文章编号：1003-0077(2011)00-0192-13

基于稀疏主成分分析的非正式语词的心理-人格特征研究

钟 毓，费定舟

（武汉大学 哲学学院心理系，湖北 武汉 430072）

摘 要：针对社会媒体中非正式文本的数据分析经常出现的稀疏数据矩阵，在应用文本分析工具的基础上使用稀疏主成分分析这一特征，降维分析方法分析现实情况下聊天文本中非正式语词表现的认知语用特征、描述非正式语词与人格的关系。使用短文本主题模型、心理距离问卷、大五人格问卷测量人格和背景变量，使用计算机文本分析工具对被试提供的即时聊天文本内的语词计频，使用简体中文版语词查询与字词计数字典和认知语用学对稀疏主成分分析后非正式语词维度进行特征表征。在非正式语词降维上，稀疏主成分分析比主成分分析在因子载荷数上更稳定，在累积方差解释率上也相对更优(24.54% >23.40%)；降维所得的 6 因子中“主观评价”与宜人性正相关($r_{0.05}$=.16，p=.03<0.05)，“随意社交”与宜人性负相关($r_{0.05}$=−.16，p=.03<0.05)，“认知愉悦”与性别显著正相关($r_{0.05}$=.43，p=.00<0.001)。使用稀疏主成分分析对非正式语词的降维效果较好，并且比较简体中文版语词查询与字词计数字典的非正式语词维度和降维后所得非正式语词维度，两者在和人格的相关上是相符的，且后者能探索出更多信息。

关键词：文本分析；稀疏主成分分析；非正式语词

中图分类号：TP391　　**文献标识码**：A

Judging Personality by Informal Words: a Sparse PCA Approach

ZHONG Yu, FEI Dingzhou

(Psychology Department of Philosophy School, Wuhan University, Wuhan, Hubei 430072, China)

Abstract: In this paper, a new method is presented to identify personality with dimension reduction by sparse principal component analysis (SPCA). Based on categories of linguistic inquiry and word count dictionary (LIWC), informal words usage and psychological trait in instant chat is analyzed, and the relation between informal words and personality is described. Biterm Text Model (BTM), psychological distance questionnaire and Big Five personality questionnaire are used to measure personality and related variables. The informal words dimensions are explained based on simplified Chinese version of linguistic inquiry and word count dictionary and cognitive linguistic usage. It is shown that the numbers of load factors gotten by the SPCA more stable than the numbers of traditional principal component analysis(PCA), and the cumulative explained variances are better (24.54%>23.40%). With respect to 6 dimensions, “subjective evaluation” was positively related to agreeableness ($r_{0.05}$=.16, p=.03<0.05), “casual socializing” was negatively related to agreeableness ($r_{0.05}$=−.16, p=.03<0.05), while “cognitive pleasure” and gender were significantly positively related ($r_{0.05}$=.43, p=.00<0.001). These results suggest that SPAC for dimensional reduction performs better PCA in related studied issues.

Key words: text analysis; sparse principal component analysis; informal words

1 引言

人类的语言使用，不管是正式文本还是非正式文本都与使用者的人格和心理特征有密切关系。Allport 和 Odbert 早在 1936 年就开用词汇使用与人格和其他心理特征的关系研究的先河[1]，以 Five Big Inventory(五大人格特征词汇量表)为代表的词汇—人格测量工具被认为是目前有效的研究手段之一[2]。心理语言学的计算模型的一个重要任务是提

收稿日期：2016-09-16　**定稿日期**：2016-11-08

供比较有效的统计工具，如聚类分析(Cluster Analysis)，主成分分析(Principal Component Analysis，PCA)或其他计算模型，如概念空间理论(Concept Space Theory)。这些方法的使用必须同时与词汇分析的意义抽取方法相配套，即从文本中使用意义抽取方法抽取相关词汇语义，再使用统计工具和其他计算模型来做进一步分析。目前，沿这个方向出现的问题里有两个方面值得注意：一是使用正式文本还是非正式文本来作为处理材料；二是由于大多数数据是稀疏的，从稀疏数据中如何提取有效的人格描写维度是具有挑战性的任务，这就需要找到合适的统计方法。

本文选取非正式文本(如微信或 QQ 文本等)，统计模型引进和采用稀疏主成分分析(Sparse Principal Components Analysis，SPCA)．本文的贡献体现在以下两方面。

(1) 语词在不同语境下就会有不同的意思；人们对语词的使用又可以表现出所交谈的对象的身份是陌生人还是熟人，是上司还是下属；语词使用也可以表现出一个人的年龄、性别，甚至人格特征[3-4]。非正式语词在本研究中是以 LIWC 的非正式语词类别作为研究基础。研究非正式情境下的语词，是为了研究自然状态下的个体特征；而研究非正式语词，是为了研究非实词类语词能否表现个体特征。

现实中基于网络聊天的文本数据受到多个变量的影响，人们聊天的主题、人们聊天的对象同样会影响到语词的使用。为了使非正式语词词典更准确，更能体现实际网络聊天情况下的人格特征，必须将一些比较重要的背景变量纳入研究。

本研究将直接使用短文本主题模型中的 P(topic|doc)来表现主题分布情况。主题模型的逻辑是每一篇文档都有一系列主题，这些主题符合一定的概率分布，依据这个概率分布随机选择一个主题，然后再从这个主题里按照另一个概率分布选择一个词。而之所以使用短文本主题模型是因为研究材料是聊天文本，聊天文本具有短、多、主题分散的特点。

(2) 比较不同特征降维方法(PCA，SPCA)在非正式语词中的效果，择优对语词进行降维以及命名；再使用计算机文本分析工具对被试的聊天文本进行分析，分析比较特征降维与经验分类下的非正式语词与性别、年级、人格的关系。

因子旋转下的主成分分析或经验分类常用于研究语词特征，但这些方法应用在非正式语词中是否高效可信，并且能够表现出语词的认知语用特征和心理特征，是否有更高效的特征提取方法呢？结果表明，SPCA 的效果普遍好于 PCA，这意味 SPCA 是一个处理文本意义抽取的有效统计模型。

本研究引入了一个新的语词心理表征方案，意图解决现实情况下聊天文本所表现的认知语用特征和心理—人格的关系。先对即时聊天情况下所会出现的非正式语词进行扩充；在背景变量的测量上使用一种短文本主题模型检测即时聊天文本的主题，使用心理距离问卷来描述双方亲密度。在比较不同降维方法的优劣后，比较较优的降维方法与经验分类方法得到的非正式语词维度与人格的相关。

2　研究基础

2.1　语词使用与背景变量的关联研究回顾

研究者认为语言、主题、听众这三者是具有相互作用的，这也是之后大多数研究所确定的事实[5-8]。然而，Pennebaker 认为语言使用具有跨时间、跨主题的信度[9-10]。尽管如此，很多研究在研究文本时还是对文本内容上有所限定[11-12]。

交流文本还涉及到交流双方的关系状态，即语词使用受到交流双方关系的影响。Holtgraves 认为语言的变化是人格特征和关系状态的函数[13]。Holtgraves 研究中对关系状态是以现实关系、喜爱程度和亲密程度的七点量表来测量的，并且以 1-5、6、7 三个级别做 Welch's F 检验；对俚语、缩略语和表情的计数是计算包含这些语词的短信条目百分比。

2.2　非正式语词的特征降维方法

主成分分析(Principal Components Analysis，PCA)被广泛应用于数据处理和维数降低，这一方法常用于社会科学的特征降维。然而，主成分分析中各主成分是所有原始变量的线性组合，不仅难以解释，而且对数据有一定要求。使用其方差旋转后因子载荷矩阵使成分矩阵的结果更明确，但是其主

成分方差不是依次递减的。有一个解决方案是借鉴套索(LASSO)或弹性网(Elastic Net)的思路,如果对载荷系数绝对值的和设定一个阀值,如果载荷的绝对值小于这个值,就认为载荷为零,这样简化了主成分的计算,而且解释性更明确。这样的解决方法产生了稀疏负荷的主成分,称为稀疏主成分分析(Sparse Component Analysis,SPCA)[12]。在实际应用上,稀疏主成分分析可以应用在人脸识别、图像处理等。从技术来讲,它把线性回归模型中的变量前面的系数(构成主成分时每个变量前面的载荷)通过约束(可以由问题背景设定阀值,比如二次惩罚和一次惩罚)变得稀疏,也即是把大多数系数都变成零,我们就可以把主要的要素凸现出来,这样主成分就会变得较为容易解释。在文本处理涉及说话人的心理特征和人格特征时,可以实现显著性变量的选择和对应参数的估计,且都有实际的心理学涵义,结果稳健且计算相对简单。

2.3 性别、年龄、人格对非正式语词使用的影响

语言查询与字词计数(Linguistic Inquiry and Word Count,LIWC)是一种可以对文本内容的词语类别(尤其是心理相关类别)进行量化分析的软件。已有的LIWC使用经验表明,它在各式各样的实验中有表现注意力,情绪,社会关系,思维方式和个体差异的能力[14]。LIWC所代表的是一种为特定词汇计算频数的分析方法(算法程序)和相应特定词汇的词典。

LIWC的某些类别与人格呈相关。在应用LIWC的人格研究中,LIWC在人们表达情感的文章中显示出语词能够反映个性风格,在学生所写的个人计划中显示出LIWC类别中语词使用频率的个体差异与大五人格特质有关。这些对个人写作的文本分析,可以说明语言使用和人格之间具有一定联系[12,15-17]。研究者对交流双方的短信进行文本分析,部分LIWC类别与人格相关,例如,外向性与人称代词,神经质与负面情绪词和宜人性与正面情绪词[13]。对邮件文本进行的文本分析显示出艾森克人格量表与LIWC某些类别相关[18]。

LIWC2015版的词典相较于LIWC2007版对非正式语词有所增加,这一部分主要是根据经验对非正式语词进行分类。它将非正式语词分为脏话、应和词、填充词、停顿词、网络语言五个类别[19](类别词汇见附录1)。LIWC2015版中的非正式语词主要是从twitter这样的非正式文本中获得,这些词汇能够代表非正式文本除了主题以外的特征。

非正式语词在不同个体上具有不同的使用特点[20]。也有研究者认为外向者会更多地使用非正式语词[21]。话语标记词(在本研究中为应和词、填充词、停顿词)相对于其他词汇来说用得不少,但话语标记词通常只在日常会话中出现,而且大多情况下没有像实词所表示的意义,这也决定了话语标记词的研究相对较繁琐。但是特定的语词确实与人格特征相关[13,16,22]。比如,自觉性高的人可能会使用更多一些看似无意义的话语标记[1]。在网络语言方面,研究者认为使用网络语言多的人在人格特征上自觉性更低,开放性较低,但情绪更稳定,而网络语言的内容对人格没有显著的影响[23]。

SC-LIWC起步时间较晚,所以有很多值得人们去验证、增删补词典的工作。黄金兰的研究团队在LIWC2007版基础上建立了繁体中文版的C-LIWC词典[24]。2013年他们进行了小幅修正,增加了抓取的精准度,并正式提供繁体中文字典修正版1.1。此外,2013年又以繁体中文版LIWC为本进行繁简转换以及两岸用语校正等步骤的修订,建立了简体中文版SC-LIWC。但是现如今的SC-LIWC在非正式语言这一类别上依然不成熟,主要是因为中文非正式语词不仅具有繁杂、与时俱进的特点,还有中国文化与认知的特点[25]。比如,“嘛”这个词不仅有缓和标记功能、命题表态功能、形象构建功能,而且人们在说话的时候会有意使用“嘛”之类的词汇来减缓、弱化说话语气的刚性,更好地传递会话意图,进而构建积极的个人形象身份[26]。本研究将会对SC-LIWC中的非正式语词做出相应扩充,以期适应本研究,得到更准确的非正式语词结果。

3 研究方法

3.1 被试

在武汉大学随机选取的100名被试,77%的被试为有效被试,其中23%的被试因为中途放弃,或者因为即时消息文本没有达到20条而被视为无效被试。其中77名有效被试中有40名男生,37名女

生，其中大一占 3.9%，大二占 22.1%，大三占 2.6%，大四占 71.4%。由于研究的特殊性，主试会对每个被试强调个人信息绝对保密，并且不会将聊天文本用于科研以外，对所采用的研究方法进行一定解释。在被试确定参与研究后发放问卷，并监督被试完成问卷以及聊天文本发送，一定程度上确保被试的有效性。

3.2　数据收集

被试被要求挑出最近和朋友的聊天记录，将聊天记录中你所发的 20～30 条微信或 QQ 文本消息发至主试的邮箱。所得到的即时交流文本需要是被试个人发出的文本信息。在进入文本分析工具之前，研究者先需要删除被试误发的他人聊天文本，删除个人标注，之后使用 NiuParser 中文语义分词软件对文本进行分词[27]。分词后的文本使用 LIWC 这种分析特定词汇词频的方法，计算整理出的特定词汇在被试文本中的词频（本研究中使用扩充后的非正式语词词典）。

根据对所有主题模型适用情境的大致了解，并且尝试使用不同主题模型对文本进行分析。作者名字选择短文本主题模型[28]（Biterm Topic Model，BTM）和 Matlab 的 topic Toolbox 1.3 描述文档主题分布，研究不同主题下的语词使用情况[29]。

让被试在主观心理距离问卷中对聊天对象与自己的心理距离打分。在心理距离的测量上，作者名字使用人际关系亲密度量表[30]，这一问卷能够用于测量个体与他人之间的心理距离。

同时，让被试填写性别、年级，完成中文版大五人格量表[31-32]。

3.3　非正式语词的扩充

由于 Pennebaker 等人增加了对 Twitter 等非正式情境下的研究，所以 LIWC2015 版对之前版本的在非正式语词有较大的补充和改动。为了增加 SC-LIWC 的非正式语词部分，使其适应研究目的，研究者对新增的非正式语词部分进行了翻译，发现非正式语词具有当地的特性，不能简单翻译来增加这一部分内容。

所以首先在 SC-LIWC 已有的部分非正式语词基础上增加非正式语词，得到本研究能使用的字典。其次由于 LIWC 中的同意词、停顿词、填充词（词汇见附录 1）是以话语标记词作为分类基础[3]。所以本研究使用常用中文话语标记词[33-34]与 SC-LIWC 互为补充。之后，由于本研究是以 SC-LIWC 为基础，所以未被词典查询到的词汇将被记录下来，再从这些词汇中挑选出脏话词和网络语。这样挑选脏话词和网络语的原因是，脏话词在即时交流的本次样本中较少，而网络语在本研究中因为样本较小，有可能一个网络语只在一个被试的文本中出现过一次。最后，由于亚洲文化与西方文化上的差异，亚洲文化会更多的使用表情符号以及颜文字[35]，本研究不具体研究表情，而将表情的频数百分比作为一个词条项。

3.4　非正式语词特征降维

通过对非正式语词的整理，研究者需要将整理后的非正式语词作为词典，使用计算机文本分析工具查询每个被试文本中各个非正式语词占每个文本总词汇数的比例。

在本研究中，研究者比较稀疏主成分分析与主成分分析在语词降维上的应用，检测其降维效果。首先要将 SPCA 所得因子载荷矩阵和 PCA 旋转后因子载荷矩阵作比较，从因子载荷数、调整后方差观察两种不同的降维结果，得出较优异的因子载荷结果。其次，使用简体中文版的语言查询与字词计数字典（SC-LIWC）和认知语用学对降维后非正式语词维度进行描述、命名。

4　结果

4.1　背景变量对非正式语词使用的影响

使用 R 中 tm 包对聊天文本进行删除停用词、删除数字、删除单个字之后，对处理后的文本进行描述统计。图 1 是聊天文本词云图，聊天语料的稀疏性在 98%，非稀疏词汇（文本中有所重复的词汇）所占比是 2 483/133 471。

由于收集到的聊天语料的稀疏性，BTM（短文本主题模型）在短文本上相较于 LDA、PLSA 有较好的分类精度，所以可以使用 BTM 来对聊天文本进行主题的区分。通过对文本的分析，图 2 可以看出离群主题的文档编号。由于聊天文本数量和被试特点，聊天文本没有像预想的那样在不同主题上有一定数量的分布，也因此在聊天主题上具有稀疏性，所以本研究将不标注聊天文本的主题，而是直接比较八个特殊的聊天文本与其他 69 个聊天文本的语词使用。

图1 聊天语料词云图

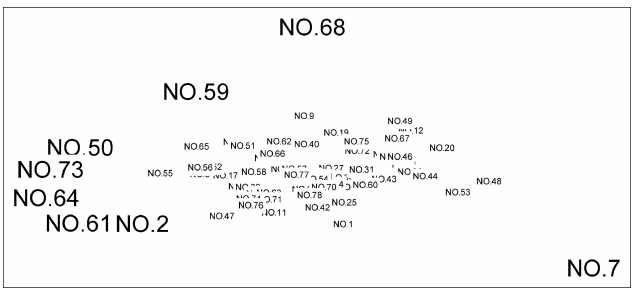

图2 文档可视化

使用两个独立样本 Mann-Whitney 检验对主题集中文本和主题离群文本进行非参数分析。主题集中文本在特定人称代名词($U_{0.05}$(40.86,23.00)=148.00,$p<.05$)、第一人称代名词($U_{0.05}$(41.62,16.44)=95.50,$p<.01$)上显著高于主题离群文本,而在金钱词上显著低于主题离群文本($U_{0.05}$(36.43,61.13)=99.00,$p<.001$)。

在语词使用与心理距离的关系上,介词与心理距离呈负相关($r_{0.05}=-.28$,$p<.05$)、应和词与心理距离呈正相关($r_{0.05}=.28$,$p<.05$),即介词在聊天文本中使用得越多,被试报告的心理距离得分越低,而应和词在聊天文本中使用得越多,被试报告的心理距离得分越高。非正式语词词频与心理距离不呈线性相关,并且差异不显著。

4.2 非正式语词的心理特征提取

为了比较使用稀疏主成分分析和使用主成分分析对非正式语词降维的效果,本研究使用 SPCA 和 PCA 对非正式语词降维。与 PCA 的旋转后因子载荷矩阵相比,SPCA 所得因子载荷矩阵在 6 因子载荷总数量上明显较优,在累计调整后方差上也相对较高,见表 1。当因子数=6,Δ=inf,stop=0.4,spca 函数返回的因子载荷情况如表 2 所示,累积的调整后方差为 24.54%,这一结果在自然语言特征降维已经足够好[12]。

因此,在对非正式语词的降维研究上可以使用稀疏主成分分析,在之后的研究中我们将非正式语词经过稀疏主成分分析所得的结果称之为 SPCA-非正式语词维度(SPCA-Informal,SPCA-I)。

表1 PCA 与 SPCA 在对非正式语词降维上的对比

		PC1	PC2	PC3	PC4	PC5	PC6
PCA	因子载荷数	9	6	6	4	4	3
	调整后方差	4.88	4.39	3.87	3.55	3.45	3.26
	累积调整后方差	4.88	9.27	13.15	16.70	20.14	23.40
SPCA	因子载荷数	8	6	4	6	7	5
	调整后方差	5.27	4.87	3.79	4.06	3.19	3.36
	累积调整后方差	5.27	10.15	13.93	17.99	21.19	24.54

注:对 PCA 的因子载荷结果进行方差旋转处理,SPCA 未处理。

根据语词在认知语用上的不同,以及 SPCA-非正式语词维度与 SC-LIWC 类别的相关,对降维后的六个主成分进行认知语用和心理意义上的双重表征。

PC1 中使用频率最高的词汇是“好”,“好”这个词汇根据语言学的观点,它能表现很多功能用法,其核心意义为“对事物的积极评价”。“好”在社交中具有建立,维持,延长,中断谈话内容的功能,在认知上表达了说话人的语气,具有主观评价意义[36-37]。PC1 与悲伤词正相关($r_{0.05}=.32$,$p=.00<0.01$)。

PC1 命名为“主观评价”。PC2 除了“嘻嘻”都是“好”加上一个语词助词，语气助词能够缓和说话人的语气，而通过使用人数比例可以看出 PC2 中的语气助词都不是常用的语气助词，使用稀有的语气助词在认知上能够引起注意或者使语气舒缓[26]，PC2 中的语词能够通过社交情境来表现肯定、状态的持续等。PC2 与后置词正相关（$r_{0.05}=.24$，$p=.02<0.05$）。因此，将 PC2 命名为“舒缓语气”。PC3 中最常使用的语词是“嗯嗯”，“嗯嗯”是“嗯”的叠连，在社交中是对对方话语的应和，在认知上表现了不置可否的心理[38]。PC3 与连词正相关（$r_{0.05}=.28$，$p=.001$），与特定人称代词负相关（$r_{0.05}=-.24$，$p=.00<0.01$），与填充词正相关（$r_{0.05}=.62$，$p=.00<0.001$）。因此 PC3 中的语词偏向社交，并且没有深层的认知，可以将 PC3 命名为“社交填充”。PC4 中最常使用的语词是“啊”和聊天文本中经常出现的表情符号。“啊”在社交、认知上与“嗯”具有相似的功能，但是相比于 PC4 中的“嗯嗯”，“啊”由于没有叠连，在社交偏向上更弱。由于表情符号是聊天文本特有的表达方式，对表情符号的研究认为使用者更不自觉，开放性更低[23]。PC4 与第一人称复数代名词正相关（$r_{0.05}=.34$，$p=.001$），与时间词正相关（$r_{0.05}=.20$，$p=.03<0.05$）。因此，作者名字将 PC4 命名为“随意社交”。PC5 最常使用“哈哈”、“哈哈哈”，这两个词在认知上存在失谐、失谐探测、失谐消解的过程，在情感上表达出愉悦[20]。PC5 与功能词正相关（$r_{0.05}=.29$，$p=.00<0.001$），与焦虑词正相关（$r_{0.05}=.20$，$p=.03<0.05$），认知过程词正相关（$r_{0.05}=.31$，$p=.00<0.001$），与应和词正相关（$r_{0.05}=.23$，$p=.01$）。本研究对 PC5 命名为“认知愉悦”。PC6 最常使用的是“就”这个词，“就”作为副词具有小量的主观量，“的话”与其相似，即“就”具有少量的主观评价意义[33]。PC6 与停顿词正相关（$r_{0.05}=.32$，$p=.000<0.001$）因此，对 PC6 命名为“少量主观”。

表 2　非正式语词的稀疏主成分分析

语词	平均数	标准差	>0(%)	SPCA-非正式语词维度					
				主观评价	舒缓语气	社交填充	随意社交	认知愉悦	少量主观
恩恩	0.03	0.30	1	**−0.50**	0.00	0.00	0.00	0.00	0.00
明白	0.01	0.10	1	**−0.50**	0.00	0.00	0.00	0.00	0.00
棒	0.04	0.24	4	**−0.45**	0.00	−0.15	0.00	0.00	0.00
哇	0.03	0.15	4	**−0.37**	0.00	0.00	0.00	0.00	0.00
哈	0.09	0.29	12	**−0.35**	0.00	0.00	0.00	0.00	0.00
行	0.22	0.50	26	**−0.13**	0.00	0.00	0.00	0.00	0.00
好	1.00	1.07	66	**−0.12**	0.00	−0.07	0.00	0.00	0.00
OK	0.08	0.30	12	**−0.10**	0.00	0.00	0.00	0.00	0.00
好呀	0.05	0.41	1	0.00	**−0.51**	0.00	0.00	0.00	0.00
好嗦	0.02	0.20	1	0.00	**−0.51**	0.00	0.00	0.00	0.00
嘻嘻	0.02	0.12	3	0.00	**−0.44**	0.00	0.00	0.00	0.00
好哒	0.05	0.28	4	0.00	**−0.44**	0.00	0.00	0.00	0.00
好嘞	0.03	0.22	3	0.00	**−0.24**	0.00	0.00	0.00	0.00
好的	0.09	0.31	9	0.00	**−0.20**	**−0.18**	0.00	0.00	0.00
好啦	0.01	0.13	1	0.00	0.00	**−0.58**	0.00	0.00	0.00
嗷	0.05	0.38	3	0.00	0.00	**−0.57**	0.00	0.00	0.00
等等	0.04	0.20	5	0.00	0.00	**−0.49**	0.00	0.00	−0.25
嗯嗯	0.23	0.69	19	0.00	0.00	**−0.18**	0.00	0.00	0.00
呵呵	0.01	0.05	1	0.00	0.00	0.00	**0.57**	0.00	0.00

续表

语词	平均数	标准差	>0(%)	SPCA-非正式语词维度					
				主观评价	舒缓语气	社交填充	随意社交	认知愉悦	少量主观
漂亮	0.01	0.05	1	0.00	0.00	0.00	**0.57**	0.00	0.00
同意	0.01	0.05	1	0.00	0.00	0.00	**0.57**	0.00	0.00
表情	2.85	5.49	43	0.00	0.00	0.00	**0.11**	0.00	0.00
啊	1.03	1.13	62	0.00	0.00	0.00	**0.05**	0.00	0.00
唉	0.03	0.20	4	0.00	0.00	0.00	**0.05**	0.00	0.00
欧克	0.01	0.09	1	0.00	0.00	0.00	0.00	**−0.55**	0.00
嘿嘿	0.05	0.26	4	0.00	0.00	0.00	0.00	**−0.55**	0.00
干	0.10	0.33	9	0.00	0.00	0.00	0.00	**−0.46**	−0.11
噢	0.03	0.17	4	0.00	0.00	0.00	0.00	**−0.36**	0.00
哈哈	0.44	1.03	29	0.00	0.00	0.00	0.00	**−0.12**	0.00
嗯啊	0.01	0.09	1	0.00	0.00	0.00	0.00	**−0.11**	0.00
哈哈哈	0.49	1.14	26	0.00	0.00	0.00	0.00	**−0.03**	0.00
okay	0.02	0.14	1	0.00	0.00	−0.04	0.00	0.00	**−0.59**
的话	0.04	0.19	5	0.00	0.00	0.00	0.00	0.00	**−0.56**
其他消息	0.04	0.20	4	0.00	0.00	0.00	0.00	0.00	**−0.49**
就	1.17	1.08	73	−0.05	0.00	0.00	0.00	0.00	**−0.17**
傻瓜	0.01	0.11	1	0.00	0.00	0.00	0.00	0.00	**−0.03**

注：以上词汇是非正式语词的一部分，限制 SPCA 降维后得到六个成分，上表是六个成分的因子载荷情况。其中平均数、标准差以及 SPCA 的结果都是基于特定词汇的词频所得出的，>0(%)是使用特定词汇的人数百分比。

结合上面关于表 1 内容的命名讨论，从表 2 中我们看到它对主成分变量的解释与词的使用经验基本上是相容的，这些解释分别把 PC1-PC6 命名为“主观评价”，“舒缓语气”，“社交填充”，“随意社交”，“认知愉悦”，“少量主观“，每一个命名对应一个主成分。在表 2 中，很明显看到非零载荷系数（表中用黑色数字标出）相当于一个聚类，每个非正式词都属于其中一个，值得注意的是，表 2 中这些词在一个心理维度上载荷系数都为正，或都为负，这意味这些词的使用是负向关联的，上面关于命名的讨论较为具体描述这种关联。有些词，在表 2 中，比如“好的”，同时属于两个成分，表明这个词的使用有含糊性。表 2 中词所对应的载荷系数大部分只在一个维度上不为零，其他五个维度为零，这是非正式词语使用的稀疏特征，对于正式词语的使用，也常常观察到类似的数据特征。这是使用稀疏主成分分析所带来的易解释优点，如果使用 LIWC，由于这个系统是用通常的主成分分析，很可能得不到这种易解释的主成分载荷系数的稀疏性。

4.3 性别、年级、人格对非正式语词使用的影响

SC-LIWC 总体 Cronbach's Alpha 为 0.46，对于文本分析而言，这个数字足够高。相关所得到人格维度与 SC-LIWC 大部分类别不相关，只有表 3 中的 SC-LIWC 类别、SPCA-I 维度与人格维度呈现相关。

表 3 将人格维度与 SC-LIWC 类别、SPCA-I 维度显著相关（$p<0.05$）的部分列出。外向性、开放性与动词负相关，宜人性与介词显著负相关，神经质与悲伤词显著正相关，自觉性与认知过程词负相关。即那些人格中外向性或开放性得分较高的人更少使用动词，那些在宜人性上得分高的人会更少使用“从、依照、把”等介词，神经质的分高的人更多使用“心痛、沮丧、无力”等悲伤词，在自觉性上得分高的人使用较少的认知过程词。

值得注意的是，后置词、量词是本土化后才有的类别。所以，尽管表 3 将已有相关列出，但是部分相关是难以解释，并且存疑的。在 SC-LIWC 的非正式语词类别中，SC-LIWC 的脏话类别词在文本中检

测到的比例只有 12%，即 77 人中只有九个人使用了这一类别词，而应和词为 100%。所以，神经质和脏话的正相关相对较不可信。宜人性和应和词正相关，宜人性得分越高，会使用越多的应和词。观察 SC-LIWC 字典，发现脏话类别词没有包括很多聊天文本中出现的脏话。SPCA-非正式语词维度与人格维度之间大部分维度之间是不相关的，只有少数相关具有显著性。宜人性与主观评价正相关（$r_{0.05}$ = .16，p = .03 < 0.05），与随意社交负相关（$r_{0.05}$ = −.16，p = .03 < 0.05），即那些更多使用主观评价语词的被试在宜人性上为自己打分很高，而那些更多使用随意社交语词的被试在宜人性上得分很低。

表 3　人格维度与 SC-LIWC、SPCA-I 相关表

		人格维度				
		外向性	宜人性	自觉性	神经质	开放性
verb	动词	−.21**	.01	.06	.14	−.18*
preps	介词	−.05	−.22**	−.04	.14	−.04
quant	概数词	.18*	.08	.04	.12	.12
number	数字	−.05	−.08	.18*	−.04	.05
prepend	后置词	.00	.04	.24**	.01	.04
specart	特制定词	.16	.18*	.13	−.09	−.00
quantunit	量词	.02	.08	.08	.07	.22**
anx	焦虑词	.04	−.08	.00	.19*	.11
sad	悲伤词	.01	−.12	.09	.23**	.12
cogMech	认知历程词	−.08	−.00	−.17*	.09	−.10
inclusive	包含词	.02	−.26***	−.00	.24**	−.05
exclusive	排除词	−.06	.16*	−.06	−.02	−.02
percept	感知历程词	−.11	−.04	−.06	.17*	−.13
hear	听觉词	−.28***	.00	−.05	.13	−.16
body	身体词	−.16*	−.07	−.06	.14	−.18
sexual	性词	.00	−.09	.05	.22*	.07
assent#	应和词#	−.04	.18*	−.03	−.08	−.13
swear#	脏话#	−.11	−.15	−.09	.20*	.00
Sub Eval+	主观评价+	−.02	.16*	.01	−.06	−.06
Casu Soc+	随意社交+	.12	−.16*	−.03	−.04	.13

注：上表数据是使用 kendall 相关得出的结果，# 表示 SC-LIWC 中非正式语词类别，+ 表示 SPCA-I 维度，*** p<0.001，** p<0.01，* p<0.05。

表 4 列出了性别、年级、心理距离与 SC-LIWC 类别、SPCA-I 维度存在的显著相关部分。性别与身体词积极情绪词、情感历程词正相关，即女生会更多的提到“脖子、皮肤、肠胃”等身体词，也会更多使用积极情绪词和情感历程词。心理距离和应和词正相关，即与聊天对方的心理距离越远，应和词会使用得越多。性别、年龄、心理距离与 SPCA-非正式语词维度只有性别与认知愉悦呈显著正相关（$r_{0.05}$ = .43，p = .00 < 0.001），即在聊天文本中更多表达认知愉悦语词的被试多是女性。

采用自然断点法（K-均值聚类）以区间分隔人格各维度的数值，将五个人格维度分为三个区间（1-低分组，2-中等组，3-高分组）。图 3 是在人格区间之间的总体不同质的部分，上排是外向性、开放性、自觉性，下排是神经质与宜人性。在神经质的自评分上，中高组在认知愉悦词汇的使用上，有较大的不同（M2 = .05，M3 = .30，M3 − 2 = .25，SD = .08，p = .001），在神经质上为自己打分极高的人更多使用认知愉悦语词。在宜人性上，为自己打分高的群体使用少量主观、社交填充、主观评价语词相对较

多，而在随意社交语词上使用较少（M2＝.95，M3＝.45，M3－2＝－.50，SD＝.25，p＝.04＜.05）。相关和LSD检验的结果与之前SC-LIWC中应和词与宜人性相关的结果是相符合的，并且深挖了非正式语词维度与人格维度之间的关系。

表4 性别、年级、心理距离与SC-LIWC、SPCA-I相关表

		性别	年级	心理距离
preps	介词	－.09	.07	－.28＊
quantunit	量词	.23＊	－.08	－.18
multifun	多用途词	.20	.12	－.23＊
futureM	未来时态标定词	.03	.27＊	.12
affect	情感历程词	.24＊	.03	.14
posEmo	正向情绪词	.24＊	.08	－.09
inhibition	限制词	－.03	.16	.24＊
inclusive	包含词	－.01	－.01	－.26＊
bio	生理历程词	.28＊	.05	.02
body	身体词	.31＊＊	－.03	－.04
space	空间词	－.25＊	.08	－.08
assent#	应和词#	.10	－.01	.28＊
Cog Plea+	认知愉悦+	.43＊＊＊	－.00	.05

注：上表数据是使用spearman相关得出的结果，#表示SC-LIWC中非正式语词类别，+表示SPCA-I维度，＊＊＊p＜0.001，＊＊p＜0.01，＊p＜0.05。

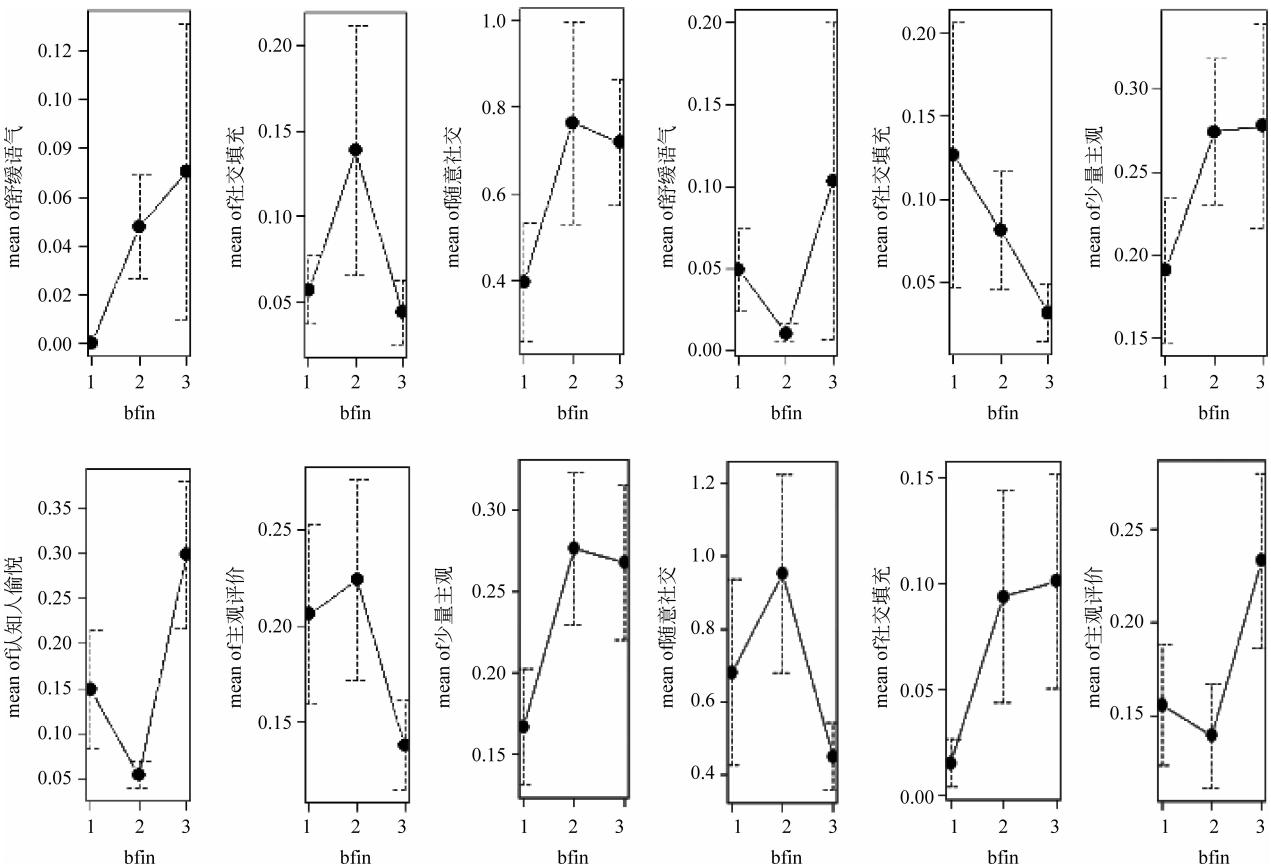

图3 均值标准误图

5 讨论

5.1 非正式情境下的语词—人格研究较少受到主题和心理距离的影响

以往研究更多关注于正式情况下的语词使用受到交流主题和交流双方关系的影响，或是实验下的语词使用，而现有研究Twitter、微博等非正式情境下的语词使用情境的较少。由于时代变迁且主题对语词使用的影响还有争议，所以本研究使用一种主题确定方法确定文档主题分布来研究网络即时交流下的SC-LIWC类别是否会受到主题不同的影响。

本文使用的判别主题方法不同于以往人们使用问卷或主观分类所得到的结果，它更客观地表现出

主题的不同，得出不同主题对 SC-LIWC 类别的影响。从之后人格与 SC-LIWC 类别的相关中可以获知，主题对有人格表征的 SC-LIWC 类别没有明显影响。

文本主题明显不同的情况下，结果显示主题集中相同的文本与主题明显不同文本之间的 SC-LIWC 类别较少。这一定程度上支持了 Pennebaker 在早先所提出的 LIWC 可以应用于跨主题的文本的理念[9-10]。非正式语词虽然和心理距离不相关，但是 SC-LIWC 中的应和词、介词和心理距离相关。在 Holtgraves 的研究中，认为最亲密的对象会使用最多的表情符号[13]。而在本研究的相关描述中并没有看出这一关系。可能是因为在测量心理距离这一变量时，主试所要求的“朋友”被被试泛化地理解，因此也需要记录两者现实关系。

5.2　稀疏主成分分析的非正式语词维度能挖掘出更多人格信息

非正式语词在本研究中主要针对 SC-LIWC 的非正式语词类别。在非正式语词对人格的表征上，语词可以降维后研究，也可以根据经验类别研究。但是降维方法有多种，需要一种适合非正式语词的降维方法；进一步来说，这种具有优势的降维方法与经验类别相比，是否更具有表征人格的能力？

在降维方法比较上，稀疏主成分分析所得的因子载荷相比主成分分析所得的方差旋转后因子载荷更优；更进一步，在对人格的表征上，稀疏主成分分析所得的非正式语词维度相比经验分类的非正式语词维度能挖掘出更多人格信息。总体上，在本研究中，稀疏主成分分析具有降维方法和人格表征上的优势。

由于本研究是基于非正式语词，并且具有小样本的特点。在对本研究中的非正式语词进行降维处理时，稀疏主成分分析相较于主成分分析，不仅因子载荷数更多，而且累积调整后方差更大。在对降维后的成分进行归纳并赋予意义时更容易解释。对于更大的样本，由于稀疏主成分分析的特点，载荷系数的稀疏性会保持，但是也许这些词会归于不同的主成分分类，模糊性的出现也增多，我们可以把选择非大学生群体作为被试来测试表 2 的主成分分类是否保持多少。不过可以预见，有些词，像日常生活中经常用到的“好的”，都属于主观评价这个主成分类别（在本研究的大学生被试中），即使在不同群体的被试中，很可能也属于主观评价这个主成分类别，这是因为有些词具有较稳定的社会交流功能，即各个阶层，年龄和职业的被试使用这些词都可以归于某个主成分类别名下。本研究由于涉及隐私，选择被试收到影响。所以，我们期望有进一步的研究来印证或检验。

性别与认知愉悦词正相关，即在交流过程中，女性比男性说更多愉悦词。这个结论与以往研究相符[39]。语词与人格的相关显示，在 SPCA-非正式语词维度中，宜人性与主观评价正相关、与随意社交负相关；在 SC-LIWC 非正式语词类别中，宜人性与应和词正相关。结合来看，宜人性得分高者可能会使用更多应和词，但是这些应和词可能更多是主观评价词，而且这些人会更少使用随意社交词。以往研究发现宜人性与确定相关词的正相关[40]，本研究则进一步确定了宜人性和主观评价词的关系。另外，得到的维度没有脏话词的原因可能是社会期望致使部分被试没有发送脏话，使得收集到使用脏话词的人很少。也有可能是在中国文化中，人们在与对方书面交流（即时交流）时碍于面子不会使用过于直白的脏话[41]，而有可能改用表情符号或者谐音。这样也能够解释宜人性与随意社交的负相关。这种由于文化不同而导致语词使用有所差异的现象值得进一步研究，研究者可以使用计算机化文本分析工具研究不同文化下的某类文本中语词使用的差异。

此外根据对人格维度的区间划分，大部分 SPCA-非正式语言维度在不同人格维度的区间上是不同质的，这种以均值标准误图直观地体现出人格维度高中低分组的语词使用特点。在神经质上为自己评分高的被试会比中低得分组更多使用愉悦词汇。同样有对 Facebook 的研究也发现神经质和愉悦词有较高的相关但因为没有现实线索所以使用线索判别神经质的准确性较低[42]。人类的笑是一种社会现象，被试在交流中的笑意味着他敏感地注意到社会信号并作出反应[43]，这种敏感特质可能是被试为自己在神经质上打分高的一个依据[44]。敏感特质是否是认知愉悦词和神经质之间的中介变量值得进一步研究，这可以帮助找到神经质的现实线索。

5.3　创新与局限

本研究的亮点在于不仅将主题模型融入非正式情境下语词使用的研究，还处理稀疏文本数据提取出非正式语词的心理—人格特征。当然，本研究也有局限性，首先是数据有较强的隐私性，因为这一数据来源于个体与他人的即时交流文本；其次，由于只测量了被试与某一朋友的心理距离而没有测量交流双方的实际关系，导致研究得到心理距离与语词使

用无关。

对非正式语词的研究不仅丰富了语词的心理—人格特征，而且帮助人工智能模拟、辨识非正式交流情境下的自然语言。

参考文献

[1] Allport G W, Odbert H S. Trait-names: A psycho-lexical study[J]. Psychological monographs, 1936, 47 (1): i.

[2] Saucier G, Goldberg L R. Assessing the Big Five: Applications of 10 psychometric criteria to the development of marker scales[J]. Big five assessment, 2002: 29-58.

[3] Laserna C M, Seih Y T, Pennebaker J W. Um... Who Like Says You Know Filler Word Use as a Function of Age, Gender, and Personality[J]. Journal of Language and Social Psychology, 2014: 0261927X14526993.

[4] Irvine C A, Eigsti I M, Fein D A. Uh, Um, and Autism: Filler Disfluencies as Pragmatic Markers in Adolescents with Optimal Outcomes from Autism Spectrum Disorder[J]. Journal of autism and developmental disorders, 2015: 1-10.

[5] Ervin-Tripp S. An analysis of the interaction of language, topic, and listener[J]. American Anthropologist, 1964, 66(6_PART2): 86-102.

[6] Fishman J A. Who speaks what language to whom and when? [J]. La linguistique, 1965, 1(Fasc. 2): 67-88.

[7] Chomsky N. Knowledge of language: Its nature, origin, and use[M]. Greenwood Publishing Group, 1986.

[8] Bybee J. Phonology and language use[M]. Cambridge University Press, 2003.

[9] Pennebaker J W, King L A. Linguistic styles: language use as an individual difference[J]. Journal of personality and social psychology, 1999, 77(6): 1296.

[10] 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.

[11] Slatcher R B, Chung C K, Pennebaker J W, et al. Winning words: Individual differences in linguistic style among US presidential and vice presidential candidates[J]. Journal of Research in Personality, 2007, 41(1): 63-75.

[12] Chung C K, Pennebaker J W. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language [J]. Journal of Research in Personality, 2008, 42 (1): 96-132.

[13] Holtgraves T. Text messaging, personality, and the social context[J]. Journal of research in personality, 2011, 45(1): 92-99.

[14] Tausczik Y R, Pennebaker J W. The psychological meaning of words: LIWC and computerized text analysis methods[J]. Journal of language and social psychology, 2010, 29(1): 24-54.

[15] Pennebaker J W, Graybeal A. Patterns of natural language use: Disclosure, personality, and social integration[J]. Current Directions in Psychological Science, 2001, 10(3): 90-93.

[16] Lee C H, Kim K, Seo Y S, et al. The relations between personality and language use[J]. The Journal of general psychology, 2007, 134(4): 405-413.

[17] Hirsh J B, Peterson J B. Personality and language use in self-narratives[J]. Journal of research in personality, 2009, 43(3): 524-527.

[18] Oberlander J, Gill A J. Language with character: A stratified corpus comparison of individual differences in e-mail communication[J]. Discourse Processes, 2006, 42(3): 239-270.

[19] Pennebaker J W, Boyd R L, Jordan K, et al. The Development and Psychometric Properties of LIWC2015[J]. UT Faculty/Researcher Works, 2015.

[20] Zhao D, Rosson M B. How and why people Twitter: the role that micro-blogging plays in informal communication at work[C]//Proceedings of the ACM 2009 international conference on Supporting group work. ACM, 2009: 243-252.

[21] Mairesse F, Walker M. Words mark the nerds: Computational models of personality recognition through language[C]//Proceedings of the 28th Annual Conference of the Cognitive Science Society. 2006: 543-548.

[22] Küfner A C P, Back M D, Nestler S, et al. Tell me a story and I will tell you who you are! Lens model analyses of personality and creative writing[J]. Journal of Research in Personality, 2010, 44(4): 427-435.

[23] Fullwood C, Quinn S, Chen-Wilson J, et al. Put on a smiley face: textspeak and personality perceptions [J]. Cyberpsychology, Behavior, and Social Networking, 2015, 18(3): 147-151.

[24] 黄金兰，Chung C K，Hui N，et al. 中文版[语文探索与字词计算]词典之建立[J]. The Development of the Chinese Linguistic Inquiry and Word Count Dictionary]. 中华心理学刊，2012，54(2)：185-201.

[25] 殷树林. 现代汉语话语标记研究[M]. 中国社会科学出版社，2012.

[26] 李成团. 话语标记语“嘛”的语用功能[J]. 现代外语，2008 (2)：150-156.

[27] Zhu J, Zhu M, Wang Q, et al. NiuParser: A Chinese Syntactic and Semantic Parsing Toolkit[J]. ACL-IJCNLP 2015, 2015: 145.

[28] Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts[C]//Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2013: 1445-1456.

[29] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, 2004: 487-494.

[30] 牛忠辉，蒋赛，邱俊杰，等. 社会距离对他人行为表征的影响：评价内容效价的作用[J]. 应用心理学，2011，16(4)：291-300.

[31] John O P, Donahue E M, Kentle R. 'The ''Big Five [J]. inventory—version 4a and, 1991, 54.

[32] John O P, Naumann L P, Soto C J. Paradigm shift to the integrative big five trait taxonomy[J]. Handbook of personality: Theory and research, 2008, 3: 114-158.

[33] 冉永平. 话语标记语的语用学研究综述[J]. 外语研究，2000 (4)：8-14.

[34] 孙利萍，方清明. 汉语话语标记的类型及功能研究综观[J]. 汉语学习，2011 (6)：76-84.

[35] Karen Tao Lok Sum. A study of the non-verbal politeness strategies in online chat conversations[D], 2013.

[36] 张明宇. 汉字“好”的语义功能研究[D]. 上海外国语大学硕士学位论文，2008.

[37] 姜其文. 试论主观增量标记“好”及其语用功能[J]. 励耘语言学刊，2015 (2)：185-197.

[38] 殷治纲，李爱军. “嗯”，“啊”类话语标记研究[C]. 中国计算技术与语言问题研究——第七届中文信息处理国际会议论文集. 2007.

[39] Robertson K, Murachver T. Intimate partner violence linguistic features and accommodation behavior of perpetrators and victims[J]. Journal of Language and Social Psychology, 2006, 25(4): 406-422.

[40] Hirsh J B, Peterson J B. Personality and language use in self-narratives[J]. Journal of research in personality, 2009, 43(3): 524-527.

[41] Tiejun W. A Review on the Study of the Concept of Mianzi and Its Function[J]. Psychological Science, 2004, 4: 040.

[42] Hall J A, Pennington N, Lueders A. Impression management and formation on Facebook: A lens model approach[J]. New Media & Society, 2013: 1461444813495166.

[43] Panksepp J, Burgdorf J. “Laughing” rats and the evolutionary antecedents of human joy? [J]. Physiology & behavior, 2003, 79(3): 533-547.

[44] Smolewska K A, McCabe S B, Woody E Z. A psychometric evaluation of the Highly Sensitive Person Scale: The components of sensory-processing sensitivity and their relation to the BIS/BAS and “Big Five”[J]. Personality and Individual Differences, 2006, 40(6): 1269-1279.

[45] Zou H, Hastie T, Tibshirani R. Sparse principal component analysis[J]. Journal of computational and graphical statistics, 2006, 15(2): 265-286.

附录

LIWC2015 与 SC-LIWC 的非正式语词表

	LIWC 词汇	SC-LIWC 词汇
脏话	af、arse、arsehole、arses、asf、ass、asses、asshole、asswipe、badass、bamf、bastard、biatch、biotch、bitch、bloody、bollock、boob、bs、bullshit、bumfuck、butt、buttfuck、butts、cock、cocks *、crap、crappy、cunt *、dammit、damn *、dang、darn、dick、dickhead *、dickhole *、dickish *、dicks、dickwad、dilf *、dipshit *、doofus、dork *、douche *、dtf、dufus、dumb、dumbass *、dumber、dumbest、dumbfuck *、dummy、effin、effin'、effing、fag、faggot *、faggy、fatties、fml、freak *、friggin、friggin'、frigging、fu、fuck、fuckboy *、fucked *、fucker *、fuckface *、fuckh *、fuckin *、fucks、fucktard、fucktwat *、fuckwad *、geek *、goddam *、half-ass *、halfass *、heck、hell、hella、hoe、hoebag、hoes、homo、homos、horseshit *、idgaf、idiot *、ignoramus、jackass、jeez、lmao、lmfao、mf、mf' *、mfs、milf *、mofo *、moron *、motherfucke *、motherfuckin *、nigga *、nigger *、omfg、piss *、prick *、pussies、pussy *、retard *、screw *、shit *、skank *、slut *、sonofa *、stfu、suck、sucked、sucks、tit、tits、titties、titty、twat *、wank *、whore *、wtf、wuss、wussy	瘪三、畜生、屌、屌、废物、狗屎、怪咖、花痴、机车、拷、靠、靠背、米共、孬、孬种、尼玛、娘儿们、娘娘腔、屁、泼妇、去你的、人妖、人渣、傻瓜、讨厌鬼、窝囊废、无耻、嘘、夭折、杂种、智障、笨蛋、不长眼、呆子、荡妇、怪胎、贱、白痴、蹩脚、操、放屁、该死、干、狗屁、混蛋、贱人、垃圾、烂货、妈的、呸、屎、他妈的、王八蛋、下流、淫妇、愚蠢

续表

	LIWC词汇	SC-LIWC词汇
网络语	:)、:(、4ev*、abt、af、afk、ahh*、aight、app、apps、asap、asf、atm、aw、aww*、b、b4、bae、bai、bamf、bb、bby、bc、bday、bf、bff*、bfs、biatch、biotch、boi、boo、brb、bruh、bs、btw、cc、cced、coz、cuz、d8*、da、dank、danker、dankest、dankness、dat、dawg*、deez、dilf*、dis、diss、dm、dope*、dotcom、dtf、dunno、eh、em、ep、ew、eww*、exbf*、exgf*、fam、fav、fb、finna、fml、fomo、frenem、ftw、fu、fwb、gf*、gg、gn、gon、gonna、gotta、grl*、gunna、gurl*、ha、hah、haha*、hai、hashtag、heh*、hella、heyy*、hii*、hm*、holla、ht、huh、idc、idk、ikr、ily*、ima、imma、j/k、jk、jus、juss、juz、k、kik、kinda、kk、l8*、lil、lil'、lmao*、lmfao*、lol、luv、meh、mf、mf'*、mfs、mhm*、milf*、mm、mmk*、mmo*、mo'、mofo*、msg、muah、mwah、nah*、nbd、nm、noes、np、nsfw、nvm、o、omfg、omg、oomf、ooo*、op、op's、outta、pic、pinterest、pls、plz、ppl*、prob、probs、prolly、r、reddit*、retweet*、rite、rofl*、rp、rt、selfir*、smh、sms、smsed、snapchat、sooo*、sp、stfu、tbh、teehe*、tf、tha、thanx、thnx、tho、tho'、thx、tix、tldr、tryna、tumblr、twitter、txt*、ty、u、unfollow*、unfriended、unfriend、unfriending、unfriends、ur、vid、w/*、wat、wattt*、whatt*、wif、woot、wtf、wut、xox*、ya、yaas*、yah、yea、yep*、yess*、yo、yu、yup	无
应和词	absolutely、agree、ah、aight、alright*、aok、aw、awesome、cool、duh、indeed、k、kk、mhm*、mmh*、mmk*、ok、okay、okey*、rt、uh-hu*、uhhu*、whoo、whooo*、woo、woohoo*、wooo*、yaas*、yah、yass*、yay*、yea、yeah、yep*、yes、yup	不错、呵、家家户户、就是说、酷、厉害、瞭、噢、唔、赞、啧啧、厂、厂厂、棒、屌、屌、嗯、嗯、嗯哼、哈、哈哈、呵呵、了不起、哦、水、喔、当然、对、好、嘿、了、了解、漂亮、是、哇、耶、真的、行、可、可以、明白、同意
停顿词	ah、ahh*、er、hm*、huh、mm、mmm*、oh、ohh*、sigh、sighed、sighing、sighs、ugh、uh、um、umm*、well、zz*	呃、啊、恩、就是、就是、然后、嗯、那
填充词	anyway*、blah、dunno、idk、idontknow、imean、ohwell、rr*、whoa、woah、y'kn*、yakno*、ykn*、youknow	等等、对了、恩、恩、话说、就是、就是、像是、之类、就

钟毓(1994—),硕士研究生,主要研究领域为社会情感分析与自然语言处理,心理语言学建模。
E-mail: y_chung@whu.edu.cn

费定舟(1969—),博士,教授,主要研究领域为计算社会科学,社会情感分析与自然语言处理。
E-mail: feeding_psy@whu.edu.cn