

我国自出版平台的作品内容管理研究

白广达 翁子扬

(武汉大学信息管理学院, 武汉 430072)

摘要: 我国自出版行业的快速发展让众多的新作品得以面世,但同时也带来作品内容良莠不齐的问题,亟需自出版平台对作品进行有效管理,而平台方在作品内容管理方面面临作品数量众多、违规文本变形、外链审查困难等挑战,对此本文从技术和管理的层面提出3点建议,分别是完善平台内容管控流程、构建多层次策略的自然语言处理(NLP)能力、对外链内容和外链域名检测以实现外链审查。

关键词: 自出版;自出版平台;内容管理

中图分类号: G23

DOI: 10.3772/j.issn.1673-2286.2020.05.009

1 自出版及自出版平台

自出版(self-publishing)是指图书或者其他形式的出版物在没有出版商按其常规出版流程进行策划、出版、发行的情况下,由作者主导推动进行的特殊出版业态^[1]。它强调网络平台的重要作用,形成了“作者-平台方-读者”的传播路径^[2]。与传统出版业态相比,自出版最突出的特点是作者在出版过程中自主性强、对出版各环节的掌控程度高。

早在19世纪,自出版就已经在英美等国家出现萌芽,其主要是通过作者自费以及包购包销的形式为被传统出版机构拒绝的作品争取出版机会^[3]。进入21世纪,随着互联网技术、数字出版和社交网络的发展,现代意义上的自出版从美国开始兴起,以KDP(Kindle Direct Publishing)、LuLu、Smashwords为代表的自出版平台为作者提供了更快的出版速度、更高的销售收入分成和更强的作品控制力,从而吸引了大量作者采用自出版模式^[4],2008年美国出版史上第一次出现自出版图书种数超过传统出版图书的情况,2009年全美76%的新书为自出版图书^[5],自出版已成为美国数字出版业的重要分支。

我国的自出版发展时间尚短,但也出现了一批具有本土化特色、形态各异的自出版平台,可分为4类,即网络文学自出版平台、专业性自出版平台、传统出版企业

自出版平台和自媒体自出版平台。

(1) 网络文学自出版平台。以起点中文网为代表的网络文学自出版平台脱胎于本土的网络文学网站,这些网络文学网站在自出版出现之前已广为流行,并且在发展中逐渐形成“初始免费+后续付费+纸质出版”的自出版模式。网络文学自出版平台上的作品主要采取连载发布的形式,读者与作者的互动密切,使得作品的“读者本位”色彩非常明显,是兼具争议和热度的一类作品,其中典型代表有南派三叔的《盗墓笔记》以及当年明月的《明朝那些事儿》。

(2) 专业性自出版平台。这一类型的自出版平台脱胎于以电子书为主要产品的数字出版平台,目前发展态势较好的有豆瓣阅读、网易云阅读等,它们最初的业务是为读者提供纸质出版物的数字化资源,后来为发展原生数字出版物引入了作者认证系统,直接由平台方签约作者,使其作品直接面对读者并获取收益。该类自出版平台和亚马逊公司的KDP平台相似,主要流程为作者按照规定上传书稿,在数字编辑审核之后在平台上进行发布,并且作者在出版物价格上也有比较高的自主权^[6]。

(3) 传统出版企业自出版平台。自出版行业的快速发展也吸引了一些传统出版企业开展自出版的尝试,这一模式的典型代表是知识产权出版社推出的“来出书”自出版平台,由于背靠知识产权出版社的行业资源,

“来出书”自出版平台可以为作品提供完整的三审三校环节并帮助作品取得书号。相较于其他的自出版模式,传统出版企业自出版平台下作者对作品的主导地位稍弱、出版商参与程度更高。

(4) 自媒体自出版平台。自媒体自出版平台的内容主要来自以博客、微博、微信为代表的社交媒体,目前流行的“博客书”“微博书”“微信书”是其代表产品。以微信书为例,自媒体自出版平台以微信为入口,在获得用户许可授权之后,抓取用户的微信朋友圈或指定的微信公众号内容并批量导出,将内容套用平台模板后一键成书,可为用户提供电子版或纸质版的成品形式。这类平台的出版服务较为单一,具有比较高的可定制性和自发性。

2 我国自出版平台在作品内容管理中的难点

自出版简化了出版流程、降低了作品的出版门槛,这一方面使得众多的新作品得以面世;另一方面也不可避免地带来作品内容良莠不齐的问题^[7],自出版作品中含有淫秽色情、封建迷信等违规内容的问题时有发生^[8],这需要自出版平台对作品内容进行有效管控,而相较于传统出版业态下出版机构对出版物的审查管理,自出版平台对于自出版作品内容的管控还存在诸多难点。

首先,相较于传统出版物,自出版平台上的自出版作品数量众多。《2018年新闻出版产业分析报告》显示,我国2018年新出版图书24.7万种;阅文集团发布的2018年年报显示,截至2018年平台拥有作品数已超过1 120万部、2018年平台新增字数多达443亿,体量庞大的待审查内容使得自出版平台的内容管理任务十分艰巨。

其次,随着文本审核机制在各平台的普遍介入,违规内容也开始出现大量变形以期规避平台的审核。在自出版平台中常见的敏感词变形方式有拼音替换、拼音首字母替换、形近字替换、同音字替换、繁体字替换、调换顺序、插入无意义汉字/符号/数字/字母等,敏感词变形的种类繁多、形式多样,这对关键词硬匹配的词库策略机器审核提出了挑战。

最后,微信书等自媒体自出版作品中存在大量以外链形式组织的内容,对于外链的审查是自出版平台对作品内容管理的难点和盲点。外链审查不能单纯沿用文本内容的机器审核方法,且外链的入口链接之间可能存在多次跳转,这些都给自出版平台对外链的有效审查增

加了困难。

需要说明的是,传统出版企业自出版平台的作品数量少、内容形式相对单一,并且有完整的三审三校环节,这类自出版平台并没有面临上述作品内容管理方面的问题和挑战,因此不在本文的讨论范围之列。

3 加强我国自出版平台作品内容管理的对策建议

针对我国自出版平台在作品内容管理中遇到的难点,本文从技术和管理的层面提出了3点建议,以提升自出版平台对作品内容的管理水平。

3.1 完善平台内容管控流程

自出版平台应完善覆盖作品全过程的内容管控流程,不仅包含作品发布前的内容甄别和审查,也应涵盖作品发布后的及时阻断和复核溯源,进而有效控制自出版作品中违规内容的出现和传播。

平台方应向自出版作者明确作品内容规范,要求自出版作者的作品内容遵守平台服务协议,不得出现涉及违反国家法律法规或相关规则的内容,同时应明确对违规内容的处理机制,包括且不限于要求内容整改、删除违规作品、封禁账号或销号处理等,确保自出版作品的创作主体对于作品内容规范和处理机制应知尽知。此外,还应应对自出版作者进行实名认证,既可以更好地维护自出版作者的版权权利和经济利益,也可以加强自出版作者的责任意识和主体意识。

对于自出版作者上传的作品,先经过机器审核完成系统自动过滤,对于机器审核识别含有违规内容的作品进行删除等处理,通过机器审核的作品进入人工审核,人工审核甄别通过的作品才可以平台发布,作品内容的风险管控务必要做到先审后发,落实事前预防和审核。对于浏览量高、转发分享次数多、享有平台推荐位或资源位的自出版作品,还应进行多次人工审查、重点查验。

对于已发布的作品,平台应对用户举报作出快速反馈、对被举报作品接入人工复查,如发现是发布前审查漏审的违规内容应立即删除,做到及时阻断。

此外,平台对于因漏审导致的违规内容传播应及时复盘,了解导致漏审、错审的原因并修正,如针对违规内容变形导致的漏审应调整相应的机器审核过滤规则

和算法组合、因敏感词更新不及时导致的漏审应更新扩容敏感词词库等,通过事后复核溯源提升平台内容管理水平。

3.2 构建多层次策略的NLP能力

在对于文本内容的审查中,NLP能力是文本机器审查的关键,目前自出版平台的机器审查多停留在关键字硬匹配的阶段,属于泛化能力相对弱的词库策略,审查命中效果有限^[9],为此自出版平台方应当构建包含词库策略、先验策略、模型策略在内的多层次策略的NLP能力,提高对违规内容机器审查的识别效率。

(1) 词库策略。词库策略包括预处理和硬匹配两部分,引入预处理是为提升文本审核的泛化能力,硬匹配将待审核文本与敏感词库进行匹配以识别包含本体敏感词内容的违规文本,是一种直接而有效的策略。

预处理一般包括英文大小写归一化、繁体转换和特殊符号过滤。预处理可以过滤容易对先验策略和模型策略模块造成干扰的特殊符号,并通过英文大小写和繁简体归一化起到泛化语料、间接扩充词库的作用^[10]。

硬匹配的关键在于根据语料建立的文本审核基础词库,目前国内自出版平台主要建立并运营维护了包含违规敏感词的主词库和边缘敏感词的人工复核词库,能够提供基本的文本审核能力,但识别变体违规内容的能力较弱,因此,还应在本体词库外建立包括拼音库、首字母库、缺字词库、跳字词库等在内的变体词库。变体词库不仅能识别特定变形的违规内容,还可以与先验策略模块中的对抗规则结合进一步覆盖更广、更深的变种类型。

(2) 先验策略。先验策略主要是针对为躲避审核机制而进行的违规文本变形问题,是一种文本对抗规则的策略。

目前常见的敏感词变形策略有拼音替换、拼音首字母替换、形近字替换、同音字替换、繁体字替换、调换顺序、插入无意义汉字/符号/数字/字母等方式,针对不同类型的文本变形,可以通过分层过滤、跳字匹配、少字识别、拼音识别、首字母识别等对抗规则的组合使用,提高对敏感词变形的有效识别率。

跳字匹配、少字识别、拼音识别、首字母识别对抗规则是以跳字词库、缺字词库、拼音库、首字母库来实现的,可以由包含违规敏感词的主词库批量转化生成。分层过滤是以算法而非词库的形式来实现,对待审查文

本进行符号过滤、符号数字过滤、符号字母过滤等多层次的过滤,逐渐拓宽干扰项覆盖范围,并能够将不同干扰项相互剥离以精准定位变体敏感词。

先验策略中的对抗规则主要是通过词库构建和算法设计来实现,通过对包含违规敏感词的主词库进行拓展和对不同类型干扰项的多层次过滤,就可以有效提升对违规内容变形的识别能力,相较于词库硬匹配策略对抗能力更强,且实现的复杂度并不高,对于自出版平台而言是提升机器审查效率的一种高可用性策略。

(3) 模型策略。当一些违规文本并不会显著触发敏感词或者一些边缘敏感词需结合语境来判定是否违规时,这就需要引入各种NLP深度学习模型,运用模型策略来捕捉语义层面的违规内容。

适用于NLP领域的常见模型有transformer、BERT、FastText等,目前应用较多的是基于卷积神经网络(Convolutional Neural Networks, CNN)展开的文本分类模型^[11],一般过程是先对文本进行分布式表示,然后进行卷积操作形成张量表示,再通过最大化池操作降维,最后通过感知运算层进行预测和分类。利用深度学习模型能够进行词法分析、句法分析和篇章分析,理解文本内容的情感倾向、分类和主题,进而判断文本涉及敏感信息的程度^[12]。

模型策略的实现复杂度高,涉及模型选型、训练工作、参数调整和模型优化,还需要保证训练语料的覆盖度以实现模型的泛化能力,建议平台方结合实际需求自行选用。

3.3 对外链内容和外链域名检测以实现外链审查

作品中出现以外链形式组织的内容是自出版作品相较于传统出版作品的一个显著区别。以微信书为例,其内容主要来自用户朋友圈或指定的公众号推文汇总,微信书中常见的外链既有指向某一公众号推文的跳转,也可以是指向外部网页的链接,这些外链作为自出版作品中出现的超文本信息,体现了自出版作品作为数字出版物在作品内容格式上的丰富性。封禁恶意外链、避免给涉黄涉赌站点引流也是自出版平台的内容管理工作之一,但审查外链内容与纯文本的内容审核有一定差异,目前国内有外链审核需求的自媒体自出版平台并没有对作品中出现的外链进行审核,本文试提出了外链审查的技术路径以供参考。

外链审查的技术路径设计主要是针对外链内容和外链域名的检测。首先应细分链接的层级(URL、CGI、PATH、HOST、DOMAIN),拉取网页源代码并提取网页的文档对象模型结构(DOM)和标题内容作为特征^[13],对于拉取的特征信息以NLP手段进行文本内容审查^[14],对于机器审查出现违规内容的外链直接予以封禁,对于机器审查结果为有违规可能性的边缘性敏感信息的外链接入人工审查进行复核。

为避免对所有外链均进行抓取及内容审查带来的低效率,可以通过对外链域名的检测来进一步优化外链内容审查的技术路径。可以对链接层级细分中的域名设置白名单(建议是域名维度,CGI维度太过细化),将政府门户网站和知名互联网大站等优质站点列入域名白名单,对于这些入口链接信誉较好的外链可以不接入内容审查的程序接口,节省计算资源。

需要注意的是,互联网“黑产”也可能会利用XSS漏洞(跨站脚本攻击)将自己的恶意外链隐藏在优质站点域名中^[15],对此需要对列入白名单的站点建立跳转围墙,因为列入白名单的站点本身较为稳定、不大可能会经常跳转到业务以外的域名,所以可以通过历史跳转数据对优质站点的跳转建立围墙,加大力度对这些站点跳出的域名做监控打击以解决XSS漏洞类问题。

参考文献

- [1] 孙赫男. 自出版在传统与数字出版业态中的发展状况及对策分析[J]. 出版广角, 2013(12): 23-26.
- [2] 段淳林, 吴哲. 基于大数据的国内自出版平台化模式研究[J]. 中国出版, 2016(15): 38-41.
- [3] 李欣人, 陶庆娟. 2015年美国自出版发展概述[J]. 出版发行研究, 2016(5): 86-89.
- [4] DOMINICK J R. The dynamics of mass communication: Media in the digital age [M]. New York: Tata McGraw-Hill Education, 2010.
- [5] 刘蒙之. 美国自助出版的发展现状与存在问题评析[J]. 现代出版, 2012(11): 63-67.
- [6] 冯静. 豆瓣阅读的自由出版模式及其文化意义[J]. 编辑之友, 2013(11): 33-35.
- [7] 王蓓. 我国网络自出版发展现状与对策分析[D]. 开封: 河南大学, 2015.
- [8] MICHEL H. Internet Literature in China [M]. New York: Columbia University Press, 2015.
- [9] 宋嘉庚. 自出版管理问题研究[M]. 北京: 中国传媒大学出版社, 2016: 12.
- [10] DAI Z, YANG Z, YANG Y, et al. Transformer-XL: Attentive language models beyond a fixed-length context [EB/OL]. [2020-02-08]. <https://arxiv.org/pdf/1901.02860.pdf>.
- [11] 张志华. 基于深度学习的情感词向量及文本情感分析的研究[D]. 上海: 华东师范大学, 2016.
- [12] ZHANG Y, WALLACE B C. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification [C] // International Joint Conference on Natural Language Processing. USA: MIT Press, 2017: 253-263.
- [13] 王正琦, 冯晓兵, 张驰. 基于两层分类器的恶意网页快速检测系统研究[J]. 网络与信息安全学报, 2017, 3(8): 48-64.
- [14] 袁梁, 林金芳. 基于文档分层表示的恶意网页快速检测方法[J]. 计算机系统应用, 2019, 28(12): 226-231.
- [15] 王红梅, 曹建. 网络安全攻击方式XSS初探[J]. 信息技术与信息化, 2019(11): 120-122.

作者简介

白广达, 男, 1995年生, 硕士研究生, 研究方向: 自出版、新媒体与数字出版发行, E-mail: baiguangda@whu.edu.cn.

翁子扬, 男, 1975年生, 硕士, 副教授, 硕士生导师, 研究方向: 新媒体与数字出版发行、艺术信息管理、史地空间信息。

Research on Content Management of Self-Publishing Platform in China

BAI GuangDa WENG ZiYang

(School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: The rapid development of China's self-publishing industry has made numerous new works available, and it has also brought the problem of prohibited content dissemination, which requires effective management of the content of the work from the self-publishing platform. The self-publishing platform is faced with challenges such as the large number of works, the deformation of violations of the text, and the difficulty of reviewing external links. This article puts forward three suggestions from the technical and management level, namely: perfecting the platform content management and control process, constructing multi-level strategy NLP capabilities, external chain content and external domain name detection to achieve external chain review.

Keywords: Self-Publishing; Self-Publishing Platform; Content Management

(收稿日期: 2020-03-18)