

基于种子词的微博表情符情感倾向判定方法

王伟¹ 周咏梅^{1,2} 阳爱民^{1,2} 林江豪³ 陈昱宏¹ 曾文俊¹

(1. 广东外语外贸大学思科信息学院, 广州, 510006; 2. 广东外语外贸大学语言工程与计算实验室, 广州, 510006; 3. 广东外语外贸大学财务处, 广州, 510420)

摘要: 情感倾向明显的表情符, 容易通过人工进行标注。但是对于情感倾向不明显的表情符, 多人手工的标注结果往往难以达成一致。因此, 提出一种利用种子词自动判定表情符情感倾向的方法。该方法利用少量种子表情符自动标注情感倾向比较明显的表情符, 生成表情符标注集; 对于情感倾向不明显的表情符, 利用种子情感词和已得到的表情符标注集构建模型, 实现其情感倾向的自动判定。实验结果表明, 本文方法在微博表情符情感倾向的自动判定上有很好的效果。

关键词: 情感分类; 机器学习; 微博表情符; 种子词; 自动标注

中图分类号: TP391 **文献标志码:** A

Determination Method for Sentiment Orientation of Microblog Smileys Based on Seed Words

Wang Wei¹, Zhou Yongmei^{1,2}, Yang Aimin^{1,2}, Lin Jianghao³, Chen Yuhong¹, Zeng Wenjun¹

(1. Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou, 510006, China; 2. Laboratory for Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou, 510006, China; 3. Financial Department, Guangdong University of Foreign Studies, Guangzhou, 510420, China)

Abstract: The smileys with obvious sentiment orientation are easily annotated manually. But the annotations of the smileys with unobvious sentiment orientation are difficult to reach a consensus. A method of automatically determining the sentiment orientation of the microblog smileys with the seed words is proposed. The method automatically annotates the corpus smileys with obvious sentiment orientation using a few seed emotions. Then these smileys are used to generate the labeled smiley set (LSS). Moreover, a model is built based on the seed emotional words and LSS to determine the smileys with unobvious sentiment orientation. Experimental results show that the presented method is effective.

Key words: sentiment classification; machine learning; microblog smileys; seed words; automatic labeling

基金项目: 国家社会科学基金(12BYY045)资助项目; 教育部新世纪优秀人才支持计划(NCET-12-0939)资助项目; 广东省教育厅科技创新项目(2013KJCX0067)资助项目; 广州市社会科学规划项目(15Q16)资助项目; 广东外语外贸大学研究生科研创新项目(14GWCXM-36)资助项目; 广东外语外贸大学校级项目(14Q3)资助项目; 广东省普通高校青年创新人才类项目(299-X5122106)资助项目。

收稿日期: 2015-06-23; **修订日期:** 2015-08-21

引言

微博等社交网络平台的出现,提供了用户在线表达情绪、对产品发表评论和传播社会事件舆论等的渠道,由此产生了大量包含情感信息的微博文本。针对这些在线文本的情感分析,可以帮助企业了解客户需求,制定营销方案,进而提高竞争力;也可以帮助政府了解民众对公共事件等的情绪与态度。微博作为一种新的交流方式,也推动着网络符号语言,尤其是表情符的发展与广泛运用。表情符表现形式简单,简洁明了,能生动形象地展现个体各种各样的表情,简化了交流互动中语言文字理解和释义的过程,因而备受青年群体的青睐^[1]。微博文本中使用的表情符,往往表达了与文本内容一致的情感倾向^[2]。这为国内外学者在分析网络文本的情感倾向时提供了新的挑战与思路。Yamamoto等^[3]根据表情符的不同作用,将其划分为“强调”“削弱”“转换”和“添加”4种角色,提出一种基于表情符角色对Twitter文本进行情感分类的方法。Khan等^[4]提出一个融合表情符、SentiWordNet和种子情感词的Twitter文本意见挖掘框架,并取得了较高的准确率。Davidov等^[5]选取50个Twitter标签和15个表情符号作为情感标签,提出有监督的情感分类模型,省去了人工标注语料的工作量。文献[6,7]在对Twitter情感进行分类时考虑了微博标签以及表情符的影响。国内学者Jiang等^[8]提出表情符空间模型(Emoticon space model, ESM),实现微博文本的主观性识别、极性判别和情绪分析。庞磊等^[9]利用情绪词和表情图片两种情绪知识对大规模微博非标注语料进行筛选并自动标注,用自动标注好的语料作为训练集构建微博情感文本分类器,对微博文本进行情感极性自动分类。刘培玉等^[10]针对微博文本,利用基础情感词典、表情符词典和网络新词,提取情感词和表情符号作为微博的情感极性信息,有效提高了微博情感倾向性判断的准确度。张珊等^[11]提出一种基于表情图片与情感词的中文微博情感分析方法,利用表情图片和情感词语自动构建微博情感语料库,并构建贝叶斯分类器,实现微博情感倾向的判定。刘伟朋等^[12]利用表情符号自动构建标注语料库,结合机器学习方法训练分类器,对中文微博进行多维情感分类。这些研究均成功利用表情符对文本情感分类模型进行了优化。另外表情符同样启发了构建情感词典的新方法。桂斌等^[2]基于微博表情符号,提出一种自动构建情感词典的方法,实验结果表明,与人工标注结果进行对比,生成的情感词典具有较高的准确率。

上述研究考虑了表情符包含的情感信息,丰富了以往的研究思路。但是,大部分模型中表情符的情感倾向以人工判定为主。本文认为与种子表情符在语料库中发生共现的表情符,具有明显的情感倾向;与种子表情符不共现的表情符,其情感倾向不太明显。对于情感倾向不明显的表情符,多人手工标注的结果往往难以达到一致。因此,本文提出一种利用种子表情符和种子情感词自动判定表情符情感倾向的方法。

1 情感倾向判定方法总体结构

本文提出的微博表情符情感倾向判定方法基本框架如图1所示。该方法主要包括两个部分,分别是自动标注情感倾向明显和情感倾向不明显的表情符。表情符较少与否定词、程度副词搭配,情感倾向

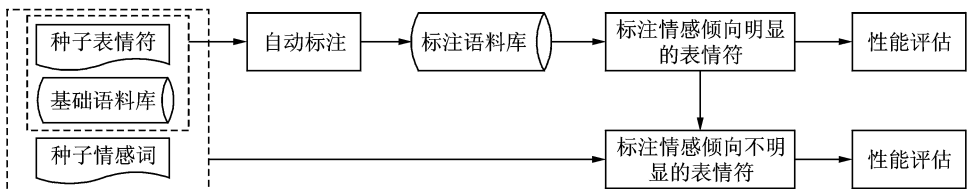


图1 基于种子词的微博表情符情感倾向判定方法基本框架

Fig. 1 Framework of sentiment determination of microblog smileys based on seed words

一般情况下不发生迁移,而且往往与微博内容的情感倾向一致。因此,表情符对微博的情感倾向具有较强的区分能力。



本文分别对情感倾向明显和不明显的表情符进行标注。首先,结合人工的方法筛选出情感倾向较强烈的种子表情符和种子情感词,同时利用种子表情符与语料的共现关系,对语料的情感极性进行自动标注,构建标注语料库。对于标注语料库中除了种子表情符以外的表情符,本文认为这部分表情符与种子表情符存在共现关系,因此具有比较明显的情感倾向。通过计算表情符与标注语料的卡方统计值,得到其情感强度,以此筛选情感强度达到阈值的表情符。然后,根据表情符分别出现在正向和负向标注语料的概率,标注其情感倾向,最后生成表情符标注集(Labeled smileys set, LSS)。对于与种子表情符不存在共现关系的表情符,本文认为其情感倾向不明显。本文提出以种子情感词作为表情符向量的特征项,构造表情符向量空间模型;将已标注极性的 LSS 作为训练集,利用机器学习的方法训练表情符情感分类器,实现表情符情感倾向的自动判定。

2 情感倾向明显的表情符的自动标注方法

2.1 种子词选取

本文的种子词包括两部分:种子表情符和种子情感词。选取的依据主要包括两个因素:情感倾向明显和文档覆盖率高。设计程序利用正则表达式和基础情感词典分别从微博文本中提取出表情符与情感词,统计词频 TF 和文档频率 DF,计算权重 $\omega = TF * DF$ 。将权重值较大的表情符和情感词作为候选种子词。然后结合经验知识,人工筛选得到本文的种子词,共包含 30 个种子表情符和 80 个种子情感词,如表 1 所示。

表 1 种子词表
Tab. 1 Seed words

种子词	正向	负向
表情符		
情感词	快乐、喜欢、哈哈、开心、幸福、期待、支持、感动、祝福、可爱、精彩、不错、成功、恭喜、美好、欢乐、健康、亲爱、温暖、美丽、漂亮、惊喜、享受、创意、兴趣、爱心、优惠、好吃、认真、优秀、浪漫、阳光、热情、平安、最佳、顺利、强大、精美、奋斗、温馨	错过、末日、可惜、可怜、严重、压力、不好、死亡、痛苦、恐怖、感冒、遗憾、担心、讨厌、麻烦、无聊、寂寞、抵制、可怕、病毒、猥琐、紧张、浪费、伤心、变态、后悔、垃圾、恶心、悲剧、无奈、事故、孤独、害怕、烦恼、失败、去世、困难、危险、紧急、对不起

2.2 标注语料库

包含转折逻辑的微博文本中,往往存在不止一种情感倾向。对于此类语料,本文暂不考虑,否则会导致模型过于复杂。因此对语料进行自动标注前,过滤掉包含转折词的微博文本。转折词包括不过、但是、但、而、然而、可是、可、只是、尽管、尽管如此、即使、即使如此、虽然。

本文提出的自动标注语料主要是利用种子表情符将微博文本标注成两类:正向和负向。判定方法基于微博文本的情感倾向由表情符的情感倾向决定的假设,采用一票否决机制,提出以下规则:

规则 1 若语料中种子表情符的情感倾向只包含正负向其中一种,则该语料的情感倾向由该类种子表情符决定。

规则 2 若语料中同时包含正负向种子表情符,则舍弃该语料。

本文对该投票机制进行数学描述,如式(1)所示

$$\text{text}P_i = \begin{cases} \text{pos} & |\mu| = N \& \mu > 0 \\ \text{neg} & |\mu| = N \& \mu < 0 \\ \text{del} & |\mu| \neq N \end{cases} \quad (1)$$

式中: $\text{text}P_i$ 表示语料的标注结果;取值 pos 表示正向, neg 表示负向, del 表示舍弃语料; N 为语料中种子表情符的数量; $\mu = \sum_{i=1}^N \frac{|\text{seed}EP_i|}{\text{seed}EP_i}$, 其中, $\text{seed}EP_i$ 表示种子表情符的情感倾向, 当种子表情符情感倾向为正向时, $\text{seed}EP_i$ 取值为 1; 相反地, 当种子表情符情感倾向为负向时, $\text{seed}EP_i$ 取值为 -1。

2.3 情感倾向明显的表情符的自动标注

情感倾向明显的表情符是指从标注语料库提取并过滤掉种子表情符的表情符。该部分表情符的自动标注包括两部分: 情感强度计算和情感倾向判定。

利用卡方统计值 χ^2 度量基础表情符 e 与文档类别 $\text{text}P_i$ 的关联程度, 得到的 χ^2 统计值作为表情符的情感强度。通过设定阈值 θ , 将情感强度达到阈值的表情符加入到 LSS。计算 χ^2 统计值为

$$\chi^2(e, \text{text}P_i) = \frac{(AD - BC)^2}{(A + B)(C + D)} \quad (2)$$

式中: $\text{text}P_i$ 表示语料标注类别; A 表示包含表情符 e 且属于 $\text{text}P_i$ 类的标注语料数; B 表示包含 e 但是不属于 $\text{text}P_i$ 类的标注语料数; C 表示属于 $\text{text}P_i$ 类但是不包含 e 的标注语料数; D 表示既不属于 $\text{text}P_i$ 类也不包含 e 的标注语料数。

对语料进行分析, 本文认为一般情况下, 正向表情符出现在正向文本的概率大于出现在负向文本的概率, 同理认为负向表情符出现在负向文本的概率大于出现在正向文本的概率。因此, 自动标注集中表情符的情感倾向可以表示为

$$\text{label}EP(e) = \begin{cases} \text{pos} & \frac{A}{A + B} > \frac{C}{C + D} \\ \text{neg} & \text{其他} \end{cases} \quad (3)$$

利用上述方法自动标注情感倾向明显的表情符, 省去了人工标注的负担, 实验部分通过人工进行校对, 验证了本文方法具有较高的准确率。

3 情感倾向不明显的表情符自动标注方法

与种子表情符不存在共现关系的表情符, 本文认为其情感不明显, 难以通过人工方式达成一致的标注结果。因此, 本文提出利用大规模语料库和表情符标注集 LSS, 训练分类器, 实现表情符情感倾向的判断。分类器的构建依赖于合理有效的表情符向量空间模型。在传统文本向量空间模型中, 通过提取组成文本的单元作为文本向量的特征, 例如单词、短语等。但是在表情符向量空间模型中, 一般情况下表情符本身就是一个词语, 难以用同样的方式抽取特征, 需要一种新的特征选择思路。因此, 本文将 2.1 节选得到的种子情感词作为特征, 以此构造表情符向量 e , 即有

$$e = \langle \text{seed}W_1, \text{seed}W_2, \dots, \text{seed}W_m \rangle \quad (4)$$

式中: $\text{seed}W_i$ 表示表情符向量的特征, 即种子情感词, 包括正向种子情感词和负向种子情感词; m 为种子情感词的数量。特征的权重 $W(\text{seed}W_i)$ 依据表情符与种子情感词在语料中的共现关系计算得到, 即有

$$W(\text{seed}W_i) = \frac{df(e, \text{seed}W_i) - \min_{i=1}^M(df(e, \text{seed}W_i))}{\max_{i=1}^M(df(e, \text{seed}W_i)) - \min_{i=1}^M(df(e, \text{seed}W_i))} \quad (5)$$

式(5)对数据进行了归一化处理, $df(e, \text{seed}W_i)$ 表示表情符 e 与种子情感词 $\text{seed}W_i$ 共现的文档数。

图 2 所示是表情符情感极性分类器的训练过程。训练集利用 2.3 节中通过自动标注方式得到的 LSS, 即表情符标注集, 省去了人工标注训练数据的负担。

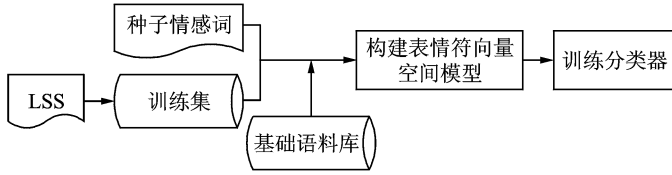


图 2 微博表情符情感倾向判定模型的训练过程

Fig. 2 Training process of sentiment determination model of microblog smileys

4 实验及结果分析

4.1 实验数据

实验数据来源于新浪微博 4 130 个用户的个人微博, 共采集 298 295 条微博, 从中筛选出包含表情符的微博共 93 131 条, 作为本文的基础语料库 BaseCopus。实验数据如表 2 所示。其中, 测试集的表情符是指情感比较不明显的表情符, 通过过滤种子表情符和第 2 节的表情符标注集得到。

表 2 实验数据

Tab. 2 Experimental data

数据集	基础语料库微博	测试表情符	
数量	93 131	正向	负向
		100	100

4.2 实验设置

设计程序时利用正则表达式提取微博文本中的表情符; 分词工具采用中科院 ICTCLAS; 实验中的情感词典通过整合 HowNet 的情感词和评价词、台湾大学 NTUSD 情感词典和大连理工大学信息检索研究院的情感词汇本体库^[13], 并去除重复项得到; 分类器选择支持向量机, 程序代码选取台湾大学林智仁开发的 LibSVM。

4.3 实验结果及分析

设置两组实验对本文方法的有效性进行验证, 一组是设定卡方统计值的不同阈值 θ , 分析情感明显的表情符自动标注的准确率; 另一组是对情感倾向不明显的表情符的情感极性进行分类, 利用 3 个训练集构造分类器, 比较分析实验结果。3 个训练集分别用 TR_1, TR_2, TR_3 表示。 TR_1 表示种子表情符; TR_2 表示表情符标注集, 由本文第 2 节得到; TR_3 表示表情符人工标注集, 即对 TR_2 的表情符进行人工标注。实验所用种子表情符和种子情感词如表 1 所示。本文采用总体准确率作为分类性能的评价指标, 计算公式为

$$\text{Overall-accuracy} = \frac{\sum_{c_i \in C} \text{Correct}(c_i)}{\sum_{c_i \in C} \text{Doc}(c_i)} \quad (6)$$

式中: Over-accuracy 代表总体准确率; $\text{Correct}(c_i)$ 表示分类为 c_i 并且正确的文档数; $\text{Doc}(c_i)$ 表示类别为 c_i 的文档总数。

4.3.1 情感明显的表情符自动标注结果分析

利用 2.2 节的方法标注基础语料库中的 93 131 条语料, 得到 33 241 条正向标注语料, 14 367 条负向标注语料, 以此构建标注语料库。使用该标注语料库计算表情符的情感强度, 以及判定其情感倾向。通过人工校对结果, 分析不同阈值 θ 下算法的性能, 实验结果如图 3 所示。表情符计算得到的卡方统计值呈现明显的分段现象, 因此阈值 θ 取值没有固定的间隔。由图 3 可以得到, 表情符情感倾向的判定取

得了较高的准确率,平均为80%左右,说明了自动标注方式的有效性。另外,随着阈值 θ 的增加,对表情符情感强度达到阈值的要求越高,正向准确率和负向准确率也有一定的提高。

4.3.2 情感不明显的表情符自动标注结果分析

将阈值 θ 取值为580, TR_1 , TR_2 , TR_3 分别作为分类器的训练集,对情感倾向不明显的表情符的情感极性进行分类。请5位相关领域研究人员,对结果进行手工校对,得到的准确率对比如表3所示。可以得出, TR_2 训练得到的分类器取得了较高的总体准确率77.5%,相对于 TR_1 有了明显的提高,说明本文方法的有效性。 TR_2 的负向准确率70%之所以明显低于 TR_1 的负向准确率85%,是因为表情符标注集LSS的正负表情符比例不平衡造成的,统计结果表明正向表情符的数量大约为负向表情符数量的2倍。对比 TR_2 和 TR_3 的总体准确率可以得到,利用表情符标注集作为训练集方法的性能接近于人工标注训练集的方法,大大减少了人工参与的负担,因此,实验表明了本文方法在判定表情符情感倾向上的有效性与优越性。

表3 不同训练集的实验结果 %

训练集	正向准确率	负向准确率	总体准确率
TR_1	45.0	85	65.00
TR_2	85.0	70	77.50
TR_3	82.5	80	81.25

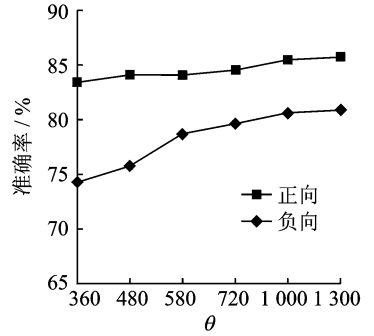


图3 不同 θ 的实验结果

Fig. 3 Experimental results with different θ

5 结束语

本文将表情符分成了两部分进行讨论,一部分是情感倾向明显的表情符,另一部分是情感倾向不明显的表情符。对此,本文提出一种利用种子词自动判定微博表情符情感倾向的方法。种子词包括种子表情符和种子情感词,通过统计与人工结合的方法筛选得到。利用少量种子表情符计算情感倾向比较明显的表情符的情感强度,并根据表情符分别出现在正向和负向标注语料的概率判定其情感倾向,以此生成表情符标注集。对于情感倾向不明显的表情符,利用种子情感词和表情符标注集构建模型,结合机器学习方法训练表情符情感分类器,实现其情感倾向的自动判定。实验结果表明,与人工标注结果对比,两部分表情符情感倾向的判定都取得很好的效果,减少了人工判定的负担。但是本文研究对于情感不明显的表情符情感强度的计算还缺少讨论;另外种子词的选择还需要人工的参与。下一步工作将针对这些问题继续研究,并且将表情符的情感信息应用到文本情感分析中。

参考文献:

- [1] 谭文芳. 网络表情符号的影响力分析[J]. 求索, 2011(10): 202-204.
Tan Wenfang. The influence analysis of network emoticons[J]. Seeker, 2011, (10): 202-204.
- [2] 桂斌, 杨小平, 张中夏, 等. 基于微博表情符号的情感词典构建研究[J]. 北京理工大学学报, 2014, 34(05): 537-541.
Gui Bin, Yang Xiaoping, Zhang Zhongxia, et al. Research on building lexicon for sentiment analysis based on the Chinese microblogging smiley[J]. Journal of Beijing Institute of Technology, 2014, 34(05): 537-541.
- [3] Yamamoto Y, Kumamoto T, Nadamoto A. Role of emoticons for multidimensional sentiment analysis of twitter[C]// Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services. [S. l.]:

ACM, 2014; 107-115.

- [4] Khan F H, Bashir S, Qamar U. TOM: Twitter opinion mining framework using hybrid classification scheme[J]. *Decision Support Systems*, 2014, 57(3): 245-257.
- [5] Davidov D, Tsur O, Rappoport A. Enhanced sentiment learning using twitter hashtags and smileys[C]//*Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. [S. l.]: Association for Computational Linguistics, 2010; 241-249.
- [6] Kouloumpis E, Wilson T, Moore J. Twitter sentiment analysis: The good the bad and the omg[C]//*Proceedings of ICWSM*. [S. l.]: AAAI Press, 2011, 11: 538-541.
- [7] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision[R]. *CS224N Project Report, Stanford Digital Library Technologies Project*, 2009, 1: 12.
- [8] Jiang F, Liu Y, Luan H, et al. Microblog sentiment analysis with emoticon space model[M]. Berlin Heidelberg: Springer, 2014; 76-87.
- [9] 庞磊, 李寿山, 周国栋. 基于情绪知识的中文微博情感分类方法[J]. *计算机工程*, 2012, 38(13):156-158.
Pang Lei, Li Shousha, Zhou Guodong. Sentiment classification method of Chinese mirco-blog based on emotional knowledge [J]. *Computer Engineering*, 2012, 38(13):156-158.
- [10] 刘培玉, 张艳辉, 朱振方, 等. 融合表情符号的微博文本倾向性分析[J]. *山东大学学报:理学版*, 2014, 49(11):8-13.
Liu Peiyu, Zhang Yanhui, Zhu Zhenfang, et al. Micro-blog orientation analysis based on emotion symbol[J]. *Journal of Shandong University: Natural Science*, 2014, 49(11):8-13.
- [11] 张珊, 于留宝, 胡长军. 基于表情图片与情感词的中文微博情感分析[J]. *计算机科学*, 2012, 39(11):146-148.
Zhang Shan, Yu Liubao, Hu Changjun, et al. Sentiment analysis of Chinese micro-blogs based on emoticons and emotinal words[J]. *Computer Science*, 2012, 39(11):146-148.
- [12] 刘伟朋, 陈雁翔, 孙晓. 基于表情符号的中文微博多维情感分类的研究[J]. *合肥工业大学学报:自然科学版*, 2014, 37(7): 803-807.
Liu Weipeng, Chen Yanxiang, Sun Xiao, et al. Multidimensional sentiment classification method of Chinese micro-blog based on the emoticon[J]. *Journal of Hefei University of Technology: Natural Science Edition*, 2014, 37(7):803-807.
- [13] 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.
Xu Linhong, Lin Hongfei, Pan Yu, et al. Construction the affective lexicon ontology[J]. *Journal of the China Society for Scientific and Technical Information*, 2008, 27(2): 180-185.

作者简介:



王伟(1991-),男,硕士研究生,研究方向:文本情感分析、机器学习和自然语言处理,E-mail: 20131010007@gdufs.edu.cn.



周咏梅(1971-),女,教授,研究方向:自然语言处理、文本情感分析和机器学习。



阳爱民(1970-),男,教授,研究方向:自然语言处理、文本情感分析和机器学习。



林江豪(1985-),男,硕士研究生,研究方向:自然语言处理、文本情感分析和机器学习。



陈昱宏(1993-),男,本科,研究方向:文本情感分析。



曾文俊(1993-),男,本科,研究方向:文本情感分析。