“智慧政务”中的文本挖掘应用

摘要

近年来，随着大数据和人工智能时代的到来，网络问政平台逐渐成为政府了解民意的重要渠道。因此运用自然语言处理和文本挖掘技术对政府提升管理水平和施政效率具有重大的意义。

对于问题 1，我们先对附件 2 中数据进行预处理，然后建立深度学习中的 LSTM 长短期记忆网络对留言文本进行多分类处理。通过 Sequential 模型构建神经网络，通过对 Embedding 嵌入层、SpatialDropout1D 层、LSTM 层设置，不断调整超参数，最终得到合适的一级标签分类模型，其 F-Score 评分为 0.84。

对于问题 2，我们首先对附件 3 中数据进行去重处理，得到没有重复的留言信息，然后利用 jieba 中文分词工具对留言主题进行分词，并且通过 TF-IDF 算法提取出留言主题中的地点和任务的关键词，且计算得到每条留言信息中的权重向量。然后根据相似度算法 TF-IDF 和 LIS 模型分别进行计算相似度，经过对比，最终使用精确度更高的 LIS 模型将同一类问题归于相同的问题 ID。最后基于 Reddit 排名算法对留言热点问题定义了四个热度评价指标：发表时间、点赞数（反对数）、问题留言数，并对热度前 5 名的留言问题给出评价结果。

对于问题 3，我们分别对相关性、完整性、可解释性进行评分。针对相关性，计算留言主题与答复意见的余弦相似度作为相关性得分；针对完整性，考察答复意见满足某种规范，通过自定义词典，将包含规范词语个数/总规范词语个数作为完整性得分；针对可解释性，根据是否含有可解释性词语对留言答复进行二分类处理，根据得到的三个指标的得分，构建对答复意见的评价模型。最后，利用评价模型对答复意见的质量进行评价。

关键词：TF-IDF 算法；LIS 模型；LSTM 长短期记忆网络模型；文本分类

Abstract

In recent years, with the arrival of the era of big data and artificial intelligence, online political platform has gradually become an important channel for the gover - nment to understand public opinion. Therefore, the use of natural language process - ing and text mining technology is of great significance for the management level and governance efficiency.

For problem 1, the long-term and short-term memory network of LSTM in deep learning is established to multi classify the message text. The neural network is constructed by sequential model. By setting the embedding layer, spatialdropout1d layer and LSTM layer, the super parameters are adjusted continuously. Finally, the appropriate first level label classification model is obtained, with F-score of 0.84.

For problem 2, We use the Chinese word segmentation tool of Jieba to segment the message subject, and extract the key words of the location and task in the message subject through TF-IDF algorithm, and calculate the weight vector in each message information. Then according to the similarity algorithm TF-IDF and LIS model to calculate the similarity, after comparison, the LIS model with higher accuracy is used to attribute the same kind of problem to the same problem ID. Finally, based on the reddit ranking algorithm, we define four heat evaluation indexes for the hot topic of message: publishing time, the number of likes (anti logarithm), the number of problem messages, and give the evaluation results for the top five heat message problems.

For question 3, select three indicators: relevance, integrity and interpretability. For the relevance index, the cosine similarity between the message subject and the response opinion is calculated as the relevance score; for the integrity index, the response opinion meets a certain specification is taken as the integrity score through the customized dictionary; for the interpretability index, the response to the message is based on whether there are interpretable words According to the scores of the three indexes, the evaluation model of the reply opinions is constructed. Finally, the evaluation model is used to evaluate the quality of the response.

Keywords: TF-IDF; LIS; LSTM; text classification

目录

1 挖掘目标

本次建模的目标是利用互联网公开来源给出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的文本数据，利用自然语言处理和文本挖掘的方法以解决下列三个目标：

1）利用文本分词和文本分类的方法对网络问政平台的群众留言数据进行文本挖掘，根据分类器的结果建立关于留言内容的一级标签分类模型，并且对分类方法进行评价。

2）根据问政平台群众留言数据，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，并且定义合理的热度评价指标，并给出最终的留言热度评价结果。

3）根据相关部门对留言的答复意见的数据,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合理的评价方案。

2 分析方法与过程

2.1 问题 1 分析方法与过程

2.1.1 流程图

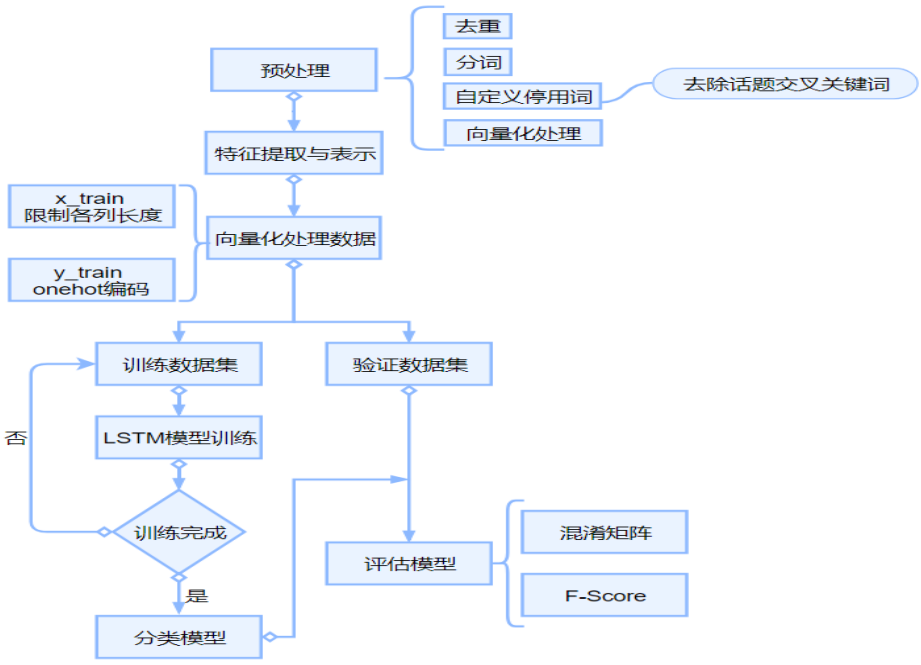


图 1 问题 1 流程图

2.1.2 数据预处理

（1） 留言数据的去重、去空

考虑到后续用 Python 处理数据时，若存在空内容或是重复文本会干扰问题的分析，因此这里采用直接过滤的方法，将这些文本进行删除处理。经过查看在全部数据中不存在重复或是为空的文本信息。

（2）中文分词

在进行留言问题的挖掘之前，需要将非结构化的文本信息转换为计算机可以识别的结构化信息，所以首先要对文本信息进行中文分词处理，本文这里采用 Python 的中文分词包进行 jieba 分词，采用分词模式中的精确模式，根据附件 2 中的数据添加了自定义分词词典。

（3）去停用词

在本文的文本处理中，停用词就是指一些没有实际含义的词，比如“的”、“了”、“呢”等，对于停用词一般在预处理阶段就将其删除，避免增加文本处理的复杂度，本文所采用的停用词是取自哈工大停用词表。

2.1.3 LSTM 建模

(1) LSTM 模型的定义与原理

长短期记忆网络（LSTM，Long Short-Term Memory）是一种时间循环神经网络，是为了解决一般的 RNN（循环神经网络）存在的长期依赖问题而专门设计出来的，所有的 RNN 都具有一种重复神经网络模块的链式形式。因为在传统神经网络中，模型不会关注上一时刻的处理会有什么信息可以用于下一时刻，每一次都只会关注当前时刻的处理。

LSTM 原理的结构图如下:

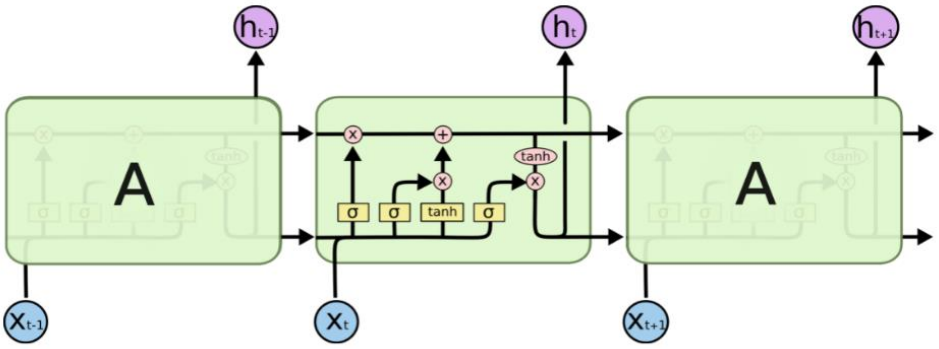


图 2 LSTM 原理的结构图

LSTMs 的关键就是下面的矩形方框，被称为 memory block（记忆块），主

要包含了三个门（forget gate、input gate、output gate）与一个记忆单元（cell）。方框内上方的那条水平线，被称为 cell state（单元状态），它就像一个传送带，可以控制信息传递给下一时刻。如下图：

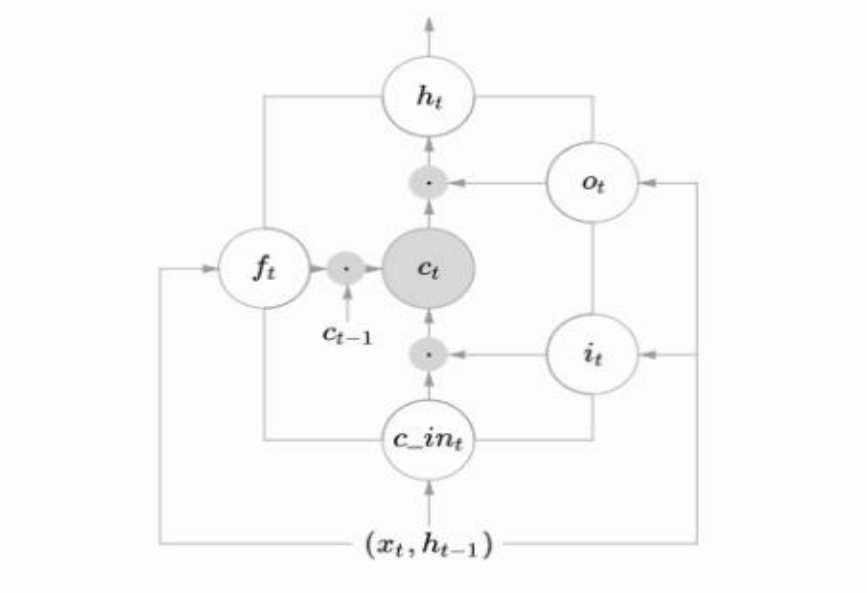
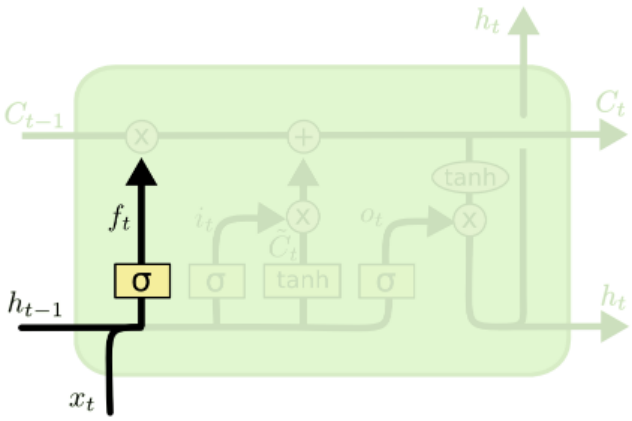


图 3 LSTM 核心思想记忆块

中心的分别为遗忘门、输入门、输出门，用 sigmoid 层表示。上图中的两个 tanh 层则分别对应 cell 的输入与输出。

具体 LSTM 模型解析如下：

1）LSTM 第一步是用来决定什么信息可以通过 cell state。这个决定由“forget gate”层通过 sigmoid 来控制，它会根据上一时刻的输出通过或部分通过。如下：

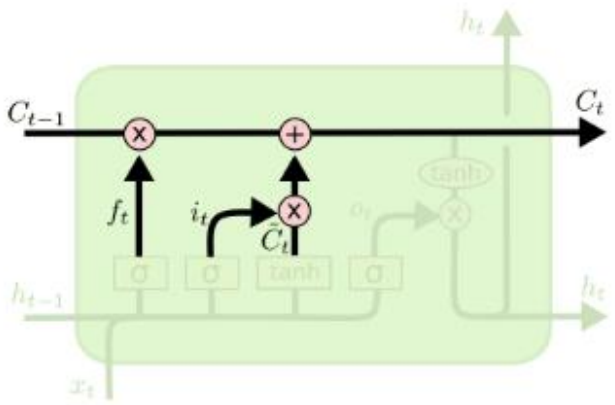


公式表示为：

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{1}$$

2）第二步是产生我们需要更新的新信息。这一步包含两部分，第一个是一个

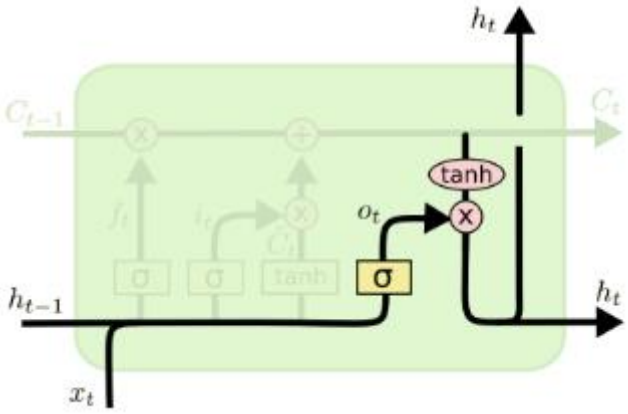
“input gate”层通过 sigmoid 来决定哪些值用来更新，第二个是一个 tanh 层用来生成新的候选值相加，得到了候选值。一二步结合起来就是丢掉不需要的信息，添加新信息的过程：



公式表示为：

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{2}$$

3）最后一步是决定模型的输出，首先是通过 sigmoid 层来得到一个初始输出，然后使用 tanh 将值缩放到-1 到 1 间，再与 sigmoid 得到的输出逐对相乘，从而得到模型的输出：



公式表示为：

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{3}$$

(2)基于留言文本构建 LSTM 分类模型

模型的建立过程：

1)首先通过 Sequential 模型构建神经网络，这里的 Sequential 更准确的应该理解为堆叠，通过堆叠许多层，构建出深度神经网络。

2)Embedding 嵌入层，针对于中文词汇把它拆成指定数量的特征维度，词表中的每一个词都用这些维度组合成的向量来表示，使用语料和模型来训练词向量

——把嵌入矩阵当成模型参数的一部分，通过词与词间的共现或上下文关系来优化模型参数，最后得到的矩阵就是词表中所有词的词向量。使用长度为 100 的向量来表示每一个词语。

3)SpatialDropout1D 层，通过 SpatialDropout1D 层帮助提高特征独立性，相当于用其取代普通 dropout。在训练中每次更新时， 将输入单元的按比率随机设置为 0，这有助于防止过拟合。

4)LSTM 层，包含 100 个记忆单元，100 表示这一层输出的维度，如果 input 的 shape 是(m,n)，那这层输出的 shape 是（m，100）dropout=0.2，使 20%比重的神经元输出（unit 的输出）激活失效。

5)全连接层，输出层主要为包含 7 个分类的全连接层，7 位该层神经单元的结点数，由于是多分类，所以激活函数设置为'softmax'，损失函数为分类交叉熵 categorical_crossentropy。

模型相关参数以及阈值设定如下：

表 1 LSTM 分类模型相关参数设定

参数	参数值
学习速率	0.01
批次大小	5
层大小	4
BatchSize	64
输入端词向量维度	100
输出端词向量维度	7

2.1.4 模型评价

本文是以留言文本分类为主的多分类问题，具体评价分类模型是以分类的查准率、查全率和 F-Score 的计算方式进行最终模型好坏的评价。计算公式如下：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{4}$$

其中， P_i 为第 i 类的查准率。

R_i 为第 i 类的查全率。

F_1 作为一个分类模型的评价指标，由分类效果的查准率和查全率决定。程序见附件“C 题 1.ipynb”。

2.2 问题 2 分析方法与过程

2.2.1 流程图

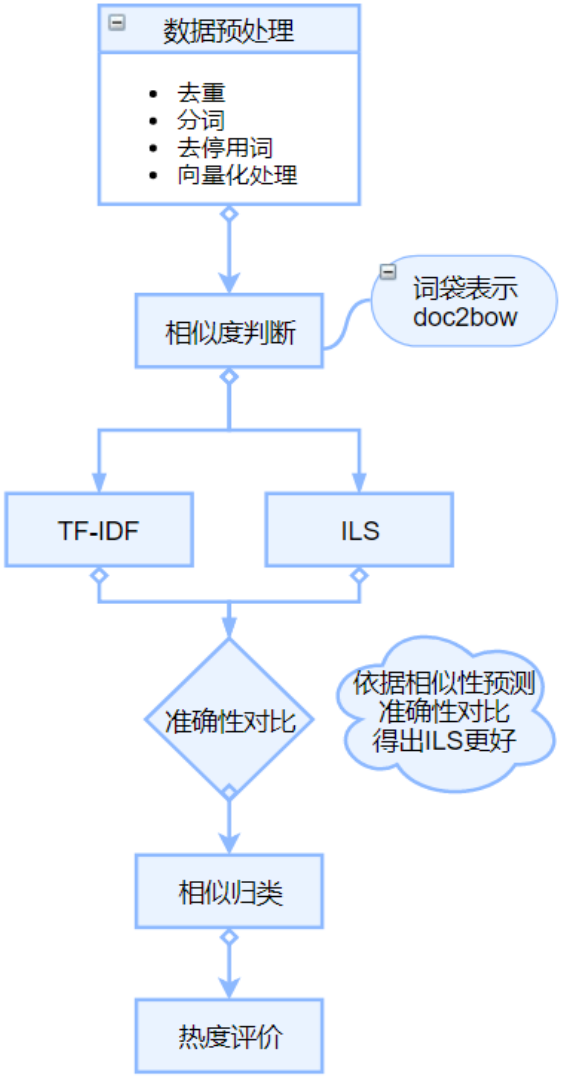


图 4 问题 2 流程图

2.2.2 数据预处理

去重、去空，中文分词，去停用词处理过程和问题 1 类似。与问题 1 不同的是，由于问题 2 要对不同地点的各类人群进行识别，所以在中文分词过程中，将一些地名类的市县区等加入了结巴词典中，便于后续分词提取关键词地址。

2.2.3 文本表示及词向量生成模型

（1）词袋模型

词袋模型（Bag of words，简称 BoW ）最早出现在自然语言处理

（Natural Language Processing）和信息检索（Information Retrieval）领域.。该模型忽略掉文本的语法和语序等要素，将其仅仅看作是若干个词汇的集合，只考虑所有词的权重，文档中每个单词的出现都是独立的。其中权重与词在文本中出现的频率有关。本文使用词袋表示，具体来说就是每个词出现的次数。连接词和次数就用字典表示。然后使用 doc2bow()函数统计词语的出现次数。

（2）TF-IDF 模型

TF-IDF 模型（又叫做词频逆文档矩阵模型），主要有词频（TF）和逆文本频率指数 IDF 构成，词频顾名思义就是这个词在所有词中出现的频率，逆文本频率指数是对这个词语重要性的度量。由于词袋模型存在一些缺陷，只能用词语出现的频率去表示这个词语的重要性，这样会导致对词频出现不是很高的重要词汇缺乏考虑，而 TF-IDF 模型就是为了解决这一缺陷。这一模型对于词频和逆文本频率的计算公式分别如下：

$$\mathrm{tf}(w_i) = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{5}$$

$$\mathrm{idf}(w_i) = log\frac{|D|}{|\{j:w_i \in d_j\}|} \tag{6}$$

其中 $n_{i,j}$ 表示该词出现的次数，$\sum_k n_{k,j}$ 表示整个文本中所有词出现的次数总和。

（3）doc2bow

由于在常见的 word2vec 模型中，词语之间的相互联系只体现在词窗口内的几个词中，具有一定缺陷，没有考虑到远距离依赖和全局信息。本文这里采用 Doc2vec 模型，它可以在语言模型的基础上，可以将能够表示段落的全局信息考虑到。Doc2bow 共有两种方法：Distributed Memory(DM) 和 Distributed Bag of Words(DBOW)，DM 试图给定上下文和段落向量的情况下预测单词的概率。在一个句子或者段落文档训练过程中，段落 ID 保存不变，共享同一个段落向量。DBOW 则在只给定段落向量的情况下预测段落中一组随机单词的概率。

本文结合词袋模型的优点运算快比较简单，首先将要对比的文本通过 doc2bow 转化为词袋模型，对词袋模型进行进一步处理，得到新语料库。然后利用 TF-IDF 模型进一步将新语料库通过 tfidfmodel 进行处理，得到 tf-idf。

2.2.4 词语相似度计算方法

上文介绍的 TF-IDF 模型也是处理文本相似度的一种办法之一，另一种计算

文本相似度的办法是 LSI 模型，它主要是根据文本和词共现以及奇异值分解（SVD）方法来得到文本主题的一种模型，对高维的向量空间进行一个降低维度的处理，简而言之就是将训练文档向量组成的矩阵 SVD 分解，并做了一个秩为 2 的近似 SVD 分解。LSI 模型主要分为以下三个步骤：

1）构建模型与优化

在 LSI 模型中使用此文档矩阵 A_{mn} 表示整个词库，它具体是指不同的词在不同的文本中的数量构成的矩阵，A_{mn}的每一列对应一个文本，对于矩阵中的元素a_{ij}不是第i 个词在第 j 条文本中出现的次数，但是这里不采用他的原始值，采用对其优化的方式，用局部和全局之间的权值乘积表示。

2）降低维度

根据公式进行奇异值分解

$$\mathrm{A} = \mathrm{U}\sum V^T \tag{7}$$

通过取左右奇异 $(U$ 和$V)$举证的前 $\mathrm{d} \ll \min(\mathrm{m},\mathrm{n})$ 列得到如下降维公式：

$$A_d = U_d \sum_d V_d^T \tag{8}$$

3） 文本相似度计算

通过矩阵计算相似度的公式如下：

$$\mathrm{SIM}(d_1, d_i, \ldots) = \frac{\sum_{i=2}^{n} d_1 \times d_i}{\prod_{i=2}^{n} \sqrt{(d_1)^2} \times \sqrt{(d_i)^2}} \tag{9}$$

关于实现识别相似留言文本的处理，本文使用 python，调用包 gensim 中的 tf-idf 模块来进一步对每个单词计算权重，在计算相似度时分别使用 TF-IDF 模型和 LSI 模型，并且用余弦相似度（问题 3 中有详细介绍）对比两个模型的准确度，发现 LSI 模型的相似性准确性跟高，于是在处理最后的文本相似归类时采用 LSI 模型。

2.2.5 热点问题挖掘

本文结合热度排行算法的思想，对利用相似性归为同一类的相似留言进行热度指数评价，建立了如下基于 Reddit 的热度评价指标：

（1）相似留言问题持续时间 T

T=留言最晚发表时间-留言最早发表时间，在 Reddit 热度排行算法中时间对于排名有很大影响，于是本文采用相似留言问题持续的时长作为热度评价的评价指标之一。（这里的留言是指被划分为同类型的留言）

（2）点赞数与反对数之差 x

x = 点赞数 - 反对数，由于 Reddit 提供了投反对票的功能，所以可以使一些具有争议的话题会排的较后。

（3）投票方向 y

y 是一个符号变量，表示对留言内容的总体看法。如果点赞数居多，y 就是+1；如果反对数居多，y 就是-1；如果点赞数和反对数相等，y 就是 0。y 是文章评价的一种定性表达，0 表示没有倾向，大于 0 表示正面评价，小于 0 表示负面评价。

（4）留言主题的受肯定程度 z

z 表示赞成票超过反对票的数量。如果赞成票少于或等于反对票，则 z 就等于 1。

结合以上几个热度评价指标，基于 Reddit 算法热度指数的最终得分计算公式如下：

$$\text{Score} = logz + \frac{yt}{45000} \tag{10}$$

其中公式中$logz$表示，赞成票超过反对票的数量越多，得分越高。这里用的是以 10 为底的对数，意味着 z=10 可以得到 1 分，z=100 可以得到 2 分。yt/45000 表示，t 越大，得分越高，即新帖子的得分会高于老帖子。它起到自动将老帖子的排名往下拉的作用。分母的 45000 秒，等于 12.5 个小时，也就是说，后一天的帖子会比前一天的帖子多得 2 分。

综上，根据热度评价指标表，将分为一类的文本进行热度指数评分，得到的结果和上述结果程序见附件“C 题 2.ipynb”，结果保存在“热点问题表”中。

表 2 热度评价指标汇总表

评价指标	具体含义	热度指数得分公式
相似留言问题持续时间 T	留言最晚发表时间-最早发表时间	$\text{Score} = logz + \frac{yt}{45000}$
点赞数与反对数之差 x	点赞数 - 反对数	
投票方向 y	符号变量	
留言主题的受肯定程度 z	赞成票超过反对票的数量	

2.3 问题 3 分析方法与过程

在这一部分中，根据附件 4，针对相关部门对留言的答复意见，从意见的相关性、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法。

2.3.1 问题 3 流程图

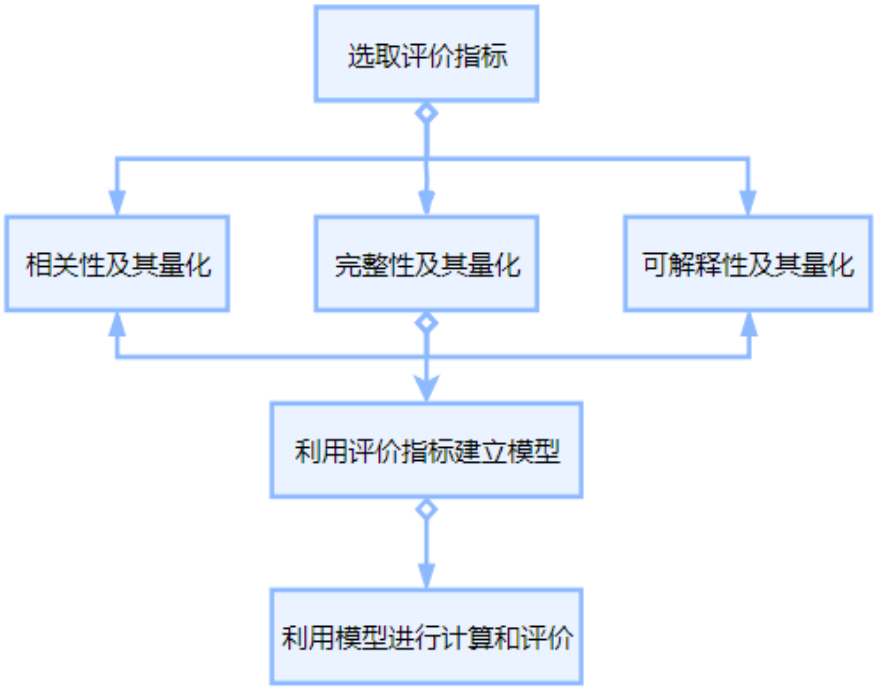


图 5 问题 3 流程图

2.3.2 选取评价指标

（1）相关性及其量化

相关性是指研究答复意见是否是针对留言的回答，相关性是评价答复意见质量最重要的一个指标，在避免出现答非所问的情况下，才能去研究答复意见质量的完整性、可解释性等其他相关评价指标。

在研究文本数据的相关性时，主要有 TF-IDF、BM25、simhash、LSA 类模型等计算方法，在本文中利用余弦相似度计算答复意见与留言主题之间的相似性。余弦相似度，又称为余弦相似性，是通过计算两个向量的夹角余弦值来评估他们的相似度。余弦相似度将向量根据坐标值，绘制到向量空间中，如最常见的二维空间。两个向量间的余弦值可以通过使用欧几里得点积公式求出：

$$a \cdot b = \|a\|\|b\| \cos\theta \tag{11}$$

给定两个属性向量，A 和 B，其余弦相似性θ由点积和向量长度给出，如下所示：

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (12)$$

其中A_i，B_i分别代表向量 A 和 B 的各分量。给出的相似性范围从-1 到 1：-1 意味着两个向量指向的方向正好截然相反，1 表示它们的指向是完全相同的，0 通常表示它们之间是独立的，而在这之间的值则表示中间的相似性或相异性。对于文本匹配，属性向量 A 和 B 通常是文档中的词频向量。余弦相似性可以被看作是在比较过程中把文件长度正规化的方法。

相关性指标主要按照以下步骤进行量化处理：

1）读取附件 4，分别提取出留言主题和答复意见两列数据；这里选择留言主题而非留言详情，主要是因为留言主题是对留言详情的高度概括，便于数据的处理与分析。

2）利用 jieba 分词和停用词表对留言主题与答复意见进行分词处理，保存在向量中；

3）用字典保存两列数据中出现的所有词并编上号；

4）根据词袋模型统计词在每篇文档中出现的次数，形成向量；

5）计算两者的余弦相似度作为相关性得分 score1；

6）将相关性得分最后结果写入 txt 文件中。

程序见附件“C 题 3.ipynb”,结果保存在 result1.txt 中。

（2）完整性及其量化

完整性即是考虑留言答复是否符合某种答复规范或答复格式，通过对全部数据中“答复意见”列的观察可以发现，绝大多数都具有格式：……，您好，……已收悉，……回复如下：……。感谢您……。×年×月×日。因此我们将符合上述格式的答复意见视作规范的完整性答复。通过自定义词典将上述格式中出现的词语纳入分词规范，对“答复意见”列进行分词处理，定义某条留言的答复意见的完整性得分为：

$$\text{score2} = \frac{\text{答复意见中包含规范词语的个数}}{\text{总规范词语的个数}}$$

其中，总规范词语的个数= 7。

程序见附件“C 题 3.ipynb”，结果保存在 result2.txt 中。

（3）可解释性及其量化

对于可解释性指标，主要考察留言答复是否具有“支撑性依据”，即意见建议是否是根据某项相关法律，是否来源于共同协商，或者是否根据实地调查而给

出。如果有“支撑性依据”就可以认为该答复意见是可解释性的，所以提取出“规定”、“通知”、“规定”、“政府发文”、“据查”、“实地考察”等一系列可以代表“支撑性依据”的词语后，建立了新的分词规范。利用新的分词规范对“答复意见”列进行分词处理，在处理后数据的基础上，根据是否含有可解释性词语对答复意见进行二分类，若含有可解释性词语则归为可解释性数据，且可解释性得分为 1，若不含有可解释性词语则归为不可解释性数据，且可解释性得分为 0，即可解释性得分定义如下：

$$\text{score3} = \begin{cases} 1 & \text{答复意见中含有可解释性词语} \\ 0 & \text{答复意见中不含有可解释性词语} \end{cases}$$

程序见附件“C 题 3.ipynb”，结果保存在 result2.txt 中。

2.3.3 利用评价指标建立模型

分别获得样本数据的相关性、完整性和可解释性得分之后，利用这 3 个得分建立有关答复意见质量的评价指标体系。出于简单实用角度考虑，直接采用线性模型去构建评价指标体系。在系数权重的确定问题上，认为相关性是评价答复意见质量最重要的一个指标，而完整性和可解释性在质量评价中居于稍次地位，因此考虑如下权重赋予方案：

（1） 权重依次分别为 0.4，0.3，0.3

质量评价得分：$\text{score} = 0.4 \times \text{score1} + 0.3 \times \text{score2} + 0.3 \times \text{score3}$

该模型对所有样本质量评价得分的程序见附件“C 题 3.ipynb”，计算结果保存在“model1.csv”文件中。

（2） 权重依次分别为 0.6，0.2，0.2

质量评价得分：$\text{score} = 0.6 \times \text{score1} + 0.2 \times \text{score2} + 0.2 \times \text{score3}$

该模型对所有样本质量评价得分的程序见附件“C 题 3.ipynb”，计算结果保存在“model2.csv”文件中。

3 结果分析

3.1 问题 1 结果分析

3.1.1 分类模型评价结果

在此模型中，通过 5 次迭代实现了后向传播寻求最小损失函数，前向传播给出更准确的信号。最终，训练损失从 1.4 降到 0.05，可以看到，验证损失在第 4、5 次迭代时，有较小的上升，而且，模型的训练准确度和验证准确度在第 4、5 次迭代时，没有显著提升，我们希望有更好的测试效果，因此将迭代次数改为 3 次。

（1）损失函数图像

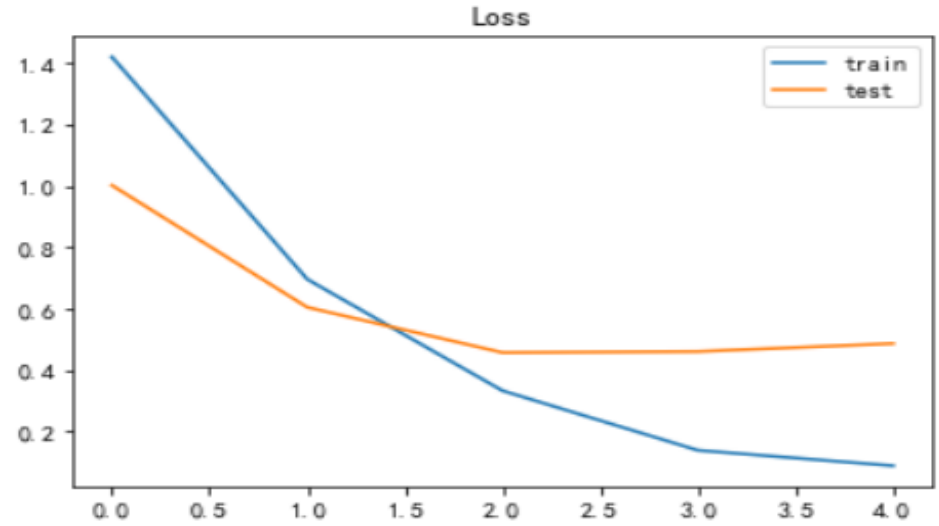


图 6 模型损失函数图像

（2）精确度

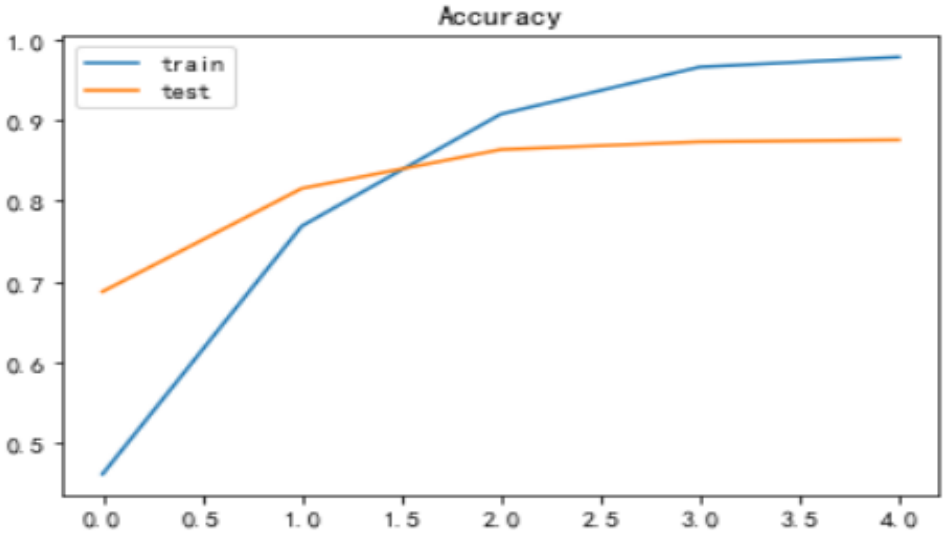


图 7 分类模型迭代精确度

根据上述图，在此模型中，通过 5 次迭代实现了后向传播寻求最小损失函数，前向传播给出更准确的信号。最终训练损失从 1.4 降到 0.05，可以看到验证损失在第 4、5 次迭代时，有较小的上升，而且，模型的训练准确度和验证准确度在第 4、5 次迭代时，没有显著提升，我们希望有更好的测试效果，因此将迭代次数改为 3 次。

（3）模型评价-混淆矩阵、F-Score 评分结果

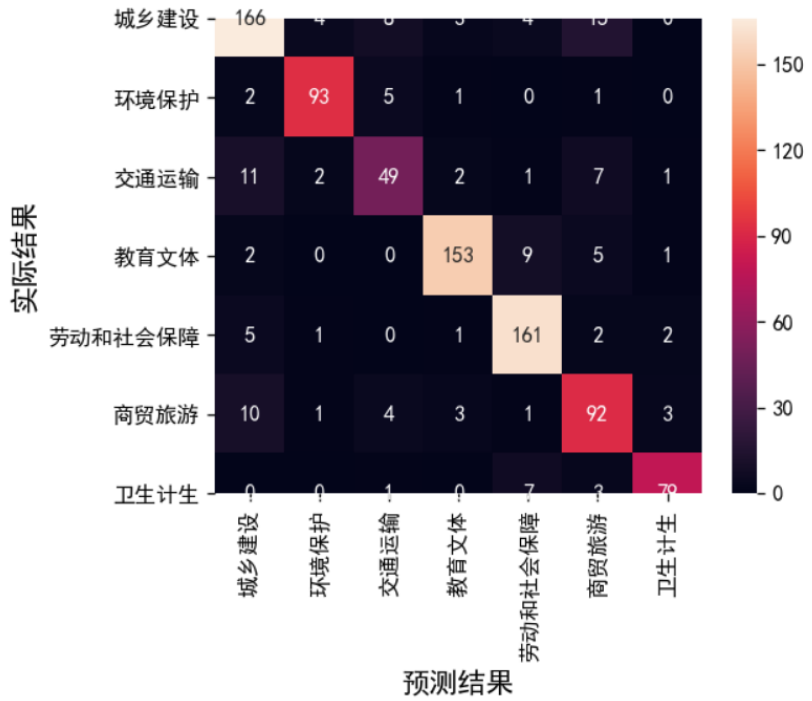


图 8 模型分类预测结果混淆矩阵图

表 3 模型分类预测结果 F1 评价

	Precision	Recall	F1-score	Support
城乡建设	0.85	0.83	0.84	200
环境保护	0.92	0.91	0.92	102
交通运输	0.73	0.67	0.70	73
教育文体	0.94	0.90	0.92	170
劳动和社会保障	0.88	0.84	0.81	172
商贸旅游	0.74	0.81	0.77	114
卫生计生	0.92	0.88	0.90	90

从分类模型评价的最终结果来看：

Precision 为查准率，例如，城乡建设的查准率为 0.85，即，被正确分类为城乡建设的数量/所有被分为城乡建设的数量。城乡建设的查准率越高，说明分得越对，其他记录被错误分类为城乡建设的越少，可以表中的 7 个类别查准率均较好。Recall 为查全率，例如，城乡建设的查准率为 0.83，即，被正确分类为城乡建设的数量/真实为城乡建设的数量，查全率越高，说明分得越全，该分类下被准确分类的越多。

总之，表中对于留言问题的一级分类标签有 6 个类别数据较好，交通运输仅

为 0.67，查全率稍低。在分析中，我们不能一味仅追求查准率或者查全率某一指标，F-Score 是综合两个指标的更好的选择，表中 F-Score 均超过 0.70，经计算，整个数据的 F-Score 为 0.84，由此可见建立的分类模型预测效果较好。

3.2 问题 2 结果展示

3.2.1 热点问题留言明细

根据前文计算每条文本之间的相似度，最终得到如下热点问题留言明细表（部分数据）：

表 4 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	留言详情	点赞数	反对数
1	217700	A909239	丽发新城小区旁的搅拌站严重影响生活	2019-12-21 02:33:21	开发商把特大型搅拌站，水泥厂……，灰尘，噪音污染严重	0	1
1	259788	A909221	A 市暮云街道丽发新城社区搅拌厂危害居民健康	2019-12-07 00:00:00	A 市暮云街道丽发新城社区（丽发新城小区）搅拌站灰尘，噪音污染严重……	0	0
1	283482	A909232	丽发新城小区附近搅拌站的一些问题	2019-12-07 09:21:32	小区旁边违建搅拌站,灰尘和噪音严重影响了居民的生活和身体健康……。	0	0
……	……	……	……	……	……	……	……
2	213584	A909172	投诉 A 市伊景园滨河苑定向限价商品房违规涨价	2019-07-28 13:09:08	投诉 A 市违规收取认购款 18.5 万，违法强行捆绑销售车位，……。	0	0
2	190337	A00090519	关于伊景园滨河苑捆绑销售车位的维权投诉	2019-08-23 12:22:00	投诉伊景园.滨河苑开发商捆绑销售位！……，强制要求职工再交 12 万的车位费，……。	0	0
……	……	……	……	……	……	……	……
3	188414	A00096844	A4 区北辰小区非法住改商问题何时能解决？	2019/8/1 7:20:31	住改商数量如此之多，且非法住改商问题很难得到有效解决，……。	0	0
3	284066	A00032346	A3 区天顶街道青青家园违规住改商问题到底谁能解决	2019/7/24 16:17:59	在 A3 区天顶街道青山新村社区青青家园违规住改商商户违反《物权法》77 条，伪造利害业……。	0	0

……	……	……	……	……	……	……	……
4	219026	A00083974	在 A 市购买二手房能享受人才新政购房补贴吗？	2019/9/27 9:17:14	我研究生毕业后根据人才新政落户 A 市，请问购买二手房能否享受研究生购房补贴？	0	0
4	244951	A00026039	反映 A 市人才租房购房补贴问题	2019/7/7 19:12:32	A 市人才租房购房补贴对吸引人才是好事，但为什么人才到政府、事业单位和高校工作就没有补贴呢？……。	0	0
……	……	……	……	……	……	……	……
5	235118	A00093448	A3 区洋湖街道一工地严重噪音扰民	2019/9/11 20:27:30	所反映的问题是紧挨小区的施工工地严重噪音扰民和违规乱停车的问题……。	0	0
5	190971	A0007349	A3 区洋湖街道白鹭社区便民服务不便民	2019/12/19 18:13:28	去洋湖街道白鹭社区办理灵活就业社保，该单位不认可 A 市公安局颁布的电子证照为由，拒绝办理……。	0	0

3.2.2 热度评价结果

根据附件 3 中数据将某一时段特定地点或人群的问题留言进行归类处理后，根据前文定义的热度评价指标：相似留言问题持续时间 T、点赞数与反对数之差 x、投票方向 y、帖子的受肯定程度 z，给出热度排名前五的热点问题的结果，如下表：

表 5 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	91	2019/4/10 10:21:47-2020-01-26 19:47:11	A 市万家丽南路丽发新城	搅拌站噪音扰民、污染环境
2	2	82	2019/7/18 20:27:40-2019-09-01 14:20:22	A 市伊景园滨河苑	定向限价商品房违规涨价
3	3	75	2019/1/2 10:24:50-2019/8/1 7:20:31	A 市	非法住改商问题
4	4	70	2019/1/16 11:58:48-2019/9/27 9:17:14	A 市	人才购房补贴申请咨询
5	5	64	2019/9/11 20:27:30-2019/12/19 18:13:28	A3 区洋湖附近	工地施工、噪音扰民

3.3 问题 3 结果展示

3.3.1 不同权重下评价指标得分结果

（1）权重依次分别为 0.4，0.3，0.3

对于计算出来的得分，进行了可视化处理，分别计算了“得分<20%”，“20%<=得分<40%”，“40%<=得分<60%”，“得分>=60%”的得分分布柱状图。

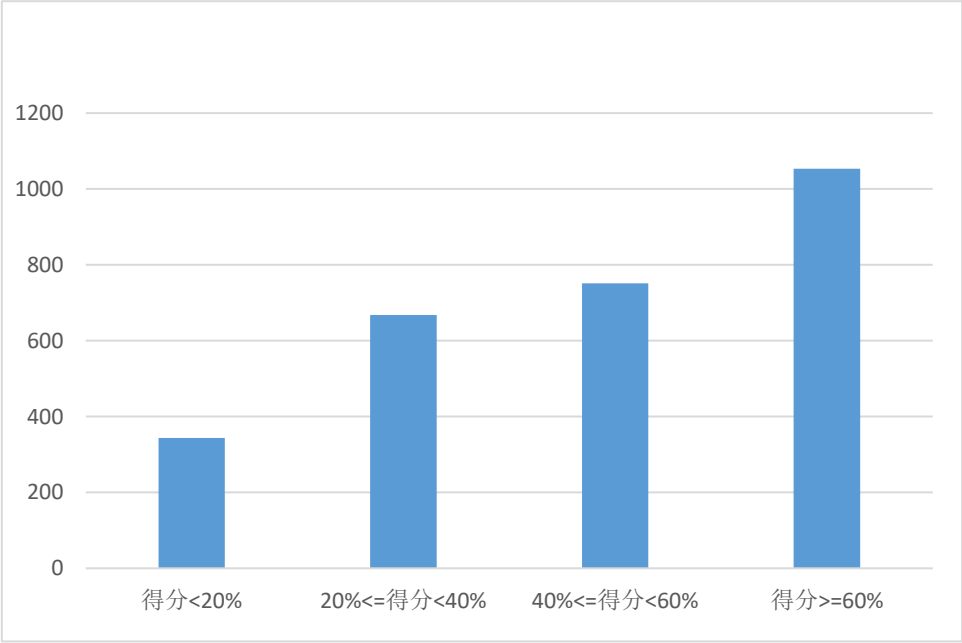


图 9 模型 1 得分分布柱状图

（2）权重依次分别为 0.6，0.2，0.2

模型 2 的得分分布柱状图如下所示：

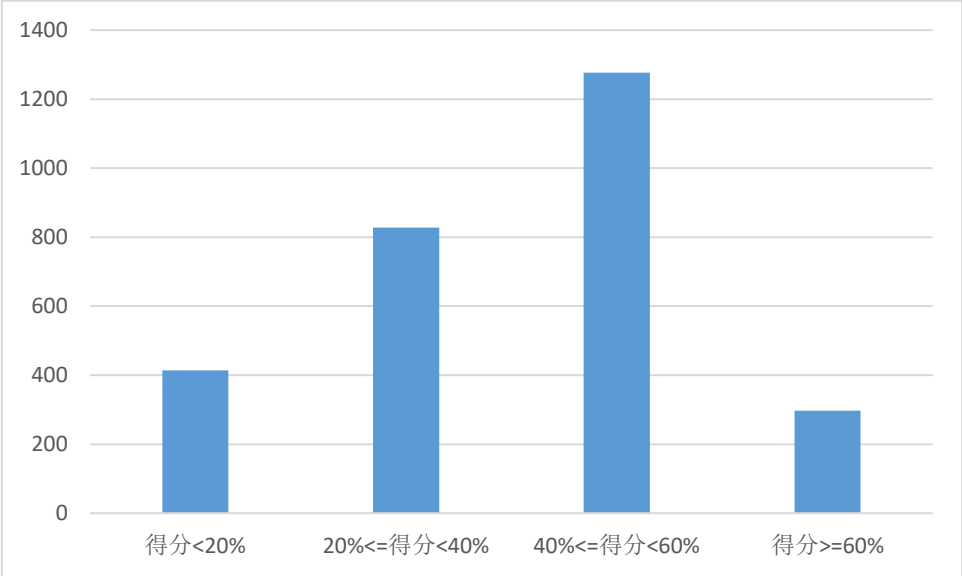


图 10 模型 2 得分分布柱状图

由于认为所给样本中答复意见的质量应该是比较好的，而从模型 2 的得分分布柱状图可以看出，“得分>=60%”数据的数目占全部样本数目的比例较小，也就是样本中的答复意见的质量评价得分也不是很高，这就不满足所认为的样本质量较好的假定，而模型 1 的柱状图中，“得分>=60%”数据的数目占全部样本数目的比例较大，说明模型 1 更加符合最初的假定，因此，选择模型 1 作为答复意见的质量评价模型。

3.3.2 评价模型的得分计算和评价结果

根据已经确定了答复意见的质量评价模型为：

$$score = 0.4 \times score1 + 0.3 \times score2 + 0.3 \times score3$$

以模型 1 中前两条留言的答复意见评价得分为例：

表 6 留言主题与答复意见的部分数据

留言编号	留言主题	答复意见	评价得分
2549	A2 区景蓉华苑物业管理有问题	答复如下：您反映的情况已收悉。答复如下：经调查了解，.....，以“意见收集方式”召开了业主大会，经业委会统计，.....，认真疏理归纳并进行了反馈，.....。在综合各方面的意见后，.....。再次感谢您对我区工作的理解和关心。2019 年 5 月 9 日	0.7255
2554	A3 区潇楚南路洋湖段怎么还没修好？	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，.....，现回复如下：.....。感谢您对我们工作的关心、监督与支持。2019 年 4 月 29 日	0.3418

首先，在相关性方面，这两条留言并没有出现答非所问的情况，均是对留言的正确回复；在完整性方面，这两条留言都较为符合我们设定的完整性规范，即在开始表明已收到留言，然后针对留言给出答复意见，最后表达对市民关心支持政府行为的感谢并给出回复时的日期；在可解释性方面，第二条留言，即留言编号为“2554”的答复意见就不如“2549”留言了，因为“2549”号留言的答复意见是有理有据的，给出了大量的“支撑性证据”——“经业委会统计”、“在综合各方面的意见后”、“依法依规召开业主大会”，而“2554”号留言的答复意见中只有简单描述，并没有提供“支撑性证据”。所以，“2554”号答复意见的

评价得分不如“2549”号也是较为合理的，由此也反映了建立的评价指标模型是合理的。

4 结论

通过对网络问政平台的群众留言数据进行分析，基于自然语言处理技术进行文本分类相似性识别从而建立智慧政务系统，很大程度上可以很好地实现智慧识别和划分，这在一定程度上可以避免依靠人工进行留言划分和热点整理时出现的处理速度慢、错误率高等问题，能显著加快问题处理速度，进一步提升政府部门的管理水平。

但是基于自然语言处理的技术，由于中文文本的繁杂性，在处理过程中存在很多具有词义歧义性的问题有待更好地提升，如果可以进一步将中文文本分类识别技术大大的提升，将能够更好地解决中文文本带来的混淆问题，从而改进会对网络问政平台的工作效率。随着大数据、云计算和人工智能等技术的发展,建立智慧政务系统已经是社会治理创新发展的新趋势，另外基于自然语言处理的技术发展前景也相对较好。