词 w，判断 w 是否为“！“，若不为”！“，则读取特征词 w 的下一个特征词 w1；若为”！“，则向前查找最近的情感词，若情感词存在，则将情感词的权值乘以”！“的权值；若无情感词，则舍弃，继续后续处理。

7、表情符号的处理

表情符号的加入不但使文本信息充满了个性化色彩，而且还为分析用户情感倾向带来了帮助。基于此，可以构建一个基于表情符号的情感倾向判别表，同时标注情感极性与权值大小，以此为依据辅助判别文本消息的情感倾向。

表 4.4 新浪微博表情符号情感表

名称	个数	权值	内容
正面情感	12	2	[好得意],[哈哈],[太开心],[鼓掌],[ok],[good][耶],[赞],[给力],[威武],[爱你],[haha]
	25	1	[bobo 抛媚眼],[红包],[呵呵],[嘻嘻],[可爱],[亲亲],[抱抱],[钱],[酷],[心],[蜡烛],[蛋糕],[话筒],[礼物],[熊猫],[兔子],[奥特曼],[互粉],[手套],[吃饭],[思考],[顶],[握手],[右抱抱],[左抱抱]
负面情感	19	-2	[怒火],[闭嘴],[鄙视],[泪],[生病],[吐],[怒],[悲伤],[抓狂],[阴险],[怒骂],[伤心],[失望],[挖鼻屎],[愤怒],[最差]
	30	-1	[可怜],[吃惊],[害羞],[偷笑],[懒得理你],[右哼哼],[左哼哼],[嘘],[衰],[委屈],[打哈气],[疑问],[馋嘴],[汗],[困],[花心],[哼],[晕],[猪头],[不要],[弱],[挤眼],[睡觉],[书呆子],[黑线],[拜拜],[感冒],[拳头],[围观],[囧],[神马],[浮云]

8、疑问句的分析

疑问句通常分为两种，一种是有疑而问，也叫询问句，另外一种是无疑而问，通常叫反问句。第一种疑问句跟本文情感分析没有多大关系，主要考虑第二种。反

问句的目的往往是加强语气，把原本的思想表达更加强烈、鲜明。它通常比陈述句表达更为有力，感情色彩也更加明显。

反问句大多带有强烈的感情色彩，而且通常以表达负面情感的居多，多为对某个事件、产品、某个人物、某个机构等等的质疑，其语气程度比较强烈。故在文本消息不包含情感词存在的情况下，可以通过判断疑问句是否为反问句而得到其情感倾向。

权值	反问标记词
-2	为什么、凭什么、难道、何必、怎能、怎么能、怎么会 怎会、哪能、能不、能没、不都、不也、不就、谁叫 谁让、就算、这算、还算、就不、还不、莫非……

9、情感倾向加权计算

经过文本预处理后的微博文本，首先识别不同极性类别的特征项，通过构建好的微博情感词表、否定词词表、程度副词词表以及反问句标记词表做相应处理，获取该条微博中每个特征项的权值，最后作求和运算，获得整条微博消息的情感倾向值，进而判别出情感倾向性。

以标点符号为分割标志，将每条文本分割为 n 个句子 S1、S2、S3…Sn,提取每个句子中的情感词 wi，如果出现程度副词 wa 修饰情感词 wi 或者该句子是包含情感词的感叹句时，则该情感词的情感倾向权值计算公式如下：

$$O_{w_j} = M_{w_a} \times S_{w_j} \tag{4-1}$$

公式(4-1)中，M_{w_a} 表示程度副词，或者感叹号"!"的权值，S_{w_j} 是句子中情感词 w_j 的权值。

当出现否定词 w_h 修饰情感词 w_j 时，为了实现其情感极性取反，则情感词的情感倾向权值公式如式(4-2)所示：

$$O_{w_j} = M_{w_h} \times S_{w_j} \tag{4-2}$$

公式(4-2)中，M_{w_h} 表示否定词的权值，S_{w_j} 是句子中情感词 w_j 的权值。

句子 S_j 中可能包含 k 个情感词，即为 W_1、W_2…、W_k，故该条句子的情感倾向度计算公式如式(4-3)所示：

$$O_{S_j} = \sum_{j=1}^{k} O_{w_j} \tag{4-3}$$

故含有 n 条句子的微博消息 d_j 最终情感倾向计算公式如式(4-4)所示：

$$O_{d_j} = \sum_{j=1}^{n} O_{S_j} \tag{4-4}$$

根据公式(4-4)得到的最终情感倾向值 O_{d_j}，将会出现下列三种情况：

$$O_{d_j}\begin{cases} > 0：\text{为正面情感} \\ = 0：\text{为中性情感} \\ < 0：\text{为负面情感} \end{cases}$$

故根据最终的情感倾向值 O_{d_j} 所处的不同情况，可以识别出该条微博消息中的文本内容所体现出的情感倾向是属于正面、负面、或者中性的。