政府数据开放平台用户评论情感差异分析

刘桂琴
（湖北师范大学图书馆，黄石 435002）

摘要：根据用户在政府数据开放平台的评论反馈，通过主题分类进行情感分析，明确平台提供服务的用户满意度及存在的问题，为优化开放数据平台的建设提供新的分析思路。利用LDA模型对武汉市政府数据开放网站的用户评论数据进行主题提取，结合深度神经网络进行评论分类，并在此基础上进行情感分析，对不同类型的评论情感差异进行探讨。LDA模型共提取9个分类主题，结合情感分析结果，2个主题的情感趋向是满意状态，7个主题的情感趋向是一般或不满意状态，根据分析结果总结平台服务中存在的不足，并提出相应的优化策略。

关键词：主题模型；政府数据开放平台；情感分析；情感差异
中图分类号：G350
DOI：10.3772/j.issn.1673-2286.2019.02.003

1 研究背景与现状

政府数据开放与传统的政府信息公开相比，在提高政府透明度，提高政府治理能力与服务效率，以及推动社会经济发展等方面具有重要意义[1]。开放政府数据资源具有巨大的应用价值和发展潜力，随着移动互联网的发展和公众对公共数据的迫切需求，政府数据开放成为必然趋势，各国纷纷建立不同层级的数据开放平台以打破“信息孤岛”，应对各行业的信息需求。2009年，美国奥巴马政府颁布《开放政府指令》，并推出全球首个政府数据开放平台Data.gov，此后英国、澳大利亚、新加坡、日本等紧跟潮流，纷纷制定政府数据开放战略并建设国家级政府数据开放平台[2]。我国为了建设服务型政府，进一步推进新型工业化、城镇化、农业现代化发展，对政府数据开放也显现出巨大的热情并投入了大量工作。2015年8月31日，国务院发布《促进大数据发展行动纲要》，明确我国将在2018年底前建成国家政府数据统一开放平台，截至2018年6月，我国已有46个地方政府数据开放平台，这些平台的成立为国家政府数据统一开放平台的建设提供了有益的探索。

我国地方政府数据平台的建设能够成为国家政府数据开放平台的蓝本，但与国外相对成熟的政府数据开放平台相比，我国地方政府数据开放的效果及其应用实践仍有不足，存在不少亟待完善的问题（如数据质量不高、数据覆盖面不均衡等），而国内学者也从多个角度进行了探讨。如余奕昊等[3]采用案例研究和对比研究方法，从功能架构的视角分析了我国地方政府数据开放平台的现状及存在的问题，并提出相应的优化策略；沈晶等[4]建立了政府数据开放发展速度评估体系，并选取5个省级政府数据开放平台计算得到政府数据开放发展速度指数；夏姚璜[5]通过文献调研发现，我国开放政府数据平台的地理空间元数据标准还不够成熟，提出应当依据国情注重地理空间元数据标准的功能开发；汤志伟等[6]基于政府创新扩散理论，构建“政治-经济-社会”三要素模型来探究政府数据开放发展水平的影响因素；刘畅等[7]基于天津市政府信息公开平台，考察了天津市政府数据开放的实践情况并提出建议；李梅等[8]通过分析政府数据开放的主要影响因素及各因素间的复杂关系，找出影响政府数据开放的直接因素、间接因素及深层次基础因素，用以全面地提升政府数据开放水平。

分析发现，已有研究多从指标确定、模型构建、实

证分析角度出发，探讨我国政府数据开放平台的建设现状并提出相应的优化策略，以提高我国政府数据开放水平。而政府数据开放平台的用户是实际体验者和数据的使用者，用户在使用过程中，对整个平台或某一项服务会产生积极或负面的情感。通过对用户的评论反馈进行情感分析，能够发现平台的服务效率、数据质量、功能完备等是否让用户满意。而已有研究较少从用户角度入手，将政府数据开放平台的用户评论进行分类，通过分析不同类型评论情感差异，进行政府数据开放平台满意度评价，发现平台服务中存在的问题。与此同时，随着人们对人工智能、机器学习的不断应用，传统实证调查受限于时间和空间，调查效率相对偏低；而大数据时代的来临，为我们创新调查方式，为更加快捷、高效、科学地评估政府数据开放平台的服务质量提供了一种可能。因此，本文从用户角度出发，以目前较为成熟的情感分析技术为手段，以国内建议反馈功能较为完善的武汉市政府数据开放平台为例，搜集用户评论数据，将主题模型与深度神经网络结合进行用户评论主题提取以及评论分类，探讨在评论分类基础上的用户情感趋势；将时序变化与情感差异结合起来，对不同类型的评论情感差异进行分析，确定用户对政府数据开放平台的满意度；结合反馈内容，快速定位痛点，并提出优化策略，以提升我国政府数据开放平台建设的水平。

2 深度神经网络简介

深度神经网络又称为深度学习，是由人工神经网络发展而来，是一种无监督特征学习和特征层次结构的学习方法，实际上是一种包含多个隐藏层的神经网络算法。Hinton在2006年首次提出深度学习的主要观点，认为多隐层的神经网络可以通过无监督的学习来逐层初始化，从而有效地标示出深度神经网络训练上的难度；其本质思想是通过海量训练数据构建多隐层的模型，来学习更有用的特征，从而提高预测的准确性[9]。

神经网络作为一种计算模型，是模拟大脑皮层的神经网络结构和功能而提出的，其表示输入的信号经由大量的神经节点构成的网络，按一定的规则进行大规模并行计算，得到输出的完整过程。其中，每个神经节点都需要进行函数计算，称为激励函数；每两个节点之间的权重用来模拟“记忆”强度，经过多个隐藏层的节点计算之后能够得到一系列的输出值，输出的结果与期望值进行对比之后可以用来调整函数和权重，从而更加接近预期结果。将学习得来的神经网络模型应用于新的输入（即进行预测），就能得到预测的结果，这样就实现了机器学习的目的[10]，其模型架构见图1。

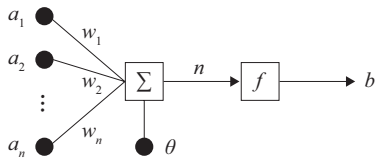

图1 神经网络模型架构

其中，a表示输入层原始数据，b为输出层的结果，w表示权重，θ表示偏移量，f为激励函数，计算过程见公式（1）。

$$b = f(n) = f\left(\sum_{a=1}^{n} a_i w_i + \theta\right) \tag{1}$$

3 政府数据开放平台用户情感分析步骤

3.1 LDA模型主题提取

（1）文本分词。首先对评论文本进行去停用词处理；然后运用公共词库，结合自定义词典对评论文本进行分词；进而将每次分词中出现的新词加入到自定义词库，经过多次词库更新，重复迭代，最终形成分词结果。

（2）确定最优主题数目。对分词处理后的评论文本进行主题建模，在利用LDA的过程中需要人工确定最优主题数目，此处要综合考虑算法的时间复杂度和算法复杂度，本文利用所有主题之间的平均余弦相似度来度量该主题结构的稳定性。计算过程见公式（2）和公式（3）。

$$Sim(T_i, T_j) = \frac{T_{AB}}{\sqrt{x}} = \frac{\sum_{i=1}^{n}(W_{Ai} \times W_{Bi})}{\sum_{i=1}^{n}\left(\sqrt{(W_{Ai})^2} \times \sqrt{(W_{Bi})^2}\right)} \tag{2}$$

$$avg - Sim(structure) = \frac{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} Sim(T_i, T_j)}{\frac{K \times (K-1)}{2}} \tag{3}$$

其中，$Sim(T_i, T_j)$表示主题T_i和T_j之间的余弦相似度；$avg-Sim(structure)$表示所有主题的平均余弦相似度，平均余弦值的取值区间在0和1之间，值越小，表

明主题结构最优。

（3）主题建模。获取最优主题结构后，利用LDA模型对评论语料库进行主题建模，产生包含*T*个主题的“评论-主题”概率矩阵，其中每一行代表一个用户的评论，每一列表示某个主题对不同评论的支持度，支持度越大，表示该评论与相应主题越相关。

3.2 深度神经网络进行文本分类

3.2.1 搭建神经网络层

①将“评论-主题”矩阵作为特征矩阵，并分割特征矩阵，20%的数据作为测试集，80%的数据作为训练集；②搭建输入层，根据特征矩阵的维数确定输入层神经元数量；③搭建隐藏层，根据输入层和输出层的神经元数量推测隐藏层中每层的神经元数量，并尝试搭建不同的隐藏层数（大于或等于3层），为了加快模型训练收敛速度，在隐藏层采用relu函数作为激活函数；④搭建输出层，选用softmax作为激活函数处理分类任务，根据输出结果获取一条评论属于每一主题的可能性，其中数值最高的作为分类结果。

3.2.2 文本分类

①将“评论-主题”概率矩阵的训练集作为原始输入，输入到深度神经网络模型中进行模型的训练；②依据主题支持度确定分类结果是否合理，结合分类精准度和损失度来调整隐藏层的层数，确定最优预测模型；③将测试集输入到已训练好的模型中，经过softmax激活函数进行分类输出之后，得到每条评论的判别结果，根据判别结果对评论进行分类。

3.3 基于评论分类进行用户情感分析

根据分类结果，对原始评论文本进行分类，进而对每一分类下的评论进行情感分析；使用python提供的SnowNLP包，获取每条评论代表正面情感的概率值；并基于ggplot绘图工具包按照时间序列绘制情感分析可视化图谱，明确用户对某一服务在不同时间段的情感变化，并汇总统计负面评价；将用户情感趋向分为3个状态，通过横向和纵向对比，展现用户对政府数据开放平台所提供服务的满意度，定位存在的不足。整个情感分析流程见图2。

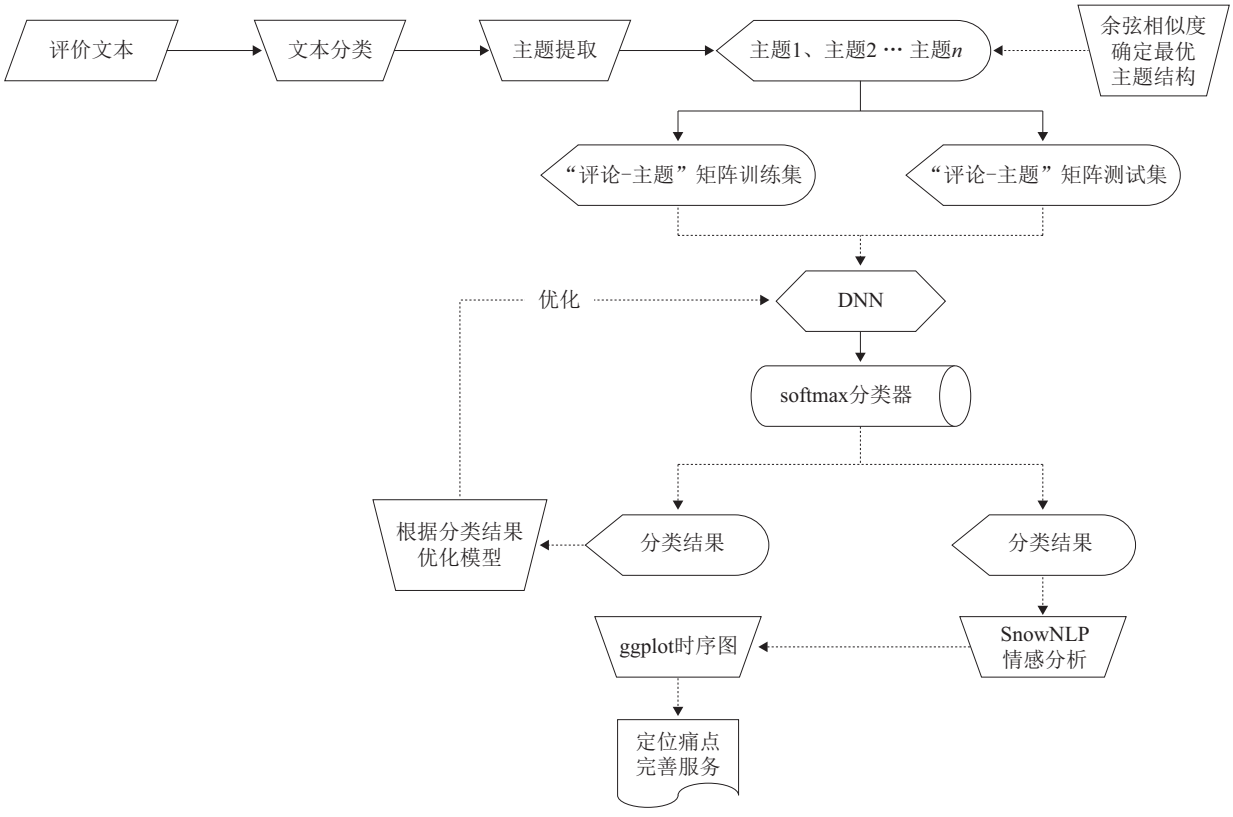

图2 图书馆用户情感分析流程

4 数据处理与结果分析

4.1 数据处理

依据《2018年中国地方政府数据开放报告》的统计结果，北京市、贵阳市、武汉市、江西省、广东省5个政府数据开放平台提供了较好的建议反馈功能模块，评分较高，对这5个平台通过电话和邮件咨询的方式请求获取用户评论数据。武汉市政务公开数据服务网提供了平台技术人员联系方式，通过申请获取到评论数据、问卷调查的统计数据，结合平台网站中显示的评论，经过数据清洗，共选取1 000条评论，内容包括评论时间和内容，时间跨度为2015年3月—2018年10月，对每条评论进行编码，编码顺序为1~1000。首先利用R语言中jiebaR包进行中文分词、去停用词等自然处理规范化过程以构建实验语料库，剔除一些常用的词语（如“的”“吗”等），并依据每次的分词结果对自定义词库进行更新，通过多次迭代获得最终分词结果。然后基于开源包JGibbLDA实现LDA主题建模，通过尝试不同的主题数目，计算主题间平均余弦相似度，获取最优的主题结构，并依据每个主题下的相关词为主题命名。而后对LDA提取的“评论-主题”概率矩阵进行分割，80%的数据作为训练集，20%的数据作为测试集；利用python的pandas包进行数据导入，采用建立在Tensorflow之上的高度抽象框架Tflearn，构建深度神经网络模型，逐步搭建神经网络层，依据工程学的通用做法，输入层和输出层的神经元数量相加除以2取整作为隐藏层的神经元数量，并尝试搭建3层以上的隐藏层，根据每条评论的主题支持度验证深度神经网络的分类结果，依据准确率和损失度，调整隐藏层神经元数量和隐藏层数，以获得最优的分类模型。最后将测试集输入到模型中，获得分类结果，并依据分类结果根据编号对原始评论文本进行分类，对于每个主题下评论文本，利用python的SnowNLP包进行情感分析，得到每条评论的正面情感概率值，并利用ggplot绘图工具包绘制时间序列情感统计图。

4.2 结果分析

根据LDA模型提取结果，结合平均余弦相似度的计算结果，当主题数为9时，对应模型结构最优，依据主题相关词对主题进行命名，9个评论主题分别为数据容量、数据更新、数据全面性、可视化、下载权限、APP应用、数据接口、网站优化、平台调研。

通过对包含800条数据的“评论-主题”矩阵进行训练，通过多次调整确定，当输入层包含9个神经元，输出层包含9个神经元，隐含层的神经元数量为9，隐藏层为3层时，分类结果最优。将包含200条数据的“评论-主题”概率矩阵导入训练好的模型中进行分类测试，根据结果对200条评论文本进行分类，并逐一进行情感分析，绘制情感分析结果数值时间序列图，结合各评论主题的时间序列图进行情感差异分析。

分析发现，关于下载权限的评论，除个别极值点，其情感分析结果数值基本处于0.2以下，表明评论没有正面情感。通过python数据框pandas显示评论里情感分析数值较低的评论发现，内容较多涉及“登录下载数据麻烦”“下载数据步骤烦琐”等，而出现个别极值点的原因是少数用户认可注册登录才能获取数据的方式，提高数据获取门槛有利于数据的安全，但多数用户认为登录后才能下载数据增加了时间成本，建议扩大数据下载权限。关于数据更新和平台调研的评论，其情感分析数值都处于0.65以上，而平台调研分值基本处于0.8以上，表明用户对这两项服务的评论都趋向强烈的正面情感。武汉市政务公开数据服务网自2015年上线以来，数据持续更新，目前公开的数据和服务有2 700多项；而针对高校、企业用户的调研需求，平台都能积极反馈，提供联系方式，以满足用户的调研需求，通过情感分析能发现用户对这两项服务有较高满意度。而关于网站优化的评论，2015—2016年用户基本趋向负面情感，由于平台刚刚上线，会出现延迟、卡顿、服务器响应失败等问题，极大地影响用户体验；而武汉市政府通过不断升级平台系统，积极向北京市、贵阳市等地方平台学习，优化布局设计与功能模块，2017年之后，用户对网站优化的评论都趋向正面情感。关于数据容量和数据接口的评论，情感分析数值基本处于0.5以下，受数据收集复杂度的影响，导致用户获取的部分数据存在缺失和容量不足，同时用户获取数据的方式存在多样性，而平台提供的数据接口难以应对需求，导致这两项评论的情感值偏低。但武汉市政务公开数据服务网提供了12个主题数据，涉及公共卫生、医疗服务等多个方面，虽然主题覆盖度仍需增加，但半数评论的情感值都呈现正面情感。对于可视化和APP应用的评论情感值，都经历了从开始的负面情感到正面情感的过渡，体现网站在不断优化，以提升用户满意度。

通过对9个主题下评论的情感分析，能够发现用户对政府开放数据平台的情感变化，部分主题如下载权限、网站优化、可视化、APP应用等，在不同时间下，其情感差异较为显著，经历了从负面情感向正面情感的转变，体现了武汉市政务公开数据服务网在不断提高用户满意度方面所做的努力。通过对情感数值较低的评价进行统计分析，能够探寻当前平台存在的不足，有利于及时抓住问题，为完善平台的相关服务、优化平台提供决策支持，9个主题下评论的情感分析结果平均数值统计见表1。

表1 9个评论主题情感分析结果平均数值统计

评论主题	均 值	评论主题	均 值	评论主题	均 值
数据容量	0.290	数据更新	0.810	数据全面性	0.450
可视化	0.637	下载权限	0.205	APP应用	0.550
数据接口	0.270	网站优化	0.653	平台调研	0.960

分析可知，根据均值不同，将用户评论的情感趋向分为3个状态：满意（均值>0.700）、一般（均值>0.400）、不满意。根据计算结果，处于满意状态的评论分类有数据更新、平台调研；处于一般状态的评论分类有数据全面性、可视化、APP应用、网站优化；处于不满意状态的评论分类有数据容量、数据接口、下载权限。根据评论分类的情感趋向划分，结合每个分类的情感分析时间序列图，能够清晰地了解在某个评论分类下，每位用户的情感趋向，根据情感分析结果均值探寻用户整体满意度，并结合极值点，统计阅读负面评论，明确用户指向，快速定位政府数据开放平台服务中存在的不足。从总体上看，情感趋向为满意状态的评论分类，是对政府数据开放平台服务工作的肯定，需要继续坚持和努力；而更为重要的是，需要重点关注情感趋向为一般或不满意状态的评论分类，通过分析，目前武汉市政府数据开放平台主要存在4个需要改进的方面：①数据容量过低，开放数据集主题覆盖率不均衡，统计数据量较多但价值较低；②可视化动态展示内容较为单一，APP覆盖领域仍需丰富；③获取数据门槛较高，数据接口较少，缺乏灵活性和便捷性；④平台缺少个性化服务，平台之间缺乏互联互通，智能化水平仍需提高。

4.3 平台优化策略

通过以上分析，根据武汉市政府数据开放平台中存在的问题，从以下4个方面制定了有针对性的优化策略，以提高平台的服务能力，真正实现政府开放数据的价值增值。

（1）培养“大数据”思维，加大数据开放力度。提高数据集质量，减少统计数据，多开放原始数据，提供完整的数据字段；增加数据主题覆盖面，鼓励各部门提高数据开放的参与度，及时根据社会需求覆盖相关主题；同时，保持数据可持续性更新，并将历史上多个批次的数据留存供用户获取。

（2）深化技术研发，提高平台工具的可利用性。充分考虑不同用户群体的需求差异和使用体验，提供更为丰富的可视化功能，动态展示多样化的内容，实现可视化工具与平台数据的互联互通；同时，政府要发挥先锋作用，充分利用物联网、移动互联网等技术平台，挖掘社会大数据，开发覆盖多领域、全部门的APP应用，并保持应用版本的迭代更新，向公众展示政府开放数据的内在价值，方便群众的同时，进一步密切市民与政府的关系，提升政府服务职能。

（3）采取数据分级分类注册的方式，提高数据接口开放率。不分数据类型和用户类型，数据获取采用严格的注册登录方式，极大地影响用户数据获取和利用数据的体验感，增加用户时间成本。因此，采取数据分级分类的注册方式，设置普遍开放、依申请开放和有限开放的数据获取门槛，降低多数用户获取数据的复杂度，以进一步推进数据开放的广度和深度。同时，平台应为每个数据集提供API接口，提升数据接口的开放率，降低数据获取的权限，便于互联网企业快速利用数据，通过政府数据开放带动创新发展，实现数据价值增值。

（4）“个人中心”实现个性化整合，提高智能化服务水平。优化平台设计，用户“个人中心”应实现各项服务的个性化整合，用户通过自定义场景功能，实现个性化搜索、下载和分析，增强用户使用体验。同时，政府应依托数据开放平台实现向“数据治理”转型，应用大数据、云计算技术，实现海量政府数据资源的互联互通；通过人工智能、机器学习等手段，在平台中为用户提供个性化数据推荐、智能交互、精准检索等功能，提高信息处理能力，使数据价值得到充分体现。

5 结语

本文以武汉市政府数据开放平台的用户评论建议

为实验数据，融合LDA模型和深度神经网络进行主题提取和文本分类，借助python的SnowNLP包进行情感分析，并绘制情感分析时间序列图，总结了武汉市政府数据开放平台建立以来存在的问题，并提出了针对性建议。

本文的研究价值在于，运用主题模型认知对象维度，利用主题模型，从1 000条用户评论中提取出9个主题，明确政府数据开放平台中用户关注的服务及其存在的问题；并通过量化的手段将定性文字转为用户情感定量评价，结合情感分析结果时间序列图，对用户评论进行定量评估，通过纵向和横向比较，明确用户的服务满意度变化；通过查阅用户评论，及时定位痛点，优化服务；通过机器学习对文本数据的挖掘和开发，也会成为传统统计调查的有益补充，二者结合将协助我们更客观地理解问题并得到更科学的评估结果；更重要的是通过本文的实验过程，也为分析其他政府数据开放平台提供一定的借鉴与参考，通过对用户评论建议进行情感分析，深度挖掘用户数据，为优化政府数据开放平台的建设提供帮助。

需要说明的是，本文的研究仅以一个地方政府数据开放平台为主，样本较少，数据源有待进一步扩展丰富；在文本分类过程中，其建模效果受机器学习参数值的影响，合理性需要进一步验证；另外，模型并没有推广实施，其适用性还有待进一步确认，这些都会对本文结论产生一定影响，我们将在下一步的研究中进行改进。

参考文献

[1] 张子良，马海群. 我国政府数据开放平台利用效果比较研究 [J]. 数字图书馆论坛，2016（6）：8-15.

[2] 武琳，吴绮琪. 英美公共安全领域政府开放数据应用进展 [J]. 情报杂志，2018，37（4）：183-186，207.

[3] 余奕昊，李卫东. 我国地方政府数据开放平台现状、问题及优化策略——基于10个地方政府数据开放平台的研究 [J]. 电子政务，2018（10）：99-114.

[4] 沈晶，韩磊，胡广伟. 政府数据开放发展速度指数研究——基于我国省级政府数据开放平台的评估 [J]. 情报杂志，2018，37（11）：156-163.

[5] 夏姚璜. 开放政府数据平台的地理空间元数据标准研究 [J]. 图书馆建设，2018（8）：40-46，53.

[6] 汤志伟，郭雨晖，顾金周，等. 创新扩散视角下政府数据开放平台发展水平研究：基于全国18个地方政府的实证分析 [J]. 图书馆理论与实践，2018（6）：1-7.

[7] 刘畅，刘璇. 天津政府数据开放：现状、问题与政策建议 [J]. 图书馆学研究，2018（11）：40-47.

[8] 李梅，张毅，杨奕. 政府数据开放影响因素的关系结构分析 [J]. 情报科学，2018，36（4）：144-149.

[9] 逯万辉. 基于深度学习的学术期刊选题同质化测度方法研究 [J]. 情报资料工作，2017（5）：105-112.

[10] 郭利敏，刘炜，吴佩娟，等. 机器学习在图书馆应用初探：以TensorFlow为例 [J]. 大学图书馆学报，2017，35（6）：31-40.

作者简介

刘桂琴，女，1984年生，硕士，馆员，研究方向：图书馆读者与情报服务，E-mail：lgq0713818@hbnu.edu.cn。

Analysis of Sentiment Differences in User Comments of Government Data Open Platform

LIU GuiQin

(Hubei Normal University Library, Huangshi 435002, China)

Abstract: According to users' comments and feedback on the open platform of government data, this paper makes sentiment analysis through subject classification, clarifies users' satisfaction with the service provided by the platform and the existing problems, and provides new analysis ideas for optimizing the construction of open data platform. Firstly, the LDA model is used to extract the topic from the user comment data of Wuhan Government Open Data Website, and the deep neural network is used to classify the comment. On this basis, sentiment analysis is carried out to explore the emotional differences of different types of comment. LDA model extracts nine categories of topics. Combining with the results of emotional analysis, the emotional trend of two topics is satisfactory, and the emotional trend of six topics is general or unsatisfactory. Based on the analysis results, the deficiencies in platform services are summarized and the corresponding optimization strategies are put forward.

Keywords: Topic Model; Government Data Open Platform; Sentiment Analysis; Sentiment Diversity

（收稿日期：2019-01-04）