基于 BP 神经网络分类、K-means 聚类及文本相似度的群众留言文本分析

摘　　要

民意内涵民智，民意关乎民生；民意力量的发挥取决于民意征集的广度和尊重民意的程度。近年来，随着网络在中国的普及率的提高，各种网络平台，如微信、微博、市长信箱、阳光热线等，也逐步成为政府了解民意、 汇聚民智、凝聚民气的重要渠道。这些渠道的增加使得相关的社情民意的文本量也增加了许多，这令政府对民意的了解也越来越多，但靠人工来进行留言划分和热点整理的相关部门的工作同时也面临着极大的挑战。网络在面向大众的同时，也在不停地发展。如今，大数据、云计算、人工智能等技术正蓬勃发展，为提升政府的管理水平和施政效率，建立基于自然语言处理技术的智慧政务系统已是个为社会治理创新发展的新趋势了。本文将对互联网公开来源的群众问政留言记录及相关部门对部分群众留言的答复意见进行挖掘与分析。

本文在本次数据挖掘过程中，首先对获取得到的评论数据利用 python 进行数据预处理，使用 jieba 分词对文本进行整理，并使用 TF-IDF 加权模型对词分配权重，实现了对评论数据的优化，提升了数据的有效程度。

接着，对提取出的数据分别建立高斯朴素贝叶斯分类和 BP 神经网络模型，对文本中的群众留言进行分类，使用 F1-score 对两种模型进行评价；高斯朴素贝叶斯分类的 F1-score 约为 68.15，而 BP 神经网络模型的 F1-score 约为 89.96。

然后，使用 k-means 聚类对留言主题进行聚类，提取留言中的热点问题。并对挖掘出的热点问题进行分析，同时提取其中关键词与时间跨度进行观察与剖析。

最后，通过文本相似度分析，建立一定的评价模型，对答复意见进行评分，评分最高达到了 86 分，最低的仅 0 分；并对得到的评分进行了分析，评分大致符合正态分布，评价模型较为贴近实际。

关键词：群众留言；jieba 分词；BP 神经网络；k-means 聚类；余弦相似度

Abstract

Public opinion implies public wisdom, and public opinion is related to people's livelihood; the exertion of public opinion power depends on the breadth of public opinion collection and the degree of respect for public opinion. In recent years, with the increasing popularity of the Internet in China, various network platforms, such as WeChat, Weibo, mayor's mailbox, and Sunshine Hotline, have gradually become important channels for the government to understand public opinion, gather people's wisdom, and gather people's popularity. The increase in these channels has increased the amount of texts related to social conditions and public opinion, which has made the government more and more aware of public opinion. However, the work of relevant departments that rely on manual to divide messages and organize hotspots also faces extreme difficulties. Big challenge. While facing the public, the Internet is constantly developing. Today, big data, cloud computing, artificial intelligence and other technologies are booming. In order to improve the government's management level and governance efficiency, the establishment of a smart government system based on natural language processing technology is a new trend for the development of social governance innovation. This article will excavate and analyze the records of the public's questioning messages from the Internet and related departments' responses to some of the people's messages.

In this data mining process, this article first uses python to preprocess the obtained comment data, use jieba word segmentation to organize the text, and use the TF-IDF weighting model to assign weights to the words to achieve the optimization of the comment data. To increase the effectiveness of the data.

Then, the Gaussian Naive Bayes classification and BP neural network model were established for the extracted data, the mass messages in the text were classified, and the two models were evaluated using F1-score; F1- of the Gaussian Naive Bayes classification The score is about 68.15, and the F1-score of the BP neural network model is about 89.96.

Then, use k-means clustering to cluster the message topics to extract the hot issues in the message. It also analyzes the hot issues excavated, and extracts keywords and time span for observation and analysis.

Finally, through text similarity analysis, a certain evaluation model was established, and the comments were scored. The highest score reached 86 points, and the lowest was only 0 points. The scores were analyzed and the scores were roughly in accordance with the normal distribution. The model is closer to reality.

Keywords: mass message; jieba word segmentation; BP neural network; k-means clustering; cosine similarity

目录

第一章 绪论

1.1 研究背景及意义

过去，政府了解民意的途径是各级人大和政协、信访部门及新闻媒体，在互联网高速发展的今天，政府想要深入了解民意，方法也变得更多样化更方便了。如今，众多网络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。民众仅需要登录这些平台进行简单的操作，就能反映自己遇到的问题了。这使得相关的文本量变得日渐庞大，此时，人工操作便显得有些无力了。

本文的目的是通过文本分析，对群众留言进行分类、聚类，并对留言答复进行评分。使得群众留言在人工处理中更加便捷，减少人工消耗从而提高对群众留言的处理效率。

1.2 研究内容

本文主要对近几年收集到的群众留言进行深入研究，分析对比已有数据后，本文进行了如下研究：

（1）群众留言分类

分析并研究给出的数据，并建立一级标签分类模型，将群众留言分类并评测模型。本文分别建立高斯朴素贝叶斯分类模型与 BP 神经网络模型对群众留言进行分类，并使用 F1-score 对模型进行评价。最终选择 BP 神经网络模型进行分类。

（2）热点问题挖掘

通过 k-means 聚类，对群众留言进行聚类，从中挖掘出群众反映较多且受群众关注的问题，并给出热点指数。最终确定排行前五的热点问题，并分析这些问题的关键关注点。

（3）答复意见的评价

针对工作人员给群众的答复意见，本文给出了评价指标，并利用文本相似性分析，给每一条答复打上了分数。并对评价模型进行了分析，确定评价是否客观。

第二章 模型假设

本文为进行文本分析提出了如下假设：

1.群众留言分类：

（1）假设留言主题为分类的主要数据。

（2）假设留言时间，留言用户，留言编号对留言分类的问题模型影响较小。
（3）假设不存在有人为组织故意刷评论的情况。

2.热点问题挖掘：
（1）热点问题主要影响因素是留言主题与留言详情，假设评论点赞数一定程度也表明了问题的热点。
（2）假设一个用户只有一个 IP。

3.答复意见评价：
（1）假设可根据留言详情与留言答复的词频，文本对比相关性、完整性等。
（2）假设留言用户，编号等对分析影响较小。

第三章 数据预处理

附件 1 是留言分级标签表，附件 2 是留言详情及分类表，附件 3 是留言详情及点赞表，附件 4 是留言详情及答复表。

群众留言由汉字构成，中文博大精深，但并不是每个字都能具有鲜明的代表性，单看“的”、“我”等字并不能知道这一句话说的是什么，如果只是简单的按照每个汉字进行文本分析，显然不能获取准确内容。故在构建模型时，本文首先进行了如下数据与处理。

3.1 jieba 分词

众所周知，中文文本与英语是有差距的，必须以词语为单位才可以获取文中的正确意思。所以在本文中，我们对文本进行了 jieba 分词。jieba 分词算法主要从以下方面展开。

1.基于 Trie 树结构实现高效的词图扫描，生成每一句中汉字所有可能成词情况的有向无环图（DAG）。

假如在群众留言中存在一句话“某学院强制实习”，本文对照已存在的 Trie 树对这句话生成了一个有向无环图，其路径如图所示：

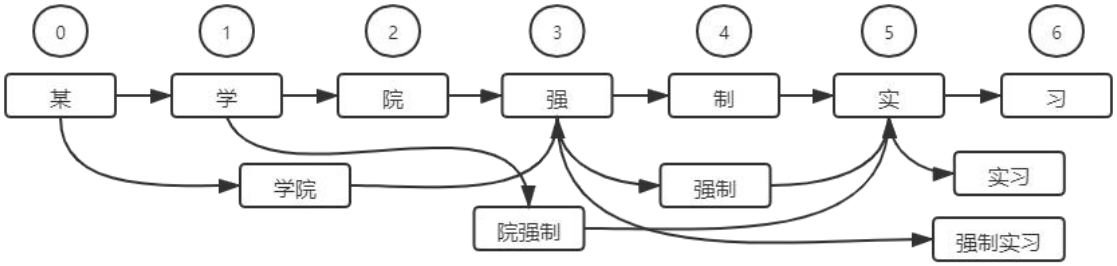

图 3.1 DAG 路径图

可能的路径有：

a.某/学/院/强/制/实/习

b.某/学院/强/制/实/习

c.某/学院/强制/实/习

d.某/学/院强制/实习

. . . . . .

2. 利用动态规划的方法查找出最大概率路径，以找出基于词频的最大切分组合。

查找未分词句子中已切分好的词语，并查找该词语出现的频率，若没有该词，就把词典中出现频率最小的那个词语的频率作为该词的频率，也就是说 P(某词语) = FREQ.get(‘某词语’,min_freq)根据动态规划查找最大概率路径的方法，对句子从右往左反向计算最大概率，即

$$P(NodeN) = 1.0 \quad (1)$$

$$P(NodeN-1) = P(NodeN) * Max(P(last))\ldots \quad (2)$$

依次类推，最后得到最大概率路径，并得到最大概率的切分组合。

3. 对于未登录词，采用基于汉字成词能力的 HMM 模型，并使用 Viterbi 算法进行处理。

未登陆词的意思是在 jieba 源码中自带的 dict.txt 里面没有被记录的词。

当我们把 jieba 分词源码中的 dict.txt 中的词语全部删除时，jieba 依然能够进行分词，但分出来的词长度多为 2，这是因为这时使用的就是 HMM 来进行分词了。

将中文的词汇按照 B（begin）、E（end）、M（middle）、S（singgle，单独成词的位置）排成一个序列，jieba 中就是以这种形式来标记中文的词语。

比如说，在群众留言中存在“路口至加油站路段人行道”，“路口”这个词语可以被标记为 BE，也就是路/B 口/E。复杂一些的词语“加油站”会被标记为 BME，即：“开始”，“中间”，“结束”。在 jieba 中对语料进行了训练从而得到了三个概率表，再加上 viterbi 算法便得到了一个概率最大的 BEMS 序列， 按照 B 打头，E 结尾的方式，对待分词的句子重新组合，就得到了分词结果。如下图所示：

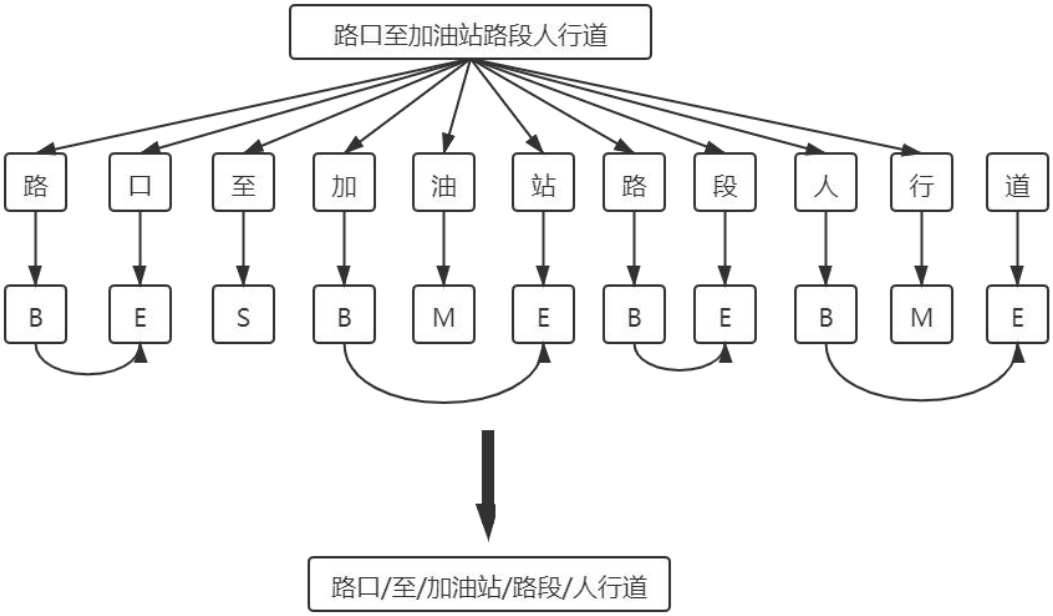

图 3.2 BEMS 序列图

通过上图可以看出，“路口至加油站路段人行道”这句话可以得到一个 BEMS 序列 ：[B,E,S,B,M,E,B,E,B,M,E]，将序列中的 BE 与 BME 组合得到一个词，S 独立出来，就可以得到分词结果：“路口/BE 至/S 加油站/BME 路段/BE 人行道/BME”，这样就得到了一个较为准确的分词结果。

为了提高模型的准确度，本文将留言详情用 jieba 分词把其中的句子切分成词语，并建立停用词表，将文本中的一些助词、语气词及一些标点符号删掉。经过分词后，得到了表 3.1 留言主题表及表 3.2 留言详情表。

表 3.1 留言主题表

序号	留言主题（分词）
1	A 市 经济学院 体育学院 变相 强制 实习
2	在 A 市 人才 app 上 申请 购房 补贴 为什么 通不过
3	希望 西地省 把 抗癌 药品 纳入 医保 范围
4	A5 区 劳动 东路 魅力之城 小区 临街 门面 烧烤 夜宵 摊
5	请 给 K3 县 乡村 医生 发 卫生室 执业 许可证
6	A 市 能否 设立 南塘 城轨 公交站
7	请求 A 市 地铁 2 线 在 梅 溪湖 CBD 处 增设 一个 站
8	请问 A 市 什么 时候 能 普及 5G 网络
.....	.....

表 3.2　留言详情表

序号	留言详情（分词）
1	区 大道 西行 便道 未管 所 路口 至 加油站 路段 人行道
2	位于 书院 路 主干道 大厦 一楼 至 四楼 人为 拆除 水电 设施
3	小区 高层 为 二次 供水 楼顶 水箱 长年 不洗 现在 自来水 龙头
4	领导 区苑 小区 位于 区 小区 物业 市程明 物业管理 有限公司
……	……

3.1.1 词云图

词云图的作用是更直观地看出每一类标签的关键词，在图中越显著的词，其在文本中出现的频率就越高。

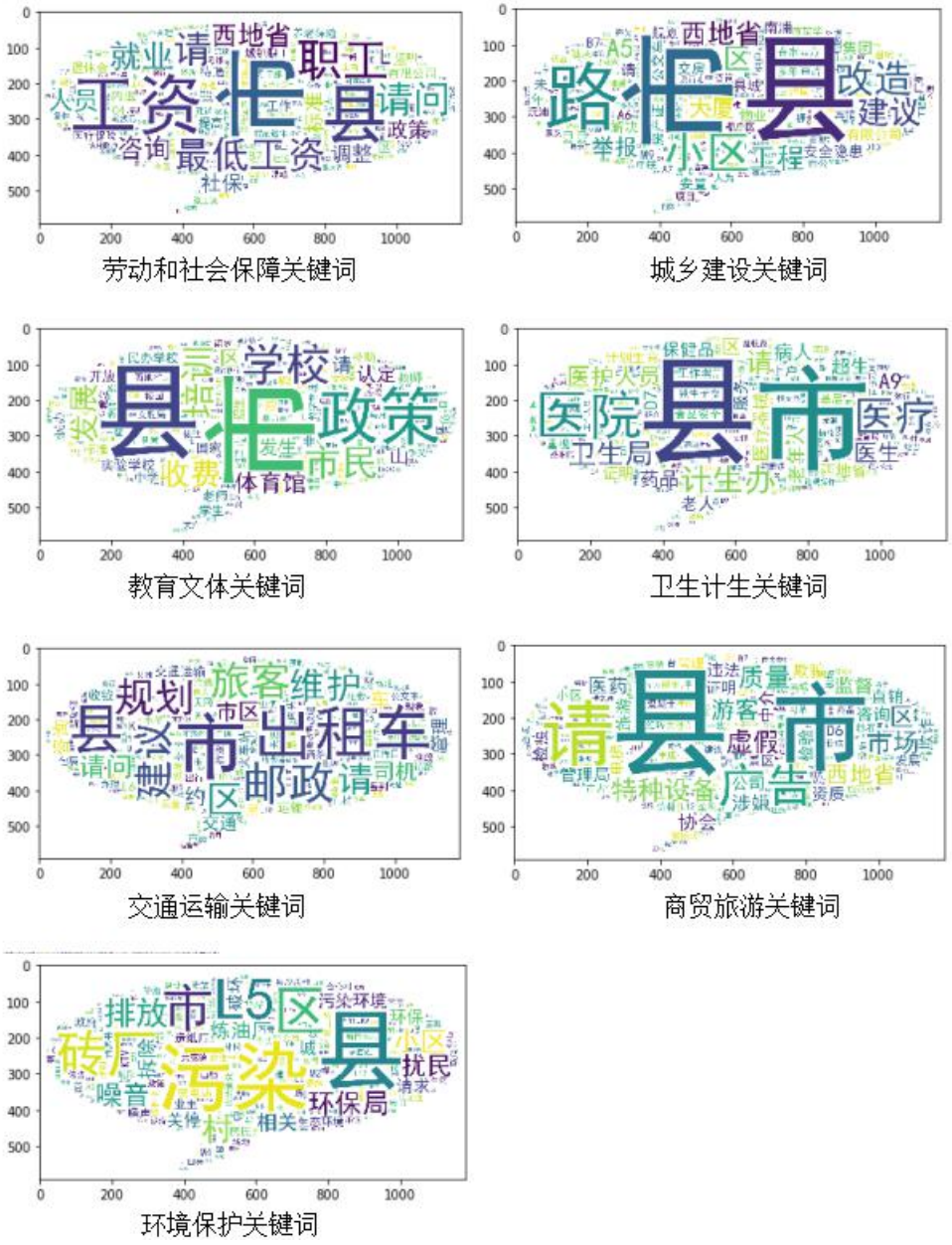

图 3.3 各类留言词云图

由图可知，像“市”“县”这样的关键词最多，这说明以上问题都是伴随着城市化发展过程产生的。

除此外，本文根据示例数据的几种问题类别，具体分析了大众反应的主要问题：

在劳动与社会保障方面，我们可以发现，反复提到的是“工资、就业，职工、最低工资，社保、政策”。

城乡建设方面主要提到的是“改造、小区、安全隐患、建议”。

教育文体方面主要反应的是“学校、政策、收费、培训、老师、学生”等关键词。

卫生计生方面主要关键词是“医院、医疗、卫生局、医生、病人、超生、老人、计生办、保健品”。

交通运输方面主要是“出租车、邮政、旅客、规划、司机、市区”等。

商贸旅游方面主要是“广告、虚假、游客、质量、违法”等方面。

环境保护方面主要是“砖厂、污染、环保局、扰民、噪音、排放”等。

我们可以根据词云图以及词频对群众反映的问题有一个大致了解，这有助于工作人员的决策，答复等工作。根据以上词云图，我们已经对群众反映的问题有了一个大致的了解，也对问题的回复与解决有了一个大致的思路。我们可以得知，群众在劳保方面主要是关注就业，社保等问题。在城建方面，主要是发展建设与民众生活有了矛盾。在教育方面，关注于教育公平性等问题；等等。

但是，仅由如此数据进行分类，必然是不准确的，故本文引入了权重的概念，使用 TF-IDF 加权模型来给文本中更重要的词增加权重。

3.2 TF-IDF 加权模型

在我们的留言详情标签表中，很多语句都出现了您好、请问等词，这些词对我们来说，是无关紧要的，不能作为分类的标准的，这时候我们就要考虑如何过滤掉这些词。

为评估每一条留言详情中的字词对于这条留言详情的重要程度，我们用到了 TF-IDF 算法。TF-IDF 算法将权重分为两部分：局部因子(词频)和全局因子(逆文档频率)，通过统计学的方法来计算术语对于文本内容的重要性[1]。当一个词在它所在的语句中出现的次数越多，它的重要性就越高；但它的重要性在语料库中随着它出现的频率的增加反而会变低。

（1）词频 TF

词频 TF 是指某个词在文本中出现的频率：

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

其中，针对本文，$n_{i,j}$ 是这个词在这个语句中的出现次数，$\sum_k n_{k,j}$ 是在所有语句总共出现的词的数目。

（2）逆向文件频率 IDF

逆向文件频率由总语句数除以包含该词语的语句的数目，再取对数得到：

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{4}$$

为防止 $|\{j : t_i \in d_j\}|$ 为 0，一般情况为给分母加一个 1，即：

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}| + 1} \quad (5)$$

当包含该词语的语句的数目越少时，IDF 就越大，此时该词的类别区分能力的对应越强。

（2）TF-IDF

TF-IDF 其实就是 tf*idf，将词频与逆向文件频率相乘，就可以得到高权重的 TF-IDF，可以轻松过滤掉一些无关紧要的词。比如今天、昨天、县、市这些词，出现的频次很高，但是由这些词并不能看出有效的信息，通过 TF-IDF 处理后，它们的权重就轻松地降低了。

第四章 研究方法与过程

4.1 总体流程

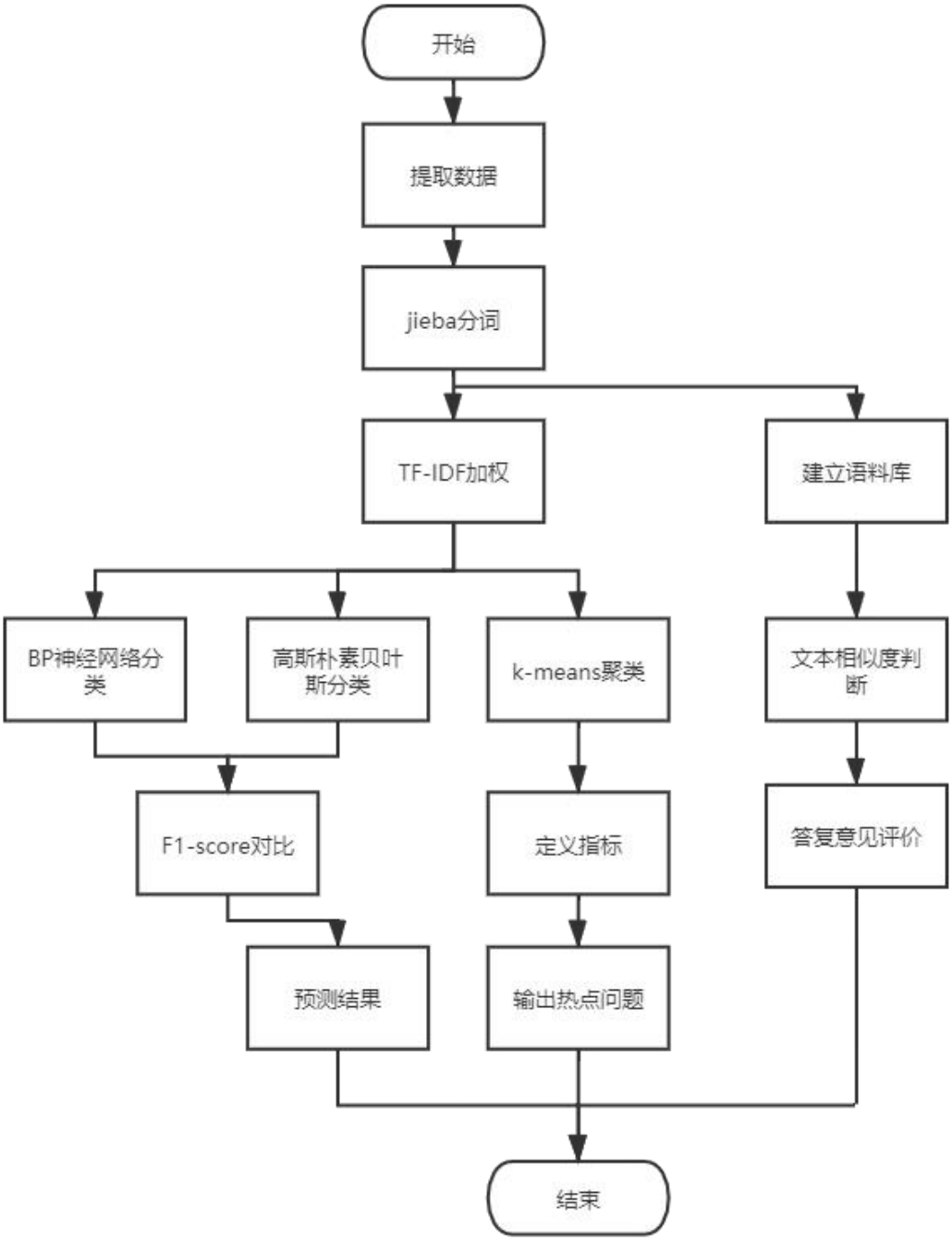

图 4.1 总体流程图

4.2 群众留言分类问题分析

留言分类问题的关键是根据已知数据，采用合适的分类方法对数据分类，以此建立一个分类器模型模型。首先，文本已经通过预处理得到了分词之后的文本，但文本依然不可量化，所以我们先使用 TF-IDF 模型构建文本特征模型。然后在此基础上，使用分类器模型进行训练。本文采取了机器学习（高斯贝叶斯）和深度学习（深度神经网络），并进行了对比，最后采取了深度神经网络。

4.2.1 高斯朴素贝叶斯模型

朴素贝叶斯的三个常用模型是高斯、多项式、伯努利朴素贝叶斯模型，在本文中我们主要用到了高斯朴素贝叶斯模型。高斯朴素贝叶斯的适用范围比朴素贝叶斯更广。

4.2.2.1 贝叶斯定理

贝叶斯定理是一个修正概率的定理。通常来说，我们在对某一事件进行推断之前，往往已经了解了关于这个事件的概率，这种实现了解的的概率我们一般称为先验概率。在后续的研究中，如果我们通过抽样调查等又获得了有关该事件的信息，就可以根据这些对之前判断的先验概率进行修正，使的先验概率变成后验概率。

$$P(A|B)=\frac{P(B|A)*P(A)}{P(B)} \tag{6}$$

P(A)：事件 A 的先验概率。

P(B)：事件 B 的先验概率。

P(A|B)：已知 B 发生后 A 的条件概率（后验概率）。

P(B|A)是已知 A 发生后 B 的条件概率（后验概率）。

4.2.2.2 朴素贝叶斯

朴素贝叶斯法是基于贝叶斯定理与条件独立假设的分类方法[2]。即在贝叶斯算法的基础上假定目标值时属性之间是相互条件独立的。针对本文来说，每一个留言之间，并不会遇到彼此的影响，它们之间是相互独立的，所以考虑采用贝叶斯分类。当遇到关键词“社保”、“费用”、“报销”等出现在同一个语句中时，程序大概率就会判断这个留言详情是属于“劳动和社会保障”类，因为通过训练，此时程序中这些词出现在该类的频次最高，当然它也有可能属于其它类，如果没

有其他的信息，这个语句就会归属于条件概率最大的类别，这就是朴素贝叶斯。

假如有文本向量 $X=\{X_1,X_2,X_3,...,X_m\}$，类标签 $Y\in\{1,2,3,...,n\}$，训练样本 s 个，则有训练样本集：

$$S=\{(x_1,y_1),(x_2,y_2),...,(x_s,y_s)\} \tag{7}$$

设预测目标为类别 a，由式（6）可知，后验概率 P（Y=a|X）为

$$P(Y=a\mid X)=\frac{P(X\mid Y=a)*P(Y=a)}{P(X)} \tag{8}$$

由条件独立性有

$$\begin{aligned}P(X\mid Y)&=P(X_1,X_2,...,X_m\mid Y=a)\\&=P(X_1\mid Y)*P(X_2\mid Y)*...*P(X_m\mid Y)\end{aligned} \tag{9}$$

根据贝叶斯定理，则有

$$y=\arg\max_k\prod_{i=m}^{m}p(X=x_i\mid Y=a) \tag{10}$$

4.2.3.3 高斯朴素贝叶斯

高斯分布又称正态分布，高斯朴素贝叶斯与朴素贝叶斯不同的是，高斯朴素贝叶斯假设条件概率是多元高斯分布。

高斯分布概率密度函数：

$$N(x\mid\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma}}\exp\{-\frac{(x-\mu^2)}{2\sigma^2}\} \tag{11}$$

则在第 a 类分类标签下的高斯分布为：

$$N(x_i\mid\mu_{i,a},{\sigma_{i,a}}^2)=\frac{1}{\sqrt{2\pi\sigma_{i,a}}}\exp\{-\frac{(x_i-{\mu_{i,a}}^2)}{2{\sigma_{i,a}}^2}\} \tag{12}$$

则条件概率为

$$P(X\mid Y=a)=\prod_{i=1}^{m}N(x_i\mid\mu_{i,a},{\sigma_{i,a}}^2) \tag{13}$$

由式（12）、（13）可知：

$$y=\arg\max_a P(Y=\mathrm{a})*P(X=x\mid Y=a) \tag{14}$$

最后进行最大似然估计法，对式中的参数进行估计，即可计算概率分布。

4.2.2 深度神经网络模型（BP 神经网络）

BP（Back Propagation）神经网络是一种按误差逆传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一。

人工神经网络模拟人的思维，是一个非线性动力学系统，其特色在于信息的分布式存储和并行协同处理。虽然单个神经元的结构极其简单，功能有限，但当大量神经元构成一个网络系统时，其所能实现的却是极其丰富的。BP 神经网络主要是在对样本数据的不断学习和训练过程中，不断修正不同特征参数的权重值，从而进行预测[3]。

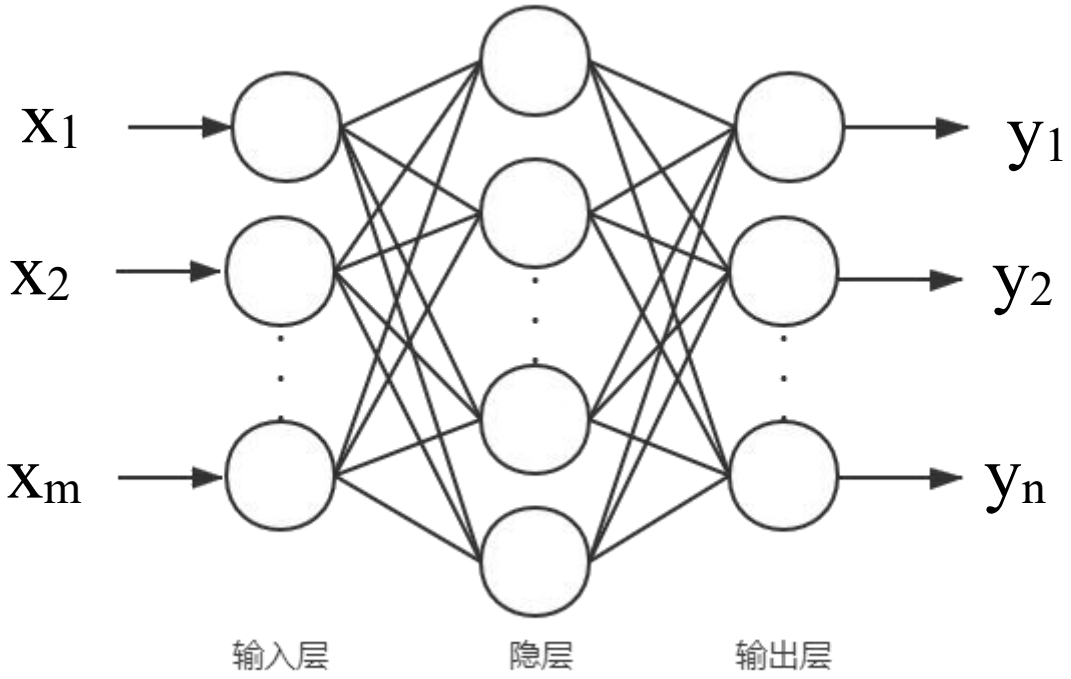

图 4.2 神经网络拓扑结构图

如图所示为 BP 神经网络的拓扑结构，一般包含三层，即输入层、隐层和输出层。它的特点是：每一层由若干个结点(神经元)组成，层间结点通过权值联接，同一层结点之间没有连接。单计算层前馈神经网络只能求解线性可分问题，能够求解非线性问题的网络必须是具有隐层的多层神经网络。

如图 4.2 所示的 BP 神经网络，输入层包含 m 个节点，输出层包含 n 个节点，可以看做是一个 m 维向量到一个 n 维向量的映射。 隐层节点的选择有一个经验公式：

$$h = \sqrt{m+n} + a \qquad (15)$$

其中 h 为隐含层节点数目，m 为输入层节点数目，n 为输出层节点数目，a 为 1~10 之间的调节常数。

4.2.2.1 BP 算法的原理

设有一个 m 层的神经网络，输入层有样本 X；设第 k 层的 i 神经元的输入总和为U_{i}^{k}，输出X_{i}^{k}；从第 k—1 层的第 j 个神经元到第 k 层的第 i 个神经元的权系数为W_{ij}各个神经元的激发函数为 f，取$\mathrm{f}(x)=\dfrac{1}{1+\exp(-x)}$，为输出节点 i 的信号，则各个变量的关系如下所示：

$$X_{\mathrm{i}}^{\mathrm{k}}=f(U_{i}^{k}) \tag{16}$$

$$U_{\mathrm{i}}^{\mathrm{k}}=\sum_{j}W_{ij}X_{j}^{k-1} \tag{17}$$

反向传播算法分两步进行：

1.正向传播：输入的样本从输入层经过隐单元一层一层进行处理，通过所有隐层后，传向输出层；在逐层处理的过程中，每一层神经元的状态只对下一层神经元的状态产生影响。最后在输出层把现行输出和期望输出进行比较，若现行输出与期望输出不符，则进入反向传播过程。

2.反向传播：把误差信号按原来正向传播的通路进行反向传回，并对每个隐层中的各个神经元的权系数进行修改，以使误差信号趋向最小。

反向传播算法描述如下：

首先定义误差函数：

$$\mathrm{e}=\frac{1}{2}\sum_{i}(M_{i}^{m}-Y_{i})^{2} \tag{18}$$

再确定梯度：

$$\frac{\partial \mathrm{e}}{\partial W_{ij}}=\frac{\partial \mathrm{e}}{\partial U_{i}^{k}}\cdot\frac{\partial U_{i}^{k}}{\partial W_{ij}} \tag{19}$$

其中

$$\frac{\partial U_{i}^{k}}{\partial W_{ij}}=X_{\mathrm{j}}^{k-1} \tag{20}$$

令 $d_i^k = \frac{\partial e}{\partial U_i^k}$，

$$\text{则} \frac{\partial e}{\partial U_i^k} = \frac{\partial e}{\partial X_{\mathrm{i}}^k} \cdot \frac{\partial X_{\mathrm{i}}^k}{\partial U_i^k} \tag{21}$$

其中

$$\frac{\partial X_{\mathrm{i}}^{\mathrm{k}}}{\partial U_i^k} = f'(U_i^k) = (\frac{1}{1+\exp(-U_i^k)}) = X_i^k(1-X_i^k) \tag{22}$$

①当 k=m 时：

$$\frac{\partial e}{\partial X_{\mathrm{i}}^k} = \frac{\partial e}{\partial X_{\mathrm{i}}^{\mathrm{m}}} = (X_i^m - Y_i) \tag{23}$$

②当 k<m 时：

$$\frac{\partial e}{\partial X_{\mathrm{i}}^k} = \sum_{\mathrm{l}} \frac{\partial e}{\partial U_{\mathrm{l}}^{\mathrm{k+1}}} \cdot \frac{\partial U_{\mathrm{l}}^{\mathrm{k+1}}}{\partial X_{\mathrm{i}}^{\mathrm{k}}} = \sum_{\mathrm{l}} W_{li} \cdot d_l^{k+1} \tag{24}$$

所以权值修改公式为：

$$\Delta W_{\mathrm{ij}}(\mathrm{t}+1) = \alpha \Delta W_{\mathrm{ij}}(\mathrm{t}) - \eta \cdot d_l^k \cdot X_j^{k-1} \tag{25}$$

其中：

$$\mathrm{d}_l^m = X_i^m(1-X_i^m)(X_i^m - Y_i) \tag{26}$$

$$\mathrm{d}_l^k = X_i^{\mathrm{k}}(1-X_i^k)\sum W_{li} \cdot d_l^{k+1} \tag{27}$$

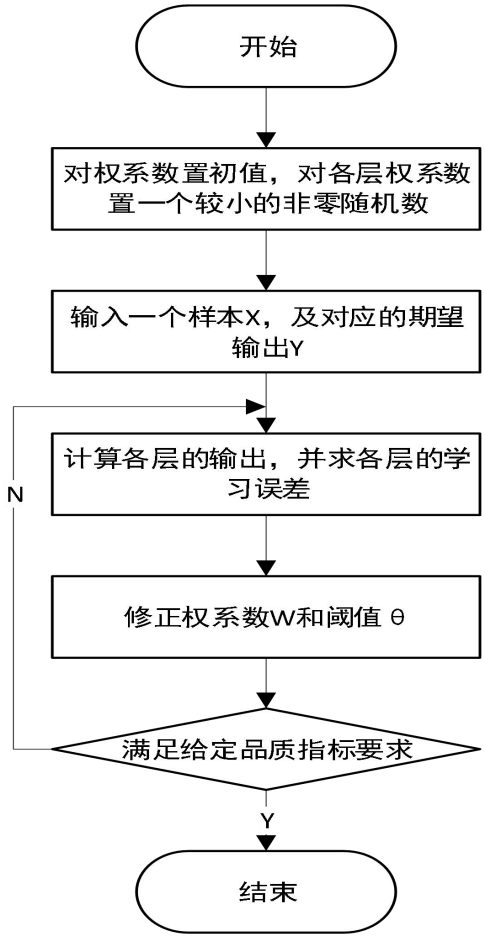

图 4.3 算法的执行步骤流程图

BP 算法的执行步骤如图 4.3 所示，当给定不满足品质指标要求时，回到计算各层输出步骤再次执行；在这个学习过程，对于任一给定的样本和期望输出都要执行，直到满足所有输入输出要求为止。

4.2.3 模型的结果与分析

4.2.3.1 模型的建立

本文先建立了文本特征模型，然后在此基础上选用分类器进行数据训练。在此之前本文先对评论的种类分布进行一个统计，如图所示：

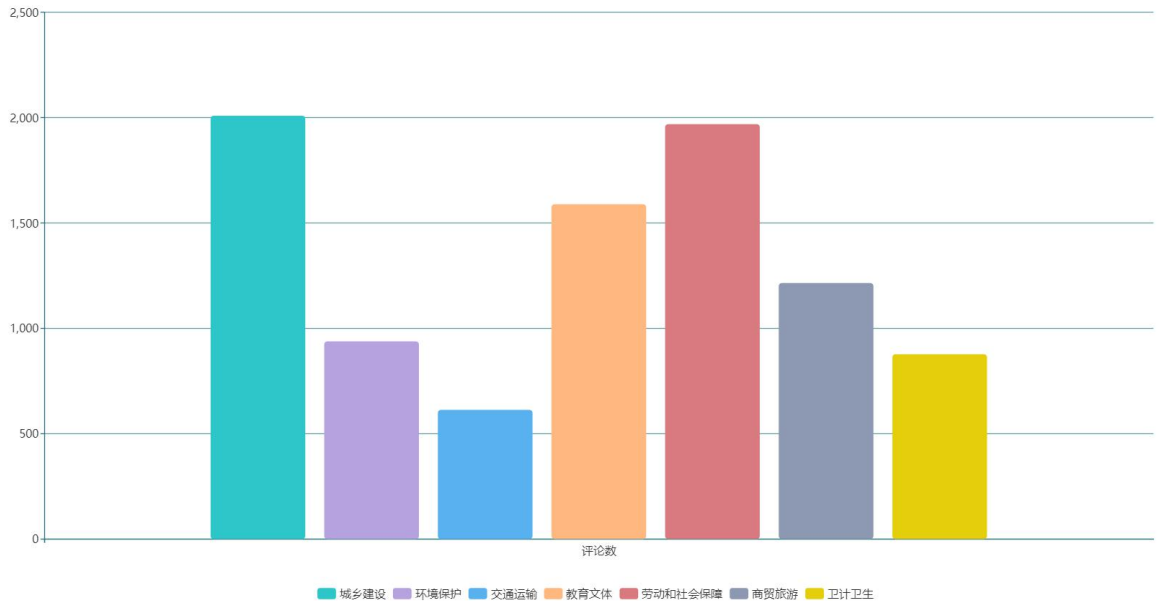

图 4.4 留言类别条形统计图

表 4.1 留言类别统计表

类别	数量	标签
城乡建设	2009	1
劳动和社会保障	1969	2
教育文体	1589	3
商贸旅游	1215	4
环境保护	938	5
卫生计生	877	6
交通运输	613	7

在数据量庞大的情况下，应当每种类型取同样数量的评论，有利于分类器的分类。但在数据量较小的情况下，这样反而会造成数据浪费。本文中数据量较高，故本文对每中类型选取了相同数量的留言（即每类取 613 条留言）进行分类器的训练。

4.2.3.2 F-Score 评价指标

针对群众留言分类问题的模型，本文使用 F-Score 进行了评价：

$$F = (1+\beta^2)\frac{P_i * R_i}{\beta^2(P_i + R_i)} \tag{28}$$

其中：P 为查全率，R 为查准率, β 是一个权值；

1. 查全率

$$P = \frac{TP}{TP + FP} \tag{29}$$

2. 查准率

$$R = \frac{TP}{TP + FN} \tag{30}$$

TP、FP、FN、TN 的解释如下表

表 4.2 参数解释表

真实情况	预测结果	
	真	假
真	TP(Ture Positive)	FN((Ture Negative)
假	FP(False Positive)	TN(False Negative)

3. β 的取值

β 的取值与查全率和查准率有关

①当查全率比查准率更重要时， β 的取值小于 1；

②当查准率比查全率更重要时， β 的取值大于 1；

③当查全率与查准率一样重要时， β 的取值等于 1；

一般情况下， β 取 1，在本文中 β 也取 1,即有：

$$F_i = \frac{2P_iR_i}{P_i + R_i} \tag{31}$$

针对留言分类问题，本文建立的模型有多个类，样本并不平衡，故采用 macro F1 给所有类赋予相同的权重，即：

$$F_1 = \frac{1}{a}\sum_{i=1}^{s}\frac{2P_iR_i}{P_i + R_i} \tag{32}$$

4.2.3.3 高斯分布下的朴素贝叶斯模型下的分类效果

本文已经建立了相关的分类器模型，接下来我们根据模型得出相关的分类结果。如图所示（每次训练的数据会有不同，但基本上趋近于一特定值，如表 4. 所示，此处数值是其中多次分类过程得出的平均数据）：

表 4.3 高斯分布下的朴素贝叶斯模型 F1-score 表

评估指标	precision	recall	F1-score
数值	68.30	68.28	68.15

$$F1-Score=2\cdot\frac{\Pr ecision\cdot \operatorname{Re} call}{\Pr ecision+\operatorname{Re} call}\approx 2\cdot(\frac{68.30\cdot 68.28}{68.30+68.28})\approx 68.15 \quad (33)$$

通过计算可以看出，高斯分布下的朴素贝叶斯模型的 F1-score 大约在 68.15 左右。

F1-score 是综合考虑了模型查准率和查全率的计算结果，F1-score 越大自然说明模型质量更高，但是若 F1-score 过高，则会影响模型的泛化能力，可能会造成过拟合。

4.2.3.4 深度神经网络模型下的分类效果

我们已经建立了相关的分类器模型，接下来我们根据模型得出相关的分类结果。如图所示（每次训练的数据会有不同，但基本上趋近于一特定值，此处数值是多次分类过程得出的平均数据）：

表 4.4 深度神经网络模型 F1-score 表

评估指标	precision	recall	F1-score
数值	89.99	89.92	89.96

$$F1-Score=2\cdot\frac{\Pr ecision\cdot \operatorname{Re} call}{\Pr ecision+\operatorname{Re} call}\approx 2\cdot(\frac{89.99\cdot 89.92}{89.99+89.92})\approx 89.96 \quad (34)$$

由数据可见，我们在 BP 深度神经网络模型下的 F1-score 高达 89.96，分类器分类应该较为准确。

4.2.3.5 分类效果对比

表 4.5 分类方法 F1-score 对比表

评估指标	precision	recall	F1-score
高斯朴素贝叶斯	68.30	68.28	68.15
BP 神经网络	89.99	89.92	89.96

通过高斯朴素贝叶斯和 BP 神经网络的 F1-score 对比发现，高斯朴素贝叶斯的 F1-score 为 68.15，而 BP 神经网络的 F1-score 为 89.96，故本文采用 BP 神经网络作为分类器模型，最后根据 BP 神经网络分类器得出模型的评估结果。

表 4.6 BP 神经网络 F1-score 表

评估指标	F1-score
测试一	89.96
测试二	90.11
测试三	89.82
测试四	90.12
测试五	89.76
平均值	89.96

注：为保证测试模型的稳定性，所以对多次运行结果求平均值。

对 BP 神经网络进行多次测试，最后取得其平均 F1-score 为 89.96。

4.3 热点问题

4.3.1 热点问题挖掘分析

对热点问题进行挖掘的关键是对评论进行聚类，我们需要采取合适的方式，让内容相似的评论，聚集到一起。本文根据聚集的簇的大小与评论的热度关注也就是点赞数，对热点问题设置一个热度指标，找出其中的热点问题，之后将相关的时间跨度，热点问题关键词一一获取，完善模型。

根据假设，热度排名有两个参考因素，同类问题的评论数 c 和同类问题的点赞总和（g），假设一个用户一个 IP 的条件下，依据日常，评论是一个人的看法，一个人点赞表示认同这种看法，我们可以把同一类相似的问题聚类，然后找出同一类问题中，民众最关注的（也就是点赞数最高的问题），就可以得到其中的热带问题。

4.3.2 k-means 聚类

聚类分析是对数据的分类，是数值分类确定与多元统计技术相结合的结果，将一个数据集根据它的特征划分为若干个组，使组内的相似性与组间相似性相比较大[4]。简而言之，这个算法的目的就是将数据中相似的放到一起，聚成一个类；与分类不同的是，在聚类结果出来之前，我们并不知道每一个类各自有什么特征。

4.3.2.1 k-means 聚类步骤

Step：

1.将数据集中的 m 个数据分为 k 个簇， 随机选取 k 个聚类质心点。

2.计算每一个样本与质心之间的欧氏距离 r。

欧氏距离：

$$r = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \tag{35}$$

3.通过欧式距离 r，将各样本归类到其距离最近的簇。

即：

$$c_j = \arg\min_i s \tag{36}$$

c_j 代表各簇中与各质心距离最近的样本点。

4.通过计算每个簇中的均值，得到新的质心。

即：

$$\eta_j = \frac{\sum_{x_i \in c_j} x_j}{|c_j|} \tag{37}$$

η_j 为新的质心。

5.设置一个阈值，当新的质心与旧的质心小于这个阈值时，可以认为聚类已经达到期望的结果，算法终止。否则，重复 step2~5。

4.3.3 模型建立

4.3.3.1 k 值选取

本文根据实际不断尝试，获取最利于热点问题获取的 k 值，本文根据示例数据的尝试，发现 30 条评论的示例数据确定 k 值为 7 时，效果最好。如下表所示：

表 4.7 测试数据聚类表

问题 ID	留言主题
1	A 市经济学院体育学院变相强制实习
1	A 市经济学院强制学生外出实习

1	A 市经济学院组织学生外出打工合理吗？
1	A 市经济学院强制学生实习
1	A 市经济学院寒假过年期间组织学生去工厂工作
4	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气，急需处理！
4	A5 区劳动东路魅力之城小区油烟扰民
4	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气
4	A5 区劳动东路魅力之城小区临街门面烧烤夜宵摊
4	A5 区劳动东路魅力之城小区底层餐馆油烟扰民

根据这个比例，本文对全部数据先确定一个选取 k 值的比例。

全部数据的 k 取值=测试数据 k 值与测试数据量比值*全部数据量

=（7/30）*4326=1009.4

根据上式，估计全部数据的 k 取值应该就是在 1009 左右，通过不断尝试，可以确定更以利于效果的 k 值。本文通过反复尝试，确定了 k 值为 1200。

4.3.3.2 热点问题

本文根据取到的 k 值，使用 k-means 聚类得到了如下热点问题表：

表 4.8 热点问题表

热度排名	热度指数	问题描述
1	2097	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
2	1767	反映 A 市金毛湾配套入学的问题
3	821	请书记关注 A 市 A4 区 58 车贷案
4	790	严惩 A 市 58 车贷特大集资诈骗案保护伞
5	669	A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到等等问题

根据热点问题表可以看出，热度指数最高的是 A 市 A5 区汇金路五矿万境 K9 县存在一系列问题，其次是反映 A 市金毛湾配套入学的问题、请书记关注 A 市 A4 区 58 车贷案等。我们我们针对“请书记关注 A 市 A4 区 58 车贷案”问题进行展开，如下表所示：

表 4.9 热点问题示例明细表

热度排名	问题 ID	留言编号	留言主题	留言时间	点赞数	反对数
4	465	217032	严惩 A 市 58 车贷特大集资诈骗案保护伞	2019-02-25 09:58:37	790	0

4	465	240554	A 市 58 车贷老板跑路美国，经侦拖延办案	2019-02-10 20:58:40	6	0
4	465	272858	A 市 58 车贷恶性退出案件为什么不发布案情进展通报？	2019-01-16 23:21:21	0	0
4	465	218132	再次请求过问 A 市 58 车贷案件进展情况	2019-01-29 19:15:49	0	0

上表是关于问题“58 车贷案的相关细节”的展开细节，可以看到，五八车的案的相关细节，这个问题包括“严惩 A 市 58 车贷特大集资诈骗案保护伞”、“A 市 58 车贷老板跑路美国，经侦拖延办案”、“A 市 58 车贷恶性退出案件为什么不发布案情进展通报？”等等留言主题，分别在 2019 年 1 月、2 月都有留言，2019-02-25 的“严惩 A 市 58 车贷特大集资诈骗案保护伞”问题获 790 个点赞，0 个反对。

4.3.4 模型结果

4.3.4.1 关键词获取

根据如上热点问题挖掘，本文将各热点问题下的留言主题进行关键词挖掘，展示各热点问题的最关键的问题描述，以便于更好地观察。

表 4.10 热点问题描述表

热度排名	热度指数	地区/人群	问题描述
1	2097	A 市 K9 县	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
2	1767	A 市/入学	A 市金毛湾配套入学的问题
3	821	A 市/车贷案	请书记关注 A 市 A4 区 58 车贷案
4	790	A 市/车贷诈骗	严惩 A 市 58 车贷特大集资诈骗案保护伞
5	669	A4 区	A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到，合理吗？

根据表 4.10 可以看出，热点指数第一的问题中，最关键的问题描述是“A 市 A5 区汇金路五矿万境 K9 县存在一系列问题”，热点指数第二的问题中，最关键的问题描述是“A 市金毛湾配套入学的问题”。

4.3.4.2 时间跨度获取

为观察热点问题是长期存在的热点问题还是短期存在的热点问题，本文对这些热点问题进行了时间跨度划分，以便观察热点问题是否在一定时间内得到解决。

表 4.11　热点问题时间范围表

热度排名	热度指数	时间范围	时间长度	问题描述
1	2097	2019/01/15 至 2019/11/11	300	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
2	1767	2019/02/11 至 2019/12/17	309	反映 A 市金毛湾配套入学的问题
3	821	2019/02/21 至 2019/08/02	162	请书记关注 A 市 A4 区 58 车贷案
4	790	2019/01/16 至 2019/02/25	40	严惩 A 市 58 车贷特大集资诈骗案保护伞
5	669	2019/01/30 至 2019/09/06	219	A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到，合理吗？

根据上表可以看出，热点问题排行前五的问题，最长存在的范围在 309 天，是“反映 A 市金毛湾配套入学的问题”，最短的是“请书记关注 A 市 A4 区 58 车贷案”，问题持续有人反映的时间长度为 40 天。热点指数高的前两个问题存在群众持续反映时间均为 300 天以上，对这两个问题观察，发现这两个问题均为一时不好处理得当的问题；而观察仅存在 40 天的“严惩 A 市 58 车贷特大集资诈骗案保护伞”与存在 162 天的“请书记关注 A 市 A4 区 58 车贷案”，发现其中存在 162 天的问题与存在 40 天的问题时间段几乎是相连的。

4.4 答复意见的评价

对答复意见的评价的关键是对答复意见进行文本相似性分析。一般来说，一条好的答复意见一般包括对问题的重述、调查结果、解决方案等；故本文对留言详情与答复意见进行文本相似度分析，构建评价模型，根据相似度高低为答复意见评分（即将相似度乘以 100 得到最终分数）。

4.4.1 文本相似度分析

针对文本相似性分析，本文分别对比了最小编辑距离算法和余弦相似度算法计算不同答复意见间的相似性，最终选定余弦相似度进行文本相似度分析。

4.4.4.1 最小编辑距离算法

狭义编辑距离：设有 A、B 两个字符串，用 ED（A，B）来表示 A 字符串转换成 B 字符串需要操作次数。即最少删除的字符数、插入和替换字符的次数。直观来说，两个串互相转换需要经过的步骤越多，差异越大。

Step：

①对两部分文本进行处理，将所有的非文本字符替换为分段标记“#”

②将较长文本作为基准文本，遍历分段之后的短文本。若发现长文本包含短文本子，则句后在长本文中移除，若未发现匹配的字句，则累加长度。

③比较剩余文本长度与两段文本长度和，比值为不匹配比率。

4.4.4.2 余弦相似度

余弦相似度算法识根据两个词向量间的余弦夹角来判断词向量之间的相似性的[5]。

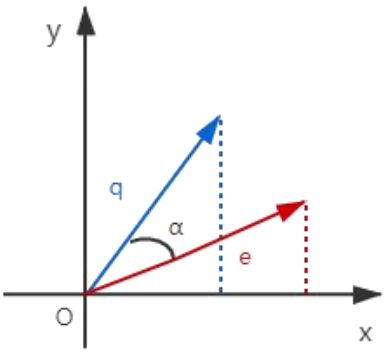

图 4.5 向量夹角图

余弦夹角如上图所示，若两个向量的夹角为 0，其余弦值为 1，相似度最高，其计算公式如下:

$$\cos\alpha = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}} \tag{38}$$

余弦的这种计算方法对 n 维向量也成立。假定 A 和 B 是两个 n 维向量，A 是 [A_1, A_2, …, A_n] ，B 是 [B_1, B_2, …, B_n] ，则 A 与 B 的夹角α的余弦等于：

$$\cos\alpha = \frac{\sum_{i=1}^{n}(A_i + B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} = \frac{A \cdot B}{|A| \times |B|} \tag{39}$$

在实际应用中，表征文本特征的两个向量的长度是不同的。因此必然需要对上述向量进行处理。根据原向量是否在新向量（归并后的向量）存在来判断，若存在，则以该词汇的词频来表征；若不存在，则该节点置为 0 。

示例：

Text1_1: It is a beautiful butterfly

Text1_2: beautiful butterfly

Text2_1: She is a beautiful girl

Text2_2: beautiful girl

Vector: beautiful butterfly girl

Vector1 = (1, 1, 0)

Vector2 = (1, 0, 1)

4.4.2 答复意见评分

通过文本相似度分析，并给答复意见评分，得到如下表格：

表 4.12　高分答复意见表

答复意见	答复时间间隔/天	分数
网友“UU008424”　　您好！您的留言已收悉。现将有关情况回复如下：　　您好！我街道接到反映后，物管办联合和馨园社区工作人员积极了解情况。现将有关情况回复如下：目前小区常住人口近 5000 人，小区内地面车位加地下车库车位仅有 900 余个，小区业主车辆达到 1100 余台，外来租户车辆达到 300 余台。小区内停车问题一直困扰着小区居民，小区消防通道经常堵塞，给住户带来了严重的安全隐患。为此，小区业主代表经过研究、协商提出，外来租户的车辆统一安排至地下停车场停放，不进入小区内部停放。　　1.和馨园保障房小区为 103 号令拆迁安置小区，并不是商品房住宅小区，属于农民安置小区，小区根据安置农民集体愿望，制定了《关于小区车辆管理的实施方案》。该方案属于村民自治组织内，《物权法》、《物业管理条例》适用于商品房住宅小区。2.社区采纳意见后，先后召开了小区业主代表大会、党员大会、中层骨干会议，以及发放调查问卷等形式，征集小区业主意见，由于小区业主普遍认为小区租户为非 103 号令拆迁人员,仅缴纳了房屋租金,未额外缴纳物业管理费用,而业主物业管理费用的免缴纳是通过每人 5 平米的门面房屋出租费用认缴，所以业主认为理应由业主优先享受小区内的停车位等配套设施，最终小区业主一致投票决定外来租户车辆不能进入小区内部停放。　　根据小区业主意见，社区居委会与业主代表共同制定了《关于小区车辆管理的实施方案》，并张榜公示。方案明确为保障业主正常权益，原则上小区业主对小区内院地上停车位享有优先使用权，租户须将车辆停入地下车库或停放于和馨园商业街沿线门店前的停车位，方案由和宇物业、高新物业在和馨园小区一期从 2017 年 2 月开始实施。　　3.业主的正常权益指的是房屋所有人的正常权益，业主并不是将所有权益都让渡给租户，比如说租户并没有权利参与社区的业主大会，但正常享有房屋租住权；公共停车位是和馨园集体经济组织所有。　　如果有租户车辆需要进入小区地下停车库停放，a 区租户到高新物业公司办理地下车库的停放手续，bc 区租户到和宇物业公司办理地下车库的停放手续。　　感谢您对我们工作的关心、监督与支持。　　2017 年 3	23	86

月 30 日		
网友“UU008835” 您好！来信收悉。现回复如下： 经镇综治办调查核实，信访人黄尊富，系我镇乌川湖村楼梯坡组村民。2010 年—2012 年期间，浏醴高速公路修建途经我镇乌川湖村，并在该村楼梯坡组设计有一涵洞，该涵洞能安全通过小轿车、农业用车等车型。黄尊富的房屋在楼梯坡涵洞边，距离高速主线有 200 多米远，不符合拆迁条件。自涵洞开工建设以来，黄尊富以及其父黄刚明以楼梯坡涵洞设计位置较低，涵洞出口接线坡度较陡，造成其本人和家人出行不便为由，多次向江背镇浏醴高速公路指挥部反映，要求解决该问题，并要求进行拆迁。 江背镇党委、政府高度重视此事，从 2011 年至今，多次协调浏醴高速公路指挥部采取措施保障黄尊富一家的安全出行，并给以一定经济补偿，具体情况如下： 1、在浏醴高速建设期间，积极解决黄尊富一家的出行问题并给以补偿。由浏醴高速八标段项目部承担原材料、人工等所有开支，由黄尊富之父黄刚明按照其家人出行方便的需求，现场指挥施工，对涵洞出口到其家门口的坡道泥路进行水泥硬化，有效减小了涵洞建设对黄尊富一家生产生活的影响，已于 2012 年 10 月完工。同时该户户主黄刚明与指挥部签订了因涵洞设计较低导致出行不便的一次性补偿协议，协议明确要求黄刚明一家领取补偿款后，不得再就涵洞出口坡度较陡造成出行不便的问题提出其它要求。2012 年 1 月 20 日黄刚明领取补偿款 21200 元。 2、黄尊富否认第一次协议效力，并以自己名义签订合约再次拿到补偿。黄尊富认为 2012 年 1 月 20 日签订协议时他不在家，是户主黄刚明签订的，否认第一次协议的效力。2012 年 12 月 4 日，指挥部再次组织乌川湖村、楼梯坡组与黄尊富协商，并与其签订了 2 万元的一次性经济补偿协议，黄尊富承诺领取资金后，不再就此事向镇政府提出其他任何的补偿要求。 3、江背镇政府积极召开信访答复会。在该次补偿后，黄尊富仍不满意，每隔一个月到江背镇政府上访，并多次向问政西地省、县长信箱、“12345”等平台提出拆迁的无理诉求。2014 年 5 月 22 日，江背镇党委、政府组织镇派出所、国土所、经贸办、综治办、浏醴高速公路拆迁指挥部等部门负责人和黄尊富本人召开了有关信访答复会，明确答复黄尊富，因其房屋不在浏醴高速拆迁红线范围内，要求拆迁房屋的诉求没有相关的政策依据，也没有现实操作拆迁的可能，由于高速公路建设造成的客观影响已得到妥善解决，如果仍然不能认同，必须合理合法的到相关部门反映情况。 4、2015 年 12 月 17 日，江背镇综治、信访、派出所、原浏醴高速指挥部成员代表、黄尊富参加了最近一次的信访答复会。会上，我镇对黄尊富进行了政策宣讲与解释，黄尊富不理解、不满意，并通过打电话、发短信等方式多次骚扰我镇工作人员的正常办公与休息，我镇明确告知信访人：一是由于浏醴高速建设对其家庭出行造成的不便，指挥部已两次给予其家庭补助共计 4 万元，信访人也在协议上签字认可；二是根据信访人家庭实际情况，不符合拆迁的标准，不在拆迁范围之内；三是征地拆迁的主体是浏醴高速建设指挥部，不是江背镇人民政府；四是信访人反映诉求必须依法依规进行，不得以发短信威胁或以打电话骚扰等非正常方式干扰、影响国家机关或国家工作人员的正常工作、生活，在本次答复会议对其训诫的基础上，如若继续违反，将承担相应的法律责任。2016 年 1 月 7 日	8	86

由表 4.12 中的高分答复可以看到，高分留言答复不仅有对问题的重述，还有对问题的实际调查，并且均给出了几条解决方案，让群众真真切切感受到进行留言后问题得到了解决。

表 4.13　低分答复意见表

答复意见	答复时间间隔/天	分数
已收悉	8	0
网友：您好！留言已收悉	32	0
“UU0081182”	37	0
你好！2019 年 10 月 10 日	18	0
您的留言已收悉，关于您反映的问题，已转市公安局调查处理。	34	0
您的留言已收悉。关于您反映的问题，已转市教育体育局调查处理。	37	0

根据上表可以看出，在低分答复意见表中，有些答复间隔超过 30 天，有些答复没有实际意义，为了回答而回答，仅仅给出“已收悉”、“你好”等答复，不能为群众群众解决问题，也没有体现真正关切群众留言与生活，这些回答模型给出的分数都是 0 分。

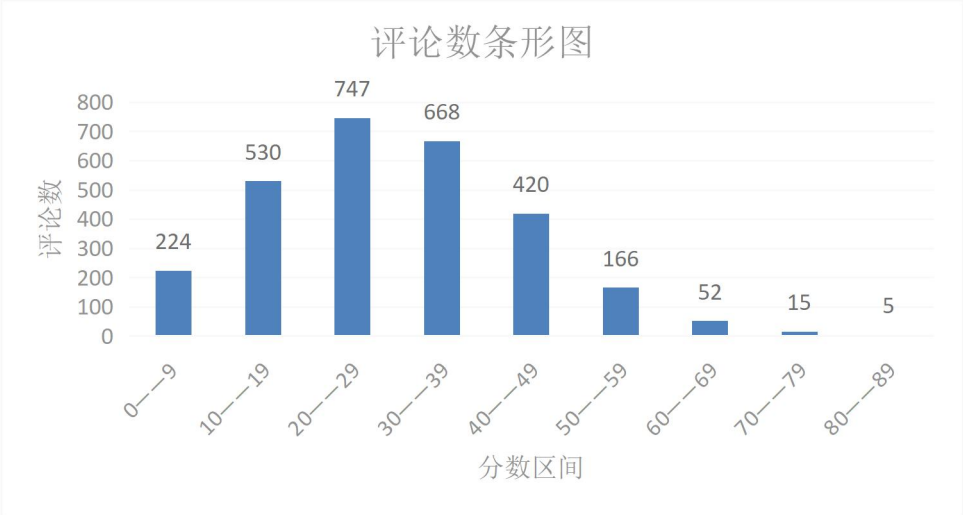

图 4.6 评论数与分数条形统计图

通过图 4.6 可以看出，在 20-29 分数段内，评论数最高，达到了 747 条，80-89 分区间内，仅有 5 条评论，整体大致符合正态分布的特征。

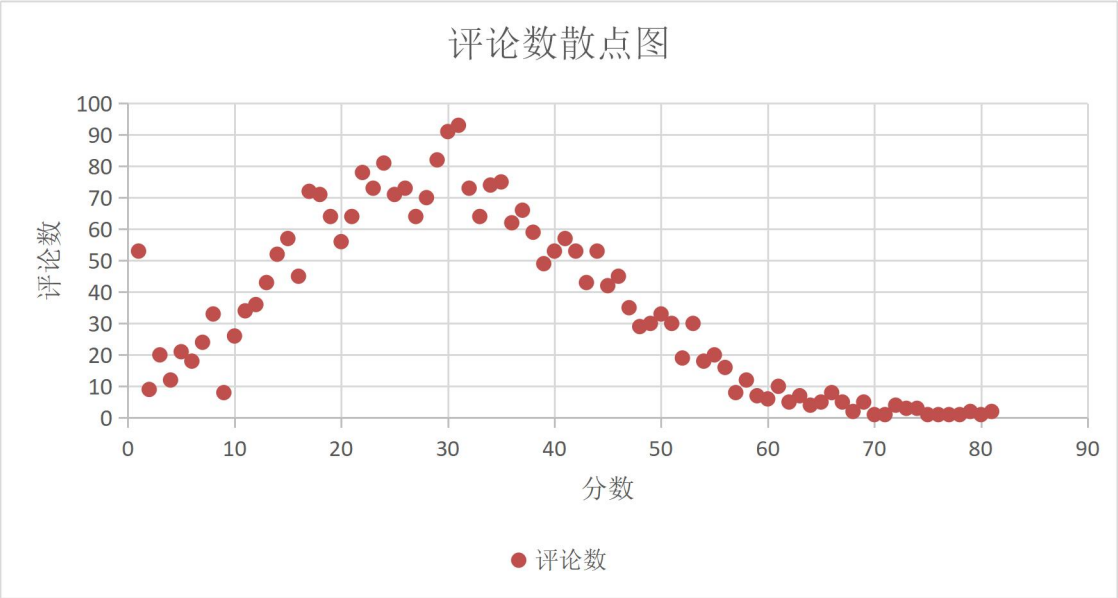

图 4.7 评论数与分数散点图

从概率统计规律来说，生活中大部分事情符合正态分布，比如说某地的高二男生的身高、某个学校的考试成绩等。在图 4.6 和图 4.7 中，可以明显地观察到，图形呈中间高，两边低且大致对称的形状，大致符合正态分布；故判断在答复意见的评价中，评分规则较为符合实际。

第五章 总结

本文的主要目的是通过数据挖掘与机器学习的方法建立起模型，从而达到对群众留言进行分类、聚类、评价的效果。对于处理大量群众留言来说，这些模型非常有效，不仅省时省力，还可以在短时间找出急需处理的热点问题，同时筛选出有益的建议。

首先，先对数据进行数据预处理，通过 jieba 分词与 TF-IDF 算法对数据进行优化，提高数据在模型中的有效性。而后本文建立了 BP 神经网络对群众留言进行分类、并利用 k-means 聚类对群众留言进行聚类、最后利用余弦相似度进行文本相似度分析从而达到给答复意见评分的效果。

本文对同一个问题分别建立几种不同的模型，同时利用多种评价、统计方法对模型进行观测，以保证模型的有效性，使得模型更具实际意义。但本文也存在着一些不足，比如说 BP 神经网络在训练时需要消耗大量的时间、答复意见的评价比较主观等。文本分析在现实生活中很具有实用性，本文在解决这些问题的时候考虑得还不够全面，应再多考虑些指标，进行相应的分析，这是我们期望去完成的方向。

参考文献

[1]古倩.基于特征向量构建的文本分类方法研究[D].西安理工大学.2019

[2]李航.统计学习方法[M].北京:清华大学出版社.2012

[3]左付山,李政原,吕晓等. 基于 BP 神经网络的汽油机尾气排放预测[J].江苏大学学报(自然科学版) .2020 (3);

[4]闫丽洁.基于 K-means 聚类方法的早期聚落规模等级研究[J].地域研究与开发.2020(04)

[5]苏东出.基于 TF-IDF 和余弦相似度的图书馆 OPAC 系统的研究和实现[J].平顶山学院图书馆.2019(07)