基于评论异常度的新浪微博谣言识别方法

张仰森 [1] 彭媛媛 [1] 段宇翔 [1] 郑 佳 [1] 尤建清 [1]

摘 要 以微博为代表的社交媒体在为公众提供信息共享平台的同时, 也为谣言提供了可乘之机. 开展微博中谣言的识别和清理方法研究, 对维护社会的安全稳定有着重要的现实意义. 本文针对新浪微博平台中谣言识别的问题, 提出了一种基于评论异常度的微博谣言识别方法. 首先采用 D-S 理论实现微博评论异常度的计算方法; 然后利用评论异常度与微博的内容特征、传播特征、用户特征对微博进行抽象表示; 最后再利用 SVM (Support vector machine) 构建一个基于评论异常度的谣言识别模型, 实现对新浪微博中谣言微博的识别. 实验表明, 本文提出的谣言识别模型对新浪微博中谣言识别具有较好的效果, 谣言微博识别的 F1 值达到了 96.2 %, 相较于现有文献的最好结果提高了 1.3 %.

关键词 谣言识别, 新浪微博, 评论异常度, D-S 理论, SVM

引用格式 张仰森, 彭媛媛, 段宇翔, 郑佳, 尤建清. 基于评论异常度的新浪微博谣言识别方法. 自动化学报, 2020, **46**(8): 1689−1702

DOI 10.16383/j.aas.c180444

The Method of Sina Weibo Rumor Detecting Based on Comment Abnormality

ZHANG Yang-Sen[1] PENG Yuan-Yuan[1] DUAN Yu-Xiang[1] ZHENG Jia[1] YOU Jian-Qing[1]

Abstract Microblog plays an important role in social network service, while providing an information communication platform for users, it also provides a loophole for rumors. It is of great practical significance to automatically detect and clean up rumors in microblogs for the security and stability of society. In this paper, a rumor detecting method based on comment abnormality is presented. Firstly, we use D-S theory to implement the calculation method of comment abnormality. And then, we combine the comment abnormality, text features, propagation features and user characteristics to abstractly represent Sina Weibo. Finally, we use SVM (Support vector machine) to build a rumor detecting model based on comment abnormality. The experimental results show that the rumor detecting model proposed can effectively improve the detecting performance. And the F-measure of the rumor detecting is up to 96.2 %, which is up by 1.3 % compared with the best value in other literatures.

Key words Rumor detecting, Sina Weibo, comment abnormality, D-S theory, SVM

Citation Zhang Yang-Sen, Peng Yuan-Yuan, Duan Yu-Xiang, Zheng Jia, You Jian-Qing. The method of Sina Weibo rumor detecting based on comment abnormality. *Acta Automatica Sinica*, 2020, **46**(8): 1689−1702

微博作为一种新兴开放的社交媒体, 凭借文本短小、传播迅速、操作灵活等特点, 迅速成为人们发布、传播和共享信息的重要传播媒介, 以极快速度影响着社会的信息传播格局. 同时, 微博的低门槛性使得微博用户的类型跨度较大, 不仅包括各种官方媒体、权威机构、知名人士, 还包括一些普通平民用户. 这些特性使微博成为社会各种话题的聚集地, 并发展成为重要的舆论载体和各类谣言发布与传播的温床.

目前, 关于 “谣言” 社会上并没有一个公认的定义. 虽然在现代汉语词典中, 谣言被解释为没有事实根据的消息, 但在现实世界中, 有很多谣言却是基于一定的事实编造出来的, 只是对事实进行了引申、歪曲、夸大甚至捏造. 为此, 本文将要讨论的谣言是那些凭空捏造、没有事实根据或虽有一定事实根据, 却由发布者进行了扭曲或夸大, 偏离了事实真相的言论. 微博谣言则是指在微博这个特定社交媒体中传播的那些凭空捏造或扭曲事实真相的言论. 微博平台中谣言的泛滥会给人们的日常生活和社会安稳造成极端恶劣的影响. 例如 2011 年 3 月, “碘盐可防辐射” 就是有人恶意夸大或歪曲碘盐作用而产生的谣言, 导致了一场全国性的 “抢盐风潮”; 2017 年 5 月, “肉松由棉花制作” 这个凭空捏造的谣言导致了人们的食品安全恐慌; 2017 年 11 月 “红黄蓝” 事件中 “老虎团” 信息则是某些捕风捉影造成的谣言, 引发了民众对军队的不信任感. 因此, 研究微博谣言的

收稿日期 2018-06-24 录用日期 2018-09-12
Manuscript received June 24, 2018; accepted September 12, 2018
国家自然科学基金 (61772081), 北京市教委科研计划 (KM201711232022) 资助
Supported by National Natural Science Foundation of China (61772081), Science and Technology Development Project of Beijing Municipal Education Commission (KM201711232022)
本文责任编委 张民
Recommended by Associate Editor ZHANG Min
1. 北京信息科技大学智能信息处理研究所 北京 100101
1. Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing 100101

识别方法, 对阻止谣言传播、维护社会稳定具有重要的现实意义.

1 相关研究

微博谣言的分析与识别已经成为互联网内容安全领域的热门研究方向之一, 其主要研究方法是提取微博中的各种特征, 并基于这些特征对微博是否为谣言进行判定. 微博谣言的分析与识别最早起源于对 Twitter 中文本的分析与挖掘, Qazvinian 等[1] 综合考虑了 Twitter 中 Tweet (推文) 的浅层文本特征、话题特征、网络行为特征, 构建了一个集成分类器, 用以判别当前 Tweet 是否属于谣言, 在其构建的包含 2 797 条普通 Tweet 和 3 803 条疑似谣言的数据集中, 召回率达到了 89.7 %. Takahashi 等[2] 对日本地震海啸后在 Twitter 中引发的谣言传播进行分析, 发现谣言的爆发时间点、相关 Tweet 的转发率和相关 Tweet 词语的使用差异性对谣言的传播具有显著的影响, 并据此构造了一个谣言检测系统. Castillo 等[3] 构建了 4 个维度的谣言评价体系, 其中有微博文本的特征、用户信息的特征、话题的特征和消息传播的特征, 并采用决策树分类方法, 对 Twitter 中话题的可信度进行分级, 并将可信度最高的话题视为普通话题, 其他等级的话题视为谣言, F1 值达到 86 %. Suzuki 等[4] 利用 Tweet 在转发时原始 Tweet 被保留或删除的特点, 计算转发过程中原始 Tweet 的保留率, 以此来评估 Tweet 信息的可信度, 实现谣言的识别. Ma 等[5] 利用循环神经网络对 Twitter 话题消息在传播过程中随时间变化的特征进行分析, 实现了 Twitter 话题消息的谣言识别, 该模型在 Twitter 数据集上 F1 值达到了 86 %, 在新浪微博数据集上 F1 值为 90.6 %.

对于中文微博谣言的识别, 由于新浪微博的结构以及汉语表达方式都与 Twitter 存在很大不同, 使得面向 Twitter 谣言识别的方法很难直接应用于中文微博的谣言识别. 国内针对中文微博开展了大量相关研究, 并取得了很多成果. 新浪微博官方于 2010 年 11 月启动微博辟谣工作, 主要针对新浪微博中存在的不实信息进行查证和辟谣, 但通过对官方辟谣账号和社区管理平台分析发现, 新浪微博的官方辟谣主要依靠工作人员、网络警察, 并结合用户举报, 人工对被举报微博进行筛选和查证, 这种方法虽然准确率很高, 但存在实用性较差、时效性不足等缺点. Yang 等[6] 在 Twitter 谣言识别基础上, 对新浪微博中的谣言进行分析, 总结了基于内容、传播和用户 3 个方面的 17 类特征, 还提出了微博发布客户端和微博发布地址两类新特征, 并利用 SVM 分类器进行谣言识别, 文中利用新浪官方辟谣平台构建了数据集进行实验, 证明其提出的新特征组合能使谣言识别准确率达到 78.7 %; 高明霞等[7] 针对中文微博的特点, 系统梳理了中文微博信息可信度的测量指标, 建立了针对文本内容、信息发布者和信息传播方式三方面特征的度量方法, 并利用多维证据理论进行特征融合, 构建了中文微博可信度评估框架 CCM-IF, 辅助谣言微博的识别和垃圾微博过滤; Wu 等[8] 依据微博在转发过程中的话题类型、转发用户、转发时间和转发内容的情感等特点构建了微博传播树, 最后与微博用户和内容相结合, 利用分类的方法实现微博谣言的识别, 并在其构建的数据集中达到了 91.3 % 的准确率. 祖琳坤等[9] 在 Yang 的基础上, 新增了微博评论的情感倾向特征, 文中先将每一条微博评论分为正向情感和负向情感, 然后对微博评论的整体情感进行评估, 最后将情感特征与微博基础特征相结合, 使得谣言识别的准确率达到了 94.9 %.

综上可以看出, 微博涉及文本内容信息、用户背景信息、用户之间的互动信息, 因而在研究微博谣言识别时, 大多数学者都采用基于微博的内容特征、用户特征和传播特征进行分析. 传播特征主要应用了用户之间的转发和点赞特征, 对用户之间的相互评论特征利用较少, 尽管也有人使用了微博评论中的情感倾向特征, 但微博评论中的其他信息并未得到充分利用, 而这些信息对判定微博内容可信度具有一定的补充作用. Mendoza 等[10] 在对 Twitter 谣言进行分析时发现, 与普通新闻媒体相比, Twitter 社区中的谣言更容易受到质疑, 谣言的评论中包含了更多的评判性或辱骂性词语. 在新浪微博中, 谣言微博的评论与普通微博同样存在较大差异, 如实例图 1、图 2 所示. 图 1 展示了同一微博用户的普通微博与谣言微博的评论数, 可以看出, 普通微博的评论数目分别为 4 条、7 条和 12 条, 而谣言微博的评论数目高达 6 104 条; 图 2 展示了同一事件中谣言微博与普通微博评论内容的对比, 可以看出, 谣言微博的评论内容包含了其他用户对该条微博的批判和质疑. 通过对大量普通微博和谣言微博比较发现, 谣言微博往往会更吸引网民大量关注并参与讨论, 导致评论数量出现异常, 同时, 在评论中使用的词语更具有批判性和质疑性, 致使词语使用度和情感倾向性都会出现异常. 受此启发, 本文在考虑微博的内容、用户和传播特征的基础上, 将微博评论中的多维信息引入到微博谣言的识别中, 提出微博评论异常度的概念, 并将微博评论异常度与微博的内容、传播及用户特征相结合, 构建基于评论异常度的微博谣言识别模型, 从而实现新浪微博中谣言的识别.

2 基于评论异常度的谣言识别模型构建

微博评论信息对判定微博内容的关注度和可信

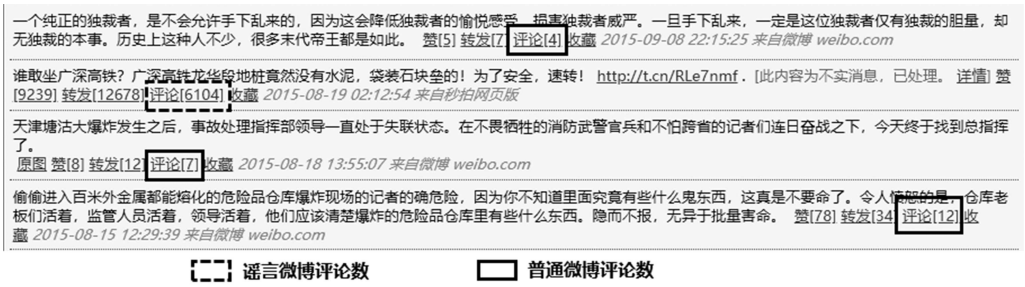


图 1　谣言微博与普通微博的评论数对比

Fig. 1　Comparison of the number of comments between rumor Weibo and ordinary Weibos

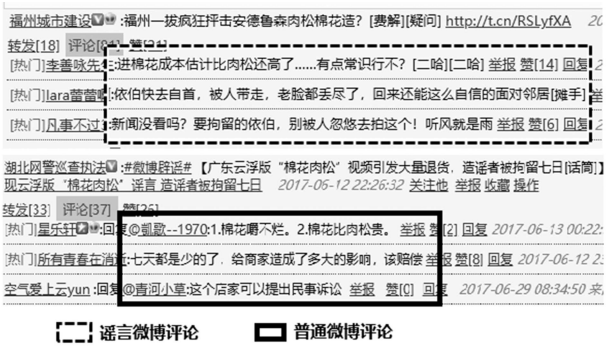


图 2　谣言微博与普通微博的评论文本对比

Fig. 2　Comparison of the comment texts between rumor Weibo and ordinary Weibo

度具有一定的补充作用. 例如评论的数量可以反映人们对微博内容的关注度, 评论的情感可以反映微博内容引导的情感, 评论的内容可以反映网民对微博内容的态度. 通过与普通微博比较发现, 谣言微博的内容多为捏造或违背事实的虚假言论, 人们无法对微博真伪进行评价时, 多被引发出负面情绪, 使用的质疑性、批判性的词语频次更多. 同时, 会引发更多人的参与讨论, 导致谣言评论数与发布者其他微博评论数相差较大. 也就是说, 微博评论的情感、用词和数量对谣言的识别具有一定的辅助作用. 本文意图利用评论的情感、用词和数量对谣言微博进行识别, 但考虑到新浪微博平台中微博评论状态的多样性, 若将上述信息作为 3 个独立特征引入到谣言识别模型, 可能会削弱这些评论信息对谣言识别的辅助作用. 因此, 本文考虑将这 3 个评论特征进行融合, 构建 “评论异常度” 特征, 用其对微博的评论特征进行整体刻画, 进而构建基于微博评论异常度的谣言识别模型. 下面将对微博基础特征的提取、评论异常度的计算以及基于评论异常度的谣言识别模型进行论述.

2.1　微博基础特征的提取

本文将微博的内容特征、传播特征和用户特征统称为微博的基础特征.

从微博内容的角度来看, 谣言微博为了体现内容的真实性, 常利用图片、视频和外部链接来辅助说明. 同时为了使其引起更多人注意、达到广泛传播的目的, 谣言微博还会参与多个话题[11], 包含较多的 “@” 符号, 并具有明显的感情倾向. 因此, 本文通过提取微博文本的长度、是否含有多媒体信息、情感倾向、“@” 数量和话题数量 5 个方面的特征来考察微博内容的真实性.

从微博传播的角度来看, 谣言微博因内容本身夸大、歪曲事实或毫无事实根据, 比普通微博更加耸人听闻, 也更容易引起他人的评论和转发, 表现为用户的参与度[12] 增加. Yang 等[6] 曾对新浪微博中的谣言微博进行统计分析, 发现大约有 71.8 % 的谣言微博是由非移动客户端发布. 因此, 本文通过提取微博的发布时间、发布平台和网民参与度来考察微博传播方面的可信度.

从微博用户的角度来看, 微博用户本身的权威度也在很大程度上影响着微博的真实性[13]. 例如, 微博中具有一定影响力的公众人物在发布微博时会比普通用户更加慎重, 发布的内容会具有更高的可靠性; 而粉丝量较少、没有影响力的普通用户发布微博时顾虑的因素少, 发布的内容可信度也相应较低. 因此, 本文通过提取用户认证情况、个人信息情况、用户的影响力、粉丝数目和发布微博数目 5 个方面的特征来考察微博用户的权威度.

综合微博的内容、传播和用户信息, 本文构建了微博谣言识别的基础特征体系, 共包含 3 个大类, 13 个小类的特征, 如表 1 所示.

微博以上的基础特征的提取方法, 在文献 [1, 3, 6, 8 – 9, 12, 14] 中已有详细研究, 本文不再进行论述. 下面对微博评论异常度的计算方法进行论述.

2.2　微博评论异常度的计算方法

微博评论信息可以从微博评论情感、用词和数量 3 个方面进行考虑. 通过上文分析, 当微博评论中负向情感越强、一些特殊用词越多、评论数量差异越大时, 微博为谣言的可能性越大. 因此, 本文分别定义 “评论情感异常度”、“评论用词异常度” 和 “评

表 1 微博谣言识别基础特征体系

Table 1 The basic feature system of Weibo rumor detecting

特征种类	特征名称	特征描述
微博内容特征	$Length$	微博文本的长度
	$Has\_Multimedia$	微博文本中是否含有多媒体信息, 如图片、视频和外部链接
	$Emotion\_Tendency$	微博的情感倾向, 分为正向情感和负向情感
	$Number\_of\_@$	微博文本中的 @ 数量
	$Number\_of\_topics$	微博文本参与的话题数量
微博传播特征	$Time\_Span$	微博发布时间与用户注册时间的间隔天数
	$Client\_Type$	发布微博的客户端类型, 包括移动客户端和非移动客户端
	$Participation$	网民参与度, 评论数和转发数两者之和与评论数、转发数和点赞数三者之和的比值
微博用户特征	Has_Verify	用户是否为认证用户
	$Has\_Description$	用户是否有自述信息
	$Influence$	用户影响力, 用户粉丝数与用户粉丝数和关注数两者之和的比值
	$Register\_Time$	用户的注册时间
	$Number\_of\_blogs$	用户的微博数量

论数量异常度" 来衡量微博评论中负向情感的强度、特殊用词的频率以及评论数量的差异性, 最后利用这 3 类评价信息计算微博评论异常度.

2.2.1 评论情感异常度

谣言发布者有意歪曲事实或捏造事实, 群众无法判别事情真伪时容易激发负面情感, 而不法分子也正利用这一特点故意煽动群众情绪. 因此, 本文将微博评论文本的情感作为衡量微博评论异常度的 1 个特征. 微博评论的情感可分为正向情感和负向情感两个类别, 通过分析微博评论中正 (负) 向情感倾向的评论数目, 可实现对微博评论情感异常度的计算. 对于单个评论文本情感倾向的分析, 考虑到评论文本一般为短文本且含有大量网络用语, 本文构建了基于微博情感词典的朴素贝叶斯情感分类模型.

1) 微博情感词典的构建

虽然现有的情感词典知识库, 例如 HowNet 情感词库、中文情感词汇本体库、《NTUSD》等都包含有丰富的情感词, 并划分了详细的情感倾向, 但是, 微博文本口语化较为严重, 且含有较多的微博表情符号和网络流行语, 使得现有的情感词典知识库很难直接用于微博评论文本的情感分析上. 因此, 本文在 HowNet 情感词库的基础上, 补充微博表情符号和网络流行语对情感词典进行扩建.

对于微博表情符号, 可通过微博官方 API 进行获取. 在新浪微博中, 微博的表情符号以短文本形式存储, 例如 "🙂" 存储为 "[微笑]", 因此, 表情符号的情感倾向可以根据其短文本内容并辅助于人工校正的方式确定. 本文通过对微博评论中常用的表情符号筛选, 共选取了 239 个微博表情符号加入到情感词典中.

对于网络流行语, 我们在 "网络用语词典网站"、"网词网" 上爬取了 1 142 条网络用语及其释义, 例如 "杯具: 原指盛水的器具, 后因与 '悲剧' 一词谐音, 成为 '悲剧' 的一种幽默的说法". 以 HowNet 情感词典为基准, 我们采用统计和人工相结合的方式选取了 1 095 条网络用语加入到情感词典.

2) 评论情感异常度计算

在对微博评论文本分词处理时, 将微博情感词典加入到用户词典当中, 以便情感词在分词时能作为整体被切分. 本文利用微博情感词典中的情感词作为评论文本的特征词, 这一特征词集合可表示 $D=\{w_1,w_2,\ldots,w_n\}$, 其中 w_i 为 D 的第 i 个特征词, n 为特征词的数目. 评论的情感倾向类别分为负面情感类别 C_0 和正面情感类别 C_1. 单个微博评论的情感倾向 C 的判定计算方法如式 (1) 所示.

$$C=\left\{C_i\,|\,C_i=\underset{C_i\in\{C_0,C_1\}}{\arg\max}\left(P(C_i)\prod_{j=1}^{n}P(w_j\,|\,C_i)^{wt(w_j)}\right)\right\} \tag{1}$$

其中, $P(C_i)$ 表示类别 C_i 的先验概率, $P(w_j\,|\,C_i)$ 表示特征词 w_j 在类别 C_i 情况下的后验概率, $wt(w_j)$ 为第 j 个特征词在待分析微博评论中的权重, 这里采用 BOOL 型权值, 即当特征词 w_j 出现在微博评论中时, $wt(w_j)$ 取值为 1, 否则为 0.

本文分别从普通微博评论和谣言微博评论抽取评论语料进行情感标注, 并利用标注语料计算出 $P(C_i)$ 和 $P(w_j\,|\,C_i)$. 为了避免某一特征词在训练样本中出现次数为 0, 导致 $P(w_j|C_i)=0$ 的情况, 本

文引入 Laplace 平滑, 如式 (2) 所示.

$$P(w_j|C_i)=\frac{Count(w_j,C_i)+\alpha}{\sum\limits_{k=1}^{n}Count(w_k,C_i)+1} \tag{2}$$

$$\alpha=\frac{1}{\sum\limits_{C_i\in\{C_0,C_1\}}\sum\limits_{k=1}^{n}Count(w_k,C_i)} \tag{3}$$

其中, $Count(w_j,C_i)$ 表示特征词 w_j 在情感类别为 C_i 的评论文本集合中出现的次数. 通过引入 Laplace 平滑, 当特征词 w_j 不存在于 C_i 类时, $P(w_j|C_i)$ 仍会有一个极小的存在概率; 当特征词存在时, 对原 $P(w_j|C_i)$ 影响较小.

利用所有单个评论的情感倾向计算微博评论的总体负向情感强度, 并将该负向情感强度作为微博评论的情感异常度 $Comment\_Emotion$, 如式 (4) 所示.

$$Comment\_Emotion=\frac{n_{C=C_1}-n_{C=C_0}}{N_{\text{Comment}}} \tag{4}$$

其中, $n_{C=C_0}$ 表示情感倾向为负的评论数, $n_{C=C_1}$ 表示情感倾向为正的评论数, N_{Comment} 为评论总数. $Comment\_Emotion$ 的取值范围为 $[-1,1]$ 之间, $Comment\_Emotion$ 越接近 1, 微博评论的正向情感越强, $Comment\_Emotion$ 越接近 -1, 微博评论的负向情感越强, 微博的状态也就越可疑.

通过对收集到的微博评论信息分析发现, 有 77% 的谣言微博的评论呈现负面情感, 有 18% 的普通微博的评论同样呈现负面情感, 如果单纯使用评论文本的情感异常度作为谣言微博的判断依据, 会为谣言识别带来一定误差. 因此, 在评论文本情感异常度的基础上, 引入评论用词异常度和评论数量异常度, 进一步丰富评论异常度的含义.

2.2.2　评论用词异常度

微博评论文本的用词情况可以反映出网民对微博内容的观点或意见. 通过对谣言微博的评论文本的分析发现, 表示批判、质疑的词语使用较为普遍, 例如 “求证”、“造谣”、“假新闻” 等. 因此, 我们可以从谣言微博评论中挖掘出常用关键词对评论的内容进行评估. 首先将采集到的大规模微博评论划分为谣言微博评论和普通微博评论两个类别, 在去除评论中出现的专有名词如人名、地名之后, 分别统计出两类微博评论中各词语的使用频次; 然后将谣言微博中词语频次与普通微博中的词语频次相减, 对出现在谣言微博和普通微博中的词语进行区分, 获取体现谣言微博特征的区分性词语集合, 采用相减的方法可以抽取到那些在谣言微博中出现频次很高但在普通微博中出现频次较低的词语; 最后, 在区分性词语集合中, 按照频次差值的大小, 选取 Top100 的词语作为谣言微博评论的区分性关键词, 为后面计算评论用词异常度奠定基础. 在所选取的区分性关键词中, Top10 的词语及其频次差值如图 3 所示.

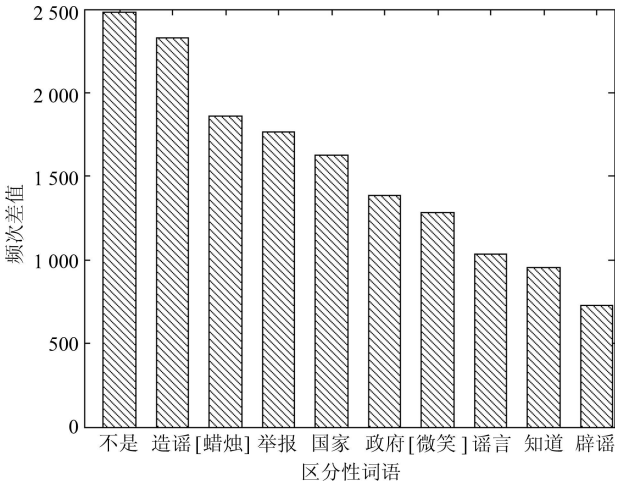


图 3　区分性词语频次差值

Fig. 3　The frequency differences of identified words

在所获得的谣言微博评论的区分性关键词集合中, 各个词在判定谣言微博的计算中所起的作用是不一样的, 有些明显表达质疑和反驳的词语, 对于判定谣言微博具有更大的权重. 为此, 我们基于 HowNet 挑选了 6 个在语义上能够直接反映微博谣言特性的词语, 构成显著区分性词集, 记为 $Zwords$. $Zwords=$ {造谣, 举报, 谣言, 辟谣, 不实, 传谣}.

从区别性关键词集合中减去 $Zwords$ 中的 6 个显著区分性词语, 得到含有 94 个关键词的集合 $Gwords$, 称为一般区分性词集. 所得谣言微博评论的显著区分性词集 $Zwords$ 与一般区分性词集 $Gwords$ 如表 2 所示.

表 2　谣言微博评论的区分性词集

Table 2　The identified word sets of rumor Weibo comments

类别	词集
$Zwords$	造谣, 举报, 谣言, 辟谣, 不实, 传谣
$Gwords$	不是, [蜡烛], 国家, 政府, [微笑], 知道, 真相, 新闻, 脑子, 真的, 祈祷, 呵呵, 智障, 没有, 恶心, 消息, 是不是, 口德, 真是, 相信, 素质, 打死, 事实, 智商, 抵制, 他妈的, 怒骂, 证据, 老百姓, [吃惊], 新浪, 不要脸, 证实, 脑残, [拜拜], 垃圾, 可怕, 小心, 尼玛, 传播, 暴力, 难道, 神经病, 法律, 公道, 记者, 媒体, 赶紧, 去死吧, 真假, 可能, 删除, 网警, 乱说, 不信, 打脸, 假新闻, 眼球, 国人, 键盘, 官方, 人性, 理智, 良心, 明显, 所谓, 民众, 不用, 无辜, 底线, 言论, 该死, 肯定, 水军, 真的假, 遭报应, 有意思, 侮辱, 生命, 央视, 闭嘴, 活该, 愤怒, 确定, 喷子, [怒], 煽动, 真实, 常识, 骂人, 缺德, 鄙视, 无知, 不删

下面将通过 $Zwords$ 和 $Gwords$ 中的词语来计算微博的评论用词异常度. 由于随着时间的推移, 微博评论者对微博中所描述的事实掌握或了解的越来越多, 其在评论中提出的质疑或反驳的可信度越来越高, 显著区分性集合 $Zwords$ 中的词语若出现在评论中, 将随着评论时间推移, 其权重应该相应增大[15]. 为此, 可计算评论中区分性词语的使用频率, 即评论用词异常度的计算公式如式 (5) 所示.

$$Comment\_Keyword = \sum_{gword \in Gwords} \frac{n_{gword}}{N_{words}} + \sum_{zword \in Zwords} \frac{\beta n_{zword}}{N_{words}} \tag{5}$$

其中, N_{words} 为当前微博评论的总词数, n_{gword} 为 $Gwords$ 词集中的词语 $gword$ 在评论中出现的总频次, n_{zword} 为 $Zwords$ 词集中的词语 $zword$ 在所有评论中出现的总频次, β 为 $zword$ 的影响因子, 其计算方法如式 (6) 所示.

$$\beta = \frac{t_{zword} - t_{\min}}{t_{\max} - t_{\min}} + 1 \tag{6}$$

其中, t_{zword} 为包含词语 $zword$ 的所有评论中最新评论发布的时间, $t_{\min}$ 为微博发布后第一条评论发布的时间, $t_{\max}$ 为微博发布后最新评论发布的时间. $Comment\_Keyword$ 越大, 代表微博评论使用的区分性词频率越高, 微博的状态越可疑.

2.2.3 评论数量异常度

相较于普通微博, 谣言微博在传播的过程中会引起大量用户的参与和讨论, 对于发布者而言, 其谣言评论的数目相较于普通微博有较大的差异, 图 4 展示的为 5 位被证实发布过谣言的微博用户, 其发布的谣言微博的平均评论数和发布谣言微博之前 10 天内发布的普通微博的平均评论数的对比. 通过实例可以发现, 同一用户的谣言微博平均评论数远大于其普通微博. 因此, 可以将微博的评论数与该用户发布的其他微博评论数之间的差异程度作为特征用于衡量微博评论的异常度.

为确保用户发布的其他微博评论数的可靠性, 本文构建一个滑动时间窗, 其长度为待分析微博发布时间的前一个月, 计算该用户当前待分析微博与其滑动时间窗内所有普通微博平均评论数之间的差异程度, 即评论数量异常度如式 (7) 所示.

$$Comment\_Number = \log_2 \left(\frac{|CCount_{\text{curr}} - \frac{1}{N_{Weibo}} \sum_{i=1}^{N_{Weibo}} CCount_i|}{\min(CCount_{\text{curr}}, \frac{1}{N_{Weibo}} \sum_{i=1}^{N_{Weibo}} CCount_i)} \right) \tag{7}$$

其中, $CCount_{\text{curr}}$ 代表当前待分析微博的评论数, N_{Weibo} 代表用户滑动时间窗内所发布的普通微博总数, $CCount_i$ 代表滑动时间窗内第 i 条微博的评论数量. $Comment\_Number$ 越大, 待分析微博的评论状态越可疑.

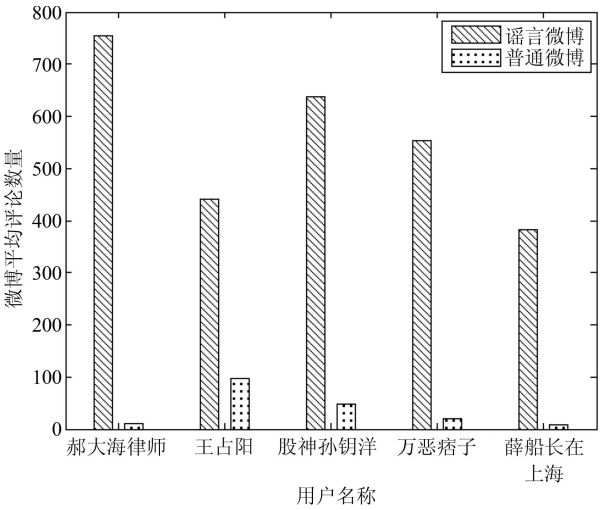


图 4 用户的普通微博与谣言微博平均评论数对比

Fig. 4 Comparison of the average number of comments between rumor Weibo and ordinary Weibo for some users

2.2.4 基于 D-S 理论的微博评论异常度计算

本文将微博的评论异常度特征作为微博谣言识别中引入的一个新特征. 但由于微博评论的多样性、不规范性, 使得评论的状态受到多个因素的影响, 对评论是否属于异常状态的评估处于一种 “不确定” 的状态, 对评论异常度的计算具有一定的模糊性. 如果依靠某一特征或直接将多个特征进行加权计算, 很难准确地评估微博评论是否异常. 因此, 本文将处理不确定性推理具有较强优势的 D-S 理论[16−17] 引入到评论异常度的计算中来, 将评论情感异常度、用词异常度和数量异常度融合, 构造评论异常度计算模型. 该模型包括评估框架的构建、信任分配函数的确定和证据的加权合成.

1) 评估框架的构建

评估框架是所解决问题的状态空间集合及其推理的证据体系, 本文需要解决的问题是对微博的评论状态进行评估, 其状态可分为正常状态和异常状态两种, 微博评论属于异常状态的概率即为微博评论异常度. 因此, 本文定义微博评论异

常度评估框架 $\Theta = \{A, N\}$, 其中 $A(Abnormal)$ 代表 "微博评论异常的状态", $N(Normal)$ 代表 "微博评论正常的状态", 并且 $A \cap N = \emptyset$. 在评论异常度的评估中, 主要考虑评论的情感异常度、用词异常度和数量异常度, 因此构建证据三元组 $E(CE, CK, CN)$, 其中 CE、CK、CN 表示评论情感、评论用词和评论数量, 以 $E_x(v(CE), v(CK), v(CN))$ 表示微博 x 证据三元组的取值, $v(CE)$、$v(CK)$、$v(CN)$ 取值分别为微博 x 的 $Comment\_Emotion$、$Comment\_Keyword$ 和 $Comment\_Number$.

2) 信任分配函数的确定

D-S 理论中, 构建基于证据理论的不确定推理的初始信任分配函数是十分关键的一步, 但这一问题尚未有一个统一的理论和方法, 目前大多应用采用的是专家直接指定的方法. 鉴于此, 本文利用训练集构建了一个包含相同数量谣言微博和普通微博的语料库, 利用隶属度函数来确定各个证据的初始信任分配值. 隶属度函数是模糊集合中的一个概念, 用来描述论域中某一确定元素属于模糊集合的概率. 本文中的证据三元组 $E(CE, CK, CN)$ 为论域, 三元组中的每个证据值是一个确定元素, 利用隶属度函数可以对当前微博评论的异常状态进行概率描述, 可以较好地处理信任分配函数中主观判断的问题. 具体过程如下:

步骤 1. 计算语料库中每条微博 x 的证据值, 并分别取语料库中每个证据中的最大值 $v_{\max}(E_i)$ 和最小值 $v_{\min}(E_i)$, 其中 $E_i \in \{CE, CK, CN\}$.

步骤 2. 将证据 E_i 的取值范围 $[v_{\min}(E_i), v_{\max}(E_i)]$ 划分为 n 个区间, 其中第 j 个区间表示为 $[v_{j-1}, v_j]$ $(1 \le j \le n)$. 统计语料库中证据 E_i 取值处于第 j 个区间的谣言微博数量 N_r 和普通微博数量 N_n, 计算 N_r 与 $(N_r + N_n)$ 的比值 $u_j(E_i)$, 作为证据 E_i 取值为第 j 个区间时的隶属度即微博评论属于异常状态的概率. 依次计算证据 E_i 在 n 个区间内的隶属度 $u_1(E_i), u_2(E_i), \cdots, u_n(E_i)$.

步骤 3. 取证据 n 个区间的中值 $v_{\text{mid}1}(E_i)$, $v_{\text{mid}2}(E_i), \cdots, v_{\text{mid}n}(E_i)$, 利用它们与相应区间的隶属度求出证据值与隶属度的定量关系, 即构造出证据 E_i 的隶属度函数 $F(v(E_i))$. 当证据 E_i 取值为 $v(E_i)$ 时, 评论状态 S_{E_i} 属于异常状态的概率为 $F(v(E_i))$, 证据 E_i 的信任分配函数如式 (8) 所示.

$$m_{E_i}(S_{E_i}) = \begin{cases} F(v(E_i)), & S_{E_i} = A \\ 1 - F(v(E_i)), & S_{E_i} = N \end{cases} \tag{8}$$

根据以上三步, 即可获取微博评论、用词和数量 3 个证据的基本信任分配.

3) 证据的加权合成规则

鉴于传统的证据理论中, Dempster 合成规则为了保持基本概率分配函数的归一性, 忽略了证据之间可能产生的冲突, 导致证据合成的结果与实际常理相悖[18]. 为此, 本文采用加权分配的证据合成法[19] 对证据进行合成. 该改进方法在证据合成前, 会依据证据之间的相关性为每个证据分配权重, 再依据权重信息对证据的信任分配进行更新, 最后将更新的证据进行融合, 具体过程如下:

证据的权重是由该证据与其他证据之间的相似系数来评估的, 如果当前证据与其他证据的相似性越高, 说明该证据越能受到其他证据的支持, 可信度越高, 权值也就越大. 两个证据之间的相似系数定义如下: 设 E_i 和 E_j 为上述证据三元组 E 中的两个证据, 其初始信任分配函数为 $m_{E_i}(S_{E_i})$ 和 $m_{E_j}(S_{E_j})$, 焦元分别为 S_{E_i} 和 $S_{E_j}(S_{E_i}, S_{E_j} \in \{A, N\})$, E_i 和 E_j 的相似度 d_{ij} 计算如式 (9) 所示, 其中, 任意两个证据的相似度是对称相等的, 即 $d_{ij} = d_{ji}$, 当 $i = j$ 时, 相似度取值为 1.

$$d_{ij} = d_{ji} = \frac{\sum\limits_{S_{E_i} \cap S_{E_j} \neq \emptyset} m_{E_i}(S_{E_i}) m_{E_j}(S_{E_j})}{\sqrt{(\sum m_{E_i}^2(S_{E_i}))(\sum m_{E_j}^2(S_{E_j}))}} \tag{9}$$

证据的相似系数 d_{ij} 取值范围为 $[0, 1]$, d_{ij} 越接近于 1, 表示两个证据相似性越高, 相互支持度越高; d_{ij} 越接近于 0, 表示两个证据冲突性越高. 根据本文的证据三元组 $E(CE, CK, CN)$ 中任意两个证据之间的相似度系数, 构造的相似矩阵如式 (10) 所示.

$$\begin{bmatrix} 1 & d_{12} & d_{13} \\ d_{21} & 1 & d_{23} \\ d_{31} & d_{32} & 1 \end{bmatrix} \tag{10}$$

将相似矩阵的每行叠加可得到各个证据对当前证据的支持度如式 (11) 所示.

$$Sup(E_i) = \sum_{j=1}^{3} d_{ij} \tag{11}$$

将各个证据的支持度归一化处理, 即可得到各个证据的权值, 如式 (12) 所示.

$$Crd(E_i) = \frac{Sup(E_i)}{Sup(CE) + Sup(CK) + Sup(CN)} \tag{12}$$

将证据的支持度作为其在合成过程中的权值, 即证据 E_i 的支持度越高, 其在证据合成过程中权值也就越大; 证据 E_i 支持度越低, 其在证据合成过程中的权值也就越小.

依据权值可对证据进行加权合成, 获取证据 E_i 对微博评论状态为 S_{E_i} 时的平均加权证据信任分配

$m'_{E_i}(S_{E_i})$, 如式 (13) 所示.

$$\begin{aligned} m'_{E_i}(S_{E_i}) = &Crd(CE) \times m_{CE}(S_{CE}) + \\ &Crd(CK) \times m_{CK}(S_{CK}) + \\ &Crd(CN) \times m_{CN}(S_{CN}) \end{aligned} \tag{13}$$

其中, $m_{CE}(S_{CE})$、$m_{CK}(S_{CK})$ 和 $m_{CN}(S_{CN})$ 分别为证据 CE、CK 和 CN 的基本信任分配.

获取 3 个证据的平均加权证据的信任分配后, 采用 D-S 合成规则对它们进行融合. 构建融合方法如式 (14) 所示.

$$P(S) = \frac{\sum\limits_{S_{E_i} \cap S_{E_j} = S} m'_{E_i}(S_{E_i}) m'_{E_j}(S_{E_j})}{1 - \sum\limits_{S_{E_i} \cap S_{E_j} = \emptyset} m'_{E_i}(S_{E_i}) m'_{E_j}(S_{E_j})} \tag{14}$$

上述公式中, $S, S_{E_i}, S_{E_j} \in \{A, N\}$, $P(S)$ 为两个证据 E_i 和 E_j 一次融合后得到的评论状态为 S 的概率. 由于本文中证据有 3 个, 需要进行 2 次融合, 即可得到微博评论状态为 A 或 N 时的概率 $P(A)$ 或 $P(N)$, 且 $P(A) + P(N) = 1$. $P(A)$ 即为微博评论状态异常的概率, 即为微博评论异常度 $Comment\_Abnormality$.

2.3 基于评论异常度的 SVM 谣言识别模型构建

在获取微博的基础特征和评论异常度特征后, 可利用这些特征对微博进行向量化表示, 然后再利用分类模型判定微博是否为谣言微博. 假设微博表示为 $Weibo$, 依据第 2.1 节和第 2.2 节的基础特征和评论异常度特征, 可将 $Weibo$ 抽象地表示为式 (15) 所示 $\boldsymbol{x}_{Weibo}$. 式中, $f_1, f_2, \cdots, f_{14}$ 为 $Weibo$ 的 14 个特征值.

$$\boldsymbol{x}_{Weibo} = (f_1; f_2; \cdots; f_{14}) \tag{15}$$

将 $\boldsymbol{x}_{Weibo}$ 归一化后作为分类模型的输入, 使模型的输出为微博类别的标签 T_{Weibo}, 如式 (16) 所示.

$$T_{Weibo} = \begin{cases} 1, & Weibo \text{ 为谣言微博} \\ 0, & Weibo \text{ 为普通微博} \end{cases} \tag{16}$$

2.3.1 分类模型的选择

在确定谣言识别模型的输入与输出后, 在众多的分类模型中选择合适的分类模型是构建谣言识别模型的关键. SVM 是一种典型的二值分类模型, 适用于小规模的训练样本, 泛化能力较强, 在文本分类领域已有广泛应用[1, 5−6, 9, 20]. 为此, 本文选择 SVM 作为谣言识别的分类模型. SVM 将样本空间中的分类问题形式化为一个求解凸二次规划的问题, 通过学习可以得到样本空间的最佳的分类超平面如式 (17) 所示.

$$\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{x} + b = 0 \tag{17}$$

其中, $\boldsymbol{w} = (w_1; w_2; \cdots; w_{14})$ 为法向量, 决定了超平面的方向, b 为位移项, 决定了超平面与原点的距离, $\boldsymbol{x} = (f_1; f_2; \cdots; f_{14})$ 为样本点的特征向量. 该超平面可使两类样本点与分类超平面之间的距离最大. 而在分类样本中, 总存在部分点是线性不可分的, 为此在求解最佳分类超平面时, 可对每个样本点引入一个松弛变量 ξ, 提升 SVM 训练过程中的容错性, 则求取问题的最优解可表示为式 (18).

$$\begin{aligned} &\min_{\boldsymbol{w}, b} \frac{\| \boldsymbol{w} \|^2}{2} + C \sum_{i=1}^{N} \xi_i \\ &\text{s.t. } T_i(\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \\ &\quad i = 1, 2, \cdots, N \end{aligned} \tag{18}$$

其中, C 为惩罚系数, ξ_i 为第 i 个样本点的松弛变量, $\boldsymbol{x}_i$ 为第 i 个样本点的特征向量, T_i 为第 i 个样本点的类别标签, N 为训练样本的个数. 根据拉格朗日对偶定理, 我们可通过求解该问题的对偶问题得到原问题的解.

2.3.2 SVM 模型中核函数的选择与参数确定

核函数的选择对于 SVM 的性能具有重要的影响. SVM 常用的核函数有线性核函数、多项式核函数和 RBF (Radial basis function) 核函数. 线性核函数是原始空间中内积运算, 当样本点在原始向量中已经线性可分并且特征数远大于样本数时, 能够取得不错的分类效果. 多项式核函数和 RBF 核函数都可将非线性分布的样本映射为线性可分, 且适用于特征数和样本数相当或特征数远小于样本数的情况, 但是 RBF 核函数要比多项式核函数的先验参数少, 模型更加简洁, 计算量更少. 本文中的微博被表示为拥有 14 维度的特征向量, 训练样本数量为 2 400, 特征数远小于样本数, 为此, 我们选择 RBF 核函数, 如式 (19) 所示.

$$K(\boldsymbol{x}, \boldsymbol{z}) = \exp(-\gamma \| \boldsymbol{x} - \boldsymbol{z} \|^2) \tag{19}$$

对 SVM 进行训练时, 需要确定两个模型参数: 惩罚系数 C 和 RBF 核函数中的 γ. 惩罚系数 C 控制了 SVM 的容错性, C 越大表明模型对错误的容忍度越小, 当 C 过高时易出现过拟合现象; C 越小则表明模型对错误的容忍度越高, 当 C 过小时又会影响模型的准确率. RBF 核函数中的参数 γ 影响着样本点映射到高维空间的分布, γ 越大, 模型在训练过程中应用到的支持向量越少, 当 γ 过大时, 易出现过拟合现象; 而 γ 越小, 利用到的支持向量越多, 当 γ 过小时, 同样会影响模型的准确率. 为此, 我们使用十字交叉验证策略来进行参数的选取. 该策略能够克服分类器的过拟合现象, 又能使模型具有较强的泛化性. 在训练前我们将训练集随机分为 10 个子

集, 每一次使用其中 9 个子集进行训练, 并在剩余 1 个子集上验证模型的准确率, 共进行 10 次训练, 得到 10 个分类模型. 最终选用准确率最接近于 10 个模型的平均准确率的模型为最优分类模型. 参数 C 和 γ 的取值通过调用 libSVM 工具包来实现, 具体的选取过程可参考文献 [21]. 经过实验, 我们得到的最优分类模型中, C 和 γ 取值分别为 2^{11} 和 2^{-5}.

微博谣言识别模型的整体构建过程如图 5 所示, 首先对训练集和测试集向量化表示; 然后利用训练数据对 SVM 进行训练; 最后利用训练好的 SVM 对测试数据进行分类, 从而识别出测试数据中的谣言微博.

3　实验结果与分析

为了验证本文提出的微博评论异常度和基于评论异常度的微博谣言识别模型, 本文设计了相关实验进行验证, 主要包括 3 个部分:

1) 评论异常度参数的确定;

2) 微博评论异常度的评估;

3) 基于评论异常度的谣言识别模型的评估.

3.1　实验数据

由于微博谣言识别方面的研究尚未有一个公开、完整的数据集, 中文微博谣言识别的研究大多是从社交媒体平台开放的 API 自行收集数据进行实验. 因此, 本文依据新浪微博平台构建了一个新浪微博谣言识别研究的实验数据集, 包括谣言微博和普通微博, 其中每条微博都包含微博内容、转发数、评论数、点赞数、全部评论文本和微博发布者个人资料、关注数、粉丝数. 谣言微博数据来源于新浪微博社区管理中心不实信息举报中的结果公示板块[1], 这一板块的微博内容已被新浪微博官方证实为谣言. 本文从中爬取了 2015 年 5 月 19 日至 2017 年 5 月 19 日的谣言微博共 9 406 条, 为确保所采集的谣言微博数据有一定的评论特征, 我们抽取了原微博未被删除且评论数量大于 100 的谣言微博数据共 1 568 条, 评论文本 50 万余条. 普通微博数据来源于从新浪微博提供的 API 随机爬取的同一时期的微博. 在真实环境中, 谣言微博的数量要远远小于普通微博的数量, 为了避免分类器将实验数据全部分为普通微博以获得高准确率的情况, 我们从普通微博中随机选取了评论数量大于 100 且文本不为纯符号、长度大于 10 的微博, 并通过人工校验删除可能为谣言的微博, 最终获得普通微博数据 1 600 条, 评论文本 43 万余条. 我们选取 1 200 条谣言微博和 1 200 条普通微博构成训练数据集用于评论异常度参数的确定和 SVM 分类器参数的确定, 剩余 368 条谣言微博和 400 条普通微博构成测试集用于微博谣言识别模型的验证.

3.2　评论异常度参数的确定

在微博评论异常度计算过程中, 本文将利用训练数据集构造 CE、CK 和 CN 3 个证据的隶属度函数, 进而得到它们的信任分配函数.

首先, 依据训练数据集分别获取 CE、CK 和 CN 3 个证据的取值范围 (如表 3 所示) ; 然后, 将它们的取值范围划分为 n 个区间, 通过训练数据集计算出每个区间的隶属度, 并将各区间的中值与区间相应隶属度作为一个数值对 $(v_{\mathrm{mid}}(E_i), u(E_i))$; 最后利用证据的 n 个数值对进行曲线拟合得到证据的隶属度函数. 在进行曲线拟合时, 我们观察发现, 3 个证据的 n 个数值对分布都与 sigmoid 函数类似, 当 $v(E_i)$ 取值越接近于两端时, $u(E_i)$ 也越接近于 0 或 1. 因此, 我们采用最小二乘曲线拟合算法, 按照式 (20) 对数值对进行拟合, 式中函数参数 a、c 和 d 通过 MATLAB 数学软件编程获取.

$$u(E_i) = \frac{a}{1 + \mathrm{e}^{-c(v(E_i)-d)}} \tag{20}$$

在对证据 CE、CK 和 CN 的值域进行区间划分时, 不同的划分方法得到的离散数值对的个数会

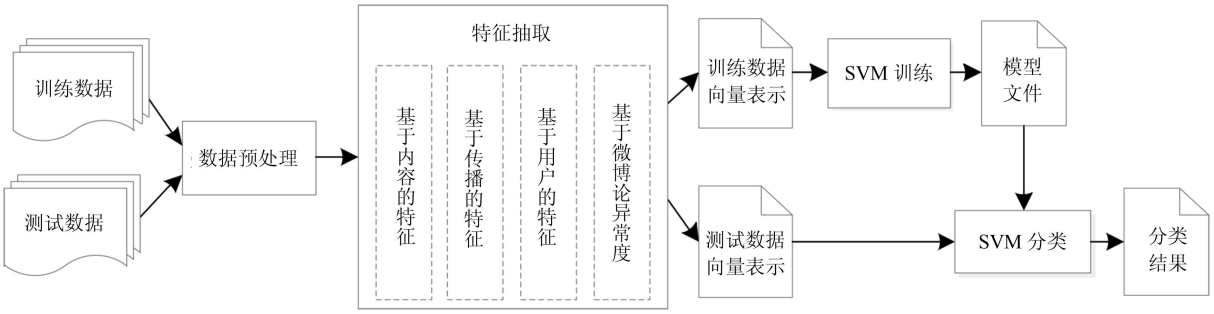


图 5　基于评论异常度的微博谣言识别模型

Fig. 5　Weibo rumor detecting model based on comment abnormality

[1]http://service.account.weibo.com/?type=5&status=4

表 3 CE、CK、CN 的取值范围

Table 3 The range of values of CE, CK and CN

证据	取值范围
CE	[−0.530, 0.782]
CK	[0, 0.178]
CN	[−7.245, 8.231]

不同, 所得到隶属度函数拟合效果也会不同. 为此, 本文在进行证据值域划分时, 对不同的区间个数 n 进行实验, 以发现较好的区间划分个数 n. 这里对 n 取值为 25、50、100 和 200 时进行曲线拟合实验, 利用误差平方和 SSE (Sum of the squared error) 对拟合效果进行评价, SSE 值越接近于 0, 曲线的拟合效果越好, SSE 随 n 值的变化情况如图 6 所示.

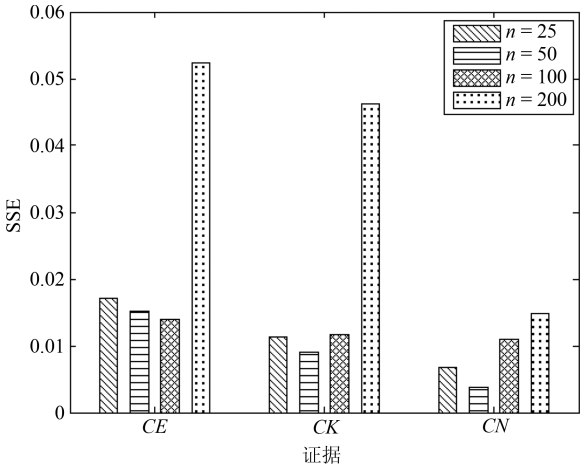


图 6 SSE 与区间个数 n 的关系

Fig. 6 Relationship between SSE and the interval number n

从图 6 可以看出, 证据 CE 的值域划分个数 $n = 100$ 时, 拟合效果最好, 所获得的拟合函数如式 (21) 所示, 函数参数分别为 $a = 0.9981$, $c = -13.5$, $d = 0.021$, 函数拟合效果如图 7 所示. 按照式 (8) 可得到 CE 的信任分配函数 $m_{CE}(S)$ 如式 (22) 所示.

$$F(v(CE)) = \frac{0.9981}{1 + \mathrm{e}^{13.51(v(CE)-0.021)}} \tag{21}$$

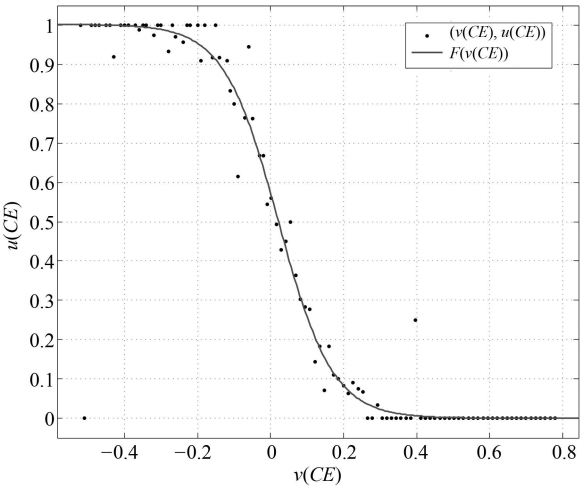


图 7 证据 CE 隶属度函数

Fig. 7 The membership function of evidence CE

从图 6 还可以看出, 当证据 CK 和 CN 的值域划分个数 $n = 50$ 时, 拟合的效果最好. CK、CN 证据的拟合效果如图 8、图 9 所示, 按照式 (8) 所计算的信任分配函数如式 (23) 和 (24) 所示.

3.3 评论异常度的评估

在构造好基本信任分配函数后, 利用改进的 D-S 理论对 3 个证据进行融合, 计算微博评论异常度. 实验对测试集微博的评论异常度进行计算, 谣言微博与普通微博评论异常度的分布如图 10 所示.

在图 10 中, 横坐标为评论异常度的值, 其取值范围为 [0, 1], 纵坐标为微博的数量. 从图 10 中可以看出, 普通微博与谣言微博的分布有较大的差别, 普通微博的评论异常度的取值大多位于 0.1 到 0.4

$$m_{CE}(S_{CE}) = \begin{cases} \dfrac{0.9981}{1 + \mathrm{e}^{13.51(v(CE)-0.021)}}, & S_{CE} = A \\ 1 - \dfrac{0.9981}{1 + \mathrm{e}^{13.51(v(CE)-0.021)}}, & S_{CE} = N \end{cases} \tag{22}$$

$$m_{CK}(S_{CK}) = \begin{cases} \dfrac{1}{1 + \mathrm{e}^{-122.5(v(CK)-0.0425)}}, & S_{CK} = A \\ 1 - \dfrac{1}{1 + \mathrm{e}^{-122.5(v(CK)-0.0425)}}, & S_{CK} = N \end{cases} \tag{23}$$

$$m_{CN}(S_{CN}) = \begin{cases} \dfrac{0.9997}{1 + \mathrm{e}^{-1.756(v(CN)-0.7689)}}, & S_{CN} = A \\ 1 - \dfrac{0.9997}{1 + \mathrm{e}^{-1.756(v(CN)-0.7689)}}, & S_{CN} = N \end{cases} \tag{24}$$

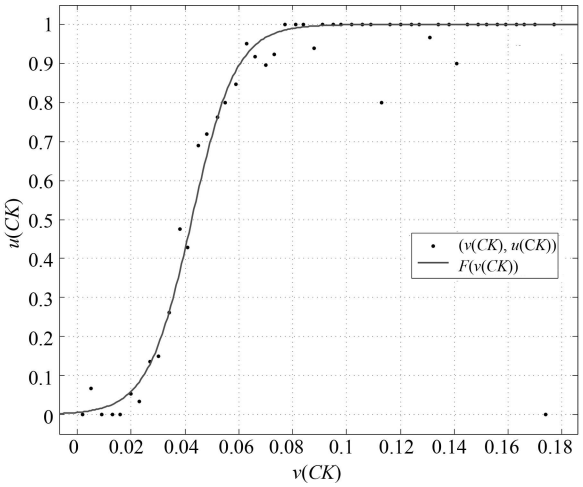


图 8　证据 CK 隶属度函数

Fig. 8　The membership function of evidence CK

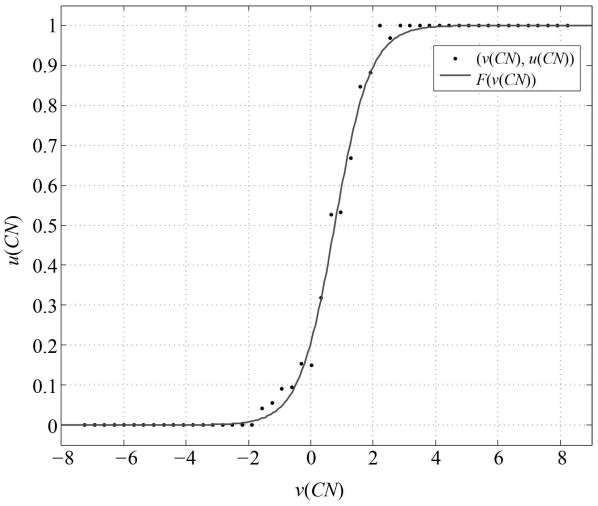


图 9　证据 CN 隶属度函数

Fig. 9　The membership function of evidence CN

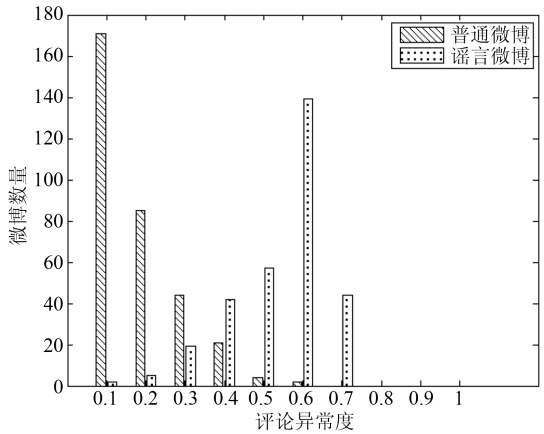


图 10　谣言微博与普通微博评论异常度分布对比

Fig. 10　Comparison of the comment abnormality distribution between rumor Weibos and ordinary Weibos

之间, 且随着评论异常度值的增加, 微博数目越来越少; 而谣言微博的评论异常度大多位于 0.5 到 0.7 之间, 并且在 0.6 时, 其数目达到了最多, 可以看出微博评论异常度对谣言微博与普通微博具有较好的区分能力.

从表 4 可以看出, 在这 14 项特征中, $Comment\_Abnormality$ 特征的准确率最高, 这与本文从网民对微博的反馈即评论信息方面进行综合考虑是密不可分的, 微博评论是网民经过主观思考后对微博内容进行的评价, 要比微博本身的内容和微博发布者所反映出来的特征具有更高的可靠性. 除此之外, 基于用户特征中的 Has_Verify 和 $Number\_of\_blogs$ 准确率位于第二、第三, 这说明微博内容本身的可信度也与微博的发布者即信息源头有着较大的关系. 而微博的文本长度、"@" 数量和话题数量的准确率相对较低, 说明其对普通微博和谣言微博的区分度有限, 也有部分原因是在本文所构建的实验数据集中, 微博的评论数量均大于 100, 这类微博的传播范围相对较广, 其中大多数都带有 "@" 符号和话题标签.

表 4　不同特征的准确率比较

Table 4　Comparison of accuracies of different features

序号	特征	准确率
1	$Length$	0.513
2	$Has\_Multimedia$	0.627
3	$Emotion\_Tendency$	0.601
4	$Number\_of\_@$	0.543
5	$Number\_of\_topics$	0.515
6	$Time\_Span$	0.633
7	$Client\_Type$	0.645
8	$Participation$	0.563
9	Has_Verify	0.671
10	$Has\_Description$	0.532
11	$Influence$	0.513
12	$Register\_Time$	0.639
13	$Number\_of\_blogs$	0.703
14	$Comment\_Abnormality$	**0.831**

3.4　基于评论异常度的谣言识别模型的评估

为验证本文所提出的基于评论异常度的谣言识别模型的效果, 设计了 4 组不同的特征集合进行验证, 根据特征组合构建微博表示向量, 采用 SVM 对微博进行分类实验, 具体特征组合情况如表 5 所示, 实验结果如表 6 所示.

从表 5 可以看出, 对照特征组合 3 即本文基于评论异常度的谣言识别模型取得了最好的效果, 其 F1 达到 0.962. 通过对基本特征组合和特征组合 1 的实验结果比较可以看出, 微博的评论情感异常度对谣言识别有较大的帮助, 这是由于微博评论的情

感是由微博内容引导的, 当微博内容对事实进行夸大或捏造时, 网民无法判断微博的真伪, 进而引发的情绪多为质疑、批判等负面情绪, 导致评论的情感倾向较为明显. 特征组合 2 相较于特征组合 1, 其召回率提升不大, 但是准确率却提升了 0.02, 说明评论用词异常度和评论数目异常度的引入, 丰富了微博评论的特征, 并能将那些内容真实但评论文本中情感倾向明显的普通微博与谣言微博区分开来. 但同时通过对比特征组合 2 和特征组合 3 可以看出, 如果将评论情感异常、评论用词异常和评论数量异常度三者进行融合可以起到更加好的效果, 这是因为利用 D-S 理论对三者融合, 可以减少评论特征之间相互冲突的影响, 加强了对评论特征整体的刻画, 同时减少了 SVM 的输入维数, 使得整体的评论特征在 SVM 训练过程中具有更好的区分性.

表 5 不同特征集合的组合情况

Table 5 Combination of different feature sets

对照实验	特征组合描述
基本特征组合	内容特征 + 传播特征 + 用户特征
特征组合 1	基本特征组合 + 评论情感异常度
特征组合 2	基本特征组合 + 评论情感异常度 + 评论用词异常度 + 评论数目异常度
特征组合 3	基本特征组合 + 微博评论异常度

表 6 不同特征集合的实验结果对比

Table 6 Comparison of experimental results of different feature sets

特征组合	准确率	召回率	F1 值
基本特征组合	0.868	0.913	0.890
特征组合 1	0.902	0.930	0.916
特征组合 2	0.928	0.937	0.933
特征组合 3	0.954	0.971	**0.962**

从谣言检测的时间段来看, 本文所提方法对于微博传播初期的谣言检测, 主要依靠微博的基础特征进行识别, 其效果优于文献 [6] 中不使用评论特征的情况. 在微博传播过程中, 其他微博用户的评论信息增多, 本文的方法对微博谣言识别的准确率在不断提高, 当评论数量超过 100 时, F1 值已高于文献 [8 – 9] 所得到的结果. 因此, 本文的方法在微博传播初期或评论达到一定数量时, 都能对谣言进行有效的检测.

在与文献 [6, 8 – 9] 进行比较时, 我们对其文献上所介绍的方法进行了实现, 并用我们构建的测试集进行测试, 由于数据集的不同以及实验参数设置的缘故, 测试效果与文献中所列的实验结果有所差距, 故在比较时, 我们采用了文献中介绍的实验结果. 本文方法与文献方法的实验对比如表 7 所示.

表 7 不同方法的实验结果比较

Table 7 Comparison of experimental results of different methods

方法	准确率	召回率	F1 值
文献 [6]	0.787	–	–
文献 [8]	0.913	0.913	0.913
文献 [9]	0.949	0.949	0.949
本文方法	0.954	0.971	**0.962**

文献 [6] 提取了新浪微博的基础特征, 采用 SVM 对微博进行分类, 准确率达到了 0.787. 文献 [8] 利用微博的转发评论, 构造了微博传播树, 使准确率有了显著提升, 达到了 0.913, 说明了传播结构所反映出的微博热度及群众的回馈对于谣言的识别是有帮助的. 文献 [9] 直接在基础特征上引入了微博评论的情感特征, 使准确率得到了进一步提升, 达到了 0.949, 说明了微博评论的情感特征能更加直观地反映微博所引导的情感和群众对微博的态度, 评论内容的潜在信息对微博分类具有更好的区分度. 本文的方法提取了更多的微博评论潜在信息, 获取了更全面的其他微博用户反馈信息, 使得准确率能够进一步提升, 达到了 0.954.

4 总结与展望

社交媒体传播的谣言对人们的生活有很强的破坏作用. 本文针对新浪微博谣言展开研究, 在现有的微博内容特征、传播特征、用户特征的基础上, 引入了微博的评论特征, 从评论的情感、用词和数量 3 个方面, 构造了微博评论异常度的计算模型, 进而实现了基于评论异常度的微博谣言识别模型, 并通过实验验证本文所提模型的合理性和有效性. 本文的主要贡献有: 从微博评论多个维度的信息进行分析, 将微博评论的多维特征引入到了微博谣言识别过程中; 构建了评论异常度计算方法, 能够对微博的评论状态进行有效评估; 将隶属度函数应用到 D-S 理论中, 为 D-S 理论中证据的初始信任分配提供了参考.

在现实环境中, 谣言微博与普通微博存在较大的不均衡性, 这可能导致有些特征在数据均衡情况下表现明显, 而在数据不均衡情况下效果变差的问题. 在下一步工作中, 我们将继续挖掘微博评论中更多具有区分性的潜在特征, 构建评论用户可信度的评估指标, 并利用深度学习的方法自动抽取微博文本与评论文本的相关性特征, 进一步提升谣言的识别效果. 此外, 在以评论为依据进行谣言识别时, 对于评论数较少的微博侦测具有一定局限性. 在进行谣言识别时, 可以考虑联系热点事件, 实时提取官方新闻等权威内容, 将权威信息与微博内容进行对比, 提升微博谣言的识别性能. 在当前研究趋势中, 社交

媒体的谣言识别逐步从静态的、时滞的、小规模的向动态的、实时的、大规模的转变[22], 在未来研究工作中, 可以利用本文的谣言识别模型融合多源数据, 从时间角度分析信息在社交媒体中的传播过程, 期望实现谣言模型能够对多源、大规模数据进行动态监测.

References

1 Qazvinian V, Rosengren E, Radev D R, Mei Q Z. Rumor has it: Identifying misinformation in microblogs. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011. 1589−1599

2 Takahashi T, Igata N. Rumor detection on twitter. In: Proceedings of the 6th International Conference on Soft Computing and Intelligent Systems, and the 13th International Symposium on Advanced Intelligence Systems. Kobe, Japan: IEEE, 2013. 452−457

3 Castillo C, Mendoza M, Poblete B. Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India: ACM, 2011. 675−684

4 Suzuki Y. A credibility assessment for message streams on microblogs. In: Proceedings of the 2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. Fukuoka, Japan: IEEE, 2010. 527−530

5 Ma J, Gao W, Mitra P, Kwon S, Jansen B J, Wong K F, Cha M. Detecting rumors from microblogs with recurrent neural networks. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA: AAAI Press, 2016. 3818−3824

6 Yang F, Liu Y, Yu X H, Yang M. Automatic detection of rumor on Sina Weibo. In: Proceedings of the 2012 ACM SIGKDD Workshop on Mining Data Semantics. Beijing, China: ACM, 2012. Article No. 13

7 Gao Ming-Xia, Chen Fu-Rong. Credibility evaluating method of Chinese microblog based on information fusion. *Journal of Computer Applications*, 2016, **36**(8): 2071−2075, 2081
(高明霞, 陈福荣. 基于信息融合的中文微博可信度评估方法. 计算机应用, 2016, **36**(8): 2071−2075, 2081)

8 Wu K, Yang S, Zhu K Q. False rumors detection on Sina Weibo by propagation structures. In: Proceedings of IEEE the 31st International Conference on Data Engineering, Seoul, South Korea: IEEE, 2015. 651−662

9 Zu Kun-Lin, Zhao Ming-Wei, Guo Kai, Lin Hong-Fei. Research on the detection of rumor on Sina Weibo. *Journal of Chinese Information Processing*, 2017, **31**(3): 198−204
(祖坤琳, 赵铭伟, 郭凯, 林鸿飞. 新浪微博谣言检测研究. 中文信息学报, 2017, **31**(3): 198−204)

10 Mendoza M, Poblete B, Castillo C. Twitter under crisis: Can we trust what we RT? In: Proceedings of the 1st Workshop on Social Media Analytics. Washington D.C., USA: ACM, 2010. 71−79

11 Yang J, Counts S, Morris M R, Hoff A. Microblog credibility perceptions: Comparing the USA and China. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work. San Antonio, Texas, USA: ACM, 2013. 575−586

12 Mao Er-Song, Chen Gang, Liu Xin, Wang Bo. Research on detecting micro-blog rumors based on deep features and ensemble classifier. *Application Research of Computers*, 2016, **33**(11): 3369−3373
(毛二松, 陈刚, 刘欣, 王波. 基于深层特征和集成分类器的微博谣言检测研究. 计算机应用研究, 2016, **33**(11): 3369−3373)

13 Zhang Yang-Sen, Zheng Jia, Tang An-Jie. A quantitative evaluation method of micro-blog user authority based on multi-feature fusion. *Acta Electronica Sinica*, 2017, **45**(11): 2800−2809
(张仰森, 郑佳, 唐安杰. 基于多特征融合的微博用户权威度定量评价方法. 电子学报, 2017, **45**(11): 2800−2809)

14 Liu Ya-Hui, Jin Xiao-Long, Shen Hua-Wei, Bao Peng, Cheng Xue-Qi. A survey on rumor identification over social media. *Chinese Journal of Computers*, 2018, **41**(7): 1536−1558
(刘雅辉, 靳小龙, 沈华伟, 鲍鹏, 程学旗. 社交媒体中的谣言识别研究综述. 计算机学报, 2018, **41**(7): 1536−1558)

15 Bordia P, Difonzo N, Schulz C A. Source characteristics in denying rumors of organizational closure: Honesty is the best policy. *Journal of Applied Social Psychology*, 2000, **30**(11): 2309−2321

16 Sevastianov P, Dymova L, Bartosiewicz P. A framework for rule-base evidential reasoning in the interval setting applied to diagnosing type 2 diabetes. *Expert Systems with Applications*, 2012, **39**(4): 4190−4200

17 Mokhtari K, Ren J, Roberts C, Wang J. Decision support framework for risk management on sea ports and terminals using fuzzy set theory and evidential reasoning approach. *Expert Systems with Applications*, 2012, **39**(5): 5087−5103

18 Li Wen-Li, Guo Kai-Hong. Combination rules of D-S evidence theory and conflict problem. *Systems Engineering-Theory & Practice*, 2010, **30**(8): 1422−1432
(李文立, 郭凯红. D-S 证据理论合成规则及冲突问题. 系统工程理论与实践, 2010, **30**(8): 1422−1432)

19 Zhao Qiu-Yue, Zuo Wan-Li, Tian Zhong-Sheng, Wang Ying. A method for assessment of trust relationship strength based on the improved D-S evidence theory. *Chinese Journal of Computers*, 2014, **37**(4): 873−883
(赵秋月, 左万利, 田中生, 王英. 一种基于改进 D-S 证据理论的信任关系强度评估方法研究. 计算机学报, 2014, **37**(4): 873−883)

20 Zhang Q, Zhang S Y, Dong J, Xiong J H, Cheng X Q. Automatic detection of rumor on social network. In: Proceedings of the 4th CCF Conference on Natural Language Processing and Chinese Computing. Nanchang, China: Springer, 2015. 113−122

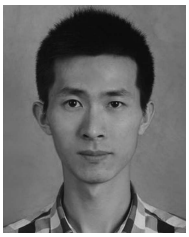
21 Chang C C, Lin C J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011, **2**(3): Article No. 27

22 Chen Yan-Fang, Li Zhi-Yu, Liang Xun, Qi Jin-Shan. Review on rumor detection of online social networks. *Chinese Journal of Computers*, 2018, **41**(7): 1648−1676
(陈燕方, 李志宇, 梁循, 齐金山. 在线社会网络谣言检测综述. 计算机学报, 2018, **41**(7): 1648−1676)

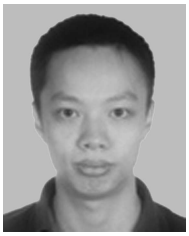
张仰森 北京信息科技大学教授. 主要研究方向为自然语言处理和人工智能. 本文通信作者.
E-mail: zhangyangsen@163.com
(**ZHANG Yang-Sen** Professor at Beijing Information Science and Technology University. His research interest covers nature language processing and artificial intelligence. Corresponding author of this paper.)

彭媛媛 北京信息科技大学硕士研究生. 主要研究方向为自然语言处理.
E-mail: pengyy0322@163.com
(**PENG Yuan-Yuan** Master student at Beijing Information Science and Technology University. Her main research interest is nature language processing.)

段宇翔 北京信息科技大学硕士研究生. 主要研究方向为自然语言处理和观点挖掘. E-mail: duanyx5173@163.com
(**DUAN Yu-Xiang** Master student at Beijing Information Science and Technology University. His research interest covers nature language processing and opinion mining.)



郑 佳 北京信息科技大学硕士研究生. 主要研究方向为自然语言处理和情感分析. E-mail: zhengjia0826@163.com
(**ZHENG Jia** Master student at Beijing Information Science and Technology University. His research interest covers nature language processing and emotion analysis.)



尤建清 北京信息科技大学讲师. 主要研究方向为自然语言处理.
E-mail: yjq@bistu.edu.cn
(**YOU Jian-Qing** Lecturer at Beijing Information Science and Technology University. His main research interest is nature language processing.)