项目描述的欺诈性与众筹投资意愿：基于文本分析的方法*

沈　倪[1]，王洪伟[2]，王　伟[3]
（1.浙江大学 管理学院，浙江 杭州 310058）
（2.同济大学 经济与管理学院，上海 200092）
（3.华侨大学 工商管理学院，福建 泉州 362021）

摘　要　以众筹市场为研究对象，采用文本分析与计量模型相结合的方法，检验项目描述的欺诈性对投资意愿的影响。采纳了内敛性、虚构性、分离性等指标进行文本欺诈性线索度量。采用线性和逻辑回归对数据进行分析，并实施鲁棒性检验。实验结果表明，描述项目的欺诈性线索与用户的在线投资意愿呈负相关。为此，投资者在选择项目时应当考虑项目描述暗含的欺诈信息，避免资产受损，而发起者在描述项目时也应当注意规避欺诈性的描述，以免造成误解，同时众筹平台应加强管理。

关键词　众筹项目，欺诈性，文本分析，投资意愿

中图分类号　C931.6

1　引言

众筹市场迅猛发展，与此同时，恶意欺骗事件层出不穷。2012 年，游戏众筹项目“星际公民”陷入卷钱跑路的传言中。2014 年，“皇冠众筹”项目虚假宣传推广，吸引 40 余人投资，投资金额达 200 余万元。随后，该平台忽然关闭，发起人因涉嫌诈骗而被捕。2015 年，上海优索环保科技发展有限公司涉嫌以“原始股”非法集资，其法人代表被批捕，其炮制的假股票骗取了上千名群众的 2 亿多元资金。

投融双方的互信是确保众筹成功的重要因素，但是由于信息不对称等原因，项目评估主要是由投资者本人完成，其风险性高。实际上，筹资者为了尽快募集资金，有时会故意夸大事实，使文本描述不真实。另外，广义上说，资金滥用也是一种欺诈，因为筹资者没有严格按照事先承诺的方式使用资金。为此，识别项目描述中的欺诈性线索亟待解决。

研究表明，产品描述会显著影响消费者的购买意愿[1]。同样地，众筹项目的陈述方式也会影响融资成功率。因此，有学者建议筹资者应该提高项目文本的展示质量，如采用更为合理的文字描述[2]。在欺诈性属性方面，有研究采用自然语言处理方法，从内敛性、虚构性、分离性、词汇复杂性、词汇多样性等方面提出多种文本欺诈性线索的识别方法[3-5]。但是，对于不同类型的欺诈性线索能否改变投资人的风险感知，以及对众筹融资结果的影响程度，还缺少系统性的分析。文献[6]基于 Kickstarter 数据，从四个维度（认知负荷、臆想情节、分离性和负面情绪）验证欺诈性线索对融资成功与否的影响。但是，鉴

* 基金项目：国家自然科学基金项目（71771177，71601082，71601119）；福建省自然科学基金项目（2017J01132）；福建省社会科学规划项目（FJ2016B075）。

通信作者：王洪伟，同济大学经济与管理学院，教授、博士生导师，E-mail：hwwang@tongji.edu.cn。

于中英文在语法和语义层面上的差异，该结论是否适用于中文语境下的众筹项目，仍有待验证。此外，即使筹资者无意欺骗，如果缺少理论指导，其筹资行为也可能会被误认为是欺骗行为。另外，鉴于融资成功率普遍偏低，众筹平台也需要向筹资者提供项目描述方面的指导，同时增强其自身对项目管理甄别的能力，而这方面的理论研究仍然薄弱。

令人欣慰的是，自然语言处理技术能够依据文本语言特征来进行欺诈分析。另外，从心理语言学角度，编造的故事与真实的故事在语言使用上存在显著差异，这也为文本欺诈线索的检测提供了思路。这涉及以下问题：①如何度量文本描述中的欺诈性线索？②欺诈性线索如何影响用户的投资意愿？为此，以众筹项目的文本描述为研究对象，基于文本分析的方法，采用不同的欺诈性度量指标，并应用线性和逻辑回归模型，证实了欺诈性语言特征与参与者支持行为呈负相关。这说明人们对于欺诈这一维度存在刻板印象，验证了一定置信范围内理论指标的正确性。研究结果有助于众筹发起者合理地描述项目，也有助于参与者和众筹平台更好地甄别项目。

2 研究假设与模型设计

2.1 研究假设

为了筹集资金，发起者会有意偏离事实，对项目进行夸张性描述，旨在获取投资者的支持。另外，发起者会将有创意但不成熟的项目拿出来筹资，导致投资者蒙受损失。这将引发一个问题：项目描述中的虚假信息对融资效果起到积极的还是消极的影响？

计划行为理论认为，影响实际行为最直接的因素是行为意向，而行为意向会受主观规范、行为态度及感知行为控制的影响。行为态度又受到行为信念和结果评价的影响。当个人对特定行为持正面的态度，认为符合其主观行为规范，且感觉已掌握采取该行为的能力和资源时，个人将产生强烈的行为意向，进而产生实际行为。

自我决定理论则把心理需求动机分为内部动机和外部动机[7, 8]。外部动机是指由外界奖励而产生的动机行为，而并非由个体自发产生。内部动机则包括：①自主需求，是指个体对于行为的自我控制需求，是一种自主选择能力的需求；②能力需求，是指个体对于行为有体现个体能力的需求，能力需求也表现为一种竞争性；③归属需求，是指个体需要和他人保持关联，以满足个体自我归属的需求。

众筹是一种借助互联网公开募集资金的方式，通过捐赠、预购商品或者获得回报等方式，对具有特定目的的项目提供资金支持。调查显示，投资者参与众筹的目的有：①获得发起者承诺的回报；②通过帮助发起者以获得成就感；③为了加入与项目或发起者有关的社交圈[5, 9-11]。

根据自我决定理论，获得回报属于外部动机，获得成就感属于内部动机的能力需求，加入社交圈则属于内部动机的归属需求。如果众筹描述文本存在诈骗信息，参与者就会对项目预期结果没有把握，会感到无法获得发起者承诺的回报，降低对投资行为的控制能力，因而参与态度不那么积极。另外，如果项目存在欺诈性，投资风险就会大大增加，投资者觉得自己所能控制的资源和机会减少，依据计划行为理论，投资意愿和投资行为就会受到负面影响。所以本文提出如下假设。

H：项目描述文本包含越多的欺诈线索，就越不容易获得参与者支持，项目融资成功率越低。

2.2 欺诈性检测模型

在众筹项目的线上展示方式中，项目的文本描述所占篇幅最大，它是用户获取项目信息的主要方式。产品描述能够显著影响产品销量，将其运用到众筹项目上来说，在互联网背景下，文本描述作为展示众

筹项目的最主要内容，也会影响投资者的投资意愿。有学者通过使用基于心理学分类的词典对项目描述文本进行分析，揭示了描述文本中特定词汇的使用可以提高项目筹资成功率。文本信息有多种维度，本文将重点关注欺诈性线索的识别及其如何影响投资者的投资意愿。Newman 和 Pennebaker[4]从心理语言学上分析，认为伪造的故事与真实的故事在语法的使用上存在差异，所以可以利用语言特征设立指标来检测文本欺诈性。已有的研究文献把欺诈检测划分为多个方面。本文将遵循已有的研究成果，针对中文的语言特征，采取以下几个指标作为欺诈检测的标准。

（1）内敛性。内敛性是指文本逻辑连贯并且完整。Graesser 等[12]发现连词数量越多，文本的内敛性越强。虚构的事件总是支离破碎的。而许琼恺[13]针对欺骗性语料的特殊性，结合现有的文献资料，提出了基于假设检验的语言学线索抽取方法，通过文本内容的欺骗特征线索抽取，发现欺骗性文本比非欺骗性文本具有更少的第一人称代词、时间信息、空间信息和感知信息。参照所收集的文本数据特征，表 1 是收集到的常用连词和代词。

表 1 连词、时间和空间代词、人称代词列表（部分）

词性	实例
连词	况且 何况 乃至 纵使 纵令 纵然 致使 无论 不论 所以 只有 只要 乃至 与其 由于 因而 因为 因此 以至 以致 不然 不仅 不但 既然 即使 尽管 何况 况且 哪怕 除非 但凡 从而 而且 反而 而况 否则 固然 故而 果然 于是 至于 此外 譬如 如同 并且
时间和空间代词	这 那 这儿 那边 各 每 这里 这会儿 那儿 那里 那会儿 天 周 年 月 日
第一人称代词	我 朕 吾 予 余 俺 我们 咱们 大家 自己 俺们
非第一人称代词	厥 之 其 彼 诸 夫 人 他 它 她 人家 别人 旁人 他们 她们 诸位 列位 各位 任何人 有人 人们 它们 别人们 某人 有些人 你 您 尔 女 汝 若 而 乃 你们

（2）虚构性、分离性。现实检测理论（reality monitoring theory）显示，从真实经历回顾的故事会包含较多的空间和时间信息。Mcquaid 等[14]发现欺诈者经常使用“他们”这样的第三人称代词和单数人称代词，而真正的故事诉求者则更多地使用“我们”这样的人称代词和其他时间代词。Knapp 和 Comaden[15]发现说谎者通常将自己和他们的语言分离开来，因为他们缺少个人真实的经历。Mehrabian 和 Wiener[16]也发现相较于说真话的人，说谎者经常不会直接提起自己，而是采用一种间接的叙述方式。Newman 等[4]发现说真话的人更偏向使用第一人称，那些说谎者为了避免承担责任，更倾向于把自己从编造的故事中分离开来。欺骗性交际的特点是第一人称代词少，这是因为编造的故事总是比较简单，没有落实到具体时间地点，这样不易因前后不一致而露出破绽。分离性是指作者在多大程度上希望与文本内容分离开。欺骗行为通常与高度焦虑和内疚有关，欺诈者由于撒谎而感到愧疚，希望能够与欺诈文本分离，所以总倾向于使用非第一人称代词。

（3）多样性。多样性反映了词汇的宽泛度，Durán 等[17]采用文本中词汇的相对频数来度量。邓莎莎等[18]结合心理学相关的欺骗理论，提出了 11 种欺骗语言线索共 3 类欺骗特征（评论的词语词频；评论内容的丰富程度，其中包括词性分布、语句多样性、时空代词和感知信息；内容信服度特征，主要是语言接近程度特征），并在由评论者分别撰写的真实评论和虚假评论语料上检验了各种欺骗组合特征集的效果。实验证明，识别欺骗评论的精度接近 80%。Lau 等[3]针对 Amazon 上对产品和服务的评论，设计试验了新的计算模型来检测虚假评论，通过语义重叠性，可以判定文本不可信的程度。Wang 等[19]利用三个节点构造网络来判断评论的虚假性，认为利用相似度可以识别虚假信息。综上所述，可以看出文本语句多样性、语义重叠性，还有语言接近程度都是度量文本欺诈性的良好指标，为此，我们通过词汇多样性来衡量文本的欺诈性程度。

文本的多样性指数越大，说明文章的层次越高，编写者文化水平越高，文本欺诈性就会越低。本文借鉴辛普森指数来计算文本的多样性：

$$D=\frac{1}{\sum\left(\frac{F_i}{\text{length}}\right)^2 \text{length}} \tag{1}$$

其中，F_i 为词语 i 出现的频数；length 为文本长度。

（4）复杂性。复杂性反映了文本在多大程度上被读者理解。Lau 等[3]根据不确定性减少理论和可能性模型，发现文章长度在 20~817 字长时，额外的描述对借贷的成功有正向的作用，语言若表现出具体性这一维度，文本中含有描述的数量信息能增加借贷的成功率。彭红枫等[20]基于 Prosper 平台上的数据，利用迷雾指数，发现在利率竞拍机制下，信用等级越低的借款人，越倾向于提供借款陈述；借款人提供借款陈述能降低借款成本，但是不一定能提高借款成功率。在利率竞拍模式下，借款陈述的迷雾指数与借款成功率呈现倒“U”形关系。迷雾指数是句子的平均长度和复杂单词所占比例的线性组合［式（2）］，用于度量借款陈述的阅读难度，迷雾指数的值越小，说明借款陈述的可读性越强。可读性过强的文本虽然生动易懂，但是在语言表达的精确性、理论的严密性等方面却相对不足。

$$\text{FogIndex}=0.4\left(\text{ASL}+100\text{ACW}\right) \tag{2}$$

其中，ASL 为句子的平均长度，由总词语个数除以句子个数得到；ACW 为复杂词语的比例，由复杂单词个数（即音节大于 2 的词语个数）除以总单词个数得到。

3 数据来源与实验结果

3.1 数据来源

实验数据来自众筹网。众筹网是一家有影响力的众筹融资平台，为大众提供筹资、投资、孵化、运营一站式综合众筹服务。2017 年，众筹网为近 1 万个项目筹款超过 1 亿元。

筹资失败的项目无法被搜索引擎直接检索到，但是项目的统一资源定位符仍旧有效。每个项目都有编号作为标识，所以可将其作为识别的线索，通过循环项目编号来采集文本，所采集到的文本中包含了失败和成功的项目，通过控制编号数量，就可以得到所需数量的项目。

利用 Python 语言编写爬虫程序，抓取众筹网的文本数据，表 2 给出一个实例。将项目信息存入文件，并通过选取的度量指标转换为数值信息，然后进行线性和逻辑回归分析，从而获得参数结果来检测模型的准确性。

表 2 文本实例展示

项目编号	项目标题	项目简介	融资结果
58207	跑跑面试：用手机做面试	功能介绍：我们追求极致的用户体验，希望用科技来解决企业在招聘面试过程中的棘手问题，为所有正处在找人难、招人难的企业人力资源管理者们带来一种新奇又激动人心的产品。创造一个新的产品实在不容易，在推出这个全新的产品理念并付诸实践的过程中存在很多困难，我们想借众筹平台与大家分享这一成果，让所有企业的人力资源管理者们能体验到全新的面试模式。希望大家能与我们同行，请多多支持！①	成功
119917	松子的呐喊：不做低头族	愿与大家一同分享那存在深山原生态的东西，若项目成功我将为大家收集新鲜的未经任何加工的松子。坐车走这样的路我从来都不敢系安全带，若……我们的选择就是及时跳车。这样的事故时有发生，特别是在雨季，若稍微大一点的车走这样的路遇到急弯一次性转不过来需要倒一次车的，驾驶员必须下车用石头把车轮边缘卡住再倒车，若不这样稍微调整不好可能就会……②	失败

① 《跑跑面试：用手机做面试》，http://www.zhongchou.com/deal-show/id-58207[2017-04-13]。

② 《松子的呐喊：不做低头族》，http://www.zhongchou.com/deal-show/id-119917[2017-04-13]。

3.2 实验结果

收集到 4317 个项目数据，除去重复和缺失数据，得到 4008 个项目数据。其中，成功项目 1851 个，失败项目 2157 个。项目简介字符数共 4 050 819 个，平均长度为 938.34 个。

采用皮尔森相关系数对欺诈度量指标所对应的自变量进行相关性分析，结果如表 3 所示。可以发现，第一人称代词、非第一人称代词和连词数量之间存在多重相关性。如果不处理，其将会影响后续分析的准确性。经调整，将第一人称和非第一人称指标合并，以人称代词（第一人称代词数量与非第一人称代词数量的差值）来代替这两个指标。再次进行皮尔森相关系数计算（表 4），处理后，变量间的相关系数下降，可进行后续分析。项目结果为二分变量（0 或 1），故首先采取逻辑回归后可得结果，如表 5 所示。

表 3 皮尔森相关系数表（一）

项目	连词	时空代词	第一人称代词	非第一人称代词	迷雾指数	多样性指数
连词	1	0.18	0.6	0.48	−0.06	−0.45
时空代词	0.18	1	0.22	0.42	−0.0094	−0.34
第一人称代词	0.6	0.22	1	0.48	−0.05	−0.49
非第一人称代词	0.48	0.42	0.48	1	−0.051	−0.45
迷雾指数	−0.06	−0.0094	−0.05	−0.051	1	0.045
多样性指数	−0.45	−0.34	−0.49	−0.45	0.045	1

表 4 皮尔森相关系数表（二）

项目	连词	时空代词	人称代词	迷雾指数	多样性指数
连词	1	0.18	0.45	−0.06	−0.45
时空代词	0.18	1	0.041	−0.0094	−0.34
人称代词	0.45	0.041	1	−0.031	−0.34
迷雾指数	−0.06	−0.0094	−0.031	1	0.045
多样性指数	−0.45	−0.34	−0.34	0.045	1

表 5 逻辑回归模型结果表

项目	回归系数	估计标准误差	Z 值	$p>\|Z\|$	95%的置信区间	
连词	0.0280	0.015	1.913	0.056	−0.001	0.057
时空代词	0.0593	0.006	10.582	0.000	0.048	0.070
人称代词	0.0055	0.003	1.916	0.055	0.000	0.011
迷雾指数	−0.0013	0.004	−0.339	0.734	−0.009	0.006
多样性指数	0.3183	0.125	2.555	0.011	0.074	0.562
截距	−0.6130	0.116	−5.303	0.000	−0.840	−0.386

就内敛性来看，文本包含的连词数量越多，就越容易获得融资。虚构分离性方面，时空代词与人称代词和融资结果也呈正向关系。词汇复杂性方面，迷雾指数与融资结果呈负相关关系，文本多样性指数则与融资呈正相关关系。

4 鲁棒线性检测

为了确保结论的准确性和稳定性，本文采集了更多的项目信息（表 6），并采用以下方法进行鲁棒性测试。采取的测试方法是更换模型和因变量指标。首先，针对融资结果，更换了鲁棒线性回归模型（表 7）；其次，针对筹资比率，将其作为连续的因变量替代了融资结果（二分变量）进行线性回归（表 8）；最后，将它们合并进行对比分析来测试上述检测指标的容错能力和稳定性（表 9）。

表 6 项目扩展信息表

编号	项目结果	支持数/个	已筹款/元	筹资比例	目标筹资/元
139426	1	83	1 513	1.01%	1 500
116665	1	141	10 796	1.08%	10 000
143752	1	48	5 863	1.01%	5 808
7078	1	380	118 833	3.97%	30 000

表 7 鲁棒性检验结果（一）

项目	回归系数	估计标准误差	*Z* 值	*p*>|*Z*|	95%的置信区间	
连词	0.0051	0.004	1.372	0.170	−0.002	0.012
时空代词	0.0131	0.001	10.781	0.000	0.011	0.016
人称代词	0.0017	0.001	2.417	0.016	0.000	0.003
迷雾指数	−0.0005	0.001	−0.677	0.498	−0.002	0.001
多样性指数	0.0778	0.030	2.592	0.001	0.019	0.137
截距	0.3442	0.025	13.705	0.000	0.295	0.393

表 8 鲁棒性检验结果（二）

项目	回归系数	估计标准误差	*Z* 值	*p*>|*Z*|	95%的置信区间	
连词	0.0066	0.003	1.909	0.056	−0.002	0.012
时空代词	0.0132	0.001	11.133	0.000	0.011	0.016
人称代词	0.0013	0.001	1.914	0.056	0.000	0.003
迷雾指数	−0.0003	0.001	−0.352	0.724	−0.002	0.001
多样性指数	0.0693	0.030	2.312	0.002	0.019	0.137
截距	0.3587	0.027	13.406	0.000	0.295	0.393

表 9 各模型系数与 *p* 值对比表

回归模型	回归参数	连词	时空代词	人称代词	迷雾指数	多样性指数
逻辑回归模型	回归系数	0.0280	0.0593	0.0055	−0.0013	0.3183
鲁棒线性回归模型	回归系数	0.0066	0.0132	0.0013	−0.0003	0.0693
普通最小二乘法回归模型	回归系数	0.0407	0.0124	0.0018	-1.387×10^{-5}	0.4224
逻辑回归模型	*p*>|*Z*|	0.0560	0.0000	0.0550	0.7340	0.0110
鲁棒线性回归模型	*p*>|*Z*|	0.0560	0.0000	0.0560	0.7240	0.0210
普通最小二乘法回归模型	*p*>|*Z*|	0.0100	0.0016	0.0000	0.9960	0.0010

表 9 显示，在 3 个不同的回归模型下，6 个检测指标的系数，符号一致，取值相近，说明所得结论具有稳定性。对 p 值而言，可以看到，除了人称代词和迷雾指数有波动外，其余检测指标系数的 p 值在 3 个回归模型下并没有太大的改变，结果仍旧显著。以筹资比率作为因变量时，迷雾复杂度则呈现较大的 p 值，迷雾指数结果将在下一章做具体分析。

通过鲁棒性测试后，可以认为逻辑回归的结果在一定置信区间内是稳定的，选取的检测指标是比较合理的。

5 假设检验结果及解释

逻辑回归及鲁棒性检测显示了欺诈性与融资结果的关系。内敛性与融资成功正相关。内敛性强，说明项目描述有逻辑性。从认知角度看，逻辑性强的文本比支离破碎的文本更有说服力，不容易产生欺骗性。

在虚构性和分离性方面，时空代词和人称代词与融资成功正相关，这是因为出现的时空代词和第一人称代词越多，说明撰写者在描述真实的事情时，倾向于从自己的视角出发，这比没有时间、地点和人物的描述更让人觉得可信。非第一人称代词多的文本会让人觉得是在讲述别人的故事，这样的文本更容易让人觉得是虚构的。

对于文本词汇复杂性来说，迷雾指数值越大，说明陈述的复杂度越高，文本可读性越弱。这说明文本词汇复杂度越高，可读性越弱，融资越不容易成功。理论上，就复杂性来说，欺骗性的文本具有较少的长句和较少的音节，即较低的复杂度，这样的文本虽然生动易懂，但是在语言表达的精确性、理论的严密性等方面相对不足，所以更容易显示出欺诈性。反倒是可读性稍弱一些，复杂性稍高的文本较易获得信任。此次验证的结果与理论上不完全符合，可能有以下两个原因：①在实际项目中，复杂性过高会令人觉得晦涩难懂，有故弄玄虚之感。通俗易懂的文本反倒更像是真实的事情，包含过多复杂词汇的文本不像在叙述真实事件。在现实中，大家会觉得复杂度高、可读性差的文章欺骗性更强，进而支持那些好理解的文本。在将来进一步的研究中，将定义词汇复杂性的分界线以便于做更细致的统计。②样本数据分为不同的类别，需要按照项目类别进行区分，不同类别的文本风格有差异。科技类的项目也许原本就比较高深莫测，引用了比较专业的术语，艺术类的项目更加注重描述，而农业类的项目风格可能比较务实，注重农产品细节的说明。为此，需要针对不同类别的项目，分别考虑它们的影响效果。有些词汇在特定的领域并不算是复杂词汇，而在文本中出现的频次却很高。这些词汇在特定分类的文本中进行研究的时候是需要进行甄别和剔除的。针对多样性指数来说，它与融资效果正相关。通常，文本的多样性越强，说明文字越复杂，句式也越复杂。比起那些单一的文本，人们会觉得这样的文本更加可靠，这也反映了撰写者的文化程度较高。那么无论针对文本本身还是撰写者，这类文本都不易显示出欺诈性，所以项目更容易得到支持。由上述各个指标的分析结果看来，除去复杂性指标与理论上有一定出入外，最初的假设在置信区间内成立，即欺诈性越高的文本越不容易获得支持，融资越不容易成功。

6 理论贡献和管理启示

本文证实了欺诈性语言特征与参与者支持行为呈负相关关系，并显著影响项目筹资效果，说明了人们对于欺诈这一维度的刻板印象，验证了一定置信范围内理论指标的正确性。同时，借鉴自然语言学和心理学的理论，提出检测欺诈性可能性的指标，构建检测模型，启示众筹发起者合理描述项目，帮助参

与者和众筹平台更好地甄别项目，这在一定程度上充实了众筹投资理论。

对于众筹投资者来说，甄别项目时，首先要关注项目描述的内容。但是，如何根据文本描述去甄别那些含有虚假信息的项目进而做出投资决定，仍旧有难度。本文提出的指标和检测模型，可以更为有效地帮助参与者审视描述性文本，通过对内敛性、分离性及虚构性、词汇复杂性、词汇多样性指标的计算，有助于参与者更加容易地找准项目定位，甄别项目背后是否暗藏欺诈的信息。

对于发起人来说，在描述众筹项目时，可能会采用错误的描述方式，文字表述过于晦涩，有时与自己的本意存在较大的偏差，甚至会令人误解为虚假信息，导致项目发布后难以获得民众支持，从而使得筹资失败。为此，可以通过检测指标进行文本分析。通过连词信息的计算来观察是否完整而连贯地叙述了故事情节；通过时空代词的计算来观察是否具体而细致地回忆了事件所发生的真实时间和地点；人称代词的数据分析可以帮助发起者观察是否引用了太多的非第一人称叙述词，使得文本像是虚构的故事；通过分析文本的复杂性和词汇多样性来观察是否包含了冗余或晦涩难懂的词汇与句子，是否运用了宽泛的词汇和句子来表达自己的想法。经过上述分析，文本能更好地展示发起人的想法和创意，从而促成项目融资成功。

对于众筹网站来说，作为项目发布的承载平台，有责任保障项目的可靠性。针对那些可能存在欺诈的项目，平台本身应该对发起人进行更加严格的资质审查或者实名认证，以便于发生纠纷时有所防范和应对。对于可信性差的项目，平台可以拒绝发起者的项目发布。此外，平台还可以向发起者提出一定的警示，引导发起人规范、合理地发布自己的项目。针对潜在投资者来说，平台也应当在一定程度上保障他们的利益，可以发出公告，提醒他们要小心谨慎地实施自己的投资支持行为，一定要注意文本暗含的内容信息，做出决策时需要理智，同时平台应做出一些限制条件和项目善后服务。

7 不足与展望

首先，本文实验数据来自众筹网，来源不够宽泛；其次，没有对项目类别进行区分，如农业、艺术等不同类别，不同项目的文本描述各有偏重，风格不同，对这些数据进行相同的处理可能导致最终结果也有所区别，如果本文的研究采用已经被验证的数据集来进行验证，或者用一批已经被证实是由欺诈性描述导致失败的项目来进行佐证，那么指标的说服力将会更大，本文目前做的是一个验证预测类实验，在日后将着重于这部分数据的收集和检验，尝试更多的数据来源，增强数据样本容量和多样性；最后，对于文本分析模型来说，本文的自然语言处理技术在欺诈性线索挖掘方面有待提高。文本的特征各式各样，此次实验选取的维度还不够宽泛，如词频和语序对于文本来说也十分重要，一句话中转折连词的位置不同就可能导致整段文字呈现出截然不同的含义，这些维度也将在日后被考虑进模型中以便做更加细致的研究。数据处理上，多重共线性的解决方案仍旧是计量模型的一个重要方面，本次采取了人称代词这一指标的变化来降低多重共线性，未来研究中应当更加深入地处理这些变量间的关系。

另外，本文给出了欺诈性信息对投资意愿的影响，但是没有从心理学等视角探讨产生这种影响的原因，未来的研究应当进一步分析数据背后所隐含的意义。

参 考 文 献

[1] Goes P，Lin M，Au Yeung C M. Popularity effect in user-generated contents：evidence from online product reviews[J]. Information Systems Research，2014，25（2）：222-238.

[2] Zhou M，Lu B Z，Fan W G，et al. Project description and crowdfunding success：an exploratory study[J]. Information Systems

Frontiers，2018，20：259-274.
[3] Lau R Y K，Liao S Y，Kwok R C W，et al. Text mining and probabilistic language modeling for online review spam detection[J]. ACM Transaction of Management Information System，2012，2（4）：1-30.
[4] Newman M L，Pennebaker J W，Berry D S，et al. Lying words：predicting deception from linguistic styles[J]. Personality and Social Psychology Bulletin，2003，29（5）：665-675.
[5] Schwienbacher A，Larralde B. Crowdfunding of Small Entrepreneurial Ventures[M]. Oxford：Oxford University Press，2010.
[6] 王伟. 项目描述的文本特征与投资意愿：基于众筹市场的研究[D]. 上海：同济大学，2016.
[7] Deci E L，Ryan R M. Intrinsic Motivation and Self-Determination in Human Behavior[M]. New York：Plenum Press，1985.
[8] Harter S. Effectance motivation reconsidered toward a developmental model[J]. Human Development. 1978，21（1）：34-64.
[9] Allison T H，Davis B C，Short J C，et al. Crowdfunding in a prosocial microlending environment：examining the role of intrinsic versus extrinsic cues[J]. Entrepreneurship Theory and Practice，2015，39（1）：53-73.
[10] Cholakova M，Clarysse B. Does the possibility to make equity investments in crowdfunding projects crowd out reward-based investments？[J]. Entrepreneurship Theory and Practice，2015，39（1）：145-172.
[11] Gerber E，Hui J，Kuo P Y. Crowdfunding：why people are motivated to post and fund projects on crowdfunding platforms[R]. Computer Supported Cooperative Work，2012.
[12] Graesser A C，McNamara D S，Kulikowich J M. Coh-Metrix：providing multilevel analyses of text characteristics[J]. Educational Researcher，2011，40（5）：223-234.
[13] 许琼恺. 基于语言特性的互联网欺骗信息的自动识别[D]. 上海：上海交通大学，2014.
[14] Mcquaid S M，Woodworth M，Hutlon E L，et al. Automated insights：verbal cues to deception in real-life high-stakes lies[J]. Psychology Crime & Law，2015，21（7）：617-631.
[15] Knapp M L，Comaden M A. Telling it like it isn't：a review of theory and research on deceptive communications[J]. Human Communication Research，1979，5（3）：270-285.
[16] Mehrabian A，Wiener M. Decoding of inconsistent communications[J]. Journal of Personality and Social Psychology，1967，6（1）：109-114.
[17] Durán P，Malvern D，Richards B，et al. Developmental trends in lexical diversity[J]. Applied Linguistics，2004，25（2）：220-242.
[18] 邓莎莎，张明柱，张晓燕，等. 基于欺骗语言线索的虚假评论识别[J]. 系统管理学报，2014，23（2）：263-270.
[19] Wang G，Xie S，Liu B，et al. Identify online store review spammers via social review graph[J]. ACM Transactions on Intelligent Systems and Technology，2012，3（4）：1-21.
[20] 彭红枫，赵海燕，周洋. 借款陈述会影响借款成本和借款成功率吗？——基于网络借贷陈述的文本分析[J]. 金融研究，2016，（4）：158-173.

The Impact of Fraudulent Clue in Crowdfunding Campaign Description on Investment Willingness through Text Analytics

SHEN Ni[1]，WANG Hongwei[2]，WANG Wei[3]
（1. School of Management，ZheJiang University，Hangzhou 310058，China）
（2. School of Economics and Management，Tongji University，Shanghai 200092，China）
（3. College of Business Administration，Huaqiao University，Quanzhou 362021，China）

Abstract Taking the crowdfunding market as the research object，this paper combines the method of text analysis and econometric model analysis to validate the relationship between fraudulent of the project description and the investment willingness. This paper adopts the following text indicator variables to detect and measure the fraudulent information in the text：cohesion，dissociation，fabrication and so on. This paper utilizes the linear and logistic regression models and test the robustness of the models as well. The experimental results show that the fraudulent clues describing the project have a negative correlation with the user's online investment willingness. For this reason，investors should consider the fraudulent information implied by the project description when choosing a project，and to avoid asset damage. The sponsors should also pay attention to circumvent fraudulent description when describing the project，so as to avoid misunderstanding，and at the same time crowdfunding platforms

should strengthen their management ability.

Key words Crowdfunding projects，Fraudulent identification，Text analysis，Investment willingness

作者简介

沈倪（1994—），女，浙江大学管理学院博士研究生，研究方向：信息管理和数据挖掘，物流供应链与物流管理优化。E-mail：rowlandshen1@163.com。

王洪伟（1973—），男，同济大学经济与管理学院教授、博士生导师，研究方向：商务智能与文本挖掘。E-mail：hwwang@tongji.edu.cn。

王伟（1982—），男，华侨大学工商管理学院副教授、硕士生导师，研究方向：金融科技与商务数据分析。E-mail：wwang@hqu.edu.cn。