# 基于网络问政平台的文本挖掘应用

## 摘 要

近年来,随着信息数据时代的迅猛发展，网络问政平台逐步成为政府了解民意的重要渠道，各类社情民意相关的文本数据量也不断攀升，传统的人工留言划分和热点问题整理工作已经难以满足数据量庞大的留言体系，由此建立基于自然语言处理技术的智慧政务系统已是刻不容缓,因此本文利用来自互联网公开来源的群众问政留言记录,及相关部门对部分群众留言的答复意见，运用文本挖掘技术对相关问题进行研究。

针对问题 1，建立关于留言内容的一级标签分类模型，首先对留言进行数据预处理，根据附件 1 的一级标签进行分类，采用数据重采样解决数据倾斜问题，对重采样后的数据进行 jieba 分词和 Word2vec 训练词向量,然后将数据以 8:1:1 的比例随机划分为训练集、验证集、测试集用卷积神经网络进行训练和验证，最后用 F-Score 对分类方法进行评价。最后得到大约在迭代 3900 步的情况下，能使分类结果达到最优，总的 F-Score 达到了 0.95。

针对问题 2，热点问题的挖掘，首先对数据进行预处理，分词，去停用词筛选出核心词汇，然后利用 TF-IDF 计算文本相似度，得到留言主题可分为 1162 个分类，进而确定以留言主题为主、点赞数和反对数为辅的热点评价指标，对分类后的模型计算热度指数并排序（热度指数 = 留言数*0.9+（点赞数-反对数）*0.1）得到前 5 个热点问题 ，并通过 Excel 简单处理输出题目所示的热点问题表和热点问题明细表;根据我们得到的 5 大热点问题发现热点问题均为涉及民生的问题，且 5 个热点问题的时间范围也较长，基本上在 1 年左右，说明人们的这些问题尚未得到解决，并且人们对这些问题是持续关注的。

针对问题 3，答复意见的评价，经过对所给的数据和实际的现实情况分析，建立从回复的相关性、完整性、时效性以及回复态度四个方面对回复的留言做出相应的评价。并且建立有关的评分指标，其中回复态度采用留言情感分析方法打分。最后对这四大类分别赋予权值为 40%、30%、20%、10%得到总评分。在得到的结果看半数以上的留言是及格的，1/10 以上的回复数达到良好以上，大部分处于中等分数的留言主要是留言相关性不大以及答复态度太差导致评价得分不太高，启示有关部门需要更加细心，耐心和热情的解决群众反映的问题，这样才能创造更加美好的明天和未来!

**关键词：**重采样, Word2vec, 卷积神经网络, TF-IDF, 留言情感分析

# Abstract

In recent years, with the rapid development of the information and data era, the online questioning platform has gradually become an important channel for the government to understand public opinion, and the amount of text data related to various social conditions and public opinions has also been increasing. To meet the huge data volume of the message system, it can be seen that the establishment of a smart government system based on natural language processing technology is urgent. Therefore, this article uses the public questioning message records from public sources on the Internet and relevant departments 'responses to some of the people 's comments Use text mining technology to study related issues.

Aiming at problem 1, establish a first-level label classification model of message content, first perform data preprocessing on the message, classify according to the first-level label of annex 1, use data resampling to solve the data tilt problem, and perform jieba on the resampled data Word segmentation and Word2vec training word vectors, then randomly divide the data into a training set, a verification set, and a test set with a 8: 1: 1 ratio for training and verification with a convolutional neural network, and finally use F-Score to evaluate the classification method. Finally, it is obtained that, in the case of iteration 3900 steps, the classification result can be optimized, and the total F-Score reaches 0.95.

For problem 2, the mining of hot issues, first pre-process the data, segment the words, select the stop words to filter out the core vocabulary, and then use TF-IDF to calculate the text similarity. The message topic can be divided into 1162 categories, and then determined to Focus on the subject of the message, supplemented by the number of likes and objections, and calculate the heat index for the model after classification (heat index = number of messages * 0.9 + (number of likes-objection) * 0.1) 5 hotspot questions, and simply process the hotspot question list and hotspot question list shown in the output questions through Excel; according to the 5 hotspot questions we found, the hotspot questions are all related to people's livelihood, and the time range of the 5 hotspot questions It is also longer, basically about 1 year, which shows that these problems have not been solved and people are paying continuous attention to these problems.

In response to question 3, the evaluation of the reply comments, after analyzing the data and the actual situation, establish a corresponding evaluation of the reply message from the four aspects of relevance, completeness, timeliness and attitude of

reply. And establish relevant scoring indicators, in which the reply attitude is scored using the message sentiment analysis method. Finally, these four categories were given weights of 40%, 30%, 20%, and 10% respectively to obtain a total score. In the results obtained, more than half of the messages are passing, and more than 1/10 of the responses are above good. Most of the messages with medium scores are mainly due to the low relevance of the messages and the poor response attitude, which leads to a low evaluation score. Enlighten the relevant departments to be more careful, patient and enthusiastic to solve the problems reflected by the masses, so as to create a better tomorrow and future!

**Key words：** Re-sampling, Word2vec, CNN, TF-IDF, Message sentiment analysis.

# 目录

# 1 挖掘目标

本次挖掘的目标是利用收集的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数据，采用 jieba 分词技术，重采样处理，Word2vec 训练词向量，卷积神经网络；TF-IDF 算法计算文本相似度，构建热度评价指标，制定答复评价指标，文本情感分析等技术分析其中的关联，从而达到以下三个目标:

1）群众留言分类：根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型，并使用 F-Score 对分类方法进行评价。

2）热点问题挖掘：根据附件 3 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，定义合理的热度评价指标，并给出评价结果，并按相关格式保存热点问题表和热点问题留言明细表。

3）答复意见的评价：针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、时效性、回复态度对答复意见的质量给出一套评价方案。

# 2 符号说明

| 符号 | 意义 |
|---|---|
| $TF$ | 词频权重 |
| $n$ | 某个词在文本中出现的次数。 |
| $N$ | 文本总词数 |
| $times(n_{\max})$ | 文本出现次数最多的词的出现次数 |
| $IDF$ | 逆文档频率 |
| $x$ | 语料库的文本总数 |
| $sum(n_x)$ | 包含该词的文本数 |

| | |
|---|---|
| $HI$ | 热度指数 |
| $Ms$ | 留言数 |
| $Ls$ | 点赞数 |
| $Ag$ | 反对数 |
| $TP$ | 样本中正确预测为正的样本个数 |
| $FP$ | 样本中错误预测为正的样本个数。 |
| $FN$ | 样本中实际为正但是预测为负的样本个数 |
| $P_i$ | 为第$i$类的查准率 |
| $R_i$ | 为第$i$类的查全率 |
| $i$ | 类别 |

# 3 分析过程与方法

## 3.1 问题一分析方法及过程

问题一是群众留言分类问题，分类的方法有很多如何选取比较好的分类方法以及如何进行自然语言处理，我们做了如下思考和分析:

首先观察原数据，先将原始数据按第一级别进行分类，得到 7 大类别的数据，发现各类别的留言数存在数据倾斜的情况，于是对数据进行重采样处理。为了方便算法的训练和检验，于是将每一类的数据重采样后，以 8：1：1 的比例随机划分为训练集、验证集、测试集。

让机器“读懂”自然语言第一步先得把文字符号数字化，因此词向量在自然语言处理中发挥着重要作用，于是通过分词，用 Word2vec 训练词向量，再将划分好的训练集和验证集利用卷积神经网络进行训练和验证，通过验证后，进行预测分类，利用 F-Score 来评价结果。

### 3.1.1 流程图

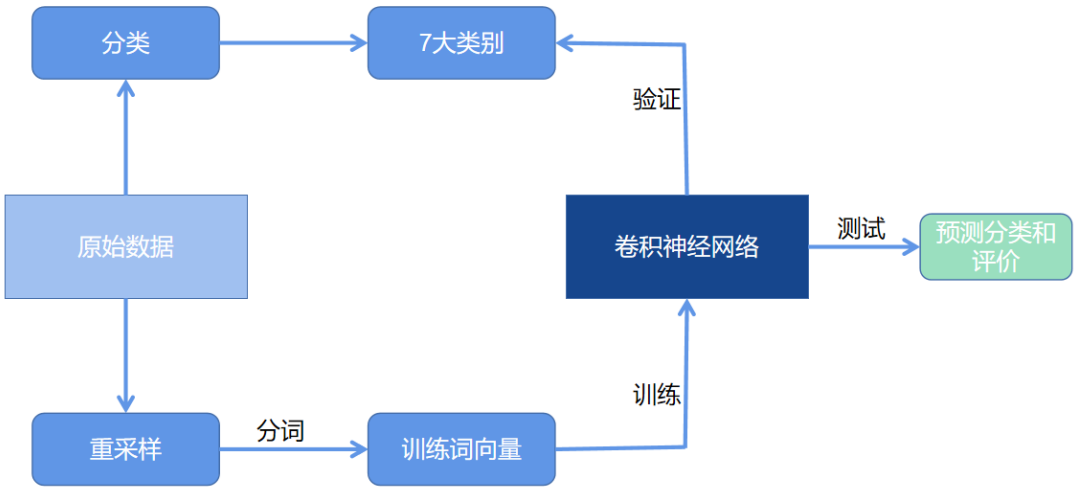



图 1 问题一的流程图

### 3.1.2 数据预处理

根据题目所给附件 2 的数据，根据附件 1 对留言以一级标签进行分类。首先我们将原始数据进行初步处理，将相同标签下的留言归结在一起，得到数据为以下几种情况：

表 1 一级标签分类情况

| 一级标签 | 留言数 |
|---|---|
| 城乡建设 | 2009 |
| 教育文体 | 1589 |
| 商贸旅游 | 1215 |
| 劳动和社会保障 | 1969 |
| 交通运输 | 613 |
| 环境保护 | 938 |
| 卫生计生 | 877 |
| 共计 | 9210 |

从表 1 可以看出，附件 2 数据共计 9210 条，并且各类别下的留言数有很大差别。其中城乡建设有 2009 条留言记录，而交通运输只有 613 条数据。相差悬殊。

通过比对附件 1 的一级标签分类，附件 1 给出的一级标签分类含有 15 个，但附件 2 数据的一级标签只有 7 个，没有包含党务政务、国土资源、纪检监察、经济管理、科技与信息产业、民政、农村农业、商贸旅游、政法八个一级标签的数据，这八个标签的数据或与党政或与经济等方面有关，由于没有给出具体的数据无法进行研究，于是下面只对已有数据的 7 大类别进行讨论研究。

### 3.1.3 数据重采样

经过数据预处理，我们发现附件 2 的数据按照一级标签分类后，各类别的留言数不仅仅有差别，并且存在数据倾斜。数据倾斜是数据挖掘中的一个常见问题，它的存在严重影响数据分析的最终结果，在分类问题中其影响更是比较大的。

数据倾斜的解决办法有过采样和欠采样：过采样是处理样本量不均衡的一个基本解决方法，其实现简单高效，通过扩展样本数少的类别的样本来解决样本不均衡问题。处理方法有：1、直接复制，即不断复制样本数少的类别样本；2、插值法，通过对样本归一化，求得样本分布，极值，均值等，然后根据样本分布，极值，均值来生成新样本来扩充样本数目。

解决数据量不均衡方法有很多，本题结合实际情况，对此我们决定使用对数据集进行重采样这一方法。过采样（over-sampling），对小类的数据样本进行过采样来增加小类的数据样本个数，即采样的个数大于该类样本的个数。具体方法是在同一类别下随机抽取一些留言，将他们进行随机拆分和组合，形成新的数据。

通过重采样之后，我们得到每个一级标签下的留言量是 2000 左右，以减小在后期的模型训练中导致的错误率。然后我们将处理后数据里面的每一类按照训练集、验证集、测试集以 8：1：1 的比例保存到 train.txt,val.txt,test.txt 文件中，保存形式为“标签+文本”的形式。以便后面进行训练。

### 3.1.4 Word2vec 训练的词向量

自然语言处理的问题要转化为机器学习的问题，第一步是要找到一种方法把这些符号数学化。词向量在自然语言处理中发挥着重要作用，所谓词向量就是通过神经网络来训练语言模型，并在训练过程中生成一组向量，这组向量将每个词表示为一个 n 维向量，除了将词表示为向量以外，词向量还要保证语义相近的词在词向量表示方法中的空间距离应当是相近的。

具体步骤如下图：

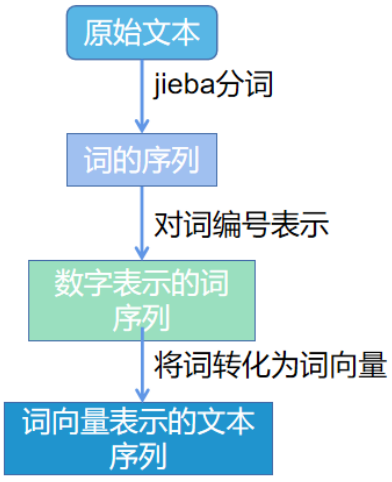



图 2 词向量生成步骤

由上图词向量生成步骤可知，首先是将原始文本分词并转换成以词的序列 ，

然后将词序列转换成以词编号(每个词表中的词都有唯一编号)为元素的序列 ，最后将词的编号序列中的每个元素（某个词）展开为词向量的形式。

### 3.1.5 CNN-卷积神经网络

卷积神经网络（CNN）在计算机视觉领域取得了极大的进展，但是除此之外 CNN 也逐渐在自然语言处理（NLP）领域攻城略地。
CNN 在文本分类基本过程如下：

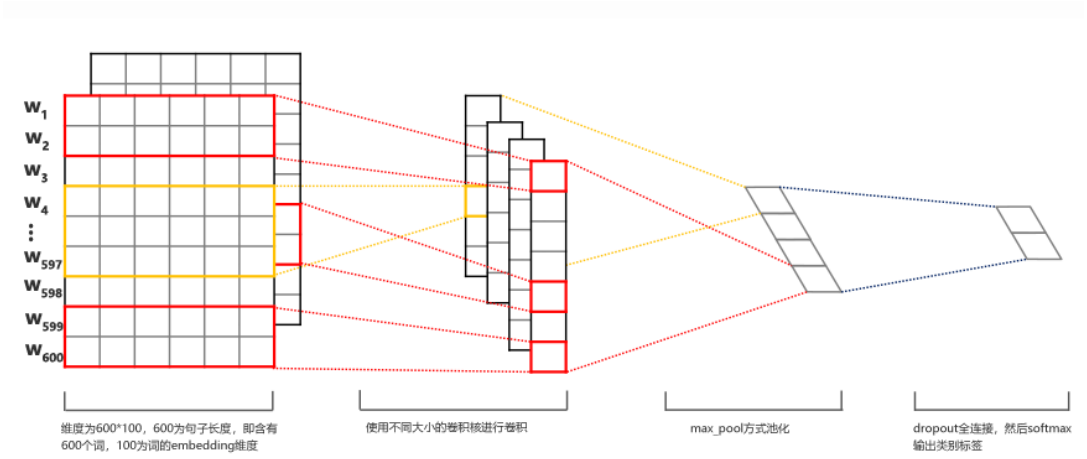



图 3 CNN 文本分类过程

卷积神经网络通常包含以下几种层：

卷积层，卷积神经网路中每层卷积层由若干卷积单元组成，每个卷积单元的参数都是通过反向传播算法优化得到的。卷积运算的目的是提取输入的不同特征，第一层卷积层可能只能提取一些低级的特征如边缘、线条和角等层级，更多层的网络能从低级特征中迭代提取更复杂的特征。

线性整流层，这一层神经的活性化函数（Activation function）使用线性整流。

池化层，通常在卷积层之后会得到维度很大的特征，将特征切成几个区域，取其最大值或平均值，得到新的、维度较小的特征。

全连接层，把所有局部特征结合变成全局特征，用来计算最后每一类的得分

使用神经网络进行文本分类，具体流程图如下：

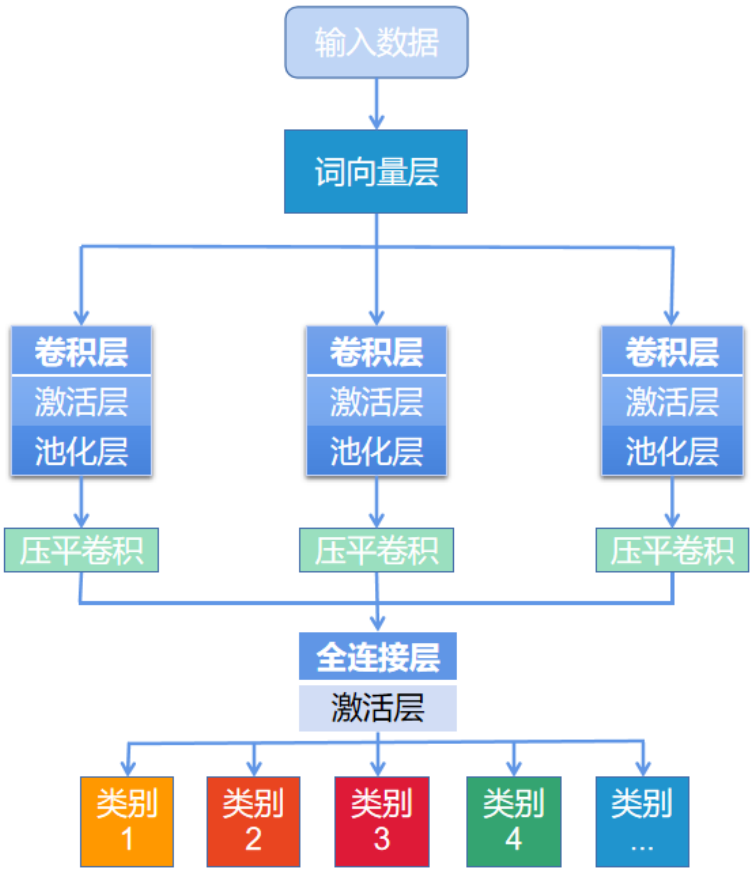



图 4 神经网络分类的流程图

上图中第一层数据输入层，将文本序列展开成词向量的序列，之后连接 卷积层、激活层、池化层 ，这里的卷积层因为卷积窗口大小不同，平行放置了三个卷积层，垂直方向则放置了三重（卷积层、激活层、池化层的组合）。之后连接全连接层和激活层，激活层采用 softmax 并输出 该文本属于某类的概率。

我们将处理后数据按照训练集、验证集、测试集以 8：1：1 的比例保存到 train.txt,val.txt,test.txt 文件中，保存形式为“标签+文本”的形式。以便进行训练。

## 3.2 问题二分析及过程

热点问题是某段时间内群众集中反映的某些问题，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。我们首先观察附件 3 的原始数据，发现有留言主题、详情，点赞数和反对数等这些指标，最终确定以留言主题为主、点赞数和反对数为辅的热点评价指标；进而对源数据进行数据预处理：将一些缺漏的地方进行数据插补处理，并且通过分词、去掉停用词，来筛选出比较核心的关键的词句以提高文本的匹配度。然后用 TF-IDF 算法对留言主题进行文本相似度计算，将主题分类，最后计算分类后的热度评价指标的指数并排序（热度指数 = 留言数*0.9+（点赞数-反对数）*0.1）得到热点问题。并且按照题目所示，用 Excel 简单整理将结果输出。

### 3.2.1 流程图

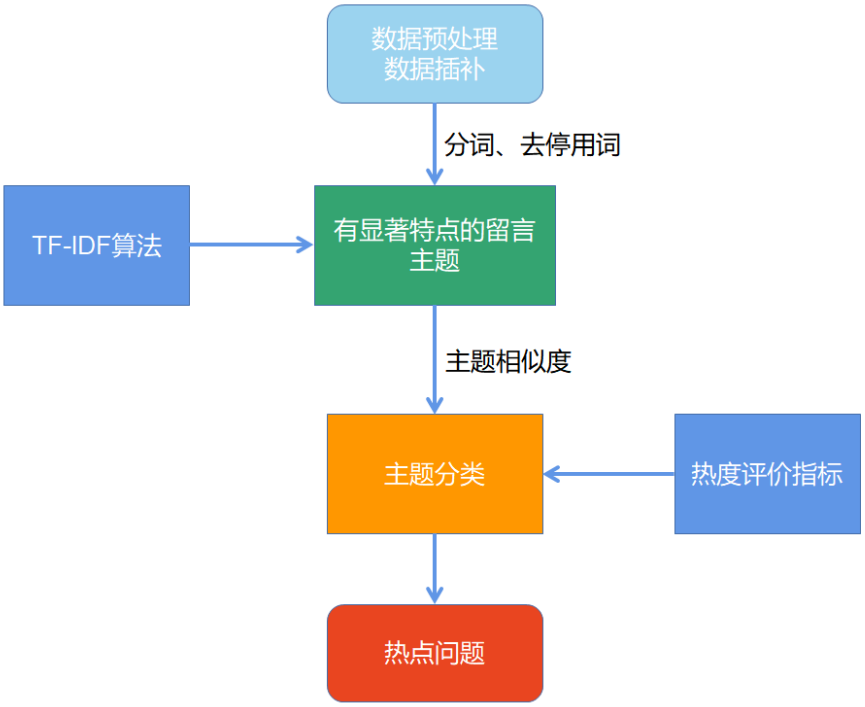



图 5 问题二的流程图

### 3.2.2 数据预处理

附件 3 的留言表共计 4326 条留言，根据数据的特点“留言主题”相对于“留言详情”来说有更明确的人物、地点、事件，所以从留言主题方面进行热点分析在自然语言处理中有更好的效果。

首先对数据进行预处理，对留言主题的个别数据进行查漏补缺，将部分文字缺失的留言进行数据插补处理，使得所有数据都以任务、事件、地点的形式进行保存，使得各个留言都有显著的特点。

接着对每条留言进行分词，便于机器理解与计算；然后去掉停用词，因为停用词在句子中是没有实际含义，相对于关键词来说是无效的，去掉停用词对数据处理的准确度会有一定的提高。

### 3.2.3 TF-IDF 算法

在数据处理之后，每条留言主题都会有简洁且鲜明的特点，如人物，地点，事件，若留言反映的是同一事件，则必然会出现相同的事件主体，这也会导致留言主题文本的相似度增高。通过计算每个留言主题与其他留言主题的文本相似度，就可以把反映相同主题的留言进行归类，在通过以留言条数为主，点赞和反对数为辅就可以大致反映出各个事件的热点程度。

TF-IDF 是一种统计方法，用以评估一个字词对于一个文件集或一个语料库

中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

TF-IDF 的主要思想是：如果某个词在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。具体算法如下步骤：

第一步，计算词频（TF），即词频权重（Term Frequency）

$$TF = n$$

其中 $n$ 表示某个词在文本中出现的次数。

考虑到文章有长短的分别，为了便于不同的文章比较，于是进行“词频”标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即：

$$TF = \frac{n}{N}$$

或

$$TF = \frac{n}{times(n_{\max})}$$

其中 $N$ 表示文本总词数，$times(n_{\max})$ 表示文本出现次数最多的词的出现次数。

第二步：计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境，IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$IDF = \log（\frac{x}{sum(n_x)+1}）$$

其中 $x$ 表示语料库的文本总数，$sum(n_x)$ 表示包含该词的文本数。

第三步，计算 TF-IDF 的值就可以判断两端文本的相似程度，进而判断是否把他们归结为同一个问题。

$$TF - IDF = TF * IDF$$

通过计算 TF-IDF 的值就可以判断两端文本的相似程度,进而判断是否把他们归结为同一个问题。

### 3.2.4 热点问题评价标准

一般来说，热点问题是由问题讨论数和阅读量来判定。对问题的留言就是一种讨论，阅读量可以由点赞数和反对数间接衡量，问题的主要热度还是由讨论数决定，所以我们给留言数和点赞数和反对数赋予一定的权数，考虑到点赞和反对数一般互为矛盾的存在，认可度一般由点赞数看出，故我们最终给出的热度评价指标为：

$$HI = 0.9Ms + 0.1(Ls - Ag)$$

其中 $HI$ 表示热度指数，$Ms$ 表示留言数，$Ls$ 表示点赞数，$Ag$ 表示反对数。

以此为标准来对 TF-IDF 分类后的结果来进行热度指数的计算，筛选出前五的热点问题。

## 3.3 问题三分析过程及方法

对于问题 3，需要制定一个评价指标对相关部门的回复做出评价，经过对所给的数据，结合实际的现实情况分析，当回复的主题与留言者一致时就可以说明这个回复是比较好的回复，同时回复的越完整，回复时间越快，回复态度越好就越能打动留言者，这种回复才是最好的回复。所以我们从回复相关性、回复完整性、回复时效性、回复态度四个方面对回复的留言做出相应的评价。

### 3.3.1 流程图

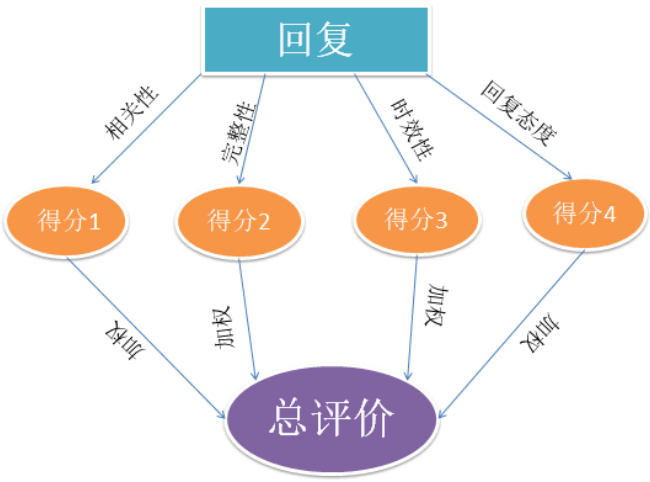



图 6 问题三流程图

### 3.3.2 回复相关性

附件四所给数据中有留言者的留言主题，留言的具体内容和回复的内容信息。留言者在反映一件事情时通常主要针对的是时间、人物、地点、发生事件做出描述，所以回复者也需要“就事论事”才可以与实际留言相关。由此我们通过对留言者的留言主题和留言内容进行分词，提取其主要部分，与回复的留言进行比对，如果这个留言是有效的，留言内容必然会包括留言者所反映的主题与事件，所以我们以留言者回复的核心内容是否在回复中作为一个指标，当回复内容讨论了留言者所反映的所有内容时，相关性就越大。

### 3.3.3 回复完整性

通过对数据进行初步分析，看到一些类似于“已收悉”，“网友：您好！留言

已收悉”等这些字数非常短的留言，反而回复性比较好的留言都具有回复字数长的特点。一般来说，字数越长，回复就越完整，字数越短，回复的完整性就不太高。所以我们以答复的字数长短作为一个完整性评价标准，评价标准得分根据实际情况如下：

表 2 完整性评价标准表

| 字数 | 15 以下 | 15-100 | 100-200 | 200-300 | 300 以上 |
|---|---|---|---|---|---|
| 得分 | 0 | 30 | 60 | 80 | 100 |

### 3.3.4 回复时效性

在所给数据中有留言者的留言时间和相关部门的回复时间。一般来说回复时间越快，这就说明相关人员办事的效率越高，对回复的评价就会越高，回复时间越慢，说明相关人员的办事效率越低，也容易导致留言者的不满。由此我们以回复时效性作为评价留言的一个指标，我们结合实际情况制定了回复时效性得分情况如下：

表 3 时效性评价标准表

| 时间/天 | 5 以下 | 5-15 | 15-25 | 25-35 | 35 以上 |
|---|---|---|---|---|---|
| 得分 | 100 | 80 | 60 | 30 | 0 |

### 3.3.5 回复态度——留言情感分析

回复的态度也是留言者评价回复的一个标准，一般来说回复者的态度越好，留言的评价性就越高。文本情感分析就是一个衡量留言回复态度的指标，分析结果可以判定这个回复是“积极的”还是“消极的”，留言者一般都想得到比较积极的回复。所以我们通过对回复文本进行情感分析，得到回复者态度的指标。态度越好，评分越高，态度越恶劣，评分就越低。

文本情感分析是对带有感情色彩的文本进行分析、处理、归纳和推理的过程。较多用于基于新闻评论的情感分析进行舆情监控。本次评价我们使用的是基于情感词典的情感分析，具体步骤如下图：

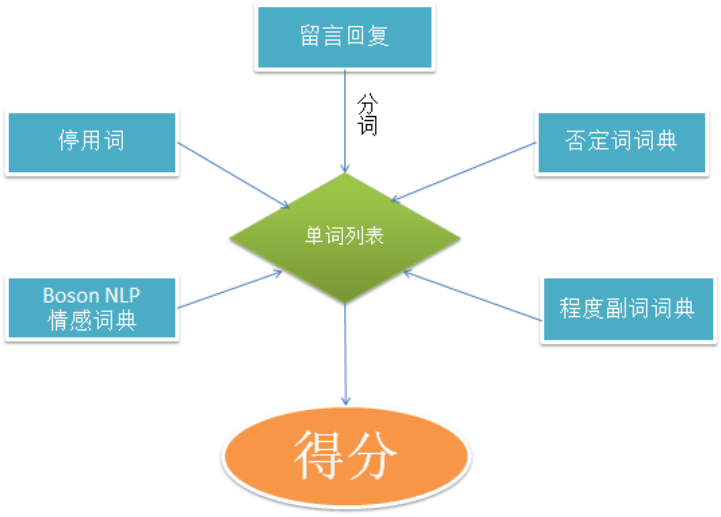



图 7 留言情感分析流程图

情感词典：来源于 Boson NLP 数据下载的情感词典，包含社交媒体词汇与相应的情感得分，积极的表示为正，消极的为负值。

否定词词典：句子中否定词的出现将直接将句子情感转向相反的方向，而且通常效用是叠加的。

程度副词词典：通过打分的方式判断文本的情感正负，那么分数绝对值的大小则通常表示情感强弱，所以就涉及到程度强弱的问题，那么程度副词的引入就是势在必行的。

停用词词典：词典为我国计算所中文自然语言处理开放平台公布的停用词表，停用词在语句中没有太多的意义，对文本情感分析没有用处，在实际的操作中应该去掉语句中的停用词。

# 4 结果分析

## 4.1 问题 1 的结果分析

### 4.1.1 训练集和验证集的训练结果

根据实际数据情况，设置卷积神经网络配置参数如下表所示：

表 4 卷积神经网络配置参数

| 参数名称 | 数值大小 |
|---|---|
| 词向量维度 | 100 |
| 词汇表大小 | 8000 |
| 序列长度 | 1000 |
| 类别数 | 7 |
| 卷积核数目 | 128 |
| Dropout 保留比例 | 0.5 |
| 学习率 | 1e-3 |
| 总迭代轮次 | 30 |
| 每次训练大小 | 64 |
| 每输出一次结果的轮次 | 100 |

在这个配置下，对 7 个分类分为训练集和验证集进行训练和验证，在整个训练过程中持续将输出的最优结果进行保存，直至迭代次数结束。

我们得到在训练过程中的训练集与验证集的错误率曲线和准确率曲线如下图：

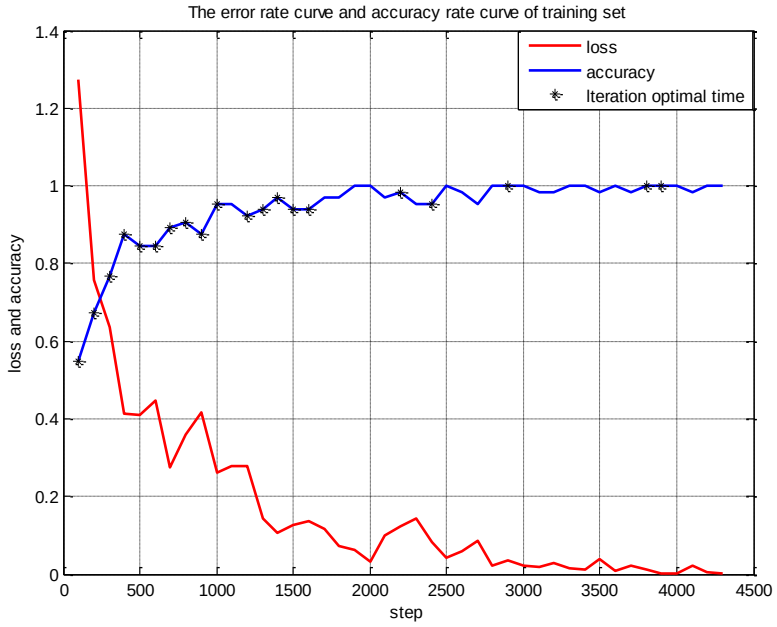



图 8 训练集的错误率曲线和准确率曲线

图中红色的曲线代表训练过程中随迭代步数增加的错误率变化趋势；蓝色的曲线则代表训练过程中随迭代步数增加的准确率的变化趋势；由图中可以看出随着迭代步数的增加，错误率是减少的，准确率是增加的，其中在前期变化得尤为明显，而且波动较大；而后期则趋于平稳；其中星号表示在前 N 次迭代中，达到最优时的步数标记，随着迭代次数的增加，它也在不断更新，一般是出现在前 N 次迭代的准确率的最高点，本次训练我们迭代了 30 次，在迭代 30 次的情况下，到 3900 步是最优的。那时候准确率达到了 100%，错误率也在很低的地方。

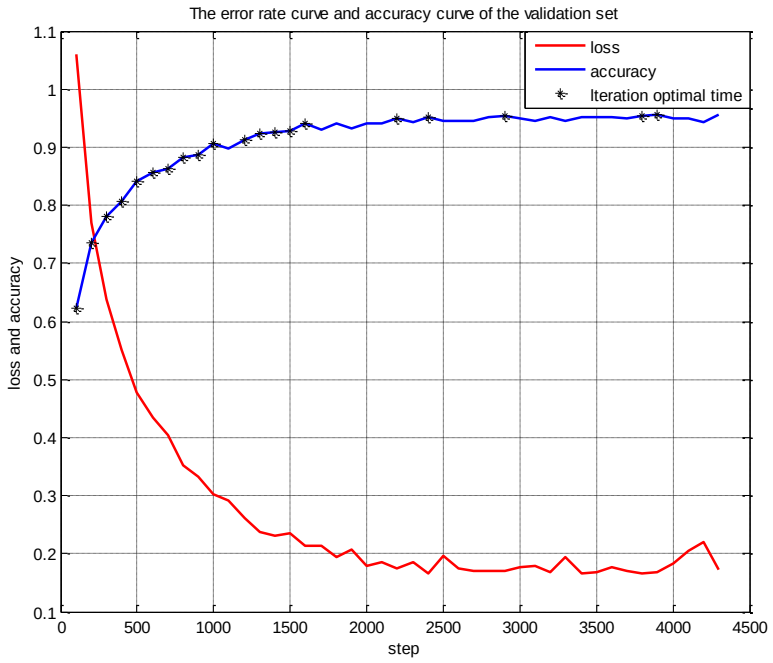



图 9 验证集的错误率曲线和准确率曲线

同理训练集的错误率曲线也是随着迭代步数的增加而减少，准确率是递增的，与训练集不同的是，验证集的错误率曲线逐步递减，准确率曲线起点也比训练集的时候高，变化趋势相对平缓许多，这说明训练的效果是很好的，也是迭代了30次，到3900步是最优的。

综上所述，由上面两个图可以看出在迭代30次的情况下，训练到3900步的时候可以达到最优。

### 4.1.2 测试集的F1-Score评价结果和分类结果

评价一个分类器的指标有很多，比如查准率，查全率，F1-Score 等，其中查准率是针对预测结果而言的，查全率是针对原样本而言的，一般来说，查准率和查全率之间是矛盾的，通常引入F1-Score作为综合评价指标，就是为了平衡查准率和查全率的影响，能较为全面的评价一个分类器。

（1）查准率（也叫精确率，Precision），即正确预测为正的占全部预测为正的比例。

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP}$$

其中$TP$为样本中正确预测为正的样本个数；$FP$为样本中错误预测为正的样本个数。

（2）查全率（也叫召回率，Recall），即正确预测为正的占全部实际为正的比例。

$$\mathrm{Re}call = \frac{TP}{TP + FN}$$

其中$TP$为样本中正确预测为正的样本个数； $FN$为样本中实际为正，但是预测为负的样本个数。

（3）F1-Score ，F1 值为算数平均数除以几何平均数，是查准率和查全率的调和平均，且越大越好。

$$\frac{2}{F_1} = \frac{1}{\mathrm{Pr}ecision} + \frac{1}{\mathrm{Re}call} \longrightarrow F_1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN}$$

其中$TP$为样本中正确预测为正的样本个数；$FP$为样本中错误预测为正的样本个数； $FN$为样本中实际为正，但预测为负的样本个数；

针对本例而言，总的平均F1-Score公式如下所示：

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

其中$P_i$为第$i$类的查准率， $R_i$为第$i$类的查全率。在测试集上进行测试，得到的F1-Score，查全率和查准率如下表

表 5 测试集的查准率，查全率和F1-Score结果

| | 查准率<br>Precision | 查全率<br>Recall | F1值<br>F1-Score |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 城乡建设 | 0.93 | 0.83 | 0.88 |
| 环境保护 | 0.96 | 0.99 | 0.97 |
| 交通运输 | 0.97 | 0.99 | 0.98 |
| 教育文体 | 0.96 | 0.93 | 0.94 |
| 劳动和社会保障 | 0.93 | 0.97 | 0.95 |
| 商贸旅游 | 0.91 | 0.94 | 0.92 |
| 卫生计生 | 0.97 | 0.98 | 0.98 |
| Avg/total | 0.95 | 0.95 | 0.95 |

由表 5 知道每个分类的查准率，查全率和 F1-Score。其中这七个类别的查准率均大于 0.9，都很高，交通运输类别和卫生计生类别的最好；商贸旅游的最低，总的平均查准率为 0.95； 而查全率的两极分化比较明显，城乡建设的查全率只有 0.83，但是环境保护和交通运输类别的查全率高达 0.99，说明这两个类别的召回程度很好，总的平均查全率也为 0.95。

最后看综合评价指标 F1 值,它的分布就比较均匀,其中城乡建设的 F1-Score 比较低， 环境保护，交通运算，卫生计生的 F1-Score 比较高；查阅原始数据记录发现， F1-Score 比较低的类别的原始记录比较多；而 F1-Score 比较高的类别的原始记录是比较少的，采取重抽样使得它的准确率高，总的 F1-Score 为 0.95，总的来说用这个方法训练得到的结果是很不错的。

表 6 测试集的分类结果

| | 城乡建设 | 环境保护 | 交通运输 | 教育文体 | 劳动和社会保障 | 商贸旅游 | 卫生计生 | 测试总和 |
|---|---|---|---|---|---|---|---|---|
| 城乡建设 | 140 | 5 | 2 | 4 | 6 | 11 | 0 | 168 |
| 环境保护 | 1 | 163 | 0 | 0 | 0 | 1 | 0 | 165 |
| 交通运输 | 1 | 0 | 168 | 0 | 0 | 0 | 0 | 169 |
| 教育文体 | 3 | 1 | 1 | 136 | 3 | 2 | 1 | 147 |
| 劳动和社会保障 | 1 | 0 | 0 | 1 | 144 | 1 | 2 | 149 |
| 商贸旅游 | 5 | 1 | 2 | 0 | 1 | 159 | 1 | 169 |
| 卫生计生 | 0 | 0 | 1 | 0 | 1 | 1 | 149 | 152 |

上图为测试集的训练结果，由前面已知，原始数据已经随机以训练集：验证集：测试集为 8：1：1 的比例分开，所以测试集的每一类大约为 150-200 个数据左右。然后代入训练好的模型来进行测试，其中城乡建设抽取了 168 条记录，环境保护抽取了 165 条记录，交通运输抽取了 169 条记录，教育文体抽取了 147 条记录，劳动和社会保障抽取了 149 条记录，商贸旅游抽取了 165 条记录，卫生计生抽取了 152 天数据。

每一行即为测试数据的预测归类情况，第一行是城乡建设类别的预测结果：城乡建设类别的 168 条记录里是预测正确的为 140 条，有 5 条错判给了保环境护类别，2 条错错判给了交通运输类别，4 条错判给了教育文体类别，6 条错判给了劳动和社会保障类别，11 条错判给了商贸旅游类别。同理，其他标签留言验证情况如上表所示。由这些结果我们可以看到这个模型的预测结果还是很不错的，大部分都预测正确了，只有少部分错误。

## 4.2 问题 2 的结果分析

首先我们通过对所有的问题进行编号，在 TF-IDF 算法计算文本相似度时通过大量数据比对，发现在同一主题下，文本相似度达到 60%及以上，在不同主题下文本相似度大部分为 10%以下，为了得到更优的结果我们以相似度为 40%为标准进行文本留言主题分类，达到标准且相似度比其他主题更高时将其归为一类。最终我们将主题归类为了 1162 类，然后根据我们给的评价标准：热度数 = 留言数*0.9+（点赞数-反对数）*0.1。计算热度指数，热度指数为前 5 的详情见下表：

表 7 热度指数前五统计表

| 问题 ID | 留言条数 | 点赞数 | 反对数 | 热度指数 |
|---|---|---|---|---|
| 1 | 59 | 0 | 288 | 24.3 |
| 2 | 44 | 1 | 23 | 37.4 |
| 3 | 29 | 1 | 47 | 21.5 |
| 4 | 26 | 2 | 33 | 20.3 |
| 5 | 23 | 4 | 42 | 16.9 |

表 7 数据显示问题 ID 为 1 的热度指数为 24.3，留言条数为 59；问题 ID 为 2 的热度指数为 37.4，留言条数为 44；问题 ID 为 3 的热度指数为 21.5，留言条数为 29；问题 ID 为 4 的热度指数为 20.3，留言条数为 26；问题 ID 为 5 的热度指数为 16.9，留言条数为 23.五个问题的留言数、点赞和反对数都是比较大的，说明有很多人在关注；而这五个问题的反对数都远远大于点赞数，其中问题 ID 为 1 的反对数高达 288 个，但点赞数为 0，在点赞数和反对数真实的情况下，这说明有两种可能：1、热点问题是有人恶意扩大，纯属谣言，这样则需要有关部分进行调查，确认热点问题的真实性；2、热点问题是真实存在的，但是被投诉的相关人员或者组织进行恶意刷反对数，这也需要有关部门尽快进行核实。总之无论热点问题真实与否，既然其是热点问题，受到很多人的关注，有关部门理应尽快解决人们反映的问题，给出官方的回应。

前五个热点问题如下表所示：

表 8 热力问题表

| 热度排名 | 问题 ID | 热度指数 | 时间范围 | 地点/人群 | 问题描述 |
|---|---|---|---|---|---|
| 1 | 2 | 37.4 | 2019/7/7 至 2019/9/26 | A 市伊景园滨河苑 | A 市伊景园滨河苑车位违法捆绑销售诈骗钱财 |
| 2 | 1 | 24.3 | 2019/11/至 2020/1/25 | A 市 A2 区丽发新城 | A2 区丽发新城附近修建搅拌厂噪音、灰尘污染严重，扰乱居民生活 |
| 3 | 3 | 21.5 | 2019/1/7 至 2020/1/2 | A7 县星沙四区 | A7 县星沙四区凉塘路的旧城改造何时开始 |
| 4 | 4 | 20.3 | 2019/2/1 至 | A 市在住宅小区 | A 市多处住宅楼里开设 |

| | | | | | |
|---|---|---|---|---|---|
| | | | 2019/11/7 | | 麻将馆扰民 |
| **5** | 5 | 16.9 | 2019/1/1 至 2019/12/18 | 有意或已办理 A 市人才新政购房补贴政策的人群 | 咨询 A 市人才购房及购房补贴实施情况等相关问题 |

由上表 8 我们可知前五个热点问题，分别是 A 市伊景园滨河苑车位违法捆绑销售诈骗钱财；A2 区丽发新城附近修建搅拌厂噪音、灰尘污染严重，扰乱居民生活；A7 县星沙四区凉塘路的旧城改造何时开始；A 市多处住宅楼里开设麻将馆扰民；咨询 A 市人才购房及购房补贴实施情况等相关问题。这五个热点问题涉及的方面全都是关乎民生问题，而且集中于住房和生活环境两大方面，可见人们对这两方面的关注度较高。另外，5 个热点问题的时间范围也较长，基本上是在 1 年左右，说明人们的这些问题尚未得到解决，并且人们对这些问题是持续关注的。

热度排名前五的部分留言细则如下：

表 9 热力问题明细表

| 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 | 188809 | A909139 | A 市万家丽南路丽发新城… | 2019/11/19 18:07:54 | A 市万家丽南路丽发新城居民区，开发商在小区旁 50 米处建搅拌站….. | 0 | 1 |
| … | … | … | … | … | … | … | . |
| 1 | 287458 | A909241 | 丽发新城小区旁建了一… | 2019/12/20 14:06 | 丽发新城小区旁建了一个大型搅拌站，运渣车吵得人精神崩溃…. | 0 | 0 |
| 2 | 190337 | A0009051 9 | 关于伊景园滨河苑捆绑销售… | 2019/8/23 12:22 | 投诉伊景园.滨河苑开发商捆绑销售车位! A 市武广新城片区下的伊景园.滨河苑是广铁集团铁路职工的定向商品房，之前…. | 0 | 0 |
| … | … | … | … | … | … | … | . |
| 3 | 188820 | A0002813 8 | A 市星沙城区旧城区棚户改造… | 2019/2/21 11:38:34 | 领导好：目前星沙城区旧城区棚户改造项目正如火如荼的进行当中，星沙 1-6 区的老建筑升级改造，水电煤气也全面提质…. | 0 | 0 |
| … | … | … | … | … | … | … | . |
| 3 | 284120 | A0003563 0 | A7 县星沙镇四区凉塘路….? | 2019/2/14 10:07:59 | 今年政府工作报告里说 2019 年要全部完成星沙的旧城改造，星沙四区凉塘路的这些旧房子什么时候可以开… | 0 | 6 |
| 4 | 285102 | A0004669 1 | A7 县楚绣城麻将馆夜夜扰民 | 2019/6/26 8:53:00 | A7 县开元路楚绣城 d05 栋 3 单元一楼麻将馆上半夜社会闲散人员聚众打牌 | 0 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | （有没有涉赌还请核查）… | | |
| … | … | … | … | … | … | … | . |
| 5 | 283494 | A0008518 5 | 关于《A 市人才购房及购房补贴实… | 2019/7/30 19:06:40 | 1.请问关于<A 市人才购房及购房补贴实施办法（试行）>这个文件现在是否还有效?2.文件第三条 在长工作… | 0 | 0 |

## 4.3 问题 3 的结果分析

我们通过回复相关性、回复完整性、回复时效性与回复的态度四个方面对留言的回复内容进行评价，在实际的评价得分中并不是取这四个指标的平均值，而是根据实际情况指定适当的权重指标。根据这四个指标的重要程度我们做出以下权重排序：相关性>时效性>完整性>回复态度，所赋权值分别为 40%、30%、20%、10%。经过计算得到所有留言的评分数据保存原始的附件中。

通过对所有评分数据进行分析，最高得分为 100 分，最低得分为 0 分，其他的得分情况如下表所示：

表 10 答复的具体得分情况

| 得分范围 | 相关性<br>留言数量 | 完整性<br>留言数量 | 时效性<br>留言数量 | 回复态度<br>留言数量 |
|---|---|---|---|---|
| <0 | 0 | 0 | 0 | 72 |
| 0-10 | 275 | 46 | 346 | 909 |
| 10-20 | 212 | 0 | 0 | 788 |
| 20-30 | 281 | 0 | 0 | 483 |
| 30-40 | 236 | 447 | 218 | 266 |
| 40-50 | 241 | 0 | 0 | 127 |
| 50-60 | 397 | 0 | 0 | 71 |
| 60-70 | 329 | 528 | 464 | 29 |
| 70-80 | 238 | 0 | 0 | 18 |
| 80-90 | 251 | 553 | 915 | 14 |
| 90-100 | 356 | 1242 | 873 | 39 |

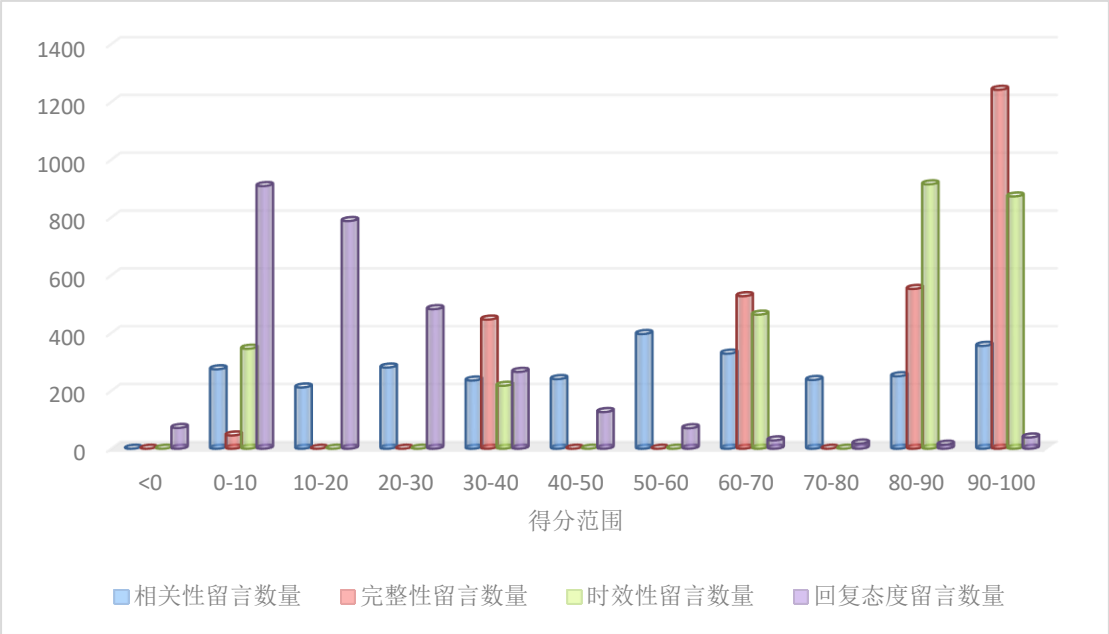



图 10 答复具体得分情况柱状图

通过具体得分的情况柱状图可以看出，以 60 分为及格线，留言的相关性在每一个得分范围里的分布比较均匀，从这一方面看，有关部门的答复情况比较一般，一部分的人群存在为了急于回复，在没看清问题情况下就快速答复或者有意逃避，答非所问；也有另一部分人群能细心的看清楚问题认真回复；从完整性的角度看，分数在合格以上的占大部分，在优秀以上的差不多是一半，说明回复的字数和信息都比较多，比较完整，看出答复人的认真。从时效性的角度看，分数在及格以上比较多，大部分的留言都是能及时回复的；而从回复态度看，一改前几个的趋势，在低分的情况比较多，甚至出现负分情况，说明有关部门的答复态度需要改进。

下面是加权后答复总得分情况表：

表 11 答复总得分情况表

| 得分范围 | 留言数量 |
|---|---|
| 0-10 | 40 |
| 10-20 | 50 |
| 20-30 | 106 |
| 30-40 | 308 |
| 40-50 | 405 |
| 50-60 | 497 |
| 60-70 | 563 |
| 70-80 | 502 |
| 80-90 | 277 |
| 90-100 | 68 |

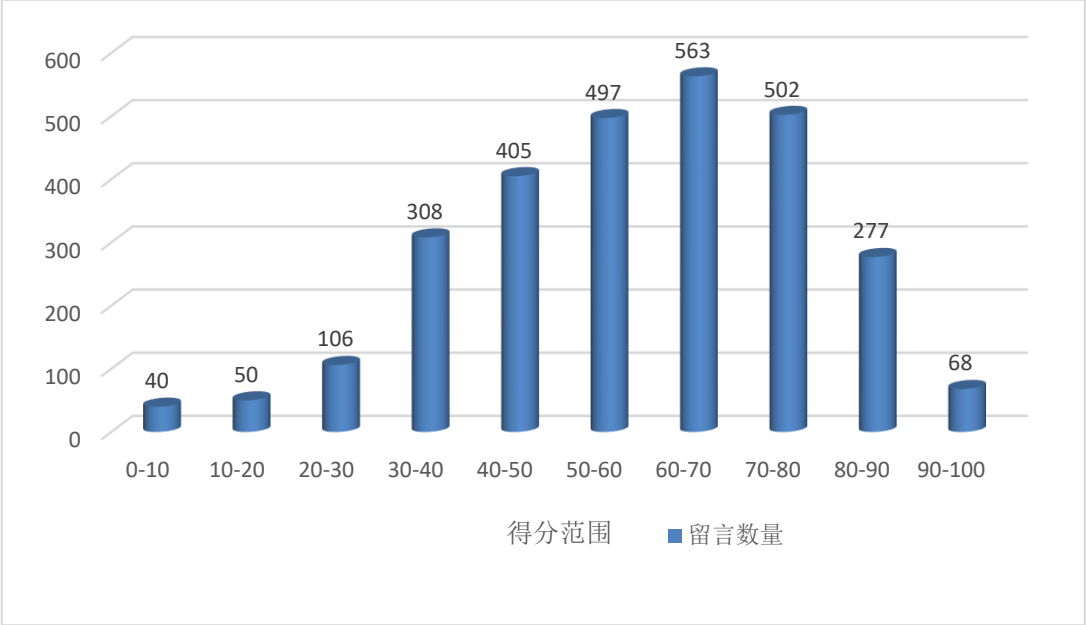



图 11 答复总得分情况柱状图

通过上图可以看出，这个柱状图左偏（即靠近右边的地方高，两头低，左右不对称，左边有一条长尾巴）。以 60 分为及格线，在这 2816 条数据里面，1410 条数据是及格的，有半数以上的留言都是及格的，其中，563 条的分值在 60-70 分之间，502 条数据是在 70-80 之间，277 条数据是在 80-90 之间，68 条数据在 90-100 之间，即有 1/10 以上的回复数达到良好以上，结合具体的得分情况可知，大部分处于中等分数的留言主要是留言相关性不大性以及答复态度太差导致评价得分不太高，启示有关部门需要更加细心，耐心和热情的解决群众反映的问题，这样才能创造更加美好的明天和未来！

# 5 结论

在这个信息与数据快速发展的时代，建立基于自然语言处理技术的智慧政务系统是一个非常重要的文本挖掘应用，传统的人工留言划分和热点问题整理工作已经难以满足数据量庞大的留言体系，本文采用 jieba 分词技术、重采样、去停用词等处理、Word2vec 训练词向量和卷积神经网络技术、TF-IDF 算法计算文本相似度、构建热点评价体系、留言情感分析以及答复评价体系对收集来的群众留言建立一级标签分类模型，最终得到的结果还是很乐观的，总体的 F1-Score 可以达到 0.95；利用热度数 = 留言数*0.9+（点赞数-反对数）*0.1 的热度评价体系，筛选出热度前五的问题，这五个热点问题涉及的方面全都是关乎民生问题，另外，5 个热点问题的时间范围也较长，基本上是在 1 年左右，说明人们的这些问题尚未得到解决，并且人们对这些问题是持续关注的；对答复的相关性、时效性、完整性和回复态度，赋予权值分别为 40%、30%、20%、10%，来得到以 60 分

为及格线，在这 2816 条数据里面，1410 条数据是及格的，有半数以上的留言都是及格的，有 1/10 以上的回复数达到良好以上，大部分处于中等分数的留言主要是留言相关性不大性以及答复态度太差导致评价得分不太高，启示有关部门需要更加细心，耐心和热情的解决群众反映的问题，这样才能创造更加美好的明天和未来！通过对本例的挖掘实现效果看，还是很不错的，但是仍有需要提升的地方，随着数据时代的发展，自然语言处理网络问政平台的文本挖掘应用还有很大的提升空间，依然是一个很热门的应用于问题。

# 参考文献


[1] 程磊,吴晓富,张索非.数据集类别不平衡性对迁移学习的影响分析[J].信号处理,2020,36(01):110-117.

[2] WANG Haitao,HE Jie,ZHANG Xiaohong,LIU Shufen.A Short Text Classification Method Based on N-Gram and CNN[J].Chinese Journal of Electronics,2020,29(02):248-254.

[3] 万磊,张立霞,时宏伟.基于 CNN 的多标签文本分类与研究[J].现代计算机,2020(08):56-59+95.

[4] 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.

[5] 杨立才,李金亮,姚玉翠,吴晓晴.基于 F-score 特征选择和支持向量机的 P300 识别算法[J].生物医学工程学杂志,2008(01):23-26+52.

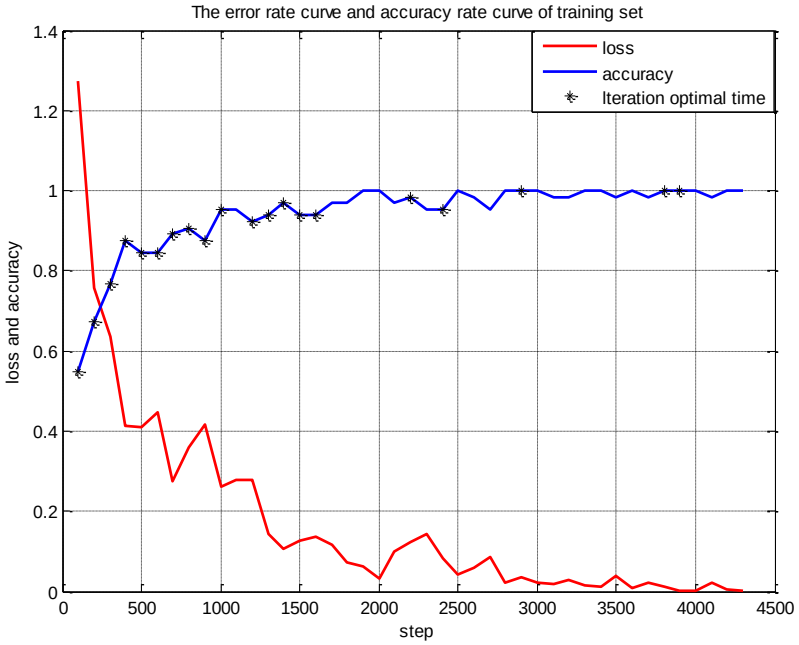
[6] Amit Kumar Sharma,Sandeep Chaurasia,Devesh Kumar Srivastava. Sentimental Short Sentences Classification by Using CNN Deep Learning Model with Fine Tuned Word2Vec[J]. Procedia Computer Science,2020,167.

# 附 录

1. 问题一的训练集的错误率曲线和准确率曲线的 Matlab 代码：

```
step=100:100:4300
loss=[1.273,0.756,0.637,0.412,0.409,0.447,0.274,0.36,0.417,0.263,0.28
,0.279,0.145,0.106,0.128,0.137,0.115,0.072,0.062,0.031,0.101,0.123,0.
142,0.084,0.042,0.06,0.086,0.021,0.036,0.021,0.02,0.029,0.014,0.013,0
.039,0.007,0.022,0.013,0.003,0.002,0.023,0.004,0.002]
accuracy=[0.547,0.672,0.766,0.875,0.844,0.844,0.891,0.906,0.875,0.953
,0.953,0.922,0.938,0.969,0.938,0.938,0.969,0.969,1,1,0.969,0.984,0.95
3,0.953,1,0.984,0.953,1,1,1,0.984,0.984,1,1,0.984,1,0.984,1,1,1,0.984
,1,1]
plot(step,loss,'r-','linewidth',2)
hold on
plot(step,accuracy,'b-','linewidth',2)
hold on
x1=[100,200,300,400,500,600,700,800,900,1000,1200,1300,1400,1500,1600
,2200,2400,2900,3800,3900]
y1=[0.547,0.672,0.766,0.875,0.844,0.844,0.891,0.906,0.875,0.953,0.922
,0.938,0.969,0.938,0.938,0.984,0.953,1,1,1]
plot(x1,y1,'k*')
grid on
xlabel('step');
ylabel('loss and accuracy');
legend('loss','accuracy','lteration optimal time')
title('The error rate curve and accuracy rate curve of training set')
```

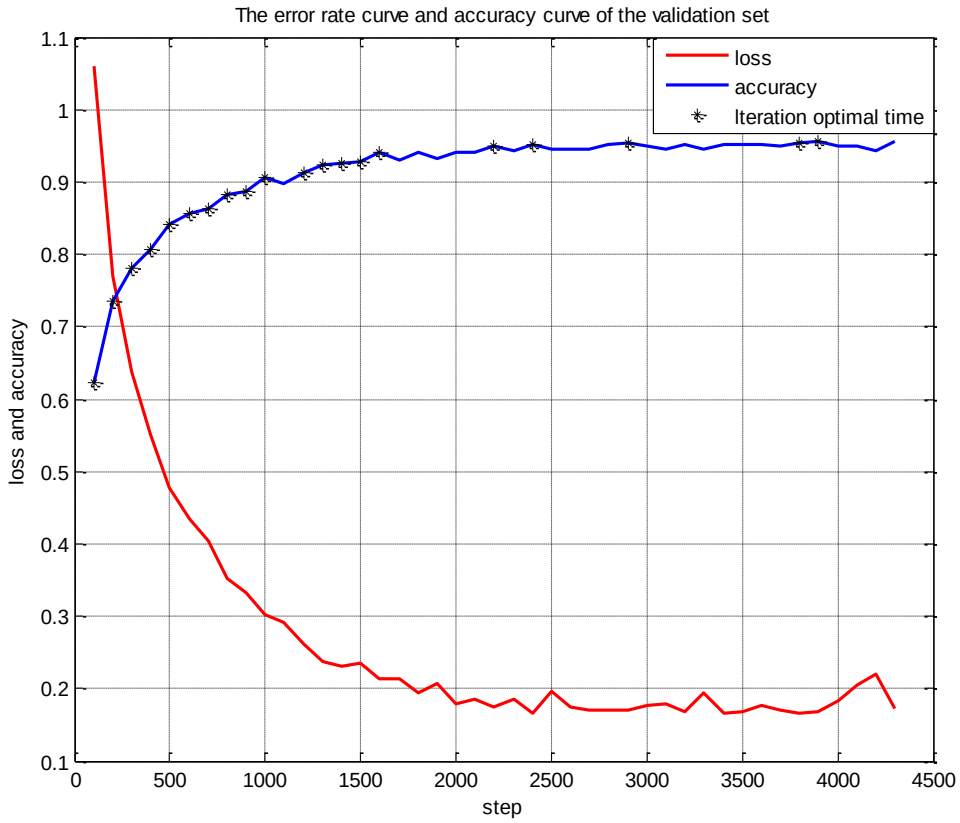
结果：




2. 问题一的验证集的错误率曲线和准确率曲线的 Matlab 代码：

```
step=100:100:4300
loss=[1.059,0.77,0.637,0.551,0.477,0.435,0.404,0.352,0.332,0.303,0.29
1,0.262,0.238,0.232,0.236,0.214,0.214,0.195,0.207,0.18,0.186,0.175,0.
186,0.167,0.197,0.174,0.171,0.171,0.171,0.176,0.18,0.168,0.194,0.165,
0.169,0.177,0.171,0.166,0.169,0.183,0.206,0.22,0.172]
accuracy=[0.622,0.736,0.781,0.807,0.841,0.856,0.862,0.883,0.887,0.906
,0.897,0.912,0.924,0.925,0.927,0.941,0.931,0.94,0.932,0.94,0.941,0.94
9,0.944,0.951,0.946,0.946,0.946,0.951,0.953,0.949,0.946,0.952,0.946,0
.952,0.952,0.952,0.949,0.954,0.956,0.95,0.949,0.942,0.956]
plot(step,loss,'r-','linewidth',2)
hold on
plot(step,accuracy,'b-','linewidth',2)
hold on
x1=[100,200,300,400,500,600,700,800,900,1000,1200,1300,1400,1500,1600
,2200,2400,2900,3800,3900]
y1=[0.622,0.736,0.781,0.807,0.841,0.856,0.862,0.883,0.887,0.906,0.912
,0.924,0.925,0.927,0.941,0.949,0.951,0.953,0.954,0.956]
plot(x1,y1,'k*')
grid on
xlabel('step');
ylabel('loss and accuracy');
legend('loss','accuracy','Iteration optimal time')
title('The error rate curve and accuracy curve of the validation set')
```

结果：




3. 其余代码和数据结果详见作品附件。