“智慧政务”中的文本挖掘应用

摘要

随着互联网技术的快速发展，网络问政平台成为了民众提出建议和民众诉求的地方。也因为如此，各类文本数据量大幅度攀升，给以往依靠人工来进行信息的处理带来了极大挑战，本文就此建立基于自然语言处理技术的智慧政务系统，分别解决如下三个方面的问题。

针对问题一的群众留言分类问题，我们利用了“一对全”（OVA）构造的多分类支持向量机 SVM 模型，并结合了随机梯度下降法（SGD）进行模型训练。首先进行分词等相关的常规预处理后，采取 TF-IDF 算法进行词的向量化。将数据集按 7：3 划分为训练集与测试集。经测试调优后，在测试集上的准确率达到了 91.26%，F1-Score 达到了 91.23%，取得了较好的分类效果。

针对问题二的是热点问题挖掘，我们首先对相似的热点问题进行聚类，采取的是对留言主题内容进行重复二分聚类算法（基于 HanLP），是 K 均值聚类算法的效率加强版，句子向量化采取的是 word2vec 预处理。对于自动判断聚类个数 K，我们通过调试，确定了基于余弦距离的准则函数的增幅阈值为 1.2 能够达到较好的聚类效果。同时我们采取了基于余弦相似度的方法来对聚类后的结果进行适当地修正，将各类中相似度高于 86%的认为是聚类效果较好的类别。

对于各类热点问题的热度指标的衡量，我们划分了二级指标体系，二级指标中分别从留言字数，留言问题存在的时长，留言内容的消极概率，点赞数，反对数，该类问题的留言数量的六个指标维度进行热度指标的定义。对于各类指标的权重的分配，为了避免人为分配的主观性，同时考虑到热点问题更应以群众的角度来去评判。所以我们采取问卷调查的方式，调查对象为网民和 NLP 领域的学者，根据投票者对与热点问题的各类指标的投票数进行合理的权重分配。

在留言文本消极概率的指标中，我们采用的是 Bi-LSTM 模型进行情感的分析以得出倾向于消极的概率。最后对经过热度指数排名前五的热点问题，进行特定地点或特定人群的提取，我们利用了 Hanlp 的分词和词性标注工具，结合正则表达式和一些逻辑判断，同时针对所给数据信息，我们构建了一个小型词表，新增了一些地点词性的词，如“小区”、“路口”等词，以便进行特定地点的提取。而对于问题描述即留言问题的概括，我们首先对某类热点问题的其中所有留言主题进行合并，接着采取 TextRank 算法进行关键句的摘取作为问题描述内容，并从结果可看出，问题描述概括度高，因此验证了该方法的合理性。

针对问题三的是答复意见的评价，我们从三个角度：答复的相关性、完整性、可解释性出发，其中可解释性包含了答复的及时性，答复态度积极性和答复意见的字数（具体性）三个方面。以此对答复意见的质量指定了一套评价方案。对于相关性我们同样采用的是对留言详情和答复意见进行 word2vec，然后计算句子词嵌入的平均值来计算相似度，用于反映相关程度。完整性中我们结合正则表达式对答复的规范书信格式进行了评判与评价。在可解释中，我们分了三个方面进行量化：答复的及时性，我们根据答复与留言时间的差值进行了相应分值的量化与分配。在答复态度积极性中，同样采用了 Bi-LSTM 模型对答复意见进行了情感分析得出留言意见倾向积极性的概率从而进行分值的量化。根据答复的字数（具体性），我们将其合适的字数范围进行划分并将其量化为具体数值得分。综合以上三个角度来对答复意见进行评价，可以较合理、较全面地阐述与评价留言答复的质量。

关键词： 多分类 SVM 模型、TextRank、word2Vec、重复二分聚类、Hanlp、Bi-LSTM

Abstract

With the rapid development of the Internet technology, the online questioning platform has become a place for people to make suggestions and appeals. Therefore, the amount of various types of text data has risen sharply, which has brought great challenges to relying on manual processing of information in the past. In this paper, a "Smart Government System" based on NLP technology is established to solve the following three problems.

For the problem of mass message classification in problem 1, we used the multi-class SVM model constructed by OVA and combined with SGD method for model training. After the conventional preprocessing such as word segmentation, we use the TF-IDF algorithm to vectorize the words. Divide the data set into training set and test set according to 7: 3. After testing and tuning, the accuracy rate on the test set reached 91.26%, and the F1-Score reached 91.23%, which achieved a good classification effect.

Aiming at the problem 2 is the hotspot problem mining, we first cluster similar hotspot problems, and take the repeated bipartite clustering algorithm (based on HanLP) on the subject content of the message. Vectorization uses word2vec preprocessing. For the automatic judgment of the number K of clusters, we determined through debugging that the increase threshold of the criterion function based on the cosine distance is 1.2, which can achieve a better clustering effect. We adopted a method based on cosine similarity to appropriately modify the results after clustering, and considered the categories with a similarity higher than 86% to be the categories with better clustering effect.

For the measurement of the heat index of various hot issues, we divided a secondary index system: the number of words in the message, the length of the message problem, the negative probability of the message content, the number of likes, the number of oppositions, the number of messages. In order to avoid the subjectivity of artificial distribution, the distribution of the weights of various indicators should be judged from the perspective of the masses. Therefore, we take the form of a questionnaire survey. The subjects of the survey are netizens and scholars in the field of NLP. We assign weight according to the number of votes for various indicators of hot issues.

In the indicator of the negative probability of the message text, we use the Bi-LSTM model for sentiment analysis. Finally, for the top five hot issues, we extract specific locations or specific groups of people by using Hanlp's word segmentation and part-of-speech tagging tools, combined with regular expressions and some logical judgments. In order to extract specific locations, we build a small vocabulary to add some part-of-speech words such as "community", "junction", etc. For the problem description, that is, the summary of the message problem, we first merge all the message topics of a certain type of hot issue, and then use the TextRank algorithm to extract the key sentences as the problem description content, and from the results, we can see that the problem description is summarized The degree is high, so the rationality of the method is verified.

Problem 3 is the evaluation of reply opinions. We start from three perspectives: the relevance, completeness and interpretability of the reply. The interpretability includes the timeliness of the reply, the enthusiasm of the reply attitude and the number of words of the reply. Based on this, a set of evaluation schemes is specified for the quality of the responses. For relevance, we use word2vec for message details and reply, and calculate the average value of sentence word embedding to calculate similarity. In the completeness, we combined regular expressions to judge and evaluate the standard correspondence format of the reply. In the explanation, we divided three aspects to quantify the timeliness of the reply, we quantified and allocated the corresponding score according to the difference between the reply and the message time. We use Bi-LSTM model to analyze the sentiment of the reply to obtain the probability of the message opinion tendency to quantify the reply. And we select the appropriate number of words and divide range to quantify as a specific numerical score. Based on the above three perspectives, the quality of the message responses can be explained and evaluated more reasonably and comprehensively.

Keywords: Multi-class SVM, TextRank, word2vec, repeated dichotomy clustering, Hanlp, Bi-LSTM

目录

一、简介

1.1 挖掘背景与意义

近年来，随着互联网、微博、阳光热线等网络问政平台的兴起，越来越多的民众开始通过网络来表达各种意见和诉求，这也成为了政府了解民意的重要渠道。一方面，公民权利意识在更为开放的网络信息环境中逐渐觉醒，公众参与社会热点问题的治理和需求增加；另一方面，政府越来越重视民意在网络的表达，并通过网络问政平台进行回应。因此网络问政的时代已经到来。

各类社情民意相关的文本数据量不断攀升，对过去主要依据人工来进行留言划分和热点问题整理的相关部门的工作带来了极大挑战。与此同时，人工智能、大数据等技术的迅猛发展，机器处理大数据代替人工处理必然是一种趋势，建立基于自然语言技术的智慧政务系统也是时代所需。

对于热点问题的发现与对热点问题热度评价指标的合理构建显得十分重要，及时发现热点问题，将有助于迅速掌握热点问题的发展变化趋势，更好地了解和引导网民心态，消解其不良影响。同时规范政府工作人员的答复意见，确保高质量的答复也是网络问政环节十分关键的一步，政府部门如何对待、如何解决、答复情况如何，都应有明确的问责标准。这样，才有利于行成长效的机制，推动网络问政的持久化、常态化。

1.2 挖掘目标

我们要构建一个高准确度的留言文本分类模型，根据附件 1 提供的三级标签体系，对附件 2 给出的数据，建立关于留言内容的一级标签分类模型。该模型能根据留言文本内容预测出该内容的一级分类标签，从而解决了依靠人工经验处理，工作量大、效率低且差错率高等问题。

同时，我们应该构建一个热点问题聚类模型，并为热点问题定义一个较为合理的热度评价指标，根据热度评价指标能够及时发现热点问题以解决民众所需。根据热度评价指标对热点问题进行排序，取出排名前 5 的热点问题，并利用命名体识别技术提取出热点问题发生的特地地点或特定人群，同时提取出简要概括该事件的问题描述，从而导出完成“热点问题表”的和 “热点问题留言明细表”。

对于附件 4 里的留言答复意见，我们要综合考虑各方面的因素，制定出一个合理且全面的答复意见评价方案，用于规范、评价政府工作人员的答复内容，从而确保网络问政平台的健康发展。

二、 问题分析

2.1 问题一的分析

在问题一中，参考附件 1 提供的内容三级标签体系和附件 2 已标注好的一级标签内容，为了将文本进行量化，我们应先对留言详情进行了分词、去停用词等预处理操作，接着按 7：3 划分数据集为训练集和测试集，然后采取 TF-IDF 算法[1]将句子预处理结果进行向量化，并应设置训练集和测试集应具有相同的维度。接下来进行模型的训练和调优，并用 F1-Score 来对模型进行评估最后经过模型的对比与筛选，我们采取了多分类 SVM[2]结合随机梯度下降法进

行一级标签的分类模型的构建。

2.2 问题二的分析

针对问题二，这是一个热点问题挖掘任务，那么首先是应该对附件 3 中的各类相似热点问题进行聚类；考虑到传统的 K 均值聚类算法效率较低，对于大数据聚类时间过长，且需提前确定聚类个数 K。所以我们采取了重复二分聚类[3]，是 K 均值聚类的效率加强版。为了能够自动确定聚类个数 K，我们通过设置余弦准则函数的阈值来进行判断。由于聚类效果少许有些偏差，我们采取了余弦相似度对聚类结果进行修正。接着应构建一个热度评价指标来衡量热点问题的紧要程度，我们定义了二级指标体系，第二级指标由六个维度构成：留言字数，该类问题存在的时长，留言内容的消极概率，点赞数，反对数，该类问题的留言数量的六个指标维度进行热度指标的定义，对于文本消极概率我们采取 Bi-LSTM[4]来进行情感分析。由于各指标量纲不同，所以需要对各指标进行无量纲化，然后将各指标进行线性加权得出热度评价指标得分。

对于权重的分配，考虑到目前较为主流的主成分分析法[5]，其主成分的解释一般带有一点模糊性，不像原始变量的含义那么清楚、确切，且需保证提取的前几个主成分的累计贡献率达到一个较高的水平。同时分析了层次分析法[6]，存在分配权重的主观性较高，且指标多时工作量大，权重难以确定，特征值和特征向量的精确求法比较复杂。因此，我们以群众的角度来衡量权重的分配，这样更能反映出民众关心的问题以何种形式成为了热点问题。我们通过问卷的方式来调查热度评价指标与我们定义的六个指标的相关程度，统计对象为网民和 NLP 领域学者，根据各指标投票数来进行合理的权重分配。

最后为了完成“热点问题表”和“热点问题留言明细表”们还需利用命名体识别技术，提取出相应热点问题的特定地点和特定人群，考虑到附件所给数据中都是类似“A 市 A5 小区”这样的用字母和数字代替具体的地方名字的数据，我们利用了 Hanlp[7]的分词与词性标注工具，并结合正则表达式和一些逻辑判断，并针对性地构建了一个小型词表，增加了一些地点词性的词如：“小区”，“路口”等以便对特定地点进行提取。对于问题描述，我们利用各类热点问题对应的所有留言主题进行合并，然后采取 TextRank 算法[8]提取出关键的一句作为问题描述。

2.3 问题三的分析

对于问题三，我们要构建一套对答复意见评价较为合理且全面的评价方案。我们分别从答复的相关性、完整性、可解释性三个维度出发，对答复意见进行一个综合的评价。留言内容与答复意见的相关性我们同样采取了对句子进行 word2vec[9]后，计算两句子词嵌入平均值，通过余弦相似性来衡量。对于完整性即答复的意见的规范格式，我们结合正则表达式来对格式的规范与否进行判断。可解释性我们分为三个方面进行量化：答复及时性：对留言时间与答复时间之差的长短进行划分，并定义了合理的得分来衡量及时性。答复态度积极性的衡量我们采用 Bi-LSTM[4]来对答复意见内容进行情感分析。答复的具体性中，我们根据答复的字数范围量化为相应得分。

三、模型假设与符号说明

3.1 模型假设

假设 1 假设各附件中留言内容中均为网民真实留言，不存在机器留言、虚假留言。

假设 2 假设各附件中留言主题与留言详情相匹配和相对应。

假设 3 假设附件 2 中给出的留言内容对应的一级分类标签准确无误，以便正确进行模型的训练。

3.2 符号说明

符号	定义
$\boldsymbol{F_{11}}$	留言字数指标
$\boldsymbol{F_{12}}$	留言问题存在时长指标
$\boldsymbol{F_{13}}$	留言内容消极概率指标
$\boldsymbol{F_{21}}$	点赞数指标
$\boldsymbol{F_{22}}$	反对数指标
$\boldsymbol{F_{23}}$	留言数量指标
$\boldsymbol{F'_{ij}}$	F_{ij} 指标的无量纲化指标
$\boldsymbol{s_1}$	答复意见相关性量化分值
$\boldsymbol{s_2}$	答复意见完整性量化分值
$\boldsymbol{s_3}$	答复意见可解释性量化分值
$\boldsymbol{s_{31}}$	答复意见及时性量化分值
$\boldsymbol{s_{32}}$	答复意见字数（具体性）量化分值
$\boldsymbol{s_{33}}$	答复意见态度积极性量化分值

表 1：符号说明表

四、数据预处理

4.1 数据的读写与处理

对于数据的读写操作与处理，我们使用的是 pandas 模块[10]，它是一个功能强大的分析结构化数据的工具集，常用于数据挖掘和数据分析，同时有数据清洗功能。其中的 Series，DataFrame 等数据结构很方便我们进行数据的统计与处理。

4.2 对文本数据进行分词操作

考虑到中文分词的特点是词与词之间没有明显的界限，所有本文采用了基于 Python 开发的一个简单实用的中文自然语言处理分词库——jieba 分词[11]，其属于概率语言模型分词。其原理是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。通过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。对于未登录的词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

利用 jieba 分词库，我们可以很好进行对文本数据的切分。

4.3 去停用词

在文本数据的处理中，停用词是指没有实际含义且功能极其普遍的词，比如中文中的“我、你、的、了、吗”等。对于类似留言这种短文本数据，这些词则会对其造成一定程度上的负面影响。所以在文本预处理阶段就需对停用词进行去除。

4.4 word2vec

首先，为了将文本数据表示成计算机能够理解并处理的数字形式，我们采取 Word2Vec 来对文本进行数值化。

word2vec 是谷歌一个由 Mikolov 领导的团队发明的一套进行词嵌入的工具[9]，该模型能够有效且快速地训练向量空间模型。作者的目标是从海量数据中学习高质量的词向量，可以很好地度量词与词之间的相似性。word2vec 包含了计算词向量的 CBOW 模型和 Skip-gram 模型，如图 1。其中可看出，CBOW 模型是利用上下文对当前词进行预测，而 Skip-gram 则是利用当前词预测其上下文。

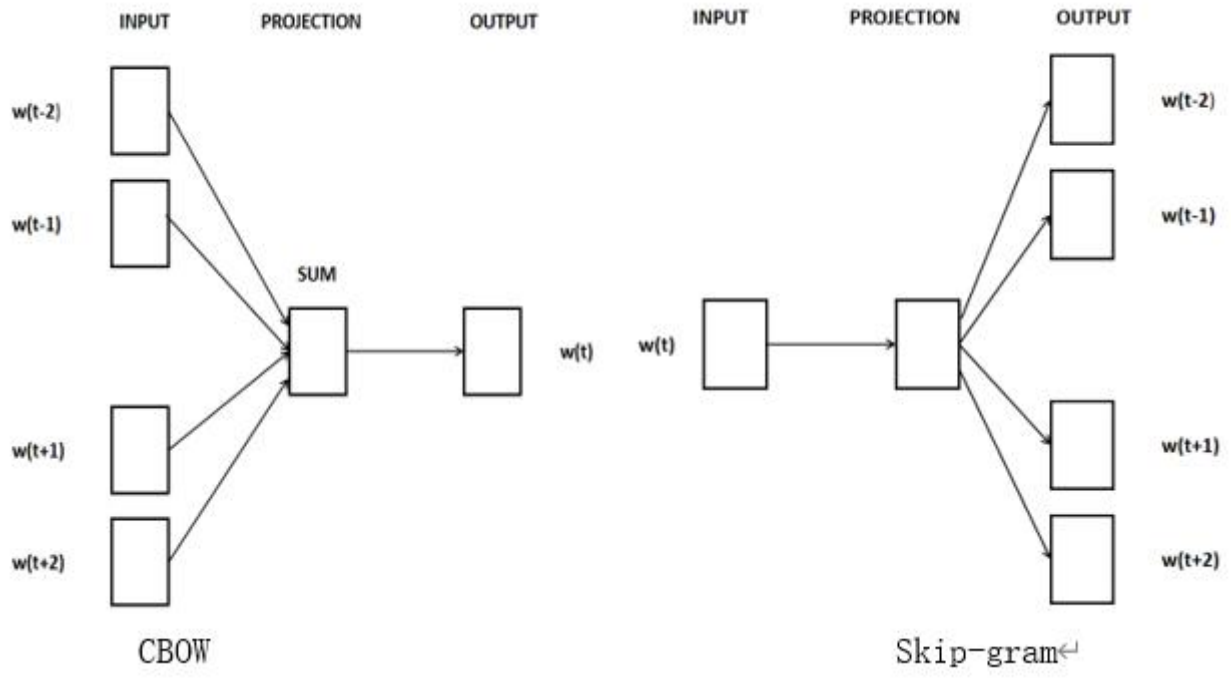

图 1: 两种 word2vec 算法模型示意图

本文采用了 text2vec 开源库[12]，其是一个文本向量化表示工具，包括词向量化、句子向量化。该库是基于 Skip-gram 模型进行了训练。Skip-gram 模型的训练目标就是使下式取得最大值：

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c}\log p\left(w_{t+j}|w_t\right) \qquad (1)$$

其中，c 为窗口大小，T 是训练文本的大小。基本的 Skip-gram 模型中条件概率的计算如下式：

$$p(w_O|w_I)=\frac{exp(v'^{T}_{wO}v_{wI})}{\sum_{w=1}^{W}exp(v'^{T}_{w}v_{wI})} \qquad (2)$$

其中 v_w 和 v'_w 是单词 w 的输入和输出向量，W 是词典的大小。

对于文本向量表示，字词粒度，通过腾讯 AI Lab：开源的大规模高质量中文词向量数据，获取字词的 word2vec 向量表示。句子粒度，通过求句子中所有词的词嵌入平均值计算得到。

这里，在进行 word2vec 词嵌入的时候，里面用的分词并没用到停用词表，因为停用词也是有一定意义的，去除停用词可以说是一般做特征提取时比较粗糙简单的做法，在样本数量少时会起到一定作用。且词向量这种浅层语义表示，在训练样本足够的情况下，停用词的词义是可以区分出来的。

通关导入 text2vec 库，执行语句 text2vec.encode()我们可以很方便地获取字词的 word2vec 向量表示。

五、任务一：群众留言分类

5.1 数据集的构建

5.1.1 数据集的分类数目分布

由附件 1 提供的内容分类三级标签体系可知，一级标签共有 15 类。我们的任务是根据附件 2 建立关于留言内容的一级标签分类模型，而附件 2 中的一级标签数据只有 7 类，所以我们需要解决的是一个 7 分类问题。

现在我们来统计下各类别的数据量，如图 2：

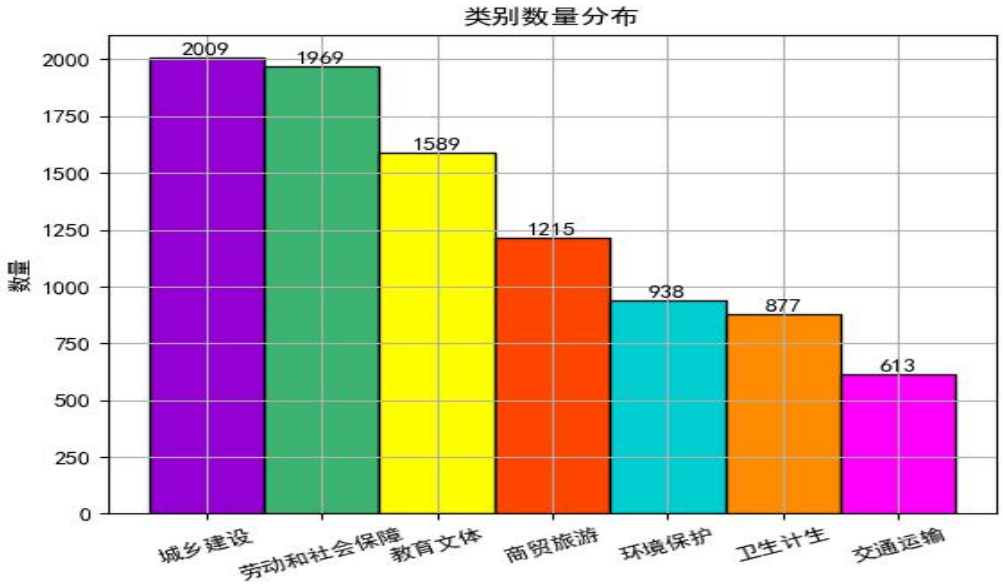

图 2: 附件 2 中的一级标签类别数目分布图

通过图 X，我们可以知道，属于一级标签类别的城乡建设、劳动和社会保障数目都为 2000 条左右，而交通运输的数据量只有 613 条，各类别数目分布不太均匀。

下一步，我们把留言详情作为文本数据，把一级分类标签的文本标注上对应序号（0-6），这样便于后面的分类模型的训练，把其作为标签来进行数据集的构建。

5.1.2 TF-IDF 词频向量化

首先，我们对文本数据进行常规的文本预处理操作，如前文所述即分词与去停用词，整体处理思路图，如图 3：

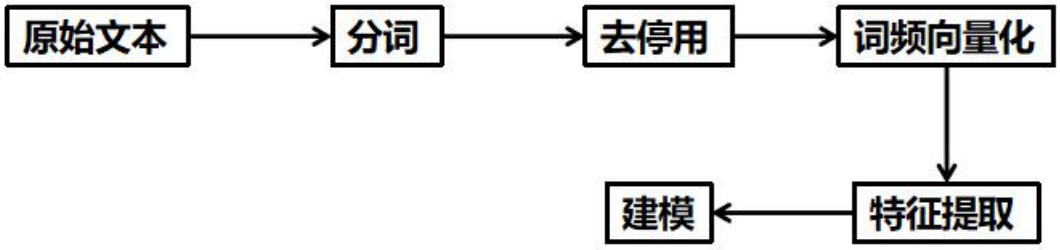

图 3: 整体流程思路图

这里我们采用 Scikit-learn(sklearn) 模块[13]来进行数据集的处理与模型的构建，其是机器学习中常用的第三方模块，对常用的机器学习方法进行了封装，可以十分高效的用于模型的建立。

那么，为了能够让计算机理解文本数据，这就需要对进行预处理后的文本进行向量化。我们采取词带模型进行词频的统计，将各个文本样本的这些词与对应的词频放在一起即进行了向量化。这里我们采取了 sklearn 中的 CountVectorizer()来进行词频向量化。

下一步就是采用 TF-IDF 算法[1]进行特征的权重的修正，再将特征进行标准化。TF-IDF

即“词频-逆文本频率”。那么 TF 即上面所说的词频，我们做的向量化也就是做了文本中各个词的出现频率统计，并作为了文本特征：

$$TF(x) = N/M \tag{3}$$

其中 N 是单词在某文档中的频次，M 为该文档的单词数，x 为某词。

而 IDF 反应了一个词在所有文本中出现的频率，如果一个词在很多的文本中出现，那么它的 IDF 值应该低，IDF 就是来帮助我们来反应这个词的重要性的，进而修正仅仅用词频表示的词特征值，其计算公式为：

$$IDF(x) = log\frac{D+1}{D_w+1}+1 \tag{4}$$

其中 D 是总文档数， D_w 是出现了该词的文档数， x 为某词

因此，TF-IDF 的计算公式为：

$$TF-IDF(x) = TF(x) * IDF(x) \tag{5}$$

通过 sklearn 中的 TfidfTransformer()类，我们方便地进行了 TF-IDF 预处理。

进行完文本数据集的数值化后，我们将数据集通过 7：3 的比例划分为了训练集和测试集，由于训练集和测试集在进行数值化时可能存在维度不一致的问题，所以在测试集进行 CountVectorizer()时通过设置参数 vocabulary 的值，从而确保训练集和测试集的向量共享维度。

最后，数据集构建完毕，接着就是对模型的选取、训练和评估。

5.2 常用分类模型的比较与选择

5.2.1 朴素贝叶斯模型

对于文本分类问题，在机器学习中，我们较常使用的是朴素贝叶斯和支持向量机 SVM 模型。

朴素贝叶斯法[14]是基于贝叶斯定理与特征条件独立假设的分类方法。对于给定的训练数据集，首先基于特征条件独立假设学习输入输出的联合概率分布；然后给定输入 x，利用贝叶斯定理求出后验概率最大的输出 y。具体地，朴素贝叶斯分类器表示为：

$$y = arg\ \max_{c_k} P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)}|Y = c_k) \tag{6}$$

其中c_k 是某个类的标签，$x^{(j)}$ 是数据的第 j 个特征，y 为输出

对于上式的先验概率和条件概率，我们会使用贝叶斯估计去计算，计算公式如下：

$$P_\lambda(X^{(j)} = a_{jl}|Y = c_k) = \frac{\sum_{i=1}^{N} I\left(x_i^{(j)} = a_{jl}, y_i = c_k\right) + \lambda}{\sum_{i=1}^{N} I(y_i = c_k) + S_j\lambda} \tag{7}$$

$$P_\lambda(Y = c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k) + \lambda}{N + K\lambda} \tag{8}$$

其中 a_{jl} 是第 j 个特征可能取的第 l 个值，I 为指示函数，K 是类的个数，S_j是第 j 维特征的最大取值。

由以上可知，朴素贝叶斯算法简单且高效。对于普通的二分类问题，如垃圾短信分类，经过训练调优，准确率很容易提升到 90%以上。对于像我们这里的多分类问题，经过模型的训练，准确率和 F1-Score 大概在 70%到 80%之间，这并不是一个很好的分类效果。所以，朴素贝叶斯算法的缺点就是对于多分类问题，分类的性能不一定很高。因此，我们采用决定采用比较适合多分类问题的多分类 SVM 模型。

5.2.2 基于随机梯度下降（SGD）训练的多分类 SVM 模型

SVM，即支持向量机，它的基本模型是定义在特征空间上的间隔最大的线性分类器，SVM 还包括核技巧，这使它成为了实质上的非线性分类器。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。下面大致叙述下 SVM 算法原理：

SVM 学习的基本想法是求解能够正确划分训练数据集并且集合间隔最大的分离超平面，如下图 4 所示，$\omega \cdot x + b = 0$ 即为分离超平面，对于线性可分的数据集来说，这样的超平面有无穷多个，但是几何间隔最大的分离超平面确实唯一的

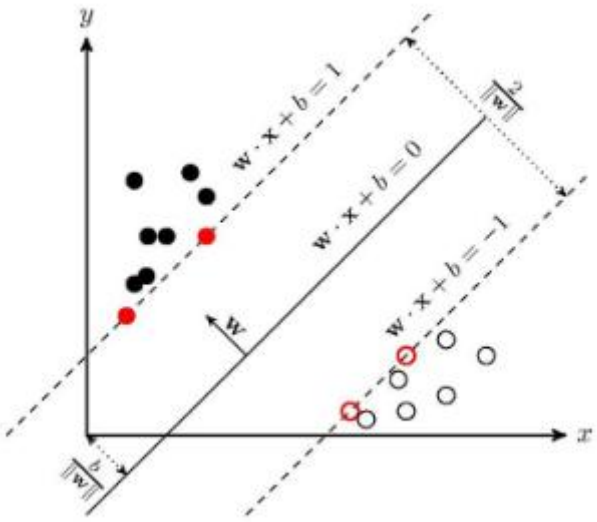

图 4 ： 分离超平面图[20]

SVM 学习算法大致如下：

输入：训练数据集 $T = \{(x_1, y_1), (x_2, y_2),\ldots,(x_N, y_N)\}$

输出：分类超平面和分类决策函数

(1)选择惩罚项 C > 0, 构造并求解凸二次规划问题：

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i \qquad (9)$$

$$s.t.\quad \sum_{i=1}^{N}\alpha_i y_i=0\quad,\quad 0\le\alpha_i\le C,\ i=1,2,\ldots N$$

得到最优解 $\alpha^*=(\alpha_1^*,\alpha_2^*,\ldots\alpha_N^*)^T$

(2)计算:

$$w^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i \qquad (10)$$

选择 α^* 的一个分量 α_j^* 满足 $0\le\alpha_j^*\le C$，计算：

$$b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_j(x_i\cdot x_j) \qquad (11)$$

（3）求分离超平面

$$w^*\cdot x+b^*=0 \qquad (12)$$

分类决策函数：

$$f(x)=sign(w^*\cdot x+b^*) \qquad (13)$$

SVM 最初只用于解决二分类问题，缺乏处理多分类问题的能力。后来随着需求的变化，需要 SVM 处理多分类。

目前构造多分类 SVM 分类器的主要方法有两类：一类是“同时考虑所有分类”方法，另一类是组合多个二分类解决多分类问题。

本文中，我们采取的是第二类构造方法中的“一对全”（OVA）方法，通过组合多个二分类支持多分类[2]，对于每个类别，则学习一个二分类器，该分类器区分所有这些分类。

我们采用了随机梯度下降法（SGD）用于线性分类器在凸损失损失函数下的判别学习。SGD 已经成功地应用于文本分类和自然语言处理处理中经常遇到的大规模稀疏机器学习问题。SGD 的优点是：高效率，易于实施。

我们利用了 sklearn 模块中的 SGDClassifier(),通过设置参数 loss = ‘hinge’即为上述叙述的基于 SGD 训练的多分类 SVM，它是用 mini-batch 来做梯度下降，在处理大数据的情况下能够很快地收敛。

5.3 多分类 SVM 模型的评估与分类结果

我们将上述构建的数据集，按比例 7：3 划分为训练集和测试集后。基于上述通过 sklearn 模块中的 SGDClassifier()构建的多分类模型，需要设置一些超参数。经过模型的调优与测试，

我们设置惩罚方式 penalty=’l2’进行 L2 正则化惩罚，惩罚参数 alpha = 8e-5，迭代次数 max_iter=50，学习率 learning_rate=’optimal’，默认根据 alpha 计算得到。

对于该模型分类方法的评价，我们使用的是 F1-Score:

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \qquad (14)$$

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

通过模型的训练与调优，我们通过混淆矩阵可视化，预测标签和实际标签之间的差异，如下图 5：

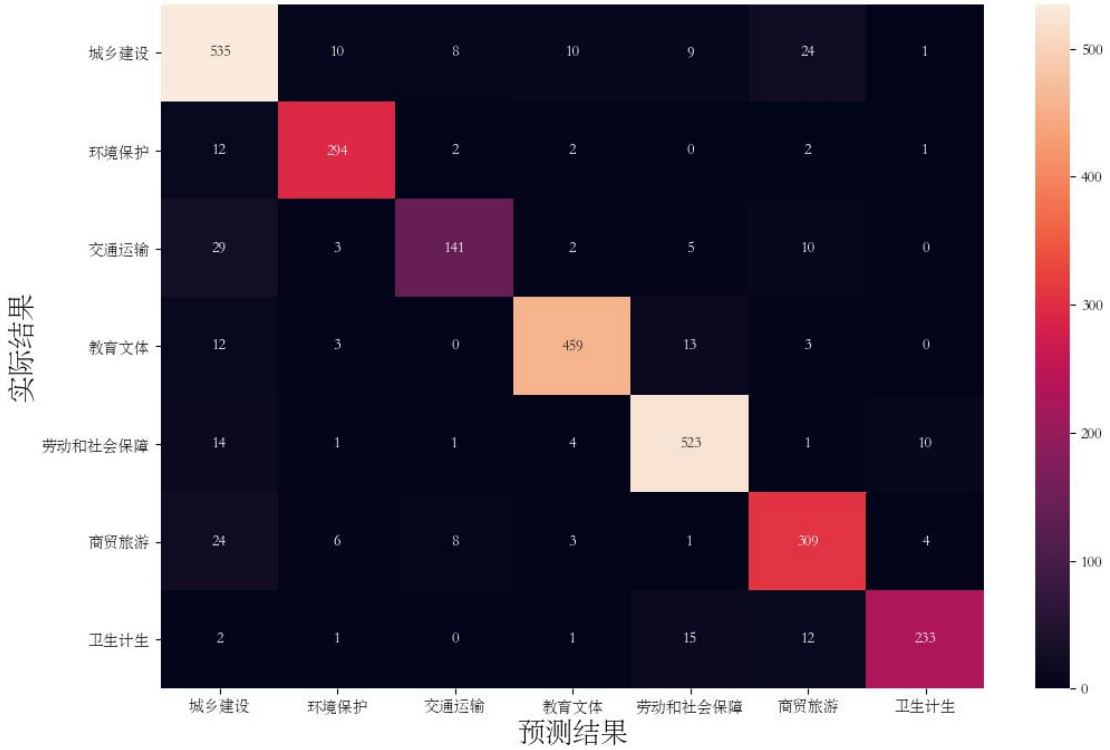

图 5 ： 可视化混淆矩阵图（可清晰看出预测和实际标签之间的差异）

在混淆矩阵中，主对角线表示测试集中预测正确的数量，主对角线以外即是预测错误的数量，主对角线以外，颜色普遍趋于黑色较深，即预测错误数量较少，分类效果较好。

经过多次模型的试验，在测试集上，模型的准确率 accuracy 和 F1-Score 大致稳定如下数值：

Accuracy	91.26%
F1-Score	91.23%

表 2：准确率与 F1-Score

F1-Score 达到了 91.23%，这是一个十分不错的分类效果。据此，我们可知基于随机梯度

下降（SGD）训练的多分类 SVM 模型，适合于附件 2 所示的留言内容的一级标签分类，且采取这种组合多个二分类 SVM 的“一对全”（OVA）方法，进行多分类 SVM 的构造，训练分类器个数较少，分类速度较快，效率高。

六、任务二：热点问题挖掘

6.1 热点问题聚类

6.1.1 重复二分聚类算法

问题二的任务是根据附件三将某一时间段内反映的特定地点或特点人群问题进行聚类，然后定义一个合理的热度评价指标去衡量某类问题的热度。

问题的开始，我们采用了传统的 K 均值聚类算法[21]进行试验，其是一种简单实用的聚类算法。具体地，K 均值算法需解决的问题是：对给定的 n 个向量 d_1 至 d_n，以及一个整数 k，要求找出 k 个簇 S_1 到 S_k，以及各自的质心 c_1 到 c_k，使下式最小：

$$\min \mathcal{L}_{Euclidean} = \sum_{r=1}^{k} \sum_{d_i \in S_r} ||d_i - c_r||^2 \qquad (15)$$

其中的$||d_i - c_r||$ 是向量与质心的欧拉距离，此函数称做聚类的准则函数。K 均值就是以最小化各向量到质心的欧拉距离的平方和为准则进行聚类。而质心即簇内数据点的几何平均，计算公式如下：

$$s_i = \sum_{d_i \in S_i} d_j \qquad (16)$$

$$c_i = \frac{s_i}{|s_i|} \qquad (17)$$

其中，s_i 是簇 S_i 内所有向量之和，称合成向量。

K 均值聚类是一种迭代式算法，每次迭代根据上一步结果进行优化：

（1） 选取 k 个点作为 k 个簇的初始质心。

（2） 将所有点分配给最近的质心所在簇。

（3） 更新各簇的质心。

（4） 继续迭代至质心不再发生变化。

通过实验我们发现，对于附件三的大约 4000 条数据，聚类效率很低，耗费时间长，且需人为确定类数 K，所以我们采取了重复二分聚类算法。

重复二分聚类算法[3]是 K 均值算法的效率加强版，具体聚类流程如下：

（1） 选定一个簇进行划分

（2） 采用 K 均值将簇划分为 2 个子集

（3） 重复（1）、（2）步骤，知道产生足够舒朗的簇。

类似如图 6 的二叉树结构：

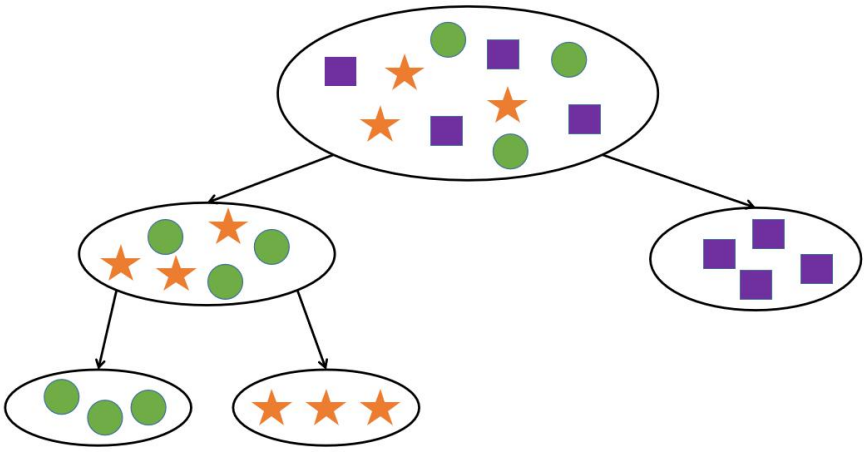

图 6 ： 重复二分聚类算法步骤图

虽然每次划分都是基于 K 均值，由于每次二分都仅仅在一个子集上进行，输入数据少，所以算法效率有所提高。

并且我们改用了更快的准则函数，一种基于余弦距离的准则函数，计算公式如下：

$$\min \mathcal{L}_{cos} = \sum_{r=1}^{k} \sum_{d_i \in S_r} cos(d_i, c_r) \tag{18}$$

具体地，经过相关步骤后，准则函数为：

$$\mathcal{L}_{cos} = \sum_{r=1}^{k} ||s_r|| \tag{19}$$

该式说明了，余弦准则函数等于 K 个簇各自合成向量的长度之和，对于 $\mathcal{L}_{cos}$ ，由于发生改变的只有原簇和新簇两个合成变量，只需求两者的程度即可。可见计算量下降了不少。

6.1.2 自动确定聚类个数

由于 K 均值聚类需要人为设定分类数 K，但实际中 K 这个超参数很难准确估计。在重复二分类聚类算法中，有一种变通的算法：通过给准则函数的增幅设定阈值来自动判断 k。此时算法停止条件变为：当一个簇的二分类增幅小于该阈值时，不再对该簇进行划分。所有簇不可分时，即聚类结束。因此该算法不需要人工进行类数 K 的设定。

对于算法的实现，我们基于 Hanlp[7]开源的自然语言处理模块的一个 python 接口包 pyhanlp[15]来实现，其中的 K 均值聚类和重复二分聚类经过了该库开发人员的标准化评测，评测选取的是搜狗实验室提供的文本分类语料的一个子集，每个类目有 1000 篇文章，共计 5000 篇。评测结果如下表：

算法	F1-Score	耗时
K 均值	83.74	67 秒
重复二分聚类	85.58	24 秒

表 3：K 均值与重复二分聚类评测结果表[3]

对比两种算法，重复二分聚类不仅准确率比 k 均值更高，而且速度是 k 均值的 3 倍。

在 pyhanlp 库的二分重复聚类算法中，采用的是二分准则函数的增幅最大为策略，每产生一个新簇，都试着将其二分并计算准则函数的增幅。然后对增幅最大的簇执行二分，直到算法满足停止条件。主要实现步骤：

（1）通过接口函数 ClusterAnalyzer()创建函数类，我们通过遍历附件 3 的文本

（2）利用该类的 addDocument()方法即可载入需要聚类的文本，在这里由于附件 3 中的留言主题能够较好地反映出该类留言的主要事件，所以我们采用留言主题文本进行聚类。

（3）执行该类的 repeatedBisection()传入阈值参数，即可实现上文叙述的自动判断聚类数量的重复二分类聚类算法。

（4）通过调试与观察聚类的结果，我们最终确定了步骤（3）中的阈值为：1.2。

聚类完成后，我们给每类标上类别记号“flag” ，如图 7 所示(这里我们随机抽取两类)：

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	flag
220616	A0001301	A3区枫林三路向日葵旁边工地一直	2019/10/7 1:25:25	A市A3区枫林三	0	0	47
235521	A0006920	A3区枫林三路涉外经济学院外街理	2019/10/15 18:59:08	A市A3区，枫林	0	0	47
216679	A0004264	A3区共和世家底商新日图文半夜开	2019/1/16 11:19:35	地点A市A3区枫	0	0	47
282978	A0005355	A3区奥园城市天骄楼盘半夜泵车浇	2019/10/21 18:00:07	连日来，A市A3	0	0	47
217998	A0001867	A3区枫林三路凌晨一点拖土车扰民	2019/9/9 15:21:17	涉外和第一师	0	0	47
262339	A0007401	A3区大道西湖建筑集团占道施工扰	2020/1/6 11:51:37	A3区大道西行	0	1	47
278791	A0002481	A市雅礼中学强制高二学生周六补课	2019/3/1 15:01:54	A市雅礼中学违	0	0	144
266368	A0003892	A市涉外经济学院寒假过年期间组织	2019/11/22 14:42:14	关于西地省A市	0	0	144
360110	A110021	A市经济学院寒假过年期间组织学生	2019-11-22 14:42:14	关于西地省A市	0	0	144
360113	A3352352	A市经济学院强制学生外出实习	2018-05-17 08:32:04	A市经济学院强	3	0	144
215733	A0005906	A市一中双语实验学校周末强制有偿	2019/9/16 14:26:24	A市一中双语实	0	0	144
360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	各位领导干部	0	0	144
360114	A0182491	A市经济学院体育学院变相强制实习	2017-06-08 17:31:20	书记您好，	9	0	144
195917	A909119	A市涉外经济学院组织学生外出打工	2019/11/05 10:31:38	一名中职院校	0	1	144
360111	A1204455	A市经济学院组织学生外出打工合理	2019-11-05 10:31:38	一名中职院校	1	0	144
255719	A0006081	A市商贸旅游职业技术学院强制学生	2019/4/28 17:32:51	各位领导干部	0	1	144

图 7 ： 部分聚类结果图

6.1.3 基于 word2vec 计算相似度对聚类结果修正

由于可能有些类别并没有很好的聚类，对于好的聚类，里面的留言问题应该反映的是同一类热点问题，其文本之间应该具有较高的相似度，所以我们在每个类间，进行了两两相似度的计算，最后得出每类中的平均相似度，用来反映该类中文本是否正确归为一类。通过设定一个合适的相似度阈值，摘取聚类效果较好的类别进行下一步的操作。因为较为突出的热点问题，一般留言条数较多，且通过聚类更容易聚在一起。与此同时，某类留言数量少的也比较容易和某类不相关的少数留言数量的问题聚在一起。而任务二中，我们的任务是挖掘出较为突出的热点问题，因此我们这样做，存在一定合理性。

对于文本相似度的计算我们采取了，计算两句子之间的余弦相似性的方法。对于聚类后的各类中的留言主题的文本数据，我们先采用预处理所说的 word2vec[9]预处理进行字词的向量表示，然后计算句子所有词的词嵌入的平均值，从而计算两句子词嵌入之间的余弦相似性，以

估计两句子之间的语义相似度。具体流程步骤如图 8：

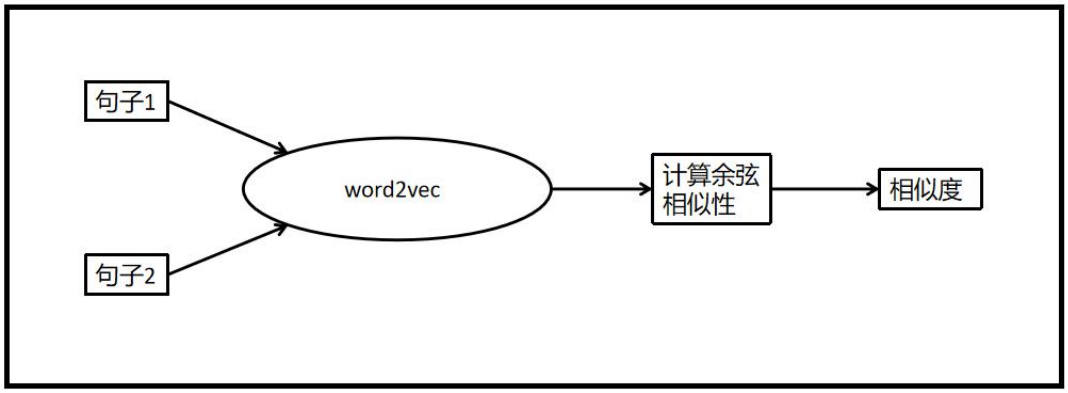

图 8：计算相似度步骤图

这里，相似度的计算公式为：

$$\cos\theta = \frac{X \cdot Y}{|X| \cdot |Y|} \qquad (20)$$

其中，X 和 Y 为两个比较相似度的句子的 word2vec 的平均词嵌入向量。当 $\cos\theta$ 的值越接近于 1，说明两句子的相似度越高。

本文中我们利用了 text2vec 模块[12]中的 Similarity.get_score()方法来计算句子之间的相似度，该方法就是基于上述原理实现的。尽管该方法计算相似度很简洁，但是用该平均词嵌入求余弦相似度的表现非常好，经前人实验，有以下结论：

（1）简单 word2vec 嵌入比 GloVe 嵌入表现的好

（2）在用 word2vec 时，尚不清楚使用停用词表或 TF-IDF 加权是否更有帮助。在 STS 数据集上，有一点儿帮助；在 SICK 上没有帮助。仅计算未加权的所有 word2vec 嵌入平均值表现得很好。

（3）在使用 GloVe 时，停用词列表对于达到好的效果非常重要。利用 TF-IDF 加权没有帮助。

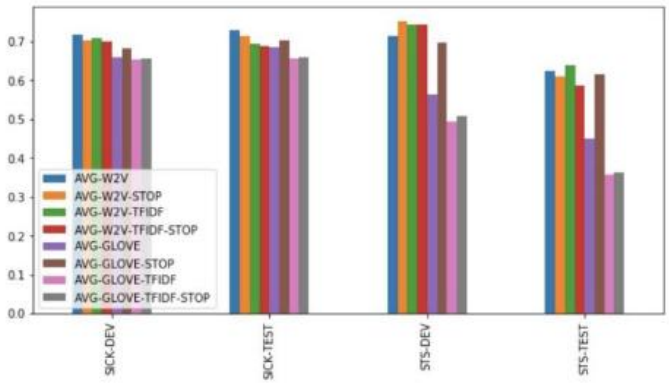

图 9：word2vec 与 Glove 效果比较图[12]

计算完聚类后的各类的留言问题间的平均相似度后，经过调试并观察结果，我们将相似度

阈值设定为 0.87，将高于 0.87 的类筛选出来，这些类更能反映出较为突出的热点问题。经过筛选，较为突出的热点问题类数为 54 类，我们将这 54 类数据存入一个新建的表“附件 3 筛选表”中，以便后续的处理。

图 10 为从“附件 3 筛选表”中抽取的一类数据：

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	flag
212323	A0002070	广铁集团要求员工购房时必须同时购买车	##########	尊敬的领	0	0	0
244528	A909235	伊景园滨河苑开发商强买强卖！	##########	A市广铁集	0	2	0
190337	A0009051	关于伊景园滨河苑捆绑销售车位的维权提	##########	投诉伊景	0	0	0
205982	A909168	坚决反对伊景园滨河苑强制捆绑销售车位	##########	我坚决反	0	2	0
289473	A0001034	反对滨河苑房子和车位捆绑销售	##########	现有伊景	0	0	0
251844	A909167	投诉伊景园滨河苑项目违法捆绑车位销售	##########	投诉广铁	0	1	0
289950	A0004475	投诉A市伊景园滨河院捆绑销售车位	##########	提问A市政	0	0	0
223247	A0004475	投诉A市伊景园滨河苑捆绑销售车位	##########	关于铁广	0	0	0
230554	A909174	投诉A市伊景园滨河苑捆绑车位销售	##########	投诉A市伊	0	0	0
285897	A909191	武广新城伊景园滨河苑违法捆绑销售车位	##########	我们是广	0	0	0
224767	A909176	伊景园滨河苑车位捆绑销售！广铁集团借	##########	伊景园滨	0	0	0
207243	A909175	伊景园滨河苑强行捆绑车位销售给业主	##########	您好！A市	0	0	0
234633	A909194	无视消费者权益的A市伊景园滨河苑车位	##########	伊景园滨	0	0	0
260254	A909173	投诉A市伊景园滨河苑开发商违法捆绑销	##########	投诉A市伊	0	0	0
205277	A909234	伊景园滨河苑捆绑车位销售合法吗？！	##########	广铁集团	0	1	0
258037	A909190	投诉伊景园滨河苑捆绑销售车位问题	##########	尊敬的领	0	0	0
196264	A0009508	投诉A市伊景园滨河苑捆绑车位销售	##########	A市伊景园	0	0	0
276460	A909170	A市伊景园滨河苑捆绑销售车位是否合理	##########	尊敬的领	0	0	0
286304	A909196	无视职工意愿、职工权益的A市伊景园滨	##########	广铁集团	0	0	0
268299	A909193	惊！！A市伊景园滨河苑商品房竟然捆绑	##########	伊景园滨	0	0	0
218709	A0001066	A市伊景园滨河苑捆绑销售车位	##########	伊景园滨	0	1	0
222209	A0001717	A市伊景园滨河苑定向限价商品房项目违	##########	广铁集团	0	0	0
236301	A909197	和谐社会背景下的A市伊景园滨河苑车位	##########	广铁集团	0	0	0
283879	A0004475	A市伊景园滨河苑项目捆绑销售车位	##########	关于铁广	0	0	0
244243	A909198	关于伊景园滨河苑捆绑销售车位的投诉	##########	广铁集团	0	0	0
279070	A0009508	投诉A市伊景园滨河苑开发商违法捆绑销	##########	投诉A市伊	0	0	0
255507	A909195	违反自由买卖的A市伊景园滨河苑车位捆	##########	广铁集团	0	0	0
250514	A0003200	广铁集团强制职工购房时捆绑购买车位	##########	广铁集团	0	0	0

图 10：利用相似度对聚类后修正的部分数据图

采取这样的方法对聚类后的结果进行了适当地筛选，较为突出且留言数量多的热点问题更容易通过这种方式被挖掘出来。

6.2 热度评价指标体系的构建

上一步，对相似热点问题进行了聚类，接着我们应该定义一个合理的热度指标来评价热点问题的热度，以便相关部门及时发现、有针对性地处理热度较高的热点问题。

那么，指标体系设计的思路：考虑到热点问题的作用机理与形成热点问题的原因，我们从留言内容特征影响力与留言传播特征影响力出发，构建了二级指标体系，如下表 4：

一级指标	二级指标	指标内涵
留言内容特征热度影响力 F_1	留言字数 F_{11}	某类热点问题的平均留言字数
	留言问题存在的时长F_{12}	某类热点问题出现与截止时间之差
	留言内容的消极概率 F_{13}	留言内容倾向于消极的概率
留言传播特征热度影响力F_2	点赞数 F_{21}	某类热点问题的点赞数
	反对数 F_{22}	某类热点问题的反对数
	留言数量 F_{23}	某类热点问题的总留言数量

表 4：热度评价指标体系

为了合理且全面地考虑到影响热度评价指标的各因素，我们从六个维度：留言字数、留言问题存在的时长、留言内容的消极概率、点赞数、反对数、留言数量，对热度评价指标进行了综合的构建。

我们把经过聚类后的每类热点问题分别存储在 pandas 模块定义的数据结构”DataFrame”中，以便我们进行各类指标的统计。

对于各类指标的计算公式如下所示：

留言字数指标 F_{11}：

$$F_{11} = \frac{某类热点问题留言总字数}{某类热点问题留言条数} \tag{21}$$

留言问题存在时长指标 F_{12}：

$$F_{12} = 某类热点问题截止时间 - 某类热点问题出现时间 \tag{22}$$

点赞数指标 F_{21}*:*

$$F_{21} = 某类热点问题的总点赞数 \tag{23}$$

反对数指标 F_{22}*:*

$$F_{22} = 某类热点问题的总反对数 \tag{24}$$

留言数量指标 F_{23}*:*

$$F_{23} = 某类热点问题的总留言条数 \tag{25}$$

对于各指标的量化，由于除了留言内容的消极概率指标是依据文本进行量化，其余五个指标对经过聚类后的每类热点问题进行统计计算均可较易获得。下面，我们针对留言内容的消极概率指标F_{13}进行分析。

6.3 基于 Bi-LSTM 模型进行文本情感分析

我们需要对留言内容的消极概率指标 F_{13} 进行计算量化，这是一个文本情感分析类问题。

在介绍 Bi-LSTM 之前，我们先概括一下 RNN 与 LSTM 的原理[4]。

RNN 即循环神经网络，它与传统神经网络的不同之处在于，它能够将权重传递给自己，如图 11（上），将其在时间上展开如图 11（下）：

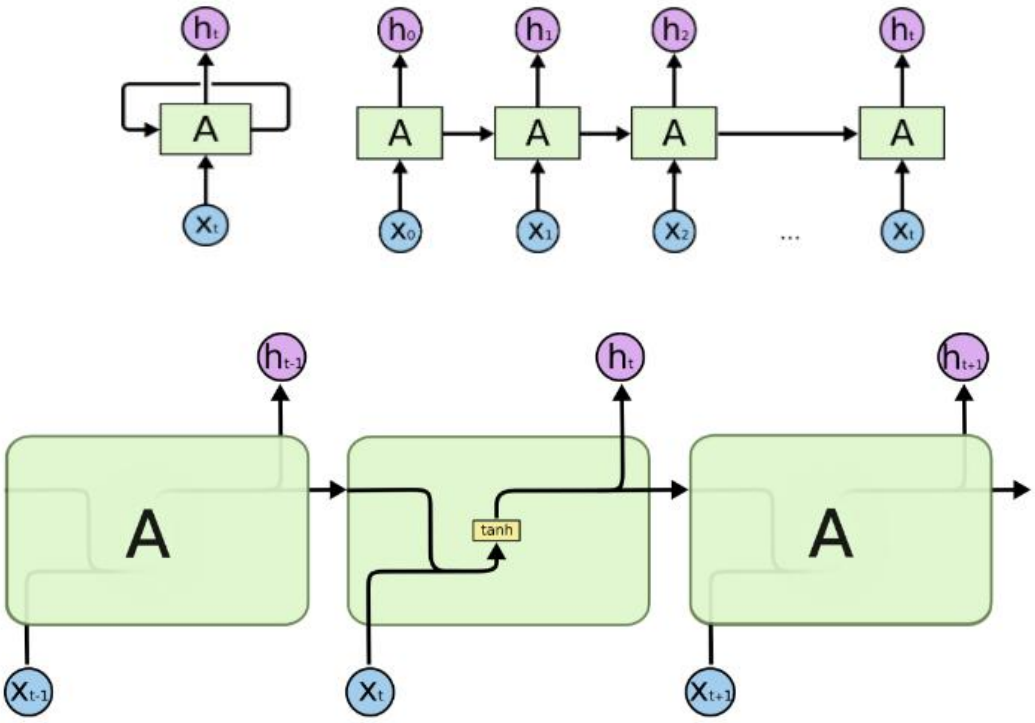

图 11：RNN 结构图[4]

不过 RNN 存在一些较为严重的缺点，如过拟合、梯度消失和梯度爆炸的问题。所以解决这些问题更好的思路就是更改 RNN 的 Cell 结构，LSTM 也就因此而出现。

LSTM 即长短期记忆网络，它是 RNN 的变种，由于其设计的特点，非常适合用于对时序数据的建模，如文本数据。使用 LSTM 模型可以较好地捕捉到长距离字词之间的依赖关系，因为 LSTM 通过训练过程可以学到记忆和遗忘某些信息。

一个 LSTM Cell 内部有三个门，分别为遗忘门、输入和输出门，每一个 Cell 如图 12：

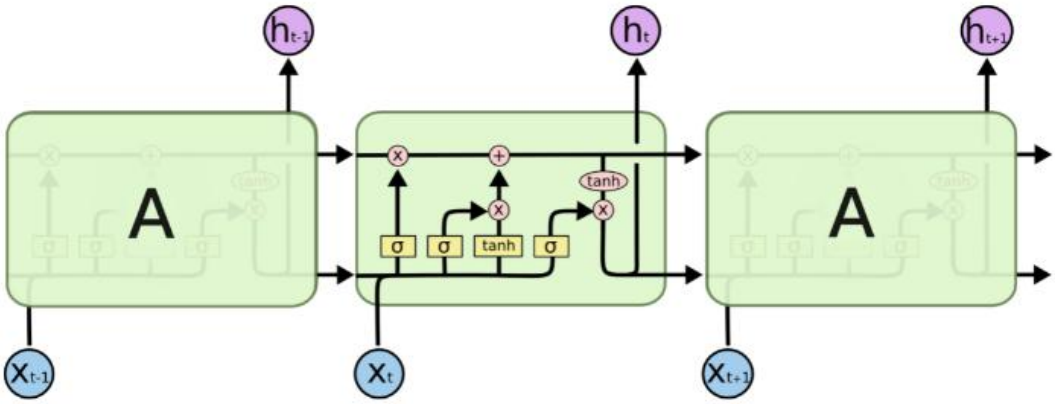

图 12：LSTM 总体框架图[4]

LSTM 中信息的传递则是有设计专门的通道，如图 13。此通道上全部是线性运算，信息持续传递。

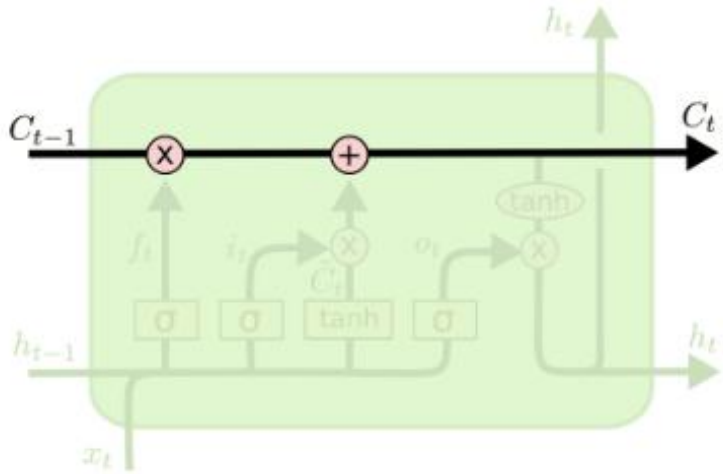

图 13：线性传递通道图[4]

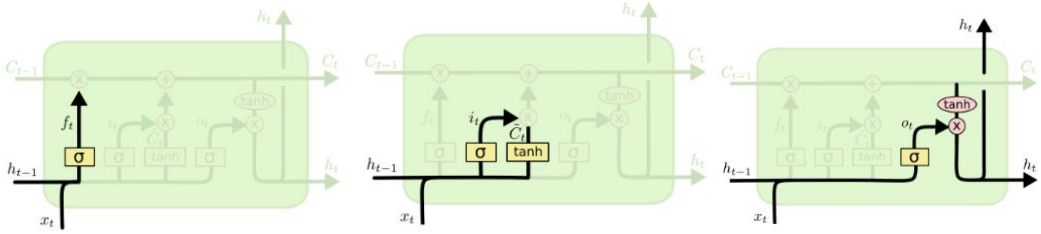

图 14：遗忘门、输入门、输出门[4]

在某一 t 时刻，输入门 i_t 和遗忘门 f_t 都以都以输入变量 x_t、上一时刻 t-1 的输入变量 h_{t-1}和偏置 b 作为输入，借着通过激活函数得到响应值。

遗忘门如图 14（左），此门决定了 LSTM 何时从系统状态丢弃信息，其计算公式为：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{26}$$

输入门如图 14（中），此门决定了 LSTM 要把何种信息存进新细胞状态中，涉及的计算公式为：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{27}$$

$$\bar{c}_t = \tan(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{28}$$

$$c_t = f_t * c_{t-1} + i_t * \bar{c}_t \tag{29}$$

输出门如图 14（右），其决定 LSTM 需要输出的值，涉及计算公式如下：

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{30}$$

$$h_t = o_t * \tan(c_t) \tag{31}$$

分析完 LSTM 模型之后，我们发现 LSTM 无法编码从后到前的信息，如句子：“这个小区的秩序乱得不行”，这里的“不行”是对“乱”得程度地一种修饰。所以更好地，我们需要使用能够进行双向编码的 Bi-LSTM 模型，全称为双向长短时记忆网络，其可以更好地捕捉双向的语义依赖。

Bi-LSTM 就是基于前向的 LSTM 与后向的 LSTM 的结合形成的。采用了 Bi-LSTM 的模型经过实验检验，比普通的 LSTM、RNN 效果有明显提升，它可以更完整的对句子的深层语义进行编码，中间编码向量既包含了句首的语义信息，也保留了句尾的语义息。因此能更好的表征语句的深层语义，比如，我们对“我爱泰迪”这句话进行编码，其模型如图 15:

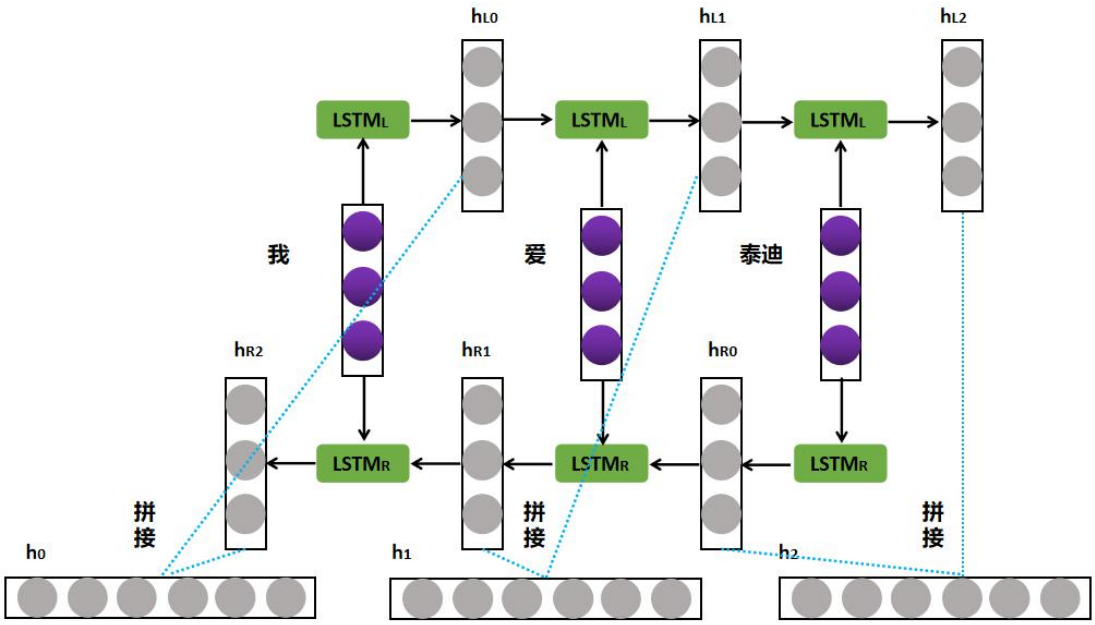

图 15：Bi-LSTM 编码句子

前向的 LSTM 依次输入“我”，“爱”，“泰迪”得到三个向量｛hL0, hL1, hL2｝，后向的 LSTM 依次输入“泰迪”，“爱”，“我”得到三个向量｛hR0, hR1, hR2｝。最后将前向和后向的隐向量进行拼接得到｛hL0, hR2｝，｛hL1, hR1｝，｛hL2, hR0｝，即｛h0, h1, h2｝。

对于 Bi-LSTM 模型的搭建，我们采用 PaddleHub 模块预训练模型 Senta[16]完成，Senta 即情感倾向分析的英文缩写，针对带有主观描述的中文文本，可自动判断该文本的情感极性类别并给出相应的置信度，可以很方便地帮助我们分析热点问题和网络舆情的监控。paddlehub 模块支持预测和 Fine-tune。通过 paddlehub.Module(name=’ senta_bilstm’)引入 Bi-LSTM 模型。该模型是百度基于海量数据训练好的模型，经过在基于开源的情感倾向分析分类数据集 ChnSentiCorp 进行评测，和进行 fine-tune 之后可以得到更好的效果，具体数据如下：

模型	dev	test	模型（finetune）	dev	test
BOW	89.8%	90.0%	BOW	91.3%	90.6%
CNN	90.6%	89.9%	CNN	92.4%	91.8%
LSTM	90.0%	91.0%	LSTM	93.3%	92.2%
GRU	90.0%	89.8%	GRU	93.3%	93.2%
BI-LSTM	88.5%	88.3%	BI-LSTM	92.8%	91.4%
ERNIE	95.1%	95.4%	ERNIE	95.4%	95.5%
ERNIE+BI-LSTM	95.3%	95.2%	ERNIE+BI-LSTM	95.7%	95.6%

图 16：各模型在 Paddlehub 上的测评图[17]

下面我们基于附件 3 的留言详情里的部分文本数据进行情感分析预测，处理结果如图 17：

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	flag	negative_prob
239227	A00030831	恳请A7县依法拘留星沙	2019-07-26 09:59:39	A7县领导：我叫李正安。19	0	6	6	0.9954
227223	A00025574	举报A7县星沙镇星沙派	2019-10-17 17:29:29	沈书记您好！本月15日举报	1	0	6	0.9838
219334	A00084138	A7县星沙商业乐园c栋	2019-04-18 17:46:12	尊敬的政府领导：我是商业	0	0	6	0.9919
232381	A00011712	A7县星沙灰埠大市场16	2019-08-14 19:09:58	领导好：星沙灰埠大市场16	0	0	6	0.9641
246288	A00010033	A7县星沙尚鑫海悦小区	2019-07-04 14:30:24	A7县星沙尚鑫海悦小区1栋二	0	5	6	0.8691
254873	A00040338	A7县星沙灰埠小区同一	2019-11-08 01:16:08	位于A7县星沙灰埠小区同一	0	0	6	0.9649
198950	A00074795	请拆除A7县星沙一区86	2019-11-04 10:51:08	星沙一区86栋东边住户违规	0	2	6	0.9968
286187	A00042014	A7县星沙三区有店面喝	2019-06-04 13:59:52	A7县星沙三区27栋楼面朝32	0	0	6	0.8506
211058	A00050010	A7县星沙望仙路两侧梧	2019-05-31 08:27:23	尊敬的领导，你好:我是居住	0	2	6	0.3012

图 17：部分文本的情感分析预测

由图 17 预测留言详情文本的消极概率可知，该 Bi-LSTM 模型的情绪预测准确度较高，可供我们实现“留言内容的消极概率 F_{13}”指标的计算。

到此，热度评价指标的 6 个指标的计算量化工作已完成。

6.4 指标的无量纲化

由于 6 个指标中，各指标的量纲并不相同，且有正指标和逆指标之分。所谓正指标，就是指指标值越大，热度评价指标就越大；而逆指标则相反。而采取逆指标正向化，有取倒数和构造函数的方法，这里我们采用指标无量纲化的方法[18]，计算公式如下：

正指标：

$$F'_{ij} = \frac{F_{ij} - min(F_{ij})}{max(F_{ij}) - min(F_{ij})} \qquad (32)$$

逆指标：

$$F_{ij}^{'} = \frac{max(F_{ij}) - F_{ij}}{max(F_{ij}) - min(F_{ij})} \quad (33)$$

式中的 $max(F_{ij})$，$min(F_{ij})$分别为指标F_{ij}的最大值和最小值。

6.5 指标权重的分配

目前，我国学者采用较多的指标体系评价方法主要分为两种：层次分析法（AHP）[6]和主成分分析法（PAC）[5]。AHP 指标体系的建立有很大主观性，且容易将悬殊的要素放在同一层次比较，影响结果精度。而主成分分析法其主成分的解释一般带有一点模糊性，不像原始变量的含义那么清楚、确切，且需保证提取的前几个主成分的累计贡献率达到一个较高的水平。

本文，我们采取了较为客观，且合理的方式来进行指标权重的分配，我们通过开放式问卷来统计调查对象对各指标的偏向程度，以下是调查问卷的具体内容：

问卷题目	你认为网络问政平台群众反映的热点留言问题应该和以下哪种或哪几种指标相关？（选项设置为文中定义的 6 个指标，每个人可以选取 3-4 个）
调查对象	不限年龄段
调查范围	网民、NLP 领域的学者
调查有效人数	366 人
总选项投票数	1226 票

表 5：调查问卷内容说明表

指标选项投票结果统计图如下图 18：

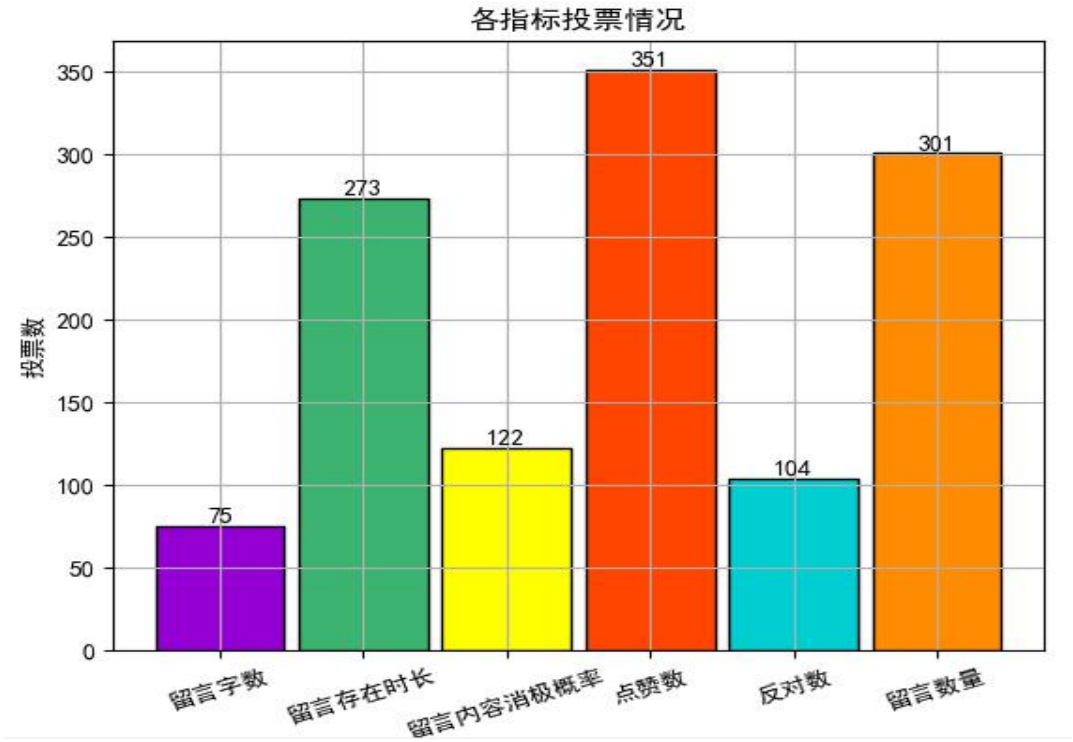

图 18：各指标投票情况

从图 18 可知，各指标投票数在总选项投票数中的占比如下表：

留言字数 $\boldsymbol{F_{11}}$指标	0.061
留言问题存在的时长$\boldsymbol{F_{12}}$指标	0.223
留言内容的消极概率 $\boldsymbol{F_{13}}$ 指标	0.099
点赞数 $\boldsymbol{F_{21}}$ 指标	0.286
反对数 $\boldsymbol{F_{22}}$ 指标	0.085
留言数量 $\boldsymbol{F_{23}}$ 指标	0.246

表 6：各指标投票占比表

因为热点问题的产生的主体是网民，他们经常性地在网络问政平台发表个人的见解，议论和评价，从而表达自己的情绪，态度和意见。而热点问题调控的主体为政府，政府作为社会热点问题的重要管理者，有责任尽快化解危机，减小事件对社会公众造成的影响和损失。所以对于热度评价指标中各指标权重的分配，我们更应该站在网民的角度，有针对性地、合理地去统计网民更关心的热点问题指标，热点问题应该是以网民群体共同希望引起政府去重视的问题。

所以我们采用统计的各指标占比作为热度评价指标的各指标权重，这样，根据以上论述，

将各指标进行线性加权所得数值即为热度评价指标。热度评价指标的计算公式如下：

$$热度评价指标 = \sum_{i=1}^{2}\sum_{j=1}^{3} w_{ij}F'_{ij} \qquad (34)$$

其中 F'_{ij} 为无量纲化后的各指标数值，w_{ij} 为各指标对应的权重，各权重之和为 1。

计算完各类热点问题的热度评价指标后，根据指标数值的大小即可确定排名前五的热点问题。然后根据热点问题留言出现时间到目前截止时间即可得出该类热点问题的时间范围。

6.6 基于 Hanlp 的命名体识别

统计出排名前五的热点问题之后，我们需要提取出相应热点问题在某一时段内发生的特定地点或特定人群。

对于特定地点的提取，我们采用的是自然语言处理工具 Hanlp[7]的 python 接口工具包 pyhanlp[15]。

首先在分词和词性标注之前，考虑到文本数据中的地址数据都是类似“A 市”,“A7 县”“A5 区”等这样用字母和数字表示表示的地址，这样的地址不利于词性标注对其进行地址的识别。所以我们利用正则表达式提取出这样类型的字母或字母和数字的组合，这里采用 re 模块[22]进行正则表达式的提取：re.findall(‘[A-Z][0-9]?’, sentence)，提取出后，为了便于词性标注工具对地址命名的识别，这里我们采取了“伪地址”对正则表达式提取的数字或数字和字母的的组合进行了替换，伪地址我们统一采取了“广州”(这里伪地址随机选取，并没刻意)，以便词性地名的标注。同时，针对像“小区”,“栋”等这样的词性是被标注为名词的，所以为了方便对地点的提取，我们根据数据，将这种词的词性修改为“ns”标注即地点标识，以便地点的识别。最后提取出特定地点后，再将“伪地址”替换回来即可。

替换完“伪地址”,接着我们利用了分词与词性标注方法 Hanlp.segment()得到分词后的词性标注。而对于特定地点的提取，我们利用了词性标注的结果结合我们设计的一个逻辑判断函数 get_location()来进行提取。基本思路如下图 19（左），这里我们随机抽取一句文本，处理流程效果如下图图 19（右）所示：

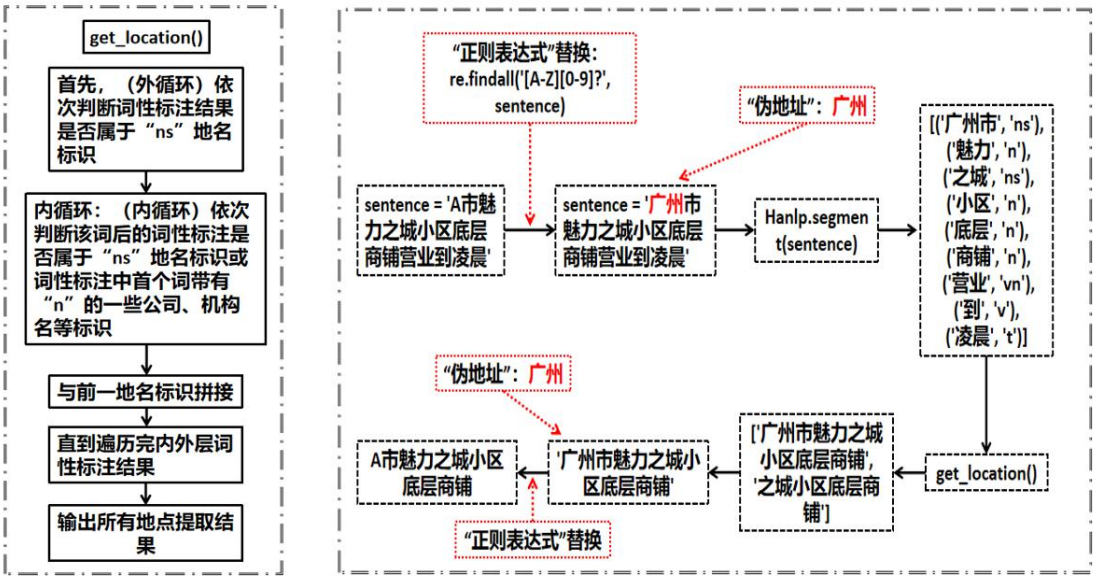

图 19：提取特定地点算法思路与处理流程图

对于有句子中有多个需要替换成“伪地址”的，我们将其依次替换，然后将原地址进行存储，提取完地点之后，再将“伪地址”依次替换为原地址即可。

采用这种提取方法，我们能够较好的提取出了留言问题所反映的特地地点。

对于特定人群的提取，我们采取了与提取特定地点相类似的方法，利用 Hanlp 的分词和词性标注，结合正则表达式。类似图 X 中的 get_location()方法，我们将其中的“ns”改为了“nnt”即职务名称,像类似“学生”,“老师”,“司机”等类似的词的词性标注即为“nnt”。在识别到“nnt”词性标注的词后，我们往该次前去找出修饰该词的词，终止条件为遇到词性标注中带有“v”开头的各种类型的动词即停止。这样就能够较好地提取出特地地点。具体流程如下图 20（这里从附件数据中随机抽取一句）：

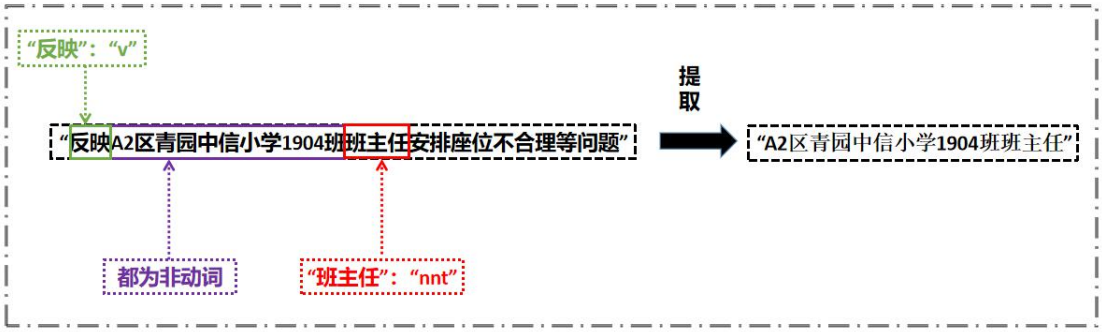

图 20：提取特定人群基本流程图

6.7 基于 TextRank 算法实现热点问题概括

在讨论 TextRank 算法[8]前，我们先来说下 PageRank 算法[8]，PageRank 在早期是用来计算网页的重要性的。整个网络可以看作一张有向图，各节点为网页。若网页 A 中存在到网页 B 的链接，那个有从网页 A 指向网页 B 的有向边。例如下图 21，从直觉可看出网页 A 的重要性

最大，因为其存在网页 B 和 C 的网页链接。

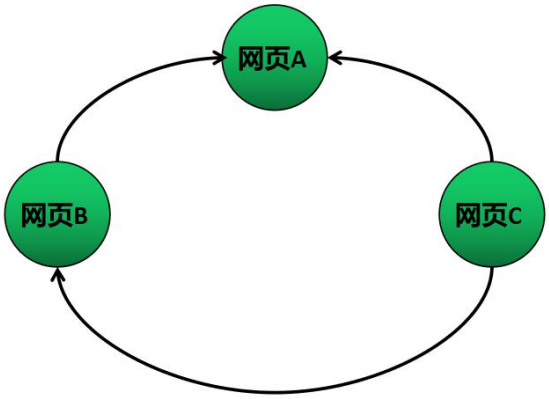

图 21：网页连接图（举例）

根据此原理构造完有向图后，使用下面的公式计算其中网页的重要性：

$$S(V_i) = (1-d) + d * \sum_{j \epsilon In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \qquad (35)$$

其中 $S(V_i)$ 是网页 i 中的重要性(PR 值)。d 是阻尼系数，一般设置为 0.85。$In(V_i)$ 是存在指向网页 i 的链接网页的集合。$Out(V_j)$ 是网页 j 中的链接存在的链接指向网页的集合。$|Out(V_j)|$ 即为集合中元素个数。

PageRank 算法利用上面的计算公式迭代一定次数后即可得到结果。初始化每个网页的重要性为 1。

那么，基于 PageRank 的算法思想，我们可以联想到借用此思想来提取文本中的关键语句。因此 TextRank 算法就是基于 PageRank 算法进行关键词或关键语句的提取。将每个句子看成图中的一个节点，若两个句子之间有相似性，则这两个节点之间存在一个无向有权边，权重为相似度。

句子中相似度采用的计算公式如下：

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)} \qquad (36)$$

分子是在两个句子中都出现的单词的数量。$|S_i|$是句子 i 的单词数。

由于是有权图，所以 TextRank 算法对对 PageRank 公式略做修改：

$$WS(V_i) = (1-d) + d * \sum_{V_j \epsilon In(V_i)} \frac{w_{ji}}{\sum_{V_k \epsilon Out(V_j)} w_{jk}} WS(V_j) \quad (37)$$

对于排名前五的热点问题，由于各类问题中的留言详情中普遍字数偏长，不利于相关部门工作人员快速、及时处理热点问题。所以我们需要对各类热点问题进行概括性描述。

经过我们对“附件 3 筛选表”中数据进行分析与观察，发现其中的留言主题的数据能够帮助我们很好地进行概括性地描述。但是如果只简单地抽取相应热点问题的一条留言主题作为问题描述，容易出现错误性概括。基于此，我们可以这样做，将某类热点问题的全部留言主题各自成为一个句子，然后将各句子进行合并。接着，我们采取 TextRank 算法来实现提取其中最为关键的一句作为问题描述。具体地我们采取了 TextRank4ZH 模块[19]进行 TextRank 算法的实现：

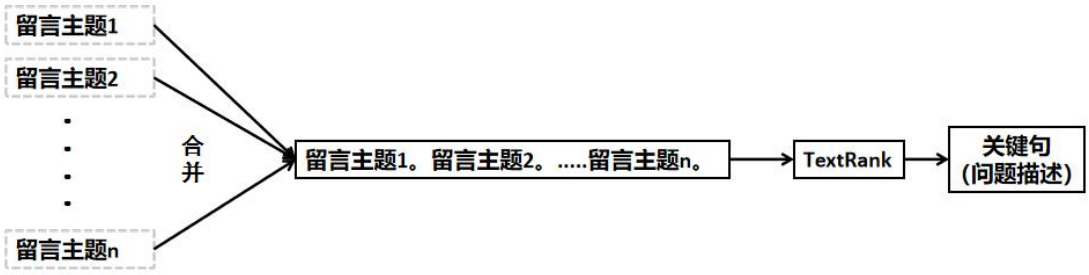

图 22：问题描述提取流程

因为采取此种方法，能够提取出某类热点问题中所有留言主题中最为关键的一个主题，用这个主题作为问题描述能很好地概括该类热点问题的内容，根据某类热点问题的部分数据，进行该处理，提取结果如下图 23：

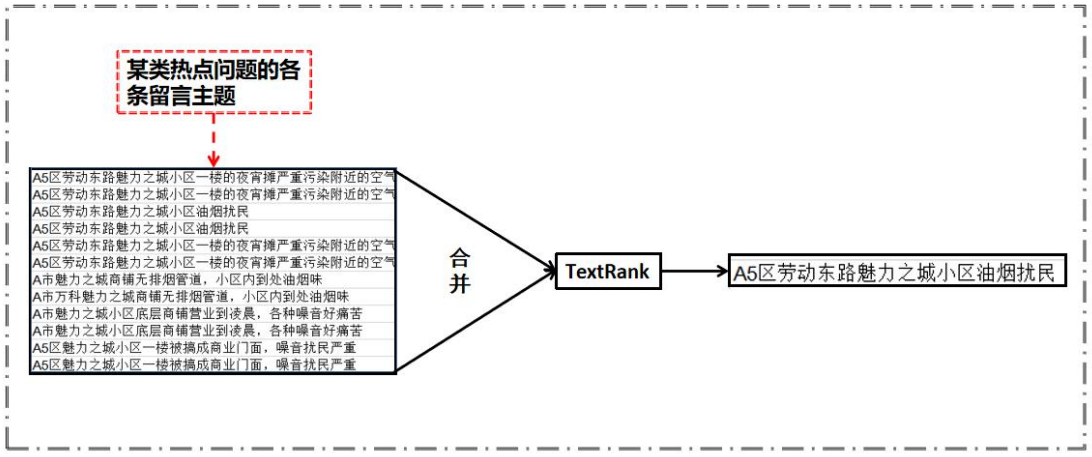

图 23：部分数据的问题描述提取效果

到此，我们已经完成了问题二的分析与求解，通过利用 pandas 模块处理好上述结果并整理完成了“热点问题表”，“热点问题留言明细表”。

七、任务三：答复意见的评价

7.1 评价方案的制定

在任务三中，我们所要完成的是从不同角度对附件 4 中的答复意见的质量制定一套合理且全面的评价方案，我们充分利用附件 4 中已有数据，挖掘出与答复意见的质量相关的因素，经过分析与挖掘，我们决定从以下 3 个角度：相关性，完整性和可解释性，对答复意见的质量做出评价：

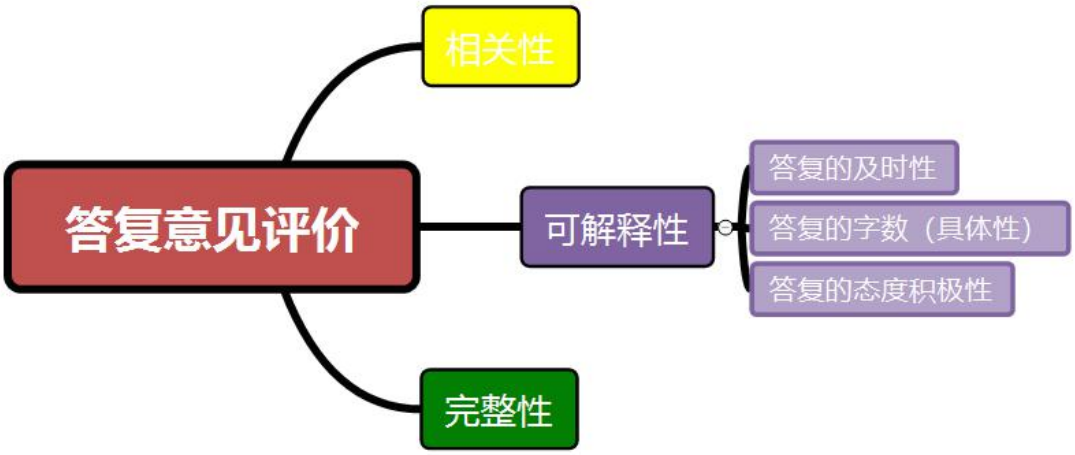

图 24：对答复意见各角度评价的思维导图

对于每个角度方面的评价，我们将其量化为具体的数值得分，分值越高，代表评价质量越高。

7.2 不同角度评价的实现与量化

7.2.1 相关性

对于相关部门的答复意见，应该与网民所反映的留言问题相关，而不是进行无关的答复。我们对留言详情和答复意见进行 word2vec[9]预处理，然后计算句子词嵌入的平均值来计算相似度(采用余弦相似度计算)，用于反映答复意见与留言问题的相关程度。因此，我们将答复意见的相关性量化如下：

相关性量化分值 s_1	相关性
5	0.9-1
4	0.8-0.9
3	0.5-0.8
2	0.3-0.5
1	0-0.3

表 7：答复与留言相关性量化分值表

将附件 4 的答复意见与留言详情做相关系计算后，相关性数值分布直方图如下图 25：

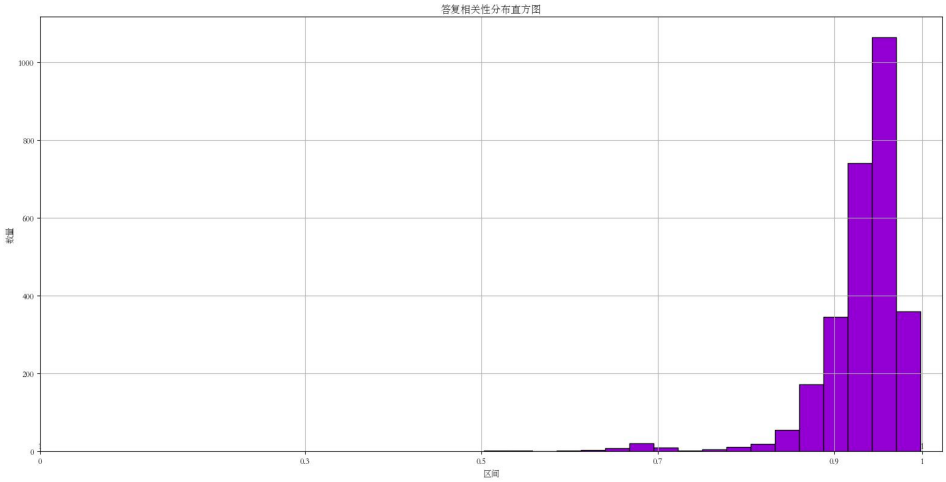

图 25：答复意见与留言问题相关性分布图

由图 X 可知，答复相关性数值在 0.9 以上的占多数，说明相关部门对于留言问题能够针对性地处理，给出与留言所反映的问题具体的处理措施。存在一些相关性较低的回复的原因可能是存在部分类似“网友：您好！留言已收悉”的答复。

7.2.2 完整性

通过查阅资料发现，各地区的网络问政平台对于答复意见的格式有一定规范。一般以类似于书信的格式进行答复。观察附件 4 中的数据，大部分以类似图 26 这种格式进行答复：

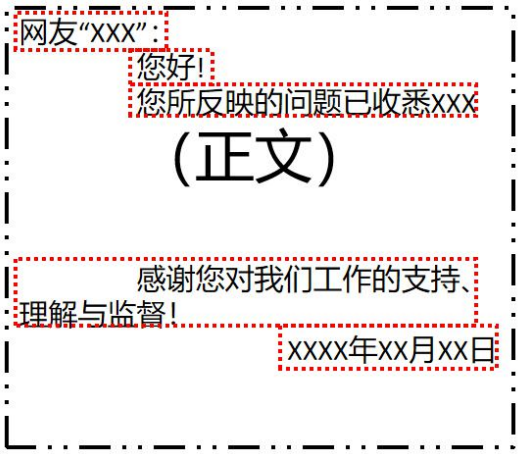

图 26：规范答复意见的格式图

对此，为了对答复意见的完整性做出评价，我们将图 26 中的红框部分分别作为评判答复意见完整信的五个部分，并将答复意见中满足其中一个部分的得 1 分，不满足得 0 分。

对于如何判断答复意见是否满足以上格式，我们利用的是正则表达式的去匹配和判断，

开头格式中：可能存在多个表达，如“已收悉”，“已查阅”等，我们通过添加类似多个相关关键词进行匹配。

结尾格式：通过在结尾匹配“感谢”，“谢谢”等词即可容易判断出是否满足第四部分。而对于年月日，通过匹配多种日期格式即可判断。

因此，我们将答复意见的完整性量化如下：

答复意见完整性分值 s_2	答复格式（开头/结尾）规范
5	满足 5 处规范
4	满足 4 处规范
3	满足 3 处规范
2	满足 2 处规范
1	满足 1 处规范
0	满足 0 处规范

表 8：答复意见完整性量化分值表

完整性量化分值分布图如图 27：

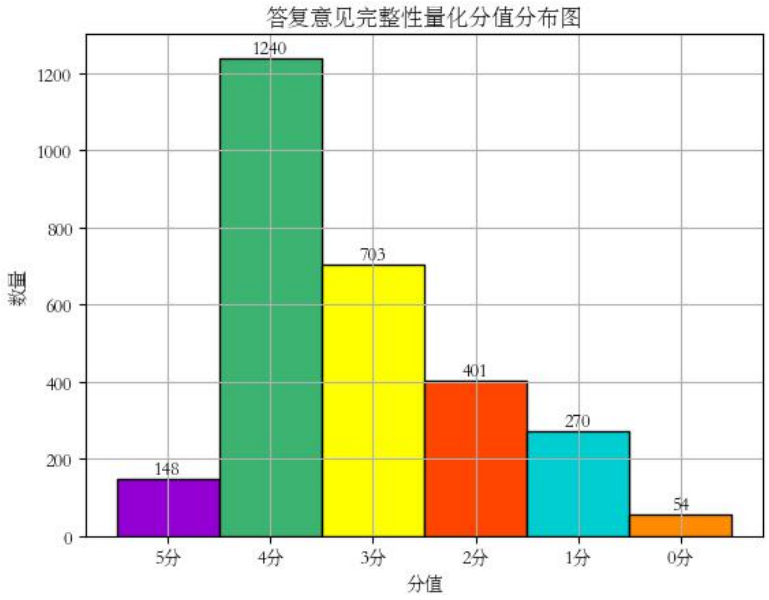

图 27：答复意见的完整性量化分值分布图

从图 27 可知，格式不规范的答复意见占多数，相关部门在答复意见的完整性即规范答复意见格式上仍需提高。

7.2.3 可解释性

对于答复意见，可解释性是指留言的答复意见中应包含足够的留言人所需要的可以理解的信息，能够较好的帮助留言人了解或解决其反映的问题。

我们通过观察数据和分析，确定了对我们定义答复的可解释性相关的三个方面的信息：答复时间的及时性，答复意见的字数，答复意见情绪倾向积极的概率。

从附件 4 中可知，通过人为判断，答复意见中存在一些可解释性不高的答复，一些类似于图 28 所示的答复，在实际中可能容易让网民引起歧义，误以为是敷衍的回复，从而导致对政府部门的信任度下降。

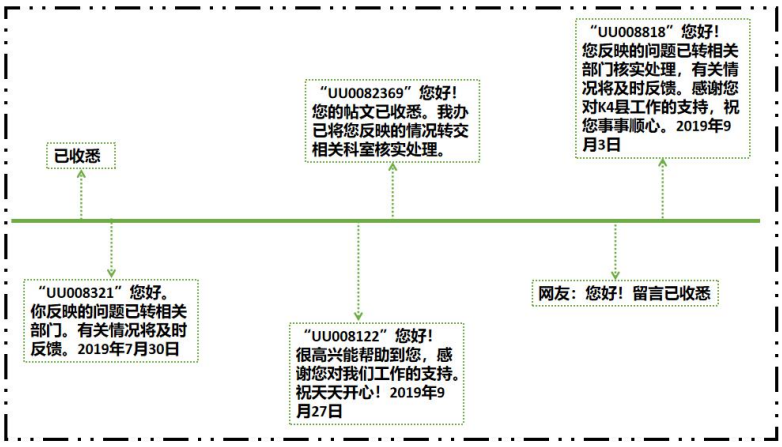

图 28：部分解释性不高的答复意见

这种回复过于简单，未详尽政府具体处理情况，工作进度不透明，类似于这种称为“万能回复”的答复，应该避免，以免引起群众反感和不信任。所以，相关部门的答复应该具体说明接下来网民反映的问题应该怎样处理，而不是简单地回复下类似“已交由先关部门去处理”的答复。

答复的及时性，在一定程度上能够引起网民的好感，让网民认为自己的问题得到了重视。我们利用了留言时间与答复时间的差值天数作为考虑点，通过将天数转为成 1-5 分的数值。经过调查调查和查阅相关资料，发现各地区政府办公厅都有相关规定：关于做好留言板回复工作的意见。其中某省对于加快回复办理时限中有规定“信访类留言办理一般不超过 30 天，咨询及其他类留言办理一般不超过 7 天。对限期不能答复的，要及时予以说明。”

因此，我们认为在 7 天以内的答复，都是及时性较高的回复，而对于超过两个星期及一个月以上的，认为是留言答复不及时。对于答复意见的及时性量化如下表 9， ：

答复意见及时性量化分值 s_{31}	时间差/天数
5	0-3
4	3-7
3	7-14
2	14-30
1	30 以上

表 9：答复意见及时性量化分值表

通过对附件 4 中整体的答复时间可视化如图 29：

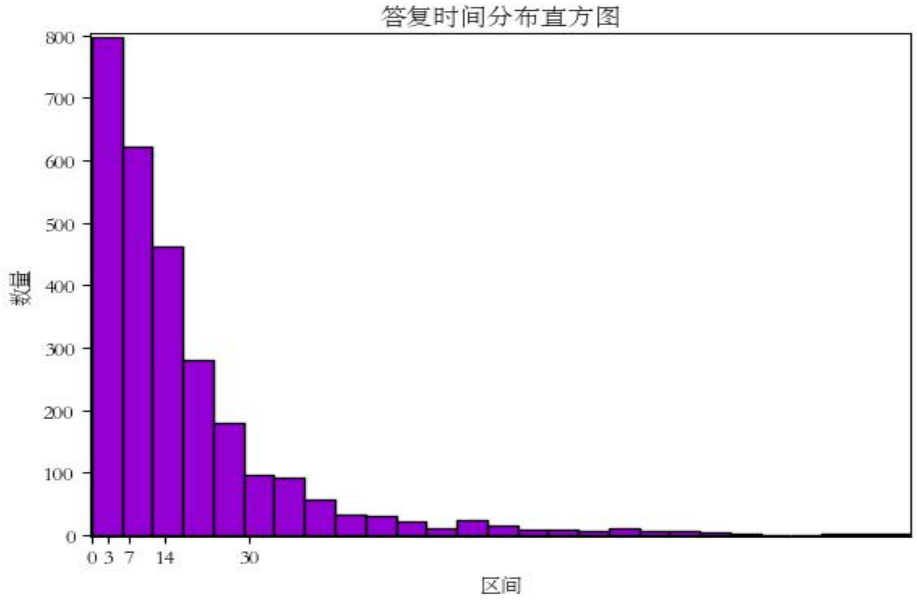

图 29：答复时间分布直方图

整体来看，对于网民的留言问题，政府部门大多数在一个星期以内进行了回复，且多数在基本在一个月内进行了回复，少数回复较慢。因此政府部门对于网民的留言问题还是较为关注且及时进行答复的。

而答复意见的字数，也能够从侧面反映出政府部门对网民反映的问题的重视，并给予了确切且具体的处理答复意见，我们将其量化为如下表所示的分值：

答复意见字数量化分值 s_{32}	字数
5	500 以上
4	100-500
3	50-100
2	10-50
1	0-10

表 10：答复意见字数（具体性）量化分值表

通过对附件 4 中整体的答复意见的字数量化分值可视化如图 30：

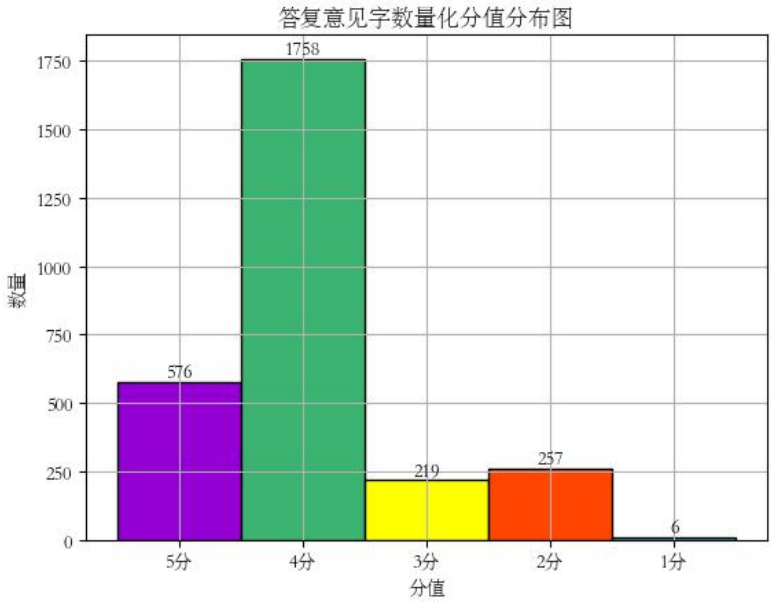

图 30：答复意见字数量化分值分布图

由图 30 可知，答复的意见字数大部分在 100 字以上，其反映出对于群众的留言问题，相关部门能够认真耐心回答。

答复意见情绪倾向积极的概率，能够反映出政府部门耐心且愿意认真考虑留言问题的态度，在一定程度上，是对单纯从答复字数上判断答复详细与否的修正。因为假若答复字数少，相关工作人员能够以较好的态度去回答的留言问题，也能够让网民心里有底，信任政府部门能够处理好该问题。所以这样的答复不一定代表了可解释性低。且一般类似这种回复，只是政府部门

已经收到了问政的请求并开始进行相关处理的通知，并非正式回复，从受理到办结一般会有60天的期限，中间可能正在转办等步骤，而且还要进一步核查网民的留言问题是否正确。

对于答复意见情绪倾向积极性的量化，这是一个情感分析问题，我们利用了前文所述的Bi-LSTM模型[4]进行答复意见态度倾向于积极的概率计算，然后根据此数值转换成1-5分，我们将态度积极性的量化如下：

答复意见态度量化分值 s_{33}	态度倾向积极概率
5	0.9以上
4	0.7-0.9
3	0.5-0.7
2	0.3-0.5
1	0-0.3

表11：答复意见态度量化分值表

将附件4的答复意见进行情感倾向分析，得出量化分值分布图，如图31：

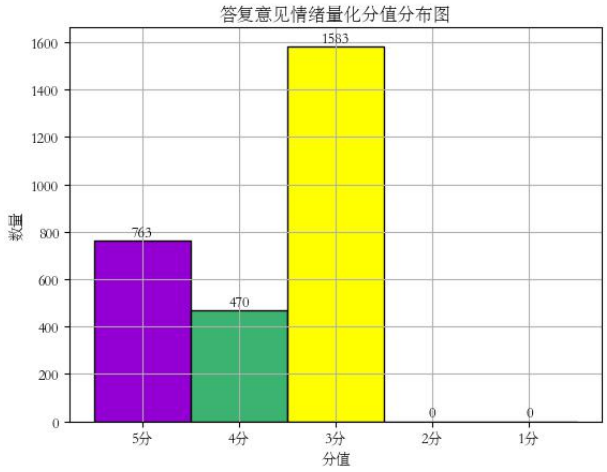

图31：答复意见情绪量化分值分布图

由图X可知，不存在态度倾向消极的答复意见，处于中性情绪的答复占多数，同时答复意见中表现出积极态度的答复意见也占多数。因此，可知相关部门在服务态度方面做的较好。

综合上述三个方面：答复时间的及时性，答复意见的字数，答复意见情绪倾向积极的概率，我们将答复的可解释性量化为三者量化分值之和：

$$s_3 = \sum_{i=1}^{3} s_{3i} \qquad (38)$$

其中，s_3 即为答复可解释性的量化分值。

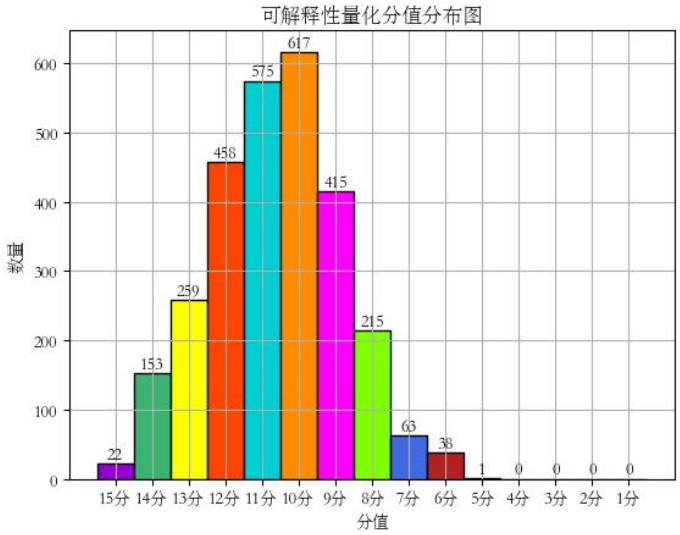

图 32：可解释性量化分值分布图

从图 32 可知，附件 4 中所反映的相关部门在答复意见的可解释上处于中等偏上水平的占多数，对于网民的留言问题，能够较积极认真地去答复。但也存在部分可解释性较低的答复，综合前面对可解释性的三个方面的分析，相关部门在答复的及时性以及答复的详细程度上有待提高，且应提倡以积极的态度去答复网民的留言问题。

7.3 评价方案小结

对答复意见的三个方面：相关性、完整性和可解释性进行量化后，即可通过以上的方案对答复意见的质量做出较为合理的评价，以此规范政府相关部门对于网民留言问题的处理机制，及时、有效地处理好网民所反映的问题，并及时将处理结果告之群众，让网络问政平台成为一个利民的高效平台。

八、总结

8.1 模型的优点

我们采用了多分类 SVM 模型结合 SGD 进行留言内容的一级标签分类模型的构建与训练，取得了较好的分类效果，SVM 泛化能力强，分类速度快且结果易于解释。

对于聚类问题我们采用的重复二分聚类算法在性能和速度上均高于传统的 K 均值算法。并利用了余弦准则函数的增幅阈值，自动确定聚类个数。采用相似度对聚类结果进行修正，达到了较好的聚类效果。

我们定义了二级指标体系对热度指标进行了衡量，较为全面地构建了热度评价指标。对于文本情感分析，我们采用了 Bi-LSTM 模型，能够很好的捕捉文本的双向语义依赖。相似性中利用 word2vec 处理后计算句子平均词嵌入计算相似性表现效果较好。进行在命名体识别中我们结合词性标注，正则表达式与逻辑判断，很好的提取出了特定地点或人群。关于问题描述采用了组合留言主题进行 TextRank 提取关键句的方法。能够快速准确的提取出组合留言主题中最

为关键的一句作为问题描述。

从三个维度：相关性、完整性和可解释性并将其量化为具体数值得分，对答复意见的质量进行较为全面的评价，以此评价方案对答复意见进行评价，能够较好地规范、监督相关部门的答复网络问政工作，让更多网民参与到网络问政平台中来。

8.2 模型的缺点与未来改进

对于“一对全”这种多分类 SVM 模型，虽然只需训练若干个（个数较少）的分类器，但是若面临大规模的训练的样本，训练速度可能会急剧减慢。且由于 SVM 是借助二次规划来求解支持向量，其中会耗费较大的存储空间和运算时间。因此，未来将会考虑如何提高多分类 SVM 算法的训练速度和性能，或者采取更适合大规模数据的模型，可考虑采取当今较为火热的一些深度学习模型，以提高训练速度和准确度，同时能够免去一些人为提取特征的繁琐步骤。

二分重复聚类算法中对于自动确定分类的数目，采取的是基于余弦距离的准则函数的增幅阈值作为聚类算法停止聚类的依据。即使如此，这个阈值也还是需根据人为调试去设定如何达到一个较好的聚类效果。因此，对于聚类算法如何尽量减少人为的主观性去调试和寻找更适合此类文本数据的模型是我们接下来的工作。

我们采用的情感分析 Bi-LSTM 模型，是 PaddleHub 开源的一个预训练模型，我们是直接利用了此预训练模型进行文本情感的分析，因为也能够达到我们所需的预测标准。考虑到，应该提高模型对我们这类留言文本的更好的预测能力与精度，接下来我们的工作是在此预训练模型的基础上，利用有关此类留言问题的数据去训练更适合我们的预测模型，以提高训练模型的效率和准确度。

在热度评价指标体系构建中，关于指标权重的分配问题，我们是通过网民的角度去调查和热点问题相关的指标，以网民对指标的投票数作为了权重的分配。此种做法有一定合理性，但是同时缺乏了一定的理论支撑和说服性。因此，未来我们将会考虑结合更为客观且利用机器而不是人为来进行指标权重的统计。

最后，希望我们的工作可以成为人工智能黄金时代的一朵小小浪花，为人类事业做出贡献。同时十分感谢泰迪杯官方构建出的“智慧政务”中的文本挖掘应用赛题。

九、参考文献

[1]. https://blog.csdn.net/asialee_bird/article/details/81486700

[2]. https://scikit-learn.org/stable/modules/sgd.html

[3]. https://github.com/NLP-LOVE/Introduction-NLP

[4]. https://zhuanlan.zhihu.com/p/47802053

[5]. 韩小孩,张耀辉,孙福军,王少华.基于主成分分析的指标权重确定方法[J]. 四川兵工学报,2012,33(10):124-12

[6]. 郭金玉, 张忠彬, 孙庆云. 层次分析法的研究与应用[J]. 中国安全科学学报, 2008,

18(5):148-153.
[7]. https://github.com/hankcs/HanLP
[8]. https://www.letianbiji.com/machine-learning/text-rank.html
[9]. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
[10]. https://pandas.pydata.org/
[11]. https://github.com/fxsjy/jieba
[12]. https://github.com/shibing624/text2vec
[13]. https://scikit-learn.org
[14]. 李航. 统计学习方法. 北京：清华大学出版社，2012
[15]. https://github.com/hankcs/pyhanlp
[16]. https://www.paddlepaddle.org.cn/hublist?filter=en_category&value=Sentiment Analysis
[17]. https://aistudio.baidu.com/aistudio/projectdetail/215814
[18]. 叶宗欲.关于多指标综合评价中指标正向化和无量纲化方法的选择[J].浙江统计, 2003(4):24 -25
[19]. https://github.com/letiantian/TextRank4ZH
[20]. https://zhuanlan.zhihu.com/p/31886934
[21]. https://www.jianshu.com/p/e4d5a0fbcefe
[22]. https://docs.python.org/3/library/re.html