未来一代计算机系统 130 (2022) 114-127

内容列表可在ScienceDirect上找到

未来一代计算机系统

杂志主页: www.elsevier.com/locate/fgcs

SIMPA。从文本到项目匹配的人格评估☆

Matej Gjurković [a,*], Iva Vukojević [b,a], Jan Šnajder [a]

[a] 萨格勒布大学，电气工程和计算机学院，文本分析和知识工程实验室，Unska 3, 10000 Zagreb, Croatia
[b] 萨格勒布大学，人文和社会科学学院，心理学系，Ivana Lučića 3, 10000 Zagreb, Croatia

我的朋友们，

你们知道吗？

文章的历史。
2021年5月15日收到
2021年12月8日收到修改后的版本 2021年12月19日接受
可于2021年12月23日上网

关键词。
基于文本的人格评估 自然语言处理
文本分析 社交媒
体文本
现实的准确性模型 个性
预测

ABSTRACT

基于文本的自动人格评估（ATBPA）方法可以分析大量的文本数据并识别细微的语言人格线索。然而，目前的方法缺乏标准问卷工具所提供的可解释性、可解释性和有效性。为了解决这些缺陷，我们提出了一种结合基于问卷和基于文本的人格评估方法。我们的*声明-项目匹配人格评估*（SIMPA）框架使用自然语言处理方法来检测目标人物文本中的自我参照人格描述，并利用这些描述进行人格评估。该框架的核心是目标人物自由表达的语句和问卷项目之间的特质约束的语义相似性概念。概念基础是由现实准确性模型（RAM）提供的，该模型描述了准确的人格判断过程，我们用反馈循环机制对其进行扩展，以提高判断的准确性。我们在社交媒体网站Reddit上展示了SIMPA在ATBPA上的简单概念验证。我们展示了该框架如何直接用于无监督地估计目标人物的Big 5分数，并间接地为有监督的ATBPA模型产生特征，展示了Reddit上人格预测任务的最先进结果。

1. 简介

人格是指个体在思维、感觉和行为模式上的稳定差异，被称为*人格特征*[1]。这些差异已被证明与许多生活结果相关，如伴侣选择、工作选择、宗教或政治倾向以及个人兴趣等[2,3]。人格通常用人格调查问卷来评估[4]。这些问卷由一组自然语言陈述组成，称为*项目*，这些*项目*与特质有正向或负向关联。例如，"喜欢阅读具有挑战性的材料 "和 "避免哲学讨论 "是两个这样的项目，它们分别与智力或经验的开放性有正面和负面的联系。受试者以李克特式量表回答问卷，表明对陈述的同意程度，也就是说，这些陈述对他们的描述程度。

广泛使用的人格问卷，如NEO-PI-R[5]和BFI[6]，专门测量五种特质--外向性、自觉性、合群性、神经质和经验开放性--被称为大五人格[7]。

虽然人格问卷是评估人格的一个成熟的工具，但人格心理学的研究也将语言作为人格线索的另一个来源来研究（例如[8,9]）。人格和语言之间的联系早已被认可--词汇假设[10]提出了特征的重要性与描述它的语言中的词汇数量之间的相关性，而五大特征最初是由语言中与人格相关的描述的潜在结构得出的。随着大量的用户生成的文本在网上的出现，最近的研究已经研究了如何利用这些文本来进行人格评估。特别是社交媒体，用户在解释他们的行为、想法或情绪时描述他们自己和他们的个性，已经被认为是个性评估的一个重要来源。这就催生了结合人格心理学和计算机科学的研究，试图从大量的文本中实现人格评估的自动化。*基于文本的自动人格评估*（ATBPA）的潜力在于，它不仅可以有效地分析大量的

☆ 这项工作得到了克罗地亚科学基金会IP-2020-02-8671 PSYTXT（"基于文本的人格预测和分析的计算模型"）项目的部分支持，以及欧洲地区发展基金KK.01.1.1.01.0009 DATACROSS项目的部分支持。

* 通讯作者。
*电子邮件地址：*matej.gjurkovic@fer.hr（M. Gjurković），ivukojev@ffzg.hr（I. Vukojević），jan.snajder@fer.hr（J. Šnajder）。

https://doi.org/10.1016/j.future.2021.12.014

ATBPA不仅可以识别大量的文本数据，还可以识别人类通常无法察觉的细微个性线索，如语言风格，同时提供远远超出人类能力的一致性。例如，ATBPA避免了诸如受访者疲劳等问题，这限制了与结果质量相关的项目数量[11]，或受访者提供社会上不受欢迎的回答[12]。这表明，ATBPA方法可以被用来作为标准人格评估工具的补充。

目前的ATBPA方法依赖于自然语言处理（NLP），而这又在很大程度上依赖于机器学习（ML）。流行的方法是基于使用封闭和开放词汇的监督性ML[13,14]。近年来，使用深度学习模型已成为首选的方法[15-17]。有监督的方法是基于个性标签的数据来学习识别文本中的个性语言相关因素。然而，这类模型的明显局限性是需要标记的数据。这就造成了一个实际问题，因为公开可用的数据集很少，而那些可用的数据集也有很多缺陷，比如用户数量少，每个用户的文本数量少，或者缺少人口统计学数据。虽然缺乏数据集肯定会带来实际的挑战，但ATBPA也存在更严重的概念上的缺陷。一个基本的弱点是缺乏可解释性和有效性。直到最近，这些问题还没有在ATBPA研究中得到重视，大大限制了ATBPA方法的使用。然而，另一个弱点是只关注人格的高级建构，主要是领域（如五大领域的OCEAN），而不是更具体的低级特征，如方面[18]，面[5]，或细微差别[19]。在这项工作中，我们提出了一种新的ATBPA方法，旨在解决上述的弱点。它通过可解释和可说明的方式，提供比现有的ATBPA方法更多的有效性证据，并且能够输出per-sonality特质分类法中较低层次的构面的估计。在概念上，

我们的方法结合了基于问卷和文本的人格评估方法。现有的ATBPA方法侧重于预测的准确性，这在心理学上相当于收敛效度，并且只构成效度证据的一个来源[20]。相比之下，问卷调查有望满足所有的心理测量特性，尽管它们也有弱点，比如填写问卷所需的时间和精力。所提出的方法是试图取两者之长：一方面，我们从人格问卷中抽取项目；另一方面，我们搜索目标人物在社交媒体文本中写的相应语句。使用可靠的、可与更具体的人格特征相联系的问卷项目，提供了背景知识，使其具有可解释性和可说明性。更具体地说，将目标人物的陈述与问卷项目相匹配提供了可解释性（即，为什么模型会做出特定的决定），而项目与特质的预先建立的联系提供了可解释性（即，为什么这个决定是有意义的）。此外，声明的使用避免了词语中固有的模糊性。词语一直是目前ATBPA方法的主要分析单位，但词语在不同的语境中可能有不同的含义。此外，它们可以表示一个以上的特征，而且它们的使用通常取决于讨论的主题。我们的ATBPA方法通过依赖整个声明来减轻单个词所固有的模糊性。一旦检测到所有与问卷项目相对应的目标陈述，就可以在任何特质层面（如细微差别、面、方面、维度）进行汇总，同时考虑到关联的极性（项目关键），类似于标准人格评分的方式。问卷调查。

我们将所提出的方法发展成一个框架，重新称为"*声明-项目匹配人格评估*"（SIMPA）。该框架通过将这些声明与人格问卷中的项目相匹配，实现了在目标文本中找到自我报告的人格描述的想法。作为SIMPA的概念基础，我们使用了现实准确性模型[21]，该模型定义了准确的人格判断所需的四个连续阶段：线索的相关性、可用性、检测和利用。采用RAM使我们能够与人类法官评估撰写过文本的目标人物的方式相提并论。特别是，从与特质相关的问卷项目开始（相关性），我们从特定的文本来源中获取目标人物的陈述（可用性），找到与项目对应的陈述（检测），并将这些检测到的状态用于人格判断（利用）。与原始的RAM不同的是，判断是基于线索在四个阶段中的一次通过，SIMPA允许通过反馈回路机制进行多次通过。反馈回路的目的是通过提高检测相关线索的灵敏度，使该框架适应特定的数据来源（如社交媒体文本）。

在技术方面，SIMPA的主要挑战是检测与问卷项目相对应的目标陈述。如果声明的语义与问卷项目的语义非常相似，从而使其显示出相同的特质，那么该声明就是与问卷项目的良好匹配。我们用一个*特质约束的语义相似性的概念*来构思这个想法，它将语义相似性与关于某种特质如何表现的知识相结合。自动确定这种相似性与几个公认的NLP任务有关，包括转述识别、语义文本相似性、文本嵌套和自然语言推理。最近NLP的发展，尤其是那些基于深度表征学习的发展，在这些任务中产生了显著的进展。我们从这一进展中获益，并使用最先进的（SOTA）NLP模型，使语句与项目的匹配达到几年前不可能达到的准确度，使我们能够提出一个框架，旨在提供可解释、可说明和有效的人格评估。

为了证明SIMPA的可行性，我们提出了一个简单的概念验证，用于社交媒体网站Reddit上的ATBPA实施。我们展示了该框架如何直接用于无监督地估计目标人物的五大得分，并间接地为有监督的ATBPA模型产生特征。我们还研究了如何用更多的白话语句来扩展问卷项目集，以提高人格评估的准确性，以及如何通过SIMPA的RAM阶段的多次传递来进一步扩展项目集。我们的概念验证实现表明，即使有许多简化的假设，SIMPA也能在人格预测任务中取得SOTA的结果。

总而言之，这项工作的贡献有两个方面：我们
(1) 定义SIMPA，这是第一个基于RAM的ATBPA框架，为实现有效的ATBPA的可解释和可说明的模型奠定了基础；(2) 在Reddit上演示了ATBPA的SIMPA框架的实施。概念验证实施的令人鼓舞的结果表明，用于ATBPA的无监督模型可以是准确的、可解释的和可说明的，这使得它们比以前的方法在实践和研究方面都更有吸引力。尽管取得了令人鼓舞的结果，这项工作仅仅代表了这个方向的第一步，我们希望能鼓励对这个主题的进一步研究。

文章的其余部分组织如下。在第2节，我们描述了背景和相关工作。第3节和第4节描述了

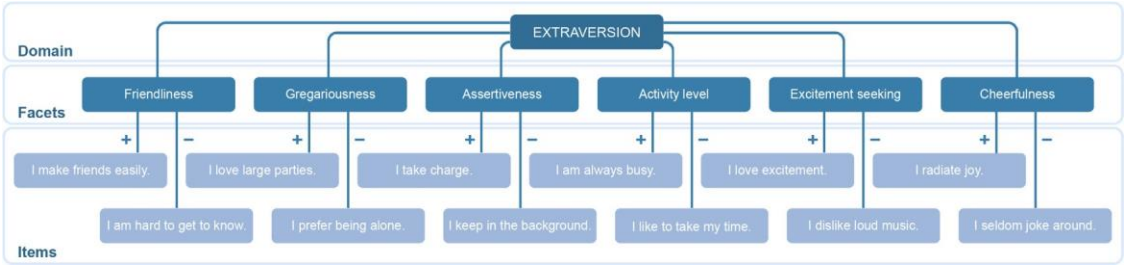


图1.IPIP-NEO问卷中外向性领域的特质层次结构。问卷项目与面相联系，而面相又被归类为领域。项目和面之间的联系是正向（+）或负向（−）的关键。

在第5节中，我们回顾了SIMPA通过四个RAM阶段和概念验证的实施情况。在第5节中，我们回顾了概念验证的实施和SIMPA框架的局限性和挑战，并讨论了未来工作的选择。第6节是结论。

2. 背景和相关工作

2.1. 性格模型

选择要评估的人格特征的前提是选择一个人格模型。基于不同的人格理论，采用不同的方法得出的人格模型有很多。有五个[5,22]或六个人格领域[23]的分层模型在人格心理学研究中应用最为广泛。这五个领域被称为 "大五"[5]，包括外向性、自觉性、合意性、神经质（或情绪稳定性[22]）和体验开放性（或智力[22]），而HEXACO模型[23]则将谦虚/谦卑作为第六个领域。一些模型还确定了每个领域内的层次性特征结构，包括各个方面、面和细微差别。更确切地说，每个领域由两个方面组成[18]。面是比方面更窄的特质，根据不同的模型，它们被归入方面[18]或直接归入域[5]。最窄的特质形式是细微差别，它对应于概念上冗余的问卷项目组，甚至是单个项目[19]。例如，一个在外向性方面具有平均分的人可能在自信方面很高，但在开朗方面很低。这个人可能会在小组任务中负责，但在任务中保持严肃，而如果分数颠倒过来，这个人就会等待别人带头，但成为小组的开玩笑者。

人格调查表是最常用的每...
子性工具[4]。这类问卷由一组项目组成--具体的、明显的变量，用于测量特质（细微差别、面、方面、领域），即潜在变量或结构。每个项目都标有它所测量的特质，一个 "键 "表示它所测量的特质的极性（"'表示特质名称的极性，'"表示相反极性+; 见图1）。由于问卷项目最初是为了适应预先确定的具有最高内部一致性的人格领域而构建的，因此可能有些细微差别没有被纳入其中，而且目前正在努力构建更多的项目，试图包括更多的细微差别[24]。许多项目在最大的人格库中是公开可用的。
项目，IPIP[25]，目前包含3320个项目。[1] 其中一些
衡量五大领域及其相应面的项目，如IPIP-NEO清单中的项目[25]。图1显示了IPIP-NEO清单中外向性领域的特质层次的一个例子。

[1] https://ipip.ori.org/。

我们的SIMPA框架的灵感来自于人格问卷的项目和评分过程，以及在不同层次的特质层次上评估特质的可能性。我们利用问卷中的项目来寻找目标人物文本中的语句，这些语句对应于如果对目标人物进行问卷调查时，目标人物对该项目的答案是什么。因此，项目被用作检测目标人物自由表达的语句的道具，这些语句有效地表明了人格特质。与现有的ATBPA方法不同，SIMPA允许在多个层次的特质上做出判断。该框架还能够扩大项目和细微差别的集合。

2.2. 现实准确性模型（RAM）

替代自我报告的人格评估是基于观察者的报告，要么是他人报告（在这种情况下，评判者通常与目标密切相关），要么是零距离判断（观察性研究中，评判者通常与目标是陌生人）。在这种情况下，理解人们如何对他人的特质做出准确的判断就变得非常重要[21]。Funder[21]提出的现实准确性模型（RAM）对这个过程进行了全面的概念化。该模型规定，从线索到准确的人格判断的过程分为四个阶段：线索必须是（1）相关的和（2）可用的，之后它必须被（3）检测到，最终（4）被法官利用。目标主要影响相关性和可用性阶段的成功，而法官主要负责检测和利用阶段的成功。到目前为止，该模型已被用于研究人们如何根据当面（如[26]）、录像（如[27]）或在线行为线索（如[28]）来判断他人的个性。在评估ATBPA方法的背景下，它也被考虑过（例如，[9,29]）。然而，据我们所知，目前还没有专门以RAM为模型的ATBPA方法。

RAM模型为困扰现有ATBPA方法的有效性问题提供了有用的见解[30]。有效性是指一个工具是否能测量相互之间的结构[31]。在实践中，有效性是通过支持对所获分数的拟议解释的证据来证明的[32]。在目前的ATBPA方法中，有效性通常是通过使用问卷中的人格分数作为基本事实、缺乏基于内容的有效性证据以及有限的可推广性来结束的[20,30]。迄今为止，大多数ATBPA研究都使用了有监督的ML模型，从而将人类的判断（来自自我报告或其他报告的问卷分数）作为人格的基础真理。因为人类的判断容易出错（例如，反应偏差和认知谬误），利用这种判断作为基础事实会使这些错误传播到预测的分数上。此外，正如Tay等人[30]所指出的，使用调查问卷的分数作为训练和评估监督预测模型的基础事实，会引起问题的循环，这使得预测的准确性在评估时没有用处。

模型的有效性。另一个问题是缺乏基于内容的有效性证据，这与ATBPA方法经常依赖语言信号有关，而这些语言信号只是人格特征的代名词（如主题兴趣），而不是真实、稳定和持久的人格模式。最后，将ATBPA应用于一个特定的社交媒体平台的数据，限制了研究结果对其他平台的推广，因为每个特定平台的特点都会影响到用户的个性特征的表达方式，这反过来又会影响到个性判断。

与现有的ATBPA工作不同，我们提出的SIMPA方法特意以RAM为模型，目的是确保可解释性和有效性，从而解决现有ATBPA方法的上述三个限制。与现有方法的另一个有趣的分歧点是，SIMPA同时使用了自我报告和观察研究的元素，因为它依靠自我报告问卷中的项目来检测观察数据中对这些项目的反应。

2.3. *基于文本的自动人格评估*

Bleidorn和Hopwood[20]描述了ATBPA的三代演变。第一代ATBPA研究[33,34]对个性在文本中的表现提供了初步的见解，通常是基于小的作者和文本样本，并使用简单的相关性或特征权重。第二代研究的特点是[15,35]，旨在提高对更大样本的预测能力，使来自有效工具（如Facebook myPersonality数据集[36]）的自我报告人格分数具有更大的统计能力。最后，第三代研究[9,30,37]关注不同的有效性和可靠性来源，以及ATBPA与传统人格评估方法相比的附加价值。我们的工作为第三代ATBPA方法做出了贡献。

RAM框架[21]对ATBPA方法提供了一个正交的视角。具体来说，三代ATBPA方法之间的主要区别点涉及他们使用的相关文本线索的类型和文本数据的来源。之前的工作已经考虑了广泛的相关线索（在ML术语中称为 "特征"），可以大致分为内容、风格或两者的组合特征。最常用的内容特征是单词、短语、单词类别，以及来自预先指定的列表（封闭式词汇法）或从文本本身提取的主题（开放式词汇法）[13,33]。相比之下，文体特征可以捕捉到文本的语言风格，通常包括子词级特征（如字符语法）、标点符号和特殊符号（如表情符号、感叹号），以及话语级指标（如可读性指数、内聚力指标、类型-符号比率）。最近的ATBPA方法依赖于深度学习代表，通常与上述语言学特征相结合[17,38,39]。第二个重要因素是数据的来源。ATBPA已经被用于电子邮件[40]、论文[41]、论坛[42]，以及最近的社交媒体平台，如Twitter[43]、Facebook[9,15]和Reddit[44,45]。数据的来源直接推动了线索的可用性，从而推动了不同类型特征的可用性。例如，作为线索的表情符号在商业邮件中的数量并不多，而在社交媒体的文本中则不同。SIMPA框架对数据源的选择是不可知的，同时也提供了一种手段，通过使用反馈循环机制来检测特定来源的相关线索。

在概念上，与我们最相似的是Vu等人[46]和Yang等人[47]的工作。他们提出的方法也将目标人物的文本与问卷项目联系起来，试图根据目标人物的文本直接预测他们将如何回答问卷项目。这与我们的方法不同，我们的方法是找到目标人物对自己表达的最相似的陈述。然而，关键的区别是，我们提出了一个人格评估的一般框架，它与技术实现和预期的应用无关。

到目前为止，有效性和可靠性问题在ATBPA研究中还没有得到太多的关注，有限的工作包含了人格心理学研究的标准做法，如检查不同的有效性来源（如内容、犯罪因素）和可靠性（即测试-恢复）[9,30,37,48]。在ML中，有效性与模型的可解释性和可解释性有关。可解释性是指在模型中可以观察到因果关系的程度，而可解释性反映了模型的内部运作可以用人的语言来解释的程度[49,50]。ATBPA模型应该既可解释又可说明，因为我们当然希望关于人类主体特征的决定是公平的，有可靠的证据支持，并且在需要时容易解释。使用文本作为人格线索的来源使得满足这些要求更加具有挑战性，因为语言的使用通常受到作者的人格和社会人口的影响，如年龄、性别和文化背景。然而，现有的ATBPA方法太容易被这些因素在数据样本中的分布差异所左右，导致模型的预测是基于语言特征和人格特征之间的虚假关联。可解释性的缺乏和混杂变量的存在会导致不确定的或有偏见的结果，甚至会带来道德上的挑战。应对这些挑战尤为重要，因为人格识别模型正在成为广泛使用的服务的一个组成部分，如对话系统[51]和推荐系统[52]。

我们的框架满足了人们对可解释和可说明的模型的需求，这些模型比现有的ATBPA方法提供了更多的有效性证据。此外，与其他方法不同的是，我们框架的实现既可以用于无监督的设置，也可以用于有监督的设置。它也适合于从每个作者的大量文本中提取人格指示性线索，并利用这些线索来获得不同层次的特质判断。

3. SIMPA框架

陈述-项目匹配人格评估（SIMPA）框架是基于将目标的陈述与问卷项目或类似的特质相关状态相匹配的想法。这种匹配是在一对陈述之间进行的：一个特质指示性陈述和一个特质相关陈述。我们将特质指示性*陈述*（TIS）定义为从目标人物的文本中提取的可以作为人格评估线索的陈述，我们将*特质相关陈述*（TRS）定义为与特定特质有已知和有效联系的陈述。TRSes的集合最初只包括问卷中的项目，但正如我们在下面讨论的那样，随后可以扩展到包括被认为与某一特定特质相关的TISes。

图2展示了SIMPA框架的概况。该框架以RAM为基础，将人格判断过程分为四个阶段：线索的相关性、可用性、检测和利用。相关线索是指TISes，它们在源文本中必须有足够的数量，才有可能做出准确的判断。检测阶段包括根据特质约束的语义相似性的概念将TISes与TRSes相匹配。最后，在利用阶段，检测到的TISes被赋予相关性分数，这些分数被汇总以产生特质级分数。

采用RAM作为SIMPA的基础也有助于识别管道中的薄弱点，即有多少和什么样的错误从一个阶段传播到下一个阶段。特别是。

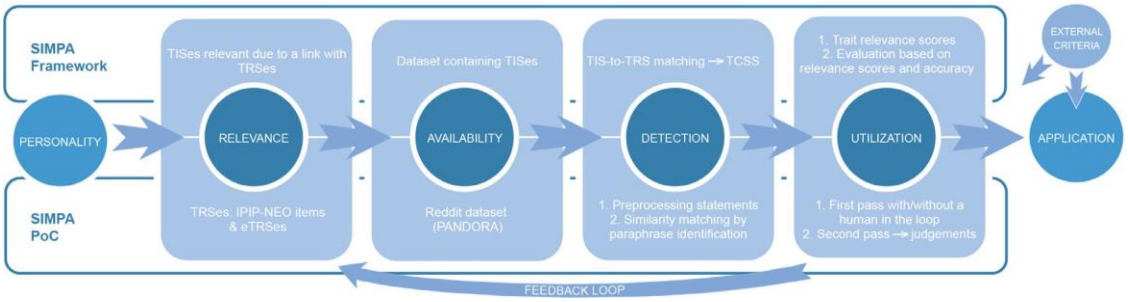


图2.图中显示了SIMPA框架（第3节）和SIMPA的概念验证（第4节），分为四个阶段，即相关性、可用性、检测和利用，伴随着反馈回路和对所获得的人格判断的应用。因此，SIMPA传达了相关文本线索（TISes）通过所有阶段的（反复）流动的理念，从而产生了可用于不同应用的人格判断。

我们将SIMPA中的错误分为两类：*不作为的错误*和作为的*错误*。前者涉及到没有检测到相关的线索或没有将相关的线索与特质联系起来，而后者则涉及到检测到不相关的线索或将相关的线索与不正确的特质联系起来。这两种类型的错误都会降低人格判断的准确性。我们设计的框架允许减轻这些错误，它包括一个反馈循环机制，使多次通过四个RAM阶段来获得特定数据的TRSes，从而增加检测TISes的可能性。

接下来我们将更详细地描述SIMPA框架的四个RAM阶段。

3.1. 相关性

相关性是RAM的起点；一个人要想准确判断一个人的特征，被判断的人必须做与该特征相关的事情[21]。例如，一个外向的人需要有与外向有关的思想和情感，或者采取与外向有关的行动；否则，即使不是不可能，也很难判断这个人是外向的。在SIMPA中，我们认为相关性是TIS的内在属性，并进一步假设TIS的相关性只能通过求助于作为TIS有效性证据的外部知识来确立。这与迄今为止的大多数ATBPA研究相反，在这些研究中，线索相关性被等同于为预测问卷分数作为标签而训练的监督模型的特征权重，这导致了低可推广性和可解释性，并危及这种方法的有效性。SIMPA通过将TISes与TRSes相匹配来回避这个问题，其中TRSes是由心理学家精心设计和验证的针对特定人格特征的问卷项目。其原理是，如果目标人物使用的语句与问卷项目所表达的意思相同，那么法官就可以认为该语句是有基于内容的有效性证据支持的有效人格线索。换句话说，一个TIS与某一特质相关，是因为它与与该特质相关的TRS很匹配。

我们将TISes区分为两大类：*人格的自我概念*和*人格表现*。个性自我概念的陈述是描述性的陈述，在这些陈述中，被试者明确地用个性来定义自己（例如："我是一个有纪律的人"）。相比之下，人格表现不像自我概念那样可以自我分类，可以进一步分为*特质参考*和*个人行为*。我们将特质参照定义为：说话者对已知的人格信号的情感、认知或行为进行可概括的参照（例如，"我从不拖沓 "作为高自律性的提示）。与此相反，个人行为是与目标描述的人格指示性行为有关的陈述（被称为*行为报告*；例如，"我在中午前完成了所有的家务 "作为高自律性的线索

例如，"我下班后会回答你 "作为高自律性的提示）或目标已经执行的（称为*行为观察*；例如，"我下班后会回答你 "作为高自律性的提示）。

与以前的ATBPA工作中使用的其他文本线索相比，TISes至少有三个原则性的优势：(1)明确而强烈的人格信号，(2)不容易受到上下文意义的改变，以及(3)与领域无关。有些TISes（特别是人格的自我概念和特质参考）是非常明确和强烈的人格信号。然而，即使是不太明确的TIS类型也与TRSes相对应，因此与目标的自我暗示有关。作为线索的TIS不容易发生意义上的改变，因为它们通常包括足够的上下文来可靠地确定TIS和TRS之间的语义关系，即TIS复制TRS的程度。例如，"我喜欢聚会 "这个说法比 "聚会 "这个词更不容易发生意义上的改变，后者也可能是 "我不是一个喜欢聚会的人 "这个句子的一部分。最后，TISes在某种程度上是独立于领域的；也就是说，同样的TISes出现在不同的社交媒体领域，尽管它们的分布肯定取决于社交媒体平台的特性。我们接下来讨论这个问题和其他方面如何影响TIS的可用性。

3.2. 可利用性

相关线索需要提供给法官，以便用于性状判断。在ATBPA中，相关线索的可用性主要取决于文本数据的来源，理想情况下，这将最大限度地提高相关线索的数量和质量。在SIMPA中，这相当于为一组给定的TRSes找到相关TISes的可能性最大化。TISes的可用性受到与交流环境、目标和特质有关的几个因素的影响。

当社交媒体文本被用作数据来源时，与环境相关的因素包含了一系列与交流平台（例如，对文本交流的依赖、对文本格式的限制、主题焦点、交流的同步性）及其用户（例如，匿名性、人口统计学）相关的特征[53]。例如，在Twitter和Reddit上，用户大多通过文本信息进行交流，而在Instagram和TikTok上，交流大多是视觉的。此外，一些平台对信息长度有限制（如Twitter），而其他平台则没有这种限制（如Reddit）。用户在平台上的匿名性也会影响到TIS的可用性；当用户认为自己是匿名的，他们往往会在网上提供更多关于自己的信息[54,55]。另一个重要的特征是平台是否以主题为重点，这可能会影响可用的TIS的分布。例如，LinkedIn通常用于与工作有关的目的；因此，我们可以预期与工作有关的TIS在LinkedIn上比在Reddit这样一个主题多样的平台上更容易获得。专题重点可以

与另一个影响TIS可用性的因素有关，即自我介绍的需要，因为某些话题的讨论需要用户提供更多关于他们自己的背景信息。例如，在关于个人关系的讨论中，用户通常会提供更多与个人有关的自我描述。TIS的可用性也受到平台上不同社会规范的影响，以及在其不同的子社区内。例如，在一些社区（如Facebook上的初为人父母的群体），互动的一个重要部分是互相鼓励，这可以增加个人行为类型的TIS的可用性。最后，平台的选择也会影响到用户的人口统计学，而这又会与个性差异以及TIS的可用性相关[30]。例如，Reddit的平均用户是年轻男性[56]，这可以增加与这个年龄或性别群体相关的TISes的可用性（例如，"我喜欢成为喧闹人群的一部分"）。

影响TIS可用性的其他因素有：与目标相关的因素。这些因素，如社会期望值和焦油发现自己的情况。有些人比其他人更容易以社会理想的方式表达自己[12]。这影响了人格评估[12]；因此，我们可以期望对这些目标有更多的社会理想的TIS。此外，某些情况会唤起某些特质[57]；因此，在线交流中出现的情况可能会影响线索的可用性。例如，被指责说谎会触发合作方面的线索，这些线索要么是正向的，要么是负向的。

TIS的可用性也取决于目标人物的特质。一般来说，与许多高度可见和经常可用的线索相关的特质，比与不太可见和不太可能可用的线索相关的特质判断得更准确[26]。在TISes的情况下，我们有理由期待并非所有的特质，或所有特质的两个关键，都有相同的可能性在在线文本交流中表现出来。可能是一个人本身就很难意识到某个特定的特征，从而将其表达为TIS。一个例子是自我意识：这个特质低的人不关心别人如何看待他们的行为，我们可以预期自我意识的负键TIS比正键的TIS少。其他与特质相关的可得性因素可能涉及某些特质的社会不可取性（例如，负键的自律性），甚至是对理想特质的吹嘘。

以上确定的影响TIS可用性的因素可能是也是相互影响的。例如，匿名可以鼓励用户谈论不同的，甚至是可耻的话题，从而降低社会期望值效应[58]。社交网络Reddit是一个很好的例子，在这个平台上，用户是匿名的，这可能会让他们在行为选择上更加灵活（例如，展示他们的弱点）。

3.3. 检测

在RAM中，检测阶段只是关于法官注意到相关的和可用的个性线索[59]。在SIMPA中，检测阶段的目标是在目标的文本中检测出尽可能多的TISes。这是通过尝试将每个目标的陈述与TRSes集合（最初是问卷项目）中的每个陈述进行匹配来实现的。这里的挑战是TISes和TRSes之间的语言差距。TISes是从自然主义文本中提取的白话语句（例如，"尴尬的情况是我的游戏"），而TRSes则是精心措辞的语句，传递了与人格相关的情感、认知或行为模式的核心含义（例如，"我不为不同的社交场合所困扰"，来自IPIP-NEO）。由于这种差距，仅仅依靠TISes和TRSes的逐字匹配会导致许多遗漏的错误。

在SIMPA中，我们将TIS与TRS的匹配操作为两个陈述之间的相似性：如果一个陈述与一个给定的TRS相同或高度相似，那么它很可能是一个TIS，并且与该特定的TRS的相同特征相关。请注意，有关的相似性是一个分级的概念，包括但也超越了两个陈述的语义等同性（即转述关系）。更确切地说，一个TIS和一个TRS之间的相似性必须(1)涵盖两个陈述之间的语义相似性和(2)确保这两个陈述都表明了相同的基本特征。例如，"我只是讨厌拥挤的地方 "这一陈述既与TRS "我避免拥挤 "高度相似，又以相同的关键指示了相同的外向性面（聚集性）。另一方面，"我在拥挤的地方工作"只是与TRS相似，但不表示相同的特质。相反，"我不喜欢星期六在商场里 "的陈述是对同一特质的指示，但在语义上与有关的TRS不太相似。我们把这种涵盖语义相似性和同一特质的指示性的相似性称为*特质约束的语义相似性*（TCSS）。

一般来说，人类所感知的两个语句之间的语义相似性取决于两个状态的上下文[60]。然而，我们将TCSS概念化为一种与语境无关的相似性度量，对应于两个语句在其 "典型 "或 "默认 "语境中使用时的语义相似性。即使有这样的限制，我们也注意到TCSS应该捕捉人类和专家知识的各个方面，包括语言知识（自然语言语义和语用学）、常识性知识和推理，以及与评估人格特征有关的社会文化知识。因此，TCSS可以被理想地设想为测量对一个问卷项目（TRS）的李克特量表的同意程度，当回答被表达为一个自然语言陈述（TIS）时。换句话说，当目标的陈述表达了对TRS的同意时，TCSS最好是高，否则就是低。由于涉及许多相互作用的语言现象，为TCSS提供一个更具体的定义是困难的。虽然原则上可以通过将TCSS的理想属性形式化为相似性度量（例如，相似性应该在否定的情况下恢复，或者如果声明是专门的或认识论上的对冲，它应该减少）来得到一个更精确的特征，但我们将这个有趣的方向留给未来的工作。

在存在词汇空白的情况下，确定两个自然语言语句之间的语义相似程度是NLP中一个被充分研究的问题，它被不同程度地归结为转述检测、语义文本相似性和识别文本连带关系[61-63]。成功地解决这些任务需要语言理解能力，涵盖广泛的语言现象，如否定、同义和反义、语义关系和词义连带关系，以及逻辑和常识推理[64,65]。最近关于深度表征学习的研究[66,67]已经为这些任务开发了高度精确的模型（例如，[68]）。这样的模型产生了SOTA的结果，并且可以在SIMPA框架的这个阶段使用现成的模型，因为它们可能会很好地减少遗漏和错误。然而，由于这些模型不是专门为TCSS设计的，一些错误将不可避免地漏掉。虽然这可以通过调整现有的SOTA模型以适应TCSS任务来缓解（例如，通过转移学习[69]），但这种调整并非易事，而且需要一个标有TCSS分数的数据集。

TCSS的NLP模型的选择直接决定了遗漏和失误的数量。当一个相关的线索没有被检测到时，就会发生遗漏错误，我们预计这主要发生在模型未能分配一个

如果一个白话语句在语义上等同于一个特定的TRS，就会出现高的TCSS得分。相反，当模型通过给不相关的语句赋予过高的TCSS分数而将其检测为相关语句时，就会发生委托错误。我们预计这种情况会发生在措辞非常相似的语句拥有不同含义的情况下。明显的例子是包含否定词或反义词的语句；例如，"我爱艺术 "可能会错误地与 "我不喜欢艺术 "匹配。在这种情况下，TIS并不指示与它相似的特定TRS的相同特质，但它仍然是特质指示性的。这表明，委托错误可以进一步分为两种类型：（1）*信息性错误*，当TIS与现有的TRS不能很好地匹配时发生，但它对法官来说仍然是特质指示性的，因此，甚至可以被认为是一个新的TRS；（2）*非信息性错误*，它不能为法官提供任何有用的特质指示性信息（例如，一个TIS "我感觉受到欢迎 "与TRS "我让人感觉受到欢迎 "匹配）。在SIMPA的利用阶段，我们利用前一种类型的委托错误来反复改进线索检测过程。

3.4. 使用情况

SIMPA框架的最后一个RAM阶段是利用，由目标产生的TIS被用来准确判断目标的个性特征。这里的主要挑战来自于以下事实：TIS与某一特定特质的相关性是一个程度问题；几个TIS可能作为一个特质的线索，因此它们的相关性必须以某种方式结合起来；低层次特质的线索相关性结合成高层次特质的线索相关性；如果要使判断尽可能准确，必须考虑到早期RAM阶段的错误。

考虑到这一点，SIMPA的利用分两步进行。

(1) 对于每个目标，我们递归地确定特质层次结构中各级特质的相关分数，从最低级别的TISes开始，一直到特质域的级别，同时考虑到低级别特质的相关和数量，以及（2）根据获得的相关分数和外部标准，通过反馈循环机制，决定是否再通过RAM阶段。这些步骤说明了这样一个事实，即人格特征是一个分层次的建构体，它由底层的细微差别组成，这些细微差别与面相联系，而面又与各个方面相联系，最后与特征领域相联系（参见图1）。此外，这些步骤还能让我们了解到每个低层次特质的贡献，因为对高层次特质的判断是基于对低层次特质的综合判断来进行的。

相关性评分。我们将某一性状线索的相关性分数定义为一个评分函数的输出，该函数将性状层次中较低层次的性状线索的相关性分数汇总（例如，通过计算总和、平均值或加权平均值）。最低级的是TRSes，每个TRS的相关性分数是通过汇总匹配的TISes的相关性分数得到的，我们将其定义为TCSS分数（参见3.3节）。直观地讲，对应于单个TRS的线索的相关性是由与该TRS相匹配的TISs的综合证据决定的，同时考虑到TCSS得分所表明的该匹配的强度。然后，TRS的这种相关性分数可以进一步汇总，以产生与这些TRS相关的更高层次（如面或方面）的特征相关的线索的相关性分数。通过这种方式，线索的相关性分数可以沿着层次结构一直传播到性状域的层面。更重要的是，法官可以选择她希望获得相关性分数的特质的层次。

除了与较低级别的特征相关的相关性分数外，相关性分数一般取决于背景，这我们认为这包括线索的数量以及语言以外的环境（例如，子社区的社会规范、人口因素）。数量在这里很重要，因为线索的数量会影响判断的信心；法官对每个目标的每个特征的判断的TIS越多，她的判断就越有把握。例如，如果为一个目标检测到的TISes包括 "我读过尼采"、"我读过一篇关于该主题的科学论文 "和 "我刚读完一本霍金的书 "等语句，与只检测到其中一个TISes相比，法官可以更确定该目标在智力方面是高的。然而，低线索数量可以通过高相关性来弥补。例如，与TRS "我喜欢阅读有挑战性的材料 "相匹配的单个TIS是一个足够强大的线索，可以判断目标在智力方面是高的。

反馈回路。为了提高在TISes和TRSes之间存在词性差距的情况下判断的准确性（参见第3.3节），SIMPA通过增加一个从利用阶段回到可用性阶段的反馈回路来扩展RAM。这使得它有可能在RAM的四个阶段中进行迭代，直到满足某个标准。更确切地说，反馈回路有两个相互交织的目的：(1)通过扩大带有TIS的TRS集，使TIS检测适应源文本语言；(2)在没有达到预期的信心标准的情况下增加判断的信心，例如在特征层次结构的某个层次上每个特征检测到的TIS的最低数量或相关性分数的最低水平。是否回环的标准通常是相关性分数和判断准确性的函数。所需的准确性水平由外部标准决定，例如，基于与评估同一特质的其他变量（如自我报告）的相关性的有效性证据。

在回环之前，TRSes的集合被扩展为那些在文本中检测到的TISes，这些TISes带有足够的特质相关的形成，可以作为TRSes使用。这里的直觉是，在RAM管道的下一次迭代中，更大的TRSes集合将检测到更多的相关和可用线索。具体来说，在两种情况下，TIS可以被提升为TRS：（1）具有足够高的TCSS的TIS，这基本上是TRS的转述，和（2）TIS是检测阶段的信息错误的结果（参见第3.3节），即TIS是高度指示性的特征，但与现有的TRS不是很匹配。第一种情况可以被看作是领域适应，因为它使得用源文本特有的语句来扩展TRSes的集合成为可能。第二种情况可以被看作是涵盖某些特征的新的细微差别的一种方式。然后，扩大的TRSes集被用于下一次通过RAM阶段，其中可用的相关线索再次被检测为TISes并被用来确定相关性分数。这个循环可以多次启动，直到没有新的TISes可以被提升为TRSes或达到所需的信心标准。

4. 概念验证的实施

本节介绍了我们在社交媒体平台Reddit上对ATBPA的SIMPA框架的概念验证实施。我们特别选择了Reddit，因为它的特点可能会促进TISes的可用性。该实施方案使用来自人格调查问卷（IPIP-NEO）的问卷项目作为相关的背景知识（TRSes）。然而，我们还调查了由专家编制并适应Reddit语言的额外TRSes是否能提高检测到的线索的相关性和数量。通过TIS-TRS匹配来检测相关线索（TISes），是使用SOTA NLP模型进行转述检测，作为TCSScomputation的代理。我们还在利用阶段引入了一些简化的假设。我们通过管道做了两遍，以证明反馈循环机制的效用，使TRSes适应Reddit上使用的语言。在

第二遍，我们利用检测到的TISes来获得五大特征的线索相关性分数。最后，我们在两个不同的应用中使用这些分数：作为五大特征分数的原始估计，展示SIMPA在无监督ATBPA中的应用；作为有监督ATBPA模型的特征，我们在Reddit上建立了一个新的人格预测的SOTA结果。

4.1. Reddit是一个数据来源

Reddit[2]是一个流行的社交媒体平台，估计有超过1,000万用户。4.3亿的 用户。[3] 有几个特点使Reddit在其他社交媒体网站中脱颖而出，成为人格研究的宝贵数据来源：（1）用户的匿名性，鼓励用户更自由地表达他们的想法；（2）主题的多样性，因为Reddit包括200多万个被称为subreddits的子社区，在其中可以表达人格的许多方面；（3）每个用户的文本质量高，数量大，这增加了提供人格指示性线索的可能性。

在我们的概念验证中，我们使用了PANDORA[45]，这是一个数据集，包含了Reddit用户的自我报告人格分数，通过挖掘他们的文本中披露的人格问卷结果获得。用户自我报告的五大人格分数是非常有价值的，因为它们为新的ATBPA方法提供了收敛有效性证据，并且能够直接比较其预测准确性。在PANDORA中，有自我报告的大5分的用户总数为$n\ 1$，608。这些用户共写了130万条评论，包括总共1430万个句子。从每个用户如此多的文本中提取和汇总TISes是一个公开的挑战。然而，我们假设这不是我们方法的障碍；相反，SIMPA在更多的数据中才能更好地工作，因为更多的数据增加了TISes的可用性。

4.2. 相关性

SIMPA框架使用诸如TRSes这样的问卷项目，这些项目都是经过精心设计的，并由领域内的前辈们验证过的，以针对特定的人格特征（参见3.1节）。概念验证的实施使用了IPIP-NEO调查问卷中的300个项目[25]。为了促进TIS与TRS的匹配，我们在每个项目的开头加上代词 "我"，将其转换为一个自我参照的句子（例如，"经常感到忧郁 "变成 "我经常感到忧郁"）。我们已经确定，这样做可以提高对相关线索的检测。

4.3. 可利用性

在SIMPA中，TISes的可用性取决于与环境、目标和特质相关的几个因素（参见第3.2节）。与环境有关的因素使Reddit特别适合于ATBPA：匿名性（它减轻了社会欲望的影响，增加了自我介绍的需要），用户的数量，每个用户的文本数量，以及主题的多样性，所有这些都与TISes的预期数量有正相关。为了获得数据中线索可用性的感觉，我们将TISes的候选数量近似为包含代词 "我 "的句子的数量。我们发现有将近520万个这样的句子，相当于数据集中的36.3%的句子。这个相对较高的比例表明，Reddit上的文本可能确实包含大量的TISes。

4.4. 检测

TIS检测包括通过TCSS（参见第3.3节）将目标的陈述与TRS相匹配。我们通过两个步骤来实现这一点。(1)文本预处理和(2)相似性匹配。

在预处理中，我们首先使用Spacy[70]的Sentencizer将每个Reddit用户的评论分成句子，将每个句子视为可能的TIS候选者。虽然句子可能更符合TRSes，但我们使用整个句子来简化和加快预处理，因为否则就需要进行句法分析。此外，大多数SOTA NLP模型都是在句子级别的数据上训练的。然而，我们预计这种简化会导致遗漏和错误的增加。我们还过滤了那些包含代词 "我 "的句子。

一旦我们获得所有目标的句子集，我们就进入相似性匹配步骤。我们选择将TCSS操作为转述检测和语义文本相似性，以便能够使用现成的预训练模型完成这些任务。这种简化使我们能够避免模型的微调，因为微调需要一个标记有正确的TIS-TRS匹配的数据集。然而，这也导致了行为错误（因为一些TIS会被错误地匹配到TRS）和遗漏错误（因为一些TIS不会被正确地匹配到TRS）的增加。虽然存在大量用于这些任务的NLP模型，但我们测试了SentenceTransformer软件包[68]中实现的三种独特而常用的模型：（1）基于连体网络的语义文字相似性模型[68]。(2) Komnios word2vec模型[71]，以及（3）基于RoBERTa的转述检测模型[68]。我们选择这三个模型是因为它们在概念上是不同的，因此在我们的任务上可能产生不同的表现。

使用这些模型，并与当前NLP[68]的做法一致，我们计算TIS候选者和TRS之间的相似性（TCSS得分）如下。我们将每个TIS编码（即 "嵌入"）为高维向量空间中的一个向量，从而使语义相似的TIS具有相似的向量。然后，我们将所有TIS-TRS对的语义相似性简单地计算为其相应向量之间的余弦。我们将这些相似性存储在一个矩阵中，其行对应于目标的句子，其列对应于TRSes。对于每个目标（Reddit用户），我们为三个NLP模型各建立一个这样的矩阵。因此，目标评论中的每个句子将与每个NLP模型的所有TRSes的相似性分布相关。然而，为了简单起见，我们只考虑每个句子中相似度最高的TRSes作为匹配。实际上，这意味着我们允许每个TIS只与一个TRS匹配。

以上描述的检测阶段的实现，尽管是简化的，但仍然涉及一些设计决定。两个主要的设计决定涉及到用于匹配的NLP模型和余弦相似度的阈值。为了进一步研究这个问题，我们为每个模型抽取了100个具有最高相似度分数的句子，并对其性能进行了定性分析。这一分析提供了两个关键的见解：（1）相似度分数的阈值直接影响到相关性和通过该阈值的TIS的数量，以及（2）不同模型的相似度分数在不同的TRS中差异很大。特别是，与其他两个模型相比，在固定的相似性阈值下，转述检测模型在不同的TRSes中表现出最稳定的相关性和数量。释义检测模型也是三个模型中捕获TCSS最好的，并且总体上有一个更好的信息性错误和非信息性错误的比例。为了进一步简化实施，我们决定只使用转述检测模型。

2 https://www.reddit.com。

3 https://backlinko.com/reddit-users#reddit-statistics。

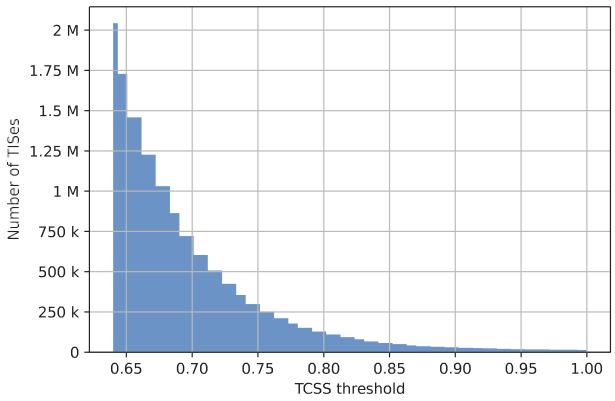


图3. 在不同的TCSS阈值下检测到的TISs的数量。

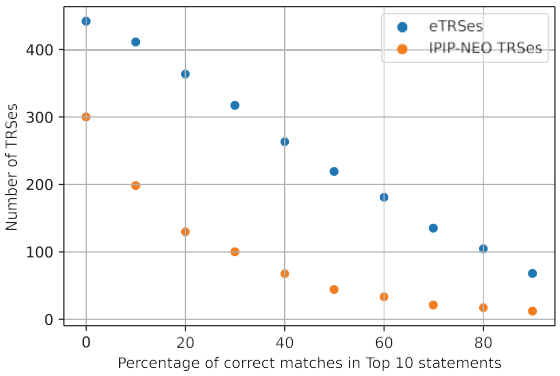


图4. 在TCSS最高的前10个TIS候选人中，不同比例的正确匹配的TRSes（橙色为IPIP-NEO TRSes，蓝色为eTRSes）的数量。例如，在按TCSS排名的前10名TIS中，有100个IPIP-NEO TRSes有超过30%的正确匹配TISes。

4.5. 使用情况

回顾一下，SIMPA中的利用目标是根据检测到的TISes及其TRS对应物来估计特定目标的相关性分数（参见第3.4节）。因此，第一步是根据TCSS和上下文来计算检测到的TISes的相关性分数。在选择使用哪些TISes以及如何计算其相关性分数方面，我们试验了两种情况。在第一种情况下，我们引入了一个简化的假设，即相关性分数等于TCSS，有效地忽略了
TISes的任何背景。然后，我们选择一个阈值并宣布
所有相关性得分高于该阈值的声明，以

是TISes。图3显示了TISes的数量与TIS的函数关系。
相似性阈值，表明阈值的选择
在很大程度上影响了检测到的TIS的数量。

上述情况既简化了检测（通过使用
语义相似性作为TCSS的代表）和利用率（通过等价交换的方式）。
与TCSS的相关性得分）。这就提出了一个问题，即
如果有更多的人参与，判断的准确性可以提高多少？
在这两个阶段使用了复杂的方法。我们在
在第二种情况下，通过在循环中加入人类专家（人格心理学家）来验证这一点。对于集合中的每个TRS

在最初的300个IPIP-NEO项目中，专家被要求对以下内容进行注释
根据余弦相似度，从数据集中找出与之最相似的20条语句。

所用的注释方案基本上是一个二元方案，区分正确和不正确的匹配，后者进一步细分为六种类型，五个信息性错误和一个非信息性错误。这样就有了总共七个类别：（1）正确的匹配（TIS与TRS具有相同的一般性和相同的极性），（2）具有相同的一般性和相同的极性，（3）不那么一般性和相同的极性，（4）不那么一般性和相反的极性，（5）指向平均得分项目，（6）同一领域的其他项目/方面，以及（7）其他（一个非信息性错误）。作为一个例子，考虑TRS"我总是有准备"。七个类别的相应TISes的例子如下。(1) "我总是为任何事情做好准备"，(2) "我从不准备"，(3) "我有准备"。
(4) "我没有准备就来了"，（5）"我从来没有完全准备好，但也不是没有准备"，（6）"我总是按部就班地安排事情"，（7）"我准备了一餐"。

图4显示了在TCSS最高的前10名TIS候选人中，不同比例的正确匹配的TRSs数量，作为TRSs质量的代表。每个TRS检测到的TISes的质量有相当大的差异。在基于IPIP-NEO的300个TRS中，只有44个具有超过50%的正确匹配。表1显示了有注释的TIS候选人的比例（前20条声明与TRSes匹配的比例为

表1
被注释的TIS候选人的比例（通过TCSS与IPIP-NEO TRSes和eTRSes相匹配的前20条陈述），按每个五大领域的注释类别细分（O-开放性；C-自觉性；E-外向性；A-合群性；N-神经质）。类别（列）。OK--正确的匹配；G--与TRS的一般程度相同，但极性相反的陈述；LG--不那么一般的陈述，极性相同；LG
- 平均值--指向 平均TRS表达的陈述，SD--与同一领域的其他项目/面相关的陈述；NOK--非信息性错误。

			领域TRSes信息性					
		好的1	G-错误	LG+234	LG-	平均5分	SD6	7挪威克朗
O	IPIP-NEO	24.0	7.7	21.2	5.0	1.1	3.8	37.2
	eTRS	36.2	4.5	16.1	4.0	2.6	2.3	34.3
C	IPIP-NEO	13.2	1.8	27.9	4.5	0.3	2.8	49.5
	eTRS	33.6	5.2	22.9	5.1	1.2	2.4	29.6
E	IPIP-NEO	21.2	3.4	25.1	5.6	0.6	0.8	43.2
	eTRS	33.9	5.7	23.5	4.6	4.0	1.1	27.2
A	IPIP-NEO	18.9	2.8	17.8	5.9	0.2	1.4	53.1
	eTRS	33.5	5.0	20.1	5.3	0.6	2.5	33.0
N	IPIP-NEO	22.8	1.9	35.2	4.3	1.2	1.0	33.5
	eTRS	29.1	3.6	32.7	4.8	2.8	0.6	26.3

考虑到每个人格领域的所有注释类别，我们观察到在IPIP-NEO TRSes检测到的人格领域中，正确匹配的比例以及信息性和非信息性错误的比例存在明显的差异。我们观察到，在用IPIP-NEO TRSes检测的人格领域中，正确匹配的比例以及信息性和非信息性错误的比例有明显的差异。这些差异可能是由于PANDORA[45]中特质分布的倾斜，或者仅仅是因为在某些领域，与社交媒体文本中的TISs相匹配的IPIP-NEO TRSes较少。使TRSes适应社交媒体语言可能是缓解后者的一种方式。

在这两种情况下，我们在RAM阶段进行了两次传递，在进行第二次传递之前，我们扩大了最初的TRSes集合。在第一种情况下，我们使用0.7的阈值来决定哪些TISes要提升为TRSes，而在第二种情况下，我们只提升被专家标记为与TRS正确匹配的TISes。在这两种情况下，我们将新的TRSes与它们所匹配的TRSes的特征联系起来。在第二步之后，如果语句的相似度分数高于0.8的阈值，我们就宣布它们是TISes。与第一步一样，我们将TISes与最相似的TRS的特征相联系。

在得到TISes之后，我们继续计算性状层次中不同层次的性状的线索相关性分数。在这里，我们引入另一个简化，将所有TISes的相关性分数值固定为1。然后，我们将所有TIS的分数相加，成

为每个关键的值的分数（±1）；最后，我们对

领域特征。为了获得数据集中目标的性状的相对表达，我们根据其他目标的得分计算每个目标的相对百分位数得分。我们通过计算每个特质的正向和负向键入分数的比例，按照比例分数对目标进行排序，然后根据这些比例分数为每个目标和每个特质分配百分位分数。

4.6. 延长TRSes

在第二种情况下，基于专家的注释的结果显示，许多原始的IPIP-NEO项目往往导致正确匹配的TISes的比率较低。因此，我们研究由IPIP-NEO项目组成的TRSes集是否可以用额外的专家制作的声明来扩展。 我们把这套附加项目称为eTRS。一位心理学专家根据对构成IPIP-NEO面的项目的解释，以及人们期望在每个面得分高或低的人在社交媒体平台（如Reddit）的评论中写下的自我描述，编制了eTRSes。eTRSes可以分为三种类型：(1)精确的IPIP-NEO项目的转述（例如，"我不努力成功 "用于IPIP-NEO项目 "我没有高度的成功动机"），(2)与现有细微差别相关的新项目（例如。"我不努力 "用于可能存在的细微差别，包括 "我只做足够的工作来过日子 "和 "我在工作中投入很少的时间和精力"），和 (3) 提及新的细微差别的项目（例如，"我不在乎赢或输 "作为代表对胜利漠不关心的新细微差别的一个项目）。在制作电子描述符时，专家遵循以下原则：电子描述符应该是（1）一个强烈的面的信号，（2）比标准项目更白话化，（3）一般性的（只有在特定性使线索更相关时才具体化，例如。例如，"我生气的时候会打东西"），（4）措辞要增加依赖语义相似性的NLP模型正确匹配TISes的可能性（即避免使用隐喻，要简洁），以及（5）不仅仅是通过否定来否定（例如，"我总是迟到"，而不是 "我从不准时 "来否定低自律性）。因此，eTRSes大多是以特质参考和自我人格概念的形式出现的（参见3.1节）。由于eTRSes被写成了特质参考，我们希望它主要涵盖特质参考的TISes，而在较小的程度上涵盖个人行为和自我概念（例如。一个eTRS "我读了很多非小说 "可以链接到一个相当的自我参考TIS，一个行为报告TIS "我读了2020年最好的非小说书"，一个行为观察TIS "我建议你多读非小说"，和一个自我概念TIS "我是一个非小说类的人"）。作为自我概念的eTRSes被期望与自我概念TISes相匹配，但它们也可能与特质参考相匹配，并在较小的程度上与个人行为相匹配（例如，eTRS "我是一个有创造力的人 "可能与自我参考TIS "我做了很多创造性的DIY "相匹配）。请注意，与IPIP项目相比，自我概念类型的eTRSes是全新的。

我们进一步完善了eTRSes，以提高其检测正确匹配的TISes的能力。我们通过两个步骤来实现这一目标，即更密切地关注eTRSes和TISes之间的语义相似性（1）。
(2)在eTRSes和PANDORA中被保留下来的一组语句之间，包括没有Big 5基础真实分数的目标的评论（n8684）。在第一步中，我们考虑eTRSes之间的相似性分数，并在eTRSes的集合中加入与两个不同面或领域相关的语义高度相似的句子。这减少了检测与同一领域的其他TRSes或面相关的TISes的信息错误，并提高了正确匹配的检测率。在第二步中，我们从持有的Reddit数据集中抽取eTRSes和一组句子，并通过余弦相似度对这些句子进行排序。然后，如果这些句子具有相同的含义，但措辞不同，我们就用这些句子扩展现有的TRSes集合。我们考虑到语义相似性和TRSes是否正确地与预设的特征相联系，删除或改进对某一特征不是很好的TRSes（即大多数类似的句子都不是很好地击中）。

结果是453个与五大领域和30个IPIP-NEO面相关的新项目的清单。创建和完善eTRSes的总工作量约为100小时，而注释工作则花费了约40小时。虽然，随后的结果（例如，基于与这些项目匹配的人格分数和黄金标记的分数之间的相关性）表明，该列表是有效的，但它应该在进一步的研究中得到更仔细的验证。

我们将使用原始IPIP-NEO项目中的TRSes检测到的TISes的相关性和数量与使用eTRSes检测到的TISes进行比较。相关性是根据检测到的与两组TRSes相匹配的TISes的注释（参见）和不同TCSS阈值下TISes的数量来估计的。图4显示，eTRSes在正确匹配的TISes数量方面优于IPIP-NEO项目。例如，在300个IPIP-NEO项目中，只有44个（14.7%）在前10个最相似的语句中有50%或更多的正确匹配，而453个eTRSes中的219个（48.3%）也是如此。表1显示了对IPIP-NEO和eTRSes检测到的TIS-es的相关性的更详细的比较。与IPIP-NEO TRSes相比，除了检测到更多正确的TISes外，eTRSes产生的非信息性错误更少，正确匹配、信息性错误和非信息性错误之间的比例更平衡。这些结果表明，创建更适合于在社交媒体上用语言表达个性的TRSes是一项值得努力的工作。

4.7. 应用

我们考虑了SIMPA框架在ATBPA任务中的两种应用：(1)一种无监督的设置，我们使用利用阶段的百分位数分数作为目标人物的Big 5分数的估计值；(2)一种有监督的设置，我们用基于利用阶段相关性分数的特征来扩展基于文本的人格预测的回归模型。后者在PANDORA Reddit数据集上实现了SOTA性能[45]。

无监督的设置。在第一个应用中，我们使用五大特征中每个特征的百分位数分数作为总的相关性分数（参见第4.5节）。我们通过与PANDORA中目标人物的自我报告的五大特征分数进行比较，来检验基于这样一个简单的相关性评分函数的五大特征估计是否提供了有效性证据。

与概念验证的实施一样，我们简化了实施并固定了一些参数的值。具体来说，我们只使用在最相似句子的前10名中至少有一个正确匹配的TRS。 我们还将TCSS的阈值设定为0.6。最后，我们只考虑那些对至少一个五大特征有超过10个检测到的TISes的目标，因为检测到的TISes的数量在同一目标的不同特征之间高度相关（$r>0.85$）。图5显示了在所有目标中检测到的TISes的数量，按所有五大特征的正负键进行细分。我们观察到，在大多数面和特征中，正向和反向键的TISes数量是不平衡的。例如，对于开朗的一面，积极的TIS比消极的TIS多六倍。这种差异可能表明，某个特定键的TISes的可用性较高，数据集中目标人格特征的分布不平衡，或者某些面的两个键之一的TRSes的覆盖率不足。

表2显示了与TIS匹配的TIS的估计分数和地面真实的Big 5分数之间的皮尔逊相关系数。

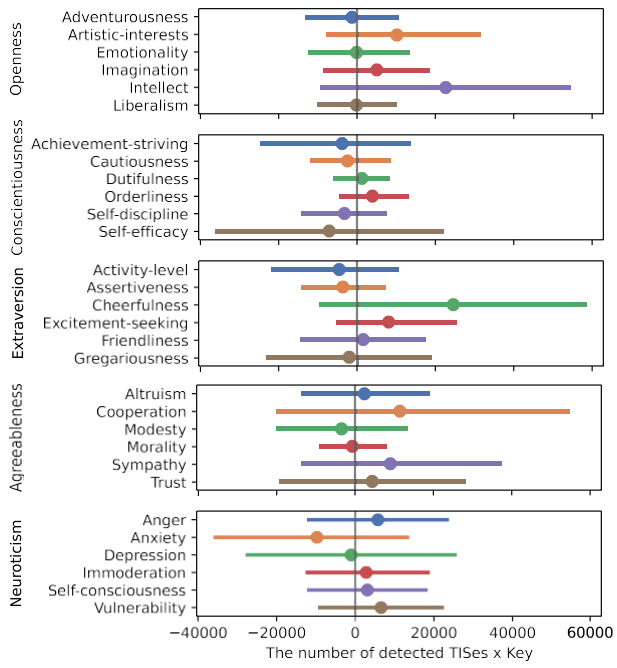


图5.在所有目标中，每个领域和面所检测到的TISes数量。点代表每个面的正向和负向键入的TISes总数之间的差异。线条显示了正向和负向键入的TISes数量的范围。

表2

使用不同的TRSs获得的百分位数估计值与五大领域（O- openness；C-conscientiousness；E-extraversion；A-agreableness；N-neuroticism）的ground truth score之间的Pearson相关系数。显著的相关关系（$P < .05$）以黑体字显示。

	TRSesDomains					OCEAN
IPIP-NEO TRS	O	.099	.055	.020	.136	**.203**
	C	**.231**	.179	**.227**	**.190**	**.220**
	E	.066	.113	**.285**	.075	.013
	A	.003	.010	.053		**.204**
	N	−.088	−.088	**−.186**	−.146	**−.253**
eTRS	O	**.143**	.045	.074	.084	−.061
	C	.006	**.160**	.018	.052	.104
	E	.040	.064	**.241**	.034	.122
	A	.012	.054	.067		**.154**
	N	.036	−.025	−.003	−.072	.121
合并的	O	**.141**	.013	.067	**.165**	.066
	C	**.126**	.102	**.102**	−.003	−.089
	E	.070	**.075**	**.226**	**.138**	−.041
	A	.027	−.021	**.106**	**.185**	.036
	N	−.025	−.064	−.063	−.086	**.106**

基于IPIP-NEO项目的TRSes，eTRSes，以及两者的组合。由于上述限制，目标的样本量小于开始的n 1，608个有自我报告的大5分的目标。更具体地说，每次实验的目标数量，IPIP-NEO TRSes为155个，eTRSes为280个，如果我们考虑两套TRSes，则为399个。结果表明，即使在SIMPA框架的检测和利用阶段的实施中引入了许多简化措施（参见第4.4和4.5节），也有证据表明收敛性和判别性的有效性。具体来说，只使用eTRSes，除了神经质之外，我们

表3

有和没有SIMPA 基于TIS的特征的PANDORA最佳表现预测模型的Big 5分数和地面真实分数之间的皮尔森相关系数。相关性的显着差异（$P < .05$）显示在黑体中。

领域	PANDORA-最佳	PANDORA-best+SIMPA
开放性	.265	.285
自觉性	.273	.304
外向性	.387	**.458**
认同度	.270	.287
神经质	.283	.312

有监督的设置。在监督设置中，我们使用目前在PANDORA数据集上表现最好的模型[45]，并使用基于TIS的特征来扩展它。该模型是一个使用Tf-Idf加权单字和基于MBTI和Enneagram分类器的预测作为特征的Ridge回归。[4] 我们将所有在与由IPIP-NEO TRSes和eTRSes组成的初始TRSes匹配时TCSS至少为0.6的句子宣布为TIS。我们还将TCSS高于该阈值的TISes纳入到基于人类专家验证的正确匹配语句的反馈循环迭代之后的TRSes扩展集中。作为模型特征，我们使用相关性评分函数的输出，计算为所有五大面和领域的正向和负向键入的TISes之和。然后，我们将PCA（有10个主成分）分别应用于原始相关性分数矩阵和所有目标的行归一化分数矩阵。这使我们能够为数据集中的每个目标获得一组固定的20个密集特征，缓解了TIS的稀疏性问题。为了能够直接比较结果，我们采用了与[45]相同的交叉验证程序和相同的折线。

表3显示了原始PANDORA最佳表现模型和我们基于SIMPA的扩展模型的所有目标（n 1，608）的五大特征预测分数和地面真实分数之间的皮尔逊相关系数。我们的最终模型，只使用了20个额外的基于TIS的特征，在所有五个五大特征上都取得了新的SOTA结果。对于外向性特征，我们取得了统计学上的重大改进（P 0.018，双尾Steiger's检验依存关系[73]），其相关度达到了.458.这些结果是令人鼓舞的，原因有二。首先，它们表明基于TIS的特征是对单词单格和基于其他人格模型预测的特征的补充。第二，考虑到在检测和利用阶段引入的许多简化，这些结果表明，在预测的准确性方面仍有很大的改进空间。

5. 讨论

设法在所有感兴趣的特质上实现了显著的正相关（收敛效度），而特质之间没有显著的相关（判别效度）。然而，效果大小表明，还有很大的改进空间。

拟议的SIMPA框架依赖于通过它们与TRSes的匹配来检测TISes，两者都是用自然语言表达。要在嘈杂的社交媒体数据上实现这一目标，需要解决新颖而不简单的NLP任务。在概念验证的实施过程中，我们证明了该方法的可行性，并在此过程中引入了几个简化的假设。更确切地说，在检测阶段，我们简化了预处理，将TISes建模为句子，而句子则更为合适。我们将TCSS的计算操作化为一个转述检测任务，使TIS与TRS的匹配一般不受特质约束的影响。我们还将匹配工作分解为两个步骤（嵌入句子

4 迈尔斯-布里格斯类型指标（MBTI）模型[72]和恩纳格是基于类型的人格模型。虽然在大众中被广泛使用，但它们却被大多数人格心理学家所不齿。

然后再计算余弦相似度），尽管联合进行这些工作可能更有效率。在利用阶段，相关性评分功能过于简化：TCSS分数被用作TISes的相关性分数，而且只考虑了最大的TCSS分数，而不是法官可用的整个相似度分布。此外，每个TRS、每个面的TRS和领域层面的面的预期TISes的基本比率没有被考虑在内，因为权重被固定为1。

虽然是可行的，但概念验证的实现过于简单，在实际应用中并不实用。未来的实施应该考虑建立在更复杂的模型上，以更好地适应任务的复杂性。努力的重点应该是相关性评分功能和TCSS。这两项任务都是新颖的，其难度也是未知的，但开发出接近人类水平的模型可能需要大量的努力。相关性评分功能应该包括语言外的上下文信息，而有效的TCSS的实施必须考虑到话语层面的现象，如共同参考和可在句子层面解决的歧义。一个有希望的方法是在专家注释的数据集上使用度量学习[74,75]来训练TCSS模型，类似于Reimers和Gurevych的工作[68]。

关于整个SIMPA框架，其局限性和适用性还有待彻底测试。然而，某些挑战和未来工作的机会已经显现出来了。主要的挑战来自于这样一个事实，即与每个参与者对每个项目都做出反应的自我报告相比，我们的方法并不包括每个参与者的所有TRS的TISes。这损害了内容有效性的证据，即线索与结构相关并完全覆盖结构（例如，TISes应该与一个面相关并应该覆盖所有面的细微差别）[32]。后者在我们的框架中很难实现。另一方面，由于该框架提供了关于线索和有效项目之间匹配的信息，因此在线索与建构相关的意义上，该框架有利于基于内容的证据。

未来的工作可以研究该框架如何被用于不同的领域以及它的心理测量特性。可以通过使用多个领域的数据来探索领域适应性。同样，可以通过比较框架在不同时间段的数据上的表现来评估可靠性的心理测量特性。这种测试-评估的设置进一步导致了框架在长期研究中的可能扩展。未来工作的另一个有趣的方向是将该框架应用于评估人格以外的结构。任何已经开发出来的问卷项目，原则上都可以用SIMPA来评估。
- 例如，态度、心理健康，甚至是身体健康的症状。这为该框架在不同研究领域的应用指出了机会。

6. 总结

声明-项目匹配人格评估（SIMPA）框架结合了基于问卷和文本的人格评估方法。它使用自然语言处理方法来检测目标人物文本中对个性的自我参照描述，并利用这些描述进行个性评估。据我们所知，这是第一个以声明为基础的文本人格评估方法，重点是确保可解释性、可解释性和有效性，也是第一个直接以人格判断的现实准确性模型（RAM）为模型的框架。我们通过在Reddit上进行基于文本的人格评估的概念验证，证明了SIMPA的可行性。我们用它来直接获得我们将目标人物的5大评分的估计值与基于自我报告的5大真实评分进行了关联，对感兴趣的特质实现了正相关，并证明了收敛性的有效性证据。SIMPA的第二个应用是一个有监督的设置，在这个设置中，我们扩展了一个回归模型，并通过基于SIMPA人格判断的特征实现了SOTA结果。未来的工作应该致力于建立更复杂的NLP模型，作为这个框架提出的新的和复杂的NLP任务的解决方案。未来工作的另一个有趣的方向是使用SIMPA来评估人格以外的建构，即任何可通过问卷调查项目来测量的建构。

CRediT作者的贡献声明

Matej Gjurković:概念化、方法学、验证、形式分析、调查、数据整理、写作-原稿、可视化、软件。**Iva Vukojević:**概念化，调查，数据整理，可视化，写作-原稿。**Jan Šnajder:概念**化，写作-原稿，调查，监督，项目管理，资金获取。

竞争性利益的声明

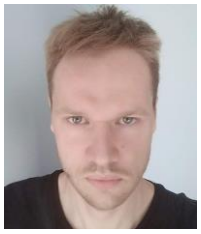
作者声明，他们没有已知的竞争性财务利益或个人关系，可能会影响本文所报告的工作。

参考文献

[1] D.C. Funder, Accurate personality judgment, Dir.Psychol.Sci. 21 (2012) 177-182.
[2] L.M. Larson, P.J. Rottinghaus, F.H. Borgen, 六大兴趣的元分析 和五大人格因素，J. Vocat.Behavior.61 (2002) 217-239.
[3] C.J. Soto, 人格特征与连续生活结果之间的联系有多大的可复制性？性格复制的生活结果 项目，Psychol.Sci. 30 (2019) 711-727.
[4] R.J. Larsen, D.M. Buss, 《人格心理学》。关于人性的知识领域 ，麦格劳-希尔出版社，2009。
[5] J.P.T. Costa, R.R. McCrae, Domains and facets: Hierarchical personality assessment using the revised neo personality inventory, J. Personal. Assess. 评估。64 (1995) 21-50.
[6] O.P. John, E.M. Donahue, R.L. Kentle, Big five inventory, J. Personal. Psychol. (1991).
[7] L.R. Goldberg, 《语言与个体差异》。在个性词汇中寻找 普遍性，Beverly Hills，1981年，第141-165页。
[8] Y.R. Tausczik, J.W. Pennebaker, 词语的心理学意义。LIWC和计算机化文本分析方法，J. Lang.Soc. Psychol.29 (2010) 24-54.
[9] G. Park, H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, M. Kosinski, D.J. Stillwell, L.H. Ungar, M.E. Seligman, Automatic personality assessment through social media language. , J. Personal.Soc. Psychol.108 (2015) 934.
[10] G.W. Allport, H.S. Odbert, Trait-names:A psycho-lexical study, Psychol.Monogr. 47 (1936) i.
[11] M.Galesic, M. Bosnjak, Effects of questionnaire length on participation and indicators of response quality in a web survey, Public Opin.q. 73 (2009) 349-360.
[12] R.R. Holden, J. Passey, 人格评估中的社会理想反应。不一定是假的，也不一定是真的，Pers.Individ.差异。49 (2010) 446-450.
[13] A.H. Schwartz, J.C. Eichstaedt, M.L. Kern, L. Dziurzynski, S.M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M.E. Seligman, et al., Personality, gender, and age in the language of social media:开放词汇 ，PLoS One 8 (2013) e73791。
[14] J.W. Pennebaker, R.L. Boyd, K. Jordan, K. Blackburn, The Development and Psychometric Properties of LIWC2015, Technical Report, University of Texas at Austin, 2015.
[15] V.Lynn, N. Balasubramanian, H.A. Schwartz, 用户个性预测的分层建模。信息层面的注意力的作用，在:计算语言学协会第58届年会论文集，计算语言学协会，2020，第5306-5316页，在线。
[16] D.Xue, L. Wu, Z. Hong, S. Guo, L. Gao, Z. Wu, X. Zhong, J. Sun, Deep learning-based personality recognition from text posts of online social networks, Appl. Intell.(2018) 1-15.

[17] Y.Mehta, N. Majumder, A. Gelbukh, E. Cambria, Recent trends in deep learning based personality detection, Artif.Intell.Rev. 53 (2020) 2313-2339.
[18] C.G. DeYoung, L.C. Quilty, J.B. Peterson, 在面和域之间。10 方面的大五，J. Personal.Soc. Psychol.93 (2007) 880.
[19] R.R.McCrae，对可靠性的更细致的看法。特质的特殊性 层次结构，Pers.Soc. Psychol.Rev. 19 (2015) 97-112.
[20] W.Bleidorn, C.J. Hopwood, Using machine learning to advance personality assessment and theory, Pers.Soc. Psychol.Rev. 23 (2019) 190-203.
[21] D.C. Funder, On the accuracy of personality judgment: a realistic approach, Psychol.Rev. 102 (1995) 652。
[22] L.R. 戈德堡，人格的另一种描述。The big-five factor structure, J. Personal.Soc. Psychol.59 (1990) 1216.
[23] K.Lee, M.C. Ashton, HEXACO 人格的心理测量特性 inventory, Multivar.Behavior.Res. 39 (2004) 329-358.
[24] R.Mõttus, T. Bates, D.M. Condon, D. Mroczek, W. Revelle, Leveraging a more nuanced view of personality:Narrow characteristics predict and explain variance in life outcomes, 2017, PsyarXiv.
[25] L.R.戈德堡，一个宽泛的、公共领域的、测量几个五因素模型的低层次面的人格量表，蒂尔堡，，荷兰，1999，第7-28页。
[26] T.D. Letzring, S.M. Wells, D.C. Funder, Information quantity and quality affect the realistic accuracy of personality judgment, J. Personal. Soc. Psychol.91 (2006) 111.
[27] D.R. Carney, C.R. Colvin, J.A. Hall, A thin slice perspective on the accuracy of first impressions, J. Res. Personal.41 (2007) 1054-1072.
[28] S.D. Gosling, A.A. Augustine, S. Vazire, N. Holtzman, S. Gaddis, 在线社交网络中的个性表现。自我报告的脸书相关行为和可观察的个人资料信息，网络心理学 Behav.Soc. Netw. 14 (2011) 483-488.
[29] W.Youyou, M. Kosinski, D. Stillwell, 基于计算机的人格判断比人类的判断更准确, Proc.Natl. Acad.Sci. 112 (2015) 1036-1040.
[30] L.Tay, S.E. Woo, L. Hickman, R.M. Saef, 人格评估的机器学习方法的心理测量和有效性问题。关注社会 媒体文本挖掘，Eur.J. Pers.34 (2020) 826-844.
[31] D.Borsboom, G.J. Mellenbergh, J. Van Heerden, The concept of validity, Psychol.Rev. 111 (2004) 1061。
[32] A.E.R.协会，A.P.协会，N.C. on Measurement in Education，Standards for Educational and Psychological Testing，美国教育 Research Association，华盛顿特区，2014。
[33] J.W. Pennebaker, L.A. King, 语言风格。语言使用是一种个人 ，J. Personal.Soc. Psychol.77 (1999) 1296.
[34] S.Argamon, S. Dhawle, M. Koppel, J. Pennebaker, 人格类型的词汇预测因素, in: Interface和分类协会联合年会论文集，2005年，第1-16页。
[35] M.Kosinski, S.C. Matz, S.D. Gosling, V. Popov, D. Stillwell, Facebook作为社会科学的一个研究工具。机遇、挑战、伦理 ，以及实用指南。，Am.Psychol.70 (2015) 543.
[36] M.Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, Proc.Natl. Acad.Sci. 110 (2013) 5802-5805.
[37] V.Kulkarni, M.L. Kern, D. Stillwell, M. Kosinski, S. Matz, L. Ungar, S. Skiena。H.A. Schwartz, Latent human traits in the language of social media:An open-vocabulary approach, PLoS One 13 (2018).
[38] Y.Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, S. Eetemadi, Bottom-up and top-down:用心理语言学和语言模型特征预测人格，在:2020年IEEE国际数据挖掘会议，ICDM，2020，1184-1189 页，http://dx.doi.org/10.1109/ICDM50108。2020.00146.
[39] A.Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, E. Cambria, Personality trait detection using bagged SVM over BERT word embedding ensembles, 2020, http://arxiv.org/abs/2010.01309[arXiv:2010.01309] 。
[40] J.Oberlander, A.J. Gill, 有特色的语言。电子邮件交流中个体差异的分层语料库比较，话语 过程。42 (2006) 239-270.
[41] K.Luyckx, W. Daelemans, Personae:一个用于从文本中预测作者和个性的语料库，在:第六届语言资源与评估国际会议论文集（LREC'08), 2008, pp. 2981-2987.
[42] F.Iacobelli, A.J. Gill, S. Nowson, J. Oberlander, 博客的大规模人格分类, in:情感计算和智能交互, Springer, 2011, 第568-577页。
[43] P.-H。Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, V. Sinha, 25条推文来了解你。一个用社交媒体预测人格的新模型，在:第十一届AAAI网络和 社会媒体国际会议论文集，2017，第472-475页。
[44] M.Gjurković, J. Šnajder, Reddit:个性预测的金矿，在:Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018, pp.87-97。
[45] M.Gjurković, M. Karan, I. Vukojević, M. Bošnjak, J. Šnajder, PANDORA Talks:Reddit上的个性和人口统计学，in:第九届社会自然语言处理国际研讨会论文集 媒体，计算语言学协会，2021年，第138-152页。
[46] H.Vu, S. Abdurahman, S. Bhatia, L. Ungar, Predicting responses to psychological questionnaires from participants' social media posts and question text embeddings, in:Computational Lin-guistics协会的研究结果。EMNLP 2020, 计算语言学协会, 2020, pp. 1512-1524, Online.
[47] F.Yang, T. Yang, X. Quan, Q. Su, Learning to answer psychological questionnaire for personality detection, in: 计算语言学协会的研究结果。EMNLP 2021，计算协会 语言学，Punta Cana，多米尼加共和国，2021，第1131-1142页。
[48] P.Novikov, L. Mararitsa, V. Nozdrachev, 推断与传统人格 评估。我们预测的是同一件事吗，2021年。
[49] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations:An overview of interpretability of machine learning, in:2018年IEEE第五届数据科学与高级分析国际会议，DSAA，IEEE，2018，80-89页。
[50] D.Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z.Yang, XAI-Explainable artificial intelligence, Science Robotics 4 (2019).
[51] S.Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E.M. Smith, Y.-L. Boureau, J. Weston, 构建开放域聊天机器人的配方, in:计算语言学协会欧洲分会第16届会议论文集。主卷，协会 计算语言学，2021年，第300-325页，在线。
[52] S.Dhelim, N. Aung, M.A. Bouras, H. Ning, E. Cambria, A survey on personality-aware recommendation systems, Artif.Intell.Rev.（2021年）。
[53] L.A. McFarland, R.E. Ployhart, Social media:A contextual framework to guide research and practice, J. Appl. Psychol.100 (2015) 1653.
[54] A.N. Joinson, 计算机中介交流中的自我披露。 自我意识和视觉匿名性的作用，欧洲。J. Soc. Psychol.31 (2001) 177-192.
[55] W.-B.Chiou, Adolescents' sexual self-isclosure on the internet:Deindividuation and impression management, Adolescence 41 (2006).
[56] M.Duggan, A. Smith, 6%的在线成人是Reddit用户，Vol. 3, Pew Internet & American Life Project, 2013, pp.1-10.
[57] J.J. Denissen, M. Luhmann, J.M. Chung, W. Bleidorn, Transactions between life events and personality traits across the adult lifespan. , J. Personal.Soc. Psychol.116 (2019) 612.
[58] L.Berdychevsky, G. Nimrod, Let's talk about sex:老年人 在线社区中的讨论，J. Leis.Res. 47 (2015) 467-484.
[59] T.D. Letzring, D.C. Funder, The realistic accuracy model, in:The Oxford Handbook of Accurate Personality Judgment, 2019.
[60] E.Pavlick, T. Kwiatkowski, Inherent disagreements in human textual inferences, Trans.Assoc. Comput.Linguist.7 (2019) 677-694.
[61] I.Dagan, B. Dolan, B. Magnini, D. Roth, Recognizing textual entailment:理性，评估和方法，J. Nat.Lang.Eng.4 (2010).
[62] I.Androutsopoulos, P. Malakasiotis, A survey of paraphrasing and textual entailment methods, J. Artificial Intelligence Res. 38 (2010) 135-187.
[63] E.Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, SemEval-2012 task 6: A pilot on semantic textual similarity, in:*SEM 2012。The First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Vol. 2: Proceedings of the Sixth International Workshop on Semantic Evaluation（SemEval 2012）, 2012, pp. 385-393.
[64] P.LoBue, A. Yates, Type of common-sense knowledge needed for recognizing textual entailment, in:计算语言学协会第49届年会论文集。人类语言技术，2011，第329-334页。
[65] E.Cabria, B. Magnini, Decomposing semantic inference, in: Linguistic Issues in Language Technology, Vol. 9, 2014-Perspectives on Semantic Representations for Textual Inference, 2014.
[66] Y.Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans.Pattern Anal.Mach.Intell.35 (2013) 1798-1828.
[67] Y.Goldberg, 自然语言处理的神经网络方法, Synth.Lect.Hum.Lang.Technol.10 (2017) 1-309.
[68] N.Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in:Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th Interna- tional Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 3973-3983。
[69] S.Ruder, M.E. Peters, S. Swayamdipta, T. Wolf, 自然语言处理中的迁移学习, in:2019年计算语言学协会北美分会会议论文集。教程，2019年，第15-18页。
[70] M.Honnibal, I. Montani, S. Van Landeghem, A. Boyd, Spacy: Industrial-strength natural language processing in python, 2020.

[71] A.Komninos, S. Manandhar, Dependency based embeddings for sentence classification tasks, in:2016年计算语言学协会北美分会会议论文集。Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp.1490-1500.
[72] I.B.迈尔斯，M.H.麦考利，A.L.哈默，《类型介绍》。A Descrip- tion of the Theory and Applications of the Myers-Briggs Type Indicator, Consulting Psychologists Press, 1990.
[73] J.H. Steiger, 比较相关矩阵元素的测试, Psychol.Bull.(1980) 245-251.
[74] P.Neculoiu, M. Versteegh, M. Rotaru, Learning text similarity with siamese recurrent networks, in:1st Workshop on Representa- tion Learning for NLP, Association for Computational Linguistics, Berlin, Germany, 2016, pp.148-157.
[75] E.Hoffer, N. Ailon, Deep metric learning using triplet network, in:A. Feragen, M. Pelillo, M. Loog (Eds.), Similarity-Based Pattern Recognition, Springer International Publishing, Cham, 2015, 84-92.



Matej Gjurković是萨格勒布大学电子工程和计算机学院的博士生和研究助理。他曾作为研究工程师参与过几个国家和欧盟项目。他的研究兴趣集中在使用自然语言处理方法来预测和分析基于社交媒体文本的个性。

Iva Vukojević是萨格勒布大学人文和社会科学学院的心理学博士生，也是萨格勒布大学电子工程和计算机学院（FER）的研究助理。她的研究兴趣围绕着探索基于文本的数字足迹形式的人类行为，重点是社交媒体文本的个性分析和预测。

Jan Šnajder于2010年在萨格勒布大学电子工程和计算机学院获得计算机博士学位。从2016年起，他在FER的电子、微电子、计算机和智能系统系担任副教授职务。他的研究兴趣集中在自然语言处理和机器学习方面。他一直是克罗地亚科学基金会资助的几个项目的主要调查员。