doi:10.3772/j.issn.2095-915x.2022.05.008

基于深度学习的互联网虚假信息识别研究

刘建强[1] 卢为党[2] 黄国兴[2] 马宁[1]

1. 军事科学院战争研究院 北京 100091；
2. 浙江工业大学信息工程学院 杭州 310023

开放科学（资源服务）标识码（OSID）

摘要： [目的 / 意义] 随着互联网技术的高速发展，虚假信息混杂成为近些年来在高新科技领域凸显出的网络舆情样式。虚假信息可作为影响科技战略态势的重要“武器”，会阻碍我国正确高效获取有价值的各类数据，从而给国家和社会造成不可估量的损失。[方法 / 过程] 本文针对互联网虚假信息的基本概念进行梳理，并结合近年相关研究工作，提出一套针对虚假信息识别的软件平台设计。同时，为提升软件平台虚假信息识别的准确度、加快算法收敛，提出一种基于生成对抗网络的虚假信息筛选算法。[结果 / 结论] 提出的虚假信息识别平台有效提升了信息识别的精准度。

关键词： 案例分析；互联网虚假信息；识别；深度学习

中图分类号： G35; TP391

Identifying Internet Fake Information based on Deep Learning

LIU Jianqiang[1] LU Weidang[2] HUANG Guoxing[2] MA Ning[1]

1. Academy of Military Science of the PLA, Beijing 100091, China;
2. School of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

Abstract: [Purpose/Significance] With the rapid development of Internet technology, the mixing of fake information has become a pattern of network public opinion highlighted in the field of high-tech in recent years. Fake information can be used as an important “weapon” to affect the strategic situation of science and technology, which will prevent my country from obtaining various types of valuable data correctly and efficiently, thereby causing immeasurable losses to the country and society. [Methods/Processes] This paper sorts out the basic concepts of Internet fake information, and proposes a set of software platform design for fake information identification based on related research work in recent years. At the same time, in order to improve the accuracy of fake information recognition on the software platform and speed up the algorithm convergence, a fake information screening algorithm based on generative adversarial network is proposed. [Results/Conclusions] The experimental results show that the proposed fake information identification platform effectively improves the accuracy of information identification.

Keywords: Case study; Internet fake information; identification; deep learning

作者简介 刘建强（1981-），助理研究员，研究方向为指挥自动化；马宁（1980-），硕士，助理研究员，研究方向为计算机应用技术，E-mail：2091970448@qq.com；卢为党（1983-），教授、博士生导师，研究方向为包括无线通信、智能通信、无人机通信、移动边缘计算；黄国兴（1986-），副教授、硕士生导师，研究方向包括雷达信号处理、模拟信号的欠奈奎斯特采样技术、超宽带 UWB 通信技术、过程层析成像。

引用格式 刘建强 , 卢为党 , 黄国兴 , 等 . 基于深度学习的互联网虚假信息识别研究 [J]. 情报工程 , 2022, 8(5): 86-100.

引言

互联网虚假信息是近年来在高新技术领域凸显出的网络舆情样式[1]，主要以国外机构和研究人员夸大、伪造研究成果的影响力，质疑、诋毁国内科研成果和专家等形式呈现，也包括国内相关专家学者编造夸大本人及团队研究效用，误导国家科技管理部门在立项、评奖、人才培养等方面产生决策误差，进而产生了巨大损失。近年来互联网虚假信息成为学术界研究的热点，Hindman 等[2]以推特为例，对互联网虚假信息在社交媒体中的成因和影响进行了介绍，详细阐述了互联网虚假信息可能造成的严重后果。Bradshaw[3]针对互联网虚假信息对于国家安全的危害进行了分析，文中表明许多国家都成立了专门从事舆论干涉的网军，用以掌握国际媒体言论，从而误导他国国家战略的目的。文献 [4-6] 针对互联网虚假信息的概念和诱导决策方法进行了梳理。互联网虚假信息的主要迷惑对象是国家情报部门和领导决策层，随着大数据分析、推荐算法等技术的快速发展，互联网虚假信息的迷惑性得到极大增强，并且可以精准定位受众目标。同时由于自媒体、短视频等传媒方式的兴起，舆论的发布愈加轻量化。这些因素都无形间增加了预防互联网虚假信息的成本。

当前，网络虚假信息的检测手段主要是使用特征工程的手段，即根据专家或者经验总结设计的虚假信息特征，如语言特征、传播特征等，再采用支持向量机、随机森林等机器学习方法对信息进行真假分类。这种基于特征提取的方法可以充分地利用专家总结的经验和知识，但缺点在于这种方法需要人工手动提取特征，无法自动从大规模互联网数据中自动挖掘特征。而网络虚假信息与垃圾邮件或广告类似，其技术、手段和形式也在不断更新换代，而这些专家总结的特征很难做到与时俱进，应对新出现的虚假信息形式。另一方面，随着近年来深度学习的迅猛发展，使用自然语言处理利用分布式架构学习大规模数据集合信息的低维特征向量表示。低维向量空间中的位置信息，以及学习到向量之间的相对距离反映了原始对象（如词、句子、文档）的语义相关度。通过聚类分析等方法对信息中潜在的威胁进行分析预警。但此种方法存在以下几个缺陷，首先对于同样使用神经网络进行的深度伪造（Deep Fake）信息识别准确度低，其次，只能检测数据库已收录的虚假信息种类，尚未收录的新种类仍然需要通过人工检测。另外，人工智能模型潜在的算法偏见、缺乏算法透明度和可解释性的缺陷可能导致识别出错。

当前国内外针对网络虚假信息的识别方法，主要集中在建立用户画像以及针对基于网络评论的信息识别。马超[7]使用 Facebook 数据集，利用随机游走算法，基于主题模型对社交网络中的用户画像分析方法进行研究，实现了对于重点用户的特征构建。李雅坤[8]基于微博平台和用户数据，利用大数据技术建立用户画像模型，构建群体特征，可以对敏感用户群体进行定位，同时生成特定的用户标签，方便对于用户的精准把控。特征画像的问题在于，需要为重点人群、群体提供多方位、全面的用户数据，数据收集成本较高。同时该方法对于长期保有用户的账号效果较好，对于网络水军短期快速

doi:10.3772/j.issn.2095-915x.2022.05.008

注册的机器人新账号，效果较差。

相比之下，基于评论的信息识别，对于网络水军账号的信息识别，普适性更强。基于评论的网络虚假信息识别，可分为基于评论内容的信息识别、基于评论行为的信息识别、基于评论关系的信息识别。基于评论内同的信息识别，主要侧重于评论文本挖掘，由于评论文本为用户伪造，所以在语言细节上会有破绽，例如文字重复率高、语言模型异常，协作风格单一等。Jindal 等 [9] 以亚马逊 580 万条评论为研究对象，针对相似度进行逻辑回归模型构建，实现了虚假评论的识别。Ott 等 [10] 构建了负向情感词库，一次对虚假评论文本进行分析，取得了高于人工识别的效果。Fusilier 等 [11] 提出利用 PU-learning 结合 n-gram 词袋特征进行虚假意见检测的方法，对 1600 余条评论进行分析，实验结果表明，该方法在正面和负面欺骗意见的检测上具有较好的学习效果。Yun 等 [12] 采用语体分析方法，对真实和虚假评论进行对比分析，构建贝叶斯和支持向量机模型，达到了 90% 准确率。针对评论行为的识别，由于虚假评论发表者的行为不同于正常用户，会在短时间写出大量的评论，并在评分上与正常用户产生偏离，所以可针对以上特征进行检测。Lim 等 [13] 提出两种虚假评论行为模型，即基于目标的虚假评论模型和基于偏差的虚假评论模型，通过结合两个模型，提升了虚假评论信息识别的准确率。Mukherjee 等 [14] 提出了虚假评论团体的概念，并针对团体行为等方面特征对虚假信息评论进行建模和识别。针对评论关系的虚假信息识别，主要着眼于评论发表者所存在的异常关系，可通过建立评论发表者、评论、时间之间的关系网络图模型，对此进行识别。Wang 等 [15] 提出一种异构评论图模型，将虚假评论识别问题转化为异常关联模式挖掘，补充了先用方法的不足。基于评论信息的虚假信息识别缺陷在于只能针对文本评论进行分析，当前虚假信息已不局限于文本，在视频音频以及其他复合格式文件，都存在虚假信息。

综上所述，当前国内外互联网虚假信息的识别，对于网络水军机器人账号，以及除文本外的虚假数据类型还有待进一步研究。虚假信息识别需要综合数据采集、数据分析、特征工程、神经网络拟合、专家系统预警等多种关键技术的组合支撑，是一个系统性的工程。为了解决上述问题，本文提出一种基于深度学习的互联网虚假信息平台设计，以及一种基于生成对抗网络的数据筛选算法。本文的主要共献有以下几点：

（1）提出了一种基于深度学习的互联网虚假信息平台设计。该平台设计有效的解决了当前多源异构数据无法有效进行特征提取的信息识别的问题；并且要素齐全，包含了数据收集、数据预处理，以及数据分析预警等多个功能模块，减少了人工数据收集和标定的工作量，同时构建自然语言处理架构，可对各类信息进行文本分类、语法分析、语义分析。

（2）针对收集到的各类异构信息，平台集成了综合预警模块和信息挖掘模块，可对可以信息进行持续的深度、迭代挖掘，对其关联的话题和历史信息进行热点分析、观点倾向分析。高效可靠的实现了虚假信息识别和溯源。

（3）为提升平台在处理海量数据时，识别精准度较低、预处理算法收敛困难，且针对

Deep Fake 消息识别效果差的问题，本文提出一种基于生成对抗网络的改进虚假信息初筛方法，用于提升虚假信息识别平台在数据分析、预警上的效能。

1 互联网虚假信息基本概念

苏鹏等 [16] 将互联网虚假信息分为 5 类，分别是 Disinformation、Misinformation、Malinformation、Fake News、Deepfakes。Disinformation 是专业人士有意构建的假消息，目的在于影响目标人群的行为，进而达到自身目的 [17, 18]。Misinformation 是误报的消息，是由于信息接收者在解读信息时出现了误判，导致理解出现偏差 [19, 20]。Malinformation 是被恶意传播的真实消息，包括恶意泄密、卖密等都属于此类 [21, 22]。Fake News 是新闻工作者或自媒体人为了吸引眼球，或能力素质不足导致传播的假新闻 [23, 24]。Deepfakes 则是通过大数据分析、深度学习等智能方法制造的假消息，利用推荐算法可以对目标人群进行精准的误导 [25-27]。互联网虚假信息，目的性强，构造逻辑严密，是一种为了特定利益，利用传统和智能手段构造虚假信息，并通过多种形式媒体进行主动宣传的手段。它的主要特点是蓄意性、虚假性、传播性、误导性。

互联网虚假信息的传播媒介可分为以下几类：

(1) 传统媒体。主流的媒体如各国的权威报纸、官方媒体网站、新闻电视台等，在信息传播上扮演着权威、官方的角色。目标群体在潜意识中对传统媒体具有更高的信赖度，所以在传统媒体上发布的虚假信息，更容易对公众和目标群体产生干扰。但另一方面，传统媒体对于虚假信息审核较为严格，所以不易发布。

(2) 社交媒体。社交媒体如国内的微博、国外的 Facebook 等是当前互联网虚假信息的重灾区。由于社交媒体用户注册门槛低，用户构成复杂，审核力度不足等原因，极易形成虚假信息。并且由于当前社会人们更趋向于利用短频快的方式获取信息，所以社交媒体上的虚假信息会对大量人群造成影响。

(3) 网络水军。网络水军指的是由国家或特定组织，利用僵尸网络或隐蔽身份小号组成的虚假信息构建团体。网络水军可能是真实的人，也有可能是被控制的自动机器人，主要用途就是在网络上针对特定话题，以极高频率散播谣言，迫使民众和目标群体无法获取准确信息源，从而出现信息和形势的误判。

(4) Deepfakes。Deepfakes 是指利用深度学习等人工智能算法作为互联网虚假信息生成工具，利用大数据分析、推荐算法等手段对目标受众进行精准定位，并高效推送虚假信息的手段。由于当前信息泄露严重，每个人的身份信息和偏好极易被特定群体所掌握分析，结合智能算法和大数据手段，Deepfakes 成为了当前最高效的虚假信息生成方法。

互联网虚假信息的传播模型揭示了虚假信息从产生到扩散的方式，准确的掌握传播模型对于设计算法识别虚假信息，进而阻断虚假信息的扩散有重要的意义 [28, 29]。当前大多数研究集中于信息传播机制和信息传播规律演化的预测 [30]，主要的模型包括线性阈值模型 [31]、传染病行为动力学模型 [32] 和基于博弈论的复杂网络

doi:10.3772/j.issn.2095-915x.2022.05.008

模型 [33] 等。传染病模型是信息传播领域的重要模型，研究人员基于此衍生开发了多种算法。文献 [34] 提出了 SLIR 模型，通过引入潜在的用户节点，结合平均场理论针对传播网络中的重要性进行了评估。文献 [35] 基于博弈论的方法，构建动力学传播模型，在不仅可以准确描述网络中信息传播趋势，同时可以揭示不同影响因素对信息传播的影响。但问题在于，算法面向的场景都是单消息场景，针对多消息场景的效果不佳。基于此，文献 [36] 利用演化博弈理论，将信息传播的方式类比为生态系统中病毒传播方式，对演化动力和进化策略进行分析，揭示了传播过程中信息之间的促进和抑制关系。

2 基于深度学习的互联网虚假信息识别平台架构

为了能够有效识别互联网虚假信息，需要结合数据收集、数据处理、数据分析预警等多重手段，综合构建一套互联网虚假信息识别平台，用以高效处理多源异构的互联网信息，准确识别虚假信息并做出及时的预警。本节对基于深度学习的互联网虚假信息识别平台的架构进行介绍，该平台利用爬虫工具对重要信息进行采集抓取。利用数据融合、自然语言处理等方法对多源异构的信息数据进行预处理，最后通过情感分析、聚类、数据分析等手段分析数据，并对异常数据进行报警。图 1 是上述三个流程的关系图。

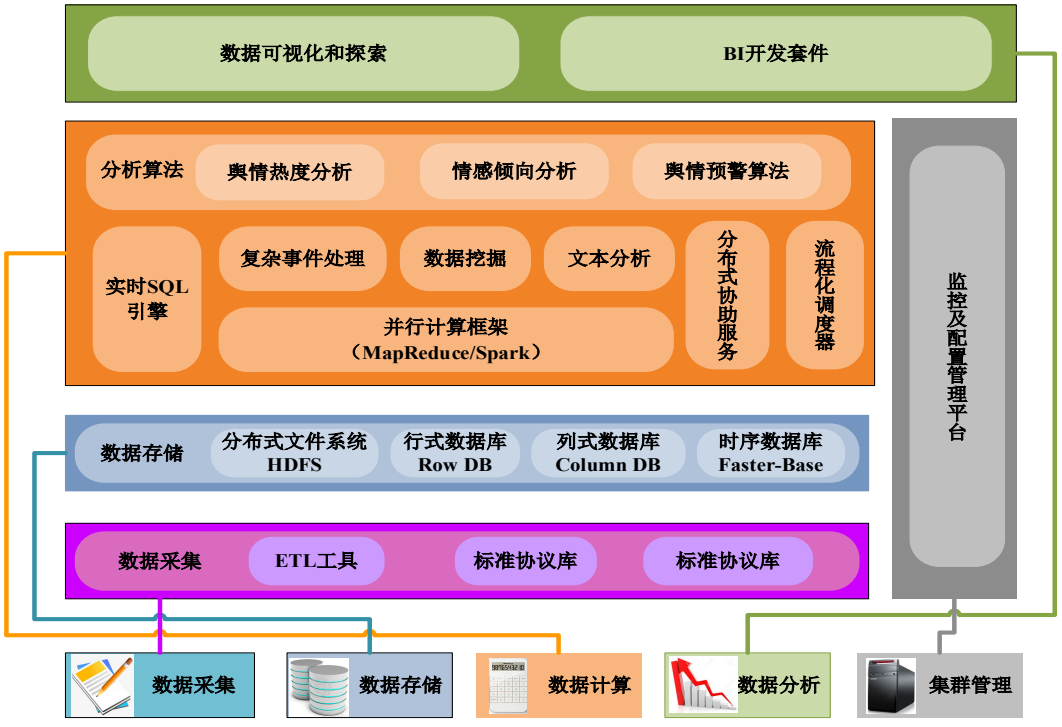

图 1 基于深度学习的互联网虚假信息识别平台架构图

2.1 基于网络爬虫的信息实时采集模块

网络爬虫 [29] 是一种高效的信息抓取工具，它集成了搜索引擎技术，并通过技术手段进行优化，用以从互联网搜索、抓取并保存任何通

过超文本标记语言进行标准化的网页信息，其流程如图 2 所示。

目前动态网页（例如 AJAX 等技术所实现）的流行，在实际中还需要基于事件驱动技术来获取动态网页的信息，这需要解决三项技术：（1）JavaScript 的交互分析和解释；（2）DOM 事件的处理和解释分发；（3）动态 DOM 内容语义的抽取。考虑到信息爬取效率，可以使用分布式爬虫系统，协同多台计算机终端来进行协同爬取网页信息 [37]。

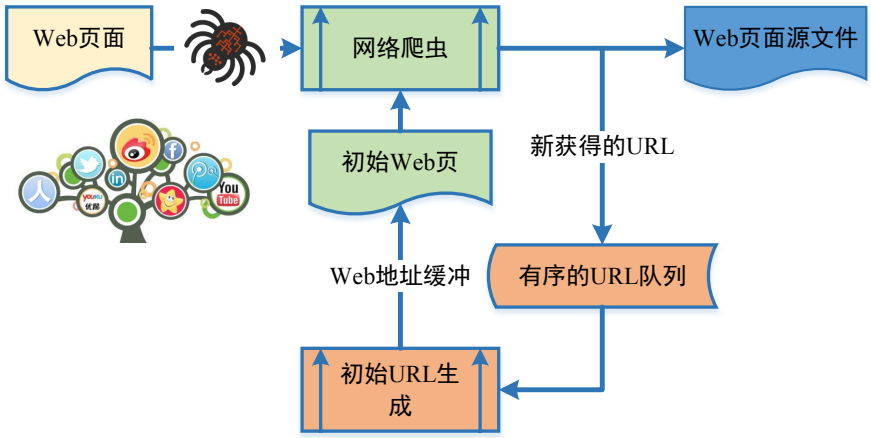

图 2 虚假信息采集流程

通过网络爬虫采集技术，可以获得丰富的用于分析的数据，并构建高质量的数据特征，具体包括：基础属性特征集、行为特征集、场景特征集、关联特征集。

基础属性特征集包括身份属性、经济属性、文化属性、社群属性等。为了获取到更精准的目标用户特征，对每类属性进行细化，得到通用属性下的二级属性，具体如下。（1）基础属性：性别，年龄，文化程度，人种，语种，国家，民族，职业，地域，行业；（2）经济属性：经济收入，可支配收入，付费方式；（3）文化属性：所处文化圈，文化喜好，个性化需求；（4）社群属性：交友需求，异性交往需求，归属需求，领导需求，合作需求等。

行为特征集是基于用户点击流、操作行为轨迹等数据等提炼加工的用户行为特征集，包括：（1）用户资料输入时长，如：联系人输入时长、工作单位输入时长；（2）操作频次，如：最近一月登录次数，最近一月提现次数等；（3）间隔时长，如：注册到申请的时长，两次操作之间的最大间隔时间；（4）用户生物探针特征，如：手机陀螺仪位置偏好、用户点击屏幕位置偏好、屏幕点击速度、屏幕点击强度偏好等；（5）用户影像拍摄偏好，如大头照拍摄次数、是否使用美颜、是否裁减等。

场景特征集主要指用户习惯操作的场景，包括：（1）用户设备信息，如：设备型号、设备语言设置、设备 APP 列表及类型等；（2）用户操作时点场景，如：工作日或节假日操作，早中晚操作；（3）操作空间场景，即用户操作的地理位置信息，如：公司或家庭，商场场或旅游区等；（4）IP 环境，如：网络类型，WIFI 或 4G 等；（5）运营商情况，如：运营商类型、在网时长、消费套餐等。

doi:10.3772/j.issn.2095-915x.2022.05.008

关联特征集主要指用户核心属性关联的情况，包括一级关联和多级关联。核心属性包括：手机号、账号、身份证号、设备号、IP、GPS 地址、微信号、第一联系人电话、第二联系人电话、公司名称、家庭地址等。其中一级关联指的是，主维度属性关联另一个维度的个数，如：最近一年设备关联的手机号个数、1 天内 IP 关联的手机号个数等；而多级关联指的是多个核心属性之间的关联情况，如二级关联：同 IP 关联的设备，这些设备关联的手机号个数。

2.2 基于自然语言文本处理方法的数据预处理模块

互联网虚假信息识别过程中，对互联网虚假信息的关键词提取是决定识别率的主要因素。因此通过自然语言处理手段主要完成分词的处理。首先分词模型将输入的语句进行词语分隔，然后把分隔的结果进行词性标注和命名实体，其目的在于提取本文中的有意义的词语并对其语义进行分析。在完成基础处理之后，可以构建更深入的自然语言处理，如文本分类、信息热度分析、观点倾向分析等。网络数据预处理流程图 3 所示。

用于分词的机器学习模型和方法主要分为两大类：一类是基于字符标记的，也就是对每一个字单独进行分段信息的标注；还有一类就是基于词的，也就是对词进行整体的标注和建模。基于单个字符标记方法的核心就是对每一个字所属词中的位置进行一个标注。对于任何一个字来说，它可以是一个词的开始、一个词的中间、一个词的结尾，或者本身就是一个单字的词，这也就是在序列标注中常用的分类。这类方法比较典型的是最大熵马尔科夫模型 [38] 和条件随机场 [39]。

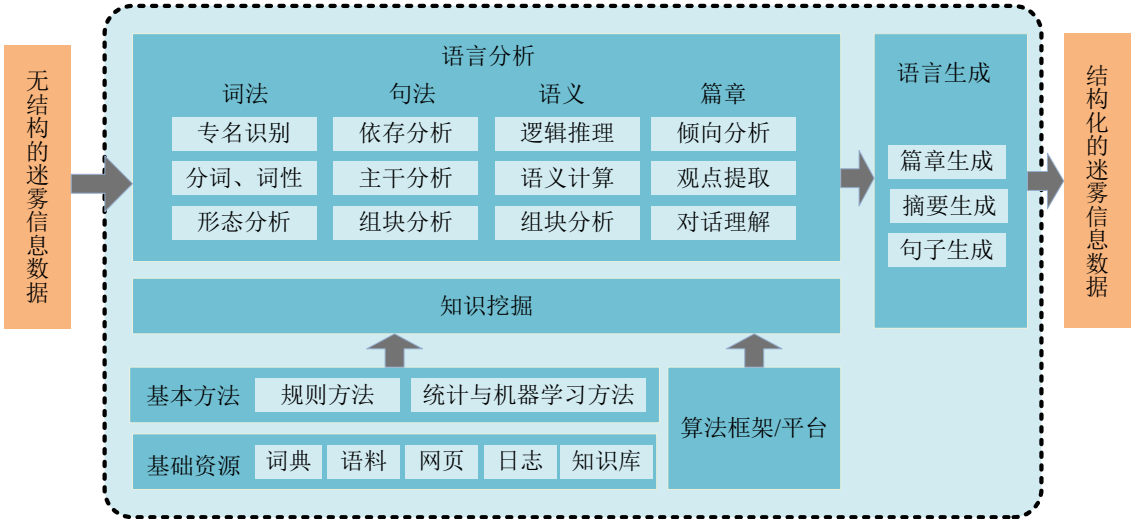

图 3 网络数据预处理流程图

虽然基于单个字符的模型对于抽取字具有较好的效果，但该模型无法直接建立相邻词之间的相关性，也无法直接看到当前整个词所对应的字符串。而基于词的模型能够很好的解决这个问题，这种模型用类似基于转换的句法分析去解决分词的问题。基于词的模型是一种渐进式、自下而上的语法分析办法，一般以从左向右的方式处理逐字处理文本的输入，并在运

行过程中通过一个堆栈去保存到当前为止得到的不完整的分词结果，并且通过机器学习的方法去决定如何整合当前的分析结果，或是接收下一个输入去拓展当前的分析结果。基于词的算法存在的问题是堆栈上保存的到当前位置的分析结果的数量会非常大，需要进行修剪来控制搜索空间的范围。

2.3 互联网虚假信息数据分析与预警模块

虚假信息分析模块是系统中最为关键的处理模块，主要利用文本分类和聚类等方法对预处理后的虚假素材信息进行分析挖掘，实现虚假信息的热点发现和跟踪。

（1）热点分析

热点发现算法从本质上来说是属于数据挖掘中的文本聚类算法。算法的实现过程如下：将预处理后的文本信息归入不同的话题，并在需要的时候建立新的话题，热点发现的目的就是要按照话题将文档进行聚类，从一组文档集中发现新热点，由于没有关于新热点的先验知识，需要建立新的主题簇。热点事件跟踪是为了用户能够跟踪自己所关心的类型事件而进行的操作，用户可以将已获得的事件的样本信息通过系统学习的方式交给系统，然后系统通过文本挖掘技术对不断到来的信息进行分类，判断是否为用户感兴趣的内容，将判断为是的信息交给用户。同时系统可以通过用户对获得的信息的反馈，不断地修正系统的学习结果，使得系统可以获得越来越接近用户所希望的信息。 因此，热点事件跟踪是一种特殊的二元分类问题。

（2）观点倾向分析

敏感信息检测是在海量的互联网信息中，识别出虚假信息。在进行敏感信息识别时需要考虑规模和正负面程度两方面，需要找出在一段时间内的上升较快，或参与规模较大的虚假信息。规模可以通过聚类后的相关网页数判断，负面程度通过中文情感分析技术识别。中文情感分析技术旨在发现用户对热点事件的观点和态度。传统的实现方式是使用 SVM[40]、条件随机场等传统机器学习算法根据手工标注情感特征对文本情感进行分析。最新的实现方式则利用深度学习实现。采用递归神经网络来发现与任务相关的特征，避免依赖于具体任务的人工特征设计，并根据句子词语间前后的关联性引入情感极性转移模型加强对文本关联性的捕获。基于深度学习的方法在性能上与当前采用手工标注情感特征的方法相当，但节省了大量人工标注的工作量。目前，情感分析的主要研究方法还是一些基于机器学习的传统算法，如 SVM、信息熵 [41]、条件随机场等。这些方法归纳起来有三类：有监督学习、无监督学习和半监督学习。当前大多数基于有监督学习的研究都取得了不错的成绩，但是由于有监督学习依赖于大量人工标注的数据，使得基于有监督学习的系统需要付出很高的标注代价。半监督学习则是采取综合利用少量已标注样本和大量未标注样本来提高学习性能的机器学习方法，它兼顾了人工标注成本和学习效果，被视为一种折中方案。无监督学习不需要人工标注数据训练模型，是降低标注代价的解决方案。基于深度学习的方法在性能上与当前采用手工标注情感特征的方法相当，节省了大量人工标注的工作量。

（3）综合预警

网络综合虚假信息预警模块的研发主要包

doi:10.3772/j.issn.2095-915x.2022.05.008

括以下三方面：（1）建立预警指标体系。有学者认为网络虚假信息的产生、发展过程会通过一系列关键指标体现，并将这些指标按照一定的科学方法确定关键指标构成、指标维度、指标层次、指标量化方法等，从而建立预警指标体系。（2）基于网络的数据挖掘的预警。这种方法就是从网络中提取与目标相关的数据，构成目标数据集。其任务是对网络数据进行网页特征提取、基于内容的网页聚类[42]、网络间内容关联规则的发现等，从其中得到与网络的挖掘目的相关的数据。利用相应的工具和技术对挖掘出的数据进行分析、解释，并通过分析结果对网络虚假信息进行预警。（3）基于观点倾向性观点分析技术的预警。采用这种方式进行预警的学者认为网络虚假信息预警能力主要体现在是否能够从海量的网络言论中，发现潜在危机的隐患。到目前为止，对观点倾向性分析主要包括“赞同”“反对”“中立”三种态度。

3 基于深度神经网络的增强互联网虚假信息初筛方法

在第三节介绍的互联网虚假信息识别平台中，数据分析与预警模块通过热点分析和观点倾向分析对海量的互联网信息进行虚假信息筛选。而当前在热点分析和观点分析中大多使用基于大数据的无监督学习方法，如文本聚类、降维分割等，或是使用基于传统机器学习，如支持向量机、条件随机场等方法。这类算法在处理海量数据时，识别精准度较低、预处理算法收敛困难，且针对Deep Fake消息识别效果差。因此，本文提出一种基于生成对抗网络的改进虚假信息初筛方法，用于提升虚假信息识别平台在数据分析、预警上的效能。

互联网虚假信息传播存在多种不同类型的信息，包括不同的源（微博、知乎等）、不同的表现形式（不同格式的文本等），为了准确分析并研判可能的虚假信息，在互联网虚假信息传播模型中，对不同的信息来源可定义不同的网络节点，例如对新浪微博、知乎、论坛等，网络用户可设为复杂网络中的节点，其输入经处理后，表示为传播信息的语义向量X。

基于生成对抗网络(GAN)的多层耦合网络构建

对输入矩阵$X=X_{n\times t}$，n表示节点数量，t表示采样时间，训练采用无监督学习，如图4所示。

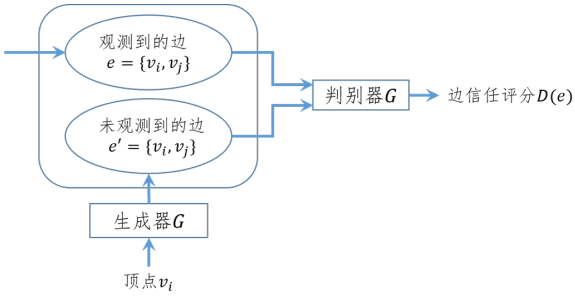

图4 构建多层耦合网络框架

该框架第一层是词向量表示层，输入句子矩阵的列和行分别是词向量的维度和序列长度；第二层是卷积层，主要通过卷积操作来提取句子的局部特征；第三层进行最大池化操作，提取关键特征，舍弃冗余特征，生成固定维度的特征向量，最后将池化层学习到的特征与注意力文本连接并作为全连接层输入特征的一部分，经过全连接层后得到特征表示结果。CNN特征提取具体过程如下：将词$W(i)$利用word2vec转化为对应的词向量$E(W(i))$，其中$E(W(i))\in R^k$代表句子中第i个词，词向量为K维，文本矩

阵表示为

$$\{E(W(1)), E(W(2)), \ldots E(W(m))\} \quad (1)$$

用 $h\times k$ 的滤波器对文本矩阵执行卷积操作，得到局部特征为

$$c_i=f(F\cdot E(W(i{:}i+r-1))+b) \quad (2)$$

式中：F 代表 $h\times k$ 滤波器，b 代表偏置量，f 代表通过 RELU 进行非线性操作的函数，$E(W(i{:}i+h-1))$ 为从 i 到 $i+h-1$ 共 h 行向量，c_i 为通过卷积操作得到的局部特征。随着滤波器依靠为 1 的步长从上往下进行滑动，走过整个句子，得到局部特征向量集合 $C_i \in R^t$。采用 n 个不同的滤波器对短文本中连续单词的 h 个窗口重复卷积运算，得到 $C_{1:m-h+1} \in R^{(m-h+1)\times n}$，采用 VALID 方式进行 padding 操作，获得与原输入相同长度的特征向量 $C_{1:m} \in R^{m\times n}$。

在生成器和判别器内部，使用堆叠式自编码机，其目的在于构建多层耦合网络，最后生成的复杂网络如图 5 所示。其中，$c_{i,t,n}$ 是得到的自编码，i 表示第 i 个网络节点，t 表示时刻，n 表示第 n 个隐藏层。定义顶点之间的欧氏距离（也可用余弦距离）

$$d_{i,j,t_1,t_2,n_1,n_2}=\left\|c_{i,t_1,n_1}-c_{j,t_2,n_2}\right\| \quad (3)$$

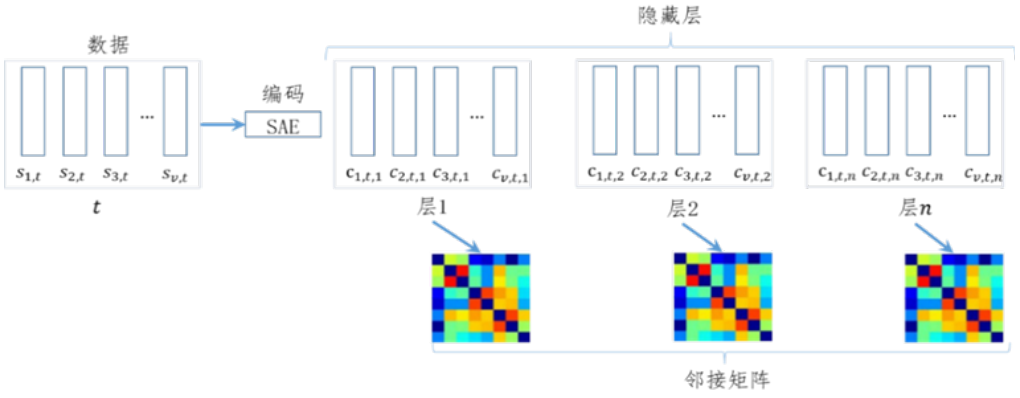

图 5 SAE 生成框架

（1）当 $n_1=n_2=const$ 时，构建的是时间上的耦合网络。

（2）当 $t_1=t_2=const$ 时，构建的空间粒度上的耦合网络。SAE 的输出编码是期望能够恢复源输入，不同隐藏层的维数为设定为不同，如果从第 1 层到第层中的节点逐渐减小，可以视为空间的粒度从细到粗的过程。因此，隐藏层的物理意义表示了空间粒度的不同，同时，SAE 的使用也相当于对复杂网络数据抽取进行了节点降维，通过使用 GAN 来生成虚假信息传播网络，其中生成器 G 试图生成顶点对，而鉴别器 D 试图区分生成的顶点对与网络实际连接的顶点对，采用 Wasserstein GAN 网络来训练，其目标函数为：

$$\min_{G_\theta}\max_{D_\phi}(\mathbb{E}_{e\sim\mathcal{P}_E}[D(e)]-\mathbb{E}_{e'\sim\mathcal{P}_{E'}}[D(e')]) \quad (4)$$

对于各个节点，可以沿用一般 SIR 模型对节点状态的定义，可将网络中的节点划分为以下四种状态，即节点标签可为节点标签 $Y=\{S, E, I, R\}$，其中易感状态 S、接收状态 E、传播状态 I、免疫状态 R。易感状态是指节点从未接收过网络中传播的虚假信息，即对该虚假信息处于未知时的状态；接收状态表示节点已经接收到网络中传播的虚假信息，但还未将该信息传播出去时所处的状态；传播状态是指节点已将网

doi:10.3772/j.issn.2095-915x.2022.05.008

络中传播的虚假信息传播出去后所处的状态；免疫状态是指节点完全不再接收网络中传播的虚假信息，并将不会再对其进行传播时所处的状态。通过对互联网传播信息的初筛，可以有效筛选出互联网信息的统计信息特征，为精准互联网虚假信息识别提供先验支撑信息。

4 实验仿真分析

本节针对提出的基于深度学习的互联网虚假信息识别平台进行实验验证，并对比提出的基于生成对抗网络的虚假信息筛选方法与传统方法的效能。近期由于国内外重大事件频发，各大互联网平台都出现了严重的舆论引导和虚假信息，致使官方频繁进行辟谣，并出台显示IP归属地的策略。实验验证以微博、微信朋友圈、知乎等平台作为主要信息源收集数据，并利用专家系统、基于自然语言处理的方法以及本文提出的基于深度学习的互联网虚假信息识别平台分别对数据进行识别。本文提出的系统基于Windows10系统，使用python3.8及C等语言进行开发。数据采用爬虫的方式获取，并根据网络信息对虚假信息进行人工标注。

表1展示了不同识别方法对不同数据类型虚假信息的识别准确率。当前的虚假信息可大致分为单一文本数据、单一视频数据、混合文本视频数据。所谓单一文本数据，即通过文章、评论、留言等方式在社交平台进行虚假信息传播的方式。所谓单一视频数据，即通过短视频、或AI换脸等智能手段篡改真实视频源的方法进行虚假信息传播的方式。所谓混合文本视频数据，即通过DeepFake等方法，通过文本和视频的方式，相互印证、传播，这类数据造成的危害相较于前两种更大。

表1 识别方法准确性比较分析(%)

识别方法 数据类型	专家系统	基于NLP的方法	基于深度学习的虚假信息识别平台
单一文本数据	85.2	90.1	95.7
单一视频数据	67.5	78.9	90.3
混合文本视频数据	54.2	67.5	86.4

实验结果表明，基于专家经验的识别方法，对于传统的文本数据有着较高的识别准确率，但由于专家信息库更新速度较慢，对于伪造的视频数据识别率较低。而对于多源异构的伪造数据，由于人工特征提取效率较低，所以识别准确度低，不能满足要求。而基于NLP的方法，利用深度神经网络提取文本和视频中的语音特征，可以较好的对单一来源的文本和视频数据中的虚假信息进行识别。但对于精心构造的DeepFake数据，由于算法缺少逻辑判断和基于经验的信息比对能力，效果也不理想。相比之下，本文提出的基于深度学习的虚假信息识别平台，在系统化的设计下，会对收集得到的数据先进行语义分析、预处理，再利用深度学习的方法进行信息聚类、识别。所以针对单一来源数据和多源异构数据都有较高的识别准确率。

图6展示了不同方法针对不同数据类型，在单位时间可以处理的数据量对比。可以看到由于专家系统依靠写定的规则进行判断，所以可对单一文本数据进行高速处理，但其对于视频数据特征提取效率低，并且由于专家信息库更新迭代慢，所以对于视频数据和混合数据的

处理速度都较慢。基于 NLP 的方法，需要通过神经网络对数据进行处理，在进行判定前需要对特征进行提取并进行语义分割、聚类等操作，所以对于三类信息的处理速度都较慢。而本文提出的架构，由于对采集得到的数据会进行渐进式、自下而上的语法分析，同时在运行过程中通过堆栈去保存有效数据，并根据语义分析的结果对搜索空间进行修剪。所以可以极大的提升信息特征提取和识别的速率，对于三类数据均有较高的处理速度。

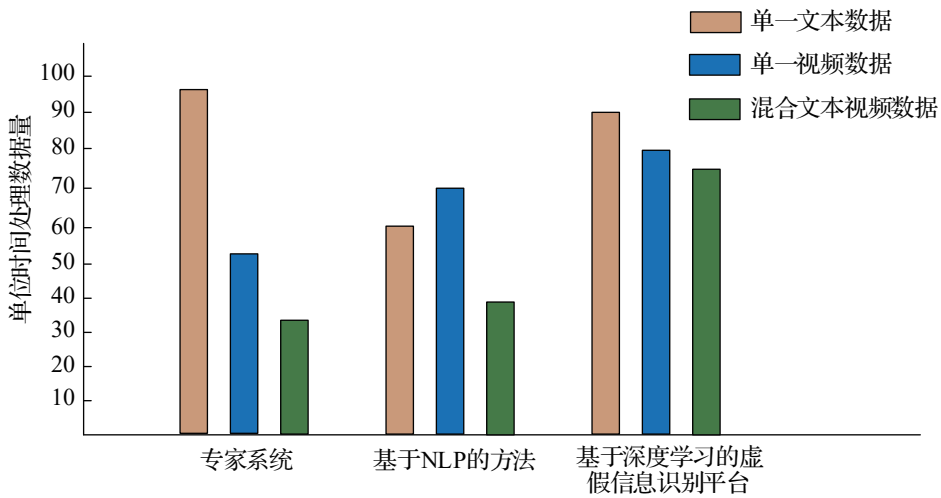

图 6 单位时间处理虚假数据量对比图

为了验证本文提出的基于生成对抗网络的信息筛选算法与传统的使用基于大数据的机器学习方法的性能对比。本文选取不同网络平台的数据，在经过相同的预处理操作后，分别用传统的方法和本文提出的方法进行虚假信息识别，结果如表 2 所示。改进后的算法在利用各网络平台采集得到的数据进行预筛选、倾向分析和虚假信息预警方面，相较于原算法，准确率都得到了较大的提升。原因在于基于生成对抗网络的利用多层耦合网络的架构，极大的提升了针对陌生虚假数据的识别的鲁棒性。同时基于 SIR 模型的判定模式，也使得算法可以有效生成对于信息特征的统计表示，为识别虚假信息提供先验经验。

表 2 筛选算法准确性比较分析 (%)

模型	原算法			改进算法		
	预筛选	倾向	预警	预筛选	倾向	预警
新浪微博	75.34	78.46	76.87	84.92	86.53	85.71
微信	75.98	79.63	77.76	85.14	84.02	85.57
抖音	80.71	81.37	81.04	87.86	86.91	87.38
豆瓣	83.15	87.26	85.16	89.13	87.55	88.33
知乎	88.53	92.57	90.50	92.82	93.35	93.08

5 结论

本文针对互联网虚假信息的基本概念、传输媒介、传播模型进行梳理和总结。结合近年研究工作，提出了基于深度学习的互联网虚假信息识别平台的架构，从样本采集、预处理、

doi:10.3772/j.issn.2095-915x.2022.05.008

信息识别和预警三个方面对该架构进行介绍。为了改进传统算法在信息识别上收敛速度慢、准确率较低的问题，本文提出一种基于生成对抗网络的虚假信息识别方法，并应用于虚假信息识别平台。实验表明本文提出的虚假信息识别平台相较于传统的专家系统和基于 NLP 的方法，在准确率和处理效率上都有较大的提升。而基于生成对抗网络的信息识别方法，相较于传统的基于大数据的机器学习方法，在预筛选、倾向分析、虚假信息预警上也取得了更好的表现。

参 考 文 献

[1] 肖慧鑫 . 拨开美军空间 “ 大数据 ” 的迷雾 [J]. 解放军生活 , 2018(9):78-79.

[2] Hindman M, Barash V. Disinformation, fake news and influence campaigns on Twitter[EB/OL]. [2018-10-04].https://kf-site-production.s3.Amazonaws.com/media_elements/files/000/000/238/original/KF-Disinformation Report-final2.pdf.

[3] Bradshaw S. The global disinformation order:2019 global inventory of organised social media manipulation[EB/OL].[2019-09-15].https://comprop.Oii.Ox.Ac.uk/wp-content/uploads/sites/2019/09/CyberTroop-Report19.pdf.

[4] Fallis D. A conceptual analysis of disinformation[J]. This Definition Comes from the American Heritage Dictionary of, 2009, 14(2):51-57.

[5] Bellutta D, King C, Carley K M. Deceptive accusations and concealed identities as misinformation campaign strategies[J]. Computational and Mathematical Organization Theory, 2021, 27(3):302-323.

[6] Boghardt T. Operation INFEKTION:soviet bloc intelligence and its AIDS disinformation campaign[J]. Studies in Intelligence, 2009, 53(4):1-2.

[7] 马超 . 基于主题模型的社交网络用户画像分析方法 [D]. 合肥：中国科学技术大学 , 2017.

[8] 李雅坤 . 用户画像与精准营销 —— 以微博为例 [J]. 财讯 , 2017(16):99-100.

[9] Jindal N, Bing L. Opinion spam and analysis[C]. Proceedings of the 2008 International Conference on Web Search and Data Mining. 2008:219-230.

[10] Ott M, Cardie C, Hancock J T. Negative deceptive opinion spam[C]. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013:497-501.

[11] Fusilier D H, Montes-y-Gómez M, Rosso P, et al. Detection of opinion spam with character n-grams[C]. International Conference on Intelligent Text Processing and Computational Linguistics. Springer, Cham, 2015:285-294.

[12] Yun K H, Yang W I, Lin T, et al. Judgment criteria for the authenticity of internet book reviews[J]. Library & Information Science Research, 2012, 34(2):150-156.

[13] Lim E P, Nguyen V A, Jindal N, et al. Detecting product review spammers using rating behaviors[C] Proceedings of the 19th ACM Conference on Information and Knowledge Management, 2010:939-948.

[14] Mukherjee A, Liu B, Glance N. Spotting fake reviewer groups in consumer reviews[C]. Proceedings of the 21st International Conference on World Wide Web. 2012:191-200.

[15] Wang G, Xie S, Liu B, et al. Review graph based online store review spammer detection[C]. 2011 IEEE 11th International Conference on Data Mining. IEEE, 2011:1242-1247.

[16] 苏鹏 , 王延飞 . 对信息迷雾的情报观察 : 概念 , 形成与应对 [J]. 情报理论与实践 , 2021, 44(3):7-9.

[17] Lanoszka A. Western Intelligence and Counter-intelligence in a Time of Russian Disinformation.[EB/OL]. [2016-07-15]. https://openaccess.city.ac.uk/id/eprint/16239

[18] Nakov P, Alam F, Zhang Y, et al. QCRI's COVID-19 Disinformation Detector: A System to Fight the COVID-19 Infodemic in Social Media[J]. Infopreneurship Journal, 2022, 28(5):114-120.

[19] Skarzauskiene A, Maciuliene M, Ramasauskaite O. The digital media in lithuania: combating disinformation and fake news[J]. Acta Informatica Pragensia, Preprint, 2021, 36(7):45-60.

[20] Bryant A G, Narasimhan S, Bryant-Comstock K, et

al. Crisis pregnancy center websites: Information, misinformation and disinformation[J]. Contraception, 2014, 90(6):601-605.

[21] Tucker J A, Guess A, Barbera P, et al. Social media, political polarization, and political disinformation: a review of the scientific literature[J]. SSRN Electronic Journal, 2018(64):310-321.

[22] Bushman B J, Anderson C A. Media violence and the American public. Scientific facts versus media misinformation.[J]. American Psychologist, 2001, 56(6-7):477-489.

[23] Keshavarz H. How credible is information on the web: reflections on misinformation and disinformation[J]. Infopreneurship Journal, 2014, 1(2):1-17.

[24] Riedel B, Augenstein I, Spithourakis G P, et al. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task[J]. Library & Information Science Research, 2017, 53(2):427-435.

[25] Oates S, Gray J. Kremlin: using hashtags to analyze russian disinformation strategy and dissemination on Twitter[J]. Social Science Electronic Publishing, 2020, 16(9):438-450.

[26] Pennycook G, Rand D G. Assessing the effect of "disputed" warnings and source salience on perceptions of fake news accuracy[J]. SSRN Electronic Journal, 2017, 19(12):376-389.

[27] Qawasmeh E, Tawalbe H M, Abdullah M. Automatic identification of fake news using deep learning[C]. 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE, 2019:147-162.

[28] Katarya R, Massoudi M. Recognizing Fake News in social media with deep learning: a systematic review[C]. 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP). 2020:81-102.

[29] Botha J, Pieterse H. Fake News and deepfakes: a dangerous threat for 21st century information security[C]. 15t International Conference on Cyber Warfare and Security. 2020:187-201.

[30] 于海洋 . 基于社交影响力的信息溯源及传播机制研究 [D]. 重庆 : 重庆邮电大学 , 2019.

[31] Chen D B, Lü L Y, Shang M S, et al. Identifying influential nodes in complex networks[J]. Physica A Statistical Mechanics & Its Applications, 2012, 17(6):208-217.

[32] Barbieri N, Bonchi F, Manco G. Topic-Aware social influence propagation models[J]. Knowledge and Information Systems, 2012, 37(3):45-55.

[33] Yao Z, Adiga A, Saha S, et al. Near-Optimal algorithms for controlling propagation at group scale on networks[J]. IEEE Transactions on Knowledge & Data Engineering, 2016, 28(12):3339-3352.

[34] Nowzari C, Preciado V M, Pappas G J. Analysis and control of epidemics: a survey of spreading processes on complex networks[J]. IEEE Control Systems, 2016, 36(1):26-46.

[35] Xiao Y, Wang Z, Li Q, et al. Dynamic model of information diffusion based on multidimensional complex network space and social game[J]. Physica A: Statistical Mechanics and its Applications, 2019, 521:578-590.

[36] Su Y, Zhang X, Liu L, et al. Understanding information interactions in diffusion: an evolutionary game-theoretic perspective[J]. Frontiers of Computer Science, 2016, 25(20):389-401.

[37] 李学勇 , 欧阳柳波 , 李国徽 , 等 . 网络蜘蛛搜索策略比较研究 [J]. 计算机工程与应用 , 2004(4):128-131.

[38] 林亚平 , 刘云中 , 周顺先 , 等 . 基于最大熵的隐马尔可夫模型文本信息抽取 [J]. 电子学报 , 2005, 33(2):236-240.

[39] 洪铭材 , 张阔 , 李涓子 . 基于条件随机场 (CRFs) 的中文词性标注方法 [J]. 计算机科学 , 2006, 33(10):148-151.

[40] Saunders C, Stitson M O, Weston J, et al. Support vector machine[J]. Computer Ence, 2002, 1(4):1-28.

[41] 谢铁 , 郑啸 , 张雷 , 等 . 基于并行化递归神经网络的中文短文本情感分类 [J]. 计算机应用与软件 , 2017, 34(3):8-15.

[42] 谢红薇 , 颜小林 , 余雪丽 . 基于本体的 Web 页面聚类研究 [J]. 计算机科学 , 2008, 35(9):3-9.