基于 NLP 与深度学习的网政平台文本数据挖掘

摘　要

本文旨在利用自然语言处理技术和深度学习技术对网络问政平台文本数据进行挖掘，通过深度学习的方法建立群众留言的分类模型，同时构建相关评价体系对留言热度和答复意见质量进行相关评价。

在本次数据挖掘过程中，对于获取的留言数据，我们首先利用 excel、python 等工具进行数据预处理、统一规范数据格式、清除异常值、分词以及停用词等过滤操作，实现数据的优化并提升其可用性，为建模做准备。

针对问题一，本文首先运用 EDA 数据增强技术对留言数据集进行文本增强，扩大了语料库的信息容量，减少类别分布不均衡的影响。然后再使用卷积神经网络进行深度监督学习，对留言详情的特征进行提取，然后采用 Softmax 回归系数对所提取的特征进行分类，建立一个多分类模型。同时使用分层抽样抽取 70%作为训练集，30%作为测试集，最后得到模型的 F-Score 得分为 0.9874，模型的分类效果较优。

针对问题二，本文首先利用基于 Levenshtein Distance 算法的模糊匹配识别相似留言并归类，再通过正则表达式构造语义规则提取留言中的地址和人群信息，并构建了留言热度评价模型识别出某一时间段内反应出反映特定地点或人群的热点问题。

针对问题三，本文从答复意见的相关性、完整性、可解释性、时效性、丰富性五个维度构建了答复意见评价体系，构建了文本相似度、答复次数、答复格式、事实依赖、文本充实度、响应时长 6 个量化指标。使用模糊-层次分析法确定各指标的权重，然后用灰色关联度分析法计算答复意见的关联度，并根据关联度将答复意见划分为优、良、差三个等级，并赋予相应的标签。

关键词：留言数据、卷积神经网络、分类器、模糊匹配、灰色关联分析

ABSTRACT

The purpose of this paper is to use natural language processing technology and deep learning technology to excavate the text data of the network political platform, to establish the classification model of mass messages by means of deep learning, and to construct the relevant evaluation system to evaluate the popularity of messages and the quality of responses.

In this data mining process, we first use excel, python and other tools for data preprocessing, unified specification data format, removal of outliers, word breakers and de-stop words and other filtering operations, to achieve data optimization and improve its availability, in preparation for modeling.

In response to question one, this paper first uses EDA data enhancement technology to enhance the message data set, expand the information capacity of the corpus, reduce the impact of the uneven distribution of categories. Then the in-depth supervision study is carried out by the convolutional neural network, the features of the message details are extracted, and then the extracted features are classified by Softmax regression coefficient, and a multi-classification model is established. Using a hierarchical sampling of 70% as a training set,30% as a test set, and finally the F-Score score of the model was 0.9874, which is better classified.

In view of question two, this paper first uses the fuzzy matching based on the Levenshtein Distance algorithm to identify similar messages and classify them, then extracts the address and crowd information in the message by constructing semantic rules of regular expression, and constructs the message heat evaluation model to identify the hot issues that reflect a particular location or population in a certain time period.

In view of question three, this paper constructs the evaluation system of response opinions from the five dimensions of relevance, completeness, interpretability, timeliness and richness of the reply opinions, and constructs 6 quantitative measures of text similarity, number of replies, reply format, fact dependence, text fulfillment, and response duration. Use - analysis to determine the weight of each indicator, and then use the gray correlation analysis method to calculate the relevance of the reply comments, and according to the degree of association, the reply comments are divided into three levels of excellence, good, difference, and assigned the corresponding labels.

Keywords: message data, convolutional neural network, Classifier, fuzzy matching, grey relational analysis.

目录

一、 研究背景和意义

近年来，随着互联网技术的普遍应用，大数据、云计算、人工智能等技术的发展，我国政务处理电子化趋势明显增强。网络问政平台是互联网时代下，政府为群众提供服务的主要平台，也是政府与群众信息交流的主要方式。提升公共服务与管理功能、建设智能化政府网站是未来政府网站发展的主要目标。而目前，大部分电子政务系统还是依靠人工进行数据整理，不断攀升的文本数据量对相关部门的工作带来了极大的挑战。

自然语言是人们进行通信和交流的主要工具，也是现代信息科学和技术研究不可或缺的重要内容。在互联网和大数据时代下，存在海量的中文自然语言描述数据，如何利用自然语言处理和文本挖掘技术，有效提炼出有效信息，具有重要研究意义和实用价值。

二、 问题分析

针对问题一，为了解决人工处理群众留言记录分类存在的效率低、工作量大的问题，本小组采用深度学习的方式实现对留言进行归类。文本被分入多个类别，每一条记录属于一个类别，属于典型的单标签多分类问题。考虑留言详情对分类标签的影响，本文搭建了基于卷积神经网络（CNN）的标签分类器，采用分层抽样的方式划分训练测试集，进行有监督学习，将关键词向量作为卷积神经网络模型的输入，经过卷积层和池化层提取高层特征，输出层接分类器得出分类结果。通过不断的训练测试、模型优化与参数调整之后，得出分类结果并对分类器的分类结果做出评价，验证该方法的有效性。

针对问题二，由于中文自然语言地址描述数据的非结构化，句法对词语约束性差，词语搭配随意性很强，且没有遵循统一的语法规则。考虑到文本的不规范性和数据量的约束，使用地址语料库或标注建立地址语料库进行机器学习的效果并不理想。因此，针对留言主题采用模糊匹配法将留言进行归类，利用正则表达式提取归纳出各类地址，定义热度评价指标，分别对每一类问题进行热度指数计算并进行热度排名，得到热点问题明细和热点问题。

关于问题三，针对相关部门对留言的答复意见，构建一套评价答复意见质量的综合评价指标体系。在观察数据并进行预处理之后，可从相关性、完整性、可解释性、时效性、丰富性五个维度进行评价体系的构建，并设计指标的量化细则，采用灰色关联分析法进行综合评价。灰色关联分析法的基本思想是根据行为序列曲线几何形状的相似性来确定序列之间联系的紧密程度，我们将这一思想应用到确定留言答复意见质量的指标

权重上。首先在层次分析法的基础上进行改良，利用改良之后的模糊—层次分析法剔除影响因子之间的相关关系来优化因子的权重，接着利用灰色关联度对留言的答复意见进行综合评价，评价方法客观准确，可操作性强。最后根据灰色关联度大小将答复意见质量划分为优、良、差三个等级，将计算结果离散化归类赋予标签。

三、 变量说明

符号	意义
S	总留言数
E	总点赞数
D	总反对数
a_j	第 j 类留言的留言数量
e_j	第 j 类留言的点赞数
d_j	第 j 类留言的反对数
h_j	第 j 类留言的热度指数
q_i	第 i 条留言记录中的留言详情
r_i	第 i 条留言记录中的答复意见
s_i	第 i 条留言记录中答复意见与留言详情的相似度
$k1_i$	第 i 条留言记录中的留言详情的关键词文本向量
$k2_i$	第 i 条留言记录中的答复意见的关键词文本向量
t	答复的响应时长（$t>0$）
ω_j	各指标对于留言答复质量的贡献权重（$1\leqslant j\leqslant 6$）
r_i	第 i 条留言记录中答复意见的灰色关联度

四、 问题一

4.1主要流程

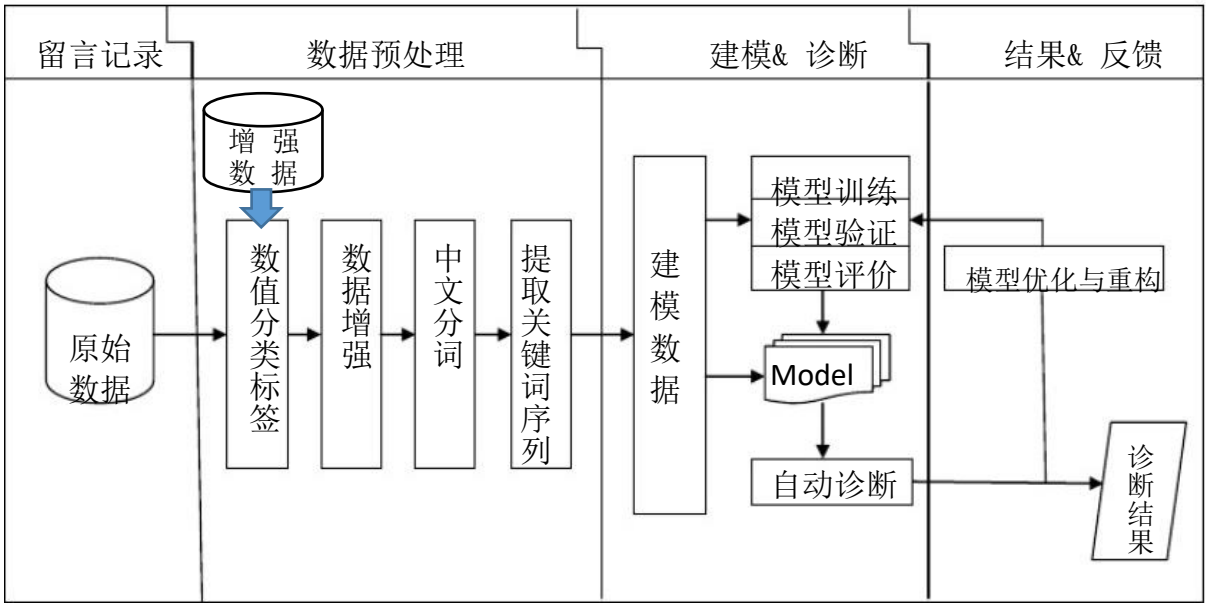

图1 问题一主要流程图

4.2 具体步骤

结合问题一总体流程，对每一步骤进行详细说明。

4.2.1 数据预处理

1. 数据增强

卷积神经网络需要大量的训练数据才能获得比较理想的效果，而所给的数据集存在数据量小，类别分布不均衡问题。

ICLR2019 workshop介绍的EDA（Easy data augmentation）（参考文献：《EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks》）技术是一种基于随机词替换的数据增强方法。EDA主要包括：同义词替换、随机插入、随机交换和随机删除四种操作，经研究表明，原始数据集经过EDA变换，不仅扩大了数据量，同时还保持了原有的标签，有效扩大了原始样本集的信息容量。

因此，本文采取EDA技术对原始数据集进行增强，同时设定替换比例α为0.05，每句的拓展句为8，得到了82872条数据，有效地扩大了信息容量。

2. 创建数值分类标签

统计数据中留言记录的文本标签数量，赋予每类文本标签新的数值标签，构建新的

标签体系结构，为建模可用。具体如下表所示：

文本标签	数值标签
劳动和社会保障	0
环境保护	1
商贸旅游	2
城乡建设	3
卫生计生	4
教育文体	5
交通运输	6

表1 留言一级标签分类表

3.中文分词

鉴于中文的特殊性，需要采取中文的分词算法对其进行分词，本文采取的是python的jieba词库对每一条留言记录的留言主题和留言详情的进行中文分词。同时把一些未登录词添加整合到jieba词库里，使得词库更加完整，分词更加准确。针对停用词，中文中的停用词如“的、和、等一下”等不能带来任何信息量，这里使用了停用词表将之去除。

4.提取关键词序列

本文在分词后首先运用TF-IDF算法对留言详情文本进行了关键词提取，TF-IDF的计算方法实际上是词频与逆文档频率的乘积，即TF×IDF。TF和IDF的计算公式如（1）和（2）所示。

$$\mathrm{TF(t,d)}=\frac{f(t,d)}{\sum_k f(w_k,d)} \qquad (1)$$

$$\mathrm{IDF}_t=\log\left(\frac{N}{1+df_t}\right) \qquad (2)$$

其中 $f(t,d)$ 表示词条 t 在文档中出现 d 出现的次数，df_t 表示语料库中包含词条 t 的文档数量，N 表示语料库中全部的文档数量。

首先提取出每条留言详情中排名前50的关键词，整合所有每条留言详情的关键词形成关键词库。

然后，建立一个2000词的token词典。先对文本中单词出现的次数做统计并排序，从而将每一条留言的文本分词列表替换成数字列表。并截长补短使得所有记录的关键词序

列的长度为50。针对82872条留言记录最终得到一个82872*50的数组，为模型做数据准备。

5.划分测试训练集

由于不同类别留言记录的数量差距大，在十折交叉验证的基础上，采用分层采样，、的方法、依照标签的比例来抽取数据，用于交叉验证，提高算法的准确性。

数据预处理后得到的数据格式如下表：

Data	Shape	Data	Shape
x_train	[58010,50]	y_train	[58010,1]
x_test	[24862,50]	y_test	[24862,1]
x_total	[82872,50]	y_total	[82872,1]

表2 预处理数据结果表

4.2.2 建模和诊断

1.搭建卷积神经网络

卷积神经网络是一种带有卷积结构的深度神经网络，卷积结构大大减少了深层网络占用的内存量，它的全值共享有效减少了网络的参数个数，缓解了过拟合问题。本文使用keras框架搭建了两层卷积神经网络，提高神经网络的准确率。卷积层和池化层是卷积神经网络特征提取的核心模块，该模型采用了自适应移动估计算法（Adam）对网络中的权重参数逐层反向调节，使得损失函数值最小，通过不断的迭代训练提高神经网络的精度。

全连接层对提取的特征进行非线性组合以得到输出，这里采用了归一化指数函数softmax输出分类标签。

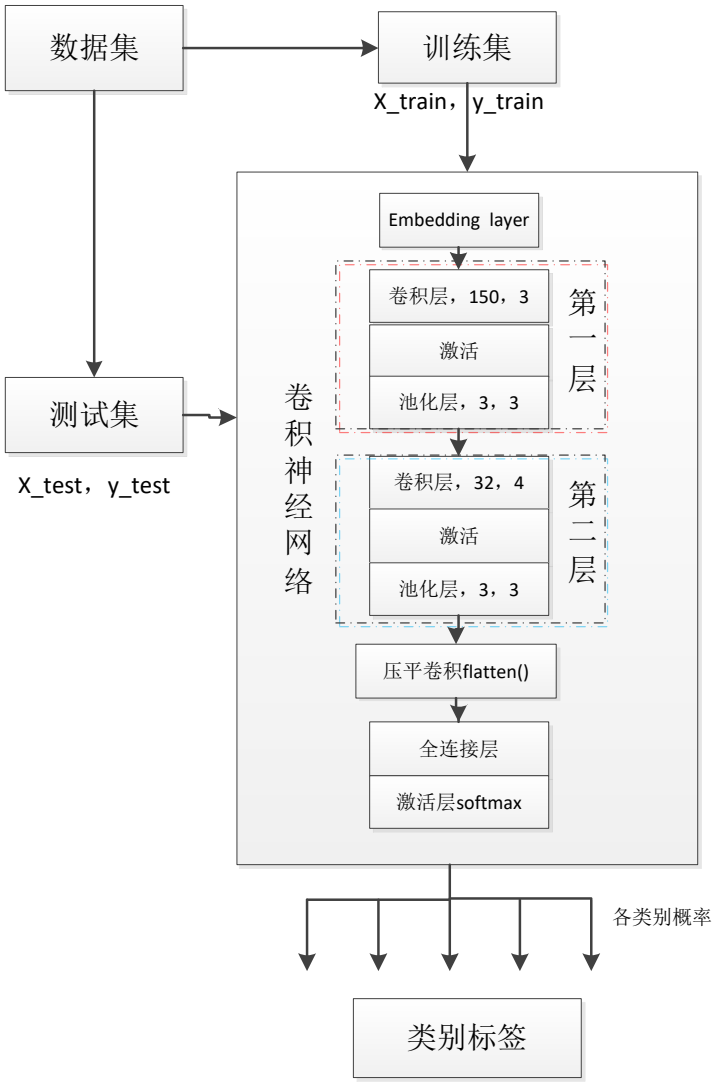

图2 卷积神经网络详情图

1)embedding层

直接初始化embeddings,基于语料通过训练模型网络来对embeddings进行更新和学习。从而将输入的数字列表转换为词向量，最后得到58010x100的词向量矩阵，每个词的维度为100。

2）卷积层

经过embedding层之后，每一条留言记录（留言详情）由高质量特征线性表示,将其输入卷积层，对输入数据进行特征提取，包含多个卷积核，每个卷积核对应了一个权重系数和一个偏差量，类似于一个前馈神经网络的神经元，实验过程中，第一层的卷积核大小为3，第二层卷积核的大小为4，卷积核在工作时，会有规律地扫过输入特征，在感受野内对输入特征做矩阵元素乘法求和并叠加偏差量。同时采取relu作为激励函数不断迭代。

3）池化层

池化层进行降维操作，降低文本的向量维度，同样也是一层特征选取和信息过滤，由池化大小、步长和填充控制来确定池化区域，实验过程中，针对两层池化层，取池化大小pol_size=3，步长stride=3，填充控制padding=same。

4）全连接层

神经网络的最后一层，采用全连接层的方式，第二层K_max池化层处理后的文本特征向量经过矩阵的concat和reshape之后变成一维数组，送入Softmax分类器，计算类别概率，预测输出分类标签。

5）相关参数调节

通过对dropout和batch_size参数的不断调节，降低模型的过拟合现象，同时设置合理的迭代次数epoch，以获得较好的收敛效果。在模型的训练中设定了validation_split作为从训练集中选取作为验证集的比率。具体参数如下表所示：

参数	取值
dropout	0.2
Batch_size	256
epoch	20
validation_split	0.2

表3 具体参数取值

2. 训练评估

1）模型训练

训练集上经过20次迭代可得到模型的最佳分类效果为98.84%，其中准确率和误差如图所示：

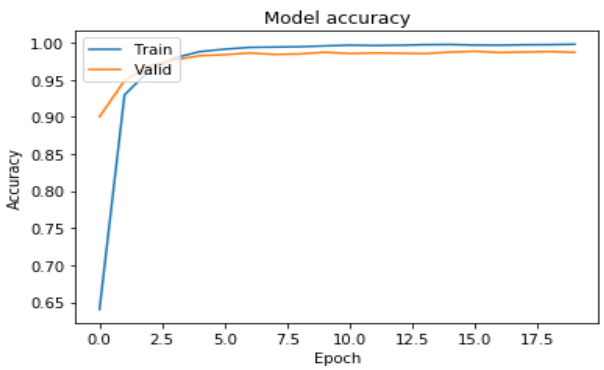

图3 模型准确率

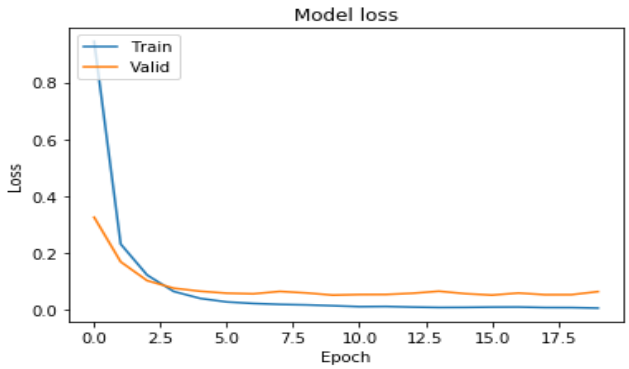

图4 模型误差率

2）模型评估：采用F-score方法对模型进行评价。

根据分类结果建立混淆矩阵：

判别	真正属于该类	实际不属于该类
判别属于该类	A	B
判别不属于该类	C	D

表4 混淆矩阵

（1） 计算精确率，精确率是分类器正确判断属于该类的样本数与分类器判断属于该类的样本总数的比值，则针对第i类，有：

$$P_i = \frac{A_i}{A_i + B_i} * 100\% \quad (3)$$

（2） 计算召回率，召回率是分类器正确判断为该类的样本数与分类器真正属于该类的样本总数的比值，则针对第i类，有：

$$R_i = \frac{A_i}{A_i + C_i} * 100\% \quad (4)$$

（3） 结合基本指标衍生出综合指标F_1测度值，针对第i类，有：

$$F_{1i} = \frac{(\beta^2 + 1) P_i R_i}{\beta^2 P_i + R_i} \quad (5)$$

β为调整参数，用于调整精确率P和召回率R在计算时的比重，这里取β=1

将n类F1测度值求和平均，得到分类器的综合评价指标：

$$F_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{2 P_i R_i}{P_i + R_i} \quad (6)$$

4.2.3 结果和反馈

针对该模型，计算出各分类的F_1值如下表所示：

类别	标签	F1值
劳动和社会保障	0	0.990943
环境保护	1	0.994479
商贸旅游	2	0.9847
城乡建设	3	0.98799
卫生计生	4	0.985362
教育文体	5	0.991968
交通运输	6	0.979813

表5 各分类F_1值

代入数据，得到最终分类器的综合评价指标 $F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_i R_i}{P_i+R_i}$=0.98642

基于深度学习，采用卷积神经网络构建的分类器模型和传统的分类方法相比，分类的精确度更高，通过不断的迭代学习，精确度能够达到95%以上，而一般传统的机器学习模型的准确率在85%以下，对于数据量大、特征难以提取的文本数据，用深度学习的算法更为合适，卷积神经网络能够提取更高层的特征，提高准确率。将模型训练结果的权重文件保存，预测数据时可以直接导入得到预测分类结果，更为高效。

五、 问题二

5.1主要流程

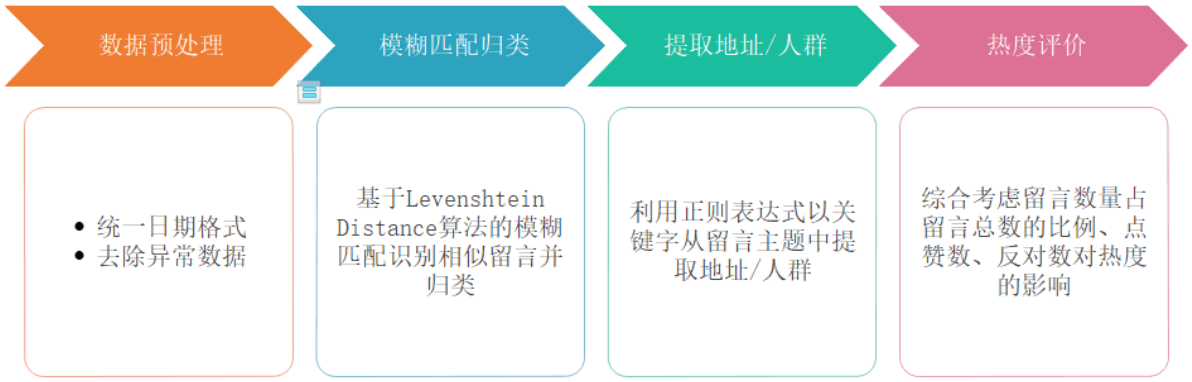

图5 问题二主要流程

5.2具体步骤

5.2.1 数据预处理

将原始数据中出现的不同的日期格式统一，并精确到日，为时间范围的计算做准备。清除留言数据中的异常数据（留言中只出现标点符号、留言内容为空的记录等）。

5.2.2 模糊匹配归类

中文自然语言地址的非结构性特征显著，大多地点由具体的命名实体构成（如某乡镇某道路、具体店名等），句法对词语约束性差，词语搭配随意性很强，没有遵循统一的语法规则。观察数据可以发现，留言记录中没有显示具体真实的地理信息，因此无法得到大规模已标注的真实地址语料库进行地址匹配。若要使用机器学习的方法，需要手工标注大量地点名称进行训练学习，不断迭代测试，而数据总量仅有 4000 多条，难以达到很好的效果。且考虑到地址文本的不规范性，建立完备的地名词典也难以实现。

由于留言主题表达精简，其中地址的字符数占比很高，又考虑到地址表达的随意性，

可以采用模糊匹配的方法将相似留言进行归类，允许了一定的容错性。具体算法步骤如下:

1.Levenshtein Distance 算法

模糊匹配基于 Levenshtein Distance 算法计算字符串之间的编辑距离，一般来说，编辑距离越小，两个字符串的相似度越大。

设有两个字符串 s_1、s_2，字符串 s_1 长度为 m，字符串 s_2 长度为 n，建立一个(m+1)*(n+1)大小的矩阵，初始化第一行第一列，使得 $d_{[i,0]}=i$，$d_{[0,j]}=j$，求出矩阵中其他元素。有下列公式：

$$d_{[i,j]}=\begin{cases}0 & i=0 \text{ or } j=0\\ min(d_{[i-1,j]}+1, d_{[i,j-1]}+1, d_{[i-1,j-1]}) & x_i=y_i\\ min(d_{[i-1,j]}+1, d_{[i,j-1]}+1, d_{[i-1,j-1]}+1) & x_i\neq y_i\end{cases} \quad (7)$$

其中 $d_{[i-1,j]}+1$ 代表字符串 s_2 插入一个字母，$d_{[i,j-1]}+1$ 代表字符串 s_1 删除一个字母，当 $x_i=y_i$ 时，不需要代价，与上一步 $d_{[i-1,j-1]}$ 代价相同，否则+1，$d_{[i,j]}$ 取以上三者中最小的一项。

$d_{[m,n]}$ 即为字符串 s_1、s_2 之间的编辑距离。

2.模糊匹配

利用 python 的 FuzzyWuzzy 模糊字符串匹配工具包，基于 Levenshtein Distance 算法进行留言主题模糊匹配，将结果大于 50 的留言归为一类，计算时间范围并对每一条留言贴上类别标签作为区分。

5.2.3 地点识别

本文使用正则表达式提取地址，利用正则表达式以关键字“省、市、区、街、道等”多次匹配，从各条留言主题中提取地址，部分结果如下所示：

留言主题	地址
建议 A 市 1 路车的终点站延伸到碧沙湖的地铁站	A 市 1 路车的终点站
A1 区 A 市乐敏食品有限公司严重侵犯员工的合法权益	A 市 A1 区乐敏食品有限公司
蒙华铁路 A9 市段一季度工程款还未拨付	A9 市
咨询 A9 市高铁站选址的问题	A9 市高铁站
A9 市淮川街道税务中心机构解散，人员分流面临失业	A9 市淮川街
请落实潜江至韶关输气管道 A9 市段农田临时租赁后的恢复工作	A9 市
A9 市永安供电所强占我田德多年	A9 市永安供电所
关于 A9 市河防洪（曙光垸退耕还垸）的建议	A9 市
A9 市河畔数万业主每天在重油烟中度日	A9 市
A9 市河旁一枝黄花泛滥	A9 市

表6 部分留言主题地址

5.2.4. 热度评价

1.热度评价指标

由于数据集中分布在 2019 年（98.26%），其他年份留言数据少且时间跨度大，因此各类留言占各自时间范围内留言数量的比例情况不能反应热度情况。且热点问题可以是短时间内爆发的问题，也可以随着时间变化愈演愈烈的问题，因此忽略时间范围对各类别的热度指数的影响，综合考虑留言数量占留言总数的比例、点赞数、反对数对留言热度的影响。

设共有 n 个类别，第i类留言的留言数量为a_j(天)、点赞数为e_j、反对数为d_j，总留言数为 S、总点赞数为 E、总反对数为 D，热度指数为h_j。

针对第i类留言：

$$h_j = 10(\frac{a_j}{S} + \frac{e_j - d_j}{E + D})*100\% \qquad (8)$$

根据公式计算各类留言热度指数，进行热度排名，根据热度排名赋予问题 ID,得到热度指数排名前五的热点问题，根据正则表达式提取出来的地址总结归纳各类别地点，对于地址描述粗略的类别（例如：A 市、C 市）结合主题识别人群，并进行各类问题描述得到热点问题如下：

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	181.39%	2019/1/29 2019/5/28	至	A市A4区	A市A4区58车贷案
2	2	162.57%	2019/5/5 2019/9/10	至	A市A5区k9县	A市A5区K9县五矿万房屋存在质量问题
3	3	137.08%	2019/1/3 2019/11/27	至	A市小区业主	A市小区业主反映小区房屋问题
4	4	53.76%	2019/8/23 2019/9/5	至	A 市绿地外滩小区	A市绿地外滩小区高铁对小区的影响问题
5	5	18.82%	2019/6/19 2019/6/19	至	A 市富绿物业丽发新城	A市富绿物业丽发新城强行断业主家水

表7 热点问题表

2.热点问题明细

由于问题ID与留言类别标签一一对应，识别出各条留言的问题ID即可得到热点问题留言明细表。部分数据如下图所示：

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	194343	A000106161	承办A市58车贷案警官应跟进关注留言	2019/3/1 22:12:30	经侦并没有跟进市领导	733	0
1	217032	A00056543	严惩A市58车贷特大集资诈骗案保护伞	2019/2/25 9:58:37	股东、苏纳弟弟苏吕	790	0
1	218132	A000106090	再次请求过问A市58车贷案件进展情况	2019/1/29 19:15:49	与知办案警官毛浚时，	0	0
1	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	案情消息总是失望，因	821	0
1	226265	A000106448	恳请A市经侦公正办理58车贷案件，还我们受害人一个公道	2019/5/28 15:08:51	侦，但这样的经侦我们	3	0
2	208069	A00094436	A5区五矿万境K9县的开发商与施工方建房存在质量问题	2019/5/5 13:52:50	质量是否能符合五矿	2	0
2	208636	A00077171	A市A5区汇金路五矿万境K9县存在一系列问题	2019/8/19 11:34:04	狗咬人，请问有人对养	2097	0
2	234086	A00099869	A市五矿万境K9县房子的墙壁又开裂了	2019/6/20 9:30:44	商不支持赔偿，考虑到	6	0
2	275491	A00061339	A市五矿万境K9县负一楼面积缩水	2019/9/10 9:10:22	购房时也未做任何说	0	0
3	200316	A00018085	再次反映A市金茂府一房二卖问题	2019/10/24 15:58:46	行为。A市住建委在201	0	0

图6 部分热点问题明细

六、 问题三

6.1 主要流程

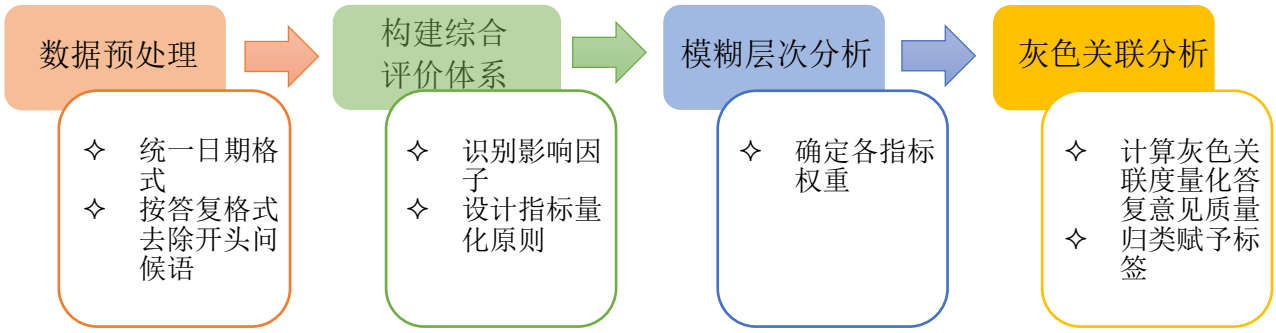

图7 第三题主要流程图

6.2 答复意见综合评价体系

在观察数据并进行预处理之后，本文拟从答复意见的相关性、完整性、可解释性、时效性、丰富性五个维度，构建包含5个要素、6个指标的评价体系，对答复意见进行定量评价。

6.2.1设计指标量化原则

1. 相关性

答复意见的相关性主要反映是否与群众提问的问题相关，是能否真正解决群众问题的关键所在。关于相关性指标的定量化，我们使用Python软件做了如下处理。

用 q_i 表示第 i 条留言记录中的留言详情，r_i 表示第 i 条留言记录中的答复意见；$k1_i$ 表示第 i 条留言记录中的留言详情的关键词文本向量，$k2_i$ 表示第 i 条留言记录中的答复意见的关键词文本向量；s_i 表示第 i 条留言记录中答复意见与留言详情的相似度。

1）中文分词

首先，对原始数据进行预处理，利用经问题一整合的更完整的jieba库和停用词表，对每一条数据的留言详情 q_i 和答复意见 r_i 进行分词、去除停用词。

2）IT-IDF算法提取关键词向量

利用Gensim库中的TF-IDF模型，计算对应的 q_i 和 r_i 中各分词的词向量得到各分词权重TF-IDF值，取出留言详情中TF-IDF值大于8的词语得到关键词文本向量 $k1_i$，取出答复意见中TF-IDF值大于10的词语得到关键词文本向量 $k2_i$。部分数据如图所示：

0	['连廊', '消防', '天楚', '拆除', '业主', '中苑', '高层建筑', '封堵', '门窗', '相邻', '消防安全', '封闭', '单元', '法规', '部分', '高发期', '落成', '物模', '潜', '据为己有', '廊', '三栋', '设计', '楼', '强制', '小区', '保证', '报告'
1	['开元', '东路', '警务', '辅快', '塘处', '另东', '主动脉', '东线', '深坑', '螺丝', '大坑', '交通部门', '路况', '向东', '来左右']
2	['乡', '公路', '最烂', '差是', '将宁', '宁朱', '他处', '特征', 'a县', '朱', '好路', '朱良桥', '显著', '破烂', '年内', '宣布']
3	['黄茅园', '小城镇', '镇', '邮政储蓄', '李', '龙潭', '乡镇', '花岗石', '产业基地', '资源', '山泉水', '农业银行', '请求', '取款机', '葛竹', '信用社', 'L县', '网点', '石材', '两边', '自来水厂', '南大门', '自动', '几条', '优势', '辖区
4	['教师', '退休', '具体', '分配', '杨', '还落', '教书', '安沙派', '听说', '局所', '住房', '长期以来']
5	['紫晶', '城', '许多', '每天', '马路', '上班', '拨原', '设个', '上千', '站台', '高速公路', '栏杆', '解决', '跨']
6	['高压', '黄星', '辐射', '肯定', '塘路', '入地', '螺丝', '楼盘', '高压线', '看上去', '大道', 'KV', '西边', '这段', '那条']
7	['养老金', '月份', '靠', '元月份', '农村', '周岁']
8	['拆迁', '安置', '泼河']
9	['人行天桥', '地下通道', '轨捷', '几棋', '几横', '兔子', '行人', '路口', '星沙', '安全感']

图8 前十条留言详情关键词文本向量

0	['封堵', '连廊', '户型图', '天楚', '业主', '中苑', '消防', '消防大队', '经开区', '上门', '窗户', '模型', '梨江', '主卧', '认为', '偷过', '个人隐私']
1	['沉坑', '开元', '东路', '八线', '路面', '东', '沉降', '放晴', '铺油', '最快', '天内', '天一', '大坑', '先对']
2	['交通运输', '前期工作', '厅']
3	['黄茅园', '金融机构', '建设', '镇', '领导', '网点', '两级', '镇党委', '全县', '难办', '有目共睹', '提意见', '才行', '存取款', '产业基地', '将来', '四在', '公益事业', '仅仅']
4	['教师', '公职人员', '常委', '住房问题', '县委', '其他', '布局', '教育局', '统筹', '全县', '离退休', '优越感', '尊严', '以县', '购房']
5	['路时', '汽配城', '西路', '公交线', '开元', '紫晶', '物贸', '楚绣城', '路口', '改线', '后路']
6	['电力设施', '电力', '楼盘', '近距离', '保护权', '输电', '不良反应']
7	['养老金', '发放', '城乡居民', '元年', '月份', '首发式', '养老保险金']
8	['货币', '安置', '限价', '拆迁', '房', '小组', '村民', '拆迁户', '水塘', '垸', '我镇', '梅', '丽', '塘村', '万家', '安沙镇', '北路', '谭坊', '村毛塘', '月始', '方案', '人民政府', '书面材料', '三合', '县政府']
9	['城市', '天华', '人行', '小学', '地下', '基础设施', '通道', '预见', '路龙喜', '路处', '路东业', '星沙盼', '已建好', '路口']

图9 前十条答复意见关键词文本向量

3）模糊匹配

接下来，依据 Levenshtein Distance 算法计算两个序列之间的差异。Levenshtein Distance 算法，又叫 Edit Distance 算法，是指两个字符串之间，由一个转成另一个所需的最少编辑操作次数。许可的编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。一般来说，编辑距离越小，两个串的相似度越大。

使用基于Levenshtein Distance算法的python模糊字符串匹配工具包——FuzzyWuzzy，输入关键词列表进行留言详情与答复意见关键词的模糊匹配，这里采用了忽略顺序匹配方法，去除关键词出现的先后顺序的影响，得到相似度 s_i。由于得到的 s_i 分布过于密集，对其采取处理，依据公式：

$$s_i = s_i(2 - \frac{s_i}{100}) \tag{9}$$

进行三次循环处理并将结果取百分数形式得到最终的相关度 s_i。处理之后，不改变数据的相对关系，扩大数值便于计算且可以使结果值更加平稳。

2. 完整性

由答复次数和答复格式两个二级指标反映。

在整理数据时发现，针对部分群众反映的留言问题，网政平台工作人员无法一次性给与完整性答复，需交由相关部门解答或需要进行调查核处再次完善答复。每次答复均以“＊＊＊＊＊＊，您好/你好”格式开头。答复的次数很大程度上反映了答复意见的完整性，因此将其纳入二级指标。

除此之外，完整性主要衡量答复意见是否符合某种规范，进行数据预处理时不难发现，答复意见主要有四种格式，针对大部分答复（1740条答复），有非常规范的留言答复以“＊＊＊＊＊＊，您/你好！＊＊＊＊＊＊回复/答复如下：”开头；另一种相对规范的格式（127条答复）以含有关键字“如下”的答复开头，如“＊＊＊＊＊＊，您/你好！＊＊＊＊＊＊解释如下/如下汇报/如下回复：”；还存在一种普通回复格式（704条）以“＊＊＊＊＊＊，您好/你好！”开头，然后进入正题；最后一类（245条），答复开头无问候语，直接直奔主题进行回复。现针对答复意见的格式特征，将答复格式划分为非常规范、比较规范、一般、较差四个等级。

答复格式	等级
非常规范	3
比较规范	2
一般	1
较差	0

表8 答复格式等级划分表

3.可解释性

可解释性是合理的说明问题的原因、解决问题的依据，尽可能阐述自己观点的可行性和合理之处。衡量网政平台答复意见的可解释性，主要依据答复意见是否引用相关法律条例、政策规定对群众问题进行解答。将答复意见中出现关键词“根据”、“政策”、“规定”、“按照”、“《》”的可解释性记为1，没有出现上述关键词的记为0。

4.充实性

答复意见的信息充实度由答复意见的有效长度占比来反映。去除文本的标点符号及空格，去除答复格式为非常规范、 比较规范、一般的三类答复文本开头问候语得到答复意见的有效长度。一段话中有效文字越多，其表达的信息越充实。

文本充实度=有效长度/答复意见总长度

5.时效性

时效性很大程度上反映了网政平台的管理水平，网政平台上，一问一答之间，政情得到沟通，疑问得到解答。网民反映的大部分问题都是关系到人民群众切身利益的事情，留言回复是否及时至关重要，因此网民咨询类诉求响应时效是评价留言答复质量的又一重要指标。对数据的时间格式进行统一化处理，精确到日，针对一次答复的留言，将答复时间减去留言时间得到各条留言的响应时长（以天为单位）。针对两次答复的留言，取第二次解决疑问之后完善答复的时间减去留言时间得到响应时长。

一次答复的响应时长=答复时间-留言时间

两次答复的响应时长=第二次答复时间-留言时间

将响应时长t按10为步长划分为从1～20共20个等级，如下表所示：

响应时长	等级
$0 \le t < 10$	20
$10 \le t < 20$	19
$20 \le t < 30$	18
…	…
$180 \le t < 190$	2
$t \ge 190$	1

表9 响应时长等级划分表

6.2.2构建答复意见质量综合评价体系

根据各要素指标构建的答复意见综合评价体系及各指标内涵如表所示：

	一级指标	二级指标	指标内涵
网政平台答复意见质量综合评价体系	相关性	文本相似度 w_1	答复意见与留言详情文本关键词的相似度
	完整性	答复次数 w_2	网政平台工作人员答复回应的次数
		答复格式 w_3	划分为四个等级，体现答复内容的规范性。
	可解释性	事实依赖 w_4	依赖于政策法规、规定事实的程度
	充实性	文本充实度 w_5	有效长度与答复意见总长度的比值
	时效性	响应时长 w_6	划分为16个等级，反映答复时间与留言时间的时间间隔

表10 答复意见综合评价体系

根据答复意见综合评价体系计算得到的部分数据结果如下表所示：

留言编号	响应时长	回复次数	回复格式	有效长度	相似性	可解释性
2549	19	1	3	0.13	0.584	0
2554	19	1	3	0.53	0.7661	0
2555	19	1	3	0.59	0.6268	0
2557	19	1	3	0.5	0.6935	0
2574	19	1	3	0.52	0.6789	0
2759	17	1	3	0.52	0.5336	0
2849	16	1	3	0.5	0.4746	0
3681	18	1	3	0.63	0.7795	1

表11 部分答复意见指标评价结果

6.3 模糊层次分析

具体步骤如下：

1. 建立指标层次结构

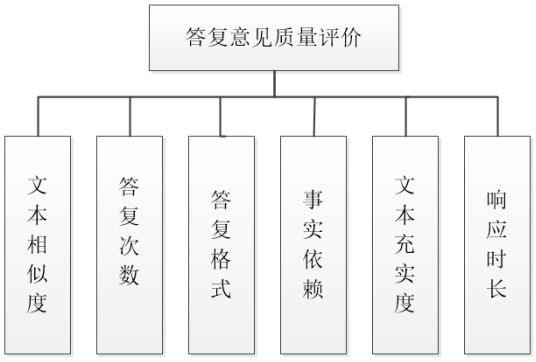

图 10 答复意见质量影响因素

2. 构建判断矩阵：

通过对各因素之间两两对比的关系，构建因素之间的判断矩阵，表示为：

$$F = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ f_{31} & f_{32} & \cdots & f_{3n} \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{bmatrix} \tag{10}$$

采用下表进行数量标度：

标度	含义
0.5	表示两个因素相比，具有相同重要性
0.6	表示两个因素相比，前者比后者稍重要
0.7	表示两个因素相比，前者比后者明显重要
0.8	表示两个因素相比，前者比后者强烈重要
0.9	表示两个因素相比，前者比后者极端重要
0.1,0.2, 0.3,0.4	若因素 i 与因素 j 的重要性之比为 a_{ij}，那么因素 j 与因素 i 重要性之比为 1- a_{ij}

表 12 相关标度

3.转换为模糊一致矩阵：

（1）先将比较判断矩阵按行和列进行求和，即：

$$f_i = \sum_{j=1}^{n} f_{ij} \qquad i = 1,2,3,\cdots,n$$

$$f_j = \sum_{i=1}^{n} f_{ij} \qquad j = 1,2,3,\cdots,n$$

f_i, f_j 分别代表第 i 行和第 j 列的和，f_{ij} 表示第 i 行的第 j 列数值。

（2）再进一步对矩阵中每个元素进行变换，得到模糊一致矩阵，变换方法如下：

$$f_{ij} = \frac{(f_i - f_j)}{2n} + 0.5 \qquad i, j = 1,2,\cdots,\mathrm{n} \tag{11}$$

（3）对得到的模糊一致矩阵进行一致性检验

在实际决策分析中，由于问题的复杂性，人们在问题的认知上往往是不够全面的，这就使得判断矩阵不具有一致性，因此在得到模糊一致矩阵时，我们还需要对矩阵进行一致性检验，检验和调整的方法如下：

选择矩阵中相对来说数据准确度最高的行向量作为基准，将该行向量中的每个元素去减去其他行向量的元素，若差值不是定值，则对对应行的元素进行调整，不断重复上述步骤，直到满足所有行向量相减差值为定值这一条件。该模糊矩阵便通过了一致性检验。检验结果如下：

4.确立各个指标的权重

一致性检验通过后可根据模糊一致矩阵计算各指标 a_i 对答复意见质量的贡献权重。方法如下：

$$w_i = \frac{1}{n} - \frac{1}{2\alpha} + \frac{1}{n\alpha}\sum_{k=1}^{n} f_{ik} \qquad i = 1,2,\cdots,n \tag{12}$$

n 为影响指标的个数，α 为调节参数，通常取 $\alpha = \frac{(n-1)}{2}$

5．代入数据求解

将答复意见的相关数据代入，此时 α =2.5，得到各指标所占权重如下：

$$\begin{cases} w_1 = 0.206667 \\ w_2 = 0.123333 \\ w_3 = 0.136667 \\ w_4 = 0.2 \\ w_5 = 0.173333 \\ w_6 = 0.16 \end{cases}$$

6.4 灰色关联分析

灰色关联度分析，是一种多因素统计分析的方法。根据因素之间发展趋势的相似或相异程度，即“灰色关联度”，作为衡量因素间关联程度，寻找各因素之间的定量数值关系。算法步骤如下：

1.无量纲化处理

由于各指标因素的物理意义不同，数据的量纲和数量级也不尽相同，不便于比较，即使比较了也难以得到正确的结果。为了保证结果的可靠性，在进行灰色关联度分析时，一般要对原始数据进行无量纲化的处理。这里对原始数据进行均值化处理。

设影响因子个数为 n，数据集中共有 m 行数据。

$$X'_{ik} = \frac{X_{ik}}{X_i} \qquad i = 1,2,...,\mathrm{m},\ \ k = 1,2,...,n \tag{13}$$

2.参考序列选取

用参考序列来反映行为特征，选取 j 个影响因素指标后的最优值作为参考序列，其他作为比较序列，设参考数列为 Y，则

$$Y = \{Y(k) \mid k = 1,2,...n\}$$

*Y(k)*为第k个指标的最优值

同样进行均值化无量纲处理;

$$Y'(k) = \frac{Y(k)}{Y} \qquad k = 1,2,...,n \tag{14}$$

3. 灰色关联系数

关联程度是曲线间几何形状的差别程度。将曲线的差值大小作为关联程度的衡量尺度，对于一个参考数列$Y^{'}(k)$，有若干个比较序列$X^{'}_1, X^{'}_2, X^{'}_3, ..., X^{'}_m$。

令$M=\{1,2,...,m\}$，$N=\{1,2,...,n\}$，有：

$$\triangle\min=\min_{i\in M}\{\min_{k\in M}|Y^{'}(k)-X^{'}_{i(k)}|\}$$

$$\triangle\max=\min_{i\in M}\{\max_{k\in N}|Y^{'}(k)-X^{'}_{i(k)}|\}$$

$$\Delta_{oi}(k)=|Y^{'}(k)-X^{'}_{i(k)}| \tag{15}$$

用ξ_{ik}的大小描述$X^{'}_i$与$Y^{'}_i$在k点的关联系数，则：

$$\xi_{ik}=\frac{\triangle\min+\rho\triangle\max}{\Delta_{oi}(k)+\rho\triangle\max} \qquad k=1,2,...,n \tag{16}$$

其中ρ为分辨系数，一般在0~1之间取值，取ρ=0.5

4. 灰色关联度

将灰色关联系数与各个影响因子权重$W=(w_1,w_2,...,w_n)$相乘，满足$\sum_{k=1}^{n}w_k=1$，加权计算X_i与Y的总关联度r_i，计算公式如下：

$$r_i=\sum_{k=1}^{n}w_k\xi_{ik} \qquad i\in M,k\in N,w_k\geq 0 \tag{17}$$

5. 排序归类

根据灰色关联度大小进行排序，灰色关联度越大，代表答复意见质量越高。这里按答复意见灰色关联度从大到小降序排列，彰显答复意见的有用性。

取出灰色关联度排名前10的结果如下表所示：

留言编号	关联度
133536	0.973071
58535	0.944404
124085	0.927656
129136	0.926071
133680	0.925816
39565	0.924286
119872	0.923566
5439	0.923427
84789	0.922134
114582	0.922113

表 13 灰色关联度 top10

根据灰色关联度的大小，将答复意见质量划分为高、中、低三个等级，具体如下表所示：

灰色关联度	答复意见质量
$0.9 \le r_i \le 1$	优
$0.7 \le r_i < 0.9$	良
$0 \le r_i < 0.7$	差

表 14 答复意见质量等级划分

根据上表规则将答复意见归类赋予标签，得到部分结果如图所示：

留言编号	留言主题	答复意见	响应时间(天)	等级
133536	L6县金塘村村民生活困难，请求政府帮助！	“UU0081558”：你好，对于你在平台反映的问题，现将有关情况回复如下：1、至今我乡还未收到异动2018年农村低保的文件通知。根据2017年5月L6县开展农村低保和社会保障兜底脱贫对象认定清理整顿的规定，我县在低保分类标准中将因重病、重残、灾害及意外事故等原因造成家庭成员基本或部分丧失劳动力，家庭支出负担沉重影响基本生活的家庭且生活条件特别恶劣，贫困程度较大需长期保障的家庭评定为二类低保。根据《关于贯彻落[政府发文]30号文件进一步做好全省医疗救助工作的通知》[政府发文]25号）文件规定：优先将儿童先天性心脏病、儿童白血病、乳腺癌、宫颈癌、肝移植、肾移植、恶性肿瘤、重性精神病（精神分裂、分裂性感情障碍、偏执性精神障碍、双相情感障碍、癫痫所致的精神障碍、严重精神发育迟滞）、艾滋病机会性感染等9种疾病纳入重特大疾病救助范围。杨彪虎患得是心肌梗塞、脑梗塞，不属于重特大疾病范围，且杨彪虎子女都已满16周岁，具有正常劳动能力，故其一户按现有政策不符合享受低保的条件。2、根据文件规定，需经本人申请、民主评议、初评结果公示、乡镇复查审核、家庭经济状况核对、县级审批、审批结果公示等程序，才能最终确定是否可以享受农村低保。因杨彪虎多年在外，且2018年以前身体健康，之前未曾申请过低保。只因杨彪虎在1月份突发心肌梗塞，才提出希望申请成为低保的诉求，不符合评定低保的工作程序。3、因2018年异动农村低保的工作还未实施，待2018年上级异动低保的方案启动，杨彪虎可以按程序申请。4、杨彪虎在生病住院期间，我乡民政办依据其本人申请，已按规定给予一定的临时困难救济。二0一八年五月三日网友：您好，您所反映的问题已于4月27日转澧滨侗族苗族乡人民政府调查核实。2018年5月2日	10	优
76358	投诉F市通岳电气有限公司违反劳动法	您的留言已收悉。关于您反映的问题，已转市人社局调查处理。	16	良
12274	请求解决蒋垄家火车噪音问题	网友：您好！留言已收悉	21	差

图 11 答复意见质量评价表

七、 结论

为了将群众的留言进行归类，我们首先对附件2中的数据进行相应的预处理，先去除无效字符，再通过jieba分词和TF-IDF算法提取关键词，然后用token字典将文本转化为数字列表。再使用embedding层将数字列表转换为向量矩阵，从而使用卷积神经网络模型进行分类器的训练，最后得到了群众留言一级标签分类模型。经验证，模型的F1得分为98.87%，分类效果较优。从而得到**附件：群众留言分类模型.h5。**

在热点问题的识别上，我们首先对附件3中的数据进行了预处理，统一时间格式，去除无效字符和缺失值，然后用基于Levenshtein Distance算法的模糊匹配对留言主题进行归类，识别出相似留言，再通过正则识别出留言的地点和人群。最后综合考虑留言数量占留言总数的比例、点赞数、反对数指标建立热度评价体系，识别出热点问题。得到了**附件：热点问题表.csv与热点问题明细表.csv。**

为了对政府的答复意见进行评价，我们首先对附件4里的答复意见数据进行了预处理，去除缺失值，统一时间格式，去除无效字符。然后我们从答复意见的相关性、完整性、可解释性、时效性、丰富性五个维度，构建了文本相似度、答复次数。答复格式、事实依赖、文本充实度、响应时长6个量化指标。通过模糊—层次分析法确定指标权重，再利用灰色关联分析法计算出每条答复意见的关联度并排序，最后根据关联度将答复意见划分为优、良、差三个等级，从而得到附件：**回复意见等级标注结果.csv**

八、 参考文献

[1]刘志荣. 电子政务的数据挖掘研究[J]. 广东技术师范学院学报, 2008(03):12-14.
[2]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.
[3]穆瑞,张家泰.基于灰色关联分析的层次综合评价[J].系统工程理论与实践,2008(10):125-130.
[4]张谦,高章敏,刘嘉勇.基于Word2vec的微博短文本分类研究[J].信息网络安全,2017(01):57-62.
[5]陈巧红,王磊,孙麒,贾宇波.卷积神经网络的短文本分类方法[J].计算机系统应用,2019,28(05):137-142.
[6]郭顺利,张向先,李中梅.面向用户信息需求的移动O2O在线评论有用性排序模型研究——以美团为例[J].图书情报工作,2015,59(23):85-93.
[7]赵卫锋,张勤.非结构化中文自然语言地址描述的自动识别[J].计算机工程与应用,2016,52(23):19-24.
[8] Wei J W , Zou K . EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks[J]. 2019.
[9] 薛彬, 陶海军, 王加强. 针对民生热线文本的热点挖掘系统设计[J]. 中国计量大学学报, 2017(3).
[10] 陈醉. 地方政府网上回应的现状及影响因素研究[D].