所学即所为？地方官员教育背景与产业发展——基于文本匹配与语义建模的实证研究

李宏扬

（浙江大学公共管理学院城市发展与管理系，杭州，310058）

摘要：干部的教育背景是其知识结构和政策倾向的重要基础。本文聚焦一个尚未被系统检验的问题：地方党政官员的教育专业是否会影响其任职期间推动的产业发展方向。为此，本文构建了一个覆盖 2005 至 2024 年间中国地级市主要官员教育背景的结构化数据库，并匹配了其任职时期对应的政府工作报告文本。我们提出了一套基于 Transformer 语言模型的文本匹配框架，首先通过 TF-IDF 模型筛选出文本中可能包含产业关键词的候选句，然后使用 BERT 模型进行语义相似度计算，提取出与中类产业相关的高相似度语句，从而实现句子级的产业标签赋予。在此基础上，本文构建了“产业关注向量”，量化每份报告中每类产业的关注程度。通过对比官员教育专业与其任职期间高关注产业之间的匹配程度，我们发现：官员的教育背景确实对产业政策表现出显著影响，特别是具有工科背景的官员更倾向于推动其专业相关的产业领域。本研究不仅丰富了地方治理中“人—政策”关系的理解，也展示了 NLP 技术在政治文本分析中的应用潜力。

关键词：地方官员；教育背景；产业政策；自然语言处理；文本匹配

What you learn is what you do? Local officials' educational background and industrial development

LI Hongyang

(Department of Urban Development and Management, School of Public Affairs, Zhejiang University, Hangzhou, 310058)

Abstract:

Keywords:

目录

1 引言

在中国地方政府治理实践中，推动区域经济发展、优化产业结构始终是政府工作的核心任务。每年发布的《政府工作报告》不仅反映了地方施政重心，也折射出主政者的政策偏好与能力风格。在诸多影响治理行为的因素中，官员的教育背景作为其知识结构与能力构成的重要基础，是否也在潜移默化中影响其产业决策方向？这一问题至今仍缺乏系统实证研究。

近年来，关于地方官员背景对治理绩效的研究逐渐增多，涵盖了出生地、成长经历、专业技术职称等多个维度。然而，“官员所学专业是否会影响其主政期间推动的产业领域”这一问题，在当前学界尚属空白。一方面，教育背景可能通过“专业路径依赖”或“知识惯性”影响官员的思维模式与政策关注点；另一方面，地方政府工作报告的政策用语、产业规划也提供了定量观察这种潜在关联的可能窗口。

本研究以“所学即所为”为出发点，尝试回答一个根本性问题：地方官员的教育专业是否会影响其任职期间的产业政策关注方向？具体而言，我们结合全国多个地级市的官员履历数据与历年政府工作报告文本，提出一种双轨方法：一方面对官员的教育背景进行专业分类与标准化处理，另一方面借助自然语言处理（NLP）技术对报告文本中的产业关注进行结构化建模。通过匹配两者之间的语义关联，建立产业匹配指数，进而量化“专业—产业”的契合程度。

在方法上，我们引入了基于 Transformer 的预训练语言模型（如 BERT）对政策文本进行语义匹配，构建“产业关注向量”，并结合 TF-IDF、句法依存分析、情感评分等技术指标多维刻画地方政府对特定产业的关注程度。此外，我们设计了一套基于教育部专业门类划分的分类系统，结合嵌入表示与语义相似度，较为精准地对官员教育背景进行类别划分。

本文的研究在理论上有助于深化我们对干部背景与政策行为之间关系的理解；在方法上，展示了 NLP 技术在社会科学研究中的具体应用场景；在实务层面，也为地方政府干部选拔、人才结构优化提供了一定的政策参考。

2 理论背景与文献综述

干部的教育背景作为其知识结构与政策偏好的重要组成部分，长期以来被视为影响治理风格与施政行为的重要变量。在组织社会学与政治心理学的框架下，教育被视为一种社会化过程，深刻塑造个体的认知框架、政策偏好与决策方式（March & Olsen, 1989）。在中国地方治理实践中，这一影响可能表现为：具有工科背景的官员更倾向于推动制造业升级，经济与金融出身的官员更关注招商引资和金融服务业，而农业院校背景的干部则可能更重视“三农”领域政策。

已有研究部分验证了上述机制。例如，Chen et al.（2019）发现，拥有技术职称或理工科背景的地方官员更可能在其主政区域推动科技类产业发展；Wu & Wang（2022）则指出，经济与金融背景的市长在其任期内更倾向于推动服务业占比提升。然而，这类研究大多以结构化数据进行经济绩效回归分析，缺乏对政策文本中“产业倾向”表达的直接捕捉，也未能将官员的教育背景与政策文本中的语义内容相联系。

与此同时，政府工作报告作为一种高度规范化与制度化的政策文本，是地方政府在特定年份内政策重点与价值取向的集中表达。近年来，越来越多的研究开始尝试对政府工作报告进行结构化处理，利用文本挖掘技术揭示地方政策变化趋势与议题偏好（Liu et al., 2021）。尽管已有研究使用关键词提取、词频分析或 LDA 主题模型来识别政策议题，但如何将这类文本与具体官员的知识背景建立映射，仍是尚待解决的挑战。

自然语言处理（NLP）技术的快速发展，特别是 BERT 等基于 Transformer 的预训练语言模型的出现，为社会科学提供了新的方法工具。BERT 模型能够有效捕捉句子间的语义相似性，弥补传统关键词方法无法处理“表达变异”与“同义扩展”的不足，已被广泛用于法律文本、政府文书、新闻舆情等语境下的智能分析任务（Devlin et al., 2018）。在中文环境下，各类预训练处理模型被证实在短句匹配、句子分类与实体识别任务中具有良好的语义泛化能力。

基于上述理论与方法背景，本文提出“所学即所为”的研究假设，尝试回答以下关键问题：地方官员的专业教育背景是否会影响其任内政府工作报告中产业政策的关注方向？本文在方法上融合 TF-IDF、BERT 语义相似度匹配、句法主语识别与情感分析等多项技术，构建“教育专业—产业语义匹配”关系，并进一步提出“产业关注向量”以刻画政策文本中对不同行业的关注强度与表达态度。

本研究的理论贡献在于拓展了地方干部研究的路径，将教育背景与政策输出在文本层面进行联结；方法上则展示了预训练语言模型在公共管理研究中的适用性，为未来的“人—政策”匹配研究提供了可复制的范式。

3 数据来源与变量构建

3.1 官员数据说明

本研究所使用的官员履历数据来源于多个**公开的人事信息数据库与地方政府官方网站**，涵盖了中国地级及以上城市党政主要负责人的基本信息。数据包含姓名、性别、出生年份、职务类别、任职时间、任职地区、学历层次、毕业院校、所学专业等字段，具有较强的结构化特征，适合进行标准化处理与统计建模。

为确保数据的广泛性与代表性，我们构建了一个跨地区、跨年份的动态官员样本库，具体包括：

- **时空范围**：覆盖全国绝大多数省份下辖的地级市，样本时间区间集中在 2005 年至 2024 年；
- **职位范围**：以市长与市委书记等主要“一把手”为核心，结合部分市政府常务副职；
- **学历信息**：数据重点记录官员所获得的最高学历，其专业字段大多为自然语言表达（如“飞行器设计”“国际贸易系”），需进一步归类处理；
- **任职年份**：通过对比简历信息与年度政府工作报告，实现精准的报告与官员任期对齐。

为避免姓名重复或模糊问题，本文在处理过程中还引入了地区 + 任职时间 + 职务的三元标识进行唯一匹配，最大程度提升了数据的可靠性与精度。

在初步清洗之后，我们保留了约 **7400 余条有效履历样本**，构成后续“教育背景—产业关注”匹配建模的基础数据。其中，专业字段经过统一预处理与分类归一，详见第 3.2 节“教育背景标准化分类”。

3.2 教育背景标准化分类

官员教育背景是构建“所学—所为”关系的核心变量。原始履历数据中的专业字段大多以自然语言表述，不同年代、院校与记录风格存在较大差异，如“国际贸易系”“工民建”“经济法方向研究生”等，直接使用将导致信息冗余与分类歧义。因此，本文对专业字段进行了系统的清洗、归一与标准化分类，主要包括以下步骤：

（1）文本预处理与专业清洗 首先对专业字段进行中文文本规范化处理，主要包括：

- 删除末尾修饰词，如“专业”“系”“班”等，仅保留核心专业名；
- 若文本中包含“系”字，则仅保留其后的专业内容（如“自动化系”→“自动化”）；
- 使用 `jieba` 中文分词工具对专业词组进行切分，统一格式；

- 去除常见停用词、空格与模糊词，如“研究方向”“继续教育”等。

（2）专业门类归类与映射 在预处理后，我们参考教育部《普通高等学校本科专业目录》（最新版）将专业归入 13 类一级学科门类，包括：哲学、经济学、法学、教育学、文学、历史学、理学、工学、农学、医学、管理学、艺术学、军事学。部分模糊或跨学科专业根据主干词义人工归类。

为了提高效率与一致性，本文还构建了一个包含近 200 条关键词的“专业-门类映射词典”，并采用**关键词规则匹配 + 语义相似度补充匹配（BERT 向量）**的双重策略，具体如下：

- 若某专业名与词典中关键词完全匹配，直接赋予对应门类标签；
- 若无法匹配，则使用 Sentence-BERT（`paraphrase-multilingual-MiniLM-L12-v2`）对专业名与每个门类关键词集合进行语义向量编码，计算余弦相似度；
- 若最大相似度超过设定阈值（如 0.5），则赋值该门类；否则标记为“其他”。

（3）分类质量验证与样例展示 经过处理后，约 92.4% 的专业字段被成功归入明确的一级学科门类。我们对其中 500 条样本进行人工验证，匹配准确率约为 94.6%，说明该方法在保持通用性的同时具有较高可靠性。

表 1展示了部分归类示例：

表 1: 专业字段分类示例

原始专业字段	清洗后专业	归类门类
国际经济与贸易系	国际经济与贸易	经济学类
飞行器设计与工程	飞行器设计与工程	工学类
中共党史研究生	党史	历史学类
公共卫生与预防医学	公共卫生与预防医学	医学类
农业经济管理	农业经济管理	管理学类
法律硕士	法律	法学类

3.3 产业匹配数据

为了刻画地方政府在不同时期的产业关注方向，本文选取各地市级政府公开发布的《政府工作报告》作为分析文本来源。该类报告通常由市长在年度人代会上发布，是对过去一年政府工作总结和未来工作部署的全面表述，具有高度的政策代表性、结构规范性和文本完整性。

研究样本覆盖若干地级市 2020 年至 2024 年间的工作报告文本，文本长度普遍在 10,000 字以上，段落结构清晰，包含经济发展、产业规划、生态治理、民生服务等多个板块。由于工作报

告原始文本以自然语言表达为主，存在用词灵活、主语不固定、产业描述不一致等问题，因此在进行产业提取之前需进行系统的文本预处理与结构化处理，主要步骤如下：

（1）文本预处理与句子切分 报告文本首先按自然段进行粗分段，再使用正则表达式与中文断句规则（包括标点“。!?;”等）进行句子级切分。随后，我们采用词性识别与依存句法工具（如`jieba.posseg`）清除虚词、保留名词动词等实词信息，并对每个句子标记其所在段落位置（如前言、主体、总结）以备后续位置权重建模使用。

（2）产业标签体系构建 为了进行标准化匹配，我们构建了一个包含超过 1800 条细分产业标签的产业关键词体系，参考《国民经济行业分类（GB/T 4754-2017）》并结合实际工作报告中常见用语进行了多级补充。该词表被组织为三层结构：大类（如“制造业”）、中类（如“电气设备制造业”）、小类/关键词（如“新能源电池制造”、“半导体芯片设计”）等，统一存储于`industry_keywords.json`文件中。

（3）多标签匹配策略设计 考虑到实际文本中可能同时涉及多个产业领域（例如同一句中既提及“半导体”也提及“软件服务”），本文采用一种**多标签匹配**机制：每个句子允许匹配最多 4 个中类产业，匹配结果基于语义相似度打分自动判定。

匹配流程如下：

- **候选召回**：使用 TF-IDF 模型将每个句子与全部产业关键词构成的“产业语料”进行相似度计算，筛选前 Top-K（如 500 条）句子作为候选；
- **语义匹配**：采用预训练语言模型 BERT（`paraphrase-multilingual-MiniLM-L12-v2`）分别编码句子与各中类名称，计算句子向量与中类向量之间的余弦相似度；
- **Top-N 过滤**：对每个句子保留语义相似度最高的前 4 个中类（如相似度 ≥ 0.25），并记录其所属的大类作为完整路径。

与传统的关键词“是否包含”方法不同，BERT 语义匹配机制能够识别同义表达、语序变体与跨短语语义延伸，有效提升匹配准确率和鲁棒性。

（4）匹配结果结构与样例展示 匹配结果被组织为“句子–中类–大类”三元组集合，其中每个句子可能被分配 1–4 个产业中类标签，适用于后续构建“产业关注向量”。以“湖州市 2024 年政府工作报告”为例，报告共分句 1421 条，平均每条匹配产业中类数量为 1.8 个，部分结果如下：

表 2 展示了典型的匹配样例：

表 2: 产业匹配结果示例（多标签匹配）

文本句子	**匹配中类**	**匹配大类**
推动新能源电池、半导体芯片与数字制造联动发展	电气设备制造业，电子元器件制造，软件服务业	制造业，信息服务业
加快布局现代农业科技园与绿色种养一体化试点	农业服务业，蔬菜种植业	农、林、牧、渔业
发展文旅融合产业，打造“夜经济”消费圈	旅游业，文化创意服务业	文化、体育和娱乐业

4 方法设计：文本匹配与关注度建模

4.1 匹配方法比较与选择

为实现对政府工作报告文本中各类产业关键词的准确识别与分类匹配，本文设计了一套“语义召回 + 语义匹配”的双阶段文本匹配流程，旨在兼顾召回率与语义准确性，并兼容政策语言中大量非标准化表达（如同义、隐喻、省略等现象）。

（1）候选召回：TF-IDF 语义初筛 首先将报告文本分句后的结果构建为一个句子级文档集合，使用 `TfidfVectorizer` 对句子集与产业关键词集合进行向量化编码。在此过程中，产业词表被拼接为一个“产业语料文档”，与所有句子构成的语料库一并向量化，进而计算余弦相似度。我们选取每篇报告中得分最高的 Top-K 句子（如 $K = 500$）作为候选句集合，显著缩小后续精匹配所需的计算量。

TF-IDF 召回的优点在于能够有效识别产业词频较高、或在文本中反复出现的重点语义区域，起到“热区定位”的作用。

（2）语义匹配：BERT 多标签精匹配 在候选句集合中，进一步使用多语种预训练语义模型 `paraphrase-multilingual-MiniLM-L12-v2`（Sentence-BERT）对每个句子与所有中类产业标签进行编码，并计算其向量之间的余弦相似度。考虑到一句话可能同时涉及多个产业，我们设置每个句子允许最多匹配 4 个产业中类标签，并保留其匹配得分。

具体做法如下：

- 构建所有中类产业标签的向量索引（向量预计算，避免重复编码）；
- 将每个句子一次性编码为句向量，使用向量广播计算其与所有中类向量的相似度；
- 对相似度结果进行排序，保留前 Top-4 得分的中类，且仅保留得分高于预设阈值（如 0.25）的匹配对；
- 返回三元组结果（句子，中类，大类）及其匹配得分。

（3）模型优化与加速策略 为提升匹配效率并支持大规模报告处理，本文采取了如下优化手段：

- **向量预计算**：所有中类产业名称在运行前即编码为固定向量，避免每次重复；
- **GPU 加速**：BERT 向量计算运行在 NVIDIA CUDA 环境下，匹配效率提升 10 倍以上；
- **Top-N 剪枝匹配**：每句只保留相似度最高的 N 个产业，显著减少冗余计算；
- **批量句子处理**：模型编码阶段采用批量 `encode(sentences)`，降低向量化开销；

- **双阈值策略**：召回阶段使用较低 TF-IDF 门槛保证覆盖率，精匹配阶段使用相似度门限保证精准性。

该流程实现了从原始文本到结构化产业标签的高效、稳健、多标签提取，为后续构建“产业关注向量”提供了可靠的语义基础。

4.2 BERT 向量匹配模型实现

4.3 产业关注向量构建

5 实证结果分析

5.1 描述性统计

5.2 回归分析与机制验证

6 结论

附录