

基于在线评论的青年群体就业信息挖掘研究

——以豆瓣小组“上班这件事”为对象的分析

叶雨鑫, 程雨琦

(上海大学文化遗产与信息管理学院, 上海 200444)

摘要: 为更好地适应新形势下的就业特点,对青年就业群体的就业需求和就业观念进行了更深层次地剖析。获取豆瓣小组“上班这件事”中用户发帖信息,采用 LDA(latent Dirichlet allocation)主题模型、ROST CM6.0、SnowNLP 方法,进行主题识别与情感分析。结果发现,当前青年群体关注的就业相关主题可分为 35 类,情感倾向总体呈现为负面。最后结合研究结果,针对劳动市场信息建设、青年劳动群体就业指导、信息渠道建设等方面提出了建议。

关键词: 青年群体; 文本挖掘; 就业信息; 情感分析

中图分类号: F241 **文献标志码:** A **文章编号:** 1671-1807(2024)22-0062-06

就业是经济发展的基础,关乎民生之本。2023 年,国务院办公厅发布《关于优化调整稳就业政策措施全力促发展惠民生的通知》,指出要从扩大容量、青年就业、强化帮扶、加强组织等方面稳就业、保就业^[1],表明当前就业形势的新动向、新问题已经成为党和国家关注的重点领域。科技的发展改变了人们搜寻、获取、交流知识的方式^[2],有研究称,社交媒体逐渐开始成为最主要、应用最广泛的求职信息搜索渠道^[3],越来越多的求职者通过在线社交网络搜寻就业情报。豆瓣“上班这件事”小组是创立于 2006 年的一个职场类高热度小组,是一个职场青年交流心得、交换情报、搜集就业信息、宣泄情绪的万能小站,其内容话题包罗万象,并且采用匿名发言的形式,因此内容相较于传统的职场类 App 更加贴近就业群体的真实想法,能够准确反映用户对求职、就业问题的关注焦点、情感倾向以及信息需求。

由此,本研究选取了“上班这件事”豆瓣小组的在线评论数据为研究对象,通过爬取相关数据,采用文本挖掘及 LDA(latent Dirichlet allocation)主题模型进行数据分析,并利用 SnowNLP 情感分析方法,探析青年群体在就业问题方面的关注焦点,分析用户多样的就业信息需求以及情感态度,据此提出改进就业环境和政策的建议。

1 文献回顾

专业社交媒体平台的使用优化了就业流程,拓宽了求职者的社交网络范畴,正在迅速改变人们寻找新工作和雇主寻找新员工的方式^[4],不少学者也认识到虚拟社区对求职者的求职行为具有重要影响。Norder 等^[5]认为技术和社会网络在求职者的求职过程扮演了重要角色,因此关注社交媒体平台对研究求职行为来说是必要的。范哲^[6]通过调查发现,在受挫后求职情境下,求职者会更加重视社媒平台的搜寻效能,希望通过各渠道信息及时满足自身就业需求,此时对官方渠道和社交媒体的使用比例差距不大。章陈栋等^[7]认为虚拟求职社区是有效的信息交流平台,不仅可以帮助高校学生进行信息搜集和整合,高校就业指导也可通过虚拟求职社区互动间接帮助学生进行求职心理调整。

目前,以虚拟求职社区为研究对象的文献大多集中在智联招聘、前程无忧、应届生求职网等招聘网站上,通过采集岗位信息和招聘信息等数据来分析就业需求信息^[8]。有文献指出,职场类虚拟社区的在线评论是员工了解各大公司,获得更多就业和职场帮助的有效途径^[9]。近年来,也有一些学者开始基于在线评论来分析求职、择业与职场中的信息需求。吕宏玉和杨建林^[10]基于在线评论的就业行业情报挖掘方法,以就职员工对企业的评论为数据

收稿日期: 2024-06-18

作者简介: 叶雨鑫(2000—),女,河南郑州人,硕士研究生,研究方向为图书馆学;程雨琦(1997—),男,江苏南京人,硕士研究生,研究方向为科学学。

源,挖掘不同维度下员工对企业的情感态度,为求职者的就业决策提供了行业情报信息。王雪芬等^[11]基于应届生求职论坛的用户交互数据展开内容分析和社会网络分析,发现用户最关注的信息是笔面试相关的信息与资源,且不同求职主题的交互需求有较大差异。栗琛^[12]研究了牛客网讨论区中的用户数据,通过构建用户画像、分析用户关注主题,对虚拟求职社区进行资源的分类推送提出建议。唐春勇等^[13]采集了就业信息共享网站上的员工网络评论大数据进行词频分析,挖掘出员工在工作中最为关注的 25 个因素,以此为据设计调查问卷,发现员工需求主要聚焦于职业发展、组织氛围、工作强度和可雇佣性四个方面。

前人的研究为本文提供了方法和思路,特别是针对员工在线评论的研究,为本文提供了较为成熟的模型框架。但前人研究成果主要集中在就业信息需求方面,尚缺乏对就业信息内容本身及情感倾向的研究,因此,本文在以往研究的基础上,爬取豆瓣小组“上班这件事”中的根帖回帖数据,结合当前国内青年就业群体求职、就业的现状,利用 LDA 主题模型及基于 SnowNLP 的情感分析来探究青年劳动者关于职场与就业话题的谈论焦点和情感态度,并提出相关建议。

2 研究设计

2.1 数据获取与预处理

“上班这件事”小组成立于 2006 年,是豆瓣上相当活跃的职场讨论小组,包含大量与就业高度相关的在线评论信息。以“上班这件事”小组中的根帖与回帖文本作为样本数据^[14],使用八爪鱼数据采集器进行内容抓取,采集了最新回帖时间在 2023 年 1 月 1 日至 2024 年 2 月 20 日,至少有一条回帖的全部根帖数据共 79 678 条,同时爬取根帖下回帖信息共 1 815 682 条。为保证研究的科学性和数据质量,筛选 2023 全年的根帖数据,同时进行去重,清洗“已删贴”“已解决”等无效数据,同时去除部分内容量较小、无法分词的根帖信息,最终得到 41 097 条根帖数据和 965 177 条回帖数据作为有效文本。

数据规范化处理完毕后,对以上数据进行 LDA 主题聚类分析和基于 SnowNLP 的情感分析,同时借助 ROST CM6 软件,进行社会网络分析,以期挖掘出蕴含在豆瓣“上班这件事”小组中的就业信息。

2.2 LDA 主题内容分析

LDA 主题模型是文本挖掘和信息检索研究的重要方法之一,在科技情报、知识挖掘、知识发现等

领域都得到了广泛应用^[15]。LDA 可对海量异构文本数据进行主题聚类或文本分析,依据文本中词与词的共现概率挖掘隐含的主题信息^[16],从而大大降低文本表示的维度。为了细粒度描述不同主题的就业情报,参考邱均平和沈超^[17]的做法,将根贴数据进行二次 LDA 主题聚类操作,以优化主题结果。

在实施 LDA 之前需确定主题数量。确定最优主题数时通常需要在对文本主题数进行初步估计后结合困惑度、一致性以及主题间相似度等指标进行计算,采用困惑度方法确定最优主题数^[18],困惑度越低说明模型收敛效果越好,主题聚类越精确。经过一次和二次 LDA 主题聚类后的结果如表 1 所示。

通过对表 1 中的内容进行分析可以发现,在经过二次 LDA 主题聚类的全部根贴中,“职场个人情绪”“职场闲聊”“同事关系”等主题占据了较大的比例,这是由豆瓣小组具有的社交属性决定的。而“考勤制度”“福利待遇”“加班情况”等主题也是组内用户关注的重点主题。此外,还可以看到“离职补偿”“劳动仲裁”“离职流程”等主题,说明也有部分组内成员关注失业保障信息。在“就业选择”大类主题中,关于新媒体、互联网行业的讨论占比较多,说明该小组成员以学生、互联网相关行业的人员为主,较少工业、农业等传统行业的求职、就业人员的参与,亦可看出社交媒体平台主要用户群体分布。

表 1 一次和二次 LDA 主题聚类后的结果

一次 LDA 主题 聚类结果	二次 LDA 主题聚类结果
离职失业交流(4 925)	试用期裁员(647)、财务报销(252)、离职补偿(637)、劳动仲裁(375)、劳动合同(367)、上下级关系(953)、失业求职(748)、离职流程(949)
薪酬福利(4 484)	福利待遇(2 104)、考勤制度(2 380)
就业选择(7 027)	工作选择(821)、新媒体行业(464)、互联网运营(436)、跨行业经验交流(1 058)、面试经历(1 452)、公司信息交流(993)、文娱行业(539)、销售行业(344)、外企外贸(285)、岗位能力要求(635)
工作地点选择(2 726)	大城市就业(1 351)、返乡就业(1 375)
职场个人情绪(4 469)	职场个人情绪(4 469)
工作内容交流(5 746)	岗位工作情况(1 563)、交接与协作(1 807)、客户关系管理(1 443)、体制内工作(933)
考勤交流(3 251)	加班情况(1 896)、休假制度(1 355)
职场关系(4 899)	同事关系(2 060)、职场闲聊(2 839)
应届生去向(3 570)	升学与就业(1 019)、校招经验交流(886)、教育培训行业(467)、专业就业问题(1 198)

注:括号内为主题数量。

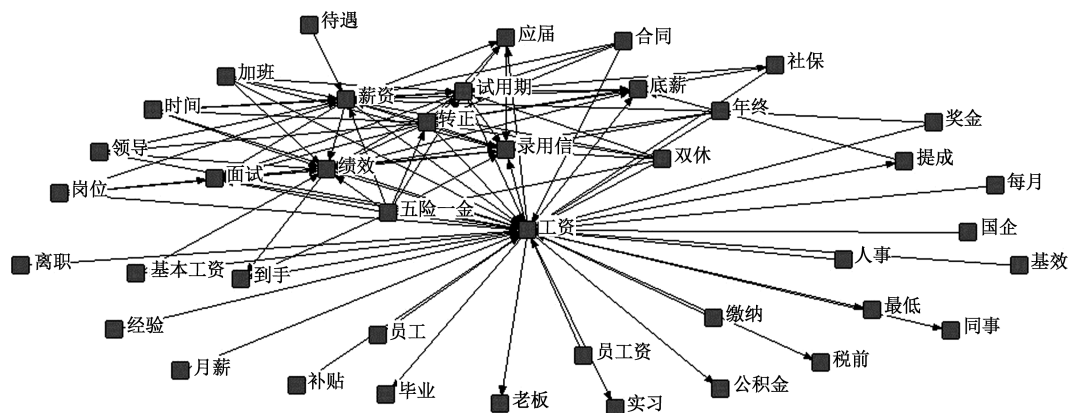


图 3 “福利待遇”ROST 语义网络

2.3 帖子主题情感分析

情感分析是自然语言处理(natural language processing, NLP)、数据挖掘和网络挖掘等领域的研究热点,其目标是从文本中分析出人们所表达的情感倾向以及观点,在舆论分析中有很大的研究意义。因此,在对帖子数据进行主题分析后,为进一步了解在线评论的情感倾向,采用 Python SnowNLP 库文件进行情感分析。SnowNLP 在分析前需要进行模型训练操作。选取数据爬取步骤中被清洗的非 2023 年的根贴数据进行人工标注,共标注 4753 条正、负极性数据,其中 493 条正面情感数据、4 060 条负面情感数据作为训练集,正面情感数据、负面情感数据各 100 条作为测试集,最后得到训练模型的准确率为 85%,说明模型可靠,可进行下一步分析。

如表 2 所示,从总体来看,各主题中根贴呈现负面情感的比例都比较高。一方面这是由于豆瓣小组的匿名性降低了人们表达负面情绪的门槛,用户在没有直接责任和后果的情况下可能更愿意分享他们的不满和抱怨;另一方面表明职场相关的话题容易引发强烈的个人情绪,尤其是在面对竞争、不公、工作与生活的平衡时,人们会产生更多的焦虑和挫折感,从而导致负面帖子数量的增多。

其中“劳动仲裁”主题较为特殊,该主题下的根贴情感体现出正面极性倾向,结合具体帖子内容可发现,这是因为该主题多是小组成员分享自己在面临劳动纠纷时进行劳动仲裁的经历,并且收到了一定的正面反馈。同时,主题的回帖情感则呈现负面,这表明大量的回帖成员对劳动仲裁问题仍具有一定困惑,无法从根贴中受到积极情绪的影响。

“大城市就业”“返乡就业”两个主题也呈现不同的态势,其中“大城市就业”主题下根贴呈现中性情感,“返乡就业”主题则呈现负面情感。此外,两个主题下的回帖信息虽然同样是负面情感,但“大城

表 2 “上班这件事”帖子细分主题情感统计

主题	回帖数量	根贴情感	回帖情感
岗位能力要求	9 387	0.71	0.37
劳动仲裁	7 464	0.63	0.39
工作选择	15 977	0.61	0.34
公司信息交流	20 591	0.60	0.33
面试经历	26 087	0.56	0.35
外企外贸	5 229	0.56	0.36
文娱行业	8 869	0.54	0.37
福利待遇	48 856	0.51	0.35
互联网运营	5 865	0.51	0.35
大城市就业	38 560	0.51	0.36
教育培训行业	10 536	0.49	0.33
专业就业问题	32 405	0.48	0.37
跨行业经验交流	25 588	0.48	0.34
新媒体行业	6 431	0.46	0.32
考勤制度	56 730	0.41	0.29
劳动合同	6 641	0.40	0.33
校招经验交流	14 548	0.40	0.32
升学与就业	31 093	0.39	0.33
离职补偿	12 154	0.39	0.33
上下级关系	24 216	0.35	0.30
试用期裁员	10 259	0.34	0.34
销售行业	5 126	0.32	0.30
返乡就业	54 686	0.31	0.30
财务报销	4 709	0.30	0.32
失业求职	12 908	0.28	0.28
岗位工作情况	26 141	0.25	0.26
加班情况	47 345	0.23	0.27
休假制度	28 782	0.22	0.26
职场闲聊	82 820	0.22	0.27
职场个人情绪	127 191	0.22	0.27
体制内工作	24 218	0.22	0.26
同事关系	53 168	0.21	0.27
客户关系管理	27 506	0.18	0.24
离职流程	15 540	0.18	0.26
交接与协作	37 551	0.15	0.23

市就业”的回帖情感中正向情感的比例总体高于“返乡就业”。进一步分析帖子内容,可以观察到在“大城市就业”主题下,“租房”“工资”“选择”等词语与主题关联紧密,表明用户更多地关注大城市的高

薪和职业发展,同时也对高昂的生活成本和激烈竞争感到担忧,这种复杂情感导致他们在小组中对大城市工作的讨论呈现期待又焦虑的态度。其次,与“返乡就业”关联紧密的词语包括“父母”“毕业”“生活”等,体现用户在选择返乡就业时多是考虑到家庭因素,也有不少帖子表达了自己在大城市遭遇挫折因此考虑返乡,因此更容易在谈论该主题时表达出失望和敏感等负面情绪。

“同事关系”“客户关系管理”“离职流程”“交接与协作”情感值较低。其中,“同事关系”主题内容反映了用户在职场中遇到的人际关系问题,包括沟通障碍、工作冲突等,这些问题易导致用户感到压力和不满。同样,通过具体评论内容也可以看出,用户在处理客户关系时会碰到如客户投诉、销售压力等困难和挑战,在离职过程中也会面对烦琐的流程和不公平的待遇,在交接工作和团队协作时遇到沟通不顺畅、交接不清等问题,这些都导致了用户感到焦虑和不满,从而在相关主题下表现出更多的负面情绪。

3 结论与建议

3.1 结论

本文通过采集豆瓣“上班这件事”小组上的根帖与回帖评论,运用 LDA 主题模型进行了主题聚类分析,通过 ROST CM6.0 绘制语义网络图展示各主题间词语的关系,同时使用 Python 中的 SnowNLP 库进行情感分析,分辨帖子内容表达的情感,情感分析的结果有助于识别用户对不同主题的情感态度以及回帖反馈情况。

首先,豆瓣“上班这件事”小组用户的讨论热点集中在 35 个主题,其中“职场个人情绪”主题数量占根帖总数量比重最大,同时也具有最高的回帖量,其情感表达比较负面;“考勤制度”“福利待遇”“加班情况”也有很高的讨论度,表明用户在进行工作选择时首要考虑的要素仍是工资待遇和福利情况,同时对试用期的具体规定和转正流程等制度信息有较高需求;“离职补偿”“离职流程”是用户进行离职失业交流时常讨论的内容,员工在面临离职时最常关注的是工资结算和失业赔偿,同时会涉及不少劳动纠纷和仲裁;“升学与就业”主题表明应届生对于就业具有很多困惑,需要大量指导。

其次,情感分析的结果表明用户对于职场话题的总体态度是负面的。具体来讲,“劳动仲裁”主题的回帖内容呈现正向情感,多是用户分享自己仲裁成功的经验,而回帖情感为负向,表明仍有许多组员缺乏劳动仲裁相关知识,在此方面还需要更多指导;用户在

讨论“大城市就业”和“返乡就业”时产生了明显的情感差异,在大城市就业的意愿更高,结合帖子内容发现虽然在大城市就业会产生较高的生活成本,但是相比之下返乡就业会面临工作机会少、就业形势差、工资待遇低等问题,因此用户对“返乡就业”表现出更多失望和敏感等负面情绪;用户在面临职场中人际关系时会产生较大的负面情绪,具体体现在同事日常相处、客户关系处理、团队间交接协作等方面,用户经常产生焦虑和不满的负面情绪。

3.2 建议

根据以上研究结果,不难看出当前青年群体关于就业信息的关注焦点和情感倾向呈现以下特征:第一,信息获取渠道不畅,对就业关键信息,如公司制度、劳动仲裁流程等,有强烈的信息需求;第二,在整体社会氛围影响下,对就业的焦虑感和无力感较为明显,对职业选择较为迷茫;第三,针对工作中遇到的实际问题,如人际交往、职场霸凌等,缺少必要的疏导和宣泄途径,总体情绪较为负面,需要建立更加完善的职业生涯规划体系和心理疏导途径。

据此,从就业信息服务的角度,可从以下几个方面入手,完善就业信息系统和内容建设,回应青年就业群体关切,缓解就业压力,为改善就业环境,提升就业质量做出贡献。

(1)建立健全劳动市场信息建设。相较于“大城市就业”“返乡就业”的话题情感并不是十分积极,说明三、四、五线城市的劳动力市场存在着相当大的信息不对称现象。一方面,用人单位存在着招工难等情况;另一方面则是求职者信息获取困难。对此,社会各方应当建立健全劳动市场的信息化建设,减少就业情报的获取难度,构建起求职者与用人单位之间的桥梁。而对于裁员、失业等人员,建设并完善再就业平台渠道,帮助再就业者快速获取信息和再就业机会。此外,“劳动仲裁”主题反映仍有不少员工在面对劳动仲裁时会因为缺乏足够的法律知识而感到无助,因此加强对劳动法等法律法规和劳动仲裁流程的宣传普及,帮助更多劳动者了解自己的权利和合法申诉途径。

(2)加强对青年劳动者的就业指导和价值宣教。对“应届生去向”相关的四个主题进行分析研究,发现当前青年劳动者对就业形势存在认知不足,同时也存在缺乏求职经验、对未来规划不清等问题,对此相关教育机构应强化对青年劳动者的就业指导服务,特别是要对应届毕业生给予更多关注和援助,帮助他们建立对自身能力和就业形势的客

观认识,并引导他们树立正确的就业观念,以促进他们的就业意愿和就业效率。此外,对于选择灵活就业的高校毕业生,国家和地方应完善相关法律法规,加强建设劳动者权益保护机制,以减少灵活就业者的就业风险,从而提高毕业生整体就业质量和稳定性。

(3)重视社交媒体及非正式沟通渠道建设。豆瓣小组凭借其自由的讨论氛围和匿名特性,吸引了大量用户前来宣泄情绪、分享经验、寻求帮助,其信息内容丰富但又良莠不齐,因此需要重视社交媒体及非正式沟通渠道的建设。研究过程中发现,目前职场类虚拟社区上仍存在着大量虚假、造谣信息,这一方面是由于部分用户为了引流,蓄意编造夸张内容吸引眼球;另一方面也是由于虚拟社区的监管制度不完善导致。因此,建议虚拟社区对用户实行公司认证、实名认证等措施,加强监管与审核力度,提高用户评论的真实性、可信度、和准确性。此外,社区管理者可通过在线问卷回访、线上奖励回馈等措施加强对用户的跟进,延续话题讨论,同时提高用户粘性。最后,社交媒体平台还可以进一步优化搜索功能,提供高级检索模式帮助用户更精准地查找信息。

参考文献

- [1] 国务院办公厅关于优化调整稳就业政策措施全力促发展惠民生的通知[EB/OL]. [2024-04-26]. https://www.gov.cn/zhengce/content/2023-04/26/content_5753299.htm
- [2] 吕美娇,李青青. 社会化问答用户知识交互行为影响因素研究[J]. 情报科学, 2022, 40(6): 141-148.
- [3] 范哲,张雨婷,孙晓宁. 高校学生求职搜寻中信息的采纳行为研究[J]. 图书情报工作, 2020(18): 105-113.
- [4] RUPAREL N, DHIR A, TANDON A, et al. The influence of online professional social media in human resource management: a systematic literature review[J]. Technology in Society, 2020, 63: 101-105.
- [5] NORDER K, EMICH K, KANA A, et al. A house divided: a multilevel bibliometric review of the job search literature 1973-2020[J]. Journal of Business Research, 2022, 151: 100-117.
- [6] 范哲. 不同求职情境下信息搜寻行为影响因素研究[J]. 现代情报, 2021, 41(11): 80-90.
- [7] 章陈栋,陆少琪,黎晓丹. 高校就业指导对求职心理调整的影响:虚拟求职社群互动的中介作用[J]. 教书育人(高教论坛), 2022(33): 49-52.
- [8] 王荣成,周镇忠. 我国劳动力市场研究中大数据应用现状与前景:基于范围综述的方法[J]. 社会建设, 2023, 10(6): 110-124.
- [9] 林文浩. 基于文本挖掘的新生代知识型员工离职成因研究[D]. 大连:大连海事大学, 2022.
- [10] 吕宏玉,杨建林. 基于在线评论的就业行业情报挖掘研究——以图书情报专业求职者为例[J]. 情报理论与实践, 2020, 43(9): 63-71.
- [11] 王雪芬,朱庆华,常李艳,等. 虚拟求职社区的用户交互行为特点——以应届生求职论坛为例[J]. 图书情报工作, 2018(10): 62-69.
- [12] 栗琛. 基于用户画像的虚拟求职社区推送服务研究[D]. 上海:上海师范大学, 2022.
- [13] 唐春勇,刘东冉,敬永春. 新环境下的员工需求维度构建——基于员工网络在线评论的大数据分析[J]. 管理现代化, 2018, 38(6): 106-110.
- [14] 李睿智,田佳玉,赵悦. 网络学术社区用户参与行为的类型及偏好研究[J]. 情报探索, 2022(7): 29-36.
- [15] 曾金,张耀峰,黄新杰,等. 面向用户评论的主题挖掘研究——以美团为例[J]. 情报科学, 2022, 40(11): 78-84.
- [16] 林轶,曹清芳. 基于 LDA 模型及情感分析的电影主题公园评论文本研究[J]. 科技和产业, 2023, 23(22): 93-98.
- [17] 邱均平,沈超. 基于 LDA 模型的国内大数据研究热点主题分析[J]. 现代情报, 2021, 41(9): 22-31.
- [18] 张东鑫,张敏. 图情领域 LDA 主题模型应用研究进展述评[J]. 图书情报知识, 2022, 39(6): 143-157.
- [19] 杨九龙,鲍慧璐. 用户在线评论省级公共图书馆内容挖掘与情感分析[J]. 图书馆学研究, 2021(21): 10-17.

Employment Information Mining of Young People Based on Online Comments: An Analysis of “the Matter of Going to Work” in Douban Group

YE Yuxin, CHENG Yuqi

(School of Cultural Heritage and Information Management, Shanghai University, Shanghai 200444, China)

Abstract: In order to better adapt to the employment characteristics under the new situation, the employment needs and concepts of the youth employment group were analyzed at a deeper level. User posting information from the Douban group on “Working” was obtained. LDA topic model, ROST CM6.0 and SnowNLP method were used for topic recognition and sentiment analysis. It is found that the employment related themes that the current youth group is concerned about can be divided into 35 categories, with an overall negative emotional tendency. Finally, based on the research results, targeted suggestions are put forward in areas such as labor market information construction, employment guidance for young labor groups and information channel construction.

Keywords: youth group; text mining; employment information; sentiment analysis