“智慧政务”中的文本挖掘应用

摘要

近年来，随着互联网社交网络的发展，社会公共事务交流与传播逐渐网络化，公众通过微信、微博、市长信箱、阳光热线等网络问政平台参与公共事务讨论，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文本分析的工作带来了挑战。本文基于自然语言处理相关技术完成文本挖掘任务，对促进智慧政务的发展有着重要意义。

针对问题一，首先对附件 1 中信息进行抽取相关文本数据，并对留言文本进行分词、去停用词等操作，利用过采样策略对不平衡的数据集进行数据增强。接着，利用 Word2vec 构建词向量，放入双向长短期神经网络 BiLSTM 模型中提取出文本上下文信息，再接上卷积神经网络 CNN 提取文本局部语义信息输出分类结果。经过多次实验结果表明，本文所提出的 BiLSTM-CNN 方法的 F1-Score 值达到了 95.8%，优于传统的 SVM、单一的 CNN、LSTM 网络，训练效率也大大提高。

针对问题二，我们首先通过 K-means 算法对附件 3 中的留言进行两次聚类，聚类的依据分别是留言主题文本，和留言主题与留言详情拼接文本。再根据制定的热度指标，计算并筛选出两次聚类结果中热度排名前十的类别。将两个筛选的结果进行合并，去除错误分类留言，相同留言留一条，一个聚类作为一个问题。从每一类中选出一条合适的留言作为问题内容，整理出新的问题表。以问题表中的留言为中心，通过计算留言余弦相似度的方法，将附件 3 中其他相似度达到阈值的留言归类入问题表中的问题，然后计算个问题热度，选出 5 个热点问题。

针对问题三，首先对附件 4 中群众的诉求概况做了基本分析，根据答复时间、答复内容的完整性、评估了答复的效率以及答复是否解决了留言群众提出的问题，并辅以留言与答复意见的相似度计算，评价了解决情况的占比多少，最后根据分析结果给予了平台相应的改善建议。

关键词：Word2vec　BiLSTM　卷积神经网络　k-means 聚类　评价

Abstract

In recent years, with the development of social network, the communication and dissemination of social public affairs are gradually networked. The public participates in public affairs discussion through WeChat, Weibo, Mayor's mailbox, Sunshine Hotline and other online political platforms. The increasing amount of text data, related to various social and public opinions, has brought challenges to the previous work of text analysis mainly relying on manual work. This paper is based on natural language processing technology to complete text mining, which is of great significance to promote the development of intelligent government.

For the first problem, relevant text data were extracted from the information in annex one, and word segmentation and deleting stop word were carried out on the message text. Data enhancement was carried out on the unbalanced data set by using oversampling technology. Then, Word2vec is used to construct the word vector, and the text context information is extracted from the BiLSTM model of the two-way long and short term neural network, and then the convolutional neural network CNN is used to extract the local semantic information of the text and output the classification results. The results of many experiments show that the F1-Score value of the method used in this paper reaches 95.8%, which is better than the traditional SVM, single CNN and LSTM network, and the training efficiency is also greatly improved.

For question two, we firstly used the K-means algorithm to cluster the comments in attachment three twice. The clustering was based on the message subject text, and text spliced by the message subject with the message details text. Then, according to the developed heat index, the top ten heat categories in the clustering results from two times were calculated and selected. Merge the results of the two filters, remove the misclassification message, leave the same message, a cluster as a problem. Choose a suitable message from each category as the content of the question and make a new list of questions. Taking the comments in the problem table as the center, by calculating the cosine similarity of the comments, the comments in annex three whose similarity reached the threshold were classified into the problems in the problem table, and then the heat of each problem was calculated to select five hot issues.

In view of question three, this paper analyzes the appeal of the masses in appendix four, evaluates the efficiency of the reply according to the time of reply, the completeness of the reply content, and whether the reply has solved the problems raised by the masses. By calculating the similarity between comments and replies, the paper evaluates the proportion of solutions, and finally gives Suggestions for the improvement of the platform according to the analysis results.

Key words: Word2vec, BiLSTM, CNN, k-means, evaluate

目录

1. 绪论

1.1 问题背景

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，庞大的信息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。同时，也使得政府不能很好地了解到民众反应的热点问题，相关部门能否及时地给予反馈回复，与他们是否真正了解民众所表达的信息相关。

目前，大数据、云计算、人工智能等技术的发展，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。

1.2 挖掘目标

本次挖掘任务主要有三个，第一是群众留言分类，需要利用相关自然语言处理技术对留言文本进行一级分类，该模型要能在后续的留言集中实现自动分类。第二是热点问题挖掘，需要对数据集中挖掘出群众集中反映的特定地点或特定人群的热点问题，并定义适当的评价指标，得出评价结果。问题三是答复意见的评价，通过建立相关的评价体系，对相关部门的答复情况进行评价。

1.3 本文工作

针对本次的三个文本挖掘问题，构建了挖掘任务流程图，如图 1.1 所示。

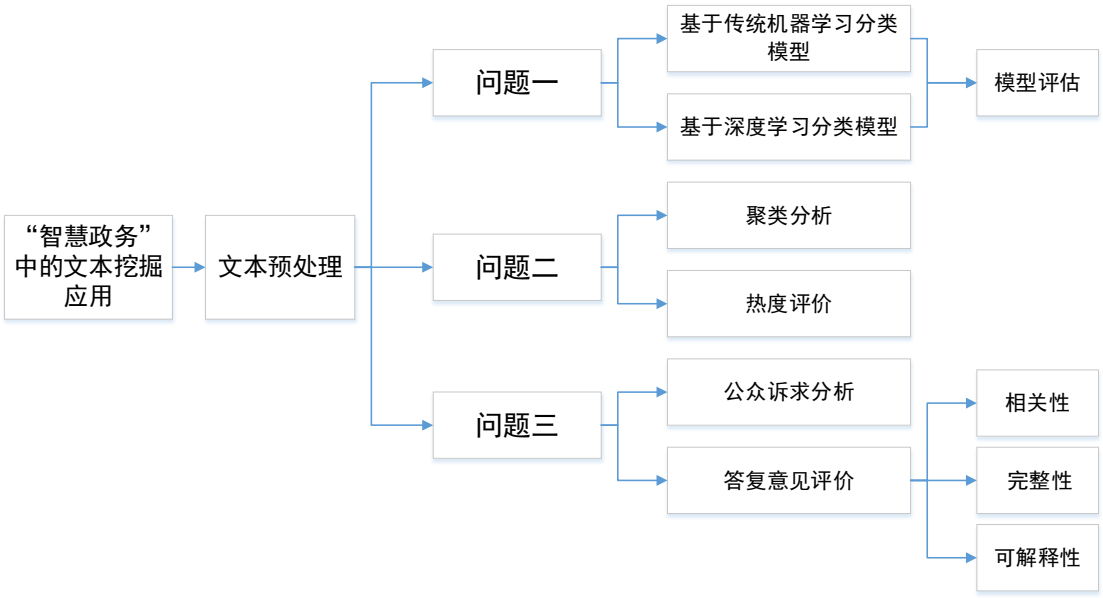


图 1.1 工作流程图

2. 文本预处理

在对文本进行分析之前，我们必然要先对文本数据进行预处理操作，包括文本筛选、去除特殊字符、文本分词、去停用词（包括标点、数字、单字和其它一些无意义的词）、文本表示等操作，如图 2.1 所示。

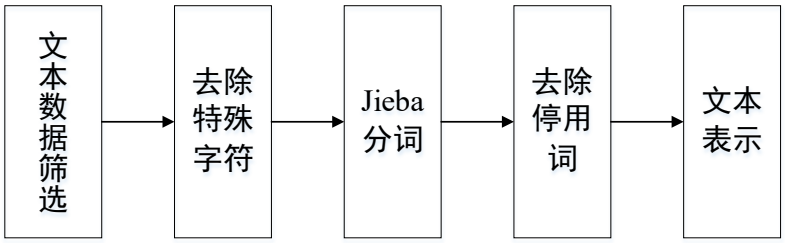


图 2.1 文本预处理流程图

2.1 文本分词

一般来说，对于句子或短文的分析，我们通常会对其进行分词操作，以获得更多的信息特征。英文文本每个单词就有其明显的含义，并且两个词之间天然的就以空格为分割。对于中文文本，由于词与词之间没有空格，因此，对中文文本进行分词操作是该工作的前提，中文分词的目的是将一句话按照一定的分词标准将其分成一个个具有独立含义的词。

本文主要运用了现在比较流行且开源的分词工具结巴 jieba 分词器，它是基于 python 开发的一个分词模块。Jieba 分词结合了基于规则和基于统计两种方法，首先基于前缀词典进行词图扫描，前缀词典可以快速构建包含全部可能分词结果的有向无环图，这个图中包含多条分词路径。其次是基于标注语料，使用动态规划的方法可以找出最大概率路径，并将其作为最终的分词结果。对于未登录词，Jieba 用了基于汉字成词的 HMM 模型，并采用 Viterbi 算法进行推导。

除此之外，jieba 中还提供了分词、词性标注和自定义词典等功能，本文增加了如“A 市、M 市、K4 县”等代表地点的分词词语，对留言进行了分词，部分分词结果如图 2.1 所示。

```
7774    [M市, 人民, 餐桌上, 的, 鸡鸭, ，, 许多, 来自, 这个, 地下, 黑, 交易,...
9036    [K4县, 端桥, 铺, 镇, 办理, 准生证, 要求, 是, 什么, ？, \n,  , ...
1311    [L10县, 王井家, 搬迁户, 房屋, 补贴, 何时, 发, ？, \n, \t, \t,...
701     [F市, 房地产, 管理局, 如此, 官僚, 工作作风, ！, \n, \t, \t, \t...
3485    [全, 大通, 湖, 市区, 最差, 的, 一条, 路, \n, \t, \t, \t, \...
955     [再次, 投诉, K8县, 住建, 局局长, 阻挡, 潇楚, 山水, 城, 更改, 规划, ...
7131    [A市, 银杏, 嘉园, 9, 栋, 的, 4, 台, 电梯, 故障, 频发, ！, \n,...
3560    [I市, 重点, 文物, “, 文昌阁, ”, 寂寞, 地, 荒废, 了, 几十年, (, ...
2484    [K3县, 白水镇, 工业园, 内东驶, 纺织厂, 制塑, 废气, 殃及, 百姓, \n, ...
5386    [三一, 重工, 无视, 法律, ，, 无故, 解聘, 员工, \n,  , ,  , ...
115     [A市, 公共汽车, 何时能, 进, 机场, 候机楼, ？, \n, \t, \t, \t,...
6519    [关于, 要求, 解决, 02, 年, A市, 统一, 招考, 录用, 事业, 编制, 人员...
2286    [F7县, 南江, 镇, 重金属, 污染, 严重, \n, \t, \t, \t, \t, ...
6044    [反映, C市, 江滨, 农贸市场, 工程项目, 的, 农民工, 工资, 问题, \n,  ...
8565    [K8县, 清水, 桥, 孩子, 晚上, 发个, 烧, ，, 都, 没有, 地方, 看病, ...
4580    [C3县, 克扣, 2017, 年, 新, 教师工资, \n, \n,  ,  , 2017...
2309    [请, 取缔, I2区, 欧江岔镇, 高平, 村, 的, 非法, 水泥, 预制厂, \n, ...
3444    [I6, 市, 的士, 收费, 乱象, 何时能, 治理, ？, \n, \t, \t, \t...
```

图 2.2 部分分词结果

2.2 去停用词

由分词后的结果可知，分词后还存在着大量的标点、特殊字符等表达无意义的词，对后续分析必然会造成影响，接下来进行去停用词操作。

停用词是指功能普遍而实际上并没有什么实际意义的词，包括一些副词、介词和连接词，如“的，了，吗”等。这些词不仅不会对文本分析任务的效果带来提升反而会降低准确度，因此，这些词在文本预处理阶段就得剔除。目前的停用词表有哈工大停用词表、百度停用词表、四川大学机器智能实验室停用词表等等，本文选择了哈工大的停用词表。除此之外，还对文本中的特殊字符及标点剔除。

经过去标点，去特殊符号，去停用词等操作后的部分结果如图 2.2 所示。

```
7774    M市 人民 餐桌上 鸡鸭 来自 地下 黑 交易 窝点 尊敬 杨 市长 地地道道 M市 本地人...
9036    K4县 端桥 铺 镇 办理 准生证 要求 K3县 人 老公 K4县 端桥 铺 镇 人 K4县...
1311    L10县 王井家 搬迁户 房屋 补贴 发 王井家 搬迁户 都 迁出去 几年 至今 还 不能 ...
701     F市 房地产 管理局 官僚 工作作风 尊敬 盛华荣 市长 您好 最近 F市 正在 开展 加强...
3485    全 大通 湖 市区 最差 一条 路 全 大通 湖区 最牛 一条 路 柳 杨村 路 房子 建 ...
955     再次 投诉 K8县 住建 局局长 阻挡 潇楚 山水 城 更改 规划 审批 程序 尊敬 李慧 ...
7131    A市 银杏 嘉园 9 栋 4 台 电梯 故障 频发 A市 A2区 老虎 塘路 65 号 银杏...
3560    I市 重点 文物 文昌阁 寂寞 荒废 几十年 图 文物 国家 不可 再生 文化 资源 我市 ...
2484    K3县 白水镇 工业园 内东骏 纺织厂 制塑 废气 殃及 百姓 尊敬 局长 大人 您好 K市...
5386    三一 重工 无视 法律 无故 解聘 员工 三一 重工 无视 法律 无故 解聘 员工 合同期 ...
115     A市 公共汽车 何时能 进 机场 候机楼 黄花 机场 A市 窗口 89 年 通航 至今 公共...
6519    要求 解决 02 年 A市 统一 招考 录用 事业 编制 人员 待遇 请示 依据 政府 发文...
2286    F7县 南江 镇 重金属 污染 严重 兹 F市 F7县 南江 镇 黄 裴村 老屋 组与 杨家...
6044    反映 C市 江滨 农贸市场 工程项目 农民工 工资 问题 胡 厅长 C市 江滨 农贸市场 工...
8565    K8县 清水 桥 孩子 晚上 发个 烧 都 没有 地方 看病 尊敬 张 厅长 您好 清水 桥...
4580    C3县 克扣 2017 年 新 教师工资 2017 年 招聘 新 教师 240 名 八月份 ...
2309    请 取缔 I2区 欧江岔镇 高平 村 非法 水泥 预制厂 尊敬 市 领导 您好 I2区 欧江...
3444    I6 市 的士 收费 乱象 何时能 治理 I6 市 的士 收费 乱象 曾 西地 省台 新闻 ...
```

图 2.3 停用词过滤后结果

观察可以发现，去掉停用词后的数据显得比较“干净”，去停用词本质上也是一种特征选择。

2.3 文本表示

经过一定的文本处理后，还需要将文本表示成计算机能直接处理的形式，即文本数字化。常见的文本表示方法有传统的词袋模型(BOW, Bag Of Words)、TF-IDF(词频-逆文档频率)和向量空间模型(Vector Space Model)等。

2.3.1 词袋模型

词袋模型(BOW, Bag Of Words)是将所有词语看成一个装满词的袋子，词是在袋子里随机放着的。不考虑其词法和语序的问题，即每个词语都是独立的。

例如：

1）Tom wants to go to Beijing

2）Jack wants to go to Chongqing

根据上面 2 个例句，可以构建一个语料的列表：

[Beijing, Chongqing, go, Jack, to, Tom, wants]

利用该语料列表可以构建长度为 7 的向量，下标与映射数组的下标相匹配，数字代表该词语出现的次数：

1）[1, 0, 1, 0, 2, 1, 1]

2）[0, 1, 1, 1, 2, 0, 1]

从以上例子可以看出，词袋模型的建立比较简单，只是简单的进行了统计词频信息，在实际问题中准确率往往比较低，对于文本中出现的词一视同仁，语序关系已经完全丢失，不能体现不同词在一句话中的不同的重要性。

2.3.2 TF-IDF

TF-IDF（Term Frequency-Inverse Document Frequency）即“词频-逆文本频率”。它由 TF 和 IDF 两部分组成。TF 指的是词频，在文本向量化中利用词袋模型做了文本中各个词的出现频率统计，并作为文本特征。IDF 指的是逆文本频率，如之前例子出现的“to”其词频虽然高，但是重要性却应该比出现词频低的" Beijing "和“Chongqing”要低。对一个文本中，如果包含词条 w 的文档越少, IDF 越大，说明该词条具有很好的类别区分能力，IDF 就是来帮助我们来反应这个词的重要性的，进而修正仅仅用词频表示的词特征值。

TF 计算公式如下：

$$TF_{\mathrm{w}} = \frac{\text{某一词条w 出现的次数}}{\text{所有的词条数目之和}} \tag{2-1}$$

IDF 计算公式如下：

$$IDF = \log\left(\frac{\text{语料库的总文档数}}{\text{包含词条w的文档数+1}}\right) \tag{2-2}$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语，TF-IDF 计算公式如下：

$$TF - IDF = TF * IDF \tag{2-3}$$

TF-IDF 算法的优点是简单快速，结果比较符合实际情况。缺点是单纯以“词频”衡量一个词的重要性，不够全面，有时重要的词可能出现次数并不多。而且，这种算法无法体现词的位置信息，出现位置靠前的词与出现位置靠后的词，都被视为重要性相同，这是不正确的。

2.3.3 Word2vec

Word2vec 是 Google 团队在 2013 年发表的词向量工具，它可以将词与词之间进行向量化，这样词与词之间就可以定量的去度量他们之间的关系，挖掘词之间的联系。word2vec 工具主要包含两个模型：连续词袋模型(continuous bag of words, 简称 CBOW)和跳字模型(skip-gram)。

CBOW 是 Continuous Bag-of-Words 的缩写，与神经网络语言模型不同的是，CBOW 去掉了最耗时的非线性隐藏层。从图 2.3 中可以看出，CBOW 模型预测的是 $p(w_t \mid w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2})$，由于图中目标词 w_t 前后只取了各两个词，所以窗口的总大小是 2。假设目标词 w_t 前后各取 k 个词，即窗口的大小是 k，那么 CBOW 模型预测的将是 $p(w_t \mid w_{t-k}, w_{t-(k-1)}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+(k-1)}, w_{t+k})$。

Skip-Gram 的模型图与 CBOW 恰好相反，如图 2.4 所示。Skip-Gram 模型预测的是 $p(w_{t-2} \mid w_t), p(w_{t-1} \mid w_t), p(w_{t+1} \mid w_t), p(w_{t+2} \mid w_t)$，由于图中词 w_t 前后只取了各两个词，所以窗口的总大小是 2。假设词 w_t 前后各取 k 个词，即窗口的大小是 k，那么 Skip-Gram 模型预测的将是 $p(w_{t+p} \mid w_t)(-k \le p \le k, k \ne 0)$。

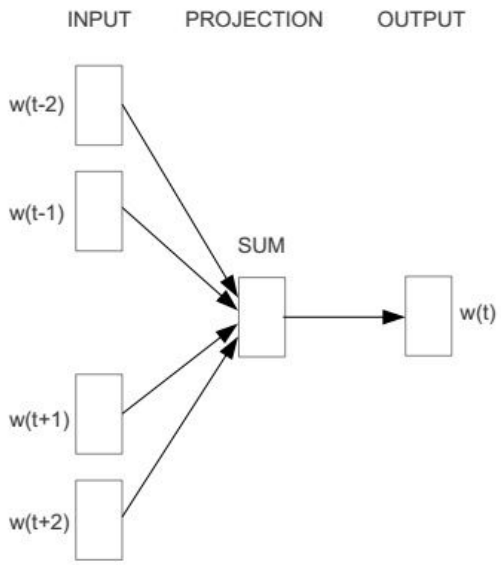


图 2.4 CBOW 模型

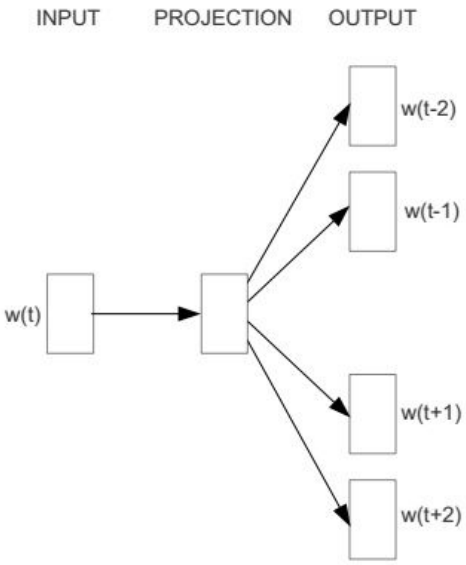


图 2.5 Skip-Gram 模型

Word2vec 可以将词从高维空间分布式映射到低维空间，而且保留了词向量之间的位置关系，从而解决了向量稀疏和语义间的联系问题。

词向量模型需要大量的语料库进行训练，且耗费时间长，本文采用了腾讯 AI 实验室的预训练模型，他们使用了汉字词句嵌入语料库，该语料库为超过 800 万个中文单词和短语，利用 Skip-Gram 模型，训练出了 200 维的词向量表示（也称为词嵌入），这些单词和短语已在大规模高质量数据上进行了预训练。这些向量捕获了中文单词和短语的语义，可以广泛应用于中文处理任务以及进一步的研究中。

3. 问题一分析及求解

3.1 问题一分析

问题一主要是对附件 2 中的留言内容进行一级分类，由于是对文本内容进行分类，还需对文本进行预处理操作，包括分词、去停用词、文本特征表示等等。最后建立相应的模型实现留言内容分类，并利用 F1-Score 来评估模型。整个流程如图 3.1 所示。

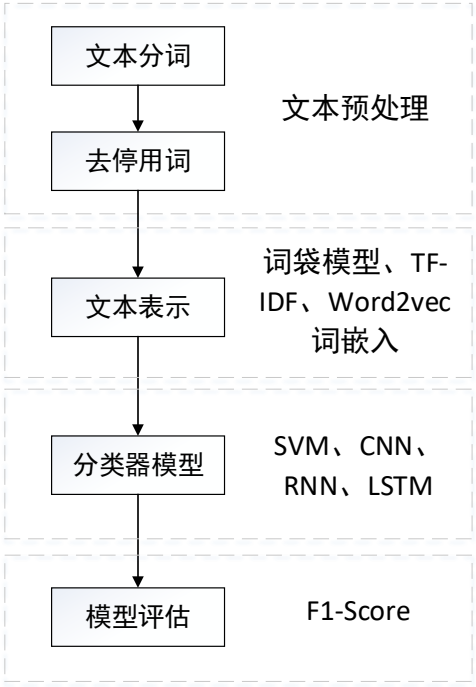


图 3.1 问题一流程图

3.2 文本分类模型

3.2.1 基于传统机器学习的文本分类

传统机器学习的文本分类通常是利用词袋模型或 TF-IDF 提取文本特征，然后利用机器学习中的贝叶斯、支持向量机等进行文本分类。

（1）支持向量机

支持向量机(Support Vector Machines, 简称 SVM)，是一种建立在统计学理论基础上的机器学习方法，常用于分类、回归问题。支持向量机就是在特征空间中找出一个超平面作为决策边界，使模型在数据上的分类误差尽量最小。

假设训练数据集为 $T=\{(x_1,y_1),(x_2,y_2),\cdots(x_n,y_n)\}$，其中，$x_i$ 为第 i 个特征向量，y_i 为 x_i 对应的标签；(x_i,y_i) 即为训练样本 T 中的第 i 个样本点。

分类超平面公式为 $w^T \bullet x_i + b = 0$，x_i 为第 i 个输入向量，w 为 x_i 对应的权值向量，b 为偏移向量，如图 3.2 所示。对于线性可分的数据集来说，这样的超平面有无穷多个，但是几何间隔最大的分离超平面却是唯一的。经过分离后得到决策函数 $f(x)=sign(w^T \bullet x + b = 0)$。

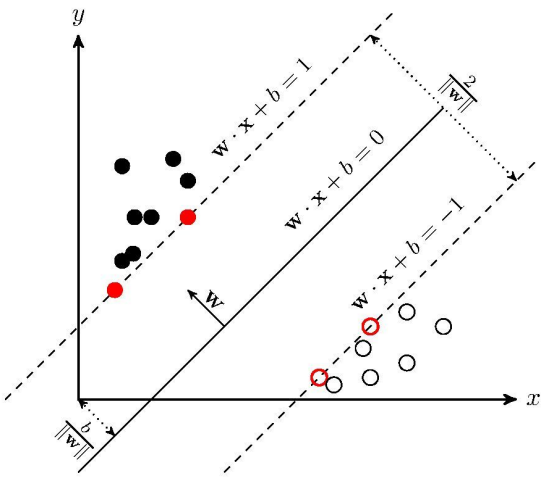


图 3.2 超平面划分

对于非线性可分的样本，可以通过非线性变换将它转化为某个维度特征空间中的线性分类问题，在高维特征空间中学习线性支持向量机，核函数就是将低维度空间的样本映射到高维度空间，使得数据集可分，核函数公式为 $k(x,z)=(\varphi(x)\bullet\varphi(z))$，式中：$\varphi$ 表示从输入空间到高维空间的非线性映射过程，最终得到非线性样本的决策函数：

$$f(x)=sign\left\{\sum_{i=1}^{l}a_i y_i(\varphi(x_i)\bullet\varphi(x))+b\right\} \qquad (3\text{-}1)$$

式中 a_i 、b 为可调节参数，以获得最优分类平面。

3.2.2 基于深度学习的文本分类

3.2.2.1 长短期记忆神经网络

LSTM 是一种特殊的循环神经网络(RNN)，被广泛应用于处理时序问题中，文本本身具有上下文语义。RNN 模型在处理较长的句子时，往往只能够理解有限长度内的信息，对于较长范围的文本，其有用信息不能很好地利用起来。LSTM 就是用来解决 RNN 中存在的问题，可以说是 RNN 的一种变体，由遗忘门，输入门，输出门三个门控制整个时间序列中的信息，如图 3.3 所示。

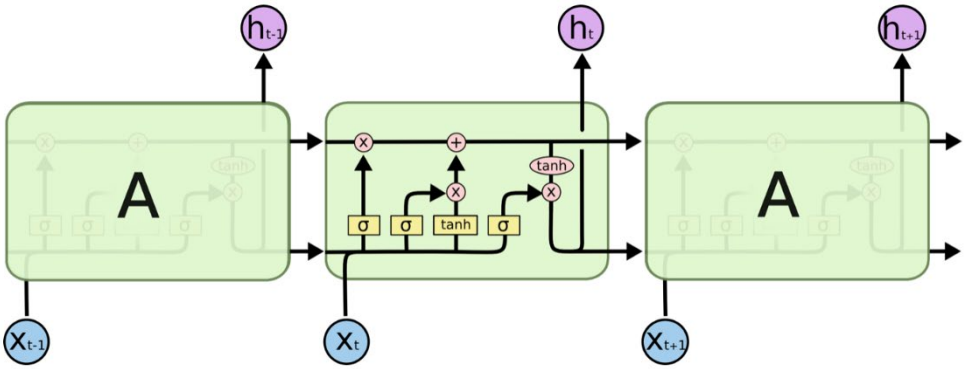
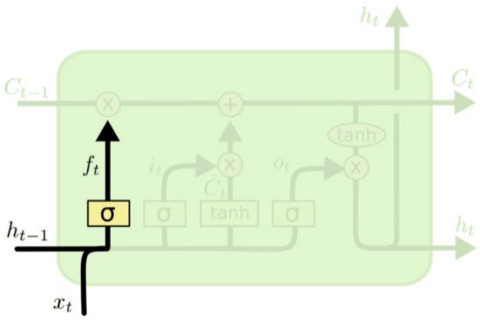


图 3.3 LSTM 网络结构

（1）遗忘门

遗忘门就是对上一节点传来的输入进行选择性的忘记，将当前信息和隐藏信息同时传递到 Sigmoid 函数中，其输出值为 0-1 之间，越接近 1 则选择保留，越接近零则选择丢弃，如图 3.4 所示。

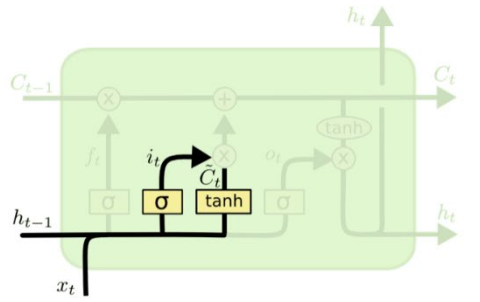


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

图 3.4 遗忘门

（2）输入门

输入门控制接受程度。首先由上一层的输入和当前信息同时传递到 Sigmoid 函数中，其输出值为 0-1 之间，越接近 1 则选择保留，越接近零则选择丢弃。其次，还要将上一层的输入和当前信息同时传递到 tanh 函数中，将输入变为[-1,1]的标准化区间。最后，将 sigmoid 的输出值与 tanh 的输出值相乘，作为下一阶段的输入。其中，sigmoid 的输出值将决定 tanh 的输出值中哪些信息是重要且需要保留下来的，如图 3.5 所示。

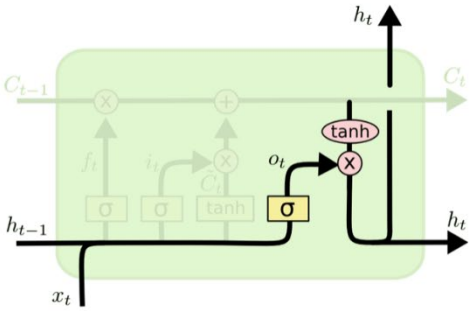


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

图 3.5 输入门

（3）输出门

不像 RNN 模型，LSTM 内部状态的输出并不会直接用作模型的输入，而是又在输出门的控制下进行选择性输出。首先，将上一个隐藏状态信息和当前输入信息传递到 sigmoid 函数中，然后将新得到的细胞状态与 tanh 的输出相乘，最后把新的细胞状态和新的隐藏状态传递到下一个时间步长中去，如图 3.6 所示。



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

图 3.6 输出门

3.2.2.2 卷积神经网络

卷积神经网络(Convolutional Neural Network, CNN)是一种前馈型的神经网络，在图像处理领域取得很好的成绩。卷积神经网络由输入层、卷积层、池化层、全连接层四部分组成，如图 3.7 所示。

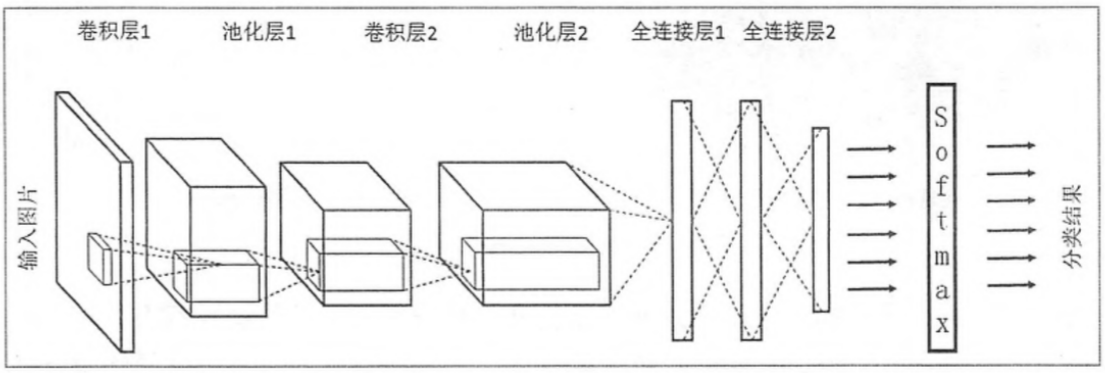


图 3.7 卷积神经网络结构

2013 年 Kim 提出的 Text-CNN 模型作为文本分类模型，如图 3.8 所示。

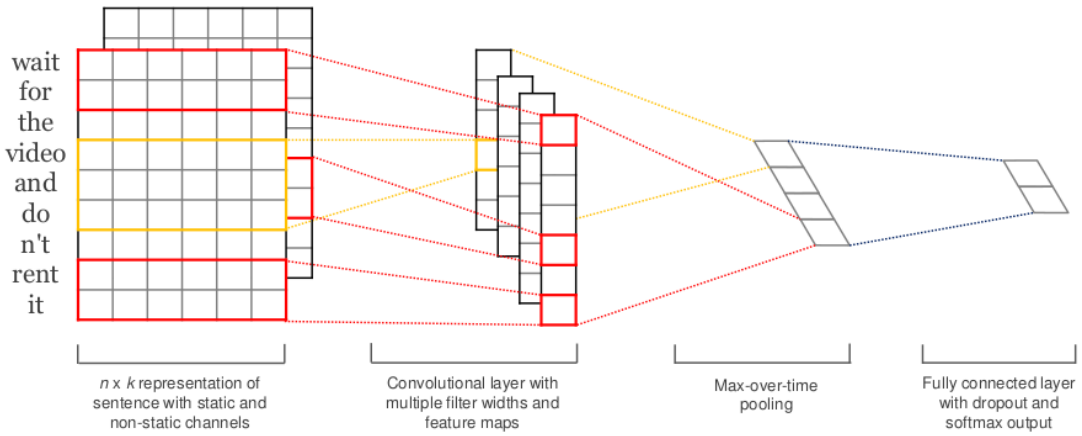


图 3.8 Text-CNN 模型结构

（1）输入层

在应用卷积神经网络模型时，文本处理与图像处理的输入层不同，Kim 在其

论文中提出了 4 中输入层的词向量表达方式:static（静态词向量）、non-static（非静态词向量）、multiple channel（多通道）和 CNN-rand（随机初始化）。其中 static（静态词向量）是利用 Word2vec、fastText 或者 Glove 等词向量工具，在开放领域语料库数据上进行无监督的学习，获得词汇的具体词向量表示方式，并将其拿来直接作为输入层的输入，在 TextCNN 模型训练过程中不再调整词向量。本文利用腾讯 AI 实验室开源的词向量模型，每个词语对应 200 维的词向量作为输入。

除此之外，我们还需对每条文本进行定长处理。对于文本分类任务，和图像处理不同的是，Text-CNN 模型的输入层需要输入一个定长的文本序列，我们需要通过分析语料集样本的长度指定一个输入序列的长度 L。如果文本长度小于 L，需要对其进行填充(一般填充为 0)，如果文本长度大于 L，则在长度 L 处进行截取。最终输入层输入的是文本序列中各个词汇对应的分布式表示，即词向量。

（2）卷积层

文本处理的卷积核一般只进行一维的窗口滑动，即卷积核的宽度与词向量的维度宽度相等。在 Text-CNN 模型中可以使用多个不同尺寸的卷积核，提取局部信息。卷积核的高度，即窗口值，可以理解为 N-gram 模型中的 N，即利用的局部词序的长度，窗口值也是一个超参数，一般选取 2-8 之间的值。

（3）池化层

在 Text-CNN 模型的池化层中一般使用 Max-pool（最大值池化），减少了模型的参数，又保证了在不定长的卷基层的输出上获得一个定长的全连接层的输入。

卷积层与池化层在分类模型的核心作用就是特征提取的功能，从输入的定长文本序列中，利用局部词序信息，提取初级的特征，并组合初级的特征为高级特征，通过卷积与池化操作，省去了传统机器学习中的特征工程的步骤。

TextCNN 有一个明显缺点，在利用卷积、池化操作时，会丢失文本序列中的词汇的顺序、位置信息，难以捕获到文本序列中的否定、反义等语义信息。

3.2.3 BiLSTM-CNN 组合模型

长短期记忆神经网络(LSTM)虽然解决了传统循环神经网络(RNN)的梯度消失或梯度爆炸问题。但在文本分类中，LSTM 模型只能学习到上文的信息，无法利用下文的信息，因为一个词的语义不仅与上文有关。还可能与下文有关，因此，利用 BiLSTM(Bi-directional Long Short Term Memory)可以很好地解决这个问题，它是由前向的 LSTM 与后向的 LSTM 结合而成，如图 3.9 所示。

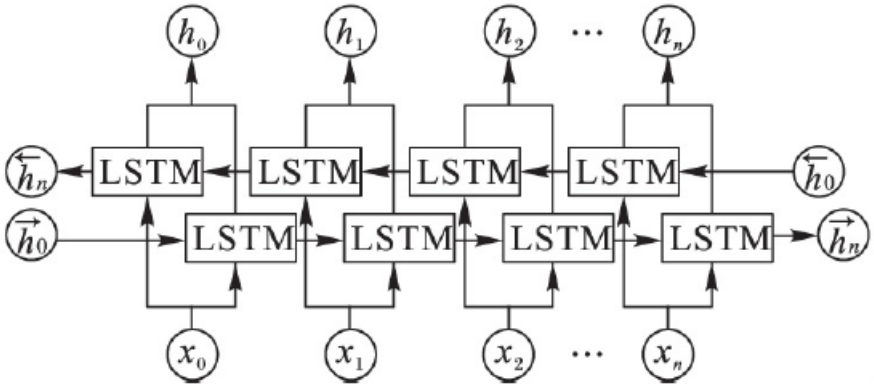


图 3.9　BiLSTM 模型结构

在 BiLSTM 模型中，$\overrightarrow{h_t}$ 表示 t 时刻 LSTM 的正向输出，$\overleftarrow{h_t}$ 表示 t 时刻 LSTM

的反向输出，h_t 表示 t 时刻 BiLSTM 的输出，x_t 表示 t 时刻的输入。其计算式为:

$$\overrightarrow{h_t}=LSTM(x_t,\overrightarrow{h}_{t-1}) \quad (3-2)$$

$$\overleftarrow{h_t}=LSTM(x_t,\overleftarrow{h}_{t-1}) \quad (3-3)$$

$$h_t=w_t\overrightarrow{h_t}+v_t\overleftarrow{h_t}+b_t \quad (3-4)$$

其中，w_t 表示正向输出权重矩阵，v_t 表示反向输出的权重矩阵，b_t 表示 t 时刻的偏置。

本文选择的模型由双向长短期记忆神经网络和卷积神经网络组合而成，如图 3.1 所示。首先利用 BiLSTM 模型获取文本信息特征，相比以往的单向 LSTM，可以最大程度地捕捉上下文信息。接着，利用卷积神经网络对 BiLSTM 模型的输出进一步提取，捕获文本中的关键组件，解决了 BiLSTM 模型无法获取文本局部特征信息和单个卷积神经网络模型无法得到词语在上下文中语义的问题。我们的模型结合了 LSTM 的结构和 CNN 结构的最大池化层，利用了循环神经网络模型和卷积神经模型的优点。

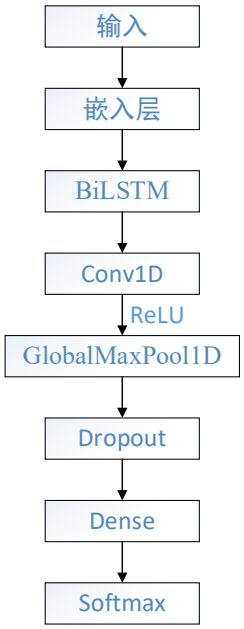


图 3.10　BiLSTM-CNN 模型结构

3.3 实验分析

3.3.1 实验环境

本文的实验环境如下：操作系统为 64 位 Win10，CPU 为 Intel Core i7-9750H, GPU 为 GeForece GTX1660Ti, 显卡驱动为 NVIDIA-SMI 417.92 ,内存大小为 DDR4 16G, 开发环境为 Tensorflow2.0, 开发工具为 Jupyter Notebook。

3.3.2 实验数据

问题一的数据集为附件 2，附件 2 中包含留言编号、留言用户、留言主题、留言时间、留言详情和一级标签，根据任务要求，我们选取了留言主题、留言详情和一级标签栏进行文本分类预测。

首先是对分类类别转化为数字编码格式，并统计各分类下的样本分布情况，如图 3.11 所示。

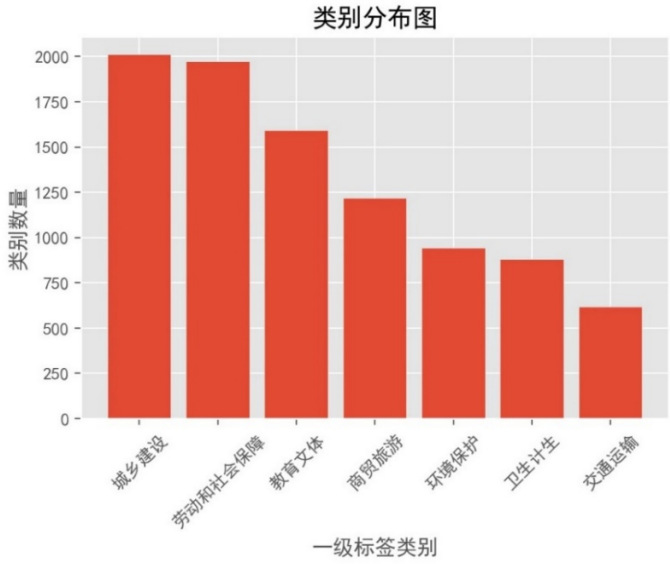


图 3.11　类别分布图

从图中可以到，样本存在着严重的不均衡问题，这必然会对后续的模型学习造成影响。例如整个训练样本中有 998 个反例，而正例只有 2 个，那么模型学习只需要返回一个永远将新的样本预测为反例的学习器，就能达到 99.8%的精度。可见，这样的学习器往往没有什么价值，因为它不能预测出任何正例。

对于样本不均衡问题，一般采用欠采样和过采样方法，欠采样即去除一些多数类中的样本使得各个样本的数目接近，然后再进行学习；过采样即增加一些少数类样本使得各个样本数目接近，然后再进行学习，有随机打乱句子、加入噪声、裁剪掉长文本的某一句和复制等方法。本文主要利用随机打乱词语间顺序进行过采样，使得各个样本分布均衡，过采样后的类别分布如图 3.12 所示。

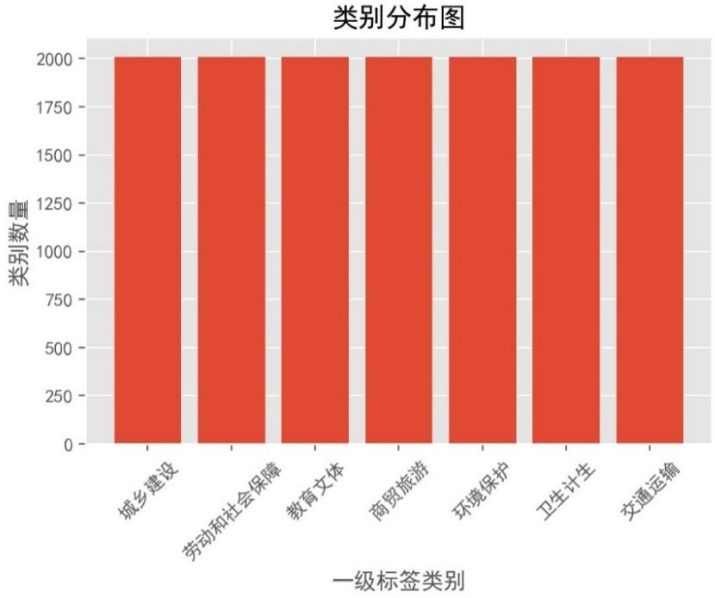


图 3.12　过采样后的类别分布图

我们还对一级标签进行了编码，将文字型数据转换为数值型数据，以便进行建模，如表 3.1 所示。

表 3.1 特征编码

一级标签	编码
交通运输	0
劳动和社会保障	1
卫生计生	2
商贸旅游	3
城乡建设	4
教育文体	5
环境保护	6

在应用词向量前，还需确定每条留言的长度，通过遍历每一条文本词语个数，除以所有文本条数，计算得出每条留言平均有 146 个词，文本长度分布图如图 3.13 所示。本文我们选取 200 作为文本的固定长度，即小于 200 长度的文本在其末尾添 0 进行，大于 200 长度的文本在 200 位置处截断。

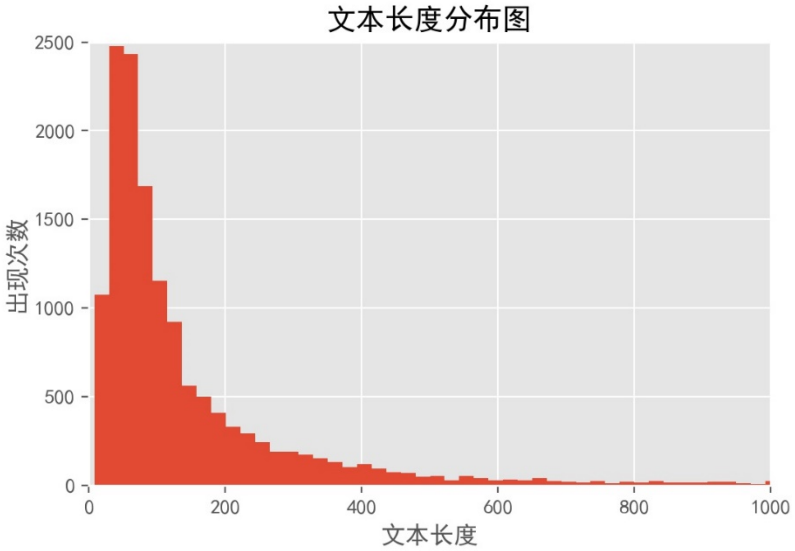


图 3.13　文本长度分布图

接着，我们利用了腾讯 AI 实验室开源的的预训练模型，该模型是利用 Skip-Gram 模型训练大量语料得到的，每个词向量维度为 200，这样就构成了整个模型的输入。

3.3.3 实验参数

实验前对数据进行拆分，0.8 比例为训练集，0.2 比例为测试集，在训练集里又分 0.2 比例为验证集，0.8 比例为训练集。将训练集中所有样本放进模型里进行训练。

由于实验参数的设定对结果影响较大，本次实验采用参数固定法，CNN 隐含层分别取 32，64，128，过滤器分别取 3，4，5；BiLSTM 隐含层分别取 64，128，进行多次实验。实验选择的优化器为 Adam，交叉熵作为损失函数。经过多次实验，发现使用表 3.2 列出的参数时，分类效果最好。

表 3.2 模型参数

参数	值
词向量维度	200
BiLSTM 隐层节点数	32
Dropout 比率	0.4、0.5
CNN 隐层节点数	128
滤波器窗口大小 kernel_size	4
池化方式	GlobalMaxPool1D
Loss	Sparse_categorical_crossentropy
Optimizer	Adam
Batch_size	128

模型的整个结构如图 3.14 所示：

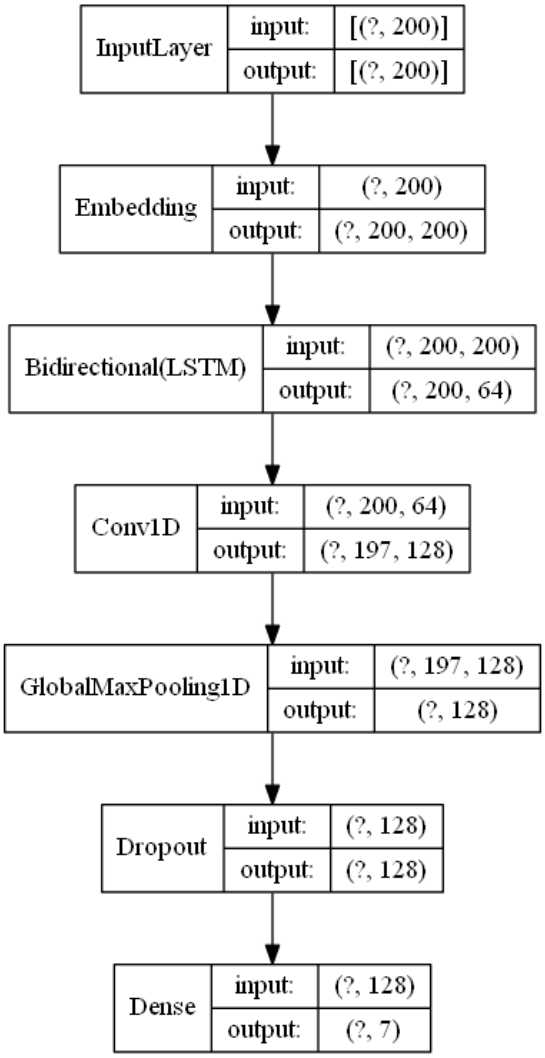


图 3.14　文本模型结构图

3.3.4 模型评价

为验证本文提出的基于 BiLSTM-CNN 模型方法的有效性，将本文方法与传统机器学习方法(SVM), CNN, LSTM 进行对比，具体实验如下：

1）传统机器学习：采用 TF-IDF 构建词向量作为输入数据，使用 SVM 进行分类。

2）CNN：采用 Word2vec 预训练词向量，利用 CNN 提取特征并进行分类。

3）LSTM：采用 Word2vec 预训练词向量，利用 LSTM 提取特征并进行分类。

4）本文方法：采用 Word2vec 预训练词向量，利用 LSTM 获取上下文信息，再利用 CNN 提取关键信息进行分类。

根据题目要求，选择 F-Score 作为分类方法的评价指标，其公式为：

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

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

实验结果如表 3.3 所示。

表 3.3 不同方法下的 F1 值、精确率和损失值对比

模型	F-Score	精确率	损失值
SVM	0.921	0.923	0.295
CNN	0.934	0.934	0.250
LSTM	0.943	0.943	0.282
本文方法	0.958	0.958	0.238

由表 3.3 可知，本文文本分类算法在 F1 指标方面，比传统 SVM、单一 CNN、LSTM 算法分别提高了 3.70 个百分点、2.40 个百分点和 1.5 个百分点，精确率和损失值明显优于传统机器学习算法、单一的 CNN 和 LSTM 模型。

分析其原因，传统的 SVM 模型需要人工提取特征向量，且 TF-IDF 后的向量为稀疏矩阵，浪费了大量的内存，不但效率低且准确率提升幅度也有限。基于 Word2vec 预训练词向量的单一 CNN 文本分类模型只考虑了局部语义特征对文本分类模型的影响，没有考虑文本上下文的关系。基于 Word2vec 预训练词向量的单一 LSTM 文本分类模型只考虑文本的下文的信息，没有考虑文本的上文信息以及局部语义对文本分类的影响，且 LSTM 在训练时较慢，迭代次数较多。

本文模型的训练结果如图 3.15 所示，随着训练轮数的增加，训练集和验证集上的准确率不断接近，达到一定值趋于平稳，训练损失值也同时不断下降，达到一定值趋于平稳，无明显的过拟合现象，这是因为我们使用了预训练好的词向量的结果。

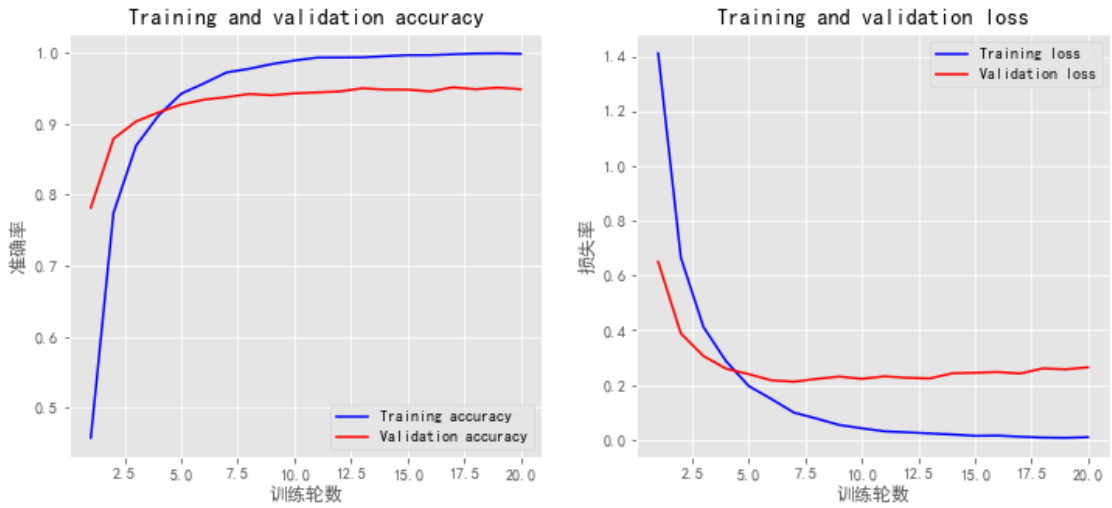


图 3.15 训练集和验证集的准确率及损失率

综合考虑文本上下文信息和局部语义特征对文本分类模型的影响，本文所提出的 BiLSTM-CNN 混合模型算法结合了单一模型的优点，使得模型训练效率提高，能更好的完成文本分类任务。

4. 问题二分析及求解

4.1 问题二分析

问题二描述：

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。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。请根据附件 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，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出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”。

问题二任务：

1、对所给留言进行归类，归类依据为特定地点或特定人群；

2、对聚类结果进行评价；

3、定义合理的热度评价指标；

4、计算各类问题的热度指数；

5、整理热度排名前 5 的留言。

整个流程如图 4.1 所示。

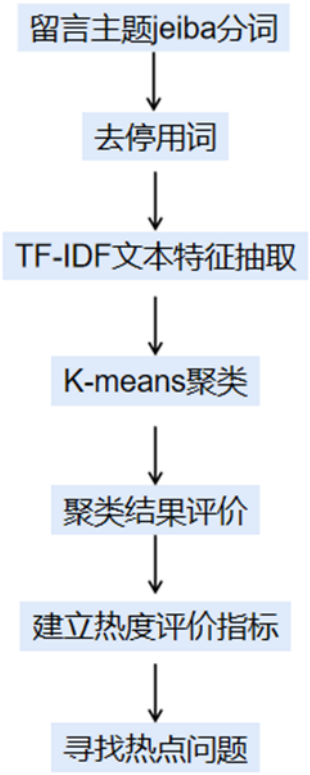


图 4.1 问题二流程图

4.2 聚类分析

4.2.1 K-means 聚类原理

聚类是把各不相同的个体分割为有更多相似性子集合的工作。聚类生成的子集合称为簇。

聚类的要求：

1、生成的簇内部的任意两个对象之间具有较高的相似度；

2、属于不同簇的两个对象间具有较高的相异度；

3、聚类与分类的区别在于聚类不依赖于预先定义的类，没有预定义的类和样本。

在本问题中，我们采取 K-means 算法进行留言的归类。

1、随机在数据当中抽取三个样本，当作三个类别的中心点（k1，k2，k3），如下图 4.2 所示。

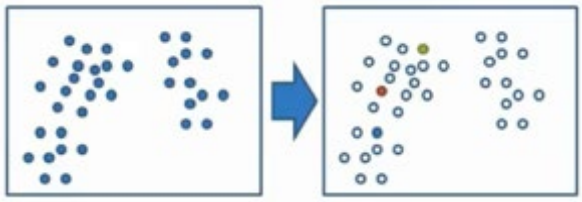


图 4.2 类别中心点

2、计算其余的点分别到这三个中心点的距离，每一个样本有三个距离（a，b，c），从中选出距离最近的一个点作为自己的标记形成三个族群，如下图 4.3 所示。

图 4.3 计算距离

3、分别计算这三个族群的平均值，把三个平均值与之前的三个旧中心点进行比较。如果相同，结束聚类，如果不相同，则把这三个平均值当做新的中心点，重复第二步，如图 4.4 所示。

图 4.4 平均值比较

4.2.2 聚类步骤：

1、随机设置 K 个特征空间内的点作为初始的聚类中心；

2、对于其他每个点计算到 K 个中心的距离，未知的点选择最近的一个聚类中心点作为标记类别；

3、接着对着标记的聚类中心之后，重新计算出每个聚类的新中心点（平均值）；

4、如果计算得出的新中心点与原中心点一样，那么结束，否则重新进行第二步过程。

4.2.3 性能评估指标

轮廓系数计算公式：

$$sc_i = \frac{b_i - a_i}{\max(b_i, a_i)} \tag{4-1}$$

对于每个点 i 为已聚类数据中的样本，b_i 为 i 到其他族群的所有样本的距离最小值，a_i 为 i 为到本身簇的距离平均值。最终计算出所有样本点的轮廓系数平均值，作为 K-means 的性能评估指标。

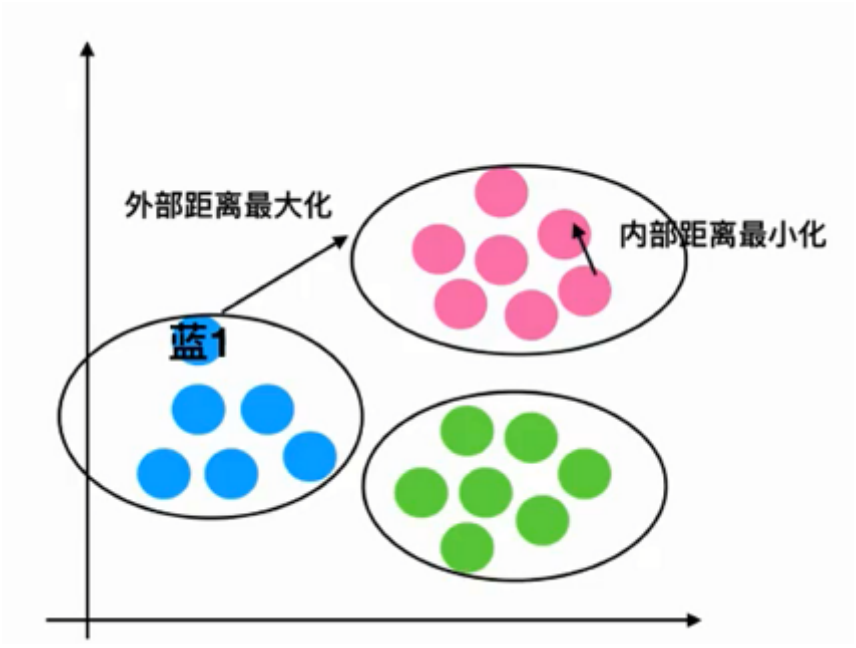


图 4.5 评估指标示意图

当 b_i 远大于 a_i 时，轮廓系数接近 1，此时的聚类效果最完美；反之，当 b_i 远小于 a_i 时，轮廓系数接近-1，此时聚类效果最差。

一般来说，轮廓系数的平均值超过 0.1 时，聚类效果就比较好了。

4.2.4 聚类结果

在本问中，通过对 K 值，即聚类数的遍历，以选取评价结果较好的 K 值，最终将 K 值确定为 3000。此时，K-means 性能评估指数的值为 0.139，基本符合要求。

4.3 热度评价

4.3.1 热度评价指标设定

一般来说，一个问题的热度，就是该问题的受关注程度。但在本问中，目的是找出某一时段内群众集中反映的某一问题可称为热点问题，因此根据受问题影响人群的数量来评判热度。

在本问中，体现民众受问题影响人群的数量的依据，主要是针对某问题的留言数量、对问题相关留言的点赞数和反对数。在所给的数据中，同一留言点赞数和反对数均没有时间分布，只有关于某一问题不同留言的时间，因此我们不考虑热度随时间衰减的情况，仅以三个依据的数量进行热度指数计算。其中，留言数和点赞数，可以在很大程度上反应受到某一问题影响，及关注这一问题的人数。而反对数则表示认为该问题没有带来负面影响的人数。根据这一思路，具体的热度计算方式如下表所示。

表 4.1 问题热度评分表

依据	一条留言	一个点赞	一个反对
得分	+1	+1	−1

热点问题计算公式如下：

$$R_i = n_i + d_i - f_i \tag{4-2}$$

其中：

R_i——第 i 类问题的热度；

n_i——第 i 类问题的留言总数；

d_i——第 i 类问题的点赞总数；

f_i——第 i 类问题的反对总数。

4.3.2 评价结果

通过 K-means 聚类算法，对留言主题进行聚类，并计算各聚类热度，初步得出热度排名前十的类别，如下表所示：

表 4.2 主题聚类、热度前 10 类表（部分）

问题 ID	留言编号	留言用户	留言主题
1	208636	A00077171	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
2	223297	A00087522	反映 A 市金毛湾配套入学的问题
3	220711	A00031682	请书记关注 A 市 A4 区 58 车贷案
4	217032	A00056543	严惩 A 市 58 车贷特大集资诈骗案保护伞
5	216316	A00097196	A4 区绿地海外滩二期业主被噪音扰得快烦死了
6	193091	A00097965	A 市富绿物业丽发新城强行断业主家水
7	234885	A00060375	A6 区月亮岛路 11 万伏高压线没用地埋方式铺设
8	284571	A00074795	建议西地省尽快外迁京港澳高速城区段至远郊
9	233542	A00080329	问问 A 市经开区东六线以西泉塘昌和商业中心以南的有关规划
10	200667	A00079480	请问 A 市为什么要把和包支付作为任务而不让市场正当竞争？

通过 K-means 聚类算法，对留言主题和留言详情进行聚类，并计算各聚类热度，初步得出热度排名前十的类别，如下表所示：

表 4.3 主题合并详情后聚类、热度前 10 类表（部分）

问题 ID	留言编号	留言用户	留言主题
1	208636	A00077171	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
2	223297	A00087522	反映 A 市金毛湾配套入学的问题
3	218132	A000106090	再次请求过问 A 市 58 车贷案件进展情况
3	272858	A00061787	A 市 58 车贷恶性退出案件为什么不发布案情进展通报？
4	217032	A00056543	严惩 A 市 58 车贷特大集资诈骗案保护伞
5	191951	A00041448	A4 区绿地海外滩小区距渝长厦高铁太近了
6	193091	A00097965	A 市富绿物业丽发新城强行断业主家水
7	218442	A00099016	A6 区月亮岛路架设高压电线环评造假，谁为民众做主
8	284571	A00074795	建议西地省尽快外迁京港澳高速城区段至远郊
9	196282	A00039390	反映 A 市地铁 3 号线松雅西地省站地下通道建设问题
9	267630	A000100648	反映 A 市地铁 3 号线松雅湖站点附近地下通道问题
10	233542	A00080329	问问 A 市经开区东六线以西泉塘昌和商业中心以南的有关规划

通过对上面两种聚类结果的比较，发现聚类结果得出的前 10 热点问题大部分相同，但聚类结果存在本类留言聚类不全，和同一问题留言被归为两类的情况，因此对两种结果进行分析总结，得出 10 个热度排名前列的问题，如下表所示：

表 4.4 整理后热度前 10 分类表（部分）

问题 ID	留言编号	留言用户	留言主题
1	208636	A00077171	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
2	223297	A00087522	反映 A 市金毛湾配套入学的问题
3	218132	A000106090	再次请求过问 A 市 58 车贷案件进展情况
4	191951	A00041448	A4 区绿地海外滩小区距渝长厦高铁太近了
5	193091	A00097965	A 市富绿物业丽发新城强行断业主家水
6	234885	A00060375	A6 区月亮岛路 11 万伏高压线没用地埋方式铺设
7	284571	A00074795	建议西地省尽快外迁京港澳高速城区段至远郊
8	239670	A00080329	问问 A 市经开区东六线以西泉塘昌和商业中心以南的有关规划
9	200667	A00079480	请问 A 市为什么要把和包支付作为任务而不让市场正当竞争？
10	196282	A00039390	反映 A 市地铁 3 号线松雅西地省站地下通道建设问题

以上表中的 10 条各类留言为中心，通过计算留言余弦相似度的方法，将其他相似度达到阈值的留言进行归类。

余弦相似度公式：

$$\cos(x_1, x_2) = (x_1 \cdot x_2) / \|x_1\| \times \|x_2\| \quad (4\text{-}3)$$

式中，x_1，x_1为要计算相似度的两个向量，$(x_1 \cdot x_2)$表示两向量的内积，$\|x_1\|$表示向量的范数。

重新计算 10 个问题的热度，然后进行排序，取热度前 5 的 5 个问题，得出 5 大热点问题明细表，如下表所示：

表 4.5 热点问题明细表（部分）

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	2113	2019/05/05 至 2019/09/19	A 市 A5 区汇金路五矿万境 K9 县	房屋质量安全存在隐患
2	2	1759	2019/03/12 至 2019/04/11	A 市梅溪湖	买房承诺的子女配套入学未实现
3	3	817	2019/01/11 至 2019/07/08	西地省 A 市 A4 区 p2p 公司	58 车贷案立案后久未解决
4	4	698	2019/08/23 至 2019/09/06	A4 区绿地海外滩小区	小区距高铁太近，噪音影响严重
5	5	243	2019/06/19 至 2019/06/19	A 市富绿物业丽发新城强	丽发新城强行断业主家水

从表中可以看出 5 个热点问题相关留言的聚类效果较好，相关留言进本包括进了相应的类别，仅有第 3 个问题混入了一条其他类别的留言。

剔除误分类留言，计算 5 个问题的热度，并进行整理，得到热点问题表，如下表所示：

表 4.6　5 大热点问题表

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	2113	2019/05/05 至 2019/09/19	A 市 A5 区汇金路五矿万境 K9 县	房屋质量安全存在隐患
2	2	1759	2019/03/12 至 2019/04/11	A 市梅溪湖	买房承诺的子女配套入学未实现
3	3	817	2019/01/11 至 2019/07/08	西地省 A 市 A4 区 p2p 公司	58 车贷案立案后久未解决
4	4	698	2019/08/23 至 2019/09/06	A4 区绿地海外滩小区	小区距高铁太近，噪音影响严重
5	5	243	2019/06/19 至 2019/06/19	A 市富绿物业丽发新城强	丽发新城强行断业主家水

4.3.3 热点问题词云图

根据热点问题明细表，在完成分词等一系列操作后，按照词频降频排列，利用 WordCloud 库可以画出热点问题的词云图，热点问题 1 和问题 2 的词云图如图 4.6、4.7 所示。

图 4.6 热点问题 1 词云图

图 4.7 热点问题 2 词云图

由词云图 4.6 可以看出词频较高为 K9 县、五矿、万境、房屋、质量等，反应了 K9 县的房屋质量存在着一定的问题。由词云图 4.7 可以看出词频较高为金毛、配套、业主、入学等，反应了金毛湾学生配套入学问题。

5. 问题三分析及求解

5.1 问题三分析

问题三是对附件 4 相关部门对留言答复的评价，观察表中数据，主要有留言编号、留言用户、留言主题、留言时间、留言详情、答复意见和答复时间。我们将留言时间进行分列，保证年月日以及具体时间分开，便于编程分析答复效率。根据答复内容“回复”、“问题”、“转交”、“移交”、“超出管辖范围”、“请致电”、“请咨询”等词提取出解决、未解决、解决中三类答复意见。利用留言内容、答复意见的相似度以及人工查找的方式判定解决效果如何。有效性、完整性根据留言与答复意见的时间、地点、人物、对象等是否一致，是否有解决方式，解决方式理想与否，判断答复意见的满意度。

5.2 公众诉求分析

从留言主题、留言详情数据分析统计得表 5.1，可以看出公众直接将问题、建议或咨询反应给书记和市长的人数较少，主要还是以统称领导或政府向政府反应有关情况，大多数人还是习惯直接提出自己的意见和建议，希望留言部门能够直接给予解决，但可能会受管辖范围等限制，得不到预期的回应。公众对书记信箱或市长信箱的选择可能受回复率的影响偏低，或者直接以领导称呼，而不是以书记或市长，导致统计出现偏差。

表 5.1 公众诉求反应的邮箱偏好

	书记	市长	领导	政府	总计
计数	395	83	871	902	2816
百分比	14.03%	2.95%	30.93%	32.03%	100.00%

在诉求类型上，投诉类最多，建议和咨询相对较少，其中，投诉类问题占比超过二分之一（55.68%），其余为建议类（17.97%）、咨询类（17.58%）以及其他（8.77%），如图 5.1 所示。

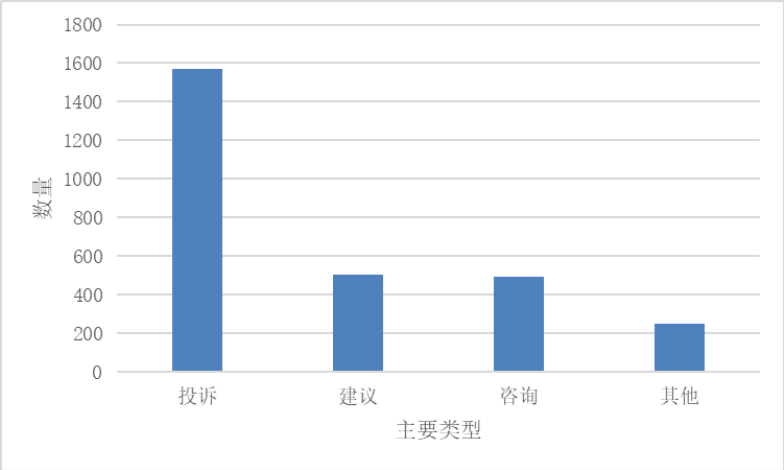


图 5.1 公众诉求类型分布

5.3 答复意见质量评估

5.3.1 评价体系建立

1）答复效率

用层次分析法对留言的答复时间按照一定的时间间隔进行分层，根据答复时间的分布来评价答复的效率，以及答复的优先级。

2）答复意见相关型、完整性

主要从相关性、完整性等对其进行评价。相关性主要针对答复意见与留言内容是否相关，答复意见是否与主题相关，排除留言与主题不相关的情况。评价相不相关主要从答复意见中的时间、地点、人物、对象等因素是否与留言能对应，可以将时间、地点等赋予一个权重，采用熵值法进行权重赋值。最后根据对应项的情况给予该项答复意见分数，并按照分数的高低排序，评价标准可以采用均方差。关于答复意见的完整性也可以从时间、地点、人物、对象以及解决的情况出发，采用打分的方式进行评价。

3）解决情况

从事实回应和价值回应两个层面，针对答复意见的相关性、完整性、效率性、可解释性等，给予答复意见解决情况评价指标体系。对于答复意见，从事实回应层面出发，将“有答复意见”和“无答复意见”两种情形。并将答复意见分为公众诉求问题得以“解决”、“解决种”、“未解决”三类。如果留言被上报的问题有确定的结果、或投诉与求助的问题得以消除或缓解，则认为留言的问题被“解决”。如果反应的问题还在调查、处理中或已经报送给相关部门等，则认为“解决中”。在其他情况下，则认为“未解决”。最后又根据解决的具体情况进行细化分类。针对答复意见的结果，从“解决”、“解决中”以及“未解决”的数量，来评价答复意见的解决率。

表 5.2 为对相关部门答复情况的进行变量赋值：

表 5.2 答复相关变量及赋值

变量	变量赋值
回应与否	0=未回应，1=回应
解决与否	11=解决，12=解决中，13=未解决
解决方式	111=按留言期望解决，112=按合理方式解决，113=协商方式解决
未解决类别	131=不能解决，132=解决困难，133=给予补偿，134=给予解决新方式
未解决原因	131=不合理，132=程序问题，133=超出管辖范围，134=其他

图 5.2 为对政府回应的情形及逻辑关系：

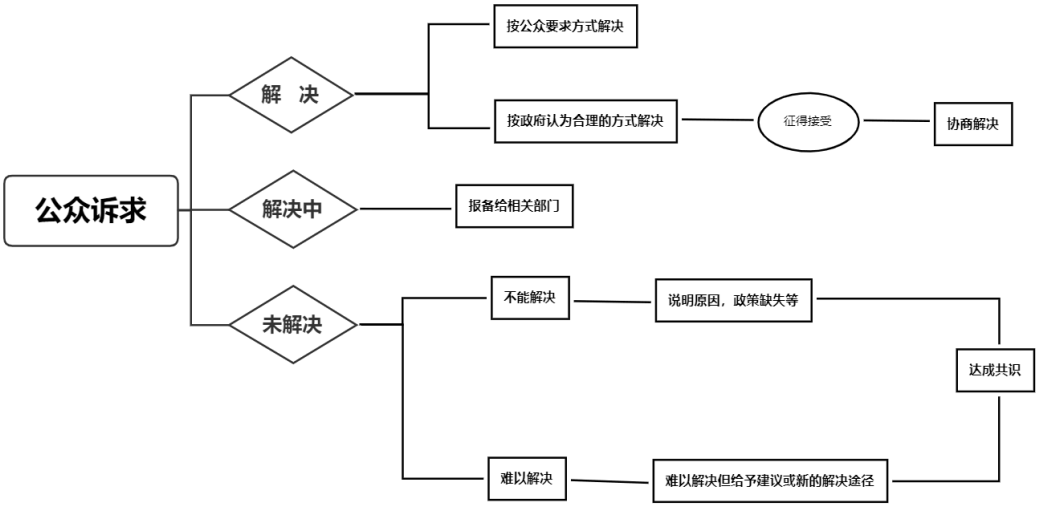


图 5.2 政府回应情形逻辑

5.3.2 评估举例

1）答复效率

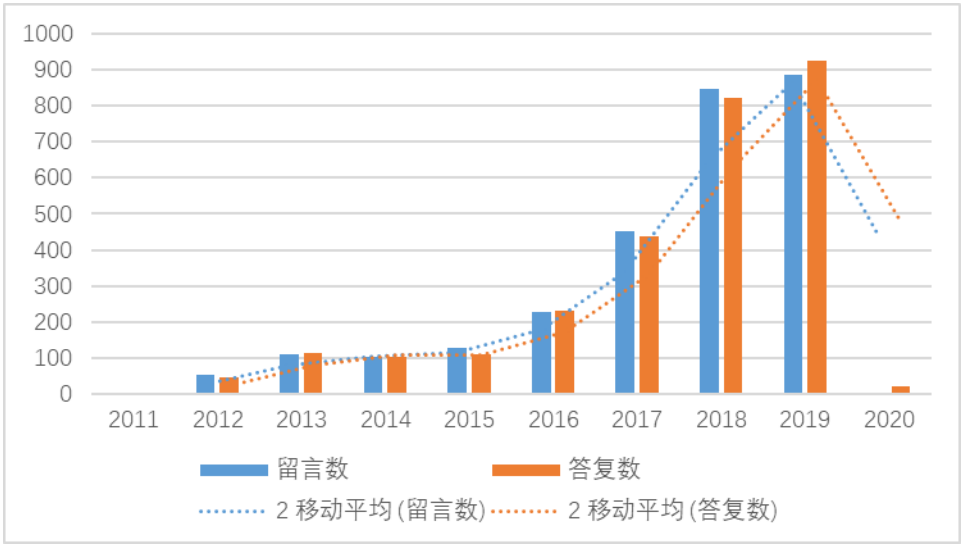


图 5.3 公众留言与政府答复数量的对比

整体来看，随着网络的发展，公众的留言数在逐年上升，在 2020 年出现了拐点。2020 年的拐点主要受两方面的影响，一是 2020 年受疫情的影响，新年之初，外出的人员较少，所以遇到的问题也就较少。二是在进行数据爬取时，爬取的时间范围比较窄。图中显示，每年公众的留言数大致与政府的回复数相等，局部出现留言数多余答复数或答复数多余留言数，前者受政府答复效率的影响，后者受未答复留言存量的影响。

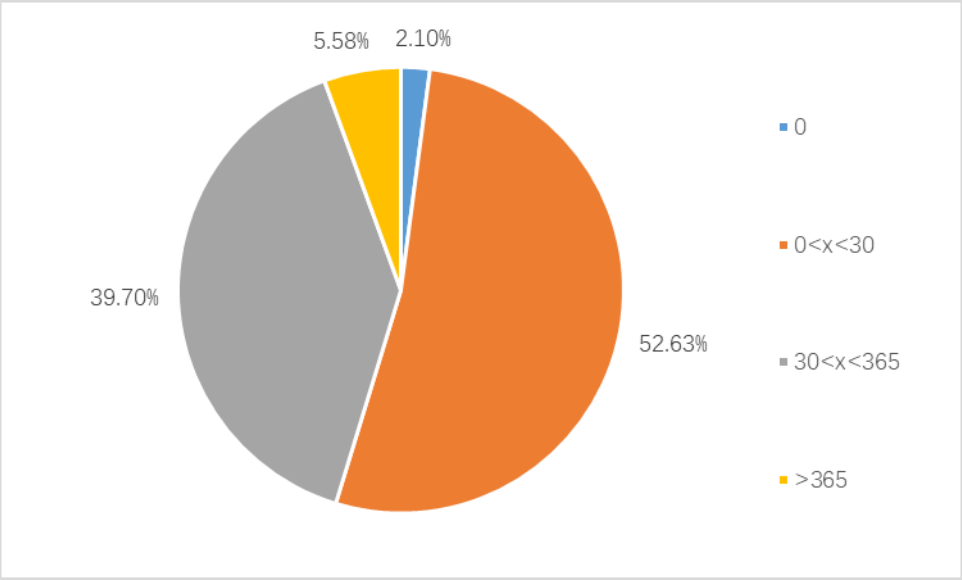


图 5.4 不同间隔时间答复数所占比例

从上图显示来看，政府在留言的当天完成答复的量仅占所有留言的 2%。政府答复时间在一月之内占据了总留言数的 52.63%，但不代表政府的答复效率比较高，反而有待提高，因为从图中还可以得悉，在所有留言中，超过一个月答复的量尽占据了总留言数的 39.7%。

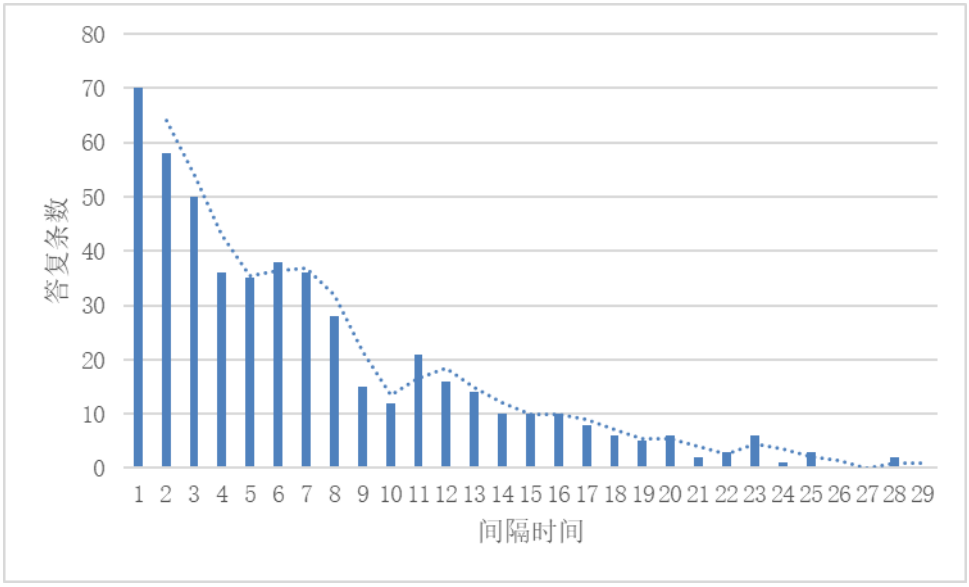


图 5.5 2019 年一个月内回复效率

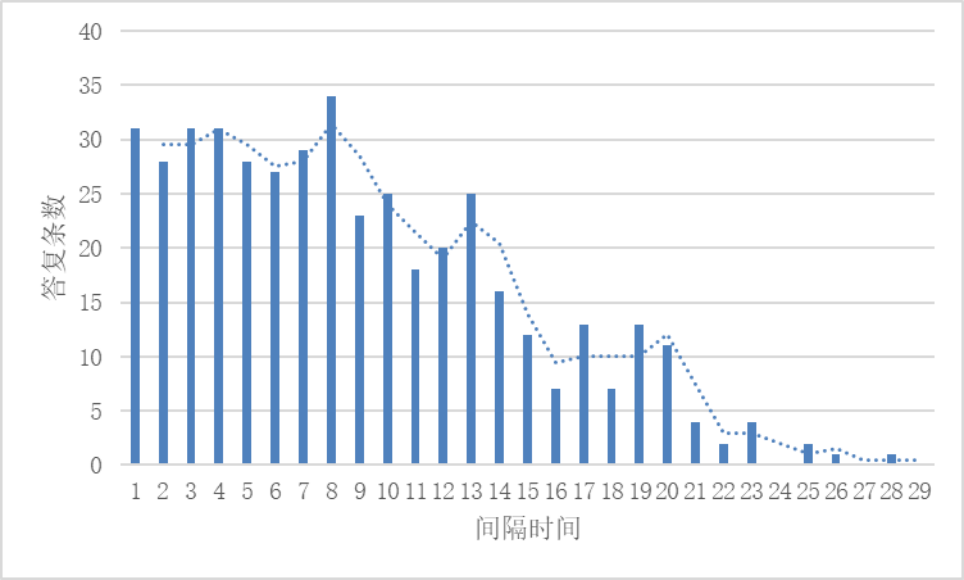


图 5.6　2018 年一个月内回复效率

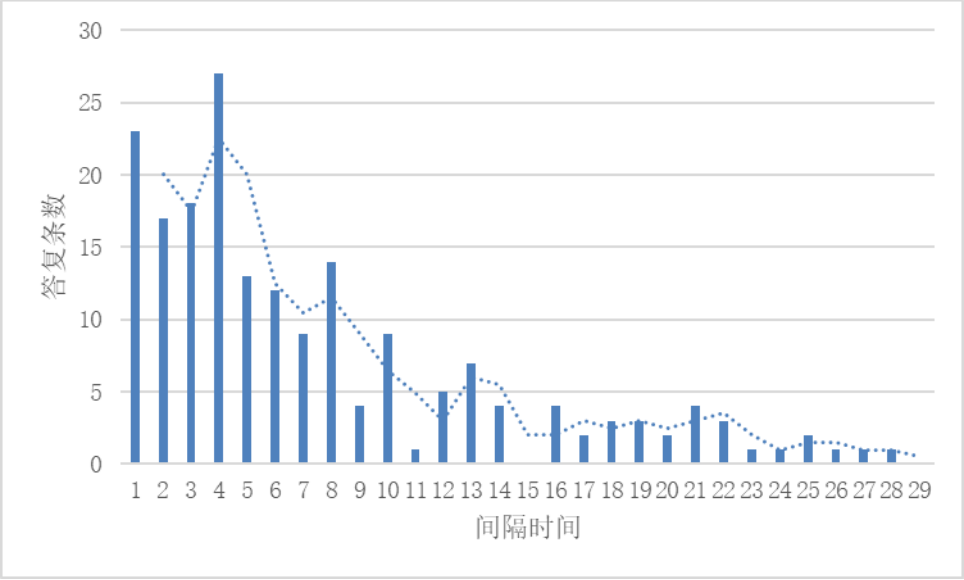


图 5.7 2017 年一个月内回复效率

对比 2017，2018，2019 年在一个月回复情况可知，政府的答复数量随时间的增长，总体是下降的，大多分布在 15 天内，考虑到处理流程所耗费的时间，说明政府的回复效率较为客观。

2）答复意见相关型、完整性

留言：**2019 年 4 月以来**，位于 A 市 **A2 区桂花坪街道的 A2 区公安分局宿舍区**（**景蓉华苑**)出现了一番乱象，该小区的**物业公司美顺物业扬言要退出小区**，因为小区水电改造造成物业公司的高昂水电费收取不了(原水电在小区买，水 4.23 一吨，电 0.64 一度）所以**要通过征收小区停车费增加收入**，小区业委会不知处于何种理由对该物业公司一再挽留，而对业主提出的新应聘的物业公司却以交 20 万保证金，不能提高收费的苛刻条件拒之门外，业委会在未召开全体业主大会的情况下，制定了一高昂收费方案要各业主投票，而投票不采用投票箱只制定表格要物业公司人员这一利害关系机构负责组织，对投票业主隐私权没有任何

保护，还对投反对票的业主以领导做工作等方式要求改变为同意票，这种投票何来公平公正公开，面对公安干警采用这种方式投票合法性在哪？

政府：现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2 区景蓉花苑物业管理有问题”的调查核实情况向该网友答复如下：您好，首先感谢您对我们工作的信任和支持，关于您在平台栏目给胡华衡书记留言，反映“**A2 区景蓉华苑**物业管理有问题”的情况已收悉。现将我们调查处理情况答复如下：经调查了解，针对来信所反映的“**小区停车收费问题**”，景蓉华苑业委会于 **2019 年 4 月 10 日至 4 月 27 日**以“意见收集方式”召开了业主大会，经业委会统计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真疏理归纳并进行了反馈，业委会制定的停车收费标准不高于周边小区价格。针对来信所反映的“**物业公司去留问题**”，5 月 5 日下午,辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议。在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局**已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序**。再次感谢您对我区工作的理解和关心。2019 年 5 月 9 日

留言与答复中所涉及的时间、地点、人物、对象以及解决方式如下：

时间：2019 年 4 月

地点：A2 区景蓉华苑

人物：居民、物业公司

对象：停车收费问题

解决方式：已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序。

3）解决情况

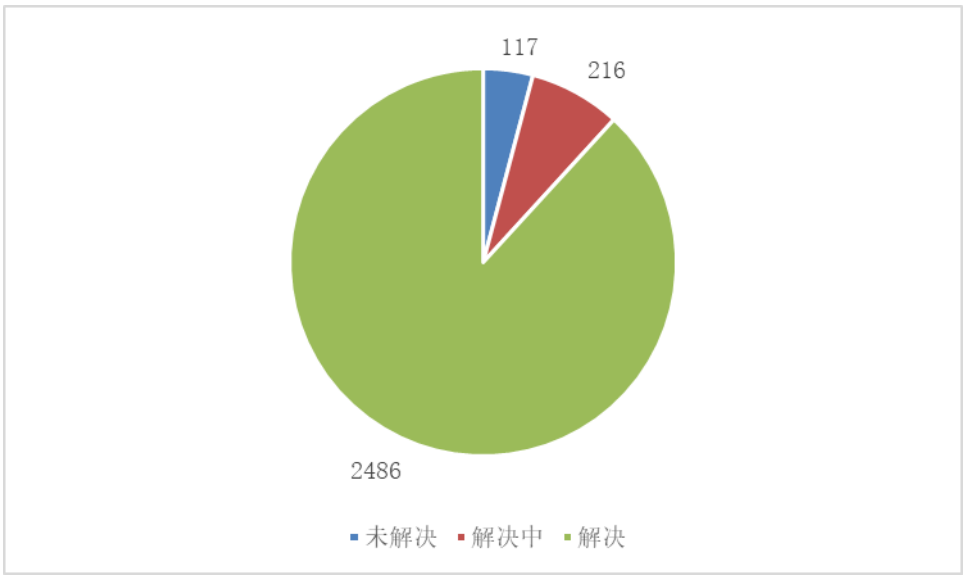


图 5.8 解决情况比例图

以相似度 0.12 为阀值，按照留言与答复的相似度来判定留言是否被解决，得到的解决结果为 2432，与上图结果大致相似。

①解决示例

留言：德雅路国防科大南门外，德雅路本来只有约 12M 人行道开了夜宵摊，

晚上特别是半晚吵闹令人无法入睡，为何清水塘城管给批同意开夜宵摊，道路本来就不宽，路边停车也严重影响交通，为何不让开到离这不远的烈士公园西门了？夜宵摊开在这里因德雅路两边均是房子，形成一个音谷，特别影响晚上休息。路二边的人特别是老年人希有关部门尽快处理。

政府：网友“UU008270”，您好！您的留言已收悉。现将有关情况回复如下：您所反映的夜宵摊位于德雅路与东风路交叉口至烈士公园北门路段，经营时段从晚上 7 点半至凌晨 2 点。为规范管理，街道和城管部门严格按照《A 市城市管理条例》和《A 市城市容貌规定》制定并张贴了规范点管理制度，且每天均派有专人定期督促经营户严格落实“三有”（有防污措施、有垃圾收集装置、有专人负责卫生）；“四不”（不影响市容、不阻碍交通、不损坏公共设施、不扰民）；“五统一”（统一经营地点、统一经营时间、统一经营范围、统一经营设施、统一标示标牌）。关于您建议将夜市规范点搬到烈士公园西门的问题，因烈士公园西门的东风路是 A 市的主要干道，不便将其作为夜市规范点。下一步，我区将加大执法力度，进一步加强对德雅路夜市规范点的管理。城管、街道等部门也将于近期组织所有经营户进行一次培训，重点学习城管相关条例法规，切实提高经营户环保意识。感谢您对我们工作的支持、理解与监督！

以上为按合理解决方式解决问题。政府答复首先从确认问题出发，然后对公众进行普法，最后通过相关原因阐述留言建议的不合理性，并给予了合理的解决办法。

②解决中示例

留言：我是居住在绿洲大道转盘处的长铺镇居民，每天早上在东方红超市门口等去老街那边的公交车去上班，短短的的两三公里路程，要坐上半个小时甚至更久才能到达，因为每次都要挨着冻在街边等上十多分钟才等得到一路去老街方向的公交车，好不容易等到公交车了，在老桥头处的红绿灯等绿灯又要等上五六分钟，因为这里的绿灯通行时间只有十几秒，而红灯时间却有六七十秒！每次都积压了大量的车辆等红灯，从而极大的影响了通行效率！去那个方向的的士也都不想去，因为都怕堵，以上种种都影响了百姓的出行便利！在这里建议公交管理部门应加大投入绿洲大道往返老街方向的公交班次，至少在每天的早、中、晚上下班和上下学高峰期间要多投入公交班次；第二建议交警部门将老桥头的红绿灯通行时间设置长一点，从而提高通行效率。以上两个问题希望能得到相关部门的立马解决！

政府：您好！您所反映的问题，我办已转交通、交警等相关部门。感谢您对我县工作的关注，欢迎您和各位网友继续进行监督！E9 县网信办。

③未解决示例

留言：敬爱的领导：我是 B2 区江南世家的高层住户。自去年（2018 年年底）自来水公司来本小区进行水改后，小区多数高层业主用水不正常，特别是晚上用水高峰期。现在天气慢慢变热，用水量会更大，为了保障自己的正常生活，很多业主，都打电话到自来水公司和 12345，但是都没有帮组解决这个关于民生的用水问题。自来水公司在水改工程没有交付工程质保金，所以我们不能制约自来水的施工方，每次电话投诉后，会有人员上来看看，但是只是看看，根本没有解决问题。基本上门检测人员都是在用水非高峰期上门。致使我们老百姓投诉无门，根本没有办法解决问题。请求有关领导帮组解决一下这个关乎正常生活的基本问题。此致敬礼！

政府：网友：您好！您反映的问题已收悉，因小区水改不属于 B2 区管辖权

限范围内，建议您向 B 市自来水公司反馈，联系电话:×××。感谢您对我区工作的支持和理解，祝您生活和工作愉快。

上述未解决留言案列中，政府相关部门说明了不能解决的原因（超出管辖范围），并给予了留言群众新的解决途径。

5.4 结论及建议

答复意见的本质是政府对群众留言做出反应的过程和结果，政府的答复率与群众留言时间、留言问题的复杂度、网站水平有很大关系，根据分析结果可发现如下结论：

第一，随着信息化、网络化的发展，公众向上反应问题的意识逐渐增强，网络留言条数随着时间呈现逐年增长趋势，到达某一年份可能会趋于平稳。网络平台下，公众参与问政热情更高，反应的问题涉及更广，但是鉴于群众对职能部门的管辖范围不清，可能出现反应对象错误，导致反应问题不能得到解决、或解决时间较长。

第二，网络下，公众可能会面临“无答复”的问题，所反应的问题不能得到有效的回应解决，一方面是由于处理难度可能超出部门管辖范围，或处理相当困难，需要更长的处理时间、另一方面则是政府没有相关的机制，保证网络问政，有问必答，导致针对一部分问题，政府工作人员“选择性回答”。

第三，网络留言时间不集中，可能导致相关留言被忽视，或者留言内容频率不高被忽视。对于公众留言，集中时间留言可能会增加答复效率，或者多人反应同一问题，也能增加答复效率。

第四，网络留言中可能出现答非所问，或敷衍作答的情况，这种情况不便于筛查，这由系统缺陷所造成，因此提高平台系统建设也是促进网络问政的一个重要因素。

对此，针对网络问证中答复问题，本文给予以下几点建议：

第一，政府应该加强网络平台建设，首先在平台首页应该给予公众各部门职能分工细则，以防群众反应问题对象错误，导致问题迟迟得不到解决。

第二，政府应该建立一套机制，保证群众留言，有问必答，且制定一定的间隔时间，在间隔时间内必须给予留言者一定的解决方式回应，不能避而不答，也不能答非所问。或者安排专人，管理留言系统。

第三，根据留言量和答复量的分布来看，两者大致相等，但回复时间存在一定差异。平台应规定留言与答复的间隔周期，集中留言，集中答复，以便政府有足够的时间处理，能给予留言群众更充分、更有价值的回应，而不是仅从职能范围评估问题可解决与否。

第四，平台系统应优化，保证时间、地点、人物、对象或事件、是否解决、解决方式、不能解决原因等要素能分开统计，便于答复质量的分析。

参考文献：

[1] Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. NAACL 2018.
[2] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.
[3] GREFF K, SRIVASTAVA R K, KOUTNIK J et al. LSTM: a search space odyssey[J]. IEEE Transactions on Neural Network and Learning Systems, 2015,28(10):2222-2232.
[4] 李航. 统计学习方法[M].北京:清华大学出版社, 2012.
[5] 周志华. 机器学习[M].北京:清华大学出版社, 2016.
[6] 祁小军, 兰海翔, 卢涵宇等. 贝叶斯、KNN 和 SVM 算法在新闻文本分类中的对比研究. 电脑知识与技术, 2019(25):226-228.
[7] 邬明强, 邬家明, 辛伟彬. Word2vec+LSTM 多类别情感分类算法优化.计算机系统应用, 2020, 29(1):130-136.
[8] 陈楠, 陈进才, 卢萍. 基于深度学习的多元文本情感研究与分析.计算机科学与应用, 2018, 8(5):669-686.
[9] 李杰, 李欢. 基于深度学习的短文本评论产品特征提取及情感分类研究. 情报理论与实践, 2018, 41(2):143-148.
[10] 熊涛. 基于长短时记忆网络的多标签文本分类[D].杭州:浙江大学, 2017
[11] 杨俊峰, 尹光花.基于 Word2vec 和 CNN 的短文本聚类研究.信息与电脑(理论版).2019(24):24-26.
[12] 何跃, 帅马恋, 冯韵. 中文微博热点问题挖掘研究[J].统计与信息论坛, 2014 29(6): 86-89.
[13] 陈巧红, 孙超红, 贾宇波. 文本数据观点挖掘技术综述[J].工业控制计算机, 2017, 30(2):94-95.
[14] 马龙军.基层政府网络回应机制创新研究——以 Z 县政府网站 2012-2018 年留言回复为例.未来与发展.2018(11):61-66.
[15] 邵梓捷, 杨良伟.“钟摆式回应”:回应性不足的一种解释——基于 S 市地方领导留言板的实证研究.经济社会体制比较.2020(01):118-126.