“智慧政务”中的文本挖掘应用

摘要

本文利用来自互联网公开来源的群众问政留言记录及相关部门群众留言的答复意见等数据，采用自然语言处理和文本挖掘等技术对数据进行分析、归类。对反映特定地点或特定人群的留言问题进行归类，分析出社情民意热度评价指标对各类社情民意留言问题的热度高低的影响，并建立有效的数学模型进行评价，有助于相关部门及时发现热点问题并进行针对性地处理，帮助其提高服务效率。

针对问题一，本文借助 Python 软件，采用 TF-IDF 算法等相关算法对数据进行去除异常数据及其分隔符等处理，按照一定的划分体系，建立关于留言内容的一级标签分类模型。

针对问题二，根据不同热度评价指标及其对留言问题热度评价的高低影响程度对其赋予权重，利用层次分析法获取其权重以及各个单项指标的评分，通过一致性检验，得出热度评价模型。

针对问题三，通过研究各地方针对以网络留言为主体的网络问政现象和网络问政活动建立的制度化的回应和办理机制，再通过采用 HNC 语义块对留言数据及相关部门的答复内容从相关性、完整性、可解释性的角度进行权重比较，利用基于贝叶斯定理与特征条件独立假设的分类方法，建立一套合理的答复意见评价模型。

关键字：群众留言　自然语言处理　文本挖掘　TF-IDF 算法　层次分析法　HNC 技术

目录

一、挖掘目标

本次建模的目标是通过研究来自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见，借助 Python、Matlab 等软件以及层次分析法达到以下目标：

（1）对指定的数据文件进行深入分析，借助 Python 软件，采取朴素贝叶斯算法和 TF-IDF 等相关算法，对数据进行去除异常数据以及分隔符等处理，并对群众留言详情进行相应的分词，提取相应的关键词，匹配与之相对应的一级分类标签。

（2）结合相关实际知识以及大多数群众对社情民意热度指标进行分析得出相对应的热度指标体系模型，采用层次分析法，对热度指标权重进行分析，利用 Matlab 层次分析技术并结合实际情况进行权重分配，进而得出热度指标模型，并对相关的留言详情匹配相对应的色度指数。

（3）对于各类社情民意，相关部门给出的答复意见，从相关性、完整性、可解释性等角度对答复意见进行分析，采用朴素贝叶斯算法和基于 HNC 句子分析得出关于答复意见的质量评价方案。

二、分析方法与过程

2.1 总体流程

（具体内容）具体建模步骤如下所示：

步骤一：针对问题，对其进行过程分析、因素分析，确定整体思路；

步骤二：采用合适的方法对抽取的数据进行预处理；

步骤三：建立数学模型进行问题求解；

步骤四：针对实验结果进行分析，最后得出结论。

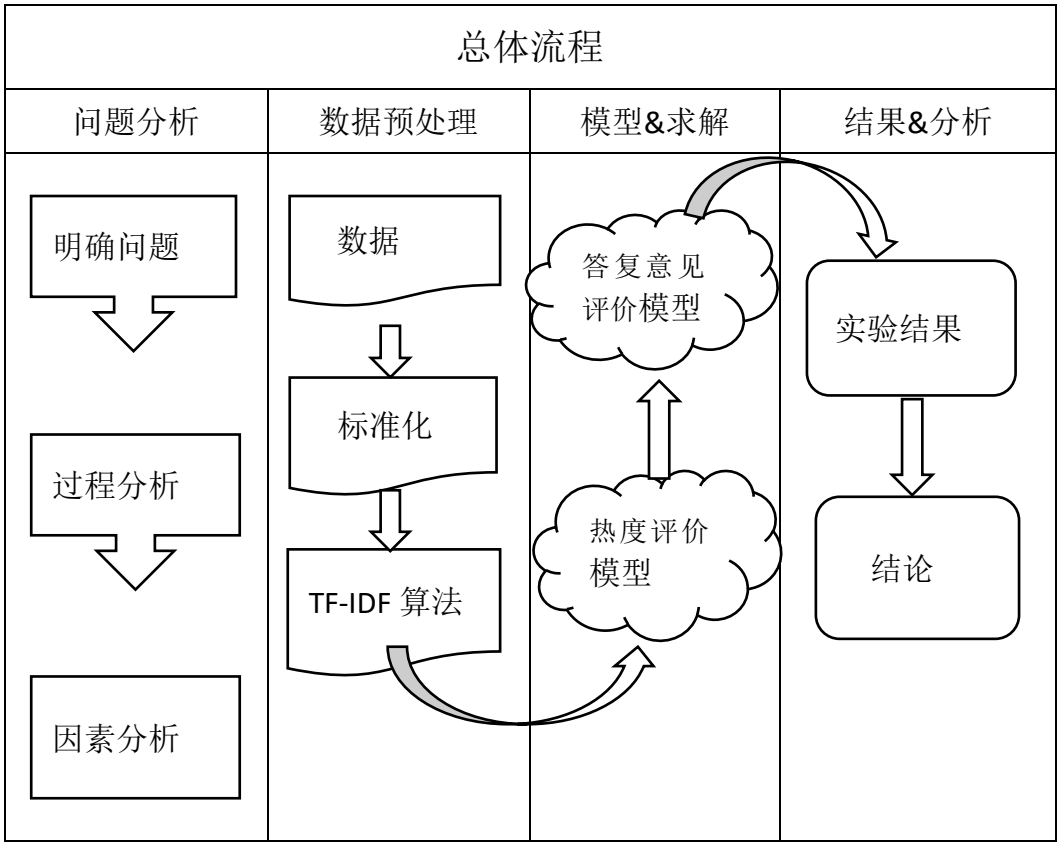

图 1 总体流程图

2.2 数据预处理

2.2.1 数据标准化

采用 Min-max 标准化方法，对数据单位进行统一的规范，便于对一级标签分类模型的计算。Min-max 标准化将数据统一映射到[0,1]区间上的数据的归一化处理的方法，公式如下：

$$y_i = \frac{x_i - \min\limits_{1 \le j \le n}\{x_j\}}{\max\limits_{1 \le j \le n}\{x_j\} - \min\limits_{1 \le j \le n}\{x_j\}} \quad (1)$$

其中，x_i为原始数据，x_j为标准化之后的数据。

2.3 社情民意一级标签分类模型

2.3.1 TF-IDF 算法

字词的重要性随着它在文件中出现的次数成正比增加，但同时会随它在语料库中出现的频率成反比下降。本次采用统计方法评估原始数据中一字词对于一个文件集与语料库中的其中一份文件重要程度，具体公式如下：

词频 TF：

公式：
$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{i,j}} \quad (2)$$

即：
$$TF_w = \frac{\text{在某类中词条 w 出现的次数}}{\text{该类中所有的词条数目}} \quad (3)$$

其中$n_{i,j}$是该词在文件中出现的次数，分母则是文件中所有词汇出现的次数总和

逆向文件频率 IDF：

公式：
$$idf_i = log\frac{|D|}{|\{j: t_i \in d_j\}|} \quad (4)$$

其中，$|D|$是语料库中的文件总数。$|\{j: t_i \in d_j\}|$表示包含词语t_i的文件数目（即$n_{i,j} \neq 0$的文件数目）。如果该词语不在语料库中，就会导致分母为 0，因此一般情况下使用 1+$|\{j: t_i \in d_j\}|$，即：

$$\text{IDF} = \log(\frac{\text{语料库的文档总数}}{\text{包含词条 w 的文档数} + 1}) \quad (5)$$

TF-IDF[8]倾向于过滤掉常见的词语，保留重要的词语。公式如下：

$$\text{TF} - \text{IDF} = \text{TF} * \text{IDF} \quad (6)$$

2.3.2 朴素贝叶斯算法

朴素贝叶斯算法[6]-[7]可应用到大规模文本集合中，具有方法简单、速度快、分类准确率高等优点。理论上，由于朴素贝叶斯算法所基于的假设太过于严格，故而其分类效果要普遍优于其他分类算法，但是在实际应用中并不能完全符合理论中的假设条件，则算法的准确率会有一定程度的下降。在类别数目较多或者类别之间相关性较小的情况下，该模型的分类性能才能达到最佳。

假设训练集中存在 j 个类别，类别集合表示为$\mathrm{C}=\{c_1, c_2, \ldots\ldots c_j\}$，文本特征词集合表示为$\mathrm{T}=\{t_1, t_2, \ldots\ldots, t_j\}$，各个文本特征对给定文本类别的影响是相互独立的。那么，类别c_i的先验概率为：

$$\mathrm{p}(c_i)=\frac{N_i}{N}, i=1,2,3, \ldots\ldots, j \tag{7}$$

其中，N_i表示属于c_i类别的文本数目，N表示训练集的文本总数。

设t_i表示文本特征集合中的第j个特征词，$\mathrm{p}(t_i|c_i)$表示特征词t_j在所有属于类别c_i的文档集中出现的概率。则未知类别文本d属于文本类别c_i的条件概率$p(\mathrm{d}|c_i)$为：

$$p(\mathrm{d}|c_i)=p\left(\left(t_1, t_2, \ldots\ldots, t_j\right)\middle|c_i\right)=\prod_{i=1}^{j} p\left(t_j|c_i\right) \tag{8}$$

根据贝叶斯定理，文本类别c_i的后验概率$\mathrm{p}(c_i|d)$

$$p(c_i|d)=\frac{\mathrm{p}(\mathrm{d}|c_i)p(c_i)}{p(d)} \tag{9}$$

$$p(d)=\sum_{i=1}^{j} p(c_i)\, p(\mathrm{d}|c_i) \tag{10}$$

其中，$p(d)$表示 d文本中所有特征词在整个文本集合中出现的概率，为常数。因此，上式简化为：

$$p(c_i|d) = \mathrm{p}(\mathrm{d}|c_i)p(c_i) \quad (11)$$

结合式（8）和（11），可得

$$p(c_i|d) = p(c_i)\prod_{i=1}^{j} p(t_j|c_i) \quad (12)$$

利用式（12）计算出的每个类别对于文档 d 的后验概率值，然后将文档 d 判定到概率值最大的那个文本类别中去

2.3.3 模型评价（F-Score）

使用 F-Score 对分类方法进行评价，公式如下：

$$\mathrm{F} = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \quad (13)$$

其中，P_i是精确率，R_i是召回率。

2.4 热度评价模型

2.4.1 热点问题

表 1-1 热点问题表

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	1	46	2019/11/02 至 2020/01/26	A 市 A2 区丽发新城	修建搅拌站，污染环境，噪音扰民
2	2	14	2019/08/18 至 2019/09/04	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民
3	3	14	2019/4/16 至 2019/11/20	A 市 A7 县金科时代	物业强制收取停车费
4	4	11	2019/1/06 至 2019/5/22	A 市 A1 区辉煌国际城	商铺非法营业
5	5	10	2019/1/06 至 2019/9/12	A 市 A3 区西湖街道茶场村	拆迁问题

表 1-2 热点问题留言明细表

问题ID	留言编号	留言用户	留言主题	留言时间	留言详细	点赞数	反对数
1	199379	A00092242	A2区丽发新城附近修建搅拌厂，严重污染环境	2019/11/25 10:17:56	A市A2区丽发新城小区近期百米范围内修建搅拌厂，严重环境污染，…	0	0
1	208714	A00042015	A2区丽发新城附近修建搅拌站，污染环境，影响生活	2020/1/20 0:00:00	尊敬的领导： 您好！作为一名居住在A2区丽发新城的业主，和小区内的每一位业主一样，…	4	0
1	213930	A909218	A2区丽发新城附近违规乱建混凝土搅拌站谁来监管？	2019/12/27 23:34:32	A市暮云街道丽发新城小区附近的搅拌站，既不符合A2区产业规划布局， …	0	0
1	215563	A909231	A2区丽发新城小区旁边的搅拌厂是否合法经营	2019-12-06 12:21:32	领导，您好！我相咨询A2区丽发新城小区旁边的搅拌厂…	0	0
…	…	…	…	…	…	…	…

从以上两个表可以看出热点问题的出现频率越高，反映的时间长度越长，则问题的热度越高；对于反映问题的地点和人群，相似的留言问题涉及的地区人群也相似；点赞数量和反对数量的多少对留言问题的热度高低有着较小的影响。

2.4.2 热度评价指标

面向各类社情民意热度评价是利用评估指标体系，通过对历史和当前数据的运算与分析，从而得出对各类社情民意热度现状的评价结果。因此，社情民意热度评价指标体系是否科学合理是至关重要的。为构建一个科学、合理和实效的评价指标体系研究社情民意热度问题，选取指标必须遵循以下原则[2]-[5]。

第一、科学性原则。选取的指标能够反映大数据环境下社情民意热度的特点，并能够正确评价其影响。

第二、可操作性原则。选取指标时，尽可能选择可以测量各类社情民意热度的定量指标，对于不能直接测量的定性指标，也要用科学的方法进行量化。

第三、全面性原则。每项指标均反映社情民意热度的某一方面情况，整个指标体系能够反映社情民意热度。

第四、层次性原则。选取的社情民意热度指标要具有层次性，指标之间相对独立且不能相互隶属，下级指标与上级指标应具有隶属关系。

基于社情民意热度作用机理及演化的规律和条件，遵循科学性、全面性、层次性以及可操作性等原则，本文从社情民意的留言数据构建了热度评价指标体系，指标体系分为三个等级，其中一级指标一个，二级指标 4 个，三级指标 8 个用于对民意热度评估。

社情民意热度评价指标如下：

一级指标	二级评价指标	三级指标
问题热度评价指标	问题频率	问题在一段时间内被反映的次数
	涉及人群地区广度	反映问题的人群
		问题涉及的人群
		问题涉及的地区
	点赞率和反对率	点赞次数
		反对次数
	时间长度	问题被反映的开始时间
		问题被反映的终止时间

（1）问题频率模型

问题频率顾名思义就是群众反映某一问题的次数在所有社情民意中的占比，关于热度指标体系，问题反映次数是量化指标的最基本要素，问题被群众反应的次数越多，则证明该问题对群众的影响较大，如果是该问题被长时间反应，则需要更加重视一点。

假设问题频率为p，群众留言总数为N，利用朴素贝叶斯算法提取群众留言的

关键词，利用留言之间两两匹配出相识度较高的留言，并归为一类，每类的问题数为n，公式如下：

$$\mathrm{p} = \frac{n}{N} \tag{14}$$

（2）涉及人群地区广度模型

涉及人群地区广度模型是指群众反应的问题所影响的人群或者地区，并且涉及人群地区的广度可分为三类：1、反映问题的人群；2、问题涉及的人群；3、问题涉及的地区。

评价模型算法相关定义：对于地区判断而言，按照中国的行政区划级别依次排下去是：第一级：省、直辖市、自治区、特别行政区第二级：地级市、自治州、旗、盟第三级：区、县级市、县第四级：乡、镇。

对于人群而言：主要利用关键词提取相关问题涉及到的人群，以提取关键字涉及到的人群计数，数字越大则涉及人群广度越大。

（3）点赞率和反对率

点赞率和反对率模型是指群众浏览相关平台，并对平台中其他群众的留言进行一种自我观点的表明的一种指标，假设点赞数为 y，反对数为n，总点赞数为Y，总反对数为 N，p_y为点赞率，p_n为反对率，p_1为该问题的点赞数与反对数的总和的占比，相应的公式如下：

$$p_y = \frac{y}{Y} \tag{15}$$

$$p_n = \frac{n}{N} \tag{16}$$

$$p_1 = \frac{y+n}{Y+N} \tag{17}$$

（4）时间长度模型

时间长度模型是描述基于贝叶斯算法提取关键词匹配后，相似问题进行归类后，对每类问题的留言时间进行提取，识别出问题被反映的开始时间和结束时间，以结束时间和开始时间的差值来作为时间长度的一个量化指标，评价模型相关算法的定义：结束时间与开始时间差值越大，则表明时间长度这个影响因素对热度指标的影响越大。

假设开始时间为st，结束时间为lt，时间长度为len，相关公式如下：

$$len = lt - st \quad (18)$$

2.4.3 构建矩阵求权重

在评价指标中，每个评价指标对于各类社情民意的热度指数结果的重要性都不同， 因此，我们需要根据不同指标对社情民意的重要程度和影响强度对其赋予权重，然后获得社情民意的相关热度指数。对社情民意热度的评价指标确定权重的方法有主观赋权评价法和客观赋权评价法两大类。权重赋值的合理性对评价结果的科学合理性起着决定性作用；若某因素的权重发生变化，将会影响整个评价结果。因此，权重的赋值需保证其科学和客观。通过对比各种方法，确定了利用发放问卷问答收集群众对不同指标的看法，并通过大量相关论文刊报总结出大量群众对各层次指标之间相对重要性的判断，根据层次分析法获得权重的评价方法。层次分析法为研究复杂的系统，提供一种新的、简洁的、实用的决策方法。其主要原理是把复杂问题分成若干层次，在每一层次逐步分析并将人们的主观判断数量化。通过加权平均的方法计算出各方案对总目标的权重值。

表 4-1~9 标度表：

标度	定义（比较因素 i 与 j）
1	因素 i 与 j 同样重要
3	因素 i 与 j 稍微重要
5	因素 i 与 j 较强重要
7	因素 i 与 j 强烈重要
9	因素 i 与 j 绝对重要
2、4、6、8	两个相邻判断因素的中间值
倒数	因素 i 与 j 比较得判断矩阵 a_{ij}，则因素 j 与 i 相比的判断为 $a_{ji}=1/a_{ij}$

利用 1~9 标度法，通过同一级别的评价指标重要程度相互比较，构建出对应的比较矩阵。比较矩阵判断矩阵如下所示：

$$A=\begin{pmatrix} 1 & 1/2 & 4 & 3 \\ 2 & 1 & 7 & 5 \\ 1/4 & 1/7 & 1 & 1/2 \\ 1/3 & 1/5 & 2 & 1 \end{pmatrix} \quad (19)$$

2.4.4 热度评价模型

根据热度评价指标及其权重，计算热度评价得分，公式如下：

$$y1 = f(x1, q) = \sum_{i=1}^{m} x1_i q_i \quad (20)$$

2.5 答复意见评价方案

2.5.1 答复意见的评价方法

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、 汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，面对各类社情民意，相关部门对部分群众留言的给予了相关的答复意见[1]。

针对群众反映的各类问题，相关部门给予相关的答复是至关重要的，并且向对应的答复内容是否能解决群众所反映的问题也是群众所关心的。

此题关于答复内容，采用朴素贝叶斯算法和基于 HNC 的句子分析，从文本的相关性、完整性、可解释性等角度对答复意见进行分析以及对群众留言详情进行文本情感分析，再通过研究各地方针对以网络留言为主体的网络问政现象和网络问政活动建立的制度化的回应和办理机制，结合两者的分析结果对答复意见的质量给出一套评价方案。

HNC 优点以及相关思路：

(1)比较分析了词语相似度计算的各种方法，实现了基于 HNC 的词语相似度计算方法，为下一步句子相似度的计算打下基础。

(2)提出了一种基于 HNC 同义词和反义词判别方法。通过词对出现的规则来判断是否是同义词反义词，由于引入了语义，简化了词语相似度的计算。

(3)在词语相似度的基础上，提出了基于 HNC 语义块的句子相似度计算方法。该方法充分考虑了语句中各个词语在语句中所处的位置以及所起的作用。

(4)把基于 HNC 语义块的句子相似度的计算方法用于相关部门答复意见和群众留言详情的句子相似度分析，通过相似度并及结合完整性以及可解释性来评价答复意见的质量。

2.5.2 答复意见的评价模型

基于 HNC 句子分析

基于 HNC 语义块的句子相似度计算（本题采用）——把句子的类型划分为作用句、过程句、转移句、效应句、关系句、状态句和判断句七大句类，每个句子是由四种主语义块和七种辅语义块构成的，根据 HNC 理论进行句子的表示和相似度计算。HNC 有自己的一套表示和计算方式，可以分析同义词，例如，句子 A：题目是对的。句子 B：题目是不错的。两个句子相似度 HNC 计算为 1。

本题所作的句子相似度研究的背景是相关部门答复内容与各类社情民意相关性，在此背景中句子相似度是一个关键的理论基础。句子相似度在主要体现在：相关部门答复内容与各类社情民意之间的相似度计算，从而得到相似度，并加以分析之间的相关性。

基于以上分析方法，为相关部门的答复意见质量建立了相关的质量评价指标：

一级评价指标	二级评价指标	指标含义
答复意见评价指标	相关性	答复意见的内容是否与问题相关
	可解释性	答复意见中的内容的相关解释
	完整性	是否满足某种规范

利用 1~9 标度法，通过对同意级别的评价指标重要程度的相互比较，构建出相应的比较矩阵，得到答复意见质量评价指标的判断矩阵为：

$$B = \begin{pmatrix} 1 & 1/2 & 3 \\ 2 & 1 & 5 \\ 1/3 & 1/5 & 1 \end{pmatrix} \qquad (21)$$

得出比较矩阵后利用规范化平均法求权重。 据答复意见评价指标的权重，计算出答复意见质量的得分得计算公式如下：

$$y = f(x, q) = \sum_{i=1}^{m} x_i q_i \quad (22)$$

其中，x 为单项指标得分，p 为单项指标对应权重值。

三、结果分析

3.1 针对问题一的结果分析

3.1.1 朴素贝叶斯分类结果与分析

本题通过对文本进行分析，用自定义字典对数据使用 jieba 进行分词，将分完词的结果用停用词表清洗停用词，然后用 TF-IDF 算法对每个留言内容进行关键词提取，最后用贝叶斯多项式模型对其进行分类 ，得到分类报告如下：

	precision	recall	f1-score	support
1	0.96	0.42	0.58	1134
2	0.06	1.00	0.12	15
3	0.00	0.00	0.00	0
4	0.73	0.95	0.82	317
5	0.96	0.61	0.74	809
6	0.07	0.91	0.13	23
7	0.03	1.00	0.05	5
accuracy			0.57	2303
macro avg	0.40	0.70	0.35	2303
weighted avg	0.91	0.57	0.66	2303

图 1 基于贝叶斯模型的分类报告

使用 F-Score 方法对模型的分类方法进行评价，得到结果为 0.34857

结论：朴素贝叶斯算法不必去除出现次数很低的词，因为出现次数很低的词的 IDF 比较大，去除后分类准确率下降，而计算时间并没有显著减少。朴素贝叶斯算法假设了数据集属性之间是相互独立的，因此算法的逻辑性十分简单，并且算法较为稳定，当数据呈现不同的特点时，朴素贝叶斯的分类性能不会有太大的差异。换句话说就是朴素贝叶斯算法的健壮性比较好，对于不同类型的数据集不

会呈现出太大的差异性。当数据集属性之间的关系相对比较独立时，朴素贝叶斯分类算法会有较好的效果。

3.2 针对问题二的结果分析

3.2.1 针对热度评价指标矩阵求权重

利用特征值法求权重，首先对列向量归一化得到特征向量 V，得到由特征值构成的对角矩阵 D。

$$V = \begin{pmatrix} -0.4657 + 0.0000i & 0.2754 + 0.4556i & 0.2754 - 0.4556i & 0.4479 + 0.0000i \\ -0.8595 + 0.0000i & 0.7860 + 0.0000i & 0.7860 + 0.0000i & -0.8903 + 0.0000i \\ -0.1090 + 0.0000i & -0.0050 - 0.1549i & -0.0050 + 0.1549i & 0.0478 + 0.0000i \\ -0.1804 + 0.0000i & -0.2612 + 0.0808i & -0.2612 - 0.0808i & -0.0662 + 0.0000i \end{pmatrix} \quad (23)$$

$$D = \begin{pmatrix} 4.0215 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\ 0.0000 + 0.0000i & -0.0056 + 0.2936i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\ 0.0000 + 0.0000i & 0.0000 + 0.0000i & -0.0056 - 0.2936i & 0.0000 + 0.0000i \\ 0.0000 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i & -0.0103 + 0.0000i \end{pmatrix} \quad (24)$$

求特征向量 V 对应的最大特征值 Max_eig：

$$Max\_eig = 4.0215$$

3.2.2 一致性检验

对构造的判断矩阵进行一致性检验，用来确定权重分配是否合理。计算一致性比例 CR，并对值进行判断：

$$CI = \frac{y_{max} - n}{n - 1} \quad (25)$$

其中y_{max}是最大特征值，n 是矩阵的维度。

$$CR = \frac{CI}{RI} \quad (26)$$

表 5：平均随机一致性指标

阶数	3	4	5	6	7	8	9	10	11	12	13	14
RI	0.58	0.89	1.12	1.26	1.36	1.41	1.46	1.49	1.52	1.54	1.56	1.58

当 CR<0.1 时，认为判断矩阵的一致性可接受范围内 CR>0.1 时，则判断矩阵不符合一致性要求，对该判断矩阵进行重新修正。

一致性指标：

$$\mathrm{CI} = 0.0072 \tag{27}$$

一致性比例：

$$\mathrm{CR} = 0.0080 \tag{28}$$

因为 CR < 0.10，所以该判断矩阵 A 的一致性可以接受!

3.2.3 指标评价权重

基于以上计算，可以得出热度评价指标权重如下表所示：

表 6 热度评价指标权重表：

一级指标	二级评价指标	三级指标	权重
问题热度评价指标	问题频率	问题在一段时间内被反映的次数	0.2884
	涉及人群地区广度	反映问题的人群	0.5323
		问题涉及的人群	
		问题涉及的地区	
	点赞率和反对率	点赞次数	0.0675
		反对次数	
	时间长度	问题被反映的开始时间	0.118
		问题被反映的终止时间	

3.3 针对问题三的结果分析

3.3.1 对利用规范列平均法求权重的结果分析

对于得到的答复意见评判指标比较判断矩阵，先对其列向量归一化得到对应的矩阵C_2， 求得：

$$C_2 = \begin{pmatrix} 0.30000 & 0.29412 & 0.33333 \\ 0.60000 & 0.58824 & 0.55556 \\ 0.10000 & 0.11765 & 0.11111 \end{pmatrix} \quad (29)$$

对C_2求其算术平均值，得到特征向量ω

$$\omega = \begin{pmatrix} -0.4629 + 0.0000i & -0.2314 + 0.4008i & -0.2314 - 0.4008i \\ -0.8711 + 0.0000i & 0.8711 + 0.0000i & 0.8711 + 0.0000i \\ -0.1640 + 0.0000i & -0.0820 - 0.1420i & -0.0820 + 0.1420i \end{pmatrix} \quad (30)$$

对应的特征值为φ：

$$\varphi = \begin{pmatrix} 3.0037 + 0.0000i & 0.0000 + 0.0000i & 0.0000 + 0.0000i \\ 0.0000 + 0.0000i & -0.0018 + 0.1053i & 0.0000 + 0.0000i \\ 0.0000 + 0.0000i & 0.0000 + 0.0000i & -0.0018 - 0.1053i \end{pmatrix} \quad (31)$$

3.3.2 对答复意见评价指标的一致性检验

对构造的判断矩阵进行一致性检验，用来确定权重分配是否合理。计算一致性比例 CR，并对值进行判断

$$\mathrm{CR} = \frac{CI}{RI} \quad (32)$$

$$\mathrm{CI} = \frac{\lambda - n}{n - 1} \quad (33)$$

对应的矩阵阶数表：

矩阵结束	1	2	3	4	5
RI	0	0	0.52	0.89	1.12

当 CR<0.1 时，则认为判断矩阵的一致性在可接受范围内，当 CR>0.1 时，则判断矩阵不符合一致性要求，对该判断矩阵进行重新修正。

计算得：

$$\mathrm{CI}_1 = 0.0018 \quad (34)$$

$$CR_1 = 0.0036 \quad (35)$$

一致性比例均小于 0.1，故权重值比较合理。

3.3.3 答复意见的指标评价权重

一级评价指标	二级评价指标	权重
答复意见评价指标	相关性	0.3090
	可解释性	0.5816
	完整性	0.1095

故通过一定的文本分析和三个指标之间的权重关系，来评价关于各类社情民意的答复意见的质量的评判。

四、结论

通过对来自互联网公开来源的群众问政留言记录及相关部门群众留言的答复意见等数据进行分类、提取和有效利用，分析出社意民意热度评价指标对各类社意民意留言问题的热度高低的影响，从而帮助相关部门及时发现热点问题并进行针对性地处理，帮助其提高服务效率。

问题热度评价指标包括问题频率指标、涉及人群地区广度指标、点赞率和反对率指标和时间长度指标。答复意见评价指标包括相关性、可解释性、完整性。

由问题热度评价指标的权重大小看来，涉及人群地区广度指标、问题频率指标、点赞率和反对率、时间长度指标这四个指标的排名可以看出其对留言问题的热度高低的影响程度，最主要的指标是涉及人群地区广度指标。由答复意见评价指标的权重大小看来，占比重最大的是可解释性指标。

参考文献

[1]贺崧. 网络问政制度化建设研究[D].吉林大学,2015.

[2]苏国强,刘芊汝,薛信朋,夏一雪.面向舆情大数据的突发事件网络民意热度评价研究[J].内江科技,2017,38(07):74-75.
[3]梁昌明,李冬强.基于新浪热门平台的微博热度评价指标体系实证研究[J].情报学报,2015,34(12):1278-1283.
[4]邓雪,李家铭,曾浩健,陈俊羊,赵俊峰.层次分析法权重计算方法分析及其应用研究[J].数学的实践与认识,2012,42(07):93-100.
[5]何跃,蔡博驰.基于因子分析法的微博热度评价模型[J].统计与决策,2016(18):52-54.
[6]李丹. 基于朴素贝叶斯方法的中文文本分类研究[D].河北大学,2011.
[7]杨鼎. 基于朴素贝叶斯的中文文本情感倾向分类研究[D].湖南工业大学,2010.
[8]张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化[J].西南科技大学学报,2020,35(01):64-69.