基于文本挖掘的智慧政务大数据分析

摘要

本文旨在利用文本挖掘和自然语言处理等方法，建立有效的数学模型来进行留言分类、热点挖掘以及答复意见评价，帮助提升政府的管理水平和施政效率。

针对问题一，首先，采用 TF-IDF，BOW 和 Word2vec 进行文本特征提取，并分别选择支持向量机分类器 (SVM)、逻辑回归分类器 (LR)、朴素贝叶斯分类器 (NB)、k-最近邻算法分类器 (KNN) 构建 9 种不同的留言分类模型。使用查准率 P、查全率 R 和 $F1$ 为模型分类性能的评价指标，比较不同分类模型在不同目标文本下分类性能的好坏。实验表明以留言详情为目标文本的 TF-IDF+ 支持向量机 SVM 分类器分类性能最好，宏 P 值为 0.9133，宏 R 值为 0.8979，宏 $F1$ 值为 0.9048。

针对问题二，首先，将留言文本按照时间升序排列，按产生时间划入不同的单位窗格（每两个月为 1 单位）。引入词的相对频度和词的增长速度两个指标，构造一个复合的权值 W 进行主题词检测。最后，使用增量聚类算法进行主题聚类，并统计各聚类主题对应的留言条数，点赞数和反对数，利用热度指数计算公式，得出排名前五的热点问题和相应热点问题的留言。

针对问题三，首先从湖南省统一互动平台网站爬取 39000 余条留言数据，从中提取了约 4000 条带有满意度打分的留言信息，作为训练模型的数据集。然后，选取了 9 个非典型文本特征，利用 χ^2 检验计算这些特征在训练集中对答复质量评价的影响权重。将爬取的数据分成 80% 的训练集和 20% 的测试集，根据权值选取特征作为模型的输入，利用随机森林进行模型的分类训练，得到的模型对于识别差评论有着较高的准确率，使用该分类模型得到附件四留言质量答复的评价结果。

关键字： 分类模型　主题词　增量聚类　χ^2 检验　特征提取

Abstract

This article aims to use text mining and natural language processing methods to establish an effective mathematical model for message classification, hot spot mining, and evaluation of reply opinions, helping to improve the government's management level and governance efficiency.

For problem one, first, we use TF-IDF, BOW(bag of words) and Word2vec for text feature extraction, and select support vector machine classifier, logistic regression classifier, naive Bayes classifier, k-nearest neighbor algorithm classifier construction 9 A different message classification model. We use the precision rate P, the recall rate R and F1 as the evaluation indexes of the model classification performance, and compare the classification performance of different classification models under different target texts. Experiments show that the TF-IDF + linear SVC model with message details as the target text has the best classification performance, with a macro P value of 0.9133, a macro R value of 0.8979, and a macro F1 value of 0.9048.

In response to question two, first of all, we will arrange the text of the message in ascending order of time, and divide it into different unit panes (1 unit every two months) according to the generation time. We introduce two indicators, the relative frequency of words and the growth rate of words, to construct a composite weight W for subject word detection. Finally, we use the incremental clustering algorithm to cluster the topics, and count the number of message posts, likes and oppositions corresponding to each clustering topic, and use the heat index calculation formula to get the top five hot issues and corresponding hot spots Problem message.

In response to question three, first of all, we crawled more than 39,000 message data from the unified interactive platform website of Hunan Province, and extracted about 4,000 message messages with satisfaction scores from it as a training model data set. Secondly, we selected 9 atypical text features, and used chi-square test to calculate the weights of these features in the training set on the evaluation of response quality. Divide the crawled data into 80% training set and 20% test set, select features as input to the model according to the weights, and use random forest for classification training of the model. The resulting model has a high accuracy rate for identifying poor reviews , We use the classification model to obtain the evaluation results of the quality response of the message in Annex IV.

Keywords:Classification Model;Subject Words;Incremental Clustering;Chi-square Test;Feature Extraction

目录

一、 问题重述

1.1 问题背景

近年来，随着微信、微博、市长信箱和阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依据人工进行留言划分和热点整理的相关部门工作带来极大挑战。同时，随着大数据、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。

根据题目附件中来自互联网公开的群众问政留言记录，以及相关部门对群众留言答复意见的文本数据，利用自然语言处理和文本挖掘的方法，建立了有效的分类和评价模型，能够更好地帮助政府提升管理水平和施政效率。

1.2 问题提出

1.2.1 群众留言分类

根据附件 1 和附件 2 给出的数据，建立关于留言内容的一级分类标签模型，使用分类评价指标 F-Score 对分类的结果进行评价。

1.2.2 热点问题挖掘

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，并给出热度评价结果。按表 1 的格式给出排名前五的热点问题，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”。

1.2.3 答复意见的评价

针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

二、 问题分析

2.1 问题一的分析

首先，对附件 2 中的中文文本数据进行了预处理，提取附件中的留言详情，留言主题和一级标签，对文本数据进行分词、去停用词、关键词提取等处理；之后，分别使用 TF-IDF(Term Frequency-Inverse Document Frequency)、BOW(Bag of words) 和 Word2vec 这三种方法进行文本特征提取，并分别选择支持向量机分类器 (SVM)、逻辑回归分类器 (Logistic Regressive)、朴素贝叶斯分类器 (Native Bayes)、k-最近邻分类器 (KNN) 构建了 9 种不同的中文文本分类模型；最后，使用查准率 P、查全率 R 和 F1 为模型分类性能的评价指标，分别选取留言主题和留言详情为目标文本，比较这 9 种文本分类模型在不同目标文本下分类性能的好坏。

2.2 问题二的分析

首先，需要对附件 3 中的数据进行预处理，对留言详情列数据进行文本清洗、去停用词、分词、词性标注以及词性过滤，词性过滤提取了能够更好地反映热点事件的动词和名词，并将留言文本按照时间升序排列，按产生时间分割成不同的单位窗格（每两个月为一个时间窗格）；之后，引入词的相对频度和词的增长速度两个定义，构造了一个复合的权值对主题词进行打分；最后，在基于增量聚类算法，对检测得到的主题词进行主题聚类，并统计各聚类主题对应的留言条数，点赞数和反对数，引入合理的热度指数计算，得出排名前五的热点问题和相应热点问题的留言。

2.3 问题三的分析

首先，使用 Python 爬虫从湖南省统一互动平台网站爬取 39000 多条的留言信息，舍弃无反馈的留言信息，得到约 4000 条带有满意度打分的留言信息，将答复满意度分为两个等级（好，差），作为训练模型的数据集。其次，选取 9 个非典型文本特征，利用卡方检验计算这些特征在训练集中对答复质量评价的重要性权值，并进行权值排序，分析不同特征对答复质量的影响。然后，将爬取的数据集分成 80% 的训练集和 20% 的测试集，将提取的三个特征作为模型的输入，利用随机森林分类算法进行答复质量评价模型分类训练。最后，再对附件四的文本数据进行同样预处理，将处理后的数据输入答复质量评价模型，得到附件四留言答复的评价结果。

三、 符号说明

符号	意义
TF	词频
IDF	词语的逆文档频率指数
$TF-IDF$	文本中特征词的特征得分
P	查准率
R	查全率
TP	预测为正例且预测正确的文本数
FP	将真实为反例的文本预测为反例的文本数
FN	将真实为正例的文本预测为反例的文本数
$F-score$	某一类文本的分类评价指标
$F1$	所有文本类别的 $F-score$ 值的平均数
n	文本集中类别总数
G_{ij}	某个文本时间窗格 j 中词 i 的词频增长速度
F_{ij}	词 i 在当前窗格 j 中出现的频率
F_{iu}	词 i 在之前某个窗格 u 的频率。
K	在当前时间窗格的之前的所有窗格总数
RF_{ij}	表示词的相对频度
$max(F_j)$	某个时间窗格 j 中出现的频率最高的词的词
W_{ij}	某个时间窗格 j 中词 i 的重要程度
a	对主题词进行选择的权重
$P(e\|h)$	词语 e 和 h 的条件概率
$F(e,h)$	词 e 和词 h 同时出现的留言文本数
F_h	词 h 出现留言文本数
H_i	热度指数

符号	意义
m_i	问题 i 的留言总数
o_i	问题 i 所有留言点赞数之和
n_i	问题 i 所有留言的反对数之和
b	留言总数指标的权重
f	点赞总数与反对总数之和这一指标的权重
w	文本中某个词
C	文本中某个主题词的集合
C_i	簇 C 中的某个词 i

本表未涉及的符号在文中有具体说明。

四、 数据预处理

4.1 问题一的数据预处理

对文本数据进行预处理是进行自然语言处理任务最基本的一步，在进行文本分类之前，对文本数据进行合适的处理可以提高分类模型的性能。对题目所提供的附件 2 中的文本数据进行如下处理：

1. 文本清洗，去除掉文本的无意义、多余的部分。在留言数据中，会有一些与文本分类无关的字符，如：html 字符，乱码数据等，它们在一定程度上会影响文本的有用的特征的提取。

2. 分类标签数值化。将数据中的一级分类标签转化为数值，便于后续实验中使用机器学习的方式进行模型训练。

3. 去除中文停用词。文本中一般会出现一些没有实际意义的连接词，如：“了”、“啊”、“的”等等，去除掉这些停用词，可以一定程度降低特征空间的维度，提高分类的效果。

4. 分词处理。由于一大段文本无法作为特征项进行文本分类，所以需要对文本进行切分。在中文文本中可以选择字、词、词组作为特征项，但词更有优势 [1]，所以采用 Python 自然语言处理工具 jieba 来对文本进行词的切分，如一个待分词的句子：“这几天都在学自然语言处理。”，进行中文分词后为：“这/几天/都/在/学/自然语言/处理。”

以上为文本数据预处理的步骤，流程图如图 1 所示：

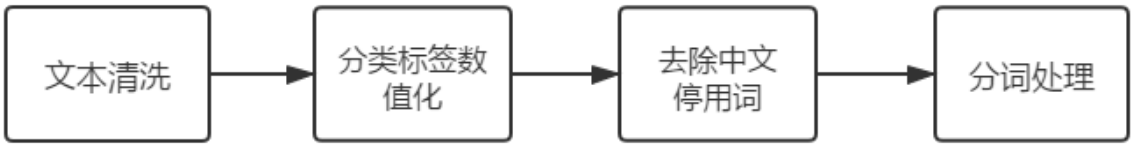

图 **1**　问题 **1** 文本数据预处理流程

4.2 问题二的数据预处理

对附件三的文本数据进行预处理，经过文本清洗，去除中文停用词，分词处理之后还需要进行词性标注处理，这样处理之后可以进行词性过滤，只取名词和动词，名词和动词能够更好的描述热点事件的特征，这也是文本维数消减的一种形式，可以更好地挖掘得到热点事件。

本题的数据预处理流程如图 2 所示：

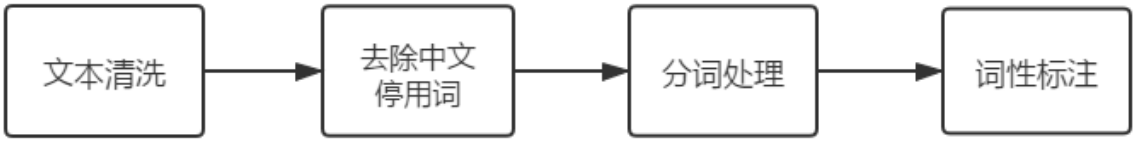

图 **2**　问题 **2** 文本数据预处理流程

4.3 问题三的数据预处理

对附件四的文本数据进行文本清洗，去除中文停用词和分词，利用 python 工具对附件四的留言长度、答复分词后的词汇数、答复分词后词汇数（去停用词）、留言问题与答复意见的相似度等 9 个非典型文本特征进行计算，并对数据进行统计。

其中，留言问题与答复意见的相似度定义为余弦相似度，余弦相似度算法是通过测量两个词向量的夹角的余弦值来衡量词向量之间的相似性，当夹角 θ 为 0 时，余弦值为 1，相似度最高，而其他任意角度的余弦值都不大于 1，余弦相似度的计算公式如下 [15]：

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

算法步骤为:

1. 采用 TF-IDF 特征提取算法分别对留言详情和其对应的答复意见进行特征提取分别得到两个文本中的关键词。

2. 将各自文本中的关键词合成一个并集作为一个词集。

3. 分别统计留言详情和答复意见文本中出现词集中每个词的词频，只保留词频 Top5 的词，生成词频向量。

4. 利用余弦相似度计算公式计算两个词频向量的余弦值，余弦值越接近 1，代表该留言详情和对应的答复意见具有很高的相关性。

五、 问题一的求解

5.1 群众留言分类模型的建立

建立群众留言分类模型的大致流程如图 3 所示:

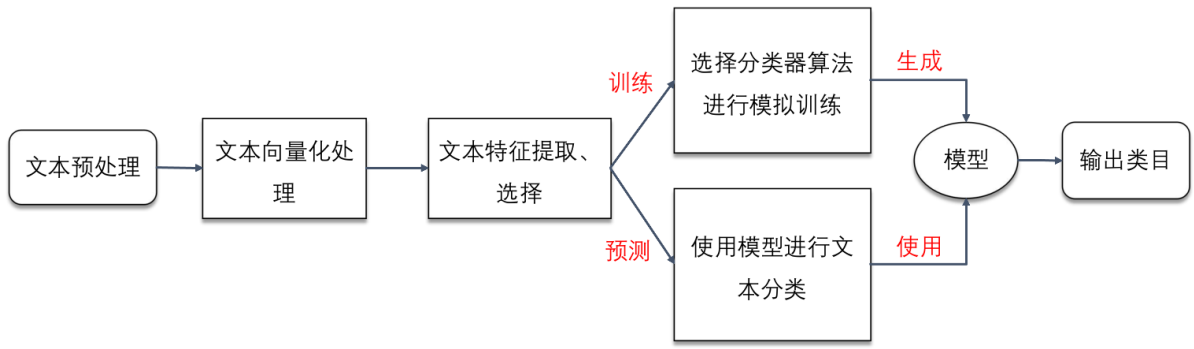

图 **3**　留言分类模型流程图

5.2 特征提取算法的选择

5.2.1 TF-IDF 算法（Term Frequency-Inverse Document Frequency）

TF-IDF 是一种用于信息检索与数据挖掘的常用加权技术，用于评估字词对于一个文本集或者一个语料库中的其中一份文本的重要程度。TF 是词频（Term Frequent）指的是某个给定词语的在该文本中出现的频率，再将其归一化，防止偏向长的文本。IDF 是逆文本频率指数（Inverse Document Frequency）可以理解为一个词语普遍重要性的度量 [2]。

算法的核心思想：如果特征 t 在文本集中出现的频率的低，但是在某一类的文本中出现的频率较高，即也能说明该特征 t 具有较好的类别区分能力，可以赋予较高的权重 [3]。

TF-IDF 的计算方式如下：

$$\text{本文本总词数}(TF) = \frac{\text{某个词在文本中出现的次数}}{\text{本文本总词数}} \tag{2}$$

$$\text{逆文档频率}(IDF) = lg\frac{\text{文本集中的文本总数}}{\text{包含该词的文本数}+1} \tag{3}$$

$$TF-IDF = \text{词频}(TF)\cdot\text{逆文档频率}(IDF) \tag{4}$$

5.2.2 词袋模型 (Bag of words)

词袋模型是一种基础的特征提取方式，基本原理就是将目标文本看作无数个词的集合，文本集中所有的词在目标文本中是否出现可以作为文本的特征，若出现则为 1，否则为 0。词袋模型对于一个目标文本可以忽略文本中词的次序和句子的语法，文本中任意一个词出现在文本中的任意一个位置，都不会受到其他因素的干扰。利用词袋模型可以实现从文本中提取特征项形成特征项矩阵被用于文本分类 [2]。

5.2.3 Word2vec 特征提取算法

为了将文本集输入神经网络进行训练，需要先将文本数据化表示成计算机能够理解的形式，word2vec 通过大量的文本集中以无监督学习的方式去学习自然语义，就是通过嵌入一个空间使得语义相近的词在空间内距离近，利用向量空间中的距离来体现词与词之间相似性。主要方法就是直接使用一个稠密的实数向量来表示一个词, 向量一般为 50 维或 100 维, 每个维度代表了该词的某个方面的特征, 有实际意义的语义特征, 其值都是 0 至 1 之间的一个数值，再取文本中所有词的向量的平均值 [4]。

5.3 分类器的选择

5.3.1 线性支持向量机 (Linear SVM)

SVM 分类器是线性分类器中的一种，可以将线性不可分的样本通过核函数映射到高维空间，并最小化损失函数，找到最优的分割平面，完成对样本的分类。后续实验使用 scikit-learn 算法实现 SVM 分类器功能，使用高斯核作为核函数，hingeloss 作为损失函数。

SVM 算法最初是为二值分类问题设计的，当处理多类问题时，就需要构造合适的多类分类器，本文通过间接法中的一对多法（one-versus-rest, 简称 OVR SVMs）构造 SVM 多类分类器。算法的核心为：训练时将某个类别的样本归为一类，其他剩余的样本归为一类，这样附件 2 中 7 个类别的样本就构造出 7 个 SVM。分类时，将未知留言样本分别放入这 7 个 SVM 进行测试，每个测试都有一个分类函数值结果，将未知样本分类到具有最大分类函数值的那类 [5]。

5.3.2 逻辑回归 (Logistics Regression)

逻辑回归主要用来解决二分类问题, 其算法步骤为：

1. 确定一个预测函数，即预测出一个值来判断归属哪一类，可定义预测值大于某个阈值判断为一类，反之为另一类；

2. 为了计算参数，需要定义一个损失函数，损失函数用来衡量真实值和预测值之间的差异，这个差值越小说明预测效果越好

3. 用梯度下降法计算参数，使损失函数不断减小，得出参数后，带入预测函数就可以来进行预测了。

为解决多分类问题，需要在逻辑回归的基础上采用 OVR(One Vs Rest) 算法。附件 2 有 7 类留言文本数据，首先依次将每一类和剩余的六个类比较作为二分类问题，7 个类别进行 7 次分类，得到 7 个二分类模型；其次，当给定一个新的待分类留言文本时，求出每个二分类对应的概率，概率最高的一类作为新文本的预测结果 [6]。

5.3.3 朴素贝叶斯 (Native Bayes)

朴素贝叶斯（Naive Bayes）算法是基于贝叶斯定理与特征条件独立假设的分类方法，它的基本思想是：首先计算出各个类别的先验概率，再利用贝叶斯定理计算出各特征项属于某个类别的后验概率，通过选出具有最大后验概率 MAP（MaximunA Posteriori) 估计值的类别即为最终的类别。

首先根据附件 2 的 7 个一级标签，设为类别集合 $C = C1, C2, C3, \cdots, C7$, 并计算每个类别的先验概率 P；其次，将一个待分类的留言文本表示为 $D = t1, t2, \cdots, t_m$,t 为该文本的特征项；最后计算待分类的留言文本属于哪一类的概率，即后验概率 $P(C1|D)$, $P(C2|D), \cdots, P(C7|D)$, 将待分类的留言文本划分为后验概率最大的类别中 [3]。

5.3.4 K-最近邻算法 (k-NearestNeighbor)

K-最近邻算法算法的分类方式是通过查询类似文档的分类情况, 来判断新文档与已知文档是否属于同一类别. 该算法的基本思想是: 给定一个新文本, 由算法搜索模式空间即训练文本集, 找出与新文本距离最近 (最相似) 的 K 篇文本, 最后根据这 K 篇文本所属的类别判别新文本所属的类别。

首先，根据留言文本的特征项集合描述训练的留言文本向量；其次新的留言文本到达后，将新的留言文本表示为文本向量形式，并且在训练集中选出与新文本最相似的 K 个文本, 这里 K 值先选取一个初值, 然后根据测试结果进行调整；最后根据与新文本最近的 K 个邻居的类属关系, 计算新文本属于每类的权重 [7]。

5.4 分类器的性能评价指标

评价分类模型性能的通用指标为查准率（Pression）, 查全率（Recall）及 F1 值（F-Score）。在本题采用这三个评价指标对分类模型性能进行评价。

查准率 P 是指分类器预测为正例且预测正确的文本数与所有预测为正例的文本数之比，计算公式为 [8]：

$$P = \frac{TP}{TP + FP} \tag{5}$$

其中，TP 是指预测为正例且预测正确的文本数，FP 是指将真实为反例的文本预测为正例的文本数。

查全率 R 是指分类器预测为正例且预测正确的文本数与所有真实为正例的文本数之比，计算公式为 [8]：

$$R = \frac{TP}{TP + FN} \tag{6}$$

其中，FN 是指真实为正例预测为反例的文本数。

F-score 值是一种综合考虑了上述两种评价指标的混合评价指标，F-score 值越高，分类性能越好。取所有文本类别的 F-Score 值的平均值 (即宏平均值 F1) 也作为分类方法的评价指标，计算公式为：

$$F_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{2P_i R_i}{P_i + R_i} \tag{7}$$

其中，n 为总的类别数，Pi 为第 i 类的查准率，Ri 为第 i 类的查全率。

5.5 分类性能对比

取数据中的留言主题、留言详情和一级标签三个指标，在经过数据预处理之后，将每个类别的数据都分为训练集（80%）和测试集（20%）。将 TF-IDF、词袋模型（BOW）和 Word2vec 这三种特征提取算法，分别与 SVM 分类器、NB 分类器、KNN 分类器、LR 分类器进行搭配构建 9 种不同的留言文本分类模型，并且通过查准率、查全率和 F1 值比较不同模型的文本分类性能的优劣，以及对比留言主题和留言详情哪个指标作为目标文本的分类效果更好，来得出最优的留言文本分类模型。

1. 比较不同目标文本的分类效果

当目标文本为留言主题时，各模型分类效果如下图 4 所示，分类模型的平均 F1 值约为 0.7857，平均查准率 P 为 0.7765，平均查全率 R 为 0.7108，分类效果最好的模型 F1 值达到 0.8429；

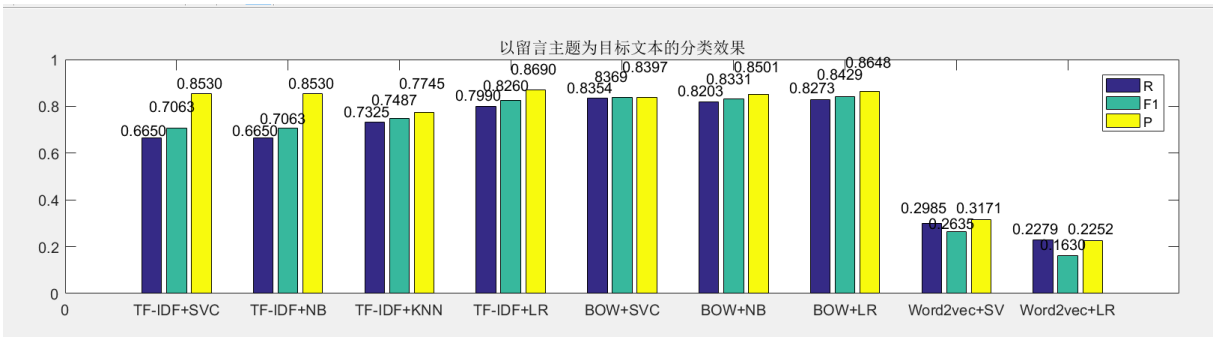

图 **4**　以留言主题为目标文本的 **9** 种模型分类效果

当目标文本为留言详情时，各模型分类效果如下图 5 所示，分类模型的平均 F1 值约为 0.8272，平均查准率 P 为 0.8404，平均查全率 R 为 0.7944，最高的 F1 值达到 0.9048。

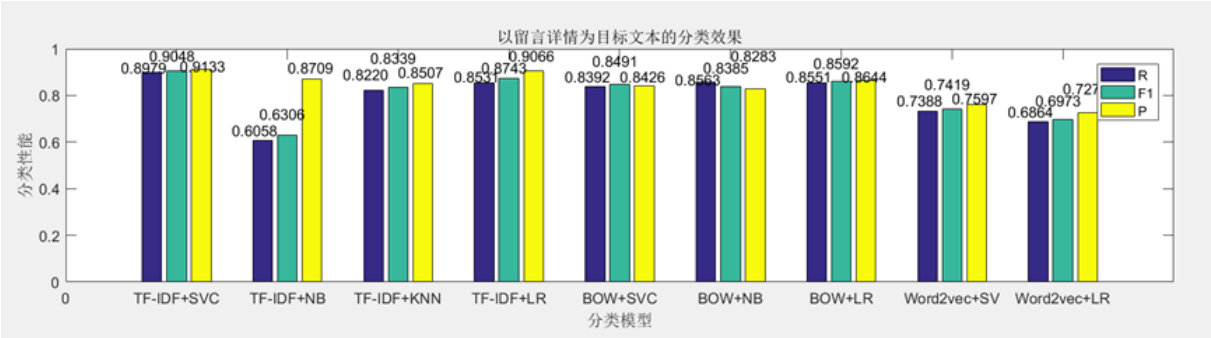

图 **5**　以留言详情为目标文本的 **9** 种模型分类效果

从实验分析结果可得，以留言详情为目标文本的分类模型总体分类效果更好，因此选择留言详情为目标文本。

2. 比较以留言详情为目标文本，比较 9 种分类模型的分类性能

本题的数据集的留言文本总数为 9210 篇，共分为 7 个类，分别为城乡建设、环境保护、交通运输、教育文本、劳动和社会保障、商贸旅游、卫生计生。各类别分布如图 6 所示。

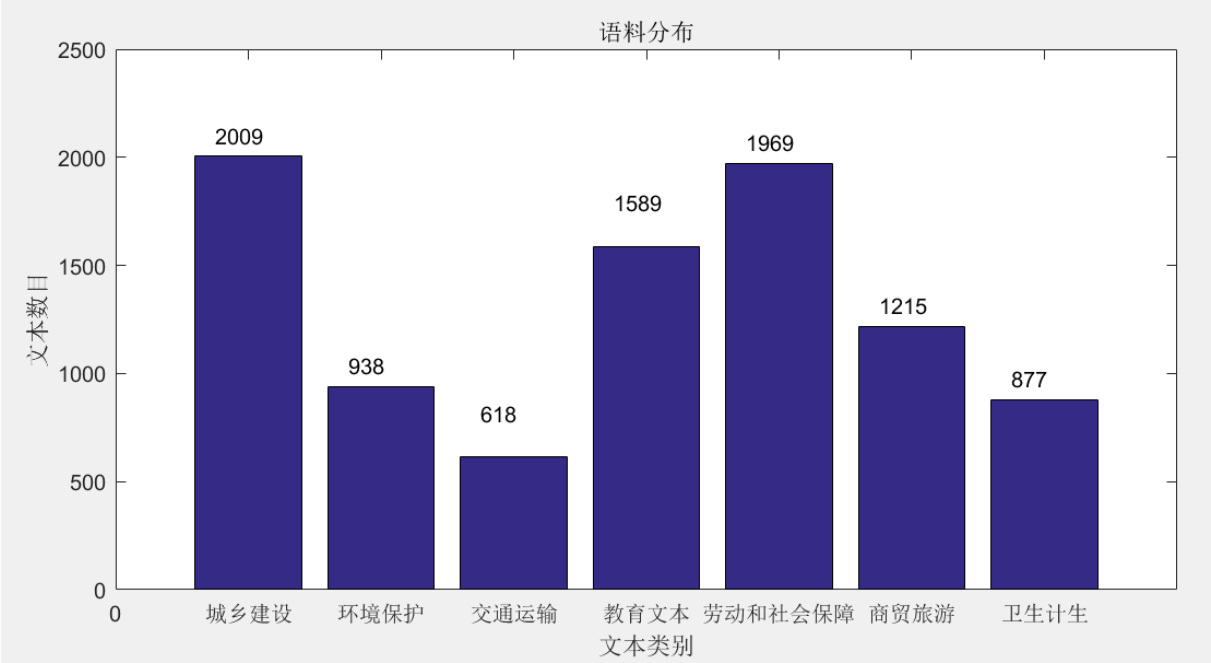

图 **6**　本题数据集中的不同文本类别的文本数目

从预处理好的留言文本中提取 80% 作为训练文本集,20% 作为测试文本集, 通过查准率、查全率、F1 值直观地比较九种分类模型的分类效果。如下图 7 所示，结果显示 TF-IDF+SVM 分类器模型分类效果要优于其他模型。

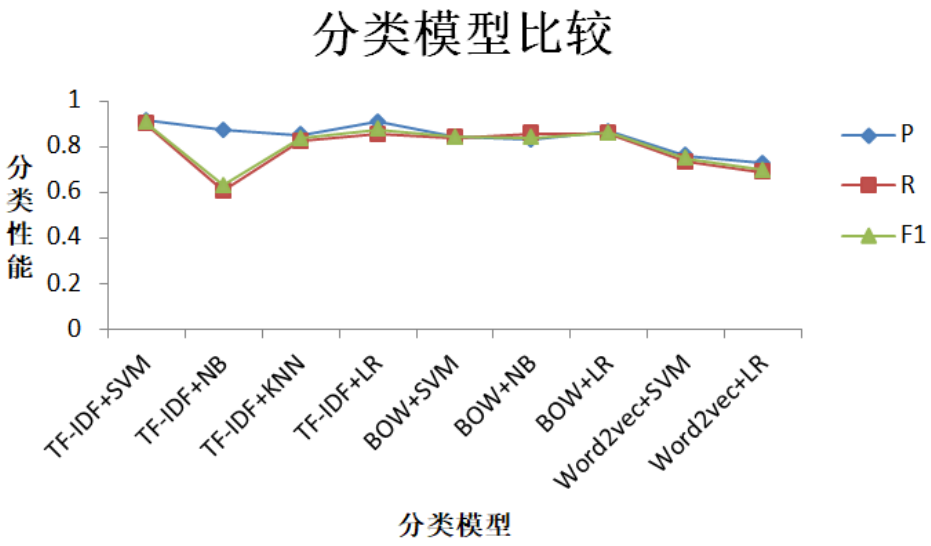

图 **7**　以留言详情为目标文本 **9** 种分类模型的分类效果

因此，选择 TF-IDF+SVM 分类器模型进行留言文本分类。实验结果如表 1（精确到 0.0001）：

表 **1**　**TF-IDF+SVM** 分类模型针对不同文本类别的分类效果

	$F-score$	P	R	$support$
城乡建设	0.8828	0.8601	0.9067	407
环境保护	0.9235	0.9141	0.9330	194
交通运输	0.8632	0.9100	0.8211	123
教育文体	0.9390	0.9404	0.9375	320
劳动和社会保障	0.9309	0.9150	0.9472	398
商贸旅游	0.8870	0.9315	0.8465	241
卫生计生	0.9073	0.9221	0.8932	159
宏 F1 平均值	0.9048	0.9133	0.8979	1842

使用 TF-IDF+SVM 分类器模型的分类效果见表 1，查准率为 0.9133，查全率为 0.8979，F1 值为 0.9048，分类效果比较好，其混淆矩阵如图 8 所示。

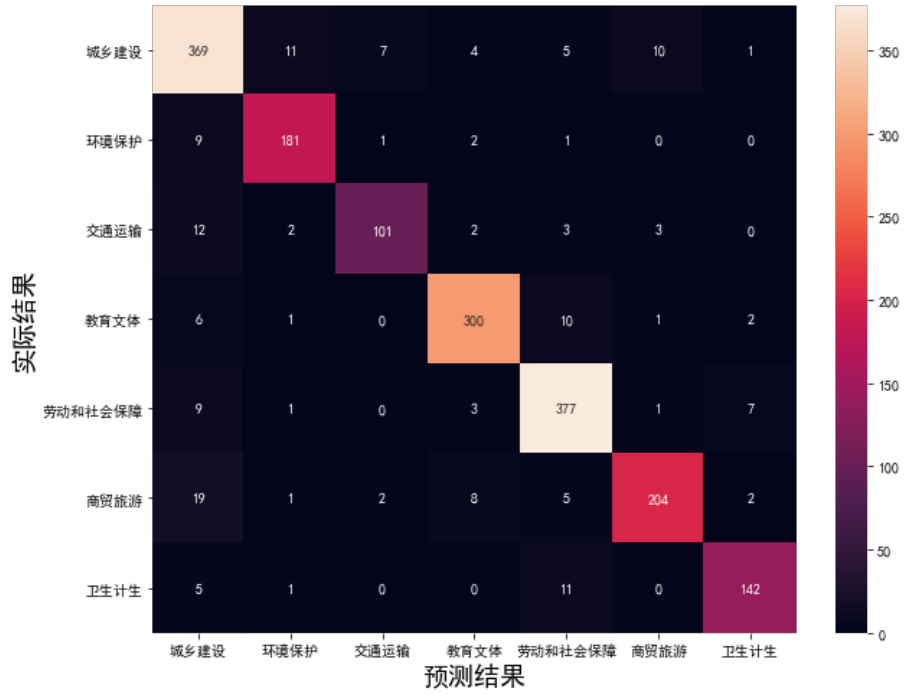

图 **8**　**TF-IDF+SVM** 分类模型的混淆矩阵

5.6 小结

我们首先介绍了数据集的选择以及分词处理，接着实验对比验证文本分类中经典的三种特征选择算法 TF-IDF, 词袋模型 (BOW) 和 Word2vec, 分类模型选择 KNN、线性 SVM、NB 和逻辑回归。通过实验对比，以留言详情为目标文本的 TF-IDF+SVM 分类器模型分类效果最好，查准率为 0.9133，查全率为 0.8979，F1 值为 0.9048。因此，选用 TF-IDF+SVM 分类器模型对留言文本进行分类，在此题背景下能够达到比较好的分类效果。

六、 问题二的求解

建立热点问题挖掘模型的大致流程如图 9 所示：

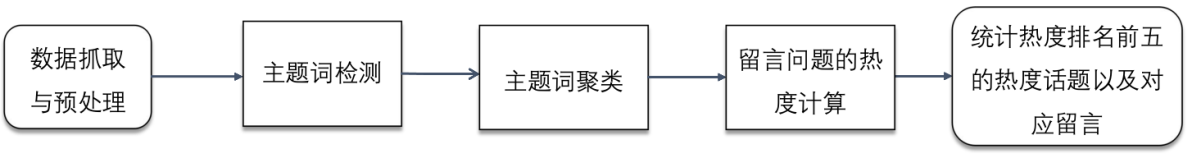

图 **9**　热点问题挖掘流程图

6.1 主题词检测

热点问题意味着人们关注的事件和活动，这种关注可能是长时间的，持续的；也可能是短期的、突发的；所以又可以划分为长期活跃的热点问题和突发的热点问题。由于这两种热点问题的形式不一样，其对应的词汇特征也有所不同，所以在提取热点主题词时要对两种情况综合考虑，所以传统的特征提取算法不太适用，可以通过计算词频的增长速度和词的相对频度两个指标来对热点主题词进行提取 [9]。

由于热点问题具有很强的时序性，所以对热点主题词的提取有别于静态的特征提取，传统的 TF-IDF 并不适用。为了能更好提取热点主题词，引入增长系数 G_{ij} 来表示在某个 j 窗格词 i 的词频增长速度，定义为当前窗格中该词的频率除以之前 K 个窗格中的频率平均值，计算公式为 [10]：

$$G_{ij} = \frac{F_{ij}}{F_i} = \frac{F_{ij} \cdot K}{\sum_u^k F_{iu}} \tag{8}$$

其中，G_{ij} 为某个时间窗格 j 中的某个词 i 的词频增长速度，F_{ij} 表示词 i 在当前窗格 j 中出现的频率，K 为在当前时间窗格的之前的所有窗格总数，F_{iu} 为词 i 在之前某个窗格 u 的频率。

因为热点问题不仅仅与词的增长速度还和词的相对频度有很大关系，为了了解词 i 在某个时间窗格 j 的重要程度，所以对词的相对频度 RF_{ij} 进行具体的计算，计算公式为 [9]:

$$RF_{ij} = \frac{log(F_{ij})}{log(F_{max})} \tag{9}$$

其中，RF_{ij} 表示词的相对频度，$max(F_j)$ 为某个时间窗格 j 中出现的频率最高的词的词频。

由于热点问题的主题词与词的增长速度和词的相对频度都具有相关性，所以衡量和确定热点问题需要结合这两个指标，可以构造参数 a 来调节两个指标对主题词的选择，计算公式 [9]：

$$W_{ij} = a \cdot log(G_{ij}) + (1-a) \cdot log(RF_{ij}) \tag{10}$$

其中，W_{ij} 定义为某个时间窗格 j 中词 i 的重要程度。

6.2 主题词聚类

按照 W 的大小对主题词表进行降序排列，然后对排序后的词增量进行聚类。将带有权值的主题词列表作为增量聚类模型的输入，以簇列表作为模型的输出，增量聚类模型的算法步骤为 [10]：

1. 将第一个主题词作为初始簇；
2. 输入下一个主题词，计算它与每个已有簇的距离；
3. 当距离它最近的簇的距离大于阈值 D 时，将该词作为一个新簇，否则，把它放入距离最近的簇中；
4. 继续输入下一个主题词，重复步骤 2 到步骤 4，直到所有词都输入完毕。

为判断步骤 2 中的主题词是否属于某个簇，需要对词语的相似度进行了合理的定义。把两个词语 e 和 h 同时出现在一条留言文本里作为词相似的依据，引入两个词的条件概率 $P(e|h)$，作为两个词的相似度 [10]。

$$P(e \mid h) = \frac{F_{e,h}}{F_h} \tag{11}$$

其中，$F_{e,h}$ 代表词 e 和词 h 同时出现的留言文本数，F_h 代表词 h 出现留言文本数。

为了定义词与簇的相似度（簇为一个词的集合），词 w 到簇 C 的距离为条件概率最大值的倒数。因此，词 w 到簇 C 的距离定义为 [10]：

$$d_{w.c} = \begin{cases} \frac{1}{\max\{P(c_i|w)|c_i \in C\}}, (\max\{P(c_i|w)|c_i \in C\} > 0) \\ \infty, (\max\{P(c_i|w)|c_i \in C\} = 0) \end{cases} \tag{12}$$

当簇 C 存在 c_i, 且词 c_i 在含有词 W 的留言文本出现概率很高，那么词 w 距离簇 C 较近，要把词 w 归入 C，反之则不归入 C。

当完成主题词的增量聚类模型之后，会得到若干个含有一个或多个主题词的留言话题, 通过这些主题词的组合可以挖掘出热点话题。

6.3 留言问题的热度指数

每个热点问题的都具有相对应的热度，需要运用热度指数去评价人们对该热点问题关注度的高低，通过比较不同热点问题的热度指数来得到热度排名。热度指数的影响因子主要有：与该问题相关的留言总数、对该问题留言的反对数、对该问题留言的点赞数等等。可以知道留言数越多，表示人们对该问题的关注度越高，则热度指数也就越高，所以确定留言数可以作为衡量问题热度指数的一个指标。而且人们对具体问题的留言的反对数和点赞数也可以很好地反映该问题的热度，点赞数和反对数越高，表示人们对这个问题的参与度和关注度也就越高，所以考虑将该问题人们的反对数与点赞数之和作为衡量问题热度指数的另一个指标，参考文献 [11]，对其热度计算公式进行适当的改变且综合两个指标的影响和本题数据，得到热度指数的计算公式如下：

$$H_i = bln(m_i) + fln(o_i + n_i) \tag{13}$$

其中，H_i 为热度指数，m_i 为问题 i 的留言总数，o_i 为问题 i 所有留言点赞数的和，n_i 为问题 i 所有留言反对数的和，b 为留言总数指标的权重，f 为点赞总数与反对总数之和这一指标的权重。

对本题数据分析，经过实验，当 b 为 0.7，f 为 0.3 时，能够较好的体现热点问题的热度。

6.4 实验过程与结果

6.4.1 数据预处理

实验使用的数据集是题目所提供的附件 3 中的留言文本数据，时间跨度为 2017 年 6 月 8 日到 2020 年 1 月 26 日，共有 4326 条留言数据。数据包括留言编号，留言用户，留言主题，留言详情，留言时间，点赞数和反对数。

经过文本清洗和去除停用词之后，利用中分分词工具 jieba 对文本数据中的留言详情进行分词，并做词性标注，只保留对主题表达和辨识作用最大的名词和动词，其他词性的词忽略，最后进行词频统计，将出现次数前 2000 的词汇用于主题词检测。

6.4.2 主题词检测

因为热点问题具有时序性，进行主题词提取时，所以先对数据的时间列进行升序处理，对时间窗格进行分割。本文实验将时间窗口设为两个月：将留言文本根据留言时间划分，每两个月的留言文本同处在一个时间窗格（其中，2019 年之前的数据非常少，仅有 3 条，划分在同一个时间窗口。2019 年之后的数据也较少，也划分为同一个时间窗格），总共将数据划分为 8 个时间窗格，对热点主题词进行提取。

公式10中包含有可供调节的参数，参数 a 的作用是调节两个系数对主题词提取结果的影响，不同的 a 取值对主题词识别的影响较大。以 2019 年 7 月 9 月这个时间窗格为例，附件表 4 中给出了 a 取不同的值所得的 Top20 热点主题词。

当 a 为 0 的时候，仅有词汇相对词频作用于主题词提取；当 a 为 1 的时候，仅有词汇增长速度作用于主题词提取，为了在相对词频和增长速度之间取得平衡，在后续试验中，a 参数值设为 0.5[10]。

统计每个时间窗格出现次数前 2000 的词语，利用主题词检测公式计算他们在这段时间段内的的 w_{ij} 值，然后按照阈值 T(T 取 5) 选出 $w_{ij} > 1$ 的主题词，即得到一个主题词表。这些被选出的词将被聚类产生各个留言话题。

6.4.3 主题聚类

应用增量聚类算法，对上述筛选出的留言主题词作主题聚类。主题聚类的部分结果见图 10。

消息时间	留言话题	对应留言热点事件
2019年7月-9月	市梅|嘉顺苑|迫于|裂缝|修复	A市梅溪湖嘉顺苑安置小区房屋质量差
2019年7月-9月	魅力|熏|烧烤店|晾晒|临街|时段|权力|死|烧烤	A5区劳动东路魅力之城小区油烟扰民
2019年7月-9月	滨河|定向|铁路职工|捆绑|景园|苑|认购|职工|资格|商品房|销售	投诉伊景园滨河苑项目违法捆绑车位销售
2019年7月-9月	混凝土|执法人员|局面|开建|宁静|表现|启用|小于	A市高新区公安局旁一工地深夜浇筑混凝土，噪音非常大
2019年7月-9月	望江|米处|纵观|发声|想象|专家|案例|杜绝|证实|隧道|施行|挖	反对加固A4区楚雅路西望江公寓2栋，请求拆迁！

图 **10**　主题聚类的结果片段

6.4.4 留言问题的热度计算

对主题聚类并筛选得到的人们参与度和关注度相对较高的热点问题，统计热点问题的留言数、点赞数和反对数，应用留言问题的热度指数计算公式 (13) 对热点问题进行热度指数计算，得出热度排名前五的热点问题如下图 11 所示。

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	4.1258	2019/1/11至2019/7/8	西地省	58车贷案件立案之后毫无进展
2	2	3.888	2019/7/7至2019/9/1	A市伊景园滨河苑	购房捆绑销售车位
3	3	3.421	2019/7/21至2019/9/10	A5区魅力之城小区	小区夜宵摊油烟扰民
4	4	3.407	2018/11/15至2019/12/2	A市	在该市申请人才补贴不通过
5	5	2.863	2019/1/8至2019/11/12	A市	地铁7号线的修建建议

图 **11**　热度排名前五的热点问题相对应的部分留言结果

与热度排名前五的热点问题相对应的部分留言结果如下图 12 所示：

问题ID	留言编号	留言用户	留言主题	留言详情	点赞数	反对数
1	220711	A00031682	请书记关注A市A4区58车贷案	万分，急切盼望有案情消息总是失望，四处诉求也无效。此种	821	0
1	217032	A00056543	严惩A市58车贷特大集资诈骗案保护	说大股东苏纳和小股东、苏纳弟弟苏吕是挂名；说担保公司注	790	0
1	194343	A000106161	承办A市58车贷案警官应跟进关注留	但是，A市A4区经侦并没有跟进市领导的留言，案件调查进	733	0
5	243808	A00053304	强烈建议将地铁7号线南延至A市生	乘坐五六公里公交车赶到“尚双塘”，再换乘地铁或到“A1区	31	0
1	268251	A000106090	西地省58车贷立案近半年毫无进展	是不抓捕控制平台高管和资产，这种情况，要求还受害人支	25	0
5	202847	A00018309	被地铁遗忘的A市经开区泉塘片区	街道，如今泉塘已逐渐成为星沙主城区的商业次中心。泉塘街	13	1
2	191001	A909171	A市伊景园滨河苑协商要求购房同时	这让我非常苦恼，作为一名退休职工，就攒了一些养老钱，	12	1
5	256590	A00091600	关于申请增加A7县泉塘地铁支线的	四路，在开元路和人民东路间设置地铁支线，在主干道盼盼路	11	1
2	200085	A000104234	A市市政建设开发有限公司对广铁职	么？还跟我们说定价12万一个的车位费是成本价，而周边小区	9	2
4	206983	A00049301	A市人才新政补贴最近两个月的怎么	突然之间就停了，大家都没收到。诉求：1，具体是什么原因	8	0
4	247736	A00085481	A市人才新政的补贴已经快两个月没	月没发了，之前都按时发的，为什么这次拖了这么久？是什	7	2
5	239005	A00057814	建议A市地铁7号线一期工程线网规	经开区和星沙飞速发展，建议地铁7号线一期就要考虑延长到	7	1
1	240554	A00029163	A市58车贷老板跑路美国，经侦拖延	不力，纵容嫌犯。这是涉嫌保护伞直观表现。那速夫妇潜逃	6	0
3	272122	A909113	A5区劳动东路魅力之城小区一楼的	分居民还是觉得要维护社会和谐稳定，合法维权。为此我们	6	0
4	225657	A00051791	关于A市人才购房补贴的疑问	的人员，政策鼓励其在A市安家发展，而对于先买房后获证的	6	2
3	236798	A00039089	A5区劳动东路魅力之城小区油烟扰	目前油烟机清洗也没有。每天油烟直排。熏死树木。对环境	4	0
1	254532	A000106062	A市58车贷恶性退出立案近半年没有	委会为其脱罪。于是，在58官网上挂出选举的公告和参选人员	3	0
1	226265	A000106448	恳请A市经侦公正办理58车贷案件，	官说，要相信经侦，但这样的经侦我们怎么相信呢？还有就	3	0
2	214975	A909182	关于房伊景园滨河苑销售若干问题	购买。据说车子是以价值12万的成本价卖给我们的，但是对于	3	0
3	195095	A00039089	魅力之城小区临街门面油烟直排扰	烟烧烤熏死人。一天24小时都是烟。请政府关闭处理这个地	3	0

图 **12**　热度排名前五的热点问题相对应的部分留言结果

6.5 小结

本问题我们提出了一种基于时序热点问题文本的检测方法，这一方法包括五个步骤：文本预处理、词性过滤、主题词检测、主题聚类、热度指数计算。在词性过滤中，只选取了名词和动词，大大减少了留言文本中的噪音数据；在主题词检测中，通过构造一个复合权值 w 对词汇进行评分，决定其是否为留言文本的主题词；最后再综合增量聚类算法和热度指数计算公式，得到热度排名前五的热点问题以及相应的留言信息。

七、 问题三的求解

7.1 爬取带标注的数据

针对本题，使用 Python 爬虫从湖南省统一互动平台爬取了 39000 多条留言及答复信息，舍弃掉其中无反馈的留言信息，共得到 4000 条有满意度打分留言数据，作为答复意见评价训练模型的数据集。通过对数据集分析，发现留言满意度为好的数据有 1184 条，差的数据有 2783 条，一般的仅有 415 条，占比不到 10%，因此将答复满意度分为两个等级（好，差）。数据占比分布如图 13 所示。

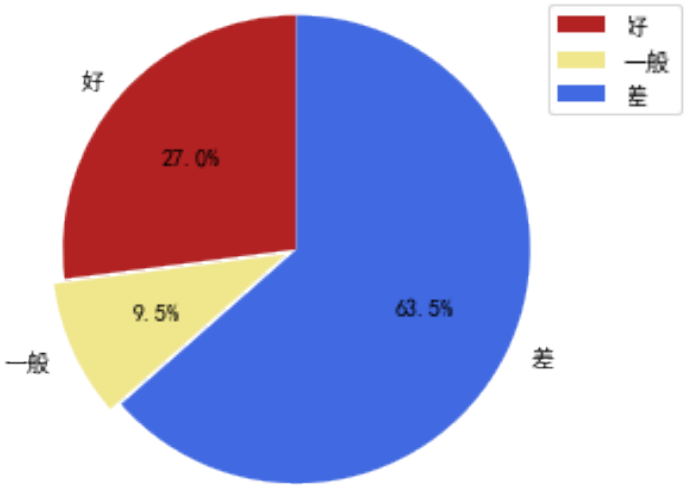

图 **13**　满意度分类数据占比图

7.2 提取表面语言特征和时序特征

参考胡泽 [12] 的研究，得知表面语言特征包含了问答对统计特征，这些特征反映了问答对之间的关系，例如 SLF4 和 SLF6 可能代表留言与答复的相关程度。而时许特征是通过相关部门对留言的响应时间去衡量的，所以推测表面语言特征和时许特征对于评价在线留言答复意见可能会有一定的帮助。本题总共提取了 8 个表面语言特征和 1 个时许特征（如表 2 所示），所有特征均被正则化为 0~1 之间，各个特征的详细介绍如下 [12]：

SLF1：答复长度。该特征易被提取，而且崔敏君等人 [13] 的研究也说明了答复长度对预测答复质量有显著的表现。

SLF2：答复分词后词汇数。该特征表示分词后答复文本中的词汇个数。

SLF3：答复分词后词汇数（去停用词）。该特征表示分词且移除停用词后，答复文本中的词汇个数。

SLF4：问答对重叠个数。该特征反映了答复意见和留言问题的相关性，一个好的答复意见和留言问题具有很大的相关性。

SLF5：问答对重叠个数（去停用词）

SLF6：问答对相似度（余弦相似度）。该特征给出了一个留言问题和它的答复意见之间的余弦相似度，一个好的答复意见中应该包含较大比例出现在留言问题中的词语。

SLF7：问答对长度之比。该特征表示留言问题与它的答复意见的原始长度之比。好的答复意见中通常有足够长的文本长度，能够包含一些详细的解释保障它的可信度。

SLF8：问答对长度之比（去停用词）

TF1：问答时间差。该特征反映了相关部门回复的及时性，一个好的答复有一个较小的响应时间。

表 **2**　非典型文本特征表

变量名	特征名称
SLF1	回答长度
SLF2	回答分词后的词语个数
SLF3	回答分词后的词语个数（去停用词）
SLF4	问答对重叠词个数
SLF5	问答对重叠词个数（去停用词）
SLF6	问答对分词相似度
SLF7	问答对长度之比
SLF8	问答对长度之比（去停用词）
TF1	时间差

7.2.1 卡方 (χ^2) 检验

卡方检验的思想是通过观察实际值与理论值的偏差来观察理论是否正确，卡方检验在文本分类领域方面可以实现特征选择，当特征数量较多时可以采用卡方检验进行特征选择，有利于达到好的分类效果。

具体原理为：如果一个特征 K_i 与一个类别 Y_j 相互独立，那么可以认为该特征对 Y_j 没有任何表征作用，用特征 K_i 与类别 Y_j 没有相关性作为原假设，卡方值越大，说明与原假设的偏离越大，则也说明卡方值越大相关度越高，得到的卡方值进行降序从而得到特征排名 [14]。

7.2.2 典型非文本特征重要性分析

通过对爬取的文本数据预处理以及分析和计算，得到 9 个特征所对应的数值，将 9 个特征对应的数值构成一个特征向量。将该向量中的数值正则化到 0 到 1 之间，其目的是统一不同的典型非文本特征的量纲，方便数据比较和共同处理，同时可以有效防止过拟合，并提高模型的泛化能力。

为了计算每个特征对答复质量评价的相对贡献，将正则化的特征向量运用卡方检验得到卡方值，将卡方值正则化再进行特征重要性排名 [12]。其结果如下表 3 所示。

表 **3**　典型非文本特征重要性排名表

典型非文本特征	特征重要性
SLF5	0.4394
TF1	0.2191
SLF4	0.1551
SLF8	0.0764
SLF7	0.0726
SLF1	0.0120
SLF6	0.0101
SLF2	0.0062
SLF3	0.0002

通过分析观察 9 个典型非文本特征重要性排名表，可以得出问答对重叠词个数（去停用词）（SLF5）的特征重要性最高，其次分别为时间差（TF1）、问答对重叠个数（SLF4），表明它们对答复意见的评价都有相对较大的贡献。实验得到的结果与前人的工作结果不同，他们的结果显示回答长度这个特征对于回答质量的预测是十分重要的，而我们的实验结果则显示回答长度这个特征对于回答质量的预测并没有显著的影响，反而是问答对之间相似性的特征在预测回答质量时更加显著。另外，问答对时间差（TF1）在回答质量预测上也十分重要。

为了对典型非文本特征深入分析, 对特征重要性排名 Top5 的特征在数据集中的分布进行作图观察：

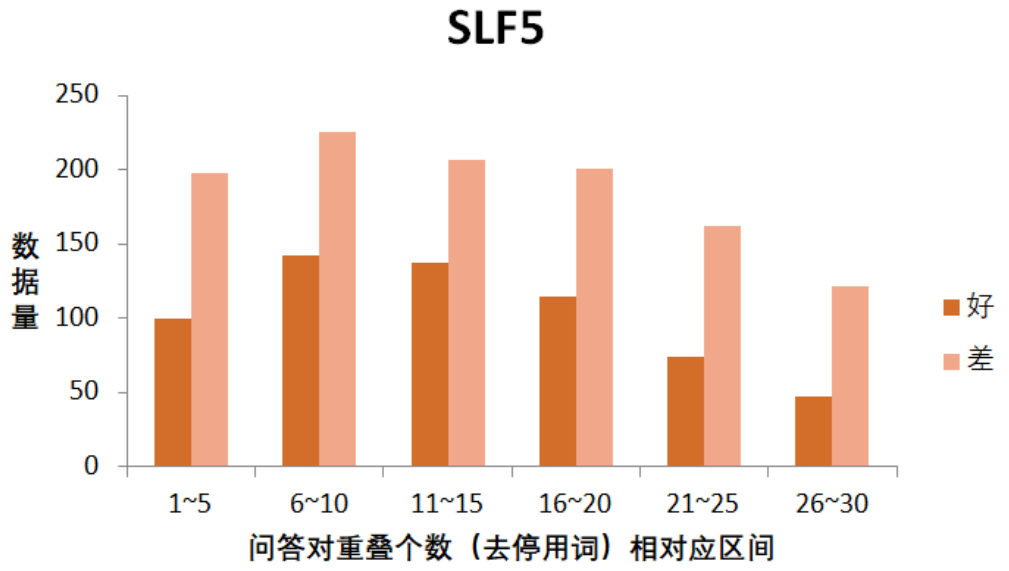

图 **14**　问答对重叠个数（去停用词）特征在数据集中的分布

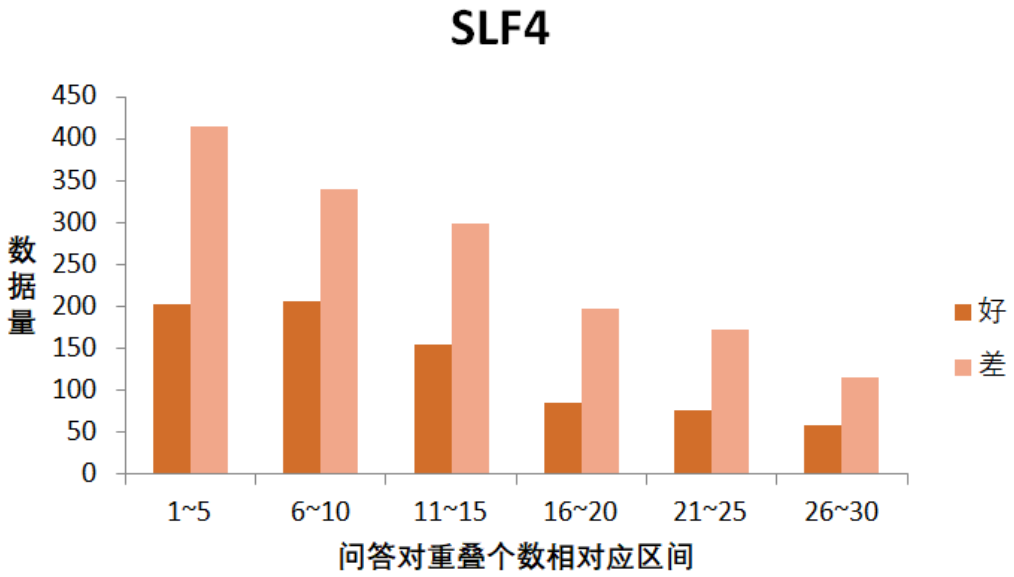

图 **15**　问答对重叠个数特征在数据集中的分布

从图 14 和图 15 可知，在问答对重叠个数少的情况下，答复质量为差的文本远远高于答复质量好的文本；在问答对重叠个数为 11~15 区间时，答复质量为好的文本与答复意见为差的文本的差值比例最少，表明问答对重叠个数在该区间答复质量为好的文本的概率更大，比较图 11 和图 12 可知，去停用词的问答对重叠个数作为特征比未经过停用词去除的问答对作为特征得到答复质量为好的文本的数据量更多，概率更大。

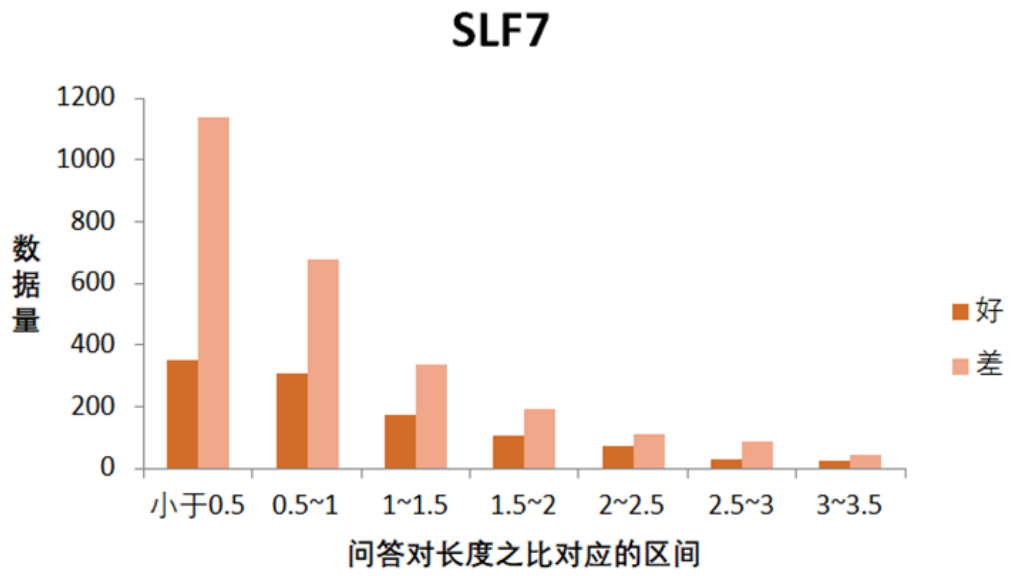

图 **16**　问答对长度之比特征在数据集中的分布

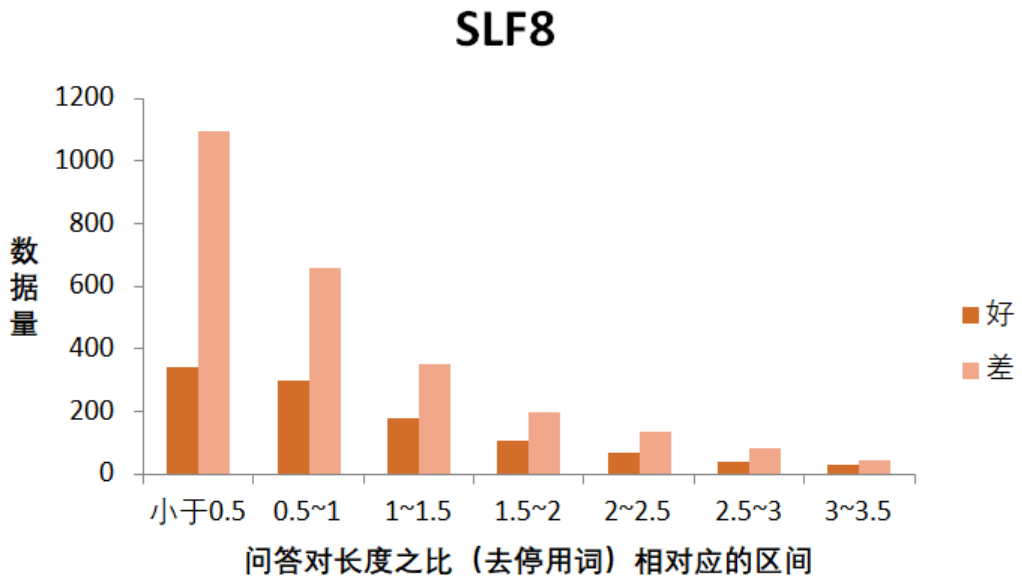

图 **17**　问答对长度之比（去停用词）特征在数据集中的分布

从图 16 和图 17 可知，在问答对长度之比较小的情况下，答复质量为差的文本远远高于答复质量好的文本，随着问答对长度之比的增长，答复质量为差的文本数量与答复质量好的文本相差的比例逐渐减小。比较图 16 和图 17 可知，问答对长度之比作为特征和问答对长度之比（去停用词）作为特征相比差别不大。

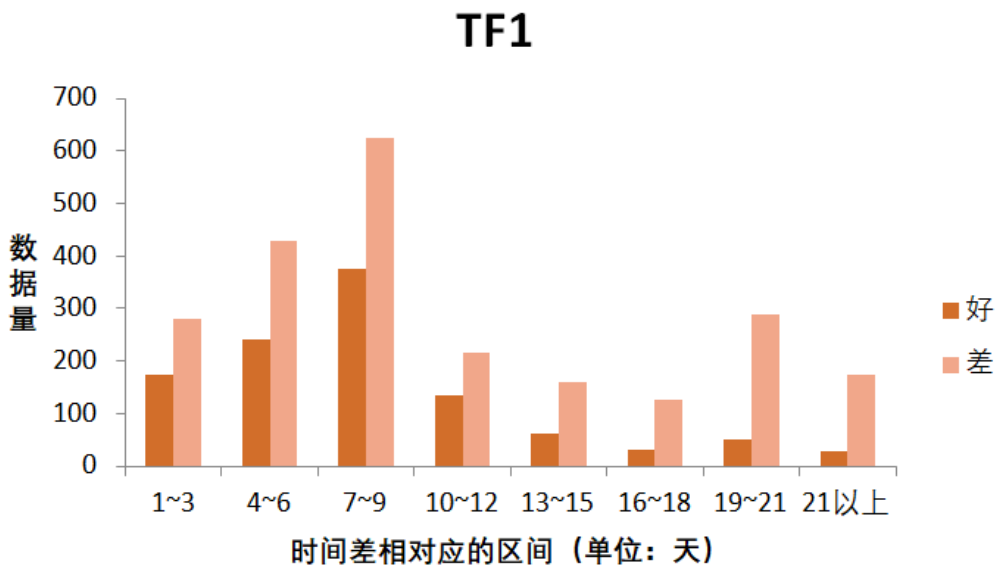

图 **18**　时间差特征在数据集中的分布

由图 18 可知，虽然总体分析，答复质量为差的文本数量都比答复质量为好的文本数量要多，但在时间差较小时，可以明显看出答复质量为差的文本与答复质量好的文本相差的比例是最小的，时间差较大时，两者数量比例也较大。主要原因是热点问题具有时效性，在留言与答复时间差较小时，可以快速高效打消人们所关注热点问题的疑惑，大大增加了答复意见为好的概率。

以上表明在网络问政平台上，留言用户更关注相关部门工作人员回答的相关性和高效性。一个和留言问题高度相关，并且能够高效率回应的答复，通常被考虑成好的回答，而与问题相关性不强的回答，即使长度很长，也不能有效解决留言用户的问题，通常会被用户看作低质量的回复。

7.3 答复质量评价模型

7.3.1 随机森林分类器

随机森林（RandomForest）是在 bagging 算法基础上更进一步，bagging 算法是从所有文本中重采样出 n 个文本构建分类器，然后重复 m 次此过程获得 m 个分类器，最后根据这 m 个分类器的投票结果决定文本属于哪一类。随机森林进行分本分类步骤如下：

1. 使用 Bootstrap 随机选取 n 个留言文本;
2. 第二步使用这 n 个文本的构建决策树;
3. 重复上述两步构建 m 棵决策树；
4. 每个决策树为一类投票，得票多者为最终结果。

7.3.2 模型的训练

将爬取得到的 4000 条有满意度打分的留言数据分成 80% 的训练集和 20% 的测试集，根据表 2 的特征排名结果选取非文本典型性文本特征作为答复质量评价模型的输入，利用随机森林分类算法进行答复质量评价模型的分类训练。

根据表 2 的排名选择了得分比较高的文本特征和全部文本特征，分别作为答复质量评价模型的输入，通过对比发现输入全部文本特征时模型的分类效果更好，可得宏 F1 值为 0.6，宏 P 值为 0.69，宏 R 值为 0.6，并且发现训练得到的分类模型对区分差评论效果更好，宏 F1 值达到了 0.85。

7.4 答复意见评价

利用 Python 对附件四的留言长度、答复分词后的词汇数、答复分词后词汇数（去停用词）、留言问题与答复意见的相似度等 9 个非典型文本特征进行计算，并将得到的特征数值输入答复意见评价模型进行满意度分类，部分预测结果如图所示。

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	predict
185986	UU008363	强烈呼吁	2011-10-3	朱公路。	“UU00836	2012-2-28	差
185799	UU008785	燃油税费	2012-9-4	底多少费	西地省平	2013-1-6	差
184423	UU0082115	对G7县文	2018-10-11	购买相关	“UU00821	2018-10-24	好
181603	UU008194	强烈反对	2018-6-12	务表现而	“UU00819	2018-7-4	差
181267	UU008766	汽车北站	2018-12-12	周围居民	您的留言	2019-1-8	差
180537	UU0082287	为促进张	2019-1-21	，中湖乡	你好，你	2019-1-28	差
179893	A000106666	H市有哪些	2018-9-9	久都没人知	您好，您	2018-9-28	差
179880	A00091124	H市学院名	2018-9-24	子和一支笔	您好，您	2018-9-28	差
179855	A000105818	建议H市旦	2018-10-24	以后，路上	您好，您	2018-10-30	差
179661	UU008403	咨询H市址	2019-6-20	地点，条	您好，你	2019-6-24	好
179652	UU0082147	关于H市电	2019-6-28	想问问咱H	您好，您	2019-7-1	好
179601	UU008609	咨询身份	2019-9-2	怕人用身	“UU00860	2019-10-8	差
179586	UU008369	咨询H市户	2019-9-16	要把孩子的	“UU00836	2019-10-8	差
179560	A0006565	H市招投标	2019-10-15	改革政策背	您好，你	2019-10-23	差

图 **19**　附件 **4** 中部分预测结果

八、 模型的评价

本文模型的建立更多是基于合理推导和理性的分析，并且对于部分重要公式和算法给出了保证的定义出处和计算步骤，对基于自然语言处理技术建立的三种不同功能模型进行了全面的分析和讨论，使得模型具有较多的理论支持和较好的容错性。

其中留言分类模型的建立是通过构建9种不同分类模型进行比较,得到TF-IDF+SVM分类模型分类效果最好，宏 F1 值达到了 0.9084，并且在城乡建设、环境保护、交通运输等 7 个类别 F1 都较高，说明该模型具有很好的实用性。但该留言分类模型都是基于传统的特征提取算法和分类器来构建的，还有一定的改进空间。

针对热点问题挖掘模型，热度指数较高的热点问题相对应的留言数、点赞数和反对数也相对较高，即人们对该话题的参与度和关注度也越高，这表明我们构建的热度指数模型在一定程度上可以反映热点问题的热度，具有一定的使用意义。但该模型是通过查阅一定的文献来确定影响热度指数的指标，并依据参考文献和实验分析确定影响热度指数指标的权重，具有一定的局限性。

针对答复意见评价模型，由于爬取的数据中数据量不够大，基于满意度分类的数据不平衡，满意度为差的数据比较多，导致模型分类效果对满意度为差的留言信息分类效果更好，这在一定程度上限制答复意见质量评价分类模型的推广和应用。

参考文献

[1] 贺科达, 朱铮涛, 程昱. 基于改进 TF-IDF 算法的文本分类方法研究 [J]. 广东工业大学学报,2016,33(5):50-53.

[2] 黄春梅, 王松磊. 基于词袋模型和 TF-IDF 的短文本分类研究 [J]. 软件工程,2020,23(3):1-3.

[3] 任世超. 基于机器学习的文本分类算法研究 [D]. 中国四川成都: 成都信息工程大学,2019.

[4] 阮光册, 谢凡, 涂世文. 基于 Word2vec 的图书馆推荐系统多样性问题应用研究 [J]. 图书馆杂志,2020,3:124-132.

[5] SVM 多分类两种方式 [EB/OL]. https://blog.csdn.net/xfchen2/article/details/79621396,2018-03-21.

[6] 逻辑回归算法（二分类）[EB/OL]. https://www.cnblogs.com/xyp666/p/9085699.html,2018-05-24.

[7] 贾会强. 基于 KNN 算法的藏文文本分类关键技术研究 [J]. 西北民族大学学报 (自然科学版),2011,32(83):24-26.

[8] 段丹丹，唐加山，温勇，袁克海．基于 BERT 的中文短文本分类算法的研究．计算机工程. https://doi.org/10.19678/j.issn.1000-3428.0056222

[9] 郑斐然, 苗夺谦, 张志飞, 等. 一种中文微博新闻话题检测的方法 [J]. 计算机科学,2012,39(1):138-141.

[10] 庄婷婷, 王平, 程齐凯. 一种时间情境依赖的微博话题抽取方法 [J]. 信息资源管理学报,2013,3:40-46.

[11] 宋逸群, 王玉海, 聂梅, 等. 大数据透视下的京津冀协同发展民生热点问题探究 [J]. 领导之友,2017,239:61-68.

[12] 胡泽. 在线问诊服务回答质量评价方法研究 [D]. 中国哈尔滨: 哈尔滨工业大学,2019.

[13] 崔敏君, 段利国, 李爱萍. 多特征层次化答案质量评价方法研究 [J]. 计算机科学,2016,43(1):94-102.

[14] 利用卡方检验进行特征选择及实践 [EB/OL]. https://blog.csdn.net/Johnxyz/article/details/82494707,2018-09-07.

[15] 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]. 内蒙古科技与经济,2019,21:67-69.

表 **4**　不同的 **a** 取值下得分最高的前 **20** 个主题词

$W(a=0.1)$	$W(a=0.2)$	$W(a=0.3)$	$W(a=0.5)$	$W(a=0.8)$	$W(a=1)$
雅居乐	临路	临路	临路	临路	临路
望江	敬老院	敬老院	敬老院	敬老院	敬老院
铁路线	给到	链	链	链	链
和稀泥	研	悲剧	悲剧	悲剧	悲剧
租房子	节省	给到	成年人	成年人	成年人
给到	链	研	给到	给到	给到
研	悲剧	节省	研	研	研
节省	雅居乐	成年人	节省	节省	节省
按计划	望江	半岛	半岛	半岛	半岛
太极	铁路线	废	废	废	废
大伙	成年人	天麓	天麓	天麓	天麓
刘	和稀泥	雅居乐	雅居乐	雅居乐	雅居乐
删除	半岛	望江	望江	望江	望江
半岛	废	铁路线	铁路线	铁路线	铁路线
废	天麓	和稀泥	和稀泥	和稀泥	和稀泥
天麓	租房子	租房子	租房子	租房子	租房子
链	按计划	按计划	按计划	丁字湾	丁字湾
悲剧	太极	太极	太极	按计划	按计划
成年人	大伙	大伙	大伙	太极	太极
敬老院	临路	刘	刘	大伙	大伙