

基于LDA主题模型的职业教育分层分类人才培养模式分析

杜 刚, 严沛萌

北京信息职业技术学院实训中心, 北京

收稿日期: 2025年5月16日; 录用日期: 2025年6月20日; 发布日期: 2025年6月27日

摘 要

为深入探究职业教育分层分类人才培养模式的社会认知特征与舆论关注焦点, 本文基于新浪微博平台采集的相关文本数据, 构建高质量语料库, 运用LDA (Latent Dirichlet Allocation)主题模型对公众舆情进行挖掘与可视化分析。研究首先通过文本清洗、分词处理、特征提取等步骤对数据进行预处理, 随后结合困惑度分析与聚类可视化确定最优主题数为5, 并据此识别出“技能育人”、“职教改革”、“区域合作”、“数智赋能”、“技能强国”五大主题。结果显示, “技能育人”与“技能强国”主题占据主导地位, 体现出公众对职业教育人才质量提升与国家战略协同的高度关注; 同时, “职教改革”与“数智赋能”反映了体制创新与数字转型的实践趋势; “区域合作”关注度相对较低, 但在多元主体协同育人中仍具有重要战略价值。本文构建的主题关系网络图及主题画像体系, 系统呈现了职业教育舆情结构与语义特征, 为优化人才培养路径、提升政策回应性与精准性提供了理论支持与数据参考。

关键词

职业教育, 分层分类, 人才培养模式, LDA主题模型, 舆情分析

Analyzing Tiered and Classified Talent Development Models in Vocational Education Using LDA Topic Modeling

Gang Du, Peimeng Yan

Training Center, Beijing Information Technology College, Beijing

Received: May 16th, 2025; accepted: Jun. 20th, 2025; published: Jun. 27th, 2025

Abstract

To gain a deeper understanding of the social perceptions and public opinion focus regarding tiered and classified talent training models in vocational education, this study constructs a high-quality corpus based on textual data collected from the Sina Weibo platform. Utilizing the Latent Dirichlet Allocation (LDA) topic model, the study performs topic mining and visual analysis of public discourse. The data was preprocessed through text cleaning, word segmentation, and feature extraction, followed by perplexity analysis and cluster visualization to determine the optimal number of topics, which was set at five. The analysis identified five core themes: “Skills-Oriented Education”, “Vocational Education Reform”, “Regional Cooperation”, “Intelligent Empowerment” and “Strengthening the Nation through Skills”. The results show that the first and fifth topics dominate public attention, reflecting strong societal concern for improving vocational talent quality and aligning with national strategies. Meanwhile, the themes of reform and digital transformation highlight ongoing institutional innovation. Although “Regional Cooperation” attracted relatively less attention, it remains strategically significant in the context of multi-stakeholder collaborative education. The topic relationship network and thematic profiling developed in this study offer a systematic view of the structure and semantics of public discourse on vocational education, providing theoretical insights and data support for optimizing talent development strategies and enhancing the precision and responsiveness of related policies.

Keywords

Vocational Education, Tiered and Classified, Talent Development Model, Latent Dirichlet Allocation (LDA) Topic Modeling, Public Discourse Analysis

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

高等职业教育在国家教育体系中占据日益重要的地位，其人才培养模式正面临深刻改革。然而，目前职业教育人才培养仍存在诸多结构性问题。有研究指出，我国职业教育的生源结构复杂、起点不一，培养目标定位不够准确，教学过程尚未摆脱传统学科体系的窠臼[1]。这些问题导致人才培养模式尚不完善，难以保证培养质量，进而影响职业教育的可持续发展。为此，教育领域不断倡导通过分类培养、分层分流等举措来优化人才培养模式，以更好地适应学生多样化背景和行业多元化需求。

与此同时，社会公众对于职业教育改革的关注也日趋高涨，微博等社交媒体已成为公众参与讨论的重要平台。海量的微博言论汇聚了来自“不同行业、不同时段”的多元观点，成为真实反映民意的重要载体。但是，这些用户生成内容往往篇幅短小、碎片化，对同一话题的讨论分散在众多帖子和评论之中。传统的研究方法(如问卷调查或人工内容分析)难以及时、全面地梳理出其中蕴含的主要关注焦点。尽管如此，微博舆论中所体现的民间关注视角对于高职教育政策和人才培养机制的完善具有不可忽视的现实意义。相较于宏观层面的官方报告或专家研究，微博中的讨论更加直接地反映了学生、家长及社会各界对高职人才培养的期待和评价，这为政策制定者和教育工作者提供了宝贵的决策参考。

为了系统挖掘微博舆论所关注的“职业教育人才培养”主题并服务于政策优化，本研究在方法上引入了潜在狄利克雷分配模型(Latent Dirichlet Allocation, 以下简称 LDA)这一数据驱动的分析工具[2]。LDA

主题是一种无监督的概率主题模型, 可以从海量文本数据中自动识别潜在主题。相比早期的主题模型(如 PLSA、BTM), LDA 在主题分类任务中性能更优, 避免了过拟合等问题, 能够生成精细且有判别力的低维语义表示。此外 LDA 模型无需人工标注训练数据, 因而非常适合处理微博等短文本语料, 在一定程度上缓解了中文语境下的一词多义和语义稀疏等难题[3]。基于上述优势, 本文利用新浪微博平台相关文本数据, 运用 LDA 主题模型进行主题识别, 挖掘公众舆论中关于职业教育分层分类人才培养模式的核心关注主题, 从而为优化职业教育人才培养策略提供有力支撑和研究建议。

2. 研究设计

本文基于 LDA 主题模型构建职业教育人才培养模式的文本挖掘研究框架, 形成“数据采集层”、“文本预处理层”、“主题建模与可视化分析层”的三层结构研究路径。首先, 基于政策文件、研究文献、官方报告和新闻评论等多源文本数据, 构建语料数据库; 其次, 运用 Python 中的 jieba 分词工具对文本进行分词处理, 并构建自定义词典和停用词表, 以提升分词准确性; 再者, 借助 TF-IDF 方法提取关键词, 并基于词频构建词袋模型; 最后, 应用 LDA 主题模型对文本进行建模, 识别职业教育语境中的潜在主题结构, 通过关键词分布、文档主题概率、主题强度分析、pyLDAvis 可视化等方式, 对各主题的内涵进行解释与命名, 从而实现对职业教育人才培养模式内容结构的系统刻画。

2.1. 数据基础层

文本数据是研究的核心材料, 初始数据主要来自新浪微博。使用八爪鱼采集器抓取了与职业教育人才培养相关的微博数据, 包括标题、内容、发布时间、点赞数、评论数和转发数等, 由此构成本文的初始数据源。采集的文本内容涵盖了“专业设置”、“人才培养方案”、“产教融合模式”、“中高职贯通”、“分层分类教学”、“行业参与”等关键词, 具有较强的主题针对性。由于来源多样, 原始数据存在格式不统一、内容冗余、标签和脚本代码混入等问题, 本文在此基础上采用数据筛选、清洗、术语归一、停用词处理和数据整合等步骤对数据进行加工。经过上述处理, 最终构建了包含 2534 条行业报道与评论的高质量语料库, 为后续的分词、TF-IDF 提取及 LDA 主题建模奠定了坚实的数据基础。

2.2. 主题画像层

在数据基础层构建完成的基础上, 将通过提取主题特征词并构建主题关系图谱来描绘主题特征。本文通过 LDA 主题模型对职业教育分层分类人才培养相关的文本语料进行潜在主题挖掘与画像构建, 以揭示语料的核心话题结构及其分布特征。

2.2.1. 分词预处理与特征提取

采用 Python 中的 jieba 工具对清洗后的文本语料执行分词, 并结合以下步骤反复迭代优化: 首先, 进行初始分词, 剔除标点符号、数字及通用无意义词。接下来形成自定义词典, 即依据领域术语与政策关键词(如“现代学徒制”、“产教融合”等)补充词条, 以提升专业词汇识别率。然后, 进行停用词扩展, 在通用停用词基础上, 新增“全文”、“链接”、“收起”等无效抓取词, 去除爬虫噪声。最后, 构建词袋模型并计算 TF-IDF 权重, 选取高 TF-IDF 词汇作为 LDA 输入特征。TF-IDF 权重计算公式如下[4]:

$$W_{x,y} = tf_{x,y} \times \log \frac{N}{df_x}$$

其中, $tf_{x,y}$ 表示词项 x 在文档 y 中的出现频次, N 为语料库中文档总数, df_x 为包含词项 x 的文档数量。

2.2.2. LDA 模型构建与主题数确定

基于上述词袋表示, 本文采用 `gensim` 库构建 LDA 模型, 并通过困惑度和可视化聚类两种指标优化主题数 K 。困惑度用于比较不同 K 下的模型困惑度值, 识别困惑度拐点; 使用 `pyLDAvis` 生成主题距离映射, 评估主题间的分离度与重叠程度, 完成可视化聚类。综合困惑度曲线与聚类可视化结果, 本文最终将主题数确定为 5, 以兼顾主题语义凝聚力与差异性。

2.2.3. 主题画像构建与命名

在确定主题数后, 提取每个主题的高频特征词(Top-20), 并结合典型语料片段进行深度解读, 构建五个主题画像并予以命名, 分别是技能育人(学生、技能、培养、企业、实践)、职教改革(教育、建设、教材、体系、改革)、区域合作(合作、经济、地方、协议、共享)、数智赋能(数字化、人工智能、平台、创新)、技能强国(国家、战略、高校、人才、发展)。此外, 利用 LDA 模型输出的文档-主题分布概率, 计算各主题在整体语料中的权重占比, 并通过主题权重雷达图、主题分布柱状图及主题关系网络图等多维可视化手段, 全面呈现各主题的关注度差异与语义关联, 进而为职业教育分层分类人才培养的内涵构建提供数据支撑。

3. 职业教育人才培养主题画像分析

基于上述研究设计, 本章将进一步具体呈现职业教育分层分类人才培养模式的主题画像分析结果。首先对数据收集与预处理的实施过程进行简要说明, 随后依次呈现 LDA 模型的主题数量确定过程、主题特征的提取与命名结果, 并对各主题的分布特征及主题间的语义关系进行深入分析与可视化展示。

3.1. 数据收集与预处理

为聚焦职业教育分层分类人才培养模式的网络舆情特征, 本文以新浪微博平台作为主要数据源, 利用八爪鱼采集器收集与职业教育人才培养相关的微博数据。首先对关键词进行筛选。根据“职业教育”、“现代学徒制”、“产教融合”、“中高职贯通”、“分层分类”等核心术语构建检索词表, 在微博搜索接口中批量抓取发布时间自 2019 年至 2024 年间的相关微博。接下来确定数据字段。每条微博数据包括: 微博文本、发布时间、用户昵称、认证类型、点赞量、评论量、转发量及所属话题标签等。然后进行初步清洗, 去除广告、无关推广及明显机器人账号发布的内容; 清除 HTML 标签、URL 链接、特殊符号、表情及非汉字字符; 统一文本编码为 UTF-8。最后, 进行停用词与分词准备。构建职业教育领域专有术语停用词表, 剔除“全文”、“未完待续”、“转发”、“微博”等无实际意义的抓取噪声; 使用 `jieba` 分词工具对清洗后文本进行分词, 并结合自定义词典补充“校企协同”、“技能本位”、“双元制”、“工学结合”等专业术语; 迭代优化词典与停用词表, 直至分词结果准确、无漏分或错分现象。最终, 本文累计采集并清洗有效微博 2534 条。完成上述数据收集与预处理后, 下一步将在 3.2 节中对微博文本进行 LDA 主题建模分析, 并进一步提取和构建职业教育人才培养模式的主题画像。

3.2. 主题画像分析

对职业教育分层分类人才培养模式的主题画像进行深入剖析, 有助于精准把握公众在微博舆论中对高职教育各维度的关注重点, 为优化人才培养策略提供依据。本节首先采用困惑度与可视化聚类两种手段确定主题数, 然后提取每个主题的高频关键词并赋予命名, 最后通过主题聚类可视化分析图、主题权重雷达图及主题关系网络图等可视化图表呈现各主题在语料中的分布特征与相互关系。

3.2.1. LDA 主题数确定

基于已清洗并分词的微博语料, 使用 Python 中 `gensim` 库训练 LDA 模型, 并在主题数范围 $K = 3$ 至

10 内计算困惑度(Perplexity), 结果如图 1 所示。一般而言, 困惑度越低表示模型拟合效果越好; 但过低的困惑度可能导致主题过度细分, 应结合研究目标综合判断。从图 1 可见, 当 $K > 5$ 时困惑度下降幅度趋于平缓; 同时, 借助 pyLDAvis 生成的主题距离映射(图 2), 在 $K = 5$ 时各主题间既保有适度间隔, 又无显著重叠, 同时结合图 2 的主题聚类可视化分析结果, 当主题个数为 5 时, 各主题覆盖的内容区分较好, 故本文最终将主题数 K 确定为 5。

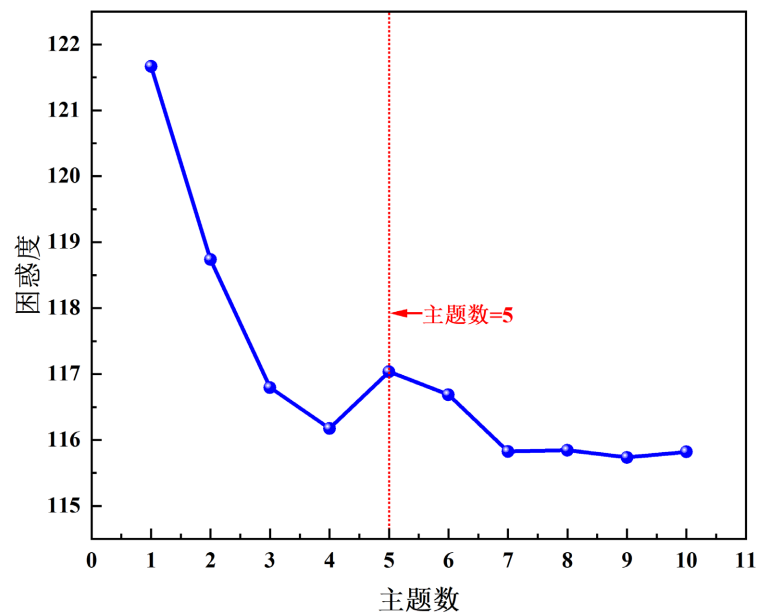


Figure 1. The theme confusion curve of the LDA model
图 1. LDA 模型的主题困惑度曲线

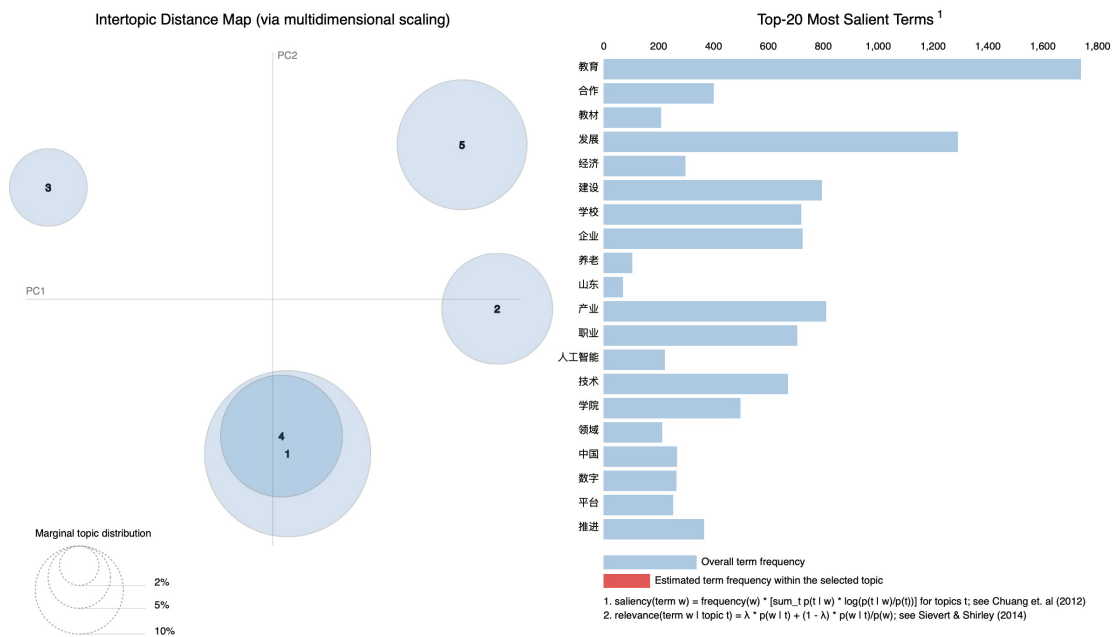


Figure 2. Visual analysis of topic clustering
图 2. 主题聚类可视化分析

3.2.2. 主题特征提取与命名

基于以上确定的 5 个主题，运用 Python 软件提取每个主题的前 20 个高频词汇，具体结果如表 1 所示。Topic 1 中的高频词汇“学生、发展、企业、职业、产业、专业、学校、技术、技能、培养、人才、学院、人才培养”等，都与技能育人涉及的主体密切相关，因此从 Topic 1 中提取的主题名称为技能育人。Topic 2 中的高频词汇“教育、建设、学校、发展、教材、专业、推进、提升、工作、职业、推动、产教融合、体系”等，则与职教改革关系密切，因此从 Topic 2 中提取的主题名称为职教改革。Topic 3 中的高频词汇“发展、合作、经济、教育、职业、技术、山东、学院、养老、产业、领域、低空、提供”等，都是描述与区域合作相关的词汇，因此从 Topic 3 中提取的主题名称为区域合作。Topic 4 中的高频词汇“产业、企业、教育、发展、人才、技术、创新、服务、人工智能、平台、数字、产教融合、专业”等，都与数智赋能直接相关，因此从 Topic 4 中提取的主题名称为数智赋能。Topic 5 中的高频词汇“教育、发展、建设、创新、专业、科技、人才、推进、中国、高校、推动、培养、国家”等，都体现出技能强国主题的内涵，因此将 Topic 5 命名为技能强国。

Table 1. Topic recognition result statistics
表 1. 主题识别结果统计

主题编号	主题名称	代表关键词示例
Topic 1	技能育人	学生、发展、企业、职业、产业、专业、学校、技术、技能、培养、人才、学院、人才培养、教育、就业、创新、需求、教学、实践、技能人才
Topic 2	职教改革	教育、建设、学校、发展、教材、专业、推进、提升、工作、职业、推动、产教融合、体系、学生、改革、中等职业学校、教学、教师、数字、支持
Topic 3	区域合作	发展、合作、经济、教育、职业、技术、山东、学院、养老、产业、领域、低空、提供、协议、服务、建设、人才、双方、中国、北京
Topic 4	数智赋能	产业、企业、教育、发展、人才、技术、创新、服务、人工智能、平台、数字、产教融合、专业、建设、推动、县域、国家、数字化、学院、体系
Topic 5	技能强国	教育、发展、建设、创新、专业、科技、人才、推进、中国、高校、推动、培养、国家、职业、劳动、大学、改革、学校、国际、合作

3.2.3. 主题分布与关系可视化

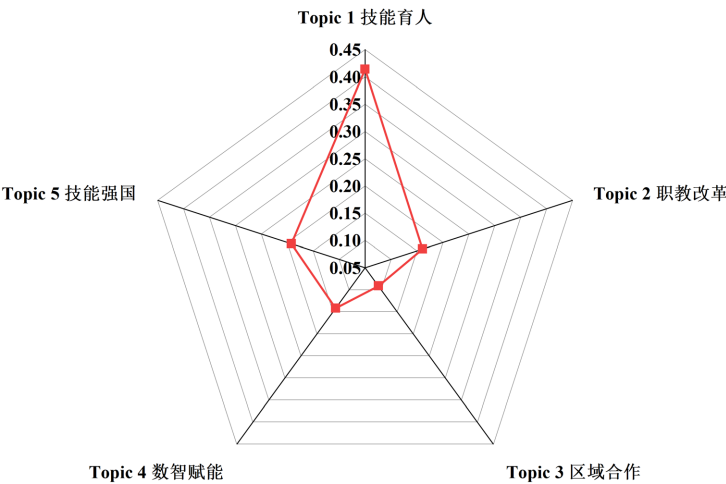


Figure 3. Theme weight radar chart
图 3. 主题权重雷达图

本文共采集了 2534 条微博数据。结合 LDA 模型对每条微博进行主题分类后, 可计算出每个主题对应的微博数量, 并据此得出各主题的权重系数, 绘制了如图 3 所示的主题权重雷达图。在职业教育分层分类人才培养相关的微博文本数据中, 关注度最高的主题是“技能育人”, 其权重达到了 0.4138, 其次是“技能强国”, 其权重为 0.1919, 这两个主题覆盖的微博数量占比达到了 60.57%, 超过微博数据总量的一半, 可知在分层分类人才培养模式舆论中, 公众对“技能育人”和“技能强国”的关注度较高。“职教改革”和“数智赋能”两个主题的权重分别为 0.1607 和 0.1420。二者反映了职业教育体系创新和数字化转型的需求, 因此也得到较高关注。“区域合作”的权重系数则小于 0.1, 表明公众对该主题内容的关注度相对较低。

图 4 为基于 LDA 主题模型构建的职业教育分层分类人才培养模式的主题关系网络图, 展示了各主题词之间的关联结构与聚类关系。从图 4 中可以看出, 图中的节点代表关键词, 节点大小表示该词的重要性或频率, 颜色表示不同的主题聚类。网络中存在多个明显的主题中心, 如“职教”、“企业”、“区域”、“合作”等, 每个中心分别聚集了对应语义场景下的一组关键词。例如, “职教”与“发展”“模式”“改革”等紧密相连, 反映了职业教育改革与发展模式的讨论; “企业”主题下聚集了“岗位”、“需求”、“平台”等词, 强调了企业在人才培养中的角色; “合作”主题则涉及“校企合作”、“区域发展”等, 说明跨区域协作和多元参与在人才培养中的重要性。整体来看, 该网络图直观呈现了职业教育人才培养在政策、区域、校企合作、技术应用等多个维度的主题分布与交互关系, 为深入理解分层分类培养机制提供了数据支撑。

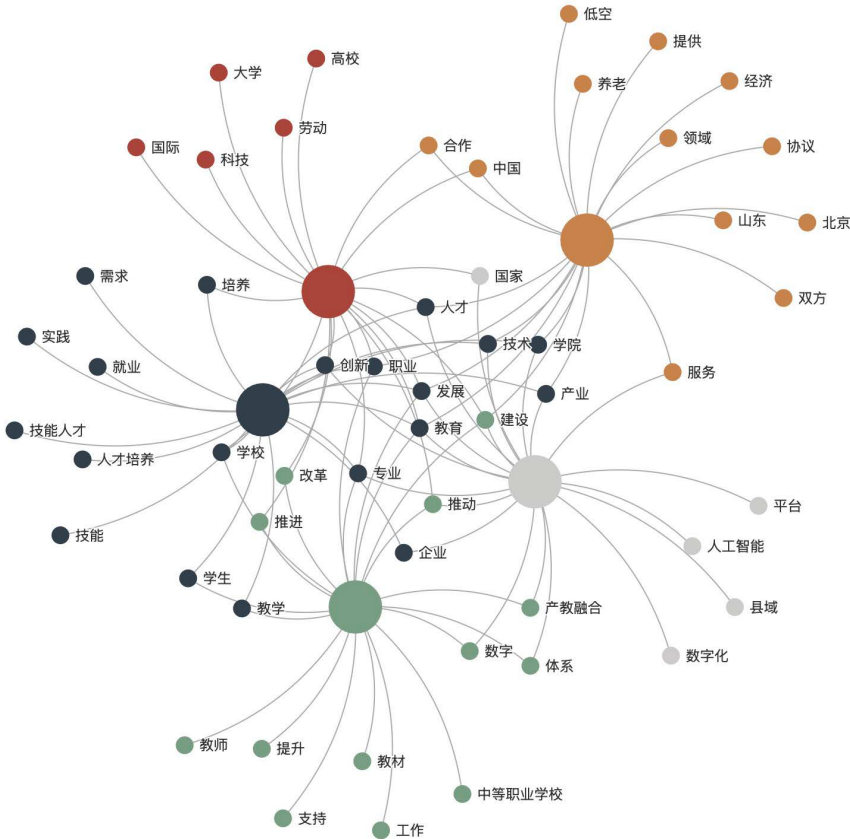


Figure 4. Theme relationship network diagram
图 4. 主题关系网络图

4. 研究建议

基于 LDA 主题模型对公众舆论的系统分析结果, 本文提出以下四方面的研究建议, 以进一步完善职业教育分层分类人才培养模式, 推动产教融合向更高层次发展。一是聚焦“技能育人”, 推动人才培养由“学历导向”向“能力本位”转型。公众对“技能育人”关注度最高, 说明职业教育应更加注重学生实践能力与岗位胜任力的提升。建议在教学设计中引入项目化教学、工作过程导向课程体系, 强化企业真实任务融入课堂实践, 构建符合岗位需求的能力标准体系。二是深化“数智赋能”, 加快数字技术与人才培养全过程融合。“数智赋能”成为重要话题, 彰显出社会各界对数字化教育手段的高度重视。建议进一步建设智能教学平台、虚拟仿真实训系统和数字资源库, 利用大数据、人工智能等技术开展学习分析、过程评价和个性化培养, 实现“数据驱动型”人才培养机制。三是强化“职教改革”政策落地, 推动制度创新与评价机制完善。职教改革主题显示出公众对政策执行成效和体制机制创新的持续关注。建议在政策制定与实施过程中, 加强对多元育人路径、学分互认制度、双师型队伍建设等方面的研究与试点, 完善评价反馈机制, 推动“分层分类”模式在实践中的有效运行。四是拓展“区域合作”广度与深度, 提升职业教育服务地方经济能力。尽管“区域合作”关注度相对较低, 但其在推动校企协同和产业对接中的战略意义不容忽视。建议鼓励职业院校牵头建设区域性产教联合体, 推动“行业-院校-地方政府”三方协同, 形成跨区域资源共享和人才共育机制, 提升服务区域经济转型和高质量发展的能力。五是强化“技能强国”战略支撑。将职业教育人才培养与国家战略紧密衔接, 确保职业教育的发展与国家产业转型和技能强国战略同步推进。

5. 结语

本文以职业教育分层分类人才培养模式为研究对象, 基于新浪微博平台获取的舆情文本数据, 运用 LDA 主题模型识别和提取公众关注的核心议题, 构建了五大主题画像, 包括“技能育人”、“职教改革”、“区域合作”、“数智赋能”和“技能强国”。研究发现, 公众对技能导向和国家战略层面的关注最为突出, 显示出社会各界对职业教育高质量发展的殷切期望。同时, 基于主题关系网络图和主题权重分析, 本文进一步揭示了不同话题之间的语义联结与聚类结构, 为理解分层分类人才培养的社会认知图谱提供了可视化依据。研究不仅从数据驱动视角深化了对职业教育分层分类人才培养机制的理解, 也为政策制定者、教育实践者提供了基于舆情反馈的优化路径。未来可进一步丰富语料来源, 并引入动态主题演化模型等方法展开深入研究, 推动理论研究与社会需求的深度对接, 助力我国职业教育体系更加科学、精准和可持续地发展。基于以上发现, 本文提出以下具有可操作性的建议: 推动“技能育人”转型落地、强化“数智赋能”基础设施建设、加快“职教改革”政策落地实施、完善“区域合作”产教协同机制、构建“技能强国”战略支撑体系。

基金项目

2022 年北京市职业教育教学改革项目“分层分类精准育人的人才培养模式研究与实践”(编号: IG2022001), 主持人: 杜刚。

参考文献

- [1] 梁克东, 王亚南. 基于“三教改革”的职业教育人才培养与评价改革创新路径[J]. 中国职业技术教育, 2019(28): 28-34+41.
- [2] 杜燕萍. 基于 LDA 主题建模的教师队伍建设改革政策文本分析[J]. 系统科学与数学, 2022, 42(6): 1411-1422.
- [3] 吴华君, 何聚厚, 陈其铁, 等. 面向职业教育在线精品课程评价的情感分析与主题挖掘[J]. 中国职业技术教育, 2022(2): 55-63.
- [4] 杨洋洋. 数据驱动下突发公共事件公众诉求主题画像与归因分析[J]. 情报理论与实践, 2024, 47(4): 126-133.