“智慧政务”中的文本挖掘应用

摘要

近些年信息网络技术飞快发展，许多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，各类社情民意相关的文本数据量不断攀升。用文本形式表示的信息已经越来越多，仅仅单纯依靠人工劳力达到高效率地获取到海量信息中的关键内容已成为不可能。为了解决这个问题，基于自然语言处理技术的智慧政务系统研究势在必行。本文主要的研究内容如下：

针对问题一：处理数据，建立分类模型，使用 F-Score 对分类方法进行评价。首先，对留言详情进行预处理，基于 Transformer 的双向编码器的 **BERT 方法**表示词向量。其次，建立**长短期记忆网络（LSTM）分类模型**对留言数据进行分类，并与两种传统的机器学习进行对比。最后，利用评价指标中的 F_1 对分类方法进行评价。基于 Logistic 回归、支持向量机（SVM）和 **LSTM** 建立的分类方法求得的 F_1 分别为 0.81、0.78、**0.86**。

针对问题二：根据留言问题进行聚类，定义合理的热度评价指标，并给出评价结果。首先利用正则化表达式进行地址匹配，将有相同地区名的留言详情放在一起形成同一地区留言问题的集合，基于 TFIDF 算法提取出留言详情的关键词后，利用 K-means 聚类算法对留言数据多的集合进行聚类，最后根据评价指标建立热度评估模型，计算出排名前五的热点问题描述分别为：**58 车贷案件、房子存在严重安全隐患、高铁地铁建设问题、小区距离高铁太近、110kv 高压电线架设存在问题。**其对应的热点指数分别为 **0.0406、0.0346、0.0224、0.0111、0.0056**。

针对问题三：建立相关部门答复意见质量评价模型，并根据实验结果对答复意见进行分析。首先建立了判断答复意见质量的优劣准则，并从答复**是否套用固定模板、相关性、完整性和可解释性**等4个方面进行量化描述；其次，提取答复意见的相关性、完整性和可解释性三个指标建立了质量评价模型；最后，计算出答复意见质量评价函数值，详见附件“**答复意见质量评价函数值明细表.xls**”，答复意见质量最好的留言编号为 **6891**。

关键词：LSTM 分类模型；BERT；K-means 聚类算法；热度评估模型；质量评价模型

Abstract

In recent years, with the rapid development of information network technology, many online political platforms have gradually become an important channel for the government to understand public opinion, pool public wisdom and pool people's spirit. As more and more information is expressed in text form, it is impossible to obtain the key content of mass information efficiently only by manual labor. In order to solve this problem, the research of intelligent government system based on natural language processing technology is imperative. The main contents of this paper are as follows:

For problem 1: Process the data, establish the classification model, and use f-score to evaluate the classification method. Firstly, preprocess the message details, and express the word vector by BERT method of two-way encoder based on Transformer. Secondly, a long and short term memory network (LSTM) classification model is established to classify the message data and compare it with two traditional machine learning methods. Finally, the classification method is evaluated by using the evaluation index F_1. The classification methods based on Logistic regression, support vector machine (SVM) and LSTM were 0.81, 0.78 and 0.86, respectively.

Aiming at question two: cluster according to message questions, define reasonable heat evaluation index, and give evaluation results. Firstly, regularization expression is used to match the address, and the message details with the same region name are put together to form a set of message problems in the same region. After extracting the key words of message details based on TFIDF algorithm, k-means clustering algorithm is used to cluster the set with more message data. Finally, the heat assessment model was established according to the evaluation index, and the top five hot issues were calculated as follows: 8 car loan cases, the house has serious security risks, the construction of the high-speed railway subway, the community is too close to the high-speed railway, the 110kv high-voltage line erection problems.The corresponding hot spot indexes are 0.0406, 0.0346, 0.0224, 0.0111 and 0.0056, respectively.

Aiming at question 3: establish the quality evaluation model of the replies from relevant departments, and analyze the replies according to the experimental results. Firstly, a criterion is established to judge the quality of the replies, and the quantitative description is made from four aspects: whether the replies are based on fixed template, relevance, completeness and interpretability. Secondly, the quality evaluation model is established by extracting the three indexes of relevance, completeness and interpretability. Finally, the evaluation function value of reply comments was calculated. Please refer to the “detailed list of evaluation function value of reply comments. xls” for details. The best quality comment number is 6891.

Key word： LSTM; BERT; K-means; Heat assessment model; Quality evaluation model

目录

1 研究背景与意义

随着我国社会经济的发展、人口数量的增加、城市化进程的加快，城市的发展面临一系列困境，如人口流动、环境污染、能源不足、交通拥堵、疾病传播和自然灾害频繁等，其中，人口快速增长与城市承载力不足之间的矛盾、社会管理服务水平不高于公众需求日益增长之间的矛盾已经成为制约我国大中型城市健康发展的重要因素，迫切需要创新发展模式，打破城市发展瓶颈。

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。

大数据时代的发展背景下，数据量的急剧增加远不是传统城市规划研究方法所能解决的，城市作为大数据的加工处理和应用的主要空间场所，需要发展一种基于大数据深度挖掘的综合分析方法，来探寻城市复杂系统中的多层次要素间的适应性相互作用规律。因此，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。

2 变量说明

符号	符号意义
$W^{(i)}$	表示第 i 则文本的向量矩阵；
$TFIDF\left(d_i,t_j\right)$	表示 t_j 在当前留言详情 d_i 中出现的次数；
$DF(t_j)$	表示在文本数据集中出现 t_j 的文本个数；
$T_i(topic)$	表示留言内容关注度；
$O_i(topic)$	表示其他用户关注度；
TP	表示被分类器正确分到类别 C 的个数；
FN	表示被分类器错误的分到类别 C 的个数；
$TEXTSIM(T_k,T_p)$	表示留言问题 T_k 和答复内容 T_p 的一个相似度指标；

3 问题归纳分析

根据题目要求及所提供数据，对所提的三个问题简要归纳如下：

第一、建立关于附件2中留言内容的一级标签分类模型。

在问题一中，根据附件1提供的内容分类的三级标签体系，对附件2中留言内容建立一级标签分类模型。并使用 F_1 对分类方法进行评价。

第二、根据附件3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。

在问题二中，利用 K-means 聚类算法对附件3的留言进行聚类，再根据附件3中的给出的信息定义合理的热度评价指标，建立热度评价模型。并根据热度评估模型按表1的格式给出排名前5的热点问题，且按表2的格式给相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

第三、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、可解释性和是否套用模板等对答复意见的质量给出一套评价方案，并根据相关部门答复意见质量评价模型，计算出各答复意见质量评价函数值。

4 问题一分析及求解

对于问题一的解答，采取如图所示求解思路：

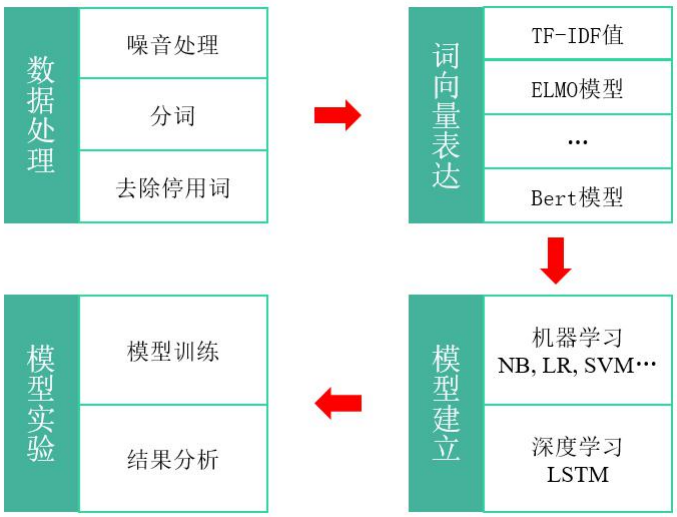

图 1 问题一流程图

4.1 数据处理

针对问题一，对附件 2 中的留言详情进行特征值提取之前必不可少的一步就是对数据进行预处理，预处理工作主要包括：（1）删除噪音数据；（2）对文本进行中文分词处理；（3）删除文本中的停用词。

4.1.1 删除噪音数据

删除留言详情中那些毫无意义的文字或者符号，有利于减小数据空间，降低错误率，提高算法性能。同时需要对这些重复的数据进行清理，减少不必要的计算。

4.1.2 中文分词

中文分词技术是话题检测过程中不可缺少的一项关键技术，是文本信息特征词汇提取的操作基础，同时中文分词也是自然语言处理中不可缺少的一部分，由于中文表述的基本单位是单个的汉字，并且中文句子中的各个词语之间没有明显的分界标志，所以在对中文文本进行信息处理之前需要将文本中的词语进行分割，即进行中文分词处理。

本研究对问题一中附件 2 给出的数据采用 Jieba 分词系统对留言详情进行中文分词。Jieba 分词是 python 编程语言中的一个中文分词包，可以对中文语料素材进行词语切分、关键词提取等，它不但具有较高的分词速度和分词精度，而且还能够开发自定义词典。如果数据直接使用 Jieba 进行分词，会得到不好的分词结果，不利于后续的改善；若分词工具里面加入一些专业词汇（如地名），则分词效果将得到一般的分词工具都不具备某个领域的专业词汇，对附件 2 中的留言详情的直接分词结果和添加人工词典后分词结果举例说明：

表 1 添加人工词典前后分词结果对比

直接分词结果	添加人工词典后分词结果
K8/县冷/江东/路/蓝波/旺/酒店/外墙/装修/无人/施工	K8 县/冷江东路/蓝波/旺/酒店/外墙/装修/无人/施工
A6/区润/和/紫/郡/用电/的/问题/能/不能/解决	A6 区/润和紫郡/用电/的/问题/能/不能/解决
A3/区雨/敞坪/镇/喜/发塘/为何/至今/不通/公路	A3 区/雨敞坪镇喜发塘/为何/至今/不通/公路

从直接分词结果可以看出“冷江东路”、“润和紫郡”、“雨敞坪镇喜发塘”这几个专业词语被拆分开了，则需要自定义词典，把涉及到的专业词汇汇总在一个文件里，并加入到 Jieba 分词器里面。

4.1.3 去除停用词

停用词在这里指的是既不能描述也不能互相区分的词语，为节省存储空间、提高搜索效率，在文本处理时，将标点符号及以下几类字词汇总，生成停用词表，分词时自动过滤。其中包括：超高频的常用词：的、地、得；虚词（介词 / 连词等）：只、当、从。这些词语在文中出现的频率很高，对文本表达没有实际的贡

献，反而会增加文本特征向量的维度，增加计算的时间复杂度和空间复杂度，因此必须将这些停用词删除。

本文根据附件 2 中留言详情和网上收集的停用词相汇集在一起形成一个比较全面的停用词表，该表用于过滤附件 2 的文本分词结果中的停用词。去停用词的思想是对分词结果集合中的词与停用词表中的词进行匹配，如果匹配成功，则删除分词结果集合中相对应的词。

4.2 基于 BERT 词向量

对于文本分类任务，首先要将附件 2 中的留言详情用向量表示，这一过程称为词嵌入（Word Embedding）。当前自然语言处理领域使用的语言模型中，Word2vec 是使用最广泛的词向量训练工具。借助于 Word2vec 方法，研究者可以高效快速的得到语料库中词语的分布式向量表征，相比于传统的独热编码表征形式，分布式向量表征可以对词向量表征的维度进行设定，有效避免了维度灾难问题。但是，研究者发现，Word2Vec 本身是一种浅层结构价值训练的词向量，所“学习”到的语义信息受制于窗口大小，并且不能区别不同语境中同一个词语的不同语义，因此后续有学者提出利用 BERT 模型预训练词向量。

Matthew 等人提出了 Embeddings from Language Models (ELMo)，ELMo 是一种双层双向的 LSTM 结构，其训练的语言模型可以学习到句子左右两边的上下文信息。除此之外，Alec Radford 等人提出了 GPT 利用 Transformer 的编码器作为语言模型进行预训练，下游的自然语言处理任务在其基础上进行微调即可使用。本研究借鉴谷歌的 Jacob 的研究思想，采用基于 Transformer 的双向编码器表示 BERT 方法进行词向量训练。

BERT 模型（如图 2）所示采用的是双向的 Transformer 结构进行编码，其中“双向”意味着模型在处理一个词时，其可以根据上下文的语义关系，表征字在上下文中的具体语义。

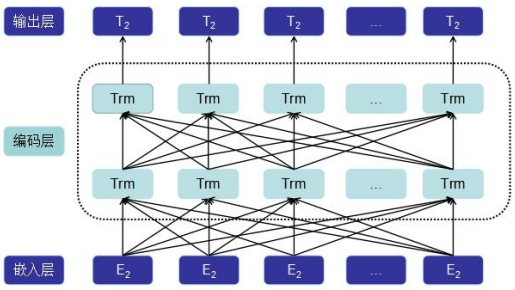

图 2 BERT 模型基本结构示意图

预训练，是 BERT 模型的一个重要阶段，通过对海量语料的训练，使得单词学习到很好的特征表示。通过 BERT 模型训练得到文本的向量表示W：

$$W^{(i)}=\left\{w_1^{(i)},w_2^{(i)},w_3^{(i)},...,w_n^{(i)}\right\}, \tag{4-1}$$

其中$W^{(i)}$表示第i则文本的向量矩阵，$w^{(i)}$表示单个字的表征向量，n表示最大句子长度。

基于 BERT 模型表达附件 2 给出的数据得到的词语是一个 768 维向量，给出一个用 BERT 模型表示词向量的例子如表 2 所示

表 2 词向量表示示例

词	词向量（768维）
A市	[-0.36583388 0.42533386 -0.1564433 -0.38958162 0.23911935 ... 0.07977315 0.08457733]
西湖	[-0.3872587 0.05136739 -0.15343817 -0.33904982 0.47121018 ... -0.37548974 -0.10423942]
建筑	[-0.76753443 0.79983723 -0.5346387 -0.248481 0.5657689 ... -0.04695757 -0.13061032]
集团	[-0.0666619 0.52873003 -0.45179442 -0.34141824 0.6382152 ... -0.09500006 -0.11596178]
占道	[-0.32712218 0.5154143 -0.01553455 -0.2168933 0.11243176 ... -0.3265405 -0.11241053]
施工	[-0.05335409 0.27290833 -0.0572087 -0.14039427 0.21650603 ... -0.5159939 0.17543451]
有	[-0.5860557 0.89086056 -0.37774506 0.0463677 0.3090775 ... 0.16612048 -0.42941812]
安全隐患	[0.5411794 0.06472475 -0.7758526 -0.43303332 0.522739 ... -0.19984733 0.04841183]

4.3 基于传统机器学习的分类方法

机器学习的常用算法根据训练样本是否含有标签可以分为有监督算法和无监督算法两大类。分类问题常常通过有监督学习方法来解决，即利用特征向量和对应标签输出组成的数据集构建并训练一个性能最佳的分类模型，再利用这个模型计算待分类样本的标签，以达到分类的目的。本文要求对附件 2 中的留言详情建立一级标签分类模型，这就是典型的多分类问题。在文本分类领域，常用的分类算法有 Logistic 回归、支持向量机等。

4.3.1 Logistic 回归

（1） Logistic 分布定义

设X是连续随机变量，X服从逻辑斯谛分布是指X具有下列分布函数和密度函数：

$$F(x)=P(X\le x)=\frac{1}{1+e^{-(x-\mu)/\gamma}} \tag{4-2}$$

$$f(x)=F'(x)=\frac{e^{-(x-\mu)/\gamma}}{\gamma(1+^{-(x-\mu)/\gamma})^2} \tag{4-3}$$

μ为位置参数，$\gamma>0$为形状参数。

（2） Logistic 回归模型

二项逻辑斯谛回归模型是一类分类模型，由条件概率分布 $P(Y|X)$ 表示，形式为参数化的逻辑斯谛分布。这里随机变量 X 取值为实数，随机变量 Y 取值为 1 或 0。

二项逻辑斯谛回归模型是如下的条件概率分布：

$$P(Y=1|x)=\frac{\exp(w\cdot x+b)}{1+\exp(w\cdot x+b)} \tag{4-4}$$

这里，$x\in R^n$ 是输入，$Y\in\{0,1\}$ 是输出，$w\in R^n$ 和 $b\in R$ 是参数，w 称为权值向量，b 称为偏置，$w\cdot x$ 为 ω 和 x 的内积。

将权值向量和输入向量加以扩充，记为 w，x，即 $w=(w^{(1)},w^{(2)},...,w^{(n)},b)$，$x=(x^{(1)},x^{(2)},...,x^{n},1)$。逻辑斯谛回归模型为：

$$P(Y=1|x)=\frac{\exp(w\cdot x)}{1+\exp(w\cdot x)} \tag{4-5}$$

$$P(Y=0|x)=\frac{\exp(w\cdot x)}{1+\exp(w\cdot x)} \tag{4-6}$$

（3）多分类 Logistic 回归

将二项的 Logistic 回归推广位多项逻辑斯谛回归模型，用于多项分类。假设离散随机变量 Y 的取值集合是 $\{1,2,...,K\}$，那么多项 Logistic 回归模型是：

$$P(Y=k|x)=\frac{\exp(w_k\cdot x)}{1+\sum_{k=1}^{K-1}\exp(w_k\cdot x)}\quad k=1,2,...K-1 \tag{4-7}$$

$$P(Y=K|x)=\frac{1}{1+\sum_{k=1}^{K-1}\exp(w_k\cdot x)} \tag{4-8}$$

其中，$x\in R^{n+1},w_k\in R^{n+1}$。

4.3.2 支持向量机

支持向量机（Support Vector Machines, SVM）最初由 Vapnik 在 1995 年提出，它作为统计学习理论中较为新颖的一个内容，用于数据分类具有很好的效果。支持向量机的基本原理是：对于一组给定的样本集 $D=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$，$y_i\in\{1,-1\}$，是隶属于两个不同类别的样本数据，通过训练，构建一个间隔超平面，同时在它的两边建立两个和它平行且有一定距离的超平面，尽可能最大化距离，这样最终能得到总误差最小的分类结果。除此之外，在低维空间中的样本可利用核函数映射到高维空间，并在其中使原本线性不可分的样本找到最优线性分类超平面。二维空间最优分类超平面如图 3 所示

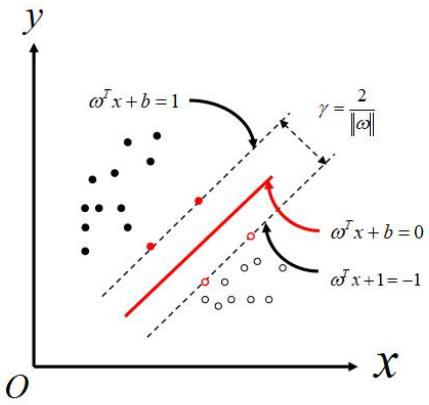

图 3 最优分类超平面

隶属于不同类别的训练样本（正例和反例）用空心点和实心点表示，用以下线性方程描述划分超平面：

$$\omega \cdot x + b = 0 \tag{4-9}$$

其中超平面方向由法向量 $\omega = (\omega_1, \omega_2, ..., \omega_d)$ 决定；超平面与原点间距离由位移项 b 决定。若超平面 (ω, b) 能正确分类文本，即 $(x_i, y_i) \in D$，则正反例的判别：

$$\begin{cases} \omega^T x_i + b \geq +1, & y_i = +1; \\ \omega^T x_i + b \leq -1, & y_i = -1. \end{cases} \tag{4-10}$$

图 3 所示，几个距离超平面（图 3 红线）最大样本点（图 3 两条虚线上的点）使公式（4-10）中的等号成立，它们被称为“支持向量”，它们中的两个正反类到超平面距离（间隔）之和为 $\gamma = 2/\|\omega\|$，使其最大值化，就能最大间隔划分分类超平面，分类误差就越小。核函数 $K(d, d_i)$ 将用特征权值表示特征分类向量 d 和支持向量 d_i 映射到高维线性空间，点积 (d, d_i) 的计算变成核函数计算，再转化成对偶问题计算，从而构成 $\sum \alpha_i^* y_i K(d, d_i)$。最终分类器决策函数：

$$g(d) = \text{sgn}\left\{ \sum_{i=1}^{n} \alpha_i^* y_i K(x_i \cdot d) + b \right\} \tag{4-11}$$

4.4 基于 LSTM 的分类模型

传统的机器学习方法主要是浅层的学习算法，在对留言内容分类任务中不能很好地抽取文本中高层抽象特征，因此为了有效提高对留言分类的准确率，本文引入长短期记忆（LSTM）模型来解决传统机器学习模型在文本分类任务中难以抽取文本中高层语义信息的问题。并将分类结果与上一章的传统机器学习效果进行对比。

4.4.1 LSTM 算法

长短期记忆网络（LSTM）是 RNN 的一种改进，他通过引入门机制构建特殊神经单元，从而解决 RNN 不能实现信息长期依赖问题。LSTM 结构如图 4 所示，其包括输入门i_t、输出门o_t、遗忘门f_t等门结构，这些门结构通过以下的递归方程来更新细胞状态C_t，同时激活从输入到输出的映射。LSTM 模型的记忆单元结构如图 4：

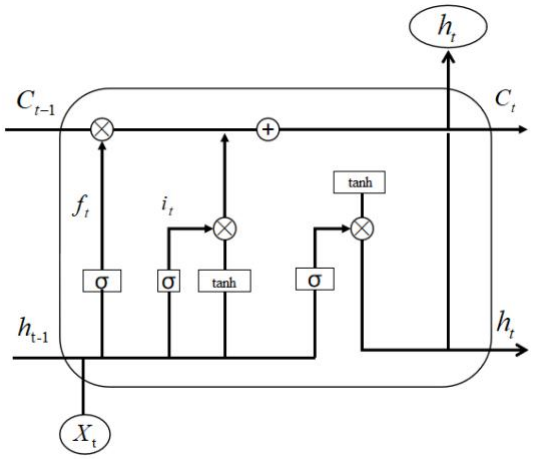

图 4 LSTM 记忆单元结构图

LSTM 的公式如下所示：

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \tag{4-12}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \tag{4-13}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \tag{4-14}$$

其中，f_t、i_t、o_t分别表示t时刻的遗忘门、输入门、输出门。σ表示激活函数。W和b分别表示权重矩阵和偏置。h_{t-1}表示$t-1$时刻 LSTM 细胞的输出，x_t代表t时刻的输入。

遗忘门f_t决定神经单元遗弃哪些信息，该门层通过读取h_{t-1}和x_t的状态，输出一个 0 到 1 之间的值，0 代表完全舍弃，1 代表完全保留。输入门i_t决定神经单元要更新的值，它从遗忘门筛选完的信息中，通过 tanh 函数更新神经单元状态。最终神经单元输出的状态由输出门o_t决定，其先用 sigmoid 层决定要输出的神经单元状态，然后将这些状态用 tanh 函数压缩在到 1 之间。

4.4.2 模型的构建

该模型采用预训练的 BERT 词向量构成的文本序列矩阵作为输入，能够较好的获得词的语义和句法信息，然后利用其特有的记忆单元结构来抽取蕴含留言内

容上下文信息的特征表达，最后将该特征向量作为 softmax 分类层的输入，来建立留言内容一级分类模型。

基于 LSTM 原理，构建基于 LSTM 分类模型，如下图所示，该模型分为五层：其中输入层表示留言分词后的词序列；第二层表示 LSTM 层；全连接层 1 有 64 个神经元；全连接层 2 有 32 个神经元输出层：最后一层中的 7 个神经元对应 7 种留言分类：

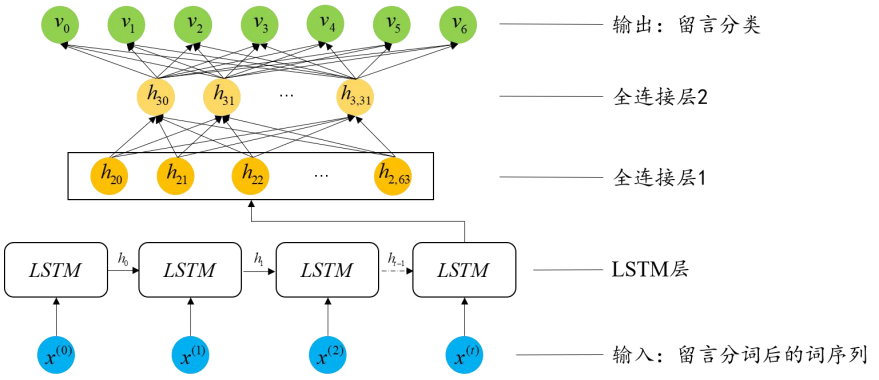

图 5 LSTM 模型的神经网络结构图

4.5 分类系统实验及结果分析

4.5.1 硬件与软件情况

（1）仿真系统硬件

本文的仿真实验均在笔记本电脑上进行，具体参数为英特尔 i5 处理器，12G 运行内存， 2G 独立显卡，Windows10 操作系统。

（2）仿真系统软件

1）编程语言：本文实验均采用 Python 语言进行。

2）相关工具包：Jieba 工具包，主要用来对文本进行分词。Matplotlib 工具包，主要用来画图操作。Scikit learn 工具包，用于传统机器学习建模。Tensorflow 工具包，用于训练 LSTM 网络。

4.5.2 仿真系统的设计和搭建

针对问题一的文本分类系统，主要包括对附件 2 中留言内容的预处理、文本词向量表示和文本分类模块。分类器模块流程图如图 6 所示：

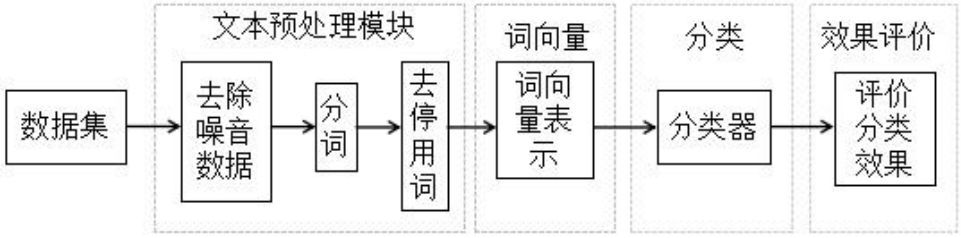

图 6 分类器流程图

4.5.3 分类训练集

在对留言数据分类过程中，数据集中的数据对象是留言内容。数据集共包括9211 条留言数据。通常，把数据集分为训练集和测试集，训练集主要用来构造分类模型，测试集主要是测试经过训练集训练后构造的分类模型的性能。常见的训练集和测试集的构建方法是将数据集的 80%作为训练集，剩下的 20%作为测试集。

本文中的数据集是附件 2 中提供的留言数据。主要的一级分类包括：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 7 个类别。

4.5.4 评价指标

任何研究都需要一个指标用于客观评价，留言内容分类也不例外。分类性能评价的度量包括召回率、精确率、 *F* 值等。其中， *TP* 指的是被分类器正确分到类别 C 的个数； *FN* 指的是实际不属于类别 C 却被分类器错误的分到类别 C 的个数。

（1）召回率（Recall）

召回率，又称为查全率，是通过算法检测出来的相关文档的数据集与原始的文档数据集总数的比率，可以用于衡量检索系统或者分类算法的查全率。其定义：

$$R=\frac{TP}{TP+FN} \tag{4-15}$$

（2）准确率（Precision）

准确率，又称精度、正确率，即通过算法检测出来的的相关文档数据集与所有被检测出来的文档数据集的比率，可以用于衡量检索系统或者分类算法的查准率。其定义：

$$P=\frac{TP}{TP+FN} \tag{4-16}$$

其中召回率和准确率是两个互相矛盾的衡量指标。

（3）F 值

对于系统的评价指标就存在一个问题，若分类器倾向于给出一个更加精确的结果，那么准确率必须很高，反之，若分类器偏向于找出更多相关的内容，那么高的召回率又是必要的。因此，如果分类器仅仅用召回率或准确率对系统性能作评价，那么评估的结果显然是不理想的。因此我们引入 F 值，同时考虑召回率和准确率，其定义：

$$F_{\beta}(P,R)=\frac{(\beta^2+1)PR}{\beta^2P+R} \tag{4-17}$$

其中 β 是一个调整参数，用于以不同权重综合查全率和查准率。在本章中，$\beta=1$，即表示查全率和查准率被平等对待，且此时 F 值又被称为 F_1 指标，定义：

$$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{4-18}$$

F 值是信息检索与分类中常用的一个参数，它可以平衡地反映召回率和准确率。

4.6 实验结果与分析

4.6.1 实验结果

经过上述模型的构建和分析，以及实验数据的选择和处理，本文进行了基于 LSTM 模型的留言详情分类实验。同时，为了保证实验的客观性，本文选取了在文本分类任务中较为常用的两种传统机器学习模型来进行对比实验，其结果见表 3（保留两位小数）。

表 3 不同算法的性能比较

	召回率	准确率	F_1
Logistic	0.82	0.82	0.82
VSM	0.80	0.76	0.78
LSTM	0.88	0.86	0.86

图 7 的直方图更直观的反映了实验结果的对比：

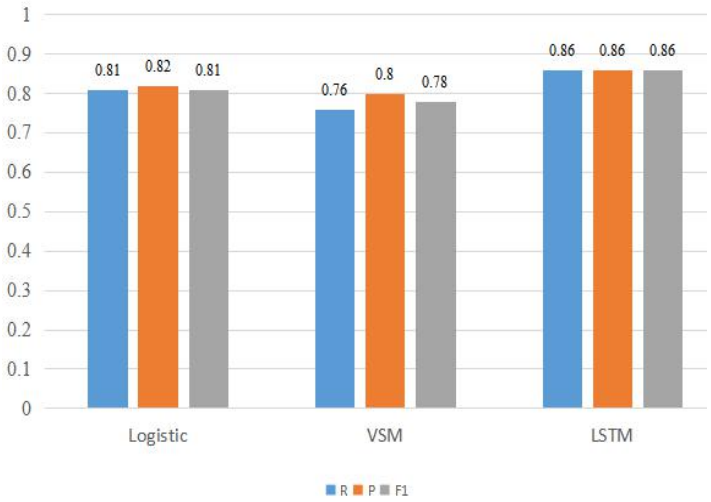

图 7 算法性能对比图

4.6.2 实验结果分析

由上表 3 和图 7 可以看出，相比于传统的机器学习模型，基于深度学习的 LSTM 模型在留言详情分类的召回率、准确率和 F_1 上有较为明显的优势。其中 Logistic 算法的 F_1 要比 SVM 算法的表现好。而 LSTM 算法是三者中最好的。该结果表明，在相同条件下 LSTM 模型在文本分类任务中比传统机器学习模型有着更高的准确率。分析其原因，一方面由于 LSTM 模型的记忆单元结构能够很好的对文本序列建模，来存储文本中的上下文信息，并从中提取文本中更为丰富的高层特征，进而提升文本分类任务的准确率。另一方面，本文采用的是更优秀的词向量来进行留言主题的向量化表示，与传统分类算法所采用的词袋模型文本表示法相比，词向量在预训练的过程中就已经对文本中的语义信息进行了一定程度的提取，且其向量结构较为稠密，能够更好的进行运算，因此将 BERT 词向量与 LSTM 模型相结合来进行留言主题分类任务取得了更为理想的效果。

图 8 展示了 LSTM 网络训练过程中模型的准确率，损失函数在训练集和测试集上的变化，模型在训练 15 次左右后逐渐收敛。

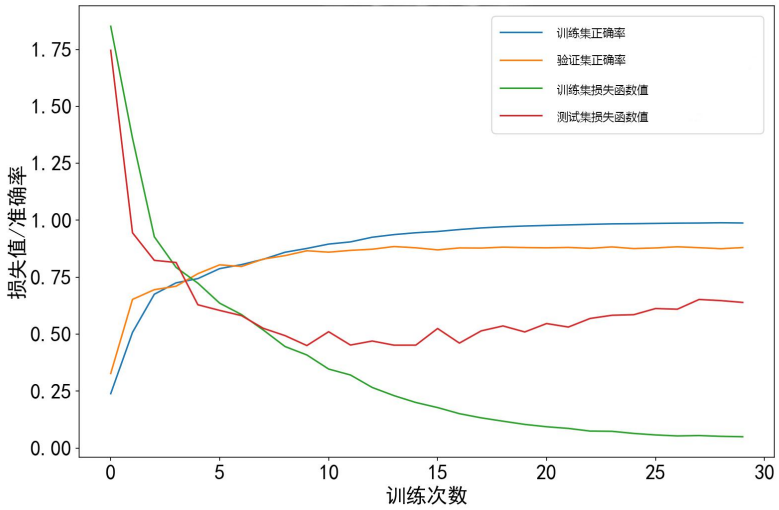

图 8 LSTM 网络训练过程

5 问题二分析及求解

对于问题二的解答，采取如图 9 的求解思路：

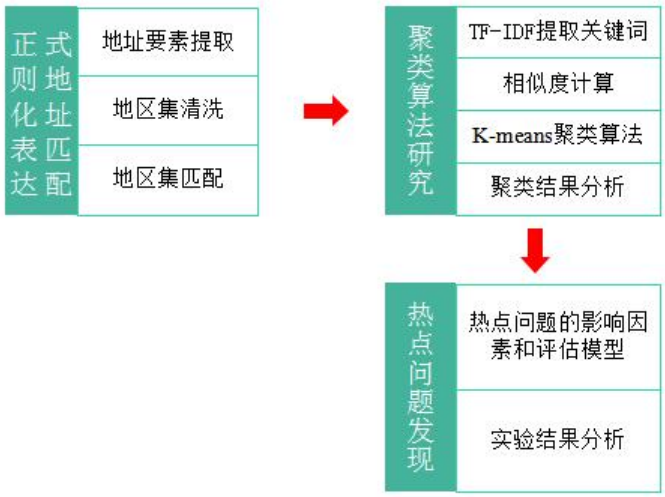

图 9 问题二流程图

5.1 基于正则表达式地址匹配

5.1.1 地址要素的提取

本文首先对留言详情中涉及的地址进行提取，构造地区集，鉴于留言详情中包含的地址文本长度较短，同时具备较强的层级结构信息，因而本文提出借助正则表达式和基于词典分词的地址提取方法。

由于省、市、区、县等高级别地址要素的提取较为容易，而构建镇、村、街道、社区等细类地址要素词典则相对困难。因此，本文提一种基于规则匹配和分

词工具相结合的方法完成对地址的提取。但留言数据中只出现过一次“西地省”，即使用基于正则表达式的方式完成市、区、县等各级地址要素的提取。然后对于镇、村、街道、社区等地址要素，需先利用问题一 Jieba 分词工具对留言数据进行分词，再基于正则表达式对地址要素进行匹配，最后共有 644 个地区名。匹配规则如表 4 所示。

表 4 匹配规则

地址要素级别	匹配规则
市	‘[A-Z]市’和‘[A-Z][0-9]市’
区	‘[A-Z]区’和‘[A-Z][0-9]区’
镇	‘[\u4e00-\u9fa5]{1,3}?镇’
村	‘[\u4e00-\u9fa5]{1,3}?村’
街道	‘[\u4e00-\u9fa5]{1,3}?街道’
小区	‘[\u4e00-\u9fa5]{1,7}?小区/社区’

5.1.2 地区集清洗

基于规则匹配和分词工具相结合的方法构造地区集过程中难免发生错误及疏漏，数据质量较低等问题，因此需要总结数据集中的常见错误形式，并根据后续工作的要求完成对数据的清洗工作。将地区集中一些不是具体地名的小区删除如：“毁坏小区”，“投诉小区”等。

5.1.3 地区集匹配

为了将有相同地区名的留言详情放在一起形成针对某一地区的留言集合，在留言详情与地区集进行匹配时，为了使留言详情不被重复匹配，则首先从最低级别进行匹配，为了解释这一情况，给出一个示例，如“A7 县楚龙街道龙塘路（断头路）何时拉通”，则对楚龙街道进行匹配过后，该条留言内容不再对 A7 县进行匹配。为了分析地区集匹配结果，给出匹配的楚龙街道的留言内容集合如表 5 所示：

表 5 楚龙街道留言内容集合

序号	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
0	190754	A0008998	A7县万家丽北路楚龙街道长期拥堵问题求解决	2019/3/1	本人家住	0	0
1	205574	A0001084	A7县楚龙街道龙塘路（断头路）何时拉通	2019/8/2	请问A7县	0	3
2	206725	A0001654	A7县楚龙街道红树湾小区的消防设施瘫痪了	2019/2/1	A7县楚龙	0	0
3	207651	A0005779	A7县楚龙街道汽配城路交通安全隐患重重	2019/11/	尊敬的沈	0	11
4	211006	A0001084	A7县楚龙街道山水湾小区大量粪池车长期乱停	2019/7/2	A市万家丽	0	1
5	222385	A0003622	A7县楚龙街道楚瑞家园小区白蚁防治治标不治本	2019/3/2	A7县楚龙	0	0
6	222668	A0007687	询问A7县楚龙街道拆迁的问题	2019/2/2	张县长，	1	1
7	224045	A0006300	A7县楚龙街道南塘路交通状况整治问题	2019/1/8	位于楚龙	0	1
8	234210	A0008789	有没有人管管A7县楚龙街道红树湾小区	2019/4/8	楚龙街道	0	0
9	234989	A0007902	A7县楚龙街道山水湾小区房屋漏雨，物业不处理	2019/2/2	县领导好	0	0
10	243062	A0001028	A7县楚龙街道山水湾童话里幼儿园何时进行改成公立幼儿园	2019/1/7	尊敬的书	0	5
11	248372	A0003034	反映A7县楚龙街道高沙村土地确权的问题	2019/5/2	看到很多	0	1
12	248547	A0009806	A7县楚龙街道三一街区小区13栋103违法住改商，开设快递站	2019/8/9	A7县楚龙	0	0
13	252244	A0001069	A7县楚龙街道盛地春天里商铺业主维权难	2019/10/	楚龙街道	0	0
14	256983	A0001130	反映A7县楚龙街道楚核星城居民房屋安全问题	2019/12/	位于楚龙	0	0
15	260479	A0006178	A7县楚龙街道龙塘安置区内有多个不正规“洗脚按摩场所”	2019/6/2	龙塘小区	1	0
16	263602	A0001067	A7县楚龙街道社区村干部高息借贷不还，百姓讨要无门	2019/8/1	欠债还钱	0	0
17	263841	A0001084	A7县楚龙街道山水湾小区电梯存在安全隐患	2019/9/1	针对山水	0	4
18	275019	A0007687	询问A7县楚龙街道拆迁的问题	2019/2/2	领导，您	0	0

5.2 基于留言内容聚类算法研究

地址匹配后的某一留言内容集合内的数据很多，但留言详情反映了不止一种问题，如表 5 所示。因此分别对每个集合的留言内容进行聚类处理。为了能使文本聚类的结果更加简洁而直观地表达出来，需要利用词频统计 TF-IDF 算法提取的关键词。

5.2.1 基于 TF-IDF 算法提取关键词

词频统计规律应用于关键词提取，提出基于词频统计的文本关键词提取方法。首先分别对每个留言内容集合留言数据进行关键词的提取。TF-IDF 算法是经典的关键词提取方法，算法流程分为三大模块：

（1）文本预处理模块

对于输入留言详情 d，首先进行分词等预处理操作，然后得到 d_i，即把文本 d 的内容看成由特征项（词、词组）组成的集合，留言详情 d 可以用特征项表示为：

$$d_i = (t_1, t_2, ..., t_j, ..., t_D) \tag{5-1}$$

其中 t_j 是特征项 $1 \le j \le D$。

（2）权重计算模块

根据各个项 t_j 在文本 d_i 中的重要性给其赋予一定的权重 w_i，TF-IDF 算法通过特征词的词频（Term Frequency, TF）和反文档频率（Inverse Document Frequency, IDF）来计算特征词 t_j 的权重 w_i，对于留言详情 d_i 中的特征词 t_j，其权重 w_j 计算公式如下

$$TFIDF\left(d_i,t_j\right)=TF\left(d_i,t_j\right)\times IDF(t_j)=TF\left(d_i,t_j\right)\times\lg\left(\frac{M}{DF(t_j)}\right) \quad (5\text{-}2)$$

其中特征词$TFIDF\left(d_i,t_j\right)$表示$t_j$在当前留言详情$d_i$中出现的次数，$DF(t_j)$表示在文本数据集中出现$t_j$的文本个数，$M$为文本数据集总数，$IDF(t_j)$是反文档频率，即$\lg\left(\frac{M}{DF(t_j)}\right)$。

（3）提取关键词提取

Sort(d_i)即按照权重w_j从大到小对特征词t_j进行排序，选取前k个词作为留言详情d最终关键词。

5.2.2 相似度计算

文本相似性是文本聚类的依据。设有n个留言的集合$X=\{x_1,x_2,...,x_n\}$，其中每个文件有n个属性，则第i个留言$x_i=\{x_{i1},x_{i2},...,x_{in}\}$，用$dist(x_i,x_j)$表示留言$x_i$和$x_j$的距离，本文利用欧式距离公式来计算$x_i$和$x_j$之间的距离：

$$dist(x_i,x_j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+...+(x_{in}-x_{jn})^2} \quad (5\text{-}3)$$

由公式可知，欧氏距离的取值范围为 0 到∞，欧式距离的值越小，表明两个向量之间的相似度越大。

5.2.3 K-means 算法

K-means 算法是一种基于划分的聚类算法。该算法的最终执行结果是将数据集合中数据划分成 K 个聚类。同时，K-means 算法也是一种动态聚类算法，使用 K-means 算法对数据集合进行聚类处理的过程中包含多次迭代操作。当完成一次迭代操作时，即对数据集合中的数据对象完成一次聚类处理，需要检测每个数据对象是否均被划分到最佳的聚类中，如果划分结果出现错误，就要对该数据对象重新划分并将其划分到最佳的聚类中。当所有的数据对象都完成检测和最佳划分后，需要对重新划分后的聚类的聚类中心进行调整，然后重复执行上述迭代过程，直到检测错误划分的过程中不存在被划分到错误聚类的数据对象，即每个数据对象均被划分到最佳的聚类中，则表明数据集合的已经成功的被划分成K个聚类，K-means 算法执行结束。

K-means 算法的描述过程中主要包含以下几个专有名词：

（1）聚类中心：代表聚类内所有数据对象的中心位置，一般使用几何中心位置进行描述，计算方法：

$$Z_j(I)=\frac{1}{n_j}\sum_{i=1}^{n_j}x_i^{(j)} \qquad j=1,2,...,k \qquad (5\text{-}4)$$

（2）与聚类之间的距离：数据对象与聚类之间的距离采用该对象与该聚类的聚类中心之间的距离进行描述。这个距离数值可以用于衡量一个数据对象是否可以被划分到一个聚类中，数值越大，说明该数据对象越有可能被归属该聚类中。该距离的计算公式：

$$D\left(x,Z_j(I)\right)=\sqrt{\left(x_1-x_{j1}\right)^2+\left(x_2-x_{j2}\right)^2+\cdots+\left(x_n-x_{jn}\right)^2} \qquad (5\text{-}5)$$

其中，$x=\left(x_1,x_2,\cdots,x_n\right)$表示数据对象，$Z_j(I)=\left(x_{j1},x_{j2},\cdots,x_{jn}\right)$表示聚类 j 的聚类中心。

（3）准则函数：该函数用于衡量各个聚类之间的相似度。K-means 算法的最终执行结果是将数据聚合划分为 k 个聚类，并且每个聚类之内的各个数据对象之间具有很高的相似度，而各个聚类之间具有较小的相似度。如果准则函数的计算结果小于预先设定的阈值，表面该算法的执行结果是一个比较理想的结构，此时可以结束算法的执行过程。该准则函数为：

$$J_c(I)=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left\|x_i^j-Z_j(I)\right\|^2 \qquad (5\text{-}6)$$

（4）迭代终止条件：用于衡量 K-means 算法的执行结果是否为一个比较理想的结果，即是否可以终止 K-means 算法的执行过程。K-means 算法的迭代终止条件如下：

$$\left|J_c(I)-J_c(I-1)\right|<\xi \qquad (5\text{-}7)$$

5.2.4 K-means 算法执行流程

首先根据设定的 k 值从输入的包含 n 个数据对象的数据集合使用随机函数选取 k 个数据对象作为初始聚类中心，计算数据集合中每个数据对象与 k 个聚类之间的距离，根据最近划分规则将数据集合划分为 k 个分类，完成对数据集合的初始划分；然后根据聚类中包含的数据对象定义该聚类的聚类中心作出调整；最后计算每个数据对象与聚类中心之间的距离，并根据最近划分规则对数据进行重新调整，循环执行上面的步骤直至分类集合中的各个分类满足迭代终止条件。其算法的具体步骤描述为：

表 6 K-means 算法的具体流程

输入：包含 n 个数据对象的数据集合，最终获得的聚类个数 k

输出：使准则函数满足迭代终止条件的 k 个聚类

执行流程：

Step 1：从包含 n 个数据对象的初始数据集合中随机选择 k 个数据对象，将这 k 个数据对象作为初始聚类中心。

Step 2：将数据集合中每个数据对象与选定的 k 个聚类进行比较，并利用

$$D\left(x,Z_j(I)\right)=\sqrt{\left(x_1-x_{j1}\right)^2+\left(x_2-x_{j2}\right)^2+\cdots+\left(x_n-x_{jn}\right)^2}$$

计算它们之间的距离，根据最近分配原则将数据对象合并到最佳的聚类中，完成对数据集合的聚类划分过程。

Step 3：利用公式

$$Z_j(I)=\frac{1}{n_j}\sum_{i=1}^{n_j}x_i^{(j)}\quad j=1,2,\cdots,k$$

重新计算刚划分好的 k 个聚类的聚类中心。

Step 4：重复执行步骤 Step2 和 Step3，直到划分的聚类能够使准则函数满足迭代终止条件，此时执行结束，最后获得的 k 个聚类即为所需要的结果。

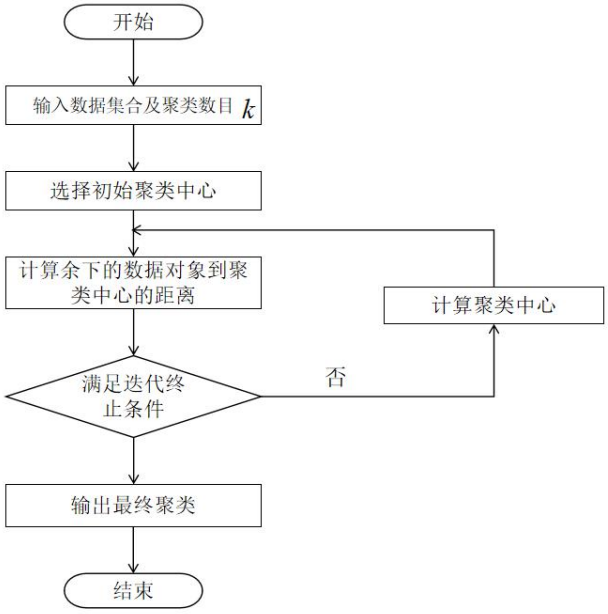

图 10 聚类算法流程图

5.2.5 聚类结果

基于 K-means 聚类算法处理后，每一类地址里面的留言详情又会根据提取的关键词的相似度分割成 k 个聚类，每个聚类包含的所有留言详情均指向同一个话题，针对 A7 县楚龙街道为例，进行聚类后的结果如表 7：

表 7　A7 县楚龙街道聚类结果

序号	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	label
0	190754	A0008998	A7县万家丽北路楚龙街道长期拥堵问题求解决	2019/3/1	本人家住	0	0	0
1	206725	A0001654	A7县楚龙街道红树湾小区的消防设施瘫痪了	2019/2/1	A7县楚龙	0	0	0
2	207651	A0005779	A7县楚龙街道汽配城路交通安全隐患重重	2019/11/	尊敬的沈	0	11	0
3	211006	A0001084	A7县楚龙街道山水湾小区大量粪池车长期乱停	2019/7/2	A市万家丽	0	1	0
4	222385	A0003622	A7县楚龙街道楚瑞家园小区白蚁防治治标不治本	2019/3/2	A7县楚龙	0	0	0
5	234989	A0007902	A7县楚龙街道山水湾小区房屋漏雨，物业不处理	2019/2/2	县领导好	0	0	0
6	248547	A0009806	A7县楚龙街道三一街区小区13栋103违法住改商，开设快递站	2019/8/9	A7县楚龙	0	0	0
7	252244	A0001069	A7县楚龙街道盛地春天里商铺业主维权难	2019/10/	楚龙街道	0	0	0
8	260479	A0006175	A7县楚龙街道龙塘安置区内有多个不正规“洗脚按摩场所”	2019/6/2	龙塘小区	1	0	0
9	263602	A0001067	A7县楚龙街道社区村干部高息借贷不还，百姓讨要无门	2019/8/1	欠债还钱	0	0	0
10	263841	A0001084	A7县楚龙街道山水湾小区电梯存在安全隐患	2019/9/1	针对山水	0	4	0
11	205574	A0001084	A7县楚龙街道龙塘路（断头路）何时拉通	2019/8/2	请问A7县	0	3	1
12	234210	A0008789	有没有人管管A7县楚龙街道红树湾小区	2019/4/8	楚龙街道	0	0	1
13	243062	A0001028	A7县楚龙街道山水湾童话里幼儿园何时进行改成公立幼儿园	2019/1/7	尊敬的书	0	5	1
14	222668	A0007687	询问A7县楚龙街道拆迁的问题	2019/2/2	张县长，	1	1	2
15	224045	A0006300	A7县楚龙街道南塘路交通状况整治问题	2019/1/8	位于楚龙	0	1	2
16	248372	A0003034	反映A7县楚龙街道高沙村土地确权的问题	2019/5/2	看到很多	0	1	2
17	256983	A0001130	反映A7县楚龙街道楚核星城居民房屋安全问题	2019/12/	位于楚龙	0	0	2
18	275019	A0007687	询问A7县楚龙街道拆迁的问题	2019/2/2	领导，您	0	0	2

5.3 热点问题发现

网络问政平台一段时间内受到用户广泛关注并引起大量讨论的事件往往会形成热点问题，对于得到的聚类结果，只是一些聚集在一起的关键词，用户只能从中了解到当前留言存在的问题，而不能直观的了解到哪些问题的影响范围更广、传播面更大、讨论的更多。因此为了给用户提供更直观的话题检测结果，还需要对聚类后的问题计算热度，从中选出具有代表性的热点问题。针对经过话题聚类之后我们得到的留言话题进行分析评估，根据其各项特征计算出该问题的热度值，进而根据话题热度值得到留言内容的热度。

5.3.1 热点问题的影响因素和评估模型

留言内容经过聚类之后，可以得到很多留言内容类簇，其中每一个类簇便是反映一个问题，因此可以把聚类结果看作是一个问题集。留言主题数量在一定程度上可以体现问题的大小，但是问题的大小并不能完全描述问题的热度。不同的问题中参与讨论的用户数量不同，用户的影响力不同，所以，本文从两个方面来讨论问题的热度，即留言内容关注度和其他用户的关注度。

（1）留言内容关注度 $T_i(topic)$：本文当前留言内容出现的频率来计算留言内容的关注度，留言内容出现的频率与其热度成正比，其公式：

$$T_i(topic)=\frac{N_i(topic)}{N_{total}} \tag{5-8}$$

其中，$N_i(topic)$ 表示当前内容 i 所包含的留言数量，N_{total} 表示当前数据中所有的留言数量。

（2）其他用户关注度$O_i(topic)$：本文用留言内容的赞成数和反对数来综合计算话题的用户关注度，留言内容赞成数和反对数与其热度成正比，其公式：

$$O_i(topic)=\frac{N_i(agreement)}{N_{agreements}}+\frac{N_i(objection)}{N_{objections}} \quad (5\text{-}9)$$

其中，$N_i(agreement)$表示该类留言内容用户赞成数；$N_{agreements}$表示当前所有留言的赞成总数；$N_i(objection)$表示该类留言内容的反对数；$N_{objections}$表示当前所有留言的反对总数。

（3）评估模型

因此，本文根据不同的影响因子对留言问题热度的不同贡献程度，提出了热度评估模型：

$$H_i=a\frac{N_i(topic)}{N_{total}}+\theta\frac{N_i(agreement)}{N_{agreements}}+\varphi\frac{N_i(objection)}{N_{objections}} \quad (5\text{-}10)$$

分别给予不同影响因子的权重，令$a=0.6$，$\theta=0.2$，$\varphi=0.2$。

5.3.2 实验结果分析

本文实验数据来源于附件 3 中的留言内容，实验采用 Python 编程语言，根据热度计算公式对收集的实验数据进行处理和计算，进而根据热度值得到留言内容中的热点问题，通过计算得到热度值最大的问题是“58 车贷案件”，并罗列了热度排名前五的留言内容，如表 8 所示。且按照规定的格式给出相应热点问题对应的留言信息，详见附件“热点问题留言明细表.xls”。

表 8 热度排名前 5 的热点问题

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	0.0406	2019/1/11 至 2019/7/8	A 市 A4 区	58 车贷案件
2	2	0.0346	2019/5/5 至 2019/9/19	五矿万境 K9 县	房子存在严重安全隐患
3	3	0.0224	2019/5/21 至 2019/10/23	A 市 A9 市	高铁地铁建设问题
4	4	0.0111	2019/8/23 至 2019/9/6	A4 区绿地海外滩小区	小区距离高铁太近
5	5	0.0056	2019/3/26 至 2019/4/15	A6 区月亮岛路	110kv 高压电线架设存在问题

表 8 显示了热度排名五的留言内容和它们的详细信息，包括热度排名、问题 ID、时间范围、地点/人群、问题描述。

6 问题三分析及求解

为了更好地服务社会，相关部门愿意通过网络对留言内容进行答复。然而，相关部门在该平台的答复意见质量存在很大差异。因此从附件 4 中相关部门对留言的答复意见的相关性、留言内容的完整性和可解释性等对答复意见给出一套评价方案，并尝试实现。

6.1 相关部门答复意见质量优劣的特征

如何识别出最佳答复，衡量相关部门的专业水平和回复态度，是咨询服务平台发展需要解决的重要问题。根据专业咨询网站给出的相关在线回复标准和规范，我们对相关部门答复意见质量优劣的特征进行归纳，得到以下优劣评判特征：

（1）良好答复意见的参考特征：答复意见与对应留言问题相似度高；答复意见充实，有依据；答复意见具有可解释性；相关部门答复及时

（2）较差答复意见的参考特征：回复内容套用固定模板，例如“网友：您好！留言已收悉”“…详情咨询…”等；直接不予回复等。

6.2 相关部门答复意见质量的描述方法

答复意见的质量由很多种因素决定，必须从不同角度对其进行描述。答复意见的的优劣直接影响着相关部门回复的准确度。从 4 个方面对答复意见质量进行量化描述。

6.2.1 答复意见套用模板或直接不回复

相关部门的有些较差回复中会出现很多固定的词语或者直接不予回复，如“已获悉”，“网友：您好！留言已收悉”等。记答复意见为Q，$|Q|$为答复意见的文本长度。取正整数$K_0=13$作为阈值，当$|Q|\le K_0$时，则认为该条答复意见为坏数据，则直接记该条答复意见得分为 0。否则，再从答复意见的相关性、完整性和可解释性等 3 个方面对答复意见进行分析。

6.2.2 答复意见内容与留言的相关性

答复意见中的每一个评论都对应于语料库中的一个文档，通常用向量的形式来表达，相关部门对留言的答复意见中的每一条回复都对应留言详情中的一个问题，通常用向量的形式来表达，留言详情和答复意见会有相似的主题，因此可以计算文档之间的距离来计算其相似度。

（1）词语相似度

对留言详情和对应答复意见分别进行分词和 BERT 训练，得到该数据库的词向量和对应词语的相似度。则利用 BERT 可以很方便的计算出留言详情中的词语和答复意见中词语的相似度 $WORDSIM(w_i, w_j)$，

$$WORDSIM(w_i, w_j) = \frac{\sum_{t=1}^{n}\left(x_{ti} \times x_{tj}\right)}{\sqrt{\sum_{t=1}^{n} x_{ti}^{\ 2}} \times \sqrt{\sum_{t=1}^{n} x_{tj}^{\ 2}}} \tag{6-1}$$

其中，$w_i = (x_{1i}, x_{2i}, ..., x_{ti}, ..., x_{ni})$ 表示某个留言详情的第 i 个词语词向量，对应答复意见的第 j 个词语词向量表示 $w_j = (x_{1j}, x_{2j}, ..., x_{jj}, ..., x_{nj})$，$n$ 表示用 BERT 模型训练词向量时设定的词向量维度。

（2）文本相似度

接下来，建立文本相似度评价函数 $TEXTSIM$，并将 $TEXTSIM(T_k, T_p)$ 作为留言问题 T_k 和答复内容 T_p 的一个相似度指标，该值越大，表示 T_k 和 T_p 越相似。$TEXTSIM(T_k, T_p)$ 一般都是基于词语间的相似度 $WORDSIM(w_i, w_j)$ 的。

留言详情和答复意见由许多关键词构成，综合这些关键词可以构造出反映该文本信息的向量。设 $T_k = \{w_{k1}, w_{k2}, w_{k3}, ..., w_{ki}, ..., w_{kk}\}$ 表示其中某一个留言详情由 k 个词语组成，w_{ki} 表示 T_k 中的第 i 个词语组成的。$T_p = \{w_{p1}, w_{p2}, w_{p3}, ..., w_{pj}, ..., w_{pp}\}$ 表示留言详情对应的答复意见由 p 个词语组成，w_{pj} 表示 T_p 中第 j 个词语组成的词集。则 T_k 和 T_p 的相似矩阵记为 $S_{kp} = (s_{ij})$，

$$s_{ij} = TEXTSIM(w_{ki}, w_{pj}) \tag{6-2}$$

其中 $i = 1, 2, ..., k$，$j = 1, 2, ..., p$

建立计算留言详情和对应的答复意见的相似度方法：步骤如下

1）找出相似矩阵 S_{kp} 的第一行的最大值 m_1，假设这个最大值在文本相似矩阵 S_{kp} 的第 i 列，则去掉矩阵中的 m_1 所在的行列（即去掉矩阵 S_{kp} 的第 1 行和第 i 列）后得到的余子阵 M_{1i}^{1}。

2）找出余子阵 M_{1i}^{1} 的第一行的最大值 m_2，并去掉该矩阵中 m_2 所在的行列，可以得到余子阵 M_{1j}^{2}。

采用步骤（2）的方法继续查找上一步得到的余子阵的最大值，直到得到余子阵为空矩阵为止。这时，查找次数为 $\min(p, k)$。考虑到文本集的大小差距悬殊，本文构造留言详情 T_k 和答复意见 T_p 的相似度，

$$T_1 = TEXTSIM(T_k, T_p) = \frac{m_1 + m_2 +, ..., + m_l}{l + \frac{\max(k, p) - l}{l}} \quad (6\text{-}3)$$

其中 $l = \min(p, k)$ 。

6.2.3 答复意见的完整性

相关部门对留言的答复意见中的详细程度与回复的文本长度有直接的关系。简短答复意见信息量一般不够，评分应该较低；同时，较长文本的答复意见评分不应该过高。因此，可以考虑使用对数函数来量化答复意见的长度与评分的关系，建立“答复意见是否完整”的评价项 P_2，

$$P_2 = \log_m L_i \quad (6\text{-}4)$$

其中，L_i 为针对第 i 个问题答复意见的文本长度，m 为常数。由于为了平衡模型，对 P_2 进行最大最小归一化或相应处理得到 T_2，保证实验数据统一取值在[0,1]之间。

6.2.4 答复意见的可解释性

统计一条答复意见的关键词和句子，所述关键词就是该条答复意见中具有重要含义的名词和动词；根据所述关键词是否在相应的句子上出现计算每两个句子之间的关联关系权重，即用两个句子共同出现的关键词数目除以这两个句子所有关键词的数目，这两个句子可以不相连。

根据多个相连句子之间的逻辑关系，定义了相连句子之间的三种逻辑结构：先总后分、先分后总、和一个连一个，来计算多个相连句子之间的关联关系权重，从而度量该文本的可理解性，建立答复意见的可解释性的评价项 T_3：其特征在于操作如下步骤：

（1）分别计算三种逻辑结构上的关联关系权重。此时，用最小的两个句子（可能不相连）的关联关系权重作为该结构上多个相连句子之间的关联关系权重；

（2）取有最大权值的逻辑结构作为最可能的逻辑结构，它的权重就是多个相连句子之间句子之间的关联关系权重；

（3）依据原有的句子排列，把从最小的两个相连句子到所有数目的相连句子之间的所有关联关系权重相加得到该文本的可理解性度量。

6.3 相关部门答复意见评价模型

相关部门答复意见的关键在于如何建立对回复质量的量化评分模型，以附件4中的数据利用计算机对相关部门进行智能评价。根据上述量化方法，给出相关部门答复意见评价模型，其流程如图11所示。

下面对3项量化指标进行整合，以计算相关部门回复信息的得分情况，建立律师回复信息的质量评价函数，即构造如下的回复质量的评价函数$F(K)$：

$$\begin{cases} F(K) = M_K \cdot \lambda^T \\ \lambda = (\lambda_1, \lambda_2, \lambda_3),\ M_K = (T_1, T_2, T_3) \end{cases} \tag{6-5}$$

其中，λ^T为λ向量的转置向量，向量λ和M_K为反映回复信息不用侧面的权重向量和得分向量。

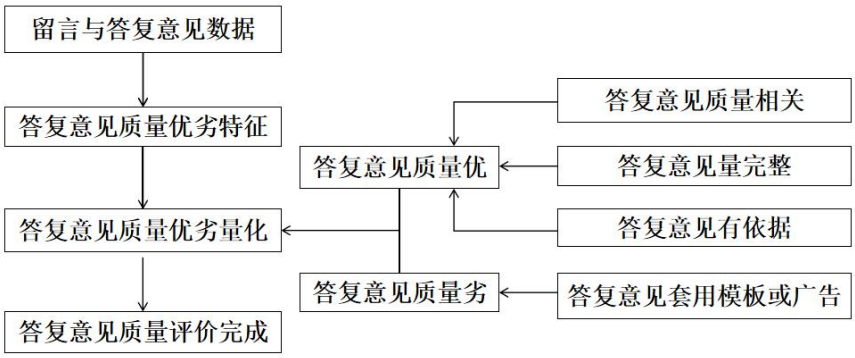

图11 相关部门答复意见评价模型

6.4 实验结果和分析

附件4数据中有2817条相关部门的答复意见。实验采用Python编程语言。

6.4.1 相关部门答复意见评价模型的计算步骤

为了给出附件4中每个答复意见的质量评分，根据答复意见质量评价函数式（6-5）得到模型的具体步骤如下。

（1）根据6.2节给出的标准计算该条答复意见的各项得分，表示向量为$M_K = (T_1, T_2, T_3)$。

（2）设置权重向量为$\lambda = (\lambda_1, \lambda_2, \lambda_3)$，计算$F(K) = M_k \lambda^T$作为第$K$条答复意见的质量评价函数。根据相关性、可解释性和完整性来判断答复意见质量的评分，其对应权重分别为0.6、0.1、0.3。如此将答复意见质量权重确定后，就可以得到答复意见的质量评价函数

$$F(K) = 0.6T_1 + 0.1T_2 + 0.3T_3 \tag{6-6}$$

6.4.2 答复意见评价模型的显示与分析

通过计算可以得到答复意见与留言内容的相似度T_1、答复意见的完整性T_2及答复意见的可解释性T_3。则2817条答复意见质量评价函数的结果$F(K)$如表所示，详见附件“答复意见质量评价函数值明细表.xls”。

表9 部分答复意见质量评价函数值

评分排名	留言编号	T_1	T_2	T_3	$F(K)$
1	6891	0.907453384	0.747901941	1.000000000	0.919262224
2	43468	0.899751158	0.681874825	1.000000000	0.908038177
3	93363	0.892216009	0.72357381	1.000000000	0.907686987
4	100846	0.881049219	0.744455405	1.000000000	0.903075072
5	81685	0.885628975	0.664078988	1.000000000	0.897785284
...	...	...	...	...	...
1861	34252	0.498367016	0.540768997	0.214285714	0.417382823
1862	42385	0.240926079	0.699545143	0.675000000	0.417010162
1863	6640	0.303549401	0.622292429	0.575000000	0.416858884
...	...	...	...	...	...
2817	159020	0	0	0	0

从表9中可以看出，可以根据评分函数清晰地判断出相关部门答复意见的质量。从附件“答复意见质量评价函数值明细表.xls”可以看出，排名第1和第2的答复意见对留言内容的回复有理有据，很好的解决了留言中的问题，满足之前提出的各项指标。而对于排名2770到2817的答复意见均采取没有意义的回答，完全没有解答留言详情中的问题。因此，评价方法对相关部门答复意见质量的评价符合规定的标准，验证了评价方法的准确性和合理性。

图12给出答复意见质量的分数分布，其中横坐标为回复质量函数值区间，纵坐标表示回复质量的评价函数值在各个区间的答复个数所占频率。从图12中可以看出，评分总体服从正态分布的形态，体现了本文答复意见评价方法的合理性。由图13可知A级答复意见是0.75-1分，B级答复意见是0.5-0.75分，C级是0.15-0.5分，D级答复意见是小于0.15分。可以看出A级和B级答复意见所占比例为55.4%，相关部门的整体答复意见还是比较好的，但仍然有9.3%的答复意见是不合格的。

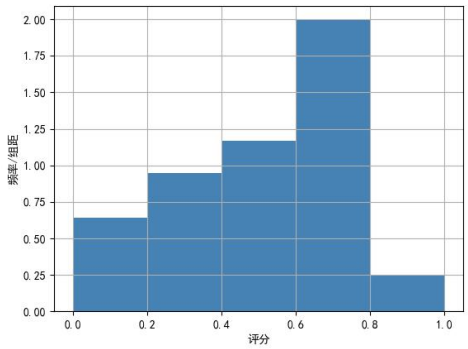

图 12 答复意见质量评价函数频率直方图

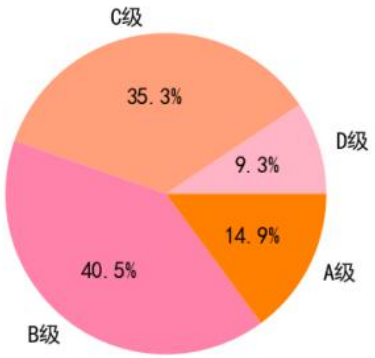

图 13 答复意见质量评分等级

7 结论

随着互联网的发展，大数据时代的来临，各类社情民意相关的文本数据量不断攀升。本文基于自然语言处理技术对数据进行处理使政府更有效的了解民意、汇聚民智、凝聚民气。

本文首先对数据进行预处理，并建立 BERT 模型来训练词向量。建立 LSTM 分类模型对留言数据进行分类，并与传统的机器学习进行对比。利用评价指标 F_1 对分类方法进行评价。通过仿真实验，证明了在 LSTM 分类器下，有比较好的分类结果。

其次，经过正则化地址匹配后，基于 TFIDF 算法提取出留言详情的关键词后，利用 K-means 聚类算法对留言数据多的集合进行聚类，并根据评价指标建立热度评估模型。计算出排名前五的热点问题描述分别为：58 车贷案件、房子存在严重安全隐患、高铁地铁建设、小区距离高铁太近、架设 110kv 高压电线。其对应的热点指数分别为 0.0406、0.0346、0.0224、0.0111、0.0056。

最后，为了对相关部门的答复意见进行评价，从答复的相关性、完整性、可解释性、是否套用固定模板等 4 个方面来建立答复意见质量评价模型。并通过仿真实验，来判断相关部门的答复意见的质量。

8 参考文献

[1] 邓楠，余本功. 基于情感词向量和 BLSTM 的评论文本情感倾向分析[J/OL]. 计算机应用研究，2018(12)：1-2.

[2] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional

transformers for language understanding [C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2017: 173-183.

[3] 祁亨年.支持向量机及其应用研究综述[J].计算机工程，2004，5(4)：6-9.

[4] Cheng X, Yan X,Lan Y, et al. BTM: Topic modeling over short texts[J]. IEEE Transactions on Knowledge and Data Engineering,2014,26(12):2928－2941．

[5] 何 跃，帅马恋，冯 韵.中文微博热点话题挖掘研究[J]. 统计与信息论坛，2014，29(6)：86-90.

[6] Chengxu Ye, Ping Yang, Shaopeng Liu. Topic Detection in Chinese Microblogs Using Hot Term Discovery and Adaptive Spectral Clustering[C]. Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 2014.

[7] Lai Jim Z C, Liaw Yi-Ching. Improvement of the k-means clus-tering filtering algorithm[J]. Pattern Recognition. 2008.

[8] 孔维泽，刘奕群，张敏，等．问答社区中回答质量的评价方法研究[J]．中文信息学报，2011，25（1）：3-9.

[9] 姚海波. 微博热点话题检测与趋势预测研究[D]. 华南理工大学

[10] 张思龙. 微博热点话题预判技术研究[D]. 解放军信息工程大学, 2013.

[11] 黎媛媛. 面向话题型微博的热点事件情感分析研究[D]. 安徽大学, 2016.

[12] 杨凯峰，张毅坤，李 燕. 基于文档频率的特征选择方法[J]. 计算机工程，2010，36(17)：33-35.

[13] 廉鑫．社区问答系统中若干关键问题研究[D]．天津：南开大学，2014．

[14] 骆祥峰，方宁，徐炜民等. 文本可理解性的度量方法：上海上大专利事务所，200910048310.X [P]. 2009-09-02.