arXiv:2302.00493v1 [cs.CL] 1 Feb 2023

你谈论什么，就是什么：为人格分析诱导评价性主题

Josip Juki´c IvaVukojevi´c Jan Šnajder萨格勒布大学电气工程与计算机学院 文本分析与知识工程实验室 乌恩斯卡街3号，10000 萨格勒布，克罗地亚 {josip.jukic, iva.vukojevic, jan.snajder}@fer.hr

摘要

对实体和概念表达态度或立场，是人类行为与人格不可或缺的一部分。近年来，随着社交媒体迅速发展，评价性语言数据变得更容易获取，从而推动了大规模意见分析。然而，令人意外的是，研究人格与评价性语言之间关系的工作寥寥无几。为弥合这一空白，我们提出评价性主题这一概念：通过对社交媒体中经过预筛选的评价性文本应用主题模型，获得评价性主题。随后，我们将评价性主题与各个文本作者关联起来，构建其评价画像。我们将评价画像应用于带有人格得分标签的 Reddit 评论，并针对评价性主题与大五人格维度之间的关系开展探索性研究，旨在进行更具可解释性的侧面层级分析。最后，我们通过观察到与人格心理学既有研究一致的相关性，验证了我们的方法。

1 引言

分享观点一直是人们日常习惯的一部分，但如今，社交媒体上大量涌现的用户生成文本使这种行为的规模不断扩大（Lee 和 Ma，2012年）。观点是人们对特定实体的倾向，对于理解人类行为至关重要。人们往往需要预测其对特定目标的情感倾向或立场，因此，观点分析在市场营销研究、社会科学与政治学以及推荐系统等领域都很有用。此外，分析我们表达观点的方式，即意见持有者的语言层面与心理倾向，对于人格分析和心理健康预测等各种下游任务也至关重要。

在公共交流等情境中，人们会使用称为评价性语言（Hunston，2010年）的语言资源来表达他们的

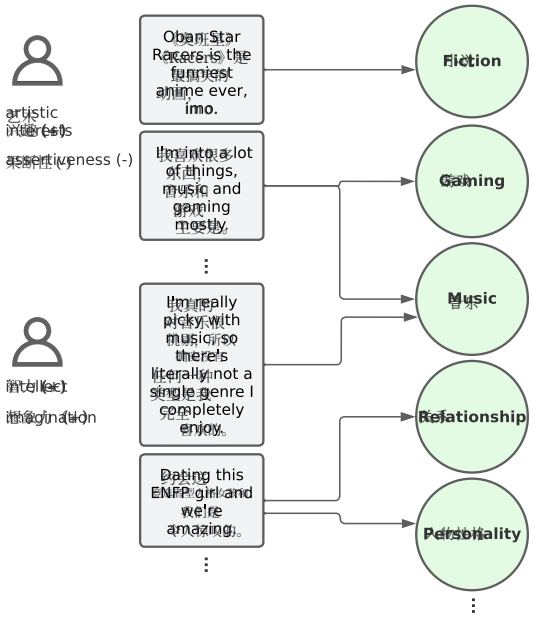

图 1：将用户陈述映射到评价性主题的示例，即从评价性文本中归纳出的主题。根据这些主题的普遍程度，我们为每位作者构建评价画像。随后，我们衡量每位作者的画像与人格侧面（如艺术兴趣、果断性、智力和想象力）之间的相关性。

态度。 Thompson 和 Hunston (2000年) 将评价定义为确定性和良好性／可取性等统一现象的统称；评价性语言则包括议长或作者对实体或命题所持态度或立场，以及对它们的情感的所有表达。由于评价与心理倾向相关（Mal-rieu，1999年），评价性语言有助于揭示人的区别性特质，即反复出现的行为、认知或情感倾向。这些源自与环境互动的行为差异属于人格心理学的研究范畴，在该领域中，这些差异通常被视为人格特质——行为、认知和情感方面具有代表性且随时间推移

保持稳定、并能跨情境持续存在的模式（Funder，2012年）。此外，研究表明，人格特质与职业偏好、政治或宗教倾向、智力、个人兴趣和观点相关（Larson 等人，2002年；Ackerman 和 Heggestad，1997；Hyman，1957年）。

据我们所知，社交媒体上的人格分析尚未研究过评价性语言。这或许可以归因于方法上的挑战。分析评价性语言的难点之一，是将信号与噪声区分开来。尽管文本中的评价无处不在，但自动区分评价性文本与非评价性文本仍是一项挑战。评价性表达形式多样，显性词汇立场标记的使用程度也各不相同。另一个难点是为人格和评价性语言选择恰当的抽象层级。如果评价对象过于具体，就很难将其与人格联系起来。不过，我们可以根据人格涵盖范围的宽窄，选择恰当的人格操作化方式，从而使其与评价对象的抽象层级相匹配。

本文研究如何利用社交媒体上的评价性语言进行人格分析。更具体地说，我们旨在探究人们谈论的主题与其人格之间的关系，如图1所示。为此，我们提出评价性主题这一概念，即由评价性文本构建的主题。为解决检测评价性文本时的信噪问题，我们开发了一种基于检测评价性表达和释义挖掘的迭代过滤技术。随后，我们利用主题模型，从预过滤后的评价性文本中提取评价性主题。为尽可能提高连贯性和多样性，我们测试了多种主题模型，包括概率模型、分解模型和神经主题模型。最后，我们将评价性主题与各个文本作者关联起来，以构建作者的评价画像。

在实验部分，我们探究评价与人格之间的关系。我们的研究聚焦于大五人格模型（Gold-berg，1981年），该模型涵盖了宽泛的人格特质，以及更具体（也可以说更容易解读）的方面，即所谓的人格侧面。我们针对 Reddit 评论开展了一项评价性主题与大五人格侧面的相关性研究。Reddit 是最受欢迎的讨论网站之一，涵盖极其广泛多样的主题，同时保护用户匿名。我们发现，评价与人格之间的相关性与人格心理学的既有研究结果一致。最后，我们提供了实证证据，表明评价性陈述与人格的关联强于非评价性表达。

总而言之，我们的贡献有三点：(1) 提出一种基于主题模型并结合评价性过滤的评价性作者画像方法；(2) 将评价性画像应用于 Reddit 评论；(3) 研究 Reddit 用户的评价性作者画像与大五人格侧面之间的相关性。

本文其余部分的结构如下。第2 节介绍背景与相关工作。第3节将介绍我们结合主题建模过滤评价性文本的方法。本节最后将阐述评价性作者画像这一概念。第4节介绍我们的实验及结果。最后，第5 节总结本文的主要结论与要点。

2 背景与相关工作

评价性语言。在语言学中，对评价性语言的研究植根于功能主义语言研究方法。这类方法旨在从语言的语义功能和交际功能角度理解并解释语言结构，并认为语言的首要功能是作为社会互动的媒介（Allan，2007年）。评价性语言研究大多仅涉及评价的某个特定领域，例如主观性（Wiebe等人，2004年）和立场（Kiesling等人，2018年）。例如，Biber和 Finegan（1989）描述了立场的词汇和语法标记，即沿着认识意义和态度意义连续体进行分类，并据此构建了表示人们态度、情感和评价的标记类别。近年来，Pavalanathan等人（2017年）基于 Reddit 数据编纂了立场标记词典。另一方面，由于情感分析在自然语言处理中无处不在，评价性语言也常通过情感倾向的视角来研究（Be- namara等人，2017年；Pang和Lee，2008）。在我们的研究中，我们尝试结合评价性语言的多个方面，即情感倾向、立场和意见标记。

目标意识。 由于表达观点的本质，评价总是针对特定目标。众多评价方法中，基于情感倾向的方法一直颇为普遍，这

促成了基于方面的情感分析（ABSA）的发展。这是一项结合方面抽取与情感分类的自然语言处理任务（Pon- tiki 等人，2014年）。近期，ABSA 研究已从传统视角发展而来，例如将方面抽取视为序列标注任务，并使用条件随机场（Toh 和 Wang，2014年），逐渐转向神经网络方法（He 等人，2017年；Luo 等人，2019年；Hoang 等人，2019年；Tulkens 和 van Cranenburgh，2020）。不过，ABSA 系统更适用于较小的领域。除 ABSA 外，立场分类（Du 等人，2017年）和主题相关的论证分类（Reimers 等人，2019年）等任务中也采用了基于目标的方法。由于我们尝试涵盖多种领域，因此选择采用主题级方法来处理目标。

自然语言处理与人格。自然语言处理已被广泛用于人格分析，最初应用于文章分析（Pennebaker 和 King，2000年），之后又用于社交网络研究（Park 等人，2014年；Schwartz 等人，2013年）。较新的研究主要基于 Facebook 数据开展。例如，Kulkarni 等人（2018年）针对 Facebook 上的大五人格体系进行了人格特质层面的因子分析。然而，据我们所知，尚无研究利用自然语言处理分析评价性语言情境中的人格。此外，自然语言处理领域鲜有研究涉及侧面层面人格分析。

3 方法论

我们提出了一个从原始文本到评价画像的三步流程。首先，我们进行评价性筛选，提取评价性文本，然后从筛选后的数据集中构建主题。最后，我们利用构建出的评价性主题创建相应的评价画像。

3.1 数据集

我们使用了 PANDORA数据集，其中包含从自我报告人格问卷结果中提取的 Reddit 用户人格得分（Gjurkovi´等人，2021年）。共有 1,608 名用户自我报告了大五人格得分。这些用户总共撰写了 130 万条评论，包含 1,430 万个句子。此外，有 127 名用户自我报告了 NEO PI-R 人格侧面的问卷得分。[1] NEO PI-R 提供了有关每种大五人格特质的六个人格侧面的信息。人格侧面是较广泛的人格特质中一个具体且独特的方面。例如，焦虑和抑郁是神经质的人格侧面，而友善和合群性则是外向性的人格侧面。完成预处理（包括将评论切分成句子）后，我们最终得到 650 万个句子。由于我们使用的是基于英语文本预训练的模型，因此过滤掉了非英语评论。有关预处理步骤的更多详情，请参见附录 A.1。

3.2 评价性过滤

尽管社交媒体上充斥着评价性语言，但要自动区分评价性文本与非评价性文本颇具挑战。为过滤掉非评价性表达，我们收集了包含评价性标记的评论——这些词汇或语法形式用于表达评价行为。我们结合现成的情感分析工具以及观点和立场词典，以涵盖评价性语言的不同方面。此外，我们还搜索了包含评价性模式的句子，即表达观点的短语。

为提取带有情感倾向的句子，我们使用了 VADER（Hutto 和 Gilbert，2014年），将 VADER 正负分数相加，以确定整体情感分数。我们使用了Hu 和 Liu（2004年）编制的观点词典，以及Pavalanathan等人（2017年）编制的立场标记词典。受Hunston 和 Thompson（2000年）启发，我们整理了一系列评价性模式，并将其编写为正则表达式，在整个数据集中进行匹配。这些模式包括诸如我喜欢/憎恨/相信/支持、就我个人而言、在我（诚实的）看来等短语。除了上述短语的各种变体外，我们还扩展了正则表达式，使其能够处理否定词和修饰语，以及在线社区中使用的口语表达（例如，在我看来——在我看来、*FMPOV*——从我的角度来看）。为收集匹配到的句子，我们设定了一项限制：每个匹配结果都必须超过情感、观点和立场分数的阈值。我们取了各类别中位于第50百分位数的匹配句子的交集，但情感类别除外；对于情感类别，我们设定了一个上限。这是因为我们发现，情感分数极高的句子通常没有明确的目标，而且若无额外语境，无法推断出目标。[2]

[1] https://ipip.ori.org/newNEO_FacetsTable.htm

[2] 我们决定不采用情感倾向高于第 90 百分位数的句子，因为这类句子通常以高度隐含的目标表达情绪（例如，一个句子中

对 PANDORA应用评价性过滤后，我们获得了 29k 个句子。由于采用了严格的模式匹配并要求较高的评价分数，我们侧重于精确率。为提高召回率，我们开发了准雪球抽样（QSB），这是一种简单的释义句挖掘技术。[3] QSB 是一种迭代过程，从一组评价性表达的种子集开始，并通过添加相似陈述来扩展种子集。我们使用句子转换器[4]（Reimers 和 Gurevych，2019年）计算上下文化表示，并利用这些表示检测释义句。我们以经过过滤的评价性句子作为种子集，初始化 QSB。随后，我们使用余弦相似度作为标准，根据相似度阈值 t_{sim}提取相似句子；该阈值会在每一步根据相似度增长因子 γ进行调整。受模拟退火启发，该因子γ 会使阈值呈指数增长，从而随着每次迭代中句子集不断扩大，实现更快收敛。每次迭代结束时，我们会将新匹配到的句子加入原有种子集。没有更多候选句子时，过程便会停止。使用 QSB，我们获得了 31 万个带有评价性标记的句子（详情见附录A.2）。

QSB 挖掘的表达式可能在目标、极性或强度上存在差异。由于我们采用了较高的相似度阈值，因此在大多数情况下，释义匹配中上述三个组成部分中只有一个会有所不同。多轮释义挖掘可能会逐步改变原始句子，从而总体上产生更宽松的匹配结果，如表1所示。

3.3 评价性主题

在第二步中，我们对预先筛选出的评价性评论应用主题建模，以获得评价性主题。我们采用了将主题定义为固定词汇表中各词项上的分布的定义（Blei和Lafferty，2009年），并进一步将评价性主题定义为由评价性语言衍生出的特定目标观点所构建的主题。为了生成评价性主题，我们试验了传统概率模型以及基于神经网络的架构。此外，由于我们处理的是 Reddit 评论，而这类评论通常较短，因此我们考虑了几种适用于短文本的主题建模方法。

潜在狄利克雷分配（LDA）。在传统主题模型中，我们选择了潜在狄利克雷分配（LDA），这是一种成熟的概率主题模型（Blei等人，2003年），也可作为有力的基线。考虑到潜在狄利克雷分配因数据稀疏问题而难以处理短文本（Hong和Davison，2010年），我们遵循Zuo等人（2016年）的做法，将评论归并成较大的文档。我们按子版块将每位作者的评论归并为伪文档；子版块即 Reddit 上专门讨论某个特定主题的论坛。

BTM。 我们测试了双词主题模型（BTM）。该模型主要针对短文本设计（Cheng 等人，2014年）。我们向模型输入评论级文本，并从评论中筛除了非评价性句子。

ABAE。 我们尝试了基于注意力的方面提取（ABAE）这一基于神经网络的架构；它是 He 等人提出的自编码器（2017年）。我们纳入这一模型，是因为它专门用于方面聚类，我们预计这与构建评价性主题的目标十分契合。ABAE 利用词嵌入来提升连贯性，并尝试将出现在相似语境中的词归为一组。此外，ABAE 使用注意力机制，在训练过程中降低无关词的相对重要性。我们采用了 ABAE，但对训练过程 做了一些修改。首先，我们在 PANDORA 上训练了定制的 Word2Vec 词嵌入（Mikolov 等人，2013年）。我们还将重构损失函数从普通的点积修改为与 $[-1, 1]$ 区间内的余弦相似度相匹配，这提高了学习的稳定性。我们向模型输入分词后的评价性句子，以学习这些主题。

CTM. 最后，我们试用了（Bianchi 等人，2021年）提出的组合主题模型（CTM）。研究表明，与其他主题模型相比，该模型能提高连贯性。CTM 将变分自编码器与句子变换器生成的嵌入相结合。[5]

主题模型通常通过主题一致性进行评价。然而，这种方法存在验证差距，即自动化的一致性评价未经人类实验验证（Hoyle 等人，2018年）。为缓解这一问题，我们统计

许多表情符号）。[3]准雪球抽样的代码可通过以下链接获取：https://github.com/josipjukic/quasi-snowballing [4]我们选择了*all-mpnet-base-v2*，来源于 https://www.sbert.net/

[5]我们使用 *paraphrase-distilroberta-base-v2* 生成上下文嵌入。

原始命中	释义匹配	相似度
在我看来，年龄是绝不可谎报的**在我看来**。	我不喜欢谎报年龄。	.86
我支持死刑，但我绝不会给.....贴上标签 我自己是反堕胎派。	我是反堕胎派，因为我宁愿不要让 人们在胎儿已经可以被合理地 视为一个人的孕晚期堕胎.....	.77
这件事让我很纠结，因为**我支持**跨性别者 百分之百，而且**我相信**跨性别女性就是女性， 故事到此为止，男性也一样。	我支持跨性别者进行性别转换 并在法律上被视为异性 sex.	.78

表1：前　子匹配的 QSB 释义匹配示例。第一列表示被匹配的评论　to
正则表达式 会话中匹配文本以粗体显示。第二列列出匹配的释义句。

	NPMI	IRBO	主题数
LDA	−.1413	.9905	20
BTM	−.2159	.8241	30
ABAE	−.0521	.9833	20
CTM	**.0628**	.9981	20

表 2：主题建模评价结果。NPMI 是取值范围为 $[-1, 1]$ 的一致性分数，其中 -1 表示主题代表词共同出现，0 表示各词的出现彼此独立，1 表示各词完全共同出现。IRBO 使用 RBO 衡量主题多样性，其中 1 表示主题完全不同，0 表示主题完全相同。我们使用不同的随机种子重复每项实验 10 次。粗体显示的分数显著优于其余模型的分数（对每对模型分别进行独立的 Wilcoxon 检验，$p < .01$，并使用 Holm-Bonferroni方法调整家族错误率）。

整个数据集中的词元共现，而不仅仅是训练数据中的词元共现；研究表明，前者与人类判断的关联更强（Ding 等人，2018年）。具体而言，我们使用归一化点互信息（NPMI）评价所诱导主题的一致性。为评价多样性，我们采用了Bianchi 等人（2021年）提出的指标，该指标定义为标准 RBO 的倒数（Terragni 等人，2021年）。[6] 我们将 NPMI 和 IRBO 作为多目标优化过程的两项准则。我们确定了训练模型的帕累托前沿，即非支配模型的集合，并从中选择主题数量最少的模型（表 2）。

3.4 评价性作者画像

在第三步中，我们利用评价性主题来生成评价性作者画像。具体而言，我们根据每个用户所涉及主题的普遍程度来描述用户，其中用户文本被表示为针对不同目标的、按情感倾向加权的主题混合—

[6]RBO 使用加权排名估算每个主题前 10 个词之间的差异程度。

合。这样，每个用户都会被分配一个平均主题普遍程度。我们计算整个用户句子集合的平均主题分布，其中每个给定句子都会针对每个主题贡献一个来自$[0, 1]$ 区间的值。

为形式化描述我们的流程，我们首先为特定文档$\mathbf{d} = [c_1 c_2 \dots c_K]^\mathsf{T}$定义主题分布，其中 $\mathbf{d}$ 表示一个文档（例如，句子、评论），而 c_k 是第 k 个主题对应的混合成分。我们汇总用户评价性句子的值，以计算该用户在每个主题上的第 n 个成分，并将汇总结果串接成向量$\mathbf{u}^{(n)}$:

$$\mathbf{u}^{(n)} = \frac{1}{N_n} \sum_{i=1}^{N_n} \mathbf{d}^{(n,i)},$$

其中， N_n 是第 n 位用户的文档数量，$\mathbf{d}^{(n,i)}$ 是第 n 位用户的第 i 篇文档。此外，我们纳入情感强度信息，以获得情感增强表示 $\mathbf{v}^{(n)}$:

$$\mathbf{v}^{(n)} = \frac{1}{N_n} \sum_{i=1}^{N_n} s^{(n,i)} \mathbf{d}^{(n,i)},$$

其中， $s^{(n,i)}$ 是根据 VADER 正负分数之和计算得出的第 n 位用户第 i 篇文档的情感强度。

用情感强度代替极性需要进一步说明。在考察情感极性时，我们可以区分两种类型：用户对某个陈述的情感极性，以及主题本身的隐含极性（Russo 等人，2015年）。由此产生了多种将情感信息纳入评价表征的方式。然而，纳入情感极性会使人难以

区分主题的情感究竟主要由隐含极性还是显性极性驱动。我们追求简约，主要是为了便于解释，因此选择仅使用情感强度。未来的研究可以探讨如何结合这两种类型的情感。

4 项实验

我们对评价性语言与人格之间关联的研究分两步进行。首先，我们根据 PANDORA 的 Reddit 评论构建评价性作者画像，具体方法如上一节所述。为了构建评价性主题，我们在筛选后的评价性句子上使用了 CTM 模型，因为 CTM 在连贯性和多样性得分方面表现最佳。根据优化过程的结果，我们推导出 20 个评价性主题。表 3 展示了部分主题及其人工赋予的标签。

接下来，我们对大五人格侧面与评价性主题之间的关系进行了相关性研究。我们采用修订版NEO人格量表（NEO PI-R），该量表旨在测量大五人格特质：经验开放性、尽责性、外向性、宜人性和神经质（CostaJr andMcCrae，1995）。在分析的第一部分，我们考察了特定主题与人格侧面之间的单独相关性。在第二部分，我们使用典型相关分析（CCA），探究一组主题整体与所有大五人格侧面之间的共同关联。我们选择关注人格侧面而非人格特质，并假设人格侧面的抽象层次与推导出的主题的细粒度更为契合。从这个意义上说，选择人格侧面符合布伦斯维克对称原则。该原则指出，所分析的构念（在本研究中为人格和主题）必须具有相近的概括性层次（Wittmann，2012年）。

4.1 成对相关

我们计算了评价性作者画像与大五人格侧面之间的偏两两相关，并控制性别这一可能的混淆变量。对于性别，我们使用了 PAN-DORA 数据集所提供的值，该数据集的得分超过了 F_1 分数。90 （Gjurkovi´c 等，2021年）。[7] 图 2显示了经假发现率校正的显著相关性-著相关性采用 Benjamini-Hochberg 方法进行了校正。我们观察到许多较小的相关性，甚至还有中等程度的相关性（$> .4$），存在于某些主题—人格侧面配对之间（Bosco 等，2015年）。鉴于基于问卷的个体差异研究中，平均相关性为 。19（Gignac 和 Szodorai，2016年），且文本数据比问卷数据噪声更大，这一结果令人意外。

为评估我们结果的有效性，我们查阅了人格心理学文献作为参考。根据心理学研究，经验开放性的各个人格侧面预计会与好奇心、食物/饮料、虚构作品、音乐、游戏、政治、人格、辩论和神秘学主题呈正相关，而与宗教主题呈负相关（Stewart 等，2022年；Ozer 和 Benet-Martinez，2006年；Soto，2019年；Intiful 等，2019年；Skimina 等，2021年；Church 等，2008年；Marshall 等，2015年；Chauvin 和 Mullet，2021年）。如图2所示，我们的大多数结果与上述既有研究一致。我们未能得到显著支持的三项关联是：与辩论呈正相关（假设方向一致，但不显著）、与游戏呈正相关（假设方向一致，但不显著），以及与神秘学呈负相关（符号相反）。不过，我们注意到，在所引用的大多数心理学文献中，研究是在人格特质层面而非人格侧面层面分析效应的。

4.2 典型相关分析

由于人格侧面本质上相互交织，我们可以通过结合这些评价主题来分析人格侧面，从而获得更深入的见解。为此，我们采用了典型相关分析（典型相关分析）。典型相关分析旨在分别对一组评价画像和一组人格侧面进行线性组合，从而使新构造的典型变量之间的相关性最大化。典型相关分析假设这两组变量均服从多元正态分布，我们通过Henze-Zirkler检验验证了这一假设（Henze和Zirkler，1990年）。[8] 我们计算了20 个典型维度，并按相关程度从高到低排序

[7] Gjurkovi´等人（2021年）的性别预测模型是一个仅在一部分用户上训练的二元分类器

自我报告属于两种二元性别类别之一。因此，对于非二元性别用户，其预测结果将不正确。不过，我们认为，这一点以及即便对于二元性别个体，预测也并非总是正确这一事实，对剔除性别影响后的相关估计值影响有限。[8]在这两种情况下，正态性均未被拒绝，其中 $p = .62$ $p = .51$ 分别对应评价变量和人格变量。

标签	代表词
open	人，很多，免费，好，许可，时间，去，开放，大，尝试
食物/饮料	味道、肉、气味、啤酒、质地、风味、咸鲜味、饮料、味觉、甜
宗教	证词、相信、真理、上帝、见证、原始、圣经、教会、耶稣、地球
贬低	屎、告诉、说话、陌生人、粗鲁、愚蠢、贱人、难过、奇怪、糟糕
投资/金融	推荐、包装、担保、选择、购买、投资、基金、票、股票、处理
虚构作品	观看、角色、阅读、电影、故事、书籍、想法、酷、新、动画
音乐	歌曲、专辑、流行音乐、最爱、音乐、声音、聆听、歌词、曲目、声音
游戏	玩、游戏、关卡、团队、伤害、技能、攻击、装备、战斗、地下城
社会问题	社会、人类、宗教、文化、权利、道德、白人、争论、受害者心态、黑人
仇恨	鄙视、憎恨、激情、立场、他妈的、该死的、书呆子、愤怒、自鸣得意、广告
日常生活	日子、工作、小時、一周、花費、食物、錢、睡眠、家、吃
人际关系	人际关系、女性、约会、孩子、伴侣、男性、性、前任、同性恋、赡养费
争论	观点、盐、颗粒、伪善的、毫无价值的、有争议的、评估、免责声明、180、细致入微的
政治	政府、政党、投票、特朗普、支持、国家、州、议长、系统、税
性与外貌	性感、头发、性感、令人厌恶、面部的、身体、有吸引力的、穿着、色情作品、两性异形
人格	类型、人格、迈尔斯-布里格斯类型指标、情绪、测试、认知的、学习、想法、大脑、荣格
辩论	理解，讨论，帖子，言论，人物，错误，思考，要点，互联网，辩论
虐待	鄙视，憎恨，学校，辍学，糟糕，骚扰，拒绝，逃课，街区，遭到殴打
神秘学	命运，风铃，猫，灵体，引导者，运气，拼命地，护身符，邀请，镜子

表3：CTM 生成的评价性主题。第一列显示的主题标签由五位标注员手动分配，取其中出现频率最高的标签。第二列显示每个主题排名前10的词语。

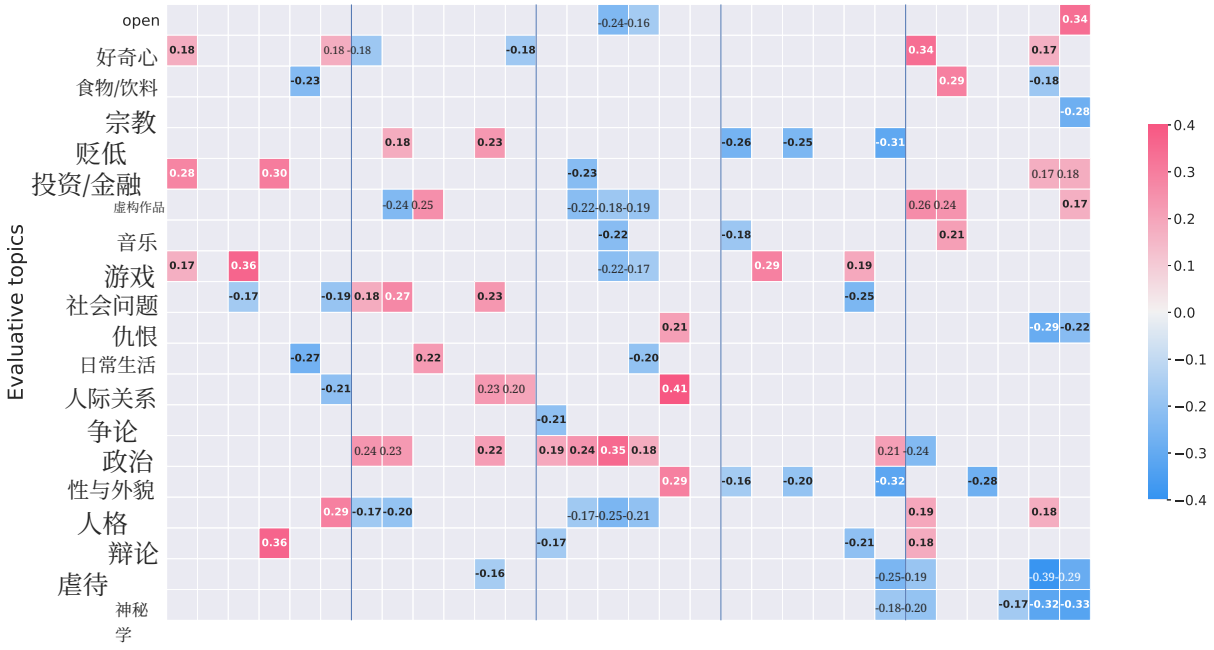

图2：控制性别后，大五人格侧面（x轴）与评价性主题（y轴）之间的两两偏相关。我们仅展示显著相关（$p < .01$），显著性通过对相关系数进行 Fisher 的 z 变换并采用 Benjamini-Hochberg 方法校正假发现率来确定。

。[9] 通过Wilks' lambda测试，我们发现前3对典型变量之间的相关性具有统计显著性。[10]

我们还想回答的一个问题是：与普通主题（由全部文本构建）相比，人格分析是否能从评价性主题（由预先过滤的评价性文本构建）中获益。为此，我们对三个不同的数据子集进行了典型相关分析：(1) 未过滤文本；(2) 通过评价性过滤获得的文本；以及 (3) 非

[9]一般而言，典型维度的数量等于变量较少的集合中的变量数量。[10]我们在以下范围内观察到显著相关（$p < .01$）：1–20、2–20 和 3–20。

评价性文本，即第一组与第二组文本之差。此外，我们还对 (4) 包含情感强度信息的评价性文本进行了典型相关分析，以考察评价性画像是否能从情感信息中获益。图3展示了四个数据子集上前三个典型变量对的典型相关。典型相关分析计算出的相关性越高，表明两组变量之间的关联越强。因此，在假设评价性画像与人格相关联的前提下，较高的相关性支持了依据我们模型构建的评价性画像的构念效度。在此处，构念效度涉及评价性主题准确评估其设计目标的程度。基于此，结果呈现出两个主要观察：(1) 与使用未筛选数据相比，评价性预过滤似乎更适合建立与大五人格的关联；(2) 情感信息会增强相关性。

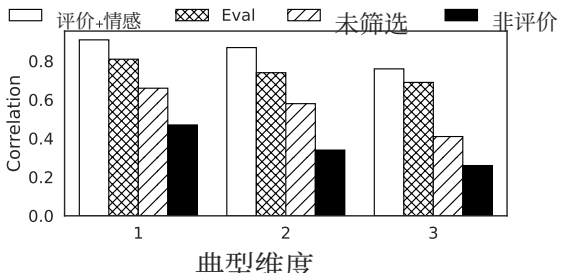

图 3：在四个数据子集上估计的前三对典型变量对（$p < .01$）的典型相关：原始文本（未筛选）、经过评价性过滤的文本（评价）、在计算评价画像时纳入情感信息的评价文本（评价+情感），以及通过从未过滤集合中减去评价集合而获得的非评价性文本（非评价）。

为进一步探究人格侧面与评价性主题之间的关系及其各自的重要性，我们将两组原始变量投影到第一和第二典型维度上，计算了典型载荷，即典型结构相关（图4）。典型载荷反映观测变量与典型变量之间的共享方差。我们遵循心理测量学中的常见做法（terBraak, 1990年），计算工具变量（评价性主题）的组内变量与典型变量相关，以及目标变量（人格侧面）的组间相关。我们识别出三个明显的主题与人格侧面聚类。第一组位于图4的右下角（音乐、虚构作品；艺术兴趣、想象力），表明经验开放性中的开放性方面。左下方的聚类（虐待、贬低；脆弱性、放纵和寻求刺激）大致对应于情绪上令人不快的聚类。最后，左上方的聚类（政治、社会问题；活动水平、果断性）可解释为社会参与。我们还通过展示人格侧面在第一和第二典型维度上的离散分布（附录中的图5和6）进一步验证典型相关分析的结果。我们发现，同一领域的人格侧面会聚集在一起，但不同领域的人格侧面之间也存在关联

不同领域，这与先前的研究结果一致（Schwaba 等人，2020）。

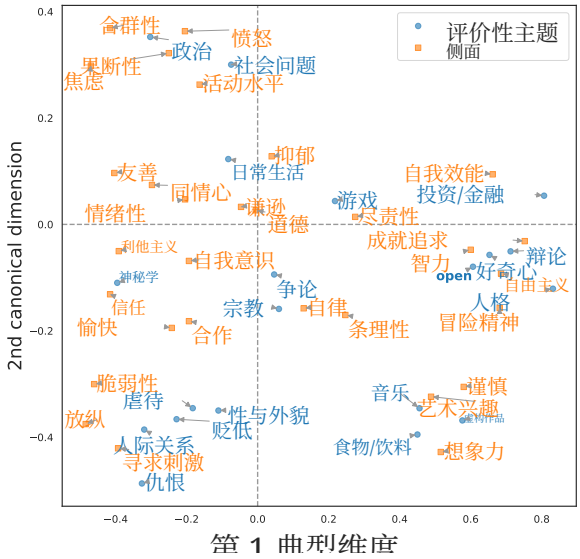

图 4：人格侧面和评价性主题前两个维度的典型载荷。各个要点之间的接近程度表示两个数据点之间关联的强弱。具体而言，当两个要点（无论其类型是主题还是人格侧面）相对于原点的角度相似且大小相近时，这表明它们对典型变量的贡献相似，且彼此关联更强。

5 结论

在自然语言处理研究中，评价性语言与人格之间的关系尚未得到充分研究。我们旨在通过提出评价性作者画像来填补这一空白——将作者文本与从经过评价性语言筛选的文本中提取的主题关联起来。我们在一个包含 Reddit 评论及自我报告的大五人格侧面得分的数据集上应用了评价性画像。通过典型相关分析，我们发现同一人格特质中的各个人格侧面在典范空间中的关联更强。

我们发现，评价性主题与大五人格维度呈中等程度的相关。此外，我们证实了这一假设：与未筛选文本相比，评价性表达对人格分析具有更高的信息价值。另外，我们发现，与评价性文本相比，非评价性文本与大五人格维度的关联弱得多。最后，我们观察到的相关性与人格心理学领域以往的研究结果一致。我们相信，本研究有助于更好地理解社交媒体上的评价性语言及其与人格特质的关系。

伦理考量

我们的研究已获学术机构审查委员会批准。

潜在危害。 我们的研究属于探索性研究，因此不能根据所得关联进行概括。否则可能导致缺乏依据的概括，从而歧视某些人群。

收集用户数据。 根据美国心理学会的伦理原则，[11] 研究人员在分析档案数据（即研究开始前收集的数据）时，可以免除知情同意；前提是，披露受试者的回答不会使其面临刑事或民事责任风险，也不会对其造成伤害。在我们的研究中，我们使用 Reddit 数据。Reddit 用户通过 Reddit 用户协议同意不向他人披露任何敏感信息，并且已获知其评论公开可见。由于用户可以选择退出并删除自己的数据，我们删除了已注销的用户账户及其评论。此外，为了进一步保护用户身份，我们以群体层面而非个人层面呈现研究结果。

潜在滥用。 原则上，评价画像可能被用于基于社交媒体文本数据在个人层面作出决策，例如用于微定向。我们强烈反对将我们的方法用于此类或其他在伦理上有问题的应用。

偏差。 我们注意到，我们使用的是 Reddit 上的自我报告数据。因此，这些数据可能并非完全准确，也可能包含各种偏差，尤其是获取偏差。我们使用的数据集可能无法代表 Reddit 用户群体，而且它当然也无法代表任何更广泛的用户群体。

局限性

尽管我们的评价性过滤技术筛选出的评价性主题与大五人格维度的相关性强于非评价性文本，但若没有额外标注，便无法单独验证过滤效果。此外，纳入情感极性等额外信息可能会损害结果的可解释性。规避这一问题的一种可能方案是使用结构主题模型（Roberts 等人，2014年），该模型可以将用户人口统计信息等额外信息作为协变量纳入建模。由于我们使用主题建模，并试图优化主题多样性和连贯性，而这一做法近来已受到批评（Hoyle 等人，2021年），因此很难为主题选择合适的代表词，而且无法保证某个主题足够连贯，能够转化为有意义的概念。从心理学的角度来看，我们仅开展了一项探索性研究，观察到的是相关性，因此无法确认人格心理学中的任何结论，只能通过关联为其提供支持。此外，由于我们进行的是维度层面研究，并且只采用同一问卷中的分数，以尽量减少不同测试带来的噪声，因此报告人格侧面分数的样本数量相对较少（$n = 127$），相比之下，报告人格特质分数的用户总数为（$n = 1,608$）。

我们的研究仅限于英语文本。若要迁移至其他语言，则需要额外的资源，即足够的词典和情感分类器，以实现评价过滤，这使得迁移并非易事。最后，我们还需要使用目标语言的文本重新训练主题模型。

致谢

感谢审稿人的评论。我们还要感谢 Irina Masnikosa 提供的宝贵语言建议。本研究完全由克罗地亚科学基金会通过项目 IP-2020-02-8671 PSYTXT（“基于文本的人格预测与分析的计算模型”）资助。

[11] https://www.apa.org/ethics/code

参考文献

Phillip L Ackerman 和 Eric D Heggestad。1997。智力、人格与兴趣：重叠人格特质的证据。《心理学通报》，121(2):219。

Keith Allan. 2007年。 语言学中的西方古典传统。EquinoxLondon。

Farah Benamara、MaiteTaboada 和 Yannick Mathieu。2017年。 超越词袋的评价性语言：语言学见解与计算应用。计算语言学，43(1):201–264。

Federico Bianchi、Silvia Terragni 和 Dirk Hovy。2021年。 预训练是一个性感的主题：上下文化文档嵌入可提升主题一致性。载于第*59*届计算语言学协会年会暨第*11*届国际自然语言处理联合会议论文集（第*2*卷：短篇论文），第759–766页，线上。计算语言学协会。

Douglas Biber 和 Edward Finegan。1989。 英语中的立场风格：证据性与情感的词汇和语法标记。文本——跨学科话语研究期刊，9(1):93–124。

David M Blei 和 John D Lafferty。2009年。主题模型。载于 文本挖掘，第101–124页。Chapman and Hall/CRC。David M. Blei、Andrew Y. Ng 和 Michael I. Jordan。2003年。潜在狄利克雷分配。《机器学习研究杂志》，3:993–1022。Frank Bosco、Kulraj Singh、James Field 和 Charles Pierce。2015年。 相关效应量基准。《应用心理学杂志》，100:431–449。Bruno Chauvin 和 Etienne Mullet。2021年。超自然信念中的个体差异：人格方面的差异作用。 《当代心理学》，40(3):1218–1227。Xueqi Cheng、Xiaohui Yan、Yanyan Lan 和 Jiafeng Guo。2014年。 BTM：短文本主题建模。*IEEE*知识与数据工程汇刊，26(12):2928–2941。A Timothy Church、Marcia S Katigbak、Jose Alberto S Reyes、Maria Guadalupe C Salanga、Lilia A Miramontes 和 Nerissa B Adams。2008年。跨文化日常行为的预测与跨情境一致性：对人格特质和文化心理学观点的检验。《人格研究杂志》，42(5):1199–1215。Paul T Costa Jr 和 Robert R McCrae。1995年。领域与侧面：使用修订版NEO人格量表进行层级人格评估。 《人格评估杂志》，64(1):21–50。

Ran Ding、Ramesh Nallapati 和 Bing Xiang。2018年。连贯性感知神经主题建模。载于 *2018*年自然语言处理实证方法会议论文集，第830–836页，比利时布鲁塞尔。计算语言学协会。

Jiachen Du、Ruifeng Xu、Yulan He 和 Lin Gui。2017年。基于目标特定神经注意力的立场分类。载于第二十六届国际人工智能联合会议论文集，*IJCAI-17*，第3988–3994页。

David C Funder。2012年。准确的人格判断。《当代心理科学》，21(3):177–182。

Gilles E Gignac 和 Eva T Szodorai。2016年。个体差异研究者的效应量指南。《人格与个体差异》，102:74–78。

Matej Gjurkovi´c、Mladen Karan、Iva Vukojevi´c、Mihaela Bošnjak 和 Jan Šnajder。2021年。PANDORA talks：Reddit上的人格与人口统计信息。载于第九届社交媒体自然语言处理国际研讨会论文集，第138–152页，线上。计算语言学协会。

Lewis R Goldberg。1981年。语言与个体差异：在人格词汇表中寻找普遍性。人格与社会心理学评论，2(1):141–165。

Ruidan He、Wee Sun Lee、Hwee Tou Ng 和 Daniel Dahlmeier。2017年。用于方面抽取的无监督神经注意力模型。载于计算语言学协会第*55*届年会论文集（第*1*卷：长篇论文），第388–397页，加拿大温哥华。计算语言学协会。Norbert Henze 和 Bernd Zirkler。1990年。一类用于多元正态性的不变一致检验。《统计理论与方法通讯》，19：3595–3617。Mickel Hoang、Oskar Alija Bihorac 和 Jacobo Rouces。2019年。使用 BERT 进行基于方面的情感分析。载于第*22*届北欧计算语言学会议论文集，第187–196页，芬兰图尔库。林雪平大学电子出版社。Liangjie Hong 和 Brian D Davison。2010年。Twitter 上主题建模的实证研究。载于首届社交媒体分析研讨会论文集，第80–88页。Alexander Hoyle、Pranav Goel、Andrew Hian-Cheong、Denis Peskov、Jordan Boyd-Graber 和 Philip Resnik。2021年。自动化主题模型评价是否失效？连贯性的不连贯性。载于《神经信息处理系统进展》，第34卷，第2018–2033页。Curran Associates, Inc.

Minqing Hu 和 Bing Liu。2004年。挖掘并总结客户评论。载于*ACM SIGKDD* 知识发现与数据挖掘国际会议第十届会议论文集，第168–177页。

Susan Hunston. 2010年。 语料库评价方法：短语学与评价性语言，第13卷。Routledge。

Susan Hunston 和 Geoffrey Thompson。2000年。 文本中的评价：作者立场与话语建构：作者立场与话语建构。牛津大学出版社，英国。

Clayton Hutto 和 Eric Gilbert。2014年。Vader：一种简约的基于规则的社交媒体文本情感分析模型。载于*AAAI*网络与社交媒体国际会议论文集，第8卷。

Herbert H Hyman。1957年。对观点与人格的探索。《世界政治》，10(1):144–153。

FredaDzifa Intiful、Emefa Gifty Oddam、Irene Kretchy 和 Joana Quampah。2019年。探索加纳某大学学生的大五人格特征与饮食习惯之间的关系。*BMC*心理学，7(1):1–7。

Scott F. Kiesling、Umashanthi Pavalanathan、Jim Fitz- patrick、Xiaochuang Han 和 Jacob Eisenstein。2018年。在线论坛中的互动立场表达。计算语言学，44(4)：683–718。Vivek Kulkarni、Margaret L. Kern、David Stillwell、Michal Kosinski、Sandra Matz、Lyle Ungar、Steven Skiena 和 H. Andrew Schwartz。2018年。社交媒体语言中的潜在人类人格特质：一种开放词汇方法。*PLOSONE*，13(11)：1–18。Lisa M. Larson、Patrick J. Rottinghaus 和 Fred H. Borgen。2002年。大六兴趣类型与大五人格因素的荟萃分析。《职业行为杂志》，61(2)：217–239。Chei Sian Lee 和 Long Ma。2012年。社交媒体中的新闻分享：满足需求与既往经验的影响。《人类行为中的计算机》，28(2)：331–339。Ling Luo、Xiang Ao、Yan Song、Jinyao Li、Xiaopeng Yang、Qing He 和 Dong Yu。2019年。利用义原进行无监督神经方面抽取。载于第二十八届国际人工智能联合会议论文集，*IJCAI-19*，第5123–5129页。国际人工智能联合会议组织。Jean Pierre Malrieu。1999年。评价语义学：语言、认知与意识形态。Psychology Press。Tara C Marshall、Katharina Lefringhausen 和 Nelli Ferenczi。2015年。大五人格、自尊和自恋作为人们在 Facebook状态更新中所写主题的预测因素。《人格与个体差异》，85：35–40。

TomasMikolov、Kai Chen、GregCorrado 和 Jef- frey Dean。2013年。向量空间中的词表示高效估计方法。*arXiv*预印本 *arXiv:1301.3781*。

Daniel J Ozer 和 Veronica Benet-Martinez。2006年。《人格与对重要结果的预测》。《心理学年度评论》，57:401。

Bo Pang 和 Lillian Lee。2008年。观点挖掘与情感分析。*Found.TrendsInf.Retr.*，2(1–2):1–135。

Gregory Park、H. Schwartz、Johannes Eichstaedt、Margaret Kern、Michal Kosinski、David Stillwell、Lyle Ungar 和 Martin Seligman。2014年。通过社交媒体语言自动评估人格。《人格与社会心理学杂志》，108。

Umashanthi Pavalanathan、Jim Fitzpatrick、Scott Kiesling 和 Jacob Eisenstein。2017年。用于人际立场表达的多维词典。载于计算语言学协会第*55*届年会论文集（第*1*卷：长篇论文），第884–895页，温哥华，加拿大。计算语言学协会。

James Pennebaker 和 Laura King。2000年。语言风格：语言使用作为个体差异。 人格与社会心理学杂志，77：1296–312。

Maria Pontiki、Dimitris Galanis、John Pavlopoulos、Harris Papageorgiou、Ion Androutsopoulos 和 Suresh Manandhar。2014年。SemEval-2014任务4：基于方面的情感分析。载于第*8*届语义评测国际研讨会论文集（*SemEval 2014*），第27–35页，爱尔兰都柏林。计算语言学协会。Nils Reimers 和 Iryna Gurevych。2019年。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。载于 *2019* 年自然语言处理实证方法会议和第*9*届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第3982–3992页，中国香港。计算语言学协会。Nils Reimers、Benjamin Schiller、Tilman Beck、Johannes Daxenberger、Christian Stab 和 Iryna Gurevych。2019年。使用上下文化词嵌入进行论证分类与聚类。载于计算语言学协会第*57*届年会论文集，第567–578页，意大利佛罗伦萨。计算语言学协会。Margaret E Roberts、Brandon M Stewart、Edoardo M Airoldi、K Benoit、D Blei、P Brandt 和 A Spirling。2014年。结构主题模型。检索于*5*月，30:2014。Irene Russo、Tommaso Caselli 和 Carlo Strapparava。2015年。SemEval-2015任务9：CLIPEval事件隐式极性。载于第*9*届语义评估国际研讨会论文集（*SemEval*

*2015*年），第443–450页，美国科罗拉多州丹佛。计算语言学协会。

Ted Schwaba、Mijke Rhemtulla、Christopher J Hop- wood 和 Wiebke Bleidorn。2020年。特质面向图谱：可视化描述人格结构中混合、核心与边缘的网络。*PLOS ONE*，15(7):e0236893。

H. Andrew Schwartz、Johannes C. Eichstaedt、Margaret L. Kern、Lukasz Dziurzynski、Stephanie M. Ramones、Megha Agrawal、Achal Shah、Michal Kosinski、David Stillwell、Martin E. P. Seligman 和 Lyle H. Ungar。2013年。社交媒体语言中的人格、性别与年龄：开放词汇方法。*PLOSONE*，8(9):1–16。

Ewa Skimina、Jan Cieciuch 和 Włodzimierz Strus。2021年。人格特质与价值观作为日常行为频率的预测因素：模型与关卡之间的比较。《当代心理学》，40(1):133–153。

Christopher J Soto. 2019年。人格特质与重要人生结果之间的关联有多大可重复性？人格与人生结果重复研究项目。《心理科学》，30(5):711–727。

Ross David Stewart、René Mõttus、Anne Seeboth、Christopher John Soto 和 Wendy Johnson。2022年。更精细的细节？大五人格领域、人格侧面和细微差异对人生结果的可预测性。《人格杂志》，90(2):167–182。

Cajo ter Braak。1990年。通过结构相关和权重的双标图解读典型相关分析。《心理计量学》，55:519–531。

Silvia Terragni、Elisabetta Fersini 和 Enza Messina。2021年。基于词嵌入的主题相似度度量。载于自然语言处理与信息系统：第*26*届自然语言在信息系统中的应用国际会议（*NLDB 2021*），德国萨尔布吕肯，*2021*年*6*月*23*日至*25*日，会议论文集，第33–45页，柏林、海德堡。施普林格出版社。

Geoffrey Thompson 和 Susan Hunston。2000年。评价：简介。载于文本中的评价：作者立场与话语建构。牛津大学出版社。

Zhiqiang Toh 和 Wenting Wang。2014年。DLIREC：方面术语抽取与术语极性分类系统。载于第八届语义评价国际研讨会论文集（*SemEval 2014*），第235–240页，爱尔兰都柏林。计算语言学协会。

Stéphan Tulkens 和Andreas van Cranenburgh。2020年。简单得令人尴尬的无监督方面提取。载于计算语言学协会第*58*届年会论文集，第3182–3187页，线上。计算语言学协会。

Janyce Wiebe、Theresa Wilson、Rebecca Bruce、Matthew Bell 和 Melanie Martin。2004年。学习主观语言。计算语言学，30(3)：277–308。

Werner W Wittmann。2012年。评估研究中的对称性原则及其对罪犯治疗的启示。反社会行为与犯罪：发展与评估研究对预防与干预的贡献，（2011）：357–368。

Yuan Zuo、Junjie Wu、Hui Zhang、Hao Lin、Fei Wang、Ke Xu 和 Hui Xiong。2016年。短文本主题建模：伪文档视角。载于第*22*届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，KDD '16，第 2105–2114 页，美国纽约州纽约。计算机协会。

A 技术细节

A.1 数据集

我们选取了 P 上填写大五人格自我报告的用户评论ANDORA。我们仅保留英语评论，并剔除了不太可能包含评价性表达的候选内容，包括过短的评论——即少于五个词的评论——以及含噪评论——即非字母数字字符占比达到或超过 50% 的评论。我们使用 spaCy 的英语处理流水线将评论切分为句子，流水线名称为*en_core_web_lg*[12]，以获得更细粒度的处理结果。我们再次剔除了过短文本，不过这次剔除的是少于三个词的句子。

A.2 评价性过滤

当我们使用 QSB 挖掘初始评价性表达集合的释义句时，我们旨在将种子集（29k 个句子）的规模扩大十倍，因此对t_{sim}和 γ进行了调优。我们使用网格搜索，参数范围为 $t_{sim} \in [0.50, 0.90]$，步长为 0。05；另一个参数的步长为 $\gamma \in [0.90, 1.10]$，同样为 0。05，这使我们得到 $t_{sim} = 0.7$ 和 $\gamma = 1.05$。QSB 生成了 310k 个包含评价性标记的句子。

A.3 主题建模

我们使用了 scikit-learn 对潜在狄利克雷分配的实现。[13]我们根据 BTM（Cheng 等人，2014年）、ABAE（He 等人，2017年）和 CTM（Bianchi 等人，2021年）的原始论文改编了代码。ABAE 和 CTM 中基于神经网络的架构分别包含 100, 020 和 19,794,280 个参数。我们对 ABAE 训练了 5 个 epoch，对 CTM 训练了 20 个 epoch；在 10 次不同运行中，平均运行时间分别为 71.4 分钟和 590.8 分钟。

我们对所有尝试过的主题模型采用了相同的预处理方法。具体而言，我们删除了标点符号，并将文本转换为小写。我们还移除了停用词、网址、电子邮件、数字和货币符号，随后对词元进行了词形还原。我们使用不同的随机种子重复了每项实验 10 次。我们通过网格搜索，将主题数量作为超参数，在 $[5, 100]$ 范围内以 5 为步长进行搜索。对于其余超参数，我们使用默认值。

A.4 计算基础设施

我们在 $2\times$ *AMD Ryzen Threadripper 3970X 32* 核处理器 和 $2\times$*NVIDIA GeForce RTX 3090*GPU 上进行了实验，配备 24GB 内存。我们使用了*PyTorch* 1.9.0 版本和 CUDA 11.4。

B 其他分析

在图2中，某些评价主题与同一人格特质中的多个侧面呈现显著相关。例如，主题社会问题与以下愤怒、放纵和焦虑 等神经质 人格特质侧面呈显著正相关。另一方面，也有一些情况下，同一人格特质中的侧面与某一主题之间的相关性符号不同。这印证了我们的假设：主题与侧面的对应关系比与人格特质的对应关系更好。

[12]https://spacy.io/models/en#en_core_web_lg [13] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

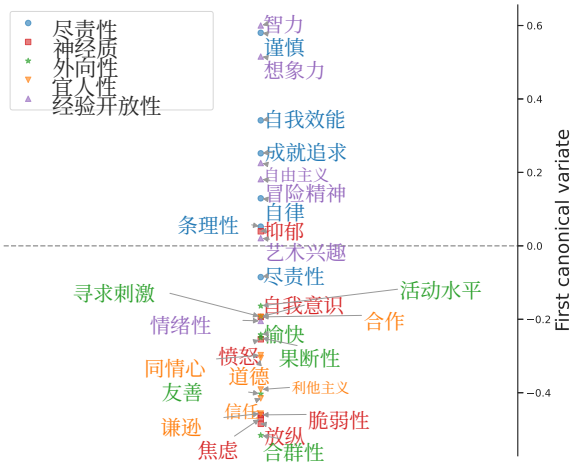

图5：侧面在第一个典型变量上的投影。y轴上的数值表示侧面在第一个典型维度上的典型载荷。

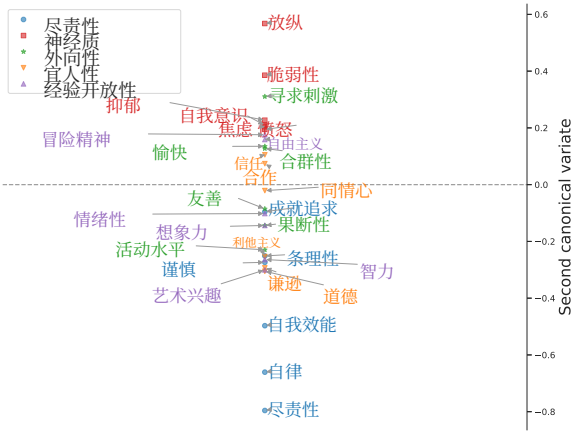

图6：侧面在第二个典型变量上的投影。第二个维度将尽责性和神经质完全分离，同时包含其余人格特质的混合侧面。