

“智慧政务”中的文本挖掘应用

摘要

互联网快速发展使网络问政平台愈发成为政府了解民意的重要渠道。面对群众众多政务留言，传统人工划分留言和标记热点耗时耗力且错误率高。本文主要利用自然语言处理和文本挖掘技术，对收集自互联网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进行分析。

对于问题一，本文将留言转化为字向量后，采用基于 n -gram 优化的 Fast-Text 算法对留言数据进行一级标签的划分，通过调节多个参数，设置禁用词和增加文本规则，达到好的分类效果。测试集正确率 91.4%，模型 MRR 得分 0.95， $F_1 - score$ 得分 0.91，模型质量较好。

对于问题二，实验发现使用 TF-IDF 等文本相似度算法效果不佳，本文选择了先分层再聚类的方式提取热点话题。先根据地点聚类，两两比对留言地点名称，计算地点相关性，提取相似度高的留言对，并聚类得到相同地点的留言。再根据语义内容聚类，提取留言中除地点外的关键字，计算两两语义相关性，提取相似度高的留言对，并聚类得到相同地点相似问题的留言；最后合并各地区相似问题，形成热点话题的方式。又结合留言内容涉及的话题、地域和充实性，以及问题时效性和点赞反对数给出热度计算方式，并展示了热度前 5 的热点话题。

对于问题三，本文分析了相关部门的部分答复，建立评价模型。结合留言与答复的文本挖掘，点赞数反对数、时间等信息，从相关性、完整性、可解释性、可行性、及时性几个角度评价了答复意见质量。综合考量了答复意见的表达规范、内容切题、有理有据、积极行动等方面，能较好体现答复意见的质量。

关键字： Fast-Text 模型 n -gram 热点话题 热度计算指标 答复评价模型

目录

一、挖掘目标	4
二、分析方法	4
三、问题一·群众留言的一级标签分类	5
3.1 思路与方法分析	5
3.2 数据预处理	6
3.3 Fast-Text 模型	6
3.4 Word n-gram 算法	8
3.5 模型优化	9
3.5.1 禁用词	9
3.5.2 语义规则	9
3.6 模型评价	9
3.6.1 F-score	10
3.6.2 MRR	10
3.6.3 探索发现	10
四、问题二·热点问题提取与分析	13
4.1 思路与方法分析	13
4.2 一级标签分层	14
4.3 留言地点分层	14
4.4 地点相似聚类	16
4.5 语义相似聚类	18
4.6 热度指数计算	18
4.6.1 话题领域	19
4.6.2 覆盖地域	19
4.6.3 内容充实	20
4.6.4 问题时效性	20
4.6.5 点赞和反对	21
4.6.6 热度计算	21
五、问题三·答复意见质量评价	23

5.1 思路与方法分析	23
5.2 相关性	23
5.3 完整性	24
5.4 可解释性	24
5.5 可行性	25
5.6 及时性	25
5.7 评价示例	25
六、总结与建议	28
参考文献	29

一、挖掘目标

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台的兴起，各种社情民意相关的文字数据量不断上升，电子政务平台逐渐成为了解民情民意、汇聚民智民情的重要渠道。这就给了传统的人工划分信息、整理热点的方式给我们带来了很大的挑战。与此同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政府系统的建立，已经成为社会治理的一大挑战和政府管理创新发展的新趋势。对提高政府管理水平和政府行政管理效率具有极大的推动作用。

附件中给出了从网上公开信息源收集到的留言记录，以及相关部门对部分留言的答复。我们采用自然语言处理和文本挖掘的方法来根据实际情况，建立数学模型，完成以下三项工作：

问题一·群众留言分类：建立关于留言内容的一级标签分类模型，解决人工根据经验分类工作量大、效率低，且差错率高等问题。

问题二·热点问题挖掘：及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。对某一时段内特定特点特定人群反映的问题进行归类，并建立合理的热度评价指标。

问题三·答复意见评价：建立对相关部门的留言答复评价模型，从多个角度评价答复质量。

二、分析方法

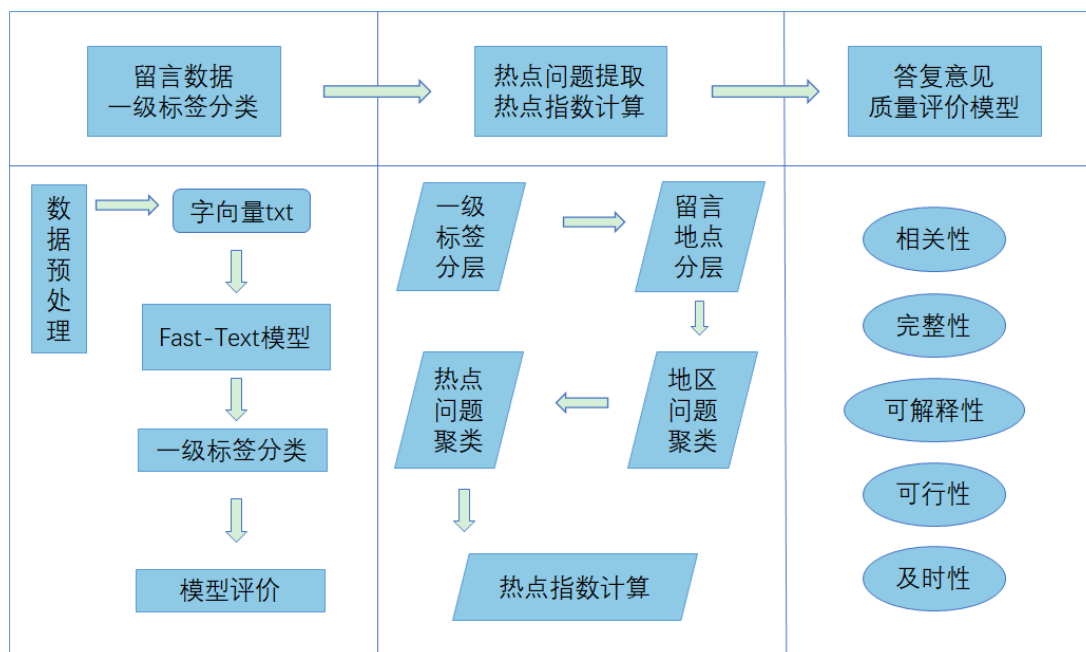


图1 全文脉络图

三、问题一·群众留言的一级标签分类

3.1 思路与方法分析

传统处理网络问政平台的群众留言依赖人工。工作人员需首先按照一定的划分体系对留言进行分类，以便后续将群众留言分派给相应的职能部门处理。这种方式存在工作量大、效率低，且差错率高等问题。本文选用 Fast-Text 模型进行文本分类，建立关于留言内容的一级标签分类模型。流程图如下：

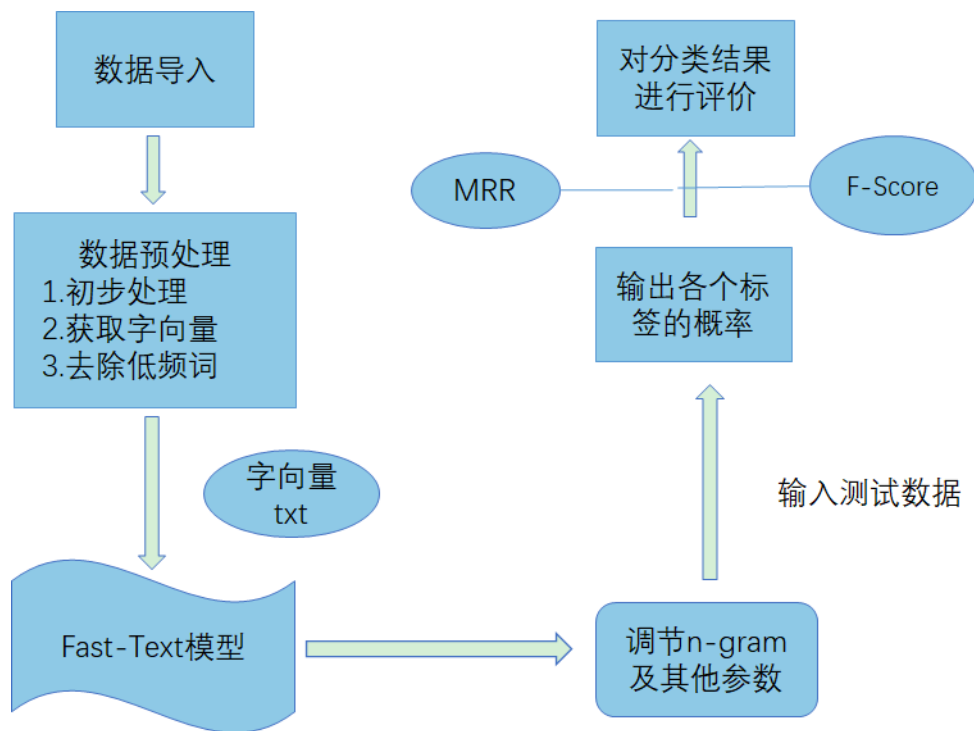


图2 思路流程图

3.2 数据预处理

step1: 初步处理

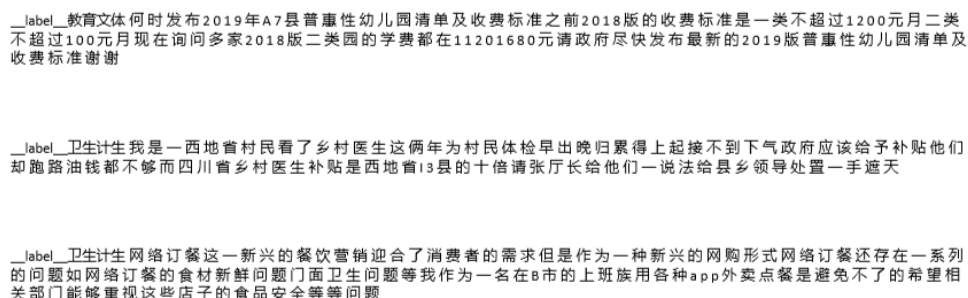
将附件 2 数据导入，数据内的“留言时间”需转换为 `datetime` 格式，以便后期处理。另发现“留言详情”中文本存在空格，将空格删去避免影响后续字向量的提取。

step2: 获取字向量

2003 年，Bengio 等提出利用神经概率语言模型训练词的分布式表示，即词向量。词向量表示了语言的深层语义，用稠密的向量解决了传统 `one-hot` 表示带来的维度灾难和词汇鸿沟问题。词向量可以获得文本中的语义信息、语句结构信息、句法信息等等丰富的内容。字作为中文文本最小组成单位，字包含了丰富的语义信息，特别是成语这样的超短文本，基于字向量进行分析能够获得更好的语义信息。因此，使用字向量来进行中文自然语言处理具有重要的实践意义，本文选取基于字向量的文本挖掘方式。

step3: 去除低频词

以留言的一级标签为 `label`，留言详情的每一个以空格隔开作为字向量，存入 `txt` 文件中。为增加关键字密度、提高分析效率，将“的”这类无意义字和出现频率小于 9 的低频字作为 `Stop Words` 不放入 `txt` 中。



__label__ 教育文体 何时发布2019年A7县普惠性幼儿园清单及收费标准之前2018版的收费标准是一类不超过1200元月二类不超过100元月现在询问多家2018版二类园的学费都在11201680元请政府尽快发布最新的2019版普惠性幼儿园清单及收费标准谢谢

__label__ 卫生计生 我是一西地省村民看了乡村医生这俩年为村民体检早出晚归累得上起接不到下气政府应该给予补贴他们却跑路油钱都不够而四川省乡村医生补贴是西地省13县的十倍请张厅长给他们一说法给县乡领导处置一手遮天

__label__ 卫生计生 网络订餐这一新兴的餐饮营销迎合了消费者的需求但是作为一种新兴的网购形式网络订餐还存在一系列的问题如网络订餐的食材新鲜问题门面卫生问题等我作为一名在8市的上班族用各种app外卖点餐是避免不了的希望相关部门能够重视这些店子的食品安全等等问题

图 3 txt 内留言格式示意

3.3 Fast-Text 模型

本文选用 `Fast-Text` 模型进行文本分类。`Fast-Text` 是一个高效的有监督文本分类模型，与 `word2vec` 中的 `CBOW` 模型有相似的架构，但 `CBOW` 预测文本中间词而 `Fast-Text` 预测标签。`CBOW` 模型将上下将上下文关系转化为多分类任务再训练逻辑回归模型，预测中心词。`Fast-Text` 模型结构如下图，其中 $x_1x_2\cdots x_{n-1}x_n$ 表示一个文本中的 `n-gram` 向量（`n-gram` 的说明见下文）。`Fast-Text` 模型用所有的 `n-gram` 为文本预测指定类别。

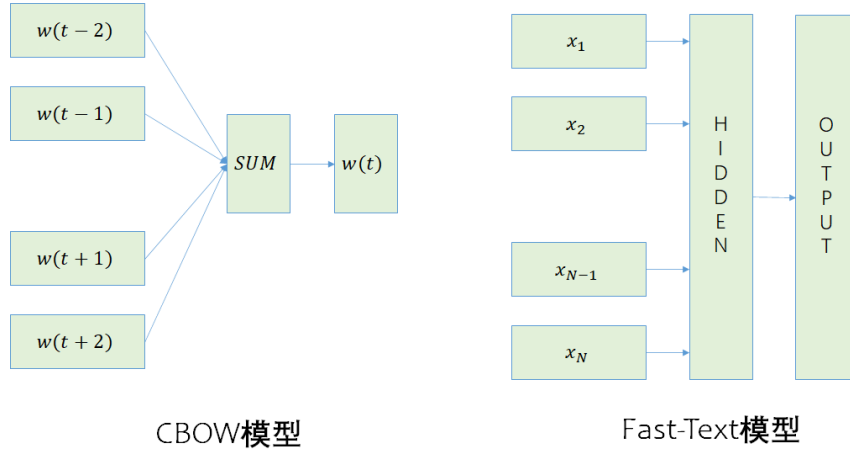


图 4 CBOW 模型与 Fast-Text 模型架构图

总结来说，相对而言 Fast-Text 有以下优点：

1. Fast-Text 在保持高精度的情况下加快了训练速度和测试速度
2. Fast-Text 不需要预训练好的字向量，Fast-Text 会自己训练字向量

Fast-Text 中的 h-softmax(hierarchical softmax) 函数根据标签和频率建立霍夫曼 (Huffman) 树，高效解决了大规模多分类问题。原理为将输入层中的词和词组构成特征向量，再将特征向量通过线性变换映射到隐藏层，隐藏层通过求解最大似然函数，然后根据每个类别的权重和模型参数构建霍夫曼树，将霍夫曼树作为输出。霍夫曼树的每一个叶子结点代表一个标签，而每个非子叶结点需要做一次二分类。

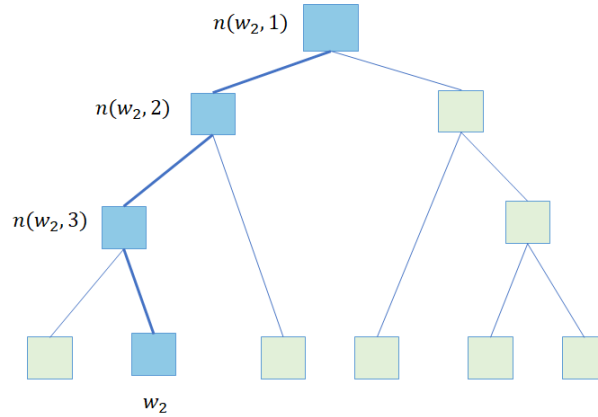


图 5 h-softmax 函数示例

非子叶结点表示为 $n(w, j)$ ，向量形式为 $\text{sigma}_n(w, j)$ ， h 表示模型隐藏层输入值，可由输入字向量得到。 $\text{sign}(w, j)$ 取 1 时表示向左，取 -1 时表示向右。

分别计算非叶子结点走左边和走右边的概率：

$$p(n, left) = \sigma(\theta_n^T \cdot h) \quad (1)$$

$$p(n, right) = 1 - \sigma(\theta_n^T \cdot h) = \sigma(-\theta_n^T \cdot h) \quad (2)$$

则以标签为 w_2 为例，概率如下：

$$p(w_2) = p(n(w_2, 1), left) \cdot p(n(w_2, 2), left) \cdot p(n(w_2, 3), right) \quad (3)$$

总结标签为 w 的概率为：

$$p(w) = \prod_{j=1}^{L(w)-1} \sigma(\text{sign}(w, j) \cdot \theta_{n(w, j)}^T h) \quad (4)$$

又有所有标签的概率之和为 1，即 $\sum_{i=1}^n p(w_i) = 1$ 。当模型是条件概率分布，损失函数可用对数函数表示，经验风险最小化等价于极大似然估计。最终得到结果：

$$\theta_j^{(new)} = \theta_j^{(old)} - \eta (\sigma(\theta_j^T h) - t_j) h \quad (5)$$

3.4 Word n-gram 算法

n-gram 算法是对 Fast-Text 模型的一个重要优化。它是一种考虑句子中字词顺序的算法，输入的是一句话——字词顺序序列，输出是这句话的概率，即这些字词的联合概率。它的第一个特点是某个词的出现依赖于其他若干个词，第二个特点是我们获得的信息越多，预测越准确。

n-gram 算法的基本思想是将文本内容按照子节顺序进行大小为 n 的窗口滑动操作，最终形成窗口为 n 的字节片段序列。直观来讲， n 越大也就是分析的词越多，获得的信息量也越大，预测应该更准确。然而当 n 变大时，会出现稀疏问题，即某些 n-gram 从未出现。通过反复测试，本文选取 5-grams 特征，以五个字为窗口去滑动。

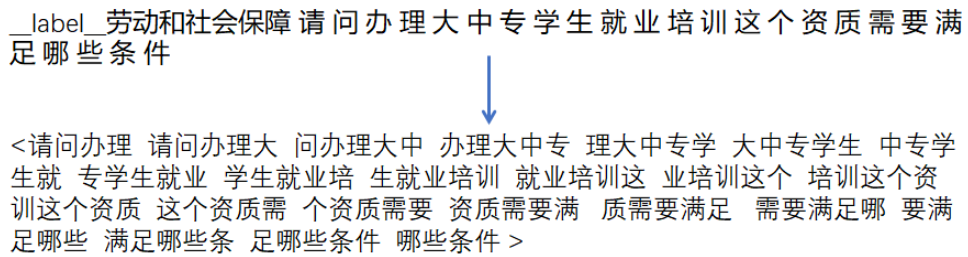


图 6 n-gram 算法示意

总结来说，n-gram 有以下优点：

1. 根据上面的字粒度的 n-gram 示例来说，即使这个词出现的次数很少，只要组成词的字和其他词有共享的部分，就可以由其他词优化生成新词
2. n-gram 可以让模型学习到字序所包含的信息，即上下文信息

3.5 模型优化

3.5.1 禁用词

留言中存在一些字会对计算机识别造成干扰，对于这些字词我们需要进行相关处理。比如”像……“；”不是……“；”违反……“之类字词，其之后的语义只是类比，可能与留言反映的问题本身没有关系。比如下例中，他是想说医疗保险问题，只是类比，实则与住房公积金并无关系，机器会误以为”劳动和社会保障“与”住房公积金“也有关系。

label_劳动和社会保障 干部职工的医疗保险费能否像住房公积金那样让人民能在网上查询知道自己的医保费到底有多少

3.5.2 语义规则

在训练模型时，观察误判文本，发现一些语义组合会对计算机产生误导。比如，有不少归入“教育文体”标签的留言内容是在讲教育缴费，内容中教育只是背景，缴费和钱提到的比较多，机器误将涉及钱的也当作“教育问题”。通过统计哪几个一级标签容易被误判成别的标签，我们总结制定一些语义规则，来帮助机器判断。

比如涉及“钱”和“教育”时，就偏向教育；只涉及“房子/建筑”时就偏向城乡建设，涉及购买房子等偏向商贸……

3.6 模型评价

对本文中的模型效果做出评价，我们采用以下几个评价指标：能体现精确率召回率的 F1-Score、MRR.

建立分类效果的混淆矩阵：

表 1 混淆矩阵

	相关 Relevant	无关 Nonrelevant
被检索到 Retrieved	TP	FP
未被检索到 Not Retrieved	FN	TN

3.6.1 F-score

精确率为所有被正确检索的占有所有实际被检索到的比例。 P_i 第 i 类的精确率，即第 i 类中正确分类的留言占实际分入第 i 类的留言的比例。计算公式为：

$$P_i = \frac{TP}{TP + FP} \quad (6)$$

召回率为所有被正确检索的占有所有应该被检索到的比例。 R_i 为第 i 类的召回率，即第 i 类中正确分类的留言占应该属于第 i 类的留言的比例。计算公式为：

$$R_i = \frac{TP}{TP + FN} \quad (7)$$

F 值为精确率和召回率的调和平均值，计算公式为：

$$F_1 = \frac{1}{n} \sum_{i=1}^n F_i = \frac{1}{n} \sum_{i=1}^n \frac{2P_i R_i}{P_i + R_i} \quad (8)$$

对本文模型进行评价，求得 F-score 约为 0.91，可以说模型质量较高。

3.6.2 MRR

MRR (Mean reciprocal rank) 是一个国际上通用的对搜索算法进行评价的机制，本文用其进行分类效果的评价。分数 $rank$ 为：对于留言的标签，第一个标签就匹配，计分数为 1，第二个匹配计分数为 $\frac{1}{2}$ ，第 n 个匹配分数为 $\frac{1}{n}$ ，如果没有匹配的句子分数为 0。最终的分数为所有得分之和。计算公式为：

$$MPR = \frac{1}{n} \sum_{i=1}^n MPR_i \quad (9)$$

$$MPR_i = \frac{1}{q} \sum_{j=1}^q \frac{1}{rank_j} \quad (10)$$

求得本文 Fast-Text 模型得分 0.95，模型质量较高。

3.6.3 探索发现

模型的效果受模型的训练数据影响，我们分析原数据集中由人工标记的标签存在一些问题，以下列举两类：

1. 关于推销虚假医疗产品划分为“商贸旅游”还是“卫生计生”，这类留言有许多则，划分却不统一。

label_卫生计生 在A9市金沙北路253号二楼有一家华晟医疗器械他们售卖的医疗床垫号称包治百病不吃药不住院不手术只需要买张两万元的床垫就能健健康康长命百岁很多老人对他们的宣传信以为真我的岳母就是其中一位认为不准她买这个床垫是天大的不孝不惜以断绝母女关系要挟请求有关部门查处华晟医疗器械这种虚假宣传坑骗消费者的行为

label_卫生计生 A9市太平桥镇新菜市场一群无良德之人以体验高端位理疗仪夸大其功效施以小礼品宣传养生保健知识推销西地省海德制药公司生产的松花粉能治百病此举实为传销受害人群都是老年人都是辛苦集攒的血汗钱棺材本此等无良之人团伙政府职能部门一定要严惩

label_商贸旅游 尊敬的杨厅长我是一名广播听众投诉举报935和1028频道肆意播放虚假医药广告明显的把保健品当药品宣传打着健康快车养生节目的名义招揽演员当听众欺骗老人上当购买对社会产生巨大的危害连小学生都知道是播的广告为何广播电台就看不出来呢广电厅的监管职能是如何体现的呢我等待您的回复

图 7

2. 关于职工工资所属标签的划分也存在问题。按照附件一中的标签，仅在”教育文体“一级标签下有，”教育文体——教师队伍和待遇——工资福利“这一条，可认为教师的工资方案划分到”教育文体“一栏。而在”卫生计生“和”城乡建设“中，并未涉及相关岗位职工工资，应该归入”劳动和社会保障“

label_卫生计生 领导好最近我发现西地省中医院ICU的护士很辛苦本来按照规定是一名护士管3位病人但是现在两名护士管着十几位病人每天上班都是超负荷工作都是在拿生命上班都是些年轻的妹子每天这么上班真的会累坏去工作又累现在工资方案还改了工资也少了【劳动与社会保障】

label_劳动和社会保障 如此改革令人心寒关于本次工资调整的几点看法最近G市教师工资进行了调整捂了十五个月之久的教师工资方案终于犹抱琵琶半遮面地与老师们见面了顿时在老师中掀起了轩然大波堪称史上差别最悬殊的工资方案令一线教师无比震惊职称的差别直接导致月工资差……【教育文体】

label_城乡建设 胡书记您好感谢您百忙之中查看这份留言我的父亲51在A6区金星北路明发国际工地工作57在工地进行施工时发生泥土塌方受伤至今仍在治疗期间工地现在拒绝支付医疗费用并且态度恶劣工地的安全生产问题得深究另我为我的父亲维权也为民工维权希望民工在工地工作能够得到应有的保护希望能够得到政府的帮助感谢【劳动与社会保障】

图 8

另外，还有一些存在争议的标签，比如”物流城存在政府渎职问题“和”物流（交通运输）“没有关系；”冒用他人身份骗取资格证“和”城乡建设“关系也不大，只是举报人是个建筑业从业者……

label 商贸旅游 我想问 特种设备安装维修许可证 电梯方面的在哪里办理需要准备哪些提交的材料

label 城乡建设 我叫涂愈是一名普通的建筑业从业者在这里求助也请求相关部门调查西地省建望集团及西地省辉东安建工程有限公司涉黑涉恶的违法行为事情是这样的2012年我通过了二级建造师执业资格考试并在2013年取得二级建造师……注册行为西地省辉东安建工程有限公司或存在以下违法行为1伪造聘用劳动合同违法行为2伪造他人签名骗取国家机关违规注册行为3冒用身份信息行为4挂证违法违规行为

图 9

这些问题也说明了传统依靠人工根据经验处理的电子政务系统，在标签上存在差错率高等问题。

四、问题二·热点问题提取与分析

4.1 思路与方法分析

对留言进行标签分类只是留言处理的基础工作，对留言反应的问题进行分析，及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。

热点问题就是某一段时间内群众集中反映的某一问题。本文先分层次划分留言以缩小处理量，再提取高相关性的留言。尝试发现 TF-IDF 等常规相似度算法在本题中效果不佳。本文选择了先根据地点聚类，得到相同地点的留言；再根据语义内容聚类，得到相同地点相似问题的留言；再合并各地区相似问题，形成热点话题的方式。最后依据关注度和话题内容等建立热度评价指标并给出排名前 5 的热点问题。

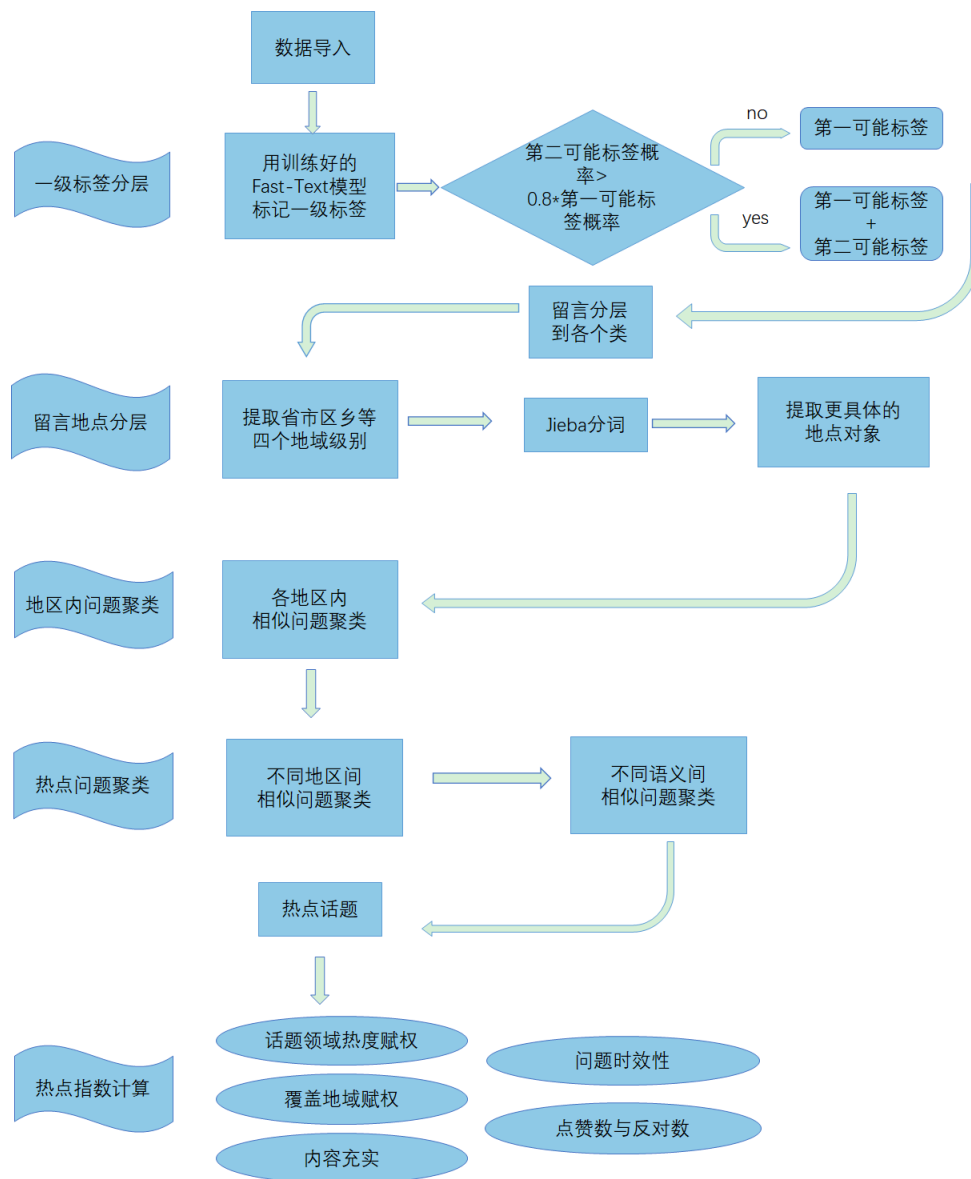


图 10 思路流程图

4.2 一级标签分层

直接用所有留言信息计算相关性来找到热点话题的运算量无疑是非常大的。热点话题是某一时段内反映特定地点或特定人群的话题，这说明属于同一个热点话题的留言在地点、人群、内容上都具有相似性。本文先采取分层缩小留言量，再聚类提取热点问题的方式。

首先使用前文的 Fast-Text 模型对留言文本进行一级标签的分类。按照一级标签构建的思路，同一标签的留言内容都是属于同一类，更具有相似性，即更可能反应的是相似的问题。故认为热点话题内的留言应当属于同一个一级标签。

在原先确定留言的一级标签时，实则计算了该留言匹配各个标签的概率，分层中为了避免一级标签分类错误，当第二可能的一级标签概率大于 0.8 倍第一可能的一级标签，我们将第二可能的一级标签也列入考虑范围。

4.3 留言地点分层

根据一级标签的分层结果，对每一个类内的留言文本内的地点继续提取。按照地区对留言再次分层，对小区域内留言进行聚类 and 关键词提取。获得小区域内的话题聚类，为热点话题的聚类提供更易计算的数据。

首先观察留言数据，推测各级区域脱敏的命名规则如下：

表 2 各级区域命名规则

省	西地省
市	大写字母 + 市 (ed.A 市)
区、县	大写字母 + 阿拉伯数字 + 区/县 (ed.A1 区)
乡、镇	实名

按照上述思路，分析留言时发现数据中存在一些脱敏时地名错误，见下图。A 市的“C 市路”，“F 市朝阳烧烤”，“五矿万境 K9 县”这些属于路名、店名、小区名中包含字词与其他市区镇的名字，在脱敏处理时被一起替换了。在后续的地点分层时要考虑到这类问题。

胡局长您好，A市C5市路由南往北方向早上经常堵车……
A2区F市朝阳烧烤龙湾店老板打伤我，还在逍遥法外。
A市五矿万境K9县负一楼面积缩水。
关于五矿万境K9县负一楼面积缩水的反馈。我是有该楼盘的业主，……

图 11

进一步提取更具体的地点，采取的方式为，对地域标志字“省”“市”“区”等字后至多 20 个字符放入 list 内进行分析（遇到句中句末标点则提前停止）。具体方式如下：

首先依据日常表达习惯认为紧跟着的双引号内的文字即为地点主体。另外留言文本中许多信息做了脱敏处理，其中部分会影响地点提取。如“地铁 2#”可知是“地铁 2 号线”的意思，在此对其进行还原。

大部分的地点提取采取此种词性分析的方式。首先使用 python 的 jieba 进行分词，遍历 list 进行词性分析。

经过多次测试，发现以 ['小区','小学','村','公寓','公司','公园'] 结尾地点较多，提取时效果也比较好。而“城”，“路”则非地名词过多，会造成干扰。以 ['小区','小学','村','公寓','公司','公园'] 为标志，将前面的字提取出来组合作为地点。

如果没有这几个关键词，则遍历由地域标志字“省”“市”“区”等后的 20 字内组成的 list，遇到词性为 n（名词）有关的字词，就放入一个字符串。为提高正确率，又引入 j(简称略语)、m(数词)、s(处所词)、l(习用语)、f(方位词)、b(区别词)。

将得到的可能地名放入到 set(集合中)，因为集合具有不重复性。对每一个词计次，选出出现次数最高的，如果有次数并列则选字符串长的作为最后的地点名程。

4.4 地点相似聚类

接下来，对分层后的数据根据地点相似和语义相似分步进行聚类。

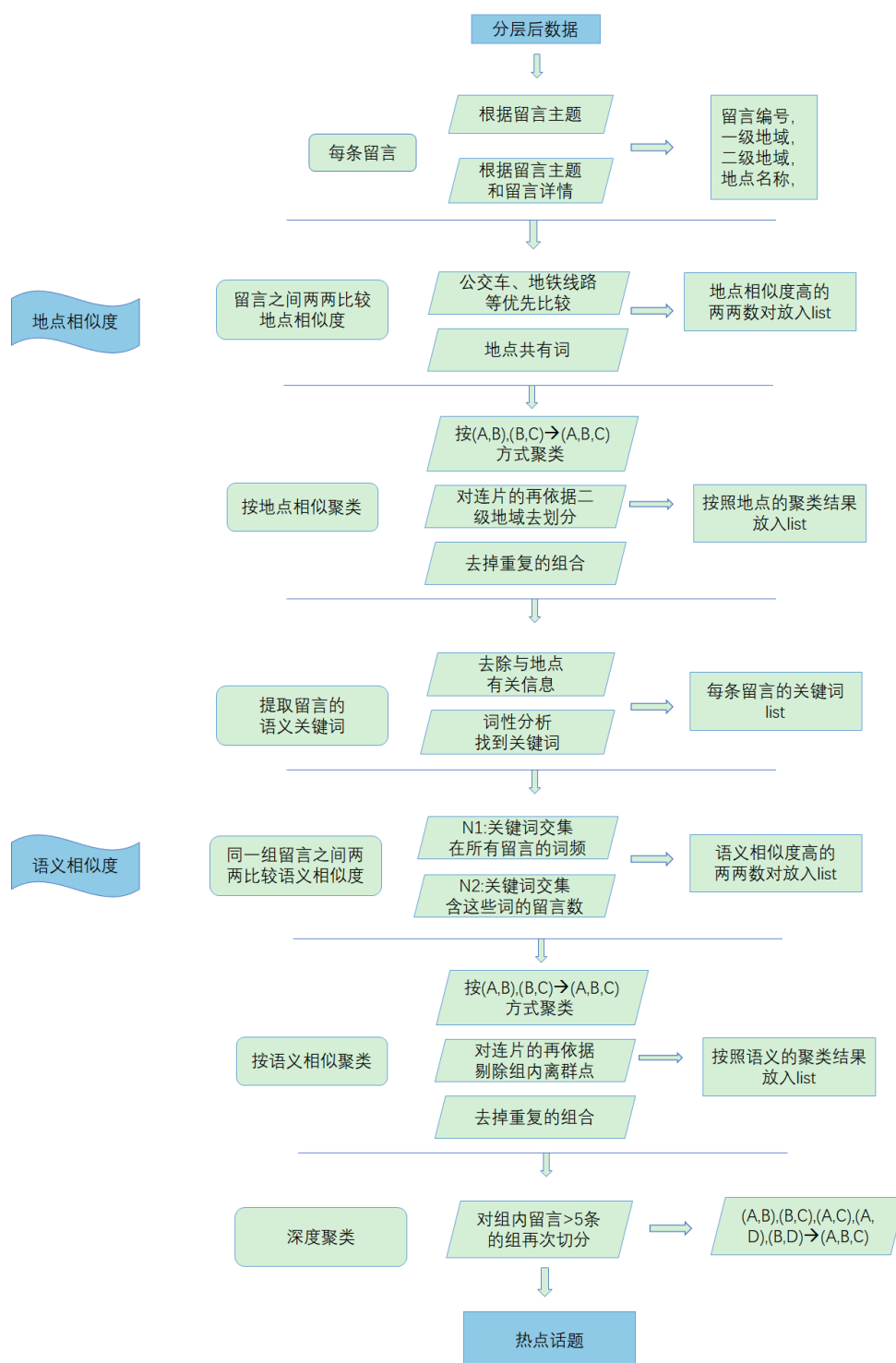


图 12 思路流程图

按照前面两次分层处理之后，每一个一级标签内的留言，应该包含一级地域、二级地域、地点名称等信息，考虑到留言主题和留言详情有一定差异，故分别提取留言主题

和留言详情加主题的信息，具体如下：

按照留言主题得到的 [一级地域，二级地域，地点名称，留言编号]

按照留言主题和留言详情得到的 [一级地域，二级地域，地点名称，留言编号]

对同一标签内所有留言，两两比较地点相似度对公交车、地铁路线优先区分，出现了公交车和地铁路线却数字不一致的直接将地点相似度记作 0。没有提到这些明确的数字时，通过统计两条留言的地点名称共同用字来计算。相关性为公用字占到较长一条留言地点名称的总字数比例，且当公用字有 3 个字连续，则再相关性结果上乘 1.2，有 4 个连续，则直接认为相关性为 1。见下例：

eg1: 留言1地点名称：劳动东路魅力之城小区 留言2地点名称：魅力之城 共同字数：4 留言1地点字数：10 留言2地点字数：4 则相关性=4/10=0.4 但这4个字连续，可直接认为相关性=1	eg2: 留言1地点名称：亚厦风荷园 留言2地点名称：风荷园小区 共同字数：3 留言1地点字数：5 留言2地点字数：5 则相关性=3/5=0.6 但这3个字连续，相关性=0.6*1.2=0.72
--	--

图 13 地点相关性计算示例

实验中发现例 2 这种“3+2”（风荷园 + 小区）的组合比较多，所以相关性的阈值取 0.72，实验中也证明了取 0.72 效果最佳。当相似度高于阈值，则认为两个地点名词代表的同一个地点。将这两个留言的留言编号以数对形式放入列表。按照以下规则进行合并：

$$\begin{aligned}(A,B),(B,C) &\rightarrow (A,B,C) \\ (A,B),(B,C),(A,D) &\rightarrow (A,B,C,D)\end{aligned}$$

此步之后可以得到大致分组。但由于只要与某一组中的一个留言有较高相关性，就会被放入该组。可能它与留言 A 的部分地点相似，A 与组内其他留言相似的却是其余部分，因此事实上这个留言不应该被放入该组。故对留言数很多（本文选取大于 5 条）的组，进行再一次切分检查。

提取这些留言的二级地域，计数并排序，将这些留言按二级地域分组。如果留言内不包含二级地域，则查看之前的相关性数对列表，该留言是否有直接相关留言在某个组内，有则将该留言划分入那一组。分组完毕后，去掉离群的留言，输出这些组。对于一个一级标签内所有组，去掉重复的组合。得到按地点相关性聚类的结果。

4.5 语义相似聚类

进一步进行语义聚类，以得到相同地点相似问题的留言话题。

首先类似前面地点相关性分析时的地点名词，此步需要先获取每个留言的语义关键词。首先因为目前得到的组内留言在地点上都比较相似，将所有一级地域、二级地域、地点名词去除以避免干扰相似度的计算。对剩下的部分进行词性分析，去除['x','eng','c','uj','ul','d','p','r','m','y','k','u']这些在句子中不太重要的成分，剩下的放入一个 list。

两两判断留言语义相关性，通过计算两个分值。

```
m1: 关键词得分1为: 找到关键词的交集, 查找在所有留言中的词频, 取倒数后相加
另外对两个留言内的关键词, 分别查找在所有留言中的词频, 取倒数后相加, 得到分数1和分数2
m1=关键词得分1/max (分数1,分数2)

m2: 关键词得分2为: 找到关键词的交集, 查找包含关键词的留言条数, 取倒数后相加
另外对两个留言内的关键词, 分别查找包含关键词的留言条数, 取倒数后相加, 得到分数1和分数2
m2=关键词得分2/max (分数3,分数4)

相似度得分= m1* m2*1000
```

图 14 语义相关性计算示例

认为相似度大于 0.03 则具有相关性，相似度高的留言两两以数对形式放入列表。用类似地点相关性聚类提到的规则进行聚类。得到多个组。每个留言在组内和一个以上（不包括一个）留言在语义上有相关性即可，如不满足则需剔除。去除重复的组合。

对于组内留言数大于 5（不包括 5）的组合，进行深度聚类。规则如下：

$$(A,B),(B,C),(A,C),(A,D),(C,D),(E,F),(E,C),(F,B) \rightarrow (A,B,C)$$

拆分完成后，得到了热点问题。

4.6 热度指数计算

一个话题的热度与该话题本身涉及内容和话题的关注度都有关系。本文对于计算问题的热点指数主要考虑了以下几个方面。

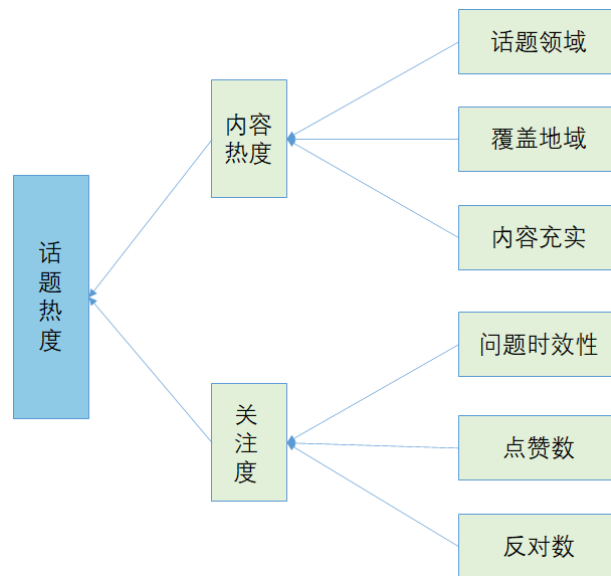


图 15 话题热度指数

4.6.1 话题领域

话题的热度首先与它的内容有关，不同类别的话题受用户的关注度也不同。参考了 2019 不同领域的电子新闻浏览量统计等资料，本文为 15 个一级标签的留言内容制定了内容热领域度赋权系数，以区分不同类别话题的易带来的热度。

表 3 内容领域热度赋权系数

一级标签	内容领域热度赋权系数
交通运输\城乡建设\劳动和社会保障	1.4
党务政务\纪检监察\政法\民政	1.2
科技与信息产业\教育文体\经济管理\商贸旅游	0.8
国土资源\环境保护\农村农业\卫生计生	0.6

4.6.2 覆盖地域

一个话题的热度还与它的地域级别有关，如果是整个省各地都面临类似问题，热度应该高于一个城市内的同样问题。所以在这里给热点话题设计的地域级别也赋权。通过统计话题内留言的地点，找到覆盖所有地点的地域级别，确定该话题的覆盖地域等级。

表 4 覆盖地域赋权

地域级别	覆盖地域赋权系数
省级	1.4
市级	1.2
区、县级	0.8
乡、镇、街道级	0.6

4.6.3 内容充实

留言内容的丰富性在一定程度上可以表征该问题的复杂性和重要性，但为了排除有些留言者的重复类似语句的问题，我们以涉及的字符多样性为依据评判内容的充实度。具体做法为计算留言内容中非重复字数，并依次确定权重系数：

表 5 内容充实性赋权

字数	内容充实性赋权系数
100 个以下	0.8
101 到 200 个	1
201 到 300 个	1.2
300 个以上	1.4

4.6.4 问题时效性

一个话题内的留言数量能体现该话题受群众的关注度，但考虑到热点话题往往具有时效性，随时间流逝，话题热度会衰退。以话题内留言密度最大值为峰值，峰值处距今约越近，时效性越强，热度也就越高。

留言密度定义为一个时间点前后 n 天内的留言数量和。取留言密度最大值处为峰，如有并列，则取距离当前时间近的时间点。

根据峰距今的天数来确定权重系数：

表 6 问题时效性赋权

峰距今时间	问题时效性赋权系数
1 个月内	1.4
3 个月内	1.2
6 个月内	1
1 年之内	0.8
1 年之外	0.6

4.6.5 点赞和反对

留言的互动关注度可以通过点赞、反对数反映。在计算关注度时，反对数和点赞数应该相加作为关注数，因为无论是赞成还是反对都是对这个话题的关注，有些话题越是存在争议越说明热。

根据关注数来确定权重系数：

表 7 关注度赋权

关注数	关注度赋权系数
0 条	1
1 到 10 条	1.2
11 到 50 条	1.4
51 条以上	1.6

4.6.6 热度计算

综合以上几个指标，得到热点话题热度计算公式如下：

$$\begin{aligned}
 \text{话题热度} &= \text{内容热度得分} * \text{关注度得分} \\
 \text{内容热度得分} &= 40 * \text{内容领域热度赋权系数} + 30 * \text{覆盖地域赋} \\
 &\quad + 20 * \text{问题时效性赋权系数} + 10 * \text{内容充实性赋权系数} \\
 \text{关注度得分} &= \sum \text{关注度权重系数}
 \end{aligned}$$

以下为热度指数前五的话题：

五、问题三·答复意见质量评价

5.1 思路与方法分析

政务系统分析留言反应的问题，最终还是要去答复解决这些问题。为合理评价相关部门的答复质量，本文从相关性、完整性、可解释性、可行性、及时性多个维度建立评价方案。

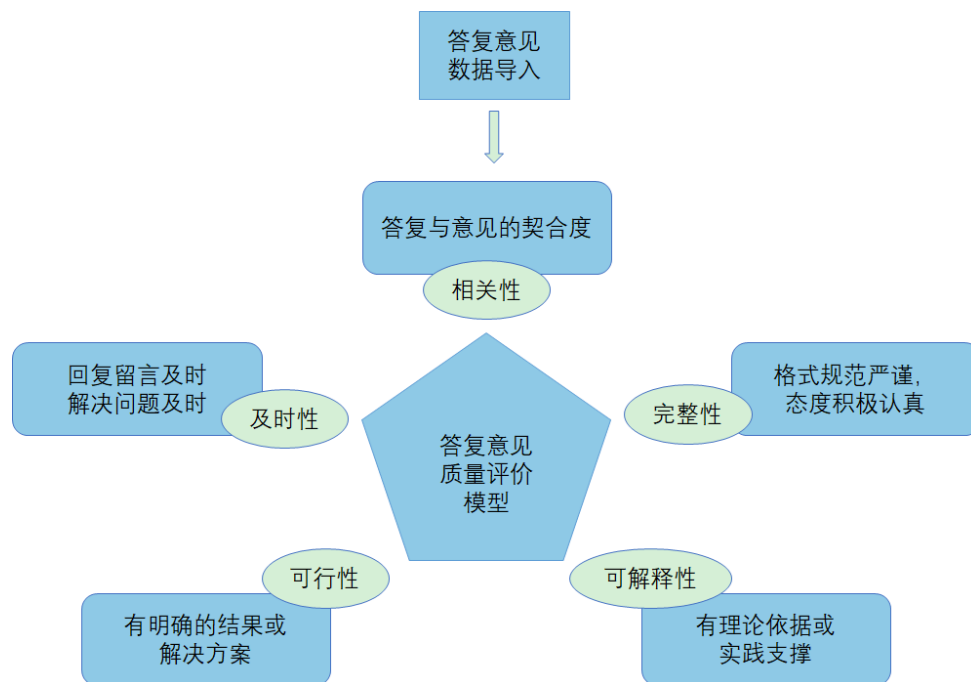


图 16 思路流程图

5.2 相关性

相关性表现为相关部门答复是否高度契合留言反应的问题。主要通过分别提取留言问题和答复意见中的关键词，进行相关性的计算。相关性高，说明部门的答复契合度高，有在真正回答且回答了留言的复的主体内容，不可只有空话；第四，在答复意见中，提供联系电话或地址以供提问群众进一步了解问题解决方案也凸显了积极的工作态度；第五，答复意见需具有规范的结束语，如“感谢您的关注和支持”。所有问题；相反，相关度低则表示可能存在答非所问的情况。

【留言】

‘反映A市医学院污水直排农田，造成几百亩不能耕种。书记您好，我是A6区大泽湖东马社区一名农民，反映A市医学院污水直排农田，造成几百亩不能耕种’\xa’反映A市医学院污水直排农田，造成几百亩不能耕种。书记您好，我是A6区大泽湖东马社区一名农民，反映A市医学院污水直排农田，造成几百亩不能耕种’\xa e，深井水不能饮用，给农民造成了无法估计的损失。反映到有关部门，解决方案是搞一条污水管道，请问农民的田塘坝深井水怎么解决，农民生活怎样解决，政府一字不提农民补偿，污水管道能解决农民生存还是生活。是浪费钱财，以后农民的田，怎么办没有了水源。靠污水耕种土地不是上策，没有水是万万不能。望领导重视。’

【答复】

‘\u3eee\u3eee感谢您对我们工作的关心、监督与支持。’

图 17 相关性负面例子

5.3 完整性

答复意见的完整性是指答复意见是否符合某种规范。具体来说，第一，答复意见需具有规范的开头（包括问候和收悉）；第二，答复意见需具有回答开始的标志，如”现答复如下:”；第三，答复意见需具有答完整性指标高体现的是部门严谨又积极的工作态度，是一篇答复意见需要达到的基本要求。

网友“A000110735”: 您好! 您在平台《问政西地省》上的留言已收悉, 市住建局及时将您反映的问题交由市房屋交易管理中心办理。现将相关情况回复如下: 按照《A市人才购房及购房补贴实施办法(试行)》第七条规定: 新落户并在A市域内工作的全日制博士、硕士生(不含机关事业单位在编人员), 年龄35周岁以下(含), 首次购房后, 可分别申请6万元、3万元的购房补贴。 “首次购房”是指在A市限购区域内首次购买商品住房(含住宅类公寓)。因此, 如购买商业性质公寓(非商品住房), 则不可申领购房补贴。以上情况, 望您知晓和理解。如您还有疑问, 建议可拨打市房屋交易管理中心咨询电话0000-00000000详询。特此回复! 2019年4月30日

网友“A000100804”: 您好! 针对您反映A3区教师村小区盼望早日安装电梯的问题,A3区住建局高度重视, 立即组织精干力量调查处理, 现回复如下: 为了完善住宅使用功能, 提高我区既有多层住宅居民的宜居水平, 2018年6月7日, A市A3区人民政府办公室下发了《关于A市A3区既有多层住宅增设电梯实施方案》的通知。该方案明确了增设电梯条件及流程。详情可咨询政务服务中心住建局受理申请老旧小区加装电梯窗口, 咨询电话: 0000-00000000。感谢您对我们工作的关心、监督与支持。2019年4月2日

图 18 完整性正面例子

5.4 可解释性

可解释性主要指答复意见是否有理有据，即意见是否有理论支持，引用政策法规条例来说明；或是有现实依据，依据在实际环境中的实践结果；又或是语句结构分明，逻辑清晰。

高解释性的答复更能让群众理解和接受，对解决问题非常重要。

【现实依据】网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题，A3区洋湖街道高度重视，立即组织精干力量调查处理，现回复如下：您反映的为潇楚大道西线道路工程项目，该项目位于坪塘老集镇，**目前正在**进行土方及排水施工。**因**该项目为城市次大道，设计标准高，该段原路基土质较差，需整体换填，**且**换填后还有三趟雨污水管道施工，施工难度较大，周期较长。**加之**坪塘集镇原有管线、排水渠道较多，**需**先处理管线和渠道才能进行道路施工，**且**因近期持续雨天，**为**保证道路施工质量，**需**在晴好天气才能正常施工。**目前**该项目已完成75土方及50排水，预计今年8月底将完工通车。感谢您对我们工作的关心、监督与支持。2019年4月29日

【理论依据】网友“A000100804”：您好！针对您反映A3区教师村小区盼望早日安装电梯的问题，A3区住建局高度重视，立即组织精干力量调查处理，现回复如下：为了完善住宅使用功能，提高我区既有多层住宅居民的宜居水平，2018年6月7日，A市A3区人民政府办公室下发了《**关于A市A3区既有多层住宅增设电梯实施方案**》的通知。**该方案明确了**增设电梯条件及流程。详情可咨询政务服务中心住建局受理申请老旧小区加装电梯窗口，咨询电话：0000-00000000。感谢您对我们工作的关心、监督与支持。2019年4月2日

图 19 可解释性正面例子

5.5 可行性

相关部门给出的答复措施是否可行机器是难以判断的，但最起码答复中需要给出结果或是解决方案。本文评价方案中的可行性就是可以根据留言是否给出明确的结果或方案，以及给出的方案的详细程度来评价。

5.6 及时性

一份好的答复必然是及时的，故计算答复意见的时间与留言发布的时间差可以反应相关部分回应的速度。而如果同样的问题被多次反应，又体现了相关部门的办事速度。这两个速度都很快才说明解决问题及时。

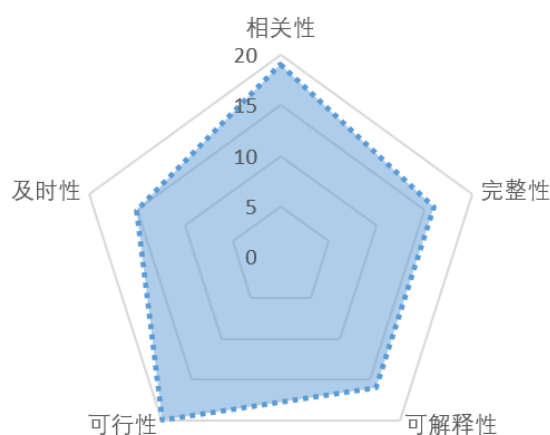
5.7 评价示例

6290 例答复五项得分分别为 [20,16,16,20,15]，总分 87。该答复在各项指标的表现都很好，惟因为回复时间稍慢，在及时性这一项上稍低。可以看到答复格式上，开头结尾完整，有使用敬辞等；内容中，切合问题，且包含了大量理论和现实依据，最终也给出了方案建议。

网友“UUee81e32”您好！您的留言已收悉。现将有关情况回复如下：经核实，A5区A5区亭街道于2016年5月收到新星小区业主递交的《新星小区成立业主大会申请人签名表》，街道社会事务办（物管办）按流程向新星社区移交《新星小区成立业主大会申请人签名表》。新星社区在收到签名表当日，制定了《关于新星小区成立业主大会申请人名单初审方案》。2016年6月30日要求原项目开发商提供了业委会成立的部分相关资料。2016年7月5日，在街道社会事务办负责人指下组建了由街道社会事务办（物管办）、社区及原项目开发商工作人员组成的初审小组，召开了审核工作调度说明会。初审小组于2016年7月9日对新星小区成立业主大会申请人签名表初审工作推进进行了小结。因部分业主对初审小组中开发商介入存有疑议，街道于2016年7月20日左右对业主签名表再次审核。2016年7月29日，街道将情况汇总后将审核结果告知递交签名表的业主。在筹备的过程中，街道于2016年8月24日收到部分业主提交的《告知确认函》，确认函表示“鉴于新星小区业委会前期筹备工作进展推进情况，为防止小区部分业主出于不明目的，违规利用原热心业主签名资料，另行组建新星小区业委会筹备组，造成违原80户签名业主的意愿，导致小区不稳定，小区业委会发起筹备的热心业主及80户广大业主强烈要求，如要启动新星筹备组工作，需筹备原发起人中至少5人联名签名同意提出。”在此之后，街道及社区再未收到任何反馈信息。一直以来，街道及社区对新星小区成立业委会予以支持，在成立业委会的过程中，按照相关政策及文件规定，街道及社区仅承担指导和协助工作。如需街道及社区的指导和协助，建议小区热心业主向A5区亭街道社会事务办（物管办）提交书面申请，共同推进业委会成立的相关工作。感谢您对我们工作的支持、理解与监督！2018年4月19日’

图 20 高分答复意见例子：6290

编号6290留言回复评价

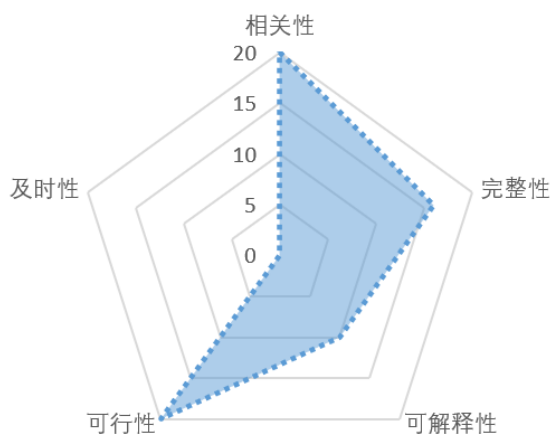


6309 例答复五项得分分别为 [20,16,10,20,0]，总分 66。这条答复比较特殊，它的答复时间与留言时间隔了三年多，及时性为 0。虽然它在其他几项指标中表现不错，但再好的答复不具有及时性一样不能解决群众问题。

网友“UUee81727”您好！您的留言已收悉。现将有关情况回复如下：据查，A6区已多渠道收到有关勤达诚境界城的各类投诉，区委、区政府高度重视，多次组织相关职能部门和开发商召开专题会议协调有关问题。4月17日，区发改、住保、城管、市场等相关部门组成房地产市场专项执法小组赴项目现场展开调查。4月至5月，区住房保障局多次约谈项目负责人，督促做好问题整改，依法依规进行楼盘销售。关于贴瓷砖的问题。经核实，项目合同中并未明确约定在楼栋大厅贴瓷砖为交房标准。经区政府及相关部门多方协调，项目方承诺将征询业主意见后决定是否加贴瓷砖；5月12日，项目方与业主代表达成一致意见，在项目标准层层间厅贴瓷砖。关于虚假宣传的问题。2018年5月3日，区市场局赴项目进行核实，根据现场检查情况，项目方在售楼过程中存在涉嫌欺骗误导消费者的情节。区市场局已于5月14日立案调查，将进一步调查核实并依法依规处理。关于“认筹金”的问题。经调查，勤达诚境界城开发商委托A市尚同家有房地产经纪有限公司进行销售服务。房产销售过程，部分业主与尚同家有公司签订书面协议，约定向尚同家有公司缴纳会员费后可享受看房、选房等服务以及一定房款的优惠，协议中约定，客户成功购房后，已收取的会员费不再退回。下一步，A6区住房保障局及有关部门将密切关注勤达诚境界城项目问题的进展情况，做好沟通协调等工作，力争纠纷得到协商和平解决。感谢您对我们工作的支持、理解与监督！2018年5月24日’

图 21 低分答复意见例子：6309

编号6290留言回复评价



六、总结与建议

随着互联网快速发展，人们群众通过各网络问政平台发布社情民意相关的相关文本不断增多，这给传统依靠人工进行的留言划分和热点整理工作带来挑战。人工划分不仅耗时耗力，在分析中我们也看到部分留言存在划分争议，以及部分错误划分。

同时，大数据、人工智能等技术也快速发展，建立基于自然语言处理技术的智慧政务系统已是社会治理创新发展的新趋势。本文主要利用自然语言处理和文本挖掘技术，对收集自互联网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进行分析。

本文采用基于 n -gram 优化的 Fast-Text 算法对留言数据进行一级标签的划分，通过设置停用词和增加文本规则，达到好的的分类效果，能够解决人工划分存在的耗时耗力错误率高等问题。划分问题可以帮助相关部门针对性处理，提取热点问题并计算热度则可以帮助政府部门认识到最“热”最迫切的民意。使用 TF-IDF 等文本相似度算法效果不佳，本文选择了先分层再聚类的方式，根据文本的地点、语义来提取热点问题。又结合留言内容和关注度给出热度计算方式，并展示了热度前 5 的热点话题。最后分析相关部门的部分答复，从相关性、完整性、可解释性、可行性、及时性几个角度评价了答复意见质量。

网络问政平台愈发成为政府了解民意的重要渠道，借助计算机技术实现“智慧政务”，对提升政府的管理水平和施政效率具有极大的推动作用。

参考文献

- [1] 何颖刚, 王宇. 一种基于字向量和 LSTM 的句子相似度计算方法 [J]. 长江大学学报 (自然科学版).2019,16(1)88 94
- [2] 陈永洲, 马静. 融合多粒度信息的文本向量表示模型 [J]. 数据分析与知识发现.2019 年第 9 期
- [3] 代令令. 基于 fastText 的中文文本分类 [J]. 计算机与现代化.2018 年第 5 期
- [4] 谢成东. 基于 SPARK 的中文文本特征提取及分类方法研究与实现 [D]. 电子科技大学,2017.6
- [5] 杨丹浩. 吴岳辛. 范春晓. 一种基于注意力机制的中文短文本关键词提取模型 [J].Computer Science 计算机科学 Vol.47,No.1,Jan.2020
- [6] 头条指数. 数说政务头条号.[2018.11.27].
<https://dydata.io/datastore/detail/1679040996693905408/>
- [7] 极光.2019 年新闻资讯行业研究报告.[2020.03.09].
<https://dydata.io/datastore/detail/1848655340695064576/>