硕士学位论文

基于社交媒体的用户行为和心理研究

RESEARCH ON USER BEHAVIOR AND PSYCHOLOGICAL CHARACTERISTICS BASED ON SOCAIL MEDIA

罗晶欣

哈尔滨工业大学

2020 年 6 月

国内图书分类号：C37　　学校代码：10213
国际图书分类号：311　　密级：公开

应用统计硕士学位论文

基于社交媒体的用户行为和心理研究

硕士研究生：罗晶欣
导　　师：夏志宏教授
申请学位：应用统计硕士
学　　科：应用统计
所在单位：南方科技大学
答辩日期：2020 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index: C37
U.D.C: 311

Dissertation for the Master Degree in Applied Statistics

RESEARCH ON USER BEHAVIOR AND PSYCHOLOGICAL CHARACTERISTICS BASED ON SOCAIL MEDIA

Candidate： Luo Jingxin
Supervisor： Prof. Xia Zhihong
Academic Degree Applied for： Master of Applied Statistics
Speciality： Applied Statistics
Affiliation： Southern University of Science and Technology
Date of Defence： June, 2020
Degree-Conferring-Institution： Harbin Institute of Technology

摘 要

随着社交媒体的兴起，社交网络逐渐成为日常生活的延展，用户在社交媒体平台上设置基本信息进行自我展示，发布内容进行自我表露，点赞互动评论进行社交网络交流，在自我与社交媒体的交互过程中吐露心声、宣泄情绪、获得社会支持。在使用社交媒体的过程中，用户将自己的用户画像——人格、主观幸福感等心理特征的痕迹印刻下来，积累的海量数据被广泛应用于广告、推荐、搜索等各个业务领域。

基于用户数据如此丰富的社交媒体平台，是否可以建立预测模型以评估用户的心理健康状态，是本文的主要研究目标。本文以新浪微博用户作为研究对象，提取微博行为特征与用户的大五人格、主观幸福感和自尊特征，建立用户心理特征的预测模型，并对心理特征各维度的影响因素进行分析。

本文利用 Python 网络爬虫收集微博的数据，在数据处理的过程中，综合考虑情感词典和机器学习等情感分析技术，对比发现基于朴素贝叶斯算法的情感分析模型效果最好，准确率可达 80.5%。另外，本研究从用户特征、微博行为特征和微博情感特征三个方面对用户心理特征分别建立了回归模型和分类模型。本文结果显示，多元线性回归模型在大五人格的尽责性维度和开放性维度有一定的解释效果，Logistic 回归模型能较好地预测大五人格的宜人性、开放性和主观幸福感的消极情感频率。决策树模型对用户心理特征的整体预测效果最好，随机森林则分别在大五人格的外向性维度、主观幸福感的消极情感频率和自尊的预测上表现较好。

关键词：社交媒体；情感分析；主观幸福感；自尊；大五人格

Abstract

With the rise of social media, social networks have gradually become an extension of daily life. Users post up basic information about themselves on social media to display themselves, post weibo for self-disclosure, interact with others for social network communication. During the interaction, they express themselves, post their emotions, and obtain social support. While using social media, users imprinted their user portraits—personality, subjective well-being and other psychological characteristics. The massive data is widely used in various business fields such as advertising, recommendation, and search engine.

Based on a social media platform with a lot of user data, whether a predictive model can be established to assess the user's mental health is the main research goal of this article. This article takes Sina Weibo users as the research object, extracts the Weibo behavioral characteristics and the user's big five personality, subjective well-being and self-esteem characteristics, establishes a prediction model of users' psychological characteristics, and analyzes the influencing factors of each dimension of psychological characteristics.

This article uses Python web crawler to collect Weibo data. In the process of data processing, comprehensively consider sentiment analysis techniques such as sentiment dictionary and machine learning. The comparison found that the sentiment analysis model based on Naive Bayes algorithm has the best effect and the accuracy rate is up 80.5%. Finally, regression model and classification models were established for user psychological characteristics from three aspects: user characteristics, microblog behavior characteristics and microblog emotional characteristics. The results of this article show that the multiple linear regression model has a certain explanatory effect in the conscientious and openness dimensions of the Big Five personality. The Logistic regression model can predict the negative emotional frequency of the subjective well-being and agreeableness, openness of the Big Five personality. The decision tree model has the best overall prediction effect on the user's psychological characteristics. The random forest performs better in predicting the extraversion of the Big Five personality, the negative emotional frequency of subjective well-being and self-esteem.

Keywords: social media, sentiment analysis, subjective well-being, self-esteem, big five factor model

目 录

第 1 章 绪　论

1.1 课题来源

二十一世纪以来计算机信息技术飞速发展，互联网各个领域不断迭代出新，想尽千方百计挖掘人们的需求，并针对多种多样的需求推出了覆盖日常社交、饮食、出行、支付、娱乐等各个方面的产品。社交媒体就是其中最为重要的“发明”之一。根据针对的用户群体和需求的差异，已经衍生出了众多不同类型的社交媒体平台和产品，全球最大的社交媒体网站 Facebook 就是其中最为典型的代表。2004 年 Facebook 正式成立，经过十年的光速发展，该平台的单日用户数在 2015 年已经突破十亿，象征着社交媒体已经深入覆盖众多人群，逐渐成为人们日常生活必不可少的一部分。

社交媒体以其实时性、传播性、交互性、丰富性等特点逐渐改变了人们沟通交流、互动、合作、获取信息、投票、认知等的方式。通过访问 Facebook、Twitter、Instagram、新浪微博、微信、QQ 等社交媒体平台，用户在虚拟网络世界形成的社交网络中传递信息、建立联系、进行自我展示。用户生成内容（User-generated content）和基于用户关系的互动是社交媒体的核心内容，比如新浪微博的文本内容、照片、视频等用户生成内容和用户之间的关注、点赞、评论、转发等社交行为，共同构成了新浪微博这类社交媒体的内容生态。国内的新浪网在 2009 年推出了新浪微博，提供微博客类型的社交媒体服务，并击败了众多同类型产品，在短信息社交媒体平台中一枝独秀。用户可以通过网页端、手机移动端等方式访问新浪微博，通过个人账号发布动态、上传图片和视频，实现即时分享新鲜事和传播互动。截至 2019 年底，新浪微博月活跃用户已达到 5.16 亿，其中 16-25 岁用户群体占比 61%，图片日均发布量为 1.2 亿，微博长文的日均发布量达到 48 万多篇，微博文字内容日均发布量达到 1.3 亿条。新浪微博已成为政治、社会、经济、文化各领域的意见交换和内容聚集的主要平台之一。

随着技术的更新换代，大数据时代背景下的社交媒体平台积累了海量、种类繁多、客观的用户行为数据，从中挖掘的信息资源已被广泛应用于金融、电子商务、广告、出行、安全等诸多领域。近年来，信息技术与社会学、心理学等社会科学领域结合的研究开始得到广泛关注。心理学致力于探讨人类的心理特征和行为规律，社交网站（Social network site, SNS）的使用行为（如使用动机、使用频率、自我表露程度、线上社会支持、反馈等）一直是心理学领域内广泛关注的研究方向。传统的心理学行为研究主要通过行为实验、问卷、访谈、

自我报告等线下的方式开展，收集数据费时费力，并且数据容易存在偏差。社交媒体的兴起及其产生的海量用户行为数据，一方面为传统社会科学领域进行大样本调查研究提供了可能；另一方面，用户行为特征由后台自动记录获取因而具有客观性，并且包含许多通过线下行为实验无法获取的数据，通过信息技术获取的数据质量和数据广度将可能有所提高。

社交媒体用户行为是用户日常行为的一部分，以新浪微博为例，该平台的社交媒体账号下包含个人展示等自我表露信息和大量的文本语言信息。本研究拟采集用户在微博平台的行为数据，与其心理学特征——人格特质、自尊、主观幸福感结合起来，探讨通过社交媒体平台数据建模预测用户心理特征的方法。

1.2 课题背景及研究意义

1.2.1 研究背景

社交网络日益成为人们日常生活的一部分，用户在社交媒体上的自我呈现行为亦是其日常生活的延伸。自我呈现（Presentation of self）又称为自我表现，普遍存在于人际互动过程中。按戈夫曼的说法，自我呈现是努力展现自我从而使他人按照自己的愿望看待自己。从传播心理学的观点来看，互联网的自我呈现是一场“自我”的展览会，社交媒体平台鼓励一种“陈列”式的自我展示文化[1]，从其命名可见一斑，如你的 Tube（YouTube）、我的空间（Myspace）、脸书（Facebook）。这类平台从产品命名、交互设计、算法设计等方面针对用户的痛点做出了种种研究，极大鼓励用户在平台上展示自己、分享自己生活的方方面面，从而攫取由用户及用户之间社交网络关系带来的巨大流量。

在社交媒体平台，人们可以更有策略地进行自我呈现，有研究者指出社交媒体用户面对的是“想象的观众”[2]，呈现的内容是经过人们精心设计的“理想自我”或可以实现的“未来自我”[1]58。Liu 等人[3]的一项基于 MySpace 平台 12 万多份个人档案的研究发现，用户更倾向于在职业的选项中填写自己的理想职业和副业。但也有研究者认为，手机及其包含的社交媒体账号在塑造自我身份、提升主观幸福感的过程中逐渐被内化为用户自我延伸（Self-extension）的一部分[4]，与用户的情感需求、生活和记忆紧密相连，从而在拍照、社交互动、记录生活等交互的过程中影响着人们的自我概念[5]。一方面，社交媒体有助于帮助用户提升人际关系质量，同时为用户提供情感支持[6]；另一方面，被动进行社交浏览的社交媒体用户更容易感受到消极情绪，从而影响其心理健康[7]。

自我表露（Self-disclosure）是指个体将与自己相关的信息透露给他人，属

于自我呈现的一种，有助于建立深厚的亲密关系。通过自我表露，人们能够真实地展现自己，并通过他人的反馈认识到真实的自己是可以被他人接受的，从而提升幸福感，同时也打开了他人了解自己内心世界的大门，建立彼此之间的信任关系。随着关系深浅的不同，自我表露的程度往往也不同。在社交媒体时代，记录个人状态、上传照片、给喜欢的微博点赞等都属于自我表露的行为。我们在社交媒体上暴露的信息远比自己认为的要多。Kosinski 等人[8]在 2013 年的研究发现，可以通过 Facebook 的点赞数准确预测用户的个人隐私属性，比如用户的年龄、性别、性取向、种族、政治倾向和人格特质等。

自我表露与人格特质和心理健康状况密切相关，心理健康的人能够自由地进行自我表露，而不必害怕别人发现真实地自己，同时自我表露也能够提高心理健康水平。李林英和陈会昌[9]关于大学生自我表露与心理健康的研究表明，人们的自我表露程度与人格特质的外向性有着显著相关，自我表露程度低的个体往往比自我表露程度高的个体体验到更强烈的孤独感。因此可以合理推测，用户在社交媒体平台上进行自我呈现和自我表露的行为数据与其人格特质、心理健康状态等有着密切的联系。Kosinski 等人[8]5804 基于近 6 万 Facebook 用户的社交媒体数据建立了预测模型，研究结果发现该模型能够正确预测大五人格特质中的开放性维度，且预测准确度与标准化人格测验的重测效果相似。

网络社交媒体的自我呈现和自我暴露，在一定程度上反映了用户的心理状态，使得通过社交媒体数据建模预测用户人格等心理特征的方法成为可能。

1.2.2 研究意义及应用

人们的行为与其人格有着密切的联系。人格是源于个体身上的稳定行为方式和内部过程，一般认为人格具有跨时间、跨情境的稳定性[10]。我们可以预期，今天外向开朗的人，在明天也是外向开朗的；面对着同样的情境，不同的人会有差异性的反应方式。由于人格与行为的密切相关性，因此基于社交媒体用户的行为数据预测用户的心理特征有着广阔的应用前景。

用户画像是互联网领域频频出现的热门概念，是一种描述用户特征从而指导业务发展的工具，常见于互联网产品设计、数据分析、用户研究、运营等各个流程中。基于各大社交媒体平台用户画像的数据，目前已经被广泛应用于精准营销、广告投放、个性化推荐及其他领域，比如微信朋友圈出现的定向广告、豆瓣读书和电影基于好友喜欢内容的推荐、淘宝商品页面的“猜你喜欢”等。人格可以说是个体自我的用户画像。Rentfrow 等人[11]的研究表明，不同人格特质的用户对音乐有着不同的偏好。个性化推荐领域有协同过滤和关联规则两种

经典算法，其背后的原理究其根本就是认为人们和自己相似的人有着相似的喜好，以及人们对于某类事物的喜好具有稳定性[12]。

心理健康是公共医学和社会科学等领域重点关注的范畴，当个体的心理活动处于正常状态时，个体能够正常开展日常活动，当心理特征出现亚健康或者异常状态时，就应当引起注意和预警，从而及时采取干预手段。随着社交媒体平台的出现，用户越来越倾向于在社交网络进行自我呈现和自我表露，这为动态识别大众心理特征提供了可能。人们通过语言来表达自己内在想法和感受，语言的某些表达甚至具有跨文化的稳定性，因此，研究者开始寻找社交网络用户在社交媒体平台上的语言表达与心理特征之间的关联。Schwartz 等人[13]关于语言特征与人格关联的研究表明，人们表现出来的语言特征与其人格特质类型密切相关。

目前的研究已证明社交媒体特征能够预测人格特质，而作为底层心理特征的人格特质与主观幸福感、自尊密切相关。主观幸福感和自尊都是衡量心理健康状态的重要心理特征，其中自尊水平能很好地预测个体的心理健康和社会表现，主观幸福感在个人和社会层面都具有重大意义。相较于人格特质而言，主观幸福感和自尊水平对个体心理健康程度的检测更为直观。

如果能够对社交媒体用户的行为特征与心理特征之间联系建立模型，就可以实现实时计算大众的心理特征。由于人格、主观幸福感格、自尊等心理特征测评数据的收集费时费力，如果在保证施测精度和用户隐私的前提下，可以通过社交媒体平台等网络数据实现大范围、大规模地获取用户心理特征数据，将具有极大的社会价值。

1.3 国内外研究现状

1.3.1 国外研究现状

国外研究者早期多通过问卷收集心理特征数据，同时通过问卷对社交媒体的网络行为进行调查。随着技术手段的发展，大数据采集成为常见研究范式，基于大数据的心理学行为研究越来越受到研究者的关注，在情绪、人格特征、幸福感、种族研究等各个领域产出了一批具有现实意义的研究成果。Kramer 等人[14]基于 Facebook 平台用户的实验研究提出了社交媒体平台的情绪可通过情绪传染机制进行传播，研究发现处在社交网络环境的人们会无意识地体验到与好友相同的情绪状态。Yu 等人[15]分析了 2014 年世界杯期间的推特数据，发现用户发布推文的情绪与赛场上的实际情况相吻合。

Whitty 等人[16]基于 Twitter 和 Facebook 的研究发现，人格特质能够预测用户在社交媒体平台的个人主页页面进行自我展示的情况，其中，在尽责性维度得分较高的用户更愿意更改其主页的个人资料，在外向性维度得分较低的用户更可能将个人头像修改为自己的照片。Back 等人[17]通过自我报告、理想自我评价和旁观者评价三种方法收集了大五人格调查问卷数据，结果发现 Facebook 简介呈现的是用户的真实人格特征，而非理想自我，这表明 Facebook 账号所体现出来的特征与用户本人具有一致性。

社交媒体平台的用户生成内容种类非常多，包含着丰富的用户语料数据，用户的语言特征与人格的相关研究是一个热门领域。Schwartz 等人[13]8 收集了 7.5 万名志愿者在 Facebook 上发布的语言特征，并对志愿者进行了人格测试，发现用户所使用的语言在人格、年龄和性别方面有着很大的差异，其中，在外向性维度得分高的用户更经常使用社交活动相关词汇，在神经质性维度得分高的用户往往过度使用消极词汇。

前人关于 Facebook 和主观幸福感的研究结果没有达成一致。一些研究表明社交媒体使用与幸福感呈正相关，浏览个人 Facebook 主页能够引发积极的自我评价和增强自尊[18]，主动浏览朋友资料可以增加个体的积极情绪[19]，Facebook 的使用可以提升个体的主观幸福感[20]。同时有研究表明，社交网站中的自我呈现对个体的生活满意度有着显著的预测效果[21]。Kross 等人[22]的研究持相反态度，他们通过在线问卷调查用户的主观幸福感、自尊、抑郁及社会支持情况，结果发现相比于线下交流和电话沟通，Facebook 的使用可能会降低人们的当前感受和生活满意度，从而降低人们的主观幸福感。前人研究结果表现出差异的原因可能是由于存在多个变量的影响，比如自尊、抑郁水平、社会支持程度。

除了通过问卷调查网络行为和主观幸福感的关联以外，国外研究还流行使用基于社交媒体平台的大数据来测量、研究和提升主观幸福感。Luhmann[23]认为大数据测量主观幸福感虽然很方便，但仍然存在很多可改进之处，其数据质量目前还比不上传统的自我报告法。Kosinski 等人[8]5804 通过 Facebook 点赞情况来预测生活满意度，研究发现预测值与用户自我报告的生活满意度的相关性较低，交叉验证的相关系数只有 0.17。Seder 和 Oishi[24]的另一项研究将个人资料图片的微笑表情与自我报告的生活满意度相联系，取得了比前人更好的结果，相关系数为 0.34。其他数据源还包括利用智能可穿戴设备，基于位置、手机使用模式或者噪音、光线等环境信息进行实验，这种非自我报告类、对瞬时情绪的预测结果还不错，但是目前大都是小样本研究[23]30。

1.3.2 国内研究现状

国内关于社交媒体行为与心理特征的研究目前主要还是通过传统问卷的形式进行，通过用户自我报告的方式收集社交媒体行为数据，结果存在一定的主观性。利用社交媒体大数据开展的研究目前还不是很多。

国内利用社交媒体数据的研究起步较晚，社交媒体大数据与大众情绪是研究方向之一。Tao 等人[25]基于微博的相关研究发现，空气质量越差，人们在社交媒体上就谈论得越频繁，而当空气污染变得十分糟糕时，微博发布的频率反而趋平。乐国安和赖凯声[26]通过采集微博平台的文本数据，分析了快乐、悲伤、厌恶、愤怒和恐惧五种微博情绪对现实社会中热点事件的反应，结果发现微博情绪对重大事件、节假日的反应都呈现出了较为灵敏的特点，并基于以上研究更新了微博客情绪词库工具。

针对社交媒体数据和用户心理特征的研究还不算很多，较早的是 Bai 等人[27]在 2012 年开展的关于人人网与人格特质的研究，该研究根据用户发布的人人状态，从中提取行为特征并计算与用户大五人格特质之间的相关性，结果呈显著相关。中国科学院大学心理所的一批科研工作者借助网络进行大规模的心理健康预测，基于用户网络行为建立了心理健康预测模型。Bai 等人[28]在 2014 年的研究表明，新浪微博用户的使用记录能够客观地预测用户的大五人格维度。李昂等人[29]基于社会媒体用户的网络行为数据，建立并评估了基于社会媒体的用户人格、心理健康、主观幸福感的预测计算模型，研究结果显示主观幸福感的模型效果良好。田玮、朱廷劭[30]在 2018 年的最新研究通过建立多层神经网络模型，实现了基于微博文本来评估用户自杀的可能性，对公共领域的自杀预防做出了贡献。

张磊[12]1883 等人综合比较了历年来基于 Facebook 和 Twitter 平台的大五人格模型相关研究，总结了与五个维度人格特征相关的语言特征和非语言特征，发现外向性维度与社交平台朋友数呈正相关，高神经质的用户经常使用负面情绪单词和感叹号，而宜人性和尽责性与积极情绪相关。

总体而言，关于社交媒体特征和人格的研究并没有达成完全一致的结果。涉及主观幸福感、自尊等其他心理特征的研究目前还较少，国内已有的研究也仅仅是提取微博文本以外的社交网络行为特征用于建模预测，忽略了微博文本所蕴含的特征。在对微博文本的情感分析上，大多研究采用情感词典的方法，没有结合机器学习技术开展情感分析。另外，在建模方法的选择上，国内已有研究主要使用线性回归等传统方法建立心理特征的预测模型，在模型拓展方面

还有很大的探索空间。

1.4 本文的主要研究内容

本章通过对课题来源和研究背景的阐述，明确了社交媒体与心理特征交叉领域的研究意义，通过对国内外研究现状的分析，综合对比了已有研究的不足之处，从而并确定了本文的研究目标：建立社交媒体用户行为特征数据与用户心理特征数据之间的联系。

本文将重点对社交媒体文本内容进行情感值提取，从用户特征、微博行为特征和微博情感特征三个方面构建特征数据集，对用户心理特征进行建模预测，以及分析。目前主流的用户心理特征建模主要为人格特质，在此基础上，本研究拟加入其他两个重要的、对心理健康有预测价值的心理特征——主观幸福感和自尊，在已有模型和技术的基础上，探索性研究能否将人格领域的模型成果推广到其他心理特征的建模工作中。

基于上述总体研究目标，本文将其拆分成四个方面的研究内容，具体阐述如下：

（1）本研究所关注的三个取向的心理特征是什么、有何理论基础和现实意义？前人做过哪些研究？在前人研究中是否已经形成合理、规范的研究范式？本研究需总结前人在社交网络、社交媒体用户心理特征领域的研究方法，形成一套可操作的、规范的实验设计方案。

（2）本研究需要哪些数据、所需的数据如何获取？对于本研究来说，关注的数据主要分为自变量和因变量，其中因变量特征为用户心理特征，需要编写问卷、发放问卷、回收问卷、剔除无效回答、计算心理特征得分等步骤获取。用于建模的自变量则为社交媒体平台属性的特征，包含的维度众多，需要选择有解释意义的特征变量、编写 Python 网络爬虫程序、抓取所需的自变量数据并存储入 MySQL 数据库中。

（3）如何进行文本预处理和情感分析？数据采集完毕后，利用 MySQL 数据库的 SQL 语句对问卷数据、用户在微博平台的行为数据进行分值计算，剔除数据中无用的元素，形成清理的数据。对于微博文本数据，编写 Python 程序，基于网络上已有的情感词典进一步进行构建，调用 jieba 中文分词接口对文本进行初步的分词。目前常见的情感分析方法主要分为情感词典和机器学习两种，本研究将综合对比不同模型的效果，并选择最优模型进行情感分析。最终，再次通过 MySQL 数据库的 SQL 语句，将微博维度的特征聚合成用户维度的特征，作为后续建模和数据分析的完整数据集。

（4）根据社交媒体平台的用户行为特征数据，进行特征构造和选择，建立用户心理特征的预测模型。本研究拟采用经典的线性回归、Logistic 回归模型，并在此基础上引入决策树、随机森林等模型，为心理特征建模方法提供参考。

第 2 章 社交网络与心理特征

2.1 基本概念介绍

2.1.1 大五人格模型

人格（Personality）是个体身上的稳定行为方式，是个体内部一系列复杂而独特的心理过程，其具有跨时间、跨情境等特点。可以说，人格影响着个体的特征性行为模式，其中既包括内隐行为，也包括外显行为。人格理论是对个体人格的假设性说明，其中大五人格模型理论的应用最为广泛，其理论结构得到了普遍的验证。

大五人格模型（Big five factor model）是人格特质领域最主流、稳定的模型，主要从开放性、尽责性、外向性、宜人性以及神经质性这五个维度来建立人格特征模型，从而全面描述个体的人格特质。关于这五个维度的特质描述如表 2-1 所示。多年来，研究者们开发了许多专业量表来测量大五人格特征。

表 2-1 大五人格因素[31]

人格维度	对应的特征
神经质性（Neuroticism）	烦恼——平静
	不安全感——安全感
	自恋——自我满意
外向性（Extraversion）	喜欢社交——不喜欢社交
	爱娱乐——严肃
	感情丰富——含蓄
开放性（Openness）	富于想象——务实
	寻求变化——遵守惯例
宜人性（Agreeableness）	自主——顺从
	热心——无情
	信赖——怀疑
	乐于助人——不合作
尽责性（Conscientiousness）	有序——无序
	谨慎细心——粗心大意
	自律——意志薄弱

大五因素在跨时间、跨群体上表现出很好的稳定性。其中的尽责性维度在实践应用中被认为是预测员工工作绩效的最佳指标[10]108。在神经质性维度上得分高的人更容易因为日常生活的压力而感到烦躁、体验到更多的消极情绪，在神经质性维度上得分低的人自我调适能力良好，大多表现为平静，不易出现极端情绪反应[10]101-103。关于大五人格与其他心理特征的研究众多，其中研究表明，大五人格能够很好地预测主观幸福感，其中，外向性和低神经质的预测效果更好[23]28。

需要注意的是，对大五人格自陈式量表各个维度的解读不能仅仅看测评的分数，一方面要将个体的得分放在更广大的人群中进行比较才有意义，另一方面行为不仅仅是由人的内在特征所决定，需要综合考虑人和情境的交互作用。总体来说，大五人格理论为人格特质和个体行为之间的联系提供了有说服力的模型。

2.1.2 主观幸福感

主观幸福感（Subjective well-being）是评价者根据自定的标准对其生活质量的整体性评估，是一种重要的心理健康指标。一般认为，主观幸福感的认知因素为个体的生活满意度，情绪体验因素包括个体体验到的积极情绪和消极情绪。这两个基本维度的情绪在情感性、强度和表达上并不相同，并具有个体差异。稳定体验到积极情绪的个体表现为随和、镇静，更乐于参加社交活动，同时也更容易对关系感到满意；而稳定体验到消极情绪的个体则常常伴随着压力、紧张、焦虑和愤怒等体验。综合而言，主观幸福感可以理解为个体对生活的满意程度并伴随相关的情绪体验。在研究方法上，一般可以通过自陈式问卷、词汇应用、面部表情和他人评价等方法来研究个体情绪。

影响主观幸福感的因素有生活事件、人格和收入等。其中，仅仅近三个月发生的事件对主观幸福感有影响。杨秀君和孔克勤[32]的研究表明，人格特质中的外向性和神经质性与主观幸福感的情感成分显著相关，且其对主观幸福感的认知成分的影响受情感成分的调节。主观幸福感中的积极情绪体验可以导致促成更好的交际关系，而爱好交际正是外向性特征的特征[33]。

社会支持（Social support）是维持个体心理健康的一个重要影响因素，它指的是人们感受到的来自他人的关心、尊重和支持，包括情感支持、资产支持和信息支持等形式，能够帮助人们更好地应对压力。研究表明，社会支持与个体的主观幸福感呈正相关，具有良好社会支持的个体表现为高生活满意度和积极情感、较低的消极情感[34]。

2.1.3 自尊

自尊（Self-esteem）是对一个人自我的概括性评价，也是自我理论的一部分。如果对自己持有积极的自我概念，一般称之为高自尊，而如果对自己持有消极的自我概念，一般称之为低自尊，低自尊可以被描述为对自我的偏低肯定。自尊被认为与人格特质存在紧密的联系，且对学业、工作表现、应对挫折的能力等有着良好的预测效果，低自尊个体可能在抑郁、物质滥用、行为过失等问题上表现得更脆弱，而高自尊个体更倾向于乐观和体验积极情绪，同时也会在面临威胁时表现出更多的敌意和不友好行为[10]206。

网络中的自我呈现有两种策略，一种是积极自我呈现，个体在网络中选择性地呈现积极正面的个人信息；一种是真实自我呈现，即个体按在自我表露的过程中会选择如实呈现的方式，这种策略往往更为深入。牛更枫等人[35]的研究表明，社交网站中的积极和真实的自我呈现对自尊有显著预测效应，其中，真实的自我呈现还能在社会支持的中介效应下对自尊产生影响。

2.1.4 微博用户行为与心理特征的相互作用

综合以上基本概念的阐述，可以得出微博的自我呈现行为和情感特征与自尊和主观幸福感相互作用的关系，如图 2-1 所示。

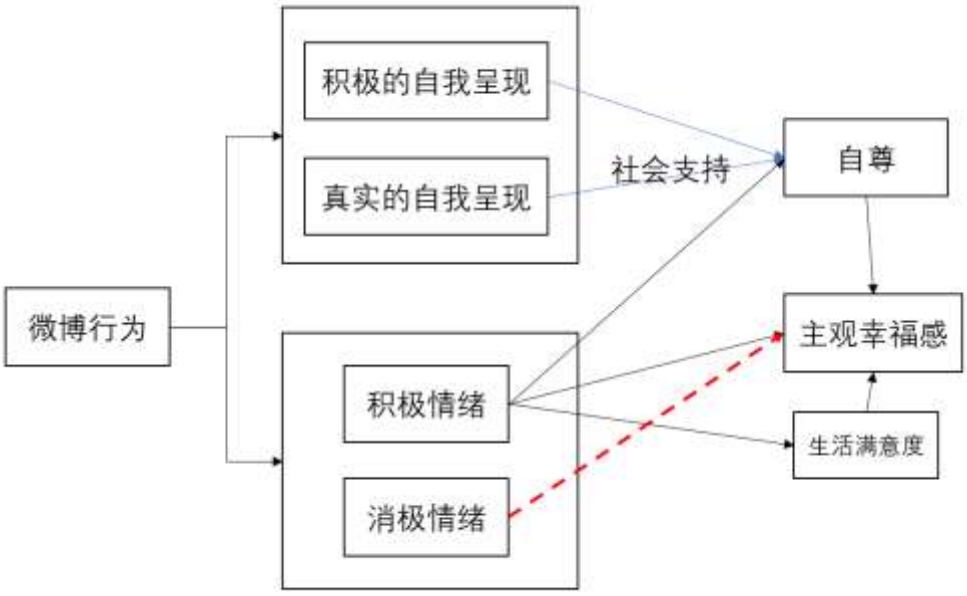


图 2-1 微博用户行为与心理特征的相互作用

微博用户在社交媒体上的用户行为主要包括自我呈现的特征和其表达的情绪、情感的特征，通过直接作用或社会支持的中介作用间接影响着用户的自尊、主观幸福感等心理特征。通过微博用户行为与心理特征的相互作用模型，可以初步确定用户在微博的自我呈现行为和情感数据是预测其心理特征的关键特征，为后续数据集收集和特征处理做了理论准备。

2.2 社交媒体用户的心理特征预测

国内外基于社交媒体平台的社会心理学探索工作主要集中在人格特质、主观幸福感等方面。社交媒体用户的人格分析与预测是大数据心理学的热门方向，在国内外目前已有许多研究成果。Gosling 和 Mason[36]的研究表明，通过社交媒体平台在线问卷收集的数据质量也很高，和传统自我报告法所获取的数据质量不相上下，且数据种类更为多样。

2.2.1 数据采集方法

要想建立人格预测模型，需要将用户通过填写心理测评量表所获得的人格特征属性与其在社交媒体中的数据关联起来。虽然获取用户在社交媒体属性的数据在技术上很容易实现，但是在涉及到用户心理特征的研究中通常存在用户隐私的研究伦理问题，需要用户进行授权才能获取，因此存在一定的困难。

除了邀请用户填写问卷并授权外，现有研究还常采用第三者阅读社交媒体平台的信息后进行主观评价的方法，或者使用公开数据集。Facebook 的一个应用 MyPersonality 是目前最大的公开数据集，已有 750 万用户参与填写人格问卷，在提交问卷的同时授权采集其个人社交媒体账号相关信息。

国内研究以中国科学院心理研究所为代表，他们开发了在线实验平台来采集新浪微博用户的数据，通过将心理测量问卷系统接入新浪微博 API，可以实现大规模的用户心理特征和社交媒体网络行为特征的数据收集。

2.2.2 研究范式

Golbeck 等人[37]调查了 55 名推特用户，收集用户填写的大五人格问卷，并采集用户最近 2000 条或所有的公开推文、粉丝数、关注数、@用户数、评论数和话题数等个人页面公开信息，通过 LIWC（Linguistic inquiry and word count）和 MRC Psycholinguistic Database 对推特文本进行处理，采取皮尔逊相关分析研究推文特征与用户大五人格特质的直接相关性，最后建立回归模型预测用户人格特质得分。此后社交网络人格预测的研究范式基本与此相似，通过发放问卷

标记人格特质类型，爬取用户社交网络数据，对文本进行分词和处理，最后建立模型对人格特质进行预测。

中国科学院心理研究所申请了一项专利，提出了基于微博用户行为的人格预测方法：（1）获取微博活跃用户 id，通过私信功能向被试用户发放在线填写的人格问卷并收集心理特征数据；（2）下载填写问卷的用户的微博数据，从中提取微博行为特征并形成特征数据集；（3）使用逐步回归算法进行特征选择，建立人格预测回归模型。

关于主观幸福感的研究已在国外研究现状中进行了阐述，一般研究范式与人格预测类似，目前存在的问题是对主观幸福感的认知成分——生活满意度的预测模型效果不佳。基于社交媒体数据预测自尊的研究目前还没有看到。

人格特质与生活的方方面面相关，可以说是一个人心理特征的底色，而更为表层的主观幸福感、自尊则与心理健康有着密切联系。随着互联网和社交媒体的发展，越来越多的人在社交媒体展示自己的生活、表达自己的情感，若能通过社交媒体数据对用户心理健康相关的心理特征建立预测模型，将会对抑郁识别、自杀识别及后续干预等社会工作产生很大的价值。

2.2.3 存在的问题

随着社交媒体用户行为数据被广泛采集并应用于商业和科研各领域，用户的个人隐私保护也迎来了巨大的挑战。从用户的角度来看，这些社交媒体平台为自己的社交需求提供了产品和服务；而从另一个角度——社交媒体平台来看，它可以不费劲地招募“免费劳工”，为自己的产品全天候、全方位地产生数据，并从中获取商业利益。一方面，用户在社交媒体平台的数据虽然是公开的，但是不经用户授权而直接爬取数据依然可能存在着法律问题。另一方面，一旦用户心理特征预测模型的有效性得到验证，那些原本属于用户个人隐私范畴的个人心理特征数据将很可能被应用于商业领域。未来的方向可能是推动相关法律条款的修改，以保护用户在社交媒体网络的个人数据隐私，在政府、互联网运营商、研究者和用户的共同努力下，限制数据的使用途径和场景，提供最大限度的保护措施。

除了用户隐私的隐患之外，大数据获取心理特征等数据存在着用户偏差的局限性。能够上网、经常发布微博的用户本身就是个有偏群体，根据微博的官方数据，微博平台的 16-25 岁用户群体占比高达 61%。在大数据时代，如何做到科学取样、尽可能覆盖科学研究的目标群体，也对科研人员提出了新的挑战。

目前国外的大多数研究都使用 Twitter 或 Facebook 的数据，国内的研究集

中在微博，除了微博之外，还有相当大的群体可能分布在 QQ、抖音、微信等社交媒体平台，未来的研究应该考虑扩大数据来源。

2.3 本章小结

本章首先介绍了本研究关注的三种心理特征——大五人格特质、主观幸福感和自尊，并介绍了大五人格特质的五种维度和主观幸福感的认知和情感两种成分。结合已有研究，将微博平台的行为划分为自我呈现行为和包含情感的行为（微博正文等），构造了用户的微博行为和主观幸福感、自尊之间的相互作用机制，为后续数据处理和特征构建作了理论准备。另外，本章还总结了前人研究中关于社交媒体用户心理特征预测的研究方法，包括数据采集方法和经典研究范式，并反思了社交媒体用户数据收集和心理特征预测相关研究中存在的问题。

第 3 章 数据收集与预处理

3.1 实验设计框架

本文的研究内容主要是通过抽取社交媒体用户行为特征建立用户心理特征预测模型，对用户的大五人格、主观幸福感、自尊这三个取向的心理特征进行预测，为大数据时代的用户心理健康预测提供合理、可行的研究方法，同时也希望对社交媒体平台的发展提供建议。基于以上目标，本文的研究方法部分主要包括三个方面的核心问题:（1)社交媒体平台的用户可以获取哪些特征？（2）如何对微博文本进行情感分析？（3）获取数据后，如何基于社交媒体平台的用户行为特征对其心理特征进行建模？

根据本文的研究目标及三个核心问题，确定了以下实验设计框架，如图 3-1 所示。首先，研究者通过私信的方式向微博用户发放调查问卷，收集用户的心理特征数据。在用户填写问卷并对微博数据用途进行知情同意后，通过 Python 网络爬虫采集用户的微博数据，形成用户的社交媒体平台行为数据，并对采集的微博文本内容进行情感分析。最后，将问卷数据和微博数据相关联，得到最终的数据集，从而进行建模工作。

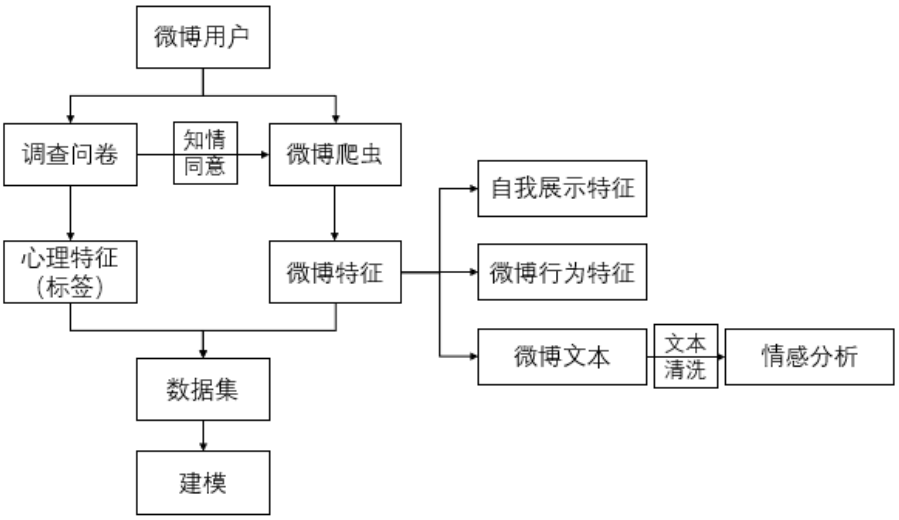


图 3-1 实验设计框架

3.2 数据收集

社交媒体用户在使用社交网络的过程中产生了大量的数据，这些数据相对于传统的心理学问卷数据而言更容易获取。在业界，用户数据常常被用于业务、商业相关领域，比如推荐系统、关联规则等常见算法的应用。如果能通过社交媒体属性的用户数据建立合理的心理特征预测模型，将会极大拓展心理学等社会科学领域的研究范围。

本研究的用户特征数据分为两部分：（1）问卷层面的用户心理特征，包含人格、主观幸福感、自尊在内的数据；（2）微博层面的用户行为特征，即用户在微博平台交互过程中产生的记录数据。其中，心理特征属性的数据可以通过填写调查问卷获得，微博属性的数据则可以通过爬虫获取。

3.2.1 问卷设计与发放

在传统心理学领域，心理特征可以通过自我报告法、行为测量和观察法等方式进行测量，其中自我报告法主要包括自陈式问卷和访谈两种形式。大五人格特质的评估技术已经发展多年，形成了 NEO-PI-R、BFQ、BFI（Big-five inventory）等成熟量表，在正常群体的测量中表现出很好的信效度。NEO-PI-R 等人格量表通常包含几百道题目，耗时久，且需要付费或授权后才能使用。John 等人于 1991 年建构并不断修订的 BFI 量表是公开的，已被翻译成多国语言在不同的文化下广泛应用于研究领域。BFI 通常使用的版本共计 44 个项目，在大五人格模型的特质形容词基础上编制而成，使用短语式的问题来评估五个人格维度。因此，从施测时间、内容覆盖度和信效度等方面综合考虑，BFI 目前已经成为最为通用的大五人格模型测评工具[38]。

主观幸福感的测量工具整体生活满意度量表、夏普量表、情绪平衡量表、正性负性情绪量表等。Diener 等人在 1985 年编制的主观幸福感问卷包括整体生活满意度、积极情感频率、消极情感频率三个分量表，具有很好的信效度，题目清晰，应用广泛。自尊的测评工具包括 Rosenberg 自尊量表和 Coopersmith 自尊量表，其中前者题目简单明了、测评方便，后者则包含多维度的自尊测评。Rosenberg 自尊量表是测量外显自尊应用最广泛的工具，该量表已被证明具有良好的信效度。

综上，本研究所使用的问卷结构如下：

（1）BFI 大五人格问卷[39]：44 个条目，采用 5 点计分，对大五人格模型的五个维度分别进行测量。其中外向性、尽责性维度题目各计 8 道，宜人性、尽

责性维度题目各计 9 道，开放性维度题目共计 10 道。

（2）主观幸福感问卷[40]：19 个条目，采用 7 点计分。包括整体生活满意度题目 5 道、积极情感频率题目 6 道、消极情感频率题目 8 道。

（3）Rosenberg 自尊量表[41]：10 个条目，采用 4 点计分。其中，5 道题正向记分，5 道题反向记分，总的分数相加来衡量自尊的指标。在正向陈述的题目上做肯定评定，同时在反向陈述的题目上做否定评定的，属于积极的自尊得分。最后的分值越高，表明自尊水平越高。

（4）筛选题：共计 3 道，用于问卷质量的筛选，测试用户填写时的注意力是否集中、是否认真答题。

本研究筛选发布微博数大于 300 条的用户，发放调查问卷，用户在提交问卷前已被告知本研究将通过爬虫获取其微博数据，并进行匿名处理。通过筛选题剔除不认真的回答者后，最终回收的有效问卷共计 70 份。经过计算，问卷数据包括大五人格 5 个维度的数据、主观幸福感 3 个维度的数据和自尊 1 个维度的数据，共计 9 个维度的因变量标签数据。

3.2.2 爬虫数据获取

本研究主要采用 Python 爬取新浪微博数据，将结果信息写入 MySQL 数据库进行存储。目前主要有两种方式来爬取微博数据：

（1）基于微博所提供的 API 接口的方法。微博提供了官方 API 接口，API 使用一套非常标准的规则生成数据，使用 get 请求获取数据，用 JSON 或 XML 格式表示数据。这种方法的缺点是微博系统会对爬虫请求的限制次数和速率，如果爬虫速度过快，很容易被系统限制，只能等待一段时间直到限制自动解除。

（2）基于微博网页解析的方法。每当网页代码有改动，响应的抓取方式也必须要随之改变。另外，如果要爬取大量的数据，账号或 IP 地址容易被封，需要使用多个不同账号、代理 IP 等方式来破解微博的反爬虫机制，此方法存在一定的难度。

鉴于本研究需要爬取的用户量不大，因此主要采取第一种方式进行爬虫，针对微博用户主页的 HTML 页面进行爬取。首先，通过 Python 的 requests 模块请求网页，获取网页中的 JSON 数据。然后通过设置 sleep time，加入随机等待机制模拟真实用户的操作，从而降低被微博系统限制的风险。最后通过 etree 模块解析网页中的文本内容，并将获取的数据写入 MySQL 数据库。

写入数据库的信息主要有用户信息和微博信息两大类。用户信息主要包含用户昵称、简介、关注数、粉丝数、微博数等；微博信息主要包含微博正文内

容、发布时间、发布工具、评论数、点赞数、转发数等。其中，微博正文内容将是后续数据预处理的重点对象。

本研究的爬虫数据部分在 2020 年 3 月 25 日至 3 月 28 日期间完成，共爬取 70 名微博用户的数据，其中微博正文共计 107094 条。

3.2.2.1 用户信息

用户信息的基本特征主要包括用户简介特征、用户用于自我展示的特征和微博使用情况特征。

（1）用户简介特征：用户在注册微博过程中主动填写或自动产生的信息，包括用户 id、性别、生日、所在地（省、市）等。

（2）自我展示特征：用户在使用微博期间主动设置的信息，包括用户昵称、用户简介、教育经历、公司、是否半年可见等。

（3）微博使用情况特征：用户在使用微博期间的行为记录特征，包括微博数、关注数、粉丝数、注册时间、阳光信用、用户微博等级、会员等级、是否认证等。

3.2.2.2 微博信息

微博信息主要以用户的单条微博为单位，包括微博发布行为特征、微博正文、社交网络特征。

（1）微博正文信息：用户发布微博产生的文本信息和行为记录信息，包括正文文本内容、发布时间、发布工具、是否包含图片或视频。

（2）微博正文：用户在发布原创微博或转发他人微博所产生的文本内容。其中蕴含着用户特有的情感特征，用于情感分析的具体方法将在下一章节进行阐述。

（3）社交网络特征：用户在发布微博前后与他人互动的信息，包括微博@用户数、微博话题（如“#带着微博去旅行#”）、点赞数、评论数、转发数。

3.3 数据预处理

由于爬虫爬取的微博数据包含用户信息和微博信息两部分，其中微博正文是非结构化的文本数据，带有用户的情绪特征，需要对文本进行情感分析。同时，微博正文文本包含特殊表情符号、网页链接等很多噪声，需要进行预处理，才能更为准确地计算出每条微博的情感倾向。

对微博正文文本的处理方式一般包括文本正则化、分词、停用词过滤等，预处理的具体流程如图 3-2 所示。

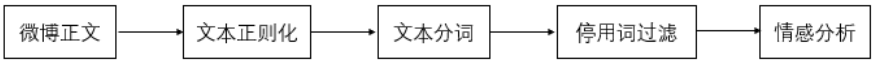


图 3-2 数据预处理流程

3.3.1 文本正则化

微博正文文本是用户微博账号的核心内容。本研究中，通过数据清洗对微博正文的文本内容进行预处理，去除可能对情感分析结果有影响的字符和无关信息，从而为特征构造和建模提供支持。

本研究利用 Python 的 re 模块调用正则表达式，通过匹配指定字符串，对转发微博去除“//@”以后的内容，仅保留用户自己发布的内容。另外，微博正文常常含有“@用户名”的内容，但是大部分微博用户的用户名毫无规律性，对分词有较大影响，因此需要剔除。微博中还经常包含活动自动生成的内容，经观察，这类微博一般含有 url、微博话题 tags 和特定句式，比如“我在 XXX 活动中……”。因此可通过文本正则化进行清理，从而达到剔除噪声数据的目的，示例如表 3-1 所示。

表 3-1 文本正则化

原微博文本（包含噪声）	正则化处理后的文本
我最喜欢的龙猫和哈尔的移动城堡!!@Neverland123//@静临 for:哇…//@动漫酱:了不起！	我最喜欢的龙猫和哈尔的移动城堡!!
我就是想闭上眼睛，好好睡一觉。忘记所有烦恼。http://t.cn/A6ZlSDUs（《Rain》@QQ 音乐）	我就是想闭上眼睛，好好睡一觉。忘记所有烦恼。（《Rain》）

微博文本可能包含一句话或者一段话，本研究通过正则化功能识别位于句末的标点符号，如“？”、“！”、“。”等，将每一条微博分割成单独的句子，以便于下一步进行文本分词。

3.3.2 文本分词

文本分词是指将一个文本序列按照一定的规范切分成单独的有意义的词的过程。对于英文文本，通常以空格将字与字隔开成最小单元，而中文文本通常使用词作为最小的情感单元。微博文本的情感分析首先要以中文分词为前提，常见的中文分词方法主要有两种：基于字符串匹配和基于统计及机器学习的方

法。目前流行的开源中文分词工具主要有 jieba、NLPIR、IKAnalyzer 等，此外还包括 n-gram 等模型。jieba 分词是一款开源在 GitHub 上的中文分词工具，提供了 Python、java 等多种编程语言的接口，而且能够识别繁体字。NLPIR 汉语分词系统是中国科学院计算机研究所开放的产品，提供中文分词、词性标注和用户词典功能。IKAnalyzer 是一款开源的 java 分词工具，结合了词典分词和语法分析算法的分词组件。本研究中将采用开源的 jieba 进行分词。

jieba 分词基于随机条件场和隐马尔可夫来进行分词，支持四种分词模式：精确模式、全模式、搜索引擎模式和 paddle 模式。其中，精确模式分词最适合应用于文本分析场景，可以将句子切割成最精确的单位；全模式分词的优点在于能快速扫描出所有可以组成词的词语，并将其全部列出，缺点是不能解决歧义；搜索引擎分词模式一般应用于搜索引擎分词的场景中，主要是在精确模式的基础上进一步切分长词；paddle 模式利用 PaddlePaddle 深度学习框架训练模型实现分词，同时支持词性标注。综合考虑，本研究选择精确分词模式，同时利用 jieba 提供的自定义分词词典功能，将常用微博词语、情感词典添加进自定义分词词典，以提升 jieba 对文本情感词的分割和提取准确度。情感词典的构建过程将在下一章节的情感词典部分进行阐述。分词处理结果如表 3-2 所示。

表 3-2 文本分词

正则化处理后的文本	分词处理后的文本
我最喜欢的龙猫和哈尔的移动城堡!!	我/ 最/ 喜欢/ 的/ 龙/ 猫/ 和/ 哈尔/ 的/ 移动/ 城堡/ !/ !
我就是想闭上眼睛，好好睡一觉。忘记所有烦恼。（《Rain》）	我/ 就是/ 想/ 闭上眼睛/ ，/ 好好/ 睡一觉/ 。/ 忘记/ 所有/ 烦恼/ 。/ （/ 《/ Rain/ 》/ ）

3.3.3 停用词过滤

停用词（Stop words）是指对在日常语言表达中经常使用、但对文本中表达的意义并不起什么作用的词语，包括人称代词、助词、副词、介词、连词等。在搜索引擎优化（Search engine optimization, SEO）中，搜索引擎在搜索时会自动忽略停用词，以节省存储空间和提高检索速度。对于搜索引擎来说，这类词语无法保证准确的搜索结果，还会降低效率。在情感分析中也是如此，为了提高情感分析模型的准确性，本研究也需要对微博文本进行停用词过滤。

本研究采用的是 Github 自然语言处理项目中常用的开源停用词词表，包含

中文停用词、表情、符号等。通过 jieba 分词结果查看词频，发现名人、明星的名字存在错分情况；对于转发行为，用户在转发行为中存在仅转发、没有个人表达的情况。因此，手动将常见名字、“转发微博”添加进停用词表中。另外，网络用语中的感叹号经常有表达程度的含义，因此将感叹号从停用词表移除。停用词过滤处理结果如表 3-3 所示。

表 3-3 停用词过滤

分词处理后的文本	过滤停用词后的文本
我/ 最/ 喜欢/ 的/ 龙/ 猫/ 和/ 哈尔/ 的/ 移动/ 城堡/ !/ !	最/喜欢/哈尔/城堡/!/!
我/ 就是/ 想/ 闭上眼睛/ ，/ 好好/ 睡一觉/ 。/ 忘记/ 所有/ 烦恼/ 。/ （/ 《/ Rain/ 》/ ）	想/闭上眼睛/好好/睡一觉/忘记/烦恼/Rain

3.4 本章小结

本章主要介绍了情感分析和建模前的准备工作，包括问卷数据收集、爬虫数据获取、微博文本预处理等内容。

数据获取的核心在于通过编写 Python 网络爬虫程序，设置随机等待时间来模拟真实用户的操作，以避免因爬虫速度过快而被系统限制。通过爬虫获取的微博初始数据集经过数据处理步骤后，将与用户的心理特征数据进行匹配，从而对心理特征进行建模。

数据预处理的部分详细介绍了文本数据清洗的具体过程，通过正则表达式功能、jieba 分词工具和停用词词典等步骤，对微博正文文本内容进行降噪和清洗，目标是为情感分析步骤准备符合要求的高质量文本数据集。

第 4 章 微博文本情感分析

中文文本情感分析常见应用于电子商务、社交网站网络评论等领域，比如京东商品评价分析等。文本通过分词和停用词过滤后，被切割成一个个字或词语，为后续的情感分析做准备。目前情感分析技术主要有两种方法，其中一种是基于情感词典的方法，在已有情感词典的基础上对经过分词处理的词汇进行打分，从而得到文本的情感值。另一种方法是基于机器学习的情感分析方法，首先将分好的词转换为词向量，输入到支持向量机、朴素贝叶斯、LSTM 等模型，通过大量标注好情感倾向标签的文本语料训练模型，从而输出测试集语料的情感特征。

本章节将对比情感词典、机器学习这两种情感分析技术，通过 Github 公开的带情感标签的微博文本数据集进行训练和测试，选择效果更好的模型，从而对本研究所获取的微博文本进行情感分析。对于最终选取的情感分析模型，以微博用户为单位，统计每个用户正向、负向情感倾向的微博占比，作为情感特征添加到后续模型中。

4.1 基于情感词典的情感分析

本部分使用基于情感词典的方法，对微博文本进行情感分析。首先构建一个较为完善的情感词典，其中包含程度级别词语、否定词、正向词语、负向词语。其次，基于微博文本分词结果中各类词语出现的统计频次，建立公式计算单条微博文本对应的情感值。在上述步骤基础上，使用 Github 公开的带情感标签的微博文本数据集进行训练和测试，以评测情感词典模型的性能。

4.1.1 情感词典构建

由于微博文本兼具日常用语表达规范与网络用语的特性，本研究基于情感词典的情感分析采用否定词词典、知网情感词典、BosonNLP 情感词典和弹幕多维情感词典相结合的方式。否定词词典采用 Github 自然语言处理项目中常用的词典。知网情感词典是指由知网发布的情感分析用词语集（beta 版），包含中文、英文情感分析用词，分为程度级别词语、正面评价和正面情感词语、负面评价和负面情感词语等类别，主要覆盖了常用的规范文本表达。其中否定词、程度级别词语需要与正、负面情感词语结合起来使用，共同判定文本的情感值。知网的程度级别词语分为六种类别，本研究人为地根据其强烈程度的不同进行赋

值，比如对表示程度强烈加强的词语“百分之百”、“极其”赋值为 3，对表示程度减弱的词语“稍微”、“不怎么”赋值为 0.6。

BosonNLP 情感词典共包括 114767 个词语，对正面、负面词语分别进行不同大小的数值标注，作为词语表达情感强烈程度的大小。该词典的优势在于其语料是从微博、论坛、新闻等社交媒体网站采集分析而成，涵盖很多网络用语和非规范文本，适合用于社交媒体情感分析。然而，BosonNLP 词典中有很多过时表达和无意义表达，其情感值的参考价值有限，在本研究中，主要利用该词典所覆盖的正面、负面词语，而不采用这些词语对应的情感标注。

微博作为社交媒体网站，其用户所使用的语言表达是不断发展变化的，弹幕评论是网络用语的常见聚集地，因此将网络上常见的弹幕多维情感词典添加至本研究的情感词典中作为补充。弹幕多维情感词典在情感维度的分类上与上述两种词典有所不同，郑飏飏等人[42]主要采用 7 分类的方法，将弹幕语料分为乐、好、怒、愁、惊、恶和惧。在本研究中，人为地将弹幕维度的乐、好两种词语划分为正面情感类别，将怒、愁、惊、恶和惧五种词语划分为负面情感类别，分别添加至正面情感词典和负面情感词典中。

至此，本研究的情感词典已构建完毕，包括程度级别词语、否定词、综合多个词典而成的正向情感词语和负向情感词语。

4.1.2 情感词典模型

4.1.2.1 情感值计算方法

在文本分词和情感词典的基础上，需要将微博文本所蕴含的用户情感特征转换为可以用数值表示的情感值。关于如何计算中文文本情感值，主要有两种方法，一种是基于词汇语义相似度进行计算，另一种是基于统计情感词出现的频数进行计算。本研究主要采取第二种方法。在前人的研究中，基于统计情感词出现的频数计算情感值也有不同的方法：统计单条文本中各类情感词的数量，正负求和从而量化情感值[43]；将单条文本中所有情感词的情感值进行平均作为情感值[44]；考虑文本中否定词、双重否定词的出现情况对情感词的影响，从而判段单条文本的情感极性[45]等。

本研究综合考虑前人的研究成果，通过编写 Python 程序实现判断和匹配情感词、否定词、程度级别词语，综合计算情感值。首先，逐一匹配单条微博词语是否在情感词典中出现，并判断其情感属性是正面还是负面，若为正面情感词，则计 1 分，若为负面情感词，则计-1 分。其次，逐一匹配该情感词前后是否出现否定词，若某情感词前出现奇数个否定词，则在情感词分值的基础上乘

以-1，若出现偶数个否定词，则在情感词分值的基础上乘以 1。然后，逐一匹配该情感词前后是否出现程度级别词语，若出现某一程度词，则在该情感词分值的基础上乘以程度级别词语对应的赋值权重。最后，将单条微博文本所计算出来的所有情感值求和，作为单条微博的最终情感值。

本研究的情感值计算方法如公式（4-1）所示。其中，m 表示单条微博中的情感词个数，i 表示单条微博中第 i 个情感词，E_i 表示第 i 个情感词的分值（取值为 1 或-1）；N_i 表示第 i 个情感词与前一个情感词之间的否定词个数，W_i 表示第 i 个情感词与前一个情感词之间的程度级别词语的权重。

$$wb\_emotion\_score = \sum_i^m (-1)^{N_i} W_i E_i \tag{4-1}$$

4.1.2.2 情感词典模型的效果

在上述步骤的基础上，可以建立情感词典模型，为微博正文文本标注对应的情感值。在正式实验之前，为了测试该情感词典模型的准确率，使用 Github 常用的公开数据集《weibo2018》（https://github.com/dengxiuqi/weibo2018）进行测试实验，该数据集共包括 10500 条带情感标注的微博文本语料数据，其中正向情感的微博文本有 5841 条，负向情感的微博文本有 4659 条，正向、负向文本语料的比例约为 5:4。将该数据集按照 7:3 的比例划分为训练集和测试集，其中训练集和测试集中正向、负向文本语料的比例约保持为 5:4。训练集的数据将用于后续基于机器学习的情感分析模型训练，测试集的数据将用于情感词典、机器学习模型的效果评估。

本研究的情感词典模型在测试集上的准确率为 64.7%，其他评价指标如表 4-1 所示。

表 4-1 情感词典模型的效果

label	precision	recall	f1-score
负向	0.65	0.47	0.54
正向	0.65	0.79	0.71

测试结果表明，对实际情感标签为正向情感的微博文本，该情感词典模型判别的召回率为 0.79；而对实际情感标签为负向情感的微博文本，该情感词典模型判别的召回率仅为 0.47。

可以看出，基于情感词典的情感分析模型能够更好地判断出正向情感的文本，而对负向情感文本的识别能力较差。这与语言表达的复杂性有很大关系，

人们在日常生活和社交媒体中，对负面情绪的表达有着多种方式和技巧，比如可以通过反讽、衬托、阴阳怪气等方式进行表达，这就给情感分析技术应用于实际场景带来了一定的困难。

总体来说，情感词典模型的优点是可以根据公式（4-1）计算出每条微博文本的情感值，在准确率达标的前提下，可以很方便地区分微博文本所表达的情感强度，能够为情感分析结果提供更多信息。然而基于情感词典的情感分析方法非常依赖于人力，需要根据文本特点人为地调整情感词典所包含的词汇，耗时耗力，且实验效果不一定是最优的。因此，接下来将与基于机器学习的方法进行对比，选择准确率最高的模型进行情感分析。

4.2 基于机器学习的情感分析

基于机器学习的情感分析方法大多属于监督学习，需要对数据进行情感标注，其分类效果取决于模型选择以及数据集的质量。人工标注的过程非常耗时耗力，在特定领域的文本标注还需要很多专业知识。这种方法主要通过选取情感词作为特征词，将文本进行向量化，从而利用机器学习算法进行文本情感分类。微博短文本需要词向量转化，相似的词会有相似的向量表示，从而方便挖掘文本中词语和句子之间的隐藏特征。好的词向量表示可以提升分类器的性能。词向量的构造方法主要包括 TF-IDF、BiGram、TriGram、Word2vec、One-hot 等。

4.2.1 支持向量机

支持向量机（Support vector machine, SVM）是一种分类器，常用于处理二分类决策问题，如文本分类、图像识别等领域。其原理是通过求解一个凸二次规划最优化问题，寻找分类的最优超平面，从而将分类间隔最大化。当训练样本线性可分时，支持向量机通过硬间隔（Hard margin）最大化，习得一个线性分类器，不存在分类错误；当训练样本近似线性可分时，通过软间隔（Soft margin）最大化允许支持向量机在一些样本点上出错，从而训练分类器；当训练样本线性不可分时，支持向量机通常利用核函数（Kernel function）将数据从低维空间映射到高维空间，使得样本在这个特征空间内线性可分，从而将在低维空间中的非线性问题转化为高维空间下的线性问题求解。常用的核函数包括线性核、多项式核、高斯核、拉普拉斯核以及 Sigmoid 核等等，也可以将这些核函数组合使用，以达到最优线性可分的效果。

在分类决策模型中，支持向量机的学习能力和泛化能力比较好，在数据集上应用最基本的支持向量机分类器就可以得到错误率较低的分类结果。其优点

是泛化错误率低，计算开销小，结果容易解释，可以很好地处理高维数据集。支持向量机的缺点在于对参数调节和核函数的选择较为敏感，如果选择的核函数不合适，则可能将样本映射到了一个不合适的特征空间，从而导致性能不佳。另外，支持向量机对噪声是很敏感的，且其原始分类器仅适用于处理二分类问题。

4.2.1.1 生成词向量和句向量

生成词向量的方法包括基于统计学的方法和基于神经网络的语言模型方法等，目前已经有很多成熟的词向量模型。Word2vec 词向量主要是通过词的上下文得到词的向量化表示，包括通过上下文预测目标词的 CBOW 方法和通过目标词预测上下文的 Skip-gram 方法。TF-IDF 是一种基于统计的加权方法，分为词频（Term frequency，TF）和逆文件频率（Inverse document frequency，IDF）两部分，计算公式分别如公式（4-2）和公式（4-3）所示。在公式（4-2）中，$n_{i,j}$是某个词语在文档D_j中出现的次数，分母则是文档D_j中所有词语出现的次数总和。在公式（4-3）中，$|D|$是语料库中的文档总数，$\left|\{j: t_i \in D_j\}\right|$表示包含某个词语$t_i$的文档数，log 项的分母加 1 是为了避免该词语不在语料库中从而导致分母为 0 的情况。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4-2}$$

$$idf_i = log\frac{|D|}{|\{j: t_i \in D_j\}|+1} \tag{4-3}$$

TF-IDF 可以根据字词在文本中出现的频数和在整个语料中出现的文档频率来计算一个字词在整个语料中的重要程度，从而过滤掉一些常见却无关紧要的词语，保留影响整个文本的重要词语。其主要思想是，如果某个词语在文本中出现的频率高，并且在其他文档中很少出现，则认为该词语具有很好的类别区分能力。

FastText 是一个快速文本分类的算法，也可以用于训练词向量，将文本转化成向量用于后续的自然语言处理任务。FastText 最大的特点是模型简单，只有一层的隐层以及输出层，因此训练速度非常快。

本部分采用 FastText 结合 TF-IDF 的方法生成词向量和句向量，具体实现过程如下：首先用 FastText 训练词向量，只保留出现次数大于 3 次的词语；用每个词的 TF-IDF 作为权重，对 FastText 生成的词向量进行加权，最多只保留 TF-

IDF 最高的前 30 个词；得到表征每个句子的向量，用于训练支持向量机模型，从而对微博文本进行情感分类。

4.2.1.2 模型效果

对于支持向量机模型，同样使用 Github 常用的公开数据集《weibo2018》进行模型的训练和测试。实验结果发现，本研究的支持向量机模型在测试集上的准确率为 77.7%，其他评价指标如表 4-2 所示。

表 4-2 支持向量机模型的效果

label	precision	recall	f1-score
负向	0.79	0.67	0.73
正向	0.77	0.86	0.81

测试结果表明，和情感词典模型一样，支持向量机模型在情感标签为正向的微博文本中分类表现更好。支持向量机模型对微博文本的情感分类整体表现优于情感词典模型。

4.2.2 朴素贝叶斯

贝叶斯分类是一类以贝叶斯定理为基础的分类算法的总称，其中朴素贝叶斯（Naive Bayes）的应用最为广泛，常用于垃圾过滤、文本分类等场景。其分类原理是通过某个对象先验概率和贝叶斯公式计算出其后验概率，即该对象属于某一类别的概率，选择具有最大后验概率的类作为该对象所属的类别。朴素贝叶斯方法的一个假设是，对于已知类别所有特征之间相互条件独立，即一个特征或词语出现的可能性与其是否与其他单词相邻没有关系，这个假设可以有效地降低满足概率分布所需的样本数。朴素贝叶斯的另一个假设是，每个特征同等重要。虽然这两个假设在实际中并不一定正确，但是朴素贝叶斯的实际效果却往往很好。

朴素贝叶斯模型具有稳定的分类效率，在数据较少的情况下仍然有效，能够处理多分类任务，适合增量式训练，即可以实时地对新增的样本进行训练。另外，朴素贝叶斯算法的计算速度快，即使使用大规模数据集也能具有较高的速度。朴素贝叶斯的缺点是对输入数据的表达形式较为敏感，在某些时候会由于假设的先验模型的原因导致预测效果不佳。

4.2.2.1 独热编码进行向量转换

与其他方法相比，朴素贝叶斯并没有对高质量词向量的需求，因此直接用

One-hot 编码进行文本特征提取。One-hot 编码又称独热编码，属于词袋模型（Bag of words）的一种，其方法是对语料库中的每个词进行编码，假设在一个语料库集合中一共有 n 个不同的词，则可以使用一个长度为 n 的向量表示每一个词，若第 i 个词出现，则向量位置 i 处的值为 1，其他位置的值均为 0。同理，可以使用 One-hot 对句子进行向量化表示。One-hot 编码不考虑词与词之间的顺序，并且假设词与词相互独立，最终得到离散稀疏的特征词向量，因此存在维数过高、矩阵稀疏、不能保留语义等缺点。

4.2.2.1 模型效果

本研究的朴素贝叶斯模型在测试集上的准确率为 80.5%，查准率、召回率等其他评价指标如表 4-3 所示。

表 4-3 朴素贝叶斯模型的效果

label	precision	recall	f1-score
负向	0.80	0.74	0.77
正向	0.81	0.85	0.83

测试结果表明，朴素贝叶斯分类模型的效果相比前两个模型而言表现更好。同样地，该模型在情感标签为正向的微博文本中分类表现更好。

4.2.3 基于朴素贝叶斯的情感分析实验

综上所述，基于机器学习的情感分析方法在评价指标上的效果普遍高于基于情感词典的情感分析方法，其中，基于朴素贝叶斯算法的情感分类模型效果最佳。因此，最终选取朴素贝叶斯模型作为本研究的情感分类模型，对本研究收集的微博文本数据进行情感分析。实验结果如表 4-4 所示。

表 4-4 微博正文情感倾向占比

情感倾向	负向	正向
所占比例	37.8%	62.2%

在用户层面，根据以上分类标准可以计算出单个用户的“负向”和“正向”这两种类型的微博数占比，同时统计微博文本中第一人称、其他人称代词出现的频率，经过计算，作为该用户微博的情感特征加入到数据集中。

4.3 特征处理

经过以上数据收集、预处理、情感分析的过程，完成了本研究的数据集构建。在用于建模前，还可以对已经获取的所有特征进行加工，构建新特征，并对部分特征进行 log 函数变换和标准化等处理，以方便分析。

最终获取的微博数据可以划分为以下维度的特征用于建模，用户层面的特征如表 4-5 所示。

表 4-5 用户特征

分类	特征	计算/获取方式
用户层面特征（用户简介、自我展示）	性别	爬虫获取
	所在地	通过所在地字段计算用户在几线城市
	年龄	通过生日字段计算
	阳光信用	爬虫获取
	用户微博等级	爬虫获取
	用户会员等级	爬虫获取
	注册天数	通过注册时间字段计算
	用户昵称长度	通过用户昵称字段计算
	用户简介长度	通过用户简介字段计算
	隐私-开放程度	通过是否填写生日、所在地、教育、公司信息和是否半年可见字段计算

微博层面的特征包括微博行为特征和微博情感特征，具体内容和计算方式分别如表 4-6、表 4-7 所示。

表 4-6 微博行为特征

分类	特征	计算方式
微博行为特征（使用情况、发布情况）	微博数	爬虫获取
	关注数	爬虫获取
	粉丝数	爬虫获取
	爬虫微博数	由于部分用户设置了半年可见，因此一些特征需要用爬虫微博数进行构造

表 4-6（续表）

分类	特征	计算方式
微博行为特征（使用情况、发布情况）	爬虫时间范围（天）	爬虫获取的最后一条微博时间减去第一条微博时间
	爬虫原创微博数	爬虫获取+MySQL 计算
	爬虫点赞数	爬虫获取+MySQL 计算
	爬虫评论数	爬虫获取+MySQL 计算
	爬虫转发数	爬虫获取+MySQL 计算
	发布图片的微博数	爬虫获取+MySQL 计算
	发布视频的微博数	爬虫获取+MySQL 计算
	活跃度（每天发布微博数）	微博数/注册天数、爬虫微博数/爬虫时间范围
	原创微博总字数	通过微博正文字段计算
	单条原创微博字数	原创微博总字数/爬虫原创微博数
	原创微博占比	爬虫原创微博数/爬虫微博数
	爬虫微博占比（可反映用户隐私情况）	爬虫微博数/微博数

表 4-7 微博情感特征

分类	特征	计算方式
微博情感特征	负向微博占比	负向微博/用户爬虫微博数
	正向微博占比	正向微博/用户爬虫微博数
	提到“我”的微博占比	提到“我”的微博/用户爬虫微博数
	提到其他人称的微博占比	提到其他人称的微博/用户爬虫微博数
	提到“我”的微博中正向微博占比	提到“我”的正向微博/提到“我”的微博
	提到“我”的微博中负向微博占比	提到“我”的负向微博/提到“我”的微博
	提到其他人称的微博中正向微博占比	提到其他人称的正向微博/提到其他人称的微博

表 4-7（续表）

分类	特征	计算方式
微博情感特征	提到其他人称的微博中负向微博占比	提到其他人称的负向微博/提到其他人称的微博

4.4 本章小结

本章主要介绍了建模前的情感分析工作，对比分析了目前常见的技术：基于情感词典的方法和基于机器学习的方法。其中，基于情感词典的方法主要是通过利用已有公开数据集和词典，构建完备的情感词典，对文档分词，找出文档中的情感词、否定词以及程度副词，最终计算出每条微博文本的情感值。本研究中基于机器学习的方法主要包括支持向量机、朴素贝叶斯这两种算法，其中基于支持向量机的情感分析算法使用 FastText 生成词向量，结合 TF-IDF 生成句向量，从而用于训练情感分类模型；基于朴素贝叶斯的情感分析算法使用 One-hot 编码对微博文本进行向量转换。通过对比情感词典模型、支持向量机模型、朴素贝叶斯模型的准确率等评价指标，最终确定使用朴素贝叶斯方法对微博文本进行情感分类，提取出用户对应的微博情感特征。

在以上过程获取的数据集的基础上，对已有特征进行合理加工，形成用户特征、微博行为特征和微博情感特征的标准化数据集，为后续数据分析和建模作了准备。

第 5 章 数据分析与结果

本研究的三个核心内容是：获取用户在微博平台的行为特征和用户心理特征；对微博文本进行情感分析；利用社交媒体平台的用户行为数据对用户心理特征进行建模。前两个问题已在前两个章节得到了解决，本章主要讨论如何建立预测模型。

在社交网络用户人格预测领域，已有研究表明经典的机器学习模型就可以取得较好的预测效果[12]1886，比如多元线性回归、支持向量机、逻辑回归、随机森林等模型，使用多元线性回归模型与其他复杂的模型差别不大[46]。由于问卷获取的用户心理特征数据为连续变量，且数据量不大，本研究主要通过经典的线性回归模型进行建模预测；此外，亦可将用户心理特征按其数值大小转换为类别型变量，将本研究转化为分类问题。在分类算法方面，本研究通过 Logistic 回归、决策树、随机森林三种方法建立分类模型，对用户心理特征的高低得分类别进行预测。本研究通过 Python 的 sklearn.model_selection 模块，使用 train_test_split 功能按 7:3 的比例将数据集随机划分为训练集和测试集。

5.1 多元线性回归模型

5.1.1 多元线性回归

线性回归是最基础的回归模型，涉及到两个或两个以上自变量的线性回归问题被称为多元线性回归，可用公式（5-1）来表示因变量 y 与自变量 x_1，x_2，…，x_k 以及误差项ε之间的一般关系：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{5-1}$$

其中，β_0，β_1，β_2，…，β_k 是多元线性回归模型的参数，也称为回归系数。

本研究在数据处理部分完成了特征提取工作，共得到 34 个特征变量，可用于建立预测模型。这些特征变量彼此之间可能存在着相关性，容易导致多重共线性。为了更好地训练模型，需要对特征变量进行降维处理，减少冗余信息和噪音信息。主成分分析（Principal component analysis, PCA）是本部分所使用的降维方法，其原理是将样本点在新的低维空间中的超平面上进行投影，若能够使样本点地投影尽可能分开，就能够实现降维。因此，需要使投影后样本点的方差最大化，这也是主成分分析的优化目标，从而通过降维达到去噪的效果。

经过 PCA 降维后的特征数为 10 个，它们的方差占比为 97%。

5.1.2 模型结果

回归模型的评价指标主要包括绝对平均误差（Mean absolute error, MAE）、均方误差（Mean squared error, MSE）、均方根误差（Root mean squared error, RMSE）、多重判定系数（R^2）。本研究主要使用 MSE 和 R^2 这两个指标对多元线性回归模型在测试集上的表现效果进行评价。其中，MSE 的计算公式如（5-2）所示，R^2 的计算公式如（5-3）所示。

$$MSE = \frac{1}{n}\sum_i^n w_i(y_i - \hat{y}_i)^2 \tag{5-2}$$

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2} \tag{5-3}$$

通过多元线性回归模型预测微博用户大五人格特质的五个维度，其结果如表 5-1 所示。

表 5-1 多元线性回归预测大五人格特质的模型结果

评价指标/心理特征维度	MSE	R^2
外向性	0.63	0.22
宜人性	0.11	0.38
尽责性	0.16	0.61
神经质性	0.39	0.17
开放性	0.36	0.54

通过多元线性回归模型预测微博用户主观幸福感的三个维度，其结果如表 5-2 所示。

通过多元线性回归模型预测微博用户的自尊得分，结果如表 5-3 所示。

由上述评价指标可以看出，多元线性回归模型在大五人格取向的尽责性维度和开放性维度有一定的解释效果，该模型分别能够解释 61%的尽责性维度和 54%的开放性维度的取值变差；对神经质性等其他维度的大五人格取向、主观幸福感、自尊这些心理特征的总体预测效果还不够理想。

表 5-2 多元线性回归预测主观幸福感的模型结果

评价指标/心理特征维度	MSE	R^2
整体生活满意度	1.36	0.20
积极情感频率	1.45	0.27
消极情感频率	0.96	0.25

表 5-3 多元线性回归预测自尊的模型结果

评价指标/心理特征维度	MSE	R^2
自尊	0.53	0.37

5.2 Logistic 回归模型

虽然用户的心理特征得分为连续变量，但是其得分的绝对值没有实际意义，必须将用户的得分放在常模群体中进行比较，才能产生价值。常模（Norm）是一组在标准化条件下接受测试的群体表现的概要总结，通常包括参照群体的平均数和标准差，以及如何把原始分数转换为百分等级的有关信息。在实际应用中，很多情况下并不需要对每一个个体都预测出准确的心理特征得分，只需要判断其是否属于危险或边缘群体即可。由于本研究的样本量不大，通过线性模型预测心理特征得分的效果有待提升，因此将心理特征变量转换为离散型数据，从而将本研究转为分类问题，利用分类方法建立预测模型。

根据问卷填写结果，参与本研究的 70 名微博用户的平均年龄为 22 岁。本研究所用的 BFI 大五人格问卷提供了美国群体的常模数据作为参照，可以将用户得分与美国 22 岁常模群体的得分进行比较，若用户在大五人格特质的某一维度得分高于常模得分平均值，则将其划分为该维度的高分组，反之，则划分为该维度的低分组。对于主观幸福感和自尊取向的数据，目前没有找到成熟、一致的常模数据作为参考，因此采用本研究用户在这些心理特征上的得分平均值作为参考进行分组，根据问卷填写结果，计算用户在这些心理特征上的得分平均值，若用户在某一维度的得分高于平均值，则将其划分为该维度的高分组，反之，则划分为该维度的低分组。

5.2.1 Logistic 回归

机器学习中包含很多分类方法，如 Logistic 回归、支持向量机、随机森林等，每种分类方法都有其优缺点。本部分采用经典的 Logistic 回归方法建立分类模型，对用户心理特征得分高低的二分类进行预测。线性回归模型对数据的分布有较高的要求，需要假定因变量与自变量之间具有线性关系，且误差项是一个服从正态分布、期望值为 0 的随机变量。而 Logistic 回归模型则克服了线性模型的这些缺点，可以在无需假设数据分布的前提下直接对分类可能性进行建模，从而避免了数据不服从假设分布所带来的问题。

二项 Logistic 回归模型的条件概率分布如公式（5-4）和公式（5-5）所示。其中，x 是自变量或输入变量，Y 是取值为 0 或 1 的因变量或输出变量，w 和 b 分别为权值向量参数和偏置参数。

$$P(Y=1|x)=\frac{exp(w\cdot x+b)}{1+\exp(w\cdot x+b)} \tag{5-4}$$

$$P(Y=0|x)=\frac{1}{1+\exp(w\cdot x+b)} \tag{5-5}$$

根据二项 Logistic 回归模型的条件概率分布公式，可以求得 Y 的两个类别的近似概率，通过比较概率值的大小，将输入变量划分到概率值更大的那个类别，从而实现分类的目标。

在本研究的分类模型中，利用 AIC（Akaike information criterion）准则结合向前逐步回归法进行特征筛选。AIC 可以作为衡量模型拟合优度和复杂度的标准，其公式如（5-6）所示，其中 k 表示进入模型的自变量个数，L 为似然函数。

$$AIC=2k-2\ln(L) \tag{5-6}$$

向前选择法的起点是从模型中没有自变量开始，逐步将自变量一个个添加入模型中，直到增加自变量不能导致模型的残差平方和显著减小为止。在建立模型前通过对自变量进行一定的筛选，去掉不必要的噪音变量，这样，不仅使建立模型变得更容易，也使模型更具有可操作性，也更容易解释。

5.2.2 模型结果

回归模型常见的评价指标包括精确率（Precision）、召回率（Recall）、F-score、

准确率（Accuracy）、ROC（Receiver operating characteristic）曲线和 AUC（Area under ROC curve）等。本研究主要通过准确率和 AUC 指标来评估分类模型的效果。其中，准确率是指所有样本被准确分类的比例，可以衡量模型的整体分类能力。AUC 被定义为 ROC 曲线下的面积，其取值通常在 0.5 到 1 之间，AUC 越接近于 1 则说明模型效果越好。

利用 Logistic 回归模型预测微博用户大五人格特质五个维度的分类，得到的结果如表 5-4 所示。

表 5-4 Logistic 回归预测大五人格特质的模型结果

评价指标	Accuracy	AUC
外向性	0.62	0.69
宜人性	0.62	0.81
尽责性	0.52	0.55
神经质性	0.67	0.58
开放性	0.71	0.73

经过向前逐步法进行特征选择后，最终对用户每个心理特征维度分别建立了预测模型，通过变量筛选后在模型中留下的自变量来预测心理特征。如表 5-5 所示，研究结果发现，微博用户大五人格特质的外向性维度与该用户在微博上的隐私-开放程度（open_to_others）呈显著正相关，与用户发布的微博数（statuses_count）和单条原创微博字数（yc_len_1wb）呈正相关，但没有达到显著水平；外向性维度与用户的注册天数（regst_day）和活跃度 2（active2，即爬虫微博数/爬虫时间范围）呈显著负相关。

表 5-5 Logistic 回归预测大五人格特质-外向性的模型结果

	coef	std err	z	P>|z|
open_to_others	2.15	0.86	2.52	**0.01***
regst_day	-7.53	3.02	-2.49	**0.01***
active2	-0.96	0.44	-2.17	**0.03***
statuses_count	3.12	1.67	1.87	0.06
yc_len_1wb	4.13	2.80	1.48	0.14

微博用户大五人格特质的宜人性维度与爬虫微博占比（spider_pct，可反映用户隐私情况）呈显著正相关，如表 5-6 所示。

表 5-6 Logistic 回归预测大五人格特质-宜人性的模型结果

	coef	std err	z	P>|z|
spider_pct	2.62	1.21	2.18	**0.03***

如表 5-7 所示，尽责性维度与用户的粉丝数（followers_count）、活跃度 1（active1，即微博数/注册天数）、所在地（location_level，即一线城市用户尽责性更高）、正向微博占比（pos_pct）呈显著正相关，与评论数（comments_count）呈正相关但不显著；与用户的负向微博占比（neg_pct）、会员等级（mbrank）呈显著负相关，与发布视频的微博数（video_wb_num）呈负相关但不显著。

一般认为，尽责性高的个体表现为自律、谨慎、负责、成就感，可以猜想这类个体倾向于维持工作和生活的条理性，并能持之以恒，间接导致了较少的消极情绪体验和较高的积极情绪体验。

表 5-7 Logistic 回归预测大五人格特质-尽责性的模型结果

	coef	std err	z	P>|z|
neg_pct	-60.55	19.24	-3.15	**0.00****
followers_count	7.75	3.22	2.41	**0.02***
active1	2.79	1.20	2.31	**0.02***
mbrank	-1.59	0.59	-2.67	**0.01***
location_level	1.94	0.93	2.09	**0.04***
pos_pct	16.60	5.54	3.00	**0.00****
video_wb_num	-3.48	1.89	-1.84	0.07
comments_count	2.09	1.57	1.33	0.19

神经质维度则与原创微博占比有一定的负相关，但是没有达到统计意义的显著水平，如表 5-8 所示。

表 5-8 Logistic 回归预测大五人格特质-神经质性的模型结果

	coef	std err	z	P>|z|
yc_pct	-1.90	1.52	-1.25	0.21

本研究结果还显示，开放性维度与用户发布的微博数（statuses_count）呈显著正相关，与提到其他人称的微博占比（u_pct）呈正相关但不显著；与活跃度 2（active2，爬虫微博数/爬虫时间范围）呈显著负相关，与关注数（follow_count）

呈负相关但没有达到显著水平，如表 5-9 所示。

表 5-9 Logistic 回归预测大五人格特质-开放性的模型结果

	coef	std err	z	P>|z|
u_pct	4.49	2.82	1.59	0.11
statuses_count	4.80	1.76	2.71	**0.01***
active2	-1.26	0.48	-2.62	**0.01***
follow_count	-2.64	1.43	-2.85	0.06

利用 Logistic 回归模型预测微博用户主观幸福感三个维度的分类，得到的结果如表 5-10 所示。

表 5-10 Logistic 回归预测主观幸福感的模型结果

评价指标	Accuracy	AUC
整体生活满意度	0.57	0.66
积极情感频率	0.52	0.55
消极情感频率	0.67	0.65

结果表明，微博用户的整体生活满意度与用户的微博等级（urank）有显著的负相关，如表 5-11 所示。这说明对生活满意度高的用户使用微博的时间和投入度更少，可能反映了其在线下生活更为充实。

表 5-11 Logistic 回归预测主观幸福感-整体生活满意度的模型结果

	coef	std err	z	P>|z|
urank	-2.30	1.12	-2.06	**0.04***

在考察主观幸福感的情感维度方面，实验结果发现主观幸福感的积极情感频率与用户的关注数呈负相关但不显著，如表 5-12 所示。

表 5-12 Logistic 回归预测主观幸福感-积极情感频率的模型结果

	coef	std err	z	P>|z|
follow_count	-1.56	1.01	-1.56	0.12

主观幸福感的消极情感频率与用户提到其他人称的微博中负向微博占比（u_neg_pct）呈显著正相关，与单条原创微博平均字数（yc_len_1wb）和用户

简介长度（description_lenth）呈一定的负相关但不显著，如表 5-13 所示。

表 5-13 Logistic 回归预测主观幸福感-消极情感频率的模型结果

	coef	std err	z	P>|z|
yc_len_1wb	-4.73	2.48	-1.91	0.06
u_neg_pct	6.86	2.79	2.45	**0.01***
description_lenth	-1.92	1.03	-1.86	0.06

利用 Logistic 回归模型预测微博用户自尊得分高低的分类，得到的结果如表 5-14 所示。

表 5-14 Logistic 回归预测自尊的模型结果

评价指标	Accuracy	AUC
自尊	0.62	0.84

研究结果发现，微博用户自尊水平的高低与评论数呈显著正相关，而与用户的活跃度 2（active2，爬虫微博数/爬虫时间范围）、粉丝数（followers_count）呈显著负相关，如表 5-15 所示。也就是说，自尊水平高的用户拥有更多来自社交网络关系中他人的评论数，同时每天发布的微博数量相对于自尊水平低的用户来说更少。前人关于孤独感的研究也许能够提供一定的解释效果：孤独与人格特质、自尊相关，孤独者的自尊得分更低，缺乏良好的亲密关系和社交关系，且更容易出现健康问题[10]203。研究表明，情绪表达性强的个体能够在自尊上的得分要显著高于情绪表达性低的个体[10]134，这为今后关于自尊与社交媒体用户行为的关系提供了可能的研究方向。

表 5-15 Logistic 回归预测自尊影响因素的模型结果

	coef	std err	z	P>|z|
active2	-1.27	0.48	-2.65	**0.01***
comments_count	3.28	1.14	2.89	**0.00***
followers_count	-3.37	1.17	-2.89	**0.00***

综合来看可以发现，微博用户的心理特征与其发布微博的情感特征、微博活跃程度或使用微博的投入程度、粉丝数或评论数等社交网络互动特征以及其他行为特征有着一定的联系。

5.3 决策树

5.3.1 决策树

决策树（Decision tree）算法既可以应用于分类问题，亦可以应用于回归问题，被广泛应用于金融、电商、医疗等各个行业。其本质是通过训练集中的经验信息，按照树状结构生成多个分支条件进行判断，最终输出结果。

在生成决策树的过程中，会经历构造和剪枝两个阶段。其中，构造的过程就是从训练集中选择合适的特征作为树的节点的过程，包括最开始的根节点，中间经历的内部节点，以及对应决策结果的叶节点。剪枝的过程主要是为了防止过拟合，去除太多的分支，从而提高模型的泛化能力。剪枝可以分为预剪枝和后剪枝。

在决策树的构造过程中，选择根节点可以通过纯度和信息熵这两个指标来进行判断。其中，纯度是指让目标变量的分歧达到最小；信息熵表示信息的不确定度，当不确定性越大时，其中所包含的信息量就越大，对应的信息熵也就越高。

根据纯度—不纯度的不同，构造决策树的方法可以分为三种：

（1)信息增益,通常也称为 ID3 算法,其目标是通过划分带来纯度的提高，使信息熵下降。通过计算不同特征作为根节点时每个节点的信息增益，将信息增益最大的节点作为父节点，从而得到纯度高的决策树。其优点是计算方法简单，缺点是对噪声敏感，容易产生分类错误。

（2）信息增益率，即 C4.5 算法。在 ID3 算法中，取值多的属性往往带来更多的信息增益，因此更倾向于被选择。C4.5 采用信息增益率的方法，通过将信息增益除以其属性熵。当某个属性有许多取值时，虽然信息增益变大了，但是其属性熵相对也会变大，从而避免了 ID3 算法的问题。在决策树构造之后，C4.5 采用悲观剪枝这种后剪枝方法，通过递归来估算每个内部节点的分类错误率并在简直前后进行比较，从而决定是否剪枝。另外，C4.5 还会离散化处理具有连续值的属性。但是 C4.5 算法的效率相对较低。

（3）基尼指数，也就是 CART（Classification and regression tree）算法，其全称是分类回归树，即既可以作为分类树，又可以作为回归树。在前两种算法中，可以生成二叉树或者多叉树，而 CART 算法只支持二叉树。基尼系数在经济学中被用来衡量一个国家内部收入差距，也可以用来反应样本的不确定度。基尼系数越小，说明样本之间的差异越小，包含较低的不确定度，也证明样本更稳定。因此 CART 算法在构造分类树时会选择基尼系数最小的属性作为划分。

另外，CART 算法主要通过代价复杂度这种后剪枝方法进行剪枝。

基尼系数的计算公式如（5-7）所示。其中$p(C_k|t)$表示节点 t 属于类别 C_k 的概率，这个公式表明，节点 t 的基尼系数等于 1 减去其所属各类别概率的平方和。

$$GINI(t) = 1 - \sum_k [p(C_k|t)]^2 \tag{5-7}$$

在本研究中，主要基于 CART 算法来构造分类树。

5.3.2 模型结果

利用决策树预测用户大五人格特质的结果如表 5-16 所示。相对于 Logistic 回归模型的结果而言，决策树在对外向性、尽责性、神经质性三个维度的预测效果有所提升。

表 5-16 决策树预测大五人格特质的模型结果

评价指标	Accuracy	AUC
外向性	0.76	0.63
宜人性	0.64	0.68
尽责性	0.71	0.71
神经质性	0.71	0.65
开放性	0.73	0.70

利用决策树预测用户主观幸福感的结果如表 5-17 所示，决策树模型在主观幸福感的三个维度上的整体表现优于 Logistic 回归模型的结果。

表 5-17 决策树预测主观幸福感的模型结果

评价指标	Accuracy	AUC
整体生活满意度	0.71	0.73
积极情感频率	0.67	0.68
消极情感频率	0.71	0.77

利用决策树预测微博用户自尊得分高低的分类，得到的结果如表 5-18 所示。相比于 Logistic 回归模型的性能有一定的提升。

表 5-18 决策树预测自尊的模型结果

评价指标	Accuracy	AUC
自尊	0.76	0.76

5.4 随机森林

5.4.1 随机森林

基于决策树算法还诞生了许多其他不同的算法，随机森林（Random forest）就是其中一种 Bagging 集成学习方法，具有容易实现、效率高等特点，并且性能强大。随机森林以多个决策树为基学习器对样本进行训练，这些决策树之间没有关联，并且都会参与分类决策，以多数分类投票作为最终的输出结果。

随机森林的随机性主要包括随机抽样和随机子集两个方面。在训练过程中，随机森林采用自主抽样法从训练集中有放回地随机重复采样，并根据每个采样的训练集训练树模型。通过在不同的样本中训练每棵树，从而降低树模型之间的关联性。尽管每棵树可能对特定训练集具有高方差，但是总体平均而言，整个随机森林将具有较低的方差。

在决策树的基础上，随机森林进一步引入随机特征选择，即从基决策树的每一个节点的特征集合中随机选择一个子集，再从中选择一个最有特征用于划分节点。因此，这种随机性经常导致随机森林起始性能相对较差，但是随着学习器的增加，它又表现出更强的泛化能力。

一般而言，随机森林在数据集上表现良好，易于并行化，能够处理特征很多的高维度数据，而无需做特征选择。

5.4.2 模型结果

利用随机森林预测大五人格特质的结果如表 5-19 所示，该模型在外向性维度的预测效果较好。整体来说，该模型的表现不如决策树模型。

利用随机森林预测用户主观幸福感的结果如表 5-20 所示。该模型在消极情感频率上预测的评价指标表现较好，整体优于 Logistic 回归模型，但是效果不如决策树模型好。

利用随机森林预测微博用户自尊得分高低的分类，得到的结果如表 5-21 所示。随机森林模型在自尊这个心理特征的预测上表现最佳。

表 5-19 随机森林预测大五人格特质的模型结果

评价指标	Accuracy	AUC
外向性	0.71	0.60
宜人性	0.57	0.58
尽责性	0.62	0.63
神经质性	0.57	0.60
开放性	0.67	0.66

表 5-20 随机森林预测主观幸福感的模型结果

评价指标	Accuracy	AUC
整体生活满意度	0.62	0.61
积极情感频率	0.53	0.55
消极情感频率	0.71	0.69

表 5-21 随机森林预测自尊的模型结果

评价指标	Accuracy	AUC
自尊	0.76	0.79

5.5 本章小结

本章的主要内容是在已有的经典模型基础上，对用户行为数据集和心理特征进行建模，首先采用经典的多元线性回归模型和 Logistic 回归模型，接下来引入决策树和随机森林模型，分别从回归和分类的角度进行数据分析，并对比模型对各个心理特征维度的准确率和 AUC 值，评估模型效果。

在多元线性回归模型中，采用主成分分析的方法对自变量特征进行特征选择；在 Logistic 回归模型中，采用基于 AIC 的向前选择逐步回归对自变量特征进行筛选。在分类模型方面，对已有研究的常用方法进行拓展，加入了决策树、随机森林模型进行对比。

最后分别输出回归和分类模型的结果，并结合用户心理特征和微博行为特征对结果进行了分析和解释。结果发现，用户的心理特征与其微博的情感特征密切相关。

结　论

本文以微博为例，将社交媒体用户的心理特征预测作为研究内容，对微博行为特征和用户心理特征进行建模，在已有社交网络用户人格预测研究范式的基础上加入了主观幸福感和自尊这两个与心理健康密切相关的心理特征。

本研究主要完成的工作有：

（1）提出了微博平台的自我展示、情绪与用户的主观幸福感、自尊等心理特征交互作用的模型，并在此基础上进行实验框架设计和数据分析工作，有机地将爬虫、情感分析、机器学习等技术应用于心理学领域。

（2）完成了数据的收集、预处理和情感分析工作。通过编写 Python 网络爬虫程序，收集了填写问卷的微博用户对应的大量微博特征，包括但不限于：用户层面的用户自我展示特征、用户简介特征、微博行为特征和微博情感特征，以及用户对应的大五人格特质、主观幸福感、自尊等心理特征。利用 MySQL 和 Python 的正则化模块进行文本清洗，利用 jieba 中文分词工具提供的 Python 接口对微博文本进行分词处理。在情感分析技术方面，综合对比了常用的情感词典方法和支持向量机、朴素贝叶斯这两种机器学习方法。

（3）建立了用户心理特征的回归预测模型和分类预测模型，实现建模预测用户心理特征的目标。在数据分析过程，发现用户的心理特征与微博行为特征、情感特征有着密切的联系。数据分析结果部分支持了前人的研究结论。

本研究仍然存在着不足之处，一方面用户数据量过小，数据分析的过程和结果证实了对社交媒体用户心理特征建模是可行的，后续工作需要扩大样本量，使模型结果更有说服力。另一方面，本研究分别对比了基于情感词典和机器学习的情感分析方法，并从微博文本中提取了第一人称、其他人称词频的信息作为特征，但在实际操作过程中仍然遗漏了微博文本的许多特征信息。后续应考虑利用深度学习的方法进行情感分析和垃圾文本过滤。最后，本研究的背景基于心理健康领域，试图通过建模预测用户的心理特征状态，从而达到预警、干预的效果。涉及到用户心理层面的建模如何在现实生活中进行应用，仍然存在很大的商榷空间。

参考文献

[1] 董晨宇, 丁依然. 当戈夫曼遇到互联网——社交媒体中的自我呈现与表演[J]. 新闻与写作, 2018(01): 56-62.

[2] Marwick A E, Boyd D. I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience[J]. New Media & Society, 2011, 13(1): 114-133.

[3] Liu H, Maes P, Davenport G, et al. Unraveling the Taste Fabric of Social Networks[J]. International Journal on Semantic Web and Information Systems, 2006, 2(1): 42-71.

[4] Han S, Kim K J, Kim J H. Understanding Nomophobia: Structural Equation Modeling and Semantic Network Analysis of Smartphone Separation Anxiety[J]. Cyberpsychology, Behavior, and Social Networking, 2017, 20(7): 419-427.

[5] 董晨宇, 段采薏. 我的手机哪里去了：手机失联焦虑症的两种研究取向[J]. 新闻与写作, 2018(03): 40-44.

[6] 张晏宁. 大学生社交网站自我表露与社会资本的关系：积极反馈和社会比较的链式中介作用[D]. 陕西师范大学, 2018.

[7] Wilson R E, Gosling S D, Graham L T. A Review of Facebook Research in the Social Sciences[J]. Perspectives on Psychological Science, 2012, 7(3): 203-220.

[8] Kosinski M, Stillwell D, Graepel T. Private Traits and Attributes Are Predictable from Digital Records of Human Behavior[J]. Proceedings of the national academy of sciences, 2013, 110(15): 5802-5805.

[9] 李林英, 陈会昌. 大学生自我表露与人格特征、孤独及心理健康的关系[J]. 中国临床康复, 2004, 8(33): 7568-7570.

[10] Burger J M. 人格心理学[M].中国轻工业出版社, 2010.

[11] Rentfrow P J, Gosling S D. The Do Re Mi's of Everyday Life: the Structure and Personality Correlates of Music Preferences[J]. Journal of Personality and Social Psychology, 2003, 84(6): 1236.

[12] 张磊, 陈贞翔, 杨波. 社交网络用户的人格分析与预测[J]. 计算机学报, 2014, 37(08): 1877-1894.

[13] Schwartz H A, Eichstaedt J C, Kern M L, et al. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach[J]. PLoS One, 2013, 8(9): e73791.

[14] Kramer A D I, Guillory J E, Hancock J T. Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks[J]. Proceedings of the

National Academy of Sciences, 2014, 111(24): 8788-8790.

[15] Yu Y, Wang X. World Cup 2014 in the Twitter World: A Big Data Analysis of Sentiments in U.S. Sports Fans' Tweets[J]. Computers in Human Behavior, 2015, 48: 392-400.

[16] Whitty M T, Doodson J, CREESE S, et al. A Picture Tells a Thousand Words: What Facebook and Twitter Images Convey about Our Personality[J]. Personality and Individual Differences, 2018, 133: 109-114.

[17] Back M D, Stopfer J M, Vazire S, et al. Facebook Profiles Reflect Actual Personality, not Self-Idealization[J]. Psychological Science, 2010, 21(3): 372-374.

[18] Gonzales A L, Hancock J T. Mirror, Mirror on My Facebook Wall: Effects of Exposure to Facebook on Self-esteem[J]. Cyberpsychology, Behavior, and Social Networking, 2011, 14(1-2): 79-83.

[19] Wise K, Alhabash S, Park H. Emotional Responses during Social Information Seeking on Facebook[J]. Cyberpsychology, Behavior, and Social Networking, 2010, 13(5): 555-562.

[20] Ellison N B, Steinfield C, Lampe C. The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites[J]. Journal of Computer-Mediated Communication, 2007, 12(4): 1143-1168.

[21] Kim J, Lee J E R. The Facebook Paths to Happiness: Effects of the Number of Facebook Friends and Self-presentation on Subjective Well-being[J]. CyberPsychology, Behavior, and Social Networking, 2011, 14(6): 359-364.

[22] Kross E, Verduyn P, Demiralp E, et al. Facebook Use Predicts Declines in Subjective Well-being in Young Adults[J]. PLoS One, 2013, 8(8): e69841.

[23] Luhmann M. Using Big Data to Study Subjective Well-being[J]. Current Opinion in Behavioral Sciences, 2017, 18: 28-33.

[24] Seder J P, Oishi S. Intensity of Smiling in Facebook Photos Predicts Future Life Satisfaction[J]. Social Psychological and Personality Science, 2012, 3(4): 407-413.

[25] Tao Z, Kokas A, Zhang R, et al. Inferring Atmospheric Particulate Matter Concentrations from Chinese Social Media Data[J]. PLoS One, 2016, 11(9): e0161389.

[26] 乐国安, 赖凯声. 基于网络大数据的社会心理学研究进展[J]. 苏州大学学报(教育科学版), 2016, 4(1): 1-11.

[27] Bai S, Gao R, Zhu T. Determining Personality Traits from Renren Status Usage Behavior[C]//Computational Visual Media. Berlin, Heidelberg: Springer, 2012: 226-233.

[28] Bai S, Yuan S, Hao B, et al. Predicting Personality Traits of Microblog Users[J]. Web Intelligence and Agent Systems: An International Journal, 2014, 12(3): 249-265.

[29] 李昂, 郝碧波, 白朔天, 等. 基于网络数据分析的心理计算:针对心理健康状态与主观幸福感[J]. 科学通报, 2015, 60(11): 994-1001.

[30] 田玮, 朱廷劭. 基于深度学习的微博用户自杀风险预测[J]. 中国科学院大学学报, 2018, 35(01): 131-136.

[31] Mccrae R R, Costa P T. Clinical Assessment Can Benefit from Recent Advances in Personality Psychology[J]. American Psychologist, 1986, 41(9): 1001-1003.

[32] 杨秀君, 孔克勤. 主观幸福感与人格关系的研究[J]. 心理科学, 2003(01): 116-118.

[33] Diener E, Suh E M. Culture and Subjective Well-being[M]. MIT press, 2003.

[34] 张羽, 邢占军. 社会支持与主观幸福感关系研究综述[J]. 心理科学, 2007(06): 1436-1438.

[35] 牛更枫, 鲍娜, 范翠英, 等. 社交网站中的自我呈现对自尊的影响:社会支持的中介作用[J]. 心理科学, 2015, 38(04): 939-945.

[36] Gosling S D, Mason W. Internet Research in Psychology[J]. Annual Review of Psychology, 2015, 66(1): 877-902.

[37] Golbeck J, Robles C, Edmondson M, et al. Predicting Personality from Twitter[C]//2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. 2011: 149-156.

[38] 陈基越, 徐建平, 黎红艳, 等. 五因素取向人格测验的发展与比较[J]. 心理科学进展, 2015, 23(03): 460-478.

[39] John O P, Naumann L P, Soto C J. Paradigm Shift to the Integrative Big Five Trait Taxonomy: History, Measurement, and Conceptual Issues[M]. New York Guilford Press, 2008.

[40] Diener E, Emmons R A, Larsen R J, et al. The Satisfaction With Life Scale[J]. Journal of Personality Assessment, 1985, 49(1): 71-75.

[41] Rosenberg M. Society and the Adolescent Self-Image[M]. Princeton University Press, 2015.

[42] 郑飏飏, 徐健, 肖卓. 情感分析及可视化方法在网络视频弹幕数据分析中的应用[J]. 现代图书情报技术, 2015(11): 82-90.

[43] Hu M, Liu B. Mining and Summarizing Customer Reviews[C]//Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004: 168-177.

[44] Yu H, Hatzivassiloglou, V. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences[C]//In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2003: 129-136.

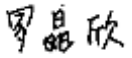
[45] Kim S-M, Hovy E. Determining the Sentiment of Opinions[C]//Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland: Association for Computational Linguistics, 2004: 1367-es.

[46] Bachrach Y, Kosinski M, Graepel T, et al. Personality and Patterns of Facebook Usage[C]//Proceedings of the 4th Annual ACM Web Science Conference. Evanston, Illinois: Association for Computing Machinery, 2012: 24-32.

哈尔滨工业大学与南方科技大学联合培养研究生学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于社交媒体的用户行为和心理研究》，是本人在导师指导下，在学校攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：  日期： 2020 年 6 月 7 日

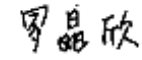
学位论文使用权限

学位论文是研究生在学校攻读学位期间完成的成果，知识产权归属南方科技大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为南方科技大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：  日期： 2020 年 6 月 7 日

导师签名：  日期： 2020 年 6 月 7 日

南方科技大学联培学位论文原创性声明和使用授权说明

南方科技大学联培学位论文原创性声明

本人郑重声明：所提交的学位论文是本人在导师指导下独立进行研究工作所取得的成果。除了特别加以标注和致谢的内容外，论文中不包含他人已发表或撰写过的研究成果。对本人的研究做出重要贡献的个人和集体，均已在文中作了明确的说明。本声明的法律结果由本人承担。

作者签名：罗晶欣　　日期：2020 年 6 月 7 日

南方科技大学联培学位论文使用授权书

本人完全了解南方科技大学有关收集、保留、使用学位论文的规定，即：

1. 按学校规定提交学位论文的电子版本。

2. 学校有权保留并向国家有关部门或机构送交学位论文的电子版，允许论文被查阅。

3. 在以教学与科研服务为目的前提下，学校可以将学位论文的全部或部分内容存储在有关数据库提供检索，并可采用数字化、云存储或其他存储手段保存本学位论文。在本论文提交后，同意在校园网内公开论文的在线浏览并下载全文。

4. 保密的学位论文在解密后适用本授权书。

作者签名：罗晶欣　　日　期：2020 年 6 月 7 日

指导教师签名：　　日　期：2020 年 6 月 7 日

致　谢

读研第一年的某个晚上，被好友邀请一起去一科报告厅看《大佛普拉斯》，被其中的一句旁白击中：“虽然现在是太空时代，人类早就可以坐太空船去月球，但永远无法探索别人的内心世界。”现在回想起来，大概当时的一些感受已经为毕业论文选题埋下了伏笔。顺便说，这部电影很棒，值得一看。

在深圳的两年很快就过去了，我还记得和朋友们坐在草坪上聊天的那些夜晚和星星，还记得和好友一起看过的无数海边日落（看日落真是一件非常疗愈的事情），也还记得找工作、收数据和写论文期间的焦虑。感谢这座城市和在这座城市遇见的人。

感谢夏志宏老师这两年里在学习和生活上的耐心指导和关心，谢谢钱江老师从毕业论文的选题、方法确定到论文撰写期间提供的指导和提出的修改意见，让我的毕业论文能够顺利完成。

感谢我的家人的支持和照顾。由于疫情的影响，毕业论文的大部分工作都是在家期间完成的，家人给予了我莫大的理解，在生活上关心我，让我感受到了特别多的爱意和幸福感。

谢谢实验室的每一位同学，我们共同度过了学习、找工作的时光，今后或许分散在各地，希望大家一切顺利。

特别感谢 The Ataris，回过头翻看写毕业论文期间的歌单，发现他们的《The Hero Dies in This One》占据了我的听歌排行榜单第一名。最初还是从一个姓御景的少年那里知道这首歌的。歌词里说道：“那些艰难的时光只会让我们更强大。”在循环 94 次的时间里，它陪我度过了大段焦灼的时间。最后借用这首歌结尾的歌词祝愿自己，希望在今后的漫长人生中，stay who you are.