基于表情和语气的情感词典用于弹幕情感分析

邱全磊,崔宗敏,喻 静
(九江学院 信息科学与技术学院,江西 九江 332005)

摘 要:现有的方法没有考虑颜文字表情对情感分析的影响,同时没有考虑语气词在情感表达中的作用。这影响了弹幕情感分析的效果,降低了情感分析的准确率。为此,构建了一种新的基于表情和语气的情感词典用于弹幕情感分析。首先提出了一种新的应用于弹幕的情感词典构建方法,该方法在以往构建情感词典方法的基础上新增加了弹幕表情词典以及弹幕语气词典的构建方法;然后,提出了一种新的基于弹幕情感词典的情感值计算方法,该方法在对弹幕进行情感值计算时,考虑了表情和语气对情感分析的影响;最后,将提出的方法与现有的方法进行对比实验。实验结果表明,提出的方法在召回率、表情分析、语气词分析等方面比现有的方法在弹幕情感分析领域具有更好的性能。

关键词:弹幕;表情;语气;情感词典;情感分析
中图分类号:TP391.5 **文献标识码**:A **文章编号**:1673-629X(2020)08-0178-05
doi:10.3969/j.issn.1673-629X.2020.08.031

Emotional Dictionary Based on Emoticons and Modal for Barrage Sentiment Analysis

QIU Quan-lei, CUI Zong-min, YU Jing
(School of Information Science and Technology, Jiujiang University, Jiujiang 332005, China)

Abstract: The existing methods do not take into account the influence of emoticons on sentiment analysis and the role of modal particles in emotional expression, which affects the effect of the sentiment analysis of the barrage and reduces the accuracy of sentiment analysis. To this end, we construct a new emotion dictionary based on emoticons and modal for barrage sentiment analysis. Firstly, a new method of constructing the emotion dictionary applied to the barrage is proposed. On the basis of the previous methods of constructing emotional dictionary, this method adds a new method of constructing a dictionary of emoticons and a modal dictionary. Then a new method based on the barrage sentiment dictionary is proposed, which takes into account the influence of emoticons and modal on sentiment analysis when calculating the sentiment value of the barrage. Finally, the proposed method is compared with the existing methods. The experiment shows that the proposed method has better performance than the existing methods in term of recall rate, expression analysis, modal analysis and so on in barrage sentiment analysis.

Key words: barrage; expression; mood; sentiment dictionary; sentiment analysis

0 引 言

近年来,随着网络视频行业的快速发展,网络视频用户规模的不断扩大,弹幕评论越来越受到人们的欢迎。弹幕是一种新兴的,及时更新的互动评论系统,它以滚动字幕的方式直接显示在视频界面上,有助于加深观众对视频内容的理解,也可以促进观众之间的交流。随着弹幕功能在各大视频网站的流行,弹幕中的情感信息越来越具有普遍性和参考性,这些情感信息能准确地反映用户在观看视频的即时情感和褒贬评价。

目前,国内外对于弹幕的研究取得了一定的研究成果,但是主要是从传播角度出发,关注用户心理、传播结构和运营模式等[1-4]。由于弹幕本身的特点,比如文本内容较短,口语化现象突出,网络用语较多,用语不规范等,所以对弹幕进行精准的情感分析仍然存在很大的挑战。

现有的对弹幕进行情感分析的方法中[5-9],没有考虑颜文字表情对情感分析的影响,颜文字表情在文本预处理阶段经常会被过滤掉,同时也忽视了语气词在情感表达中的作用,语气词通常被认为是没有意义可以被省略的停用词,这影响了情感分析的准确率。

为了解决以上问题,构建了一种新的基于表情和

收稿日期:2019-09-11 **修回日期**:2020-01-12
基金项目:国家自然科学基金(61762055)
作者简介:邱全磊(1998-),男,研究方向为自然语言处理;崔宗敏,通讯作者,博士,副教授,硕导,研究方向为大数据计算。

语气的情感词典用于弹幕情感分析,即 EMBA 方法(emotional dictionary based on emoticons and modal for barrage sentiment analysis)。该方法针对弹幕中颜文字表情的大量使用情况,提高了情感分析的准确率,同时,考虑了语气词的作用,增强了弹幕情感分析的效果。实验结果表明,该方法比现有的方法在弹幕情感分析领域具有更好的性能。

1　构建情感词典

1.1　基础情感词典

文中采用 BosonNLP 情感词典作为基础情感词典,与传统的情感词典[10]相比,BosonNLP 情感词典是从微博、新闻、论坛等数据来源的上百万篇情感标注数据当中自动构建的情感极性词典。因为标注包括微博数据,该词典囊括了很多网络用语及非正式简称,对非规范文本也有较高的覆盖率。BosonNLP 情感词典收录了 114 472 个情感词汇,按照情感倾向和情感强度对情感词进行了赋权。其中,褒义情感词的权重为正,贬义情感词的权重为负,情感词的权重范围为[-7,7]。

1.2　弹幕表情词典

自从第一个表情符号“:-)”于 1982 年在 Carnegie Mellon 公告牌上创建以来,这些基于 ASCII 的表情符号已被广泛用于表达人类的情感[11]。颜文字表情能够生动形象地表情达意,在弹幕中深受人们的欢迎。文中使用的颜文字表情来自搜狗输入法颜文字表情词库,包括 21 个类别的 802 个表情符号。目前对于颜文字表情的研究主要以传播学为主[11-13],将颜文字表情应用于情感分析的研究很少,如何确定颜文字表情的权重是一个挑战。文中通过调查统计的形式让九名研究人员根据表情类别确定表情权重,最后取平均值得到表情类别对应的表情权重。最终得到了 21 类表情符号及其对应的情感值,表情词典格式如表 1 所示。

表 1　表情词典

表情	类别	权重
ヾ(≧▽≦*)o	高兴	3
=￣ω￣=	卖萌	2
ψ(`∇′)ψ	吃货	1
(@_@;)	晕	-1
┌(。Д。)┐	害怕	-2
(▼皿▼#)	生气	-3

1.3　弹幕领域词典

由于网络文化与时俱进的发展和弹幕文本的特殊性,弹幕中仍会不断出现新的领域情感词汇,这些词汇都无法在现有的情感词典中找到。因此,文中使用 SO-PMI 算法[14]构建弹幕领域词典对基础情感词典进行扩展。首先确定基准词,然后获取情感词候选词,通过计算确定候选词的情感倾向,最后将候选词汇加入弹幕领域词典中。

SO-PMI 是将 PMI 方法引入计算词语的情感倾向中,从而达到捕获情感词的目的。作为 SO 计算的一部分,Pointwise Mutual Information (PMI)对于根据正面和负面陈述计算短语之间的强度至关重要[15]。它的基本思想是计算同时出现在文本中两个单词的概率,概率越大,相关性越大,连接越接近。PMI 计算公式如公式(1)所示。

$$\mathrm{PMI}(x,y)=\log_2\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

其中,$p(x,y)$ 表示两个词语 word_1 与 word_2 共同出现的概率,$p(x)$ 是 word_1 单独出现的概率,$p(y)$ 是 word_2 单独出现的概率。如果 word_1 和 word_2 之间存在真正的关系,则 $p(\mathrm{word}_1\&\mathrm{word}_2)$ 出现的概率将远大于 $p(\mathrm{word}_1)p(\mathrm{word}_2)$,$\log(\mathrm{word}_1\ \mathrm{word}_2)$ 大于 0。

使用 SO-PMI 计算未记录单词 word_1 的情感值的公式如下:

$$\mathrm{SO-PMI}(\mathrm{word}_1)=\sum_{P_{\mathrm{word}}\in P_{\mathrm{words}}}\mathrm{PMI}(\mathrm{word}_1,P_{\mathrm{word}})-\sum_{N_{\mathrm{word}}\in N_{\mathrm{words}}}\mathrm{PMI}(\mathrm{word}_1,N_{\mathrm{word}}) \tag{2}$$

其中,P_{words} 是一组褒义词,N_{words} 是一组贬义词,这些情感词倾向性非常明显,非常具有代表性。通过 SO-PMI 值与阈值 0 的比较,将未记录词 word_1 分类成积极,中性或消极的情感词,比如 word_1 的 SO-PMI 值大于 0 时,word_1 被识别为积极的情感词。

1.4　语气词典

语气词通常被认为是没有意义的词汇,被列入停用词当中被过滤掉,然而,因为弹幕口语化、极简化的特点,弹幕中存在许多完全由语气词组成的弹幕,如弹幕“哈哈”“嗷嗷”。如果把这些语气词当作停用词过滤掉,将影响弹幕的情感分析效果。

表 2　语气词典

语气词	权重
妈呀	1
唉	-1
靠	-2
哈哈	3

文中利用 1.3 中提到的 SO-PMI 算法构建弹幕语气词典。因为语气词的情感强度低于普通的情感词强度,所以设定语气词情感值范围为[-3,3]。利用 SO-PMI 算法,从弹幕文本中提取出语气词,根据语气词的

SO-PMI 值确定其情感值:当语气词的 SO-PMI 值处于 0 到 5 范围内,语气词情感值为 1;当 SO-PMI 值大于 15 时,语气词情感值恒等于 3,以此类推。将语气词及确定的情感值加入语气词典,最终的语气词典格式如表 2 所示。

1.5 程度词典

文中采用知网提供的程度级别词典,在实际对弹幕文本进行分析时,发现弹幕里存在网络流行词汇以及非正式的词汇当作程度副词使用的情况,如“灰常”表示程度副词“非常”,“敲”表示程度副词“超”,走召”表示程度副词“超”,将这些特殊的词汇整合添加进程度词典中,以提高情感分析的准确度,最终得到了由 228 个程度副词组成的程度词典。程度副词级别及权重如表 3 所示。

表 3 程度词典

程度	权重	例子	个数
极其	3	非常,极,木及	70
超	2.5	最,超,走召	32
很	2	很,好,彳艮	44
较	1.75	较,多,车交	39
稍	0.75	一点,略微	30
欠	0.5	半点,不怎么	13

1.6 否定词典

当否定词修饰情感词时,情感倾向一般都会发生反转,文中整理了弹幕中常用的 71 个否定副词构成否定词典,否定词权重设为-1。

1.7 网络词典

随着互联网的快速发展,产生了很多网络词汇,这些词汇不同于传统的词语,它们更加精简以及口语化,部分网络词汇具有强烈的情感色彩,例如“赛高”,“笔芯”“打 call”。文中从搜狗输入法的词库中整理筛选出最常用的网络情感新词并赋予其情感值,从而完成了网络词典的创建。

2 程度计算

如果一条弹幕说“好看”,另一条弹幕说“非常好看”,还有一条弹幕说“不好看”,若这 3 个弹幕的情感值一样,显然是不合理的,因此,需要对弹幕的情感程度进行量化,用以区分不同程度的“好看”。同理,一个人发出撒花的弹幕,如果撒花后面加了感叹号,显然情感强度应该和没加的时候不同。下面给出相关定义。

2.1 情感词程度计算

定义 1(程度词)。当情感词前面被程度词修饰时,情感词修正权重的计算规则为:

$$W = W_{\deg} * W_k \tag{3}$$

定义 2(否定词)。当情感词前面被否定词修饰时,情感词修正权重的计算规则为:

$$W = (-1)^n * W_k \tag{4}$$

情感词前面同时出现负面词和程度词的情况分为两类,一类是“否定词+程度词+情感词”,这种表达方式对情感强度的影响较弱。另一种是“程度词+否定词+情感词”,这种表达方式对情感强度有增强作用。这两种方式对句子情感权重有一定的影响。例如,“不太好看”和“太不好看”,显然,第一句话的情感强度弱于第二句话。

定义 3(程度词+否定词)。当情感词前面被程度词+否定词修饰时,情感词修正权重的计算规则为:

$$W = (-1)^n * W_{\deg} * W_k * 2 \tag{5}$$

定义 4(否定词+程度词)。当情感词前面被否定词+程度词修饰时,情感词修正权重的计算规则为:

$$W = (-1)^n * W_{\deg} * W_k * 0.5 \tag{6}$$

其中, W 是修正以后的情感词情感值, $W_{\deg}$ 是程度词对应的修正系数, W_k 是情感词情感值, n 为否定词的个数。

2.2 句型程度计算

定义 5(弹幕句型)。不同句型的弹幕对应的情感强度各不相同,定义句型影响系数 X , X 默认为 1。

规则 1:如果弹幕类型为感叹句,即弹幕里出现了“!”或“!”, $X=2$。

规则 2:如果弹幕类型为疑问句,即弹幕里出现了“?”或“?”,且弹幕中没有出现反问标志词(例如“难道”), $X=1$。

规则 3:如果弹幕类型为反问句,即弹幕出现了“?”或“?”,且弹幕中出现了反问标志词(例如“难道”), $X=1.5$。

综上所述:弹幕句型修正计算公式如下:

$$M_i = S_i * X \tag{7}$$

其中, M_i 为经过句型修正之后的第 i 个句子的情感值, S_i 为弹幕中第 i 个句子的初始情感值, X 是句型影响系数。

3 弹幕情感值计算

在第一章构建好情感词典和第二章确定程度计算规则的基础上,下面对弹幕的情感值进行计算。

3.1 句子情感值计算公式

$$S_i = \sum W + \sum E_m \tag{8}$$

其中, W 是修正后的情感词的情感值, E_m 是颜文字表情的情感值, S_i 是弹幕中第 i 个句子的情感值。

3.2 弹幕情感值计算公式

设弹幕的最终情感值为 C ,最终弹幕情感值 C 的

计算公式如下:

$$C = \sum M_i \tag{9}$$

如果 $C>0$,则将这条弹幕判定为积极的弹幕;如果 $C=0$,则将这条弹幕判定为中性的弹幕;如果 $C<0$,则将这条弹幕判定为消极的弹幕。

4　实验分析

4.1　实验数据

文中爬取了哔哩哔哩网站动画,番剧,音乐,舞蹈,科技,生活,鬼畜,娱乐,影视,放映厅等 10 个类别里截止 2018 年 3 月 30 日近期热度最高的前三个视频的弹幕数据,共获得 30 个视频的 63 006 条弹幕。通过对这些弹幕进行预处理,去除完全由标点符号构成的噪音弹幕之后,得到高质量的弹幕文本数据。从每个类别的弹幕里面随机选取 100 条弹幕,共选取 1 000 条弹幕作为测试数据。通过人工标注测试数据的情感极性,将测试数据标注为积极、中性、消极三种类别。最终标注的测试数据类别统计情况如表 4 所示。

表 4　弹幕测试数据统计

积极弹幕数量	消极弹幕数量	中性弹幕数量	测试弹幕总数
719	261	20	1 000

4.2　实验性能评估指标

文中采用在自然语言处理领域被广泛认可和使用的准确率(precision)、召回率(recall)以及 F 值作为实验性能的评估指标,分别定义如下:

$$\text{precision} = \frac{P_c}{P_a} \tag{10}$$

其中,P_c 表示判断正确的该类别弹幕数量,P_a 表示判断为该类别的弹幕数量。

$$\text{recall} = \frac{R_c}{R_a} \tag{11}$$

其中,R_c 表示判断正确的该类别弹幕数量,R_a 表示应该判断为该类别的弹幕数量。

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{12}$$

4.3　实验结果与分析

为了验证文中提出的表情和语气对情感分析的影响以及 EMBA 方法的有效性,通过表 5 中的方法对测试数据进行了测试,六组实验的实验结果如表 5 所示。

通过以上 6 组实验,对实验结果进行如下分析:

(1)现有的方法[6]采用大连理工情感词典作为基础情感词典对弹幕进行情感分析,实验一和实验二将大连理工情感词典与 BonsonNLP 情感词典进行比较。一方面,大连理工情感词典的情感词是情感色彩鲜明的传统情感词,所以准确率更高;另一方面,因为弹幕网络用语较多,用语不规范的特点,大连理工情感词典错误地将大量积极和消极弹幕分类成了中性弹幕,正确识别的弹幕数量很少,这导致了积极和消极弹幕召回率和 F 值低于 BonsonNLP,而中性弹幕的召回率达到了 100%。实验结果表明,基于网络文本构建的 BonsonNLP 情感词典在性能上优于基于传统文本构建而成的大连理工情感词典。

表 5　实验结果

实验方法	类别	precision	recall	F
大连理工词典+语义规则	积极	0.936	0.204	0.336
	消极	0.772	0.272	0.402
	中性	0.027	1.000	0.052
BonsonNLP 词典+语义规则	积极	0.915	0.658	0.765
	消极	0.698	0.663	0.680
	中性	0.026	0.300	0.047
基础词典+表情+语义规则	积极	0.922	0.702	0.797
	消极	0.720	0.739	0.730
	中性	0.033	0.300	0.059
基础词典+表情+语气+语义规则	积极	0.930	0.757	0.834
	消极	0.757	0.762	0.760
	中性	0.039	0.300	0.070
EMBA 情感词典+语义规则	积极	0.992	0.978	0.985
	消极	0.919	0.996	0.956
	中性	0.875	0.400	0.549
ESD 情感词典+语义规则	积极	0.954	0.373	0.536
	消极	0.795	0.356	0.492
	中性	0.032	0.950	0.061

(2)对比实验二和实验三的结果可以发现,在增加了表情词典之后,情感分析的各项指标都得到了显著提升,对数据进行分析发现,在加入了颜文字表情词典之后,对于“⊙▽⊙”“(:3[▩▩]”等表情弹幕可以正确分类,从而提高了情感分析的准确度。实验结果证明了颜文字表情对于弹幕情感分析的影响,也说明了构建颜文字表情词典的必要性。

(3)通过对比实验三和实验四的结果可以发现,在增加了语气词典之后,情感分析的各项指标都得到了一定的提升,这说明语气词也有助于对弹幕的情感分析。对数据进行分析发现,在加入了语气词典之后,对于“冲呀”“嗷嗷”等弹幕,可以通过识别其中的语气词进行正确地分类。实验结果证明了语气词对弹幕情感分析的影响和构建弹幕语气词典的重要性。

(4)现有的对弹幕的情感分析研究较少,且运用情感词典对弹幕进行情感分析的方法较为简单,实际情感分析的效果较差。文中选用在微博文本情感分析领域具有影响力和代表性的 ESD 方法[16]作为对比方

法。ESD 方法的核心是通过拓展情感词典并结合语义规则对微博文本进行情感分析,与文中方法的相同之处在于都选用了现有的情感词典组成基础词典;都构建了程度词典,否定词典,表情词典,网络词典;都分析了语义规则的影响。不同点在于文中构建了能识别颜文字表情的表情词典;利用 SO-PMI 算法构建了弹幕领域词典和弹幕语气词典;利用输入法词库构建网络词典,而不是人工搜集网络词汇。实验五和实验六的结果表明,提出的 EMBA 方法在各类弹幕的性能上都优于 ESD 方法,这证明了 EMBA 方法的有效性和实用性。

5 结束语

对弹幕进行精准情感分析的关键在于情感词典的构建,情感词典囊括的情感词范围越大,准确性越高,情感分析的效果就越准确。文中构建了一种新的基于表情和语气的情感词典用于弹幕情感分析,该词典由基础情感词典、弹幕领域词典、弹幕语气词典、程度词典、否定词典、网络词典组成。该方法针对弹幕评论中颜文字表情的大量使用情况,提高了情感分析的准确率,同时考虑了语气词的作用,增强了弹幕情感分析的效果。同时,还研究了语义规则对于弹幕情感分析的影响,实验结果证明了该方法的有效性。

参考文献:

[1] ZHAO Yongang. The effectiveness of barrage use in audio-visual class of college english[C]//2018 4th international conference on social science and higher education (ICSSHE 2018). Sanya, China: Atlantis Press, 2018: 1-4.

[2] 梁 栋. 弹幕研究述评及展望[J]. 未来与发展, 2019, 43(8): 36-43.

[3] 彭 晴. 新媒体环境下的"弹幕评论"研究[J]. 新媒体研究, 2016, 2(21): 14-15.

[4] ZHAO C, LI Y, HONG R, et al. Supervision of webcasting-anchor behavior evaluation based on barrage emotion analysis[C]//2018 4th international conference on big data computing and communications (BIGCOM). Chicago, USA: IEEE, 2018: 66-71.

[5] 邓 扬, 张晨曦, 李江峰. 基于弹幕情感分析的视频片段推荐模型[J]. 计算机应用, 2017, 37(4): 1065-1070.

[6] 洪 庆, 王思尧, 赵钦佩, 等. 基于弹幕情感分析和聚类算法的视频用户群体分类[J]. 计算机工程与科学, 2018, 40(6): 1125-1139.

[7] 郑飏飚, 徐 健, 肖 卓. 情感分析及可视化方法在网络视频弹幕数据分析中的应用[J]. 现代图书情报技术, 2015(11): 82-90.

[8] 庄须强, 刘方爱. 基于 AT-LSTM 的弹幕评论情感分析[J]. 数字技术与应用, 2018, 36(2): 210-212.

[9] 陈浩然, 庞 华. 旅游综艺节目的策划要素解析——基于弹幕情感分析的实证研究[J]. 东南传播, 2017(12): 112-115.

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

[11] TANG Y, HEW K F. Emoticon, Emoji, and sticker use in computer-mediated communication: a review of theories and research findings[J]. International Journal of Communication, 2019, 13: 2457-2483.

[12] 曹 婧. 网络表情符号的传播与使用[J]. 新媒体研究, 2016, 2(23): 4-5.

[13] ABIDIN C, GN J. Between art and application: special issue on emoji epistemology[J]. First Monday, 2018, 23(9): 1-2.

[14] TURNEY P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]// Proceedings of the 40th annual meeting on association for computational linguistics. Philadelphia, USA: Association for Computational Linguistics, 2002: 417-424.

[15] NAIK M V, VASUMATHI D, KUMAR A S. An enhanced unsupervised learning approach for sentiment analysis using extraction of tri-co-occurrence words phrases[C]//Proceedingsof the second international conference on computational intelligence and informatics. Washington DC, USA: Springer, 2018: 17-26.

[16] ZHANG S, WEI Z, WANG Y, et al. Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary[J]. Future Generation Computer Systems, 2018, 81: 395-403.