政务信息的文本挖掘应用

摘要：随着网络的发展，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，给主要依靠人工进行的留言划分与热点分析的相关部门的工作带来极大的挑战。本文将基于自然语言处理技术与文本挖掘技术对互联网公开来源的群众问政留言记录及相关部门对群众留言的答复意见等信息进行处理，解决群众留言分类和热点问题挖掘的问题，并针对相关部门对留言的答复意见给出一套答复质量评价方案。

关键词：多类分类模型；文本相似度；命名实体识别；自然语言处理

Text mining application of government affairs information

【**Abstract**】With the development of the Internet, the Internet questioning platform has gradually become an important channel for the government to understand public opinion, gather people’s wisdom and public opinion. The amount of text data related to various social conditions and public opinions continues to increase. The work of relevant departments brings great challenges. Based on natural language processing technology and text mining technology, this article will deal with the information of the public's public information sources on the Internet and the response of the relevant departments to the public's messages, etc. Give a set of reply quality evaluation plan to the reply comments of the message.

【**Key words**】 Multi-class classification model; Text similarity; NER; NLP

目　录

1. 挖掘目标

随着网络的发展，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，各类与社情民意相关的数据不断攀升，给主要依靠人工进行留言划分与热点分析的相关部门工作带来了极大的挑战。

本文利用自然语言处理技术与文本挖掘技术对互联网公开来源的群众问政记录及相关部门对部分群众留言的答复意见进行建模，以解决以下三个问题：

问题1：利用群众问政记录中经过人工分类后的群众留言数据，建立基于留言划分体系（一级标签）的多类分类模型，并使用F-score方法对该多类分类模型进行评价。

问题2：利用群众问政记录中的群众留言记录，将反映相同问题的群众留言进行归类，定义合理的热度评价指标，完成热点问题的识别。将热点问题的时间范围、特定地点、特定人群与问题描述进行提取，并生成热点问题表与热点问题留言明细表。

问题3：针对相关部门对部分群众留言的答复意见，利用文本相似度、多标签分类模型与词典等方法量化答复意见的相关性、完整性与可解释性，构建答复意见质量评价方案。

2. 分析方法与过程

2.1. 问题 1 的分析方法与过程

2.1.1. 处理流程

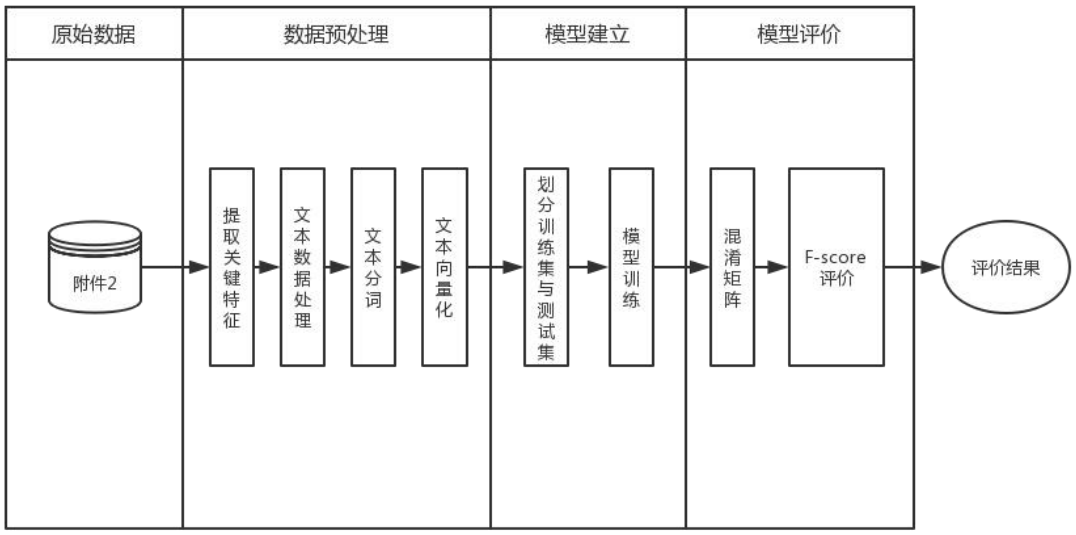

图 1　群众留言分类流程图

2.1.2. 数据预处理

2.1.2.1 数据描述

观察附件 2 中数据，数据中每条记录有 6 种特征，分别为留言编号、留言用户、留言主题、留言时间、留言详情与一级标签。

其中留言主题、留言详情与一级标签的数据为非结构化的文本信息，对于非结构化的文本需转化为结构化的数据。且留言主题与留言详情存在大量空格与无意义网址等干扰信息，需先进行文本清洗，避免影响后期模型的训练。

2.1.2.2 缺失值/重复值处理

对数据进行缺失值与重复记录查找，数据中并无缺失值与重复记录。

2.1.2.3 标签数量分布情况

对数据中各个标签类别的数据进行统计，统计结果如图 2 所示：

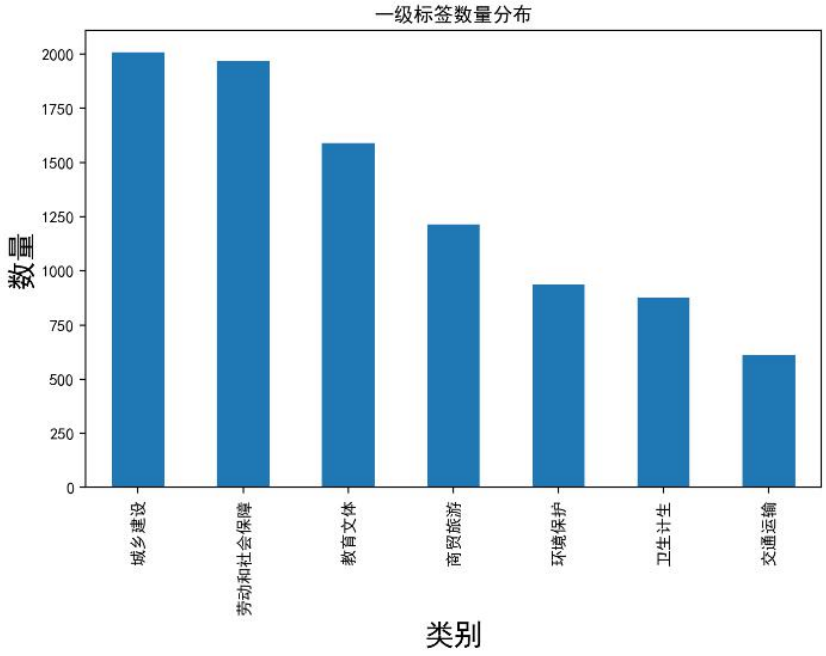

图 2　一级标签数量分布图

其中，城乡建设类数量最多，共有 2009 条记录；而交通运输类数量最少，仅有 613 条记录。各个类别数据数量并不均衡。

2.1.2.4 提取关键特征

在数据的 6 种特征中，留言编号、留言用户与留言时间对留言的分类并没有任何影响，本文将处理目标集中于留言主题、留言详情与一级标签等特征之中。因此提取留言主题、留言详情与一级标签特征进行下一步处理。

2.1.2.5 文本数据处理

(1) 特征数值化

将一级标签所在列的数据进行去重处理，提取数据中存在的所有标签类别，以 key 值保存

至字典中，并为每个类别设置对应的 value，每个标签都有其对应的数值，将非结构化文本标签转化为结构化的数值标签。

(2) 文本清洗

对于留言主题与留言详情中的文本进行处理，文本中存在的网址、文本序号、脱敏后的手机号码、结尾日期与其他的无意义的字符，都会影响后续文本的分词与模型的训练。

在此，本文使用正则表达式对数据的留言主题与留言详情依次进行过滤，删除无意义且影响结果的字符，达到文本清洗的效果。

2.1.2.6 文本分词

在训练分类模型之前，将非结构化的文本信息转化为计算机能够识别的结构信息，为了方便结构转化，作者对经过文本清洗后的留言主题与留言详情进行中文分词。这里采用基于 python 的 jieba 分词进行分词。jieba 分词涉及的算法：

- 采用基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG）
- 采用了动态规划查找最大概率路径，找出基于词频的最大切分组合
- 针对未登录词，采用了基于汉字成词能力的隐马尔可夫（HMM）模型

此外，jieba 分词还支持用户自定义词典与关键词提取等功能。

为了节省存储空间和提高搜索效率，在转换词向量之前还需要对分词后的留言主题与留言详情进行停用词过滤。停用词有两个特征：一是极其普遍，出现频率高；二是无明确意义，包含信息低。

本文使用 hanlp 提供的停用词表进行过滤，将分词后的主题与详情与停用词表中的词语进行匹配，若匹配成功则删除该词语。

2.1.2.7 文本向量化

由于留言主题与留言详情在一定程度都能表达与标签的联系，因此将分词过滤后的留言主题与留言详情的词语进行合并。

在留言信息经过文本分词后，将词语转换为词向量，完成非结构化文本信息转化为结构信息的过程。这里采用 TF-IDF 算法，将留言信息转换为权重向量。TF-IDF 算法计算步骤：

第一步，计算词频：

$$\text{词频(TF)} = \text{某个词在文章中的出现次数}$$

考虑到文章有长短之分，为了便于不同文章的比较，进行“词频”标准化：

$$\text{词频(TF)} = \frac{\text{某个词在文章中的出现次数}}{\text{文章的总词数}}$$

第二步，计算逆文档频率：

这时，需要一个语料库(corpus)，来模拟语言的使用环境。如果一个词语越常见，那么分母就越大，IDF 就越小越接近于 0。分母之所以加 1，避免分母为 0 的情况（即所有文档都不包含该词）。

$$\text{逆文档频率(IDF)} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数} + 1}\right)$$

第三步，计算 TF-IDF：

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

本文以分词后留言信息的所有词语建立语料库，计算每个词的 TF-IDF 权重。文本的每个词语都由特定的权重来表示。词语的重要性随着它在当前文本的出现次数成正比增加，但同时随着它在语料库的出现频率成反比下降。

2.1.3. 模型建立

群众留言分类问题是常规的多类分类问题，数据中共有 7 种不同的标签。对于解决多类分类问题，一般可分为三种方案：

一是直接构建多类分类器。该方案由于变量过多，导致计算速度慢且精度不高。

二是构建一对一分类器。该方案由于类别两两配对构建分类器，当分类类别过多时并不适用。

三是构建一对多分类器。本文将使用该方案构建基于线性支持向量机的多类分类模型。

2.1.3.1 SVM 算法

支持向量机（Support Vector Machines，SVM）是一种二分类模型，其基本模型是定义为特征空间上的间隔最大的线性分类器。

SVM 的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。

SVM 的学习算法就是求解凸二次规划的最优化算法。

2.1.3.2 1-v-r SVMs 算法[1]

该方法依次用一个两类 SVM 分类器将每一类与其他所有类别区分开来，得到 k 个分类函数。分类时将未知样本分类为具有最大分类函数值的那个类别。

2.1.3.3 划分训练集与测试集

将留言信息的词向量特征与对应的标签按一定比例划分为训练集与测试集，划分后类别的数据分布保持不变，避免数据量少的标签被全部划分进训练集或测试集，影响分类器的训练。

本文使用 sklearn 库的 train_test_split()函数进行划分，训练集与测试集的比例为 8：2。

2.1.3.4 模型训练

本文使用 sklearn 库的 LinearSVC 模型对划分后的训练集进行训练。LinearSVC 模型是基于 liblinear 库实现的 1-v-r SVMs 分类模型。

2.1.4. 模型评价

利用划分后的测试集对训练完成的 LinearSVC 分类模型进行评价。由于每种标签的数据量并不相同，当数据不平衡时，我们无法使用准确率来评价模型的质量。本文将使用混淆矩阵观察预测标签与实际标签之间的差异，并构建 F-Score 方法评价模型的质量。

2.1.4.1 混淆矩阵

根据测试集预测结果生成混淆矩阵，结果如图所示：

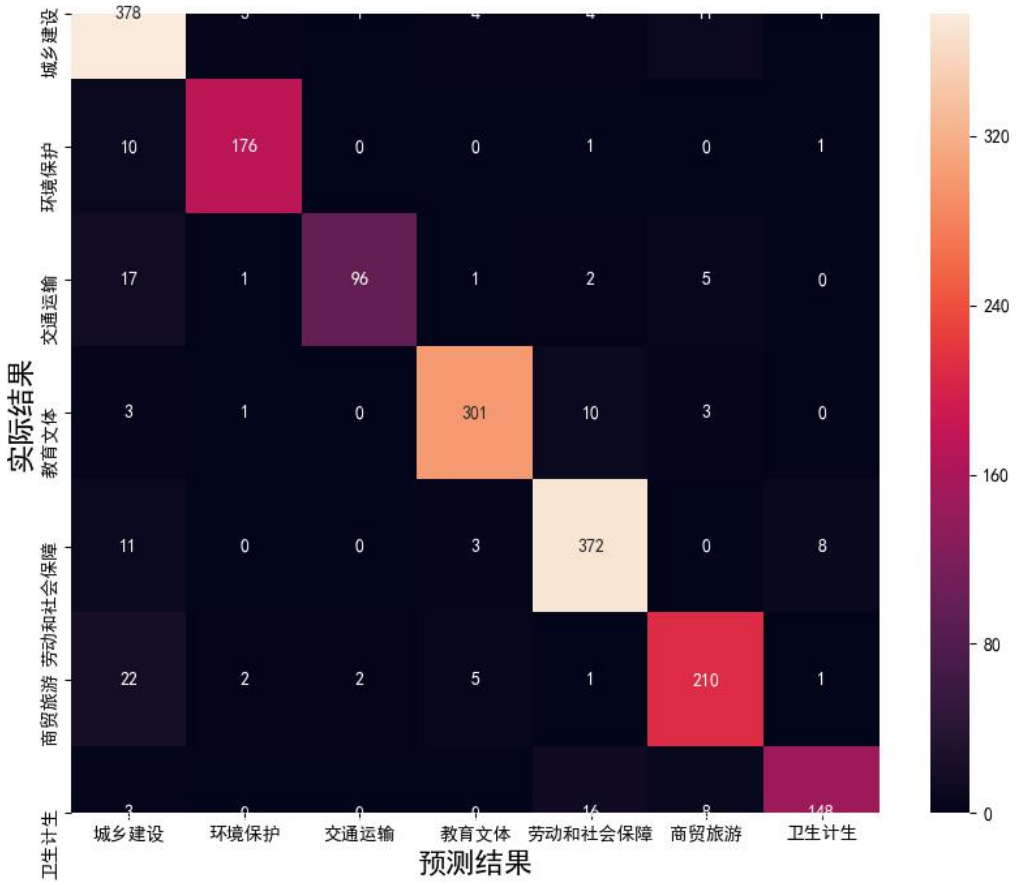

图 3　群众留言分类模型混淆矩阵

其中，城乡建设与劳动和社会保障两种标签的预测错误率较高。

2.1.4.2 F-Score 评价指数

对于多类分类模型的 F-Score 评价公式为：

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

其中 P_i 为第 i 类的精确率，R_i 为第 i 类的召回率。

利用 sklearn 库的 classification_report 函数显示各个类别的主要分类指标文本报告。

	precision	recall	f1-score	support
城乡建设	0.85	0.94	0.89	402
环境保护	0.96	0.94	0.95	188
交通运输	0.97	0.79	0.87	122
教育文体	0.96	0.95	0.95	318
劳动和社会保障	0.92	0.94	0.93	394
商贸旅游	0.89	0.86	0.88	243
卫生计生	0.93	0.85	0.89	175

图 4　分类指标文本报告

获取文本报告的精确率与召回率，计算该分类模型的 F-Score 评价指数，最终该模型的 F-Score 指数约为 0.9079。

2.2. 问题 2 的分析方法与过程

2.2.1. 处理流程

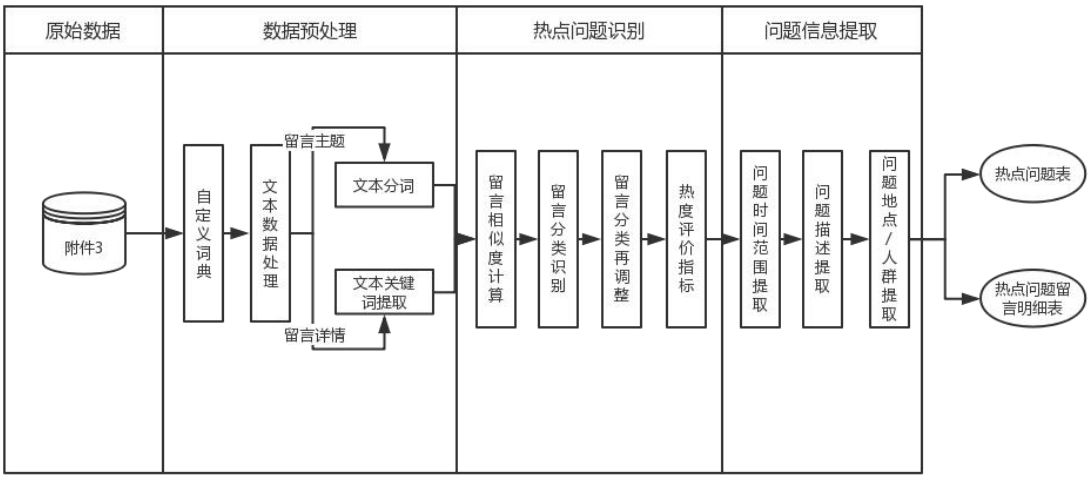

图 5　热点问题挖掘流程图

2.2.2. 数据预处理

2.2.2.1 数据描述

观察附件 3 中数据，共有 4326 条留言记录，每条记录有 7 种特征，分别为留言编号、留言用户、留言主题、留言时间、留言详情、反对数与点赞数。

其中留言主题、留言详情的数据为非结构化的文本信息，对于非结构化的文本需转化为结构化的数据，且留言主题与留言详情存在大量空格与无意义网址等干扰信息，需进行文本清洗后才能进

行处理，避免影响后续文本相似度计算。

2.2.2.2 缺失值/重复值处理

针对数据进行缺失值与重复记录查找，数据中并无缺失值与重复记录。

2.2.2.3 自定义词典

人工提取留言记录中出现的地点，例如：西湖街道、茶场村等，构建完善政务地点词典。

收集代表人群的词语，例如：业主、学生等，构建完善政务人群词典。

2.2.2.4 文本数据处理

对于留言主题与留言详情中的文本进行处理，文本中存在的网址、文本序号、脱敏后的手机号码与其他的无意义的字符，会干扰后续文本的分词与相似度计算。

在此，本文使用正则表达式对数据的留言主题与留言详情依次进行过滤，删除无意义且影响结果的字符，达到文本清洗的效果。

2.2.2.5 文本分词和文本关键词提取

使用基于 python 的 jieba 分词对文本进行分词与关键词提取。

加载自定义的政务地点词典，对经过文本处理后的留言主题进行分词，并使用 hanlp 提供的停用词表进行停用词过滤。

使用 jieba 分词中基于 TextRank 算法的关键字提取对经过文本处理的留言详情进行分词。提取每条留言详情中排名前 20 的关键词，且筛选词性为('n','ns', 'nr', 'nt', 'v', 'vn','x')，即名词、地点名、人名、机构名、动词、动名词与非语素字（由于政务地点词典主要是针对后续使用 hanlp 进行命名实体识别而训练，因此并未词典中设置对应词性，jieba 分词默认把词典中没有词性标注的未登录词划分为非语素字）。

2.2.3. 留言相似度

BOW 模型（词袋模型）：词语出现的次数表示，不考虑关键词的顺序，仅将文档看成关键词出现概率的集合，每个关键词之间相互独立。

TF-IDF 模型：词语的重要性随着它在文本出现的次数成正比增加，但同时随着它在语料库中出现的频率成反比下降。

相似度模型构建思路：首先，根据待分析的所有文本构建对应的 BOW 模型；其次，基于 BOW 模型构建语料库，训练基于词袋语料库的 TF-IDF 模型；最后，对 TF-IDF 权重转化后的词袋语料库进行稀疏矩阵相似度（基于余弦相似性）计算。

本文将留言相似度的计算问题划分为主题相似度的计算问题与详情相似度的计算问题。

首先，提取分词后的留言主题信息训练主题相似度模型和提取分词后的留言详情信息训练详情相似度模型。其次，利用训练好的 BOW 模型与 TF-IDF 模型将待计算的主题/详情向量化，并在稀疏矩阵中进行匹配。最终，将匹配结果按相似度大小进行排序，返回该主题/详情与其余主题/详情的相似度结果。

2.2.4. 热点问题识别算法

热点问题识别算法分为三个步骤。

首先利用留言主题与留言详情数据训练主题相似度模型和详情相似度模型；其次，对所有经预处理后的数据项依次进行留言分类识别（即算法 1）；最后，再对经过第一次识别后的数据项依次进行第二次调整（即算法 2）。

第一次分类识别会粗略的将所有主题相似度或详情相似度满足阈值的数据项划分为一类，优先匹配主题相似度。

第二次分类识别会将所有非匹配最大相似类别的数据项重新进行匹配。

算法 1 留言分类识别算法

第 1 步：设置主题阈值与详情阈值

第 2 步：根据主题相似模型与详情相似模型，获取主题相似列表与详情相似列表

第 3 步：判断当前数据项是否已分类。如果数据项已分类，跳出执行下一个数据项的识别；如果数据项未分类，执行第 4 步

第 4 步：查找当前数据项的最大相似已分类项（满足阈值且已分类的相似度最大的数据项），将当前数据项与最大相似已分类项划分一类。优先在主题相似列表中匹配，若主题相似列表无满足条件的数据项，则进一步在详情相似列表匹配

第 5 步：若当前数据项没有最大相似已分类项，则单独划分为一类

第 6 步：遍历主题相似列表与详情相似列表，查找相似度满足阈值且未分类的数据项，与当前数据项划分为同一类

算法 2 留言分类调整算法

第 1 步：判断当前数据项是否为最大相似项。如果是，跳出执行下一个数据项；如果不是，执行第 2 步

第 2 步：以当前数据项的最大相似项递归调用当前算法。确保当前数据项的最大相似项已经与其最大相似项划分为一类

第 3 步：判断当前数据项是否与最大相似项划分为一类。如果不是，则将当前数据项与最大相似已分类项划分一类

2.2.5. 热度评价指标

某一时段内群众集中反映的某一问题可称为热点问题。问题的热度往往取决于群众对问题的关注度。在网络问政平台上，群众对某问题的关注度直接呈现于该问题的留言数量与对该问题留言的评价（即点赞数或反对数）。因此，本文从留言数量与留言评价的角度出发，构建了基于留言得分的热度评价指标。留言得分主要分为两个部分：

（1）问题的留言得分

由于网络问政平台存在用户重复发布留言的可能性，因此留言得分又分为无重复发布留言得分

与重复发布留言得分。

假设在同一问题中，无重复发布留言用户的总留言数为 m，第 i 个重复发布留言用户的总留言数为 m_i，每条留言的得分为 ∂：

那么，无重复发布留言得分为 $m*\partial$

由于考虑到同一用户就同一问题不断发布相同留言，构成类似“刷票”的行为，会极大的影响评价指标的平衡，因此对该部分留言数量进行降权处理。

故，重复发布留言得分为 $\sum_{i=1}^{k}[\log_2(2+m_i)*\partial]$

综上所述，问题的留言得分为 $\sum_{i=1}^{k}[\log_2(2+m_i)*\partial]+m*\partial$

（2）问题的评价得分

问题的评价程度体现于问题留言的点赞数与反对数，考虑到同一个问题可能存在多条留言，且留言的评价程度可能各不相同，因此我们提取同一问题的所有留言中最大的反对数与最大点赞数进行评价。

假设在同一问题中，留言最大点赞数为 $g_{\max}$，留言最大反对数为 $b_{\max}$，每个点赞/反对的得分为 ε。

那么，问题的评价得分为 $(g_{\max}+b_{\max})*\varepsilon$。

为了加强留言得分的权重，将 ∂ 值设置为 1，ε 值设置为 0.005。故热度评价公式：

$$f=\sum_{i=1}^{k}[\log_2(2+m_i)*\partial]+m*\partial+(g_{\max}+b_{\max})*\varepsilon(\partial=1,\varepsilon=0.005)$$

2.2.6. 问题信息提取

提取热度评价排名前 5 的热点问题留言，提取每个问题的时间范围、地点/人群与问题描述。

2.2.6.1 时间范围提取

筛选同一问题的留言信息，将信息的留言时间转化为时间序列。提取时间最小的日期与时间最大的日期，若两个日期相同，则直接返回该日期；若两个日期不相同，将时间较小的日期放置在前，时间较大的日期放置在后，中间用字符串类型“至”连接，将该结果返回。

2.2.6.2 问题描述提取

问题描述提取属于 NLP 中的自动摘要问题，摘要的实现方法有基于抽象的摘要和基于抽取的摘要两种。

基于抽象的摘要，系统将试图去理解文本的内容，将视为重要的信息进行抽取，同时抽象释义出源文本的内容，该方法技术难度大且效果欠佳。因此本文将使用基于抽取的摘要来实现问题描述的提取。

基于抽取的摘要，系统将抽取整个对象集合，但并不对其进行修改。例如，关键词提取。但关键词提取是选择单词或者短语进行抽取，而问题描述提取将抽取出一个关键句作为问题的描述。由于留言主题在很大程度上完整概括了留言的信息，因此本文将使用 TextRank 算法提取同一问题不同留言主题中的关键主题作为问题描述。

TextRank 算法的打分思想是由 PageRank 的迭代思想衍生过来，公式如下所示：

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

$WS(V_i)$ 表示第 i 个句子的权重，$\sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$ 表示每个句子对所在文本的贡献程度。求和公式的分子上的 w_{ji} 表示两个句子的相似程度，分母则是文本中相对应的部分的句子的权重之和，而 $WS(V_j)$ 代表上次迭代 j 的权重，整个公式是一个迭代的过程。

对于两个句子的相似程度计算，本文使用 BM25 算法进行计算。

当 TextRank 对问题的所有留言主题进行打分后，筛选排名第一的主题作为问题描述。

2.2.6.3 地点/人群提取

问题信息的提取阶段需要将问题的特定地点与特定人群提取处理，此部分内容涉及命名实体识别问题，因此本文使用 HanLP 自然语言处理工具进行命名实体识别。

HanLP 是一系列模型与算法组成的 NLP 工具包，其具备功能完善、性能高效、架构清晰、预料时新，可自定义的特点。

本文将利用 HanLP 的地点识别功能并结合在数据预处理阶段训练的政务地点词典，实现热点问题地点的识别。再通过 HanLP 的自定义词性功能新建人群词性 np 并结合在数据预处理阶段训练的政务人群词典，实现问题人群的识别。

由于留言主题在很大程度上完整概括了留言的信息，而在问题描述提取阶段，我们使用 TextRank 算法提取出最有表征性的主题作为问题描述，因此在这个阶段我们对问题描述进行问题地点识别与问题人群识别，将词性为‘ns’与‘np’的词语提取，生成地点+人群的结果。

最终，将提取的信息按热度排名生成为“热点问题表.xls”，并提取问题对应的留言信息生成为“热点问题留言明细表.xls”，部分生成结果如下图 6、7 所示：

	A	B	C	D	E	F
1	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
2	1	32	54.69386	2019-07-05至2019-09-01	武广新城伊景园滨河苑	武广新城伊景园滨河苑违法捆绑销售车位,求解决
3	2	33	53.11	2019-11-02至2020-01-26	A2区丽发新城小区	举报A2区丽发新城小区附近仍存在非法搅拌站
4	3	50	21.06	2018-11-15至2019-12-15	A市人才	反映A市人才租房购房补贴问题
5	4	843	11.485	2019-08-19	A市A5区汇金路五矿万境K9县	A市A5区汇金路五矿万境K9县存在一系列问题
6	5	827	10.025	2019-02-14至2019-09-09	A7县星沙四区凉塘路旧城	A7县星沙四区凉塘路的旧城改造要拖到何时？

图 6　热点问题表.xls 截图

	A	B	C	D	E	F	G	H
1	问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
2	32	188801	A909180	投诉滨河苑针对广铁职工购房的霸王规定	2019-08-01 00:00:00	尊敬的张市长，您好！我	0	0
3	32	190337	A0009051	关于伊景园滨河苑捆绑销售车位的维权投诉	2019-08-23 12:22:00	投诉伊景园.滨河苑开发商	0	0
4	32	191001	A909171	A市伊景园滨河苑协商要求购房同时必须购买	2019-08-16 09:21:33	商品房伊景园滨河苑项目	12	1
5	32	195511	A909237	车位捆绑违规销售	2019-08-16 14:20:26	对于伊景园滨河苑商品房	0	0
6	32	195995	A909199	关于广铁集团铁路职工定向商品房伊景园滨	2019-08-10 18:15:16	尊敬的市政府领导，您好	0	0
7	32	196264	A0009508	投诉A市伊景园滨河苑捆绑车位销售	2019/8/7 19:52:14	A市伊景园.滨河苑现强制	0	0
8	32	199190	A0009508	关于A市武广新城违法捆绑销售车位的投诉	2019/8/1 22:32:26	武广新城为铁广集团的定	0	0
9	32	200085	A0001042	A市市政建设开发有限公司对广铁职工住宅工	2019-08-19 11:34:11	我是广铁的一名职工，对	9	2
10	32	204960	A909192	家里本来就困难,还要捆绑买卖车位	2019-08-21 18:12:20	我是广铁集团铁路职工,因	0	0
11	32	205277	A909234	伊景园滨河苑捆绑车位销售合法吗？！	2019-08-14 09:28:31	广铁集团强制要求职工购	1	0
12	32	205982	A909168	坚决反对伊景园滨河苑强制捆绑销售车位	2019-08-03 10:03:10	我坚决反对伊景园滨河苑	2	0
13	32	207243	A909175	伊景园滨河苑强行捆绑车位销售给业主	2019-08-23 12:16:03	您好！A市武广新城片区自	0	0
14	32	209506	A909179	A市武广新城坑客户购房金额并且捆绑销售车	2019-08-02 16:36:23	您好！由A市广铁集团发起	0	0
15	32	209571	A909200	伊景园滨河苑项目绑定车位出售是否合法合	2019-08-28 19:32:11	广铁集团铁路职工定向商	0	0
16	32	212323	A0002070	广铁集团要求员工购房时必须同时购买车位	2019-07-11 00:00:00	尊敬的领导：您好！我是	0	0
17	32	213584	A909172	投诉A市伊景园滨河苑定向限价商品房违规销	2019-07-28 13:09:08	投诉A市伊景园滨河苑定向	0	0
18	32	214975	A909182	关于房伊景园滨河苑销售若干问题的投诉	2019-08-22 00:00:00	尊敬的领导，您好！感谢	3	0
19	32	218709	A0001066	A市伊景园滨河苑捆绑销售车位	2019/8/1 22:42:21	伊景园滨河苑作为广铁集	1	0
20	32	218739	A909184	A市伊景园.滨河苑欺诈消费者	2019-08-24 00:00:00	A市伊景园滨河苑强行捆绑	0	0
21	32	220534	A0007509	投诉武广新城伊景园滨河苑为广铁集团的定	2019-08-12 12:37:28	投诉：武广新城片区伊景	0	0
22	32	222209	A0001717	A市伊景园滨河苑定向限价商品房项目违规捆	2019-08-28 10:06:03	广铁集团与伊景园滨河苑	0	0
23	32	223247	A0004475	投诉A市伊景园滨河苑捆绑销售车位	2019/7/23 17:06:03	关于铁广集团铁路职工定	0	0
24	32	224767	A909176	伊景园滨河苑车位捆绑销售！广铁集团做个	2019-07-30 14:20:08	伊景园滨河苑车位捆绑销	0	0
25	32	225479	A0004380	A市市政建设开发有限公司违规操作铁广职工	2019/7/5 1:55:26	A市市政建设开发有限公司	0	0

图 7　热点问题留言明细表.xls 截图

2.3. 问题 3 的分析方法与过程

2.3.1. 相关性参数

2.3.1.1 相关性

在群众问政留言记录中，相关部门对群众留言的答复意见是否与群众留言内容相关。

2.3.1.2 相关性量化流程

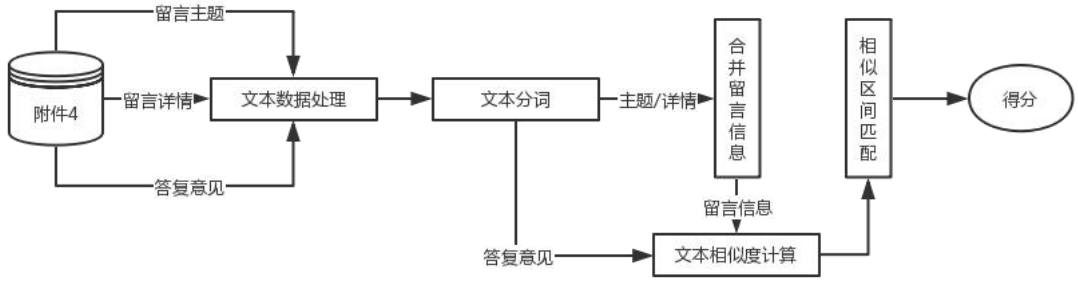

图 8　相关性量化流程图

2.3.1.3 文本数据处理

对于留言主题、留言详情与答复意见进行处理，文本中存在的网址、文本序号、脱敏后的手机号码与其他的无意义的字符，会干扰后续文本的分词与相似度计算。

在此，本文使用正则表达式对数据的留言主题与留言详情依次进行过滤，删除无意义且影响结果的字符，达到文本清洗的效果。

2.3.1.4 文本分词

使用基于 python 的 jieba 分词对经过文本数据处理的留言主题、留言详情与答复意见进行分词，并用停用表过滤文本中的停用词，最后将分词后的留言主题与留言详情进行合并，以便后续留言信息与答复意见的文本相似度计算。

2.3.1.5 文本相似度计算

BM25 算法是一种用来评价搜索词和文档之间的相关性的算法，是基于概率检索模型提出的算法。本文使用 BM25 算法计算留言信息与答复意见之间的相似度。

2.3.1.6 相似区间

根据计算所得的留言信息与答复意见的相似度，将相关性分为五个等级区间(+∞,0.1)、[0.1,0.05)、[0.05,0.01)、[0.01,0)、[0,0]分布对应 4 分、3 分、2 分、1 分、0 分。

2.3.2. 完整性参数

2.3.2.1 完整性

在群众问政留言记录中，相关部门对群众留言的答复意见是否满足答复性案语规范。

2.3.2.2 答复规范

相关部门对群众留言的答复意见在一定程度上属于答复性政府公文，因此需满足答复性案语的规范。

答复性案语规范如下：

- 开头格式。例：网友“xxxxxx”您好！您的留言已收悉。现将有关情况回复如下：
- 感谢致辞。例：感谢您对我们工作的支持、理解与监督！
- 结尾日期。例： 2018 年 12 月 3 日

2.3.2.3 完整性量化流程图

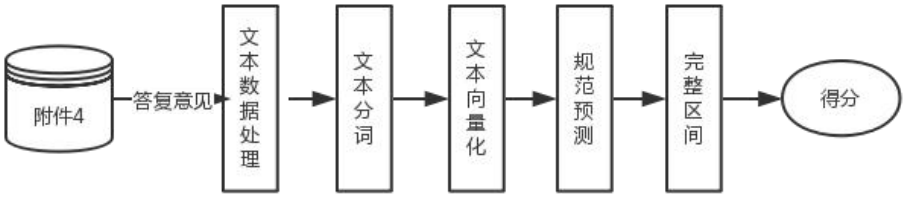

图 9　完整性量化流程图

2.3.2.4 规范预测

答复意见的完整性体现于答复意见是否满足答复规范。因此本文将答复意见的完整性量化为满足答复规范的数量。答复意见每满足一个规范获得 1 分，完整区间为[0,3]。

对于答复意见的预测问题，可转化为多标签文本分类问题。因此，本文训练了一个基于多项式朴素贝叶斯的多标签分类模型。

训练流程如下图所示：

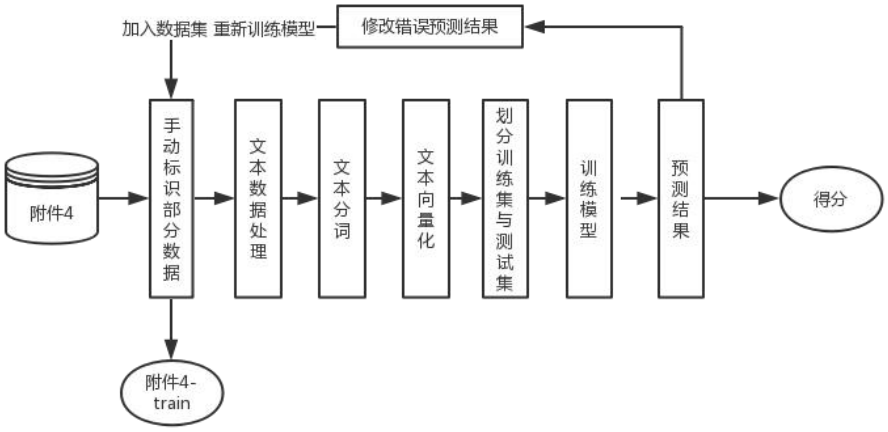

图 10　预测模型训练流程图

2.3.3. 可解释性参数

2.3.3.1 可解释性

在群众问政留言记录中，相关部门对群众留言的答复意见内容中的相关解释或理论支撑。

2.3.3.2 可解释性量化流程图

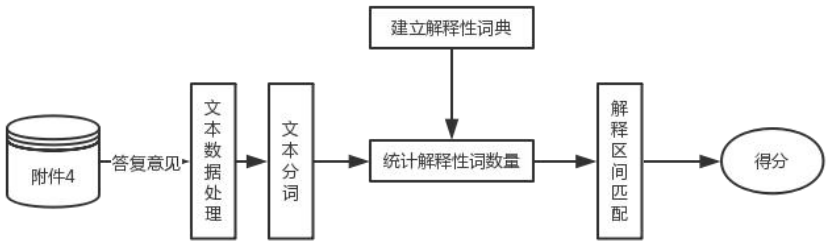

图 11　可解释性量化流程图

2.3.3.3 解释性词典

在群众问政记录中，对于群众留言的答复意见常会引入相关解释或理论支撑来增强答复意见的解释性。在引经据典之前，人们通常会有使用相关的连词来将相关解释引入。因此，文本提出一种基于构建解释性词典来量化文本可解释性的方案。通过统计文本中可解释性数量来体现文本的可解释性。

常用的解释性词语:根据、据此、依照等。

2.3.3.4 解释区间

根据文本中含有的解释性词语的数量进行划分，将可解释性分为三个等级（+∞,4]、(4,2]、(2,1]、(1,0]，分别对应 3 分、2 分、1 分、0 分。

2.3.4. 评价指标构建

整合相关性参数得分、完整性参数得分与可解释性参数得分，将所有得分相加后即为答复意见的质量得分。

得分公式： $f = f_{相} + f_{完} + f_{解}$

3. 结论

本文对群众问政留言记录进行自然语言处理，构建了基于线性支持向量机的留言分类模型，模型的评价指数约为0.9079，具有较高的应用价值。

本文利用文本相似度识别相同问题留言，使用命名实体识别技术与自动摘要技术来自动提取问题关键信息，定义了基于留言数量与留言评价的热度评价指标来筛选热点问题，达到了减少相关部门处理网络留言人力支出、提高工作效率的目的。

本文针对相关部门对留言的答复意见，从答复的相关性、完整性与可解释性三个角度出发，通过文本相似度、基于规范的多标签分类模型与解释性词典等方法对答复的三个特性进行量化，构建了答复意见质量评价方案。

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，因此利用自然语言处理技术与文本挖掘技术处理群众问政留言记录是社会治理创新发展的新趋势，对未来政府的管理水平与效率具有极大的推动作用。

4. 参考文献

[1]刘志刚,李德仁,秦前清,史文中.支持向量机在多类分类问题中的推广[J]. 计算机工程与应用,2004(07):10-13+65.

[2]毛雪岷,丁友明. 基于语义引导与支持向量机的中文文本分类[J]. 情报志,2007(11):56-58.

[3]齐乐,张宇,刘挺. 基于关键信息的问题相似度计算[J]. 计算机研究与发展, 2018,55(7): 1539-1547.

[4]郑飘飘,万健,司华友.基于评论的热点新闻事件识别方法研究[J].浙江科技学院学报,2019,31(05):392-399.

[5]熊大平,王健,林鸿飞. 一种基于 LDA 的社区问答问句相似度计算方法[J]. 中文信息学报, 2012,26(5):40-46.

[6]连冬阳. 基于深度学习的新闻评论热度预测研究[D]. 哈尔滨工业大学, 2018.

[7]周怡.新闻报道当中的解释性语言[J].新闻界,2006(02):117-118.