基于自然语言处理的智慧政务文本挖掘

摘要

这些年来，网络平台的兴起，使得这些平台已成为了相关部门了解社情民意的窗口和密切联系群众的重要渠道。对留言文本的合理处理可以大大提高相关部门办事效率和管理水平。因此，采用自然语言处理技术对文本数据进行处理是十分重要。

针对任务一，首先对初始文本数据进行分词、去停用词以及构建留言词典库预处理处理，再采用 TF-IDF 选取主要特征。分别构建线性 SVM 分类模型、朴素贝叶斯模型、感知器模型进行分类，采用查准率、召回率和F_1值评价指标比较分类效果。其中，线性 SVM 模型效果分类效果相对较好，在此基础上，不断改变非线性 SVM 模型的核函数、SVM 集成算法以及投票试验进行对比分析和深入探究，发现 k=50 时的 SVM-bagging 集成算法的性能最好，查准率为 0.889，召回率为 0.861，F_1为 0.873。

针对任务二，首先，对留言进行类似任务一中预处理，对处理后数据采用 LDA 主题模型计算概率密度函数，采用 JS 距离作为文本相似度计算公式。采用模型困惑度对主题数进行选取，经过计算，最优主题数为 29。最后，流式聚类算法 Single-Pass 聚类对留言进行聚类，得到热点问题。构建留言热点评价指数，对热点问题进行排序。

针对任务三，首先从答复意见的相关性、完整性、可解释性以及及时性构建 6 个指标，指标分别为回复相关性、文本字数、答复框架完整性、逻辑的解释性、信息的解释性、回复时间差。为便于后续分析，对指标进行标准化和正向化。其次，采用熵值法计算指标权重，其中复框架的完整性和逻辑的解释性对留言回复质量的评价权重最大，权重最小的是时间差。采用 TOPSIS 模型对答复意见质量进行量化，获得分值。为确定阈值，对每条留言回复计算所得分值进行 K-means 聚类，根据比例获得阈值 0.198、0.0793。最终有 85 条留言回复为“优秀”，590 条留言回复为“中等”，2141 条留言回复为“较差”。

关键词：文本分类、SVM-bagging 集成算法、SinglePass 聚类、TOPSIS 法

目录

图目录

表目录

第1章 问题背景与问题分析

1.1 问题背景

随着微信、微博、QQ 邮箱等网络平台的不断兴起，这些平台已成为了相关部门了解社情民意的窗口和密切联系群众的重要渠道，也是群众合理表达诉求的方式，同时，大数据时代下，各种信息爆发式增长，各种社情民意的留言文本数据也不断增加。通过把留言分类可以大大提高相关部门办事效率和管理水平，对于当地解决问题有着重要作用。然而，大量留言文本数据使得人为进行留言划分和热点整理变得十分困难，因此，采用自然语言处理技术对文本数据进行处理是十分重要的。

1.2 问题分析

1.2.1 问题一的分析

根据附件一中留言数据对文本构建分类模型。首先对初始文本数据进行分词、去停用词以及构建留言词典库预处理处理，再采用 TF-IDF 选取特征，对特征进行降维，选取主要特征。分别构建一系列文本分类模型，采用常见的查准率、召回率和F_1值作为模型的评价指标。其中，线性 SVM 模型效果分类效果相对较好，在此基础上，不断改变非线性 SVM 模型的核函数、SVM 集成算法进行对比分析和深入探究，以期得到更优的模型。

1.2.2 问题二的分析

任务二需要对热点问题进行挖掘。首先对留言进行预处理，包括分词与去停用词。为增加后续聚类的准确度，建立留言词典库。对处理后数据采用 LDA 主题模型，计算出每条留言的概率密度后，采用 JS 距离作为文本相似度计算。使用流式聚类算法 Single-Pass 聚类对留言进行聚类，得到热点问题。构建留言热点评价指数，对热点问题进行排序。

1.2.3 问题三的分析

任务三需要对构建评价体系对群众留言答复进行评价。首先从答复意见的相关性、完整性、可解释性以及及时性构建 6 个指标。为便于后续分析，对指标进行标准化和正向化。接下来，采用熵值法计算指标权重。采用 TOPSIS 模型对答复意见

质量进行量化，获得分值。为确定阈值，对每条留言回复计算所得分值进行 K-means 聚类，根据比例将留言质量分为“优秀”、“中等”、“较差”。

第2章 模型假设与符号说明

2.1 模型假设

1.数据集中人工标注的一级标签基本准确。

2.每个留言只能分为一个特定的类。

3.留言者均能较准确的表达自己的问题。

2.2 符号说明

表 1 符号说明表

H_i	聚类后第 i 类问题的热度
a_j	第 j 条留言的点赞数
b_j	第 j 条留言的反对数
w	任务三中指标权重
C	留言质量分值
P	LDA 主题模型困惑度

第3章 群众留言分类探索分析

3.1 任务一分析框架

对于第一问群众留言分类问题，本文探究分析框架如图 1 所示，具体来说，首先将文本集合进行预处理，包括对文本的分词处理与去停用词处理，为提升后续分类效果，对分词后的数据编程得到留言词典库。采用 TF-IDF 方法进行特征选取，再进行文本表示，最后建立模型，文本分类。建立模型时，首先对常用文本分类算法对比性能，其次比较不同核函数的 SVM 模型，最后采用 SVM 集成算法进行探究。

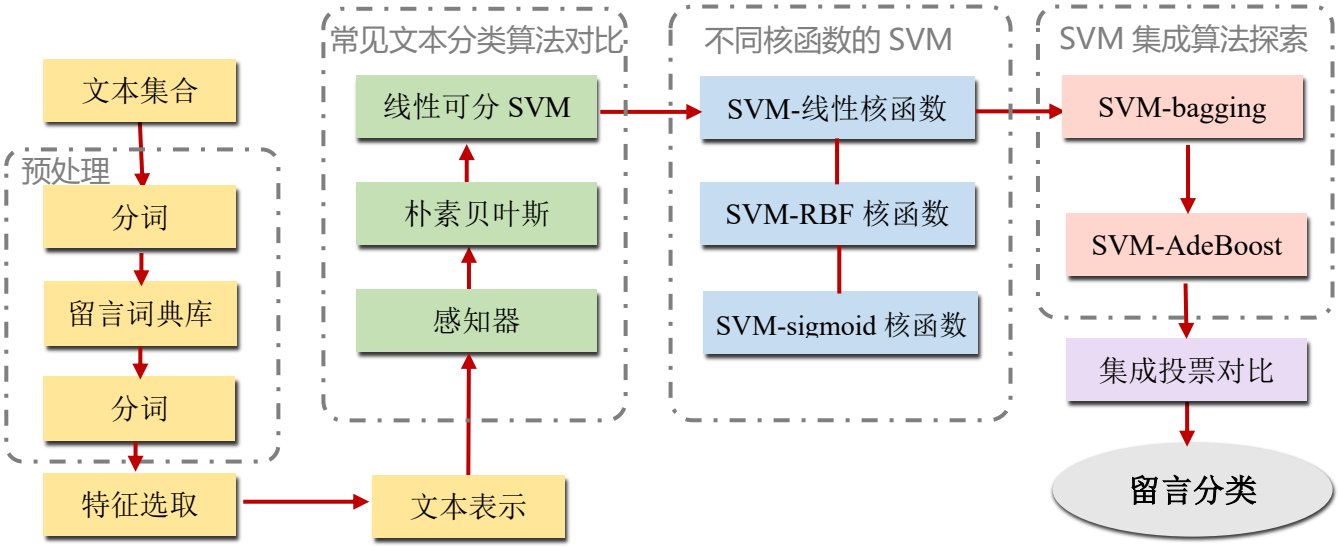

图 1 留言分类分析框架图

3.2 数据预处理

在数据挖掘过程中，数据的预处理至关重要，对于后续的特征提取，模型建立与分类预测起着关键作用。因此，在本题中，我们采用在自然语言处理中较为传统的数据预处理流程，进行分词和去停用词处理。

3.2.1 数据分词

在进行数据清洗之后，便是需要对数据进行分词处理。分词对于中文文本的分类具有重要作用，可以说，分词的表现直接影响到文本分类的性能，目前中文分词技术已经不断成熟。小组采用基于 Python 的 jieba 分词。

3.2.2 留言词典库

经过 python 的 jieba 分词后，留言变成了一个一个的词语。但是直接采用 jieba 分词，分词的效果达不到最优。例如：“安全隐患”该词足足出现了 67 次，分词结果为“安全”、“隐患”，这会增加特征数量。因此本文考虑词汇间的上下语义关系，对上下文共现词进行合并，得到留言词典库，从而提高后续分类模型效果。

本文规定：如果两个特征词连续出现在数据集的次数大于等于 5，则称两个特征词为共现词。具体算法步骤如下：

输入：文本集 $H = \{S_k | k = 1,2,\ldots,N\}$

输出：新词 $D = \{D_{ij}\}$

如果 $frequency(S_{xy}) > 2$

那么 $D = D + S_{xy}$

3.2.3 去停用词

在文本处理中，为了为节省存储空间和提高搜索效率，在处理文本之前会过滤掉某些字或词。过多的停用词会对文本的有效信息进行干扰，使得我们的模型效果变差。通常来说，去停用词分为两大类：1. 在文本中出现频率高，但其实无实际意义的词语，例如“问题”等词语。2.使用过于广泛的词语，例如“我们”、“是”等词语。3.文章中的标点符号。停用词库示例如图 2：

不怕 不惟 不成 不拘 不择手段 不敢 不料 不断 不日 不时 不是 不曾 不止 不止一次 不比
不消 不满 不然 不然的话 不特 不独 不由得 不知不觉 不管 不管怎样 不经意 不胜 不能
不至于 不若 不要 不论 不起 不足 不过 不迭 不问 不限 与 与其 与其说 与否 与此同时
专门 且 且不说 且说 两者 严格 严重 个 个人 个别 中小 中间 丰富 串行 临为 为主
为了 为什么 为什麼 为何 为止 为此 为着 主张 主要 举凡 举行 乃 乃至 乃至于 么 之

图 2 停用词示意图

去除后的留言信息如图 3 所示：

留言编号	留言用户	留言主题	留言时间	留言详情	一级标签	jiebacutterstopwords
24	A00074011	A市西湖建	2020/1/6	A3区大道西行便	城乡建设	['A', '市A 市 西湖 建筑 集团 占 道 施工 安全隐患
37	U0008473	A市在水一	2020/1/4	位于书院路主干	城乡建设	['A', '市A 市 在水一方 大厦 人为 烂尾 多年 ， 安全隐患
83	A0006399	投诉A市A1	2019/12/3	尊敬的领导：A1	城乡建设	['投诉', 投诉 A 市 A1 区苑 物业 违规 收 停车费
303	U0007137	A1区蔡锷南	2019/12/6	A1区A2区华庭小	城乡建设	['A1', '[A1 区 蔡锷 南路 A2 区华庭 楼顶 水箱 长年 不洗
319	U0007137	A1区A2区华	2019/12/5	A1区A2区华庭小	城乡建设	['A1', '[A1 区 A2 区华庭 自来水 好大 一股 霉味
379	A0001677	投诉A市盛	2019/11/2	我在2015年购买	城乡建设	['投诉', 投诉 A 市 盛世 耀凯 小区 物业 无故 停水
382	U0005806	咨询A市楼	2019/11/2	由于西地省地区	城乡建设	['咨询', 咨询 A 市 楼盘 供暖 一事

图 3 去停用词后效果示意图

3.3 文本特征选取

词频（term frequency, TF）：指某个词 w 在这个条目 i 中出现的频率，即该词

出现的次数占该条目总词数的比例，公式：

$$TF_{wi} = \frac{\text{第 i 条目中 w 词的个数}}{\text{第 i 条目总词数}}$$

虽然一些词的词频很高，但是这些词并不是这个条目的关键词汇，因为在其他条目中也出现了很多次，使得这些词汇在条目中的辨识度不高，因此需要计算逆向文本频率来说明词汇的重要性。词权重越大，说明词对主题的辨识度越高，所以引入逆向文件频率。

逆向文件频率 (inverse document frequency, IDF)：包含 w 词的条目数越高，权重越低，公式如下：

$$IDF_w = log(\frac{\text{总条目数}}{\text{含有 w 词的条目数} + 1})$$

分母加上 1，防止分母为 0。

所以词汇在该条目中是高频词语，但是在整个数据集中是低频词语可以产生较高权重的 TF-IDF,公式如下：

$$TF - IDF = TF * IDF$$

通过这个方法可以过滤掉不重要的词汇。

3.4 文本表示

文本表示利用了 CountVectorizer 原理，主要思想是把文本转化为向量矩阵 A。首先对每个整个数据集出现的词汇进行编号，再通过 TF-IDF 值算出每个词的权重，比如第 i 个条目下 w 词对应的编号是 m，w 词的权值是 n，那么有：

$$A_{im} = n$$

3.5 模型性能评估

3.5.1 模型性能评价指标

文本分类中，通常采用查准率 P、查全率 R 以及 F1 值的评价标准。以二分类为例：

表 2 二分类结果

模型预测 / 真实情况	TRUE	FLASE
TRUE	TP	FN

FALSE	FP	TN

① 查准率 P

查准率 Pi 为预测为 i 类，实际也为 i 类的比例。即模型的准确度。查准率公式如下：

$$P = \frac{TP}{(TP + FP)}$$

② 查全率/召回率 R

查全率为 Ri 为实际为 i 类，预测也为 i 类的比例。查全率公式如下：

$$R = \frac{TP}{(TP + FN)}$$

③F_1值

F 值的计算公式如下：

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

3.5.2 K-折交叉验证

由于数据集划分的随机性，将数据集简单分为测试集和训练集往往不能真实反映分类模型的准确性，模型的准确度会随着样本抽取的不同而不同。K-折交叉验证可用于提高模型泛化能力。本文采用 10-折交叉验证。具体过程如下：

第一步，将初始数据不重复抽样，随机分为 10 份。

第二步，每次挑选其中 1 份作为测试集，剩余 9 份作为训练集用于模型训练。

第三步，第二步重复 10 次，每个子集都有一次机会作为测试集，其余时候均作为训练集。

第四步，在每个训练集上训练得到一个模型，测试集上该模型测试，计算并保存模型的评价指标。

第五步，计算 10 组测试结果的平均值作为模型精度的估计。

3.6 常见文本算法对比试验

对于分类模型建立上，小组考虑模型准确度、计算时间等多个因素，分别采用SVM 线性分类模型、感知器模型、朴素贝叶斯模型等多个模型，再进行集成学习，综合比较，以期望得到一个更优的模型。

3.6.1 SVM 分类模型建立及结果

支持向量机是一类按监督学习方式对数据进行二元分类的广义线性分类器。对于二分类问题，若实例 X 只具备两个属性A_1和A_2，如图 4 所示，可以看出该二维数据是线性可分，可以通过画一条直线将两类分隔开。

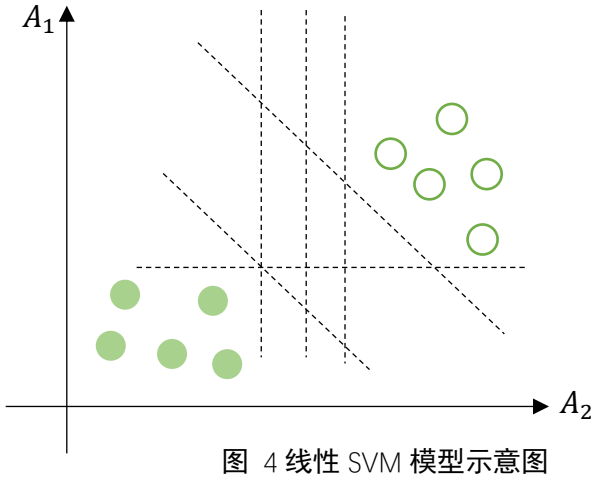

图 4 线性 SVM 模型示意图

图 4 中，可以画出无限多条直线将两个类别分隔开。对于二维的样本，就是寻找最好的一条直线将两个类分开； 对于多维样本，就是寻找最佳的分离平面；推广到 n 维样本，就是寻找最佳的决策边界——一个超平面。

分离超平面记作：

$$W \cdot X + B = 0$$

其中 W 是权重向量，$W = \{W_1, W_2, \ldots, W_N\}$；n 是维度；b 是标量，称为偏置（bias）。假定 X=（X_1,X_2），X_1和X_2是 X 在属性A_1和A_2观测值。因此，可以将分离超平面写作：

$$H_1：W_0 + W_1X_1 + W_2X_2 \geq 1，\ 对于y_i = +1$$

$$H_2：W_0 + W_1X_1 + W_2X_2 \leq 1，\ 对于y_i = -1$$

即落在H_1上面的实例，类别为+1，落在H_2下面的实例，类别为-1，结合上述两式，得到综合式子如下：

$$y_i(w + wx + wx) \geq 1，\ \forall i$$

在 SVM 中，不仅需要找到最大分离超平面，并且该超平面要求最大化地将类别分离。超平面到 H_1上任意点的距离是$\frac{1}{\|w\|}$，到H_2上任意点的距离也是$\frac{1}{\|w\|}$，因此，超平面最大边缘为$\frac{2}{\|w\|}$。

在实际问题中，SVM 分为线性可分与线性不可分两种情况，本节采用线性 SVM 模型进行计算模拟，对于非线性问题，需引入核函数，将非线性可分问题从原始的特征空间映射至更高维的希尔伯特空间，从而转化为线性可分问题，在本文 3.7 进行讨论。

本节中，线性 SVM 分类效果如表 3 所示：

表 3 线性 SVM 模型分类结果

分类器 / 类别	线性 SVM		
	查准率	召回率	F_1值
城乡建设	0.91	0.81	0.85
环境保护	0.90	0.90	0.90
交通运输	0.88	0.83	0.85
教育成本	0.87	0.83	0.85
劳动和社会保障	0.77	0.84	0.81
商贸旅游	0.87	0.96	0.89
卫生卫计	0.91	0.86	0.88
平均性能	0.873	0.861	0.861

3.6.2 朴素贝叶斯模型建立及结果

贝叶斯分类是一种基于贝叶斯定理的分类模型。朴素贝叶斯是一种简单的贝叶斯分类法，它假定一个类别下的各个属性相互独立，互不影响。

朴素贝叶斯分类的工作原理如下：

对于给定的文本集合，其中每一个文本实例使用一个 n 维度的属向向量 X 表示，X=($x_1, x_2, \dots, x_n$)给定的类别集合为 C=($c_1, c_2, \dots, c_n$),对于给定的文本实例，朴素贝叶斯分类法将 X 预测为在条件 X 下具有最大后验概率的类别 ci。当且仅当

$$P(c_i|X) > P(c_j|X) \quad 1 \le j \le m, j \ne i$$

得到最大的后验概率 P(ci|X)。

根据贝叶斯定律可知，

$$P(c_i|X) = \frac{P(X|c_i)P(c_i)}{P(X)}$$

由于对于给定的数据集合，P(X)一定，因此需要最大化$P(X|c_i)P(c_i)$，而$P(c_i)$可以通过计算得出，所以，只需最大化$P(X|c_i)$即可。

一般而言，文本信息的特征较多，计算 P（Xi|ci）较为复杂，因此朴素贝叶斯分析假定了属性之间条件独立。

$$P(X|c_i) = \prod_{k=1}^{n} P(x_k|c_i) = P(x_1|c_i)\,P(x_2|c_i)\cdots P(x_n|c_i)$$

经过模型计算后，我们得到朴素贝叶斯分类器结果如表 4 所示：

表 4 朴素贝叶斯模型分类结果

分类器 / 类别	朴素贝叶斯		
	查准率	召回率	F_1值
城乡建设	0.45	0.56	0.50
环境保护	0.72	0.67	0.69
交通运输	0.68	0.69	0.68
教育成本	0.60	0.67	0.63
劳动和社会保障	0.72	0.60	0.65
商贸旅游	0.67	0.72	0.69
卫生卫计	0.68	0.75	0.71
平均性能	0.646	0.666	0.65

3.6.3 感知器分类模型建立与模型结果

感知机是一种线性分类模型，输入为数据的特征向量，输出为数据的类别。模型希望将输入空间中的数据用一张超平面分离开。为求出该超平面，建立了基于误分类的损失函数，利用梯度下降法对损失函数进行最优化。

假设输入空间记为 x，权重记为 w，偏置项记为 b。则模型可表达为：

$$f(x) = sign(w * x + b)$$

上面公式中的$sign()$是激活函数，当输入的数值大于 0 时，函数值为+1；当输

入的数值小于 0 时，函数值为-1。

假设训练数据集是线性可分的，感知机学习的目标就是求得一个能够将训练数据集中正负实例完全分开的超平面，为了找到这个超平面，即确定感知机模型中的参数 w 和 b，需要定义一个损失函数并通过将损失函数最小化来求 w 和 b。模型中选择的损失函数是误分类点到分类超平面 S 的总距离。输入空间中任一点x_0到超平面 S 的距离为：

$$\frac{1}{||w||}|wx_0+b|$$

其中，||w||为 w 的 2--范数。

为方便求解，固定||w||为常数，并将其略去。同时对所有错误分类的点求和。得到最终的损失函数为：

$$L(w,b)=-\sum y(wx+b)$$

感知机的学习过程即是求解上述最优化问题的过程。最优化的方法是随机梯度下法，这种算法的基本思想是：当一个实例点被误分类，即位于分类超平面错误的一侧时，则调整 w 和 b，使分类超平面向该误分类点的一侧移动，以减少该误分类点与超平面的距离，直到超平面越过该误分类点使其被正确分类为止。

同时可以从数学理论上证明，而对于线性可分的数据集，算法一定是收敛的，即经过有限次迭代，一定可以得到一个将数据集完全正确划分的分类超平面及感知机模型。

感知器模型分类结果如表 5 所示：

表 5 感知器模型分类结果

分类器 / 类别	感知器		
	查准率	召回率	F_1值
城乡建设	0.81	0.87	0.84
环境保护	0.86	0.91	0.88
交通运输	0.81	0.84	0.83
教育成本	0.81	0.76	0.78
劳动和社会保障	0.82	0.80	0.81
商贸旅游	0.88	0.88	0.88

卫生卫计	0.87	0.83	0.85
平均性能	0.837	0.841	0.839

3.6.4 结果分析

将上述三种分类模型结果汇总如图 5 所示：

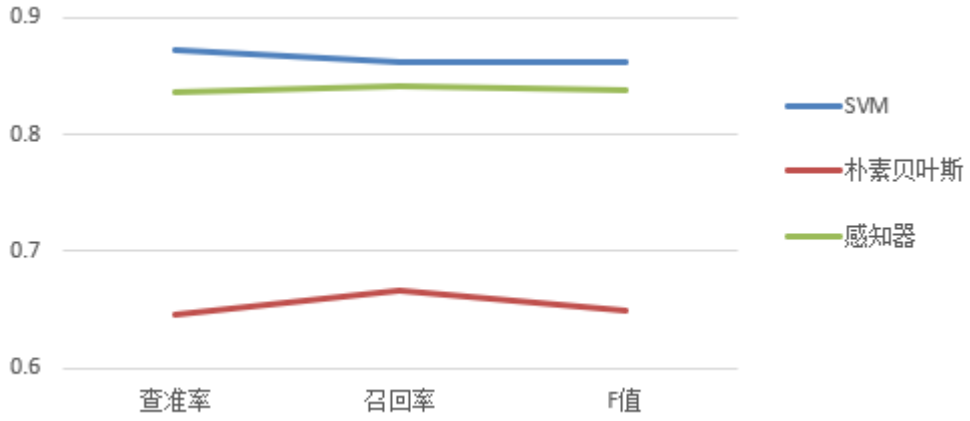

图 5 三种分类器性能对比图

虽然三种分类器通过调整参数都存在优化的可能，但是在一般条件下，橡树比较也能说明问题。感知器模型性能最弱，朴素贝叶斯模型与 SVM 模型相差不大，查准率、召回率、F_1值上，均是 SVM 更优，因此，三种分类器中，线性 SVM 分类器效果最好。因此采用 SVM 模型分类是一个较好的选择。

3.7 不同核函数的 SVM 效果探索

在本文 3.5.1 中采用了线性 SVM 模型，为此，小组继续探究 SVM 线性不可分的情况，在本节中，引入不同核函数，对比试验效果。

对于线性不可分的情况，如图 6 所示，

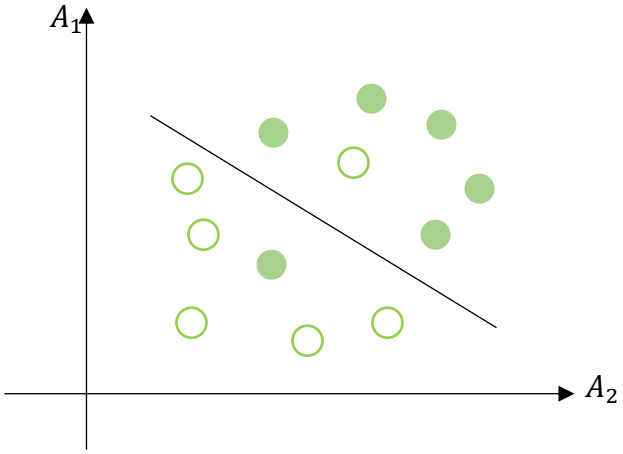

图 6 线性不可分 SVM 示意图

可以看出，图中有一个实心圆圈和空心圆圈无法正确分类，为解决该问题，可

以引入二次曲线，形式如下：

$$g(x) = c_0 + c_1 x + c_2 x^2$$

也可写为：

$$g(x) = a^T y = \sum_{i=1}^{s} a_i y_i$$

其中，

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix},\ a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix},$$

该方法可以解决 SVM 线性不可分的问题，同时也带来了新的问题。将原来的数据变换到新的高维空间，在新的高维向量空间里寻找最大分离超平面。维数的增加，使得计算量大大增加。在预测未知类别的实例时，必须计算每个支持向量的点积，在训练支持向量机的时候也会多次计算点积。由于计算量巨大，因此引入核函数代替点积运算。

SVM 模型中，常见的核函数有线性核函数、RBF 核函数、PUK 核函数和 Sigmoid 核函数等，具体形式如下：

线性核函数：$K(x_i, x_j) = \langle x_i \cdot x_j \rangle$

Poly 核函数（多项式核函数）：$K(x_i, x_j) = (\langle x_i \cdot x_j \rangle + 1)^q$

RBF 核函数（径向基核函数）：$K(x_i, x_j) = e^{-r\|x_i - x_j\|^2}$

本文分别采用上述三种核函数的 SVM 模型进行分类，比较性能。

3.7.1 线性核函数的 SVM 分类结果

Linear 核函数的分类结果如表 6 所示：

表 6 SVM-Linear 核函数分类结果

分类器 / 类别	Linear kernel		
	查准率	召回率	F_1 值
城乡建设	0.89	0.82	0.85
环境保护	0.85	0.91	0.88
交通运输	0.86	0.82	0.84

教育成本	0.87	0.77	0.82
劳动和社会保障	0.79	0.89	0.83
商贸旅游	0.89	0.85	0.87
卫生卫计	0.94	0.85	0.9
平均性能	0.87	0.844	0.856

3.7.2 RBF 核函数的 SVM 分类结果

RBF 核函数的分类结果如表 7 所示：

表 7 SVM-RBF 核函数分类结果

分类器 / 类别	RBF kernel		
	查准率	召回率	F_1 值
城乡建设	0.72	0.71	0.71
环境保护	0.77	0.73	0.75
交通运输	0.6	0.8	0.68
教育成本	0.63	0.62	0.62
劳动和社会保障	0.49	0.73	0.59
商贸旅游	0.91	0.53	0.67
卫生卫计	0.91	0.49	0.64
平均性能	0.719	0.659	0.666

3.7.3 Sigmoid 核函数的 SVM 分类结果

Sigmoid 核函数的分类结果如表 8 所示：

表 8 SVM-sigmoid 核函数分类记过

分类器 / 类别	Sigmoid kernel		
	查准率	召回率	F_1 值
城乡建设	0.15	0.92	0.26
环境保护	0.69	0.67	0.68
交通运输	0.85	0.41	0.55
教育成本	0.46	0.63	0.54
劳动和社会保障	0.8	0.18	0.29
商贸旅游	0.95	0.3	0.46
卫生卫计	0.91	0.53	0.67

平均性能	0.687	0.52	0.493

3.7.4 结果分析

将非线性可分的不同核函数的 SVM 模型结果与线性可分 SVM 结果对比，结果如图 7 所示：

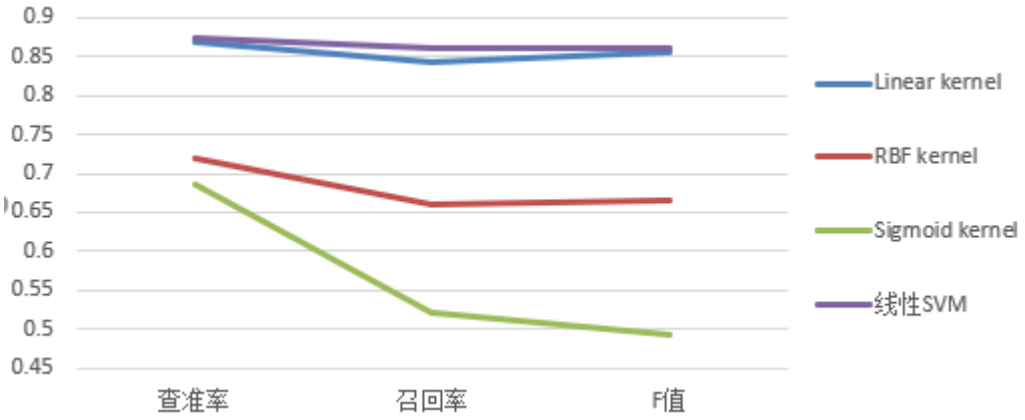

图 7 不同核函数的 SVM 分类效果对比图

由图中可以看出，Linear kernel 的 SVM 与线性可分 SVM 的分类效果最好，而 RBF kernel 与 Sigmoid kernel 的 SVM 模型效果较差。

3.8 SVM 集成算法探索

根据上述结果，文本发现，对于留言详情分类问题，线性 SVM 的分类效果更好，基于上述结果，本文继续探究集成算法的 SVM 模型。

集成学习是一种显著提高分类器模型效果的基础。它由多个分类器组合在一起，形成复合模型，而达到提高模型泛化能力的目的。集成学习方法如图 8 所示：

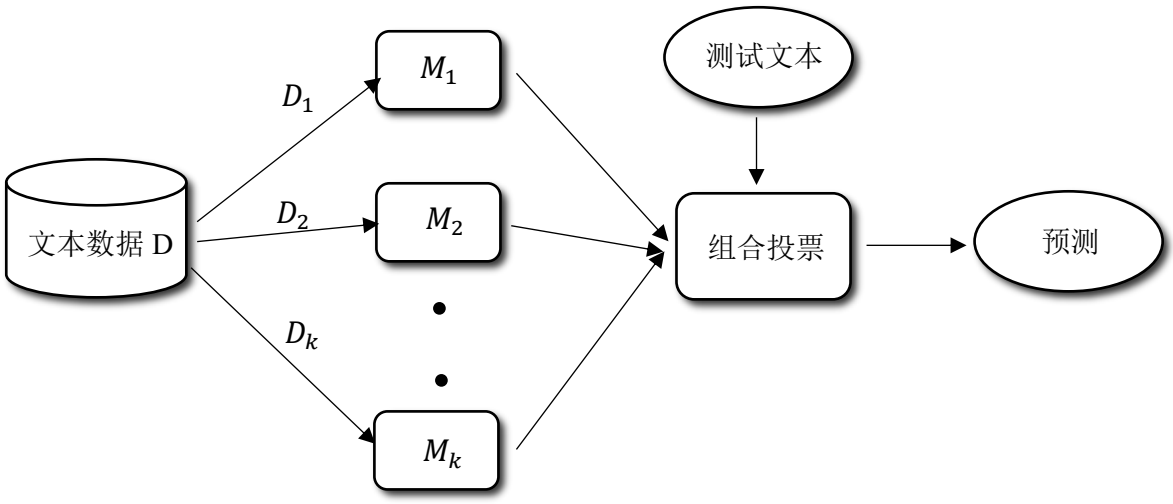

图 8 集成学习框架图

3.8.1 SVM-bagging 集成算法

bagging 是一种代表性的集成学习。它的基本思想是：从原始样本集 D 中有放回的抽取训练集$D_i(i = 1,2,3 \dots k)$，每个训练集D_i相互独立。训练集 D_i经过学习得到分类模型 M_i，共得到 k 个模型。对于本文留言详情分类问题，基分类器M_i对文本实例 X 进行分类，并返回其预测结果作为一票，bagging 分类器对 X 的分类结果统计票数，将票数最高的类标号作为 X 的类别。

本文训练$K = 20$，$K = 50$，$K = 100$时的 SVM-bagging 模型，模型分类结果如表 9：

表 9 k 不同取值的 SVM-bagging 模型分类结果

类别	K=20			K=50			K=100		
	查准率	召回率	F1 值	查准率	召回率	F1 值	查准率	召回率	F1 值
城乡建设	0.89	0.84	0.87	0.93	0.83	0.87	0.91	0.84	0.87
环境保护	0.88	0.89	0.88	0.90	0.92	0.91	0.87	0.9	0.89
交通运输	0.86	0.84	0.85	0.90	0.81	0.86	0.94	0.8	0.86
教育成本	0.88	0.8	0.84	0.87	0.82	0.84	0.83	0.77	0.8
劳动和社会保障	0.8	0.86	0.83	0.79	0.89	0.84	0.8	0.87	0.84
商贸旅游	0.88	0.9	0.89	0.92	0.87	0.89	0.88	0.89	0.88
卫生卫计	0.86	0.82	0.84	0.91	0.89	0.90	0.87	0.83	0.85
平均性能	0.864	0.85	0.857	0.889	0.861	0.873	0.871	0.843	0.856

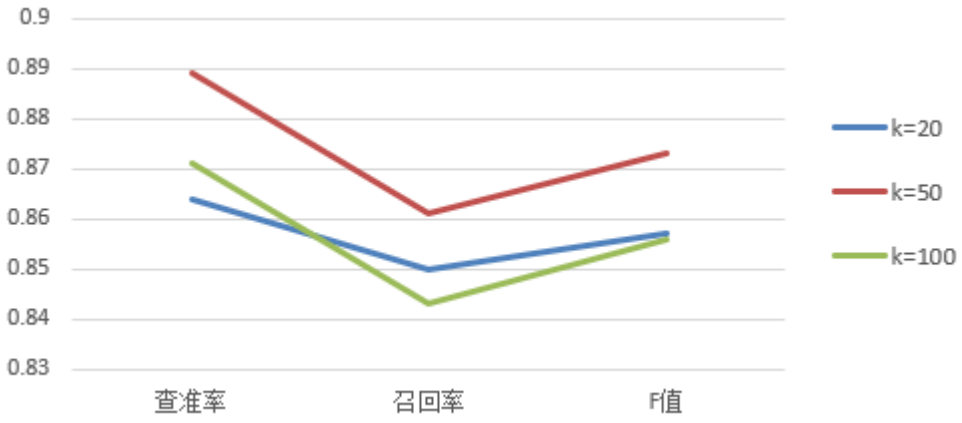

图 9 不同 k 取值的 SVM-bagging 模型分类结果对比图

从图 9 可以看出，k=50 时，SVM-bagging 效果更好，因此，在进行后续分析时，采用 k=50 时 SVM-bagging 分类。

3.8.1 SVM-AdeBoost 集成算法

Boosting 算法，是一种传统的集成学习算法。简而言之，它的思想是：创建一个模型进行分类预测，然后找到被错误分类的实例。在这些实例上加权来为下一轮建立新的模型创建训练数据集，因此后面的模型总是基于前面的模型而建立的。

Boosting 算法在解决实际问题时有一个重大的缺陷，即他们都要求事先知道弱分类算法分类正确率的下限。在此基础上，便有了 AdaBoost。

AdaBoost 它的思想是：给定包含 d 个主题的训练数据集 D，每个训练文本实例 X 赋以相同的权重 1/d，迭代 k 次产生 k 个基分类器。在第 i 次迭代时，从数据集 D 中抽出样本大小为 d 的训练集D_i。对D_i进行学习，得到基分类器M_i，使用D_i作为测试集计算基分类器M_i的误差。若样本被错误地分类，则该实例入样的权重会增加；若正确分类，该实例入样权重减少。接下来使用这些权重为下一次迭代的分类器产生训练样本。

与 bagging 相比，AdaBoost 侧重误分类的实例，通过错误分类的实例，调整实例入样的概率，以升模型效果。AdaBoost 算法分类效果如表 10 所示:

表 10 SVM-AdaBoost **模型分类结果**

分类器 / 类别	SVM-AdaBoost		
	查准率	召回率	F_1值
城乡建设	0.94	0.85	0.89
环境保护	0.87	0.91	0.89
交通运输	0.93	0.79	0.86
教育成本	0.83	0.8	0.81
劳动和社会保障	0.78	0.87	0.82
商贸旅游	0.88	0.84	0.86
卫生卫计	0.87	0.85	0.86
平均性能	0.871	0.844	0.856

3.8.2 结果分析

将 K=50 的 SVM-bagging、SVM-AdeBoost 和线性 SVM 三种模型的分类结果进行比较，结果如图 10 所示：

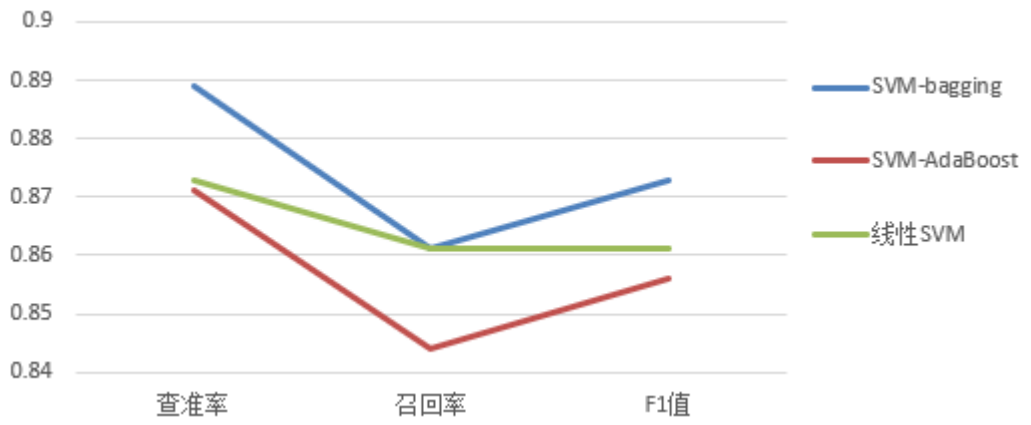

图 10 不同集成算法 SVM 模型结果

从图 10 看出，总的来说，SVM 与它的两种集成算法 bagging 和 AdaBoost 的模型效果并没有很大差别，从这里看出，SVM 本身的分类效果较好。理论上来说，SVM-bagging 比 SVM 的效果更好，泛化能力更强，表现更稳定，实验数据也验证了这一点。然而，SVM-AdaBoost 的性能却下降了，本文认为，这可能是因为 SVM-AdaBoost 过分关注错误分类实例，而这些错误分类实例本身就是不易分类的，因此发生了过拟合，模型性能下降。

从 3.6 可以看出。SVM-Bagging 算法在留言文本分类中的综合表现优于单个 SVM 分类器和 SVM-AdaBoost 集成算法。

3.9 基于 SVM-Bagging 集成的投票对比探究

基于上述层层探索，不断深入，本文得到一个想法：首先训练得到不同核函数的 SVM 分类器，采用 bagging 进行装袋，最后采用投票的方式进行预测。用 Vote1 代表 SVM、采用 linear kernel 的 SVM-Bagging 和 SVM-AdaBoost 组合分类器，Vote2 表示不同核函数的 SVM-Bagging 的组合分类器。Vote1 结果如表 11 所示：

表 11 vote1 分类效果表

分类器 / 类别	Vote1		
	查准率	召回率	F_1 值
城乡建设	0.95	0.86	0.90
环境保护	0.87	0.91	0.89
交通运输	0.86	0.86	0.86
教育成本	0.88	0.76	0.82

劳动和社会保障	0.81	0.89	0.85
商贸旅游	0.90	0.90	0.90
卫生卫计	0.94	0.85	0.89
平均性能	0.887	0.861	0.873

Vote2 结果如表 12 所示：

表 12 vote2 分类效果表

分类器 / 类别	Vote2		
	查准率	召回率	F_1 值
城乡建设	0.61	0.84	0.70
环境保护	0.71	0.90	0.79
交通运输	0.86	0.72	0.78
教育成本	0.82	0.67	0.74
劳动和社会保障	0.69	0.72	0.71
商贸旅游	0.86	0.73	0.79
卫生卫计	0.91	0.67	0.77
平均性能	0.78	0.75	0.754

从表 11 和表 12 可以看出，vote1 分类效果明显高于 vote2，选择基分类器最好是互补，且性能也较好的。

3.10 模型分类结果

将线性可分 SVM 与 SVM-bagging、vote1、对留言文本分类效果进行比较，结果如图 11 所示：

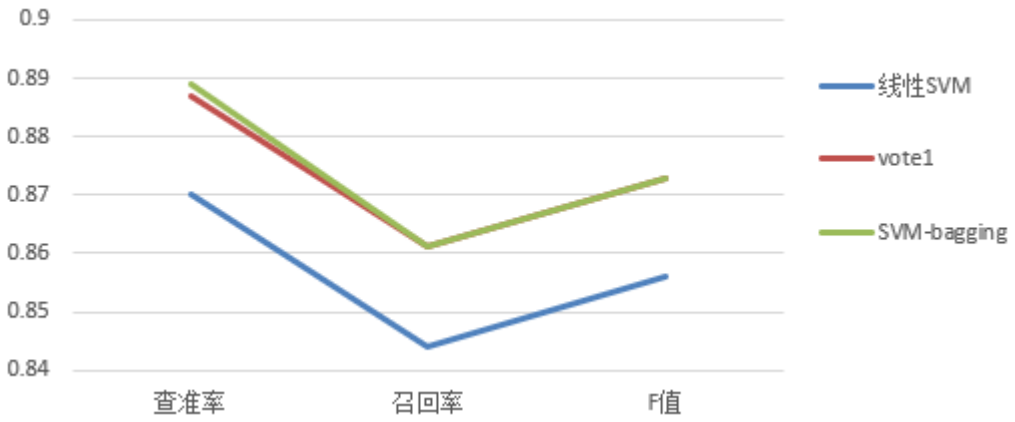

图 11 模型对比图

从图中可以看出，SVM 分类模型本身性能较优，模型性能已经达到查准率 87%左右。从三个指标来看，分类效果最好的是 SVM-bagging，k=50 时，具体分类结果见附件 4。

第4章 热点问题探索分析

4.1 任务二分析框架

任务二需要对热点问题进行挖掘，任务二分析框架如图 12 所示。首先对留言进行预处理，包括分词与去停用词。为增加后续聚类的准确度，建立留言词典库。对处理后数据采用 LDA 主题模型，计算出每条留言的概率密度后，采用 JS 距离作为文本相似度计算。使用流式聚类算法 Single-Pass 聚类对留言进行聚类，得到热点问题。构建留言热点评价指数，对热点问题进行排序。

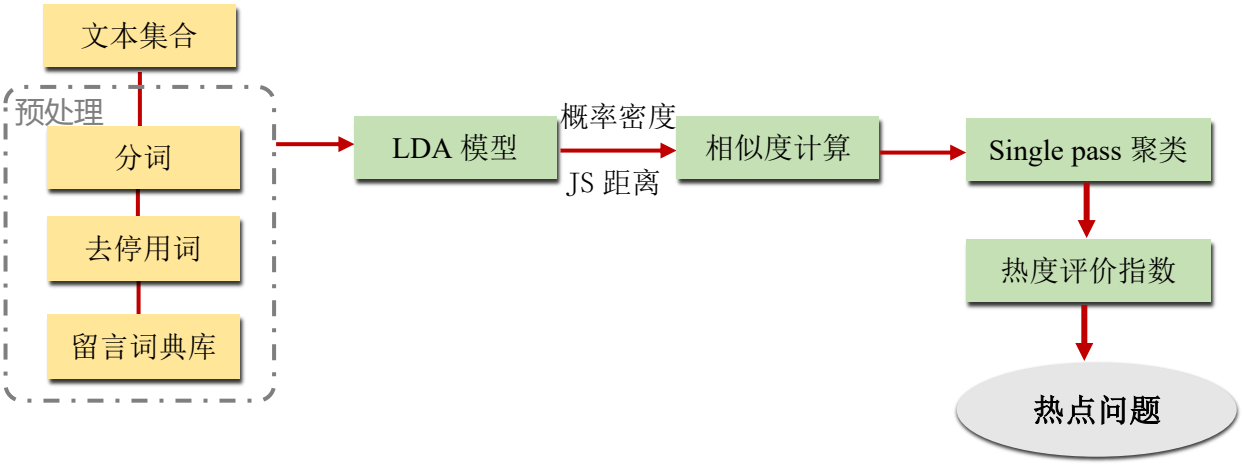

图 12 任务二分析框架

4.2 数据预处理

对于问题文本的预处理与任务一中预处理步骤相同，在这里不再赘述，具体预处理结果请看附件 1。

4.3 LDA 主题模型建立

LDA 是一种文档主题生成模型，也被称为一个“词-主题-文档”的三层贝叶斯概率模型。LDA 模型可以对文本集合进行大规模降维处理，从而识别文本中的主题信息。它将每一个文本信息看作一个词频向量，一篇文章的每个词都是以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语。因此，每一篇文本代表一些主题的概率分布，每一主题也代表着大量词汇组成的概率分布。

LDA 主题模型表示如图 13 所示：

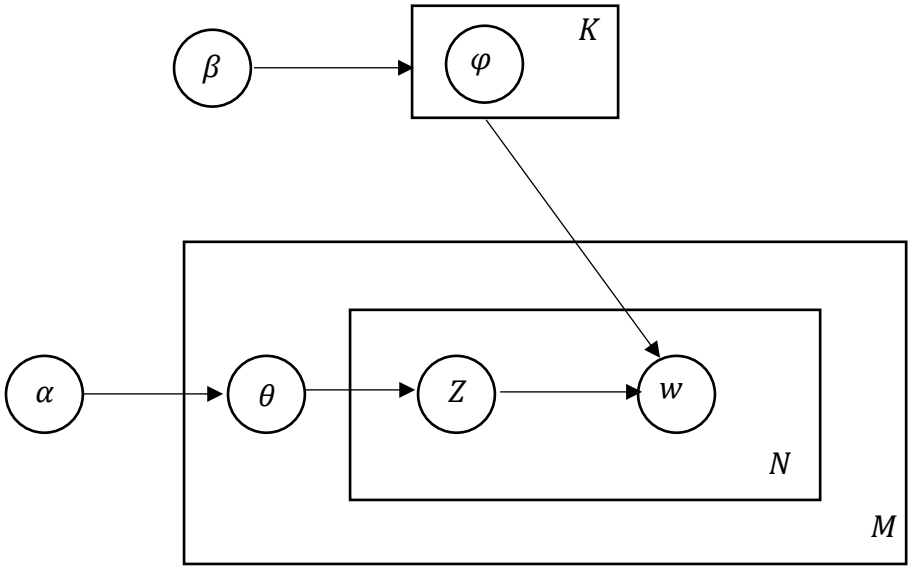

图 13 LDA **主题模型**

图 13 代表的概率模型是：

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta)$$

其中，Z 表示潜在主题，α和β是 LDA 模型中给的 Dirichlet 先验分布，N 表示要生成的文档的单词的个数，W_n表示生成的第 n 个单词 w，α是文本集中含有的所有主题分布的先验，β是所有主题中含有的全部词汇分布的先验，θ_d代表文本 d 中包含的所有主题的多项式分布。

4.3.1 主题数确定

主题数 T 对 LDA 主题模型结果的好坏有着重要作用。通常情况下，随意设定主题数 T，模型效果不会最优。因此便需要提前确定最优的主题数 T。本文采用模型困惑度确定最优主题数 T。

模型困惑度可以理解为：对于给定文本 d，模型判断文档 d 属于 Ti，判断的不确定性就是模型困惑度。因此困惑度度越低，聚类的效果越好。模型困惑度计算公式如下：

$$perplexity = e^{\frac{-\sum log(p(w))}{N}}$$

本文利用取不同主题数 T，观察困惑度 P 的变化，确定最优主题数。其中横坐标表示主题数，纵坐标表示模型困惑度，结果如图 14 所示。

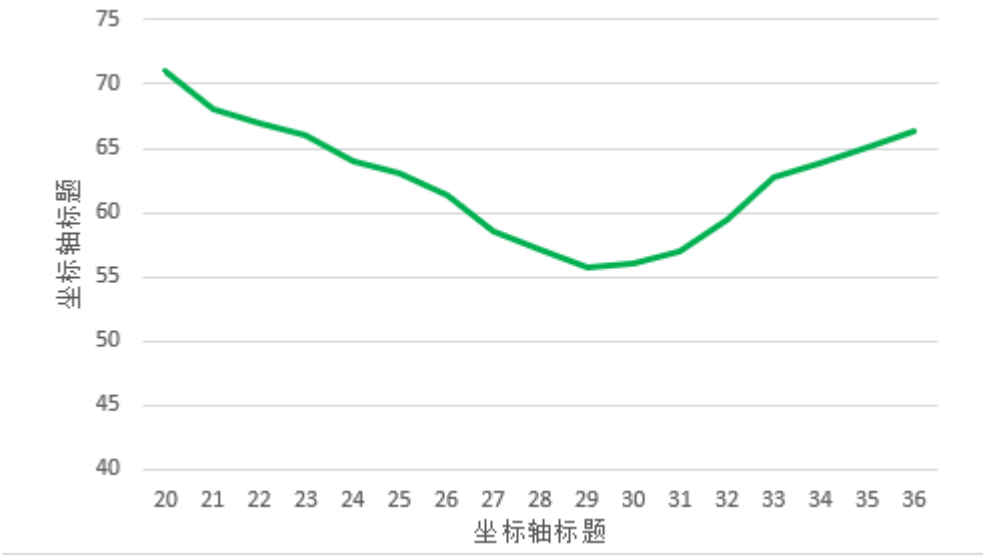

图 14 LDA 模型困惑度随主题数变化趋势

可以看出，当主题数为 29 时，模型困惑度最小，因此本文主题数选 29。

4.3.2 模型结果

LDA 主题模型部分结果如图 15，全部结果请看附件 7。

0 "A市","搅拌站","油烟味","A3区","商铺","A9市","A2区","噪音","杉树","部落","小区","空地","导致","新民","混凝土","浇筑","破坏","砂石","临街","社区","修建","坪镇","雨厂","营业","紫金","泉水","睡不着","吵得","底层","烟花","污水","乡村","小道","西湖路","全县","禁止","江西","地区","玫瑰园","贫水","浦镇","水土","九丰","餐厅","柴味","外滩","门面","二号线","满车","苏宁","饭店","A6区","新华联"

1 "A5区","污染","A市","学校","A2区","劳动","空气","东路","灰尘","解决","无证","噪音","街道","A8县","歌厅","恳请","变电站","暮云","问题","居民楼","扰民","A3区","搅拌站","小区","亚洲","菜市场","世贸","天城","环保部门","环境","房子","大道","白马","龙江路","蓉园","反应","政府","花园城","锅炉","A7县","热电厂","烟囱","天宁","老百姓","何谈","没水","住宅小区","西路","环境污染问题","举报","家属区","外环","A6区"

2 "A市","A8县","步伐","影响","搅拌站","力度","生活","米业","投诉","A9市","企业","居民","垃圾焚烧","饭店","塘村","油烟","望江","反映","野狗","校区","A7县","善化","小区","部门","失职","垃圾站","燃放烟花","爆竹","公寓","国际","经营","污染","光源","领地","散发","楷林","便利店","物流园","镇头镇","建材厂","金牌","A4区","一楼","下水道","修建","服务业","转型","升级","覆车","排到","排出","螺丝","华中"

3 "A1区","歌厅","A市","扰民","污染环境","环境污染","砖厂","黄花镇","宾馆","大厦","航道","回复","中学","轮胎厂","聚友","长房","合平路","沙场","搅拌站","隧道","飙车","电线杆","A5区","绿灯","设定","太短","白沙湾","靠近","分行","废品收购","医院","人民","专用","烟道","培训","培优","高压","A7县","西学","巷东","庙坪","乱倒垃圾","兴隆","架设","镇畔","花钱","D7县","广场","荷塘","丰国","高铁","音板","跳马"

图 15 LDA 模型部分结果

4.4 文本相似度计算

在利用 LDA 模型对文本进行建模后，文本被表示成主题下的概率分布，文本向量维数降低，因而两个文本的相似度可以通过文本间的主题相似性来反应。文本

的主题相似度是通过计算与文本对应的主题概率分布间的相似度来实现的。文本主题概率密度之间的差距越小，则文本间相似程度越大。

常用 KL 距离衡量两个概率分布的差异情况，也叫做相对熵。假设 P (x) 和 Q (x) 是 X 上的两个概率密度函数，它们之间的 KL 距离定义如下：

$$D_{KL}(P||Q) = \sum_{x\epsilon X} P(x)\, log\frac{1}{Q(x)} - \sum_{x\epsilon X} P(X)\, log\frac{1}{P(x)} = \sum_{x\epsilon X} P(x)\, log\frac{P(x)}{Q(x)}$$

由上式可知，KL *距离*具有不对称性，即$D_{kL}(P\|Q) \neq D_{kL}(Q\|P)$。

JS 距离是对 KL 距离的改进，将距离定义在[0,1]之间，因此，本文采用 JS 距离衡量文本间的相似度，JS 距离定义如下：

$$JS(z_i\|z_j) = \frac{1}{2}D_{kL}(z_i||\frac{z_i+z_j}{2}) + \frac{1}{2}D_{kL}(z_j||\frac{z_i+z_j}{2})$$

其中，z_i和z_j分别是文本d_i和d_j的概率密度。

综上，文本d_i和d_j间的主题相似度如下：

$$sim_{JS}(d_i, d_j) = JS(z_i\|z_j)$$

4.5 Singles-Pass 聚类

4.5.1 算法原理

Single Pass 聚类算法是一种流式的聚类算法，每个样本只会参与一次样本聚类，对样本的先后顺序有一定的依赖关系。简言之，对于某个未知新样本d_i，如果与现有的某个类足够相似，那么就放入这个类，否则自成一类。算法具体步骤如下：

Step1：第一篇输入的文本d_1作为初始问题类别。

Step2：计算该文本后的每一篇到达文本与已有问题类别 Ti 的主题相似度，

$$sim(d_k, T_i) = (sim(d_k, d_1) + sim(d_k, d_2) + \ldots + sim(d_k, d_h))/h$$

Step3：找到与d_i最大相似度的类别。

Step4：

① 如果$sim(d_k, T)$大于给定阈值ε，则将该文本分为此类。

② 反之，则新建立一个话题类别。

Step5：重复步骤 2、3、4，直到聚类结果。

4.5.2 聚类结果

部分聚类结果如图 16。

1	theme	class
2	A市A3区中海国际社区三期与四期中间空地夜间施工	0
3	A市万科魅力之城商铺无排烟管道，小区内到处油烟	1
4	A市万科魅力之城商铺无排烟管道，小区内到处油烟	1
5	A市魅力之城商铺无排烟管道，小区内到处油烟味	1
6	A市中航城三期复式楼34层高楼用1.2的栏杆作为生命	2
7	A市环保科技园园区振兴工业实体经济扶植奖励未按	2
8	投诉A2区丽发新城附近建搅拌站噪音扰民	3
9	投诉A2区丽发新城附近建搅拌站噪音扰民	3
10	投诉A2区丽发新城附近建搅拌站噪音扰民	3
11	A市丽发小区建搅拌站，噪音污染严重	4
12	A市丽发小区建搅拌站，噪音污染严重	4
13	A4区华章路路灯光源污染严重影响新领地小区居民生	5

图 16 single pass 聚类结果

4.6 热点问题汇总

4.6.1 热度评价指标定义

得到聚类结果后，定义热度评价指标。本文认为留言的点赞数与反对数都是该留言热度的体现，因此，第 i 个类的热度值定义如下：

$$H_i = 5n_i + \sum_{j=0}^{n_i}(a_j + b_j)$$

其中a_j为第 j 条留言的点赞数标准化后的数据，b_j为第 j 条留言的反对数标准化后的数据，n_i为第 i 类留言的条数标准化后的数据。

4.6.2 热点问题展示

将每类留言的点赞数与反对数代入计算，得到结果如表 13 和表 14：

表 13 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	0.667	2019/3/5至2020/1/7	A市地铁六号线	噪音扰民
2	2	0.374	2019/2/3至2019/4/22	A市	加快一圈二场三道建设
3	3	0.367	2019/3/28至2019/9/18	A市温斯顿英语培训学校	拖延退费
4	4	0.354	2019/9/8至2019/11/28	A市梅溪湖嘉顺苑安置小区	经常停水
5	5	0.303	2019/4/23至2019/9/12	A市五矿万境K9县	存在安全隐患

表 14 热点问题明细表

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
1	188399	A0009793	A市利保壹号公馆项目夜间噪声扰民	2019/7/3 6:23:25	您好，我想	0	2
1	188119	A0003502	对A市地铁违规用工问题的质疑	2019/5/27 16:04:44	我是一名在	0	0
1	196886	A0009548	A市地铁六号线人民东路站连续几个星期通宵施工了	2019/3/5 11:44:42	A市地铁六	0	2
1	199072	A0001110	A市新桥沁园夜宵街扰民严重	2020/1/7 22:15:02	我是新桥沁	0	2
1	213478	A0006245	A市地铁六号线施工噪音扰民	2019/8/28 10:06:25	我是保利麓	0	0
2	240662	A0003161	请加快A市一圈二场三道建设力度	2019/4/18 16:38:15	地处银杉路	0	2
2	237116	A0003161	请A市加快一圈二场三道建设力度	2019/2/3 14:15:28	作为人口密	0	2
2	275514	A0003161	请A市加快一圈二场三道建设力度	2019/4/22 17:29:22	地处银杉路	0	2
3	188467	A0005018	投诉A市温斯顿英语培训学校拖延退费！	2019/3/28 19:57:19	退费之日起	0	2
3	254068	A0005045	投诉A市温斯顿英语培训机构拖延退费	2019/9/18 18:19:41	本人2019年	0	3
4	197122	A0005541	A市梅溪湖嘉顺苑安置小区经常停水	2019/9/8 19:09:20	尊敬的各位	0	1
4	198372	A0001677	投诉A市盛世耀凯小区物业无故停水	2019/11/28 9:08:38	我在2015年	0	2
5	197453	A0001354	A7县好望谷胜益画室非法培训，存在安全隐患	2019/4/23 18:24:45	A7县贺龙体	1	0
5	215507	A0001032	A市五矿万境K9县存在严重的消防安全隐患	2019/9/12 14:48:07	预交房23栋	0	1

第5章　答复意见评价

5.1 任务三分析框架

任务三需要对构建评价体系对群众留言答复进行评价，任务三分析框架如图 17 所示。首先从答复意见的相关性、完整性、可解释性以及及时性构建 6 个指标。为便于后续分析，对指标进行标准化和正向化。接下来，采用熵值法计算指标权重。最后采用 TOPSIS 模型对答复意见质量进行量化。

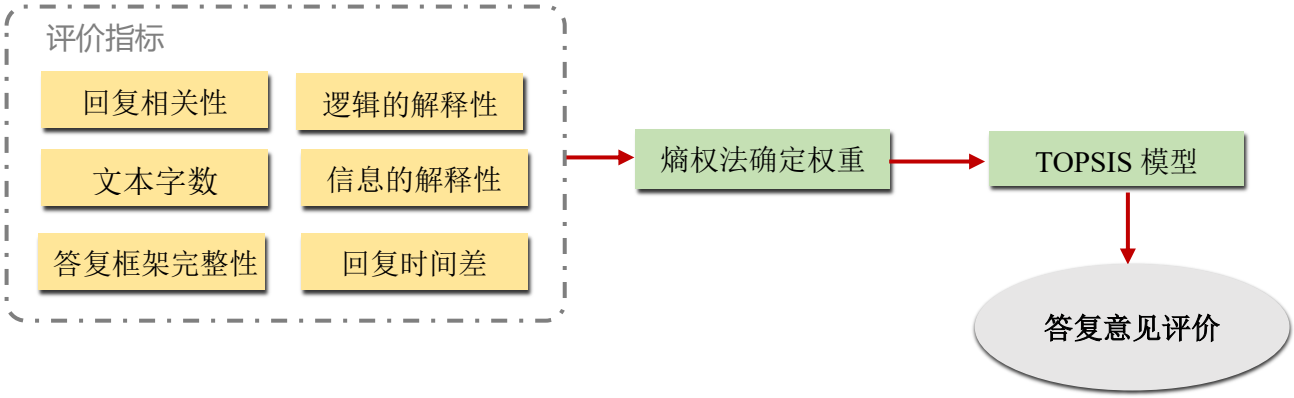

图 17 任务三分析框架

5.2 评价指标确定

对于任务三，本文从留言答复意见的相关性、完整性、可解释性以及及时性构建 6 个评价指标，指标体系如图 18 所示：

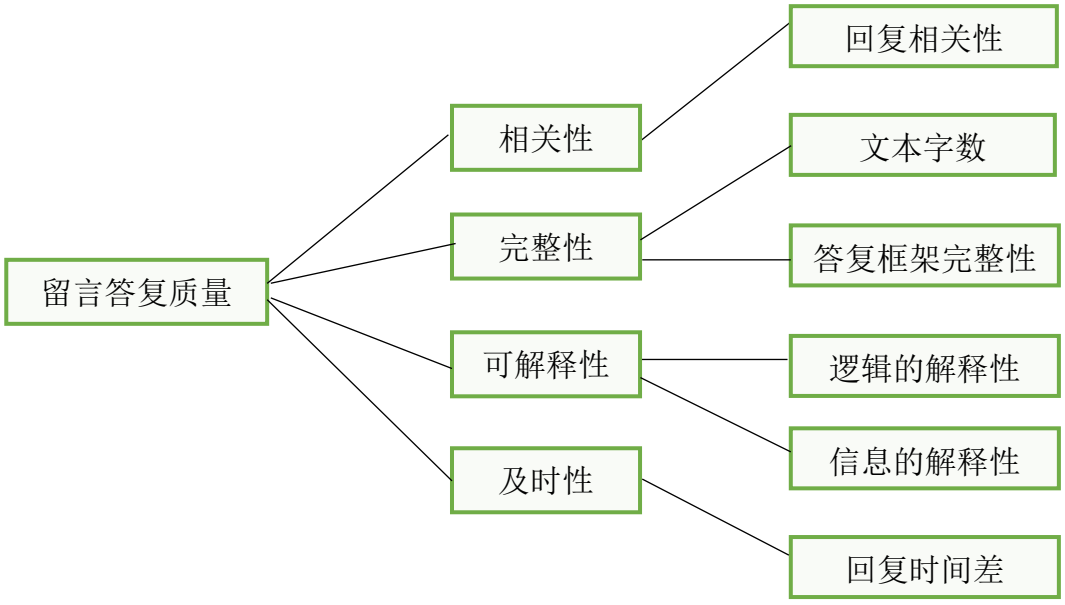

图 18 评价体系

1. **相关性：**相关性是留言答复与留言问题的是否高度相关的衡量，是否答非所问。一个相关性强的留言答复应该与留言问题紧密相扣。

① 回复相关性

本文采用任务二中使用的 JS 距离作为相关性度量，即群众留言与相关部门回复的 JS 距离。JS 距离越小，回复相关性越大。

2.完整性：一条合格的留言应内容完整，需要有问候语和结束语以及恢复内容。

②文本字数

本文认为，一般而言较为完整的留言答复字数不会太少，当回复字数较少时，小组认为该答复不完整。在全部的 2816 条留言答复中，最短的留言答复仅有 3 个字，最长的留言答复为 7883 个字，字数的多少会影响着答复的完整性。

③答复框架完整性

本文通过观察附件 4 中大量回复，小组认为较为完整的回复具有相似的回复框架，举例如下：

> **你好！**8 月 1 日你在平台的问政西地省版块发名为“G4 县档案局退休干部独生子女补贴多年不发”的帖子我局已**收悉**，我局领导非常重视，并召开了专会研究，现**答复如下**：档案局现在职 11 人，退休 9 人，属财政全额拨款，资金确实不足。近两年有 5 个退休干部需要发放独生子女补贴，该支出项目金额较大且由单位自筹解决，我们一直在积极筹措。今年初，局务会初步研究决定，2013 年一定要落实退休干部的独生子女补贴。所以，对 A000112959 网友发的“退休干部独生子女补贴”一事，我们将会在近期内解决，**恳请理解。**

图 19 回复框架举例

一个较为完整的留言，应具备四个部分：首先是问候词，例如“您好！”、“你好！”等词汇以表示答复的尊重；其次，重复群众提出的问题或者表达受到问题，例如“回复已收悉”、“收到回复”等词语；再次，相关部门应就提出问题作出一定回复，例如“答复如下”等词语：最后，相关部门应表达理解和感谢，例如“恳请理解”、“感谢您的留言”等词语。

因此小组根据附件 4，建立答复框架词库，根据具体答复是否有相关词以判断答复框架的完整性。

3.可解释性：对于一些留言而言，虽然相关部门已经回复，但是回复无逻辑，无法让留言者看懂，这也不会是一条好的留言。

④逻辑的解释性

通过一些连词或副词“因此”、“所以”等词语可以帮助增强回复的前后逻辑性和可解释性，除此之外，一些序数词也可以大大增加留言的条理性，例如“一、”、“首先”等词语。

> 西地省平台《问政西地省》栏目组： 民在贵栏目留言，咨询中央转移支付我省燃油税资金情况，现函复如下： 一、关于转移支付市州公路养护费用资金情况 2009 年燃油税费改革以来……二、关于中央转移支付我省燃油税资金增量情况 中央转移支付我省燃油税资金并不是网民所说的按 **5%至 8%**的基数固定递增，而是中央根据上年燃油税实际征收情况及当年燃油税资金增长情况酌情安排增量资金。2012 年中央转移支付我省的燃油税资金较 2011 年仅增长了 **2.32%。** 三、关于燃油税支付资金文件下发情况…….最后，感谢贵栏目组对西地省交通运输工作的关心。

图 20 解释性文本举例

⑤信息的解释性：

本文认为，回复留言时，提及相关数字可以大大增加留言的可解释性，让留言者较为直观的理解回复。除此之外，提及相关政策，例如“《西地省物业服务收费管理办法》”也可以有助于提升留言的可解释性。

4.及时性：相关部门需要尽快做出回复，减少群众等待时间。

⑥时间差：通过留言回复的时间差来衡量相关部门回复是否及时，一般而言，时间越短，回复越及时。

5.3 指标标准化

在进行评价时，一般需要将对指标属性进行正向化和标准化处理。在同一指标体系中，不同类型的原始数据无法直接表现被研究对象的优劣情况，也不利于进一步数据分析。通常采用成本型指标向效益型指标转化。

1.成本型指标：对于该类型指标而言，值越小越好，因此标准化形式为：

$$x'_{ij} = \frac{x_j^{max} - x_{ij}}{x_j^{max} - x_j^{min}}$$

2.效益型指标：对于效益型指标，值越大越好，因此标准化形式为：

$$x'_{ij} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}}$$

3.中性型指标：对于这类指标而言，值不宜过大，也不能过小，需要在一定区间$[s_1, s_2]$之间，因此标准化形式为：

$$x'_{ij} = \begin{cases} 1 - \dfrac{s_1 - \mathrm{d}_{ij}}{\max\{s_1 - d_j^{min}, d_j^{max} - s_2\}} & \mathrm{d}_{ij} < s_1 \\ 1 & \mathrm{d}_{ij} = s_1 \\ 1 - \dfrac{\mathrm{d}_{ij} - s_2}{\max\{s_2 - d_j^{min}, d_j^{max} - s_2\}} & \mathrm{d}_{ij} > s_1 \end{cases}$$

5.4 熵值法确定权重

基于信息论的熵值法是根据各指标所含信息有序程度的差异性来确定指标权重的客观赋权方法，仅依赖于数据本身的离散程度。在信息论中，熵是对不确定性的一种度量。信息量越大，不确定性就越小，熵也就越小；信息量越小，不确定性越大，熵也越大。

熵值法具体步骤如下：

Step1：对原始数据矩阵按列进行归一化处理，公式如下：

$$\mathrm{p}_{ij} = \frac{\mathrm{x}_{ij}}{\sum_{i=1}^{n} \mathrm{d}_{ij}}$$

Step2：计算各指标的熵值，公式如下：

$$\mathrm{e}_j = -k \sum_{i=1}^{n} \mathrm{p}_{ij}\, ln\mathrm{p}_{ij}\,, (j = 1,2,\ldots,m)$$

其中，k=$\frac{1}{\ln n} > 0$以及$e_j > 0$。

Step3：计算熵冗余，公式如下：

$$d_j = 1 - e_j$$

Step3：计算各指标权重，公式如下：

$$w_j = \frac{d_j}{\sum_{j=1}^{n} d_j}$$

通过结算，本文得到上述 6 个指标权重如表 15：

表 15 指标权重表

指标	权重
回复相关性	0.194
文本字数	0.096
答复框架完整性	0.351
逻辑的解释性	0.319
信息的解释性	0.039
回复时间差	0.002

从表 15 指标权重表中可以看出，答复框架的完整性和逻辑的解释性对留言回复质量的评价最为重要，其次是回复相关性。权重最小的是时间差，将时间差数据分布如图 21：

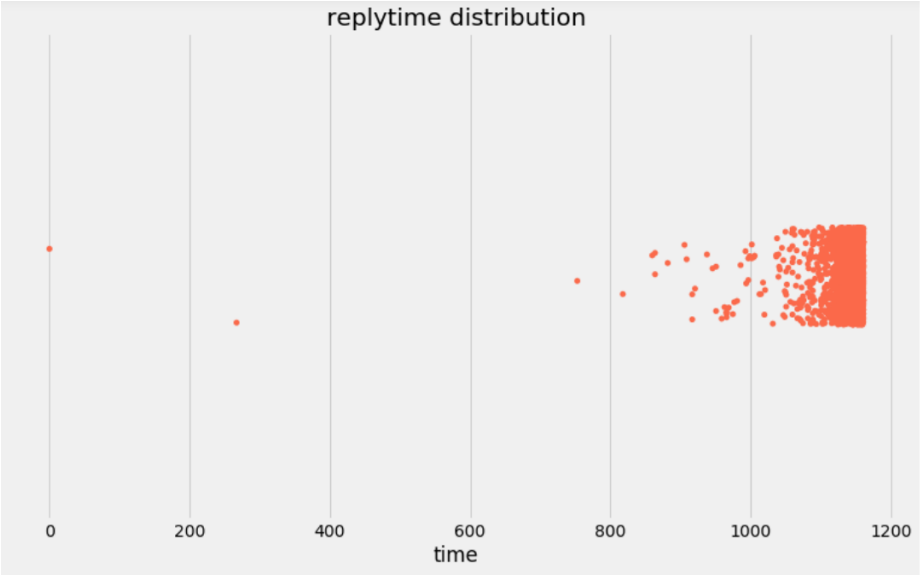

图 21 时间差分布图

从图中可以看出，时间差分布相当集中，这可能是造成时间差指标权重小的原因。

5.5 TOPSIS 评价模型建立

1. 构建初始评价指标矩阵

在对 2816 条回复留言的 6 个指标进行统计分析后，本文给出答复意见质量指数的评价矩阵 Z：

$$\mathrm{Z}=\begin{pmatrix} z_{11} & \cdots & z_{16} \\ \vdots & \ddots & \vdots \\ z_{2816,1} & \cdots & z_{2816,6} \end{pmatrix}$$

由于上述矩阵数据已经经过标准化，因此不再需要处理。

2. 确定最优方案和最劣方案

最优方案Z^+由Z中每列元素的最大值组成：

$$Z^+=(\max\{z_{11},z_{21},\dots,z_{2816.1}\},\max\{z_{12},z_{22},\dots,z_{2816.2}\},\dots,\max\{z_{16},z_{26},\dots,z_{2816.6}\})$$
$$=(Z_1^+,Z_2^+,\dots,Z_6^+)$$

最劣方案Z^-由Z中每列元素的最小值组成：

$$Z^-=(\min\{z_{11},z_{21},\dots,z_{2816.1}\},\min\{z_{12},z_{22},\dots,z_{2816.2}\},\dots,\min\{z_{16},z_{26},\dots,z_{2816.6}\})$$
$$=(Z_1^-,Z_2^-,\dots,Z_6^-)$$

3. 计算各评价对象与最优方案和最劣方案的接近程度

$$D_i^+=\sqrt{\sum_{j=1}^{6}w_j(Z_j^+-z_{ij})^2}$$

$$D_i^+=\sqrt{\sum_{j=1}^{6}w_j(Z_j^--z_{ij})^2}$$

其中w_j为第 j 个指标的权重即重要程度。

4. 计算每条留言回复与最优方案的接近程度C_i

$$C_i=\frac{D_i^-}{D_i^++D_i^-}$$

5.5.1 结果展示与分析

本文通过 TOPSIS 法，得到留言答复质量评价的部分结果（全部结果在附件 11），部分结果如图 22 所示：

留言编号	综合得分指数	排序
19133	0.623155204	1
88069	0.580588278	2
4331	0.467923503	3
119363	0.411281762	4
9744	0.404105123	5
99213	0.360279783	6
108139	0.345907968	7
132104	0.345374617	8
96762	0.340805589	9
88359	0.340757413	10

图 22 部分留言答复质量评价结果图

在本问题中，我们希望不仅可以算出每个留言质量分值，还可以得到具体分类结果。本文将留言质量分为优秀、中等、较差，确定阈值进行划分。但阈值的确定往往十分困难，主观判断给定阈值往往划分不准确，因此，寻找一种方法确定阈值是十分重要的。

在本文中，小组采用 K-means 聚类将留言质量分值聚 3 类，通过聚类分析结果确定阈值。聚类分析是一种无监督学习，可以根据数据集中样本的相似性分出不同的簇，是一种较为客观的方法。

聚类分析结果如图 23：

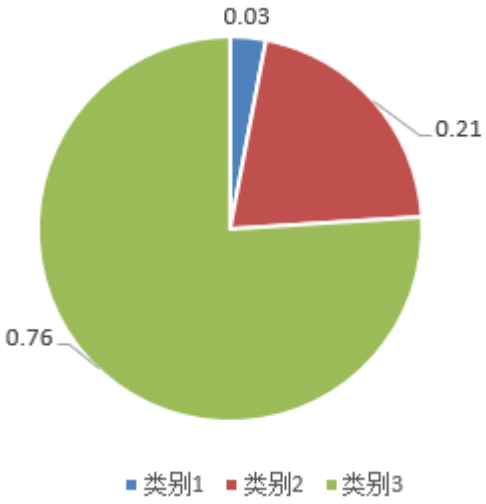

图 23 类别个案个数分布图

表 16 最终聚类中心表

类别	1	2	3
综合得分	0.264	0.119	0.365

由图 23 可知，3%的留言数据类别 1，21%的留言数据类别 2，76%的留言数据类别 3，且根据表 16 最终聚类中心表，类别 1、2、3 的留言质量依次递减。因此，我们认为排序后占前 3%的留言回复为“优秀”，有 85 条，3%~24%的留言回复为“中等”，有 590 条，24%以后的留言回复为“较差”，有 2141 条。

根据比例找到阈值分别为 0.198、0.0793。因此，当C_i>0.198 时，该留言质量为“优秀”、0.0793<C_i<0.198 时，留言质量为“中等”、C_i<0.0793 时留言质量为“较差”。

第6章 模型优缺点

6.1 模型优点

1. 对于任务一和任务二中预处理，为提升后续模型性能均编程构造新词库，减少特征，提高模型准确度。
2. 任务一中不断比较不同模型分类效果，更改 SVM 的核函数、SVM 的集成算法，为得到更优模型。
3. 任务二中文本相似度采用 JS 距离而余弦距离，有助于提高聚类效果。
4. 任务三中采用 K-means 确定阈值，而非主观随意确定，更具可靠性。

6.2 模型缺点

1. 集成学习以及交叉验证时，需要多次训练模型，使得整体运行时间较长。
2. 个别超参数的确定没有固定方法。

参考文献

[1]姜鹤. SVM 文本分类中基于法向量的特征选择算法研究[D]. 上海交通大学.

[2]李芸初. 基于支持向量机的文本分类[J]. 中国新技术新产品, 2019, 383(01):28-29.

[3]皮丽琴. 基于 AdaBoost-GASVM 算法和 LDA 主题模型的短文本分类研究[D].

[4]王振振, 何明, 杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学, 2013, 040(012):229-232.

[5]高扬. 基于 LDA 主题模型的 TFIDF 算法改进及应用[D]. 2015.

[6]张培伟. 基于改进 Single-Pass 算法的热点话题发现系统的设计与实现[D]. 2015.

[7]李廷辰, 杨艳. 基于分词聚类技术的微博热点问题挖掘[J]. 教学与科技, 2013(1):8-13.

附录

本文附件名如下，具体见附件：

1. 附件 1：任务一附件预处理后数据
2. 附件 2：任务一新词库
3. 附件 3：去停用词库
4. 附件 4：SVM-bagging 集成算法模型结果
5. 附件 5：任务二附件预处理后数据
6. 附件 6：任务二新词库
7. 附件 7：LDA 主题模型结果
8. 附件 8：任务问题表
9. 附件 9：热点问题留言明细表
10. 附件 10：任务三指标数据
11. 附件 11：TOPSIS 计算分值
12. 附件 12：逻辑用语词汇
13. 附件 13：礼貌用语词汇
14. 附件 14：singlepass 聚类结果