基于文本特征融合的衍生性网络健康谣言识别模型研究 *

陈燕方 [1]　周晓英 [2]
[1] 中国人民大学图书馆　北京 100872　[2] 中国人民大学信息资源管理学院　北京 100872

摘　要：[目的 / 意义] 衍生性网络健康谣言生成门槛低，周期性强，危害影响深远，是网络健康谣言识别与治理中需要优先解决的重点问题之一，也是重要突破口。[方法 / 过程] 借助深度语义表征和聚合方法，探索衍生性网络健康谣言文本内容的六要素特征；通过结合网络健康谣言的分布式语义特征预训练模型，构建包括六个类别、6 287 个词汇的网络健康谣言文本内容要素词库；在将健康谣言标题特征、内容文本六要素特征以及主体内容文本特征进行统一的向量空间表示与融合后，构建面向多源文本特征融合的网络健康谣言识别模型。[结果 / 结论] 模型的实证研究表明：与已有的对照模型相比，本文所提出的文本特征融合模型使衍生性网络健康谣言识别的准确率有较好的提升，且丰富的可拓展健康谣言要素词库可为后续的研究提供较好的资源支持。

关键词：网络健康谣言　健康谣言识别　文本特征　文本挖掘
分类号：R-05　G206
DOI：10.13266/j.issn.0252-3116.2023.14.008

网络谣言一直是社交网络信息传播中的痛点问题之一，其中，以食品安全、健康养生等主题为代表的网络健康谣言传播尤为广泛，且影响较恶劣。网络健康谣言主要包括两类：一类是被已有科学研究证明为捏造、子虚乌有，或被证实为典型的夸大其词言论，如“大蒜炝锅致癌”；一类是在业界和科学共同体内部尚未得到认可、明确缺乏支持依据的言论，如“PM2.5 三年堵死三分之一个肺”。

与一般网络谣言不同，网络健康谣言在传播过程中存在大量的衍生性谣言 [1]。所谓衍生性健康谣言指的是在既有健康谣言基础上，通过更换、增减、拼凑、调序等多种方式，进一步衍变而形成的。例如，通过更换既有谣言的陈述主体，事件发生时间、地点、人名或相关数字，进而再次翻新传播的系列谣言，“XX 市人民医院有人感染 SK5 病毒死亡”就是典型的衍生性健康谣言。该健康谣言前后出现过“临汾市人民医院”“宝鸡市人民医院”“湛江市中医院”等不同医院名称，而谣言主体也先后出现过“千万别吃猪肉 / 西瓜 / 梅子 / 李子 / 酸菜鱼”等不同食物名称。

衍生性健康谣言的典型特征是：谣言文本内容之间存在较强关联性 [2-3]；谣言传播存在一定的周期性 [4-5]。衍生性健康谣言看似是新谣言，实则是典型的换汤不换药型旧谣言，因为其传播逻辑以及词语的主题分布并未发生太大的变化，故而文本内容存在较强关联性。例如由“酸碱体质”衍生出的“酸性体质易生病”“碱性食物可抗癌”等系列谣言。同时，衍生性健康谣言每经历一个传播周期后，只要遇上合适的时机，就会被翻谣者（指对原健康谣言内容进行重新编排的传播者）“乔装打扮”，重返社交媒体，再次进入公众视野并流传开来，因此具备一定的周期性。根据奥尔波特等提出的谣言传播公式：谣言的流行性 =（事件的）重要性 X（事件的）模糊性 [6]。衍生性健康谣言的周期性传播原因主要在于两个方面：一是健康谣言信息事关公众健康，总能引起人们的关注（重要性）；二是辟谣与治理的不到位，导致公众对健康谣言的认识始终是模糊的（模糊性）。此外，再加之刺激因素的出现，比如季节更替、突发公共卫生事件的爆发，使得这类健康谣言被频繁加工、反复传播，继而发展成为衍生性健康谣言。

从衍生性健康谣言的生成方式及其特征来看，其生成门槛低，周期性强，导致有些流传广泛、历时久远的谣言深入人心，给谣言治理带来极大困难。同时，

* 本文系中国人民大学公共健康与疾病预防控制文理交叉重大创新平台“中央高校建设世界一流大学（学科）和特色发展引导专项资金”和国家社会科学基金重点项目“全媒体语境下的信息流行病学理论与实践研究”（项目编号：20AZD132）研究成果之一。
作者简介：陈燕方，馆员，博士；周晓英，教授，博士生导师，通信作者，E-mail: xyz-ruc@qq.com。
收稿日期：2022-12-07　修回日期：2023-04-16　本文起止页码：73-84　本文责任编辑：易飞

衍生性健康谣言文本内容的强关联性也是其被自动识别的重要突破口[3]。因此，研究人员需要密切结合衍生性健康谣言的显性特征构造具备针对性的识别方法，以优先解决由衍生性网络健康谣言带来的增量传播问题。

综上，为了有效识别衍生性网络健康谣言，本文首先借助深度语义表征和聚合方法，分析了网络健康谣言衍生性文本内容的要素特征，然后基于网络健康谣言的分布式语义特征预训练模型抽取了健康谣言文本的标题、词语和文本特征表征，同时构建了由六个类别、6 287个词汇构成的网络健康谣言文本内容要素词库；最后将健康谣言标题特征、内容文本六要素特征以及主体内容文本特征进行统一的向量空间表示与融合后，构建面向多源文本特征融合的衍生性网络健康谣言识别模型。

1 研究现状

谣言识别与预警是谣言治理中的重要一环，也是自然语言处理（Natural Language Processing, NLP）领域的研究热点之一，其研究最早可追溯到21世纪10年代前后。早期的谣言识别研究以传统机器学习内容分类法为主，即通过人工定义特征并定量化后，作为模型输入来源，然后结合相关分类算法进行分类筛选[7]。其中，定义特征主要从文本内容[8]、用户节点[9]和传播网络[10]三个维度提取并组合；分类算法则包括决策树、关联规则等基于规则的分类法和朴素贝叶斯、支持向量机等基于统计的分类方法。近年来，随着深度学习技术的不断发展与广泛应用，研究者们开始将循环神经网络（Recurrent Neural Network, RNN）[11]、卷积神经网络（Convolutional Neural Networks, CNN）[12-13]、图卷积网络（Graph Convolutional Network, GCN）[14]等模型应用于谣言识别问题，突破了以往需要人工定义特征的局限，在挖掘表示谣言传播各阶段文本的深层特征方面取得了较好效果，是谣言识别的新兴研究方法。但其不足之处在于，谣言类型多样，形式多元，既涉及社会、政治、历史、健康等多类领域，又包括主观臆测和客观篡改等多类传播形式。对于以主观臆测为代表的政治、历史类谣言是比较难以论证证实的，因此谣言自动识别研究多聚焦于现有科学技术方法和知识体系可论证、可识别的谣言范畴。

近年来，随着健康传播和健康信息学关注度的不断上升，健康领域的谣言治理问题逐渐成为谣言治理研究的重要分支。尤其自新冠疫情暴发以来，突发公共卫生事件下的谣言预警与治理研究一度成为相关学科研究的热点之一。在研究数据集上，由于公共的健康谣言语料库较少，目前研究多基于自主采集数据的识别方法[15]，且多以推特、微博等短文本数据集为主[16]。但健康谣言不同于一般谣言，并非仅是一个观点的表达，通常是谣言信息与真实信息穿插叙事的中长文本，具备更强的传播性，这也是健康谣言风靡微信公众号的重要原因之一。因此，近来对于健康谣言长文检测的研究也开始受到关注[4,17]。在抽取特征上，从传统质性分析方法[18-19]到深度学习技术[20-21]，技术越来越前沿，从语言统计特征[22]、网络特征[23]、用户特征与结构特征的结合[16]以及多维特征模型[24]到用户感知的视觉特征[25]、交互特征[26]等，特征挖掘的维度也越来越多元。但是对健康谣言区别于一般谣言的形式特征挖掘较为薄弱，这种形式特征并非简单的语言统计特征，而是在表达结构和语境方面表现出的关联特征，尤其是中文语境中。在识别方法上，与传统谣言识别中的机器学习分类法[27]相比，深度学习方法[28-29]作为后起之秀备受青睐。

综上可知，数据集、特征与模型是健康谣言识别问题的三要素。其中，特征抽取是抓手，如何将抽取的特征与合适的算法模型相结合则是关键。健康谣言的识别与治理是一个社会性的研究问题，并非一个单一的技术问题。因此，一味追求模型的前沿性，而忽视本文引言中分析的衍生性健康谣言特有的文本特征（衍生性与关联性）与传播特征（周期性与反复性），很难真正解决健康谣言泛滥传播的困境。正如朱梦蝶等[15]指出的，结合图情领域的特色与优势，实现社交媒体中健康谣言信息内容的聚类和序化，进而对健康谣言进行更深层次的处理和利用是重要的探索方向。对此，本文提出充分挖掘并利用衍生性健康谣言文本内容的关联性和主题性特征，以探索衍生性健康谣言识别的突破口。

2 数据来源与数据准备

本文构建衍生性健康谣言识别模型所需的数据主要包括两部分，一是衍生性健康谣言数据集的获取，主要用于衍生性健康谣言内容特征的分析以及识别模型测试实验；二是构建衍生性健康谣言识别模型所需的健康信息语料库。

2.1 衍生性健康谣言数据集准备

“谣言过滤器”是微信在2014年发布的辟谣官

方账号，它与人民网、丁香园、果壳网等主流媒体、商业性和公益性等专业辟谣机构合作，将诸多专业辟谣机构的辟谣内容和查证结果进行聚合，以对微信上近期传播广泛，影响恶劣的谣言进行辟谣。据“爱微帮”数据显示，“谣言过滤器”中，52.35% 的发文来自健康领域 [30]。由于微信常常成为健康谣言发源地和聚集地，且存在大量的衍生性健康谣言，将微信的谣言文本作为案例分析极具意义。因此，选取微信辟谣中心和“谣言过滤器”小程序中被辟谣的谣言文章作为衍生性健康谣言种子语料，进一步地，针对每一条种子语料，我们将其标题在搜狗微信搜索（支持搜索微信公众号和微信文章）中进行检索，会得到大量与该种子语料核心观点一致（即谣言内容所阐述的观点），但结构和表述略有差异的微信文章，即本文所定义的衍生性健康谣言。按照这一标准，共人工筛选出 502 条衍生性健康谣言原文作为样本。

2.2 健康信息数据集准备

在识别构建前，我们首先需要明确模型的应用数据场景，并为之准备适用的数据集，且由于模型均基于网络健康谣言文本内容展开，需要获取大量和健康内容主题相关的文本信息作为模型的训练语料，以此保证健康谣言相关内容要素库的高效构建。本文为获取大规模文本内容所构建的分布式网络健康谣言爬虫框架如图 1 所示：

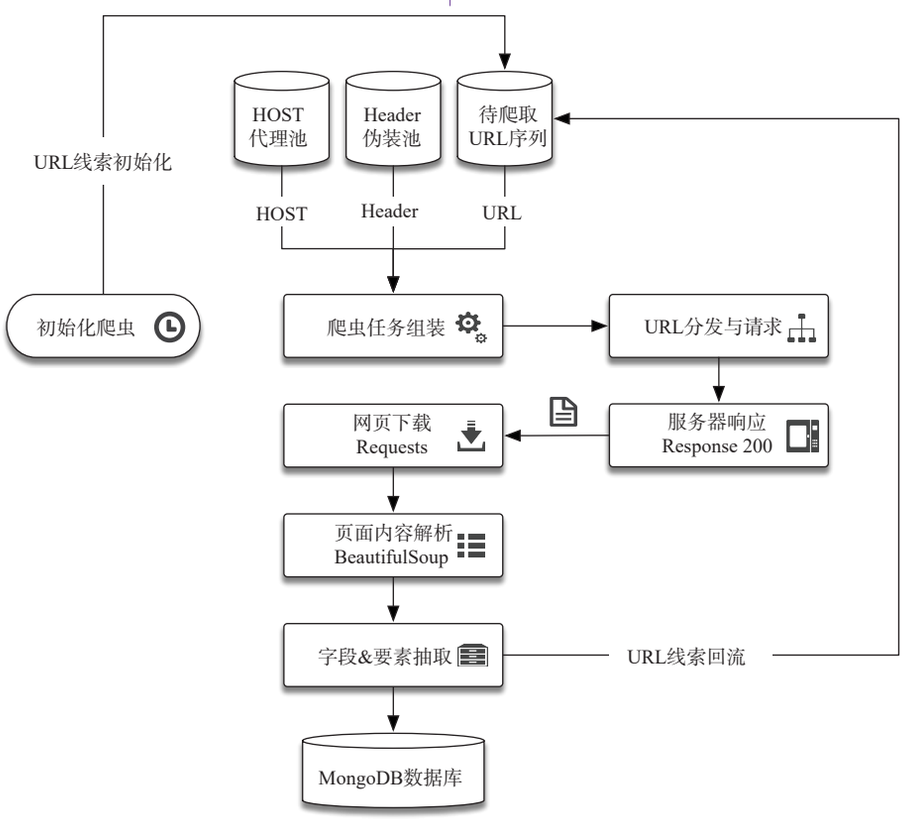

图 1 分布式网络健康谣言爬虫框架

其中，HOST 代理池和 Header 伪装池用于防止目标网站屏蔽爬虫采取的动态 IP 策略和请求头部策略，通过控制请求频次、IP、请求 Header 来源能够有效地增强爬虫的爬取效率。并且通过 Scrapy 库构建主从式分布框架实现，主节点负责爬虫任务组装、调度与管理，从节点负责数据爬取、内容解析与数据存储。基于组装后的爬虫，通过目标网站的 URL 即可请求得到对应的网页内容数据。然后，利用网页内容解析包 Beautiful Soup[31] 即可抽取得到目标字段（如文本内容、作者、时间等）和下一步爬取的 URL 线索。最终，为了保存混合得到的多种信息，如文本、图片等，笔者选取了当前流行的非结构化数据库 MongoDB[32] 进行数据存储，并开源了爬取的原始数据集（本文涉及的原始文本语料下载地址：https://pan.baidu.com/s/1URcXc_WcKQcAMoByIbl9cA?pwd=pkq2）。

如表 1 所示，典型的数据集包括 4 类：健康相关的百科数据集、新闻数据集、自媒体健康相关文章、健康相关谣言文本。数据集具体获取的有效字段包括：文档 ID、文档标题、文档内容、文档作者、文档发表时间、文档简介、文档标签、文档图片数量 / 存储地址等。

表 1 预训练数据集来源和样例

序号	数据集名称	数据内容 & 样例	文档数量 / 条	采集网址
1	健康百科	主题：健康问答相关文本文档 举例：青春期的女孩子会变声吗？ (https://baike.120ask.com/art/454451)	424 376	https://baike.120ask.com/
2	人民健康网	主题：权威健康新闻 举例：北京本来生活自营店豆腐皮铅超标（http://health.people.com.cn/n1/2019/1012/c14739-31395784.html）	13 960	http://health.people.com.cn
3	搜狐健康网	主题：健康相关自媒体新闻 举例：宝宝多大可以吃盐？辅食中加多少盐合适？（https://www.sohu.com / a/377683768_601022）	162 553	https://health.sohu.com/
4	网络健康谣言	主题：网络健康谣言原文 举例：掉头发不能吃什么食物？当心会掉得更多（https://i.7y7.com/mi p/234061.html）	502	辟谣媒体、各类网站

3 衍生性网络健康谣言文本内容的特征分析

衍生性网络健康谣言是在既有健康谣言基础上加工形成的，因此本节基于衍生性健康谣言原始语料，利用内容向量特征表示学习方法，对语料库的主题内容进行聚类分析，从而得到当前衍生性健康谣言文本内容的六要素特征框架，并进行详细的内容分析。

3.1 衍生性健康谣言文本内容主题聚类分析

3.1.1 聚类步骤

为充分了解衍生性健康谣言文本内容特征，我们对 502 条衍生性健康谣言进行主题关联聚类分析，具体主题聚类步骤如图 2 所示：

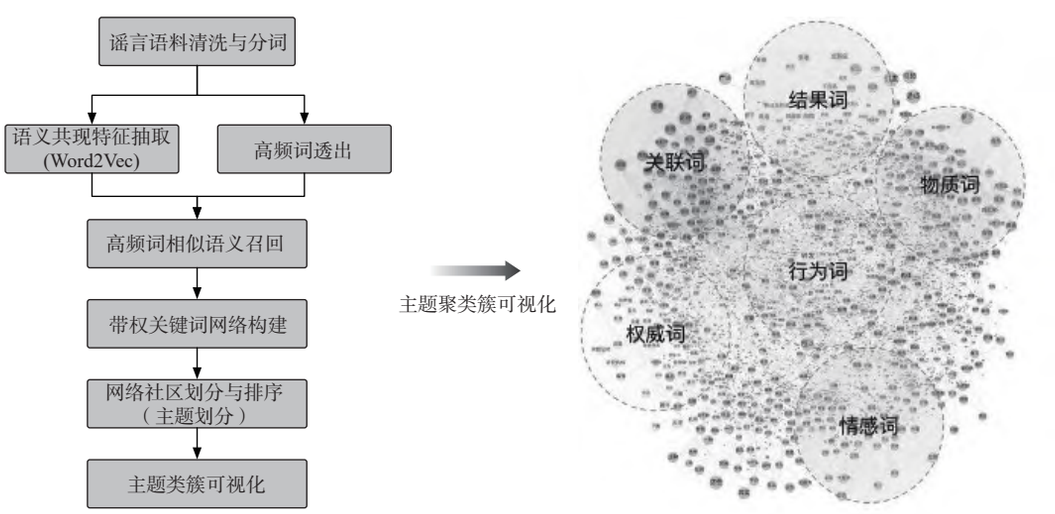

图 2 衍生性网络健康谣言文本内容的主题词聚类网络

步骤一：语料预处理与词向量表征学习。对语料进行基础的清洗和分词操作，包括编码归一（标点符号、大小写归一等预处理）、分词（使用 Jieba 分词实现）、停用词去除等，由此得到洁净的语料序列。接着，使用 Word2Vec 模型在上述语料上学习得到衍生性健康谣言词语之间的语义共现特征，并将其映射到固定的低维向量空间（Word2Vec 的相关原理详见 4.1）。

步骤二：语义关联网络构建。基于统计到的 K 个高频词语，并在衍生性健康谣言语义共现向量空间中进行语义近邻搜索，分别得到每个高频词的 N 个相似词语及其对应的相似度，每条边可记为（high_freq_words, simi_words, weight），从而得由 K*N 条边聚合成的带权语义关联网络。

步骤三：网络社区划分（主题划分）与节点重要性排序。基于得到的词语关联网络，利用模块度最大化社区划分方法（Louvain 算法）对网络中节点所属的主题社区进行分类，从而得到每个词语所属的主题簇。同时，在网络关联图中使用带权 PageRank 算法得到各节点的相对重要性排序。

步骤四：主题划分结果可视化。对得到的结果进行可视化，其中相同圈内的节点表示该词语属于相同的主题簇，同时节点越大，表示该节点的 PageRank 得分越高，即在衍生性健康谣言中的提及率和共现率越高。

3.1.2 聚类结果

在分析文本词频之后，本文对衍生性网络健康谣

言文本的主题关联网络进行了拆解，具体而言，提取了如下六类典型的主题网络，包括物质词网络、关联词网络、效果词网络、行为词网络、情感词网络和权威词网络。具体聚类结果如下：

（1）主题类型一：物质词。物质词一般是衍生性网络健康谣言中被认为影响公众健康的主体，如蛋糕、饼干中隐藏的反式脂肪酸、高铁中的辐射、不粘锅涂层中的“特富龙”、花生芽中的白藜芦醇等。从聚类结果来看，物质词可以分为三种：①原材料食品中具体包含的成分类词汇，如脂肪酸、活性元素、蛋白质等；②原材料食品类词汇，如西瓜、小龙虾、香蕉、柠檬等；③日常环境相关物质词汇，如甲醛、辐射等。

（2）主题类型二：关联词。关联词是表达某物与健康关联的一类词。衍生性网络健康谣言所传达的意图主要包括两大类，即某物有益于健康或某物有害健康。相应地，关联词可分为积极关联词和消极关联词。积极关联词主要包括增强、提高、促进、调节、治疗、排出、淡化、清理等有益于健康的动名词，消极关联词则包括阻止、导致、致（癌）、掉（发）等有害健康的动名词。

（3）主题类型三：效果词。效果词是表达导致健康结果的一类词。紧接在关联词之后的效果词通常与一些常见的疾病相关，如鼻炎、颈椎病、肾虚、体寒、脑溢血、糖尿病……当然，在效果词中占比最高的是“癌症”。由于各类癌症病因复杂且致死率高，在治疗过程中伴随着各种痛苦体验，这一系列因素都使癌症成为衍生性网络健康谣言造谣者手中屡试不爽的杀手锏。公众秉承“宁信其有，不信其无”的心理，总是谈癌色变。此外，随着人口老龄化的加剧、疾病谱（指将疾病按其患病率的高低而排列的顺序）从传染性疾病向慢性非传染性疾病转变，各类慢性疾病也成为了效果词网络图中的榜上客。

（4）主题类型四：情感词。情感词是形容状况和调动情绪的一类词。衍生性网络健康谣言为了达到较好的说服性效果，多喜欢采用绝对化、鼓吹式语气或数字来调动读者的情绪。具体包括：①“99%的人不知道”“100% 丧命”“史上最全”“千万不要……”等绝对化数字或词语；②“太恐怖了”“太棒了”“太让人意外了”“强一万倍”等夸张式语气；③此外“！”也是造谣者的杀手锏，常被用来加强语气。这些打着情感关怀旗号的强感情色彩词语，旨在通过感性压制住公众的理性，达到情绪唤醒的目的。

（5）主题类型五：行为词。行为词是引导公众做或不做某些行为的一类词。网络健康谣言造谣者的最终目的是说服或引导公众的行为，既包括日常健康生活行为也包括信息行为。健康行为即诱导公众按照健康谣言的内容去实施某些与我们健康相关的具体行为，如不吃龙虾，不同食螃蟹和水果，每天喝绿豆汤抗癌等。信息行为则主要是诱导分享，可以说绝大部分的衍生性网络健康谣言都包括了诱导分享体。

诱导分享体主要包含以下几类：①直接简单诱导分享体，如“千万人收藏了”“赶紧分享出去”等；②道德绑架类诱导分享体，如“转发可救无数人”“转给你爱的人”“不转不是中国人！”等；③迷信蛊惑类，如“转发可保全家平安”“不转发必将厄运连连”等；④物质诱惑类，即转发后可以获得一定的金钱奖励或享受某项健康产品服务，如“转发即可到店享受免费洗牙”等；⑤条件强制类，即转发了才能浏览全文或获得测试结果等。

（6）主题类型六：权威词。权威词是指为使公众相信而使用的包括来源、依据等内容的一类词，权威词、中英文专业术语词的大量出现揭示了衍生性网络健康谣言的典型特征，即伪造权威。网络健康谣言通常披着伪科学的外衣，喜欢以权威身份发言，包括假借国外专家、重点实验室身份，断章取义引用“权威”实验结论，大量堆砌专业术语，使健康谣言看起来有理有据，进而让部分用户深信不疑。以调研样本为例，“中美肿瘤专家最新发现”“食品总局发布”“CCTV 新闻”等大量伪造公文、伪造权威的词汇出现在标题或正文中。此外，还含有大量医学、生物、化学等领域的中英文专业术语。这些“有理有据”的专业化和科学化外衣，提高了公众对于网络健康谣言理性判断的门槛，人们难以根据已有的知识和经验立即判断出健康谣言的真实性，同时在碎片化阅读时代，通过网络搜索去验证信息的行为也鲜有发生。

3.2 衍生性文本内容六要素特征分析

根据上文的主题聚类结果，笔者认为一条典型的衍生性网络健康谣言通常包括六要素特征，即物质词 + 关联词 + 效果词 + 情感词 + 行为词 + 权威词。物质词包括成分相关词汇、食品相关词汇、日常环境相关词汇等；关联词则分为正向作用健康和负向作用健康两类；结果词则多是相关疾病；情感词多为夸张式语气词；行为词包含健康行为诱导词和信息行为诱导词；权威词则多为中英文专业术语、权威机构或知名学者等。

一条典型的衍生性网络健康谣言样例如图 3 所示，叙事信息为“生番茄中的龙葵碱可致死”，具体而言：物质词为“生番茄中的龙葵碱”；关联词为“导致”；效果词为“死亡”；情感信息为“震惊！”；行为信息中，健康行为信息为“不要再生吃番茄”，信息行为信息为“速速转发！”；权威信息则为“英国临床医学博士托马斯·莎耶博士”。

震惊！千万不要再吃生番茄了！据英国临床医学博士托马斯莎耶博士证实，生番茄中含有的龙葵碱可导致死亡的！为了您和您家人的健康，请速速转发！

图 3　一条典型的网络健康谣言样例

以上衍生性网络健康谣言六要素特征公式的提炼，对网络健康谣言要素特征词库的构建具有指导意义，可以更准确地判断衍生性网络健康谣言，以不变应万变。

4　衍生性网络健康谣言内容识别模型

基于以上分析，可以看到，衍生性网络健康谣言具备高度的可识别分类特征，因此通过构建基于预训练的特征表征，能够有效增强模型的识别效率和应用范围。接下来，将阐述完整的识别模型构建流程，包括：首先，使用文本预训练模型，得到网络健康谣言文本语料基础的分布式语义特征表征向量（词向量特征表征、标题向量特征表征和文本内容向量特征表征）。其次，基于预训练得到的词向量库扩展衍生性健康谣言的六要素种子词库，从而保证要素词库的丰富性。最后，基于上述准备的标题特征表征、六要素特征表征以及内容特征表征，构建文本特征融合模型，从而完成最终的识别过程。

4.1　分布式语义特征抽取

近年来随着深度学习的兴起，词语的连续特征表示学习（Distributed Representation of Words, DRW），即语言特征的预训练模型，在自然语言处理的各个领域得到了广泛的应用，并取得了令人瞩目的效果。DRW 模型的原型思路最早是由 D. E. Rumelhart 和 G. E. Hinton 等在 1985 年提出 [33]，在该文中，作者指出通过反向传播算法能够让神经网络发现目标数据的分布表示。这一重要发现也直接引发了当前有关文本、图片等多种类型的特征表示方法的提出。典型的应用包括 2013 年由 T. Mikolov 等提出的 Word2Vec[34] 模型，该模型将传统的神经网络语言模型简化后，提出基于大规模数据集的分布式词语特征表示学习方法。其本质是基于三层经典的神经网络模型学习词语间的共现特征，以此将词袋表示方法的离散高维特征空间映射到连续的低维向量空间，从而快速得到词语的特征表示。此后，Sentence2Vec/Doc2Vec[35] 等一系列的文本表示模型相继被提出，此类模型在 Word2Vec 的思路和基础上，以更加具体的表示目标，对文本的各种要素，如句子、段落、篇章等进行表示学习，从而更加有效地提取出针对下游应用所需的连续特征。

本文采用针对词语、句子和篇章对网络健康相关语料的标题、内容的文本特征表示进行学习的方法，构建针对健康文本的预训练特征集合（Pretrained Vector for Health Corpus，PVHC）。具体预训练流程如图 4 所示：

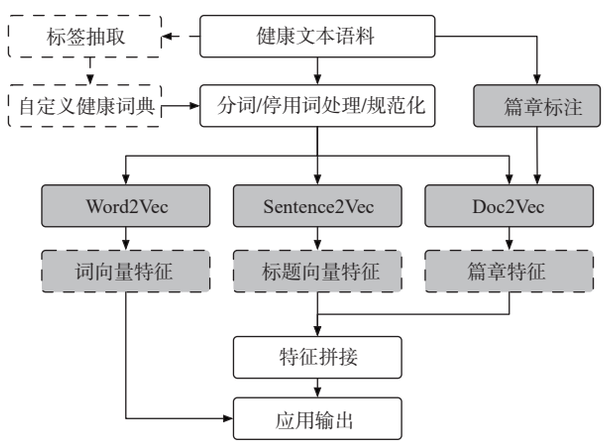

图 4　网络健康语料文本特征向量预训练流程

首先将爬取得到的网络健康谣言相关的语料文本（谣言 / 非谣言）进行语料标注、训练内容划分和数据清洗，从而得到后续训练所需的基础要素。其中由于健康相关词语的独特性，本文基于文档标签清洗得到 6 515 个与健康相关的专业词语，并以此作为用户词典参与分词，提升针对健康领域的语料分词效果。然后，基于清洗和规范化后的语料集合，利用 Word2Vec、Sentence2Vec 和 Doc2Vec 三个特征向量表示学习模型进行训练，分别得到文本词语、标题内容和文本内容连续特征向量表示，为后续模型的应用奠定基础。本文实验配置的硬件环境为：i7-4790 @ 3.60G Hz 处理器，24GB 内存，Windows7 操作系统，GPU 为 NVIDIA GeForce GTX 980Ti。软件环境为：公共基础环境为 Python 3.6 Anaconda 科学计算环境，Word2Vec 以及 Doc2Vec 模型的训练均使用 Python Gensim 2.1.0 框架处理包完成，基础分类器模型使用 Python Sklearn 机器学习包完成，GPU 加速驱动版本为 CUDA9.0 以及 cuDNN7.0.5。

4.2 衍生性文本六要素词库的构建

为提升模型的识别效果，除了对主体文本内容的全局特征进行表示外，还需要针对衍生性健康谣言相关的核心主题词语进行表示，其中重要的一步即构建谣言要素相关的词库。按照上文六大主题聚类分析结果，当前的衍生性网络健康谣言多由六类要素组成，即物质词、关联词、效果词、情感词、行为词和权威词，本文基于已训练得到的词向量集合构建了与网络谣言要素相关的底层词库，该词库的具体构建流程如图 5 所示：

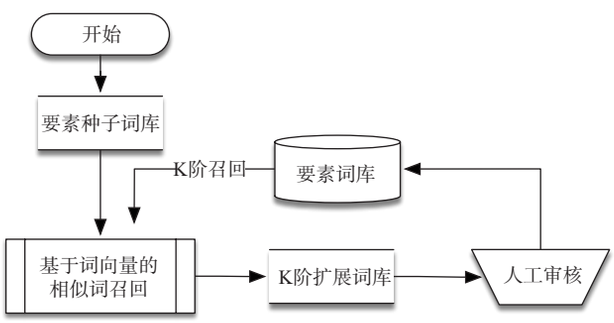

图 5 网络健康谣言信息内容要素词库构建流程

首先，基于已知的部分健康谣言内容，针对每一类网络健康谣言内容文本要素构建 10—20 个种子词语，即 SeedWords=$\{s_1, s_2, \cdots, s_n\}$，并基于得到的种子词语计算它们在已训练健康相关语料的向量空间中的 Top N 个种子相似词语（N 通常在 20—100 之间），从而得到种子词的一阶相似词语，即：

$$Recall_1 = \bigcup_{s_i \in \text{SeedWords}} \text{Sim}(s_i) \qquad \text{公式（1）}$$

其中，Sim(x) 为向量相似函数，用以获取特定词语 x 在向量空间中的相似词语列表。然后对该一阶要素词语进行人工审核，保留正确的词语，从而得到有效一阶要素词库。接下来，基于得到的有效一阶要素词库再次进行向量空间的相似性词语索引即可得到二阶要素词语库，即：

$$Recall_2 = \bigcup_{s_j \in Recall_1} \text{Sim}(s_j) \qquad \text{公式（2）}$$

以此类推，即可得到 K 阶要素词库：

$$Recall_k = \bigcup_{s_x \in Recall_{K-1}} \text{Sim}(s_x) \qquad \text{公式（3）}$$

基于上述过程进行贪心迭代，直到某轮索引后，词库有效新增词语数为 0 为止。最终，通过上述词库拓展过程，共计得到六个类别的要素词共 6 287 个。构建的网络健康谣言要素词语库的典型样例如表 2 所示：

表 2 网络健康谣言要素词语典型样例库

要素类型	样例词语举例
物质词要素	脂肪酸，龙葵碱，酸性物质，亚硝酸盐，多环芳烃，苯并芘，阿斯巴甜，麸胺酸，黄麴毒素，抗生素，钠，汞，多环芳烃，膳食纤维，氯化物，等等
关联词要素	导致，产生，引起，加重，预防，降低，稳定，远离，赶走，消灭，清除，致使，引致，使之，加深，等等
效果词要素	食道癌，胃癌，铅中毒，贫血，动脉粥样硬化，骨质疏松，老年痴呆症，高烧，头痛，气喘，抽搐，吐血，昏迷，死亡，呕吐，头晕，腹泻，等等
情感词要素	沸腾，千万不能，100%，很重要，切记，坏消息，震惊，请三思，震撼全球，顿时傻眼，这么神奇，蒙在鼓里，拯救更多人，太棒了，等等
行为词要素	吃，喝，食用，生吃，不要吃，摄取，拿来吃，转寄，少吃，喝一口，带皮吃，同食，扔，转发，必须转，等等
权威词要素	院士，研究员，博士，食品总局，科学类奖项，专家，医院，医生，卫生部，CCTV，妇科主任，北京医院，权威教授，华佗，钟南山，等等

4.3 基于文本内容特征融合的识别模型构建

基于文本内容特征融合的衍生性网络健康谣言识别模型如图 6 所示。该模型基于前两节产生的基础文本特征向量库，构建融合的特征识别框架。通过将文本的三类重要特征：标题特征、健康谣言内容文本六要素特征以及主体内容文本特征进行统一的向量空间表示后，将抽取得到的特征进行串联，最大程度保留有效信息，从而得到分类器训练所需的特征集合。最后，使用各类分类器，如逻辑回归（Logistic Regression, LR）、梯度提升决策树（Gradient Boosting Decision Tree, GBDT）或支持向量机（Support Vector Machine, SVM）等，进行有监督的分类模型训练。模型的核心步骤包括三个方面：三类文本特征的统一向量空间表示，文档特征的融合，以及衍生性健康谣言识别模型的分类器训练。

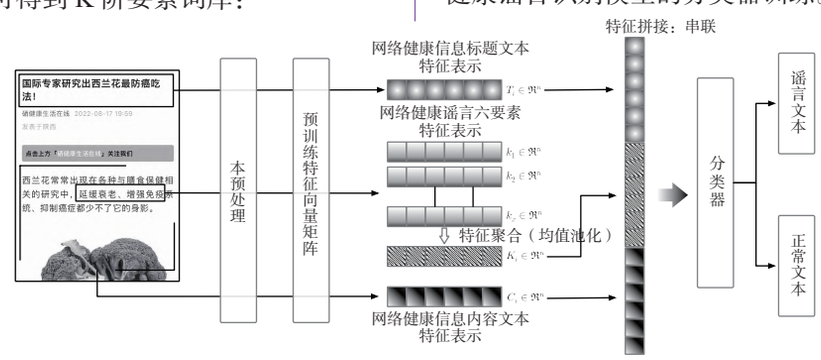

图 6 基于文本特征融合的衍生性网络健康谣言识别框架

步骤一：三类文本特征的统一向量空间表示。目标主体的特征抽取优劣对于机器学习的最终学习效果有着较为重要的影响，通过优化特征表示方法不仅能够提高模型的准确性，更能够增加模型的可扩展性。对于每一篇衍生性网络健康信息文本，本文将其文本内容特征概括为三个类别，即：标题内容特征、谣言六要素内容特征以及主体内容特征。三个类别的抽取样例如图 7 所示：

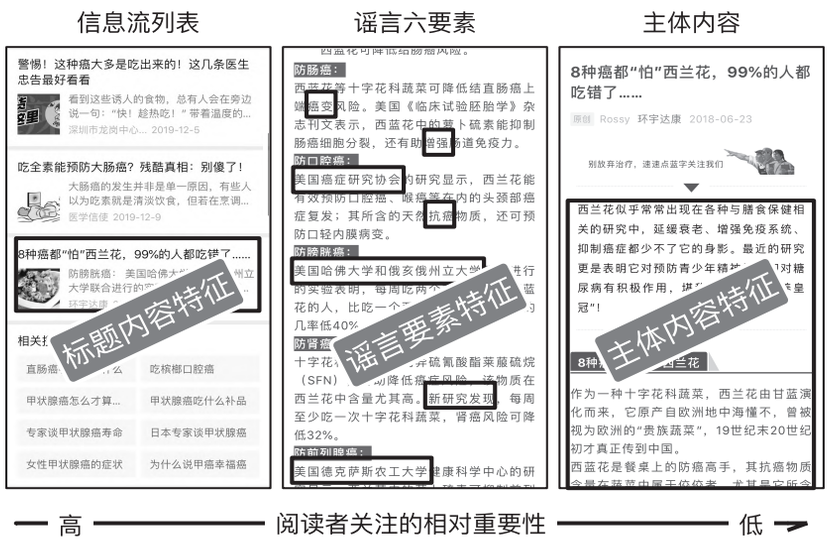

图 7　三类衍生性网络健康谣言文本内容特征的抽取样例

首先，从用户视角出发，思考和构建特征框架。用户阅读一条典型衍生性健康谣言是从标题列表页，到要素特征（即用户特别关心的一些特征词，如：致癌、不健康、专家表明等），再到具体的详细文本内容，该信息流转过程也是众多衍生性网络健康谣言“造谣者”撰写或“翻谣者”改造健康谣言的一般套路，即：首先通过设计具有吸引力的“标题内容”，诱使读者点击；其次，通过设计具有“可靠性、震撼性”的关键要素内容，让读者“关心则乱 / 丧失理智”；最后，基于具体的内容在逻辑上进行引导，让用户信任，甚至付诸实践和转发。基于这一典型的衍生性网络健康谣言产生流程，本文分别针对衍生性网络健康谣言的标题特征、要素特征和内容主体特征逐个抽取，并构建统一的特征表示。其基本思路是基于上文得到的词向量和标题（句子）向量以及篇章向量展开。对于每一篇信息内容 i，定义其特征组合如下：

$$Doc_i = \mathbf{F}(T_i, K_i, C_i),\ T_i, K_i, C_i \in \mathbb{R}^n \quad \text{公式（4）}$$

其中，Ti 为标题特征向量，Ki 为关键要素特征向量，Ci 为文档主体内容特征向量，F(*) 为特征聚合池化函数。公式（4）中，标题特征向量和文档主体的内容向量直接使用 Sentence2Vector 以及 Doc2vec 模型即可获取。本文基于如图 8 所示的流程，对衍生性网络健康谣言六要素特征向量进行计算。

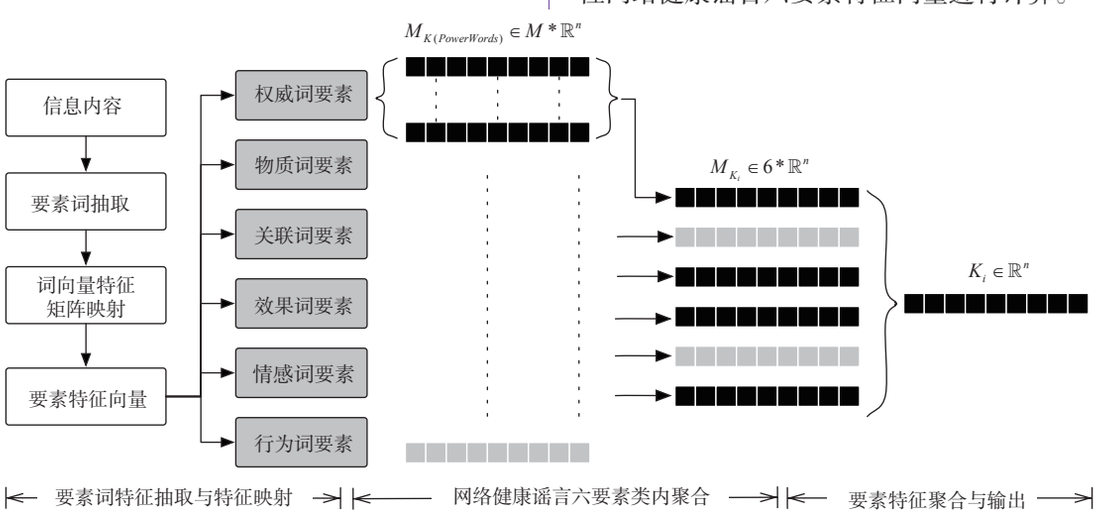

图 8　衍生性网络健康谣言六要素特征向量聚合流程示意

值得注意的是，如果健康谣言六要素中某类要素无法匹配得到，使用 0 向量作为占位向量进行补齐，使得参与聚合的 M 个特征都具有相同的特征位数。

步骤二：文档特征的融合。通过步骤一，可以得到每篇网络健康信息中的三种类型的文本特征向量，即标题特征向量 Ti、要素特征向量 Ki 和内容主体向量 Ci。接下来，基于不同的特征聚合函数 F(*) 对上述三种特征进行融合。典型的聚合方式包括均值池化聚合、最大值池化聚合以及串联聚合。为了更大程度上保留原始文本特征的有效信息，本文将使用特征串联的方式作为候选聚合方法，从而得到每篇信息文档模型训练和预测所需的特征序列，即对于每篇网络健康信息文档有：

$$Doc_i = \mathbf{F}(T_i, K_i, C_i) \rightarrow L_i \qquad \text{公式（5）}$$

步骤三：衍生性健康谣言识别模型的分类器训练。在完成标题特征向量、要素特征向量和内容主体向量的融合以后，通过选取特定的有监督模型，如 LR、SVM 等即可实现衍生性网络健康谣言识别模型的分类器训练。同时基于训练完成的模型，即可实现对新增衍生性网络健康信息的识别和分类。

5 模型实证

为了验证本文提出模型的有效性，接下来将结合真实场景的健康谣言传播数据集进行实证研究，并通过选取当前常见的谣言识别与分类模型进行对比分析，最后通过对实验结果的讨论给出有效性结论。

5.1 实验设计

为了验证模型的可行性，笔者从健康百科网站 [36] 获取的健康信息语料中随机选择了 5 000 篇健康相关的正常信息文本，同时将 502 篇衍生性健康谣言文本作为模型测试实验的数据集，并在此基础上随机选择 20% 的文本内容作为测试集、剩余 80% 的文本内容作为训练集进行实验。

对比模型：文本分类是自然语言处理领域的一个经典问题，而衍生性健康谣言识别则是一个典型的文本分类问题。因此，本节将选取四类具有代表性的文本分类方法作为对照组与本文提出的 CARD 模型进行对比，具体包括：

（1）基于经典文本分类场景下的文档向量化的分类方法（TF-IDF）。通过采用对应的文本向量化方法构建文档的特征集合，并结合已有的机器学习分类器进行文档分类，本节实验将选取朴素贝叶斯模型（Naive Bayes，NB）和逻辑回归模型（Logistic Regression，LR）作为分类器，形成 2 个经典的文本分类模型。

（2）TextCNN 模型 [37]。2014 年由 Y. Kim 等提出的、以预训练词向量为基础构建的基于卷积神经网络的深度学习模型。该模型通过构建的初始化词向量，有监督地学习文本的分类特征，由于其结构简单、效果好，在文本分类、推荐等 NLP 领域应用广泛。

（3）FastText 模型 [38]。2016 年由 Facebook 提出的基于深度学习的文本分类模型。该模型与 Word-2Vec 模型类似，但不同点在于 FastText 不仅仅使用上下文窗口内的信息作为输入，还加入了句子的 N-gram 特征，以进一步提升模型的分类效果。本节对比模型采用 bow+bigram+trigram 模型组合。

（4）Transformer 模型 [39]。2017 年由 Google 提出的基于注意力机制（Attention-based）的端到端的深度学习模型。该模型不依赖于 RNN 和 CNN 的固有结构，大幅降低了原有 Seq2Seq 模型的计算复杂度，在大规模数据集上具有良好的效果。

除了上述提出的对比模型外，还有本文提出的 CARD 模型，该模型包含两个子模型，分别是 CARD（Pure）和 CARD（TSC）模型，其中 CARD（Pure）为以标题（Title）+ 主体内容（Context）为特征来源的子模型，而 CARD（TSC）模型除上述两种特征外，还增加了衍生性健康谣言词语的六要素聚合特征（Six Key Features of Health Rumor），同时本模型将使用 LR 终端分类器。

超参数约定：本节三类深度学习模型的输入文本最大长度限定为 Max_L=500，输入 / 输出向量空间维度限定为 d=300 维，样本重复迭代次数均为 10 次。

评价指标：为了评估和对比不同模型之间的识别效果，使用三类分类模型评估中常用的经典指标作为效果衡量指标（Macro），包括 Accuracy（准确率）、Recall（召回率）以及 F1 值。

5.2 实验结果与讨论

本文提出的 CARD 模型与现有基准的谣言文本分类模型的识别效果对照如表 3 所示。

首先来看常见的文本分类方法，即以 TF-IDF 为代表的文本特征抽取方法。表 3 中，TF-IDF（NB 或 LR）模型在准确率上都取得了不错的效果，特别是基于 TF-IDF（LR）模型相比除 FastText 之外的基准对照模型，同时在召回率（Recall）和 F1-Score 指标上都有不错的表现。且由于该模型为典型的统计机器学习模型，其训练效率相比深度学习也有一定的优

表 3 CARD 模型及对比模型在 502 条中文网络健康谣言数据集上的分类效果

特征抽取模型	分类模型	评价指标 /%		
		Accuracy	Recall	F1-Score
TF-IDF	NB	90.98	74.76	79.93
	LR	95.40	82.79	87.58
FastText		95.27	85.27	89.33
TextCNN		84.69	74.38	78.13
Transformer		78.49	77.50	77.18
CARD（Pure）		91.49	86.12	88.51
CARD（TSC）		97.01	93.00	94.87

势，特别是在大规模数据场景中，分类效率较高，正因为如此，很多大型工业实践场景也常常采用这种模式作为基准模型进行粗排打分或作为人工审核的过滤标准。

进一步地，在深度学习类的模型中，以 Transformer 和 TextCNN 为代表的深层神经网络模型，并未取得相比浅层神经网络模型（FastText 和 CARD）更好的分类效果，甚至比传统的特征抽取方案（TF-IDF）都要弱，其模型分类准确性仅为 78.29% 和 84.69%。其原因在于这类模型对于文本的初始化特征有较强的依赖性，通常需要使用一定的预训练词向量模型作为基础的语料进行输入才能进一步提升模型效果，而直接基于随机化的初始化特征向量并未能有效地对文本分类特征进行建模。

此外，深度学习的效果优异性通常也依赖于大规模的训练语料和数据，谣言传播这类数据倾斜较大和样本不均衡的场景难以发挥深度学习在大规模样本学习上的优势。近年来，为了改善上述深度学习模型在小样本场景下的表现和效果，也有很多基于迁移学习或多任务学习模型被提出，用以优化单个深度学习模型在小样本分类场景中的劣势，在后续研究中，将考虑如何结合面向小样本的深度学习模型来进一步提升当前模型在更广泛数据集上的表现效果。

最后，通过结果可以看到，基于标题特征—内容文本六要素特征—主体内容文本特征三层特征融合的 CARD（TSC）模型在三个指标上都取得了最优的效果，模型的准确率和召回率分别达到了 97.01% 和 93.00%。此外，基于 CARD 模型的消融实验表明：未加入六类健康谣言词聚合特征的 CARD 模型（Pure）在 F1 值上也取得了除 FastText 模型之外的次优效果，但相比融合特征模型，准确率和召回率分别下降了 5.52% 和 6.88%。

特别地，对于谣言分类模型而言，其召回率的表现效果尤为重要（由于在实践场景中，被判别为疑似健康谣言的文本还会经过人工审核，因此召回率的效果将影响最终的健康谣言影响范围），其得分关系到是否能在最大程度上将混淆在语料库中的谣言筛选出来，从而降低疑似健康谣言混入正常渠道流通的概率。基于召回率指标来看，在本实验中，CARD（TSC）相比第二优的 FastText 模型提升了 7.73%，相比传统的文本分类模型（TF-IDF_LR）有 10.21% 的提升。此外，由于使用多阶段建模策略，本文模型在预测结果的可解释性方面相比一般的黑盒深度学习模型也更有优势。综上，本文提出的基于标题特征—内容文本六要素特征—主体内容文本特征三层特征融合的 CARD（TSC）模型针对衍生性网络健康谣言的识别具有较优的识别效果。

6 结语

网络健康谣言识别是网络健康谣言治理的重要环节，也是网络健康谣言技术治理的重要实践。对此，本文基于当前网络健康谣言传播的反复性和文本内容的衍生性与关联性特征，构建了一个基于标题特征—内容文本六要素特征—主体内容文本特征三层特征融合的衍生性网络健康谣言识别模型。实验结果表明，本模型相较传统的文本分类模型（TF-IDF）以及深层神经网络学习模型均有更好的识别效果，能有效解决衍生性网络健康谣言的增量问题。不足之处在于，由于无法获取用户的实时动态行为数据和传播链路数据，未对新型网络健康谣言的早期识别进一步展开研究。这需要平台在实际应用场景中，进一步完善和优化。同时，对于新型网络健康谣言的识别预警也是今后的研究方向之一。

参考文献：

[1] 刘鹏飞，周悦．食品安全谣言的法律处置 [J]. 中国报业，2015(13): 52-53.

[2] 黄淼，黄佩．基于知识关联特征的网络内容识别——以健康谣言为重点 [J]. 北京邮电大学学报（社会科学版），2020, 22(01): 1-6, 13.

[3] 谭励，王舸，周丽娜，等．基于多示例学习的食品健康领域长文谣言检测 [J]. 计算机工程与设计，2022, 43(11): 3101-3107.

[4] 王世海．社交媒体健康谣言特征与主要易感人群关联性研究 [J]. 记者摇篮，2022(4): 24-26.

[5] 陈昊．微信中健康类谣言的传播与治理策略 [D]. 济南：山东师范大学，2020.

[6] 奥尔波特，波斯特曼．谣言心理学 [M]. 刘水平，梁元元，黄鹏．译．沈阳：辽宁教育出版社，2003: 5.

[7] 陈燕方，李志宇，梁循，等．在线社会网络谣言检测综述 [J].

计算机学报 , 2018, 41(7): 1648-1677.

[8] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on Twitter[C]//Proceedings of the 20th international conference on World Wide Web. New York: Association for computing machinery, 2011: 675-684.

[9] WU K, YANG S, ZHU K Q. False rumors detection on sina weibo by propagation structures[C]//Proceedings of the 2015 IEEE 31st international conference on data engineering. Piscataway: Institute of electrical and electronics engineers, 2015: 651-662.

[10] KWON S, CHA M, JUNG K, et al. Prominent features of rumor propagation in online social media[C]// Proceedings of the 2013 IEEE 13th international conference on data mining. Piscataway: Institute of electrical and electronics engineers, 2013: 1103-1108.

[11] MA J, GAO W, MITRA P, et al. Detecting rumors from microblogs with recurrent neural networks[C]// Proceedings of the 25th international joint conference on artificial intelligence. Menlo Park: Association for the advancement of artificial intelligence press, 2016: 3818-3824.

[12] KALIYAR R K, GOSWAMI A, NARANG P, et al. Fndnet-a deep convolutional neural network for fake news detection[J]. Cognitive systems research, 2020, 61(6): 32-44.

[13] 汪建梅 , 彭云 , 余晨钰 . 融合时间序列与卷积神经网络的网络谣言检测 [J]. 小型微型计算机系统 , 2022, 43(5): 1020-1026.

[14] 伊静 . 面向在线社交媒体的谣言识别与传播分析研究 [D]. 济南：山东师范大学 , 2021.

[15] 朱梦蝶 , 付少雄 , 郑德俊 , 等 . 文献视角下的社交媒体健康谣言研究：特征、传播与治理 [J]. 图书情报知识 , 2022, 39(5): 131-143.

[16] SICILIA R, LO GIUDICE S, PEI Y, et al. Twitter rumor detection in the health domain[J]. Expert systems with applications, 2018, 110: 33-40.

[17] 许诺 , 赵薇 , 尚柯源 , 等 . 基于预训练语言模型的健康谣言检测 [J]. 系统科学与数学 , 2022, 42(10): 2582-2589.

[18] 张帅 . 社交媒体虚假健康信息特征识别 [J]. 图书情报工作 , 2021, 65(9): 70-78.

[19] 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(3): 294-304.

[20] 石锴文 , 刘勘 . 突发公共卫生事件中微博谣言的识别 [J]. 图书情报工作 , 2021, 65(13): 87-95.

[21] 於张闲 , 冒宇清 , 胡孔法 . 基于深度学习的虚假健康信息识别 [J]. 软件导刊 , 2020, 19(3): 16-20.

[22] LIU Y, YU K, WU X, et al. Analysis and detection of health-related misinformation on Chinese social media[J]. IEEE access, 2019(7): 154480-154489.

[23] SAFARNEJAD L, XU Q, GE Y, et al. A multiple feature category data mining and machine learning approach to characterize and detect health misinformation on social media[J]. IEEE Internet computing, 2021, 25(5): 43-51.

[24] GHENAI A, MEJOVA Y. Fake cures: user-centric modeling of health misinformation in social media[J]. Proceedings of the ACM on human-computer interaction, 2018, 2(9): 1-20.

[25] GF A, FI A, IMDD A, et al. Experts perception-based system to detect misinformation in health Websites[J]. Pattern recognition letters, 2021, 152(12): 333-339.

[26] ZHAO Y, J DA, J YAN. Detecting health misinformation in online health communities: incorporating behavioral features into machine learning based approaches[J]. Information processing & management, 2021, 58(1): 102390.

[27] SAEED F, YAFOOZ W, AL-SAREM M, et al. Detecting health-related rumors on Twitter using machine learning methods[J]. International journal of advanced computer science and applications, 2020, 11(8): 324-332.

[28] 陆恒杨 , 范晨悠 , 吴小俊 . 面向网络社交媒体的少样本新冠谣言检测 [J]. 中文信息学报 , 2022, 36(1): 135-144, 172.

[29] 赵月华 , 朱思成 , 苏新宁 . 面向网络虚假医疗信息的识别模型构建研究——一种基于预训练的 BERT 模型 [J]. 情报科学 , 2021, 39(12): 165-173.

[30] 爱微帮 . 谣言过滤器 [EB/OL]. [2023-06-14]. http://data.aiweibang.com/user/search?kw=%E8%B0%A3%E8%A8%80%E8%BF%87%E6%BB%A4%E5%99%A8.

[31] RICHARDSON L. Beautiful soup[EB/OL]. [2023-06-14]. https://www.crummy.com/software/BeautifulSoup/.

[32] DOUBLECLICK. MongoDB[EB/OL]. [2023-06-14]. https://www.mongodb.com/.

[33] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning internal representations by error propagation[R]. Cambridge: Massachusetts Institute of Technology Press, 1985: 318-362.

[34] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2023-06-14]. https://arxiv.org/pdf/1301.3781.pdf.

[35] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]//Proceedings of 31st international conference on machine learning. Red Hook: Curran Associates, 2014: 2931-2939.

[36] 珠海健康云科技有限公司 . 健康百科 [EB/OL]. [2023-06-14]. https://baike.120ask.com/.

[37] KIM Y. Convolutional neural networks for sentence classification[EB/OL]. [2023-06-14]. https://arxiv.org/pdf/1408.5882.pdf.

[38] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[OL]. [2023-06-14]. https://www.semanticscholar.org/paper/Bag-of-Tricks-for-Efficient-Text-Classification-Joulin-Grave/892e53fe5cd39f037cb2a961499f42f3002595dd.

[39] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of 31st conference on neural information processing systems. Red Hook: Curran Associates, 2017: 5998-6006.

作者贡献说明

陈燕方：数据收集，模型训练，论文撰写；

周晓英：论文修改。

Research on Derivative Online Health Rumors Identification Modal Based on Text Feature Fusion

Chen Yanfang[1] Zhou Xiaoying[2]

[1] Renmin University of China Libraries, Beijing 100872

[2] School of Information Resource Management, Renmin University of China, Beijing 100872

Abstract: [Purpose/Significance] Online derivative health rumors are characterized by low generation thresholds, strong periodicity, and far-reaching consequences. This is one of the key issues that need to be prioritized in the identification and goverance of online health rumors, and it is also an important breakthrough point. **[Method/Process]** Through the methods of deep semantic representation and aggregation, this paper explored six element features of the derivative text features of online health rumors. At the same time, combined with the distributed semantic features pre-trained model of online health rumors, the thesaurus of content elements of online health rumors (6 categories, 6287 words in total)is obtained. Finally, through the unified vector space representation and fusion of title feature, six element features of health rumors content and main content feature, a online health rumor discrimination model framework based on multi-source text feature fusion was constructed. **[Result/Conclusion]** The empirical study of the model shows that text feature fusion model proposed in this paper has a significant improvement in the recognition of derivative online health rumors compared with the control model, and the abundant and expandable thesaurus of health rumor elements provides better resource support for subsequent research.

Keywords: online health rumors health rumor detection text features text mining