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，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 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具有负相关。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，同时众筹平台则应加强管理能力。

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，使文本描述显得不真实。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。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，编造的与真实的故事在语言使用上存在显著差异，这也为检测文本欺诈线索提供了思路。这涉及到以下问题：（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？

有研究表明产品的文本描述能够在很大程度上影响产品销量[8]。同样的众筹项目的展示方式(如文本描述)会影响项目的融资成功率。因此，研究者们建议筹资者应该提高项目文

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）
作者简介：沈倪（1994），女（汉），博士生，浙江宁波，E-mail：rowlandshen1@163.com
通讯作者：王洪伟（1973），男（汉），教授，辽宁大连，E-mail：hwwang@tongji.edu.cn

本展示的质量，例如:采用更好的文字描述[15]。在文本的欺诈性属性方面，虽然已有研究提出了一些如何识别文本欺诈性线索的方法。但是，在具体采用怎样的指标度量欺诈性和欺诈性线索对项目结果的影响程度方面的研究还是非常少。并且筹资者在实际运作众筹项目时，总是期望有足够的理论研究进行指导。但是，由于众筹模式出现时间晚，发展时间不长，纵观现有的研究成果， 很难对筹资者的实际操作提供完善的理论指导。有时筹资者会无意识的在文本描述中透露出欺诈信息。另外一方面，对众筹平台来说，也面临相似的困境，尽管众筹平台发展迅速，但是成功筹资的项目比例低，众筹平台的实际需求之一也是需要提供给筹资者有关众筹项目文本描述方面的指导，同时增强自身对项目管理甄别的能力，而这方面需求在目前的理论研究中，还几乎是一片空白。

为此，针对众筹项目的文本描述，基于文本分析的方法，采用 Python 语言对网络文本进行抓取清洗，采纳内敛性、虚构、分离性、词汇复杂性、词汇多样性指标进行欺诈性度量，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负相关。这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。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

为了筹得资金，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。另一方面，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抑或消极的影响。

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，而行为意向受主观规范、行为态度以及感知行为控制所影响。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[6,10]。外部动机是指由于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，以满足个体的自我归属的需求。

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[21]。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[3, 5, 9,21]。

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

在众筹项目的在线展示方式中，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章进行分析，揭示了描述文本中特定词汇的使用可以增加项目筹资成功率。文本信息有多种维度，本文将重点关注文本中欺诈性线索的识别，来研究其如何影响投资者的投资意愿。Pennebaker[16]从心理语言学上分析，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，遵循已有的研究成果，并且针对中文的语言特征，本文将采取以下几个指标作为欺诈检测的标准。

1)内敛性：是指文本逻辑连贯并且完整。Graesser[1]发现连词数量越多，文本的内敛性越强。虚构编造的事件总是支离破碎的，而许琼恺[29]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 是收集到的常用的连词和代词。

表 1 连词、时空代词、人称代词列表（部分）

	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	你 我 朕 您 吾 予 余 尔 女 汝 若 而 乃 俺 你们 我们 咱们 大家 自己 俺们 大家
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

2)虚构、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。Sarah M[20]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词。Knapp[12]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。

Wiener 和 Mehrabsian[2]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[16]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，更倾向于把他们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（如 I, my, me）少，这是因为编造的事件故事总是比较简单，没有落实到具体时间地点，这样不易前后不一致而露出破绽。分离性是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，欺诈者由于撒谎带来的心理愧疚感，希望能够与欺诈文本分离，所以总倾向于使用非第一人称指示词。

3)多样性：反映了词汇的宽泛度，Durán 和 Malvern[18]采用文本中词汇的相对频数来度量。邓莎莎[21]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、时空信息和感知信息；内容信服度特征，主要是语言接近程度特征）；采用设计科学的方法，实现了在线欺骗识别系统，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[14]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[20]利用三个节点构造网络来判断评论的虚假性，认为相似度可以用来识别虚假信息。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良好的指标，为此，我们选取词汇多样性计算文本的欺诈性程度。

文本的多样性指数越大则说明文章的层次越高，编写者文化水平越高，文本欺诈性就会降低。本文计算多样性的公式借鉴了辛普森指数：

$$D = 1/\sum (Fi/Length)\hat{}2 * (Length) \quad (1)$$

其中 F_i：词语 i 出现频数，$Length$ 为文本长度。

4)复杂性：反映了文本在多大程度上被读者所理解。Lau[14]根据 Uncertainty Reduction 理论和 Elaboration Likelihood 模型，发现文章长度在 20 到 817 字长之间时，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时，文本中含有描述的数量信息能增加借贷的成功率。彭红枫[26]基于 Prosper 平台上数据，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，但是不一定能增加借款成功率，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 型”关系。指数是句子的平均长度和复杂单词所占比例的线性组合（见公式 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。

$$FogIndex = 0.4 * (ASL + 100 * ACW) \quad (2)$$

其中，ASL 是句子的平均长度，由总词语个数除以句子个数得到；ACW 是复杂词语的比例，由复杂单词（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”。这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 URL 仍旧有效。每个项目都有 ID 作为标识，所以可将其作为识别的线索，通过循环项目 ID 来采集文本，这一串项目中就会包含失败和成功的项目，通过控制 ID 数量，就可以得到所需数量的项目。

利用 python 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，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 ID	项目标题	项目简介	项目结果
58207	跑跑面视——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 HR 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是很容易，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 HR 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（表示项目成功）

3.2 实验结果

收集 4317 个项目数据，除去重复和缺失数据，得到有效数据为 4008 个项目。其中，成功项目 1851 个，失败项目 2157 个。项目简介字符数共 4050819 个，平均长度为 938.34 个。

对于欺诈度量指标这些自变量进行相关性分析，根据皮尔森相关系数（见表 3）发现，第一人称、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，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。再次进行相关系数计算（见表 4），这样处理后，变量间相关系数下降，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(见表 5)。

表 3 皮尔森系数表 1

	连词	时空代词	第一人陈	非第一人称	FOG 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	0.6	0.22	1	0.48	-0.05	-0.49
非第一人称	0.48	0.42	0.48	1	-0.051	-0.45
FOG 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 皮尔森系数表 2

	连词	时空代词	人陈代词	FOG 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
FOG 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

	Coef	std err	Z	P>|z|	[95.0% Conf]	[Int.]
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
FOG 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
intercept	-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，时空信息与人称代词和融资结果也是正向关系。词汇复杂性方面，FOG 迷雾指数与融资结果呈负相关，文本多样性指数则与融资呈正相关关系。

4 鲁棒性检测

为了确保结论的准确性和稳定性，采集了更多的项目信息（见表 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 robust linear 回归模型（见表 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（见表 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（见表 9）。

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复杂性 FOG 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，FOG 复杂度则呈现较大的 *P* 值，FOG 指数结果将在下一节做具体分析。

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。

表 6 项目扩展信息表

ID	项目结果	支持数	已筹款	筹资比例	目标筹资
139426	1	83	1513	1.01	1500
116665	1	141	10796	1.08	10000
143752	1	48	5863	1.01	5808
7078	1	380	118833	3.97	30000

表 7 项目结果 robust linear 回归结果表

	Coef	std err	Z	P>|z|	[95.0% Conf]	[Int.]
连词	0.0051	0.004	1.372	0.170	-0.002	0.012
时空信息	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
FOG 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	0.0010	0.019	0.137
intercept	0.3442	0.025	13.705	0.000	0.295	0.393

表 8 筹资比率线性回归结果表

	Coef	std err	Z	P>|z|	[95.0% Conf]	[Int.]
连词	0.0066	0.003	1.909	0.056	-0.002	0.012
时空信息	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
FOG 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	0.0021	0.019	0.137
intercept	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

		连词	时空信息	人称代词	FOG 指数	多样性指数
Logit	Coef	0.0280	0.0593	0.0055	-0.0013	0.3183
Robustlinear	Coef	0.0066	0.0132	0.0013	-0.0003	0.0693
Ratio OLS	Coef	0.0407	0.0124	0.0018	-1.387e-05	0.4224
Logit	P>|z|	0.056	0.000	0.055	0.734	0.011
Robustlinear	P>|z|	0.056	0.000	0.056	0.724	0.021
Ratio OLS	P>|z|	0.010	0.0016	0.000	0.996	0.001

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，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。

在虚构和分离性方面，时空信息和人称指示词与融资成功正相关，这是因为出现越多的

时空信息和第一人称，说明撰写者倾向在描述真实的事情时，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。非第一人称指示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。

对文本词汇复杂性来说，迷雾指数值越大，说明陈述的复杂度高，文本可读性越弱。结果呈负相关，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受到信任。此次验证的结果与理论上不完全符合的，可能有以下两个原因：（1）在实际项目中，复杂性过高会令人晦涩难懂，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是真的，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，将定义过强的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。科技类的文章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，而农业类的文章风格可能比较务实，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，说明文字较为复杂，句式较为丰富，比起那些单一的文本来说，人们会觉得这样的文本是更加可靠的，这也反映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

本文通过实证研究证实了欺诈性语言特征与参与者支持行为负相关，进而显著影响项目筹资结果，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，提出检测模型，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，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，仍旧有所难度。本文提出的指标和监测模型，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的找准项目定位，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，使得筹资失败。为此，可以通过检测指标进行文本分析。计算连词信息，来观察是否完整而连贯的叙述了故事情节；计算时空信息，来观察是否具体细致的回忆了事件所发生的真实的时间地点；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叙述词，使得文本像是虚构的故事；计算文本的复杂性和词汇多样性，来观察是否包含了冗余无用或晦涩难懂的词汇和句子，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，文本能更好的展示发起人的想法和创意，促成项目融资成功。

对众筹网站来说，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的发布自己的项目。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，同时做出一些限制条件和项目善后服务。

7 不足与展望

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，比如农业艺术等不同类别，不同项目的文本描述可能各有偏重，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。今后的研究将在这个方面进一步细化。实验数据选取多是作为训练集进行逻辑回归，如果要进一步进行验证，最好用已经被验证的数据集来进行验证，或者是一批已经被证实是由于欺诈性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试验，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，加强数据样本容量和多样性。

对于数据模型来说，本文的自然语言处理技术在欺诈性线索挖掘中仍需要进一步提高。文本的特征各式各样，此次实验选取的维度还不够宽泛，例如词频和语序对于文本来说也十分重要，比如一句话中转折连词的位置不同就可能导致整段文字呈现截然不同的含义，这些维度也将在日后被考虑进模型中做更加细致的研究。数据处理上，计量模型的多重共线性的解决方案仍旧是一个重要的方面，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严谨的处理这些变量间的关系。

最后，论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理论探讨为什么会产生如此影响，从而进一步分析数据背后所隐含的意义。应当对这一方面再做进一步研究和分析来强化本文结果对于管理实践的指导分析作用。

参 考 文 献

[1] 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

[2] Albert M, Morton W. Decoding of inconsistent communications [J]. Journal of Personality and Social Psychology, 1967, 6 (1): 109-114

[3] Allison T H, Davis B C, Short J C, Webb J W. Crowdfunding in a Prosocial Microlending Environment: Examining the Role of Intrinsic Versus Extrinsic Cues [J]. Entrepreneur Theory and Practice, 2015, 39(2): 53-73.

[4] Benjamin R G. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty [J]. Educational Psychology Review, 2012. 24(1):63-88.

[5] Cholakova M, Clarysse B. Does the Possibility to Make Equity Investments in Crowdfunding Projects Crowd Out Reward-Based [J]. Entrepreneurship Theory and Practice, 2015, 39(1):145-172

[6] DECI E L, RYAN R M. Intrinsic Motivation and Self-Determination in Human Behavior [M]. New York：Plenum Press, 1985

[7] Evans J R, Michael S W, Meissner C A, Brandon S E. Validating a new assessment method for deception detection: Introducing a Psychologically Based Credibility Assessment Tool [J]. Journal of Applied Research in Memory and Cognition, 2013, 2(1): 33-41.

[8] Gao Q, Lin M. Linguistic Features and Peer-to-Peer Loan Quality: A Machine Learning Approach[J]. Available at SSRN 2446114, 2013:1-58.

[9] Gerber E M, Hui J S, Kuo P Y. Crowdfunding: Why people are motivated to post and fund projects on crowdfunding platforms[C]. Proceedings of the International Workshop on Design, Influence, and Social Technologies: Techniques, Impacts and Ethics. 2012

[10] Harter S, Effectance Motivation Reconsidered. Toward a Developmental Model [J]. Human Development. 1978, 21(1):34-64.

[11] Herzenstein M, Sonenshein S, Dholakia U M. Tell me a Good Story and I May Lend you Money: The Role of Narratives in Peer-to-peer Lending Decisions [J]. Journal of Marketing Research, 2011, 48(SPL): 138-149

[12] Knapp M L, Comadena 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

[13] Larrimore L, Jiang L, Larrimore J, Markowitz D, Gorski S. Peer to Peer Lending: The Relationship Between Language Features, Trustworthiness, and Persuasion Success [J]. 2011, 39(1): 19-37.

[14] Lau R Y K, Liao S Y, Kwok R C W, Xu K, Xia Y, and Li Y. 2011. Text mining and probabilistic language modeling for online review spam detection [J]. ACM Transaction of Management Information System, 2012, 2 (4): 1-30

[15] Mollick E R. Swept Away by the Crowd? Crowdfunding, Venture Capital, and the Selection of Entrepreneurs[J]. SSRN, 2013:1-48.

[16] Newman M L, Pennebaker J W, Berry D S. Lying words: Predicting Deception from Linguistic Styles [J]. Personality and Social Psychology Bulletin, 2003, 29(5): 665-675.

[17] Pennebaker J W, Mehl M R, Niederhoffer K G. Psychological Aspects of Natural Language Use: Our Words, Our Selves [J]. Annual review of psychology, 2003, 54(1): 547-577

[18] Durán P, Malvern D, Richards B, Chipere N, Developmental Trends in Lexical Diversity[J]. Applied Linguistics, 2004, 25(2): 220-242

[19] Rebekah George B. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty [J]. Educational Psychology Review, 2012, 24 (1): 63-88

[20] Sarah M, Michael W. Automated insights: verbal cues to deception in real-life high-stakes lies [J]. Psychology Crime & Law, 2015, 21 (7): 617-631

[21] Schwienbacher A, Larralde B. Crowdfunding of small entrepreneurial ventures [M]. Handbook of entrepreneurial finance. Oxford University Press, 2010.

[22] Wang G, Xie S, Liu B, Yu P S. Identify Online Store Review Spammers via Social Review Graph [J]. ACM Transactions on Intelligent Systems & Technology, 2012, 3(4): 1-21

[23] 邓莎莎, 基于欺骗语言线索的虚假评论识别[J]. 系统管理学报, 2014(02): 263-270.

[24] 李焰, 高弋君, 李珍妮, 才子豪, 王冰婷, 杨宇轩. 借款人描述性信息对投资人决策的影响——基于P2P 网络借贷平台的分析 [J]. 经济研究, 2014, (S1): 143-155.

[25] 刘一丹. 语言视角下的欺骗识别研究 [D]. 重庆:西南大学, 2010. 62.

[26] 彭红枫, 赵海燕与周洋, 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. 金融研究 [J], 2016(04): 158-173.

[27] 王会娟, 张路. 借款描述对 P2P 借贷行为的影响研究 [J]. 金融理论与实践, 2014, (8): 34-36.

[28] 吴文清, 付明霞, 赵黎明. 我国众筹成功影响因素及羊群现象研究 [J]. 软科学, 2016, (02): 5-8.

[29] 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. 61.

[30] 张富翠. 广告用语中的欺诈性语言 [J]. 西南民族大学学报（人文社科版）, 2010, 31(7): 146-150.

[31] 张虎等, 基于集成学习的中文文本欺骗检测研究 [J]. 计算机研究与发展, 2015(05): 1005-1013.

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

SHEN Ni1，WANG Hongwei2，WANG Wei3

(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 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

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