



计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331, CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目： 基于深度学习的医患舆情多维演化仿真分析
作者： 谭旭，吴璞，蒋知义，邹凯，吕欣
网络首发日期： 2022-09-28
引用格式： 谭旭，吴璞，蒋知义，邹凯，吕欣. 基于深度学习的医患舆情多维演化仿真分析[J/OL]. 计算机工程与应用.
<https://kns.cnki.net/kcms/detail/11.2127.TP.20220926.2002.018.html>



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

基于深度学习的医患舆情多维演化仿真分析

谭旭^{1,2}, 吴璞^{1,2}, 蒋知义¹, 邹凯¹, 吕欣³

1. 湘潭大学 公共管理学院, 湖南 湘潭 411105

2. 深圳信息职业技术学院 素质赋能中心, 广东 深圳 518172

3. 国防科技大学 系统工程学院, 长沙 410076

摘要：为了探析医患关系近年来的情感演化过程，助力政府部门决策干预提供理论依据，从而更好引导网络舆论走向，促进和谐医患关系的构建。针对互联网复杂语境下大规模医患舆情文本数据，在 BERT 预训练模型下游任务中构建情感分类器，并与 LDA 主题抽取技术相结合进行多维情感演化仿真分析，最后结合 ARIMA 模型进行情感走势预测。通过实验分析表明，LDA-BERT 医患舆情多维情感分析模型的情感预测准确度达到 98%，ARIMA 医患舆情时间序列预测模型的预测平均误差低于 11.25%，证明其能够有效运用于大规模医患舆情演化的多维度监测与分析。

关键词：医患舆情；LDA-BERT 模型；ARIMA 模型；情感演化

文献标志码：A 中图分类号：TP 391.9 doi：10.3778/j.issn.1002-8331.2206-0167

Research on multi-dimensional evolution model of doctor-patient public opinion based on deep learning

TAN Xu^{1,2}, WU Pu^{1,2}, JIANG Zhiyi¹, ZOU Kai¹, LYU Xin³

1. School of Public Administration, Xiangtan University, Xiangtan, Hunan 411105, China

2. COME Center, Shenzhen Institute of Information Technology, Shenzhen, Guangdong 518172, China

3. College of System Engineering, National University of Defense Technology, Changsha 410076, China

Abstract: It can better guide the trend of network public opinion and improve the relationship between doctors and patients by analyzing the sentiment evolutionary trends of doctor-patient relationships in recent years and assisting the government in providing the theoretical basis for decision-making and intervention. In the large-scale public opinion text of doctors and patients in the complex context of the Internet, we have constructed a sentiment classifier in the downstream tasks of the pre-training model BERT. It is combined with LDA topic extraction technology for multi-dimensional emotional evolution analysis. Finally, we combine the ARIMA model to predict sentiment trends. The experimental results show that the sentiment prediction model accuracy based on LDA-BERT reaches 98%. The average prediction error of the ARIMA time series prediction of the doctor-patient public opinion model is less than 11.25%. It is demonstrated that the algorithm presented in this research can be effectively applied to the multi-dimensional monitoring and analysis of large-scale doctor-patient public opinion evolution.

Key words: doctor-patient public opinion; LDA-BERT model; ARIMA model; sentiment evolution

基金项目：广东省普通高校创新团队及特色创新项目(2020KCXTD040、2020KTSCX302)；深圳市自然科学基金重点项目(202008291726500001)。

作者简介：谭旭(1981-),男,博士,教授,鹏城学者,研究方向为网络舆情分析与智能评价;吴璞(1994-),男,硕士,研究方向为网络舆情分析;蒋知义(1972-),女,硕士,副研究馆员,研究方向为信息资源管理;邹凯(1965-),男,博士,教授,博导,研究方向为网络舆情分析、智能绩效评价;吕欣(1984-),男,博士,教授,国家杰青,研究方向为大数据挖掘与应急管理。

2021年3月11日,十三届全国人大四次会议表决通过了关于国民经济和社会发展第十四个五年规划和2035年远景目标纲要^[1],提出全面推进健康中国建设,统筹发展和安全,建设更高水平的平安中国;加快数字化发展,建设数字中国。据《中国医师执业状况白皮书(2017年)》数据显示,当前平安中国建设在医疗服务方面还不够系统和完善,不同程度的医疗纠纷、医患冲突等事件频频发生^[2]。同时数字中国建设促进了互联网应用的井喷式增长,但也掀起了网络舆论的热潮。根据中国互联网络信息中心(CNNIC)发布的第47次《中国互联网络发展状况统计报告》,截止2020年12月,我国网民规模达9.89亿,使用手机上网的比例达99.7%^[3]。紧张的医患关系和快速传播的网络载体助推医患舆情的发酵和蔓延,影响整个社会的稳定和谐,亟待政府部门妥善解决。

当前许多学者对于医患关系的法律制度构建^[4,5]、管理模式探索^[6,7]、传媒传播机理探究^[8-10]、评价决策分析^[11]等方面的研究较为广泛。围绕以医患关系为主体的社交媒体网络舆情分析的研究较为缺乏。随着大数据、人工智能技术的爆发式发展,网络舆情已成为衡量主流民众表达情绪的晴雨表,能有效表征医患关系的演化趋势及内涵特点,因此对社交媒体平台的文本进行情感分析能够直观的反映出网民的真实情感倾向。目前文本情感分析主要基于传统机器学习^[12-14]和深度学习方法^[15-17]。由于传统机器学习在情感分类任务中不能有效利用上下文语义信息,因此影响其分类的准确性。较之传统的机器学习, BERT (Bidirectional Encoder Representation from Transformers)预训练模型(Pre-trained Model)摆脱了复杂特征工程的束缚,通过注意力机制^[18]使模型充分学习上下文语义信息,从而实现更精准的情感分类结果,并通过微调在自然语言处理的下游任务中取得较好结果,是目前应用最广泛且效果显著的深度学习预训练模型^[19]。

虽然 BERT 在情感分类任务上表现突出,但在主题维度的挖掘和时间序列数据的预测方面稍显不足。为了多维度且更细粒度地探析情感演化,通过主题挖掘 (Topic Mining) 能够从海量文本信息中识别出关键词与核心主题^[20],其中 LDA 模型^[21]能取得良好的主题挖掘效果且具有灵活的可扩展性^[22-24],适合与其他模型相结合。ARIMA 模型是在 ARMA 模型^[25]的基础上进行差分处理建构的,其具有结构简单、方便操作、预测速度快等特点,相对于其他时间序列预测方法更适合实际应用^[26]。

综上,为了深度理解我国医患关系的演化趋势和

公众的关注焦点,为政府部门健全医患舆情风险预警与应对机制提供科学合理的实证分析结论。本文拟通过第1节模型构建部分进行 LDA-BERT 医患舆情多维演化分析建模和 ARIMA 时间序列预测建模。再通过第2节的实验与分析,验证模型的分类和预测效果,并从粗粒度和细粒度两个层面对全国医患舆情分布和年度医患舆情演化进行深入剖析,得到医患舆情地区、时间、主题、未来趋势的多维度刻画,最后给出实证分析结论。本研究契合政府部门对医患舆情的现实管控需求,对构建和谐医患关系、保持社会和谐稳定具有重要的现实意义。

1 模型的构建

为了深入剖析医患舆情演化,本文分别从地区、时间、主题三个维度解析医患舆情演化过程,并对医患舆情的正负向情感及总的情感演化进行预测。图1为本文所构建的医患舆情多维演化分析及预测模型框架。



图1 医患舆情多维演化分析及预测模型框架

Fig.1 Framework of doctor-patient public opinion multi-dimensional analysis and prediction

1.1 LDA-BERT 医患舆情多维演化分析建模

为提升 BERT 模型的情感分类精度以获得高质量的舆情演化分析效果,增加 BERT 模型主题维度的语义信息,不仅从粗粒度层面直观分析医患舆情全国地区分布情况,而且从结合了主题的细粒度层面对年度医患舆情的演化进行剖析,实现了对医患舆情地区、时间、主题下的多维度分析。本文将 LDA 模型主题抽取技术与 BERT 模型予以融合,以期达到在不同主题下精确表征情感演化过程,包括 LDA 模型主题提取和 BERT 模型情感倾向识别两个阶段,具体流程如图 2

所示。

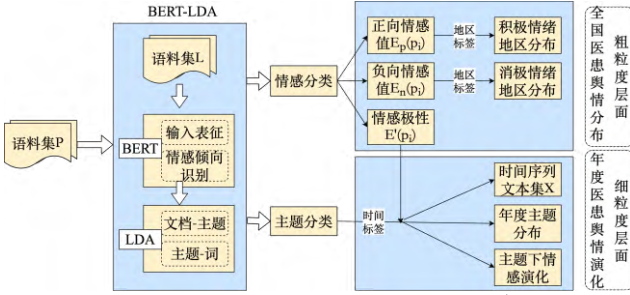


图2 LDA-BERT 医患舆情多维演化分析建模流程

Fig.2 LDA-BERT model for doctor-patient public opinion multi-dimensional analysis

针对文档的主题提取, 给定文档集 $P = \{p_1, p_2, \dots, p_N\}$, 其中, $p_i (i = 1, 2, \dots, N)$ 为若干词条组成的文本, 假设主题集为 $T = \{T_1, T_2, \dots, T_L\}$, 词语集为 $W = \{w_1, w_2, \dots, w_M\}$. 通过将文档集中每篇文档的主题以概率分布的形式给出, 分别得到“文档-主题”概率分布 θ_p 和“主题-词”概率分布 φ_t . 其中, θ_p 和 φ_t 分别满足以 α 和 β 的超参数的 Dirichlet 分布, 如式 (1) 和式 (2) 所示:

$$P(\theta_p | \alpha) = \frac{\Gamma(\sum_{l=1}^L \alpha_l)}{\prod_{l=1}^L \Gamma(\alpha_l)} \prod_{l=1}^L \theta_{p,l}^{\alpha_l - 1}. \quad (1)$$

$$P(\varphi_t | \beta) = \frac{\Gamma(\sum_{m=1}^M \beta_m)}{\prod_{m=1}^M \Gamma(\beta_m)} \prod_{m=1}^M \varphi_{t,m}^{\beta_m - 1}. \quad (2)$$

其中, $\Gamma(\cdot)$ 为 Gamma 函数, $\theta_{p,l}$ 表示主题 T_l 在文档 p 中的概率分布, $\varphi_{t,m}$ 表示词语 w_m 在主题 T_l 中的概率分布。

利用 Gibbs 采样, 可得到“文档-主题”概率 $\theta_{p,l}$ 和“主题-词”概率 $\varphi_{t,m}$, 如式 (3) 和式 (4) 所表达。

$$\theta_{p,l} = \frac{c_{T_l}^w + \alpha_l}{\sum_{l=1}^L c_{T_l}^w + \alpha}. \quad (3)$$

$$\varphi_{t,m} = \frac{c_{T_l}^p + \beta_i}{\sum_{l=1}^L c_{T_l}^w + \beta}. \quad (4)$$

其中, $c_{T_l}^w$ 表示词语 w 被分配给主题 T_l 的频数, $c_{T_l}^p$ 表示文档 p 分配给 T_l 的词语数。

而后为了实现情感的识别, 本文拟采用如图 3 所

示的 BERT^[16] 模型, 主要包含输入层、双向 Transformer 编码层和输出层。BERT 模型输入时会在词向量表示的文本的句首前添加字符 [CLS], 句尾后添加字符 [SEP], 并用 [MASK] 进行句中内容的随机遮盖。在初始词向量的基础上, 通过 Token、Segment 和 Position Embeddings 叠加嵌入来增强表达, 充分刻画每一个 Token 与句子的特征。基于双向 Transformer 架构的编码层, 能够充分学习文本语义与上下文之间的语义关系。再引入注意力机制, 对文本特征进行提取, 将获取到的特征信息进行如式 (5) 所示的加权处理:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (5)$$

其中, Q, K, V 为输入向量矩阵, d_k 是输入向量 K 的维度大小。通过 softmax 函数进行归一化处理, 得到当前单词在每个句子中每个单位位置的表示程度, 并对 V 进行加权。特征信息的权重越大, 表明其在情感分类任务中越重要。BERT 模型在输出时的文本语义融合了整个文本上下文的语义信息。本文在 BERT 原有模型的基础上, 在下游任务中构建了情感分类器, 用于执行情感分类任务 (如表 1 所示)。每个文档记为: $\{p_i, \text{score}: \{E_p(p_i), E_n(p_i)\}, \text{label}: E'(p_i)\}$, 其中 p_i 为第 i 个文档, score 为情感倾向值, label 为该文档的情感倾向标签。

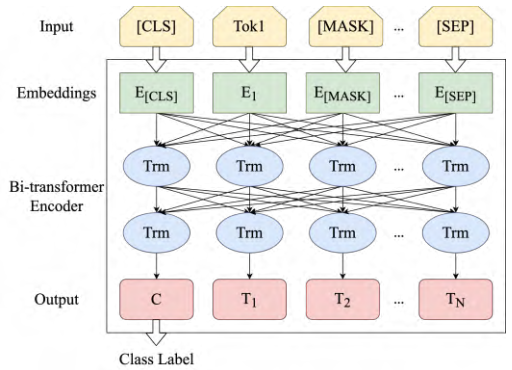


图3 BERT 情感分类模型

Fig.3 BERT model for sentiment classification

表1 情感分类器的标签设置

Table 1 Label for sentiment classifier		
情感倾向标签	0	1
含义	负向情感 (如难过、烦躁、伤心等)	正向情感 (如开心、幸福、快乐等)

由此, 改进后的 LDA-BERT 多维演化算法模型步骤如下:

Step1: 数据预处理与语料集词典构建。输入医患舆情文本语料集 P , 对语料集 P 进行分词、去停用词等预处理, 构建语料词典。

Step2: LDA 主题数确定及主题抽取。通过计算不同主题个数并结合可视化来进行文本主题分类, 确定 LDA 主题最佳个数 K 。通过主题抽取得到每一个文档 p_i 对应的主题 T_l 的“文档-主题”概率分布 $\theta_{p,l}$ 、每一个主题 T_l 对应的词语 w_m 的“主题-词”概率分布 $\phi_{l,m}$, 确定文本语料集中文档所属的主题及各主题下的词语分布。

Step3: 进行 BERT 的输入表征和情感分类器构建。输入标注语料集 L , 通过 BERT 模型将输入语料进行文本的向量化表示, 句首添加字符[CLS]、句尾添加字符[SEP]、字符[MASK]在句中以一定概率随机出现; 然后执行 MLM (Masked Language Model)、NSP (Next Sentence Prediction) 任务; 预训练模型的损失主要在 MLM、NSP 中, 因此两个任务联合学习的损失函数如式 (6) 所示:

$$L(\theta, \theta_1, \theta_2) = L_1(\theta, \theta_1) + L_2(\theta, \theta_2) \\ = -\sum_{i=1}^M \ln p(m = m_i | \theta, \theta_1) - \sum_{j=1}^M \ln p(n = n_j | \theta, \theta_2). \quad (6)$$

其中, θ 是 BERT 中 Encoder 部分的参数, θ_1, θ_2 分别是两个任务中的 Encoder 参数; 基于标注数据集对 BERT-base 预训练模型下游任务中加入新的处理类, 用于执行 SC (Sentiment Classification) 任务;

Step4: BERT 情感分类深度预训练和微调。设置预训练和微调阶段参数: 学习率、Epoch 训练次数、batch 大小、最大序列长度、初始化参数等。为避免训练过程出现过拟合现象, 采用 Adam 算法计算最佳学习率。输出 Epoch 训练结果并迁移至微调模型中, 输入语料集 P , 经过下游 SC 任务计算出各语料集中每个文档 p_i 的情感倾向信息 $\{p_i, score: \{E_p(p_i), E_n(p_i)\}, label: E'(p_i)\}$;

Step5: 将 Step4 中的每个文档 p_i 的情感倾向信息中添加地区标签得到 $\{p_i, score: \{E_p(p_i), E_n(p_i)\}, label: E'(p_i), location\}$ 从粗粒度层面对全国医患舆情分布进行分析; 将 Step2 中生成的“文档-主题”分布添加到每个文档 p_i 的情感倾向信息并融入各文档对应的时间标签得到 $\{p_i, score: \{E_p(p_i), E_n(p_i)\}, label: E'(p_i), theme, time\}$, 并从细粒度层面对年度医患舆情演化进行剖析。

1.2 ARIMA 医患舆情时间序列预测建模

根据 1.1 节 Step5 中的每个文档 p_i 的情感倾向信息, 分别构建时间序列总文本集 $X_{SUM} = \{time, Num(p_i)\}$, 正向情感时间序列 $X_p = \{time, Num(E'(p_i) = 1)\}$, 负向情感时间序列 $X_n = \{time, Num(E'(p_i) = 0)\}$, 时间是按年份/季度来计量的。假设 $X = \{X_1, X_2, \dots, X_t\}$, 为得到平稳时间序列进行差分处理, 那么 X_t 的 d 阶差分 $Y_t = (1 - B)^d X_t$, 其中 B 为滞后算子。 Y_t 服从 ARMA(p, q) 模型, 则 X_t 是符合 ARIMA(p, d, q) 模型过程的。自回归过程如式 (7) 所示:

$$Y_t = \mu_1 Y_{t-1} + \mu_2 Y_{t-2} + \dots + \mu_p Y_{t-p} + \mu_t \\ = \mu_t + \sum_{i=1}^p \mu_i Y_{t-i}, t \in N. \quad (7)$$

其中, Y_t 为时间序列文本集 X_t 的差分值; $\mu_1, \mu_2, \dots, \mu_p$ 为该方程的自回归系数; μ_t 为服从均值为 0、方差为 σ^2 的正态分布且在不同时刻的值互不相关的白噪声过程; 滑动平均过程如式 (8) 所示:

$$u_t = \delta_1 \gamma_{t-1} + \delta_2 \gamma_{t-2} + \dots + \delta_q \gamma_{t-q} + \gamma_t \\ = \gamma_t + \sum_{i=1}^q \delta_i \gamma_{t-i}, t \in N. \quad (8)$$

其中, $\delta_1, \delta_2, \dots, \delta_q$ 为该方程的滑动平均系数; γ_t 为白噪声序列;

由此可得到 ARMA 模型数学表达式 (9):

$$Y_t = \gamma_t + \sum_{i=1}^p \mu_i Y_{t-i} + \sum_{i=1}^q \delta_i \gamma_{t-i}, t \in N. \quad (9)$$

ARIMA 时间序列建模分析步骤如下:

Step1: 拆分训练集和测试集。对构建的医患舆情时间序列 X (包括时间序列总文本集 X_{SUM} 、正向情感时间序列 X_p 、负向情感时间序列 X_n) 均拆分为以 2016 年—2020 年的数据为训练集, 2021 年 1 月至 6 月的数据为测试集;

Step2: 时间序列平稳性检验。对非平稳医患舆情时间序列 X (经过 d 次差分得到平稳时间序列 Y_t);

Step3: ARIMA 模型的识别与定阶。根据自相关函数 (ACF) 与偏自相关函数 (PACF), 初步识别模型中 p 与 q 的可能取值。再基于 AIC 准则、BIC 准则和 HQIC 准则进行模型定阶, 计算出各自对应的值, 选取三者值达到最小的那一组作为理想阶数 ARMA(p, q);

Step4: 白噪声与相关性检验。对差分后的平稳序列进行纯随机性检验, 判断序列是否为白噪声序列。并对 ARMA(p, q) 模型所产生的残差做德宾-沃森 (D-W) 检验, 判断残差的自相关性;

Step5: 模型参数估计。采用最小二乘法估计模型的参

数 μ 和 δ , 对 $t+1$ 时刻进行预测, 公式如式 (10) 所示:

$$Y'_{t+1} = \mu_1 Y'_t + \dots + \mu_p Y'_{t+1-p} + u_t - \delta_1 \gamma_t - \dots - \delta_q \gamma_{t+1-q} \quad (10)$$

Step6: 检验模型性能。用测试集对模型预测结果进行检验, 计算总发文量、正负向情感的平均绝对百分比误差(MAPE)^[27]和对称平均绝对百分比误差(SMAPE), 由此分别计算三者的平均误差率。

2 仿真实验分析

2.1 数据来源

为深度探析近年来医患关系现实情况及医患舆情演化趋势, 本文以新浪微博社交平台为数据源, 构建基于 Python 的网络爬虫框架。在遵循网络爬虫规则的前提下, 通过高级搜索在“医患关系”、“暴力伤医”话题下爬取了 2016 年 1 月 1 日至 2021 年 6 月 10 日的发帖和评论内容, 共计 99841 条。经过数据预处理后最终得到数据 81203 条, 即医患舆情文本语料集 P 为 81203 条。对医患舆情文本语料集进行词云图可视化, 如图 4 所示, 发现医患舆情主要聚焦于暴力伤医、医患关系、医者仁心、伤医事件等方面。预训练语料集 L 来自新浪微博公开正负情感标注数据集约 12 万条 (正负向情感标注各 6 万条), 按 9:1:1 的比例将预训练语料集 L 拆分为训练集 95991 条, 测试集 12000 条, 验证集 12000 条。



图4 医患舆情主题词词云

Fig.4 Theme word clouds of doctor-patient public opinion

2.2 全国医患舆情分布

为剖析医患关系的本质, 本文从宏观维度以近 6 年的医患文本数据为切入点, 呈现我国医患关系在全国范围内地区分布差异性及其正负向情感偏向程度, 利用 LDA-BERT 医患舆情多维演化分析模型进行粗粒度层面的分析。根据 1.1 节算法中的 Step3 在 BERT-base 模型的基础上构建 BERT 情感分类器, 输入语料集 L , 执行 Step4 预训练过程, 设置初始学习率为 $2e-5$, 最大序列长度为 128, Epoch 训练次数为 30 次。通过训练迭代了 8999 步得到模型的 AUC 值为 0.9792, Loss 值为 0.06。为对比本文算法的情感分类效果, 采用文本情感分类经典模型 TextCNN 作为分类基线模型。基于 2.1 节语料集 L 进行训练, 训练结果如表 2 所示。通过实验对比能够清晰的看出, LDA-BERT 情感分类模型效果突出。

表2 情感分类结果对比

情感分类模型	准确率	召回率	F1 值	精确率
LDA-BERT	0.9792	0.9649	0.9750	0.9946
TextCNN	0.8750	0.8756	0.8747	0.8794

由此, 将训练好的模型迁移到情感分类微调模型中, 输入医患舆情文本语料集 P 。通过模型进行情感值计算和情感极性分类。找到语料集 P 中各文档对应的 ID 建构新的 URL 来获取对应的地区标签, 并将标签与各语料集中每个文档 p_i 的正向情感值 $E_p(p_i)$ 和负向情感值 $E_n(p_i)$ 对应, 可以得到时空维度下全国医患舆情分布情况, 分别如图 5、图 6 所示。

图 5 和图 6 分别为全国 34 个省份的医患舆情正负向情感地区分布情况, 其中, 图 5 能够直观看出贵州省在医患舆情上表现最突出的积极情绪, 其次是黑龙江省。图 6 能够清晰看出最具消极情绪的微博用户多分布在甘肃省, 其次为辽宁省和宁夏回族自治区。医患舆情在地区分布上具有明显的差异性, 地区间的差异影响着公众对医患关系的认知, 进而影响着医患关系的发展。因此, 对于医患舆情管控方面, 针对地区差异要有所侧重。对于地区间的医疗资源的不均衡, 也应有所调整。



图5 正向情感全国地区分布图

Fig.5 Regional distribution map of positive sentiment



图6 负向情感全国地区分布图

Fig.6 Regional distribution map of negative sentiment

2.3 年度医患舆情演化

为更细腻地刻画近6年来国内医患关系的变化细节,借助本文LDA-BERT混合模型的多维度智能分析和可视化呈现,从网络舆情的视角探寻总体情感极性的演化情况,如图7所示。根据1.1节算法的Step1进行数据预处理,采用中文停用词表、哈工大停用词表、百度停用词表、四川大学机器智能实验室停用词表进行更为精确的停词处理。将生成的词典经过doc2bow计算每个不同单词出现的次数,将单词转换成整数id,并将结果作为稀疏向量返回,生成语料。训练TFIDF模型^[28]并生成TFIDF矩阵,选择 $\alpha=$

'auto', $\beta = 0.01$ 的先验参数,对LDA模型进行训练,经过迭代计算,K=4时,困惑度最小,主题数最优。结合gensim调用api实现LDA的可视化交互分析。通过pyLDAvis可视化得到图8主题分布情况。验证主题数K=4时主题抽取分布较为均匀合理。图8中左侧部分表示医患舆情语料集四个主题分布,圆的大小表示每个主题出现频率的大小,圆与圆之间的距离远近表示主题间的相似性,可以看到主题4与其余三个主题之间距离较近,关系较为密切。图8右侧部分表示 $\lambda=1$ 时四个主题中排名前三十的词语分布。其中 λ 越接近1,词语与主题的相关性越高。从图中可以发现,与主题最为密切相关的词语主要集中在四个层面:医患主体层面,如“医生”、“医院”、“患者”等;法律惩治层面,如“律师”、“法律”、“严惩”等;医疗纠纷层面,如“医疗”、“医闹”、“看病”等;新闻报道层面,如“新闻”、“卫健委”、“人民日报”等。抽取各主题下排名前10的词语,得到医患舆情“主题-词”分布,见表3所示。

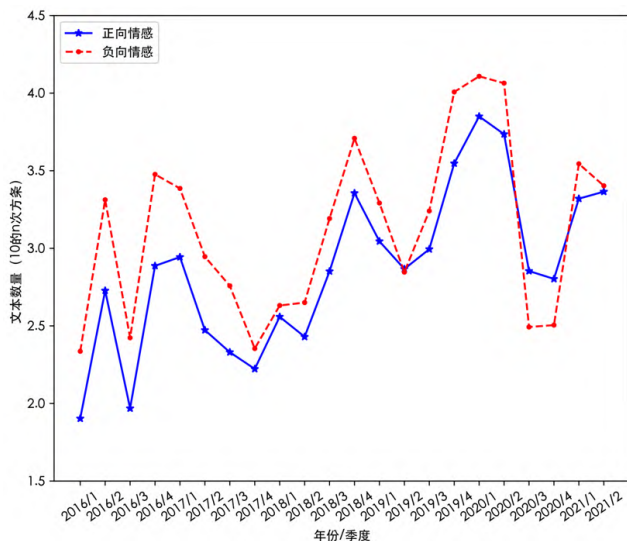


图7 不同年度下医患舆情总体情感极性演化趋势

Fig.7 Sentiment evolution trend of doctor-patient public opinion in different years

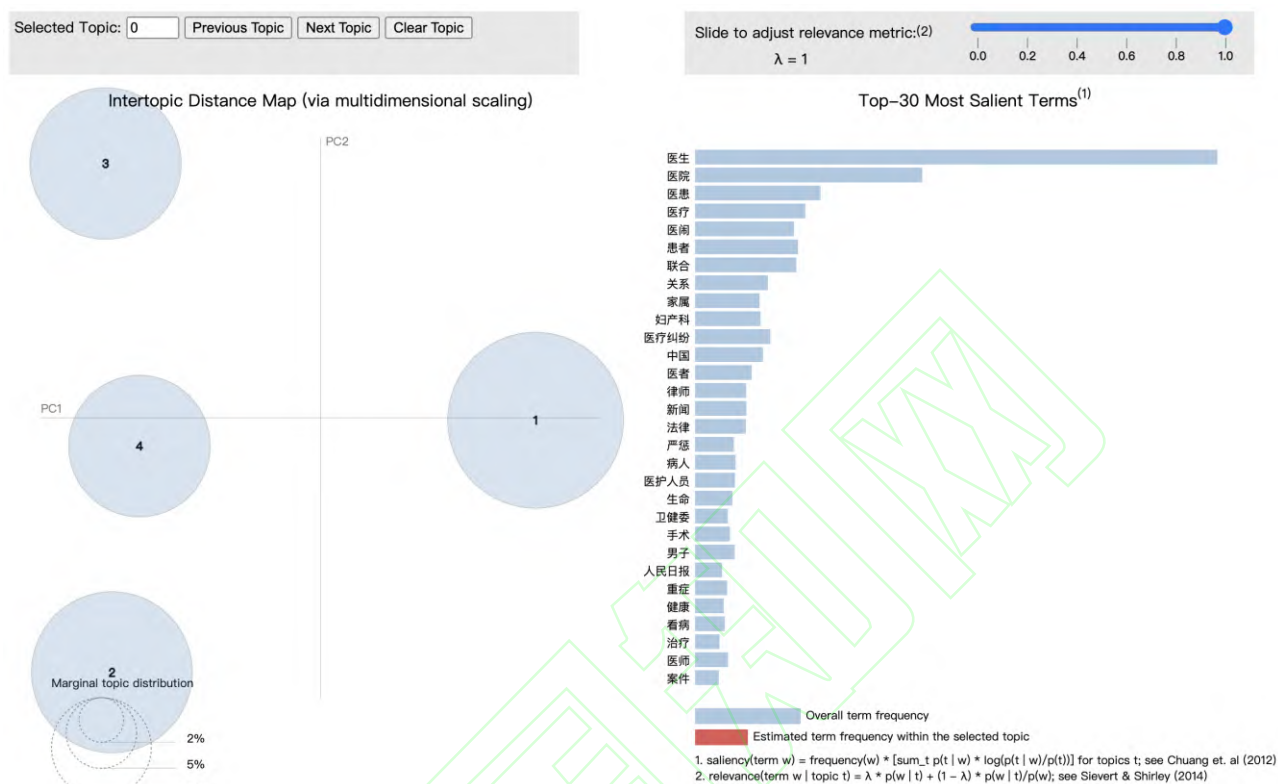


图8 主题分布寻优结果

Fig.8 Optimization results of theme distribution

表3 医患舆情的“主题-词”分布结果

Table 3 The "theme-word" distribution of doctor-patient public opinion

主题	关键词
1	医生(0.446), 联合(0.086), 医疗纠纷(0.064), 中国(0.058), 男子(0.034), 医师(0.028), 医疗事故(0.025), 仁心(0.024), 北京(0.019), 生活(0.015)
2	医疗(0.127), 患者(0.118), 关系(0.084), 妇产科(0.075), 律师(0.059), 病人(0.047), 医护人员(0.046), 卫健委(0.038), 健康(0.033), 病情(0.026)
3	医院(0.231), 医者(0.058), 新闻(0.052), 法律(0.052), 生命(0.038), 手术(0.035), 重症(0.033), 看病(0.0302), 护士(0.028), 态度(0.026)
4	医患(0.165), 医闹(0.13), 家属(0.085), 严惩(0.051), 人民日报(0.035), 治疗(0.032), 案件(0.031), 理解(0.023), 医德(0.019), 说法(0.019)

根据表 3, 按照概率大小分布, 主题 1 概率最高的关键词是“医生”, 主题 2 概率最高的关键词分别为“医疗”、“患者”, 主题 3、4 概率最高分别为“医院”、“医患”。将四个主题分别命名为“医生”、“患者”、“医院”、“医患关系”。

为了时序地演化医患舆情发展态势, 分析不同主题下的情感趋势分布如图 9 所示。在四个主题的整体趋势上, 舆情情感呈波动状态, 在 2019 年第四季度和 2020

年第一季度正负情感都处于较高态势, 在 2020 年第二季度之后正向情感开始高于负向情感, 与 2019 年底暴发的新冠疫情有关, 疫情期间对于医护人员支援武汉的各种正面报道, 给公众重新树立起医护人员医者仁心、无私奉献的形象, 对于缓解医患关系具有重要意义, 因此在疫情之后医患舆情首次出现正向情感高于负向情感。各主题在 2017 年第四季度情感差异较为明显。其中, “医生”主题下正负情感相近, “患者”主题下正负

情感差异最为明显,“医院”主题下和近年的趋势较为接近,“医患关系”主题下正向情感态度最低。这说明

公众对于“患者”与“医患关系”持消极态度,主要是对于患者袭医伤医造成各种医闹事件的不满。

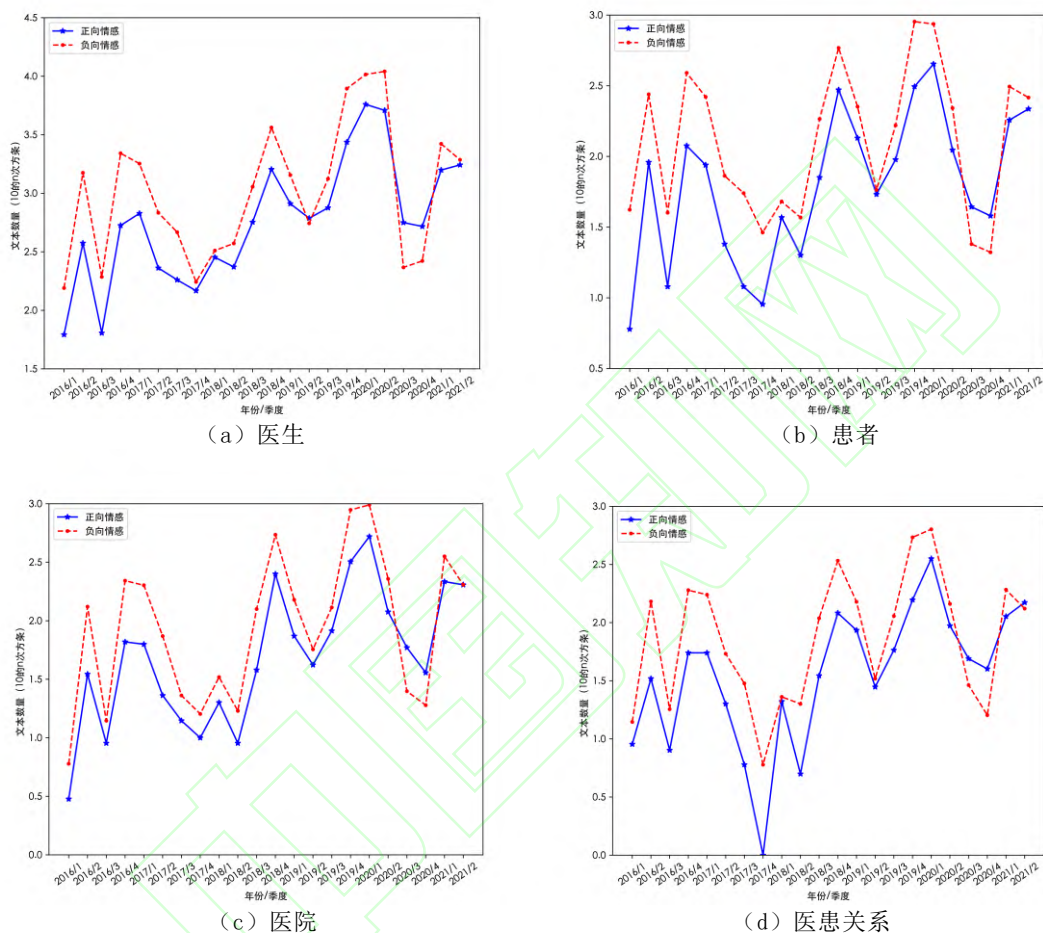


图9 不同主题下医患舆情情感演化趋势图

Fig.9 Sentiment evolution trend of doctor-patient public opinion under different themes



图10 2016—2021年度医患舆情主题词云分析结果

Fig.10 Theme word cloud of doctor-patient public opinion from 2016 to 2021

为探析各年度主题词分布,对 2016—2021 各年度进行主题词抽取并进行词云图可视化分析得到图 10。由图 10 近 6 年来主题词词云展示,将各年度主题分别归纳为:法治监管、医闹伤医、病人救治、医疗纠纷、护医措施、理解医生。发现医患关系中的关注主体由对患者主体的关注转为对医生主体的关注。结合年度整体情感演化趋势(图 7)不难发现,公众对于医患关系的关注与重视是呈现波动上升的,而对于医患关系的情感纠葛从开始的“医闹伤医”、“医疗纠纷”到最近的“保护医生”、“理解医生”是趋于缓和的发展态势。同时,各年度主题也反映了在医患关系中需要在法律层面和护医措施上做出改进。

2.4 医患舆情情感预测分析

为进一步探寻医患舆情情感演化态势,通过对医患舆情情感走势进行科学合理的前瞻和预判,对于政

府部门更好管控社交媒体平台的网络舆情,从而为制定相应舆情防范策略提供合理的理论参考。根据 1.1 节算法中的 Step5,得到时间维度上的舆情情感分析结果,代入 1.2 节算法中的 ARIMA 模型进行预测,确定不同参数下预测效果,分别计算平均绝对百分比误差(MAPE)和对称平均绝对百分比误差(SMAPE),如表 4 所示。由此确定 ARIMA 模型的阶数为(1, 1, 0)时,误差最小。总发文量预测的 MAPE 值为 9.46%、SMAPE 值为 14.15%,正向情感预测的 MAPE 值为 11.15%、SMAPE 值为 14.82%,负向情感预测的 MAPE 值为 6.72%、SMAPE 值为 11.2%。由此得到医患舆情预测模型的平均误差率不超过 11.25%,三者的预测趋势如图 11 所示,可以发现预测结果与真实结果虽然存在一定范围的滞后性,但是整体的演化趋势是一致的,能够较好地预判医患舆情未来发展趋势。

表 4 不同参数下舆情情感预测误差对比情况

Table 4 Sentiment prediction error of doctor-patient public opinion under different parameters

(p,d,q)	MAPE				SMAPE			
	negative	positive	sum	均值	negative	positive	sum	均值
(1,1,0)	6.72	11.15	9.46	9.11	11.2	14.82	14.15	13.39
(3,1,0)	94.99	55.51	66.04	72.18	51.33	43.39	45.81	46.84
(4,1,0)	94.52	58.99	67.95	73.82	57.65	48.28	48	51.31
(4,1,1)	92.54	55.25	76.88	74.89	61.71	45.86	56.14	54.57
(5,1,0)	91.42	57.78	68.11	72.43	60.14	48.32	52.38	53.61

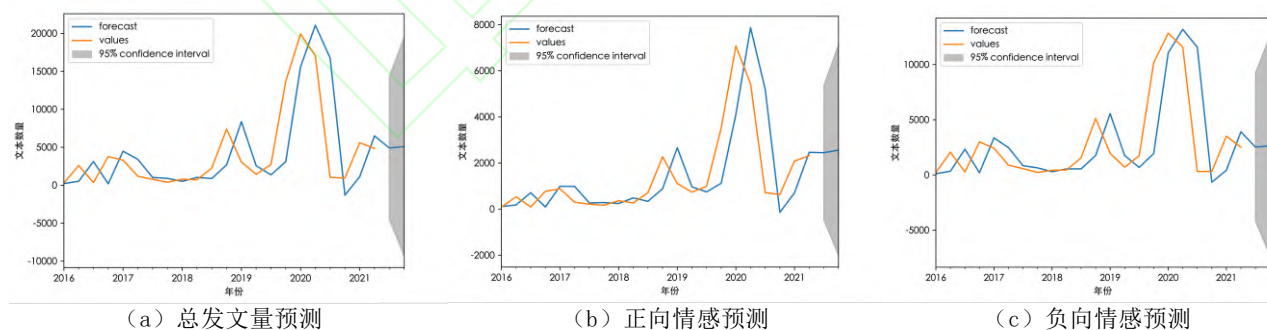


图 11 医患舆情情感预测对比分析

Fig.11 Comparison results of sentiment prediction for doctor-patient public opinion

根据图 11 可以看出,总发文量与正负情感骤升的区间为新冠疫情暴发时期,其余区间数据总体上在一个范围内波动,但 2021 年之后预测的趋势仍然处于缓慢上升态势。对比图 11 (b) 与 (c) 发现,医患舆情负向情感从高于正向情感到逐渐趋于持平,差距逐渐减小,说明医患关系有缓和迹象,但医患关系依旧是一个亟待解决和关注的问题。

3 结语

本文围绕医患关系这桩国内持续讨论的热门话题展开研究。从网络舆情的角度获取互联网复杂语境下大规模医患舆情文本数据,通过构建 LDA-BERT-ARIMA 混合深度学习模型来探寻医患舆情的多维度演化过程,并从粗粒度和细粒度两个维度深度剖析了

医患舆情在不同地区、主题、时间层面的情感演化趋势并进行未来情感走势预测。本文能够较为完整的展现 2016 年至 2021 年医患舆情情感演化和公众对于医患关系话题下重点关注的主题。实证分析表明: (1) 从地区分布来看, 医患舆情地域分布上的情感差异明显, 政府部门可根据地区特色进行针对性干预和监管。(2) 从主题分布来看, 医患舆情关注焦点在法制监管、医疗纠纷、护医措施等层面, 相关部门在法制监管的力度和广度方面还需加强, 对于护医措施方面应明确公安机关、医院、医生和患者不同主体的责任与义务。(3) 从舆情的演化与预测来看, 未来一段时间负向情感与正向情感差距适当减少, 医患关系有缓和迹象, 但仍需重视和监管负向情感舆情态势。

参考文献:

- [1] 新华社. 中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要[EB/OL]. (2021-03-13) [2021-07-13]. http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm.
The Xinhua News Agency. Outline of the 14th five-year plan for national economic and social development and vision 2035 of the PRC[EB/OL]. (2021-03-13) [2021-07-13]. http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm.
- [2] 中国医师协会. 中国医师执业状况白皮书[EB/OL]. (2018-01-09)[2021-05-20]. <http://www.cmda.net/ztwqgzdt/596.jhtml>.
The Chinese Medical Doctor Association. White paper of practice status of Chinese doctors.[EB/OL]. (2018-01-09) [2021-05-20]. <http://www.cmda.net/ztwqgzdt/596.jhtml>.
- [3] 中国互联网络信息中心. 第 47 次中国互联网络发展状况统计报告 [EB/OL]. (2021-02-03)[2021-05-20]. http://www.cnnic.net.cn/hlwfyj/hlwzbg/hlwjbg/202102/t20210203_71361.htm.
China Internet Network Information Center. The 47th statistical report on Internet development in China [EB/OL]. (2021-02-03) [2021-05-20]. http://www.cnnic.net.cn/hlwfyj/hlwzbg/hlwjbg/202102/t20210203_71361.htm.
- [4] 孟祥伟, 王澜, 吴群红等. 多维视角下和谐医患关系构建[J]. 中国公共卫生, 2020, 36(8): 1163-1166.
MENG XW, WANG L, WU Q H, et al. Establishment of
- harmonious doctor-patient relationship under multidimensional perspectives: a cross-sectional study[J]. Chinese Journal of Public Health, 2020, 36(8): 1163-1166.
- [5] 李文芬, 杨风, 罗光强等. 基于 PEST 模型构建和谐医患关系的策略研究[J]. 中国卫生事业管理, 2021, 38(4): 254-255.
LI W F, YANG F, LUO G Q, et al. Research on the strategy to construct harmonious doctor-patient relationship based on PEST model [J]. Chinese Health Service Management, 2021, 38(4): 254-255.
- [6] 王璐, 董琳, 陈明雁等. 融媒体时代医院网络舆情管理模式探索[J]. 中国医院, 2021, 25(5): 87-88.
WANG L, DONG L, CHEN M Y, et al. A preliminary study on hospitals' management mode of internet public opinion in the era of media convergence [J]. Chinese Hospitals, 2021, 25(5): 87-88.
- [7] 王士强. 新媒体环境下医院舆情管理 SWOT 分析及对策[J]. 传媒, 2020(17): 93-96.
WANG S Q. Hospital public opinion management under the new media environment based on SWOT model [J]. Media, 2020(17): 93-96.
- [8] 曹博林. 互联网医疗: 线上医患交流模式、效果及影响机制[J]. 深圳大学学报(人文社会科学版), 2021, 38(1): 119-130.
CAO B L. Internet-based medical treatment: online doctor-patient communication, effect and influence[J]. Journal of Shenzhen University (Humanities and Social Sciences), 2021, 38(1): 119-130.
- [9] 顾秋阳, 琚春华, 鲍福光. 融入用户群体行为的移动社交网络舆情传播动态演化模型研究[J]. 系统科学与数学, 2020, 40(12): 2278-2296.
GU Q Y, JU C H, BAO F G. Study on dynamic evolution model of the spread of public opinion in mobile social network incorporated with the behavior of user group [J]. Journal of Systems Science and Mathematical Sciences, 2020, 40(12): 2278-2296.
- [10] 黄远, 刘怡君. 多层多属性舆情传播网络的仿真研究[J]. 系统工程学报, 2019, 34(6): 844-854.
HUANG Y, LIU Y J. Simulation research on propagation network with multi-layer and multi-attribute public opinions [J]. Journal of Systems Engineering, 2019, 34(6): 844-854.
- [11] 陶思怡, 梁立波, 刘伟等. 网络舆情视域下公众对负面医疗信息的评价分析[J]. 中国医院, 2021, 25(2): 55-57.
TAO S Y, LIANG L B, LIU W, et al. Research on public negative medical information cognition from the perspec-

- tive of online public opinion[J]. Chinese Hospitals, 2021, 25(2):55-57.
- [12] LI J, RAO Y H, JIN F M, et al. Multi-label maximum entropy model for social emotion classification over short text[J]. Neurocomputing, 2016, 210:247-256.
- [13] KAUR S, SIKKA G, AWASTHI L K. Sentiment Analysis Approach Based on N-gram and KNN Classifier[C]//First International Conference on Secure Cyber Computing and Communication, Jalandhar, India, 2018.
- [14] RATHOR A S, AGARWAL A, DIMRI P. Comparative study of machine learning approaches for Amazon reviews[J]. Procedia computer science, 2018, 132:1552-1561.
- [15] 李文亮, 杨秋翔, 秦权. 多特征混合模型文本情感分析方法[J]. 计算机工程与应用, 2021, 57(19):205-213.
- LI W L, YANG Q X, QIN Q. Multi-feature mixed model text sentiment analysis method[J]. Computer Engineering and Applications, 2021, 57(19):205-213.
- [16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30:6000-6010.
- [17] 胡任远, 刘建华, 卜冠南等. 融合 BERT 的多层次语义协同模型情感分析研究[J]. 计算机工程与应用, 2021, 57(13):176-184.
- HU R Y, LIU J H, BU G N, et al. Research on sentiment analysis of multi-level semantic collaboration model fused with BERT[J]. Computer Engineering and Applications, 2021, 57(13):176-184.
- [18] 宋云峰, 任鸽, 杨勇等. 基于注意力的多层次混合融合的多任务多模态情感分析[J]. 计算机应用研究, 2022, 39(3):716-720.
- SONG Y F, REN G, YANG Y, et al. Multimodal sentiment analysis based on hybrid feature fusion of multi-level attention mechanism and multi-task learning[J]. Application Research of Computers, 2022, 39(3):716-720.
- [19] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//The North American Chapter of the Association for Computational Linguistics, Minneapolis, USA, 2019.
- [20] 杨秀璋, 武帅, 夏换等. 基于主题挖掘和情感分析的“新冠肺炎疫情”舆情分析研究[J]. 计算机时代, 2020, (8):31-36.
- YANG X Z, WU S, X H, et al. Public opinion analysis of COVID-19 epidemic using topic mining and emotion analysis[J]. Computer Era, 2020, (8):31-36.
- [21] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. the Journal of Machine Learning Research Archive, 2003, 3:993-1022.
- [22] XING W, MENG F Y, NING J L, et al. Comparative first-principles study of elastic constants of covalent and ionic materials with LDA, GGA, and meta-GGA functionals and the prediction of mechanical hardness[J]. Science China (Technological Sciences), 2021, 64(12):2755-2761.
- [23] 郭业才, 张浩然. 基于改进 LDA 和自编码器的调制识别算法[J]. 系统仿真学报, 2021, 33(2):494-500.
- GUO Y C, ZHANG H R. Modulation recognition algorithm based on improved LDA and autoencoder[J]. Journal of System Simulation, 2021, 33(2):494-500.
- [24] 张舒, 李慧, 施珺等. 结合用户评论与评分信息的推荐算法[J]. 陕西师范大学学报(自然科学版), 2020, 48(2):84-91.
- ZHANG S, LI H, SHI J, et al. Research on recommendation algorithm combining user comments and rating information[J]. Journal of Shaanxi Normal University (Natural Science Edition), 2020, 48(2):84-91.
- [25] PAPPAS S S, EKONOMOU L, KARAMELAS P, et al. Electricity demand load forecasting of the Hellenic power system using an ARMA model[J]. Electric Power Systems Research, 2010, 80(3):256-264.
- [26] 张利. 基于时间序列 ARIMA 模型的分析预测算法研究及系统实现[D]. 江苏大学, 2008.
- ZHANG L. Study of the time series forecasting algorithm and system realization based on the ARIMA model [D]. Jiangsu University, 2008.
- [27] 罗伟. 基于合作对策的股指时间序列组合预测[J]. 系统仿真学报, 2018, 30(6):2086-2094.
- LUO W. Combination forecasting of stock index time series based on cooperative game theory[J]. Journal of System Simulation, 2018, 30(6):2086-2094.
- [28] 叶雪梅, 毛雪岷, 夏锦春, 等. 文本分类 TF-IDF 算法的改进研究[J]. 计算机工程与应用, 2019, 55(2):104-109.
- YE X M, MAO X M, XIA J C, et al. Improved approach to TF-IDF algorithm in text classification[J]. Computer Engineering and Applications, 2019, 55(2):104-109.