doi:10.3969/j.issn.1000-7695.2025.9.004

基于 BERTopic 和长短期记忆网络（LSTM）模型的政策主题挖掘与预测研究

——以工业互联网政策为例

李 艳，辛云丽

（西安工程大学管理学院，陕西西安 710600）

摘要：通过对政策文本进行主题挖掘并预测其发展趋势，有助于明晰政策重点和趋势，为完善相关领域政策体系提供参考。首先从主题内容视角出发，引入新兴 BERTopic 模型挖掘潜在主题；其次增加时间维度，构建动态主题模型，从主题频率方面刻画主题演化趋势；再次，构建支持度指标，应用长短期记忆网络（LSTM）模型对政策热点进行定量预测，并与传统时间序列自回归移动平均模型（ARIMA）对比以验证模型拟合效果；最后以工业互联网领域 2016—2023 年发布的 1 304 篇政策为例进行实证检验。检验结果表明，工业互联网相关政策可细分为 15 个核心主题，聚焦于创新应用、网络体系、平台建设、安全保障、资金奖励五大方面，随着时间的推移，工业互联网由初期的摸索借鉴转为规模化应用，政策主题也渐趋丰富。未来，数字赋能标杆平台、新型工业化信息化、标识解析体系的贯通应用或将成为热点方向，主题热度较高。

关键词：BERTopic；政策文本；主题预测；长短期记忆网络；工业互联网

中图分类号：F420；T012；G301 **文献标志码：**A **文章编号：**1000-7695（2025）9-0031-11

Research on Policy Topic Mining and Prediction Based on BERTopic and Long Short-Term Memory（LSTM） Models: Taking Industrial Internet Policy as an Example

Li Yan, Xin Yunli

(School of Management, Xi'an Polytechnic University, Xi'an 710600, China)

Abstract: Through theme mining of policy texts and predicting their development trends, it helps to clarify the policy focus and trends and provides a reference for improving the policy system in related fields. Firstly, this paper introduces the emerging BERTopic model to mine potential themes from the perspective of theme content; Secondly, the authors increase the time dimension to dynamically portray the theme evolution trend in terms of theme frequency; Thirdly, the authors construct the support index, apply the Long Short-Term Memory(LSTM) model to quantitatively predict the policy hotspots, and compare it with the traditional time-series Autoregressive Integrat ed Moving Average Model(ARIMA) model to assess the model fitting effect.Finally, this paper takes the 1 304 policies in the industrial internet field released in 2016—2023 as an example of empirical test. The test results show that industrial internet-related policies can be subdivided into 15 core themes, focusing on five major aspects: innovation and application, network system, platform construction, safety and security, and financial incentives. With time, the industrial internet has shifted from the early stage of groping and borrowing to large-scale application, and the policy themes have become richer and richer.In the future, the application of digital empowerment benchmark platforms, the integration of new industrialization and informatization, and the interconnected application of the identification and resolution system may become hot directions with a high level of thematic popularity.

Key words: BERTopic; policy text; topic prediction; LSTM; industrial internet

收稿日期：2024-06-18，**修回日期：**2024-09-09

基金项目：陕西省软科学研究计划项目“加强建设科技强省政策评估研究”（2024ZC-WTXM-06）；陕西省社会科学基金项目“区块链协同下食品安全社会共治的概率共识逻辑研究”（2020F014）

项目来源：西安市软科学研究计划重点项目“西安市工业企业数字化转型新型能力评价与演进路径研究”（2023JH-RKXZD-0003）

0　引言

政策文本是指国家或地区的各级权力或行政机关以文件形式颁布的法律、法规、部门规章等[1]，在政策分析研究领域中占有重要地位。政策往往具有向导功能，其为社会发展和人类行为设定方向，使整个社会有序地朝着既定目标前进[2]。挖掘政策文本潜在主题，深层分析政策背后含义和趋势，进行科学有效地预测，对完善国家治理体系，优化政策制定有重要意义。

随着政府信息公开化以及互联网的蓬勃发展，对政策文本的定量分析成为当前政策文本的研究趋势[3]，已有学者在政策主题挖掘与演化方面展开了诸多研究，研究方法包括共词分析（如 Li 等[4]的研究）、网络分析（如 Wang 等[5]的研究）、内容分析（如 Gao 等[6]的研究）、文本挖掘（如 Su 等[7]的研究）等，已形成一定范式。然而，仍存在以下不足：一是主题聚类时，大多采用传统的隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）建模方法，忽略了词语之间的语义和上下文关系，导致主题表达的多样性不足；二是在研究内容上，重演化轻预测，大多数研究止步于主题演化，对主题预测的研究数量相对较少，主题预测似乎更多地被作为主题挖掘与演化研究的附属产物；三是已有政策主题预测研究中，以定性研究为主，存在主观性较强、准确性欠佳等问题。

基于此，本文将 BERTopic 模型和长短期记忆网络（LSTM）模型引入政策文本分析中，实现对政策主题的识别与预测，以期完善、改进政策文本分析方法，为相关研究提供一定的参考借鉴。首先，将 BERTopic 模型引入政策文本分析中，对政策文本数据进行主题建模，发现政策文本中的主题结构和潜在语义，弥补 LDA 建模方法忽略了词汇之间的语义和上下文关系、主题表达不丰富等不足。其次，构建动态主题模型深入探究各主题演化趋势，并得到主题强度时间序列，然后选取 LSTM 模型实现对主题热点的定量预测，与传统时间序列模型——差分自回归移动平均模型（ARIMA）模型对比，依据常见评估指标——平均绝对误差（MAE）和均方误差（MSE）分别对模型进行评估，以验证模型的拟合效果。最后以工业互联网政策为例，实证检验方法的可行性和有效性。

1　相关研究

1.1　政策文本分析

政策文本是政策研究和政府决策过程中的重要依据。随着计算机和政府信息公开化的发展，政策文本数据体量越来越大，政策文本研究主流范式也由最初的政策文献解读到语言学实证研究再到如今的量化研究[8]。

在研究方法上，当前学者主要采用政策计量法、内容分析法以及主题建模法等对政策文本进行量化分析，如刘天畅等[9]应用政策文本计量方法对我国适老化改造政策演化特征进行了研究；马晓飞等[10]采用内容分析法将政策文件内容编码，从政策工具和政策应用两个维度探索国家层面人工智能产业政策的发展；郭丕斌等[11]运用 R 语言主题模型分析我国光伏产业创新政策层级特征。

在研究内容上，政策文本分析主要思路如下：一是关注政策本身的特征与现状，如石磊等[12]从政策工具视角下探究中国科技人才的政策特征，赖莎等[13]则从法规体系、政策过程及政策工具 3 个维度分析医保基金监管政策的发展及现状；二是关注政策内容的演化对比，包括国家、省、市不同层级对比、国内外比较以及政策与文献等其他文本的内容对比等，如张涛等[14]基于文本相似度视角对我国中央及 22 个省级政府发布的大数据政策文本比较研究，关注国内中央和地方的政策对比，杨慧等[15]应用 LDA 模型综合对比分析我国与美国、欧盟的气候政策情况，关注国内外政策比较；三是关注政策效力评估，如王帮俊等[16]基于光伏产业政策文件，建立量化评估模型测算各年度政策效力，朱新超等[17]基于 PMC 指数模型对我国和美国的人工智能政策进行量化评价。

1.2　主题挖掘与演化

主题挖掘是指从文本数据中发现隐藏在其中的主题或话题的过程，用于帮助理解大量文本数据中的内容，发现其中的模式和关联[18]。主题挖掘的研究方法通常基于统计学和机器学习技术，包括文本聚类、共词分析、主题建模等方法。如武川等[19]选取文本聚类中的 K-means 聚类算法对前沿技术专利主题相似网络进行聚类，进而识别前沿技术领域。孙文婷等[20]基于共词分析技术，挖掘政策文本高频词并构建各阶段共词网络，进而对中药材产业政策展开分析。曹蓉等[21]将共词分析与主题建模相结合，采用社会网络分析方法（SNA）与 LDA 主题模型对我国慈善政策的合作网络与主题热点进行演化分析。相较于文本聚类和共词分析，主题建模能够更深层次地理解文本内容，而不仅仅是基于词语共现或表面相似性进行分析。在主题建模中，当前应用最为广泛的模型是 2003 年 Blei 等[22]提出的 LDA 主题模型。如在专利研究领域，滕飞等[23]将 Doc2vec 算法和 LDA 主题模型结合并基于专利数据进行核心技术识别；在跨学科研究领域，祁颖等[24]将 LDA 与 Word2vec 相结合对国内外人文社会科学领域跨学科

研究文献的主题进行识别分析；在政策文本研究领域，华斌等[25]应用LDA主题建模实现对中国高新技术产业政策的主题层级演化分析，Gan等[26]应用LDA主题建模实现对专利政策的主题分类。

主题演化是基于主题挖掘中识别的主题，分析主题随时间的变化[27]。出于实际研究需求，诸多学者（如Blei等[28]、Alsumait等[29]）将时间引入主题模型，提出了动态主题模型（DTM），在线版本的LDA模型OLDA（Online Latent Dirichlet Allocation）等动态LDA模型，以继续分析主题演化情况。如曹丽娜等[30]对一系列主题模型分析比较后，应用动态主题模型挖掘随时间变化的动态话题链；崔凯等[31]将LDA模型扩展到在线文本流中，建立并实现了在线LDA模型，能从互联网语料中刻画主题演化趋势。

但由于LDA主题模型是一种典型的词袋模型，其在挖掘主题时存在易忽略词间关系，无法结合语境输出文本主题等不足[32]。2022年Grootendorst[33]提出基于BERT预训练模型的主题模型BERTopic很好地解决了这一问题，该模型不仅能够更好地捕捉文本数据的语义信息，而且无需人为设置主题数量，最大限度地减少了参数选择过程的主观性，且BERTopic模型支持引入时间参数，实现动态主题模型，分析各主题随时间演化情况。虽然BERTopic模型比较新颖，但已在学术界得到广泛应用，如聂亚青等[34]基于BERTopic模型对健康信息学的3个核心领域分别进行主题挖掘与演化分析，对比分析健康信息学在不同学科中的研究异同；刘洋等[35]使用BERTopic模型，借助抖音平台的短视频发布数据挖掘短视频用户的内在行为需要。

1.3 主题趋势预测

主题趋势预测旨在预测主题未来可能的发展趋势，总体来看，大致分为定性研究和定量研究两种，相较于易受到个人主观意见和偏见影响的定性研究，定量研究则通过数据分析和统计模型来预测未来的发展趋势，该类方法更具科学性，也能够体现出主题随时间变化的趋势与惯性[36]。然而，当前主题定量预测主要应用于学科领域，如梁继文等[37]运用ARIMA模型和t-SNE算法、模糊C-均值算法实现对新生科学主题的预测，郝雯柯等[38]借助Neural Prophet模型和非参数检验方法实现对社会科学领域新兴主题的预测。而在政策文本主题预测中，仍以定性分析为主，如杜尚荣等[39]通过梳理分析“五育”并举政策的演进历程得到其内在发展规律进而总结延伸出“五育”并举的发展趋势；诸葛凯等[40]聚焦于数字经济政策，得到数字经济政策演进规律继而提出我国数字经济政策的未来动向。此类预测均是在梳理政策演化的结果上对政策未来发展动向或趋势进行总结。

目前主题预测定量研究中常用的时间序列模型包括计量经济学模型和基于神经网络的机器学习模型，计量经济学模型中的典型代表为差分自回归移动平均模型，如岳丽欣等[41]应用ARIMA模型对信息构建领域的主题时间序列进行预测。基于神经网络的机器学习模型包括BP神经网络、支持向量机（SVMs）、和LSTM等，如李静等[42]学者对比分析这3种典型机器学习算法，发现LSTM模型对热点主题未来发展趋势预测准确度最高；之后朱光等[43]应用LSTM模型对研究主题的研究热点进行预测，并以隐私领域为例检验方法的可行性。

1.4 研究述评

综上，现有研究在政策文本分析、主题挖掘与演化、主题趋势预测等领域已取得诸多成果，但仍存在以下问题：（1）政策文本主题模型选取方面，当前研究主要依赖于主流的LDA模型，然而，LDA模型存在一些明显的局限性，如易忽略词间关系，无法充分结合语境输出文本主题等不足，这导致了在主题提取和表示上的不精确性，影响了后续分析的准确性和可靠性。（2）现有政策文本研究主要集中于政策的制定、评估，侧重于对当前主题的识别和分析，而对后续政策发展走向的预测研究较少涉猎，难以满足对未来政策方向的深入理解和准确预测的需求。（3）主题定量预测在学科领域的应用相对广泛，但在政策文本分析中的应用仍然较少。目前，政策主题趋势的预测主要依赖于定性研究，存在主观性较强、结果的准确性有待商榷等不足。

基于此，本文提出基于BERTopic模型的政策主题挖掘与预测方法：（1）引入基于BERT预训练模型的BERTopic主题模型，在进行主题识别时将上下文语义考虑在内，结合语境保留较全信息。（2）在政策主题挖掘与演化的结果之上，进一步拓展主题研究框架，计算主题强度时间序列进行定量预测，将LSTM模型引入政策文本主题强度预测中，并与传统的计量经济学时间序列模型ARIMA模型对比，依据MAE，MSE指标综合选择误差最小模型对政策文本主题进行趋势预测分析。

2 研究设计

2.1 研究思路与总体框架

本研究总体框架见图1，主要包括三部分：数据获取与预处理、主题挖掘与演化以及主题强度时间序列预测。

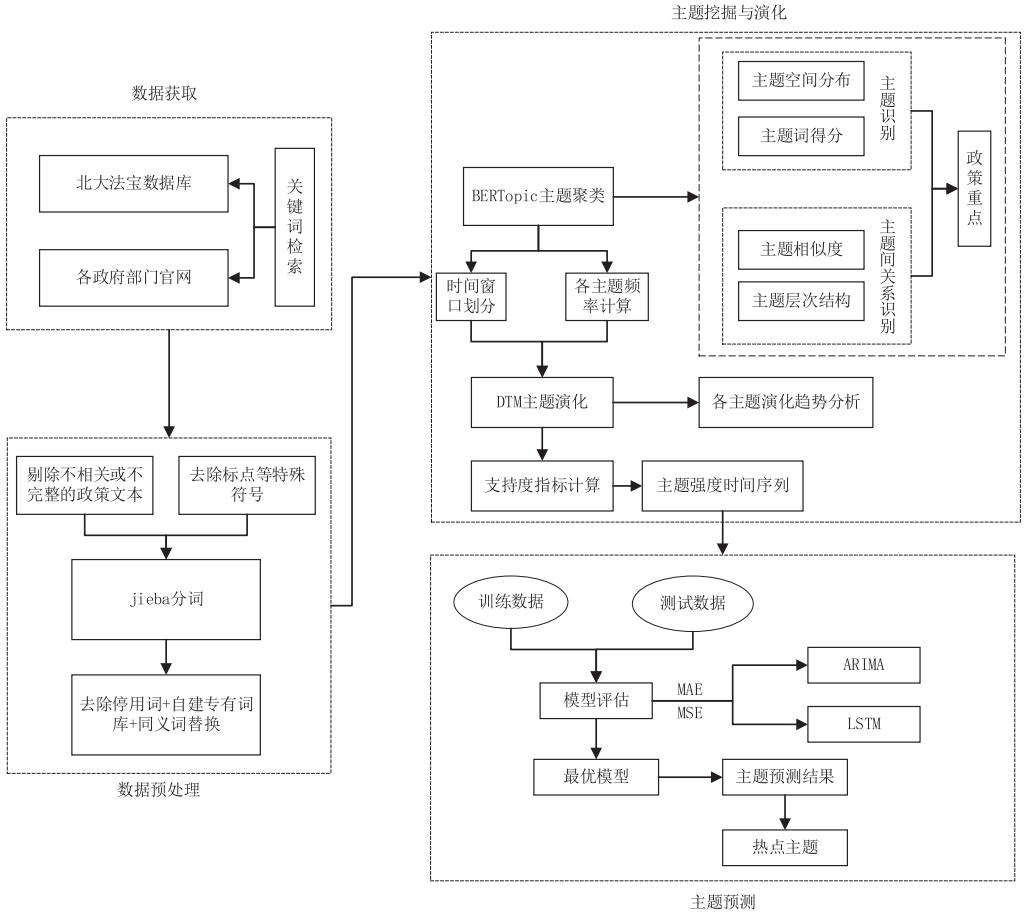


图 1　研究框架

2.2　数据获取与预处理

以北大法宝网（https://www.pkulaw.com/）及各政府官方网站为主要数据来源，按照标题关键词检索的形式进行相关政策文本收集，继而对收集到的文本进行数据预处理。首先对文本数据清洗，借助正则化表达式过滤文本中无用的 HTML 标签、数字、特殊符号、空格等；然后进行 jieba 分词操作，并综合哈工大、四川大学、百度停用词表去除停用词，再根据结果人工添加政策文本中常见的无意义的停用词，以确保数据的质量和可用性。

2.3　主题挖掘与演化

在主题挖掘上，本研究应用 BERTopic 主题模型完成政策文本主题聚类：依据主题空间分布，各主题的主题词得分对训练结果进行主题识别；依据主题相似度，主题层次结构对主题间关系识别；最后综合主题及主题间关系识别政策重点。

在主题演化上，本研究应用 DTM 动态主题模型得到各主题演化趋势：将政策文本以年为单位划分时间窗口，不同年份发布的政策文件作为 DTM 模型输入，计算得到各主题在不同时间内的主题频率，进而得到各主题频率随时间演化趋势。

2.3.1　BERTopic 主题模型

BERTopic 是一种基于预训练语言模型 BERT 的主题模型，属于主题聚类无监督深度学习模型的一种[33]。BERTopic 的原理见图 2，首先，利用 BERT 嵌入将文档转换为数字表示的向量；其次，使用 UMAP（Uniform Manifold Approximation and Projection）算法来降低文档嵌入的维度，并使用 HDBSCAN（Hierarchical Density-Based Spatial Clustering of Applications with Noise）算法创建语义相似文档的聚类；最后，采用基于类的 TF-IDT（c-TF-IDF）算法进行主题表示，即计算各语义簇（所得主题）内的词的重要性得分，挖掘出每个话题中的重要词汇，为每个文档簇生成主题 - 单词分布，同时在主题描述中保留重要单词[33]。

c-TF-IDF 算法中关于词语的重要性得分的计算方法见式（1）。

$$W_{t,c}=\mathrm{tf}_{t,c}\times\log\left(1+\frac{A}{\mathrm{tf}_t}\right) \tag{1}$$

式（1）中：$W_{t,c}$ 表示词语 t 在语义簇 c 中的重要性得分；$\mathrm{tf}_{t,c}$ 则表示语义簇 c 中词汇 t 出现的频率；tf_t 表示所有语义簇中词汇 t 出现的频率；A 表示每个

语义簇中出现词汇数的均值。

对于主题间关系，采用余弦相似度衡量主题间相似程度。余弦相似度（similarity）的计算方法如式（2）所示。

$$\text{similarity}=\cos(\theta)=\frac{\boldsymbol{A}\bullet\boldsymbol{B}}{\|\boldsymbol{A}\|\times\|\boldsymbol{B}\|}=\frac{\sum_{i=1}^{n}A_i\times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (2)$$

式（2）中：θ 表示两个向量之间的夹角，夹角越小，余弦相似度越大，主题间的相似程度越大；$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 分别表示两类主题的词频向量，A_i 和 B_i 分别表示两类主题所包含的主题词的向量特征。

与传统的 LDA、非负矩阵分解（NMF）、CTM（Correlated Topic Model）、Top2Vec 等主题模型相比，BERTopic 有以下显著优势：（1）使用 BERT 的上下文相关性来建模文本语义，较之词袋模型，能够更好地捕捉文本数据的语义信息，考虑词语之间的语义关系和上下文信息，提高了主题模型的表达能力和准确性，有助于发现文本语义层次的主题结构；（2）无需预先定义主题的数量，通过聚类方法自动确定主题数目，减少了参数选择的主观性；（3）克服了传统方法中基于密度聚类与基于中心采样之间不兼容的缺点[44]。

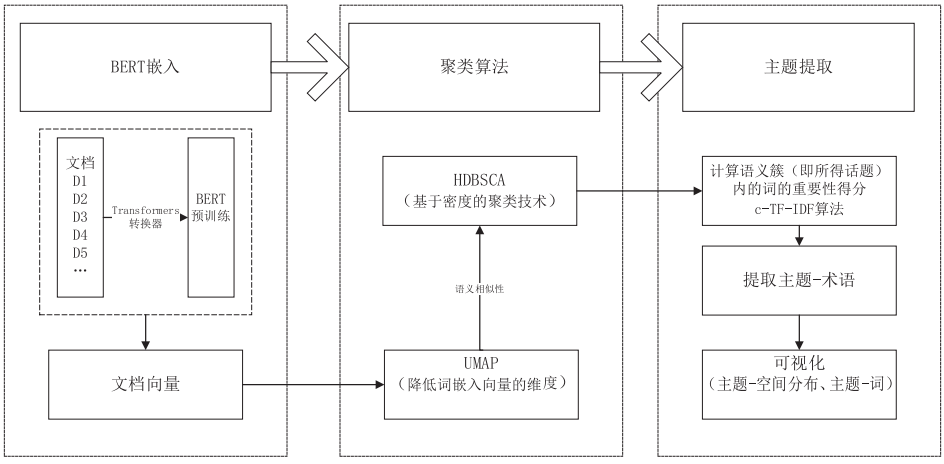


图 2 BERTopic 模型原理

2.3.2 动态主题模型（DTM）

DTM 旨在实现动态主题识别与追踪，分析主题随时间的演化过程[45]。BERTopic 通过计算每个时间片段内的主题表示来实现动态主题建模，在得到各主题后依据主题所属文档发布时间划分时间间隔，以主题内主题词作为代表，通过对相同主题的文档进行归类统计，计算各主题在不同时间段的主题频率并可视化为折线图，从而得以研究各个时间段内不同主题的演化过程[33]。

2.4 主题热度预测

主题热度预测是对主题未来发展趋势的一种预测，以主题热度来衡量和表示主题的关注程度以及主题的研究体量，并通过对主题热度数值的预测来预判未来前沿和趋势。

在指标选取上，本研究基于上述各主题概率时间序列，计算各主题支持度指标 SI（Strenth Index），作为衡量趋势指标。SI 是用于定量揭示主题强度特征的指标[43]，用于反映主题在某个时间段的受关注程度，表现为同一时间窗口下该主题所属政策文本的多少，计算方式如公式（3）所示。

$$\text{TSI}(k)_t=\frac{\text{DoC}(k)_t}{\text{DoC}_t} \quad (3)$$

式（3）中，$\text{TSI}(k)_t$ 表示主题 k 在 t 时间窗口下的支持度，分子 $\text{DoC}(k)_t$ 表示 t 时间窗口中该主题政策文本数，分母 DoC_t 表示 t 时间窗口中所有的政策文本数。TSI 值越大，表示主题强度越高。

（2）在方法选取上，本研究采取基于神经网络的机器学习模型 LSTM 进行预测训练，并与传统的时间序列模型 ARIMA 对比拟合效果。对于 LSTM 模型：1）构建数据集；2）将数据集分为训练集和测试集；3）LSTM 模型训练；4）LSTM 预测数据。对于 ARIMA 模型：1）Adfuller 单位根检验数据平稳性；2）白噪声检验；3）依据自相关函数 ACF（AutoCorrelation Function）、偏自相关函数 PACF（Partial AutoCorrelation Function）、贝叶斯信息准则 BIC（Bayesian Information Criterion）选取 ARIMA（p,d,q）参数大小（p 为自回归项数，d 为差分次数，q 为移动平均项数）；4）模型拟合效果检验（残差图检验、正态检验、自相关性检验）[41]。

最后选取常见评估指标，平均绝对误差（MAE）

和均方误差（MSE）对模型进行评估。各指标计算如式（4）—式（5）所示。

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{4}$$

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{5}$$

式（4）—式（5）：n 为样本数量；y_i 为第 i 个样本的真实值，$\hat{y}_i$ 为第 i 个样本的预测值。

3　案例分析

工业互联网作为新型工业化的战略性基础设施和发展新质生产力的重要驱动力量，为推进新型工业化提供了坚实支撑。截至 2024 年，“工业互联网”连续 7 年（2018—2024 年）被写入政府工作报告。现阶段我国工业互联网发展不断攀升新高度，相应的产业政策正在不断完善，并且开始聚焦于从整体兼顾向细节处转变，产业政策持续向好[46]，准确把握工业互联网发展热点及趋势，对政府、企业等各方面均有重要意义。故而，本研究以工业互联网政策为例，分析工业互联网领域热点主题及演化趋势，预测未来发展方向。

3.1　数据来源

在北大法宝网（https://www.pkulaw.com/）及各政府官方网站以“工业互联网”为关键词进行标题检索，得到2016年1月至2023年12月平台收录的“中央法规”与“地方法规”累计 1 304 篇政策，其中中央法规 62 份，地方法规 1 242 份，2016—2023 年每年的政策文本数量分布见图 3，可见工业互联网相关政策发布以地方工作性文件为主，且发布数量整体呈上升趋势。

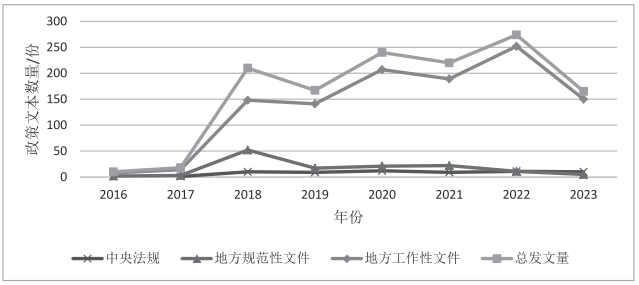


图 3　2016—2023 年工业互联网相关政策发文量年度分布

3.2　主题提取结果分析

（1）主题识别。将 1 304 篇政策文本数据预处理后输入 BERTopic 模型进行训练，共识别出 15 个主题，各主题的特征词分布如图 4 所示，可以得到，15 个主题分别是 Topic 0（工业企业发展）、Topic 1（优秀案例征集）、Topic 2（工业信息化）、Topic 3（标识解析体系）、Topic 4（项目专项资金）、Topic 5（工业互联网培训）、Topic 6（工业互联网平台）、Topic 7（跨行业、领域平台）、Topic 8（互联网安全大赛）、Topic 9（工业互联网大赛）、Topic 10（奖励支持）、Topic 11（创新项目发展）、Topic 12（网络安全）、Topic 13（资金管理办法）、Topic 14（领航案例设计）。

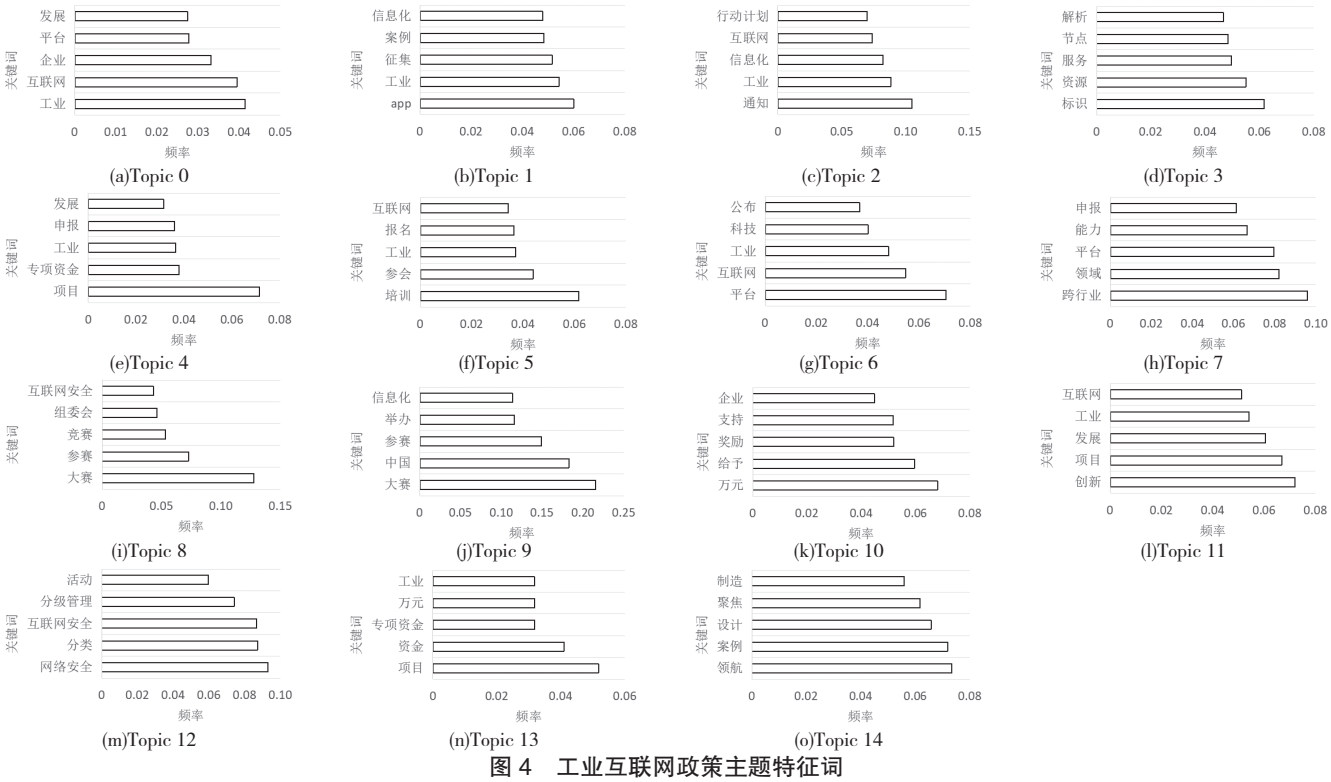


图 4　工业互联网政策主题特征词

（2）主题间关系分布。生成主题嵌入后可以创建主题相似性矩阵（见图 5），单元格颜色越深，表示单元格所在的横轴和纵轴的两个主题相似度越高。可以看到，部分主题之间相似度非常高，如 Topic 0（工业企业发展）和 Topic 2（工业信息化）语义相似度高达 0.992 305，可以推断，在讨论工业企业发展和工业信息化时，经常使用相似的词汇和语境，这是因为工业企业发展必然伴随着信息化的进程，通常需要依赖先进的技术和信息化手段来提高效率、优化生产流程和管理系统。部分主题之间则相似度相对较低，如 Topic 3（标识解析体系）和 Topic 9（工业互联网大赛），二者主题间语义差异较大，合理推测分属不同的政策重点。

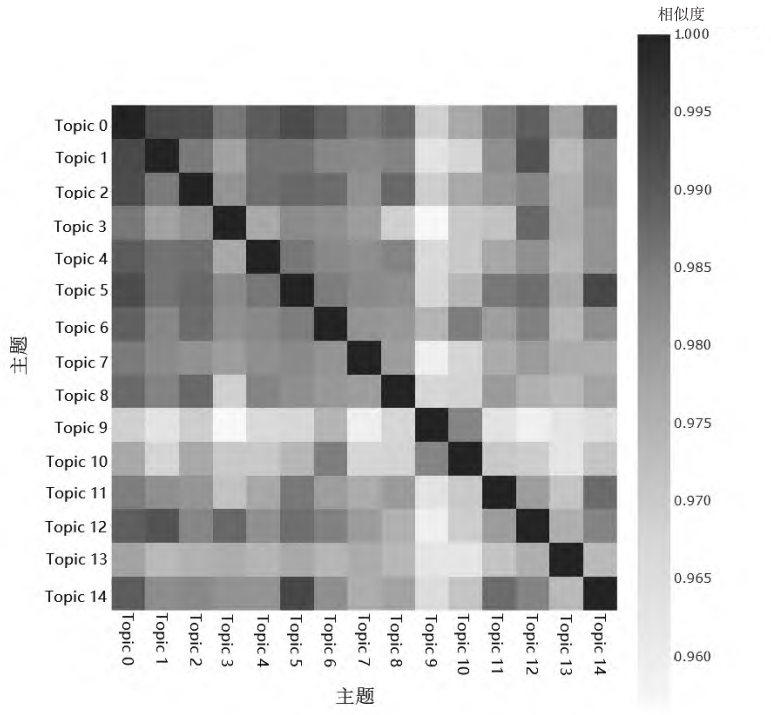


图 5　工业互联网政策主题相似度热力图

此外，为了更好地理解主题的潜在层次结构，将主题间的关系进行层次聚类，结果见图 6。图 6 中横轴表示层次聚类过程中的合并距离，距离越小，表示两个主题间的关系越密切。从中可以直观地了解各主题在不同层次上的关联，例如 Topic 8（互联网安全大赛）和 Topic 9（工业互联网大赛）有直接的紧密联系，二者均通过举办比赛推动工业互联网技术发展与应用；Topic 5（工业互联网培训）和 Topic 12（网络安全）有直接关系，表明在网络安全层面，大都通过培训促进其发展，如工业和信息化部网络安全管理局指导部人才交流中心建设了网络安全在线培训平台等；Topic 0（工业企业发展）和 Topic 11（创新项目发展）有直接的紧密联系，和 Topic 6（工业互联网平台）有间接关系，工业企业的发展直接依赖创新项目提供的技术突破与模式优化，而工业互联网平台可以作为支撑创新项目的基础设施，提供数据收集、分析、应用等功能，间接促进工业企业发展。

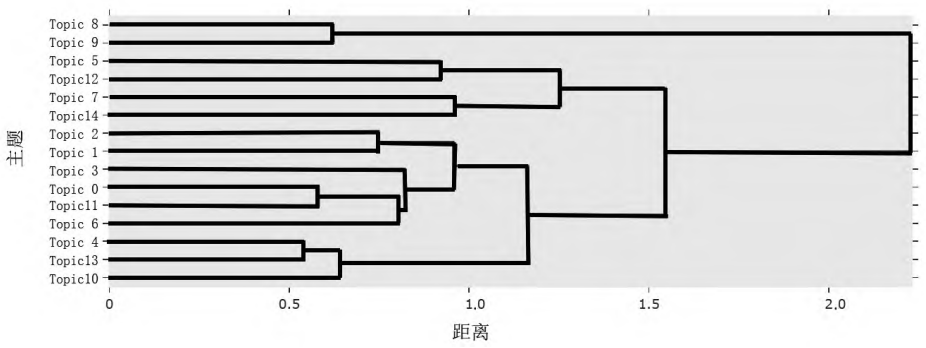


图 6　工业互联网政策主题层次结构图

综合考虑主题间相似度及主题间关系，结合工业互联网体系架构，归纳总结，可得到五大政策重点，见表1。

（1）创新应用：以促进工业企业发展为主（Topic 0），政府鼓励企业申报各类试点项目及创新项目（Topic 11），促进工业企业的信息化与工业化深度融合（Topic 2），激发企业的创新积极性；此外，通过举办评选活动，包括征集优秀APP案例、领航案例设计、工业互联网大赛（Topic 1，Topic 9，Topic 14）等吸引企业分享其在工业互联网领域的成功经验和创新成果，展示工业互联网在不同领域的应用价值和效果的同时，为其他企业提供借鉴和学习的范例。

（2）网络体系：作为工业互联网网络体系的重要组成部分，标识解析体系建设是政策领域重点关注（Topic 3），政府通过制定统一的标识解析体系标准和规范，确保不同厂商、不同系统之间的互联互通，有助于提高工业互联网网络的稳定性和可靠性，降低数据交换的成本和风险。

（3）平台建设：工业互联网平台的建设是促进工业互联网发展的关键环节（Topic 6），是工业互联网的中枢，政府尤为重视平台建设，尤其是跨行业、跨领域的平台（以下简称“双跨”平台）更是推动产业升级和创新发展的重要支撑（Topic 7）。工业互联网平台是技术创新和产业创新的重要载体，“双跨”平台则作为我国工业互联网平台最高发展水平的代表，政府鼓励和支持“双跨”平台建设，促进不同行业、不同领域间的数据共享和协同创新。

（4）安全保障：一是开展各项工业互联网安全培训、大赛等活动（Topic 5、Topic 8），涵盖了物联网安全、工业网络安全、数据安全等方面的内容，旨在提高从业人员的安全意识和技能，确保工业系统的安全稳定运行；二是重点倡导网络安全意识（Topic 12），通过教育和宣传活动，提高对网络安全的重视程度。

（5）资金奖励：在政府激励举措中，主要以资金支持为主，通过组织申报工业互联网高质量发展专项资金、发布工业互联网赋能奖补细则、提出项目配套奖励等引导促进工业互联网发展（Topic 4、Topic 10、Topic 13）。这些激励措施旨在激发企业的创新动力和发展热情，为工业互联网的研究、开发和应用提供强有力的支撑。通过建立多层次的资金引导机制，吸引更多的社会资本投入到工业互联网的创新和发展中，加速工业互联网技术的研究与应用，推动产业结构的优化升级，实现经济高质量发展。

表1　政策重点及主题分布

政策重点	主题
创新应用	Topic 0（工业企业发展）、Topic 1（优秀案例征集）、Topic 2（工业信息化）、Topic 9（工业互联网大赛）、Topic 11（创新项目发展）、Topic 14（领航案例设计）
网络体系	Topic 3（标识解析体系）
平台建设	Topic 6（工业互联网平台）、Topic 7（跨行业、领域平台）
安全保障	Topic 5（工业互联网培训）、Topic 8（互联网安全大赛）、Topic 12（网络安全）
资金奖励	Topic 4（专项资金申报）、Topic 10（奖励支持）、Topic 13（资金管理办法）

3.3　主题频率演化分析

为了更好地呈现主题演化趋势，在对工业互联网政策进行静态主题建模的基础上，引入时间维度，对政策主题进行动态演化分析，得到如图7所示的各主题的主题频率随时间演化的折线趋势图。

整体来看（见图7a），首先主题数量上，研究期前2年主题数较少，后期随着技术的进步和理论的成熟，主题逐渐丰富，Topic 0（工业企业发展）和Topic 5（工业互联网培训）是自2016年起政府最先关注的主题，2017年开始，Topic 1（优秀案例征集）、Topic 2（工业信息化）、Topic 3（标识解析体系）开始进入大众视野，其余主题则从2018年才开始陆续出现；其次，在各主题频率整体上，Topic 0（工业企业发展）和Topic 1（优秀案例征集）相对是政府更为关注的主题，主题频率相对一直处于较高位置，其余主题则大都处于较低的平稳状态。

在创新应用方面（见图7b），Topic 0（工业企业发展）是最早受关注的话题，亦始终是政府关注重点，2016—2023年都处于较高的主题频率，发展初期，势头较猛，于2018年左右达到峰值。后随着信息技术的发展，一方面，促进企业信息化和工业化深度融合（以下简称“两化融合”）成为工业互联网应用领域重点方向（Topic 2）；另一方面，鉴于工业互联网在我国处于发展初期，为提高公众对工业互联网的认知度和信任度，征集优秀案例的相关政策发布（Topic 1），以树立行业典范和领军企业，激励其他企业效仿学习。再随后举办工业互联网大赛（Topic 9），组织申报创新项目（Topic 11），遴选领航案例（Topic 14）等也发展起来，但相对处于稳定状态。

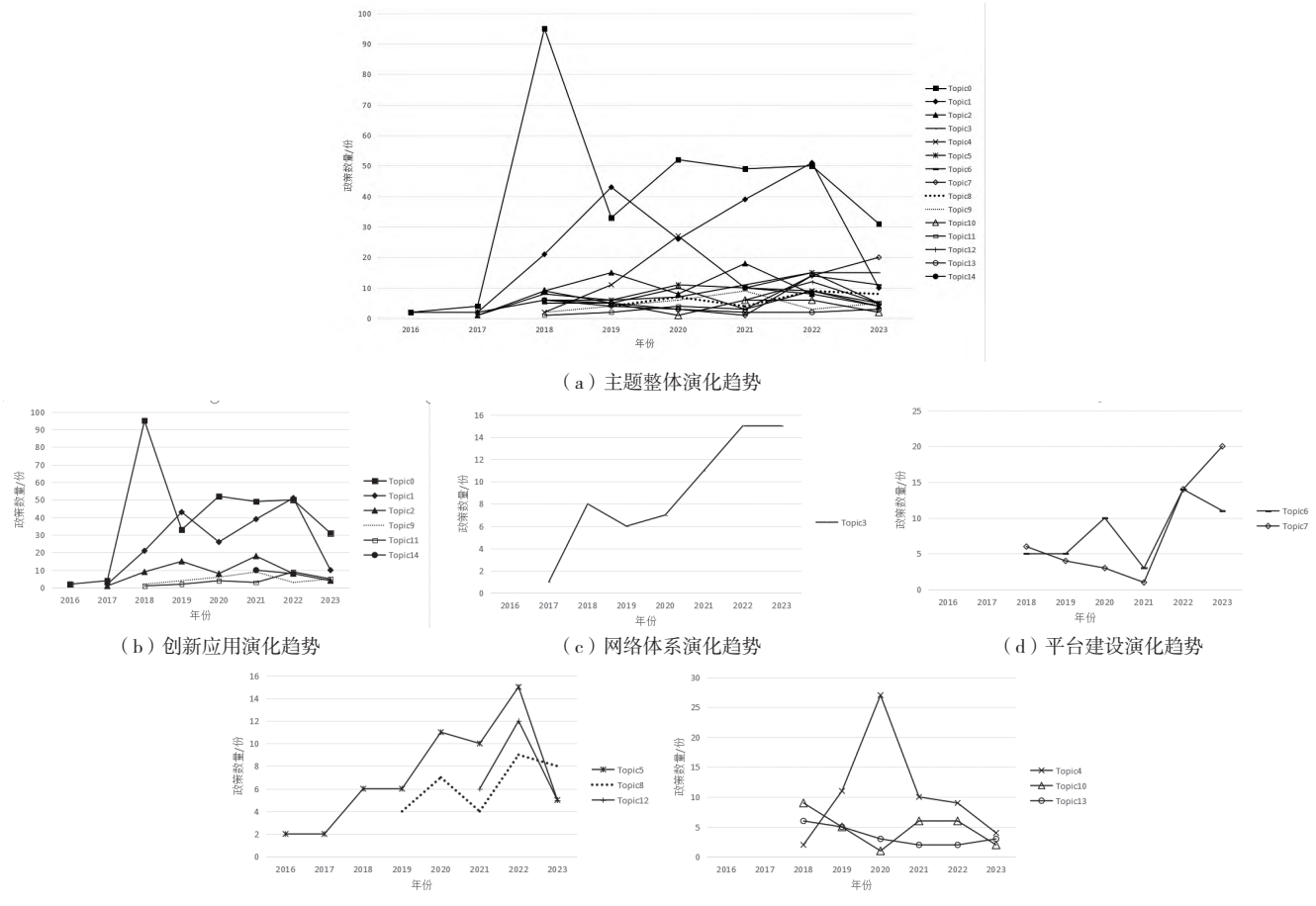
在网络体系方面（见图7c），Topic 3（标识解析体系）早受关注且整体来看主题热度呈持续上涨

趋势。工业互联网标识解析体系是工业互联网的重要组成和“神经系统”，其发展离不开政府的合理引导。自 2017 年以来，我国工业互联网标识解析体系从无到有、从小到大、“夯基架梁”工作基本完成，国家顶级节点稳定运行，接下来仍需继续推动标识解析体系由“建”到“用”，拓展在工业领域的应用广度和深度。

在平台建设方面（见图 7d），Topic 6（工业互联网平台）和 Topic 7（跨行业、跨领域平台）关注热度在 2022 年之前基本保持一致，2022 年之后，Topic 7 的热度则逐渐超过 Topic 6，可见，在平台建设中，政府关注重心逐渐向“双跨”平台倾斜。

在安全保障方面（见图 7e），Topic 5（工业互联网培训）为政府主要的安全引导机制，主题频率持续上升，政府积极组织工业互联网安全专题培训班，开展工业互联网安全专题培训活动，增强工业互联网安全防护能力。除培训外，2019 年左右 Topic 8（互联网安全大赛）也逐渐成为促进工业互联网安全发展的一种途径。相对于培训，大赛更紧密结合实际场景和工业互联网安全技术应用发展状况，政府积极支持并主办此类竞赛。2021 年，随着网络安全威胁的不断演变和加剧，网络安全（Topic 12）主题开始出现，主题频率呈现先升后降的趋势。

在资金奖励方面（见图 7f），Topic 13（专项资金管理办法）和 Topic 10（奖励支持）热度一直相对稳定，Topic 4（项目专项资金）于 2020 年达到峰值后开始回落，2021 年以来，3 个主题整体趋势一致，且频率分布无明显差异。



（a）主题整体演化趋势

（b）创新应用演化趋势　（c）网络体系演化趋势　（d）平台建设演化趋势

（e）安全保障演化趋势　（f）资金奖励演化趋势

图 7　2016—2023 年工业互联网相关政策主题演化

3.4　主题热度预测结果分析

基于上述各主题随时间演化的主题频率，依据支持度指标计算公式，得到各主题每年的主题热度时间序列。其中，Topic7、Topic8、Topic12、Topic14 时间序列过短，模型拟合效果不优，在此不做分析。采用 LSTM 模型和 ARIMA 模型分别对剩余主题进行模拟预测，选取 MAE 和 MSE 指标分别计算模型预测精度，拟合效果见表 2，综合来看，LSTM 模型拟合效果更优，误差相对更小。

表 2　ARIMA 模型和 LSTM 模型预测误差对比

主题	ARIMA 模型		LSTM 模型	
	平均绝对误差 MAE	均方误差 MSE	平均绝对误差 MAE	均方误差 MSE
Topic 0	0.160 379 19	0.025 807 47	0.097 510 01	0.015 630 10
Topic 1	0.366 652 94	0.234 028 15	0.078 750 00	0.008 914 83
Topic 2	0.219 167 43	0.093 451 92	0.028 730 73	0.001 225 76
Topic 3	0.136 207 61	0.022 822 93	0.038 004 47	0.004 196 20
Topic 4	0.372 172 10	0.150 266 32	0.019 367 34	0.000 646 16
Topic 5	0.196 112 22	0.061 642 64	0.026 564 93	0.000 907 84
Topic 6	0.285 338 30	0.142 028 26	0.020 094 75	0.000 594 51
Topic 9	0.310 235 35	0.106 423 07	0.011 194 73	0.000 170 44
Topic 10	0.227 031 35	0.061 397 33	0.010 405 53	0.001 648 08
Topic 11	0.147 599 88	0.027 172 67	0.011 782 76	0.000 280 23
Topic 13	0.090 391 54	0.009 924 17	0.012 563 87	0.000 242 23

本研究中，LSTM 模型采用 TensorFlow 框架和 Keras 模块，将每个主题前 80% 的数据作为训练集，后 20% 的数据作为测试集，预测各主题下一年主题热度，所得结果如下表 3 所示。

表 3　LSTM 模型预测结果

主题	2024 年预测值	主题	2024 年预测值
Topic 0	0.299 206 585	Topic 6	0.631 261 706
Topic 1	0.190 086 573	Topic 9	0.057 347 864
Topic 2	0.495 515 764	Topic 10	0.450 123 847
Topic 3	0.485 057 473	Topic 11	0.443 143 815
Topic 4	0.397 254 258	Topic 13	0.373 971 373
Topic 5	0.244 470 254		

按主题热度预测值排序，Topic 6（工业互联网平台）、Topic 2（工业信息化）、Topic 3（标识解析体系）是预测值排名前三位，热度较高，考虑是接下来的工作导向。结合 2024 年政府工作动态，可以得到以下结论。

首先在工业互联网平台建设方面，多省（如湖南省、广东省等）提出力争到 2025 年，引进培育多家工业互联网标杆平台或数字赋能标杆平台。平台建设依然是工业互联网领域的重中之重，随着数智时代的到来，各行各业数字化转型的不断深入，数字赋能平台是趋势亦将是主旋律，平台呈现“由广入深”的发展趋势。

其次，在工业和信息化方面，自党的二十大报告以来，推进新型工业化和信息化深度融合，建设现代化产业体系成为主流趋势。加之新质生产力成为高热度话题，广为探讨，如何推进新型工业化，培育新质生产力，向“新”而行，成为重点关注的话题，亦是各省份发展战略目标之一。

再之，在标识解析体系建设方面，一方面，多省（如广东省、湖北省、甘肃省等）发布工业互联网行动计划，提出培育发展多个工业互联网标识解析节点。另一方面，为进一步凝聚产业共识，推动标识解析体系由“建”到“用”，2024 年 1 月 21 日，工业和信息化部等 12 部门联合印发《工业互联网标识解析体系“贯通”行动计划（2024—2026 年）》，这是我国第一份针对工业互联网标识解析体系出台的政策文件，也是第一份工业互联网规模发展新阶段专项行动。打造自主可控的标识解析体系，加快推动标识解析规模化发展，实现标识的贯通应用是接下来的工作重点。

4　结论与展望

本研究将新兴 BERTopic 模型引入政策文本分析范畴，并构建了主题挖掘 – 主题演化 – 主题预测的系统框架，最后以工业互联网领域政策为例，检验了本文所提方法的可行性。

实证分析结果表明，工业互联网政策体系可细分为 15 个核心主题，重点围绕创新应用、网络体系、平台建设、安全保障、资金奖励 5 个方面展开，值得注意的是缺乏对数据要素的重视。随着时间的推移和工业互联网的发展，政府不断调整政策走向，政策主题日趋丰富，其演变轨迹反映了政策焦点的转移，其中 2016—2020 年，政策更多地聚焦于工业企业的发展，通过征集并推广优秀实践案例的形式为其他企业提供学习和借鉴的范例，带动工业企业整体发展；而 2021—2023 年，随着数字技术和信息通信技术的成熟，工业互联网由初期的摸索借鉴转为规模化应用；未来，数字赋能标杆平台、新型工业化信息化、标识解析体系的贯通应用成为热点方向，主题热度较高，广受关注。

本研究的创新之处在于，一是将深度学习 BERTopic 模型引入政策文本分析中，更好地实现对政策文本的语义理解和主题挖掘；二是拓展了政策文本量化分析范式，构建政策主题强度时间序列实现定量预测，为政治学、公共政策和信息技术领域等相关学者提供有关政策主题预测的新理论和实践视角，为政策研究和实践提供重要的参考和指导。

本研究也存在一定局限，如在预测指标选取上略显单一，仅从支持度指标这一维度构建时间序列模型，未引入其他可能影响预测结果的指标，未来

可继续引入其他指标，优化预测模型，实现更好的预测效果。

参考文献：

[1] CHILTON P, SCHÄFFNER C. Politics as text and talk: analytic approaches to political discourse [M]. Amsterdam: John Benjamins Publishing Company, 2002: 26.

[2] ZHANG X Q, YANG F. Rural informatization policy evolution in China: a bibliometric study [J]. Scientometrics, 2019, 120(1): 129–153.

[3] 郑新曼，董瑜. 政策文本量化研究的综述与展望 [J]. 现代情报，2021,41(2):168–177.

[4] LI X T, WU L, YU L N, et al. Policy analysis in the field of rare diseases in China: a combined study of content analysis and Bibliometrics analysis [J]. Frontiers in Medicine, 2023, 10: 1180550.1–1180550.14.

[5] WANG J X, ZHANG J J. The impact of policy topic networks on recombinant creation capabilities: empirical evidence from the energy field [J]. Journal of Cleaner Production, 2022, 380: 134858.1–134858.12.

[6] GAO X J, YANG S. Meteorological analysis of China's policy for women's entrepreneurship: themes, validity and prospects [J]. SN Social Sciences, 2024, 4(2): 45.1–45.20.

[7] SU J Y, XU K X, ZHOU F, et al. Analysis of the themes and evolution trends of urban renewal policies in Hangzhou:based on a text mining of the policy documents from 2002 to 2021 [C] //GUO H L, FANG D P, LU W S. Proceedings of the 26th international symposium on advancement of construction management and real estate. Singapore: Springer Nature Singapore, 2021: 1539–1550.

[8] 黄萃，任弢，张剑. 政策文献量化研究：公共政策研究的新方向 [J]. 公共管理学报，2015,12(2):129–137,158–159.

[9] 刘天畅，朱庆华，赵宇翔. 中国适老化改造政策的文本分析与演化特征研究 [J]. 情报科学，2024,42(1):84–93.

[10] 马晓飞，白雪松. 基于 34 份国家层面人工智能产业政策的文本量化研究 [J]. 北京邮电大学学报（社会科学版），2021, 23(5):19–30.

[11] 郭丕斌，施涛，吴青龙. 基于 R 语言主题模型的光伏产业创新政策层级性特征分析[J]. 科技进步与对策，2021,38(2):128–136.

[12] 石磊，熊嘉慧，李金雨，等. 政策工具视角下中国科技人才政策量化分析 [J]. 科技管理研究，2024,44(5):22–31.

[13] 赖莎，王冬，冯俊妃. 我国医疗保障基金监管政策现状及效能提升策略研究 [J]. 中国医院，2023,27(12):33–35.

[14] 张涛，马海群，易扬. 文本相似度视角下我国大数据政策比较研究 [J]. 图书情报工作，2020,64(12):26–37.

[15] 杨慧，杨建林. 融合 LDA 模型的政策文本量化分析：基于国际气候领域的实证 [J]. 现代情报，2016,36(5):71–81.

[16] 王帮俊，喻攀. 光伏产业政策效力和效果评估：基于中国 2010–2020 年政策文本的量化分析[J]. 软科学，2022,36(8):9–16.

[17] 朱新超，姜景，陈悦. 基于政策建模一致性 (PMC) 指数模型的中美人工智能政策比较研究 [J]. 科技管理研究，2024, 44(8):20–30.

[18] 张晨逸，孙建伶，丁轶群. 基于 MB–LDA 模型的微博主题挖掘 [J]. 计算机研究与发展，2011,48(10):1795–1802.

[19] 武川，王宏起，王珊珊. 前沿技术识别与预测方法研究：基于专利主题相似网络与技术进化法则 [J]. 中国科技论坛，2023(4):34–42.

[20] 孙文婷，汤少梁，段金廒，等. 基于共词分析的我国中药材产业政策演进路径探析 [J]. 中国卫生经济，2023,42(1):9–13,18.

[21] 曹蓉，刘彦芝，王铮. 中国慈善政策合作网络与主题热点演化研究：基于 SNA 和 LDA 的大数据分析 [J]. 社会保障研究，2023(1):41–52.

[22] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993–1022.

[23] 滕飞，张奇，曲建升，等. 基于专利竞争力指数和 Doc–LDA 主题模型的关键核心技术识别研究：以新能源汽车为例 [J]. 数据分析与知识发现，2024,8(11):33–46.

[24] 祁颖，张涛. 国内外人文社科领域跨学科研究：文献主题对比与中国路径选择 [J]. 情报科学，2023,41(12):81–90.

[25] 华斌，康月，范林昊. 中国高新技术产业政策层级性特征与演化研究：基于 1991—2020 年 6043 份政策文本的分析 [J]. 科学学与科学技术管理，2022,43(1):87–106.

[26] GAN J X, QI Y. Selection of the optimal number of topics for LDA topic model:taking patent policy analysis as an example [J]. Entropy, 2021, 23(10): 1301.1–1301.45.

[27] 胡吉明，陈果. 基于动态 LDA 主题模型的内容主题挖掘与演化 [J]. 图书情报工作，2014,58(2):138–142.

[28] BLEI D M, LAFFERTY J D. Dynamic topic models [C] //COHEN W, MOORE A. ICML '06: proceedings of the 23rd international conference on machine learning. New York: Association for Computing Machinery, 2006: 113–120.

[29] ALSUMAIT L, BARBARÁ D, DOMENICONI C. On–line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking [C] //GIANNOTTI F, GUNOPULOS D, TURINI F. Eighth IEEE international conference on data mining: ICDM 2008. Pisa: IEEE, 2008: 3–12.

[30] 曹丽娜，唐锡晋. 基于主题模型的 BBS 话题演化趋势分析 [J]. 管理科学学报，2014,17(11):109–121.

[31] 崔凯，周斌，贾焰，等. 一种基于 LDA 的在线主题演化挖掘模型 [J]. 计算机科学，2010,37(11):156–159,193.

[32] 滕广青，江瑶，庹锐. 基于多数据源维度的领域知识演化对比研究：以美国石墨烯领域研究为例 [J]. 情报资料工作，2023, 44(6):61–70.

[33] GROOTENDORST M. BERTopic: Neural topic modeling with a class–based TF–IDF procedure [EB/OL].(2022–03–11) [2024–09–09].https://arxiv.org/abs/2203.05794.

[34] 聂亚青，吴庭璋，王若佳，等. 基于 BERTopic 模型的健康信息学主题挖掘与发展演化研究 [J]. 情报科学，2024,42(4):98–110,118.

[35] 刘洋，柳卓心，金昊，等. 基于 BERTopic 模型的用户层次化需求及动机分析：以抖音平台为例 [J]. 情报杂志，2023, 42(12): 159–167.

[36] 霍朝光，霍帆帆，董克. 基于 LSTM 神经网络的学科主题热度预测模型 [J]. 图书情报知识，2021(2):25–34.

[37] 梁继文，杨建林，王伟. 知识单元重组视角下的科学主题预测研究 [J]. 情报学报，2023,42(5):511–524.

[38] 郝雯柯，杨建林. 基于语义表示和动态主题模型的社科领域新兴主题预测研究 [J]. 情报理论与实践，2023,46(2):184–193.

[39] 杜尚荣，舒清雅. “五育”并举演进的政策逻辑、制度选择与未来趋势：基于教育方针百年演进史的分析 [J]. 当代教育论坛，2023(3):82–92.

[40] 诸葛凯，何会涛，袁勇志. 我国数字经济政策演进与趋势分析：基于政策文献的量化考察 [J]. 经济体制改革，2024(1):24–32.

[41] 岳丽欣，周晓英，陈旖旎. 基于 ARIMA 模型的信息构建研究主题趋势预测研究 [J]. 图书情报知识，2019(5):54–63,72.

[42] 李静，徐路路. 基于机器学习算法的研究热点趋势预测模型对比与分析 :BP 神经网络、支持向量机与 LSTM 模型 [J]. 现代情报，2019,39(4):23–33.

[43] 朱光，刘蕾，李凤景. 基于 LDA 和 LSTM 模型的研究主题关联与预测研究：以隐私研究为例[J]. 现代情报，2020,40(8):38–50.

[44] EGGER R, YU J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts [J]. Frontiers in Sociology, 2022, 7: 886498.1–886498.16.

[45] LI D, YING D, XIN S, et al. Adding community and dynamic to topic models [J]. Journal of Informetrics, 2012, 6(2): 237–253.

[46] 王黎萤，李胜楠，王举铎. 我国工业互联网产业政策量化评价：基于 PMC 指数模型 [J]. 工业技术经济，2022,41(11):151–160.

作者简介：李艳（1982—），男，河北承德人，副教授，博士，主要研究方向为数据智能、数字社会、数据治理；辛云丽（2000—），通信作者，女，河南林州人，硕士研究生，主要研究方向为数据智能、科技政策。

（见习责任编辑：李洁）