大数据背景下文本语料预处理技术项目探析

许　越　黄思缘　吴佳怡　顾　秦　王　络
（上海立信会计金融学院，上海 201209）

摘　要　随着计算机智能化技术发展的提升,越来越多的人有条件利用智能设备进行网上娱乐活动。但随着用户数量的增加，评论区会出现一些不文明、不理智的发言。本项目将通过对于不文明用语的及时识别和屏蔽，降低用户在网络上与其他用户起冲突的可能性，也在一定程度上减轻了网络不文明现象可能给被攻击用户造成的负面心理影响。

关键词　文明网络交流环境　文本分析技术　人工智能
中图分类号：H0；TP311　　**文献标识码**：A　　**文章编号**：1007-0745(2021)06-0015-03

1 项目价值和意义

随着人们生活水平的不断提高，智能设备已经逐渐成为了人们生活中的必需品，越来越多的人使用智能设备在网络上通过各种软件进行线上社交活动，而在这个过程中，不可避免地会产生网络不文明用语现象。软件上的发布动态、评论、转发、聊天等功能给人们提供了一个更方便地进行思想交流的平台。但与此同时，网络的匿名性也导致了各类网络不文明现象的频发。言语上口无遮拦的攻击谩骂对网络环境和被攻击者的心理都造成了极其恶劣的影响，对于攻击者本人正确的思想道德培养也有一定的阻碍。[1]

如今，国内人工 + 机器的不文明用语审核方式虽已在一定程度上提升了识别效率，但仍无法赶上用户创造网络用语的速度，识别的准确度难以得到提升。且目前的机器识别方式仍较死板，只能通过简单的文本比对机械地识别出某个字或某个词，不能联系前后文本完整地识别出语句的准确含义，因此有时会导致原本不存在不文明用语的文本被错误地识别、屏蔽，给用户的线上交流带来一定困扰，也降低了用户的软件使用体验。而真正使用了不文明语言的文本也可能因为使用了替代词而没有被及时识别屏蔽，破坏了良好的网络语言环境。本项目将通过文本预处理、建立语料库、不文明用语库等方式，使用相似性比较，聚类分析等文本挖掘技术，实现对不文明用语更高速、更准确的识别处理。

本项目的意义可以体现在用户、网络平台、社会及人工智能发展四个方面：首先，对用户而言，本项目将通过对于不文明用语的及时识别和屏蔽，降低用户在网络上与其他用户起冲突的可能性，也在一定程度上减轻了网络不文明现象可能给被攻击用户造成的负面心理影响，同时能有效避免为防止踩中屏蔽词只能使用替代词进行交流的情况，增强用户的沟通效率，提升用户的软件使用体验，为交流双方提供一个更加健康的网络环境；其次，对有用户留言、评论、转发等各类功能的网络平台而言，本项目能为其提供更高效的用户留言管理方法和策略，创建良好的网络语言环境，减少人工审核不文明用语的成本。同时也能为用户创造一个更优秀、更文明的线上交流平台，提升用户的使用体验，使得用户愿意更频繁地使用该平台进行线上交流，为平台增加收益。本项目也能帮助平台通过某一词汇的提及度了解用户对于某一话题的关注度，为网络平台业务开展和话题建设打下良好的基础。增强平台对网络话题趋势的掌握度，更清晰地了解用户喜好，为软件的功能提升提供方向，吸引更多用户，最终达成良性循环；再次，对社会而言，网络肩负着引导舆论、成风化人的职责，使用文明规范的语言文字是传承文明、传播文化的基本要求。本项目能够减少网络上不文明用语的出现频率，从而减少不文明用语对社会风气产生的不良影响。同时，对于网络上数量庞大的未成年用户而言，一个文明和谐的网络语言环境会对他们的身心健康发展起到良性引导的作用。也在一定程度上降低了线上的语言暴力给用户的身心所带来的危害；最后，对人工智能的发展而言，人工智能本就是在不断学习中成长，通过对互联网上大量的语言识别样本进行学习意味着能够使人工智能更精确地识别出当前文本的真实语义，甚至识别出带有更强烈的情感色彩的调侃、讽刺等语气的文本含义，避免错误的识别屏蔽，达到优化用户体验的目的。提升人工智能在语言识别方面的成长进度，为未来人工智能的发展打下基础。

2 项目设计

2.1 研究对象与研究方法

项目灵感来源于大一上学期我们在思想政治课上研究的课题——上海市大学生对于网络道德的认识。该研究通过向大学生发放纸质问卷和电子问卷的方式进行调查，采取简单随机抽样的方式发放问卷。研究目的在于从整体上探究大学生对于网络道德的认知程度，从人们对于网络持有的意识态度、网上行为规范、评价选择等方面设计问卷。同时，结合了校内校外随机采访辅助前期调研，侧重对访问者在网上冲浪时对于不文明或具有煽动性的言论的真实感受。同时请大学生对于制止网络暴力给予适当的建议。希望通过丰富的问卷内容体现出大学生真实的心理状态，从而进一步探究解决网络暴力以及网络不文明现象的有效手段。

2.2 样本的概况及分布

本次研究在上海立信会计金融学院等学校共发放 150 份

★基金项目：上海立信会计金融学院大学生创新创业训练计划项目资助，项目号：S202011047069。

纸质调查问卷，有效问卷 112 份。在性别比例上，参与调查的男生占 20%，女生占 80%。其中大一学生为本次研究着重调查的对象，占据 80%。另外还有 14.67% 的大二学生，2.67% 的大三学生和 1.33% 的大四学生参与了调查。

2.3 理论综述

当下，大学生是使用网络最频繁、耗时最多的社会群体之一。根据数据统计，62.67% 的大学生平均每天会花费 4 个小时以上的时间在网络上，而在其中，有 68% 的大学生会把大部分时间花在社交媒体上，可见网上交流是大部分大学生必不可少的社交手段，如今常用的社交媒体包括在全国甚至全球关于娱乐休闲生活信息分享交流的平台。通过数据显示，82.67% 和 80% 的大学生把微信和 QQ 作为常用的社交软件。此外，还有 44%、10.67% 和 5.33% 的大学生分别把微博、贴吧和直播网站这样具有互动性、透明性、公开性的网上交流平台作为常用的社交软件。其中的互动性就体现在媒体会为那些看到信息的人提供自由评论的区域，让他们发表看法，这样的设计让互不相识的人通过网络建立起了联系，为网上冲浪增添了许多乐趣。

但是人们对待同一事物的看法不可能完全相同，有时候针对某个观点难免会起纷争。通过问卷调查的数据，41.34% 的大学生无法做到在阅读完信息后理性地判断内容的真实性再转发评论，从而导致某些不慎或者过激的言论成为扰乱网络秩序的源头，网络暴力也由此而生。

据调查，超过四分之一的大学生遭受过网络暴力，其中有 17.33% 的大学生以个人行为代替报警或举报维权进行回击，而 9.33% 的大学生只选择默默忍受或不予理睬。可见对于网络暴力的迫害，不是所有的大学生都能采用正确的渠道合理地进行解决。有专家指出，网络暴力会带来道德绑架、舆论嘲讽、虚假信息和侵犯隐私四种危害。如果没有有效的手段来治理网络暴力，势必会对大学生乃至所有网民产生严重的影响。

为了营造和谐的网上交流环境，相关的平台为用户设置了举报系统。当读者浏览到垃圾营销、涉黄信息、人身攻击、有害信息以及违法信息时，可以按类型向平台进行投诉，平台的工作人员也会马上进行反馈。

针对这一点，我们小组设计了相关问题来调查大学生是否能有效利用此类举报系统。

经数据统计，面对不良信息只有 45.33% 的大学生能够理性地举报所有他们认为的不良信息，多数大学生只是看心情举报，少数则是不予理睬或是凑个热闹，这表明只有一半不到的大学生能有效利用平台设置的举报系统。大部分的大学生理应具备识别网络暴力的能力，但为什么这类系统不能被大学生完全利用到位？提出疑问后，我们紧接着就大学生面对网络暴力所持有的态度展开调查。

根据数据显示，超过四分之一的大学生面对网络暴力表示无所谓、看热闹或是低估了网络暴力带来的伤害。由此可见，从用户角度来说，平台设置的举报系统一定程度上可以惩治发表不良言论的人，但还有一大批未能被举报的用户成为漏网之鱼，同时，仍有一部分人因为对待网络暴力的态度不同而未能及时制止使得事态恶化；从平台自身来说，举报系统的不完善同样会让部分用户利用平台的漏洞，不断散播不良信息，这两点让网络暴力的问题无法得到有效的根治。

所以，为了打造更加文明的网络环境，我们小组决定从用户发布信息的源头探究在信息发布栏里加入文本分析的技术，通过文本预处理、建立语料库、不文明用语库等方式，使用相似性比较、聚类分析等文本挖掘技术，对评论者发表的留言、评论进行识别，提取文本特征，计算其与不文明用语语料的相似性。从而能够相对快速、准确地对用户留言进行及时的处理，识别其中的不文明用语并通过限制发文、信用打分等方式对留言者进行标识和评价，从而起到一定的警告作用。

3 项目方案

3.1 项目的主要问题

3.1.1 评论数据的收集以及数据的处理

我们需要大量的数据建立屏蔽词的语料库与是否屏蔽的数据库，首先要解决的是如何获取大量真实可靠的清洁数据，而数据的处理方式需要运用大量实践去建立初步模型决定采用的预处理方式，是本次项目的重难点，需要我们运用数据科学知识找到最有效的途径。

3.1.2 建立文明用语的语料库

为了实现屏蔽机制，我们需要将网络上的各种语言分类为文明用语、不文明用语和侮辱性用语。因为数据较为庞大且存在大量的俚语、隐晦语、网络用语、符号等，如果要全面准确地识别隐藏其中的不文明用语，需要合适的文本分析挖掘方法。

3.1.3 网络环境维护方案优化策略

在识别了是否需要屏蔽数据之后，我们需要采取一种相对合适的方式来优化，例如直接屏蔽、将屏蔽部分的不文明用语替换成文明用语，并对用户进行警告，设置一定限度的禁言措施。但过度的警告措施会引起用户反感，所以需要大量数据来确定措施的力度对用户的影响，在维护网络环境的同时最大程度地保证用户对平台的驻留。

3.2 拟解决途径

3.2.1 数据的采集

我们准备选取当下在大学生中较热门的网站，比如微博、易班，在这些以评论作为主要交流方式的平台上可以更简单地获取信息，且网站中较大的流量可以获得更庞大的数据，为之后建立数据库和处理数据打下基础。运用爬虫作为搜集数据的工具可以快速准确地搜集到大量数据，减少人工搜集的难度。

3.2.2 数据预处理

首先对于被爬取的数据需要过滤污染数据，去除重复数据，并去除无关消息，得到较为干净的数据。中文语料数据大多为短文本或长文本。通过 jieba 和 HanLP 等较为简单的中文分词器与词性注解的方式将较为长的文本分为我们需要的词，运用去停止词、特征提取、TF-IDF 权值计算等方式，将文本留言转化成数据向量，使用文本相似性计算，Logistic

（下转第 27 页）

的选取原则为布设在工程影响范围以外的的永久建筑物上，结合本项目实际情况，基准点为热电厂院内冷却塔基础和原防护网基础（桥梁工程平面控制点）。基准点周围做防护标记，防止热电厂操作机械磕碰及人为损坏导致基准点失效。

加载之前的监测点标高；每级加载后监测点标高；加载至 100% 后每隔 24h 监测点标高；卸载 6h 后监测点标高。通过各监测点位的标高计算各点的沉降量、沉降差及沉降速度。在预压过程中对支架的沉降进行监测并记录。

在全部加载完成后的支架预压监测过程中，当满足下列条件之一时，应判定支架预压合格：各监测点最初 24h 的沉降量平均值小于 1mm；各监测点最初 72h 的沉降量平均值小于 5mm。

对支架的代表性区域预压监测过程中，当不满足上述规定时，应查明原因后对同类支架全部进行处理，处理后的支架重新选择代表性区域进行预压，并满足上述规定。

综合支架变形、混凝土的收缩徐变以及温度变化产生的变形、预应力作用、预制箱梁架设后的梁板荷载等因素，提前设置预拱度，为支立模板的标高提供依据。

与传统的施工方案相比，本文提出的现浇盖梁施工技术方案具有以下主要特点（见图 2）：

（1）节省了机械和人工成本、缩短了工期。利用既有承台作为承载基础。

（2）降低了施工过程中安全风险。

（3）提高了周转率，降低了施工、管理成本。钢构支架安拆均采用汽车吊整体吊装、拆卸，操作简单快捷，其装、拆需时短，周转速度快，可有效的减少支架、设备配置数量。

（4）充分利用了空间钢结构整体性好的特点，采用部分构件定型，部分联接件可旋转伸缩的钢构件来进行盖梁支架拼装连接。[4]

3 结语

城市高架桥是城市交通网络的重要组成部分，在城市高架桥的施工过程中做好城市高架桥大悬臂支架技术的应用，对于确保高架桥的施工质量及施工效率十分重要，文章在分析城市高架桥支架技术的基础上，对如何做好城市高架桥支架技术的施工与应用进行了分析阐述，对此类桥梁的施工有着重要意义。

参考文献：

[1] 周正海 . 现浇支架施工技术在城市高架桥中的应用 [J]. 科技创新与应用 ,2016(03):242.
[2] 杨锐 , 杨云峰 , 朱漪 . 基于“景观基础设施”理念的城市高架桥整治——以宁波机场高架快速路为例 [J]. 中国园林 ,2014,30(05):69-73.
[3] 张晓斌 . 合理开发利用城市高架桥下的空间资源——兼论宁波城市高架桥下空间利用的初步设想 [J]. 宁波经济（三江论坛),2012(07):17-19,27.
[4] 钟杰 . 城市立交桥连续箱梁现浇支架设计方案、施工技术及强度验算 [J]. 建筑工程技术与设计 ,2014(02):210,222.

（上接第 16 页）

回归分析等分析方法，将其与留言特征语料库中的样本数据进行比较分析，从而识别出不文明用语。[2-3] 基于 Python 的内环境，可以运用现阶段我们在 Python 上的学习经验。

3.2.3 文明语料库的建立

我们初步采用人工采集与数据分类的方式建立文明语料库，采集了约 500+ 的屏蔽词。并使用线上问卷的形式向大学生群体征集希望被屏蔽的语料，通过数据查重的形式，将重复数据清除。

3.2.4 网络环境的维护以及优化方案

参考我们小组曾经参与的“关于大学生网络道德问题调查问卷”，将问卷面向的对象扩展为群众，了解不同年龄段人群对于不文明用语及当下由网络不文明现象引发的社会热点的印象和看法，在此基础上增加受访者对于各类不文明用语的的容忍度与希望惩处的力度。在数据库建立之后接受使用者增加新出现屏蔽词，提高数据库维护的效率，使之更人性化地维护网络环境。除此之外我们希望可以通过词意解析的方式，将屏蔽词去除并替换成文明用语表达原有的意思。

4 项目特色与创新点

此项目是基于时下网络暴力造成的抑郁症和自杀的已成为热点话题的案例，针对网络环境净化热点问题采取的解决措施。进入新世纪以来，互联网带给我们的生活和工作上的改变是以肉眼可见的速度持续增加的，由于互联网的介入，工作效率越来越高、生活的便利性越来越大，但随之而来的是互联网不断发展之下网络环境的有待改善。随着网络用户不断增多，网络上的不文明行为也逐渐增加，由此而导致的网络暴力事件也屡见不鲜，希望能通过此项目阻止此类行为的发生，打造一个更文明的网络交流环境。

对大数据进行科学研究，建立文明用语语料库，活用专业技术。建立文明用语语料库，针对大学生这个特殊群体，体现该年龄层用语特色，随时更新导入网络流行用语和游戏用语，从词汇、短句应用，语言习惯，措辞方面便捷有效地阻止不文明用语的出现，并给出一定的预警和管理措施。

参考文献：

[1] 李然 , 林政 , 林海伦 , 王伟平 , 孟丹 . 文本情绪分析综述 [J]. 计算机研究与发展 ,2018(01):32-45.
[2] 石凤贵 . 基于 jieba 中文分词的中文文本语料预处理模块实现 [J]. 电脑知识与技术 ,2020(14):254-257.
[3] 李宝玲 , 郭立鑫 , 李珂 . 基于 HanLP 的档案智能检索系统研发与应用 [J]. 档案管理 ,2020(06):43-45.