计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(12):263−267 [doi: 10.15888/j.cnki.csa.007719]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于口罩评论数据的用户情感趋势与关注分析①

曾志伟, 刁明光, 王欣鹏, 何炳辉

(中国地质大学 (北京) 信息工程学院, 北京 100083)
通讯作者: 刁明光, E-mail: dmg@cugb.edu.cn

摘　要: 为了对疫情期间口罩的用户评论数据进行情感关注分析, 本文用谷歌浏览器的插件 Web Scraper 爬取了 2020 年 3 月 1 日到 4 月 11 日中淘宝网的口罩的共计 143 330 条用户购买评论数据. 为了提高情感预测的精度, 在此数据集上经过人工标注情感为积极和消极的共计 14 400 条数据后, 用 SnowNLP 情感分析模型进行了训练, 最后用训练后的语料库进行了情感预测. 从整体上可见用户评论的情感是积极的. 在用户评论的每日情感变化趋势上, 本土新增病例 (不含海外输入) 的趋势在一定程度上影响着用户每日情感趋势的整体变化, 而国内新增病例 (含海外输入) 的局部波动变化趋势也影响着每日情感局部的相应波动变化趋势. 在对预测后的评论进行分类后, 发现用户的积极评论中对口罩的关注主要集中在口罩的质量、包装、价格、厚实, 而在消极的评论中对口罩的关注主要集中在质量、包装、味道和是否为医用.

关键词: 口罩评论; 用户关注点; 情感分析; 趋势分析; SnowNLP; Python

引用格式: 曾志伟,刁明光,王欣鹏,何炳辉.基于口罩评论数据的用户情感趋势与关注分析.计算机系统应用,2020,29(12):263–267. http://www.c-s-a.org.cn/1003-3254/7719.html

Analysis of User Sentiment Trend and Concern Based on Mask Review Data

ZENG Zhi-Wei, DIAO Ming-Guang, WANG Xin-Peng, HE Bing-Hui

(School of Information Engineering, China University of Geosciences (Beijing), Beijing 100083, China)

Abstract: In order to analyze the sentimental focus of the comment data from users of masks during the outbreak of virus, we extracted 143 330 comments about the purchase from Taobao users from March 1st to April 11th, 2020 by means of the Web Scraper of Google browser. To improve the accuracy of the sentimental estimation, each comment of the total 14 400 pieces was manually marked as positive or negative emotion on this data set. And then we used SnowNLP, the sentimental analysis model to train them. At last, the trained corpus was used for sentimental estimation. The overall sentiment of the comments was proved positive. On the basis of the daily emotional variation trend of users' comments, the trend of local new cases (excluding overseas input) to some extent affects the overall change of their daily emotional trend. And the local fluctuation trend of domestic new cases (including overseas input) also affects that of the everyday emotional performance. After classifying the predicted comments, we found that users' positive comments focused on the quality, packaging, price, and thickness of masks, while negative comments focused on the quality, packaging, smell, and whether the masks were for medical use.

Key words: mask comment; user focus; sentiment analysis; trend analysis; SnowNLP; Python

在疫情期间, 人们急需获得口罩等防护物资. 而对于用户关于口罩的评论的分析研究不仅能从侧面反映出当前疫情对人们的情绪影响程度, 而且也能反映出疫情一定的发展趋势; 对于用户评论的情感关注的研

① 基金项目: 2019 大学生创新创业训练计划项目 A(X201911415126)
Foundation item: Year 2019, College Students' Innovation and Entrepreneurship Training Program A (X201911415126)
收稿时间: 2020-05-11; 修改时间: 2020-06-10, 2020-06-15; 采用时间: 2020-06-19; csa 在线出版时间: 2020-11-30

究, 还对商家提高口罩销量和评分, 有着积极作用. 因此对于口罩评论数据的分析研究, 具有一定的理论与现实意义, 并且具有很高的应用价值.

如今, 国内外的学者对于用户评论数据的研究, 大多只停留在单一的情感分析上, 而忽略了评论数据在时间纬度上所蕴含的情感趋势、对于特定事件从侧面反映出的发展趋势, 以及在关注点上对于商家发展的影响.

近年来, 对于在线商品评论数据的情感分析研究[1]在不断发展. 有研究人员[2]运用扩展的情感倾向点互信息算法, 构建了一个面向中文微博的情感词典, 从而实现了相应的情感倾向分类系统; 针对情感分析和观点挖掘而提出的词典模型[3], 包括了与观点挖掘和情感分析相关语义范畴的分类, 为态度持有者和态度的极性以及文本中不同参与者的情绪和情感的识别提供了方法; 将在线评论文本分解为评论对象–对象属性–评论描述三层体系, 并结合评论模式和评论语境提出的基于属性特征的评论情感量化分析算法[4], 提高了文本情感分类的准确性; 通过获得特定领域具有感情倾向的特征词语[5], 而后利用基准词与特征词语进行的情感分类有着较好的效果; 在文本特征中探索在线评论的有用性因素, 而建立相应的有用性影响因素模型[6], 在分类预测上, 对在线评论的有用性有较强的判别能力; 针对特定事件结合时间信息和地理位置信息而建立的舆情时空演化分析方法[7]能可视化地展示舆情的时空演化过程; 基于语义理解[8]的文本情感分类方法, 能有效地判定文本情感倾向性.

本文对于口罩评论数据的情感分析, 采用了针对口罩评论而训练的特定语料库, 并且将 Jieba[9,10]分词与 SnowNLP[11]情感分析模型结合, 对用户的情感发展趋势和关注点进行了分析. 期望得出针对口罩评论的较高情感分类准确率, 以及从中挖掘出用户对口罩评论的整体态度、影响用户对口罩不同情感关注的相关属性和疫情发展趋势对用户评论的每日情感趋势的影响.

1 研究方法

本文采用的是 Python 的类库 SnowNLP 情感分析模型对口罩的用户评论数据进行的情感分析.

SnowNLP 情感分析中运用的情感分类方法为朴素贝叶斯定理. 它是在贝叶斯定理上作出“认为每个属性各个特征是相互独立的”这一假设而得出的.

朴素贝叶斯定理在情感分类中的公式如下:

$$P(C_i|X_1,X_2,X_3,\cdots,X_n)=\frac{P(C_i)P(X_1,X_2,X_3,\cdots,X_n|C_i)}{P(X_1,X_2,X_3,\cdots,X_n)},\ i=0,1$$

在假设下可简化为:

$$P(C_i|X_1,X_2,X_3,\cdots,X_n)=P(C_i)\Pi_{j=1}^{n}P\left(X_j|C_i\right),\ i=0,1$$

其中, 随机事件C_i表示样本为C类的情感正负概率, X_n表示测试样本中某一特征词 X 出现的概率. 在计算每个语句情感正负时, 用计算出的先验概率$P(C_i)$分别乘以它的每个属性特征词的条件概率而得出的情感概率值, 取其中正负情感值较大的作为此语句的情感.

SnowNLP 的情感分析大致判断过程如图 1 所示.

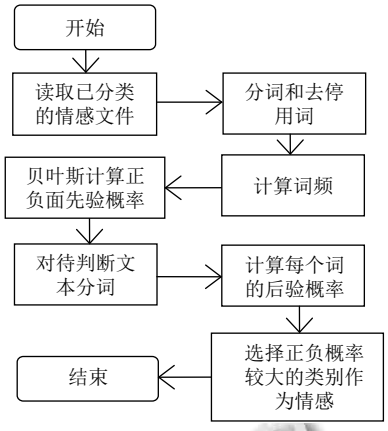

图 1 SnowNLP 情感分析流程图

由于 SnowNLP 自带的语料库本身包含的是不同种类商品评论的语料, 其语料具有局限性与滞后性, 因此情感预测准确率将会受到很大的限制, 所以并不适合本文针对口罩评论的情感分析预测. 因此, 本文通过标注情感为积极和情感为消极的共计 14 400 条评论数据后, 通过 SnowNLP 自带的贝叶斯模型进行训练生成关于针对口罩评论的语料库, 便于后续精确的情感分析预测.

2 数据处理

本文的数据来源于淘宝网站, 其内容为用户对口罩的评论文本信息.

2.1 数据采集

本文通过谷歌浏览器的插件 Web Scraper 进行数据爬取, 获得了关于口罩的用户评论文本数据. 其 Web Scraper 的采集流程如图 2 所示.

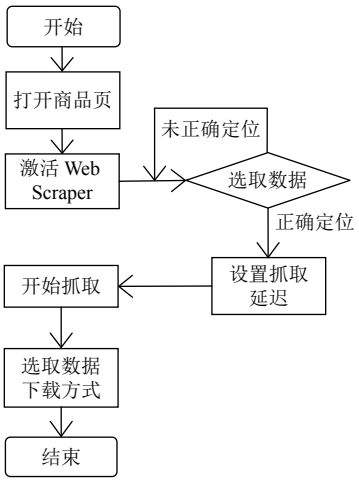

图 2 Web Scraper 采集流程图

2.2 数据预处理

获取的数据里含有许多脏数据, 因此需要进行一系列的数据预处理工作: 首先需要进行数据清洗, 清洗掉无效的表情以及“此用户没有填写评论!”这一类无效评论数据, 然后进行文本分词, 最后去除停用词.

2.3 分词比较

SnowNLP 的分词方法是基于 Character-Based Generative Model[12] 的, 其中$w_1^m=[w_1,w_2,\cdots w_m]$为特定单词序列, c_1^n为给定的包含 n 个字符的句子, $[c,t]$为$[character,tag]$的缩写, 公式如下:

$$P\left(w_1^m|c_1^n\right)\equiv P\left([c,t]_1^n|c_1^n\right)=P\left(c_1^n|[c,t]_1^n\right)\times P\left([c,t]_1^n\right)/P\left(c_1^n\right)$$

可进一步简化为:

$$P\left([c,t]_1^n\right)\approx\prod_{i=1}^{n}P\left([c,t]_i|[c,t]_{i-k}^{i-1}\right)$$

Jieba 分词则是基于 Trie 树结构实现高效的词图扫描, 生成句子中汉字所有可能成词的情况构成有向无环图, 再采用动态规划查找最大概率路径, 找出基于词频的最大切分组合, 对于未登录词 (Out-Of-Vocabulary, OOV), 则采用基于汉字成词能力的 HMM (Hidden Markov Model), 使用 Viterbi 算法, 生成按 B(Begin)E(End)M(Middle)S(Siggle) 标记的中文词汇. 并且 Jieba 分词还支持自定义字典, 对于提高分词准确率有一定帮助.

在文本分词方面, 本文没有用 SnowNLP 自带的分词而选择的是 Jieba 分词. 因为通过对比 SnowNLP 和 Jieba 的分词效果 (如表 1), 可知 SnowNLP 在分词时, 无法识别否定词, 如“不贪便宜”被分成了“不”和“贪便宜”, “不喜欢”分成了“不”和“喜欢”, 这会导致在后续的情感分析时使整体偏向的情绪与语句正确的情绪相反. 但是 Jieba 分词的效果却相对更好, “不贪便宜”和“不喜欢”都分词正确. 因为使用 Jieba 分词能调用 Jieba 分词提供的 load_userdict() 函数来自定义相应的词库 (本文为常用词词典和针对口罩评论词的结合), 优化分词效果, 如“很好”、“不敢恭维”和“买教训”等, 从而提高情感判断的准确率.

表 1 Jieba 分词与 SnowNLP 分词效果对比

原语句	Jieba分词效果	SnowNLP分词效果
口罩很好, 正规产品, 自封袋包装, 方便卫生 , 特殊时期感谢有担当有责任的企业!	口罩/很好/正规/产品/自封袋/包装/卫生/特殊时期/感谢 有担当/有责任的/企业	口罩/很/好/正规/产品/自封/袋/包装/卫生/时期/感谢/担当/责任/企业
不敢恭维, 大家看吧, 我就当花钱买教训, 以后不贪便宜了!	不敢恭维/看吧/就当/花钱/买教训/不贪便宜	不/敢/恭维/看/花钱/买/教训/不/贪便宜
不喜欢这个颜色, 哭唧唧, 还是两包都同一个颜色.	不喜欢/颜色/哭唧唧/两包/都/同一个/颜色	不/喜欢/颜色/哭/唧唧/两/包/都/一个/颜色

3 情感分析检验

在生成了针对口罩评论数据的语料库和对数据进行预处理后, 为了得到经过处理后的数据在此语料库下的情感分析的准确率, 因此本文用通过手工标注的 16 308 条数据进行了情感分析检验. 得到检验表表 2.

表 2 情感分析准确率检验表

语句情感	总数	正确数	错误数	准确率(%)
Positive	8260	7821	439	94.69
Negtive	8048	7215	833	89.65

表 2 中情感为积极的语句的准确率为 94.69%, 情感为消极的语句的准确率为 89.65%, 总语句的准确率为 92.20%. 可见语句情感分析结果较好, 因此可以用此方法对文本数据进行情感分析.

4 用户评论数据分析

通过对文本数据进行情感分析的检验后, 以下将对剩下的 112 622 条口罩的用户评论文本数据进行基于用户评论的每日情感趋势分析和情感关注分析.

4.1 基于用户评论的每日情感趋势分析

通过 SnowNLP 情感分析得到的数值分布在 0 到 1 之间, 数值大于 0.5 的评论情感为积极, 小于等于 0.5 的评论情感为消极. 其中数值越接近 1, 情感越积极, 数值越接近 0, 情感越消极. 在对用户评论的每日情感趋势进行分析的研究中, 对每天所预测出的所有情感数值做了取平均值的处理, 并与国内每日新增病例 (含海外输入) 和本土每日新增病例 (不含海外输入) 一起进行相应分析 (如图 3), 其中病例信息来自国家卫生健康委员会官方网站. 可见用户评论的每日情感数值都较积极, 但整体上情感数值有下降的趋势. 在 3 月 11 日前每日平均情感指数较高, 之后指数就呈现缓缓下降趋势, 而本土每日新增的病例数在 3 月 11 日前病例数都较高, 但是整体处于下降趋势, 而本土病例新增趋势在 3 月 11 日之后就呈现平稳态势, 这与情感指数在 3 月 11 日后整体处于下降趋势相呼应. 可见本土新增病例的趋势在一定程度上影响着情感指数整体上的趋势变化. 而国内新增病例在 3 月 10 日、14 日、16 日、23 日、30 日和 4 月 11 日的趋势上升变化导致了当日或之后一段时间每日情感趋势上升的变化, 在 3 月 12 日、17 日、24 日、31 日和 4 月 6 日、9 日的趋势下降变化也相应导致了当日或之后一段时间每日情感趋势下降的变化. 因此本土新增病例的趋势在一定程度上影响着每日情感趋势的整体变化, 而国内新增病例的局部波动变化趋势也影响着每日情感相应局部的波动变化趋势.

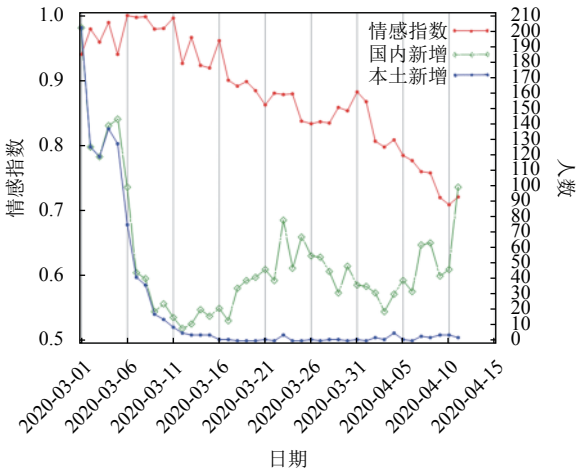

图 3 口罩评论数据的每日平均情感数值与新增病例

4.2 基于用户评论的情感关注分析

在对用户评论进行了每日情感趋势分析后, 为了进一步了解用户对于口罩的关注点, 因此本文将用 SnowNLP 情感分析得出的情感分析数值进行了分类, 分为积极情感和消极情感两类, 再分别取出出现次数前 10 的高频词分别绘制成了情感为积极的高频词柱状图 (图 4) 和情感为消极的高频词柱状图 (图 5). 图 4 中出现频率最高的词为“质量”, 共出现了 30 921 次. 其中从“质量”、“包装”、“价格”、“厚实”等词中可以看出, 影响用户评论情感为积极的因素主要为口罩的质量好、包装好、价格实惠和口罩的厚实, 其次用户也直接对其收到的口罩表达了“不错”、“好”、“挺好”等主观情感.

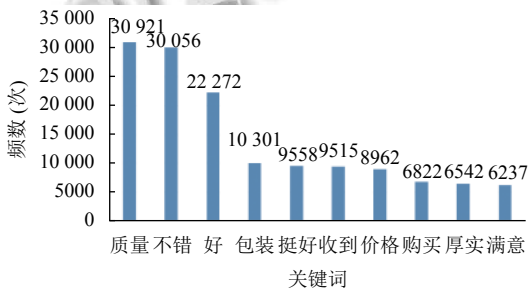

图 4 口罩评论情感为积极的高频词统计柱状图

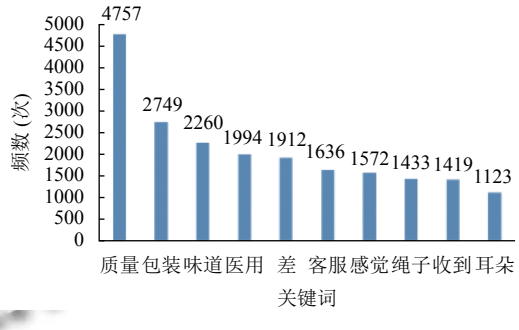

图 5 口罩评论情感为消极的高频词统计柱状图

在图 5 中, 值得注意的是, 用户的消极评论主要围绕在“质量”、“包装”、“味道”、“医用” 等关键词上, 可以看出用户对于口罩的质量差、包装差、有异味、不是医用等有着明显的消极情感. 并且客服的态度也对用户的评论是否为消极有着一定的影响因素.

为了更直观且美观的显示出用户的关注点, 因此绘制出了用户评论情感为积极的词云图 (图 6) 和情感为消极的词云图 (图 7), 图中的字体的大小代表的是词频. 从图 6 中可以看出, 用户关注的核心为口罩的质量, 其次, “包装”、“价格”、“厚实”等词的关注点也较为突出, 体现出了在疫情期间用户对店铺出售的口罩的质量、包装和价格表达了很高的赞美. 从情感为消极的

词云图图 7 中可以看出用户关注的核心依然为口罩的质量, 其次为“包装”、“味道”、“医用”、“客服”等, 体现出用户对个别店铺售卖的口罩的质量差、包装差、有异味、没有医用标准、客服态度差表达了深深的忧虑.

图 6 口罩评论情感为积极的词云图

图 7 口罩评论情感为消极的词云图

因此对于需要提高口罩评论评分的商铺, 可以从口罩的质量、包装、价格、送货速度、厚实度、是否有医用标准以及客服态度上进行改良.

5 结 论

本文对用户评论数据的分析, 是按照日期递增进行的, 并且总天数只有 42 天, 因此对于不同的季节对用户情感关注的影响以及疫情的不同发展阶段对用户对口罩评论的每日情感的发展趋势的影响的分析是不太全面的, 因此后续就需要采集时间跨度更大的数据进行相应研究.

本研究还存在着一定的缺陷, 如情感分析所采集的数据量较小, 导致情感分类准确率只达到了 92.20%, 因此在后续的研究中, 就需要采集更多的数据来对模型进行训练, 进一步提高情感分类的准确率.

参考文献

1 丁森华, 邵佳慧, 李春艳, 等. 文本情感分析方法对比研究. 广播电视信息, 2020, (4): 92–96.
2 陈晓东. 基于情感词典的中文微博情感倾向分析研究 [硕士学位论文]. 武汉: 华中科技大学, 2012.
3 Maks I, Vossen P. A lexicon model for deep sentiment analysis and opinion mining applications. Decision Support Systems, 2012, 53(4): 680 –688. [doi: 10.1016/j.dss.2012.05.025]
4 李慧, 柴亚青. 基于属性特征的评论文本情感极性量化分析. 数据分析与知识发现, 2017, (10): 1–11.
5 刘玉娇, 琚生根, 伍少梅, 等. 基于情感字典与连词结合的中文文本情感分类. 四川大学学报 (自然科学版), 2015, 52(1): 57–62.
6 郝媛媛, 叶强, 李一军. 基于影评数据的在线评论有用性影响因素研究. 管理科学学报, 2010, 13(8): 78–88, 96.
7 陈兴蜀, 常天祐, 王海舟, 等. 基于微博数据的“新冠肺炎疫情”舆情演化时空分析. 四川大学学报 (自然科学版), 2020, 57(2): 409–416.
8 闻彬, 何婷婷, 罗乐, 等. 基于语义理解的文本情感分类方法研究. 计算机科学, 2010, 37(6): 261–264.
9 张启宇, 朱玲, 张雅萍. 中文分词算法研究综述. 情报探索, 2008, (11): 53–56.
10 崔连超. 互联网评论文本情感分析研究 [硕士学位论文]. 济南: 山东大学, 2015.
11 Chen CX, Chen J, Shi C. Research on credit evaluation model of online store based on SnowNLP. Proceedings of 2018 3rd International Conference on Advances in Energy and Environment Research. Guilin, China. 2018. 03039.
12 Wang K, Zong CQ, Su KY. Integrating generative and discriminative character-based models for Chinese word segmentation. ACM Transactions on Asian Language Information Processing, 2012, 11(2): 7.