“学术伦理研究”专题

微观伦理视阈下伦理灰色地带问题探究

——以文本挖掘研究方法为例

张珍真

(广东财贸职业学院 基础教育学院,广东 广州 510445;澳门城市大学 人文社会科学学院,澳门 999078)

摘 要:微观伦理着眼于研究中情境化的具体问题,研究采用微观伦理视角,聚焦于人文社科领域中的伦理灰色地带问题,本文以当前广泛应用于各个领域的文本挖掘研究方法为例,对 100 篇相关学术文献进行内容分析,发现 52%的文献未明确提出伦理意识,而提出伦理意识的文献从参与者立场出发的较少。研究指出,文本挖掘研究方法的伦理灰色地带问题包括:参与者隐私保护、数据获取的合法性、分析的客观公正性、知情同意的确认,以及平台商业利益的维护;研究最后提出基于不同研究阶段的伦理考量与建议。

关键词:微观伦理学;应用语言学;文本挖掘;伦理灰色地带;数据隐私;知情同意

中图分类号:G40-059.1 **文献标志码**:A **文章编号**:1674-6414(2024)06-0054-14

0 前言

学术伦理问题一直备受研究者关注。随着研究主体、研究方法、研究内容、研究对象的动态发展,学术伦理的讨论愈加丰富,凯文·哈格蒂(Kevin Haggerty)(2004)就曾提出过伦理蠕动(ethical creep)的概念。随着媒介的发展,社交媒体已成为公共社交生活中重要的交流平台。近年来,Twitter、Facebook、微博和博客等社交媒体平台被广泛研究,因其实现了直接互动,产生大量用户见解,成为研究人员关注的新领域(Fiesler et al., 2018)。伴随着网页抓取技术的成熟,文本挖掘也被各个学术领域作为一项日益重要的研究方法而被使用。

同样,在应用语言学及相关领域,此类型研究也在不断增长,包括从互联网平台中获取用户的数据、内容与信息,从而进行语言教学研究、语言学习研究、社会语言学研究和话语

收稿日期:2024-07-05

作者简介:张珍真,女,澳门城市大学应用语言学博士生,主要从事应用语言学和话语分析研究。

引用格式:张珍真. 微观伦理视阈下伦理灰色地带问题探究——以文本挖掘研究方法为例[J]. 外国语文,2024(6):54-67.

分析等。它既可以进行量化的研究,也可以进行质性的分析,例如网络民族志等。此外,研究者通过网络爬虫抓取不同的社交平台的用户数据、帖子、图片、地理定位等,来探究公众态度和语言学习、身份认同等,也成为近期研究态势(Zhao et al.,2021)。

随着文本挖掘研究方法的发展,关于它的学术伦理问题的争议也备受关注。社交媒体内容因其内容公开,并且系用户自行发布,因此内容的所有权问题便变得敏感,而研究人员是否可以直接爬取社交媒体平台内容进行学术研究,学界对此则有着不同的观点。梅森等(Mason et al.,2002)就指出这其中涉及到数据的保护、用户隐私性的保护、知情同意,以及如何汇报数据等伦理困境,是学术伦理问题中的灰色地带。这种灰色地带往往难以被明确界定,却对学术的研究伦理性产生重要的影响。

本文将以文本挖掘研究方法为例,通过内容分析法,分析 100 篇使用文本挖掘方法的学术文献,探究文献中的伦理要素,以及研究者现有的伦理意识,并且根据彼得·德·克斯特(Petter De Costa,2014:414)提出作出伦理决定的三个阶段(研究前期阶段、数据收集与分析阶段、数据报告以及发表阶段),研究如何以一般的伦理原则解决文本挖掘的伦理灰色地带问题,并为学术伦理实践提供一些建议。

1　研究背景

学术伦理问题讨论由来已久,以应用语言学领域为例,从历史视角来看,自从 TESOL 研究委员会(Tarone,1980)发布《在英语为第二语言(ESL)研究领域中的学术伦理指引》以来,越来越多的学者开始从不同的角度对该领域学术研究以及实际教学中的伦理问题展开有意义的讨论。例如,将学术伦理问题逐步扩大到特定研究方法以及语言测试领域(Norris et al., 2000)。随后,随着互联网技术的发展,学界进一步探讨了数字研究的伦理问题(Spilioti et al., 2017)以及拓宽研究认知论的多样性 (Ortega, 2012),反映了伦理问题讨论的逐步深化与多样化。从理论视角来看,该领域对于伦理问题的探究集中于研究方法(Ortega, 2005),研究主体与过程(黄国文, 2024),以及关注大学院校中的伦理审查委员会(IRB)。

除此之外,学者们还从“宏观伦理”与“微观伦理”的角度上讨论此问题。关于宏观伦理,有学者指出它的概念包含两个维度:(1)程序伦理,指的是为计划中的研究项目从相关伦理委员会(如 IRB)获得的批准;(2)职业规范中所规定的伦理标准;但一般的“宏观伦理”原则可能不足以确保当前的伦理研究,需要扩大道德视角,并且制定一套更加情境化的实践准则,这便是“微观伦理”视角(Kubanyiova,2008)。卢尔德斯·奥尔特加(Lourles Ortega)(2016)则提出从微观伦理的角度入手,认为不能孤立地考虑伦理学的问题,应该从理论、方法、实践和认识论的背景下去综合考虑伦理问题。

笔者认为,在谈论学术研究的伦理问题时不仅需要遵守宏观伦理的原则,还需要根据具体的研究情景、研究方法、研究主体与对象,进行微观的探讨,特别是存疑性研究实践领域(questionable research practices area,QRPs),即伦理灰色地带,本文采用的便是“微观伦理”的视角。

道德准则就像房子的框架一般,并没有为研究过程中可能出现的每一个困境提供答案(Kennedy,2005)。这便是伦理灰色地带出现的原因。灰色伦理地带包括弱势群体,研究者角色混淆、同意、隐私、保密和匿名,以及风险的性质(Kennedy,2005)。可见,它存在于研究各个方面,涉及到研究主体、研究客体、研究方法,同时也存在于各个研究领域。就如文本挖掘所带来的伦理问题受到了越来越多学者的关注,马特·凯斯勒等(Matt Kessler et al.,2023)就曾提到对社交媒体的研究,反思性很重要,因为目前该类研究的伦理领域存在许多灰色地带。

当前,学术界对于文本挖掘的伦理性讨论主要集中于两个方面:一方面是针对其信息获取方式的伦理性与法律框架争议。例如,当我们在使用 Facebook 数据时需要从伦理角度考虑信息的私权与公权的问题,同时于法律层面需要考虑隐私与数据保护法规(Mancosu et al.,2020)。另一方面是知情同意与数据分析方式的伦理性争议。许多 IRB 考虑到公共数据不属于他们管辖范围,但在实际操作过程中,这种“公域”与“私域”的界限往往是模糊的;与此同时,大部分 Twitter 用户是希望研究前获得知情同意征求,可这类研究想要获得知情同意却是困难的(Bruckman,2014)。

可见,文本挖掘技术所涉及到的伦理问题有着广泛的关注度与讨论,但主要集中于某一个方面,鲜有学者围绕着德·克斯特(De Costa,2014)提出的作出伦理决定的三个阶段来对这项技术所涉及的灰色地带伦理问题进行完整的探究。因此,本文将基于此,研究以下三个问题:

(1)国内文本挖掘类研究的伦理意识现状如何?

(2)解决此类研究中的伦理灰色地带应遵循什么伦理准则?

(3)在这类研究中如何避免出现伦理问题?

2 内容分析:文本挖掘研究方法的伦理灰色地带

为使研究更加全面,我们在中国知网(CNKI)通过多个关键词(社交媒体+内容分析/文本挖掘/情感分析/话语分析/数据挖掘),分五次检索,收集到 2013—2024 年使用文本挖掘研究方法的文献共 3 808 篇,删除学位论文 2 876 篇,删除不符合研究目的文献,如标题不符合要求,全文阅读后不符合研究要求,以及书评类、技术类和无关主题类文献,去重后,最终确定 100 篇($n=100$)文献为研究样本。

进行内容分析的过程,需要了解这类研究是否具备伦理意识,以及出现伦理问题时如何解决,推断伦理意识是否随着时间逐渐加强,本文提出以下一个假设:

随着时间的推移,伦理意识有所增强。

根据研究目的与研究问题,笔者将这 100 篇文献按照文献基本信息、研究特征和伦理意识三个方面进行编码:

表 1　编码分类

类别	详细信息
文献基本信息	标题
	发表年份
	研究领域
研究特征	研究对象
	研究方法
	爬虫工具
	分析工具
	分析类型
	爬取平台
	样本容量
伦理意识	是否涉及特殊人群
	是否具备伦理意识
	研究样本呈现方式
	是否出现具体推文信息

在伦理意识方面,在对样本中的文献编码时,从参与者隐私性、数据准确性、知情同意、获得许可、公平与偏见和伦理声明六个评价维度来评价文献中的伦理意识。

编码时主要关注:(1)在数据爬取与数据汇报的过程中是否注意保护参与者隐私;(2)对参与者是否进行知情同意征求,是否关注数据准确性,数据真实性;(3)是否遵守相关平台协议,遵守相关法规;(4)数据分析时是否注意公平与无偏见,是否有伦理声明。如果这些在文献中有所体现,该文献则被视为具备一定的伦理意识。为了评估编码过程的一致性,本文采用 Holsti 系数与 Cohen's Kappa 系数交叉验证,保证编码的准确和一致性。

研究中,邀请两名了解内容分析法的博士生编码者对文献进行逐条编码,在编码的最开始,针对少量样本共同讨论,确定两位编码者对伦理意识有共同的认知与界定,随后这两名编码者对剩余的样本进行编码,并计算编码者间的一致性系数。

Holsti 一致性系数(CR)通过以下公式计算得出:

$$CR = \frac{2M}{N_1 + N_2}$$

其中,M 指两名编码者一致的样本数,N_1 和 N_2 分别指编码者 1 和编码者 2 编码的总样本数。

最终,编码者 1 和编码者 2 分别编码的 100 个样本中共有 92 个一致。经计算,Holsti 系数为 0.92,表示编码者之间一致性的比例很高。同时,检测 Cohen's Kappa 系数,它用于衡量编码者之间的分类一致性,考虑了一致性中的随机因素,其公式为:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

根据计算,Cohen's Kappa 系数为 0.84,表明在考虑随机因素后的编码一致性仍然非常好,接近完美一致性。综上,Holsti 检验结果显示,一致性系数(CR)为 0.92,Cohen's Kappa 系数为 0.84。两项验证表明编码者之间具有高度一致性,编码过程可靠。

2.1 结果

2.1.1 文献基本信息与研究特征

这些文献的发表时间涵盖了从 2013 年到 2024 年,其中 2021 年和 2022 年为发表高峰,分别占 22%和 18%。这些文献分布在多个研究领域,包括社交媒体(30%)、新闻传播(13%)、商业(12%)和语言学(5%)等,此外还有信息技术(5%)、社会学(5%)、公共卫生(4%)、旅游管理(4%)、体育(4%)等领域,反映了文本挖掘技术在各个研究领域的广泛应用。

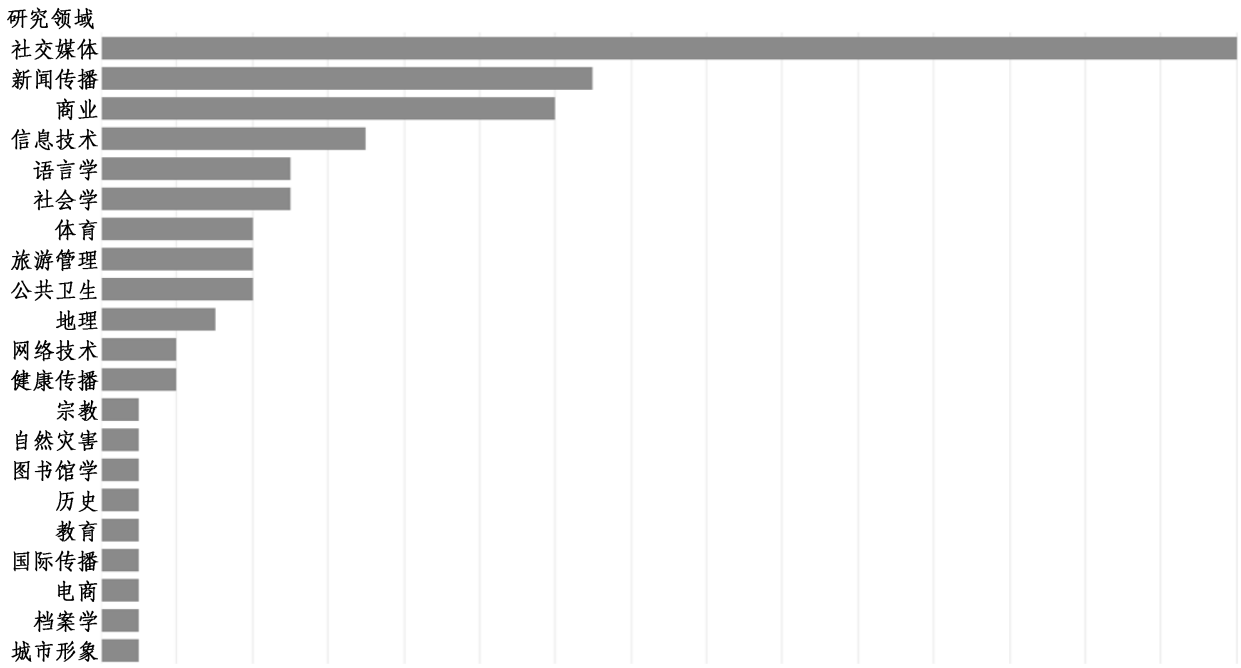

图 1 研究领域分布图(n=100)

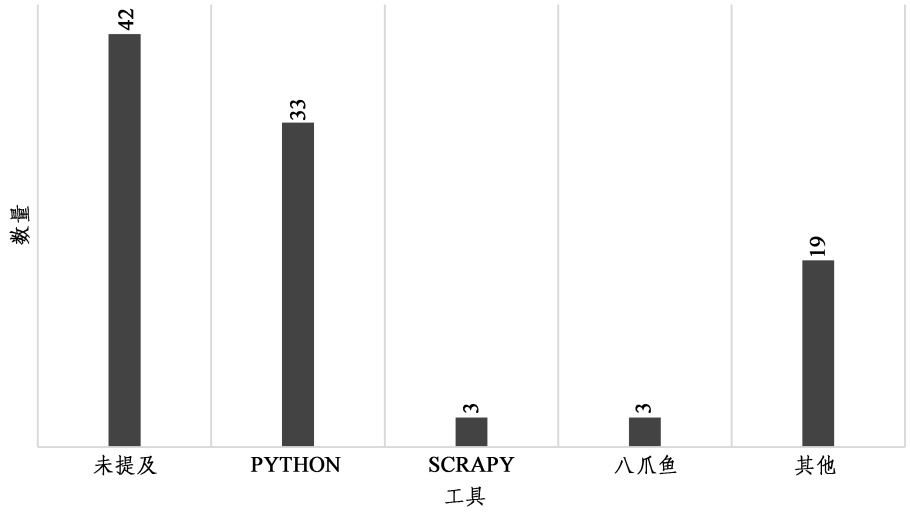

图 2　爬取的工具选择($n=100$)

微博是最常用的爬取数据来源平台,占 51%,其次是 Twitter(现为 X),占 10%,其他数据来源包括 B 站(3%)、微信(2%)以及其他平台(34%),如电子商务类网站、知乎、CNKI、京东商城、百度贴吧、YouTube 和抖音等。针对海外的研究,大多数研究者选择 Twitter 作为研究平台。数据来源的多样性反映了文本挖掘技术在不同平台上的广泛应用。在爬取工具方面,有 42%的文献未明确提及使用工具,而提及使用工具的文献中,Python 语言占比 33%,是使用最多的爬取工具,其余工具有 Scrapy(3%)、八爪鱼(3%),剩下为其他(19%)。数据爬取后,在分析类型上主要集中于主题分析(LDA)以及情感分析。

2.1.2 伦理意识

在伦理意识方面,52%的文献未明确提及伦理问题,且样本中的文献均未提及“知情同意”与“伦理声明”。其余 48%的文献涉及到了伦理意识,其中 24%从研究者立场关注数据准确性,4%关注数据真实性,4%关注数据样本量,3%既关注用户隐私也关注数据准确性,只有 8%站在参与者角度上关心用户隐私,从数据中可见研究者立场高于参与者立场。

在涉及用户隐私的研究中,有研究者强调保护用户个人信息的重要性,防止数据泄露和滥用。例如,郑文等(2020)在微博平台爬取用户信息进行分析时,因保护个人隐私,隐去用户个人信息。同时,在数据准确性方面,也有研究者在文章中提出了对爬虫爬取的数据的准确性表示一定的担忧(杨奕 等,2021)。此外,还有部分文献讨论了数据样本量的重要性,强调了大规模数据在统计分析中的优势和潜在问题。在遵守平台协议方面,罗向东等(2024)在对京东平台的用户评论进行挖掘时,提出要遵守平台的网络协议以及《中华人民共和国网络安全法》,不能对平台服务器造成压力,并要保证数据的合法性。

在涉及特殊人群的研究中($n=12$),如弱势群体、未成年人、学生等,有伦理意识的仅有四篇,一篇关注用户隐私性,三篇关注数据真实性与准确性。可见,更多的研究是从研究者

立场出发,对于特殊人群的研究伦理意识的考虑还需更加细微。

表 2 伦理意识特征

编号	是否具备伦理意识	数量	百分比
1	否	52	52%
2	是(数据准确性)	24	24%
3	是(用户隐私)	8	8%
4	是(数据真实性)	4	4%
6	是(数据样本量)	4	4%
5	是(用户隐私;数据准确性)	3	3%
7	是(平台协议)	3	3%
8	是(平台协议;互联网法规)	1	1%
9	是(利益冲突)	1	1%
	总计	100	100%

2.2 假设检验

同时,为了探讨不同年份发表的文献在伦理意识方面的不同情况,验证前文的假设,使用 SPSS 29.0 进行逻辑回归、斯皮尔曼相关性分析以及频数分布统计。

表 3 逻辑回归

项	回归系数	标准误	z 值	Wald X2	p 值	OR 值	OR 值 95% CI
发表年份	-0.066	0.08	-0.818	0.67	0.413	0.936	0.800 ~ 1.096

本文将发表年份作为自变量,将伦理意识进行编码(0 为否,1 为是)作为因变量进行二元逻辑回归分析。从表 3 可知,模型公式为:ln(p/1-p)= 132.902-0.066 * 发表年份(其中 p 代表“伦理意识”为 1 的概率,1-p 代表“伦理意识”为 0 的概率)。最终分析得出的结果是:发表年份的回归系数值为-0.066,但并没有呈现出显著性(z=-0.818,p =0.413>0.05),意味着发表年份未与伦理意识产生影响关系。

表 4 相关性分析

		伦理意识
发表年份	相关系数	-0.049
	p 值	0.630
	样本量	100

* p<0.05 ** p<0.01

随后,我们利用相关分析去研究伦理意识与论文发表年份之间的相关关系,使用斯皮尔曼相关系数表示相关性的强弱情况。从表 4 可见,伦理意识和发表年份之间的相关系数值为-0.049,接近于 0,并且 p 值为 0.630>0.05,因而说明伦理意识和发表年份之间并没有

相关关系。

表 5　不同年份伦理意识比率比较

年份	总篇数	伦理意识	百分数(%)
2013	1	0	0
2014	2	2	100
2015	2	0	0
2016	5	5	100
2017	1	0	0
2018	5	4	80
2019	6	2	33.33
2020	10	4	40
2021	22	10	45.45
2022	18	7	38.89
2023	18	8	44.44
2024	10	6	60

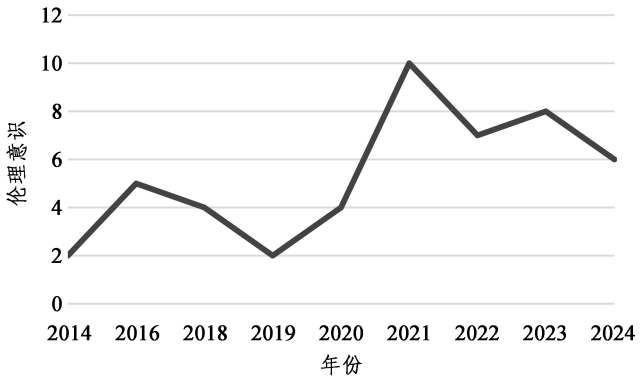

图 3　样本中具备伦理意识的文献分布曲线图

通过逻辑回归和相关性分析,可以看到在这 100 篇文献中,尽管文本挖掘技术在学术研究中的应用广泛,但对伦理意识的关注程度并未随时间发展逐渐上升,假设不成立。可见收集的这 100 篇文本挖掘类文献的伦理意识还有一定的提升空间。

2.3 假设验证

在对 100 篇文本挖掘相关文献进行内容分析后,我们从编码后的数据中可以看到,该技术的研究领域分布广泛,研究方法以量化分析为主,爬取工具主要为 Python 语言,爬取平台中,51%来自微博,海外的研究则更多是基于 Twitter 平台爬取数据。根据微博 robots 协议,用户使用爬虫进行数据爬取具有一定的合规性风险,这是大部分研究者未考虑到的。

在伦理意识方面,有 52%的文献未在文中体现伦理意识,在提出伦理意识的文章中,大多站在研究者的立场对数据准确性和样本代表性表示担忧,只有 8%从参与者立场,提出用

户隐私保护问题。同时,鲜有文献提及“知情同意”以及“伦理声明”。

综上,前文所提的假设“随着时间的推移,伦理意识有所增强”不成立,这表明当前该领域的学术伦理意识有待提升,研究者们需要更加注意研究的伦理问题。

3 讨论

基于上述内容分析,为回答研究问题二:解决此类研究中的伦理灰色地带要遵循哪些伦理准则? 我们需要明确,在进行学术研究过程中需要遵守哪些基本的伦理规范。

3.1 学术研究伦理的常识性准则:参与者立场

关于学术伦理规范问题,有诸多机构与学者提出相关的指南。1979 年,美国国家保护人类受试者委员会撰写了《贝尔蒙报告》,提出三项基本原则:尊重人(respect of persons)、仁慈(beneficence)和正义(justics)(U. S. Department of Health and Human Services,1979)。这些原则成为许多领域学术研究的基本原则与惯例。品普(Pimple,2002)将伦理研究简化为三个价值:(1)在报告和展示数据时真实;(2)在引用和使用他人作品时公平;(3)只进行有意义和有用的研究的智慧。此外,伦理审查委员会经过伦理蠕动后明确伦理审查的目的包括:评估研究活动对参与者造成的伤害类型和风险;确保参与者能够知情同意参与研究活动;监督研究者的程序以维护参与者的匿名性和保密性(Haggerty,2004)。

在这些原则中,公平(公平对待参与者、公平引用他人作品)、保护参与者(隐私、风险)、知情同意(确保参与者知情同意)等概念被反复提及,而这些原则均从参与者立场出发,在保证尊重的前提下,尽可能维护参与者的权利。因此,在进行文本挖掘技术所涉及的伦理问题时,研究者可站在参与者立场,以公平性、真实性为标准进行判断。然而,在文本挖掘研究方法中,参与者除了社交媒体平台中的用户以外,平台本身作为数据的提供方与运营方,也是重要的参与者,因此保障社交媒体平台在研究中的权利也是在研究中需要考虑的关键点。总而言之,参与者立场是常识性伦理中的重要准则。

3.2 文本挖掘的伦理性问题以及对应策略探讨

3.2.1 研究前期阶段:明确爬取数据的法律风险

在研究开始之前,研究者就需要具备伦理意识。黄国文(2024)指出,在学术研究中,研究者居于重要地位,其个人的意识形态、哲学立场和价值观会影响整个研究过程。研究人员需要评估他们的研究活动对参与者有可能造成伤害的类型和风险(Haggerty,2004)。因此,研究者在研究开始前,需要根据自身伦理判断,明确他们的研究设计、研究方法、研究目的是否符合一定的伦理准则。

对于文本挖掘来说,其特殊之处在于研究之前就需明确数据获取的合法合理性,这需

要研究人员具备伦理与法律意识,研究前熟悉一定的法律法规以及网站服务条款,确保自身研究行为不涉及法律风险。凯斯勒等(2023) 就曾指出研究者在互联网上进行人类学研究,会因网络数据的“公有权”与“私有权”的问题而面临独特的伦理挑战。其实,这类争议在学界一直没有停歇过。德莱尔等(Dreyer et al.,2013) 认为网络数据存在一种固有的悖论,一方面,网络应该是开放的且易于公众访问;另一方面,这些网络数据对于网站所有者来说是重要的资产,因此需要受到保护。从法律角度来看,网站所有者不一定拥有网站用户生成的数据,这便形成了一个灰色地带。研究人员在开始研究前,需要多加注意会涉及到的伦理与法律风险。

3.2.2 数据收集以及数据分析阶段

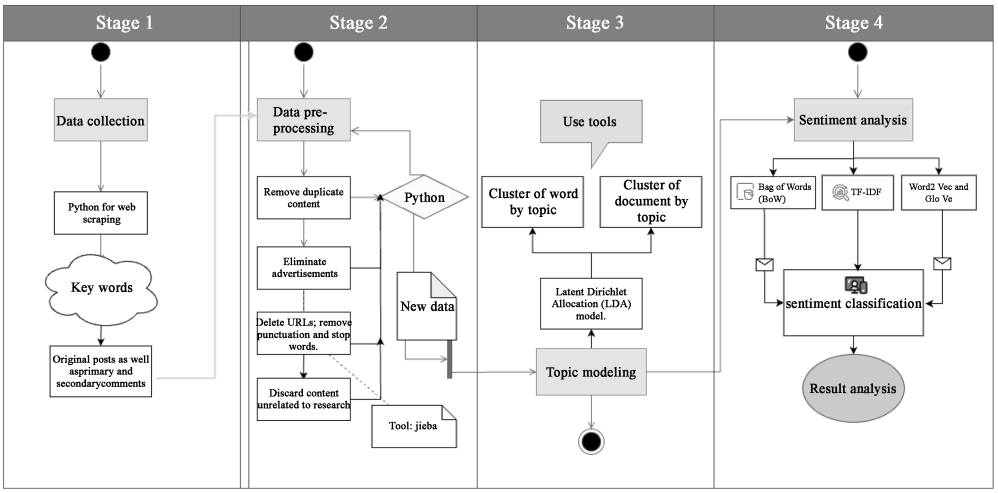

图 4　使用文本挖掘进行研究的流程图

如图 4 所示,在使用文本挖掘方法进行研究与分析的过程中,所面临的伦理灰色地带有以下几个方面:

(1)数据获取的合法性

我们从上文的内容分析中看到,文本挖掘技术的数据获取方式多数是使用 Python 等编程语言进行网络抓取,但这种方式一直在法律的边缘游走,给科研人员带来巨大的法律风险。2018 年,剑桥分析(Cambridge Analytica)丑闻发生之后,Facebook 就收紧了技术手段访问(Mancosu et al., 2020)。因此,在进行网络抓取过程中第一个伦理考虑便是查看网页是否包含禁止自动网页抓取的 robots. txt. 文件,数据获取如果不遵守此类协议,将会面临伦理风险和法律风险(Krotov et al., 2020)。

在中国的语境下也是如此,例如常用的社交平台微博,其服务使用协议(https://weibo.com/signup/v5/protocol)的第 1.3.2 条款中明确写道:未经微博运营方事先书面许可,用户不得自行或授权、协助任何第三方非法抓取微博内容,同时解释“非法抓取”是指采用程序

或者非正常浏览等技术手段获取内容数据的行为。同时,1.4 条款明确用户应当严格遵守微博运营方所发布的 robots 协议。然而,在基于文本挖掘的学术研究中,鲜有研究人员提及获取网站平台的书面许可。前文分析提及平台协议的文献只占 3%。这是商业平台与学术研究之间的一道鸿沟。非法获取数据不仅会损害用户的个人隐私,还会损害数据平台利益,这将违反学术研究伦理道德。

针对这个问题,理查德·兰德斯等(Landers et al., 2016)建议研究人员只抓取公开的、未加密的数据源,以避免法律风险。因而,研究者在研究开始前需要了解相关的伦理问题。

(2)用户知情同意书的获取

在一般伦理准则中,研究人员需要确保参与者能够在知情的情况下同意参与研究活动。在实验研究中,获得受试者的知情同意书是研究人员的共识,但在社交媒体中,这个问题就变得复杂。在此前的分析中,鲜有文献提及“知情同意”,因为往往文本挖掘技术在数据收集的阶段,获取的用户数据是海量的,都得到用户的知情同意,似乎是不现实的;并且网络用户的个人信息并不一定是真实的,因此一对一联系取得知情同意的难度很大。

是否可以放弃获取知情同意书?海伦·尼森鲍姆(Helen Nissenbaum)(2004)认为信息收集和传播应适合情境。例如,尽管研究人员普遍认为 Twitter 是公开的,但用户同意公众查看其推文,并不意味着同意研究人员收集和分析这些推文。

针对这一困境,有学者提出或许可以通过获得技术公司/平台代表的许可,并寻求批准使用其数据,因为一些网站会有一定的法律免责声明(Spilioti et al.,2017;Mancosu et al., 2020)。研究者可以通过阅读这些免责声明来合理合法地获取数据资源,但这个问题也需要根据不同平台特性,进行具体情况的分析,特别是在中国的法律框架下,遵守不同网络平台用户协议。在研究开始之前,阅读不同平台用户协议,是研究者不能忽视的步骤。

(3)数据分析时的灰色地带

《贝尔蒙报告》(1979) 指出,研究人员需要尊重人、仁慈和正义。因此,研究者在分析数据时同样需要遵守这些原则。值得注意的是,网络数据抓取的信息存在的第一个风险是分析的偏见性。研究者在分析时有可能会根据之前的偏见进行推论与编码,这会助长分析的歧视性(Krotov et al., 2020)。例如,在应用语言学领域中,常见的数据分析有对爬取的语料进行话语分析,在此过程中,语料的公正性选取与分析,以及消除研究带来的次生伤害,都是研究者需要考虑的问题。

第二个风险是在数据分析阶段需要保护用户的个人隐私。在对语料进行分析和传播过程中,保护参与者的身份是巨大的挑战。例如,截取社交媒体中的内容进行分析,若内容较为敏感与特殊,便需要通过特殊途径保护用户信息。有研究表明,基于社交网络的特性,

即便隐藏个人信息,参与者仍然可能被识别,因此需要对用户信息进行特殊编码隐藏(Kessler et al. , 2023)。此外,也不能忽视对内容中所涉及到的弱势群体、儿童的保护,在社交媒体中不缺乏对于此类人群的内容展示,而研究人员在处理这方面信息的时候,需要特别注意他们的隐私权、名誉权等。在前文中内容分析中可以看到,这是目前伦理意识较淡薄的方面。

第三个风险是数据真实性、可靠性的问题。在分析中,共有 31%的文献提到了数据的准确性与真实性,以及样本容量大小方面的担忧,因为网络数据中存在大量的不完整、不准确的信息,而研究人员仅通过分析网络数据作出战略决策,有可能导致不严谨的研究结论(Krotov et al., 2020)。又如,通过分析社交媒体帖子进行用户情感分析,其中面临的挑战是互联网语言的复杂性与情感态度的标注与分类的简单性之间的矛盾,这也会影响研究的准确性。

3.2.3 数据报告以及研究发表阶段

品普(2002)将伦理研究简化为报告和展示数据时的真实性,因此在汇报数据和发表的时候需要将数据收集与分析过程完整透明呈现。虽然 IRB 没有给具体的指导纲要,但基于常识性的伦理准则,在此阶段还需要充分站在参与者的立场来做出伦理判断,并且针对不同的参与者身份、群体而采取不同的处理方式,例如需要充分重视弱势群体的权利,对于社交媒体中的政要、社会组织、公众人物的信息处理时也需要慎重。同时,还不能忽视社交媒体网站的利益,避免因为数据爬取而引起不必要的商业纠纷。

4　结语

本文通过对 100 篇学术文献的内容分析,揭示了文本挖掘技术在不同研究领域中的应用现状和伦理意识的关注情况。结果表明,尽管文本挖掘技术应用广泛,但对伦理意识的重视程度仍需提升。例如,“伦理声明”“知情同意”等方面仍存在较大改进空间。在伦理考量上,多数研究从研究者视角出发,对参与者立场的关注较少。

在微观伦理视阈下,针对此类研究的伦理灰色地带,研究者们应从参与者立场出发,保证参与者的权利,包括知情同意权,隐私权,名誉权等;同时要保证数据来源平台的权利,避免商业利益冲突和法律纠纷;在研究开始前了解数据来源平台的相关政策与协议,尊重数据所有权。在数据分析过程中,要保证客观公正,可通过各类编码来保护参与者隐私。

总之,文本挖掘研究方法所带来的技术与伦理的反思,将促进这一研究方法在将来的发展与使用,但这需要在遵守人类社会伦理准则的前提下,符合各个研究领域的学术伦理要求。这需要研究人员共同促进。

参考文献：

Bruckman, A. 2014. Research Ethics and HCI. *Ways of Knowing in HCI*[M/OL]. *Springer eBooks*,449-468.

De Costa, P. I. 2014. Making Ethical Decisions in an Ethnographic Study[J]. *Tesol Quarterly*(2): 413-422.

Dreyer, A. & J. Stockton. 2013. Internet "Data Scraping": A Primer for Counseling Clients[J]. *New York Law Journal* (7): 1-3.

Fiesler, C. & N. Proferes, 2018. "Participant" Perceptions of Twitter Research Ethics[J]. *Social Media+Society*(1):1-14

Haggerty, K. D. 2004. Ethics Creep: Governing Social Science Research in the Name of Ethics[J]. *Qualitative Sociology*(27): 391-414.

Kennedy, J. E. 2005. Grey Matter: Ambiguities and Complexities of Ethics in Research[J]. *Journal of Academic Ethics*(3): 143-158.

Kessler, M., Marino, F. & D. Liska. 2023. Netnographic Research Ethics in Applied Linguistics: A Systematic Review of Data Collection and Reporting Practices[J]. *Research Methods in Applied Linguistics*(3):100082.

Krotov, V., Johnson, L. & L. Silva, 2020. Tutorial: Legality and Ethics of Web Scraping[J]. *Communications of the Association for Information Systems*(1):22.

Kubanyiova, M. 2008. Rethinking Research Ethics in Contemporary Applied Linguistics: The Tension Between Macroethical and Microethical Perspectives in Situated Research[J]. *The Modern Language Journal*(4):503-518.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J. & A. B. Collmus. 2016. A Primer on Theory-driven Web Scraping: Automatic Extraction of Big Data from the Internet for Use in Psychological Research[J]. *Psychological Methods*(4):475.

Mancosu, M. & F. Vegetti. 2020. What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data[J]. *Social Media+ Society*(3):.

Mason, S. & L. Singh. 2022. Reporting and Discoverability of "Tweets" Quoted in Published Scholarship: Current Practice and Ethical Implications[J]. *Research Ethics*(2): 93-113.

Nissenbaum, H. 2004. Privacy as Contextual Integrity[J]. *Washington Law Review*(79):101-139.

Norris, J. & L. Ortega. 2000. Effectiveness of L2 Instruction: A Research Synthesis and Quantitative Meta-analysis[J]. *Language Learning*(3):417-528.

Ortega, L. 2005. Methodology, Epistemology, and Ethics in Instructed SLA Research [special issue] [J]. *Modern Language Journal*(3):317-327.

Ortega, L. 2012. Epistemological Diversity and Moral Ends of Rsesearch in Instructed SLA[J]. *Language Teaching Research*(2): 206-226.

Ortega, L. 2016. Multi-competence in Second Language Acquisition: Inroads into the Mainstream[G]//Vivian Cook & Li Wei. *The Cambridge Handbook of Linguistic Multi-competence*. New York: Cambridge University Press,50-76.

Pimple, K. 2002. Six Domains in Research Ethics: A Heuristic Framework for the Responsible Conduct of Research[J]. *Science and Engineering Ethics*(8):191-205.

Spilioti, T. & C. Tagg. 2017. Ethics of Online Research Methods in Applied Linguistics [special issue] [J]. *Applied Linguistics Review*(2-3):163-167

Tarone, E. 1980. TESOL Research Committee Report: Guidelines for Ethical Research in ESL[J]. *TESOL Quarterly* (3): 383-

388.

U. S. Department of Health and Human Services. 1979. *Belmont Report*[R].

Zhao, H. & H. Liu. 2021. (Standard) language Ideology and Regional Putonghua in Chinese Scial Media: A View from Weibo [J]. *Journal of Multilingual and Multicultural Development*(9): 882-896.

黄国文. 2024. 应用语言学教学与研究中的伦理问题[J]. 外国语文(2):1-12.

罗向东,强威,张希莹等. 2024. 基于文本挖掘的跑鞋用户评价及情感分析[J]. 丝绸(6):108-119.

杨奕,张毅. 2021. 复杂公共议题下社交媒体主题演化趋势与社会网络分析——以中美贸易争端为案例的比较研究[J]. 现代情报(3):94-109.

郑文,赵偲,李泽堃,等. 2020. 基于 Web 数据挖掘的 COVID-19 流行病学特征分析[J]. 电子科技大学学报(3):408-414.

Exploring Ethical Grey Areas from a Micro-Ethics Perspective: A Case Study of Text Mining Research Methods

ZHANG　Zhenzhen

Abstract: Micro-ethics focuses on context-specific issues in research. The ethical grey areas in the humanities and social sciences are explored were from a micro-ethical perspective. 100 related academic articles was conducted with the widely applied text mining research method as an example. Findings indicate that 52% of the articles did not explicitly address ethical awareness, and those that did rarely adopted the perspective of participants. Key ethical grey areas in text mining research include the protection of participants' privacy, the legality of data acquisition, objectivity and impartiality in analysis, the confirmation of informed consent, and the protection of platform commercial interests. This paper concludes by presenting ethical corsiderations and recommendations specifically aligned with each stage of the research process.

Key words: micro-ethics; applied linguistics; text mining; ethical grey areas; data privacy; informed consent

责任编校:蒋勇军