

基于文本挖掘的智慧城市建设公众认知研究

梁壮男

(安徽建筑大学数理学院, 合肥 230601)

摘要: 社交媒体平台已成为公众获取智慧城市建设信息的重要来源。为探究公众对智慧城市建设的认知现状,对微博平台中智慧城市建设相关文本数据进行研究。运用 LDA(潜在狄利克雷分布)主题模型得到公众关注的 6 个主题,运用基于情感词典的情感分析方法得到公众对智慧城市建设的情感倾向分布,并对消极文本进行可视化。结果表明,公众关注的主题可分为数字经济、技术应用、股票市场、人工智能、社会民生、政策发展,且公众对智慧城市建设普遍持积极态度。

关键词: 智慧城市; LDA(潜在狄利克雷分布); 情感词典; 公众认知

中图分类号: F293 **文献标志码:** A **文章编号:** 1671-1807(2024)03-0092-06

智慧城市作为一种新兴的城市发展理念与模式,旨在通过新一代信息技术提升城市服务质量以及居民的生活水平,实现城市向智慧化的方向转变^[1]。

目前,我国城市正处于新旧治理模式交替阶段。城市的高速发展、人口的极速膨胀给城市治理带来了巨大的挑战,智慧城市建设是我国城市化进程中的重要举措^[2],加快智慧城市建设能够有效提升城市治理水平,有助于实现城市的可持续发展。智慧城市的发展以提高公众的生活质量和幸福感为目标,研究智慧城市建设过程中公众的认知情况,能够了解智慧城市的发展趋势,帮助政府和企业更好地理解公众的需求,从而调整智慧城市建设的方向和策略。

1 文献综述

作为城市发展领域的热点,智慧城市相关话题备受学术界关注。近年来,国内外对于智慧城市建设现状的研究涉及政策与治理研究、文献计量分析与发展水平评估等多个方面。郭昊等^[3]对中国知网和 WoS(Web of Science)数据库关于智慧城市的文献进行计量分析,发现国内外在智慧城市方面侧重点不同,为未来我国智慧城市发展指明方向。Sharma 等^[4]用潜在狄利克雷分布对发表的文献进行文本分析,梳理智慧城市发展脉络,对智慧城市领域研究的趋势进行总结。李德智和王婷婷^[5]利用文本

挖掘技术对我国智慧城市政策的主题分布及时空特征进行研究,为未来智慧城市政策制定提供建议。Tariq^[6]选取了 90 个指标代表 26 个影响因素,从多个维度上比较澳大利亚主要城市的智慧程度与发展水平,根据结果为澳大利亚提供适当的建议,以提高城市的智慧程度。

上述学者对智慧城市建设现状的研究多是对官方或者学术领域发布的相关信息进行研究,缺少从公众认知角度了解智慧城市建设现状。智慧城市建设过程中离不开人的参与,只有充分发挥人的主体作用,才能实现智慧城市的可持续发展,提升城市的整体运行效率和居民的生活质量^[7]。在信息技术高速发展的时代背景下,社交媒体以互联网为载体为网络用户提供了一个良好的信息表达渠道^[8]。因此,现以智慧城市建设为研究背景,基于微博这一社交媒体平台,运用主题挖掘与情感分析的文本挖掘方法对相关文本数据进行分析,探究公众对智慧城市建设的认知情况,为未来智慧城市建设提供建议。

2 研究设计

利用潜在狄利克雷分布(linear Discriminant analysis, LDA)和基于情感词典的情感分析方法,对智慧城市建设公众的认知现状进行研究。研究框架如图 1 所示,包括数据收集及预处理、模型建立以及结论与建议 3 个部分。

收稿日期: 2023-11-06

基金项目: 安徽省高校自然科学基金项目(KJ2020A0479);安徽建筑大学博士启动基金(2020QDZ20)

作者简介: 梁壮男(1999—),女,安徽宿州人,硕士研究生,研究方向为数据挖掘、机器学习。

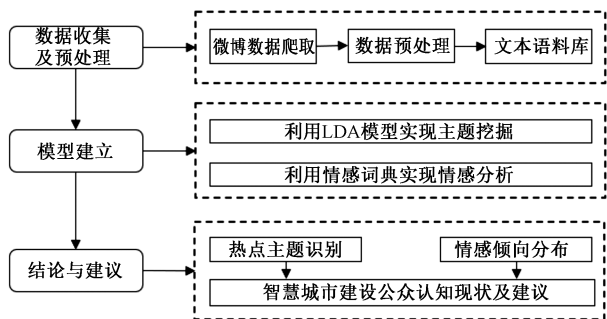


图1 研究框架

2.1 数据收集及预处理

2.1.1 数据收集

将微博平台作为数据源,以“智慧城市”为搜索关键词,运用 Python 中 Scrapy 框架爬取自 2023 年 1 月 1 日至 10 月 1 日的帖子,最终得到 16 835 条微博文本数据。爬取的内容包括微博 ID、微博正文、发布时间以及点赞数、评论数、转发数等信息。

2.1.2 数据预处理

(1)数据清洗。对文本数据进行清洗的目的是通过消除潜在的噪声源来准备文本语料库,以供后续分析^[8]。对于通过爬虫获取的网络文本数据,首先删除本次研究中不需要的信息列,如发布工具、发布位置等;然后删除微博正文中存在的空白博文以及网页链接;最后人工筛查无意义的内容,包括招聘信息、广告等。

(2)分词及去除停用词。文本分词是进行文本分析的基础性工作,借助 Python 中的 Jieba 工具包实现中文分词。将与智慧城市相关的新兴词汇添加到自定义词库,如“智慧交通”“数字经济”等。使用精确分词模式,尝试对语句进行最精确的切分,使其不存在冗余数据,提高文本分析的有效性。同时使用停用词列表去除分词后数据集中的无意义字符以及其他出现频率高但对本次研究无意义的词语。

2.2 模型建立

对通过预处理后的文本语料库,利用 LDA 模型以及情感词典实现主题挖掘与情感分析,研究智慧城市建设公众的认知现状。其中关于 LDA 模型及情感词典的介绍如下。

2.2.1 LDA 模型

LDA 模型^[9]实际上是一个 3 层贝叶斯概率模型,其组成结构包括词、文档、主题 3 种粒度,可用于推测文档的主题分布。LDA 模型认为一篇文档有多个主题,每个主题又对应着不同的关键词,可以通过词汇的概率分布来反映主题。其中主题数 K

是 LDA 模型中的重要参数,可以由困惑度确定,通常困惑度 perplexity 数值越小代表模型的训练效果越好。其计算公式为

$$\text{perplexity} = \exp \left[-\frac{\sum_d \sum_n \ln p(W_d, n)}{\sum_d N_d} \right] \quad (1)$$

式中: N_d 为文档 d 中包括的词语数量; $p(W_d, n)$ 为文档 d 第 n 个词语的生成概率; D 为文档总数。

2.2.2 基于情感词典

情感分析通常被用于从大量的文本数据中提取情感信息,以获得人们的观点、看法、态度等。基于情感词典的情感分析方法是一种常见的基于规则的方法^[10],它依赖于一个预先构建好的情感词典。情感词典是一个包含了大量单词及其情感极性(如正向、负向、中性)的词典^[11]。具体来说,对于分词后的每个词,可以通过查找情感词典中是否存在该词,并获取其情感极性。

在计算文本的情感倾向时,可以将文本中所有情感词的情感极性进行加权求和,得到文本整体的情感极性^[12]。如果总体情感极性为正向,则认为文本是积极的;如果总体情感极性为负向,则认为文本是消极的;如果总体情感极性接近于 0,则认为文本是中性的。

3 实证分析

3.1 描述性统计

在 2023 年 1 月 1 日至 10 月 1 日期间共产生了 16 835 条与智慧城市相关的微博帖子,经过数据清洗后共得到 15 926 条有效数据。其中共有 6 765 个微博用户,平均每个账号发布 2.35 个帖子。微博平台中点赞、评论等用户行为一定程度上体现了用户在该话题的参与度,利用 Python 中 Pyecharts 工具包绘制用户行为数据时序图,如图 2 所示。

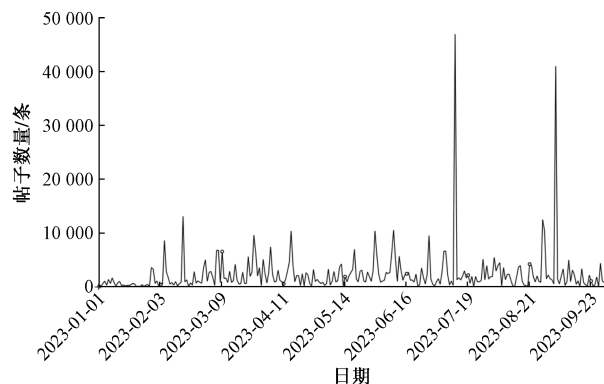


图2 点赞、转发、评论数量时序图

由图 2 可知,2023 年 1—6 月公众对智慧城市的关注度相对稳定,在 2023 年 7—9 月有部分日期公众的关注度明显增加。同时查看相关文本数据发现用户在微博平台对智慧城市相关帖子的点赞量明显多于转发以及评论。说明在新浪微博平台,网民对智慧城市相关话题有一定的关注度,对智慧城市话题感兴趣,但对相关信息的缺乏讨论。

爬取的数据中包含微博发文的具体时间,因此可以分析各个时段用户的活动强度。以点赞数量为例绘制图 3,结果显示在微博平台,平均每天的 09:00—11:00 以及 15:00—17:00 是用户点赞人数较多的时段,可以考虑利用此时段通过微博平台推广与智慧城市建设相关的信息,获取公众的想法,提高智慧城市建设过程中公众的参与度。

统计各城市在文本语料库中出现的频率,将所有城市出现的热度值做归一化处理,将热度值最大的城市作为 100 并进行同比例调整,热度值排名前

10 的地区见表 1。

由表 1 可知,2023 年以来,北京、天津、上海等经济较发达地区,智慧城市建设微博讨论热度相对较高。说明智慧城市的发展离不开经济的支持,在经济发达的省份,对该项目的投入相对较多,成果更加明显,微博讨论热度相对较高。

为了能够更清晰了解在“智慧城市”相关话题中公众讨论的主要话题,对预处理后的微博正文进行词频分析,排名前 20 的高频词见表 2。在基于词频-逆文档频率(term frequency-inverse document frequency,TF-IDF)算法提取的关键词中,选取前 500 个生成词云图,如图 4 所示。

在词云图中,所选取的 500 个关键词,出现的频次越高,字体就越大。由表 2 和图 4 可知,“高质量”“发展”“建设”等词占比较高,表现出公众对智慧城市建设具有较高的期待。“数字经济”“数字化转型”“科技”“数据”“产业”等词出现频率较高,说明智

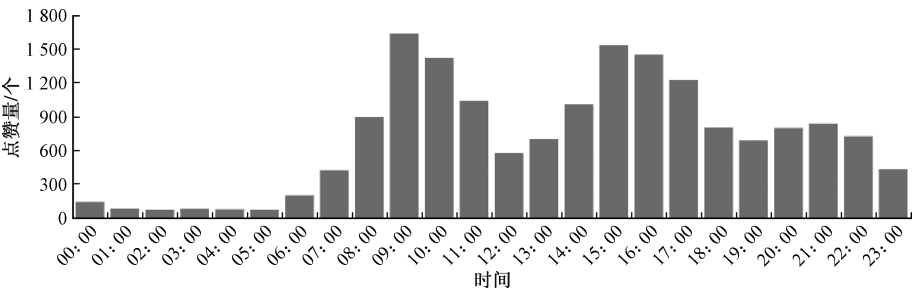


图 3 日均各时段微博用户点赞量统计

表 1 2023 年 1—9 月智慧城市各地区微博讨论热度

地区	热度值	地区	热度值
北京	100	江苏	51
天津	97	山东	38
上海	76	河北	36
浙江	58	重庆	35
广东	56	河南	32

表 2 前 20 高频词

词语	频次	词语	频次
发展	31 955	数字	12 940
建设	26 016	数据	12 902
智慧城市	24 183	智慧	12 804
城市	21 082	创新	12 190
科技	20 251	领域	12 149
新	17 237	项目	11 594
服务	16 858	合作	11 194
智能	14 075	推动	11 099
技术	13 556	平台	10 557
产业	13 140	人工智能	9 630

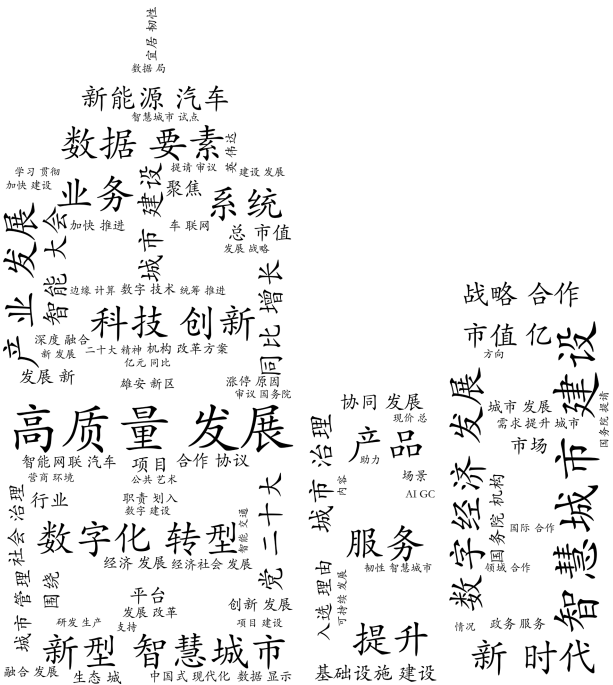


图 4 文本词云图

慧城市建设过程中离不开这些关键要素，在一定程度上代表了智慧城市发展的方向和趋势。

3.2 主题模型构建

利用 LDA 模型聚类之前，需要先确定文本主题个数 K 的值。利用 Python 中的 gensim 工具包绘制困惑度曲线，如图 5 所示。

根据图 5 可知，当主题数为 6 时，困惑度最小。利用 Python 中的工具包可视化 LDA 主题模型的结果如图 6 所示。可视化结果中圆圈之间的距离表示所识别主题之间的关联程度。当圆圈之间的距离太近或重叠时，可以看作是一个相似的主题。当主题数 $K=6$ 时，圆圈之间没有较多重合，因此将 LDA 模型的主题数量确定为 6。

LDA 模型提取的 6 个主题下的 20 个关键词见表 3。根据每个主题所包含的关键词，对每个主题进行概括。主题 1 反映公众对智慧城市建设中数字

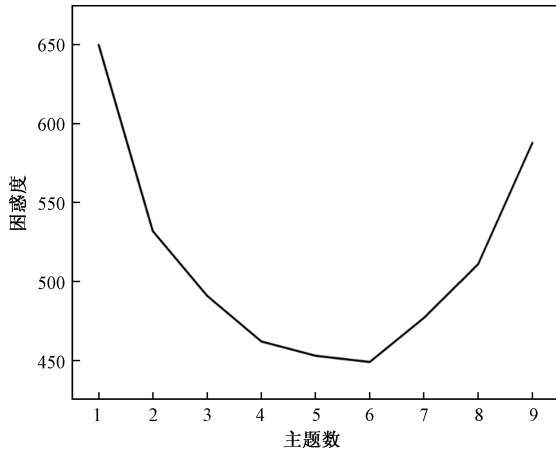


图 5 困惑度曲线

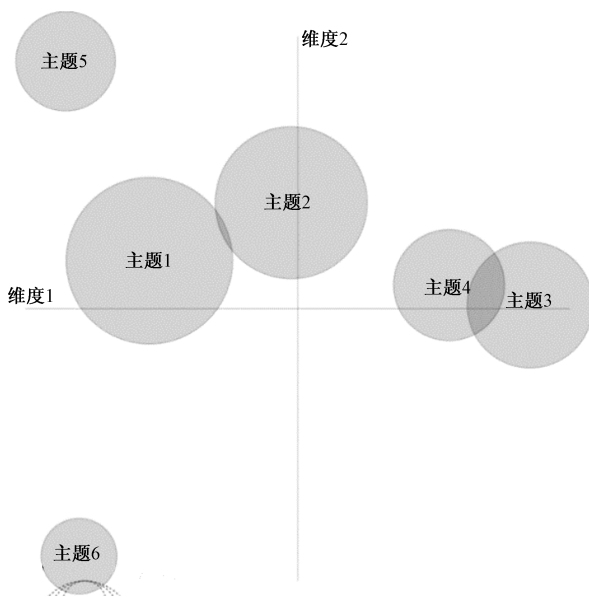


图 6 LDA 可视化结果

表 3 主题关键词

主题	主题关键词
1	发展、建设、产业、推动、创新、数字经济、项目、全球、推进、新、经济、数字、合作、高质量、提升、现代化、加快、服务、科技、工业
2	数据、应用、技术、建设、数字、平台、智慧、城市、大数据、数字化、提供、系统、场景、管理、5G、领域、信息化、行业、产品、物联网
3	概念、股份、科技、亿元、板块、涨停、业务、产品、增长、电子、汽车、市场、同比、服务、原因、新能源、个股、投资、信息、资金
4	合作、AI、科技、互联网、人工智能、技术、模型、软件、智能、发展、产品、华为、算力、市场、研发、卫星、芯片、加速、动力、支撑
5	城市、建设、生活、发展、社区、住房、文化、城市更新、改造、居民、设计、建筑、提升、小区、推进、公园、治理、打造、规划、空间
6	战略、国务院、建设、发展、职责、推进、管理、工作、协调、科技、数据、改革、统筹、中央、部门、组织、监管、数字、社会、加快

经济及相关产业的关注；主题 2 反映公众对智慧城市建设中相关技术应用及基础设施建设的关注；主题 3 反映公众对智慧城市建设相关股票市场的关注；主题 4 反映公众对智慧城市建设中人工智能技术发展的关注；主题 5 反映公众对智慧城市建设背景下社会民生发展的关注；主题 6 反映公众对智慧城市建设相关政策及战略布局的关注。分别将其概括为数字经济、技术应用、股票市场、人工智能、社会民生、政策发展。结合 LDA 主题识别图（图 6），可知在 6 个主题中，主题 1 占比最高，且主题 1 与主题 2 中有部分重合，说明数字经济及相关产业是支撑智慧城市建设的重要组成部分，与相关技术与基础设施的发展相互促进。主题 3 与主题 4 也有部分重合，这是因为智慧城市建设相关的股票市场中包括一些与人工智能相关的企业与技术。

3.3 情感分析

BosonNLP 情感词典主要来源于社交媒体文本^[12]，因此使用基于 BosonNLP 情感词典的方法进行情感分析。实验结果显示，在清洗后的 15 926 条数据中，有 13 236 条数据的情感倾向为积极，约占 83.1%；1 793 条数据的情感倾向为中性，约占 11.3%；897 条数据的情感倾向为消极，约占 5.6%。说明在微博平台中，公众对智慧城市建设基本持积极态度，大部分用户对智慧城市建设的发展是看好的，未来智慧城市建设仍有较大的发展潜力。

由于负面信息在人们的认知和情感上能产生更强烈的影响,为了更好地了解公众对智慧城市建设的看法,对消极情绪文本进行词云图可视化呈现,如图7所示。结合词云图并查阅相关文本数据可知,在智慧城市话题中公众的负面讨论主要集中在以下两个方面:①智慧城市建设成果转化方面,涉及“智慧停车”“联想”等话题。智慧城市建设过程中一些便民设施的设立如果不能确保落实,会导致公众无法享受智慧城市带来的便利和服务,从而引发抱怨。②公众隐私的安全问题。智慧城市建设过程中需要大量的数据采集和处理,存在公众个人隐私被滥用或泄露的风险。

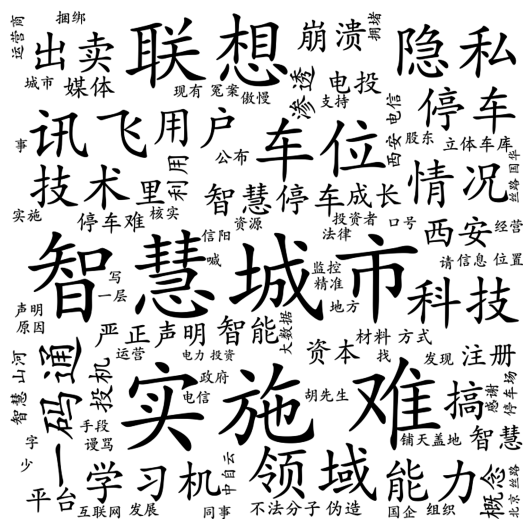


图7 消极文本词云图

4 结论与建议

4.1 结论

选取微博平台2023年1—9月与智慧城市话题相关的在线文本数据,通过LDA主题模型与基于情感词典的情感分析方法对智慧城市建设中公众的看法进行研究,得出以下结论。

(1)当前微博用户对智慧城市话题有一定的关注,多是对相关官方微博的点赞,对相关问题展开的讨论相对较少。且经济发达地区智慧城市的讨论热度相对较高。

(2)公众关注热点围绕数字经济、技术应用、股票市场、人工智能、社会民生以及政策发展6个方面,其中数字经济话题受到的关注最多。

(3)公众对智慧城市发展整体上表现出积极正面的情感态度,消极情绪主要集中在对用户隐私安全以及建设成果共享方面,同时对智慧城市实施难度表示担心。

4.2 建议

基于以上分析结果,对智慧城市建设提出以下建议。

(1)抓住大数据和新媒体时代的机遇,积极利用社交媒体平台,从多个维度展示智慧城市建设的进展和动态,加大智慧城市建设宣传力度,提高公众对智慧城市概念的认知,引发公众的积极讨论。

(2)重点解决公众普遍关注的智慧城市建设热点问题,尤其是公众关注最多的数字经济领域。同时相关部门应密切关注公众的需求,将其作为政策制定和决策的重要依据,以提供更好的公共服务从而满足公众的期望。

(3)政府部门应该及时回应公众对智慧城市建设中存在的问题的关注,引导公众理性诉求从而产生更多有效互动。同时拓宽和创新公众参与智慧城市建设的渠道,便于公众及时反映问题,提高公众参与智慧城市建设和管理的效率和效果,推动智慧城市朝着更利于人的发展方向建设。

参考文献

- [1] 张小娟,贾海薇,张振刚. 智慧城市背景下城市治理的创新发展模式研究[J]. 中国科技论坛, 2017(10): 105-111.
- [2] 路永华,朱晓荣. 智慧城市建设中有关问题研究[J]. 物联网技术, 2019, 9(7): 83-86.
- [3] 郭昊,商容轩,米加宁. 智慧城市:理论缘起、进展与未来方向:基于文献挖掘的发现[J]. 电子政务, 2022(11): 63-73.
- [4] SHARMA C, BATRA I, SHARMA S, et al. Predicting trends and research patterns of smart cities: a semi-automatic review using latent Dirichlet allocation (LDA)[J]. IEEE Access, 2022, 10: 121080-121095.
- [5] 李德智,王婷婷. 我国智慧城市政策主题分析与时空特征研究[J]. 现代城市研究, 2023(2): 38-44.
- [6] TARIQ M A U R, FAUMATU A, HUSSEIN M, et al. Smart city-ranking of major Australian cities to achieve a smarter future[J]. Sustainability, 2020, 12(7): 2797.
- [7] BLASI S, GOBBO E, SEDITA S R. Smart cities and citizen engagement: evidence from T-witter data analysis on Italian municipalities[J]. Journal of Urban Management, 2022, 11(2): 153-165.
- [8] 黄楠,肖俊,张增一. 新冠肺炎疫情期间公众对人工智能的认知与态度:基于新浪微博内容的文本挖掘[J]. 科普研究, 2021, 16(5): 33-41.
- [9] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
- [10] 陈俊,席宁丽,李佳敏,等. 融合 Skip-gram 与 R-SOPMI 的教育领域情感词典构建[J]. 应用科学学报, 2023, 41(5): 870-880.

[11] 邓东. 情感词典构建方法及其应用研究[D]. 北京: 北京交通大学, 2020.

[12] 张公让, 鲍超, 王晓玉, 等. 基于评论数据的文本语义挖掘与情感分析[J]. 情报科学, 2021, 39(5): 53-61.

Research on Public Perception of Smart City Construction Based on Text Mining

LIANG Zhuangnan

(School of Mathematics and Physics, Anhui Jianzhu University, Hefei 230601, China)

Abstract: Social media platforms have become an important source for the public to obtain information about smart city construction. In order to explore the current situation of the public's perception of smart city construction, text data related to smart city construction in microblogging platforms is studied. The LDA(linear Discriminant analysis) topic model is used to get the six topics of public concern, the sentiment analysis method based on sentiment dictionary is used to get the distribution of public sentiment tendency towards smart city construction, and the negative text is visualized. The results show that the topics of public concern can be categorized into digital economy, technology application, stock market, artificial intelligence, social livelihood, and policy development, and that the public generally has a positive attitude towards smart city construction.

Keywords: smart cities; LDA(linear Discriminant analysis); affective lexicon; public perception