基于文本挖掘的留言分类模型研究

摘要：在科技发达的21世纪，微信、微博、官方网站、市长信箱等网络平台已经逐步成为政府了解各类社情民意和解决民众基本问题的重要渠道。如何高效地实现政务投诉信息的自动分类，以帮助归口部门及时有效地处理与回复，从而提高各级政府部门的工作效率和服务水平，进而提升市民对政务服务的满意度成了目前迫切需要解决的问题。本文将基于市民问政留言信息，用机器学习实现了问政留言信息的自动分类，类别属性可精确到三级标签。同时，在精确分类标签的基础上抽取了留言的关键信息，实现了留言的快速部门归口工作，帮助归口部门提高工作效率。最后，实时统计留言热度指标，帮助政府部门快速聚焦民生热点问题，提升政务能力，并且还构造了留言政务处理的评价指标体系，用以监督政务处理工作，提升市民满意度。

针对问题一：本文首先将附件2中的非结构化数据进行数据预处理：去重、中文分词以及停用词过滤，然后基于TF-IDF算法提取留言特征值，构成词袋。由于数据还是非结构化数据，因此需要将其转为结构化数据供计算机识别，将文本数据向量化，再通过支持向量机、最近k邻近方法、朴素贝叶斯方法、随机森林分类、逻辑回归等五种分类算法对留言问题进行分类，训练出一级标签分类模型，还通过自定义字典和余弦相似度方法为每条留言上二级标签和三级标签。

针对问题二：本文对附件3中的非结构化进行一系列数据预处理，并将预处理后的数据集转为向量，运用问题一的模型和方法给留言上一级标签、二级标签和三级标签，然后把三级标签数目和热度指数作

为热度评价指标，利用 Reddit 的热排序算法和命名实体识别方法，最后得出排名前五的热点问题并给出了相应热点问题对应的留言信息。

针对问题三：本文针对相关部门的答复意见，从答复的相关性和时效性等角度分析了部门答复意见的质量。通过计算留言主题和详情与答复意见两者特征词的余弦相似度，反映答复意见与留言详情的相似性（即相关性）。再通过计算答复时间与留言时间之间的差值，利用分段函数划分时间段并得出满意度评分（即时效性）。最后根据分析出的结果，提出针对性建议。

关键词：中文分词；TF-IDF 算法；词向量；支持向量机分类；Reddit 算法；余弦相似度

Study on the message classification model based on text mining

Abstract: In the 21st century, WeChat, Weibo, official websites, mayor's mailbox and other online platforms have gradually become an important channel for the government to understand all kinds of social and public opinion and solve the basic problems of the people. How to effectively realize the automatic classification of government complaint information, to help the home control departments to deal with and respond in a timely and effective manner, so as to improve the efficiency and service level of government departments at all levels, and thus enhance the satisfaction of the public with government services has become an urgent problem to be solved. Based on the information of the public asking for political message, this paper will realize the automatic classification of the question message information by machine learning, and the category attribute can be accurate to the three-level label. At the same time, on the basis of accurate classification label extracted the key information of the message, realized the rapid department of the message, to help the home office to improve work efficiency. Finally, the real-time statistical message heat index, to help government departments quickly focus on people's livelihood hot issues, improve the ability of government affairs, and also constructed the message government processing evaluation index system, to supervise the handling of government affairs, improve public satisfaction.

In response to question one: This paper first pre-processes the unstructured data in Annex 2: de-weighing, Chinese word-sharing and de-checking, and then extracts the message feature value based on the TF-IDF algorithm to form a word bag. Because the data is still unstructured data, it is necessary to convert it into structured data for computer recognition, quantify the text data, and then classify message problems by five classification algorithms, such as support vector machine(SVM), recent k proximity method(KNN), simple Bayesian method, random forest classification, logic regression, etc., train the first-level label classification model, and also use custom dictionaries and cosine similarity methods to label two and three labels for each message.

In response to question two: This paper on the unstructured in Annex 3 for a series of data pre-processing, and pre-processed data set into vector, the use of the model and methods of question one to the message on the first-level label, secondary label and third-level label, and then the number of three-level labels and heat index as a heat evaluation index, using Reddit's hot sorting algorithm and named entity identification methods, finally to get the top five hot issues and give the corresponding hot issues corresponding to the message information.

In response to question three: This paper for the relevant departments of the response to the opinion, from the relevant and time-sensitive point of view of the response to analyze the quality of the department's reply. By calculating the

cosine similarity between the subject matter and details of the message and the characteristic words of the response to the comments, the similarity (i.e. relevance) of the response comments and the details of the message is reflected. By calculating the difference between the reply time and the message time, the time period is divided by segmentation function and the satisfaction score (i.e. timeliness) is obtained. Finally, according to the results of the analysis, the targeted recommendations are put forward.

Keywords: Chinese Word Segmentation; TF-IDF algorithm; Support vector machine classification;Reddit algorithm; cosine similarity

目录

1 挖掘背景与目标

1.1 挖掘背景

在科技发达的 21 世纪，计算机技术正在以飞快地速度和广度影响着人类社会生活的各个方面，信息化、全球化得浪潮把人类带入一个崭新的信息时代，使人们从过去以物质和能量为基础的活动平台转移到以信息和网络为基础的新平台。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。如何对群众留言信息做到自动分类并识别关键信息是当前的一个大难题。

1.2 挖掘目标

本次建模目标是利用题目提供的互联网上公开的群众问政留言信息以及相关部门对一些群众留言的答案意见文本数据，利用 TD-IDF、SVM 分类算法、Reddit 的热排序算法、余弦相似度以及分段函数，达到三个目标：

（一）利用中文分词和分类模型的方法对非结构化数据进行文本信息挖掘，根据训练出的模型的准确度，得出一级标签分类模型。

（二）运用问题一得出的分类模型，对附件 3 的留言问题预测一级标签，并查找相应的二、三级标签，把三级标签数目和热度指数作

为热度评价指标，得出排名前 5 的热点问题。

（三）根据相关部门对民众留言问题的答复意见数据，从相关性、时效性等多个角度分析答复意见的质量，余弦相似度衡量相关性，自定义分段函数及时间段阈值反映时效性，并给相关部门答复意见提供客观的、可实现的建议。

2 问题一分析

2.1 问题一分析方法与过程

2.1.1. 分析流程

以下是针对问题一的分析流程，主要经过以下几个步骤：

（1）对附件 2 进行去重；

（2）接着对文本进行中文分词和停用词过滤；

（3）利用 TF-IDF 算法提取留言主题和留言详情特征词；

（4）将提取出的特征词转化为向量，以便后续训练数据；

（5）划分训练集和测试集，训练并得出模型；

（6）最后通过结果分析，给出模型评价指标并对模型进行评价。

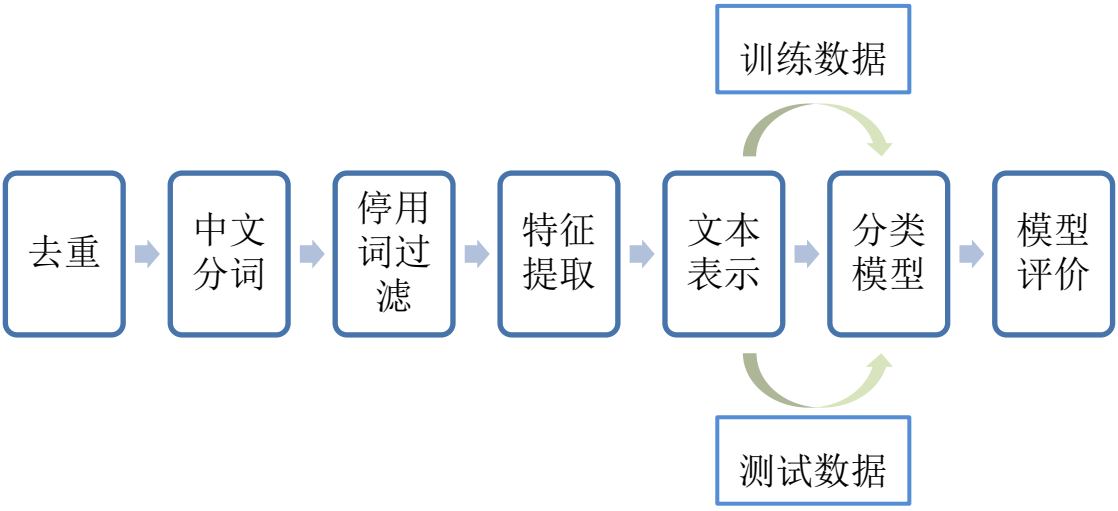

图 1 问题一流程图

2.1.2 数据预处理

（一）附件二数据表去重

对于附件 2.xlsx，每行代表了一条意见文本，但是难免会出现两个完全一样的文本，所以首先对文本数据进行了“去重”的预处理工作。去重后数据不变，说明没有完全一样的文本。

（二）对留言主题和留言详情进行中文分词

在对留言信息进行数据挖掘之前，需要先把非结构化的文本数据转换为能够让计算机识别的结构化信息。先利用正则表达式做辅助预处理。例如，英文句子 I am a teacher，用中文则为：“我是一名教师”。计算机可以很简单通过空格知道 teacher 是一个单词，但是不能很容易明白“教”、“师”两个字合起来才表示一个词。把中文的汉字序列切分成有意义的词，就是中文分词，有些人也称为切词[1]。对句子“我是一名教师”分词的结果是：我/是/一名/教师。附件 2.xlsx 以中文文

本的方式给出数据，为了便于转化，先要对留言主题和留言详情进行中文分词。目前 Python 主要用的几种分词工具有：jieba、pyltp、pynlpir、pkuseg 等。本文主要采用 python 的中文分词模块—jieba 和 pkuseg，对附件二中每一条意见的主题和详情进行中文分词。

jieba 分词模块提供分词、标注词性，支持 GBK 编码，新词发现，关键词提取等功能。部分分词结果示例如图 2。

西地省 校园 智慧 业务 用户 移动用户 阿里 公平竞争 排查 省内 访问 创业 定位 域名 省外
转让 药店 截图 举报人 推销 电话 广告 打电话 自称 此致敬礼 一个月 承诺 数十张 崇源堂 我店 中介 尾款 成邦 房管局…
学校 明阳 培训 说法 老师 县明阳 过市 十节课 要交 开发票 一节课 乱收费 含糊其辞 放学 学费 资料费 收取 乱收费 生…
工地 民工 明发 维权 受伤 态度恶劣 拒绝 支付 胡书记 父亲 百忙之中 感谢您 深究 塌方 医疗费
建造师 建望 省辉 东安 违法行为 证书 西地省 有限公司 执业资格 附图 聘用 工程 签字 及西地 挂证 刘利宏 举报人 工…
传销 数字 请市 区块 货币 假借 新贸链 链之名 传办 之实 下线 严厉打击 诈骗 省市 大力发展 女朋友 楚江 城映 江苑 …
医疗保险 查询 网上 望市 费能 干部职工 医保 公积金 职工 住房 公积金 执行力 住房 政策 出台 管理中心 贷款 买房 组…
强省 体育事业 体育 请市 有何 加快 省会市 体育场馆 严重不足 赛事 迅猛 国家级 强国 国家 市委 公共 体育馆 省运会 …

图 2 部分分词结果

（三）停用词过滤

在信息检索中，为了节省存储空间提高分析下文的正确率，在处理文本之前会自动过滤掉标点符号以及一些信息量很低甚至没有信息量的子和词，这些字和词即称为 Stop Words（停用词）。分词结果中有许多对表达句子意思没有什么实际意义的噪音词，这些噪音词应作为停用词去除。比如连词、叹词、介词、标点符号等。在留言文本中占了一定的比例，去除后对于降低文本维度有一定的帮助。利用 pynlpir 模块标注留言的所有词的词性，主要分为四大类：一是代词；二是副词；三是介词；四是标点符号。当然还有其他的一些词性的词也属于停用词的范畴。为了把留言文本中包含的停用词过滤干净，本文通过添加自定义停用词并结合《哈工大停用词表》进行辅助过滤。停用词过滤后的部分结果示例如图 3。

主题详情
['西湖', '建筑', '集团', '占', '道', '施工', '安全隐患', '大道', '西行', '便道', '未管', '路…
['在水一方', '大厦', '人为', '烂尾', '多年', '安全隐患', '严重', '位于', '书院', '路', '主干道…
['投诉', '市区', '苑', '物业', '违规', '收', '停车费', '区苑', '小区', '位于', '火炬', '路', …
['蔡锷', '南路', '区华庭', '楼顶', '水箱', '长年', '不洗', '区区', '华庭', '小区', '高层', '…
['区区', '华庭', '自来水', '好大', '一股', '霉味', '小区', '高层', '二次', '供水', '楼顶', '…
['投诉', '盛世', '耀凯', '小区', '物业', '无故', '停水', '年', '购买', '栋', '楼楼', '两层', …
['咨询', '楼盘', '集中', '供暖', '一事', '西地省', '地区', '常年', '阴冷', '潮湿', '气候', '…
['桐梓', '坡', '西路', '可可', '小城', '长期', '停水', '得不到', '解决', '胡书记', '家住', '…
['反映', '收取', '城市', '垃圾处理', '费', '不', '平等', '问题', '梅家田', '社区', '辖区', '…
['魏家坡', '小区', '脏乱差', '市政府', '您们好', '市区', '魏家坡巷', '业主', '多年', '以来', '…

图 3 部分停用词过滤后的结果

2.1.3 文本分类

（一）特征提取

经过上述文本数据预处理后，虽然已经去掉部分停用词，但是还存在大量词语，给后续文本向量化带来困难，所以提取特征值的主要目的就是在不改变原始文本核心信息的情况下，尽量减少需要处理的词数，由此来降低向量空间的维数，从而简化计算，提高分类的效率和精度。目前常见的特征选择方法有以下五种：互信息（MI）、卡方统计量（CHI）、信息增益（IG）、文本频数(DF)、TF-IDF 算法等下面对这些典型的特征方法做简单的介绍。

1、互信息（MI）

MI 是信息论中的概念,用于度量一个消息中两个信号之间的相互依赖程度[2]。在统计语言模型中，它用来表示文本特征 f 与类别 c 两个变量的相关性，文本特征 f 与类别 c 的互信息MI(f, c)的定义如下：

$$\mathrm{MI}(\mathrm{f},\mathrm{c}) = \log\frac{X\times N}{(X+Z)\times(X+Y)} \tag{1}$$

其中：X 为 f 和 c 同时出现的次数；Y 为 w 出现而 c 未出现的次数；Z 为 c 出现而 f 未出现的次数，N 为所有文档数。如果 f 和 c 相互

之间相互独立，则MI(f, c)为零。互信息的缺点是没有考虑词语发生的频度，导致特征分值受临界特征的概率影响较大。

2、卡方统计量（CHI 统计）

CHI 统计方法计算的是特征 f 与类别 c 之间的依赖关系，假设 f 和 c 之间符合具有一阶自由度的χ^2分布。对于类别 c，文本特征 t 的χ^2统计量定义如下：

$$\chi^2(f,c)=\frac{N\times(XR-ZY)^2}{(X+Z)\times(Y+R)\times(X+Y)\times(Z+R)} \quad (2)$$

其中：X 为 f 和 c 同时出现的次数；Y 为 w 出现而 c 未出现的次数；Z 为 c 出现而 f 未出现的次数，R 为 f 和 c 都没有出现的次数，N 为所有文档数。与 MI 类似，如果 f 和 c 之间相互独立，则$\chi^2(f,c)$的值为零。χ^2统计值越高，相关性越大，同时携带的类别信息也就越多。

3、信息增益（IG）

信息增益(Information Gain) 在机器学习领域被广泛使用[3]. IG 通过统计文档类别 c 中出现和不出现文本特征 f 的文档频数来衡量 f 对于 c 的信息增益。f 对 c 的信息增益定义如下：

$$\mathrm{IG(f)}=-\sum_{i=1}^{m}P(c_i)\,logP(c_i)+P(f)\sum_{i=1}^{m}P(c_i\,|f)\,logP(c_i\,|f)+P(\overline{f})\sum_{i=1}^{m}P\left(c_i\,|\overline{f}\right)logP\left(c_i\,|\overline{f}\right) \quad (3)$$

其中：$P(c_i)$为 c_i 类文档在语料中出现的概率，$P(f)$为语料中包含词条 f 的文档的概率，$P(c_i\,|f)$为文档包含词条 f 时属于c_i类的条件概率，$P(\overline{f})$为语料中不包含词条 f 的文档的概率，$P\left(c_i\,|\overline{f}\right)$为文档不包含词条 f 时属于$c_i$的条件概率，m 表示文档类别数。IG(f)的值越大则被选取的可能性就越大。

4、文本频数(Document Frequency)

文本频数是最简单的评估函数,其特征分值为训练集合中此特征出现的文本数[4]。一般常把它作为评判其他评估函数的基准[5]。

5、TF-IDF 算法

TF-IDF 向量表达是一种基于词统计的向量表达方法[6]。TF-IDF 算法的主要思想：如果某个词在一篇文章中出现的频率 TF 高，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。

TF 为词频（Term Frequency），即 TF 权重，表示某个词在文本中出现的频率。同时考虑到不同的文档有长有短，为了方便不同文档的比较 TF 值通常会被归一化，对某一特定文件里的词语t_i：

$$\mathrm{c} = \frac{\mathrm{n}_{i,j}}{\sum_k^j n_{k,j}} \tag{4}$$

上式中：$\mathrm{n}_{i,j}$ 表示该词在文件d_j中出现的次数，分母部分表示文件d_j中的总词数。

IDF 为逆向文件频率（Inverse Document Frequency），即 IDF 权重，指某个词普遍性的度量，此时需要建立一个语料库（corpus）。若计算出来的 IDF 越大，则说明该词具备很好的类别区分能力。

$$\mathrm{idf}_i = \log \frac{|D|}{|\{j:t_i \epsilon \mathrm{d}_j\}|} \tag{5}$$

其中，$|D|$语料库中的文本总数，$|\{j: t_i \epsilon \mathrm{d}_j\}|$表示包含词语$\mathrm{t}_i$的文本数目+1。若该词语不在语料库中，就会导致分母为零，所以分母一般情况下加一。

TF-IDF 就是 TF 和 IDF 值的一个乘积，即：

$$\text{tfidf}_{i,j} = tf_{i,j} \times \text{idf}_i \tag{6}$$

显而易见，TF-IDF 值与一个词在文本中出现的次数成正比，某个字或词的重要性越高，TF-IDF 值越大。本文主要利用 TF-IDF 算法来提取文本信息的主要关键词。

词云图（简称“词云”），也叫文字云，是对文本中出现频率较高的“关键词”予以视觉化的展现[7]。

图 4 是对文本提取关键词，并对关键词进行词频统计，以词云图的形式呈现，对关键词予以视觉上的突出。

图 4 词云图

（二）向量空间模型

要想计算机可以理解文本，需要将文本转化为机器可以识别的结构化数据—文本向量化。向量空间模型是处理文本数据非常有用的概念和模型，并在信息检索与文档排序中广泛使用[8]。向量空间模型（Vector Space Model，VSM）于 20 世纪 70 年代被提出并广泛用于文本检索系统，其主要思想为:将文本集中的每篇文本使用空间向量的形式进行向量表示，文本的每个特征词对应文本向量的每一维，特征

词所对应的值为该特征在整个文档集中所占的权重值[9]。（可以用 TF-IDF 等算法得到）。向量空间模型就是将文本向量化成为一个特征向量：

$V=\{（t_1，w_1），（t_2，w_2），（t_3，w_3），……\}$

其中，t_i(i=1，2，3…，n)为文档 D 中的特征项，w_i(i=1，2，3…，n)为特征项的权值，可由 TF-IDF 算法得出。

然后针对特征向量和已经标记的文本信息采用机器学习的方法训练模型，再用训练好的模型对未标记的文本进行预测进而得到分类结果。

2.1.4 分类算法

文本挖掘可以采用聚类、分类等算法。文本聚类是一种无监督的机器学习问题，文本分类是一种有监督的机器学习问题，一般分为训练和分类两个阶段。目前常用的文本分类算法有：距离测度函数分类算法，支持向量机（SVM）算法，神经网络方法，最大平均熵方法，最近 k 邻近方法（KNN 算法）和朴素贝叶斯方法（MultinomialNB）等[10]。

（一）KNN 算法

邻近算法，或者说 K 最近邻(kNN，k-NearestNeighbor)分类算法是数据挖掘分类技术理论比较成熟方法，同时也是最简单的方法之一。KNN 算法的核心思想是：在待分类样本特征空间中的 k 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也

属于这个类别并具有这个类别上样本的特性。该方法在确定分类决策上只依赖最邻近的一个或者几个样本的类别来决定这个待分类样本的所属类别。另外在 KNN 算法中所选择的邻居必须都是已经正确归类的对象。算法步骤：

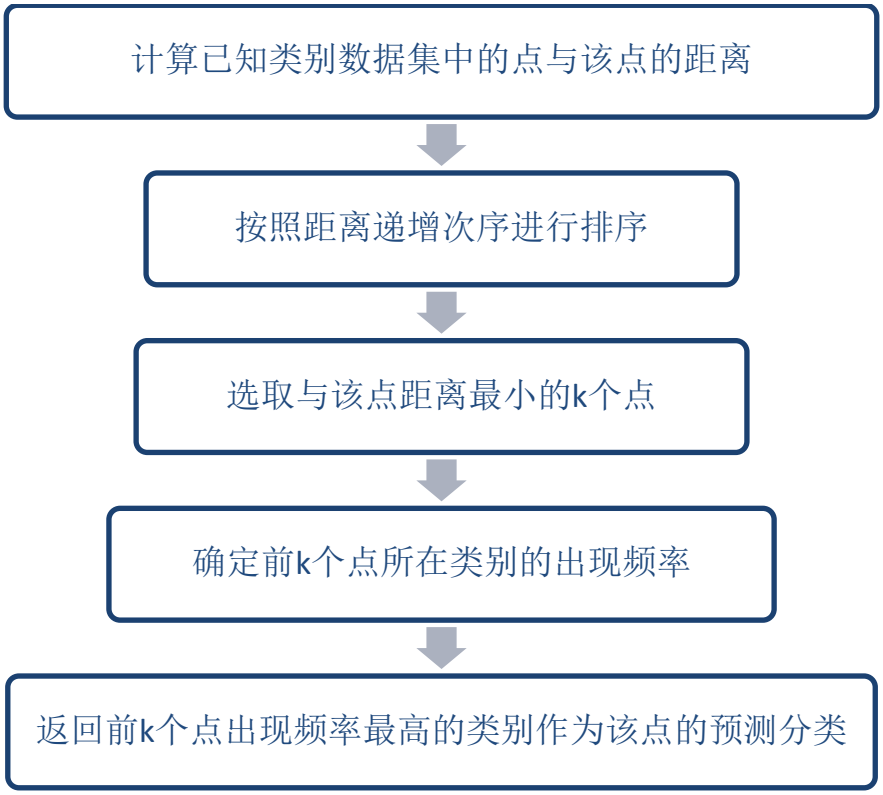

图 5 KNN 算法步骤

（二）NB 分类器

朴素贝叶斯分类是一种很简单的分类算法，朴素贝叶斯的基础思想是：对于给出的待分类项，若求解在此项出现的条件下各个类别出现的概率哪个最大，则就可以认为此待分类项就属于哪个类别。贝叶斯分类算法原理如下：

设文本向量 $X = \{X_1, X_2, \ldots, X_d\}$ 是 d 维向量，类标签 $Y \in \{1, 2, \ldots, c\}$，样本个数为n； 给定文本向量 X，预测目标为寻找类别 k，使得 $p(Y = k|X) = \frac{p(X|Y=k)*p(Y=k)}{p(X)}$ 最大；

对于某一个样本 X，p(X)的取值不变，所以预测$\max_k p(Y = k| X)$，等价于预测$\max_k p(X|Y = k) \cdot p(Y = k)$的类别 k。

条件独立性假设:对已知类别，所有属性之间都相互独立。

条件独立性假设是为了简化 $p(X|Y = k)$的估计：

$$p(X|Y = k) = p(X_1, X_2, \dots, X_d|Y = k) = p(X_1|Y = k) \cdot p(X_2|Y = k) \dots p(X_d|Y = k) \quad (7)$$

基于条件独立性假设，贝叶斯算法也就是预测：

$$\operatorname{argmax}_k p(X_1|Y = k) \cdot p(X_2|Y = k) \dots p(X_d|Y = k) \cdot p(Y = k) \quad (8)$$

一般采用极大似然估计来估计上述概率分布，方法如下：

$$p(Y = k) = \frac{n_k}{n} \quad (9)$$

$$p(X_i = s|Y = k) = \frac{n_{k,i_s}}{n_k} \quad (10)$$

其中 n_k表示标记是 k 的样本个数，n 为总的样本个数；n_{k,i_s}表示标记为k且文本向量的第 i 维的值为 s 的样本个数。

朴素贝叶斯分类器的不足：

(1) 朴素贝叶斯分类器使用的前提是，各个文本特征项之间必须是相互独立的。

(2)需要测试的文本必须是完整的或者特征项分布均衡。

（三）支持向量机（SVM）

支持向量机(support vector machine)是一种分类算法，也可以做回归，根据不同的数据做不同的模型。SVM 方法适用于大样本集的分类，特别对于文本分类，它结合降维和分类在一起。SVM 算法基于结构风险最小化原理，将原始数据集合压缩到支持向量集合， 然后用

子集学习得到新知识，同时也给出由这些支持向量决定的规则，并且可得到学习错误的概率上界[11]。假设线性分类面的形式为：

$$g(D) = \omega \cdot D + b = 0 \tag{11}$$

其中：ω为分类面的权系数向量，b 为分类阈值。将判别函数进行归一化处理，使得所有样本都满足$|g(D)| = 1$，即$y_i[(\omega \cdot D_i) + b] - 1 \geq 0$，i = l，2，…，N，$y_i$是样本的类别标记，即当样本属于类 C 时$y_i = 1$，否则$y_i = -1$；$D_i$为相应的样本。样本的分类间隔就等于$\frac{2}{||\omega||}$，这样设计的目的就是要使得这个分类间隔的间隔值最小。由此定义 Lagrange 函数：

$$L(\omega, b, a) = \frac{1}{2}(\omega \cdot \omega) - \sum_{i=1}^{n} \alpha_i \{y_i[\omega \cdot D_i + b] - 1\} \tag{12}$$

其中：α_i为 Lagrange 乘数（$\alpha_i > 0$），对ω和 b 求偏微分并令其为 0，则原问题转换为如下对偶问题：在约束条件$\sum_{i=1}^{n} y_i \alpha_i = 0$，$\alpha_i > 0$，$i = 1, 2, \ldots, n$下对$\alpha_i$求解出下列函数的最大值：

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (D_i, D_j) \tag{13}$$

若α_i^*为最优解，再由公式

$$\omega^* = \sum_{i=1}^{n} \alpha_i^* \times y_i D \tag{14}$$

得出最优分类面的权系数向量。计算如下最优分类函数，来判断某一样本是否属于类 C：

$$f(D) = sign\{(\omega^* \cdot D) + b^*\} = sign\{\sum_{i=1}^{n} \alpha_i * y_i (D_i \cdot D) + b^*\} \tag{15}$$

若$f(D) = 1$，那 D 就属于该类；否则不属于该类。对于线性不可分的情况，可以引入松弛因子，在求最优解的限制条件中加入对松弛因子的惩罚函数。

运用 sklearn 中 SVM 的 LinearSVC 训练数据并预测结果。对于多分类问题，假设有 n 个类，我们可以把每个类作为二分类进行训练，以此来帮助实现该类与其他 n-1 个类的分离。预测过程其实就是在计算每个分类器的得分，然后将最大得分的分类器选作为类标签的过程。先划分训练集和测试集，运用 sklearn 库的 model_selection 模块中提供的一函数将矩阵随机划分为训练子集和测试子集，并返回划分好的训练集测试集样本和训练集测试集标签。这里的划分比例为 8：2。

除了上述介绍的三种分类算法，本文还用到了随机森林分类和逻辑回归方法训练模型，并得出五种分类模型的准确率，最终选用准确率最高的分类模型。

2.1.5 模型评价指标

特征提取、文本表示作为分类的前处理过程,其有效性可以通过分类的效果来测试。为评价分类效果,我们采用几方面的性能评价方法: 查准率（Precision）、查全率（Recall）和 F1-score 评价。对于某一特定的类别,召回率表示该类样本被分类器正确识别的概率，即被正确分类的文本数和被测试文本总数的比率。准确率表示分类器做出的决策是正确的概率，即正确分类的文本数与被分类器识别为该类的文本数的比率。F1-score 就是 Precision 和 Recall 的加权调和平均：

$$\mathrm{F1} = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{16}$$

其中P_i为第 i 类的查准率，R_i第 i 类的查全率。

2.2 结果分析与评价

通过去重，对文本进行中文分词和停用词过滤，用 TF-IDF 算法提取特征词并文本转化后的数据集作为输入训练测试模型，本文采用留出法随机划分训练集和测试集，在模型训练过程中按照 80%和 20%的比例分配。运用常用的支持向量机（SVM- LinearSVC）算法、KNN 算法、NB 分类器、随机森林分类器、逻辑回归（LogisticRegression）等五种机器学习分类的方法训练模型预测结果。五种分类模型训练数据的查准率、查全率、F1-score 如表格 1：

表格 1 五种分类模型的查准率、查全率、F1-score

模型	precision	recall	F1-score
LinearSVC 分类	**0.91**	**0.91**	**0.91**
KNN 分类	0.86	0.85	0.85
NB 分类	0.72	0.71	0.71
随机森林分类	0.81	0.80	0.80
逻辑回归	0.89	0.88	0.88

显而易见， LinearSVC 分类器的查准率、查全率和 F1-score 数值最高，其查准率为 91%、查全率为 91%、F1-score 为 91%，说明该模型的效果相比其他几个模型更好，其混淆矩阵如图 6 所示：

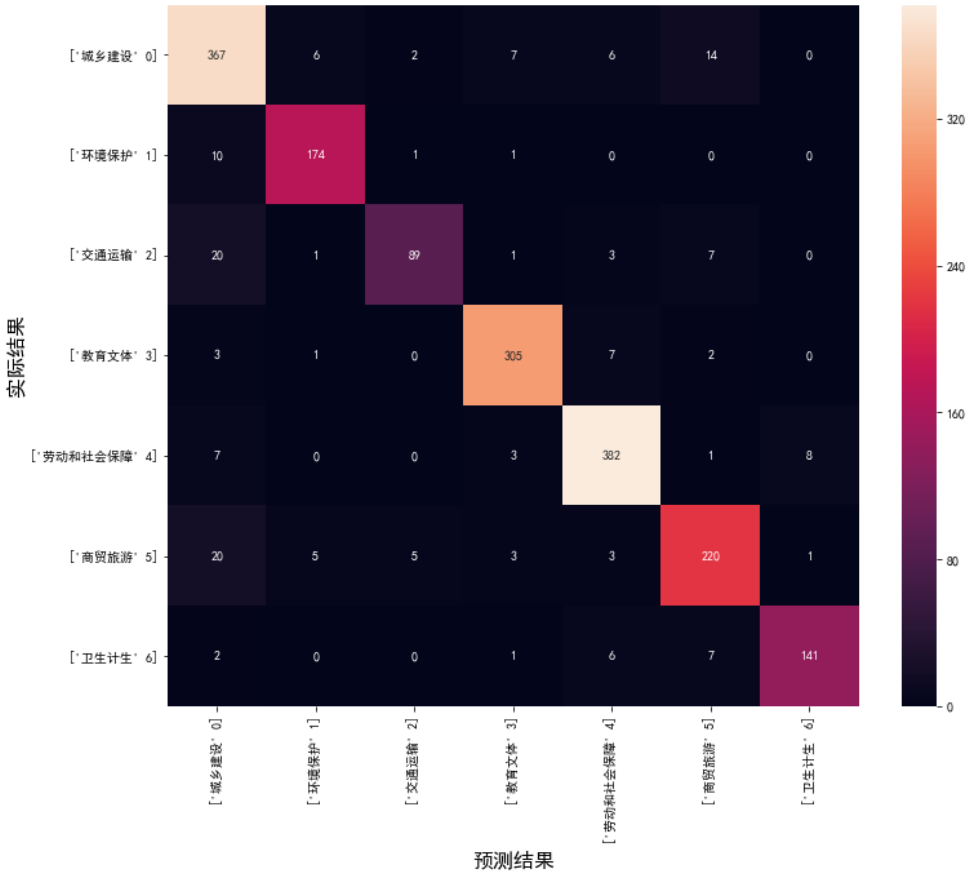

图 6 LinearSVC 分类模型混淆矩阵

因此本文选用 Linear SVC 模型进行分类预测，由此即建立留言内容的一级标签分类模型。

模型预测出每条留言的一级标签后还需要查找对应的二级、三级标签。这里利用自定义字典查找二、三级标签。首先自定义一个一级标签字典，一级标签做 key，一级标签对应的二级标签做 value；其次，再自定义一个二级标签字典，即二级标签做 key，二级标签对应的三级标签做 value。一级、二级签字典部分内容分别如图 8、图 9 所示。

```
'环境保护': ['建设项目审批', '环保管理', '其他', '环境污染'],
'交通运输': ['出租车管理', '邮政管理', '其他', '建设管理', '客货运输', '交通运输'],
'商贸旅游': ['旅游管理', '商业贸易', '其他', '市场监管', '质监检验检疫'],
'卫生计生': ['医患纠纷', '公共卫生', '其他', '医政监管', '食品药品监管', '人口计生'],
```

图 7 一级标签字典部分内容

```
'城镇居民社会保险': ['其他', '新农村合作医疗', '城镇居民医疗保险', '城乡居民养老保险'],
'城镇职工社会保险': ['职工生育保险', '职工养老保险', '职工失业保险', '职工医疗保险', '其他', '职工工伤保险'],
'工资福利': ['福利待遇', '工资调整', '其他', '最低工资标准', '住房公积金', '工资发放'],
'就业培训': ['其他', '就业和再就业', '职业技能鉴定', '职业培训'],
'劳动保护': ['劳动环境', '女工保护', '劳动安全', '安全防护', '其他', '工作时间和休息休假'],
'劳动关系': ['劳动合同纠纷', '劳务派遣纠纷', '其他', '非法用工', '农民工权益', '协议解除劳动关系', '辞职辞退'],
'社保基金管理': ['其他', '社保基金管理'],
'退休政策及待遇': ['病退及提前退休人员待遇', '其他', '退休政策', '内部退养人员待遇', '退休人员待遇', '退休金发放']}
```

图 8 二级标签字典部分内容

分别建立一、二级字典后，查找留言内容的二、三级标签流程如图 9 所示：

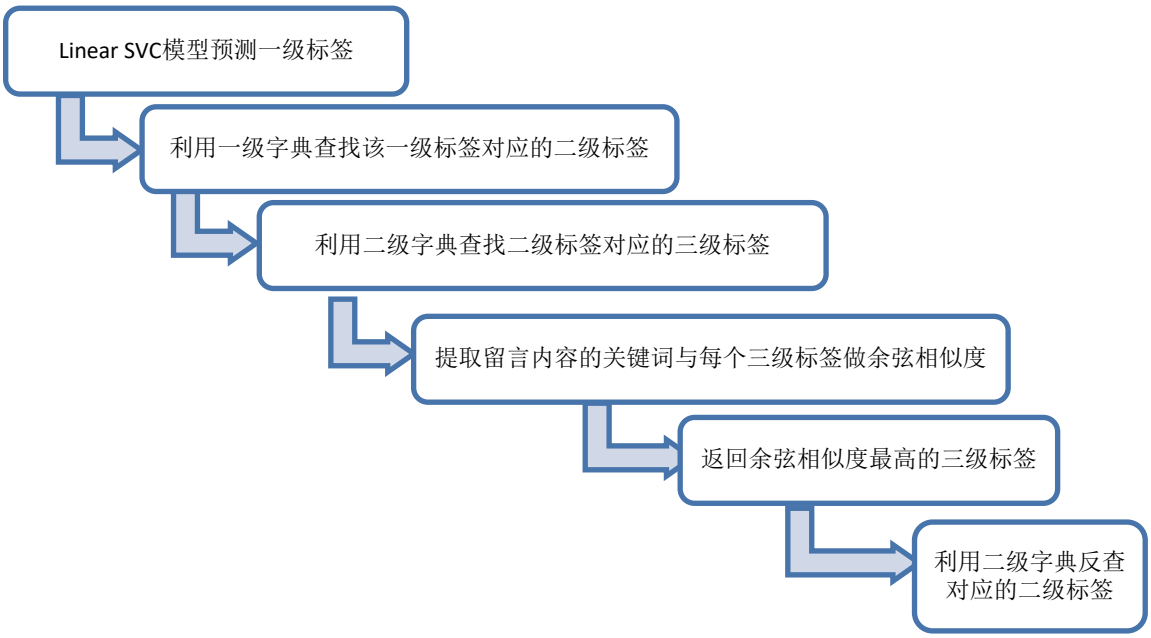

图 9 查找二、三级标签流程图

查找已经预测出留言的一级标签所对应的二级标签和三级标签，其目的就是自动将群众留言分派到相应的职能部门进行更有效、更快捷的处理。

小结

针对问题一用 TF-IDF 算法提取特征词并转为向量，构建了五个分类模型训练数据集，并比较了五种模型的 precision、recall 以及

F1-score 三个指标，最终选用三个指标最高的 LinearSVC 分类模型做为一级标签分类模型。并查找对应的二、三级标签。

3 问题二分析

3.1 问题二分析方法与过程

3.1.1 分析流程

以下是针对问题二的分析流程，主要经过以下几个步骤：

（1）对附件 3 进行去重、中文分词和停用词过滤等数据预处理工作；

（2）接着对附件 3 里的每条文档上一、二、三级标签并统计排名前 10 的三级标签；

（3）运用 Reddit 算法的计算公式计算这 10 个三级标签的热度指数并排序得出排名前 5 的热点话题；

（4）最后对这 5 个热点话题进行命名实体识别，做出问题描述。

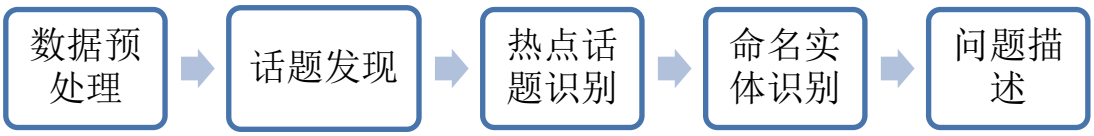

图 10 问题二流程图

3.1.2 数据预处理

文本预处理主要实现两个功能：

一、对附件 3.xlsx 的文本内容去重；

二、对附件 3.xlsx 的留言主题和留言详情进行中文分词、停用词过滤处理。

3.1.3 话题发现

把上述预处理后的数据集，利用问题一训练出来的 SVM 分类模型，预测出每条留言的一级标签并查找相应的二级、三级标签。在排序时，以每个三级标签文档数目作为初步热度问题高低的判别标准。对三级标签文档数排名前 10 名的话题再进一步计算热度指数，从而得到最终的热度话题。

3.1.4 热点话题识别

热点话题相对于其他一般性话题而言，热点话题具有关注度高、点击率高、时间长、传播快等特点，所以本文结合这些特点和留言内容，利用 Reddit 算法的计算公式进行热点排序，最后找出热点话题。

Reddit 热排序算法的公式计算公式如下：

$$\text{Score} = log_{10}z + \frac{yt}{45000} \tag{17}$$

首先计算 x，x 为点赞数与反对数的差，U表示点赞数，D 表示反对数，即$\text{x} = \text{U} - \text{D}$，从而判断 z 和 y。

$$\text{z} = \begin{cases} |x|， & |x| \geq 1 \\ 1， & |x| < 1 \end{cases} \tag{18}$$

z 表示话题受肯定的程度，x 越大，话题越受肯定。

$$y=\begin{cases}1，\ x>0\\0，\ x=0\\-1，\ x<0\end{cases} \tag{19}$$

y 是一个符号变量，表示投票方向，是文章评价的一种定性表达，x 大于 0 表示正面评价，x 等于 0 表示没有倾向，x 小于 0 表示负面评价。

$$t=A-B \tag{20}$$

t 为时间差，即话题最晚发布时间 A 与最早发布时间B的差值，单位为秒。$\frac{yt}{45000}$中分母表示 45000 秒。

本文把文档对应的三级标签数目和热度指数作为热度评价指标，结合两个指标评出排名前五的热点话题。

3.1.5 问题描述

话题与事件的概念是不同的，话题是一个核心事件或活动以及与之直接有关的事件或活动，一个事件必定属于一个话题[12]。基于留言内容，利用 pyltp 进行命名实体识别，提取每条留言里提到的特定地点或者特定人群的问题。

3.2 结果分析与评价

通过给数据集分类并查找对应的三级标签后，依据三级标签文档数目进行排序，得到排名在前 10 名的三级标签，如图 11 所示：

Index	三级标签
小城镇建设	569
小区管理	483
商标管理	423
居民服务设施	260
噪音污染	155
体育事业	137
城镇居民医疗保险	135
安全隐患	135
出租车管理	126
生态示范和模范城区创建	108

图 11 三级标签统计值排序

然后用代码实现 Reddit 算法计算这 10 个三级标签的热度指数，并对所有热度指数进行排序，得到排名前五的热点问题以及其对应的热度指数如图 12 所示，利用 pyltp 进行命名实体识别提取地点做出问题描述如图 13 所示，具体请见附件“热点问题表.xls”。

```
('城镇居民医疗保险', 800.8923483480625)
('小区管理', 747.9151055455707)
('噪音污染', 715.453174297429)
('安全隐患', 713.7663537917913)
('小城镇建设', 713.6628176179205)
```

图 12 排名前五的热点问题及对应的热度指数

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	800.8923483	2018/11/15至2020/01/06	A市	A市城镇居民医疗保险
2	2	747.9151055	2019/01/02至2020/01/26	A市	A市小区管理
3	3	715.4531743	2019/01/02至2020/01/09	A市	A市噪音污染
4	4	713.7663538	2019/01/01至2020/01/07	A市	A市安全隐患
5	5	713.6628176	2019/01/01至2020/01/07	A市	A市小城镇建设

图 13 问题描述

可以看出，排名前 5 的热点问题分别为：“城镇居民医疗保险”、“小区管理”、“噪音污染”、“安全隐患”以及“小城镇建设”。说明这五类问题对某地点或者某类人群的影响范围非常大，引起了众多民

众的关注。对热度指数排名前五的热点问题，还提取对应的留言信息（留言编号、留言用户、留言主题、留言时间、留言详情等信息），部分内容如图 14 所示，具体详情请见附件“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	253198	A00035110	对A市居民在网上办理	2019/5/26 18:33:05	A市现在办事越来越方便	0	0
1	266827	A00054222	星沙商业乐园在A7县	2019/3/16 22:09:14	尊敬的胡书记：您好！请	1	0
1	212658	A00046298	A8县外出打工的农民	2019/3/12 17:48:39	我叫唐雪中，身份证号码	0	0
1	265775	A00074401	A8县市流沙河镇大田	2019/5/10 17:06:40	尊敬的胡书记：您好！关	0	0
1	265242	A00054222	请依法处理好A7县星	2019/3/29 14:11:47	尊敬的县委书记沈谋裕先	0	0
1	198973	A00025783	A3区鑫庆医药公司拖	2019/5/28 16:20:42	西地省鑫庆医药有限公司	0	0
1	264915	A00073801	希望A市能给底层创业	2019/6/3 5:33:13	尊敬的市委政府领导你们	0	0
1	264904	A000109822	想请西地省计算机软件	2019/3/9 9:34:19	本人于2002年在校期间考	0	0
1	264735	A00075088	要求A市公安局A3区分	2019/2/18 14:51:59	申请事项：请求行使监督	0	0
1	267577	A00042288	希望西地省把抗癌药品	2019/09/08 21:01:59	让癌症病人看到光明吧，	0	0
1	212128	A000110735	在A市买公寓能享受人	2019/4/24 15:07:30	尊敬的书记：您好！我研	0	0
1	199374	A00099577	要求依法追究A3区林	2020/1/6 15:35:17	请求：一、依法追究林某	0	0
1	237798	A000107775	2020西地省城乡居民医	2019/11/20 15:11:32	2020西地省城乡居民医疗	0	0
1	262531	A00059064	西地省普通话水平测试	2019/4/8 11:42:58	本人19年3月与4月份在西	1	0
1	262248	A00063640	咨询原A7县居民的户口	2019/5/30 23:51:13	我户籍所在地为A7县干杯	0	0
1	239711	A00064002	反映A市鑫华驾校的一	2019/11/27 17:19:28	举报A市鑫华驾驶员培训	0	0
1	260596	A00036572	能不能修一下A4区政府	2019/11/2 17:08:27	政府机关大院里的篮球场	0	0
1	211189	A00013002	西地省农华园林科技发	2019/10/5 18:23:58	西地省农华园林科技发展	0	0
1	211176	A00048220	举报西地省建达鸿实业	2019/5/8 18:21:33	2007年7月，湖省农机总	0	0

图 14 热点问题留言明细表

针对利用三级标签数目和热度指数作为热度评价指标，以及运用 pyltp 工具包命名实体识别提取地点或人群都是比较客观和现实的选择。群众集中反映的问题就是热点问题，及时发现群众反映的问题，更有助于政府相关部门有针对性地处理群众反映的问题，提高服务效率和建设服务型政府。

4 问题三分析

4.1 问题三分析方法与过程

4.1.1 分析流程

以下是针对问题三的分析流程，主要经过以下几个步骤：

（1）构建答复意见评价因子多方面评价答复意见的质量；

（2）利用余弦相似度反映答复意见与留言问题的相关性；

（3）通过自定义分段函数及阈值计算时效性满意度；

（4）最后本文给出评价以及一些建议作为参考。

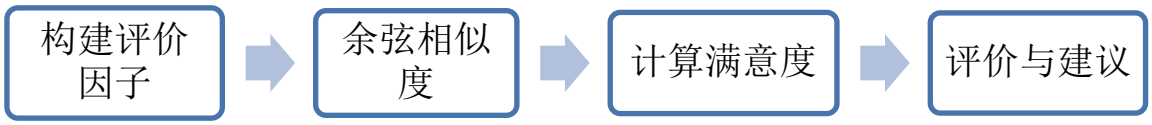

图 15 问题三流程图

4.1.2 构建答复意见评价因子

政府网站、微博、微信、市长信箱、手机客户端等网络问政平台逐渐成为各级政府和部门开展网上政民互动、了解民意的重要途径，相应的政府部门也为民众的留言给出相应的、有质量的答复意见。评价答复意见的质量，本文主要从答复的相关性、时效性等角度进行评价。

（一）相关性

分别提取 25 个留言主题和留言详情与答复意见的特征词，通过计算不同向量之间的相似度来量化文本之间的相似度，常用的计算两个向量相似度的方法是余弦相似度。

余弦相似度是通过计算两个向量的夹角余弦值来评估他们之间的相似度。对于二维空间，由向量点积公式：设向量 $W=(W_1,W_2,\ldots,W_n)$，向量 $T=(T_1,T_2,\ldots,T_n)$，那么它们之间的相似度为：

$$S(W,T)=\cos\theta=\frac{\sum_1^n(W_i+T_i)}{\sqrt{\sum_1^n W_i^2}\times\sqrt{\sum_1^n T_i^2}} \quad (21)$$

余弦值的范围在[-1，1]之间，cosθ越接近 1，代表两个向量的方

向越接近，相似度越高；cosθ越接近-1，表示两个向量的方向越相反，近乎正交。

（二）时效性指标

答复的时效性也是评价答复意见质量的一个重要因子，时效性是指答复信息的新旧程度、行情最新动态和进展，答复时间段越短，满意度得分越高。这里把相关部门答复的时间段分为小于 2 天、2~7 天、7~15 天、15~30 天、30~60 天、超过 60 天，并规定满意度 5 分满分制，建立分段函数：

$$\mathrm{f(s)}=\begin{cases} s=5，\ t\le 2 \\ s=4，\ 2<t\le 7 \\ s=3，\ 7<t\le 15 \\ s=2，\ 15<t\le 30 \\ s=1，\ 30<t\le 60 \\ s=0，\ t>60 \end{cases} \tag{22}$$

其中：t 表示答复的时间段，s 为满意度分值。例如：一留言用户的留言时间为 2019 年 4 月 25 日 9:32:09，相关部门答复时间为 2019 年 5 月 10 日 14:56:53，也就是说，相关部门在第 15 天对留言做出答复，通过代码实现得到答复的时间段 t 为 15，在$7<t\le 15$范围内，则该条留言答复的满意度 s 为 3。

除了从上述两个角度评价答复质量，本文还提出了一些相关建议，让政府和民众之间的交流和沟通更加的灵活高效。

4.2 结果分析

（一）相关性

通过计算每条留言详情与其对应的答复意见的余弦相似度，若相似度越高，则说明两者的相关性越高。部分文本相似度结果如图 16 所示：

主题详情关键词	答复意见关键词	sim
['物业公司', '小区', '投票', '业…	['业主大会', '业委会', '业主', '…	0.968386
['修好', '南路', '几下', '一段', …	['施工', '坪塘', '排水', '换填',…	0.963394
['幼儿园', '教师', '民营', '工作'…	['民办', '幼儿园', '待遇', '教师'…	0.966779
['公寓', '研究生', '新政', '购房'…	['购房', '房屋交易', '补贴', '管…	0.957582
['马坡岭', '小学', '公交站点', '…	['马坡岭', '来信', '小学', '公交…	0.958531
['泥巴', '泥泞不堪', '含浦镇', '…	['街道', '含浦', '学士', '含浦镇'…	0.966722
['电梯', '小区', '胡书记', '期待'…	['电梯', '建局', '住宅', '年月日'…	0.960931
['幼儿园', '社区', '东澜湾', '医…	['幼儿园', '黎托', '小区', '东澜…	0.965919
['业主', '美麓', '停工', '质量', …	['施工单位', '整改', '年月日', '…	0.960134
['洋湖', '绿化带', '路段', '壹号'…	['绿化带', '为洋湖', '建设', '壹…	0.966722
['养殖', '区大托', '大托', '大棚'…	['原大托村', '租金', '村民', '村…	0.963394
['人防', '鄱阳', '安置', '一万四…	['鄱阳', '人防', '平方米', '人防…	0.951457
['万国', '小区', '清水塘', '渔业'…	['收悉', '邀标', '方案设计', '年…	0.965058
['出借', '平台', '芒果', '贵省', …	['收悉', '警情', '银盆岭', '侦查…	0.956517
['增开', '公交车', '强烈建议', '…	['发车', '驾驶员', '配车', '线路'…	0.96173
['车道', '新开铺', '拆除', '通行'…	['新开铺', '披塘', '路口', '年月…	0.969244

图 16 部分文本相似度结果

结合所有数据的相似度，可以看出，两个文本间大部分的相似度都在 95%以上，也就是说，政府相关部门对留言给出的答复和留言内容本身有很强的相关性。

（二）时效性分析

通过遍历附件 4.xlsx 的每条留言的留言时间和答复时间，代码实现计算出两者相差的天数，并运用分段函数得出每条答复意见时效性满意度。结果如图 17、图 18 所示：

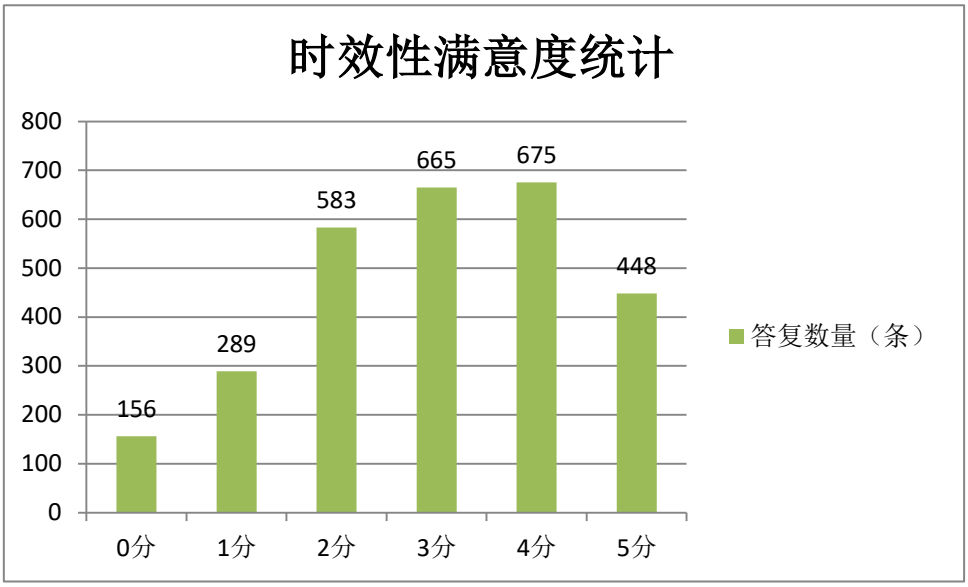

图 17 时效性满意度统计

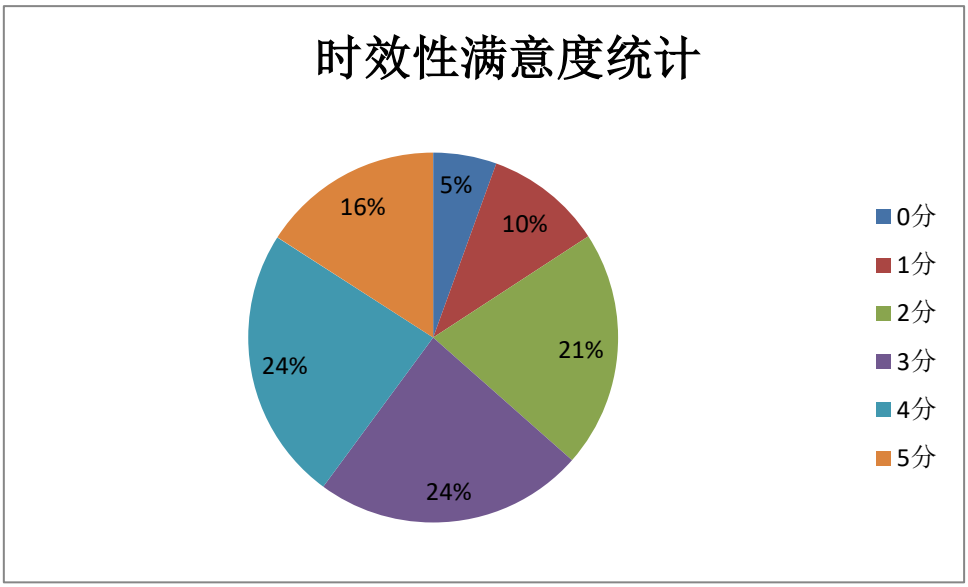

图 18 时效性满意度统计

由图表知，答复时间在 2~7 天、7~15 天的占比最大，在 2 天内答复的留言数量占比 16%。群众的满意度大多都是 3 分以上，说明相关部门对留言答复的时效性比较好。

4.3 评价与建议

（一）评价

由上述的分析，大部分的答复相似度都很高，说明相关部门对群

众留言的问题给予较为全面的答复。从留言答复的时效性满意度可以看出将近一半的留言发出后，30 天有相应的答复，意味着一些有关部门回复的时效性低，这样就不能充分发挥网上政府和群众互动中政府回应的时效性指标较强的现实意义。

（二） 建议

网上群众留言中通常涉及的内容可分为:查询咨询、举报投诉、意见建议、居民服务等，政府和民众互动的成效，直接影响到政府的形象，行政的有效性以及民众的满意度等。针对政府相关部门答复的质量提出几点建议：

1、在相关部门做出答复后，可以发起一个满意程度评分，让留言用户针对相关部门答复的内容，从内容相关性，答复时效性和答复完整性等方面，进行五星评分制。

2、答复时效性是政府和群众互动中各级政府和部门的响应速度、反应效率与服务水平的直接体现指标，建议在收到群众留言信息后的 15 个工作日内做出答复，进一步提高各级政府和部门的政府和群众互动交流，提升民众对服务型政府建设的满意程度。

3、针对比较常见的问题或者咨询，设置自动回复功能。这样更快速、更有效的解决群众留言问题，也一定程度上减轻了相关部门工作人员的工作压力。

5 总结

我们基于 TF-IDF 算法提取特征词，将其向量化，进一步通过比较

支持向量机、最近 k 邻近方法、朴素贝叶斯方法、随机森林分类、逻辑回归等五种分类算法训练的模型准确度，最终采用准确度最高的 LinearSVC 分类模型，并构建自定义动态字典返回相应留言问题的二级标签和三级标签，利用查准率、查全率、F1-score 以及得出的混淆矩阵进行模型评价；利用附件 3 的三级标签数目和热度指数作为热度评价指标，结合 Reddit 算法，得出排名前五的热点话题；从相关性和时效性等方面分析相关部门对群众留言问题答复意见的质量，可以看出答复意见与留言问题的整体相关性很高，时效性的满意度也都位于中上水平，此外我们还提出了一些其他的建议。

此外，由于没有二、三级标签的真实数据可利用，在问题一和问题二中提取留言内容的关键词与每个三级标签做余弦相似度从而返回三级标签中存在不足。若后期有二、三级标签的真实数据，可构造自定义动态三级标签字典，即三级标签是 key，提取该三级标签的留言信息的关键词为 value，在 value 不断的填充和完善下根据余弦相似度预测三级标签更具有可信度和说服力，以此弥补上述存在的不足。

6 参考文献

[1]李春玲. 文本挖掘在垃圾邮件过滤中的应用研究[D]. 中国人民大学，2008

[2] 胡佳妮,徐蔚然,郭军.中文文本分类中的特征选择算法研究[J]. 光通信研究，2005，129(3)

[3] Tom Mitchell. Machine learning[M]. McCraw Hill, 1996.

[4] Yiming Yang. A Comparative Study on Feature Selection in Text Categorization[J]. The ICML97,Nashville,1997

[5] 寇苏玲,蔡庆生.中文文本分类中的特征选择研究[J].计算机仿真，2007,24(3)

[6]杨海波. 基于微博热点话题发现的关键技术研究[D]. 兰州交通大学，2017

[7] 严明,郑昌兴. python 环境下的文本分词与词云制作[J]. 开发案例, 2018.34.021

[8]迪潘简. f 散卡尔(Dipanjan Sarka). Python 文本分析[M].北京：机器工业出版社，2018：132

[9] 吴龙峰,于璨,王峰. 向量空间模型的文本分类研究进展与应用[J] 宿州学院学报. 2019, 34(12)

[10] 刘永芬,程丽,陈志安.基于特征选择的 M-SVM 中文文本分类[J]. 国际 IT 传媒品牌，2019，40(9)

[11] 叶志刚.SVM 在文本分类中的应用[D]. 哈尔滨工程大学，2006

[12] 杨经. 网络舆情热点话题发现技术研究[D]. 福州大学，2011