伪健康信息甄别能力与文本特征研究
——合肥市中老年群体为例

钱鹏博[1]，倪悦[2]，胡柠荟[3]

摘要：随着老年人健康意识的不断增强，越来越多的老年人倾向于通过网络搜寻健康信息，有的通过网站，有的通过移动终端，也有的通过社会化媒体等渠道获取健康信息。与此同时，我们也应该注意到，新媒体平台上的健康信息看似丰富多元，实则良莠不齐，常常令公众真假难辨，有不少信息甚至以“健康资讯”之名行虚假广告之实，欺骗公众、误导消费。对提升公众健康信息素养而言，此类“伪健康信息”传播显然无益甚至可能危害公众健康。我们此次调研将会通过构建模型、分析数据的方法，多方位多角度科学的调查这些伪健康文章，找到其迷惑人的内在因素，帮助中老年人提高甄别信息的能力。我们的调研结果表明伪健康文章受众群体包括以患有高血压、高血糖等慢性疾病为主的中老年人群和重大疾病患者及其家属，针对第一类人群的伪健康文章主要以积极情感倾向为读者介绍某种改善慢性病病情的方法，针对第二类人群的伪健康文章主要以消极情感倾向从专业人士的角度为读者提供疾病预防知识及建议,将文本情感倾向与文章主题特征联系起来，为健康信息质量研究提供了新视角。

关键词：伪健康信息；LDA 主题模型；情感分析

指导老师：胡芳芳老师

团队成员：倪悦，胡柠荟，钱鹏博，张俊伟，李文雅，项馨悦，陈奕君，王贝宁，丁雨辰，刘庆局

1. 安徽大学管理学院
2. 安徽大学管理学院
3. 安徽大学石溪学院

目录

一、调研现状及问题

（一）社会背景

据 2021 年 5 月 11 日最新消息，国家统计局在国新办发布会上发布了第七次全国人口普查关键数据 60 岁及以上人口为 26402 万人，占 18.70%(其中，65 岁及以上人口为 19064 万人，占 13.50%)。，相对于 2020 年上升 5.44%。健康是老年人面临的首要问题。研究发现，越来越多的老年人倾向于通过网络搜寻健康信息，具体体现在: 老年人的健康意识比其他人强，他们关注疾病信息、医药信息、营养膳食信息等。健康是保障中老年群体生活和参与社会的重要基础，也是健康中国建设的重要组成部分。当前，积极应对人口老龄化已成为国家战略。

（二） 信息传播背景

从传播方式看，健康信息的传播既包括以报纸、广播、电视、互联网、手机等媒介为载体的大众传播活动，也包括以政府、社区、家人、同伴、医生为主体的组织传播或人际传播行为。尽管这些传播方式始终交织共存于社会的健康传播网络之中，但显见的是，越来越多的社会公众选择通过网站、移动终端和社会化媒体等渠道获取健康信息。中国互联网络信息中心的统计数据表明，医学健康类信息在网络用户关注的科普知识类别中排名高居第二，而国内知名健康资讯门户网站的月均访问人次最高可达 3 亿。这些数据彰显了新媒体在健康传播中日益突出的角色和地位。与此同时，我们也应该注意到，新媒体平台上的健康信息看似丰富多元，实则良莠不齐，常常令公众真假难辨，有不少信息甚至以“健康资讯”之名行虚假广告之实，欺骗公众、误导消费。对提升公众健康信息素养而言，此类“伪健康信息”传播显然无益甚至可能危害公众健康。

（三）市场背景

在“健康中国 2030”的背景下，促进我国老年人在线的健康交流已经成为一个重要且紧迫的问题。互联网及社会化媒体为老年人提供了随时随地获取、分享、评论信息的渠道与平台，但其中也潜藏了诸多风险。中老年网民年纪越大，越容易遭遇网络谣言, 且学历较高的老年人也同样面对谣言危害。微信数据显示, 截至 2020 年 10 月中旬，触发微信的谣言防护机制用户中，50 岁及以上用户占比超四成，远远高于其他年龄群体。老年人因自身生理退化等原因对健康话题的高度关注而成为网络中健康信息的重要消费者。尽管老年人可能受益于子代的数字反哺，但反哺停留于技能培训，而在偏重内容层面的评判、转发、分享的反哺较少，这使得老年人在信息获取、使用、处理等方面均处于不利位置，更易成为伪健康信息的易感人群，老年人容易轻信伪信息的内容并积极转发，成为伪健康信息的传播者与受害者。如 65 岁以上老年人是转发假新闻的主力，分享假新闻文章数量是最年轻年龄段的 7 倍；《2020 年网络中谣言治理报告》中显示，谣言鉴别力弱、受教育程度低、网龄短、主观幸福感低的中老年群体更易传谣。这表明老年人在信息快速流动与难以监控分析的网络传播时代，更容易受到伪健康信息的诱导，成为伪健康信息的“二传手”。

这对提高老年人生命质量、促进健康公平带来了巨大的挑战。

二、调研目标及意义

（一）调研目标

1.社会对健康信息的需求量极大却缺乏辨别和监管体制，通过研究实践完善体制。

“互联网+时代的到来，互联网已成为大众获取健康信息的主要来源”。第 47 次《中国互联网络发展状况统计报告》显示[1]，2020 年 12 月，我国网民规模达 9.89 亿，较 2020 年 3 月增长 8540 万，互联网普及率达 70.4%，同时《中国网民科普需求搜索行为报告》表明[2]，健康与医疗在网络用户关注的科普知识类别中排名第一，搜索份额占比 66.83%，可见，用户对健康医疗类科普知识尤为关注。通过健康网站或社交媒体获取健康信息，成为用户实现自我健康教育，提升健康信息素养的重要途径。然而由于互联网的开放性和便捷性，且监管网络信息的法律法规尚不健全，网络信息质量良莠不齐，真伪难辨[3]，伪健康信息逐渐成为某些社交媒体账号吸引流量的工具，由于大量用户缺乏足够的健康信息素养，面对众多伪健康信息难辨真假，严重影响了用户对健康信息的信任和有效利用。

2.针对中老年要更加重视提升他们的辨别真假信息能力

由于社交平台具有强关系性、匿名性以及信息传播的快速性，客观上为伪健康信息的传播提供了条件，且随着年龄的增长，个体认知的不确定性提升，部分伪健康信息甚至造成了较大的社会影响。在社交平台用户中，中老年人群因自身生理退化原因成为健康信息的重要消费者。由于该类群体缺乏健康信息质量的评价能力，在接受官方信息不及时的情境下，伪健康信息可能会对此类群体造成负面影响。故引导该类群体辨别网络伪健康信息工作至关重要。而构建伪健康信息文本特征对中老年为健康信息甄别能力的提升至关重要

（二）调研意义

1.从信息源头和接收者双向探究伪健康信息特征，为中老年用户提供参考

随着互联网技术的发展与普及，互联网日益成为国民的重要健康信息来源之一，通过健康网站或社交媒体获取健康信息成为社会大众实现自我健康教育、提升健康信息素养的一种重要途径。然而，相对于报纸、广播、电视等传统媒体，健康网站、社交媒体等网络平台上传播的健康信息质量良莠不齐，不乏谣言、迷信、伪科学等各种无用甚至是有害的伪健康信息。这些伪信息存在相当的隐患，已引起学者们的关注。为此，我们以微信等网络平台中传播的健康信息为样本，探究伪健康信息的特征，以期提供有效工具帮助网络用户鉴别真伪健康信息，同时既对提升我国网络用户健康信息素养有着积极的意义。

2.帮助健康网站和平台的建设者和管理者识别剔除低质量的健康信息

由于社交平台具有强关系性、匿名性以及信息传播的快速性[4]，客观上为伪健康信息的传播提供了条件，且随着年龄的增长，个体认知的不确定性提升，部分伪

健康信息甚至造成了较大的社会影响。在社交平台用户中，中老年人群因自身生理退化原因成为健康信息的重要消费者[5]。由于该类群体缺乏健康信息质量的评价能力[6]，在接受官方信息不及时的情境下，伪健康信息可能会对此类群体造成负面影响。故引导该类群体辨别网络伪健康信息工作至关重要。而构建伪健康信息文本特征对中老年为健康信息甄别能力的提升至关重要

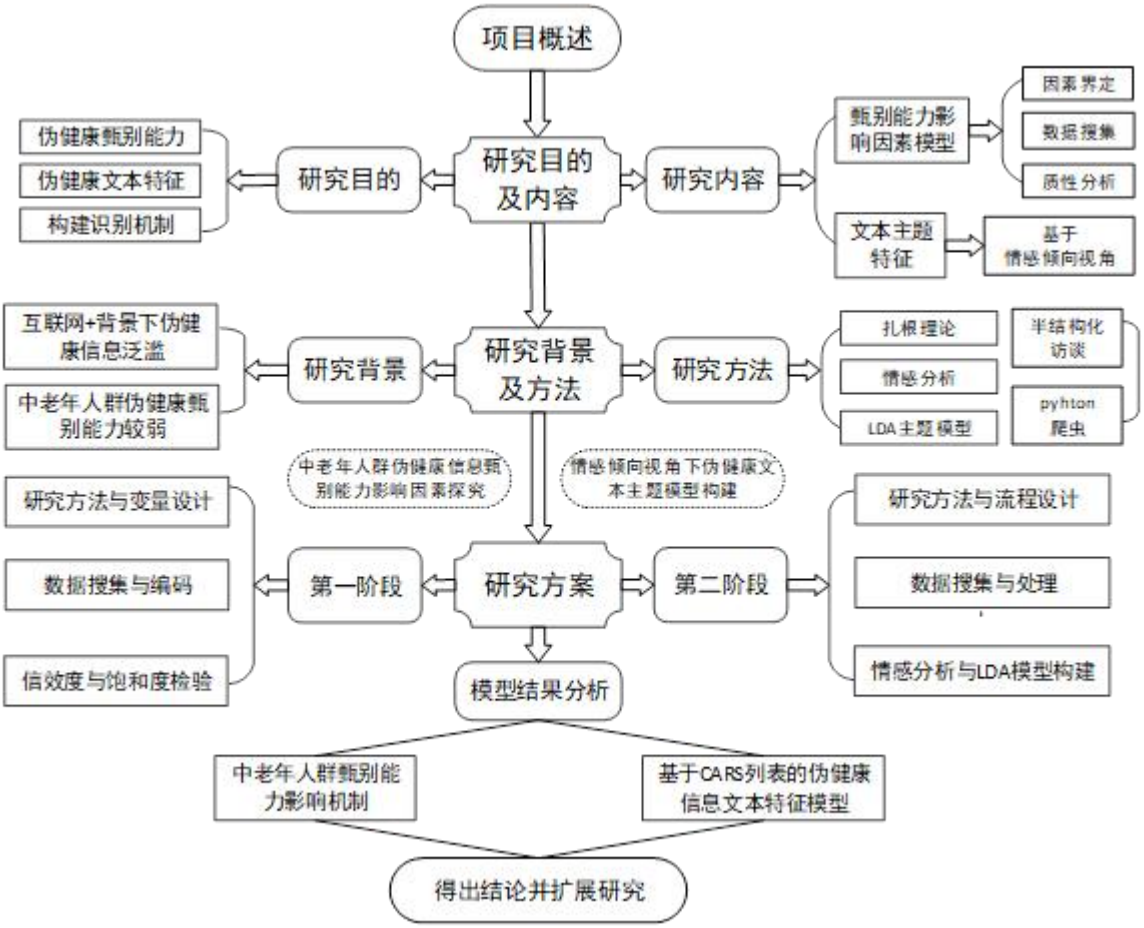
三、调研内容及方法

（一）分析调研对象

随着互联网技术的发展与普及，互联网日益成为国民的重要健康信息来源之一，通过健康网站或社交媒体获取健康信息成为社会大众实现自我健康教育、提升健康信息素养的一种重要途径。然而，相对于报纸、广播、电视等传统媒体，健康网站、社交媒体等网络平台上传播的健康信息质量良莠不齐，不乏谣言、迷信、伪科学等各种无用甚至是有害的伪健康信息。这些伪信息存在相当的隐患，已引起学者们的关注。

（二）确定调研内容

团队首先通过分层抽样和随机抽样对合肥市 200 位中老年人群进行半结构化访谈，基于扎根理论构建中老年人群伪健康信息识别能力影响因素模型，随后针对中老年群体关注的伪健康文章类型，利用 python 进行文本爬取，随后对伪健康信息文本进行情感分析，再采用 LDA（Latent Dirichlet Allocation）主题模型提取文章内容的深层次语义主题特征，分析不同情感倾向下的文本特征及对应读者群体，最后基于 CARS（credibility、accuracy、reasonableness、support）列表构建伪健康文本主题模型，分析伪健康信息的文本特征和情感特征及两者潜在关联，以帮助中老年群体正确辨识伪健康信息，提升其健康信息素养提供有益参考。研究思路及过程如下图所示。



（三）调研方法

1、LDA 主题模型

潜在狄利克雷分布模型（Latent Dirichlet Allocation）是 Blei 等于 2003 年提出的一种文档主题生成模型。由于 LDA 能够降低文本表示维度，在语义挖掘领域得到了广泛应用。LDA 模型是一个三层贝叶斯网络模型，其核心思想是每个文档对应一个服从 Dirichlet 分布 θ 主题分布，每个主题对应的词分布服从 Dirichlet 分布 ϕ，其中文档-主题分布 α 参数和主题-词分布 β 参数服从 Dirichlet 分布 α，β .设采集 M 条伪健康信息文本，共有 N 个词，文本主题个数为 K，从 Dirichlet 分布 α 中取样生成微博主题的 ϕ 词分布，根据词分布，取样生成相应的词 w。模型不断重复上述过程，直至所有微博文本采样完毕，最终得到每条微博文本的主题分布及各主题的词分布。

LDA 主题模型是一种无监督模型[7]，其中主题个数是模型重要的输入参数。本文采用困惑度（perplexity） 确定文档的最优主题数目。困惑度是用于评估模型优劣的标准，可用于调节主题个数，其计算公式如下：

$$perplexity(D) = \exp\left\{-\frac{\sum_{d=1}^{M}\log\, p(w_d)}{\sum_{d=1}^{M} N_d}\right\}$$

上式中，wd 表示词，p(wd)表示文档中词的概率，Nd 表示文档数量，D 表示文档中所有词的集合。使用困惑度进行评估时，主题越多，困惑度数值会逐渐下降；

而主题数越多，LDA 模型计算代价越大。同时为了避免模型过拟合，应综合考虑选取困惑度数值和主题数目，选择困惑度最小和主题数最少的数值作为 LDA 模型训练的最优数目。

2、情感分析

本文利用自然语言处理（简称 NLP）、数据挖掘算法等对文本语言进行情感判断，从而把握文本意见观点、态度的计算研究。本文在 python 环境中调用百度 AI 平台的开源情感分析文档将文本的情感极性划分为消极、中性与积极三个层级，便于后面对不同层级情感倾向的文章进行 LDA 主题模型训练，从而研究不同情感倾向下的文本主题的区别。

3、扎根理论

扎根理论是哥伦比亚大学的 Glaser 和芝加哥大学的 Strauss 于 1967 年共同提出，发展至今已形成 Glaser＆Strauss 的经典版本、Strauss＆Corbin 的程序化版本以 Charmaz 的建构主义版本等三大流派 (吴毅等，2016)。本研究主要借鉴程序化版本扎根理论思想开展分析。扎根理论是一种量化与质化研究相结合的定性研究方法，旨在从经验资料基础上提升理论，由经验资料的深入分析进而逐步形成理论框架。扎根理论的基本思路主要包括从资料中产生资料、对理论保持敏感、不断比较、理论抽样、灵活运用文献以及理论性评价等方面。在质化研究的资料收集方式上，扎根理论主要有观察法与访谈法两种典型方式; 在量化研究的资料编码技术上, 主要分为开放式编码、主轴编码以及选择性编码三部分程序。

四、调研过程

（一）前期准备阶段（2021 年 6 月–2021 年 7 月）

（1）组建队伍，联系指导老师。

（2）确定调研课题，明确成员分工，撰写暑期社会实践策划书。

（3）搜集相关研究文献和著作，对项目主题进行初步了解。

（4）进行前期工作准备，联系调研地相关单位。

（5）系统了解如今特定人群对网络伪健康信息的辨别方式、存在的问题以及影响因素，收集相关现有理论与模型结构，制定研究方案。

（二）实地调研阶段（2021 年 7 月）

（1）以半结构化访谈形式实地调研三里庵街道龙河社区 30 位中老年居民的网络伪健康信息识别能力的影响因素，运用扎根理论研究方法，借助质性分析软件 NVivo12 归纳出影响中老年人群网络健康信息识别因素的主范畴和核心范畴

（2）提出中老年用户网络信息识别影响因素模型，从信息源头和信息对象两个方面对伪健康信息构建特征序列和识别模型。

（三）数据处理阶段（2021 年 7 月）

（1）依照研究假设选取研究方法，明确研究流程，并按照取样的要求进行取样。基于前期取样的阶段，对文本进行去重与压缩、空行删除、停用词去除等操作来降

低噪声。

（2）此外，伪健康信息会涉及医药养生的专用词汇，所以需要构建专用词典，最后通过文本分词和关键词提取为后续的情感模型分析奠定数据基础。

（四）调研总结阶段（2021 年 8 月）

（1）在数据处理的基础上进一步进行实证结果的分析，选取 LDA 主题模型挖掘其主题特征及不同情感倾向并借助 CARS 特征列表构分析其文本特征与情感特征及两者之间的关联，得出相应参数，构建情感倾向视角下基于 CARS 列表的伪健康文本主题模型。

（2）对模型进行实证性检验，对前期问题进行改善，进行深度调研与数据挖掘，完善新的模型结构，整理情感倾向视角下基于 CARS 列表的伪健康文本主题模型，撰写初期结论报告并请相关专家进行模型的鉴定与评估。

（3）撰写调研报告。

五、调研简介

（一）第一阶段——中老年群体伪健康信息甄别能力影响因素探究

1、研究方法与流程设计

由于第一阶段研究居民的伪健康信息甄别能力，涉及到各种抽象的因素，这类因素往往具有难以量化分析的复杂性特征；同时，质性研究注重在研究过程中采用开放、弹性的方法，更有利于深入探索和研究抽象的复杂因素，因此，质性研究方法更契合本文研究的内容和需求。质性研究方法中的扎根理论，研究起点始于具体的社会现象，通过收集和分析资料，最终形成解释性的理论体系，而在本阶段中，旨在挖掘影响我国中老年群体伪健康信息甄别能力的影响因素，在此基础上构建理论框架，因此，选用扎根理论，自上而下地进行数据分析，通过归纳形成实质理论，为第二阶段文本特征模型构建理论基础。

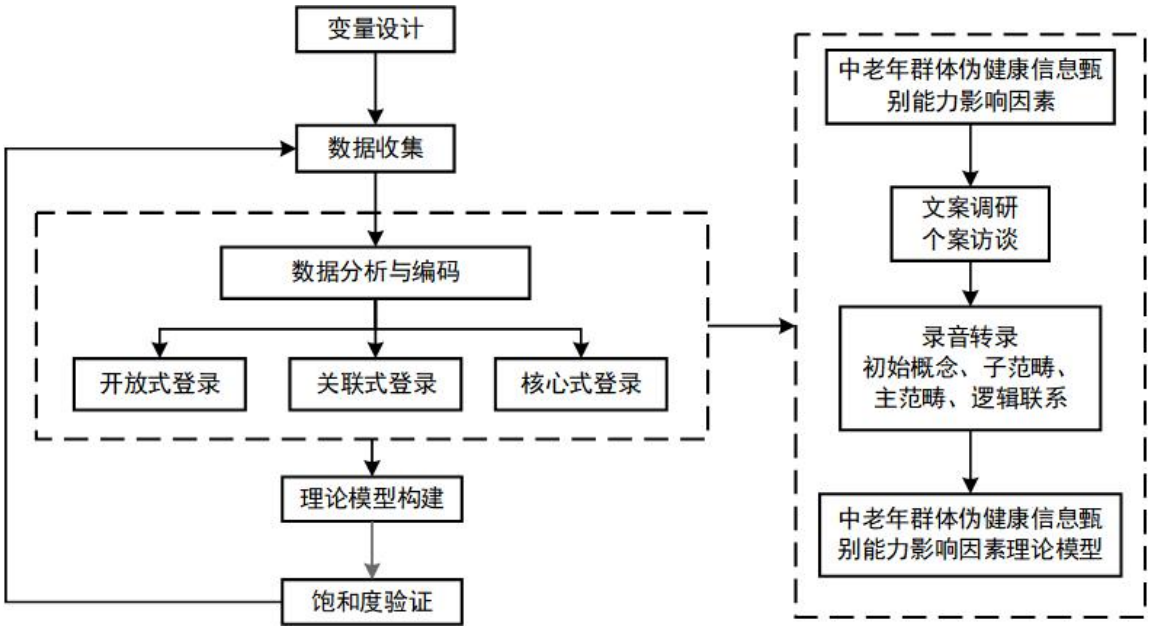


图 2 理论框架的构建

2、数据收集

在数据收集方面，扎根理论特别强调研究样本的丰富性而非样本数量的多少。本阶段为了保证样本数据来源的多样性，分别在不同的时间点、对合肥市不同城区进行数据的跟踪收集。样本数据来源的地区方面，本文选择在蜀山区、瑶海区、包河区、庐阳区进行数据的收集，因为这 4 个城区能够在很大程度上覆盖合肥市经济发展的主要特点。

本阶段计划采用焦点小组讨论和深度访谈 2 种方法来收集质性资料。焦点小组讨论法的优势在于受访者之间可以充分讨论、彼此启发，而研究者可以直观观察到受访者在讨论过程中的反应，从而获取的资料更为全面、真实;个人深度访谈法的优势在于受访者思考和发表观点的时间充分，研究者可以更深刻地挖掘资料。两种方法综合使用，可以使获取的资料在广度和深度方面形成有效互补。

表 3 调查区域时间及方法设计

地区	时间点	方法
包河区	2021 年 7 月	焦点小组讨论/个人深度访谈
蜀山区	2021 年 8 月	焦点小组讨论/个人深度访谈
庐阳区	2021 年 9 月	焦点小组讨论/个人深度访谈
瑶海区	2021 年 10 月	焦点小组讨论/个人深度访谈

根据扎根理论，理论模型的构建主要分为两步，第一步是初步构建理论模型阶段; 第二步是检验模型理论饱和度。本文研究样本数是依据模型理论饱和度来确定的。在初步构建理论模型阶段，计划对 28 人进行访谈，每个地区进行 1 次焦点小组讨论，每个焦点小组 3 人，每组讨论时间持续约为 2 小时; 每个城区进行 4 次个人深度访谈，每次深度访谈的时间持续约为 80~100 分钟(8 人为面对面访谈，8 人为视频访谈) 。在模型理论饱和度检验阶段，共有 9 人参与了访谈，其中有 6 人进行了 2 组焦点小组讨论 (时间约为 2 小时) ，3 人进行了个人深度访谈 (时间约为 90 分钟) 。所有受访者的年龄需介于 55~75 岁之间。本文旨在研究我国城市居民健康信息甄别能力的影响因素，具体以是否关注 CARS 列表 4 个维度的内容来界定健康信息甄别能力，基于此，焦点讨论小组和个人深度访谈所使用的访谈提纲如下表所示。

表 4 访谈提纲

访谈主题	主要提纲内容
个人基本情况	性别、年龄、教育程度、收入、主要社会关系网络、对健康信息关注程度
伪健康信息甄别能力	在您获取健康信息的过程中，以下内容哪些您会关注、哪些不会关注: 信息是否匿名、是否缺乏、质量监控、信息中有无错误的语法或错别字、信息用语是否过于夸张或绝对; (可信度维度)

	在您获取健康信息的过程中，以下内容哪些您会关注、哪些不会关注：信息是否标注日期、信息、是否含糊不清或不全面、信息日期是否较老、信息观点是否片面、涉及专业知识是否错误；(准确性维度) 在您获取健康信息的过程中，以下内容哪些您会关注、哪些不会关注：信息语气有无节制、信息内容是否存在冲突或矛盾；(合理性维度) 在您获取健康信息的过程中，以下内容哪些您会关注、哪些不会关注：信息数据是否缺乏来源和统计、信息是否缺少文档来源说明、其他资源能否佐证该信息 (相关支持维度)
伪健康信息甄别能力影响因素	您通常通过哪些方式来获取质量更高的健康信息呢？ 通常通过什么方式来甄别健康信息的质量？为什么？ 您在获取高质量健康信息的过程中遇到了什么困难吗？ 您认为您需要什么样的帮助来克服这些困难？

3、模型变量设计

本阶段参考 CARS 列表来界定健康信息甄别能力，具体按照是否关注以下 4 个维度来界定：第一，可信度，即健康信息的真实性和可靠性；第二，准确性，即健康信息所反映的相关专业知识的正确性；第三，合理性，即健康信息的公平、客观、适度及一致性;第四，相关支持，即辅助甄别健康信息质量的其他来源和渠道。

4、数据编码

讨论和访谈过程中受访者围绕主题自由发言，发言内容在征得参与者同意后全程录音，访谈结束后将录音逐字转化为文本。在录音转化文本过程中，对参与者的音调、语速都进行标注以提升数据的可靠性。访谈记录的书面文本整理 2 稿，一稿为原始记录，另一稿为综合了信息真实性和受访者各方面情况进行判断后、剔除掉无效内容后留下的原始语句。从方法论视角而言，扎根理论的研究方法主要有 3 个流派：格拉瑟与斯特劳斯的原始版本、斯特劳斯和科宾的程序化版本以及卡麦兹的建构主义版本，这 3 个流派在数据编码方面差异较大。本阶段选择斯特劳斯和科宾的程序化数据编码方式，对访谈收集到的文本资料进行三级编码，即开放式登录、关联式登录和核心式登录，在过程中对概念进行不断提炼、归纳和修正，直至达到理论饱和。

5、信效度检验

质性研究中，信度是指记录的数据与实际事物真实情况的吻合程度。本文确保研究信度的方法为：首先，收集数据环节，按照不同地点和时间点进行收集，来保证资料的多元性；其次，在资料获取过程中，通过受访者的协助来进一步检核记录内容；最后，选取 2 位编码员运用内容分析法分别进行编码分析，根据公式定性数据的

编码信度＝一致的编码数目/所有编码数目进行计算。

效度方面，进一步分为外部效度和内部效度 2 种。外部效度是指研究者所宣称的知识与实际事实相符的程度；内部效度是指研究者对现实的建构程度。外部效度强调研究发掘的本质规律能够推广来解释符合类似情境和时间内的其他事物；内部效度则强调研究过程中各部分、方面、层次和环节之间的平衡性和一致性。本文确保研究效度的方法为:外部审核方面，请 5 名专家分别对研究过程和初步结论进行审核，同时也请受访者检视最终形成的理论框架，依据他们的建议进行调整与修改；内部自审方面，记录过程始终提醒自己对已有理论和原始资料传递的理论均保持高度警觉，同时也注重将新发掘的理论与已有研究结果进行比较分析。

（二）第二阶段——情感倾向视角下伪健康文本主题模型构建

1、研究方法与流程设计

在第二阶段中，为深入挖掘网络伪健康信息的文本特征和情感特征，构建伪健康信息特征模型，首先利用 python 爬取微博、微信等社交平台 2015—2020 年内被曝光的伪健康文章，其次对伪健康信息的文本进行情感分析，再采用 LDA（Latent Dirichlet Allocation）主题模型提取文章内容的深层次语义主题特征，分析不同情感倾向下的文本特征及对应读者群体，最后基于 CARS（credibility、accuracy、reasonableness、support）列表构建伪健康文本主题模型，分析伪健康信息的文本特征和情感特征及两者潜在关联，以帮助中老年群体正确辨识伪健康信息，提升其健康信息素养，为改善网络健康信息环境提供有益参考。

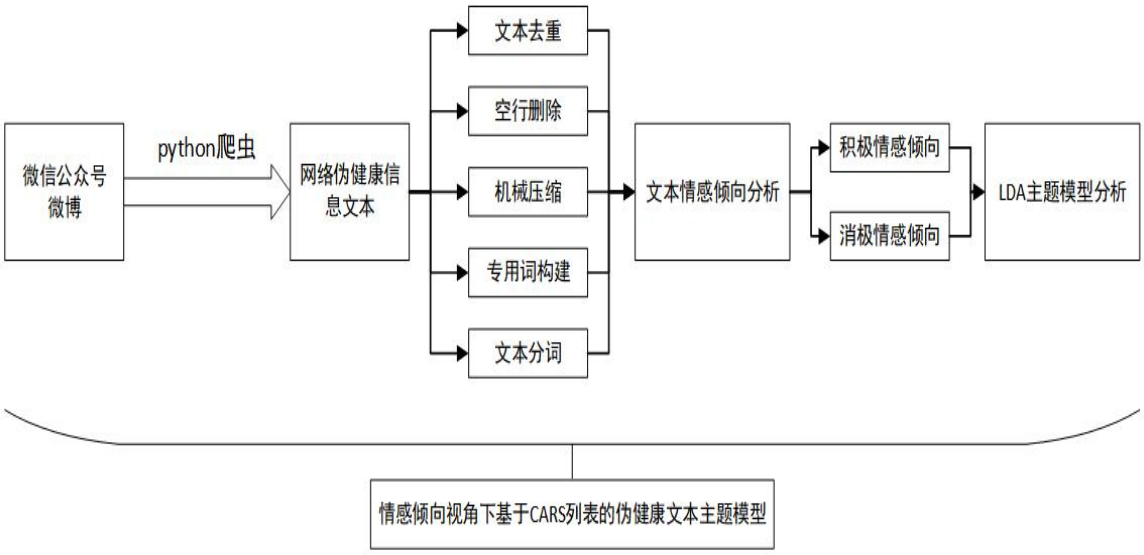


图 4 伪健康信息特征模型的构建

2、 数据搜集与处理

本阶段拟在微博平台上以健康领域关键词检索谣言文章并通过微信小程序里面的“微信谣言助手”查找健康养生类谣言，并反向对应找到公众号的相关文章，采

用 fiddler 抓取网络数据包，然后进行反序列化，解析字段，最后存取数据库，数据包括公众号名称、文章标题、文章内容。

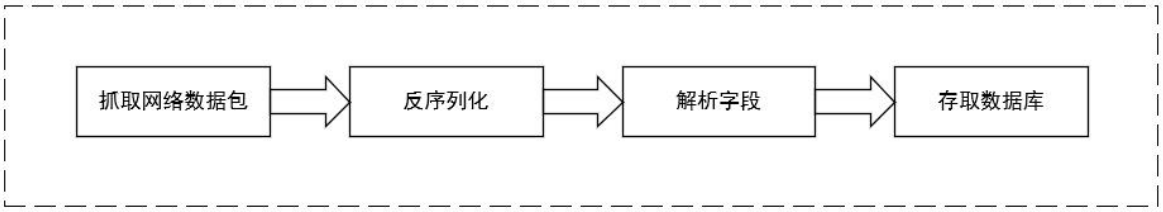


图 5 爬取文章流程

基于 python 获取的伪健康文本信息含有大量的空格、空行，以及价值低甚至无价值的评论冗余信息，这些信息会对情感分析造成干扰，所以通过文本去重与压缩、空行删除、停用词去除等操作来降低噪声。此外，伪健康信息会涉及医药养生的专用词汇，所以需要构建专用词典，最后通过文本分词和关键词提取为后续的情感模型分析奠定数据基础。

3、情感分析与 LDA 模型构建

在完成数据搜集与处理后，首先基于中文文本情感词典，计算伪健康文本的标题和内容的情感得分。情感得分取值范围为[0,1],判定积极和消极情感倾向的得分，并结合置信度对文本进行情感归类，研究伪健康文本的情感倾向分布，并将积极情感倾向文本和消极情感倾向文本区分开，分别进行 LDA 主题建模，研究不同情感倾向下文本主题特征的差异，首先对标题和内容的困惑度进行计算确定主题个数，分类标准则依据困惑度计算公式，基于 python 软件的 gensim 包计算不同主题的主题词并根据主题词权重分布选取一定量的主题词进行文本主题分析。

六、调研结果与分析

（一）第一阶段——中老年群体伪健康信息甄别能力影响因素

1.模型构建

目前中国互联网正持续向高龄人群渗透，据第 44 次《中国互联网络发展状况统计报告》，截至 2019 年 6 月，50 岁以上网民群体占比由 2018 年底的 12．5%提升至 13．6%。微信等社交媒体已成为中老年群体信息获取的主要渠道，有关生活养生、休闲娱乐、健康保健的信息正在中老年群体中广泛传播 。信息接收与信息接受的主要区别在于接收注重收到本身，而接受注重于收到并经过考虑愿意使信息内化，因此本文中的信息接受行为我们认为是用户收到健康信息后内化为自己的知识体系，进而进行实践、传播 。中老年网络用户比例不断提升，越来越多的研究者针对中老年用户这一特定群体进行健康信息行为研究。吴丹等通过实验发现老年用户的健康状况、网络熟悉程度以及信息的可信度会影响用户的信息检索行为。朱姝蓓等 研究发现个人心理因素及个人实施成本是直接影响老年用户信息搜寻行为的内在因素。王莹莹通过扎根理论发现影响老年人健康信息规避行为的因素为个人因素、信息因素和社会因素 。本阶段旨在通过扎根理论探究中老年用户对伪健康信息的甄别能

力，构建中老年伪健康信息甄别能力影响因素的理论模型。

2.研究设计

（1）研究方法

本研究认为，探究老年人群体伪健康信息甄别能力，不能仅仅局限于对相关文献的梳理，还应结合原始资料进行更为深入的调查和分析，因此本研究选择扎根理论作为主要研究方法，采用半结构访谈来收集原始资料。通过对访谈录音资料进行整理分析，探究影响中老年人群体甄别能力的影响因素。

（2）数据来源

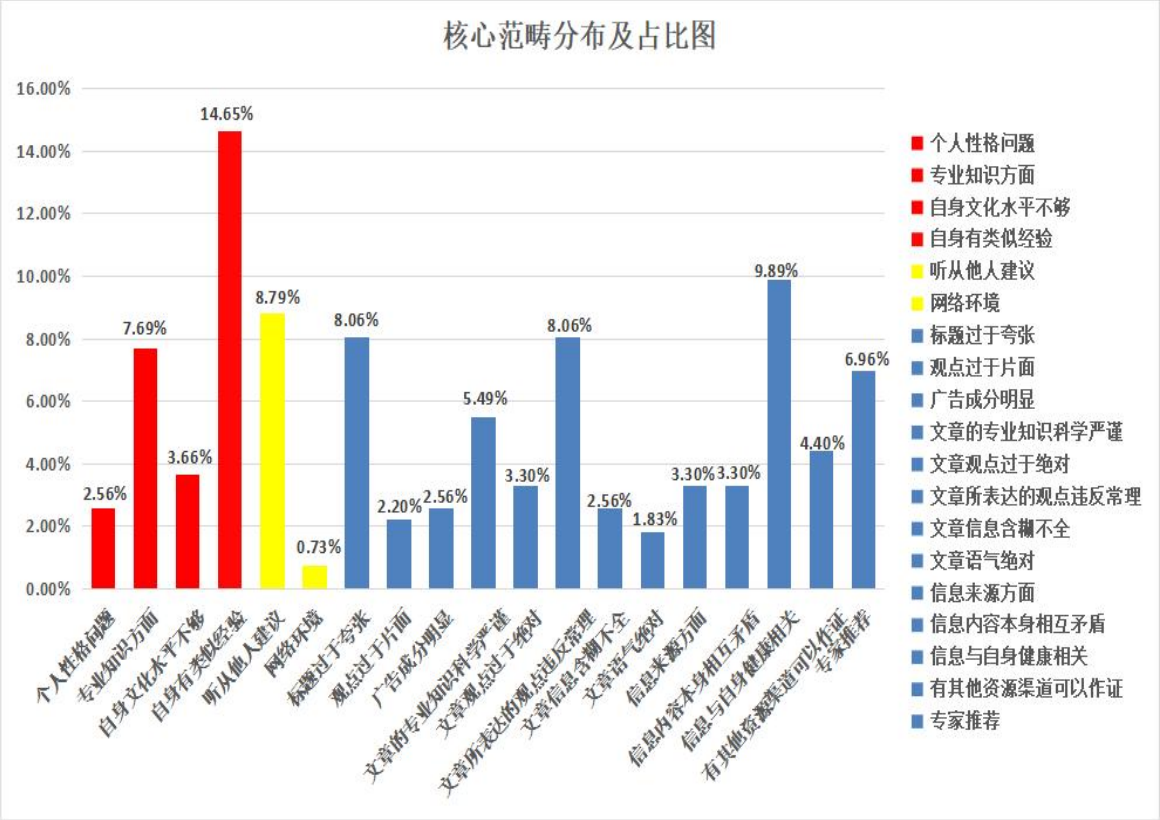
本研究通过在合肥市各类中小型社区广场选取年龄在 55 岁以上并经常使用微信的中老年人，每次访谈时间为 20 – 60 分钟，访谈过程中结合访谈提纲并根据访谈时的实际情况灵活地做出调整。访谈内容主要包括受访对象的基本信息（性别、年龄、受教育程度、职业、健康状况）以及对微信平台中存在的健康信息的态度和接受行为。

（3）编码与分析

本研究共访问 30 位受访者，受访者多为退休人员，年龄范围为 55—83 岁，平均龄为 60．38 岁，访谈结束后，将收集到的 30 份录音资料转换成近两万字的原始资料作为研究的主要资料，对原始资料进行三级编码，在研究过程中不断比较分析整合概念，直至达到理论饱和。共进行两轮编码，第一轮开放编码的目的在于构建编码表，抽取中老年伪健康信息甄别能力影响因素的常用概念，如个人性格、专业知识、文化水平、经验判断、他人影响、网络环境、标题夸张、观点片面等 19 个基本范畴，并分别对应纳入个人因素影响、外界因素、信息本身因素，最终汇总于编码表。第二轮编码由 2 名编码人员基于同一编码表分别进行，编码结束后，汇总不一致的编码，并进行讨论，保留一致意见作为最终编码结果。具体数据编码汇总如表 6 所示。

	范畴	常用概念	参考点数量（个）	参考点占比（%）	材料来源数量（个）
伪健康信息受众甄别能力	个人因素影响	个人性格问题	7	2.56%	5
		专业知识方面	21	7.69%	13
		自身文化水平不够	10	3.66%	7
		自身有类似经验	40	14.65%	21
	外界因素	听从他人建议	24	8.79%	14
		网络环境	2	0.73%	1
	信息本身因素影响	标题过于夸张	22	8.06%	16
		观点过于片面	6	2.20%	5
		广告成分明显	7	2.56%	6
		文章的专业知识科学严谨	15	5.49%	10
		文章观点过于绝对	9	3.30%	6
		文章所表达的观点违反常理	22	8.06%	16
		文章信息含糊不全	7	2.56%	5
		文章语气绝对	5	1.83%	5
		信息来源方面	9	3.30%	9
		信息内容本身相互矛盾	9	3.30%	8
		信息与自身健康相关	27	9.89%	17
		有其他资源渠道可以作证	12	4.40%	8
		专家推荐	19	6.96%	12

通过对访谈资料编码分析，明确中老年伪健康信息甄别能力各个影响因素占比，并在此基础上绘制核心范畴分布及占比图，见图 7。自身有类似经验对中老年群体伪健康信息甄别能力影响最大，占比 61.92%,其次是个人因素（28.56%）。

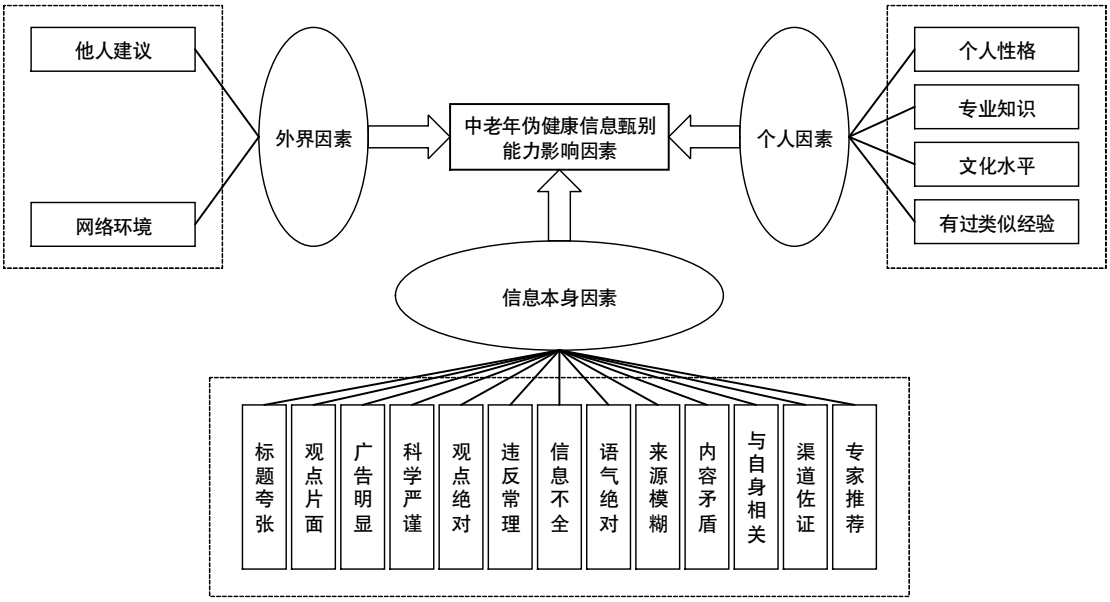


（4）研究信度、效度及理论饱和度检验

为提升研究信度，本文采用目的抽样、研究者协同编码、二次访谈等方式。在确定访谈对象时，对被试严格筛选，剔除不符合条件的人。采访和转录的研究者为 2 人一组协同编码，尽可能排除个体编码带来的不确定因素。采访完成后，对部分访谈转录文本不确定的内容与相应被试协商并进行二次访谈，确保访谈数据真实准确。研究效度即研究结果有效性程度，指研究结果准确性和可推广性。为保证研究效度，本文采用“三角检验”方法，即采用多种手段对同一案例进行多维交叉验证，以避免由于研究者主观偏见而对研究结论产生错误判断。本文通过对不同籍贯、年龄的群体进行半结构化访谈，获取智能健康手环不持续使用行为相关信息，佐之以相关领域研究者论文，确保三角检验有效实施，进而确保本文有较高效度。在以往研究基础上，本文采用 Francis 等所使用的方法：对前 16 份访谈资料进行编码，结果发现在对第 21 份资料进行编码时已没有新的基本范畴出现，随后继续对剩余 3 份资料进行编码，完成后通过重复样本编码并结合预留的 5 份样本数据编码完成理论饱和度检验。结果没有出现新范畴，被试数达到对特定主题进行探讨的充分样本量，研究建立的理论模型已达饱和，因此，研究结论有一定可靠性。

(5) 中老年伪健康信息甄别能力影响因素模型构建与分析

由图可知，中老年微信健康信息接受行为的影响因素可以归纳为个人因素、外界因素、信息本身因素 3 个主范畴。这 3 个主范畴皆为健康信息接受行为的影响因素，但它们各自对信息接受行为的影响程度和方式并不完全一致。具体阐释如下。



2.5.1 个人因素

中老年微信用户感知风险主要体现在保护个人信息及感知信息的可靠性，用户会根据自己感知的风险大小决定对于信息的接受行为。在访谈过程中，多位受访者对伪健康信息的真假甄别普遍表示否定态度，但仍对部分伪健康文章表示半信半疑，学者研究也指出，国内中老年人群的健康信息素养整体偏低，中老年对网络信息辨别能力较差，网络谣言以及网络诈骗的时有发生导致部分中老年用户上当受骗，因此，伪健康信息甄别能力是公众健康水平提升的关键因素。访谈过程中发现受教育程度较高的用户较容易甄别伪健康信息，而受教育程度较低的用户受到知识水平的限制，对一些科学名词或理论机制难以理解。研究过程中发现个人性格也是影响甄别能力的因素，如“P35 我平时就不太爱看这些文章，我自己性格也不太喜欢这类文章”。若用户有过类似经验，则容易识别伪健康信息，例如“P25 我曾经颈椎就有一些问题，有了解过专业知识，所以知道哪些是真的哪些是假的”。

2.5.2 信息因素

中老年用户的信息接受行为不仅受到个人因素的影响，同样受到外部信息因素的影响。近年来微信平台中健康养生信息泛滥，研究过程中通过分析用户关注的信息类型发现，相较于药物健康信息，中老年用户更为关注饮食类和保健类健康信息，希望从生活方面提高健康水平，因此有关饮食保健类的健康信息更容易为中老年用户所接受。微信平台上健康信息获取便利，但也存在许多信息质量问题：由于信息内容重复且信息量大，需要用户有较强的信息甄别能力，从众多信息中寻找有用信

息。但是中老年用户的健康信息甄别能力较差。部分中老年用户受到诸多低质量健康信息的影响，对微信健康信息失去信任不愿接受。过度商业化的信息服务环境的侵蚀，导致用户对于健康信息产生误解，将健康信息与商业广告相混淆。信息服务质量也影响用户的接受行为，例如“P19 有时候那个朋友圈里发，然后看一眼就过了，不敢深入，因为这种东西你只要跟他说一句话，他马上呢他就要给你反馈好多，太麻烦了，穷追猛打的那种”，这就对微信健康信息质量提出了较高的要求。信息的有用性、易于理解性、可靠性等特征对于用户接受信息产生正向影响。有学者针对伪健康信息的特征进行研究，对于用户更好的辨别信息提供了理论支持。相关健康社交媒体应针对用户易于接受信息的特征发布健康信息，同时监管部门也应加大监管力度，创造更好的信息环境。访谈过程中发现用户更愿意相信权威机构发布的信息以及好友分享的健康信息，对于信任度高的好友所分享的信息更为信任，如“P20 就是，关系比较好的朋友给我发我就看一看，关系不好的，远的发的我就忽略了”；中老年用户对于权威机构所发布的健康信息信任度较高，但是我国现在权威机构的健康信息一般以公众号形式进行发布，而中老年用户对于公众号了解甚少，用户的功能性使用能力仍有待提高。

2.5.2 外界因素

外界因素对于微信平台健康信息接受行为的影响主要体现在家庭环境和社会环境方面。研究发现中老年人在无法对信息真实性进行判断时，倾向于听从子女的建议，子女对于微信健康信息的态度会直接影响中老年人对健康信息的接受行为。如“P2 想要和亲戚们分享，但是儿女们会提醒我们小心受骗，也就不太经常转发了”。社会环境主要是受到当前社会风气影响，中老年人既习惯于传统的信息了解模式，同时又在接触社交媒体信息，线上线下混合使用，构成了一个信息技术复杂的环境，在这个复杂环境中，用户更愿意接受大众所倾向观念的信息，如“P11 根据大众的舆论上的宣传和评判，自己做一下判断。主要是根据自己所了解的情况，结合电视上，大家说的”。

（二）第二阶段——基于文本挖掘的网络伪健康信息特征及情感分析

1.研究方法与过程

伪健康文本主题模型构建的方法如图 1 所示，共分为 5 步：①采集文本数据②数据预处理，清洗无效数据③情感倾向分析并对文本分类④采用 LDA 模型训练得到伪健康信息文本主题词分类⑤基于 CARS 列表综合文本特征与情感特征构建主题模型。

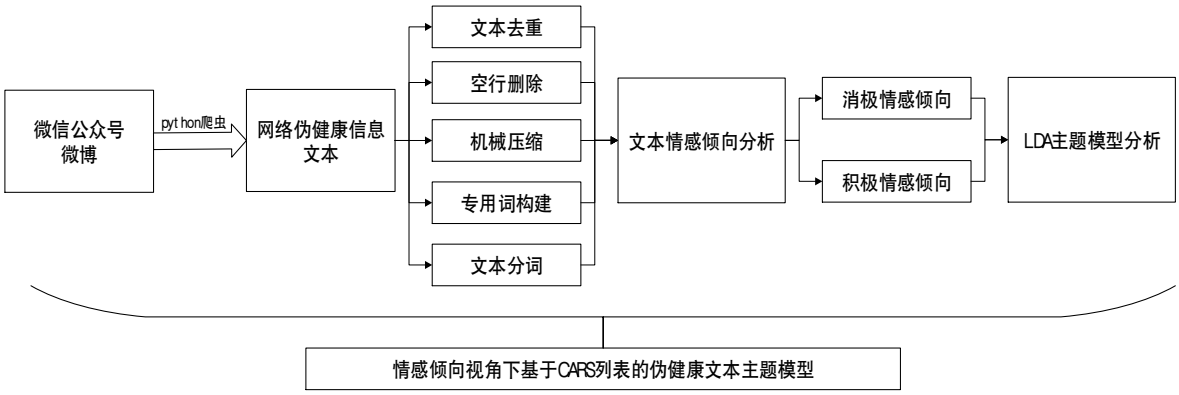


图 1 伪健康文本主题模型构建方法设计

2.数据搜集与处理

本文在微博平台上以健康相关词汇检索伪健康文章并通过微信小程序查找伪健康文章信息，反向找到对应公众号的相关文章，采用 fiddler 抓取网络数据包，进行反序列化，解析字段，最后存取数据库，共采集到 15076 篇伪健康文章，数据包括发布来源、文章标题、文章内容。

基于 python 获取的伪健康文本信息含有大量的空格、空行，以及价值低甚至无价值的评论冗余信息，这些信息会对情感分析造成干扰，所以通过文本去重与压缩、空行删除、停用词去除等操作来降低噪声。最终得到 13758 篇文本数据。此外，伪健康信息会涉及医药养生的专用词汇，需对文本数据进行 jieba 分词处理后，对专用词汇进行人工标注筛选并构建专用词典，为后续的情感模型分析奠定数据基础。

(1) 情感分析

本文利用自然语言处理（Natural Language Processing）、数据挖掘算法等对文本语言进行情感判断，从而把握文本观点、态度的计算研究。本文在 python 环境中调用百度 AI 平台的开源情感分析文档将文本的情感极性划分为消极、中性与积极三个层级，便于后面对不同层级情感倾向的文章进行 LDA 主题模型训练，从而研究不同情感倾向下的文本主题的区别。

(2) LDA 主题模型

潜在狄利克雷分布模型（Latent Dirichlet Allocation）是 Blei 等于 2003 年提出的一种文档主题生成模型，由于 LDA 能够降低文本表示维度，因此在语义挖掘领域得到了广泛应用。设采集 M 条伪健康文本，共有 N 个词，文本主题个数为 K，从 Dirichlet 分布 α 中取样生成文本主题的 Φ 词分布，根据词分布，取样生成相应的主题词 W。模型不断重复上述过程，直至所有文本采样完毕，最终得到每条文本的主题分布及各主题的词分布。

由于 LDA 主题模型是一种无监督模型，其中主题个数是模型重要的输入参数，为了保证模型构建结果的准确合理，本文采用困惑度（perplexity） 确定文档的最优主题数目。困惑度是用于评估模型优劣的标准，可用于调节主题个数，其计算公式如下：

$$\text{per}plexity(D) = \exp\left\{-\frac{\sum_{d-1}^{M}\log p(w_d)}{\sum_{d-1}^{M} N_d}\right\}$$

上式中，w_d 表示词，$p(w_d)$ 表示文档中词的概率，N_d 表示文档数量，D 表示文档中所有词的集合。使用困惑度进行评估时，主题越多，困惑度数值会逐渐下降；而主题数越多，LDA 模型计算代价越大。同时为了避免模型过拟合，应综合考虑选取困惑度数值和主题数目，选择困惑度最小和主题数最优的数值作为 LDA 模型训练的最优数目。

3.情感倾向分析

基于中文文本情感词典，计算 13758 条伪健康文本的标题和内容的情感得分。情感得分取值范围为[0,1]，判定积极和消极情感倾向的得分，并结合置信度对文本进行情感归类，研究伪健康文本的情感倾向分布，并将积极情感倾向文本和消极情感倾向文本区分开，分别进行 LDA 主题建模，根据计算结果，最终得到伪健康文本内容和标题情感倾向分布如图 2、3 所示，伪健康文本情感态度倾向两极化差异明显，伪健康文本在标题上积极情感和消极情感占比相同，在内容上积极情感占比较大，结合内容分析初步发现伪健康传播者常利用积极情感倾向的词语掩盖文章的效能不足，且主要针对中老年人等患病比例较大且获取信息渠道有限的群体。

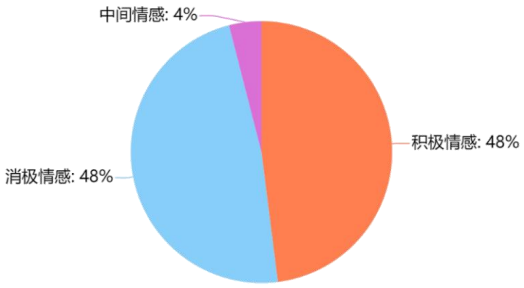


图 8 标题情感倾向分布结果

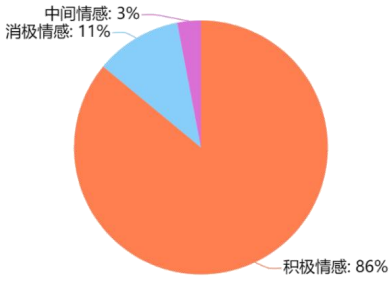


图 9 内容情感倾向分布结果

4.LDA 主题模型分析

在完成对文本的情感倾向分类后，分别对每一类文本进行 LDA 主题聚类，研究不同情感倾向下文本主题特征的差异，对标题和内容的困惑度进行计算确定主题个数，分类标准则依据困惑度计算公式，基于 python 软件的 gensim 包计算不同主题的主题词并根据主题词权重分布选取权重排名前十的主题词进行文本主题分析。

（1）消极情感倾向的伪健康文本分析

本文依据困惑度公式，计算出 2-10 区间内（间隔为 1）不同主题个数的困惑度数值，实验结果如图 4、5 所示，横轴显示主题个数，纵轴显示困惑度，由图可以看出随着主题个数的增加，困惑度波动变化。

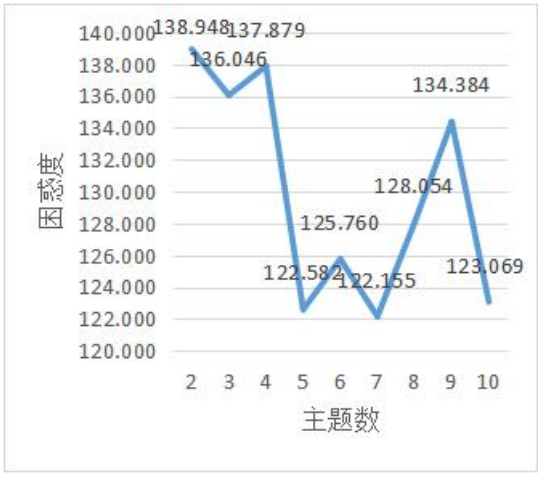


图 10 消极情感文本标题困惑度

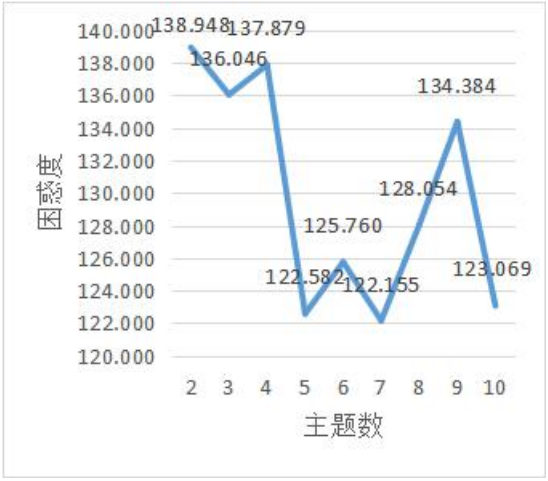


图 11 消极情感文本内容困惑度

如图 4 所示，当主题数为 3 时，困惑度最小，故最佳主题词数 K 为 3，结合最佳主题词数，筛选对文本内容描述价值最高的主题词作为关键主题词，再将每个话题下文本内容及其关键词进行汇总，具体见表 1，消极情感标题内容主要与医生、癌症有关，可见，用户认为涉及到专业知识及重大疾病带有一定负面性。该类标题主要从专业人员角度出发，给读者提出专业性较强，涉及医学知识的建议，如健康饮食可以降低患癌风险，以消极情感倾向带动读者情感波动，促使读者阅读文章的内容。如“告诉大家：医生偷偷吃的抗癌食物，每天坚持吃一点，活到百岁不得癌”这一标题从医生视角进行阐述，使用户产生信赖感，且标题将日常饮食和抗癌结合起来，诱导读者继续阅读文章内容。

主题 1		主题 2		主题 3	
词	权重	词	权重	词	权重
吃	0.053	医生	0.018	告诉	0.015
血压	0.049	吃	0.016	癌症	0.013
高血压	0.025	种	0.011	医生	0.007
医生	0.02	孩子	0.009	一定	0.007
种	0.018	提醒	0.006	农村	0.007
血管	0.007	食物	0.006	身体	0.006
喝	0.006	钟南山	0.005	千万	0.006
告诉	0.005	告诉	0.005	吃	0.005
肝	0.005	东西	0.005	女性	0.005
降压	0.005	健康	0.005	菜	0.005

如图 13 所示，当主题数为 7 时，困惑度最小，故将主题数定为 7，如下表 2 所示，可以发现消极情感的主题词多涉及各类癌症、女性患者、血压等话题。如胰腺癌这一话题文章通过夸大胰腺癌的致死率，造成读者紧张等负面情绪后进而介绍如何通过饮食料理来预防胰腺癌。

主题1		主题2		主题3		主题4		主题5		主题6		主题7	
词	**权重**	**词**	**权重**	**词**	**权重**	**词**	**权重**	**词**	**权重**	**词**	**权重**	**词**	**权重**
肺癌	0.024	患者	0.043	健康	0.228	孩子	0.237	医院	0.155	关注	0.356	乳腺癌	0.008
喝	0.02	营养	0.008	降低	0.028	女人	0.016	检测	0.010	点击	0.245	女性	0.006
获得	0.006	高血压	0.006	高血压	0.015	生活	0.010	疫情	0.008	生活	0.030	孕妇	0.006
措施	0.005	作用	0.006	飞花飞花	0.011	回忆未来1988	0.008	猪肉	0.008	本文	0.022	增生	0.006
脱离	0.005	飞花	0.005	血管	0.009	男人	0.007	人员	0.006	免费	0.020	更年期	0.005
吸烟	0.005	血管	0.005	食物	0.008	时间	0.007	口罩	0.005	妙招真帮手	0.010	年期	0.005
螺蛳	0.005	含有	0.005	患者	0.006	糊涂	0.005	病毒	0.005	孩子	0.008	宝宝	0.005
贝壳	0.005	好	0.005	导致	0.006	功能	0.005	感染	0.005	很多	0.005	雌激素	0.005
奢侈	0.005	食物	0.005	升高	0.005	关注	0.005	隔离	0.005	食物	0.005	原位癌	0.005
牵扯	0.005	功能	0.005	疾病	0.005	喜欢	0.005	肺炎	0.005	内容	0.005	乳腺癌	0.005

（2）积极情感倾向的伪健康文本分析

本文依据困惑度公式，计算出 2-10 区间内（间隔为 1）不同主题个数的困惑度数值，实验结果如图 6、7 所示，横轴显示主题个数，纵轴显示困惑度，由图可以看出随着主题个数的增加，困惑度波动变化。

图 12 积极情感文本标题困惑度

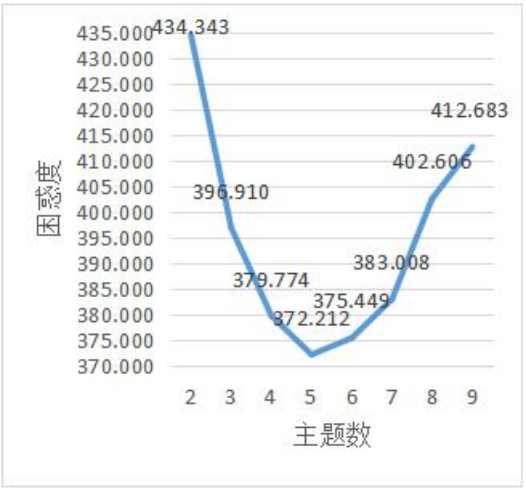


图 13 积极情感文本内容困惑度

如图 13 所示，当主题数为 4 时，困惑值最小，故最佳主题词数 k 为 4。如上文筛选方法，将每个话题下文本内容及其关键词进行汇总，详见表 3。可知积极情感标题内容主要与日常生活行为及常见心血管疾病相关有关，通过用户日常常见饮食行为和身体健康联系在一起，介绍健康饮食理念及具体食谱或其养生作用从而诱导读者继续阅读。如“高血压也有害怕的食物，每天更换着吃，清理血管，帮助降血压”这一标题将饮食养生与疾病联系起来，让读者产生好奇心理，以积极的情感倾向吸引读者继续阅读该文章。

表 14 积极情感文本标题主题词分布

主题 1		主题 2		主题 3		主题 4	
词	**权重**	**词**	**权重**	**词**	**权重**	**词**	**权重**

吃喝	0.027	血压	0.076	生肖	0.013	吃喝	0.053
年岁	0.016	高血压	0.031	家里	0.007	减肥	0.008
好吃	0.011	食物	0.02	减肥	0.005	肝脏	0.008
简单	0.01	降血压	0.012	学会	0.005	告诉	0.008
做法	0.009	天然	0.011	朋友	0.004	猝死	0.007
家常	0.008	血管	0.01	男人	0.004	食物	0.007
子宫	0.007	血糖	0.008	身体	0.004	女性	0.007
家常菜	0.007	降压药	0.007	告诉	0.006	天然	0.007
农村	0.007	稳定	0.006	早餐	0.006	健康	0.006

如图 15 所示，当主题数为 5 时，困惑度最小，故将积极情感内容主题数定为 5，如下表 4 所示，选取主题中权重比为 0.05 以上的主题词可以发现，积极情感主题内容主要和血压、血管、关注，以及其公众号名称（其中飞花飞花、妙招真帮手、回忆 1988 均为公众号名称），以某文章为例，该文章从“飞花飞花”微信公众号上爬取下来，介绍了芹菜等与降血压相关的食物，然而芹菜的降压作用非常微弱。而在本文中却过度夸大芹菜的作用和价值，并且通过介绍芹菜易于烹饪鼓动读者进行尝试，若读者刚好为高血压患者，则可能会造成延误治疗的危险。

表 4 积极情感文本内容主题词分布

主题 1		主题 2		主题 3		主题 4		主题 5	
词	权重	词	权重	词	权重	词	权重	词	权重
关注	0.558	吃	0.012	女人	0.008	血压	0.016	血压	0.021
请	0.007	好	0.007	生活	0.006	请	0.008	吃	0.015
点击	0.006	锅	0.007	男人	0.006	高血压	0.007	食物	0.008
妙招真帮手	0.006	放入	0.007	很多	0.006	活	0.005	血管	0.008
本文	0.005	适量	0.006	好	0.005	点击	0.005	作用	0.008
免费	0.005	克	0.006	喜欢	0.003	关注	0.005	身体	0.007
内容	0.005	炒	0.006	时间	0.002	本文	0.005	含有	0.007
分享	0.005	回忆未来 1988	0.006	孩子	0.002	免费	0.005	健康	0.006
天都	0.005	分钟	0.005	香烟	0.002	飞花飞花	0.005	患者	0.005
先点	0.005	做法	0.005	朋友	0.002	功能	0.005	营养	0.005

5.情感倾向视角下的伪健康文本主题模型构建

（1） 伪健康文本主题分析

5.1.1 伪健康文本标题特征归纳

（1）简洁凝练，表达形式以直观表述和设置悬念为主。在对伪健康文本标题长度进行分析，本文发现伪健康标题普遍较长，大多在 20-40 字范围内，且标题主要为

开门见山或设置悬念，通过标新立异的表达方式吸引用户关注，且大多与高血压、抗癌等话题相关，例如《钟老呼吁：被评上“一级致癌物”的东西，很多就藏在家中，找一下你家有吗？》故使得读者以消极情绪阅读文章。

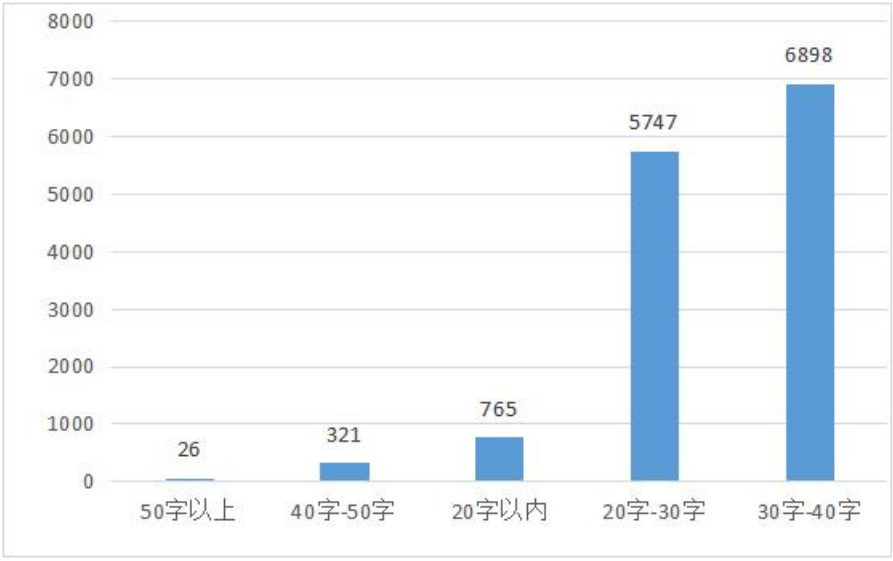


图 8 文本标题长度分布

(2) 语气词、感叹词频出。对伪健康文章标题进行分析时，发现大量标题都使用语气词进和感叹词以增强文章的主观情感，不仅不符合写作规范，同时降低了科学性和严谨性，通过感叹词和语气词的使用，对读者的心理进行试压，迫使读者续阅读下去。

(3) 常以专业人士身份进行叙述。在伪健康文章标题文本中，“一定”、“千万”、“必”等程度副词十分常见，这些程度副词常与“医生”、钟南山等专业人士相关，如《医院医生披露：心寒、肺寒、脾寒、肝寒、肾寒，一个方子，把五脏之寒统统散掉！》，增强文章的说服力，从而获取读者信任。

(4) 常使用重大或突发性疾病吸引用户关注。如“癌症”、“猝死”等突发性疾病是伪健康信息标题中高频词汇，如《3 岁男孩多器官衰竭死亡，医生： 这药别总给孩子吃，当心无药可医》等，这类与公众生命安全高度相关的疾病极易引起用户关注，用户倾向于阅读该类文章以规避文章中涉及到的行为习惯以确保生命安全，因此，关注并乐于阅读该类文章的用户大多在危机意识的驱动下，且在阅读后乐于分享、扩散该类文章。

5.1.2 伪健康文本内容特征归纳

(1) 叙事主体多为第一、二人称且身份模糊、假借权威。伪健康文本内容多采用第一人称或第二人称为叙述视角，通过营造交互氛围，增强用户对文章内容的依赖性。为增加文章说服力，标题常以专业人士名义进行阐述，从而提升文章的可信度，使得用户误以为确实是专业人士撰写该文章，并乐于认同文章观点。

(2) 叙事对象多与健康饮食相关。伪健康文章的传播者通过把控人们对于重大

疾病的关注热度和恐惧心理，将以癌症为主的疾病与时令果蔬结合，为果蔬虚构药用价值，引发读者重视关注，并进一步转发扩散文章，从而达到传播者的目的。

（3）附加社会热点，专家机构频出。在伪健康文本内容中，通常在开篇引入一些近期发生的时间或者社会轶闻热点等，从而增加文章的真实性，拉近与读者的距离，且将专家和机构的名称贯穿其中，试图利用受众对专业人士的信任来诱导其接受文章中的观点。

（2） 基于 CARS 列表的文本主题模型构建

基于 CARS 列表的四项评判指标，综合情感倾向分析和 LDA 主题聚类分析结果，本文从内容和受众两个角度将伪健康文章的主要群体分为两类，一类是患有高血压高血糖此类慢性病的中老年群体，另一类是身患重大疾病的的患者及其家属。

针对第一类读者，伪健康信息发布者主要以积极的情感态度撰写文章，为读者介绍对健康饮食及其他生活习惯对健康的积极影响，过分的夸大某种饮食方法或生活习惯对慢性疾病的功效以增强文章可信度。在准确性方面，片面的介绍某一种改善慢性病病情的方法以凸显文章准确性。在合理性方面，仅描述负面事实，含有强烈的个人感情色彩。

针对第二类读者，伪健康信息主要为消极情感，发布者多从专业人士的角度为读者提供健康知识及建议，如日常生活中有效预防及抗癌，通过声称一些独家或者机密的信息，否定读者的日常生活习惯，要求读者使用文章介绍的方法进行防癌抗癌。在准确性方面，同样具有观点片面的特征，在合理性方面，通过在标题和内容中过分申明文章观点，论述过于片面。且两类文章在相关支持方面，都存在数据缺乏来源、文本缺乏源文档来源、假借权威的问题。

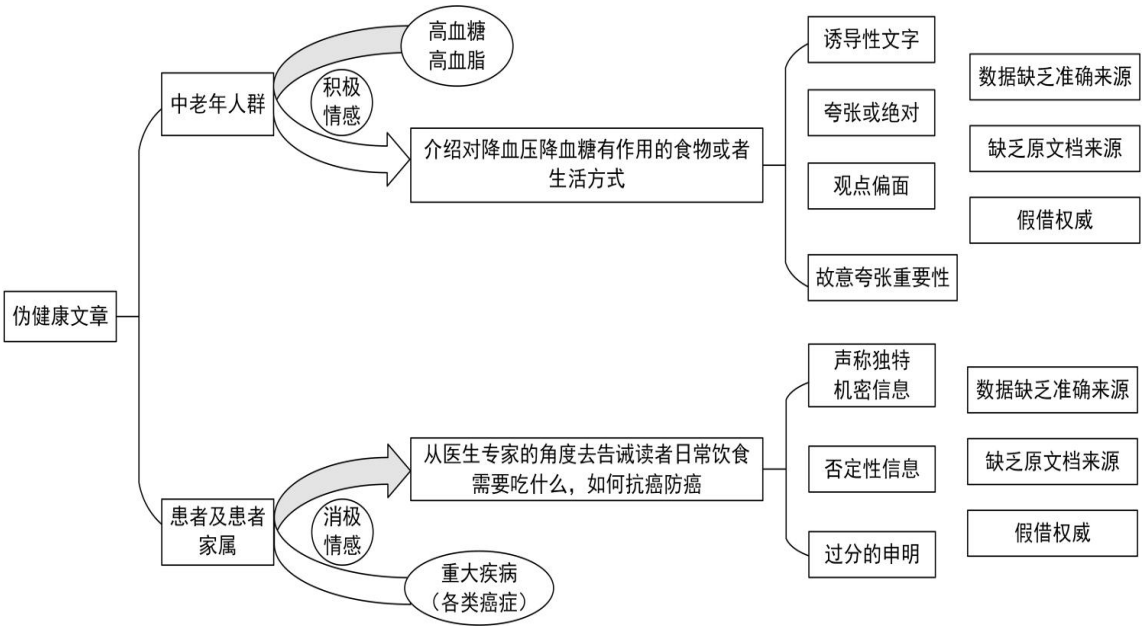


图 9 情感倾向视角下基于 CARS 列表的伪健康文本主题模型

参考文献

[1] 第 47 次《中国互联网络发展状况统计报告》发布 https://baijiahao.baidu.com/s?id=1691106020663057507&wfr=spider&for=pc

[2]中国网民科普需求搜索行为报告（2019 年第一季度） https://www.cast.org.cn/art/2019/4/26/art_1281_94546.html

[3]李月琳，张秀，王姗姗.社交媒体健康信息质量研究：基于真伪健康信息特征的分析[J].情报学报，2018,37(03):294-304.

[4]丁艳. 微信公众平台辟谣机制研究[D].山西大学,2020.

[5]王一迪. 老年人缘何分享伪健康信息？——基于社会支持视角的研究[C]. 北京大学新闻与传播学院.北京论坛·健康传播分论坛丨医疗、人文、媒介——“健康中国”与健康传播 2020 国际学术研讨会论文集.北京大学新闻与传播学院:北京大学新闻与传播学院,2020:429-437.

[6]Manafo E H，Wong S．Exploring older adults' health information seeking behaviors［J］．Journal of Nutrition Education and Behavior，2012，44(1) : 85 － 89