

## 第八届泰迪杯数据挖掘挑战赛

### “智慧政务”中的文本挖掘应用

## 摘要

随着我国改革开放的深入，特别是互联网的快速发展，新技术的运用以及随之而来的新媒体的快速普及，为我国政府的社会治理以及公共服务提出了新的要求。在 2015 年的政府工作报告中，李克强总理首次提出了“互联网+”的概念，基于互联网技术和新媒体平台的文本收集也成为政府提供公共服务的新途径。文本收集作为一种新工具正受到政府的重视和人们的关注，是创新政府公共服务的新方法，有利于民众获取更多知情权和话语权，同时也变革着政府社会管理与公共服务方式。

本文将在文本内容特征和类别多样特征的基础上，从文本挖掘的角度出发，尝试构建合理分类的模型，并重点研究热点问题挖掘和意见反馈质量评价。基于文本挖掘的度量方法，通过政务意见文本的数据集，分别利用构建模型、统计验证和预测分析方法检验了不同特征的政务文本的热度和反馈评价，发掘更高效便利的政务意见操作模型，提升意见反馈的准确度和传播效果。

针对构建分类模型，首先要对文本进行分词，中文分词是对中文进行处理的基础内容，对于输入的一段中文，有效的对其进行中文分词，中文分词后的文档可以让电脑更好的识别其语义。根据语义，我们提取出核心名词“时间”、“地点”、“人物”、“事件”作为核心指标。确立去停用词后进行文本特征选择，求得文本的相似度，最后构建出模型并进行 F-Score 检验评价。

针对热点指标构建，从众多的留言中识别出相似度留言。其次是对问题的归类，在前文对文本内容进行处理的基础上，对政务意见进行识别归类，对一定时间内特定的地点和事件进行聚类，把特定的时间和任务进行归并，对相似的问题进行归一。最后是定义合理的热度评价指标，确定热度评价指标的科学性和计算方法，并给出评价结果，提取出最重要的热点问题并进行排序，从而导出对应结果。

针对评价模型构建，首先从相关性的角度，探究答复意见的内容是否与问题相关，从完整性的角度探究意见答复是否满足某种范式。然而，作为服务型政府，对于意见的评价应该是以人民为中心。因而，在探究意见答复的可解释性和满意度时，我们认为应该是意见反馈者对于相关部门的答复做出评价。通过分层随机抽样的形式，在不同一级指标中的意见及答复中，抽取足够的样本在一定范围内以问卷的形式由不同人群进行评价，以五级量表的形式进而获得平均评价。最后是提取回复时间与意见的落款时间相比较，得出意见的处理时间，进而计算意见答复的响应率。由前述，获得五项评价指标运用动态加权综合评价方法，从实证的路径构建评价方案。

**关键词：**文本挖掘 分类模型 热点指标构建 评价模型

## Summary

With the deepening of China's reform and opening up, especially the rapid development of the Internet, the application of new technologies and the subsequent rapid popularization of new media, it has put forward new requirements for the government's social governance and public services. In the 2015 government work report, Premier Li Keqiang first proposed the concept of "Internet +", and text collection based on Internet technology and new media platforms has also become a new way for the government to provide public services. As a new tool, text collection is receiving the attention of the government and people's attention. It is a new method of innovating government public services, which is conducive to the people's right to know and speak, and it also changes the government's social management and public service methods.

This article will try to build a reasonable classification model from the perspective of text mining based on the characteristics of text content and the diverse characteristics of categories, and focus on the research of hot topic mining and feedback quality evaluation. Based on the text mining measurement method, through the data set of government opinion texts, the construction model, statistical verification and predictive analysis methods were used to test the popularity and feedback evaluation of government texts with different characteristics, and to explore more efficient and convenient government opinion operation models to improve The accuracy of feedback and communication effect.

In order to build a classification model, we must first segment the text. Chinese segmentation is the basic content for processing Chinese. For an input section of Chinese, Chinese segmentation is effectively performed. The documents after Chinese segmentation can allow the computer to recognize it better. Semantic. According to the semantics, we extract the core nouns "time", "place", "person" and "event" as the core indicators. After establishing stop words, select text features, find the similarity of the text, and finally build a model and perform F-Score test evaluation.

Aiming at the construction of hotspot indicators, similarity messages are identified from many messages. The second is to categorize the problem. On the basis of processing the text content in the foregoing, identify and classify the government affairs opinions, cluster specific places and events within a certain time, and merge specific times and tasks. Similar problems are normalized. The last is to define a reasonable thermal evaluation index, determine the scientificity and calculation method of the thermal evaluation index, and give the evaluation results, extract the most important hot issues and

sort them, so as to derive the corresponding results.

For the evaluation model construction, first of all, from the perspective of relevance, explore whether the content of the reply is related to the question, and from the perspective of completeness, whether the reply meets a certain paradigm. However, as a service-oriented government, the evaluation of opinions should be centered on the people. Therefore, when investigating the interpretability and satisfaction of opinion responses, we believe that it should be the opinion feedbackers who evaluate the responses of the relevant departments. In the form of stratified random sampling, in the opinions and responses of different first-level indicators, enough samples are taken to be evaluated by different groups in the form of questionnaires within a certain range, and the average evaluation is obtained in the form of a five-level scale. Finally, the response time is extracted and compared with the payment time of the opinion, and the processing time of the opinion is obtained, and then the response rate of the opinion response is calculated. From the foregoing, five evaluation indicators are obtained and a dynamic weighted comprehensive evaluation method is used to construct an evaluation plan from an empirical path.

**Keywords:** text mining, classification model, hot index construction, evaluation model

目录

1.引言.....6

    1.1 背景资料.....6

    1.2 研究问题.....6

    1.3 问题解析.....6

    1.4 全文框架.....7

2.模型构建.....8

    2.1 模型的假设.....8

    2.2 合理性分析.....9

    2.3 符号的说明.....9

3.群众留言分类.....10

4.热点问题指标构建.....10

5.政务回复评价模型构建.....11

    5.1 抽样设计.....12

    5.2 数据说明.....13

6 总结.....22

    6.1 改进.....22

    6.2 展望.....22

## 1.引言

### 1.1 背景资料

随着我国改革开放的深入，特别是互联网的快速发展，新技术的运用以及随之而来的新媒体的快速普及，为我国政府的社会治理以及公共服务提出了新的要求。在 2015 年的政府工作报告中，李克强总理首次提出了“互联网+”的概念，基于互联网技术和新媒体平台的文本收集也成为政府提供公共服务的新途径。文本收集作为一种新工具正受到政府的重视和人们的关注，是创新政府公共服务的新方法，有利于民众获取更多知情权和话语权，同时也变革着政府社会管理与公共服务方式。

从题目可知，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，各类与社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。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务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

因此，本文将在文本内容特征和类别多样特征的基础上，从文本挖掘的角度出发，尝试构建合理分类的模型，并重点研究热点问题挖掘和意见反馈质量评价。基于文本挖掘的度量方法，通过政务意见文本的数据集，分别利用构建模型、统计验证和预测分析方法检验了不同特征的政务文本的热度和反馈评价，发掘更高效便利的政务意见操作模型，提升意见反馈的准确度和传播效果。

### 1.2 研究问题

近年来，网络问政平台的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。因此，需要根据所搜集的相关数据，建立若干模型以研究如下问题：

（1）建立关于留言内容的一级标签分类模型，提升在处理网络问政平台的群众留言时，工作人员的工作效率和准确率，并使用 F-Score 对分类方法进行评价。

（2）定义合理的热度评价指标，将某一时段内反映特定地点或特定人群问题的留言进行归类，并给出评价结果。

（3）针对相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。

### 1.3 问题解析

#### 1.3.1 问题（1）的分析

对政务文本的分析需要用到各种机器学习及自然语言处理等各方面的技术，主要包括中文文本的预处理技术，文本分类及聚类技术。这些技术保证了对于中文文本挖掘的快速准确的处理，从而建立模型。

针对构建分类模型，首先要对文本进行分词，中文分词是对中文进行处理的基础内容，对于输入的一段中文，有效的对其进行中文分词，中文分词后的文档可以让电脑更好的识别其语义。根据语义，我们提取出核心名词“时间”、“地点”、“人物”、“事件”作为核心指标。确立去停用词后进行文本特征选择，求得文本的相似度，最后构建出模型并进行 F-Score 检验评价。

### 1.3.2 问题（2）的分析

问题（2）要求对一段时间集中爆发的问题，多人反映的热点问题及时发现，通过构建指标模型协助相关部门进行有针对性的处理，提升服务效率。

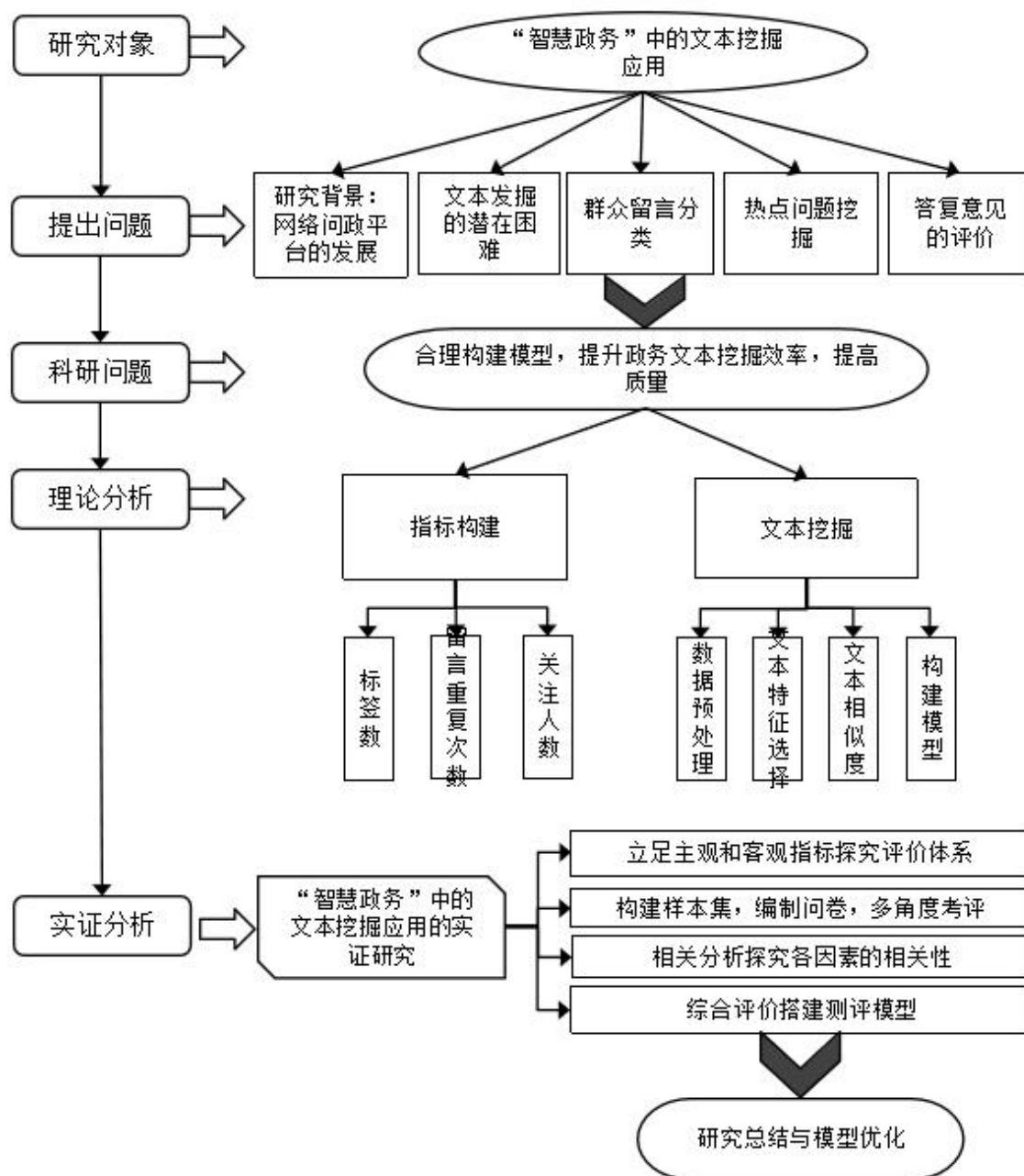
首先是对问题进行识别，从众多的留言中识别出相似度留言。其次是对问题的归类，在前文对文本内容进行处理的基础上，对政务意见进行识别归类，对一定时间内特定的地点和事件进行聚类，把特定的时间和任务进行归并，对相似的问题进行归一。最后是定义合理的热度评价指标，确定热度评价指标的科学性和计算方法，并给出评价结果，提取出最重要的热点问题并进行排序，从而导出对应结果。

### 1.3.3 问题（3）的分析

问题（3）要求针对相关部门对留言的答复，从相关性、完整性、可解释性等角度对答疑意见给出一套评价方案，并尝试实现。

作为一道开放性问题，首先从相关性的角度，探究答复意见的内容是否与问题相关，从完整性的角度探究意见答复是否满足某种范式。然而，作为服务型政府，对于意见的评价应该是以人民为中心。因而，在探究意见答复的可解释性和满意度时，我们认为应该是意见反馈者对于相关部门的答复做出评价。通过分层随机抽样的形式，在不同一级指标中的意见及答复中，抽取足够的样本在一定范围内以问卷的形式由不同人群进行评价，以五级量表的形式进而获得平均评价。最后是提取回复时间与意见的落款时间相比较，得出意见的处理时间，进而计算意见答复的响应率。由前述，获得五项评价指标运用动态加权综合评价方法，从实证的路径构建评价方案。

## 1.4 全文框架



## 2.模型构建

### 2.1 模型的假设

为了简化问题和建模的需要，给出必要的假设与合理性解释，以及通用符号的含义说明。

(1) 假设所有意见反馈的主题都高度概括网友所反馈的问题，与后续的留言详情在内容语义上相符。

(2) 假设对留言内容的反对或点赞仅涉及态度倾向，两者对于热度问题的测评均为同向影响指标。

(3) 假设政府各部门比较负责，错误分类的问题都将被转送至对应的部门进行解答，不存



在未被回复的问题。

(4) 假设反馈意见的网友都是理性人，回答问题(3)中提出的各项测评指标相互独立、互不影响。

## 2.2 合理性分析

(1) 留言详情当中存在终于情绪化的语句和无用的语气词，但作为意见反馈的主题，正常情况下是对问题的高度概括，因而我们假定两者所述的内容相同，从而选取留言主题作为文本挖掘的对象，简化步骤。

(2) 不论是对留言的态度是反对还是点赞，都是评论者对于问题的关注的一种体现，因而我们认为，两者对于热度问题的测评均为同向影响指标，从而可以在热度指标的构建问题上，对这两个指标进行归类。

(3) 因为技术问题导致的分类错误无法完全杜绝，但信息收集并不会出现缺失，因而我们假设只要各部门比较负责，即便是接收到不属于本部门管辖的问题，也会转送到对应的部门进行解决。

(4) 题目(3)中的指标分为主观性指标和客观性指标，客观性指标不会因为主观因素而受到影响，但主观性指标因人而异。考虑到部分反馈意见集中，因而假设反馈意见的网友都是理性人，排除因为受到旁人的态度而影响评价的可能性，从而让我们的问卷调查更贴切反馈意见的人群实际情况。

## 2.3 符号的说明

符号	解释
$\bar{y}$	指标概率的均值
$n$	样本总数
$y_i$	样本个体指标概率
$d$	绝对精度
$\mu_{\frac{1-\alpha}{2}}$	一定置信水平条件下的检验统计量
$s^2$	样本方差
$W_i$	分层后，层内样本数占总体的比例
$m$	对应人员数
$N$	意见总数

符号	解释
$x_1$	相关性
$x_2$	完整性
$x_3$	响应时间
$x_4$	平均可解释度
$x_5$	平均满意度
$X$	综合得分

### 3.群众留言分类

面对网络问政平台上众多的群众留言，我们从教育文体、交通运输、商贸旅游、城乡建设、环境保护、劳动和社会保障、卫生计生七个一级指标中，寻找相应的语料库，对群众的留言内容进行一一识别、归类。

具体实现方法如下：建立七个指标的相应语料库后建立三层循环，第一层对每个留言主题进行分析，第二层先使用 `readcell` 函数将语料库导入为 `cell` 格式并对七个语料库进行循环，第三层对某一留言主题使用 `contains` 函数对某一语料库中的每个关键词进行循环匹配，若匹配到任一关键词，便在此留言主题后标记相应标签。每个留言主题会经过七次循环，整个程序总共经过约 322642908 次循环。

对于该模型可以改进的地方在于在后期我们可以对标记了两种或以上类别的留言，进行再次筛选分析，逐渐提高关键词数量的导出要求，直至只剩一种类别；对于关键词频数相同的留言，最后可以进行人工识别的方法进行分类处理，该流程可以大大提高分类的效率与准确率。

### 4.热点问题指标构建

问题本身的热度是可以通过留言问题的出现次数以及评论点赞或者反对数量来热度量化的，所以在问题二中，我们第一步是分析附件二中重复出现的留言问题以及留言内容然后将其整理出来放在新的 `excel` 表格里面，详情如下图所示

B	C	D	E	F	G	H	I	J	K	L	M
U000784	A8县扶峰村民房屋倒塌多年，为何不能申请危房改造？	5/11/9 18:5	听说党的	城乡建设	500多块办证手续费，这样收费合理吗？						
U0002483	A7县北山镇新桥集棚户区改造工程存在质量问题	9/11/27 9:3	报告、险情	城乡建设	A2区向东南和沙湖桥马路变成了商贩们的自由市场						
U000784	A8县扶峰村民房屋倒塌多年，为何不能申请危房改造？	5/11/9 18:5	听说党的	城乡建设	A4区中山西路污水流街，“僵尸车”影响秩序						
U0003167	反映C4市收取城市垃圾垃圾处理费不平等的问题	/11/15 11:4	门所在的物	城乡建设	A5区华雅物业歧视农民工，违反劳动法						
U0008488	A市魏家坡小区脏乱差	/11/10 11:1	好好休息	城乡建设	A7县黄兴镇中心幼儿园学费多收不合理						
U0002783	请为A9市北盛镇居民健身休闲提供一个场地	9/4/20 20:4	一个传统	城乡建设	A7县普高老师也要拥有报名参加城区重点高中、初中、小学教师招聘的机会						
U0008264	关于加快C4市白田特色小镇建设的建议	7/4/11 22:4	政压力。	城乡建设	A8县市花明楼镇郑家冲的贵庭住工游青味太重						
U0008264	关于加快C4市白田特色小镇建设的建议	7/4/11 22:4	政压力。	城乡建设	A8县新农合医保每人交90元，可农合本上只填了60元						
U0001811	D4区雨母乡楚贵村安置房工程的工程款何时能拿到手	9/7/18 17:2	无数次的找	城乡建设	A9市进修附小经常收取各种费用，家长负担很重！						
U0006914	举报D9县建设局拖欠工程款和工资	9/1/31 8:4	少付清三台	城乡建设	A9市开心广场电梯仍未年检						
U0001469	D市住房公积金贷款门槛高	4/3/8 22:0	但是后来	城乡建设	A9市永安镇新建工厂污染严重让居民无法正常生活						
U0001161	D市廉租房（公租房）现在还可以购买共同产权吗	8/9/7 18:1	同产权房	城乡建设	A市2011年出租车燃油补贴怎么了？						
U0006341	关于D市美的梧桐庄园二期三期规划调整规划设计图的意	7/9/28 3:0	圾站建设的	城乡建设	A市A1区征地办公室违背“劳动法”侵犯劳动者权益						
U0001469	D市住房公积金贷款门槛高	4/3/8 22:0	但是后来	城乡建设	A9市永安镇新建工厂污染严重让居民无法正常生活						
U0001161	D市廉租房（公租房）现在还可以购买共同产权吗	8/9/7 18:1	同产权房	城乡建设	A市2011年出租车燃油补贴怎么了？						
U0006341	关于D市美的梧桐庄园二期三期规划调整规划设计图的意	7/9/28 3:0	圾站建设的	城乡建设	A市A1区征地办公室违背“劳动法”侵犯劳动者权益						
U0006341	关于D市美的梧桐庄园二期三期规划调整规划设计图的意	7/9/28 2:5	圾站建设的	城乡建设	A市城市道路交通如何规划						
U0006341	关于D市美的梧桐庄园二期三期规划调整规划设计图的意	7/9/28 3:0	圾站建设的	城乡建设	A市A1区征地办公室违背“劳动法”侵犯劳动者权益						
U0006341	关于D市美的梧桐庄园二期三期规划调整规划设计图的意	7/9/28 2:5	圾站建设的	城乡建设	A市城市道路交通如何规划						
U0004791	六年了，E12市温馨港湾楼盘还没有房产证	9/4/3 20:1	多次协商	城乡建设	A市的洒水时间段非常不合理，路过的车全部变脏						
U0004791	六年了，E12市温馨港湾楼盘还没有房产证	9/4/3 20:0	多次协商	城乡建设	A市东三路农贸市场油烟污染影响市容						
U000602	E4县宋家塘和顺家园廉租房的规划设计合理吗	7/6/14 21:0	在的6—7	城乡建设	A市合能枫丹丽舍拒绝公积金贷款						
U0006966	F3区农经局为何不发危房改造款？	9/6/4 16:1	借款挪用了	城乡建设	A市麓山国际实验学校教师工资低，平时每月1500元左右						
U0006966	F3区农经局为何不发危房改造款？	9/6/4 16:2	借款挪用了	城乡建设	A市仁和会计培训金犀北校区不给我们退款						

第二步，在附件3中已经提供好了每个留言的点赞数以及反对数，因为不论是点赞还是反对都是表明对这个问题的关注，如果留言一有10个人点赞，留言二有10个人反对，在这种情况下，我们认定留言一与留言二的关注人数是一样的，也就是我们认定了点赞与反对本质上是一样的，所以在附件3的分析中，我们第一步是将每一条留言的反对人数与点赞人数做一个求和处理，在这里，我们只截取了求和之后排在前五的五条留言的明细情况作为例子，详情如下图所示

194343	A000106161	A市58年贷案警官应跟进关注	2019/3/1 22:12:30	希望。但是，A市A4区经侦并没有跟进市领导的留言，案件调查进	0	733	733
217032	A00056543	A市58年贷特大集资诈骗案保	2019/2/25 9:58:37	有功；说大股东苏纳和小股东、苏纳弟弟苏吕是挂名；说担保公	0	790	790
220711	A00031682	请书记关注A市A4区58年贷案	2019/2/21 18:45:14	焦急万分，急切盼望有案情消息总是失望，四处诉求也无效。此	0	821	821
223297	A00087522	反映A市金毛湾配套入学的问题	2019/4/11 21:02:44	育局暂未将金毛湾楼盘纳入配套入学，A3区教育局向市教育局汇	5	1762	1767
208636	A00077171	汇金路五矿万境K9县存在一系	2019/8/19 11:34:04	？我们小区也曾发生过狗咬人，请问有人对养宠物的情况进行过	0	2097	2097

第三步，如果只用反对人数以及点赞的人数之和作为热点问题的衡量标准，可能会存在很大的偏差，所以我们对其做了加权处理，具体公式如下

$$\frac{\text{标签数}}{\text{总留言数}} \times \text{留言重复次数} \times (\text{反对人数} + \text{点赞人数})$$

具体的计算我们可以看这个例子：假如附件中有200条留言，留言一属于城乡建设的标签，城乡建设标签在附件中有50条记录，留言一在附件中出现了2次，点赞人数为2，反对人数为0。那么加权求和就如下所示

$$\frac{50}{200} \times 2 \times (2 + 0)$$

运用这个公式我们把附件中的留言都做了加权处理，然后用excel的筛选功能(Ctrl+Shift+L)把排列在前五的记录找了出来,并且将其保存在热点问题留言明细表中具体如图三所示

194343	A000106161	A市58年贷案警官应跟进关注	2019/3/1 22:12:30	希望。但是，A市A4区经侦并没有跟进市领导的留言，案件调查进
217032	A00056543	A市58年贷特大集资诈骗案保	2019/2/25 9:58:37	有功；说大股东苏纳和小股东、苏纳弟弟苏吕是挂名；说担保公
220711	A00031682	请书记关注A市A4区58年贷案	2019/2/21 18:45:14	焦急万分，急切盼望有案情消息总是失望，四处诉求也无效。此
223297	A00087522	反映A市金毛湾配套入学的问题	2019/4/11 21:02:44	育局暂未将金毛湾楼盘纳入配套入学，A3区教育局向市教育局汇
208636	A00077171	汇金路五矿万境K9县存在一系	2019/8/19 11:34:04	？我们小区也曾发生过狗咬人，请问有人对养宠物的情况进行过

第四步，在找出排列在前五的热点问题之后，我们用问题一用过的分类方法将五个热点问题各级标签找了出来，并且将其保存在热点问题表的附件中。

5.政务回复评价模型构建

## 5.1 抽样设计

### 5.1.1 设计概述

实验概述：在同一类意见反馈中，随机抽取 6 个样本。通过前期摸排，以附件 4 的该类文本为总体，通过观测样本的指标概率，并以此作为调查指标，估算出构建评价的最低样本容量。

样本概述						
样本编号	1	2	3	4	5	6
指标	0.42	0.25	0.38	0.40	0.33	0.38

总体概述						
总体编号	1	2	3	4	5	6
$W_i = \frac{N_i}{N}$	0.13	0.18	0.15	0.17	0.17	0.20

### 5.1.2 设计推导

$$\bar{y} = \frac{1}{n} \sum_{i=1}^6 y_i = 0.36$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^6 (y_i - \bar{y})^2 = 0.0038$$

本次抽样调查为无放回抽样，样本不是独立同分布，基于 Wald-Wolfowitz 定理，按绝对精度决定样本量。在置信水平 95%的条件下，求得绝对精度  $d$

$$d^2 = (\mu_{\frac{1-\alpha}{2}})^2 \frac{1}{n} (1 - \frac{n}{N}) s^2 = (1.96)^2 \times \frac{1}{6} \times (1 - \frac{6}{126}) \times 0.0038^2 \approx 8.80519\text{E} - 06$$

$$n = \frac{(\mu_{\frac{1-\alpha}{2}})^2 s^2}{d^2 + \frac{1}{N} (\mu_{\frac{1-\alpha}{2}})^2 s^2} \approx \frac{(1.96)^2 \times 0.0038^2}{8.80519\text{E} - 06 + \frac{1}{126} \times (1.96)^2 \times 0.0038^2} = 6$$

在实际工作中为了比较可靠地达到提出的精度要求，通常会将算出值扩大 10%，即 6.6。

将部门作为分层依据，按照奈曼方法分配样本量，在分层抽样中， $n = \sum_{i=1}^K n_i$  固定，使得

$$V(\bar{y}_{st}) = \sum_{i=1}^K W_i^2 \left( \frac{1}{n_i} - \frac{1}{N} \right) S_i^2, \text{ 达到最小的样本量分配为 } n_i = n \frac{W_i S_i}{\sum_{j=1}^K W_j S_j}, i = 1, \dots, K., \text{ 即可得到}$$

奈曼最优的最小分配量，反推出需要在 2000 多的文本中大约抽取 50 条文本作为实验样本。

## 5.2 数据说明

### 5.2.1 数据来源

#### (1) 样本选取

鉴于政务文本具有覆盖部门多、发展时间长等特点，已经形成了较为完善的运行模式，并且其能提供开放的数据获取接口，保证了数据的可得性，因此在构建样本数据时，直接从平台所得文本中随机抽取。为了使研究对象覆盖不同级别政府机构运营的政务意见，我们选取覆盖不同时间、面向不同政府官方机构的文本意见，共 50 条。

#### (2) 数据获取方法

首先利用附件 4 的文本数据，随机采集每个账号在特定时间前的意见信息作为研究对象，共近 3000 条文本数据，并提取研究假设中涉及的提问与回复时间、意见主题特征和文本内容；

其次，由于文本数据采集的限制，我们通过将抽取的文本制作成问卷的形式，间接采集了每个文本信息的可解释性和满意度评价，相比附件 4 中的文本能更准确地反映信息受众对回复的态度，共获得 147 个评论用户，并对他们的评价求取平均值；

最后，为了回复相关性和完整性的数据，对每个评论用户，利用和①类似的方法，采集其提出意见的主题内容信息，并根据相关定义求得信息。

#### (3) 数据处理

本研究对采集的原始数据进行异常值和缺失值预处理，通过数理统计和数据挖掘的清理规则校验数据，使有效数据保持一致，提高数据质量。设计本次问卷的 Likert5 量表，从多个维度测。采用 SPSS24.0 对数据进行录入和整理，并运用 SPSSModeler18.0 对数据进行预处理和数据挖掘；在相关分析和验证性因素分析环节，综合运用 SAS9.4 和 SPSS24.0 和 SPSSAU 进行分析；建模环节，运用 Amos 建立结构方程模型，多种软件发挥优势，为本次建模提供技术支持。

### 5.2.2 指标解释

本文在模型中关注了网友提问的政务文本信息，即历史提问记录与部门回复的匹配度及相关评价。假设网友发出的意见集合为  $P(g) = \{p(g)1, p2(g), \dots, pn(g)\}$ ,  $n$  为政务文本数量；每个部门回复的集合为  $U(g) = \{u1, u2, \dots, um\}$ ,  $m$  为对应人员数；每一个网友  $u$  所发的政务意见集合为集合  $P = \{p1, p2, \dots, pN\}$ ,  $N$  为总数。

**相关性：**我们对相关性的理解是基于网友对反馈意见的主题与政府部门回复的文本重合度作为衡量。通过提取二者的词频，以主题中的词句作为分母，政务意见回复中出现的相同词句为分子，计算出意见与回复之间的相关性，计算公式如下：

$$\text{相关性} = \text{词语覆盖度} = \frac{\text{回复中出现主题中的词句}}{\text{主题中的词句}}$$

**完整性：**我们对完整性的界定是政府部门对意见进行回复时，具有一定的模板和固定回复的话，通过观察，我们认为有四项，分别是：“xxx 网友您好”、“关于你对 xx 问题的提出”、“感谢您对 xxx 的关注”、“回复日期”。此四项是政务回复的基本用语，也是体现服务型政府、保证回复相关性的重要衡量。从完整性的角度来说，可分为 0、1、2、3、4 共五个等级，代表着政务回复当中是否含有这四项基本用语，越完整等级越高。

**响应时间：**我们认为对于问题的解决和答复会影响意见反馈者的满意度，因此选取了提问时间和回复时间作为响应时间的计算基础，考虑到简化问题，我们以“天”为计算单位，计算公式如下：

$$\text{响应时间} = \text{回复时间} - \text{反馈时间}$$

**可解释性：**我们对可解释性的理解是，意见反馈者能够理解政府部门的回复，因而对于可解释性的测评，我们认为应该以人为中心，设计本次问卷的 Likert5 量表，从个体主观的角度进行评价。

**满意度：**我们对可解释性的理解是，意见反馈者能够理解政府部门的回复，因而对于可解释性的测评，我们认为应该以人为中心，设计本次问卷的 Likert5 量表，从个体主观的角度进行评价。

### 5.2.3 相关性研究

相关分析用于研究定量数据之间的关系情况，是否有关系，关系紧密程度情况等。

Pearson 相关-标准格式

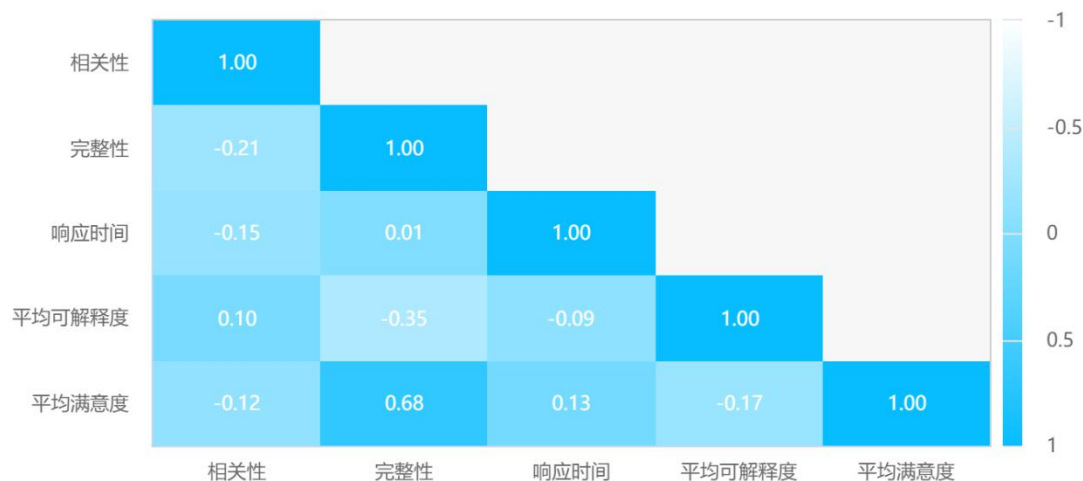
	平均值	标准差	相关性	完整性	响应时间	平均可解释度	平均满意度
相关性	0.461	0.316	1				
完整性	2.400	1.385	-0.208	1			
响应时间	24.020	40.988	-0.151	0.011	1		
平均可解释度	2.740	1.440	0.097	-0.346*	-0.086	1	
平均满意度	3.120	1.335	-0.116	0.680**	0.129	-0.175	1

\*p<0.05\*\*p<0.01

从上表可知，利用相关分析去研究相关性和完整性，响应时间，平均可解释度，平均满意度共 4 项之间的相关关系，使用 Pearson 相关系数去表示相关关系的强弱情况。具体分析可知：相

相关性、完整性、响应时间、平均可解释度、平均满意度共 4 项之间均不会呈现出显著性，相关系数数值分别是-0.208, -0.151, 0.097, -0.116, 全部均接近于 0，并且 p 值全部均大于 0.05，意味着相关性、完整性、响应时间、平均可解释度、平均满意度共 4 项之间均没有相关关系。

Pearson相关可视化图



基于此，对上述指标进行逐步回归，对比分析影响满意度的主要因素。

逐步回归分析结果-简化格式

	回归系数	VIF
常数	1.547(5.490**)	-
完整性	0.655(6.427**)	1.000
样本量	50	
R <sup>2</sup>	0.463	
调整 R <sup>2</sup>	0.451	
F 值	F(1, 48)=41.304, p=0.000	

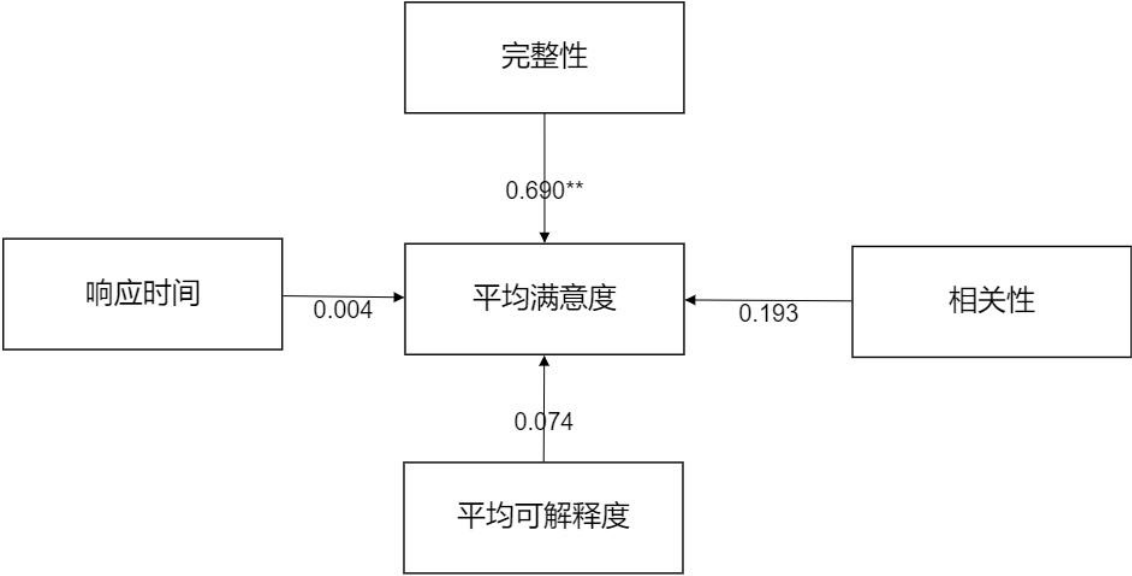
因变量：平均满意度

D-W 值：2.421

\*p<0.05\*\*p<0.01 括号里面为 t 值

将相关性、完整性、响应时间、平均可解释度作为自变量，而将平均满意度作为因变量进行逐步回归分析，经过模型自动识别，最终余下完整性一共 1 项在模型中，R 方值为 0.463，意味着完整性可以解释平均满意度的 46.3%变化原因。而且模型通过 F 检验(F=41.304, p=0.000<0.05)，说明模型有效。以及模型公式为：平均满意度=1.547+0.655\*完整性。完整性的回归系数值为 0.655(t=6.427, p=0.000<0.01)，意味着完整性会对平均满意度产生显著的正向影响关系。总结分

析可知：完整性会对平均满意度产生显著的正向影响关系。



同时，上述结果在灰色关联法中得到进一步验证。

关联度结果

评价项	关联度	排名
相关性	0.854	5
完整性	0.870	4
响应时间	0.993	1
平均可解释度	0.873	3
平均满意度	0.875	2

结合上述关联系数结果进行加权处理，最终得出关联度值，使用关联度值针对 50 个评价对象进行评价排序；关联度值介于 0~1 之间，该值越大代表其与“参考值”（母序列）之间的相关性越强，也即意味着其评价越高。从上表可以看出：针对本次 5 个评价项，响应时间的综合评价最高（关联度为：0.993），其次是平均满意度（关联度为：0.875）。

5.2.4 指标权重确立

探索研究人群分类，研究每类的特征情况如何，聚类分析使用 K-均值聚类方法进行，最终生成类别频数分布如下表。

聚类类别基本情况汇总

聚类类别	频数	百分比（%）
cluster_1	45	90.00%
cluster_2	5	10.00%



合计	50	100%
----	----	------

使用聚类分析对样本进行分类，使用 Kmeans 聚类分析方法，从上表可以看出：最终聚类得到 2 类群体，此 2 类群体的占比分别是 90.00%，10.00%。整体来看，2 类人群分布较为均匀，整体说明聚类效果较好。

聚类类别方差分析差异对比结果

	聚类类别方差分析差异对比结果(平均值±标准差)		F	p
	cluster_1(n=45)	cluster_2(n=5)		
相关性	0.48±0.33	0.29±0.13	1.617	0.210
完整性	2.42±1.41	2.20±1.30	0.114	0.737
响应时间	12.11±16.14	131.20±41.94	165.575	0.000**
平均可解释度	2.82±1.47	2.00±1.00	1.481	0.230
平均满意度	3.11±1.34	3.20±1.48	0.020	0.889

\*p<0.05\*\*p<0.01

得到聚类类别后，系统默认生成一系列数据表示聚类类别，为了探索出各个类别的具体特征，因而使用方差分析去研究各个类别群体的差异性，最终可结合各个类别特征进行类别命名。使用方差分析去探索各个类别的差异特征，从上表可知：相关性，完整性，响应时间，平均可解释度，平均满意度共 4 项并没有呈现出显著性(p>0.05)，意味着聚类分析过程中，此 4 项对聚类分析不会起作用，可考虑移出此 4 项后再次进行聚类分析。

接着对指标进行主成分分析，探索定量数据可以浓缩为几个方面(主成分)，通常用于权重计算。

KMO 和 Bartlett 的检验

KMO 值		0.509
Bartlett 球形度检验	近似卡方	40.299
	df	10
	p 值	0.000

使用主成分分析进行信息浓缩研究，首先分析研究数据是否适合进行主成分分析，从上表可以看出：KMO 为 0.509，大于 0.5，但是小于 0.6，满足主成分分析 KMO 值大于 0.5 的基本要求，同时也建议加大样本量，以及移出相关关系过强的研究项，再次进行分析，以求更加科学的研究。

方差解释率表格

编号	特征根			主成分提取		
	特征根	方差解释率%	累积%	特征根	方差解释率%	累积%
1	1.939	38.771	38.771	1.939	38.771	38.771
2	1.069	21.384	60.154	1.069	21.384	60.154
3	0.865	17.303	77.457	-	-	-
4	0.852	17.039	94.497	-	-	-
5	0.275	5.503	100.000	-	-	-

上表格针对主成分提取情况，以及主成分提取信息量情况进行分析，从上表可知：主成分分析一共提取出 2 个主成分，特征根值均大于 1，此 2 个主成分的方差解释率分别是 38.771%、21.384%，累积方差解释率为 60.154%。（提示：如果主成分提取个数与预期不符，可在分析时主动设置主成分个数）。另外，本次分析共提取出 2 个主成分，它们对应的加权后方差解释率即权重依次为： $38.771/60.154=64.45\%$ ； $21.384/60.154=35.55\%$ ；

载荷系数表格

名称	载荷系数		共同度(公因子方差)
	主成分 1	主成分 2	
相关性	-0.389	0.572	0.478
完整性	0.880	0.262	0.843
响应时间	0.235	-0.795	0.688
平均可解释度	-0.538	-0.027	0.290
平均满意度	0.818	0.201	0.709

上表格展示主成分对于研究项的信息提取情况，以及主成分和研究项对应关系，从上表可知：针对共同度而言，共涉及平均可解释度共 1 项，说明主成分和研究项之间的关系非常薄弱，主成分不能有效的提取出研究项信息。因而应该将此 1 项进行删除，删除之后再次进行分析。

成份得分系数矩阵

名称	成分	
	成分 1	成分 2
相关性	-0.201	0.535
完整性	0.454	0.245
响应时间	0.121	-0.744

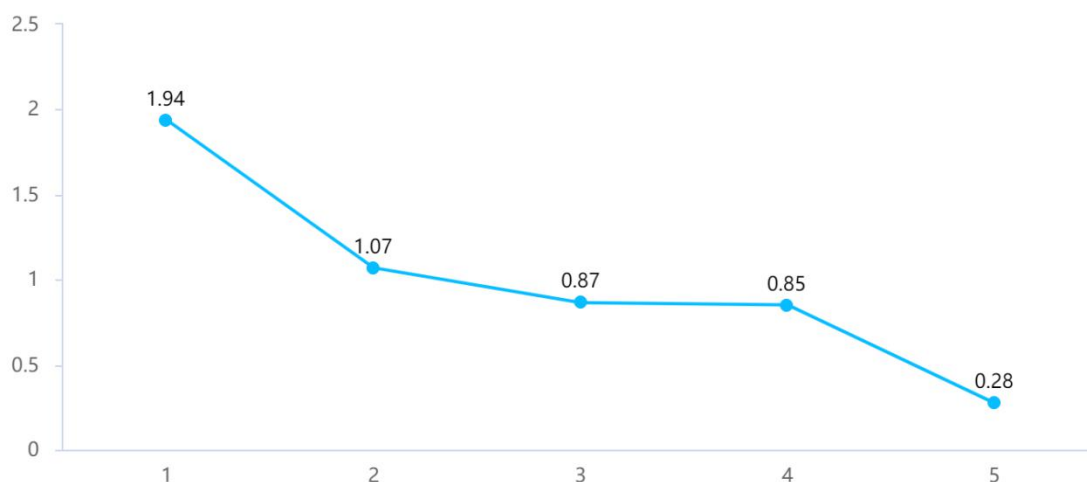
平均可解释度	-0.277	-0.025
平均满意度	0.422	0.188

使用主成分分析目的在于信息浓缩，则忽略“成份得分系数矩阵”表格。如果使用主成分分析法进行权重计算，则需要使用“成份得分系数矩阵”建立主成分和研究项之间的关系等式，如下：

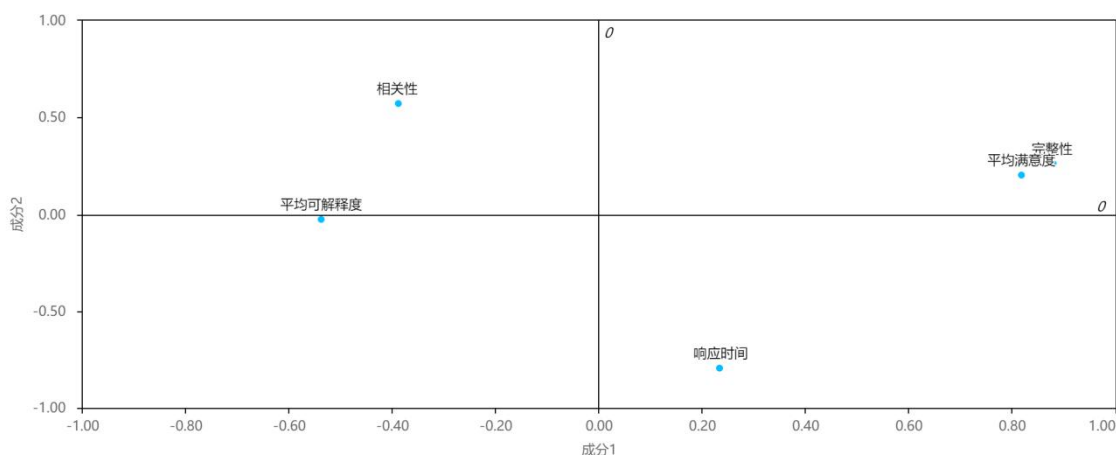
成分得分 1= $-0.201 \times \text{相关性} + 0.454 \times \text{完整性} + 0.121 \times \text{响应时间} - 0.277 \times \text{平均可解释度} + 0.422 \times \text{平均满意度}$

成分得分 2= $0.535 \times \text{相关性} + 0.245 \times \text{完整性} - 0.744 \times \text{响应时间} - 0.025 \times \text{平均可解释度} + 0.188 \times \text{平均满意度}$

碎石图



载荷图



### 5.2.5 评价方法确立

熵值是一种物理计量单位；熵越大说明数据越混乱，携带的信息越少，效用值越小，因而权

重也越小。熵值法则是结合熵值提供的信息值来确定权重的一种研究方法。对指标进行熵值分析。

熵值法计算权重结果汇总

项	信息熵值 e	信息效用值 d	权重系数 w
相关性	0.9318	0.0682	16.37%
完整性	0.9488	0.0512	12.29%
响应时间	0.7639	0.2361	56.67%
平均可解释度	0.9637	0.0363	8.72%
平均满意度	0.9752	0.0248	5.95%

使用熵值法对相关性等总共 5 项进行权重计算，从上表可以看出：相关性,完整性,响应时间,平均可解释度,平均满意度总共 5 项，它们的权重值分别是 0.164,0.123,0.567,0.087,0.060。各项间的权重大小有着一定的差异，其中响应时间这项的权重最高为 0.567，以及平均满意度这项的权重最低为 0.060。

结构方程模型 SEM 共包括两种关系，分别是影响结构关系和测量关系；回归路径系数汇总表展示因子或量表间的结构影响关系情况，对该评价体系建立模型。

模型回归系数汇总表

X	Y	非标准化路径系数	SE	z	p	标准化路径系数
相关性	平均可解释性	0.118	0.619	0.190	0.849	0.026
相关性	平均满意度	0.202	0.446	0.453	0.650	0.048
完整性	平均可解释性	-0.354	0.141	-2.511	0.012	-0.340
完整性	平均满意度	0.663	0.101	6.599	0.000	0.688
响应时间	平均满意度	0.004	0.003	1.318	0.187	0.135

从上表可知：相关性对于平均可解释性影响时，此路径并没有呈现出显著性( $z=0.190$ ,  $p=0.849>0.05$ )，因而说明相关性对平均可解释性并不会产生影响关系。

相关性对于平均满意度影响时，此路径并没有呈现出显著性( $z=0.453$ ,  $p=0.650>0.05$ )，因而说明相关性对平均满意度并不会产生影响关系。

完整性对于平均可解释性影响时，标准化路径系数值为 $-0.340<0$ ，并且此路径呈现出 0.05 水平的显著性( $z=-2.511$ ,  $p=0.012<0.05$ )，因而说明完整性会对平均可解释性产生显著的负向影响关系。

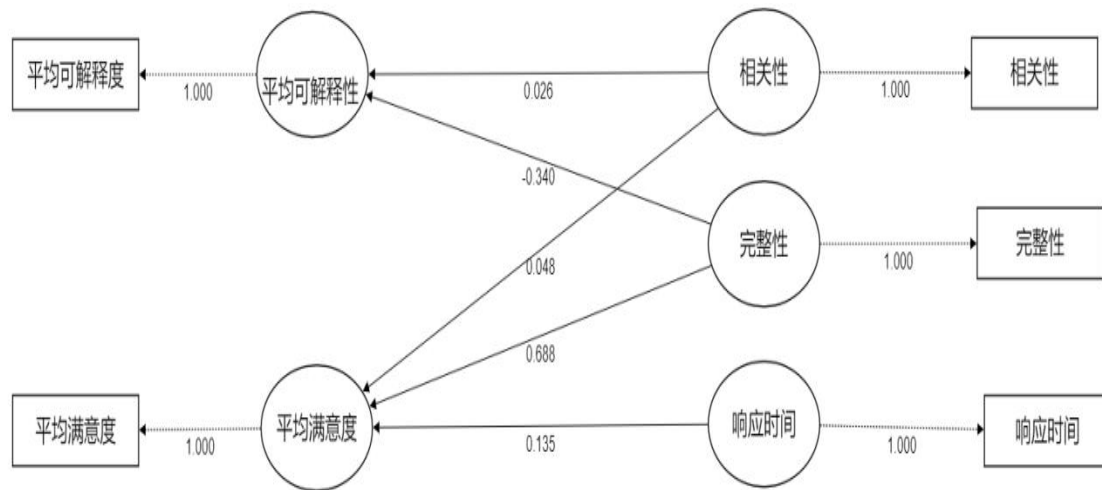
完整性对于平均满意度影响时，标准化路径系数值为  $0.688>0$ ，并且此路径呈现出 0.01 水平

的显著性( $z=6.599$ ,  $p=0.000<0.01$ ), 因而说明完整性会对平均满意度产生显著的正向影响关系。

响应时间对于平均满意度影响时, 此路径并没有呈现出显著性( $z=1.318$ ,  $p=0.187>0.05$ ), 因而说明响应时间对平均满意度并不会产生影响关系。

协方差关系表格

X	Y	非标准化路径系数	SE	z	p	标准化路径系数
相关性	完整性	-0.089	0.062	-1.439	0.150	-0.208
相关性	响应时间	-1.912	1.814	-1.054	0.292	-0.151
完整性	响应时间	0.612	7.868	0.078	0.938	0.011
平均可解释性	平均满意度	0.133	0.181	0.734	0.463	0.104



模型构建完成后, 尝试使用模糊综合评价方法进行打分。模糊综合评价借助模糊数学的一些概念, 对实际的综合评价问题提供评价, 即模糊综合评价以模糊数学为基础, 应用模糊关系合成原理, 将一些边界不清、不易定量的因素定量化, 进而进行综合性评价的一种方法。

权重计算结果

	相关性	完整性	响应时间	平均可解释度	平均满意度
隶属度	0.030	0.146	0.478	0.165	0.180
【权重】	0.030	0.146	0.478	0.165	0.180

从上表可知, 针对 50 个指标【样本量为 50 所以有 50 个指标】, 以及 5 个评语集进行模糊综合评价, 并且使用  $M(.,+)$  算子进行研究;

首先建立评价指标权重向量矩阵  $A$ , 以及构建出  $50 \times 5$  权重判断矩阵  $R$ , 最终进行分析出 5 个评语集的权重值, 分别是: 0.030, 0.146, 0.478, 0.165, 0.180。

从上表可知, 5 个评语集中响应时间的权重值最高(0.478), 结合最大隶属度法则可知, 最终

综合评价结果为"响应时间"。

综合得分计算

综合得分	相关性	完整性	响应时间	平均可解释度	平均满意度
3.113	2	3	4	3	3

显然，最后可构建出综合评分指标：

$$X = 0.03x_1 + 0.146x_2 + 0.478x_3 + 0.165x_4 + 0.18x_5$$

其中，X 越大意味着该条政务文本回复的质量越高，评价越好，影响整体评价的最重要因素是对意见反馈的响应时间。显然，提高政府部门的工作效率，及时回应百姓关注的热点问题，是提高综合得分的最主要途径。

## 6 总结

### 6.1 改进

从问题本身的角度出发，还有很多值得进一步改进和研究的问题。

(1) 从分类的角度看，意见收集时其实可以让网友选择意见投向哪一个部门，这样对于一级指标分类下的分工能够精简，减少不必要的工作量。

(2) 作为网络政务服务平台，应该立足互联网思维，当前微信公众号的测评标准除了阅读量和在看量以外，还有 WCI 指数。对于热度问题的构建，可参考相关指数。

(3) 作为服务型政府，意见反馈是一方面，但对于反馈的意见的质量评价应该有提出意见的人进行点评，当前系统并无进一步评价或提问的渠道，在日后的工作改进中可增设评价环节，这样也有利于提高各部门回复的质量和效率。

### 6.2 展望

在本次比赛的问题分析中，我们深刻地体会到了自然语言处理技术的便捷，在问题一中我们初步尝试了运用文本分类模型去处理自然语言将其分成不同的分类。在问题二中我们重新理解了热点问题的发掘，热点问题是一个社会的焦点，无论放在哪里都是对社会舆论影响颇深的，掌握社会热点问题发掘有利于行政部门的管理，找出社会焦点管理起来才会更具有针对性。在热点问题的发掘中，我们需要找到衡量热度化的标准，热点问题衡量的基础标准是关注的人数，但是除了这个基础它可能还会与其他的条件挂钩，这需要我们具体问题具体分析。在问题三我们运用模糊综合评价方法，首先对原始数据进行异常值和缺失值预处理，通过数理统计和数据挖掘的清理规则校验数据，使有效数据保持一致，提高数据质量。设计本次问卷的 Likert5 量表，从多个维度测。采用 SPSS24.0 对数据进行录入和整理，并运用 SPSSModeler18.0 对数据进行预处理和数据挖掘；在相关分析和验证性因素分析环节，综合运用 SAS9.4 和 SPSS24.0 和 SPSSAU 进行分

析；建模环节，运用 Amos 建立结构方程模型，多种软件发挥优势，为本次建模提供技术支持，最终做出评价方法。

大数据，云计算，人工智能技术的出现使自然语言处理技术日益完善，自然语言政务系统将成为社会治理创新发展的重要趋势。自然语言处理技术将完全运用到对民众的留言，建议，等的分类管理中，使其完全的代替人工分类，这不仅提高了政府管理的效率，还能进一步的了解民生，民意。

### **参考文献：**

- [1] 于秀林、任雪松（1999）：《多元统计分析》，中国统计出版社。
- [2] 王学民（2017）：《应用多元统计分析》，上海财经大学出版社。
- [3] 陈希孺（1991）：《概率论与数理统计》，中国科技大学出版社。
- [4] 德尔维奇 斯劳特（2017）：《The Little SAS Book 中文版》，清华大学出版社。
- [5] 韩中庚（2016）：《数学建模方法及其应用》，高等教育出版社
- [6] 朱明（2008）：《数据挖掘》，中国科学技术大学出版社