

在线医药电商评论情感分析

——基于 XGBoost 集成加权词向量和 大语言模型的情感识别模型

田梦影¹, 时 维²

(1. 北京外国语大学国际商学院, 北京 100089; 2. 西交利物浦大学智能工程学院, 江苏 苏州 215028)

摘要: 消费者评论是考察消费者情感的重要数据源, 对商品评论进行数据挖掘是帮助在线医药电商改善经营的重要途径。立足于在线医药电商的用户评论, 基于 SO-PMI(情感倾向点互信息)算法构建该领域情感词典, 对评论词向量进行情感加权。利用 XGBoost(极限梯度提升树)集成词向量和 LLM(大语言模型)构建情感识别模型, 最后得出评论情感指数, 从多个维度展开, 分析消费者评论中的情感趋势。实证分析表明, 构建的情感识别模型的 AUC(曲线下的面积)等验证指标较 LLM 模型相比有进一步提升, 具有一定的应用价值。

关键词: 在线医药电商; LLM(大语言模型); XGBoost(极限梯度提升树)算法; 情感指数; 情感识别

中图分类号: TP391 **文献标志码:** A **文章编号:** 1671-1807(2024)09-0128-08

随着互联网时代的到来, 电子商务呈现蓬勃发展的趋势。据统计, 2023 年“双十一”期间两大电商平台天猫和京东交易额分别达到 5 403 亿元和 3 491 亿元。与此同时, 随着线上商品的种类不断丰富, 以及在线购物的方便快捷, 越来越多的消费者选择在互联网上进行交易, 在线评论数量也随着迅速增长。对于消费者来说, 在线评论已经成为其作出购买决策的重要信息来源之一^[1]。对于商家来说, 蕴含在商品评价中的信息包含着消费者的主观情感, 通过文本情感分析可以让商家及时了解消费者对其商品和服务的反馈信息, 从而对产品和服务进行优化, 获得更多的利润。

随着国家接连出台重磅政策鼓励互联网医疗的发展, 互联网医疗正迎来新的发展机遇。自 2012 年天猫医药正式上线开始, 在线医药电商飞速发展, 行业每年的增长率超过 50%。2019 年年底突然爆发的新型冠状病毒肺炎疫情使得阿里健康、京东健康、壹药网、好药师等 B2C(business to consumer, 商对客)医药电商的交易规模呈爆发式增长^[2], 2023 年中国医药电商市场交易规模已达到 2 852 亿元, 由此可见互联网医疗的发展潜力巨大。通过对在线医药电商文本评论的情感分析, 可以帮助商家获取消费者对于药品和服务质量的主观性感受, 自动识别消费者的褒贬

态度和意见, 基于此进行有针对性的改善^[3]。

1 文本情感分析的研究述评

文本情感分析又称意见挖掘, 是指通过计算机技术自动分析某段文本的内容, 对文本的主观性、情感极性、情感类别进行信息挖掘, 从中提取更多有价值的信息^[4-5]。由于互联网的迅速发展和不断普及, 以及 Web3.0 时代的到来, 互联网应用于用户之间的互动越来越密切, 双向交流的便捷性也使得用户越来越倾向于在网络平台中留下表达自己主观感受的评论。网络文本中蕴含的信息含量急剧上升, 自然语言处理近几年也成为国内外学者的热门研究领域。

文本情感分析的研究方法主要分为基于情感词典和基于机器学习^[6]。基于词典的分析方法是先将文本进行分词处理, 再构造某个领域的情感词典, 在这个词典的基础上, 利用线性代数和统计分析的方法统计文本中积极和消极情感词的数量, 从而确定文本的情感类别^[7]。罗浩然和杨青^[8]提出了基于情感词典和堆积残差的双向长短期记忆网络的情感分析方法, 借助“教育机器人”研究领域内的专业词汇提高了分析此类文本时的精确度, 分类准确率较之前的算法提升了 4.5%。张倩男^[9]基于 Vivo 手机的用户评论数据构建了手机领域的情感词典, 研究发现领域情感词典的手机好评度高于通用情感词典的好评度。邢

收稿日期: 2024-02-07

作者简介: 田梦影(1998—), 女, 重庆人, 硕士研究生, 研究方向为商务分析与大数据; 时维(2000—), 男, 黑龙江大庆人, 硕士研究生, 研究方向为人机交互。

丹和屈仁均^[10]采用了基于词典的方法对跨境电商顾客评论进行了文本情感分析,并在此基础上提出了顾客满意度综合评价方法。

基于机器学习的分析方法是先对文本进行人工标注,并划分为训练集与测试集,利用支持向量机(support vector machine, SVM)、 K 近邻(K -nearest neighbor, KNN)、朴素贝叶斯(Naive Bayes, NB)等机器学习的算法对训练集的特征进行学习,并建立特定的分类模型,然后将分类模型应用于测试集从而进行分类准确性判断^[11]。汪梦欣等^[12]采用基于机器学习的情感分析技术训练学习产品各个属性评论的情感极性,得到正面、负面、中性的评论数量,进一步采用直觉模糊集的方法进行顾客满意度评价研究。林轶和曹清芳^[13]运用贝叶斯机器学习方法训练的SnowNLP库对游客评论文本进行情感分析,并将情感极性划分为正向、中性和负向。医药电商所属的在线健康领域,也有部分学者对医疗在线语料进行了文本分析。由丽萍和何玲玲^[14]采用词典和规则相结合的方法进行在线医疗评论的情感语义分析;高慧颖等^[15]提出了一种基于特征加权词向量的情感分析方法,在分类模型中表现出了更好的效果。

虽然目前文本情感分析和关于在线健康领域的研究正在火热地开展之中,但关于在线医药电商评论的情感分析尚未有较全面的研究,且目前尚未构建在线医药电商领域的情感词典,若使用通用情感词典进行情感分类难免会出现预测准确率不高的状况,且传统的情感分析方法并不足以满足提高

准确率的要求^[16]。因此,本文选取京东商城关键词为“999 感冒灵”的商品作为研究对象,在构建了在线医药电商领域的情感词典后,基于极端梯度提升算法(extreme gradient boosting, XGBoost)集成加权词向量和大语言模型(large language model, LLM)提出了一种新的情感分析模型,从而对消费者评论进行准确的情感极性判断,并根据分析结果构建情感指数,以此反映在线医药电商目前存在的问题,帮助商家把握在线医药电商领域的情感趋势。

2 在线医药电商评论情感识别模型

2.1 构建领域情感词典

在线医药电商情感词典的构建流程如图 1 所示。首先爬取京东大药房感冒用药分类下的商品评论构成语料库,人工对语料库进行情感极性判别,分为正向语料库和负向语料库。对语料库进行分词处理,并导入哈工大停用词表去除停用词、去标点符号。然后统计词频,分别选取出各 20 个正向情感词和负向情感词,并与人工筛选出的大连理工大学中文情感词汇本体库中 10 个正向和负向词进行融合,形成最终的情感种子词。再进一步通过情感倾向点互信息(semantic orientation pointwise mutual information, SO-PMI)算法从京东大药房评论语料库中,在情感种子词的基础上分别找出对应的情感候选词。最后将情感候选词与 3 个使用最广泛的通用情感词典进行去重融合,最终得到在线医药电商评论情感词典。

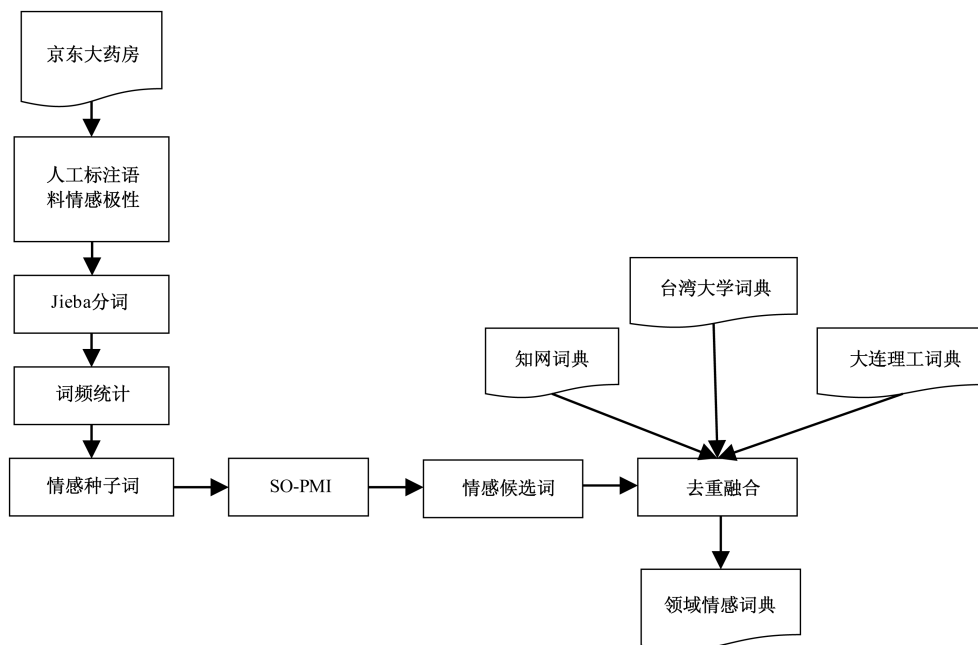


图 1 领域情感词典的构建流程

2.1.1 数据预处理

首先选取京东大药房上感冒用药分类下的商品,按销量降序排列,编写 Python 程序爬取 5 万余条评论,并根据星级得分对评论进行情感极性标注,1、2 星为消极评论,4、5 星为积极评论,从而生成积极情感和消极情感的基础语料库。语料库数据见表 1。

2.1.2 提取情感种子词

为了准确识别评论中包含的多个实体及其复杂的语义结构,最终采用了基于双向控制门单元循环神经网络(bidirectional recurrent neural network, Bi-GRU)预训练的 jieba-paddle 模型对语料库进行分词,并导入哈尔滨工业大学自然语言处理实验室发布的中文停用词表去除停用词和标点符号。分好词后进一步将情感词进行词频统计,从词频最高的词语中挑选出情感表达强烈的各 20 个正向情感种子词和负向情感种子词,再从大连理工大

表 1 基础语料库数据

序号	商品名称	商品评论	情感分类
1	丙酸氟替卡松鼻喷雾剂	过敏性鼻炎太烦人了,用这个喷剂效果很好,第一次从京东买,用了一次感觉不错!	positive (积极)
2	儿感清口服液	宝宝感冒神药,中药成分更安全	positive (积极)
3	连花清瘟胶囊	骗人的商品售前客服说有送药箱,售后又说赠完了	negative (消极)
4	蓝芩口服液	建议大家药物不要在网上买,假的和了几天没效果,就开始怀疑了,到药店买来一天见效	negative (消极)
...

表 2 正负情感种子词

正向情感种子词					负向情感种子词				
很好	方便	实惠	舒服	美丽	差评	怀疑	贵	退换	肮脏
不错	值得	喜欢	保障	安全	垃圾	退款	难受	暴力	邪恶
很快	满意	优惠	有效	正确	失望	烂	差劲	丑陋	悲惨
常备	信赖	好评	缓解	温柔	降价	假药	一星	危险	不当
便宜	推荐	品牌	性价比	卫生	严重	坏	最差	坏	落后
善良	幸福	得当	厉害	佳	过期	慢	吐	错误	脆弱

表 3 正负向情感候选词

正向情感候选词			负向情感候选词		
物美价廉	速度快	正品	乱七八糟	包装费	费事
送货上门	一模一样	老牌子	摔	烂	不良反应
宝贝	质量保证	完好无损	降价	失望	破
...

学中文情感词汇本体库中按照情感强度人工筛选出 10 对正向和负向情感种子词,最终形成构建领域情感词典的 30 对情感种子词,见表 2。

2.1.3 提取领域情感候选词

中文词语常用的情感极性判断方法是基于点互信息(pointwise mutual information, PMI)算法的计算方法^[17]。PMI 用于判断某个词汇与基准词同时出现的概率,即

$$PMI = \log_2 \left[\frac{P(\text{word1}, \text{word2})}{P(\text{word1})P(\text{word2})} \right] \quad (1)$$

式中: $P(\text{word1}, \text{word2})$ 指 word1 和 word2 同时出现在语料的概率,若两个词语相互独立, $P(\text{word1}, \text{word2}) = P(\text{word1})P(\text{word2})$,由此可得 $PMI = 0$ 。若 $PMI > 0$,表示两个词语相关,值越大相关性越强;若 $PMI < 0$,则表示两个词语不相关。

SO-PMI^[18]在 PMI 算法的基础上,判断新词汇与情感种子词的相关程度。若与正向情感种子词的相关程度大,则可以判断该词属于正向情感词汇;若与负向情感种子词相关程度大,则可划分至负向情感词汇。SO-PMI 算法为

$$SO-PMI(\text{word}) = \sum_{i=1}^{\text{num}(\text{pos})} PMI(\text{word}, \text{pos}_i) - \sum_{i=1}^{\text{num}(\text{neg})} PMI(\text{word}, \text{neg}_i) \quad (2)$$

式中: $\text{num}(\text{pos})$ 和 $\text{num}(\text{neg})$ 分别为正负向情感种子词的总数。若 $SO-PMI > 0$,可判断该词语为正向情感候选词;若 $SO-PMI = 0$,可判断为中性词语;若 $SO-PMI < 0$,则可将词语划分至负向情感候选词。用 Python 实现 SO-PMI 算法,导入待处理京东大药房评论语料库以及正负向情感种子词,最后输出正负向情感候选词,见表 3。

2.1.4 合并通用情感词典

将得到的情感候选词人工剔除极不合理的词汇,并与使用最为广泛的 3 个通用情感词典(知网 HowNet 情感词典、台湾大学 NTUSD 简体中文情感词典、大连理工大学中文情感词汇本体库)进行

去重合并,构成最终的在线医药电商领域情感词典。经统计共有 14 554 个正向和 18 684 个负向情感词。

2.2 基于 XGBoost 集成加权词向量和 LLM 的情感识别模型

本研究提出一种将情感词典与机器学习算法结合的方法来构建情感分析模型。首先,在使用词向量技术处理文本数据后,同时引入 LLM 来判断评论的情感极性。在自然语言处理领域,LLM 已经被证明在识别文本情感上极为有效^[19]。首先,LLM 在理解上下文方面表现得比传统模型更为出色,能够准确识别文本情感^[20]。其次,LLM 通常能够处理更复杂的句子结构,识别出更精细的情感层次。已有研究表明,LLM 能够捕捉到细腻的情感差异,即使是在那些含蓄或者复杂情绪表达的文本中^[21]。

为了实现精准判断在线医药电商消费者评论的情感分类,将基于医药电商情感词典的加权词向量和 LLM 判断的情感极性结果作为 XGBoost 的输入层,以此进行集成训练,得到最优的情感判断结果,如图 2 所示。

2.2.1 加权词向量

在训练情感识别模型中,首先面临的问题是将评论文本转换为模型能够有效识别的词向量。采用基于 Bi-GRU 的 jieba-paddle 模型的分词技术,通过这种精准的方法,能够有效地从大量评论中提取出关键的情感信息,同时显著降低了词向量的维度,为后续构建的情感分析模型提供坚实的基础。

在保留核心实体后,使用词频-逆文档频率技术(term frequency-inverse document frequency, TF-IDF)将评论转化为词向量。首先通过 TF-IDF 计算

每个实体的得分。TF-IDF 算法首先计算每个词语的词频和逆文档频率,然后将 TF 值和 IDF 值相乘,最终的值就代表该实体在文档中的重要性权重。

TF-IDF、TF,以及 IDF 计算公式为

$$\text{TF-IDF(词频 - 逆文档频率)} = \text{词频(TF)} \times \text{逆文档频率(IDF)} \quad (3)$$

$$\text{TF(词频)} = \frac{\text{某个词在文章中的出现次数}}{\text{文章总词数}} \quad (4)$$

$$\text{TF(词频)} = \frac{\text{某个词在文章中的出现次数}}{\text{文章总词数}} \quad (5)$$

在得到上述的词向量后,基于前文构建的在线医药电商领域情感候选词的 SO-PMI 得分,对词向量进行加权处理。鉴于评论中情感极性的判定在很大程度上依赖于特定的情感词汇,之前构建的专属领域情感词典发挥至关重要的作用。考虑到情感候选词的 SO-PMI 得分范围较广,使用 Z-score 标准化的方法,将每个情感候选词的 SO-PMI 得分映射到-1~1 的区间内,从而作为情感词的加权系数。

在加权计算中,一旦识别出评论中的情感词实体,就会根据其对应的加权系数乘以其 TF-IDF 得分,来调整最终的权重。在特征选择上,只保留了前 5 000 个最具代表性的特征,并设定了最小文档频率为 5。尽可能地减少词向量的复杂度,同时保留了最有信息量的词汇特征,最终构建了高价值的加权词向量。这种加权机制使得情感词在词向量中得到更合理的权重,从而在情感分析模型中发挥出更大的作用。

2.2.2 基于 LLM 识别情感极性

LLM 模型通常覆盖丰富的语言样本和场景,并

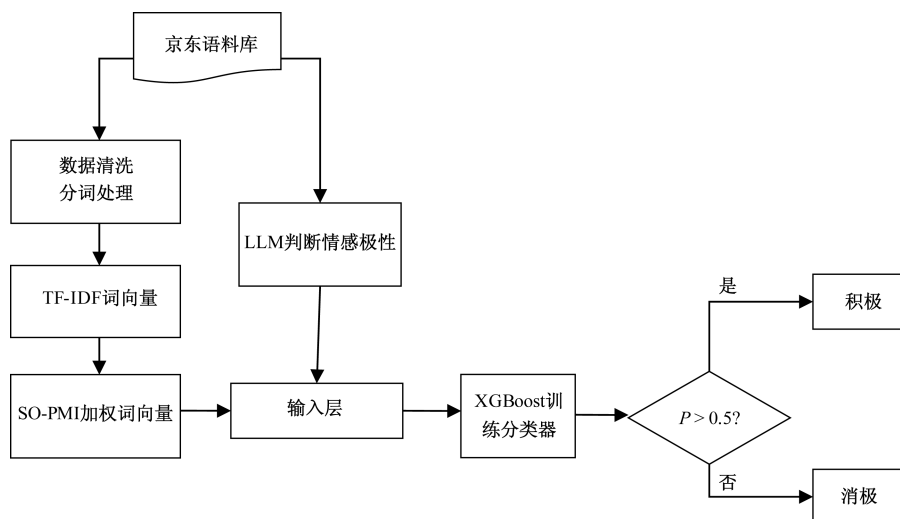


图 2 基于 XGBoost 集成加权词向量和 LLM 的情感识别模型框架

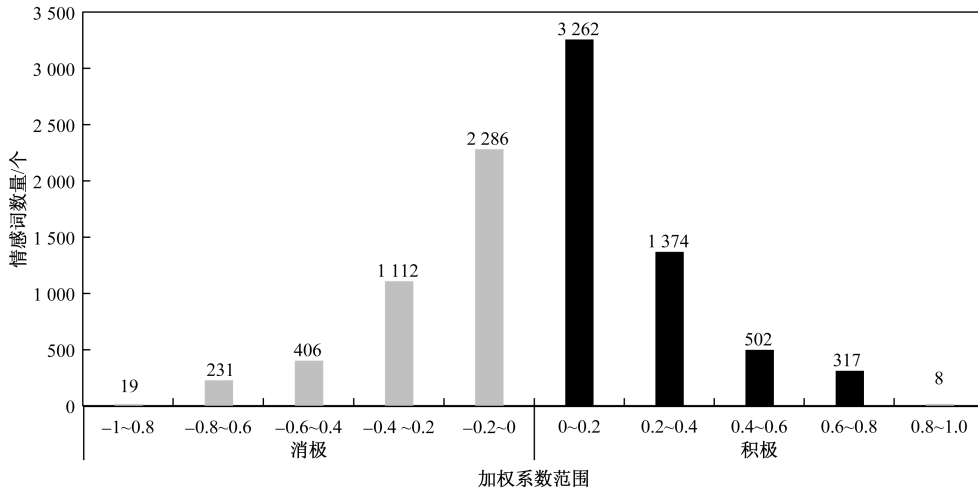


图 3 SO-PMI 标准化后的加权系数分布

且在庞大且多样化的文本数据集上进行训练,故 LLM 能够处理多种语言和表达方式,在自然语言处理(natural language processing, NLP)领域,LLM 已经证明了其卓越的性能和应用价值,无论是在理解复杂的语言结构,还是在处理大规模的文本数据方面,LLM 都展现了其关键的作用。

为了进行准确的情感极性判断,采用 3 种适用于中文的 LLM,包括 Llama-7b、Llama-7b-4bit 以及 Baichuan-7B。为了选择性能更加优异的模型,随机选取 400 个已标注情感的评论,包括 200 个积极性和 200 个消极性的评论。通过表 4 对比分析可知, Baichuan-7B 模型在判断评论的情感极性上表现更为出色,因此后续研究会在使用 Baichuan-7B 的基础上进行展开。

在使用 LLM 判断评论的情感极性时采用直接的提问方式。首先向模型提出明确的问题,引导其

判断评论的情感极性,并限制其回答仅为“积极”或“消极”。这种方法有效降低了模型在响应过程中偏离主题的风险,确保模型专注于判断评论的情感极性。同时,由于指定回答的选项,显著降低了计算资源的消耗。采用直接提问的方式能够更高效且准确地利用 Baichuan-7B 模型来判断评论的情感极性,在提高工作效率的同时又能保证情感分析的质量。

2.2.3 集成训练 XGBoost 模型

如图 4 所示,将加权词向量和 LLM 判断的情感极性结果结合起来作为 XGBoost 的输入层来进行训练,输出结果判断为积极则赋值为 1,消极为 0。通过这种集成方式,不仅利用了词向量对于情感特征的细致的捕捉能力,也充分发挥了 LLM 在理解复杂语义和情感倾向上的优势,使模型的预测能力达到最优。

2.3 实验与结果分析

2.3.1 数据准备

选取京东大药房上关键词为“999 感冒灵”且销量最高的商品作为研究对象,通过 Python 程序爬取消费者购买商品后的评价。在数据清洗后,将剩

表 4 LLM 模型性能

LLM 模型	准确率/%	精确率/%	召回率/%	F/%
Llama-7b-4bit	79.00	77.00	29.60	42.70
Llama-2-7b	88.60	87.10	98.00	92.20
Baichuan-7B	93.20	91.90	98.80	95.30

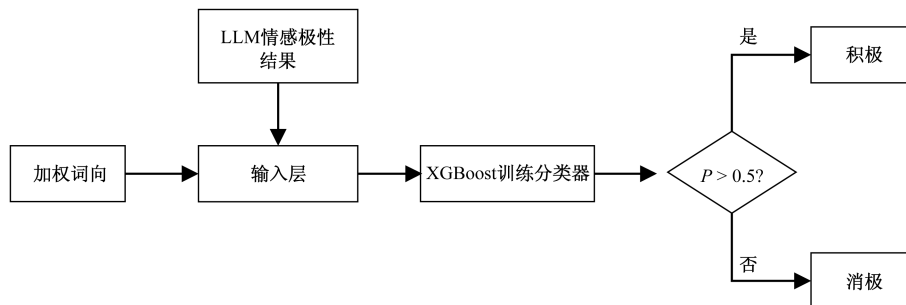


图 4 基于 XGBoost 的情感识别模型

余商品评价进行人工标注情感为 positive(积极)或是 negative(消极),最后分别有 707 条正向和306 条负向评论。表 5 为经过数据处理后的数据集情况。

表 5 京东大药房语料库

情感极性	评论内容	评论数量/条
positive (积极)	九九感冒灵,这款药我个人觉得非常不错,家庭常备药	707
	999 感冒灵活度高,价格实惠,物流速度快,服务态度好,好评!	
	性价比比较高,比药店便宜,一样的东西,很实用	
	...	
negative (消极)	快递慢了一天,发现生产日期是去年 10 月份,有效日期是明年,都已经过了一年还发货给客户,差评!	306
	差评,为什么三盒有两盒包装盒都是拆开过的,产地还不一样	
	一天一个价	
	...	

2.3.2 结果分析

分别构建 3 种模型对在线医药电商领域的评论进行情感分析。如图 5 所示,仅使用 LLM 进行情感极性的判定已经显示出极佳的效果,其精确率(Precision)高达 97%,远超过仅依赖于加权词向量的模型。随后,将加权词向量与 LLM 情感极性判定的结果集成到 XGBoost 模型中,结果显示,尽管精确率略有下降,但准确率、召回率和 AUC (area under curve,曲线下的面积)值得到了进一步的提升。AUC 作为一个衡量二分类模型性能的关键指标,不仅标志着模型在区分正负情感评论上的有效性,也突显了其在情感分析中的综合准确性。

根据以上分析结果可知,基于 XGBoost 算法集成加权词向量和 LLM 的模型训练效果证实了提出的方法在判别在线医药电商评论文本情感极性上的可行性。这种集成策略显著优化了模型的泛化能力和鲁棒性,在实际应用中展现出强大的分析能力。

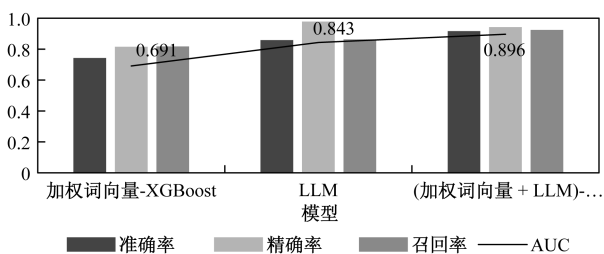


图 5 3 种模型效果对比

3 评论情感指数分析

在线医药电商情感指数是指在对商家所有用户评论进行情感极性量化的基础上,从医药电商独立店铺的维度,利用综合评价算法计算该店铺所有消费者评论的总体情感强度^[22]。在参考文献[23-24]的研究结果后,构建以下情感指数计算模型:

$$I_i = 100 \times \left[\frac{(M_{\text{pos}} - M_{\text{neg}})}{\frac{M_{\text{all}}}{2}} + 0.5 \right] \quad (6)$$

其中: I_i 为第 i 家店铺的总体情感指数; M_{pos} 和 M_{neg} 分别为正向和负向评论的数量; M_{all} 为第 i 家店铺所有的消费者评论。

在京东电商平台选取感冒用药分类下品牌自营店铺销量最高的商品评论,共爬取到 4 家在线医药京东自营店铺的 2023 年 1—12 月的消费者评论,并用前文的情感识别模型自动判断评论情感极性以及计算店铺情感指数,数据见表 6。

从表 6 中可以观察到 999 京东自营官方旗舰店和以岭官方旗舰店消费者的差评数较多,快克旗舰店和云南白药旗舰店的总体情感指数较高,能达到 90 以上。

表 6 4 家医药品牌自营店铺数据

店铺名称	商品名称	消极评论数量/条	积极评论数量/条	总体情感指数
999 京东自营官方旗舰店	999 感冒灵	96	857	89.93
以岭官方旗舰店	莲花清瘟胶囊	85	609	87.75
快克官方旗舰店	快克复方氨酚烷胺胶囊	33	655	95.20
云南白药官方旗舰店	云南白药云丰黄连上清片	26	779	96.77

3.1 情感指数店铺维度分析

运用绘图软件对评论数据集以月维度对 4 家店铺进行分析,并制作评论类型折线图,如图 6 所示。

3.2 评论分类月维度分析

999 京东自营官方旗舰店和快克官方旗舰店的评论数从 1 月到 10 月保持在一个稳定的水准,到 11 月评论数激增。而云南白药官方旗舰店的评论数全年都是保持比较平稳的趋势,在 7 月份数值有所下降。可能存在的原因是,每年 11 月正值电商“双十一”购物节,消费者喜欢在电商平台优惠时期屯这 3 家的药品,而云南白药云丰黄连上清片对消费者而言是需要用到时才会有进行购买的决策。快克官方旗舰店在除去电商节消费者购买数激增

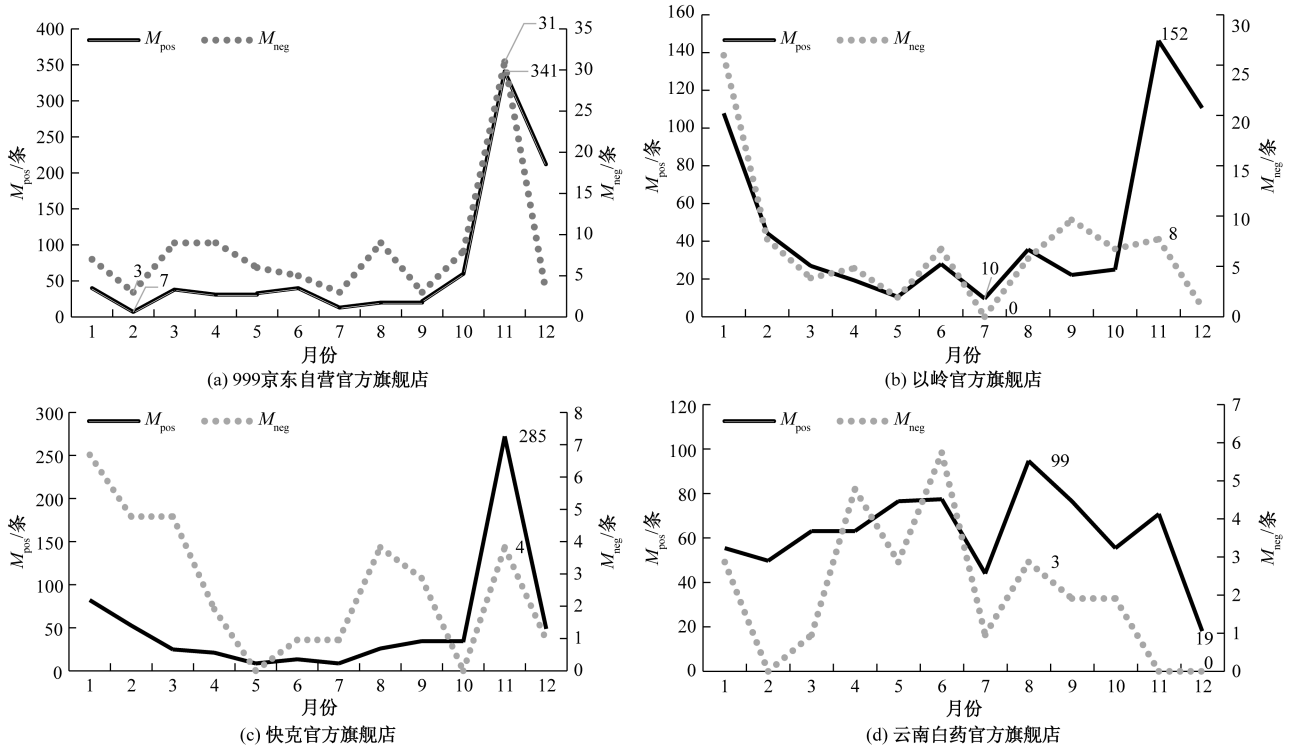


图 6 4家店铺评论数据分析结果

的因素后,差评数集中在8月和9月,云南白药官方旗舰店的差评集中在4—6月,品牌商家可以集中分析这些时间段消费者的差评内容,并针对性地提升服务质量以提高消费者满意度。

3.3 情感指数店铺月列联分析

用绘图软件对店铺和月份两个维度进行交叉列联分析,并制作折线图,如图7所示。

图7显示,云南白药的情感指数在一年内总体表现最平稳,且数值都在90以上;快克情感指数总体表现排第2,但一年之中有波动,如3月份和4月份的情感指数数值相差10;以岭官方旗舰店的情感指数在7—9月存在明显波动,7月情感指数最高可达90以上,9月只有70;而999京东自营官方旗舰店的情感指数在一年内都有波动,3月情感指数最低,只有80,11月情感指数最高,达到90以上。通

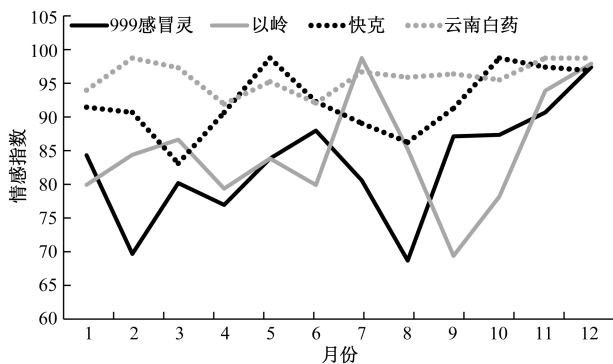


图 7 情感指数店铺月列联分析

过上述分析,能够得出情感指数可以作为衡量消费者情感趋势的重要指标。

4 总结

(1)基于 SO-PMI 算法构建了在线医药电商的领域情感词典,并采用集成加权词向量和 LLM 输出情感极性作为 XGBoost 输入层的评论情感分析模型,自动判断了在线医药电商评论的情感倾向。实验结果表明,本研究提出的情感识别模型在情感分析领域展现了巨大的潜力,不仅在理论上具有创新性,而且在实际应用中展现出了卓越的性能,为未来的研究和实践提供了新的方向。

(2)建立了情感指数,多维度分析了京东商城4家医药品牌自营店铺的总情感倾向。以京东大药房感冒用药分类下商品在线评论为研究对象的实证分析表明,构建的在线医药电商领域情感词典有较好的文本情感极性判断性能,建立的情感指数模型能够动态监测顾客情感变化,帮助品牌商家及时把握整个在线医药电商行业的情感趋势。

本文仍存在一些不足。首先抓取到的数据只是感冒用药分类下的商品评论,未来可将药品类型纳入研究模型之中。其次针对店铺的情感指数主要是基于积极评论和消极评论的数量作为计算依据,未来可以考虑纳入评论的情感强度,使得情感指数对于情感总量的刻画更为客观具体。最后,在

构建情感词典时,仍有人工干预的步骤,未来可以尝试构建领域自适应情感词典,进一步提升对任意领域海量文本进行情感判断的效率。

参考文献

- [1] 朱振中,单明辉,房晓芸. “陈述”还是“修辞”:在线评论语言风格对消费者购买意愿的影响[J]. 财经论丛, 2023(11): 92-102.
- [2] 张丽,张祯. 基于文本挖掘的新冠肺炎疫情下医药在线消费者的需求研究[J/OL]. 运筹与管理, 1-8[2024-03-09]. <http://kns.cnki.net/kcms/detail/34.1133.G3.20230901.0849.002.html>.
- [3] 黄仁,张卫. 基于 word2vec 的互联网商品评论情感倾向研究[J]. 计算机科学, 2016, 43(S1): 387-389.
- [4] 杨立公,朱俭,汤世平. 文本情感分析综述[J]. 计算机应用, 2013, 33(6): 1574-1578.
- [5] 李胤丞,刘继. 基于注意力机制和双通道神经网络的文本情感分析[J]. 信息技术与信息化, 2023(12): 71-74.
- [6] 刘爽,赵景秀,杨红亚,等. 文本情感分析综述[J]. 软件导刊, 2018, 17(6): 1-4.
- [7] 王新宇. 基于情感词典与机器学习的旅游网络评价情感分析研究[J]. 计算机与数字工程, 2016, 44(4): 578-582.
- [8] 罗浩然,杨青. 基于情感词典和堆叠残差的双向长短期记忆网络的情感分析[J]. 计算机应用, 2022, 42(4): 1099-1107.
- [9] 张倩男. 基于情感分析和机器学习的用户评论信息挖掘[J]. 科技和产业, 2023, 23(23): 121-127.
- [10] 邢丹,屈仁均. 基于主题模型和情感分析的进口跨境电商顾客满意度评价研究[J]. 科技和产业, 2023, 23(1): 58-65.
- [11] 肖慧莲,徐锐. 基于文本挖掘的生鲜电商顾客满意度研究[J]. 科技和产业, 2022, 22(1): 288-294.
- [12] 汪梦欣,耿秀丽,白玛吉姆,等. 基于在线评论与情感分析的顾客满意度评价研究[J]. 软件导刊, 2023, 22(8): 10-16.
- [13] 林轶,曹清芳. 基于 LDA 模型及情感分析的电影主题公园评论文本研究[J]. 科技和产业, 2023, 23(22): 93-98.
- [14] 由丽萍,何玲玲. 基于框架语义的在线医疗评论情感分析[J]. 现代情报, 2020, 40(3): 111-116.
- [15] 高慧颖,公孟秋,刘嘉唯. 基于特征加权词向量的在线医疗评论情感分析[J]. 北京理工大学学报, 2021, 41(9): 999-1005.
- [16] 苏兵杰,周亦鹏,梁勋鸽. 基于 XGBoost 算法的电商评论文本情感识别模型[J]. 物联网技术, 2018, 8(1): 54-57.
- [17] 叶霞,曹军博,许飞翔,等. 中文领域情感词典自适应学习方法[J]. 计算机工程与设计, 2020, 41(8): 2231-2237.
- [18] 黄创霞,温石刚,杨鑫,等. 个体投资者情绪与股票价格行为的互动关系研究[J]. 中国管理科学, 2020, 28(3): 191-200.
- [19] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv Preprint arXiv, 2018: 1810.04805.
- [20] ZHANG Z, ROBINSON D, TEPPER J. Detecting hate speech on twitter using a convolution-gru based deep neural network[C]//The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, ESWC, 2018: 745-760.
- [21] SUN C, QIU X, XU Y, et al. How to fine-tune bert for text classification? [C]//Chinese Computational Linguistics: 18th China National Conference, CCL 2019. Kunming: CCL, 2019: 194-206.
- [22] 金家华,吴浪涛,张婷婷,等. 基于情感分析的外卖商家评分研究[J]. 管理学刊, 2020, 33(2): 66-75.
- [23] 刘玉林,菅利荣. 基于文本情感分析的电商在线评论数据挖掘[J]. 统计与信息论坛, 2018, 33(12): 119-124.
- [24] 刘苗,李蔚,朱述政,等. 基于互联网文本情感分析的消费情感指数构建[J]. 统计与信息论坛, 2018, 33(8): 31-38.

Reviews Sentiment Analysis of Online Pharmaceutical E-commerce: Sentiment Recognition Model Based on XGBoost Integrated Weighted Word Vector and Large Language Model

TIAN Mengying¹, SHI Wei²

(1. International Business School, Beijing Foreign Studies University, Beijing 100089, China;

2. School of Advanced Technology, Xi'an Jiaotong Liverpool University, Suzhou 215028, Jiangsu, China)

Abstract: Consumer reviews are an important data source to investigate consumer sentiment, and data mining of product reviews is an important way to help online pharmaceutical E-commerce improve business. Based on the user review text of online pharmaceutical e-commerce, a sentiment dictionary for online pharmaceutical e-commerce based on the SO-PMI (semantic orientation pointwise mutual information) algorithm is constructed, then integrate weighted word vectors and the LLM (large language model), train classifiers are used to construct a sentiment recognition model using the XGBoost (extreme gradient boosting) algorithm. Finally a sentiment index to analyze the sentiment tendency of the online pharmaceutical E-commerce consumer reviews is established from multiple dimensions. The empirical analysis shows that the validation indexes of the constructed emotion recognition model such as AUC (area under curve) are further improved than that of the LLM model, which has certain application value.

Keywords: online pharmaceutical E-commerce; LLM (large language model); XGBoost (extreme gradient boosting) algorithm; sentiment index; sentiment recognition