借款描述的可读性能够提高网络借款成功率吗

陈　霄，叶德珠，邓　洁

[摘要] 在P2P网络借贷市场中,借款人的借款描述可作为吸引投资者而成功获取借款的一种方式，而投资者能否理解借款描述以及这种理解的程度则取决于借款描述所传递信息的可读性。借款描述的可读性能缓解借款人和投资者之间的信息不对称程度从而提高借款的成功率吗?本文试图运用来自P2P网络借贷平台的经验证据对此进行回答。本文以中国"人人贷"平台2011年1月至2014年5月128532笔订单为研究对象,基于文本复杂性和可理解性构建中文语言环境下的可读性指标，考察借款描述可读性在P2P网络借贷市场上的作用。实证结果发现:①借款描述的可读性有助于减少信息噪音,提高信息甄别速度,从而降低投资者的信息处理成本;②可读性更强的借款描述能向投资者传递积极信号,提高借款的成功率;③借款描述中句均字数每增加一个,借款成功率降低0.4%,而句均词数每增加一个,借款成功率则会降低0.7%;④常用字字数每增加一个,借款成功率提高0.14%,借款违约率下降0.12%,常用词词数每增加一个,借款成功率提高0.28%。上述结论在经过一系列稳健性检验之后依然成立。随着P2P网络借贷平台的去担保化,如何有效利用不可验证的软信息正成为借贷双方面临的问题,本文的研究结论表明,借款描述中的文本可读性能够发挥信号作用并且具有信息含量。

[关键词] P2P网络借贷；　信息不对称；　借款描述；　文本可读性

[中图分类号]F272　[文献标识码]A　[文章编号]1006-480X(2018)03-0174-19

一、引言

可读性(Readability)是文本沟通的基石(Leong et al.,2002),前期有关文本可读性的研究大多集中在财务会计(Li,2008;Loughran and McDonald 2014;Tan et al.,2015)和市场营销领域(Sawyer et al.,2008;Archak et al.,2011;Ludwig et al.,2013),这些研究发现,文本可读性能够显著影响上市公司股价和产品的销售量。那么在中国的P2P网络借贷市场中未经认证的借款描述信息是否能通过可读性的差异来传递并产生增量信息？投资者又是否能够通过文本可读性来识别借款人的质

[收稿日期] 2017-05-05

[基金项目] 国家自然科学基金项目"儒家文化与过度储蓄:微观机理、实证检验与宏观政策研究"(批准号71473102),教育部人文社会科学基金项目"文化、偏好与消费"(批准号13YJA790139)。

[作者简介] 陈霄,暨南大学经济学院博士研究生;叶德珠,暨南大学经济学院教授,博士生导师,管理学博士;邓洁,暨南大学经济学院博士研究生。通讯作者:邓洁,电子邮箱:Jiedeng0905@163.com。感谢匿名评审专家和编辑部的宝贵意见,当然文责自负。

量呢？本文基于中国具有代表性的 P2P 网络借贷平台“人人贷”中的大样本数据，力图回答这一问题，以此进一步扩展 P2P 网络借贷中借款描述的相关研究。

P2P 网络借贷是一种新型的融资方式，它能够允许借款人直接从个人投资者处获取资金。一笔借款的获得通常是由借款人首先在 P2P 网络借贷平台中提供个人有效信息进行注册，在提出借款申请后，借款人需要提供详细个人资料，随后 P2P 网络借贷平台会对这些信息进行审核，在审核通过之后，借款人的借款申请会被建立一个独立的借款页面，借款页面中包含了借款人的个人信息，如年龄、学历和收入等。在这种模式中投资者和借款人都是匿名的，并且不会建立物理上的联系，投资者进行投资决策的信息集来自于借款页面中借款人披露的个人信息，通常情况下中国的 P2P 网络借贷市场中投资者的投资下限为 50 元人民币。在 P2P 网络借贷市场中借贷双方面临的基本问题就是信息不对称问题。

P2P 网络借贷平台为借款人提供一个开放式的文本区域，在该区域内，借款人可以写下任何他们想要表达的信息（Herzenstein et al.，2011；Larrimore et al.，2011；Michels，2012），如为什么需要贷款？为什么对他们的借款进行投资是值得的？或者说为什么他们是能够被信任的？本质上，这是借款人用于说服和吸引投资者进而提高借款成功率的一种方式①。

本文基于复杂性和可理解性构建基于中文语言环境下的文本可读性指标。一方面，复杂的文本信息会产生信息噪音，加剧受众的信息处理成本（Miller，2010）。因此，本文使用借款描述中的字数和词数除以标点符号的数量代表句均含字（词）量，衡量复杂性。另一方面，人们更喜欢容易被加工处理的信息（Reber and Schwarz，1999），因为信息的可理解性越强，处理信息的过程就越流畅，因此，本文基于《现代汉语语料库字（词）频表》中 5708 个汉字和 14629 个词，为常用字（词）语料库，根据借款描述中常用字（词）数衡量借款描述的可理解性。

前期有关 P2P 网络借贷中借款描述的研究大多集中在借款描述的内容和文字的数量方面，国外的研究较为丰富（Herzenstein et al.，2011；Larrimore et al.，2011；Lin et al.，2013；Dorfleitner et al.，2016），国内的相关实证研究甚少（李焰等，2014）。Herzenstein et al.（2011）首次基于 Prosper 的数据探讨了借款描述中所反映出的身份认同（Identity Claim）信息对借贷成功率和借款利率的影响。作者定义了六种身份认同信息，实证研究发现，在借款人提供的借款描述中这些个人身份认同信息的数量越多，借款成功率越高，借款利率同样也会越高。Larrimore et al.（2011）则主要基于金融领域中的文本分析方法，通过国际英语词典预定义了 15 个词组类别，并探讨了这 15 个词组类别对借款成功率的影响。作者发现，扩展、具体和定量性的借款描述能够提高借款人的借款成功率。Gao and Lin（2015）则填补了借款描述的可读性对 P2P 网络借贷的影响在实证研究方面的空白，他们基于迷雾指数（Fox Index）证实，借款人在美国的 P2P 网络借贷平台 Prosper 中所披露的借款描述内容的可读性越高，其借款违约率越小。上述研究都是基于英语语言环境探讨在美国的网络借贷平台 Prosper 中借款描述的作用，Dorfleitner et al.（2016）首次基于德语语言环境，探讨德国网络借贷平台中借款描述的字数和反应借款人情感及社会内涵的八组关键词对借款成功率及违约率的影响。作发现，表现借款人正面情感的词语能够显著提高借款成功率，却对借款违约率没有影响。李焰等（2014）首次探讨了中国文化背景下自愿披露的借款描述对投资者决策的影响，采用质性分析法确定了 8 个借款描述性信息特征如稳定、顾家等，并发现有效特征有利于加快投资人的信息甄别速度，提高借款人的借款成功率。彭红枫等（2016）基于 Prosper 网络借贷平台的数据，考察利率形成机

① 网络借贷平台不会对这些内容进行验证，借款人可以自由撰写任何内容以提高他们借款的可信度，进而提高借款成功率。

制变动前后借款描述对借款成功率和实际借款利率的影响，并发现在利率竞拍模式和固定利率模式下,借款描述能降低借款成本,但未必能增加借款成功率。叶德珠和陈霄(2017)的研究则发现,借款描述的中文文字与标点符号数量能产生增量信息作用,标点符号数每增加十个,借款成功率相对降低 8%,而借款成本则会降低 0.51%。

上述研究为后续探讨借款描述的作用提供了许多借鉴。但仍存在下述局限性:①现有研究大多分析的是英语或德语书写的借款描述,而不同语言之间存在显著差异,因此这些研究的研究方法并不能直接应用于中文书写的借款描述中(Loughran and McDonald,2016),而国内对 P2P 平台中借款描述的实证研究少之又少,且仅限于研究文本内容与长度,尚未有文献关注中文环境中文本可读性在网络借贷中的作用。②样本规模较小。例如 Herzenstein et al.(2011)使用 Prosper 平台中 1493 个借款申请样本为研究对象,并且使用人工判断的方法归纳借款人的个人特征信息,容易产生样本的选择性偏误问题。③测量偏误。由于限定语态用词范畴有限,并且同样的单词在不同语境下的含义存在非常大的差异,因此,容易造成判断偏误(Shiller,1995;Loughran and McDonald,2011)。例如 Gao and Lin(2015)使用迷雾指数来度量借款描述的可读性,而在 P2P 网络借贷市场的借款描述中,一些多音节的单词是比较容易理解的。

本文的边际贡献主要体现在三个方面:①扩展 P2P 网络借贷方面的相关文献。近年来,P2P 网络借贷方面的研究得到了许多金融研究领域内学者们的关注,产生了很多高质量的研究成果,并发表在国际顶级刊物上,对 P2P 网络借贷领域的研究产生了很大影响(Duarte et al.,2012; Michels,2012;Liu et al.,2015;Dorfleitner et al.,2016)。本文进一步扩展了 P2P 网络借贷领域的文献,探讨借款描述中的文本可读性对借贷结果的影响，有助于理解借款描述与投资者和借款人行为之间的关联。②扩展文本分析的相关文献。文本分析的使用,大量集中在会计学领域。这些研究发现,会计文本中的语调和可读性等特征可以反映高管的心理特征或者行为动机等状态(Huang et al.,2014;Law and Mills,2015;Loughran and McDonald,2016)。本文尝试性地构建基于中文语言环境下的文本可读性指标,进一步扩展了文本分析法的应用范围。③本文的研究结果对于 P2P 网络借贷市场参与者而言具有重要参考作用:对借款人而言,在书写借款描述时应使用言简意赅和通俗易懂的表达方式;对投资者而言,在 P2P 网络借贷市场中投资那些借款描述可读性更强的订单至少不会提高他们的投资风险。2008 年的国际金融危机暴露了贷款决策标准的缺陷,定量的信用分数已经被证明是一种不可靠的预测消费者偿还无抵押贷款的方式，本文则证实借款描述的可读性能够作为评估借款人的重要指标,有助于探索中国 P2P 网络借贷市场的发展规律。

二、研究假设

在 P2P 网络借贷市场中,投资者和借款人之间的信息不对称程度更加严重:①P2P 网络借贷市场的借贷双方都是匿名的,因此他们并不清楚对方的真实身份,这无疑强化了交易中的不确定性。②投资者无法通过其他途径获取借款人更多的其他信息，只能依靠借款人在 P2P 网络借贷平台上披露的信息进行决策,而借款人有可能只选择披露那些有利于自己获取贷款的信息,同时这些被披露的信息也有可能是假的。③借款人的声誉信号只能发挥有限的作用。在中国的 P2P 网络借贷平台上,借款人的信用等级是平台方所给出的,而不是由中国人民银行提供的官方信用评价。毫无疑问,投资者会关注任何有助于他们甄别优质借款人的信息。Gao and Lin(2015)对美国网络借贷平台 Prosper 的研究发现,投资者将借款人的借款描述视为重要的参考依据。借款描述不仅能通过影响投资者对信息的理解影响投资者的行为,而且其内容还能够反映借款人的个体特征。

借款描述可读性能够影响投资者的阅读时间。当文本的可读性较强时,读者更容易以较快速度浏览并理解这段文本的意义和目的(McKeown et al.,1992),而当借款描述的可读性较差时,投资者有可能会放弃阅读。Kahneman and Tversky(1973)指出,人在特定环境下的注意力容量是有限的,当人们面对大量信息的时候,不得不对信息进行筛选,只能够选择性地阅读。Peng and Xiong(2006)的研究同样也证实投资者的注意力是有限的。投资者在 P2P 网络借贷市场中投资数额较小,他们可能并不愿意花费大量时间去阅读可读性较差的借款描述。因此,可读性较差的文本可能会使借款描述失去其吸引和说服投资者以促成借款的本质意义。

可读性是一个决定任何文本能否被有效理解的关键因素 (Rameezdeen and Rajapakse,2007)。Loughran and McDonald(2014)对美国上市公司年报的研究发现,投资者更可能投资年报可读性更强的公司。同样,Rennekamp(2012)和 Tan et al.(2014)基于实验研究的方法证明,文本可读性能够对小额投资者的判断产生显著影响。因为根据易于加工(Ease of Processing)理论,人们看起来更加喜欢容易被加工处理的信息(Reber and Schwarz,1999),因此,信息的可读性越强,投资者处理信息的过程就会越流畅,这也暗示着此类信息披露更加可靠和更值得信赖。大量的研究表明,易于阅读的文本提高了阅读者的理解速度及阅读的持续性(Klare,1976)。在 P2P 网络借贷中,大多数投资者是非专业的中小投资者,因此对于文本的阅读能力也是有限的,而借款描述能够起到有效沟通借贷双方的前提在于它能够被理解,而能够被理解的前提在于具备可读性。根据上述分析,本文提出:

假说 1:可读性越强的借款描述,借款成功率越高。

借款描述可读性能够反映借款人的个人特征。以往的研究已经考察了上市公司年报的可读性与公司业绩的关系。Li(2008)基于上市公司年报迷雾指数(Fog Index)的研究发现,年报可读性更强的公司会表现出更强的盈利能力。Abrahamson and Amir(1996)的研究则发现,上市公司管理层会在当前的信息披露中提供更多有关未来盈余的增量信息来克服信息不对称问题。混淆假说(Obfuscation Hypothesis)认为,表现不佳的公司会通过采取降低文本的可读性和使用大量不必要的复杂词汇的方法有意掩盖其文本信息的真正内容(Abu Bakar and Ameer,2011),因为资本市场会根据公司的信息做出延迟反应。所以管理层有可能操纵年报的可读性,尽量让更多的投资者难以发现管理层不想披露的信息(Bloomfield,2008)。Li(2008)、Loughran and McDonald(2014)和 Tan et al.(2015)对上市公司年报的研究均发现,当公司业绩不佳时,管理层会加长年报的篇幅,增加很多无关的信息,让投资者不得要领,而当公司业绩较好时则会言简意赅地说明公司的运营情况。

上市公司在年报中的文本可读性可以反映其财务能力,这一逻辑已被证实。那么在个人借贷市场中,本文同样可以借鉴上述逻辑推断,即借款人借款描述中的文本可读性同样能反映他们的财务和还本付息能力,当借款人还款能力强、信用质量高时,他们会使用浅显易懂的表达方式让投资者了解自己的真实情况。此外,可读性较差的文本也可能是因为书写文本的借款人本身的教育程度较低,而受教育程度越低的借款人,一般越难得到高收入的稳定工作,进而越不利于他们的债务偿还。Tausczik and Pennebaker(2010)则直接指出,文本的可读性能反映作者的教育、社会地位以及社会阶层,受教育程度较低的作者所书写的文本的可读性较受教育程度较高的作者而言更低,因此,借款描述可读性更强的文本更有可能是由受过良好教育且具有较高稳定收入的借款人所书写, 他们有着更强的还款能力(Campbell and Dietrich,1983;De Gregorio and Lee,2002)。根据上述分析,本文提出:

假说 2:可读性越强的借款描述,不会显著提高借款违约率。

三、研究设计

1. 可读性的衡量指标

为了研究文本可读性对借款成功率和借款违约率的影响，本文必须构建适合中文环境下的可读性指标。国外学者们共提出了大约 70 种测度方式(Klare,2000),而在金融学领域,影响力最广泛的就属迷雾指数(Fog Index)和字典法。迷雾指数(Fog Index)的构建公式为:0.4(Words/Sentences+100×Hard Words/Words)。这一指数由 Gunning(1969)首次提出,Li(2008)将该指数应用于上市公司年报的分析。迷雾指数(Fog Index)的构成总共有两个部分:①Words/Sentences,为每句话中单词的占比;②100×Hard Words/Words,为每百字中复杂单词的占比,复杂单词为音节数大于 2 个及以上的单词。但是这种分析方法是基于英文环境之中,由于中英文语言之间的差异性较为显著,因此上述衡量文本可读性的方法,无法直接应用于非英文书写的文本环境中。尽管迷雾指数(Fog Index)在金融学领域中的使用非常广泛,但是其准确性依然存在许多争议,在英文的环境中有许多单词虽然超过两个音节,却非常容易理解,例如,Loughran and McDonald(2014)构建了一个上市公司年报的常用字字典列表,这一常用字字典列表被广泛应用于金融学领域的文本分析。但是由于中文和英文之间的差异,直接翻译该字典来创建中文字典并不合适。

本文结合迷雾指标与词袋法(即字典法)构建了一个适用于中文环境下的文本分析方法,在引用成熟字典分析词义的基础上,期望能较有效地度量文本内容的复杂性。文本可读性指标将由复杂性和可理解性两部分组成:

(1)复杂性。人在特定环境下的注意力容量是有限的(Kahneman and Tversky,1973),根据 Miller(1956)提出的“7±2 原则”,人脑会将复杂信息分块,并且在短期内一般只能够记住 5—9 个事物。复杂的文本信息会产生信息噪音,加剧受众的信息处理成本(Miller,2010)。前期衡量可读性的方法如迷雾指数(Fog Index)等,都用到了一个指标,那就是平均每句话的长度(ASL)。微软公司在 Microsoft Word 的程序中也提供了测量可读性的功能,在报告文档可读性指标中,也包含每句话的字符数。由每句话的信息含量在可读性测度中的应用普遍性可见,显著影响读者从文本中获取所需信息的一个重要因素在于文本书写的连贯性(Vauras et al.,1992)。Roux(2008)则发现,标点符号的数量显著影响文本内容和句子长度,本文使用每个停顿处的含字(词)量来衡量文本的复杂性:①本文使用中文分词软件对借款人的借款描述信息进行中文分词处理。②标点符号能够起到分割文本的作用(Shriberg et al., 2000),对此,本文分别统计每个借款描述中所有表示句尾和停顿的标点符号,这些符号包括:句号(。)、问号(?)和感叹号(!),表示句尾;逗号(,)、冒号(:)、分号(;)和顿号(、),表示停顿。③利用分词软件统计汉字数和词数,除以上述标点符号的加总,计算每次停顿之间字词的含量。本文用汉字数除以句尾和停顿符的加总(*PC_ChiWord*)衡量句均的含字量,词数除以句尾和停顿符的加总(*PC_Voc*)。用这两个指标衡量文本的复杂性,该项指标越大,文本可读性越弱。

(2)可理解性。使用由国家教育部发布的《现代汉语语料库字(词)频表》作为文本可读性中的“可理解性”的组成部分。《现代汉语语料库字频表》的语料规模为 2000 万汉字,其根据出现的次数,囊括了 5708 个汉字为常用字。《现代汉语语料库词频表》的语料规模为 2000 万汉字,其根据出现的次数,囊括了 14629 个词为常用词。本文根据借款人使用《现代汉语语料库字频表》中的字数来衡量借款描述中的常用字(*ComWord*),根据借款人使用《现代汉语语料库词频表》中的词数来衡量借款描述中的常用词(*ComVoc*),这两个指标衡量文本的可理解性,该项指标越大,文本可读性越强。

为了更好阐明本文所构建的文本可读性指标,以“人人贷”网络借贷平台中的一条借款描述为

例加以说明:“年关将至,又一年过去了,回望这一年真的没什么成绩,都过年了,今年春节要换的东西太多,也该孝敬年老的双亲,希望能得到人人贷的资金让我自己能过个轻松年,来年再努力……”,该文本信息中共有 82 个字数,其中包括 10 个标点符号,82 个字,44 个词。经测度,该文本信息的复杂性指标中句均字数(*PC_ChiWord*)为 8.2(82/10),句均词数(*PC_Voc*)为 4.4(44/10),可理解性指标中常用字(*ComWord*)为 82,常用词(*ComVoc*)为 34。

2. 数据选取与网站特征描述

本文使用的数据来自“人人贷”网络借贷平台。选择“人人贷”的原因有三个方面:①“人人贷”在中国的市场知名度较高。“人人贷”是中国最大的网络借贷平台之一,成立于 2010 年 5 月,其服务已经覆盖全国 30 余个省的 2000 多个地区、几十万名客户,为行业内最具影响力的品牌之一。②“人人贷”在中国的市场认可度较高,2014 年和 2015 年连续两年被中国互联网协会和中国社会科学院评为“AAA 级(最高级)民营网络借贷企业”。2015 年,“人人贷”还入选由中国互联网协会与中华人民共和国工业和信息化部联合发布的 2015 年“中国互联网企业 TOP100”排行榜。③“人人贷”平台符合 P2P 网络借贷的基本特征。在“人人贷”网络借贷平台中,投资者和借款人不产生物理上的联系,是一种典型的 Peer-to-Peer 网络借贷模式。本文以 2011 年 1 月 1 日至 2014 年 5 月 31 日“人人贷”网站上发布的全部借款订单作为初始样本,并对样本做如下处理:①剔除机构担保和实地认证的订单样本;②剔除借款人所在地为中国香港、中国澳门和中国台湾的样本;③为排除极端值的影响,对借款人的年龄和借款金额在上下 1%的水平上进行缩尾处理。最终样本为 128532 个,其中,成功获得借款的样本为 9037 个,在成功获得借款之后没有按时还款的样本 719 个。

3. 变量设定

借鉴相关文献的研究设计(Michels,2012),本文将借款人的各项基本信息分为订单因素、个人因素和宏观因素。订单因素包括借款利率、借款金额、借款期限、信用等级等;个人因素包括年龄、婚姻状况、收入、受教育水平、房产、车产和工作时间等;宏观因素包括年度、地区和工作行业。各项指标具体说明如表 1 所示。

4. 变量特征分析

本文将上述变量的数据进行描述统计分析,结果经整理如表 2 和表 3 所示。表 2 为信息披露指标的描述性统计分析。从中可以看到,在文本可读性四项指标中,平均每个借款描述中句均含字量(*PC_ChiWord*)为 13.8,句均含词量(*PC_Voc*)为 7.66,所使用的常用字(*ComWord*)为 56.38,常用词(*ComVoc*)为 21.54;在所有样本中,借款人的借款成功率为 7%,而违约率为 7.9%,意味着市场中的违约风险较高;平均利率水平为 16.03%,平均借款金额为 57000 元,可以认为市场对小额借贷资金的需求较为迫切。借款人普遍信用等级较低,借款人的平均信用等级为 1.19;借款人的年龄平均不到 33 岁,表明在 P2P 网络借贷市场中,大部分有资金需求的借款人为年轻人;48.3%的借款人已婚,41%的借款人拥有房产,23.6%的借款人拥有车产。表 3 是文本可读性的差异性检验。从 Panel A 中可知,在成功获取借款的样本中,借款描述的文本复杂性均要小于失败借款的样本,并且,文本的可理解性均要大于失败借款的样本。从 Panel B 中可知,在违约的样本中,借款描述的文本复杂性与按时还款的样本并未存在显著差异,但是,文本的可理解性均要小于按时还款的样本。表 4 是文本可读性的相关系数矩阵。从 Panel A 中可知,句均含字(词)量均与借款成功率呈显著的负相关,而常用字(词)量与借款成功率呈显著的正相关。从 Panel B 中可知,句均含字量均与借款违约率呈显著正相关,而句均含词量与借款违约率的相关性并不显著。另外,常用字(词)量与借款违约率呈显著负相关性。

表 1　　变量定义

变量	变量含义
SUCCESS	借款人在平台中发布的借款订单,当借款成功时为 1,否则为 0
DEFAULT	借款人在借款成功后,未能按时偿还借款为 1,否则为 0
PC_ChiWord	借款描述中的汉字数除以句号、感叹号、问号、逗号、顿号、冒号和分号的数量之和,衡量文本的复杂性
PC_Voc	借款描述中的词数除以句号、感叹号、问号、逗号、顿号、冒号和分号的数量之和,衡量文本的复杂性
ComWord	借款描述的汉字中,《现代汉语语料库字频表》所包含的字数,衡量文本的可理解性
ComVoc	借款描述的词语中,《现代汉语语料库词频表》所包含的词数,衡量文本的可理解性

表 2　　变量的描述性统计

变量	变量名	平均值	标准差	最小值	最大值	样本量
SUCCESS	是否借款成功	0.070	0.256	0	1	128532
DEFAULT	是否违约	0.079	0.271	0	1	9037
PC_ChiWord	句均字数	13.800	17.275	0	432	128532
PC_Voc	句均词数	7.660	9.580	0	276	128532
ComWord	常用字字数	56.380	40.453	0	490	128532
ComVoc	常用词字数	21.540	16.850	0	300	128532
INTEREST	借款利率	16.030	3.921	3	24.400	128532
AMOUNT	借款金额	57000	98000	3000	500000	128532
MONTHS	借款期限	12.300	7.951	1	36	128532
CREDIT	借款人信用等级	1.190	0.816	1	7	128532
AGE	借款人年龄	32.350	6.603	23	54	128532
MARRIED	是否已婚	0.483	0.500	0	1	128532
INCOME	借款人收入水平	3.904	1.263	1	7	128532
EDUCATION	借款人教育水平	1.800	0.794	1	4	128532
HOUSE	借款人是否有房产	0.411	0.492	0	1	128532
CAR	借款人是否有车产	0.236	0.424	0	1	128532
WORKTIME	借款人工作时间	2.395	1.009	1	4	128532
Region_East	借款人来自东部地区	0.592	0.491	0	1	128532
Region_West	借款人来自西部地区	0.128	0.334	0	1	128532
Region_Northeast	借款人来自东北地区	0.062	0.241	0	1	128532
Region_Middle	借款人来自中部地区	0.218	0.413	0	1	128532
YEAR	借款年份	2013	1.006	2011	2014	128532
YEAR=2011	借款年份为 2011	0.152	0.359	0	1	128532
YEAR=2012	借款年份为 2012	0.212	0.409	0	1	128532
YEAR=2013	借款年份为 2013	0.378	0.485	0	1	128532
YEAR=2014	借款年份为 2014	0.257	0.437	0	1	128532

表 3　　差异性检验

Panel A					
变量	*SUCCESS*=0	平均值 1	*SUCCESS*=1	平均值 2	平均差
PC_ChiWord	119495	13.98	9037	11.45	2.52***
PC_Voc	119495	7.77	9037	6.35	1.42***
ComWord	119495	55.59	9037	66.94	−11.35***
ComVoc	119495	21.30	9037	24.82	−3.52***
Panel B					
变量	*DEFAULT*=0	平均值 1	*DEFAULT*=1	平均值 2	平均差
PC_ChiWord	8318	11.41	719	11.89	−0.48
PC_Voc	8318	6.33	719	6.51	−0.18
ComWord	8318	67.35	719	62.15	5.20**
ComVoc	8318	24.96	719	23.19	1.77**

注：***、**、* 分别表示在 1%、5%、10%的水平上显著，括号内为 Z 统计值。

表 4　　相关系数矩阵

Panel A					
变量	*SUCCESS*	*PC_ChiWord*	*PC_Voc*	*ComWord*	*ComVoc*
SUCCESS	1.0000				
PC_ChiWord	−0.0374*	1.0000			
PC_Voc	−0.0379*	0.9590*	1.0000		
ComWord	0.0717*	0.1328*	0.1381*	1.0000	
ComVoc	0.0534*	0.1768*	0.2181*	0.8834*	1.0000
Panel B					
变量	*DEFAULT*	*PC_ChiWord*	*PC_Voc*	*ComWord*	*ComVoc*
DEFAULT	1.0000				
PC_ChiWord	0.0083*	1.0000			
PC_Voc	0.0054	0.9721*	1.0000		
ComWord	−0.0265*	0.1085*	0.1161*	1.0000	
ComVoc	−0.0232*	0.1703*	0.1983*	0.9108*	1.0000

注：***、**、* 分别表示在 1%、5%、10%的水平上显著，括号内为 Z 统计值。

四、实证分析

1. 借款描述可读性能够影响借款成功率吗

在变量的特征分析部分，本文发现借款描述中的可读性指标与借款成功率存在显著关系，但无法据此下定结论，即可读性越强的借款描述越能够显著提高借款人的借款成功率，因为它可能受到其他因素的影响。因此，为验证假说 1，可读性更强的借款描述是否能够对借款成功率产生积极影响，本文构建如下模型：

$$SUCCESS_i=\beta_0+\beta_1 Readability\_Var_i+\beta_i Control_i+\varepsilon_i \quad (1)$$

公式(1)中 $SUCCESS_i$ 为借款人是否成功获得借款的虚拟变量，当借款人成功获得借款时取 1，否则为 0；$Readability\_Var_i$ 是借款人借款描述的可读性指标，分别为句均含字(词)量(复杂性)和常用字(词)数(可理解性)。控制变量包括订单、个人和宏观因素，ε_i 为随机干扰项。在上式中，本文需

要通过判断 $Readability\_Var_i$ 指标前的系数 β_1 是否显著来证明假说 1。表 5 列示了借款描述的可读性与借款成功率的 Logit 回归结果及其边际效应。

表 5 第(1)列是没有加入借款描述可读性时各变量对借款成功率的影响,从中可以发现,借款人信用等级、年龄、收入和受教育程度越高的借款订单,借款成功率越高,而借款利率和借款金额越高、借款期限越长的订单,借款成功率越低。这些结论与前期文献的实证结果一致(Herzenstein et al.,2011;Duarte et al.,2012;Burtch et al.,2014;Dorfleitner et al.,2016;Chen et al.,2018)。实证结果表明,利率与借款成功率之间的关系显著为负,这个负向影响的结果符合信贷理论中的信贷配给理论的逻辑。信贷配给理论认为,借款人在向银行借款时,可能存在逆向选择问题,即提出借款利率非常高的借款人,可能根本上就不准备还款,或者说借款利率高的借款人,违约率更高。出借方为防止此类借款人得到贷款,在实际贷款过程中,并不按照利率高低来配置贷款,即对那些愿意支付非常高的利率的借款人,反而不愿意贷款给他们。最终的结果是,借款人愿意支付的利率越高,越不容易得到贷款。

表 5 的第(2)、(4)、(6)和(8)列列示的是,分别加入 *PC_ChiWord*、*PC_Voc*、*ComWord* 和 *ComVoc* 后借款描述的可读性对借款成功率的影响。从中可以发现,*PC_ChiWord* 和 *PC_Voc* 项系数项符号为负,并且在 1%的置信水平上显著,而 *ComWord* 和 *ComVoc* 项系数项符号为正,并且在 1%的置信水平上显著。

表 5 的第(3)、(5)、(7)和(9)列分别列示的是对应第(2)、(4)、(6)和(8)列的边际效应。由第(3)列的结果看,*PC_ChiWord* 项的边际效应为–0.0003,并且在 1%的置信水平上显著,这意味着在控制其他因素的条件下,借款描述中的句均含字量每增加一个,借款成功率将会降低 0.4%(–0.0003/0.07)。由第(5)列的结果看,*PC_Voc* 项的边际效应为–0.0005,并且同样在 1%的置信水平上显著,这意味着在控制其他因素的条件下,借款描述中的句均含词量每增加一个,借款成功率将会降低 0.7%(–0.0005/0.07)。由第(7)列的结果看,*ComWord* 项的边际效应为 0.0001,并且同样在 1%的置信水平上显著,这意味着在控制其他因素的条件下,借款描述中的常用字字数每增加一个,借款成功率将会提高 0.14%(0.0001/0.07)。由第(9)列的结果可知,*ComVoc* 项的边际效应为 0.0002,并且同样在 1%的置信水平上显著,这意味着在控制其他因素的条件下,借款描述中的常用词词数每增加一个,借款成功率将会提高 0.28%(0.0002/0.07)。实证结果证明假说 1 成立。

2. 投资者做出了正确决策吗

借款描述一方面通过影响投资者对信息的理解影响投资者行为,另一方面,借款描述中表达的内容还能够反映借款人的特征(Gao and Lin,2015)。那么投资者是否正确识别这些内容背后的特征呢?换句话说,文本可读性更强的订单是否缓解了借贷双方的信息不对称进而减少借款人的道德风险?如果投资者做出的判断是正确的,那么文本可读性更强的订单至少不会显著提高借款违约率,因此,为验证假说 2,本文构建如下模型:

$$DEFAULT_i=\beta_0+\beta_1 Readability\_Var_i+\beta_i Control_i+\varepsilon_i \quad (2)$$

公式(2)中 $DEFAULT_i$ 为借款人在获得借款之后是否未按时还款虚拟变量,当借款人未能按时还款时取 1,否则为 0;$Readability\_Var_i$ 是借款人借款描述的可读性指标,分别为句均含字(词)量(复杂性)和常用字(词)数(可理解性)。控制变量包括订单、个人和宏观因素,ε_i 为随机干扰项,在上式中,本文需要通过判断 $Readability\_Var_i$ 指标前的系数 β_1 是否显著,表 6 列示了借款描述的可读性与借款违约率的 Logit 回归结果及其边际效应。

表 5　　借款描述的可读性与借款成功率

	(1)	(2)	(3)	(4)	(5)
	SUCCESS	*SUCCESS*	*SUCCESS*	*SUCCESS*	*SUCCESS*
PC_ChiWord		−0.0075***	−0.0003***		
		(−6.30)	(−6.29)		
PC_Voc				−0.0128***	−0.0005***
				(−5.70)	(−5.69)
INTEREST	−0.0692***	−0.0695***	−0.0028***	−0.0693***	−0.0028***
	(−21.85)	(−21.90)	(−21.70)	(−21.85)	(−21.65)
lnAMOUNT	−0.4319***	−0.4339***	−0.0176***	−0.4338***	−0.0176***
	(−34.41)	(−34.49)	(−32.21)	(−34.49)	(−32.21)
MONTHS	−0.0217***	−0.0216***	−0.0009***	−0.0216***	−0.0009***
	(−9.32)	(−9.23)	(−9.21)	(−9.22)	(−9.20)
CREDIT	0.9860***	0.9830***	0.0399***	0.9830***	0.0399***
	(58.24)	(58.18)	(65.40)	(58.18)	(65.39)
AGE	0.0392***	0.0385***	0.0016***	0.0385***	0.0016***
	(18.02)	(17.65)	(17.43)	(17.65)	(17.43)
MARRIED	0.2590***	0.2544***	0.0103***	0.2544***	0.0103***
	(7.48)	(7.35)	(7.34)	(7.35)	(7.34)
INCOME	0.3915***	0.3916***	0.0159***	0.3910***	0.0159***
	(30.64)	(30.61)	(29.91)	(30.56)	(29.86)
EDUCATION	0.1787***	0.1730***	0.0070***	0.1733***	0.0070***
	(9.80)	(9.48)	(9.40)	(9.50)	(9.41)
HOUSE	−0.0707**	−0.0686**	−0.0028**	−0.0684**	−0.0028**
	(−2.12)	(−2.06)	(−2.06)	(−2.05)	(−2.05)
CAR	0.2650***	0.2675***	0.0109***	0.2669***	0.0108***
	(7.82)	(7.88)	(7.87)	(7.86)	(7.85)
WORKTIME	0.2589***	0.2602***	0.0106***	0.2596***	0.0105***
	(16.90)	(16.97)	(16.76)	(16.94)	(16.73)
cons	−3.3459***	−3.1783***		−3.1845***	
	(−22.52)	(−21.03)		(−21.04)	
年份	YES	YES	YES	YES	YES
行业	YES	YES	YES	YES	YES
地区	YES	YES	YES	YES	YES
N	128532	128532	128532	128532	128532
r2_p	0.3818	0.3828		0.3827	
	(6)	(7)	(8)	(9)	
	SUCCESS	*SUCCESS*	*SUCCESS*	*SUCCESS*	
ComWord	0.0026***	0.0001***			
	(8.47)	(8.46)			
ComVoc			0.0052***	0.0002***	
			(6.76)	(6.76)	
INTEREST	−0.0699***	−0.0028***	−0.0700***	−0.0028***	
	(−22.09)	(−21.88)	(−22.11)	(−21.90)	
lnAMOUNT	−0.4379***	−0.0178***	−0.4354***	−0.0177***	
	(−34.74)	(−32.44)	(−34.59)	(−32.31)	
MONTHS	−0.0219***	−0.0009***	−0.0219***	−0.0009***	
	(−9.36)	(−9.34)	(−9.38)	(−9.36)	

(续表 5)

	(6) SUCCESS	(7) SUCCESS	(8) SUCCESS	(9) SUCCESS	
CREDIT	0.9836***	0.0399***	0.9845***	0.0399***	
	(58.07)	(65.31)	(58.18)	(65.45)	
AGE	0.0385***	0.0016***	0.0387***	0.0016***	
	(17.65)	(17.43)	(17.77)	(17.54)	
MARRIED	0.2617***	0.0106***	0.2607***	0.0106***	
	(7.55)	(7.54)	(7.52)	(7.51)	
INCOME	0.3888***	0.0158***	0.3903***	0.0158***	
	(30.37)	(29.68)	(30.52)	(29.82)	
EDUCATION	0.1761***	0.0071***	0.1768***	0.0072***	
	(9.66)	(9.58)	(9.69)	(9.60)	
HOUSE	−0.0697**	−0.0028**	−0.0694**	−0.0028**	
	(−2.09)	(−2.09)	(−2.08)	(−2.08)	
CAR	0.2603***	0.0106***	0.2623***	0.0106***	
	(7.67)	(7.66)	(7.74)	(7.72)	
WORKTIME	0.2613***	0.0106***	0.2611***	0.0106***	
	(17.03)	(16.82)	(17.04)	(16.82)	
cons	−3.3437***		−3.3550***		
	(−22.49)		(−22.57)		
年份	YES	YES	YES	YES	
行业	YES	YES	YES	YES	
地区	YES	YES	YES	YES	
N	128532	128532	128532	128532	
r2_p	0.3829		0.3825		

注:***、**、* 分别表示在 1%、5%、10%的水平上显著,括号内为 Z 统计值。

表 6 第(1)列列示的是未加入借款文本可读性时前文所提及的变量对借款违约率的影响。从中可发现,借款利率越高、借款金额越大、借款期限越长和借款人年龄越大的订单,违约率也越高。信用等级越高、受教育程度越高和工作年限越长的订单,其借款违约率会越小。

表 6 的第(2)、(4)、(6)和(8)列分别列示的是加入 *PC_ChiWord*、*PC_Voc*、*ComWord* 和 *ComVoc* 后借款描述的可读性对借款违约率的影响。从中可发现,*PC_ChiWord* 和 *PC_Voc* 项系数项符号为正,而 *ComWord* 和 *ComVoc* 项系数项符号为负,但是仅有 *ComWord* 项系数在 10%的置信水平上显著。这意味着借款描述中仅有常用字使用量对借款违约率有显著影响,常用字使用量越高,借款人的借款违约率越小。表 6 的第(3)、(5)、(7)和(9)列分别是对应(2)、(4)、(6)和(8)列的边际效应。由第(7)列的结果看,*ComWord* 项的边际效应为−0.0001,并且在 10%的置信水平上显著,这意味着在控制其他因素的条件下,借款描述中的常用字字数每增加一个,借款违约率将会下降 0.12%(0.0001/0.079)。综上,借款描述可读性对借款违约率不存在显著影响,即证实本文假说 2 成立,使用浅显易懂表达方式的借款人并没有显著地提高借款违约率。

五、稳健性检验

1. 改变被解释变量

本文实证分析的前提在于投资者会阅读这些借款描述信息,Gao and Lin (2015) 基于 Prosper

表 6　　借款描述的可读性与借款违约率

	(1)	(2)	(3)	(4)	(5)
	DEFAULT	*DEFAULT*	*DEFAULT*	*DEFAULT*	*DEFAULT*
PC_ChiWord		0.0004	0.0000		
		(0.13)	(0.13)		
PC_Voc				0.0012	0.0001
				(0.24)	(0.24)
INTEREST	0.1377***	0.1377***	0.0077***	0.1377***	0.0077***
	(8.11)	(8.11)	(8.21)	(8.11)	(8.21)
lnAMOUNT	0.4757***	0.4757***	0.0265***	0.4757***	0.0265***
	(8.19)	(8.19)	(8.35)	(8.20)	(8.35)
MONTHS	0.0640***	0.0640***	0.0036***	0.0640***	0.0036***
	(8.34)	(8.34)	(8.50)	(8.34)	(8.51)
CREDIT	−2.8920***	−2.8921***	−0.1613***	−2.8922***	−0.1613***
	(−8.58)	(−8.58)	(−8.62)	(−8.58)	(−8.62)
AGE	0.0283***	0.0283***	0.0016***	0.0283***	0.0016***
	(3.67)	(3.67)	(3.69)	(3.67)	(3.69)
MARRIED	0.0181	0.0182	0.0010	0.0182	0.0010
	(0.16)	(0.16)	(0.16)	(0.16)	(0.16)
INCOME	0.0596	0.0595	0.0033	0.0596	0.0033
	(1.46)	(1.46)	(1.46)	(1.46)	(1.46)
EDUCATION	−0.4068***	−0.4064***	−0.0227***	−0.4062***	−0.0227***
	(−6.45)	(−6.44)	(−6.45)	(−6.44)	(−6.44)
HOUSE	−0.0977	−0.0974	−0.0054	−0.0971	−0.0054
	(−0.98)	(−0.97)	(−0.97)	(−0.97)	(−0.97)
CAR	−0.1425	−0.1430	−0.0080	−0.1435	−0.0080
	(−1.33)	(−1.33)	(−1.33)	(−1.34)	(−1.34)
WORKTIME	−0.0991*	−0.0991*	−0.0055*	−0.0990*	−0.0055*
	(−1.83)	(−1.83)	(−1.84)	(−1.83)	(−1.83)
cons	−6.2559***	−6.2605***		−6.2655***	
	(−8.51)	(−8.52)		(−8.53)	
年份	YES	YES	YES	YES	YES
行业	YES	YES	YES	YES	YES
地区	YES	YES	YES	YES	YES
N	9037	9037	9037	9037	9037
r2_p	0.3539	0.3539		0.3539	

	(6)	(7)	(8)	(9)	
	DEFAULT	*DEFAULT*	*DEFAULT*	*DEFAULT*	
ComWord	−0.0016*	−0.0001*			
	(−1.69)	(−1.69)			
ComVoc			−0.0031	−0.0002	
			(−1.20)	(−1.20)	
INTEREST	0.1395***	0.0078***	0.1392***	0.0078***	
	(8.19)	(8.29)	(8.17)	(8.28)	
lnAMOUNT	0.4812***	0.0268***	0.4781***	0.0267***	
	(8.26)	(8.42)	(8.22)	(8.38)	
MONTHS	0.0637***	0.0036***	0.0638***	0.0036***	
	(8.31)	(8.47)	(8.32)	(8.48)	
CREDIT	−2.8955***	−0.1614***	−2.8941***	−0.1614***	
	(−8.55)	(−8.60)	(−8.56)	(−8.60)	

(续表 6)

	(6) *DEFAULT*	(7) *DEFAULT*	(8) *DEFAULT*	(9) *DEFAULT*	
AGE	0.0290***	0.0016***	0.0288***	0.0016***	
	(3.76)	(3.78)	(3.73)	(3.75)	
MARRIED	0.0202	0.0011	0.0204	0.0011	
	(0.18)	(0.18)	(0.18)	(0.18)	
INCOME	0.0606	0.0034	0.0601	0.0033	
	(1.48)	(1.48)	(1.47)	(1.47)	
EDUCATION	−0.4068***	−0.0227***	−0.4066***	−0.0227***	
	(−6.45)	(−6.46)	(−6.46)	(−6.47)	
HOUSE	−0.0950	−0.0053	−0.0967	−0.0054	
	(−0.95)	(−0.95)	(−0.97)	(−0.97)	
CAR	−0.1414	−0.0079	−0.1411	−0.0079	
	(−1.32)	(−1.32)	(−1.31)	(−1.31)	
WORKTIME	−0.1021*	−0.0057*	−0.1025*	−0.0057*	
	(−1.89)	(−1.89)	(−1.89)	(−1.89)	
cons	−6.2804***		−6.2619***		
	(−8.53)		(−8.51)		
年份	YES	YES	YES	YES	
行业	YES	YES	YES	YES	
地区	YES	YES	YES	YES	
N	9037	9037	9037	9037	
r2_p	0.3545		0.3543		

注:***、**、* 分别表示在 1%、5%、10%的水平上显著,括号内为 Z 统计值。

平台中的两次自然实验证实,当 Prosper 取消提供借款描述信息后,投资者人数显著下降,因此投资者在决策中会将借款描述视为重要参考依据。而对“人人贷”网络借贷平台而言并没有符合这种自然实验的外部冲击,对此,本文将投资者人数视为被解释变量,通过考察句均字(词)数和常用字(词)数是否对投资者人数产生显著影响进而从侧面判断投资者是否阅读了这些信息。从回归结果看,仅句均字数对投资者人数的影响不显著,其他可读性变量均能对投资者人数产生显著影响。因此可以认为本文的逻辑前提是成立的,即投资者在进行投资决策时会参考借款人提供的借款描述信息。①

2. 改变样本

(1)在实证分析中,本文假设借款人是理性人,借款人在书写借款描述的过程中不存在非正式的或者随意的表达方式,例如使用大量标点符号或者网络语言等。在稳健性检验中,本文剔除这部分存在异常的借款描述:①剔除借款描述中标点符号数量大于或等于字数的样本,以及标点符号数量等于总字符数的样本,并重新对借款成功率和借款违约率进行回归。②从标点符号数量分布看,与正态分布相比,借款描述中的标点数呈现“尖峰厚尾”的分布形态,因此本文选择对标点符号数量在右侧 1%的水平上进行缩尾处理,并根据缩尾处理后的标点符号数量构建句均字数(*PC_ChiWord*)和句均词数(*PC_Voc*)指标 *W_PC_ChiWord* 和 *W_PC_Voc*,并重新对借款成功率和借款违约率进行回归。③由于“人人贷”发展初期,借款描述书写不规范的样本数量相当多,为避免此类样本对实证结果的影响,本文仅保留 2012 年及以后的样本后重新进行回归。回归结果显示,假说

① 关于稳健性检验的详细结果,请见《中国工业经济》网站(http://www.ceijournal.org) 附件。

1 和假说 2 依然成立。

(2)根据 2016 年 8 月中国银行监督管理委员会出台的《网络借贷信息中介机构业务活动管理暂行办法》第三章第十七条,同一自然人不得在同一 P2P 网络借贷平台上融资超过 20 万元人民币。并且,“人人贷”网络借贷平台上借贷的最高年利率,被设定为同期银行借款年利率的 4 倍(24%),且随着银行借款利率的调整而调整。因此,考虑 P2P 网络借贷面向大众小额借款需求的特征(Xu et al.,2015;Serrano-Cinca and Gutierrez-Nieto,2016)和借款人设置利率范围的合法性。在稳健性检验中,本文剔除借款金额大于 20 万元和借款利率大于 24%的样本进行回归,结果显示,在控制其他因素的条件下,假说 1 和假说 2 仍然得证。

(3)本文根据中华人民共和国教育部发布的《现代汉语语料库字(词)频表》中的字或词数来衡量借款描述中的常用字(*ComWord*)及常用词(*ComVoc*),以作为文本可读性中的“可理解性”的组成部分,但是字或词频表中含有一些没有信息含量的介词、助词、代词、连词,譬如的、地、你、我、他等,该类字或词未必能提高文本的可读性。因此,为了更精确地度量文本的可理解性,本文在稳健性检验中剔除前 100 位最高频率字或词中的介词、助词、代词、连词,并重新对借款成功率和借款违约率进行回归,回归结果显示,第四部分所得结论依然成立。

3. 改变解释变量

本文用汉字数除以句尾和停顿符的加总(*PC_ChiWord*)与词数除以句尾和停顿符的加总(*PC_Voc*)两个指标衡量文本的复杂性,且根据借款人使用《现代汉语语料库字(词)频表》中的字数(*ComWord*)和词数(*ComVoc*)衡量文本的可理解性。然而,如国内外诸多文献所发现,文本的长度对投资者的投资决策也会产生显著影响(Li,2008;Loughran and McDonald,2014;Tan et al,2015;Dorfleitner et al.,2016;彭红枫等,2016)。考虑到不同文本长度的借款描述中的字和词数具有显著差异性,本文所构造的绝对数量指标可能会导致回归结果的偏误,因此,本文在稳健性检验中构建文本相对值指标:借款描述长度中句均字数占比(*PPC_ChiWord*)与句均词数占比(*PPC_Voc*),并重新对借款成功率和借款违约率进行回归,第四部分所得结论依然成立。

4. 控制文本长度

前述实证分析没有控制文本长度。一方面,由于“人人贷”网络借贷平台对借款描述的长度存在一定限制,每个借款人书写的借款描述不能够超过 500 个字或字符,这一上限就相当于控制了文本长度。另一方面,为避免多重共线性问题,前期 P2P 网络借贷有关借款描述文本内容特征的研究中均未将文本的长度作为控制变量(Herzenstein et al.,2011;李焰等,2014)。然而文本的长度能够体现借款人为获取借款的努力程度,同时 Miller(2010)的研究发现,上市公司年报的文本长度和文本可读性指标之间存在替代关系。基于这种考虑,本文在稳健性检验部分中将文本长度指标(*logDlength*)作为控制变量,并重新对借款成功率和借款违约率进行回归,实证结果显示,在控制文本长度后,假说 1 和假说 2 依然得证。

5. Tobit 回归

本文前述部分采用 Logit 二值模型做借款描述的可读性与借款成功率及借款违约率的实证回归。但由于借款人提出借款申请后,“人人贷”会进行初步预审,本文观察到的样本已存在一定“选择偏误”问题,所以,本文在稳健性检验中采用了 Tobit 模型重新进行回归,结果仍与第四部分所得结论保持一致。

6. 内生性问题

只有借款成功的样本才会被观察到是否违约,借款描述可读性并不会显著影响借款违约率。但

是,通常高质量的借款人更有可能在借款描述时更加认真,进而“自我选择”书写可读性更强的借款描述来体现自身质量,因此,“样本选择”将导致“选择性偏差”。根据 Heckman(1979)提出的两阶段模型,第一阶段,首先,需要使用 *SUCCESS* 变量(是否借款成功虚拟变量),当借款人是否违约能够被观测到时取 1,否则为 0。其次,使用 Probit 模型,将 *DEFAULT* 视为被解释变量,其他信息视为解释变量进行回归,但需要注意的是,第二阶段的解释变量应该是第一阶段解释变量的子集,否则会由于存在多重共线性而导致第二阶段估计的结果存在偏差。运用两阶段模型进行估计,其结果表明,借款描述可读性的各项系数对借款违约率的影响均不显著,这意味着在控制由选择性偏误引起的内生性问题之后,本文第四部分所得结论依然没有改变,因此,可以认为,结论是稳健的。

借款描述存在学习效应,借款人可以通过观察其他借款人书写的借款描述对其带来的影响进行学习,对成功的借款描述进行模仿。这种不可观测的遗漏变量可能导致本文的结果存在内生性问题。本文借鉴公司金融领域构建工具变量的方法,Hong et al.(2005)的研究发现相同城市的机构投资者投资操作趋于一致,Leary and Roberts(2014)基于美国上市公司的研究发现,公司的融资决策在很大程度上受到同行业公司的资本结构和融资决策的影响,特别是规模小的公司更容易受到规模大并且业绩好的公司影响。叶德珠和陈霄(2017)采用借款人年龄、受教育程度和收入水平相同的其他借款人平均使用标点符号的数量作为工具变量,使用 IVPROBIT 模型对借款成功率和借款违约率进行回归,用两阶段最小二乘法(2SLS)对借款利率进行回归,以解决学习效应引起的遗漏变量问题。因此,根据同群效应(Peer Effect),本文认为借款人在书写借款描述时有可能受到相同年龄段、受教育程度、收入水平、地区的借款人借款描述特征的影响,而相同年龄段、受教育程度、收入水平、地区的其他借款人书写借款描述的特征却无法对该借款人的借款成功率和借款违约率产生影响。

对此,本文构建了一个工具变量:①本文按照借款人的年龄段,将借款人分为四组,分别是[30及以下]、[31,40]、[41,50]和[51及以上]。②分别构建与借款人年龄相仿、受教育程度、收入水平和所在区域相同的其他借款人平均句均字(词)数和常用字(词)数变量,*Me_PC_ChiWord*、*Me_PC_Voc*、*Me_ComWord* 和 *Me_ComVoc* 作为复杂性和可理解性指标的工具变量。第一阶段的回归结果表明,工具变量均能够产生显著为正的影响。③本文使用 IVPROBIT 模型对借款成功率和借款违约率进行回归。结果表明,句均字(词)数和常用字(词)数对借款成功率和借款违约率的影响与正文第四部分所得实证结果一致,可以认为本文的结论是稳健的。①

六、结论

文字是信息传递的重要形式。在 P2P 网络借贷中,借款人书写的借款描述是特有的非结构化信息,在借贷活动中借款人可以充分运用语言能力,将个人的私有信息及时传递给投资者。已有的研究证明,在信息不对称较为严重的 P2P 网络借贷市场,借款人自愿性书写的借款描述信息能对投资者行为产生显著影响,对投资者的决策产生举足轻重的作用。本文使用中国具有代表性的 P2P 网络借贷平台“人人贷”的数据,基于借款描述的文本可读性,从文本的复杂性和文本的可理解性两个维度,探讨借款人借款描述中文本可读性的作用。本文的主要结论是,投资者可以通过借款描述的文本可读性,判断借款人是否真的试图通过借款描述来降低借贷双方的信息不对称,进而推断出借款人潜在的还款表现。

本研究在理论上具有一定意义。文本信息正逐渐成为金融领域实证研究中的热点问题,然而这些研究的文本主体大多为英文。由于不同语言在表达方式之间有内在差异,使得那些能被直接用于

① 关于内生性检验的详细结果,请见《中国工业经济》网站(http://www.ciejournal.org)附件。

英文环境下衡量文本可读性的方法,并不能直接应用于中文语言环境的文本分析。本文尝试构建能从大量中文自然语言中提取文本可读性信息的文本分析指标。该方法通过使用句均字(词)数和常用字(词)数,将文本可读性具体分解为复杂性和可理解性两大维度。在理论分析的基础上考察借款描述的可读性对借款成功率与借款违约率的影响。本研究给中文语言环境下的文本研究提供了一种新思路,尤其对中文语言环境下构建文本可读性指标方面的研究是有益补充。

本研究在管理实践上也有一定意义。本研究为传统信贷机构的风险管理提供了一种新的思路。传统信贷机构非常重视对贷款的风险管理,尤其是在对信贷风险的衡量中形成了专门的贷款风险评估模型。然而,传统信贷机构贷款风险评估模型中,借款人能够被证实的硬信息占据很大权重,例如历史还款记录、现金流和抵押物等,这使得传统信贷机构在信息采集和验证方面往往需要花费很高的成本;传统信贷机构在评估信贷风险时,主要考察借款人的还款能力,而没有有效的方法对借款人的还款意愿进行测度。本文的研究则证实,文本信息能够体现借款人的还款意愿,借款人书写的文本信息内容对既有的风险评估模型是一种有效而低成本的补充。因此,传统信贷机构或许可以重新审视目前的风险评估模型,将借款人提供的文本信息作为是否发放贷款的依据之一。

本研究在政策制定上具有一定的意义。美国证券交易委员会对于信息披露的可读性和可理解性要求可以追溯到 1933 年的《证券法》。1998 年,美国证券交易委员会专门发布《简易英语手册》,目的在于明确如何产生简练的信息披露文件。同样在 1998 年,美国证券交易委员会规定,公司在招股说明书中需要使用简易英语。2008 年,美国证券交易委员会实行一项规定,要求基金招募说明书的摘要必须要以简易英语的方式书写。美国联邦政府在 2010 年 10 月通过的《简易写作法案》规定,要求联邦行政机构在发布的每一份文件中都使用简易英语。中国银行监督管理委员会在 2016 年 8 月 24 日出台的《网络借贷信息中介机构业务活动管理暂行办法》中仅仅强调借款人披露信息的真实性,但是对于披露信息的可读性和可理解性则没有具体的规定和指引。本研究表明,借款描述的可读性能够显著提高交易效率。因此,在未来监管政策的制定中,中国银行监督管理委员会可以出台针对 P2P 网络借贷市场的简单中文信息披露手册,强制要求借款人书写浅白的文字,提高借款描述的可读性,以便于投资者浏览。

本文的主要局限有两个方面:①“空谈博弈”的测度。由于文本信息在不同的条件下表现出一定的“语言膨胀”问题,例如借款人为吸引投资者,可能在借款描述中过于夸大自身的还款能力,P2P 网络借贷平台又无法对这些借款描述信息进行验证。因此,在“空谈博弈”(Cheap Talk)条件下所产生的信息是否具有信息含量往往被质疑(Demers and Vega,2013)。本文在现有的研究设计中尚未找到恰当的方法去排除借款人在“空谈博弈”条件下书写的借款描述所产生的影响,这是本研究今后仍需深入探讨的问题;②“选择偏误”问题。由于在借款申请满标后,“人人贷”会进行预审,因此本文观察到的样本已存在一定的“选择偏误”问题,尽管本文已用 Tobit 模型进行了检验,但在未来的研究中仍有必要尝试找到更好的解决方案。

〔参考文献〕

〔1〕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.

〔2〕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.

〔3〕叶德珠,陈霄. 标点与字数会影响网络借贷吗——来自人人贷的经验证据[J]. 财贸经济, 2017,(5):65-79.

〔4〕Abrahamson, E., and E. Amir. The Information Content of the President's Letter to Shareholders [J]. Journal of

Business Finance and Accounting, 1996,23(8):1157–1182.

〔5〕Abu Bakar, A. S., and R. Ameer. Readability of Corporate Social Responsibility Communication in Malaysia[J]. Corporate Social Responsibility and Environmental Management, 2011,18(1):50–60.

〔6〕Archak, N., A. Ghose., and P. G. Ipeirotis. Deriving the Pricing Power of Product Features by Mining Consumer Reviews[J]. Management Science, 2011,57(8):1485–1509.

〔7〕Bloomfield, R. Discussion of "Annual Report Readability, Current Earnings, and Earnings Persistence"[J]. Journal of Accounting and Economics, 2008,45(2–3):248–252.

〔8〕Burtch, G., A. Ghose., and S. Wattal. Cultural Differences and Geography as Determinants of Online Prosocial Lending[J]. Mis Quarterly, 2014,38(3):773–794.

〔9〕Campbell, T. S., and J. K. Dietrich. The Determinants of Default on Insured Conventional Residential Mortgage Loans[J]. Journal of Finance, 1983,38(5):1569–1581.

〔10〕Chen, X., B., Huang., and D. Z. Ye. The Role of Punctuation in P2P Lending: Evidence from China[J]. Economic Modelling, 2018,68(1):634–643.

〔11〕De Gregorio, J., and J. W. Lee. Education and Income Inequality: New Evidence from Cross–Country Data[J]. Review of Income and Wealth, 2002,48(3):395–416.

〔12〕Demers, E., and C. Vega. Understanding the Role of Managerial Textual Content in the Price Formation Process[R]. University of Virginia Working Paper, 2013.

〔13〕Dorfleitner, G., C. Priberny., S. Schuster., J. Stoiber., M. Weber., I. de Castro., and J. Kammler. Description–Text Related Soft Information in Peer–to–Peer Lending–Evidence from Two Leading European Platforms[J]. Journal of Banking and Finance, 2016,64(3):169–187.

〔14〕Duarte, J., S. Siegel., and L. Young. Trust and Credit: The Role of Appearance in Peer–to–Peer Lending[J]. Review of Financial Studies, 2012,25(8):2455–2483.

〔15〕Gao, Q., and M. Lin. Lemon or Cherry? The Value of Texts in Debt Crowd Funding[R]. University of Arizona Working Paper, 2015.

〔16〕Gunning, R. The Fog Index After Twenty Years[J]. Journal of Business Communication, 1969,6(2):3–13.

〔17〕Heckman, J. Sample Selection Bias as a Specification Error[J]. Econometrica, 1979,47(1):153–161.

〔18〕Herzenstein, M., S. Sonenshein, and U. M. Dholakia. 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(S):S138–S149.

〔19〕Hong, H., J. D. Kubik, and J. C. Stein. The Neighbor's Portfolio: Word–of–Mouth Effects in the Holdings and Trades of Money Managers[J]. Journal of Finance, 2005,60(6):2801–2824.

〔20〕Huang, X., S. H. Teoh, and Y. L. Zhang. Tone Management[J]. Accounting Review, 2014,(89):1083–1113.

〔21〕Kahneman, D., and A. Tversky. On the Psychology of Prediction[J]. Psychological Review, 1973,80(4):237–251.

〔22〕Klare, G. R. A Second Look at the Validity of Readability Formulas[J]. Journal of Reading Behavior, 1976, 8(2):129–152.

〔23〕Klare, G. R. The Measurement of Readability[J]. Journal of Computer Documentation, 2000,24(3):107–121.

〔24〕Larrimore, L., L. Jiang, J. Larrimore, D. Markowitz, and S. Gorski. Peer to Peer Lending: The Relationship Between Language Features, Trustworthiness, and Persuasion Success [J]. Journal of Applied Communication Research, 2011,39(1):19–37.

〔25〕Law, K. K. F., and L. F. Mills. Taxes and Financial Constraints: Evidence from Linguistic Cues[J]. Journal of Accounting Research, 2015,53(4):777–819.

〔26〕Leary, M. T., and M. R. Roberts. Do Peer Firms Affect Corporate Financial Policy [J]. Journal of Finance, 2014,69(1):139–178.

〔27〕Leong, E. K. F., M. T. Ewing, and L. F. Pitt. E-comprehension: Evaluating B2B Websites Using Readability Formulae[J]. Industrial Marketing Management, 2002,31(2):125-131.

〔28〕Li, F. Annual Report Readability, Current Earnings, and Earnings Persistence [J]. Journal of Accounting and Economics, 2008,45(2-3):221-247.

〔29〕Lin, M. F., N. R. Prabhala, and S. Viswanathan. Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending [J]. Management Science, 2013,59(1): 17-35.

〔30〕Liu, D., D. J. Brass, Y. Lu, and D. Y. Chen. Friendships in Online Peer-to-Peer Lending: Pipes, Prisms and Relational Herding[J]. Mis Quarterly, 2015,39(3):729-742.

〔31〕Loughran, T., and B. McDonald. When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks[J]. Journal of Finance, 2011,66(1):35-65.

〔32〕Loughran, T., and B. McDonald. Measuring Readability in Financial Disclosures[J]. Journal of Finance, 2014, 69(4):1643-1671.

〔33〕Loughran, T., and B. McDonald. Textual Analysis in Accounting and Finance: A Survey [J]. Journal of Accounting Research, 2016,54(4):1187-1230.

〔34〕Ludwig, S., K. de Ruyter, M. Friedman, E. C. Bruggen, M. Wetzels, and G. Pfann. More Than Words:The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates [J]. Journal of Marketing, 2013,77(1):87-103.

〔34〕McKeown, M. G., I. L. Beck, G. M. Sinatra, and J. A. Loxterman. The Contribution of Prior Knowledge and Coherent Text to Comprehension[J]. Reading Research Quarterly, 1992,27(1):79-93.

〔36〕Michels, J. Do Unverifiable Disclosures Matter? Evidence from Peer-to-Peer Lending [J]. Accounting Review, 2012,87(4):1385-1413.

〔37〕Miller, B. P. The Effects of Reporting Complexity on Small and Large Investor Trading[J]. Accounting Review, 2010,85(6):2107-2143.

〔38〕Miller, G. A. The Magical Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Information[J]. Psychological Review, 1956,63(2):81-97.

〔39〕Peng, L., and W. Xiong. Investor Attention, Overconfidence and Category Learning [J]. Journal of Financial Economics, 2006,80(3):563-602.

〔40〕Rameezdeen, R., and C. Rajapakse. Contract Interpretation: the Impact of Readability [J]. Construction Management and Economics, 2007,25(7):729-737.

〔41〕Reber, R., and N. Schwarz. Effects of Perceptual Fluency on Judgments of Truth [J]. Consciousness And Cognition, 1999,8(3):338-342.

〔42〕Rennekamp, K. Processing Fluency and Investors' Reactions to Disclosure Readability[J]. Journal of Accounting Research, 2012,50(5):1319-1354.

〔43〕Roux, J. P. An Interlocutory Analysis As a Methodological Approach in Studying Cognitive-Linguistic Mediations: Interest, Difficulties, and Limits [J]. European Journal of Developmental Psychology, 2008,5(5): 609-622.

〔44〕Sawyer, A. G., J. Laran, and J. Xu. The Readability of Marketing Journals: Are Award-Winning Articles Better Written[J]. Journal of Marketing, 2008,72(1):108-117.

〔45〕Serrano-Cinca, C., and B. Gutierrez-Nieto. The Use of Profit Scoring As an Alternative to Credit Scoring Systems in Peer-to-Peer (P2P) Lending[J]. Decision Support Systems, 2016,89(C):113-122.

〔46〕Shiller, R. J. Conversation, Information, and Herd Behavior[J]. American Economic Review, 1995,85(2):181-85.

〔47〕Shriberg, E., A. Stolcke, D. Hakkani-Tur, and G. Tur. Prosody-Based Automatic Segmentation of Speech into Sentences and Topics[J]. Speech Communication, 2000,32(1-2):127-154.

〔48〕Tan, H. T., E. Y. Wang, and B. Zhou. When the Use of Positive Language Backfires: The Joint Effect of Tone, Readability, and Investor Sophistication on Earnings Judgments [J]. Journal of Accounting Research, 2014, 52(1):273-302.

〔49〕Tan, H. T., E. Y. Wang, and B. Zhou. How Does Readability Influence Investors' Judgments? Consistency of Benchmark Performance Matters[J]. Accounting Review, 2015,90(1):371-393.

〔50〕Tausczik, Y. R., and J. W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods[J]. Journal of Language and Social Psychology, 2010,29(1):24-54.

〔51〕Vauras, M., J. Hyönä, and P. Niemi. Comprehending Coherent and Incoherent Texts: Evidence from Eye Movement Patterns and Recall Performance[J]. Journal of Research in Reading, 1992,15(1):39-54.

〔52〕Xu, Y., C. Luo, D. Y. Chen, and H. C. Zheng. What Influences the Market Outcome of Online P2P Lending Marketplace? A Cross-Country Analysis[J]. Journal of Global Information Management, 2015,23(3):23-40.

Can Readability of Loan Description Promote Lending Success Rate of Online

CHEN Xiao, YE De-zhu, DENG Jie

(Finance Department, College of Economics, Jinan University, Guangzhou 510632, China)

Abstract: In Peer-to-Peer (P2P) lending market, loan description of the borrowers can be regarded as a way of attracting investors and further promoting lending success rate, and readability of loan description tends to determine the investors' understandability on loan description related soft information, As evidenced by P2P lending platform's experience, we attempt to answer the question that does readability in loan description mitigate information asymmetry between borrowers and lenders, and promote lending success rate?. The paper examines the effect of readability on P2P lending market by using the data of 128532 lists during January 2011 to May 2014 from a Chinese P2P lending platform, "RenRenDai", and constructing the index of loan description readability under Chinese language environment based on the complexity and understandability. Our empirical results show that: ①The readability of the loan description is more likely to reduce the information noise and improve the speed of information identification, thus decrease the cost of information processing of investors. ②A more readable description of the loan can convey a positive signal to investors, and improve the borrower's loan success rate. ③An increasing of average characters per sentence in loan description will cause lending success rate's falling of 0.4%, whereas an increasing of average words per sentence will result in lending success rate's falling of 0.7%. ④The characters commonly used words for each additional one, the loan would increase the success rate of 0.14%, the loan default rate would decline by 0.12%, whereas more commonly used words for each additional one, and the loan would increase the success rate of 0.28%. The aforementioned conclusions remain valid after conducting an array of robustness tests. How to efficiently use the unverified soft information is becoming the problem of both investors and borrowers as the cessation of guarantee from P2P lending platforms, to this we prove that the readability of loan description can be informative and transmit signals.

Key Words: P2P lending online; asymmetric information; loan description; text readability

JEL Classification: G11 G21 G23

〔责任编辑:湘学〕