doi:10.3772/j.issn.2095-915x.2025.03.003

基于局部线性自编码的政策文本挖掘方法研究

开放科学（资源服务）标识码（OSID）

魏超　赵伟　王弋波　吴欣雨

中国科学技术信息研究所 北京 100038

摘要：[目的 / 意义] 在政策文本数据量激增和结构日益复杂的背景下，针对政策文本复杂度高、领域性强、术语专业的特点，当前基于全局特征表示的方法在语义细节捕捉与跨领域泛化方面存在不足。[方法 / 过程] 从局部视角出发，提出一种基于局部线性自编码的政策文本挖掘方法，通过提取局部平滑的政策文本低维表示来改善政策聚类和分类效果。首先利用 BERT 表示的余弦距离确定局部政策文本近邻，然后通过联合最小化局部政策文本近邻的加权均方误差和重构误差来训练自动编码器。[局限] 主要针对中文政策数据，对于不同语言政策数据适用性存在局限。[结果 / 结论] 相较于对比方法，所提方法在谱聚类达到 85.78，提升了 10%，在 SVM 分类算法方面平均值提高到 86.88%，提升了 7.5%。结果表明，该方法可以为政策文本提供更具辨识力的低维表示，从而提升政策聚类、分类任务效果。

关键词：政策文本挖掘；自编码网络；局部线性；政策聚类和分类

中图分类号：G353；TP391

Policy Text Mining Method Based on Local Linear AutoEncoder

WEI Chao ZHAO Wei WANG Yibo WU Xinyu

Institute of Scientific and Technical Information of China, Beijing 100038, China

Abstract: [Objective/Significance] In the context of the surge in the amount of policy text data and the increasing complexity of its structure, for the characteristics of high complexity, strong domain and specialized terminology of policy text, the current methods based on global feature representation are deficient in semantic detail capture and cross-domain generalization. [Methods/Processes] From a local perspective, a policy text mining method based on local linear autoencoder is proposed to improve the policy clustering and classification effects by extracting local smooth low-dimensional representation of policy text. First, the cosine distance represented by BERT is used to determine the local policy text neighbors, and then the autoencoder is trained by jointly minimizing the weighted mean square error and reconstruction error of the local policy text neighbors. [Limitations] It is mainly aimed at Chinese policy data, and has limitations in applicability to policy data in different languages. [Results/

基金项目　中国科学技术信息研究所重点工作项目“面向战略研究的智能情报分析平台建设”（ZD2024-01）。

作者简介　魏超（1985-），博士，助理研究员，主要研究方向为深度学习、知识挖掘，E-mail：weich@istic.ac.cn；赵伟（1975-），博士，研究员，主要研究方向为科技资源管理，王弋波（1985-），硕士，副研究员，主要研究方向为科技资源管理；吴欣雨（1996-），硕士，助理研究员，主要研究方向为自然语言处理、知识组织与知识工程。

引用格式　魏超，赵伟，王弋波，等 . 基于局部线性自编码的政策文本挖掘方法研究 [J]. 情报工程，2025, 11(3): 27-40.

doi:10.3772/j.issn.2095-915x.2025.03.003

Conclusions] Compared with the comparison method, the proposed method achieves 85.78 in spectral clustering, an improvement of 10%, and the average value in SVM classification algorithm is improved to 86.88%, an improvement of 7.5%. The results show that the proposed method can provide a more discernible low-dimensional representation for policy texts, thereby improving the performance of policy clustering and classification tasks.
Keywords: Policy Text Mining; Autoencoder Networks; Local Linearity; Policy Clustering and Classification

引言

政策文本作为政府实施国家治理和公共事务管理过程中形成的规范性文献载体，不仅是政府行为的重要记录，更是开展政策内容解析与治理效能评估的基础性资料[1]。传统政策文本分析方法主要依赖分析人员的学识储备和专业经验来深入解读政策文本的核心思想和观点、演化规律、发展趋势等[2]，这种以专家知识为主导的解读模式存在主观性较强、人力成本高昂等局限。随着信息技术的快速发展和政府信息公开工作的不断深化，大数据驱动下多学科深度融合的研究范式不断发展，越来越多的学者利用文献计量学、文本挖掘等技术方法开展政策文本量化研究，由于其在处理大量政策文本数据时表现出更高的分析效率，因而成为当前政策文本分析研究主流[3]。

政策文本量化通过将政策文本中有价值信息转换成可量化的结构化数据，进而借助计算机实现政策文本内容分析[4]，按照数据转化方式，当前研究主要包括政策计量、内容分析和政策文本挖掘。其中，政策计量以文献计量学为核心，通过政策文本的结构化共现特征（如标题、关键词、机构、时间等[5]）来揭示宏观规律（如政策变迁与府际关系等[6]），内容分析基于政治话语特征，通过对政策概念体系的人工编码（如 PMC 指数模型[7]、政策工具[8]等）来解析政策语义。政策文本挖掘依托文本挖掘技术，通过向量化表示技术从大规模政策文本集合中发现隐含知识（如政策聚类分类[9]、政策关联分析[10]等）。综合来看，政策计量分析停留在浅层语义，对深层隐含语义的捕捉能力有限，内容分析由于依赖人工进行特征编码，使其人力成本随样本量增长呈指数上升，而政策文本挖掘则在政策语义的分析深度和大规模政策文本数据的分析效率之间取得了较好的平衡。

当前研究构建政策文本的特征表示主要采用主题建模和深度神经网络两种技术方法，其中主题模型（如 LDA、NMF 等）遵循词袋假设，忽略了词序和上下文语义联系，导致对语义细节的捕捉存在不足[11]。针对这一问题，一种改进策略是利用词向量（如 Word2Vec）生成的稠密词向量作为输入，然后通过基于序列模型（如 RNN、LSTM）和注意力机制（Transformer）的深度神经网络对词序和上下文语义进行建模，使文本表示向量在捕捉语义细节上取得优势。然而，深度神经网络的训练通常依赖高质量的训练数据，参数调优成本较高，尽管当前政策文本数据规模不断增加，但政策具有较强的领

域相关特性，表现为不同政策领域的主题分布存在异质性，某些领域高质量训练数据不足，这就导致模型存在泛化能力不足，分析精度退化的问题[12]。已有研究表明文本数据集中分布在一个局部线性的低维流形周围[13]，若能将该内在几何结构引入模型训练，则有助于提升模型的局部平滑性，从而增强模型在低资源环境下的鲁棒性与泛化能力[14-15]。

针对上述问题，本文以自编码网络（Auto-Encoder，AE）为基础架构，从局部线性的流形视角出发，提出局部线性自编码政策文本挖掘方法。该方法在传统自编码网络架构的基础上，通过引入局部线性约束机制，有效增强了低资源环境下政策文本语义表示中的鲁棒性与泛化能力。

与传统的自编码网络相比，本方法具有以下优势：首先，传统 AE 仅关注全局特征重构，难以保持局部语义空间的拓扑一致性；而本方法通过设计流形正则化约束项，显式地维持了政策文本在嵌入语义空间中的几何拓扑结构。其次，这种结构感知式的建模方式，弥补了标准自编码网络对局部语义平滑性忽视的缺陷，使其在处理多领域、异质性强的政策文本时展现出更优的泛化能力和鲁棒性。最后，本方法继承了 AE 的双向可逆映射特性，不仅能够实现样本内数据的高效表征，还具备优异的样本外数据适应能力，这一特性显著提升了模型在分类场景中的泛化性能[16]。

在实证研究部分，本文选取政策文本挖掘中的语义可视化、主题聚类与分类等典型任务进行系统评估。实验结果表明了本方法在表征语义细节和领域泛化能力方面的优异性能，通过改进文本向量表示质量，后续可实现 3 方面的应用价值提升：（1）在政策聚类方面，增强对跨领域政策主题的辨识能力，支持政府部门把握政策体系的结构特征；（2）在政策分类方面，提高对新兴政策领域的适应性，为政策文件自动归档提供技术支持；（3）在政策分析方面，通过 2D 可视化呈现政策文本的语义分布，辅助决策者直观理解政策关联性与演变趋势。

1 相关研究进展

1.1 政策文本挖掘

基于构建政策文本表示方法的不同，政策文本挖掘可以划分为基于主题模型、基于深度神经网络两类。

（1）基于主题模型

主题模型一般以词袋空间的计数向量作为输入，然后采用矩阵分解或变分推断等方法挖掘出揭示文本隐藏语义的因子，它能够帮助分析人员在更大的数据集上开展政策文本挖掘[17]。一些学者基于主题模型开展政策文本挖掘研究，如孔令凯等[18]利用 LDA 方法挖掘共性技术政策文本的主题内容，并对政策主题强度进行量化分析，然后分析共性技术政策组合，从综合性、一致性及均衡性探究政策组合特征。另外，为了提升主题的可理解性，一些学者将词向量技术与主题模型进行结合，如马海群等[19]选取我国数据安全政策文本，利用 NMF 方法进行主题挖掘，然后从主题内容演化和主题强度演化两个视角出发，分别结合 Word2Vec 词向量模型和主题聚类进行可视化处理与分析。闫盛枫[20]利用动态主题模型（Dynamic Topic Mod-

doi:10.3772/j.issn.2095-915x.2025.03.003

els，DTM）开展公共政策文本时序建模与可视化分析，并通过 Word2Vec 技术以后处理的方式对动态时序主题词进行同义词扩展，进一步增强了政策主题语义解释性。受限于词袋假设割裂词义联系的影响，基于主题模型的政策文本挖掘方法对政策语义细节的捕捉存在不足。

（2）基于深度神经网络

基于深度神经网络的相关研究普遍是以词嵌入向量作为输入，利用卷积神经网络模型（CNN），双向长短期记忆模型（BiLSTM），以及融合自注意力机制（Self-Attention）的 Transformers 架构，构建政策文本表示向量，然后再通过相似度计算、聚类分类等方法进行挖掘分析。如为了解读数字平台的隐私政策内容，Deotale 等 [21] 利用词向量和动态神经网络进行政策主题分类，然后又利用 skip-thought 句子向量和聚类进行摘要分析。Chang 等 [22] 考虑了词序对于文本建模的影响，并探索利用 LSTM 在政策科学文本分类任务中的性能表现。另外，胡吉明等 [23] 采用 LDA 主题模型和 CNN-BiLSTM-Attention 注意力模型提取政策文本表示特征，提升政策文本分类的效果和准确度。沈思等 [24] 使用来自国家及省市的政策发布平台的政策文本数据构建了预训练数据集，并基于 MLM 和 WWM 任务对两个中文 Transformers 架构模型 BERT-base-Chinese 和 Chinese-RoBERTa-wwm-ext 进行预训练，最终实现政策实体抽取。

（3）小结

综上，当前政策文本分析方法主要沿着两条技术路径发展：基于主题模型和基于深度神经网络。虽然这些方法在突破传统词袋模型局限和增强上下文语义理解方面取得了显著进展，但其核心缺陷在于未能充分考虑政策文本特有的流形分布特性。具体而言，现有方法存在两个关键性局限：

首先，在语义空间建模方面，传统方法往往忽视政策文本的局部几何结构特性，使得政策文本在语义空间中局部平滑性难以得到有效维持，容易造成局部近邻的聚集性分布，最终导致细粒度语义关系建模失真。

其次，在跨领域适应性方面，现有方法难以维持不同政策领域间的拓扑关联。当处理异质性较强的政策文本（如同时包含经济政策和环境政策）时，语义空间的整体结构容易发生扭曲，直接影响政策相似性和差异性的度量精度，进而削弱模型的泛化能力，降低鲁棒性。

值得注意的是，虽然流形学习在政策文本挖掘领域尚未得到系统应用，但其在相关文本分析任务（如新闻主题演化追踪 [25] 和中文文本话题分类 [26]）中已展现出卓越的结构保持能力。这些成功案例为本研究将流形学习引入政策文本分析提供了重要的方法借鉴和实践依据。

1.2 流形学习

（1）基本概念

首先，解释与流形学习相关的基本概念 [13]：

流形：设 $\boldsymbol{M}$ 是一个 Hausdorff 空间，如果给定任意一点 $p \in \boldsymbol{M}$，可以找到关于点 p 的一个开邻域 $\mathbf{U} \subset \boldsymbol{M}$ 与 d 维欧式空间 $\mathbb{R}^d$ 上的一个开集同胚，那么就称 $\boldsymbol{M}$ 是 d 维拓扑流形，简称 d 维嵌入流形。换句话说，d 维流形是局部欧式、第二可数的 Hausdorff 空间。

Hausdorff 空间：对于拓扑空间 $\boldsymbol{X}$，给定任

意不同点 $\boldsymbol{x}$ 和 $\boldsymbol{y}$，始终存在 $\boldsymbol{x}$ 的邻域 **U** 和 $\boldsymbol{y}$ 的邻域 **V**，使得 **U** 和 **V** 不相交（**U**∩**V**=⌀），称 $\boldsymbol{X}$ 是 Hausdorff 空间。

由流形定义可知，流形与欧氏空间在局部上存在着同胚映射，即在每个局部邻域上，流形和欧氏空间一样具有平滑性和线性结构（或称局部线性）。流形学习是指利用观测空间中的采集数据恢复潜在低维流形结构，同时构造出数据的低维嵌入表示，其核心是通过构造一个由观测空间到低维流形空间的连续可逆一一映射（同胚映射），实现数据的低维表示。从广义上讲，只要遵循数据集服从流形分布假设，任何以学习数据集的结构特性保持为目标的特征表示方法都可以看作流形学习的范畴。

流形学习代表算法包括局部线性嵌入（Local Linear Embedding，LLE）、拉普拉斯特征映射（Laplacian Eigenmaps，LE）。流形学习算法一般包括 3 个步骤 [27]：

1）基于相似性度量选择数据点的近邻，并构建数据集的近邻图；

2）基于步骤 1 选择的近邻，对局部拓扑结构特性进行线性估计；

3）利用步骤 2 的局部结构特性，构建同胚映射，并获得数据的低维表示；

（2）流形正则化

流形学习的核心是对局部拓扑结构进行近似估计并构建同胚映射，它不局限于具体的技术框架，随着机器学习方法的发展，流形学习与正则化方法紧密融合，催生了流形正则化技术（Manifold Regularization）。流形正则化最早由 Belkin 等 [28] 在 2006 年提出，它是在 Tikonov 正则化的基础上，基于流形局部欧式的结构特性定义泛函，从而求解数据的低维表示及构建同胚映射。一般来说，流形正则化的泛函定义如式（1）所示。

$$J=L+\lambda\sum_{i,j}W_{i,j}\cdot Dist(\boldsymbol{y}_i,\boldsymbol{y}_j) \quad (1)$$

其中 $\sum_{i,j}W_{i,j}\cdot Dist(\boldsymbol{y}_i,\boldsymbol{y}_j)$ 是关于数据集的近邻图的拉普拉斯正则项，λ 是非负正则超参数，用于控制流形正则项对整个优化目标的影响，y_i 代表数据的低维表示，而 $Dist(\cdot)$ 是一个关于某种距离测度的函数，$W_{i,j}$ 代表近邻图上的两个数据点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的连接边权值，一般根据二者是否为近邻取 0 或 1，如式（2）所示。

$$W_{i,j}=\begin{cases}1, \ if\ x_i\in N_p(\boldsymbol{x}_j)或x_j\in N_p(\boldsymbol{x}_i)\\ 0, \qquad\qquad 否则\end{cases} \quad (2)$$

这种定义背后的直观动机是在特征空间中靠近的数据点（几何近邻）在低维空间中的表示应该尽可能相似，从而保持数据在低维空间中几何结构的平滑性。受此启发，很多特征表示研究利用图拉普拉斯算子对近邻图局部样本点的几何结构进行离散近似，以正则化约束项的方式引入机器学习算法优化目标中，如将 LLEAN 整合到线性投影模型中 [29]，将 LapDMM 引入到混合迪利克雷模型中 [30]，通过对目标的最优化求解构造出保持数据在低维空间中几何结构的平滑性的同胚映射，最终构建数据的低维表示。

2　研究思路与方法框架

本方法遵循流形假设，即大规模政策文本数据通常分布在低维流形周围，通过将政策近邻文本的几何拓扑关系作为正则约束项纳入自编码网络优化架构中，构建从特征空间到低维流形空间的嵌入映射函数，从而得到政策文本

doi:10.3772/j.issn.2095-915x.2025.03.003

低维表示向量，然后利用低维表示实现对政策文本的聚类和分类挖掘。研究框架包含 3 个部分：（1）局部政策文本近邻选择，利用基于 Transformer 的预训练模型实现政策文本相似性度量，并建立政策文本近邻图选择局部近邻文本；（2）局部嵌入自编码，利用相似性度量对低维表示向量进行加权嵌入，并作为正则化约束项纳入自编码优化过程，构建低维嵌入映射函数；（3）政策聚类与分类挖掘，基于正则化自编码过程建立政策文本的表示向量，并通过聚类、分类等政策挖掘任务进行效果评估。研究框架图如图 1 所示。

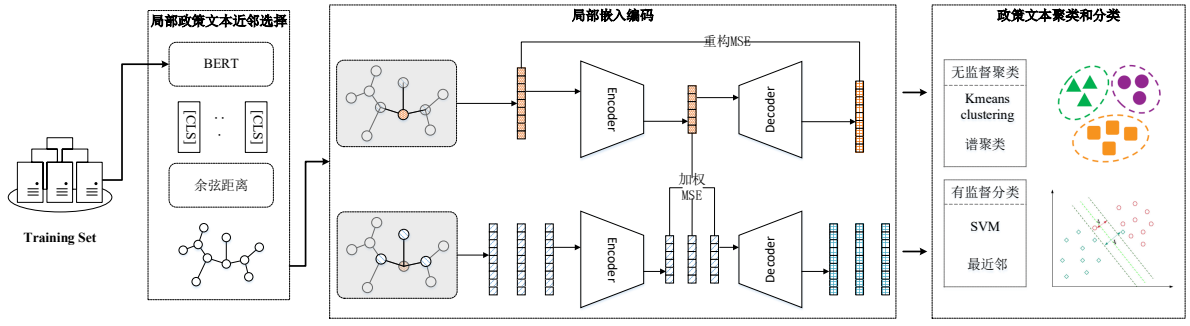

图 1　基于局部嵌入自编码的政策文本挖掘方法研究框架

2.1　局部政策文本近邻选择

常见的近邻选择方法包括 k 最近邻和 ε 最近邻两种。它们都是基于数据集上的相似性度量选择近邻，不同的是 k 最近邻选择 k 个最相似样本作为近邻，而最近邻是选择相似性大于 ε 的那一部分样本作为近邻。本文采用 k 最近邻方法选择局部政策近邻文本，因为 k 最近邻通常比 ε 最近邻表现要好[31]，其中 k 值作为经验参数，具体数值将通过后续的参数敏感性实验进行优化确认。为了增强近邻文本语义相似性和准确性，首先选取公开的无微调预训练模型生成政策文本的语义表示向量（预训练模型核心参数参见 3.2 小节），然后采用余弦距离计算不同政策文本间相似性，最后根据相似性值选取每个政策文本最相似的 k 个文本作为近邻。局部政策文本近邻选择具体处理过程如表 1 所示。

表 1　局部政策文本近邻选择具体处理过程

算法 1：	局部政策文本近邻选择。
输入：	近邻数目 k，输入数据集 $\Omega=\{\boldsymbol{d}_i\}_1^M$
输出：	局部近邻文本集合 $\Psi=\{\mathrm{K}_i\}_1^N$，其中 $\mathrm{K}_i=\{\boldsymbol{s}_j,sim(\boldsymbol{d}_i,\boldsymbol{s}_j)\}_{j=1}^k$ 表示关于 $\boldsymbol{d}_i$ 的 k 个最近邻文本集合，$\boldsymbol{s}_j\in\Omega$ 表示第 j 个最近邻，sim 表示 $\boldsymbol{d}_j$ 与 $\boldsymbol{s}_j$ 的余弦距离。
处理步骤	For 每一个 $\boldsymbol{d}_i$ 对于任意的 $\boldsymbol{d}_{j\neq i}$，计算 $\boldsymbol{x}_i$=pretrain($\boldsymbol{d}_i$) 与 $\boldsymbol{x}_j$=pretrain($\boldsymbol{d}_j$) 的余弦距离， 选取余弦距离值由小到大排序前 k 个文本得到 $\mathrm{K}_i=\{\boldsymbol{s}_j,sim(\boldsymbol{d}_i,\boldsymbol{s}_j)\}_1^k$。 End for

2.2　局部嵌入自编码

本文以自编码网络为基本架构，对近邻政策文本进行局部线性嵌入，AE 是一个 3 层神经网络，其输入输出层神经元数目相同，包括编

码（Encoder）和解码（Decoder）网络[32]。按照自编码重构过程，给定一篇政策文本 $\boldsymbol{d}_i$，经过基于 Transformer 的预训练模型生成表示向量为 $\boldsymbol{x}_i$，其编码 — 解码过程如下：

$$\boldsymbol{y}_i=f(\boldsymbol{x}_i)=\sigma(\boldsymbol{W}\cdot\boldsymbol{x}_i+\boldsymbol{b}) \quad (3)$$

首先，根据公式（3），向量 $\boldsymbol{x}_i$ 经过编码网络权重矩阵 $\boldsymbol{W}$ 的线性变换，加上偏置向量 $\boldsymbol{b}$，再通过非线性激活函数 σ 进行映射，得到关于 $\boldsymbol{x}_i$ 的低维表示向量 $\boldsymbol{y}_i$。然后，根据公式（4），解码网络将 $\boldsymbol{y}_i$ 经过编码网络权重的转置矩阵 $\boldsymbol{W}^{\mathrm{T}}$ 的线性变换，加上偏置向量 $\boldsymbol{c}$，再通过非线性激活函数 σ 进行映射，生成重构向量 $\boldsymbol{x}_i'$。

$$\boldsymbol{x}_i'=f(\boldsymbol{y}_i)=\sigma(\boldsymbol{W}^{\mathrm{T}}\cdot\boldsymbol{y}_i+\boldsymbol{c}) \quad (4)$$

其中公式（3）和（4）中激活函数采用 sigmoid 函数 $\sigma(a)=1/(1+e^a)$，编解码网络的权重矩阵 $\boldsymbol{W}$ 和 $\boldsymbol{W}^{\mathrm{T}}$，中间隐层单元偏置向量 $\boldsymbol{b}$，输出层的偏置向量 $\boldsymbol{c}$。自编码网络以网络输入作为重构目标，其优化目标是最小化 $\mathbf{s}_i$ 和 $\mathbf{o}_i$ 间的重构均方误差（Mean-Square Error，MSE），采用 mini-batch 梯度下降算法求解，为了简洁，将上述参数记为 $\boldsymbol{\theta}=\{\boldsymbol{W},\boldsymbol{b},\boldsymbol{c}\}$，其优化目标函数如式（5）所示。

$$J(\boldsymbol{\theta}|\boldsymbol{\Omega})=\frac{1}{M}\sum_{\boldsymbol{x}_i\in\Omega}\mathcal{L}(\boldsymbol{x}_i,\boldsymbol{x}_i';\boldsymbol{\theta})=\frac{1}{M}\sum_{\boldsymbol{x}_i\in\Omega}\|\boldsymbol{x}_i-\boldsymbol{x}_i'\|^2 \quad (5)$$

通过公式（5）的优化求解过程，自编码网络将专注于以更低重构误差恢复输入文本，从而使得编码网络可以捕获更能揭示输入政策文本的关键语义信息。与基本自动编码网络不同，本文假设低维表示不仅要尽可能恢复输入，同时还依赖于输入的局部政策近邻文本。因此，除了要有效捕获输入政策文本的语义信息，编码网络还要兼顾局部近邻文本的语义信息，即编码网络可以极力恢复局部政策文本中以更高概率近似每个文本及其邻居中的单词出现，最终使得编码网络对于来自近邻的政策文本更加平滑，即分配的低维表示更加相似。为此，通过局部近邻文本与输入文本间 MSE 以及对应余弦距离 d_j 进行加权，构建了正则化约束项纳入上述公式，得到新的重构误差表示如公式（6）所示。

$$J_{new}(\boldsymbol{\theta}|\boldsymbol{\Omega})=\frac{1}{M}\sum_{\boldsymbol{x}_i\in\Omega}\|\boldsymbol{x}_i-\boldsymbol{x}_i'\|^2+\lambda\sum_{\boldsymbol{s}_j\in\boldsymbol{n}_i}d_j\|\boldsymbol{x}_i-\boldsymbol{s}_j\|^2 \quad (6)$$

2.3 政策文本聚类和分类挖掘

对政策文本语义相似性或差异性的辨识能力是很多政策信息管理系统（比如检索、摘要、可视化等）的基础，因此本文选择政策语义聚类和分类两类典型的挖掘任务。具体的，首先，利用优化后的编码网络构建政策文本低维表示向量，然后，分别采用 Kmeans 和谱聚类等无监督聚类算法对政策文本的低维表示进行聚类挖掘，其中，Kmeans 是一种简单而广泛使用的聚类算法，其核心目标是使簇内样本到簇中心（质心）的平方距离之和最小化，即最小化组内平方和。它通过反复更新簇中心和分配样本的迭代优化方式，逐步将数据划分为预定义的 K 个簇。谱聚类是一种基于图论的聚类方法，它通过相似性矩阵构建一个图，然后通过图的拉普拉斯矩阵的特征分解实现非凸形状的聚类分割，其核心是通过相似度矩阵捕捉数据点间的非线性关系，解决传统方法难以处理的复杂结构问题。

最后，采用 SVM 和随机森林（Random Forest，RF）等有监督分类算法进行政策分类挖掘。SVM 是一种有监督学习方法，其核心原理基于间隔最大化准则构建最优分类超平面。

doi:10.3772/j.issn.2095-915x.2025.03.003

该方法通过寻找能将不同类别样本正确分隔且间隔最大的线性决策边界（由支持向量定义），将低维空间的非线性问题借助核函数映射至高维特征空间实现线性可分，同时引入松弛变量构造软间隔以增强对噪声数据的容忍度[33]。RF是一种高效稳健的有监督集成学习方法，它属于基于决策树的组合学习算法范畴。对于一个新的未知样本，RF会通过构建多棵独立训练的决策树形成“森林”，每棵树分别对样本进行类别判断或数值预测，最终通过集体投票（分类任务）的方式整合各树的输出，从而得出鲁棒性更强的最终预测结果[34]。

3 实验结果和分析

3.1 数据集

本文选用北大法宝v6法律数据库作为实验数据来源，以“人工智能”和“大数据”作为领域主题检索关键字，检索时间为2016—2023年，并选取中央法规、地方性法规和法律动态3种类型政策文本，确定相关文本5501篇。其中中央法规主要由全国人大及其常委会、国务院、最高人民法院、最高人民检察院等中央国家机关制定，包括法律、行政法规、司法解释、部门规章等正式法律文本，地方性法规由省级人大、设区市人大及地方政府制定，包括地方性法规、地方政府规章、规范性文件等，聚焦区域治理需求，而法律动态反映法律制定与修订过程的非正式文本，包括立法草案、工作报告（如国务院政府工作报告）、法规解读、执法检查报告等时效性资料，用于追踪立法趋势和政策导向。

两个领域数据分布情况见表2，其中“人工智能”领域包括中央法规121篇、地方性法规1020篇、法律动态53篇，共计1194篇有效法规文本，文本包含词汇平均长度为2847.72。“大数据”领域包括中央法规123篇、地方性法规4064篇、法律动态120篇，共计4307篇有效法规文本，文本包含词汇平均长度为3174.84。最后采用jieba分词并加载哈尔滨工业大学停用词表进行停用词去除等预处理操作后，得到31654个不同词汇。

表2 实验数据分布情况

	中央法规（篇）	地方性法规（篇）	法律动态（篇）	合计（篇）	平均长度
人工智能	121	1020	53	1194	2847.72
大数据	123	4064	120	4307	3174.84

3.2 实验设置

本文共设置两类实验，实验一是本文提出的表示模型的最佳参数设置实验，该实验中将根据式（4）取得的最小损失值来确定最佳模型以及相应的超参数，包括近邻数k和低维表示维度dim。实验二是政策聚类和分类效果评估实验，该实验将利用实验一得到的最佳模型构建政策文本表示向量，并分别采用聚类（Kmenas、谱聚类）和分类（SVM、随机森林）进行效果评估。为了充分验证本文模型的效果，将政策文本集随机乱序，并按照1:1比例平均分成两部分，记作 $\boldsymbol{X}_{train}$ 和 $\boldsymbol{X}_{test}$。实验一利用 $\boldsymbol{X}_{train}$ 训练本文表示模型和其他同类型对比模型，实验二利用 $\boldsymbol{X}_{test}$ 作为模型效果测试数据。

此外，为了对本文方法效果进行对比分析，文本表示对比模型分别选取传统的TF-IDF模型、LDA和LDA2vec主题模型。同时考虑到文

本平均长度大于 512，基于 Transformer 的 BERT 预训练模型存在信息丢失，因此采用支持最大 4096 输入的中文 longformer_zh 模型[①]，并将其作为基于深度神经网络的文本表示对比模型纳入对比分析中。其中，由于 TF-IDF 无需训练过程，因此通过 sckit-learn 算法包对全部数据进行权重计算，并采用 $\boldsymbol{X}_{test}$ 的 TF-IDF 矩阵进行聚类和分类评估。LDA 采用 Gensim 进行训练，其最佳主题数目依据最小困惑度值确定为 120。LDA2vec 基于 lda2vec-pytorch 实现[②]，相关核心参数设置如下：最佳主题数设为 120，lambda_const 为 300.0。longformer_zh 相关核心参数设置如下：隐层维度为 768，vocab_size 为 31654，hidden_act 为 gelu，max_position_embeddings 为 4096。

3.3 聚类和分类评价指标

对于聚类实验，采用 NMI 作为评价指标，其取值范围是 [0,1]，取值越高表明聚类分组与真实分组越相似，其计算公式如式（7）所示：

$$NMI(C,T)=\frac{MI(C,T)}{max(H(C),H(T))} \quad (7)$$

其中，$C=\{C_i\}$ 表示基于各种文本嵌入方法的文本表示经过聚类算法之后的分组结果，而 $T=\{T_i\}$ 表示文本真实类别情况。$H(C)$ 和 $H(T)$ 分别对应两种情况的熵值，$MI(C, T)$ 是两种情况之间的互信息，它的计算公式如式（8）所示：

$$MI(C,T)=\sum_{C_i\in C,T_i\in T} p(C_i,T_i)\log\frac{p(C_i,T_i)}{p(C_i)p(T_i)} \quad (8)$$

$p(C_i, T_i)$ 是文本 $\boldsymbol{x}_i\in C_i$, T_i 的联合概率，表示从文本集中随机选取一篇文本同时属于 C_i, T_i 的概率，而 $p(C_i)$ 表示随机选取一篇文本 $\boldsymbol{x}_i\in C_i$ 的概率，$p(T_i)$ 表示随机选取一篇文本 $\boldsymbol{x}_i\in T_i$ 的概率。

对于分类实验，考虑到数据中存在类别间不平衡的情况，采用加权 F 值 $\bar{F}$ 来评估最终结果的准确率，$\bar{F}$ 值越高表明分类结果越准确，其计算公式如式（9）所示：

$$\bar{F}=\frac{\sum_i c_i F_i}{C} \quad (9)$$

其中，c_i 表示类别 i 的文本占整个测试集的比例，C 表示测试集的整体大小。F_i 是关于类别 i 的 F 值，它与精确率（Precision）P_i 和召回率（Recall）R_i 密切相关，这三种评价指标的计算公式如式（10）、（11）和（12）所示：

$$P_i=\frac{|\{真实文本\}|\cap|\{预测文本\}|}{|\{预测文本\}|} \quad (10)$$

$$R_i=\frac{|\{真实文本\}|\cap|\{预测文本\}|}{|\{真实文本\}|} \quad (11)$$

$$F_i=2\frac{P_i\cdot R_i}{P_i+R_i} \quad (12)$$

3.4 参数选择实验

针对本文方法中近邻数 k 和低维表示维度 dim，利用 $\boldsymbol{X}_{train}$ 数据进行模型训练优化，其中 dim 分别设置为 90、110、130、150，k 分别设置为 11、13、15、17、19、21，并记录式（5）$J_{new}(\boldsymbol{\theta}|\boldsymbol{\Omega})$ 误差值随迭代次数 epoch 变化情况，详见表 3，当 dim 为 110，k 为 19 时，模型取得最优误差值 0.098（加粗显示）。因此，本文方法相关参数设置如下：近邻数 k 为 19，低维表示向量为 110，学习率为 1e-4，迭代数为 100，每个 batch_size 为 8 条，优化器为 Adam，实验运行环境是 python3.8，GeForce RTX 3090（24G），CUDA 版本为 12.0，pytorch 版本为 1.7.1。

① https://huggingface.co/ValkyriaLenneth/longformer_zh/tree/main
② https://github.com/TropComplique/lda2vec-pytorch/tree/master

doi:10.3772/j.issn.2095-915x.2025.03.003

表 3　不同参数组合对应模型损失值（dim,k）

（dim,k）	10	20	30	40	50	60	70	80	90	100
(90,11)	24.672	20.321	13.236	7.347	8.694	6.845	5.367	3.779	3.867	3.857
(90,13)	24.874	20.854	10.847	7.064	9.245	6.784	5.569	4.257	3.567	3.894
(90,15)	24.574	20.103	9.324	6.824	8.364	6.984	5.364	4.125	3.315	3.845
(90,17)	24.354	19.956	8.335	6.536	9.357	6.894	5.268	4.368	3.452	3.947
(90,19)	24.451	19.265	7.214	6.228	10.67	6.214	4.687	4.351	3.424	4.015
(90,21)	23.324	19.324	7.234	6.354	10.357	6.534	4.997	4.687	3.354	3.898
(110,11)	16.362	3.264	3.547	4.123	3.759	2.854	2.311	0.812	1.354	0.98
(110,13)	16.354	3.684	2.957	2.684	2.84	2.256	2.037	0.798	1.365	0.53
(110,15)	16.251	3.882	2.931	2.614	1.954	1.914	1.423	0.601	1.245	0.32
(110,17)	16.012	3.478	2.9	2.654	1.789	1.465	1.137	0.203	1.127	0.21
(110,19)	15.987	3.267	2.88	2.014	1.327	1.044	0.737	**0.098**	0.106	0.113
(110,21)	15.654	5.234	2.84	1.867	1.987	1.394	1.823	0.099	0.987	0.383
(130,11)	23.354	2.364	1.568	1.964	2.794	2.997	2.234	1.295	1.398	0.924
(130,13)	23.312	2.364	1.354	1.035	2.984	3.654	2.802	1.234	1.086	0.857
(130,15)	23.296	2.121	1.124	0.967	2.997	3.817	2.968	1.117	1.564	0.998
(130,17)	23.124	2.301	1.016	0.864	2.634	4.095	2.614	1.954	1.359	0.874
(130,19)	23.012	2.001	1.245	0.884	2.897	3.854	2.554	1.824	1.077	1.004
(130,21)	22.997	1.951	1.874	1.798	3.156	4.351	2.368	1.694	1.388	0.997
(150,11)	23.226	9.583	12.305	9.956	6.984	5.742	3.894	2.952	3.523	3.001
(150,13)	23.325	10.958	12.204	9.684	6.554	5.957	4.854	3.002	3.142	2.589
(150,15)	23.647	12.657	12.021	9.59	5.768	6.259	4.594	3.302	3.341	2.887
(150,17)	23.124	15.397	11.965	9.556	5.241	7.002	4.637	2.99	3.262	2.789
(150,19)	23.364	14.624	11.8	9.451	5.651	7.621	5.127	3.431	3.491	2.996
(150,21)	23.034	18.759	11.324	9.49	5.334	7.035	5.987	3.112	3.482	2.857

3.5　政策聚类

表 4 展示了本文方法与对比模型所构建的政策文本表示在 Kmeans 和谱聚类两种算法的 NMI 表现，相关参数设置如前所述，其中聚类数目与数据真实类别数保持一致，ncluster 设为 2。结果显示，相较于传统的 TF-IDF，本文方法和 LDA、LDAvec、longformer 构建的表示向量在两种无监督聚类设置下表现更优。此外，与其他方法相比，本文方法取得最高 NMI，在 Kmeans 达到 84.98，提升了 7%，在谱聚类达到 85.78，提升了 10%。这些证据表明本文方法构建的低维表示能够有效表征数据之间的相似性结构，从而有利于优化聚类效果。

表 4　不同表示模型在 Kmeans 和谱聚类两种聚类算法的 NMI（ncluster = 2）

	Kmeans	谱聚类
TF-IDF	58.94	61.71
LDA	63.73	63.38
LDA2vec	70.26	73.86
longformer_zh	77.68	75.36
本文方法	84.98	85.78

此外，为了展示本文方法与对比模型所构建政策文本表示的辨识性能，从“人工智能”和“大数据”两类数据中分别随机挑选 350 条数据，采用 t 分布随机邻域嵌入（t-SNE）[33] 对不同方法生成的政策表示进行二维可视化，核心参数设置如下：初始维度参数 init_dims = 30，困惑度 perplexity = 50。图 2 展示了运用 LDA、LDA2vec、longformer 和本文方法等 4 种方法构建的低维表示的二维散点图。每个二维散点表示一个政策文本，相同类别数据采用相同标记表示，标记 1 对应人工智能类别，标记 2 对应大数据类别。

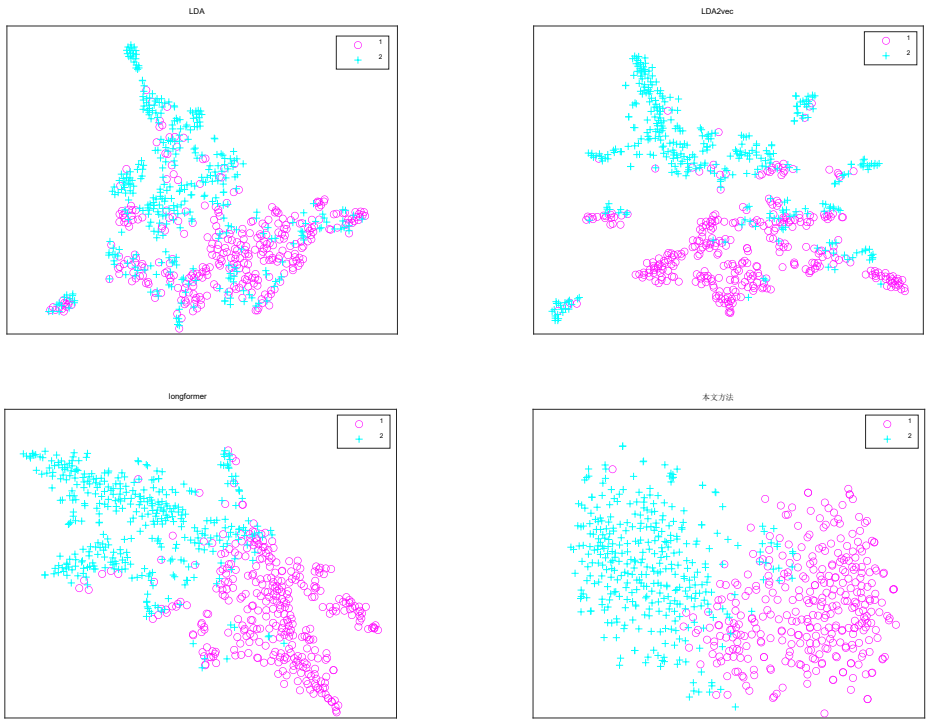

图 2　各方法的政策文本嵌入表示向量的可视化效果

综上分析可以得出以下三方面重要结论：

首先，在可视化效果方面，与其他方法相比，本文方法在二维可视化空间中展现出显著优势：其生成的类别边界更加清晰明确，数据点分布密度更为均匀。具体表现为，其他方法呈现的数据点往往表现出“星状聚集”的效果，而本方法呈现的数据点彼此间分布更加均匀，呈现凸形带状分布。这主要得益于本方法提供的政策文本表示具有优异的同类内聚性和异类分离性，同时在局部上更加平滑，从而避免数据在局部范围出现由于无法表征差异性和相似性而出现的“聚集”效果。

其次，在领域适应性方面，考虑到大数据政策涉及政务、金融、医疗等多场景，人工智能政策多集中于技术伦理、产业应用等有限维度，这种语义分布相对发散的领域政策在聚类过程中难度更大，容易出现效果退化。然而，本文方法的数据点分布呈现出良好的平滑性特

doi:10.3772/j.issn.2095-915x.2025.03.003

征，点与点之间的过渡自然平缓，有效避免了数据粘连现象，为后续的聚类和分类算法提供了良好的基础，使其能够更准确地辨识政策文本的语义相似性。

最后，在交叉语义处理方面，由于“人工智能”与“大数据”政策文本存在一定语义交叉，一些政策文本表现出较强的多主题交叉复合的语义重叠现象。例如《新一代人工智能发展规划》中提及“大数据”24 次，“人工智能”352 次，同时包含“大数据智能”“大数据驱动知识学习成为人工智能发展重点”等复合表述。这就使得在无监督聚类情况下，二者分离难度较大。但相比其他方法，本方法仍能保持较好的 NMI 值和凸形聚类边界，这进一步验证了本方法在处理复杂语义交叉问题上的优越性。

3.6 政策分类

表 5 展示了本文方法与其他对比方法所构建的政策文本表示在 SVM 和 RF 两种分类算法下的表现。从中可以看出，在所有方法中，本文方法在两种分类方法上均取得最佳分类效果，表现出更显著的分类性能提升。具体而言，本文方法将平均值提高到 86.88%（SVM）和 85.76%（RF），相较于 longformer 方法，分别提升了 7.5%（SVM）和 9.3%（RF）。这些证据表明本文方法构建的低维表示具有更强的辨识效果，有利于优化政策分类的效果。这主要得益于本文方法采用了局部视角进行模型训练，通过局部政策文本近邻的加权均方误差和重构误差的联合最小化，使得每个政策文本低维表示充分融合了来自近邻政策文本中的语义信息，增强了低维表示的局部平滑性，使得编码网络倾向于为近邻文本分配更相似的低维表示。

表 5　不同表示模型在 SVM 和 RF 两种分类算法的 $\bar{F}$(%)

	SVM	RF
TF-IDF	66.23	63.86
LDA	70.32	69.86
LDA2vec	73.17	72.64
longformer_zh	79.37	76.46
本文方法	86.88	85.76

4 结论

本文方法通过联合最小化局部政策文本近邻的加权均方误差和重构误差来训练自动编码器，使得编码网络倾向于为来自局部近邻的政策文本分配更相似的低维表示。通过融合流形学习与深度神经网络优势，有效解决了政策文本分析中的语义结构保持和领域适应性问题。本文方法在谱聚类取得 85.78，在 SVM 分类算法方面取得平均值为 86.88%。结果表明，本文方法可以为政策文本提供更具辨识度的低维表示。针对长文本处理，本文采用支持最大 4096 输入的中文 longformer_zh 模型生成固定维度的嵌入向量作为局部线性自编码的输入，从而最大化地保留原始政策文本的语义信息，对于长度超过 4096 的政策文本，longformer_zh 只能进行截断处理，这在一定程度上造成语义信息缺失，影响政策文本语义的细节刻画。值得注意的是，尽管如此，本方法的实验表现仍然优于对比方法，这也表明本文方法在表征语义细节和领域泛化能力方面的优异性能。

在后续实际应用中，该方法可用于提高对

新兴政策领域的适应性，为政策文件自动归档提供技术支持；可支持跨领域政策对比分析，例如面向不同领域的政策工具进行相似性或差异性量化分析；实现政策主题演变分析，如针对相关政策主题利用时序滑动窗口机制，可捕捉政策语义的渐进式变迁。当前研究仍存在一定局限性，包括仅基于中文政策数据训练，未来需要扩展至多语言政策验证跨语种适应性，进一步拓展方法的应用范围。

参 考 文 献

[1] 裴雷 , 孙建军 , 周兆韬 . 政策文本计算：一种新的政策文本解读方式 [J]. 图书与情报 , 2016(6): 47-55.

[2] 曹玲静 , 张志强 . 政策信息学视角下政策文本量化方法研究进展 [J]. 图书与情报 , 2022(6): 70-82.

[3] 胡吉明 . 政策文本研究：从内容计算到功能理解 [J]. 图书情报知识 , 2023, 40(4): 145-152.

[4] 金璐瑶 , 曾静静 . 政策文本量化研究：概念、方法与研究视角 [J]. 数字图书馆论坛 , 2024, 20 (4): 12-21.

[5] GYÓDI K, NAWARO Ł, PALIŃSKI M, et al. Informing policy with text mining: technological change and social challenges[J]. Quality & Quantity, 2023, 57(1): 933-954.

[6] 杨正 . 政策计量的应用 : 概念界限、取向与趋向 [J]. 情报杂志 , 2019, 38(4): 60-65, 51.

[7] 王思 , 张洪铭 . 基于 PMC 指数模型的我国大模型产业政策文本量化研究 [J/OL]. 情报科学 , 1-20. [2025-06-11]. http://kns.cnki.net/kcms/detail/22.1264.G2.20241023.1931.018.html.

[8] 王静茹 , 王东波 , 黄水清 , 等 . 政策工具视角下我国突发事件政策文本研究 [J]. 图书情报工作 , 2024, 68(24): 114-123.

[9] 胡吉明 , 钱玮 , 苗煜松 . 基于特征表示深度学习的政策精准推送模型研究 [J]. 情报杂志 , 2025, 44(2): 167-173.

[10] 武楷彪 , 郎宇翔 , 董瑜 . 融合句法结构和词义信息的政策文本关联挖掘方法研究 [J]. 数据分析与知识发现 , 2022, 6(5): 20-33.

[11] ASUDANI D S, NAGWANI N K, SINGH P. Impact of word embedding models on text analytics in deep learning environment: a review[J]. Artificial intelligence review, 2023, 56(9): 10345-10425.

[12] 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.

[13] MEILĂ M, ZHANG H. Manifold learning: what, how, and why[J]. Annual Review of Statistics and Its Application, 2024, 11(1): 393-417.

[14] LI Q, SORSCHER B, SOMPOLINSKY H. Representations and generalization in artificial and brain neural networks[J]. Proceedings of the National Academy of Sciences, 2024, 121(27): e2311805121.

[15] CERVINO J, CHAMON L F O, HAEFFELE B D, et al. Learning globally smooth functions on manifolds[C]// International Conference on Machine Learning. PMLR, 2023: 3815-3854.

[16] ZHENG Y, HE T, QIU Y, et al. Learning manifold dimensions with conditional variational autoencoders[J]. Advances in Neural Information Processing Systems, 2022(35): 34709-34721.

[17] ISOAHO K, GRITSENKO D, MÄKELÄ E. Topic modeling and text analysis for qualitative policy research[J]. Policy Studies Journal, 2021, 49(1): 300-324.

[18] 孔令凯 , 杨晓萌 , 林超然 , 等 . 基于文本挖掘的共性技术政策文本量化及政策组合特征研究 [J]. 科技管理研究 , 2023, 43(15): 35-45.

[19] 马海群 , 崔文波 , 张涛 . 我国数据安全政策文本主题挖掘及其演化分析 [J]. 现代情报 , 2024, 44(8): 28-38.

[20] 闫盛枫 . 融合词向量语义增强和 DTM 模型的公共政策文本时序建模与演化分析 —— 以 “ 大数据领域 ” 为例 [J]. 情报科学 , 2021, 39(9): 146-154.

[21] DEOTALE R, RAWAT S, VIJAYARAJAN V, et al. POCASUM: policy categorizer and summarizer based on text mining and machine learning[J]. Soft Computing, 2021, 25(14): 9365-9375.

[22] CHANG C, MASTERSON M. Using word order in political text classification with long short-term memory models[J]. Political Analysis, 2020, 28(3):

doi:10.3772/j.issn.2095-915x.2025.03.003

395-411.

[23] 胡吉明，付文麟，钱玮，等 . 融合主题模型和注意力机制的政策文本分类模型 [J]. 情报理论与实践，2021, 44(7): 159-165.

[24] 沈思，陈猛，冯暑阳，等 . ChpoBERT：面向中文政策文本的预训练模型 [J]. 情报学报，2023, 42(12): 1487-1497.

[25] 徐月梅，李杨，梁野，等 . 基于流形学习的新闻主题关系构建和演化研究 [J]. 现代图书情报技术，2016(10): 59-69.

[26] 范少萍，李迎迎，郑春厚 . 基于局部线性判别嵌入算法的中文文本分类研究 [J]. 情报理论与实践，2014(2): 131-135, 126.

[27] WEI C, LUO S, PAN L, et al. Locally weighted embedding topic modeling by markov random walk structure approximation and sparse regularization[J]. Neurocomputing, 2018(285): 35-50.

[28] BELKIN M, NIYOGI P, SINDHWANI V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples[J]. Journal of machine learning research, 2006, 7(11): 2399-2434.

[29] XUE J, ZHANG B, QIANG Q. Local linear embedding with adaptive neighbors[J]. Pattern Recognition, 2023(136): 109205.

[30] LI X, WANG Y, OUYANG J, et al. Topic extraction from extremely short texts with variational manifold regularization[J]. Machine Learning, 2021(110): 1029-1066.

[31] JEBARA T, WANG J, CHANG S F. Graph construction and b-matching for semi-supervised learning[C]// Proceedings of the 26th annual international conference on machine learning. 2009: 441-448.

[32] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[33] DENG X, LI Y, WENG J, et al. Feature selection for text classification: a review[J]. Multimedia Tools and Applications, 2019, 78(3): 3797-3816.

[34] CHEN H, WU L, CHEN J, et al. A comparative study of automated legal text classification using random forests and deep learning[J]. Information Processing & Management, 2022, 59(2): 102798.

[35] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11): 2579-2605.