DOI:10.11918/j.issn.0367-6234.201706178

一种情感分析与质量控制的异常评论识别方法

张 瑞[1,2],金志刚[1],胡博宏[1],张子洋[1]

(1.天津大学 电气自动化与信息工程学院,天津 300072; 2.天津中德应用技术大学 软件与通信学院,天津 300350)

摘 要:针对因数据量的增加以及异常评论策略的更新,以用户内容和行为为基础的传统微博异常评论识别方法效果不断下降的问题,提出一种基于情感分析和质量控制的微博异常评论识别方法. 通过将预处理后的微博评论进行情感分析,将微博评论进行量化处理,在对微博评论进行质量控制的过程中,根据异常与正常用户在时域上对热点微博的评论分布差别检测可疑时间间隔,结合用户聚类分析,设计了异常评论识别模型. 结果表明:该方法利用情感评分,对于评论文本进行较为准确的情感分类,然后通过调整边界值范围和时间阈值范围来限定异常检测等级,当边界值范围增大时,对于异常评论的检测范围扩大,容忍度下降,检测灵敏度高;当时间阈值扩大时,容忍度提高,检测灵敏度较低;适当的选择边界值和时间阈值,可以有效提高与正常评论行为相似的异常评论识别准确率.

关键词:情感分析;质量控制;微博评论;异常检测;时间阈值;识别方法

中图分类号:TP391 **文献标志码**:A **文章编号**:0367-6234(2018)09-0164-07

A spammer detection method based on sentiment analysis and quality control on comments

ZHANG Rui[1,2], JIN Zhigang[1], HU Bohong[1], ZHANG Ziyang[1]

(1.School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China;
2. Department of Software and Communication, Tianjin Sino-German University of Applied Sciences, Tianjin 300350, China)

Abstract: To avoid the poor effect of spammer detection in traditional methods based on content and user behavior due to the increase of data and spammers' updated strategy, a spammer detection method based on sentiment analysis and quality control in microblog was proposed. In the method, the pre-processed comments in microblog were quantified by sentiment analysis. Then in the process of quality control of microblog commentary, the suspicious interval was detected according to different distribution between spammers and normal users for hotspot in varying time. Then a model for spammer recognition was established by cluster analysis. The experimental results showed that the method used the emotion score to make the emotion classification more accurately for the comment text, and then adjusted the boundary value range and time threshold range to limit the anomaly detection level. When the boundary value range increased, the anomaly detection range of the abnormal comment was increased, the tolerance was reduced and the detection sensitivity was improved. When the time threshold was expanded, the tolerance was improved and the detection sensitivity was reduced. Therefore, the appropriate choice of the boundary value and time threshold can effectively improve the accuracy of anomaly recognition which is similar to the normal commentary behavior.

Keywords: sentiment analysis; quality control; microblog commentary; spammer detection; time threshold; detection method

当前伴随着社交网络的飞速发展,信息的传递模式发生了巨大改变,以微博为代表的社交网络已经成为当前最重要的媒体之一. 网络信息数量的急剧增长为网络监管带来了极大的挑战,网络水军可以通过撰写大量不合实际的内容或评论,制造舆论话题,控制舆论导向,推动各种事件或产品以获取利益[1]. 如何在微博平台上准确有效地检测异常评论,减少由网络环境带来的恶劣影响,是当今的研究热点. 目前对于微博异常评论的研究主要分析微博的转发和评论中时序特征、内容特征,从性质的角度来分析和判别[2-4].

网络水军识别问题是一个二分类问题[5],即从全体发布信息的用户中区分出正常与异常的账号. 当前的水军所发表的异常言论识别方法主要以内容特征等为根据,文献[6]通过计算评论的文本倾向性,从而发现偏离正常用户评论的水军评论及其分布,文献[7]主要分析虚假评论和正常用户之间不

收稿日期:2017-06-28
基金项目:国家自然科学基金(71502125)
作者简介:张 瑞(1986—),男,讲师,博士研究生;
金志刚(1972—),男,教授,博士生导师
通信作者:金志刚,zgjin@tju.edu.cn

同的评论模式特征，如评论语言具有大量的重复部分等. 随着用户辨识力的增强，使得原先依赖制造垃圾内容，并具有鲜明特征的水军攻击效果和影响大大降低，为了继续掌控网络评论的导向，诱导其他用户作出错误的评价，网络水军产生了新的欺骗策略[8]，但是整个网络水军的评论分布与正常用户依然有着差距，其特征不容易被用户行为所影响，而且通过在时域上检测可疑时间间隔，也有一定的效果. 文献[9]通过使用两个流聚类算法进行异常检测，可以准确识别 Twitter 中的异常用户. 文献[10]通过一种时间序列模式识别技术分析可疑时间间隔的多种因素，以识别水军，具有较高的有效性和准确性. 文献[11]根据异常用户与异常信息之间的紧密联系，将异常用户检测与异常信息检测相结合，并考虑信息之间的连接与用户二者之间的关系来优化检测结果，但该算法复杂度相对较大.

本文分析了微博用户对于某一话题的评论，运用情感分析，结合时域分析构建异常评论识别模型 SATAUCA（temporal and user correlation analysis based on sentiment analysis），通过分析上述文献中所使用的用户属性和行为特征，结合文本中的情感分析对评论进行量化处理，并通过进行时域上的变更检测及用户聚类分析进行质量控制计算，提出了面向微博的异常评论识别模型，并通过新浪微博真实数据进行仿真实验，验证了本文方法的准确性和有效性.

1　SATAUCA 模型（sentiment analysis TAUCA model）

SATAUCA 模型包括微博评论数据获取、情感分析、TAUCA（temporal and user correlation analysis）检测. 由 SATAUCA 模型构建的微博评论异常检测模型如图 1 所示.

1.1　情感分析模型

微博评论字数一般较短，大多集中在 30 字以内，且以评论者对于被评论微博内容表示赞同或者反对，喜爱或厌恶等态度为主，因此情感的表达是微博评论文本的最主要特点，微博情感评分是对评论文本中的情感和观点进行分析和挖掘来判断其中的情感倾向. 微博情感分析的基础是情感词典的构建，情感词典的覆盖程度对情感分类效果有很大的影响[12-13]，当前中文文本情感分析中，所流行使用的情感词典是知网发布“情感分析用词语集（beta 版）”[14]，台湾大学 NTUSD 中文情感极性词典[15]等，但由于以上词典欠缺情感强度标注，因此本文采用情感领域本体[16]，中文情感词汇本体的情感分类体系是在国外比较有影响的 Ekman 的 6 大类情感分类体系的基础上构建的，情感共分为 7 大类 21 小类. 使用该词典可以更为准确地计算微博评论的情感评分，该情感词汇本体中的一般格式见表 1.

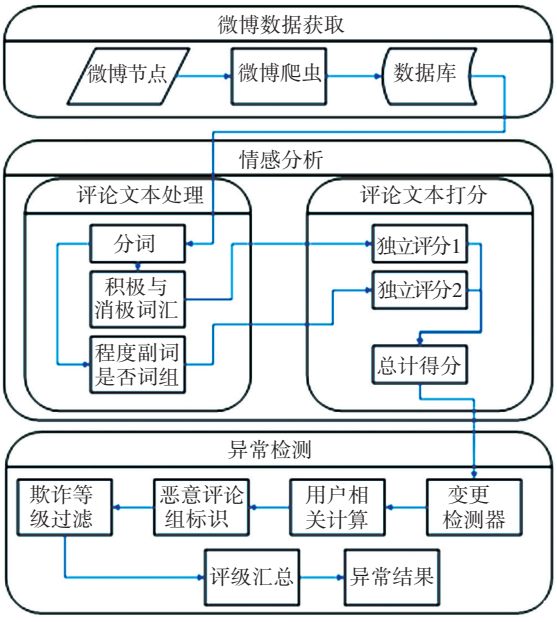

图 1　SATAUCA 模型构建的微博评论异常检测模型

Fig.1　Microblog comment anomaly detection model based on SATAUCA

表 1　情感词汇本体格式举例

Tab.1　Examples ofemotional vocabulary ontology

词语	词性种类	词义数	词义序号	情感分类	强度	极性
无所畏惧	idiom	1	1	PH	7	1
手头紧	idiom	1	1	NE	7	0
周到	adj	1	1	PH	5	1
言过其实	idiom	1	1	NN	5	2

情感分类按照文献[16]所述，情感分为 7 大类 21 小类. 情感强度分为 1,3,5,7,9 五档，9 表示强度最大，1 为强度最小. 情感分类见表 2.

情感词汇本体中的词性种类一共分为 7 类，分别为名词（noun），动词（verb），形容词（adj），副词（adv），网络词语（nw），成语（idiom），介词短语（prep）. 每个词在每一类情感下都对应了一个极性. 其中 0 代表中性，1 代表褒义，-1 代表贬义. 词汇本体共含有情感词共计 27 466 个. 同时在微博评论文本中，经常会出现诸如“特别”、“极其”、“非常”等强化情感的副词，这类副词会影响其所结合的情感词汇的权重赋值，以及“不是”、“绝不”等表示转折和否定的副词会影响其所结合的情感词会的方向赋值. 因此需要构建副词词典来对评论文本进行更为准确的打分. 这里本文根据副词的强度来进行人工标注，如有点的权重赋值为 1，比较为 2，非常为 3，

及其为4,不为-1. 另外,由于微博文本具有短文属性,即句子间的连接关系会影响微博文本的情感分类,因此这里要在分析中加入连词词典是非常必须要的. 在这里本文将连词分为两类,顺序连接:并且,而且,所以,接着,并等;转折连接:但是,然而,可是,竟然等. 根据连词性质的不同,本文定义如下运算规则,顺序连接:连词前后分句值相加;转折连接:取转折后分句的值.

表2 情感分类

Tab.2 Emotional classification

编号	大类	小类	例词
1	乐	快乐(PA)	神采、欢快、安逸
2	乐	安心(PE)	安好、安适、自在
3	好	尊敬(PD)	指导、致敬、仰望
4	好	赞扬(PH)	清峻、盛誉、雄劲
5	好	相信(PG)	信奉、确定、倚重
6	好	喜爱(PB)	爱上、心动、一见钟情
7	好	祝愿(PK)	渴慕、希望、长生不老
8	怒	愤怒(NA)	投诉、窝火、羞愤
9	哀	悲伤(NB)	苦口、伶仃、贫寒
10	哀	失望(NJ)	叹息、死心、无语
11	哀	疚(NH)	抱疚、后悔、抱歉
12	哀	思(PF)	先烈、思念、挂记
13	惧	慌(NI)	慌乱、心虚、结巴
14	惧	恐惧(NC)	疑惧、惊吓、危亡
15	惧	羞(NG)	害臊、羞涩、怯场
16	恶	烦闷(NE)	板脸、迷糊、燥热
17	恶	憎恶(ND)	伎俩、阴谋、中伤
18	恶	贬责(NN)	责备、不雅、数落
19	恶	妒忌(NK)	嫌隙、妒火、醋劲
20	恶	怀疑(NL)	动摇、可疑、多心
21	惊	惊奇(PC)	错愕、奇迹、骇然

通过以上词性分解和词典构建,可以得到文本情感算法,将分词后的词汇集合进行连词判断,如果出现了转折连词,则取转折后的词汇集合作为文本情感评分的主体,如果没有则依序将词汇集合进行求和.

本文以微博评论语句作为研究对象,首先对单条微博进行文本预处理,并以标点符号为分割标志,将单条微博分割为 n 个句子,提取每个句子中的情感类别,在本文主要采用 NLPIR 分词工具包来进行微博文本分词. 在已切分的分句中,根据句子中出现的情感类型属性词数作为情感词分类特征值. 第2步,根据获得的分句情感特征值,将原有微博文本转化为分句情感、副词、连词组成的集合. 第3步根据已获得集合,获得整体评论的情感分类. 比如"昨天中国足球队收获平局,真是无法满意的结果,但是队员们积极拼搏的态度还是让球迷看到了未来的希望"给出一个包含3个句子的微博文本示例,通过以上步骤可以分析出该微博文本包含1个无情感,1个贬责的恶情绪,1个祝愿的好情绪,但由于最后一个"祝愿"分句前有一个转折连连词,因此以该分句作为整体文本的情感数值.

接下来,本文将已打分的微博评论数据依照以上规则,按照图2所示流程和整体情感倾向分值计算规则所示计算方法进行情感赋分,生成情感评论评分数据集. 本文将处理后的所有微博评论导入到质量控制模型中,用以发现异常评论.

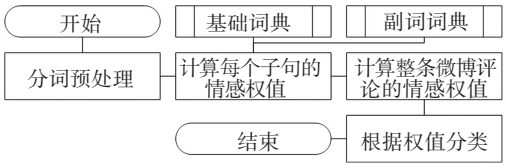

图2 微博评论打分流程图

Fig.2 Microblog comments scoring flow chart

1.2 质量控制模型

本文使用一种异常检测机制[17](temporal and user correlation analysis,TAUCA),TAUCA 模型包括3个主要步骤:1)变更探测;2)用户关联计算;3)恶意用户组识别. 其中最主要的是变更检测. TAUCA 使用改进的 CUSUM 作为变更检测器. 在众多名誉系统当中,每个项目都会有较为固定的评分分布,如果评分快速改变或者超出正常范围,则会被认为其中有恶意用户存在. 如果变更探测器被某一个项目触发了,这个项目就会被标记为被攻击状态. 在用改进的 CUSUM 中,变更发生的时间间隔被称为可疑间隔. 用 P_{θ_0}、P_{θ_1} 为变化前后的概率密度函数 PDF,用 y_k 为数据集(评分集)的第 k 个样本. 当 P_{θ_0} 为平均数为 μ_0 的高斯函数,P_{θ_1} 为平均数为 μ_1 的高斯函数,且二者的方差均为 σ^2 时,基本的 CUSUM 决策函数为

$$g_k = \max(g_{k-1} + (y_k - \mu_0 - \frac{\mu_1 - \mu_0}{2}, 0). \quad (1)$$

在改进的 CUSUM 决策函数中,假设改变前的平均值为 μ_0,改变后的平均值为

$$\mu_1^+ = \mu_0 + v(\text{或者}\mu_1^- = \mu_0 - v). \quad (2)$$

根据式(2),可以定义两个方向的检测函数分别为

$$g_k^+ = \max((g_{k-1}^+ + y_k - \mu_0 - v/2), 0), \quad (3)$$

$$g_k^- = \max((g_{k-1}^- - y_k + \mu_0 - v/2), 0). \quad (4)$$

其中 v 为未知参数,表示检测的灵敏度,v 的值越小

检测器越灵敏.

在攻击中,如果两个攻击方向同时被检测到,定义检测值较大的方向为攻击方向. 变更检测器识别出改变并且发出警告的时间即停止时间用 t_a 表示,t_a 的计算公式为

$$t_a = \min\{k: g_k \geqslant \bar{h}\}. \tag{5}$$

式中 $\bar{h}$ 为阈值. 当 $g_k \leqslant 0$,或者 $g_k \geqslant \bar{h}$ 时,CUSUM 检测器重新启动,g_k 将归为 0,同时一个新的检测周期开始运行. 即使分布不是高斯分布,上面的检测器对于平均值变化也是敏感的. 非工作时间由 t_b 来表示,作为 g_k 下降到阈值 h 的时间,即

$$t_b = \min\{k: g_k < \bar{h}, \text{且 } k \geqslant t_a\}. \tag{6}$$

式中:t_s 为改变的开始时间,t_e 为改变的结束时间,t_1 为计数开始时间. 改进的 CUSUM 可以检测恶意改变的间隔,如果这没有检测到恶意改变,$t_1 = 0$. 如果存在前期检测改变,则非工作时间为 t_b^{pre},此时 $t_1 = t_b^{pre}$. 从检测曲线 g_k 中,可以获取 t_b、t_b^{pre}. 下一个目标为估计可疑间隔的开始时间 t_s 和可疑间隔的结束时间 t_e,t_s 的最大可能估计值为

$$\hat{t}_s = \arg \min_{t_1 \leqslant c \leqslant t_a} \sum_{i=t_1}^{c-1} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)}, \tag{7}$$

t_e 的最大可能估计值为

$$\hat{t}_e = \arg \max_{\hat{t}_s \leqslant h \leqslant t_b} \sum_{i=h+1}^{t_b} \ln \frac{p_{\theta_0}(y_i)}{p_{\theta_1}(y_i)}. \tag{8}$$

图 3 表示了几个特殊时间之间的关系. 在可疑间隔内提供评分的用户为恶意用户. 对于两个恶意用户 X、Y,假设二者共有 n 个共同评分项,二者的共同评分项分别为 $\{x_1, x_2, \cdots, x_n\}$ 和 $\{y_1, y_2, \cdots, y_n\}$. 两个用户间的距离的计算公式为

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{9}$$

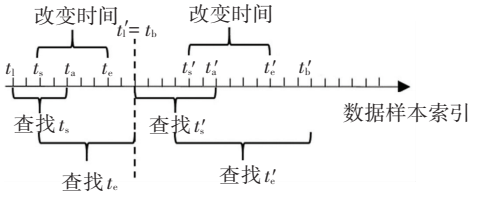

图 3　改进的 CUSUM 示意图

Fig.3　Improved CUSUM schematic diagram

然后使用 K-means 算法可以将可疑用户分为可疑用户组,对于每一个可疑用户组,定义组内对被攻击项目的平均评分为组评分(group rating). 当攻击方向为增加方向时,组评分最高的可疑用户组被标记为恶意用户组,该组中所有的用户被标记为恶意用户. 同理,当攻击方向为降低评分时,恶意用户组为组评分最低的组.

在本文中,一条微博博文的评论应具有一定的倾向性,表达用户对该博文的态度. 微博评论中选取情感评分作为用户态度的量化指标,表达其态度的倾向性. 一条微博博文的所有评论在情感上应具有较为固定的分布,当该分布出现快速改变时,或者出现评分远离该分布时,可以认为有异常评论出现了,同时根据该分布快速改变的方向可以确定异常评论所体现出的倾向性对于原有微博评论情感分布的影响. 因此质量控制模型 TUACA 适用于微博评论中的隐性评分场景.

1.3　微博热点话题和评论特点分析

微博热点话题有着热度保持强的特点,在各垂直领域发挥着聚合舆论、甚至引领行业的作用. 一方面,地方性和持续性话题始终保持良好热度,另一方面,与明星、电视剧、社会性相关的新兴话题也迅速博得网民的积极关注与热议. 同时,为更好地发挥微博的社交资源优势,微博中推出了热点话题榜这一功能,可以获取到 1 h、24 h、1 周、1 个月不同周期的热点微博. 同时根据采集的热点微博样例数据可以表明一条热点微博的持续周期通常不超过 1 个月,尤其在前 7 d 评论最为热烈和集中.

接下来对采集数据中的情感评分进行分值统计,微博评论中的情感分布呈现双峰形态,根据本文的评分规则统计得到大量评论集中于[-25,-5]和[5,25].

1.4　情感分析与质量控制的异常评论识别模型

在微博场景下,用户对于微博会发出评论,根据该评论可计算得到相应情感评分值,对于一条微博的所有评论,当该情感值分布区间集中在 $[a,b]$,该微博评论的情感方向为 $a \rightarrow b$,若在 t 时刻出现了偏离该区间的情感评论,即该评论情感评论项依照改进的 CUSUM 检测函数可被检测到. 当有多个用户的评论出现对于 $[a,b]$ 区间的情感偏离时,对这些评论进行聚类,可得到多个新的情感评论区间,其中情感评论分布数量最多的评论区间即为式(9) 所表示的可疑评论组. 同时根据可疑评论组的情感分布区间与原始评论区间的偏离值的正负性,可以确定攻击方向是增加评分还是降低评分.

2　实　验

本文将按照图 1 所示过程进行实验,实验所使用的运行系统配置为 2.4 GHz i7 处理器、16 G 内存、64 位 Windows 8.1 操作系统以及 Python2.7 版本环境. 使用新浪微博爬虫采集本实验所需数据集,在近 1 个月的时间内对于评论数在 5 000 以上的热

点微博进行了采集和整理,共获得 20 条微博及其相关评论. 将其作为评论集 $A_x(x \in [1,20])$, 其中每条评论均包含评论者的 ID,评论内容和评论发表时间. 将获取的所有微博评论和微博主题正文进行人工识别. 将采集到的微博数据进行 5 倍放大,并随机抽样成 20 份样本,保证每条微博数据在 20 份数据样本中出现 5 次. 将评论数据样本、情感词典以及评分规则派发给 20 位本学科专业研究人员进行情感评分的人工标注和异常评论的人工选取. 这样对于每条微博评论的评分可以保证 20 位研究人员中 5 位进行了情感评分和异常标注,以期得到更为客观的评分数据. 在人工标注的结果一致性上采取聚类方法进行统计,根据 High-Voting 投票原则来决定情感评分有分歧的评论分数和异常标注. 最后将评分数据进行汇总,得到所有微博评论的人工评论分数和异常检测结果.

2.1 SATAUCA 检测实验

将上一节中已经生成的标签短语文档 A_x 导入到 SATAUCA 中进行模型训练和计算:1)将 A_x 文档进行情感分析,对每条评论进行打分,得到情感评分集 B_x ;2)对该数据集进行时域质量控制检测,其中评分边界值范围取 $v \in [0,v_{max}]$, 时间阈值范围为 $t \in [0,t_{\max}]$; 3)在不同的时间阈值和边界值上,会获得该"时间阈值-边界值"下的 $B_{x_{v,t}}$, 以及该平均值 $\mathrm{avg}(B_{x_{v,t}})$; 4)分别进行质量控制后,输出可疑评论编号,形成可疑评论组 C_x.

该方法可以通过调整边界值范围和时间阈值范围来限定欺诈检测等级. 这里选取《你好乔安》和"老虎伤人"这两个近期的热点话题作为样例,与人工识别标注作对比. 由于两个热点话题均值都大于 0,因此将两个热点话题边界值设定在 0~25 的正面评价范围,根据微博评论热点持续周期特点,将时间阈值设为 0.5~30 d,其检测结果见表 3.

可以在表中观察到在"老虎伤人"的评论中,情感评分随着时间的推移,评分均值变化不大且最大值也在 5 左右,因此对于该微博的评论检测在边界阈值为[0,5]时已经可以取得较好的效果. 而对于《你好乔安》这篇微博的评论,其均值随着时间推移产生较大变化,可以看到边界值需要增大到[0,10]以上才能取得较好的检验效果.

还可以观察到,在同样的边界条件下,随着时间阈值的增大,对于样本检测嫌疑数会逐渐减少,而准确度会有一定的提高,因此可以将瞬时异常评论的影响在时间阈值的延展上体现出来,对于发现情感评分异常的微博评论有着更好的效果. 通过调整边界值范围和时间阈值范围可以限定异常检测的灵敏等级,当边界值范围增大时,对于异常评论的检测范围扩大,容忍度下降,检测灵敏度高,嫌疑数会相应增多;当时间阈值扩大时,容忍度提高,检测灵敏度较低,嫌疑数会相应减少.

表 3 ("老虎伤人")/《你好乔安》的实验数据

Tab.3 Experiment results of the "tiger accident" and *Hello*, *Joan*

情感评分均值	边界值	时间阈值	嫌疑数	恶意评论
(0.987 12) 2.007 02	0~5	0.5	(31)176	(1)147
		1.0	(31)148	(1)97
		1.5	(31)132	(1)31
		2.0	(31)123	(1)61
		2.5	(31)91	(1)42
		5.0	(30)56	(30)30
		15.0	(30)11	(30)11
		30.0	(30)11	(30)11
(1.996 87) 4.315 88	0~10	0.5	(202)1 154	(196)1 059
		1.0	(35)742	(31)649
		1.5	(31)444	(31)399
		2.0	(31)264	(31)233
		2.5	(31)210	(31)187
		5.0	(31)55	(31)37
		15.0	(30)11	(30)11
		30.0	(30)31	(30)11
(3.4808 9) 6.961 96	0~15	0.5	(92)1 683	(31)1 640
		1.0	(92)947	(31)920
		1.5	(42)793	(31)551
		2.0	(42)544	(31)523
		2.5	(42)444	(31)426
		5.0	(39)119	(31)88
		15.0	(30)11	(30)11
		30.0	(30)11	(30)11
(4.823 73) 9.325 36	0~20	0.5	(210)1 705	(31)1 685
		1.0	(64)1 465	(31)1 445
		1.5	(46)1 161	(31)1 144
		2.0	(38)1 049	(31)1 033
		2.5	(36)846	(31)833
		5.0	(32)326	(31)246
		15.0	(30)54	(30)31
		30.0	(30)11	(30)11
(5.94644) 11.96778	0~25	0.5	(979)1 898	(910)1 886
		1.0	(172)1 472	(31)1 412
		1.5	(96)1 411	(31)1 352
		2.0	(49)1 154	(31)1 103
		2.5	(38)1 039	(31)990
		5.0	(34)475	(31)446
		15.0	(31)25	(31)12
		30.0	(30)11	(30)11

2.2 性能分析

本文从目标检测和区间检测两个角度进行了性能分析,分别为:1)攻击目标检测的性能分析.检测率 $DR = n / N$,其中 n 为算法检测出的针对目标 O 的攻击次数, N 为实际(本文中为人工标注结果)针对目标 O 的攻击次数. 从表 4 和表 5 可以看到,随

着时间阈值的增大，与人工标注结果相比准确率不断趋近，取得了相当好的性能. 2）可疑区间检测的性能分析. RR（reduction rate）= 1－可疑区间评分数量/整个区间评分数量，SRIR（suspicious rating inclusion rate）= 可疑区间中干扰评分的数量/整个区间中干扰评分的数量.

表 4　《你好乔安》的检测率、RR 值、SPIR 值对比分析

Tab.4　Comparison analysis of detection rate, RR value, and SPIR value of *Hello*, *Joan*

时间阈值/D	检测率/%	RR/%	SRIR /%
0.5	14.3	67.4	67.4
1.0	15.2	76.3	76.6
1.5	22.9	80.8	80.8
2.0	24.2	84.5	84.8
2.5	25.9	87.2	87.2
5.0	48.8	94.9	94.9
15.0	97.5	98.9	98.9
30.0	98.7	99.2	99.2

表 5　“老虎伤人”的检测率、RR 值、SPIR 值对比分析

Tab.5　Comparison analysis of detection rate, RR value and SPIR value of the “tiger accident”

时间阈值/d	检测率/%	RR/%	SRIR/%
0.5	79.3	92.4	92.4
1.0	83.3	98.0	98.0
1.5	83.3	98.7	98.7
2.0	83.3	99.0	99.0
2.5	83.3	99.1	99.1
5.0	99.1	99.2	99.1
15.0	99.2	99.2	99.2
30.0	99.3	99.3	99.3

由表 4、5 可见，当时间阈值为 30 时，随着时间推移，评论数目的稳定性增强，因此在较长的时间范围内可以更加准确地检测出异常评论. 时间范围的增大可以有效消除瞬时评论对于整体均值的影响，得到更为全面的检测结果.

除此以外，对于经过异常检测后的评论评分与过滤前的评分进行了整体性能评价，公式如下所示.

RRO（recovered reputation offset）= $Z - Y$，其中 Y 为目标的真实评分，Z 为经过 SATAUCA 过滤干扰评分后评分的均值. 由图 4、5 可以看到时间阈值的增大可以使得性能指标收敛，对于不同均值和峰值的评论可以消除带来的影响，使得算法具有更好的适应性和鲁棒性.

2.3　对比试验

在基于文本层面的挖掘上，常用的有根据 Twitter 消息的浅层文本特征、行为特征和元素特征，构建多个贝叶斯分类器和集成分类器以识别 Twitter 中的谣言（identifying misinformation）等[18]方法. Feng 等[19]提出使用句法的虚假评论检测方法 CFG，在该方法中，为评论提取了不同的属性，针对评论文本的句法结构特点，利用语义树来发现评论中的隐藏信息. 本文选取该类方法中的基于文本的判断方法（content-based）和 Feng 等提出的 CFG 方法作为对比，在采集到的全部 20 个微博热点话题下的全部评论中进行测试，本文方法的准确度为 0.957，精确度为 0.962，召回率为 0.959，均高于文献[18-19]所提出的方法. Feng 等通过研究商品评论文本的特点，探索深层次的句法结构，定义了一些常规和非常规的句式来识别语义，随着定义特征的增多并加入到算法中，因此该方法对于评论文本的分析性能较传统的 IM 方法有了较大提高，精确度为 3 个方法中最高. 本文提出的 SATAUCA 方法对于微博场景中的干扰评论检测效果有一定的提高，以较少的特征值在准确率和召回率上取得了较好的效果，同时可以提供在不同时间上的检测结果，充分考虑到在微博评论中初期评论较热烈时的分布不均匀和稀疏特征，利用时间阈值来调节微博评论趋势的发现结果.

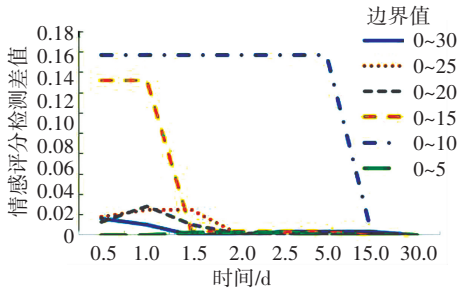

图 4　“老虎伤人”整体性能分析

Fig.4　Overall performance analysis of the “tiger accident”

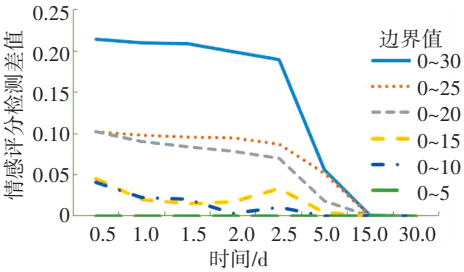

图 5　《你好乔安》整体性能分析

Fig.5　Overall performance analysis of *Hello*, *Joan*

3　结　论

1）可以在微博评论场景中，检测出干扰评论，取得了一定的效果，同时由于加入了时间条件，可以更为准确的获得某一时刻的情感分布变化.

2）该方法利用情感评分，对于评论文本进行较为准确的情感分类，然后通过调整边界值范围和时间阈值范围来限定异常检测等级，当边界值范围增大时，

对于异常评论的检测范围扩大,容忍度下降,检测灵敏度高;当时间阈值扩大时,容忍度提高,检测灵敏度较低;适当的选择边界值和时间阈值,可以有效提高与正常评论行为相似的异常评论识别准确率.

3)下一步,可以建立基于评论用户特性的社交图谱,通过积累用户行为、标签属性等,增加用户特写的维度,这样对于计算性能的要求较高,需要在今后的工作中对于建立高性能稳定的数据存取平台进行尝试,为水军用户检测的确定作进一步的工作.

参考文献

[1] GUPTA M, ZHAO P, HAN J. Evaluating event credibility on twitter [C]//Proceedings of the 2012 SIAM International Conference on Data Mining. Anaheim:Society for Industrial and Applied Mathematics, 2012:153. DOI: 10.1137/1.9781611972825.14

[2] RAVIKUMAR S, BALAKRISHNAN R, KAMBHAMPATI S. Ranking tweets considering trust and relevance[C]//Proceedings of the Ninth International Workshop on Information Integration on the Web. Scottsdale:ACM, 2012: 4. DOI:10.1145/2331801. 2331805

[3] 谢丽星,周明,孙茂松.基于层次结构的多策略中文微博情感分析和特征抽取[J].中文信息学报,2012,26(1): 73.DOI: 10.3969/j.issn.1003-0077.2012.01.011
XIE Lixing, ZHOU Ming, SUN Maosong. Hierarchical structure based hybrid approach to sentiment analysis of Chinese microblog and its feature extraction[J]. Journal of Chinese information processing, 2012, 26(1):73.DOI: 10.3969/j.issn.1003-0077.2012.01.011

[4] 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.DOI: 10.3969/j.issn.1000-3428.2012.13.046
PANG Lei, LI Shoushan, ZHOU Guodong. Sentiment classification method of Chinese microblog based on emotionalknowledge [J]. Computer Engineering, 2012, 38(13):156. DOI: 10.3969/j.issn.1000-3428.2012.13.046

[5] 程晓涛, 刘彩霞, 刘树新. 基于关系图特征的微博水军发现方法[J]. 自动化学报, 2015, 41(9):1533. DOI: 10.16383/j.aas.2015.c140906
CHENG Xiaotao, LIU Caixia, LIU Shuxin. Graph-based features for identifying spammers in microblog networks[J]. ACTA automatic sinica, 2015, 41(9):1533. DOI: 10.16383/j.aas.2015.c140906

[6] 刘鸿宇, 赵妍妍, 秦兵,等. 评价对象抽取及其倾向性分析[J]. 中文信息学报, 2010, 24(1):84. DOI: 10.3969/j.issn.1003-0077.2010.01.015
LIU Hongyu, ZHAO Yanyan, QIN Bin, et al. Comment target extraction and sentiment classification[J]. Journal of Chinese Information Processing , 2010, 24(1):84. DOI: 10.3969/j.issn.1003-0077.2010.01.015

[7] JINDAL N, LIU B, LIM E P. Finding unusual review patterns using unexpected rules [C]// ACM Conference on Information and Knowledge Management. Toronto: CIKM, 2010:1549. DOI: 10.1145/1871437.1871669

[8] 毛佳昕, 刘奕群, 张敏, 等. 基于用户行为的微博用户社会影响力分析[J]. 计算机学报, 2014, 37(4): 791. DOI: 10.3724/SP.J.1016.2014.00791
MAO Jiaxin, LIU Yiqun, ZHANG Min, et al. Social influence analysis for microblog user based on user behavior[J]. Chinese journal of computers, 2014,37(4):791. DOI: 10.3724/SP.J.1016.2014.00791

[9] MILLER Z, DICKISON B, DEITRICK W, et al. Twitter spammer detection using data stream clustering[J]. Information Sciences, 2014, 260(1):64. DOI: 10.1016/j.ins.2013.11.016

[10] HEYDARI A, TAVAKOLI M, SALIM N. Detection of fake opinions using time series[J]. Expert Systems with Applications, 2016, 58: 83. DOI: 10.1016/j.eswa.2016.03.020

[11] WU F, SHU J, HUANG Y, et al. Co-detecting social spammers and spam messages in microblogging via exploiting social contexts [J]. Neurocomputing, 2016, 201:51. DOI: 10.1016/j.neucom.2016.03.036

[12] 陆浩, 牛振东, 张楠,等. 基于句法与主题扩展的中文微博情感倾向性分析模型[J]. 北京理工大学学报, 2014,34(8):824. DOI: 10.15918/j.tbit1001-0645.2014.08.031
LU Hao, NIU Zhendong, ZHANG Nan, et al. A model for sentiment classification of Chinese microblog based onparsing and theme extension[J]. Transactions of Beijing institute of technology, 2014, 34(8):824. DOI: 10.15918/j.tbit1001-0645.2014.08.031

[13] 朱玺,董喜双,关毅,等.基于半监督学习的微博情感倾向性分析[J].山东大学学报(理学版),2014,49(11) : 37. DOI: 10.6040/j.issn.1671-9352.3.2014.136
ZHU Xi, DONG XiShuang, GUAN Yi, et al. Sentiment analysis of Chinese microblog based on semi-supervised learning[J]. Journal of Shandong University(Natural Science), 2014, 49(11):37. DOI: 10.6040/j.issn.1671-9352.3.2014.136

[14] 董振宇. 知网发布“情感分析用词语集(beta 版)”[EB/OL]. (2007-10-22)[2017-06-28].http://www.keenage.com/html/c_bulletin_2007.htm

[15] KU L W, CHEN H H. Mining opinions from the web: beyond relevance retrieval[J]. Journal of American Society for Information Science and Technology, Special Issue on Mining Web Resources for Enhancing Information Retrieval, 2007, 58(12):1838. DOI: 10.1002/asi.20630

[16] 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. DOI: 10.3969/j.issn.1000-0135.2008.02.004
XU Linhong, LIN Hongfei, PAN Yu, et al. Constructing the affective lexion ontology[J]. Journal of the China society for scientific and technical information, 2008, 27(2):180. DOI: 10.3969/j.issn.1000-0135.2008.02.004

[17] LIU Y, SUN Y. Anomaly detection in feedback-based reputation systems through temporal and correlation analysis[C]//Social Computing (SocialCom), 2010 IEEE Second International Conference on. Minneapolis:IEEE, 2010: 65. DOI: 10.1109/SocialCom.2010.19

[18] QAZVINIAN V, ROSENGREN E, RADEV D R, et al. Rumor has it: identifying misinformation in microblogs[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for computational Linguistics, 2011: 1589.

[19] FENG S, BANERJEE R, CHOI Y. Syntactic stylometry for deception detection[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers—Volume 2. Jeju Island: Association for Computational Linguistics, 2012: 171. DOI: 10.1017/CBO9781107415324.004

(编辑 魏希柱)