“智慧政务”中的文本挖掘应用

摘 要

近年来，网络问政平台成为群众反映社会热点问题、政府部门了解民情民意等的重要渠道。由此带来的大量留言文本数据给相关部门带来了极大挑战。引入自然语言处理和文本挖掘的方法解决相关问题是一种常见且有用的手段。

通过收集问政平台的留言数据，本身蕴含着大量值得挖掘的信息。但是，由于文本数据本身的高纬性和语义复杂性，要想提取价值信息，需要我们对数据进行严谨的处理和科学的分析。

针对题目中的具体问题，我们通过以下方法进行解决：

1.针对问题一，参考附件一的分类标签，我们在附件二的数据基础上进行了数据的初步分析，掌握其一般的统计特征。进一步的，我们尝试对数据进行结构上的预处理，并采用不同的文本编码方法，诸如词袋模型、改进 Bert 模型等，以生成涵盖文本特征的文本向量；下游分类任务采用 Bert 原生分类、LightGBM、朴素贝叶斯、SVM 为代表的的四类分类器，比较分类结果的 F-score，对于效果较好的 bert 原生多分类器、svm、lightgbm 基模型， 我们采用 stacking 方法进行了模型的融合，最终得到泛化能力强，分类准确率为 97.32%的留言文本分类模型，调优模型过程汇总在数据附件 5 中。

2.针对问题二，对热点事件的识别，我们参考一般舆情传播动力衡量指标，提出了衡量事件热度的四个指标：事件反映持续时间、群众留言行为强度、当事人或物社会影响力以及群众普遍关注度。在提出上诉指标之后，使用层次分析法，将定性的赋权决策过程数学化，得出各个指标权重。对于热点事件，我们采用多层聚类的方法，使用 Bert 对文本进行编码，从文本向量层次进行粗粒度聚类，再从时间维度上使用 DBSCAN 算法基于密度做进一步聚类，得到高纯度事件簇。在前述热度评价指标的基础上，获取热度 Top5 的事件并汇总到附件热点问题表中。对于热点事件明细信息要求的地点和人群信息，我们采用 Bert+BiLSTM+CRF 组合模型对实体进行识别，并在已识别实体的基础上对人群进行关联。对于热点问题描述，我们提出了 NER、详情摘要与留言主题一同作为参考的方法、进一步提高了问题描述的准确性，相关留言一并汇总到附件热点问题留言明细表中。

3.针对问题三，对相关部门的答复文本质量进行评价，我们围绕相关性、完整性、及时性、可解释性以及中立性等，在充分探索附件 4 数据的基本特征的基础上，从多层次将抽象的指标具化为可在数据集上实现的指标，如相关性转换为文本向量距离、可解释性转换为层次结构评价以及法规政策引用量等。对于多指标评价，我们用层次分析法同样给出了指标权重系数，并说明了其合理性。多指标评价结果汇总到数据附件 4 中。

关键词：特征提取 Bert 多层聚类 DBSCAN BiLSTM 层次分析法

Abstract

In recent years, the online political inquiry platform has become an important channel for the masses to reflect social hot issues and government departments' understanding of public sentiments. The resulting large amount of message text data has brought great challenges to the relevant departments. Introducing natural language processing and text mining methods to solve related problems is a common and useful method.

By collecting the message data of the questioning platform, it contains a lot of information worth mining. However, due to the high latitude and semantic complexity of the text data itself, in order to extract value information, we need rigorous processing and scientific analysis of the data.

For the specific problems in the title, we solve them by the following methods:

1. For question one, refer to the classification label in Annex I. Based on the data in Annex II, we conducted a preliminary analysis of the data to master its general statistical characteristics. Further, we try to pre-process the data structurally, and use different text encoding methods, such as bag-of-words model, improved Bert model, etc., to generate text vectors that cover text features; downstream classification tasks use Bert native classification, LightGBM , Naive Bayes, and SVM are the four types of classifiers. Compare the F-score of the classification results. For the bert native multi-classifier, svm, and lightgbm-based models with better results, we used the stacking method to fuse the models. Finally, a message text classification model with strong generalization ability and classification accuracy rate of 97.32% is obtained. The process of tuning the model is summarized in data attachment 5.

2. Regarding problem two, for the identification of hot events, we refer to the general public opinion propaganda power measurement indicators and put forward four indicators to measure the heat of the event: the duration of the event reflection, the intensity of the masses ’message behavior, the social influence of the parties or things, and the mass Attention. After the appeal indicators are submitted, the analytic hierarchy process is used to mathematically determine the qualitative weighting decision-making process to obtain the weight of each indicator. For hot events, we use a multi-layer clustering method, use Bert to encode text, perform coarse-grained clustering from the text vector level, and then use the DBSCAN algorithm to further cluster based on density from the time dimension to obtain high-purity event clusters . On the basis of the aforementioned heat evaluation index, the events of Top 5 of heat are obtained and summarized in the hotspot question table. For the location and crowd information required by the hot spot event detailed information, we use the Bert + BiLSTM + CRF combined model to identify entities, and associate the crowd based on the identified entities. For the description of hot issues, we have proposed a method of NER, summary of details, and the subject of the message as a reference to

further improve the accuracy of the problem description. Related messages are summarized in the hot question message list.

3. For question two, evaluate the quality of the response texts of the relevant departments. We focus on relevance, completeness, timeliness, interpretability and neutrality, etc., on the basis of fully exploring the basic characteristics of the data in Annex 4, from more Hierarchy materializes abstract indicators into indicators that can be implemented on the data set, such as relevance conversion into text vector distance, interpretability into hierarchical structure evaluation, and amount of laws and policies cited. For multi-index evaluation, we also use the analytic hierarchy process to give the index weight coefficients and explain their rationality. The multi-index evaluation results are summarized in data attachment 4.

Keywords: feature extraction, Bert, Multi-layer Clustering, DBSCAN BiLSTM, AHP

目 录

第一章 概述

1.1 问题描述

近年来，随着微信、微博、市长信箱、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，而目前相关处理还主要依靠人工来进行，包括留言划分和热点整理等重复繁杂的任务。

幸而在大数据、云计算等技术兴起的时代背景下，自然语言处理技术得到了丰富和发展，在问政平台上的使用将具有光明的前景。

立足本题数据，首先我们需要对原始的留言数据进行分类，利于后续分发到专职部门进行解决。文本数据一级标签共计 15 个，属于典型的多分类问题；之后，需要基于附件 3 提供的留言数据，识别出热点事件，包括时间段、发生地点或者特定的人群，并对问题进行一个简单且概括的描述。我们需要给出一个客观的热度评价指标，这一点会指导我们对热点事件的识别；最后我们需要给出政府部门对群众反映问题的进行答复的质量建立评价模型，主要涉及答复的相关性、完整性以及可解释性等。

1.2 论文结构安排

本文共分为四章，各章节内容安排如下：

第一章，对论文需要解决的问题进行描述，并简要介绍论文的行文结构；

第二章，针对问题一，首先对附件 2 数据集进行了统计分析，对整个留言语料进行了初步探索；针对数据集中存在类不均衡现象，我们采用过采样方法扩充了交通运输类以及环境保护类的留言数据；针对留言详情内容口语化，词向量建模时，会出现文本截断现象，我们采用 textrank 算法将留言数据提炼主题句，同时将留言主题与精简后的留言详情进行拼接；其次，考虑到 Bert 能够对语料库上下文进行动态表征，因此使用 Bert 生成文本向量；最后，将 bert 原生多分类器、svm、lightgbm 作为基模型，采用 stacking 的思想，将不同模型的训练集预测结果和训练集真实 label 做为已有标注数据，利用 softmax 思想将三种基模型进行融合，从而对测试数据打标签。

第三章，针对问题二，对附件 3 中的数据进行初步探索，挖掘留言信息的时间、语义特征。在此基础上提出了完备的事件热度评价指标并使用层次分析法确定定量的权值。对于热点事件的识别采用了多层聚类的手段，在语义类别、时间密度上对留言文本进行聚类，获取热点事件，依据热度评价指标进行排序并提取 TOP5 事件。对于实体识别、群体识别以及问题描述给出了我们的策略。

第四章，针对问题三，围绕相关性、完整性、可解释性、及时性等提出了一系列的指标，。使用层次分析法将定性与定量相结合，给出了判断矩阵并求出各个指标的权值，完成对答复文本的综合性评价。

第二章 留言文本数据分类

本章主要针对问题一，按照文本数据挖掘流程依次对数据统计特征进行探索、进行必要的数据预处理、文本数据向量化的多种方法分析与结果比较、下游分类的多种方法分析与结果比较以及模型融合与测试结果。

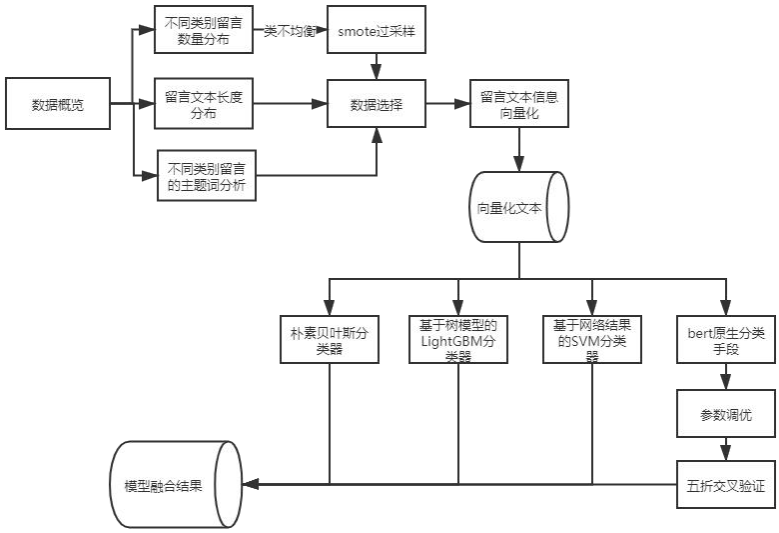


图 2.0-0：本章架构图

2.1 数据统计分析

文本数据分析能够有效帮助我们理解数据语料，快速检查出语料可能存在的问题，并对后续模型训练超参数的选择起到引导作用。因此我们从以下几个方面对数据进行统计分析。

2.1.1 总体数据概览

首先，我们对赛题附件 2 给定数据进行总体概览，结果如图所示：

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9210 entries, 0 to 9209
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   留言编号    9210 non-null   int64
 1   留言用户    9210 non-null   object
 2   留言主题    9210 non-null   object
 3   留言时间    9210 non-null   object
 4   留言详情    9210 non-null   object
 5   一级标签    9210 non-null   object
dtypes: int64(1), object(5)
memory usage: 431.8+ KB
None
```

图 2.1-1：数据的总体概览

可以看出，附件 2 中总共有 9210 条记录，不存在空值和重复字段，每条记录有 6 个属性字段分别为留言编号、留言用户、留言主题、留言时间、留言详情以及标签。其中对于本次留言分类任务而言，留言用户、时间为无关因素，可以不予考虑。

2.1.2 不同类别留言数量分布

为了分析样本数据是否存在类不均衡现象，针对一级标签字段具体取值不同，对不同类别样本数据分布进行分析，具体如下图所示：

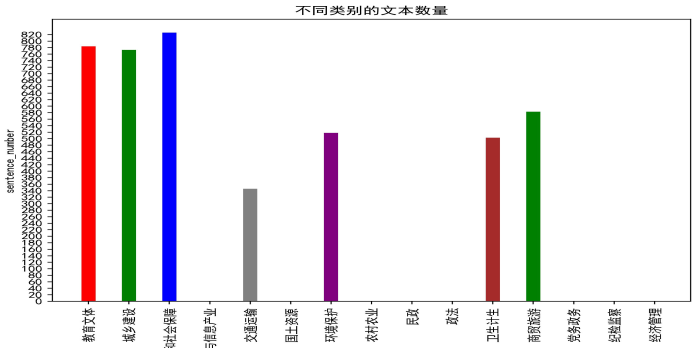


图 2.1-2：不同类别文本的数量

附件 2 中存在 7 种不同类别数据（教育文本、城乡建设、劳动与社会保障、科学与信息产业、交通运输、环境保护、卫生计生、商贸旅游），而附件 1 中给定的标签数据有 15 种，有 8 种标签数据在我们的样本数据集中没有出现。同时发现不同类别的留言数据在数量存在较大的差异，例如标记为交通运输类别的留言数据总量不及教育文体、城乡建设、劳动和社会保障数据的一半。对于这种问题，我们将根据后续模型效果来决定是否对样本数据进行类不均衡处理。

2.1.3 留言文本长度分布

由文本长度分布图可知，各类文本长度分布情况基本一致，在（0,500）区间占比较大，最长可达 12400 字。这将指导数据集的构建工作。

	商贸旅游	教育文体	劳动社保	城乡建设	交通运输	环境保护	卫生计生
均值	407	513	458	398	292	412	436
最大值	3929	12428	6237	3032	3070	2980	3026
最小值	44	30	47	45	45	57	44

图 2.1-3:各类文本长度以及总体文本长度统计分析

2.1.4 不同类别留言的主题词对比分析

为了初步掌握各类文本的标签与关键词的关系，我们运用 TF-IDF 算法对文本进行关键词统计，并绘制了词云图。

图 2.1-4：各类留言关键词词云图

由词云图可以观察出文本涉及的各类区分明确，每个类的高频关键词都有具有符合本类的鲜明特征，这对于后续分类特征构建起到指引作用。

2.2 数据集重构

2.2.1 类不均衡的问题

在数据探索部分已经发现，数据存在一定的类不均衡问题。对于这种问题我们主要采取的是过采样算法（SMOTE）。

SMOTE（Synthetic Minority Oversampling）即合成少数类过采样技术。SMOTE算法是对随机过采样方法的一个改进算法，由于随机过采样方法是直接对少数类进行重采用，会使训练集中有很多重复的样本，容易造成产生的模型过拟合问题。而 SOMTE 算法的基本思想是对每个少数类样本，从它的最近邻中随机选择一个样本 $\hat{x}_i$（ $\hat{x}_i$是少数类中的一个样本），然后在和$\hat{x}_i$之间的连线上随机选择一点作为新合成的少数类样本。

(1)对于少数类中每一个样本x_i，以欧氏距离为标准计算它到少数类样本集S_{min}中所有样本的距离，得到其 k 近邻。

(2)根据样本不平衡比例设置一个采样比例以确定采样倍率 N，对于每一个少数类样本x_i，从其 k 近邻中随机选择若干个样本，假设选择的近邻为$\hat{x}_i$。

(3)对于每一个随机选出的近邻$\hat{x}_i$，分别与x_i原样本按照如下的公式构建新的样本。

$$x_{new} = x_i + rand(0,1) \times (\hat{x}_i - x_i) \qquad （式 2.1.1）$$

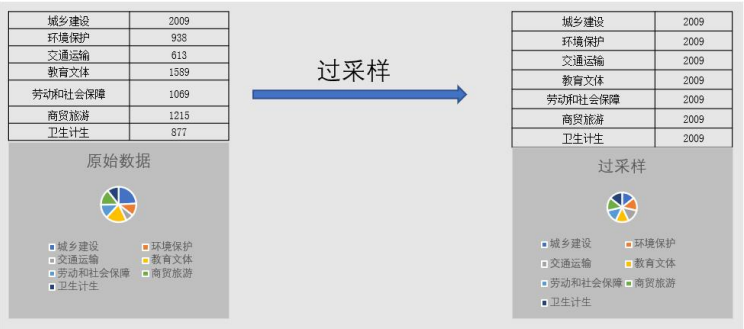


图 2.2-1：类不均衡问题的解决

2.2.2 数据选择

对于原始数据，分类的依据可以有多个，比如依据主题分类、依据留言详情分类、留言主题与详情内容融合等等。这个问题关乎模型的训练与测试数据集，我们的策略是每一种数据都进行尝试并对比结果。

图 2.2-2：数据集的重构方案

以下分别简要介绍每种方案的优缺点：首先是留言主题，主题句的特征是长度较短，具有一定的概括性，但可能存在特征表征不足的问题；对比而言的就是留言详情，从 2.1 的数据统计分析可以看出，文本长度在 200-250 字左右，内容详尽，但是由于留言的口语化和存在错别字，可能会对后续处理带来噪声；思考将两者融合，比较简单直接的方法就是进行二者拼接，但是带来的是文本数据过长的问题；为了进一步优化这个问题，尝试使用 TextRank 进行文本摘要提取，保证在最大限度下压缩文本长度而不至于失真。

具体流程如下：

- 预处理：将输入的文本内容分割成句子得 $T = [S_1, S_2, S_3, \cdots\cdots S_m]$，构建图 G = （V, E），其中 V 为句子集，对句子进行分词、去除停止词，得 $S_i = [t_{i,1}, t_{i,2}, t_{i,3}, \cdots\cdots t_{i,n}]$，其中 $t_{i,j} \in S_j$ 是保留后的候选关键词；
- 句子相似度计算：构建图 G 中的边集 E，基于句子间的内容覆盖率，给定两个句子 S_i, S_j，采用如下公式进行计算：

$$Similarity(S_i, S_j) = \frac{|\{t_k \vee t_k \in S_i \wedge t_k \in S_i\}|}{\log(|S_i|) + \log(|S_j|)} \qquad （式 2.2.1）$$

若两个句子之间的相似度大于给定的阈值，就认为这两个句子语义相关并将它们连接起来，即边的权值 $W_{ji} = Similarity(S_i, S_j)$；

- 句子权重计算：根据公式，迭代传播权重计算各句子的得分；
- 抽取摘要句：将（3）得到的句子得分进行倒序排序，抽取重要度最高的 T 个句子作为候选摘要句；
- 形成摘要：根据字数或句子数要求，从候选摘要句中抽取句子组成摘要。

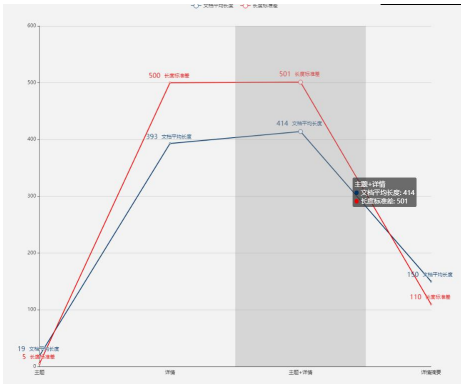


图 2.2-3：各方案的文档长度与标准差

如上图所示：相较于简单地采用留言详情或者留言主题与详情拼接，采用 TextRank 对摘要进行抽取得到的文档平均长度更短、长度标准差更小。但是与采用短小的主题相比携带更多信息。这种单文本长度将更有利于后续模型的训练与分类任务。因此，我们采用将留言主题与 TextRank 处理后的摘要句拼接，对留言详情数据集进行精炼与浓缩。

2.3 留言文本信息向量化

在自然语言处理过程中，要使文本转换为计算机认识的形式需要对文本进行向量化表示。其中文本向量化的粒度分为两种形式：

一是以词为单位，例如 word2vec、词袋模型、n-gram；

二是以句子为单位，例如主题模型，LSA,LDA 等。

然而，2018 年谷歌提出了 BERT 模型，它是一种基于双向语言模型的优秀的文本表征模型。它的优势在于它是语料库上下文进行的动态表征，不再是局限、片面的对词级特征进行识别，真正做到了完全双向关联。同时引入的 Masked 模型和 Next Sentence Prediction 对于词级预测、句级关系的 NLP 问题也都起到很好的效果。在 11 项 NLP 任务上都取得了突破，与原来的语言模型相比，BERT 几乎融合原有模型的所有优点，具体性能对比如下表所示：

模型	获得长距离语义信息程度	左右上下文语义	是否可以并行
Word2Vec	1	True	True
单向 LSTM	2	False	False
ELMo	2	True	False
OpenAI GPT	3	False	True
BERT	3	True	True

表 2-3-1 自然语言处理模型对比

Bert 本质上是一个两段式的 NLP 模型，第一阶段是 Pre-training，包括 Embedding、Transformer Encoder 和 Loss 优化，来生成一个预训练模型；第二阶段是 Fine-tuning，利用训练好的模型来完成 NLP 下游任务，支持微调。我们

基于 BERT 模型对重构后的留言数据进行向量化，在向量化的基础上在进行 NLP 下游任务。

2.4 留言文本多分类任务模型构建

下游 NLP 是典型的文本多分类问题。Bert 本身提供了文本分类的下游任务处理，同时，目前广为流行的分类模型还有基于概率模型的朴素贝叶斯、基于网络结构的 SVM 以及基于树模型的 LightGBM 算法。每种方法各有利弊，我们对每一种方法都进行了尝试，通过多种指标对分类质量、分类效率进行评估，以备考虑模型融合等后续处理手段。

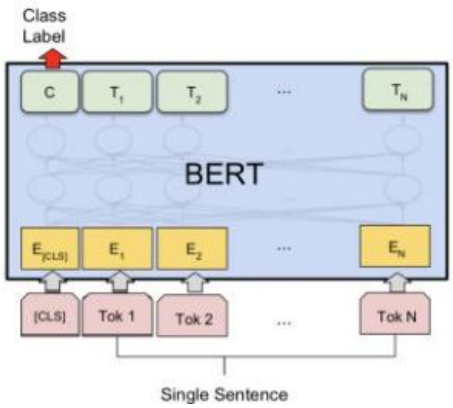
2.4.1 多分类模型

（1）Bert 原生分类手段:基于的 Token embedding 的第一个特殊符号—【CLS】的词向量值作为序列的特征向量，基本原理在于【CLS】受上下文语义影响，该特殊符可以视为汇集了整个输入序列的表征。基于 Fine-Tuning 的思想，在原有的预训练语言模型的基础上，增加少量的神经网络层，如 Softmax 层来完成文本分类的工作。

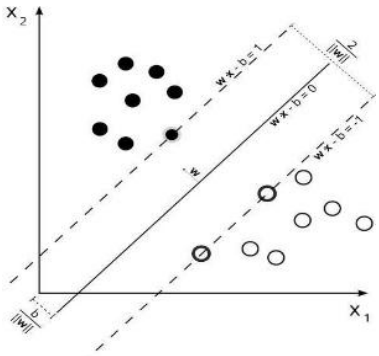
（2）基于概率模型的朴素贝叶斯分类器：朴素贝叶斯分类器是一系列以假设特征之间强（朴素）独立下运用贝叶斯定理为基础的简单概率分类器。优点在于数据集较小的情况下的仍旧可以处理多类别问题使用于标称数据。面对 Bert 生成的序列向量，朴素贝叶斯并没有很好的处理能力，主要原因是：其一，Bert 生成向量的各维度是连续属性；其二，Bert 生成向量各个维度并不是完全独立的。因此这种分类方法在原理上来讲不会具有很好的分类效果。

（3）基于网络结构的 SVM 分类方法：支持向量机的思想来源于感知机，是一种简单的浅层神经网络。但 SVM 较一般的神经网络具有更好的解释性和更完美的数学理论支撑。SVM 的目的在于寻求划分各类的超平面。当然，支持向量机可实现非线性分类，利用核函数将数据抽象到更高维的空间，进而将非线性问题转换为线性问题。

（4）基于树模型的 LightGBM 分类器：LightGBM 是对一般决策树模型和梯度提升决策树模型 XGBoost 的一种改进。内部使用 Histogram 的决策树算法和带有深度限制的 Leaf-wise 的叶子生长策略，主要优势在于更快的训练效率、低内存占用以及适用大规模并行化数据处理。



Bert 针对分类的 Fine-Tuning



SVM 中分类超平面示意

图 2.4-1：多分类模型原理图

2.4.2 参数调优及模型融合

（1）Bert 模型超参数调整

为了使得模型在我们的数据集上达到最优性能，需要依据我们的评价指标 FMacro_F1 值,寻找最优超参数组合（需要调整的超参数包括 max_seq_length、train_batch_size、learning_rate、num_train_epochs）。如下图所示：

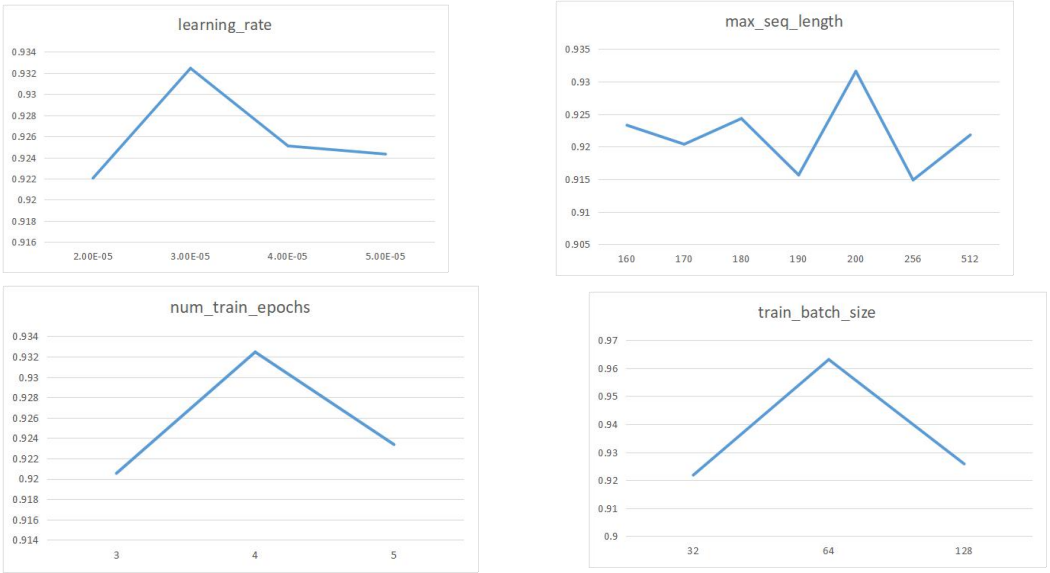


图 2.4-2：Bert 模型参数调优过程

可以看到，最优超参数组合 max_seq_length=180、train_batch_size=64、learning_rate=3e-5、num_train_epochs=4.0,其中学习率我们采用三角学习率，首先 warm_up，学习率逐渐变大，在 linear lr decay，学习率逐渐变小，有效改善训练效果。

（2）5-折交叉验证

首先，按照 8:2 的比例划分训练集和测试集；其次，在划分的训练集基础上，采用分层抽样 5 折交叉验证将数据划分成 5 份，如下图所示。每次选择不同的一

份作为模型的验证集，其中分层抽样保证了每折数据集中的各类别样本比例保持不变。

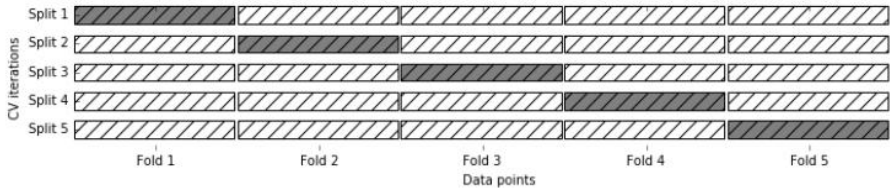


图 2.4-3：5-Fold 交叉验证示意

最后得到 5 个不同的生成模型，将 5 个生成模型的预测结果取平均，得到最终的结果。如下图所示，是我们在最优参数情况下，在 bert 原生多分类模型在验证集上得到的概率文件。

	交通运输	环境保护	教育文体	城乡建设	商贸旅游	劳动和社会保障	卫生计生
2	0.00032713896	0.00057410955	0.0006818441	0.9968797	0.0009006151	0.00042911133	0.00020737064
3	0.0011694204	0.8015326	0.0015039984	0.19136807	0.0015647836	0.0018573955	0.0010036525
4	0.00076432084	0.00052489556	0.0006641209	0.0010347302	0.99561775	0.0005653936	0.0008288095
5	0.0006468245	0.0017914898	0.9742949	0.0056827026	0.013927277	0.0016262578	0.0020304837
6	0.0005223711	0.0027361438	0.002259325	0.9932696	0.0004099459	0.0005264284	0.00027617937
7	0.00022523782	0.00034601986	0.99739397	0.0003934043	0.00038696217	0.00067567057	0.00057860836
8	0.00044765577	0.00083980657	0.000999801	0.00028358086	0.00081822457	0.0018951686	0.99471575
9	0.0004867582	0.00079138175	0.00072772754	0.9962697	0.0003705695	0.001110591	0.00024335705
10	0.00020909346	0.000245468	0.99782276	0.00051094	0.00033054297	0.000442881	0.00043848483
11	0.00024111074	0.00020442167	0.99762565	0.00030494315	0.00033590646	0.00079310185	0.0004949501
12	0.9937842	0.00059172336	0.00095558434	0.00071669	0.0020744707	0.0008229784	0.0010543033
13	0.00020835303	0.00026199577	0.99745756	0.00055005355	0.00039059264	0.0005724681	0.00055903173

图 2.4-4：Bert 模型在验证集上所得概率文件

我们选取了 Bert 原生分类模型、朴素贝叶斯、SVM 分类方法、LightGBM 分类器四种模型作为原始模型。对 4 个原始模型进行 5 折交叉验证训练，得到 5 个不同的生成模型，将 5 个生成模型的预测结果取平均，得到一个结果，将这个结果视为基模型预测的结果。

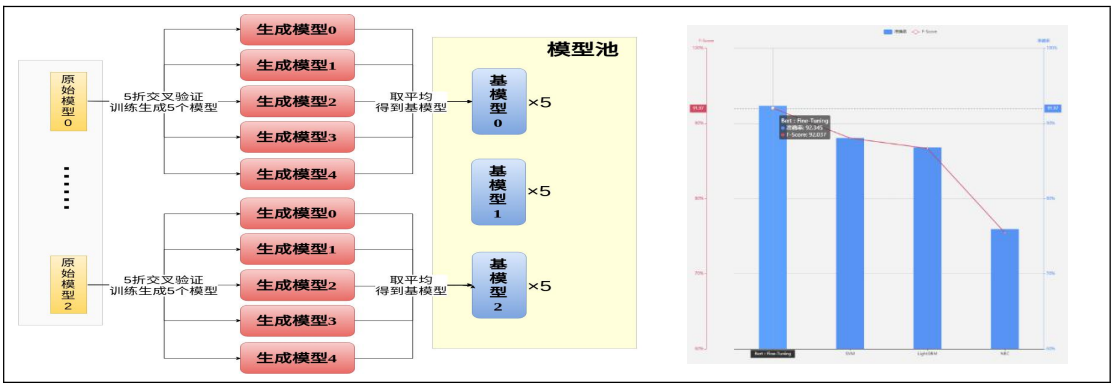


图 2.4-5：模型池构建与单模型效果对比

图 2.4-5 的右图展示了不同模型在同一测试集上分类效果（准确率与 F1 值指标计）上的对比分析：Bert 原生分类模型表现总体最优，朴素贝叶斯表现最差。

（3）模型融合

不同分类器对同一样本的分类能力不同，因此，每个分类器对每个样本都有不同的贡献能力。我们常常采用的是加权各个分类模型的结果以得到更优的分类效果。

基于前面的结果，我们考虑抛弃朴素贝叶斯分类器（NBC），因为这种分类方法在该数据集上表现不尽如人意。以下是模型融合的结构图：

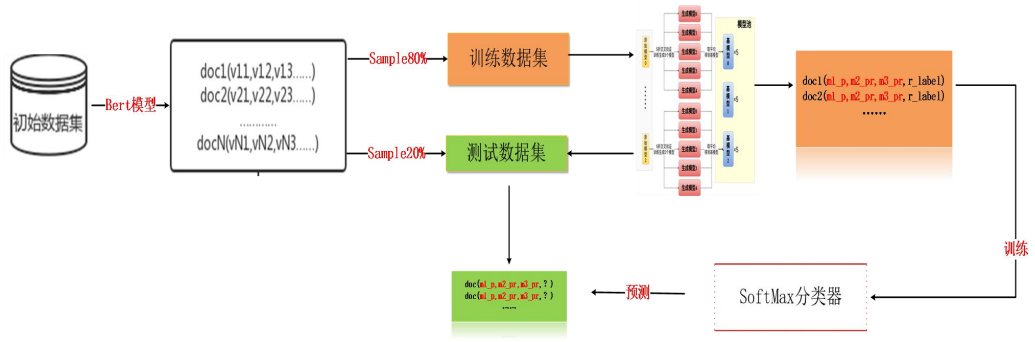


图 2.4-6：模型融合结构图

我们在三种不同 base 模型基础上进行 stacking，将三列训练集预测结果和训练集真实 label 做为已有标注数据，利用 softmax 分类器进行训练，将三列预测的结果作为测试集进行预测。下图展示了融合后模型的分类质量与单个分类器质量对比，可以看出有模型融合带来的质量(以 F-Score 衡量)提升达到 5%.

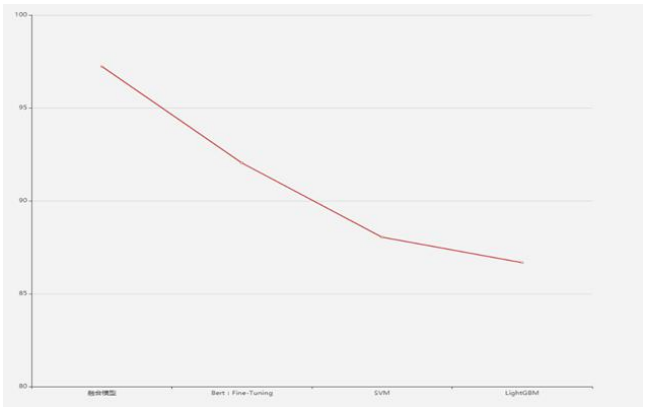


图 2.4-7：融合模型与单模型分类质量对比（最左侧为融合模型）

由于在生成基模型的时候，未使用全部数据，只是随机抽样了 80%，因此，我们可以通过设置不同的随机种子，将以上过程重复多遍（这里我们重复了 5 遍），以提高模型的泛化能力，同时从侧面反应我们模型的稳定性。左表反映的是利用不同随机种子划分的测试集的 jacard 系数，可以发现测试集之间相关性较弱。右图反映了我们 stacking 后的模型在不同测试集上 F-Score 值基本没有变化，从侧面证明了模型的稳定性。

	test1	test2	test3	test4	test5
test1	1	0.112424	0.10752	0.114789	0.112458
test2	0.112424	1	0.116662	0.114485	0.118257
test3	0.10752	0.116662	1	0.116662	0.109597
test4	0.114789	0.114485	0.116662	1	0.116895
test5	0.112458	0.118257	0.109597	0.116895	1

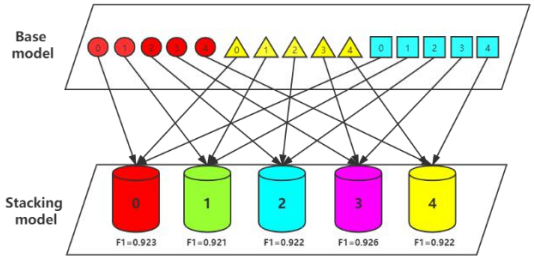


图 2.4-8： 测试集的弱相关性（左图）与融合模型稳定性（右图）

2.5 总结

本章针对问题一，首先对附件 2 数据集进行了统计分析，对整个留言语料进行了初步概览，发现数据集中存在类不均衡现象，针对此现象，我们采用过采样方法扩充了交通运输类以及环境保护类的留言数据；针对留言详情内容口语化，同时内容过长，在后面词向量建模时，会出现文本截断现象，我们采用 textrank 算法将留言数据进一步精简，提炼主题句，同时将留言主题与精简后的留言详情进行拼接以完备文本特征；考虑到 Bert 能够对语料库上下文进行动态表征，因此使用 Bert 生成文本向量（详见数据附件 1）；同时将 bert 原生多分类器、svm、lightgbm 作为基模型，采用 stacking 的思想，将不同模型的训练集预测结果和训练集真实 label 做为已有标注数据，利用 softmax 思想将三种基模型进行融合，从而对测试数据打标签，最终我们的分类 F 值提升至 97.32%，且在不同测试集上表现出较强的稳定性。

第三章 留言文本中热点事件挖掘

本章主要针对问题二，根据问题要求展开对数据的必要探索性分析，对事件热度衡量给出客观评价指标，制定某时间段内热点事件挖掘的策略，主要包括多层聚类在热点事件识别中的使用，进而抽取实体与主题，完成题述任务。

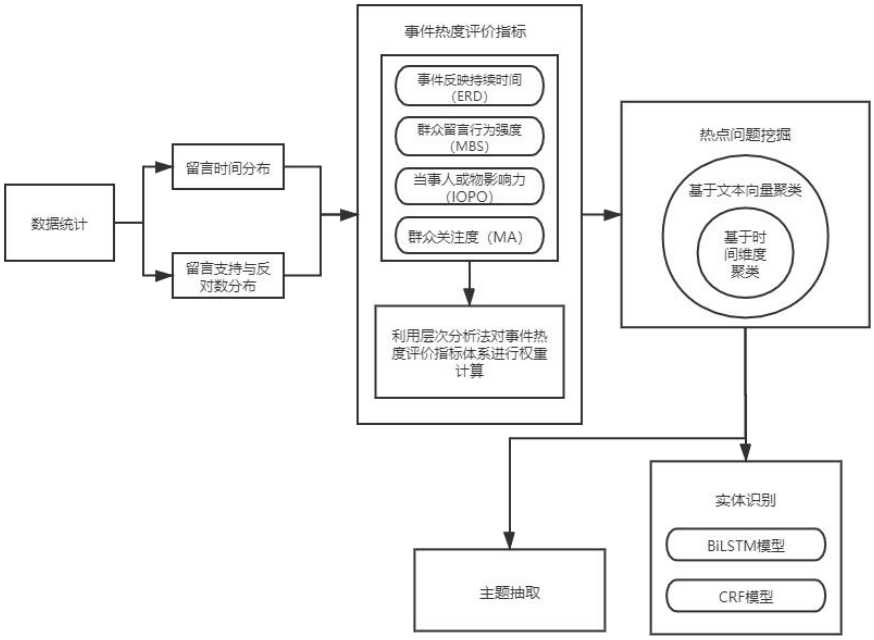


图 3.0-0：本章架构图

3.1 数据统计分析

附件 3 数据集提供了留言的多个维度信息，包括留言主题、详情、时间以及留言的反对数和支持数等。对数据集进行初步的统计分析，帮助我们发现留言实例在与热点发现问题相关的维度上表征的规律，以便于后续进行热点事件挖掘等深入分析。

3.1.1 留言时间分布

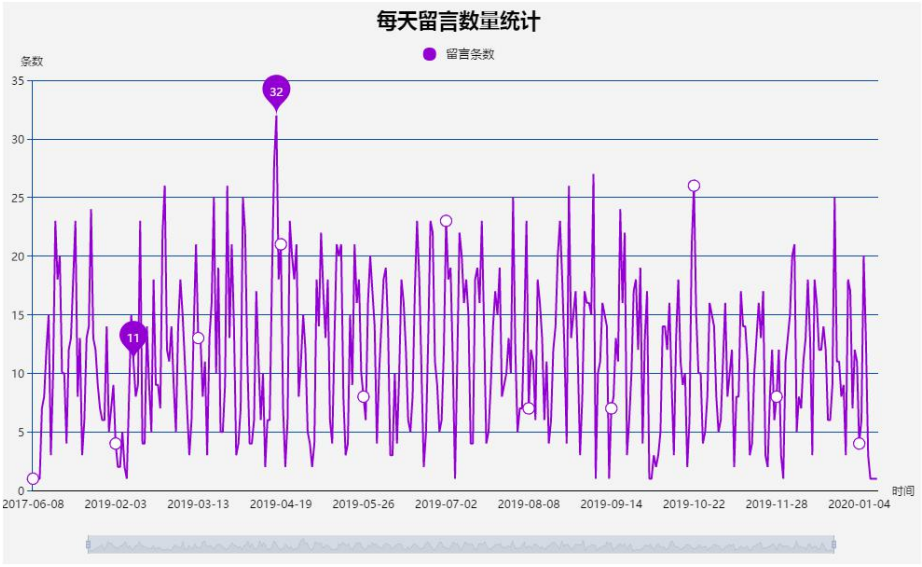


图 3.1-1：每日留言条数分布

数据集中涉及日期共计 379 天。从上图可以看出，样本数据主要集中在 2019 年度，单日留言数目峰值为 32 条，日均留言 11 条，整体趋势波动较大，区间为 [0,32]。

具体到一天的各个时间段，我们将时间划分为上午(6:00-12:00)、下午(12:01-18:00)、以及晚上（18:01-24:00,0:01-5:59）对群众集中网络留言的时间段规律进行发掘。并绘图如下：

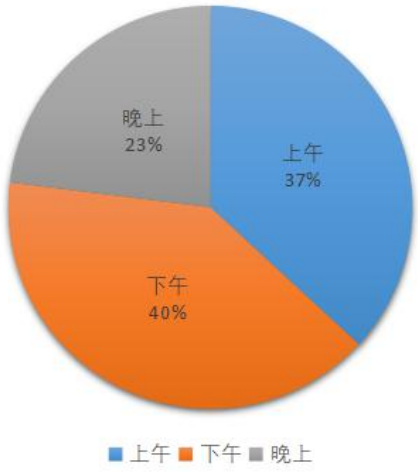


图 3.1-2：饼图展示了留言时间段分布

从饼图中可以看出，在下午时段是群众留言集中的时间段，占比高达 40%，其次是上午的 37%。尽管我们为晚上划定的时间长度为 12 小时，但是占比也仅为 23%。通过分析，我们认为，在群众较少留言的晚上如果存在集中的留言行为，发生突发性热点事件的可能性很大，是应当给予关注的。

3.1.2 留言支持与反对数分布情况

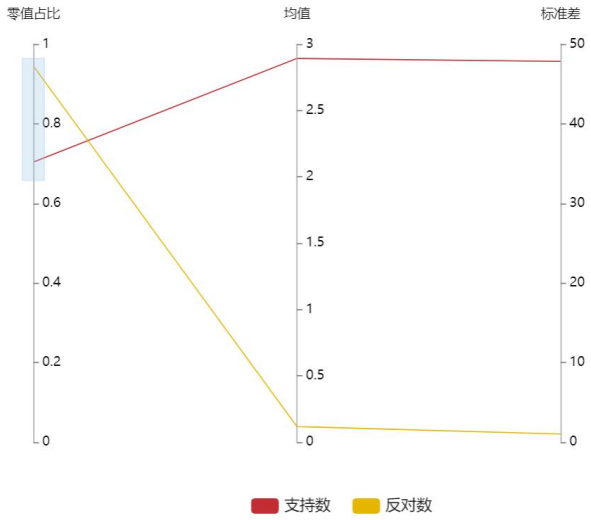


图 3.1-3：留言的反对与支持数的若干统计量

进一步了解，留言的支持和反对数对于留言的热度具有很重要的意义。如果某一个群众反映了大家都遇到的问题，那么群众大多数会选择点赞、反对等形式来表达自己相同的诉求或者相反的诉求。由平行坐标图可知，反对数为 0 的占比达 0.945 高于支持数 0 值占比 0.704，同时，反对数均值为 0.117 远低于支持数均值 2.892。但是标准差方面，支持数标准差为 47.827，远高于反对数的 1.028。说明了在网络政务留言平台上，群众并没有对其他问题进行支持或者反对的评价习惯（0 值占比很大），同时，相较而言，群众更倾向于支持已被反映的问题。

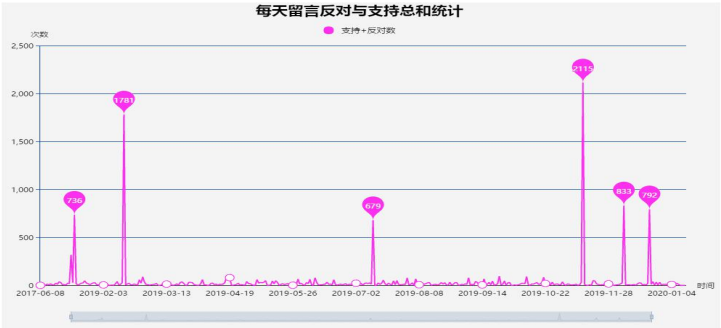


图 3.1-4：留言的反对与支持数（以天计算）

结合上图，从真实的数据集上反映出，数据集中存在几天，留言的支持和反对数之和表现异常的高，比如 2019 年 11 月 13 日，所有留言支持与反对总和高达 2115，但是由图 3.1-1，该天留言总数为 14 条，属于正常的均值水平，所以这一天的一些留言很有可能构成热点事件。

3.2 事件热度评价指标

给出事件的热度值是问题的要求，同时也是后续识别热点事件中的评价标准，本小节将围绕常见的事件热度评价指标，结合 3.1 中数据探索发掘的初步信息，给出我们的评价指标，并说明其合理性。

3.2.1 确定事件热度评价指标

借鉴舆情传播动力理论中常见的四种作用力（媒体影响力、网民作用力、政府疏导力和非常规突发事件作用力）中网民作用力和非常规突发事件作用力的概念，我们认为对于留言所反映的事件热度评价同样适用。具体分析如下：

（1）事件反映持续时间（ERD）：某一特定问题被反映的持续时间。对于时间段将其映射到[0,1]范围内，采用以下方法：

$$ERD = \frac{d}{1+d}\ ;\ d = TimeStamp_{start} - TimeStamp_{end} \qquad （式 3.2.1）$$

值得注意的是，在实际的留言平台上可能存在某一问题的多次反映，比如下水道疏通问题，可能在较长时间内多次因为拥堵而被群众反映，我们需要将每一次拥堵问题分割开来，分别计算 ERD。我们考虑虽然是同一类事件，但是相邻两次事件之间时间差距一般而言会比单一事件发生时的留言的时间间隔大。基于这一共识，在下节的热点事件提取方面，我们采用时间维度上聚类的方法，分离同类事件的多次发生带来的群众留言反映，分别计算 ERD。

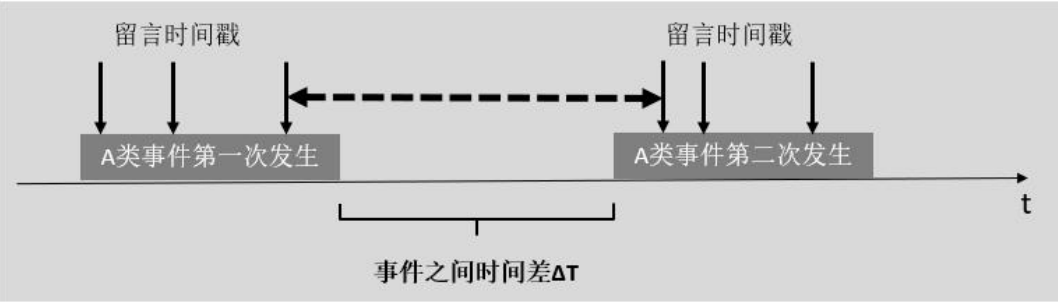


图 3.2-1：同类事件多发生情况下时间维度表征

（2）群众留言行为强度（MBS）：主要指对于某一特定问题，在一定时间段中，群众在网上问政平台上留言频次。

$$MBS_i = \frac{M_i}{M_{all}}(\text{特定时间窗口}) \qquad （式 3.2.2）$$

但是实际在操作的过程中，对时间窗口的选择存在很大的困难，而且如果时间反应的持续时间（也即上文提出的 ERD）大于时间窗口，该选择哪个时间窗口来计算 MBS 呢，这也是一个需要解决的问题。这里我们引入一种新的方法对 MBS 进行衡量：

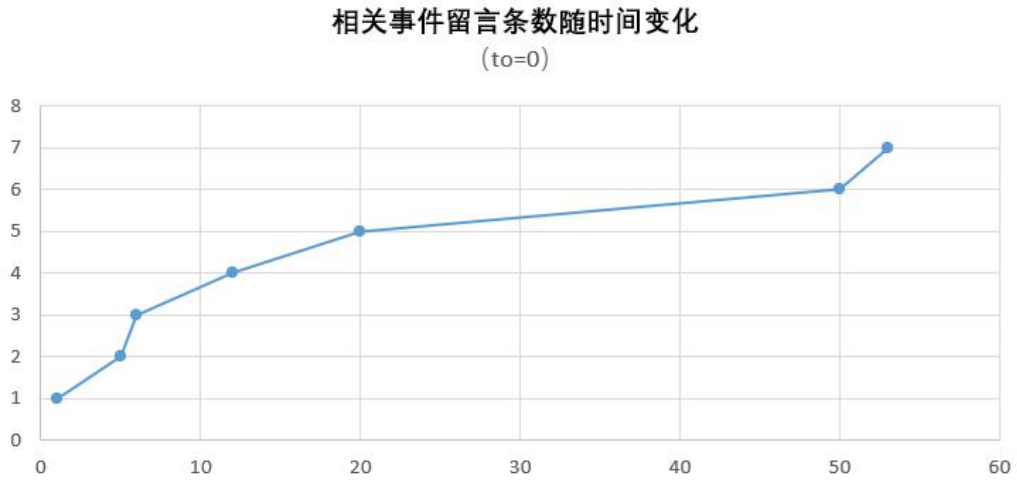


图 3.2-2：相关事件留言数量随时间变化折线图

在事件反映持续时间段内，统计每增加一条相关留言的时间长度，计算上图中每两个点之间的斜率。进一步的，计算斜率的均值（mean）和标准差（std），使用均值除以标准差，并进行归一化，得到 MBS 的值。

$$MBS = Normalized\left(\frac{\text{mean}}{\text{std}}\right) \qquad \text{（式 3.2.3）}$$

这里使用均值除以标准差而不使用一般均值的主要考虑是：减弱偶然情况下某两条留言的时间维度上的过分接近而对 MBS 计算过大带来的影响。

（3）当事人或物影响力（IOPO）：主要指涉及当事人或者单位的知名度、影响力等。在留言数据集中，事件主要分为涉及个人、涉及部分群体（如小区、教师群体等）、涉及所有人群（如道路交通瘫痪影响等）。这里我们为不同的事件涉及主题赋予不同权重，依次为 0.3，0.7 和 1。

（4）群众关注度（MA）：主要指群众对留言问题的关注程度，在这里使用反对数和支持数进行衡量。在 3.1 中已经探究过支持数和反对数（support and oppose，以下简称 S-O）的整体分布，0 值占据很大比例。所以更加凸显 S-O 值较大的留言文本所反映问题的重要性。在这里，我们不认为支持数和反对数权值存在差异，它们一同反应了问题的受关注程度。对于 MA 值的归一化，采用 Sigmoid 函数处理。

3.2.2 利用层次分析法对事件热度评价指标体系进行权重计算

对于各项指标赋予权重的方法有很多，如主观经验法、德尔菲加权法等。但是这些方法不可避免地引入了较大了主观性，精度难以保证，因而带来事件热度评价失衡的问题。这里我们采用层次分析法（AHP）对权重进行评价，这种方法的优点在于定性与定量结合，基于较少的定量信息使得赋权过程数学化。以下详细介绍层次分析法在本问题中的使用过程：

（1）基于 3.2.1 提出的三个指标，我们定性的认为群众留言行为强度（MBS）

的意义更加突出，其次是事件反映的持续时间（ERD）、最后是当事人或物影响力（IOPO）。这一点是符合逻辑的。采用 1-5 分标度法，为 IOPO 打分为 1，ERD 相对于 IOPO 打分为 2，MA 相对于 IOPO 打分为 3，MBS 相对于 IOPO 打分为 4，MA 相对于 ERD、MBS 相对于 ERD、MBS 相对于 MA 打分均为 2 分，由此构造判断矩阵如下：

	IOPO	ERD	MBS	MA
IOPO	1	0.50	0.25	0.33
ERD	2	1	0.50	0.5
MBS	4	2	1	2
MA	3	2	0.5	1

（2）计算权重值并进行一致性检验。基于以上判断矩阵，计算特征值，取最大特征值对应的特征向量，归一化后得到对应的权重向量：

$$\omega = （0.10,0.17,0.47,0.26）$$

考虑在定性构造判断的矩阵时，存在不一致的情况，比如 A 比 B 重要，B 比 C 重要，却又出现了 C 较 A 更重要的情况。因而必须进行一致性检验，一致性检验试验 CR 值进行计算，最终结果为 0.085 小于 0.1，故而通过一致性检验。

最终得出的留言问题热度值 $H = \omega \times （IOPO,ERD,MBS,MA）$。

3.3 热点问题挖掘

本节基于聚类的方法对热点事件进行挖掘，在这里我们讨论了多层聚类的方法，给出最终各个事件簇，并在各事件上使用 3.2 提出的事件热度评价指标，排序得到群众反映的热点问题。

3.3.1 第一层聚类：基于文本向量

使用 Bert 模型得到每个文本留言的向量，使用 K-means 进行第一次聚类，改变 K 值得到 SSE 随 K 值变化的趋势，以得到最佳的 K 值。

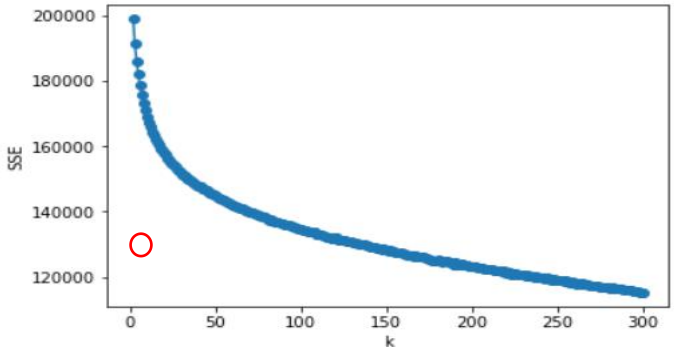


图 3.2-3：SSE-K 趋势图

由上图，最佳 K 值为 23，由此第一次聚类所得 23 个在留言内容上相近的类。通过词云图能够观察每个类的主题：

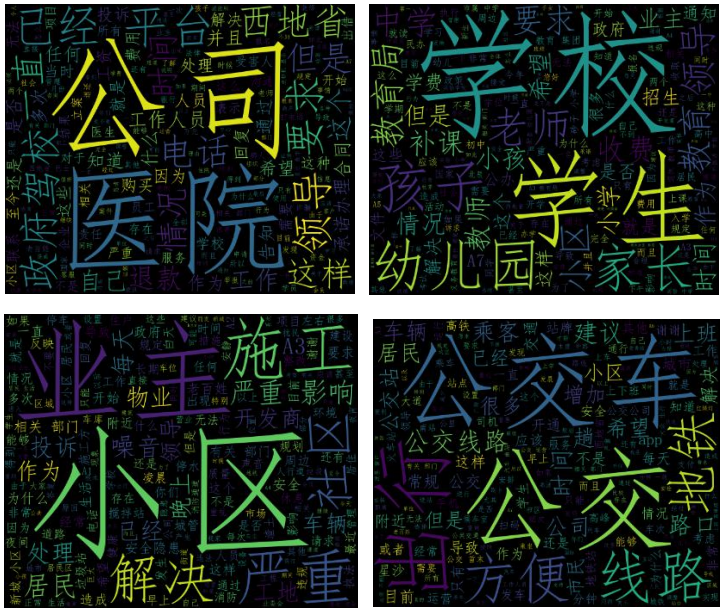


图 3.3-1：部分词云展示

词云图显示每个类有较为突出的主题，但是类与类之间仍然存在大量的共同关键词，类本身并不够纯净，通过 SSE 也能看出存在较大的误差。

再从时间维度看，以 1h 为时间窗口， 遍历整个数据集。在 1h 内，留言频次大多集中在小于 3 的范围，大于 4 条留言的时间段很少。

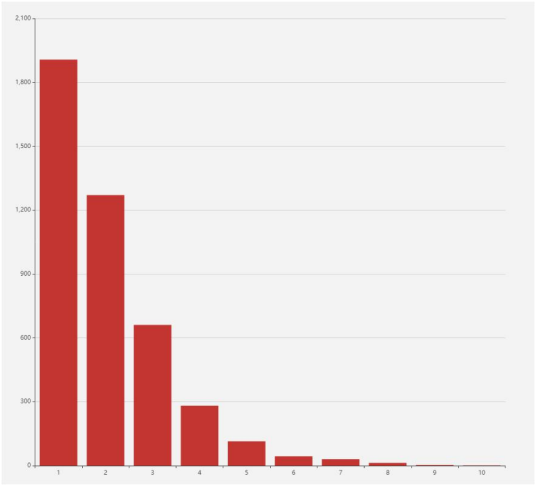


图 3.3-2：1h 时间窗口下的留言频次

这说明在时间维度上再进行一次聚类是必要的，这样可以将大多数留言文本变为离群点，逐步逼近真正的热点问题。

3.3.2 第二层聚类（基于时间维度）

对第一层聚类所得的各簇以时间维度进一步聚类，将原始数据中的时间格式转换为时间戳，这样时间将转换为一个一维空间上的点。我们期望得到的是时间

密集的同类事件被发掘，而稀疏的时间不该被重视，于是第二层聚类采用基于密度的 DBSCAN 方法，对一维数据进行聚类。下图描述了 DBSCAN 的流程：

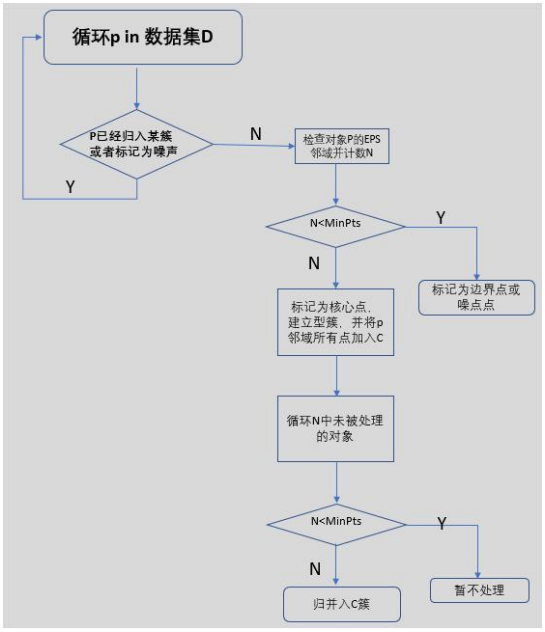


图 3.3-3：DBSCAN 算法流程图

在 DBSCAN 中存在很关键的两个参数，一者为 MinPts（邻域包含最小数目），二者为 Eps（给定的对象半径）。结合实际问题的实际意义，MinPts 的选择很简单，在 3.3.1 中，我们给出了时间维度上留言频次的分布，可以看出 MinPts 取值 4 或者 5 能过滤绝大多数的离群点。

关于 Eps 的选择问题，各簇必须是一致的。Eps 以小时计，观察同一个数据集上每个 Eps 值下的簇的个数，并绘制图如下：

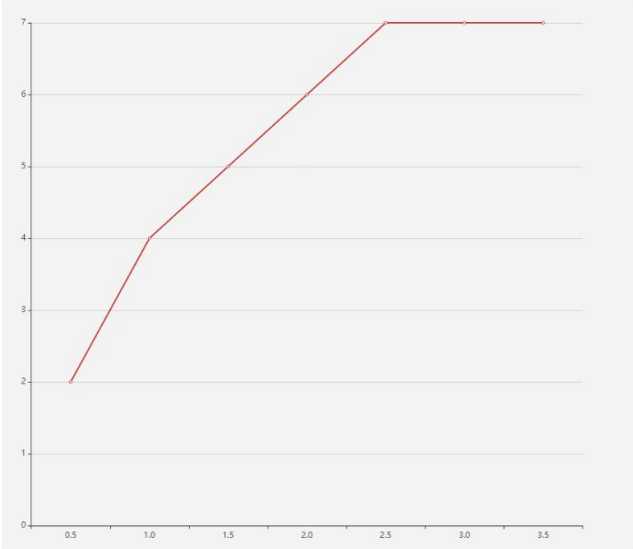


图 3.3-4：识别出的簇的个数随着 Eps 变大而变大

由上图可以看出，1h 是最好的 Eps，原因在于 Eps 不能过大而涵盖时间分散的区域，而 Eps=1h 时，曲线出现拐点，对应的平均簇的个数是 4。第一次聚类的其中一个类经过时间维度下的 DBSCAN 聚类结果如下：

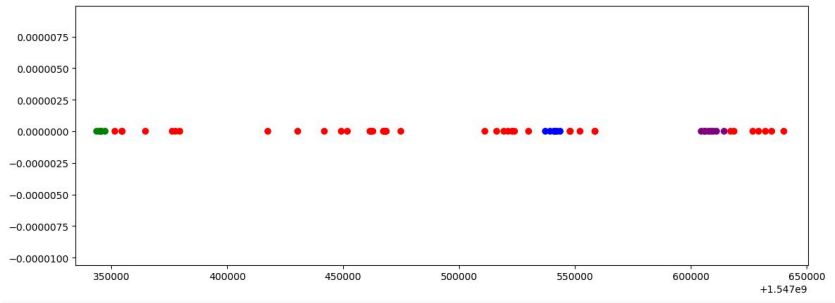


图 3.3-5：红色标记离群点，非红色标记密度高的簇（示例）

综上，两层聚类之后能够识别所有可能成为热点事件的留言簇，在留言簇上计算热度评价指标并排序，得到有最终 Top5 热点话题，如下：

ID	事件热度值	时间范围	事件内容
1	0.886	2019/1/2 至 2020/1/7	魅力之城小区餐馆油烟扰民
2	0.795	2019/3/28 至 2020/1/8	经济体育学院强制学生外出实习
3	0.781	2019/1/8 至 2020/1/6	江山帝景小区存在安全隐患
4	0.763	2019/3/28 至 2019/11/18	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
5	0.724	2019/6/12 至 2019/11/25	A 市保利麓谷林语小区旁 6 号地铁修建扰民

表 3-3-1 热点问题 Top5

3.4 热点问题的实体识别与主题抽取

本节主要介绍在热点问题确定之后对实体的识别以得到热点问题涉及到的人群和地点，以及对多留言文本进行主题提取以获取问题描述。

3.4.1 实体（地点/人群）识别

在 NLP 中，对于命名实体识别的任务，我们通常将之视为一种序列标注的任务，常用手段有传统机器学习模型，如基于隐马尔可夫模型标注和最大模式匹配方法的 HanLP；和深度神经网络模型，如以词向量为输入的 CNN、RNN 等神经网络模型。但是在实际的使用过程中，传统机器学习模型的弊端凸显，包括对于规则建立的苛刻要求、需要大量先验知识、识别准确率低、泛化能力弱等等，我们

更多选择的是深度神经网络模型。

本文主要采用 Bert-BiLSTM-CRF 的组合模型实现留言数据集上的中文实体识别。Bert 模型在这里的主要作用是基于其强大的文本特征表达能力提供输入文本向量，Bert 的基本原理在第二章已经介绍过，这里不再赘述。以下详细介绍 BiLSTM-CRF 模型：

（1）BiLSTM 模型：从循环神经网络（RNN）入手，RNN 实现了先前信息保留到当前任务中，这对于传统的深度神经网络而言是一次飞跃。但是 RNN 存在一个弊端——对长期信息保留能力差（也即无法解决好长期依赖的问题）。为此，引入了 LSTM（Long Short Term），LSTM 采用了细胞状态的概念，细胞状态就像是一条信息通道，能够将较为久远的信息保留下来。当然，对信息如何保留、更新等是基于 LSTM 内部的门控结构，这是一个很精妙的过程。LSTM 网络维持这细胞状态，确保其直接在链上运行，只存在与每个神经元模块极少的线性交互。

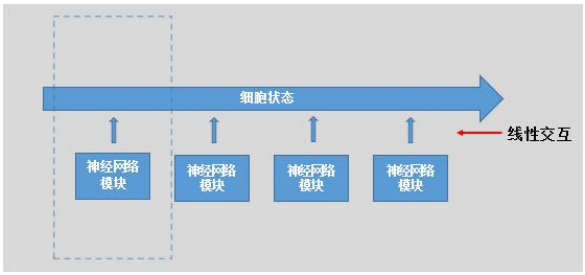


图 3.4-1：LSTM 架构图

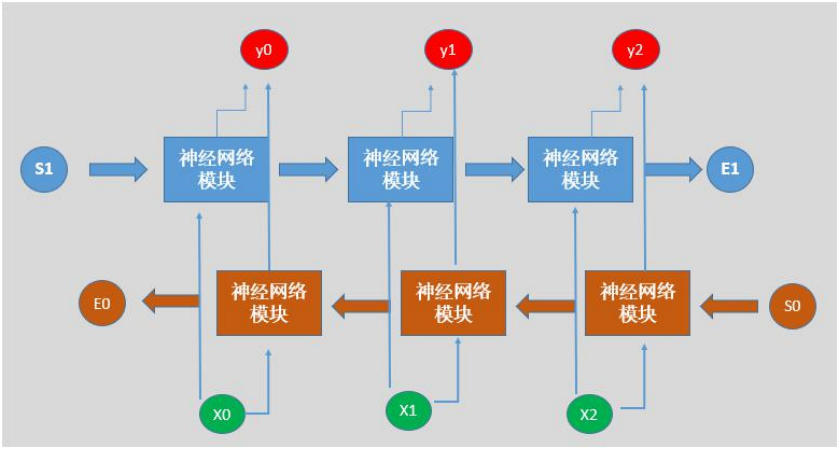


图 3.4-2：BiLSTM 架构图

LSTM 在多种 NLP 问题上表现良好，但是它与 RNN 网络一样，都只依赖之前时刻的信息来预测下一时刻的信息。但是当前的输出和过去、将来都有关联。于是就有了双向循环神经网络（BiLSTM）产生。BiLSTM 是双层 LSTM 的叠加，在对文本处理上可以分别看作是从句子开头开始输入和从句子末尾进行输入。

（2）CRF 模型：条件随机场是马尔科夫场的特例，区别在于 CRF 可以定义数量更多、种类更加丰富的特征函数、且可以使用任意的权重值，最重要的是，CRF 计算的是全局最优解，而非局部最优解，这将大大提高其在序列标注上的表现。

融合 Bert 对文本特征表达、BiLSTM 学习词语上下文信息的能力以及 CRF 基

于全局信息进行标注的特点，构建 Bert-BiLSTM-CRF 模型结构如下:

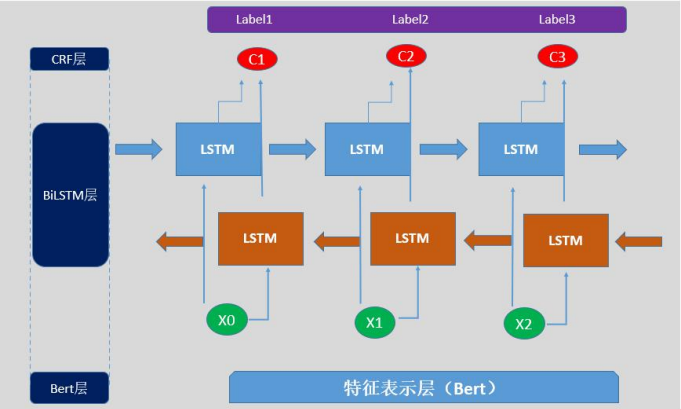


图 3.4-3：Bert-BiLSTM-CRF 多层模型结构

本文采用的 NER 训练集来自于 MSRA、人民日报和 Boson，语料集合大小分别为 46365 条、23061 条和 2000 条。

NER 基于文本数据是将前述一个热点簇中的留言详情进行合并得到。实际操作过程中，由于数据集本身的脱敏操作导致地名失真，如 A 市、A7 县等。对于这类地名前缀采用匹配的方法得到。采用以下正则匹配表达式：

[a-zA-Z][0-9]*[\u4e00-\u9fa5]{1}　　（式 3.4.1）

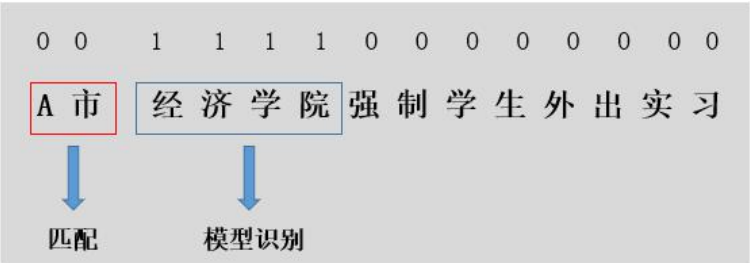


图 3.4-4：实体识别结果实例

关于群体的识别问题，现有的 NER 方法只能提取机构、地名、人名，对于群体识别并没有很好的解决方案。立足问题要求，我们提出了一种基于实体相似性匹配的方法：

图 3.4-5：联想举例

基于前述提取的机构名、地名构建专家数据集，如 3.4-5 图所示。在得到一个新的机构名称或者地点名时，根据实体之间相似性进行划分，将待映射实体对应到距离最近的实体所关联的群体。对于一个留言文本中所有提取到的实体，均

进行上述操作，得到一个针对该留言的联系群体列表，取出现次数最多即可。举例如下：

留言详情：

桐梓坡 589 号白鹤咀停车场，由聚美龙楚新能源公司建的“商学院新能源汽车充电站项目”。在没有任何告知的情况下，在商学院宿舍 5 栋外墙安装 1000 千伏安变压器。距离商学院宿舍 5 栋不到 5 米……大学生、居民，带来极大的电辐射污染、噪音污染、及消防隐患。请求进行拆除。

联系群体列表：[居民，职工，学生，学生,学生]

最终确定涉及群体：学生

```
>>>NER结果:
桐梓坡589号白鹤咀停车场 聚美龙楚新能源公司 商学院 商学院 商学院
>>>联想结果:
['居民', '职工', '学生', '学生', '学生']
Counter({'学生': 3, '居民': 1, '职工': 1})

Process finished with exit code 0
```

图 3.4-6：示例结果

3.4.2 问题描述构成

在确定了一个热点事件簇之后，我们需要得到事件的描述信息。留言主题本身提供了很好问题描述，概括性强。如下图实例：

魅力之城小区临街门面油烟直排扰民
A5区劳动东路魅力之城小区油烟扰民
A5区劳动东路魅力之城小区底层餐馆油烟扰民
A5区劳动东路魅力之城小区临街门面烧烤夜宵摊
A市魅力之城商铺无排烟管道，小区内到处油烟味
A5区魅力之城小区一楼被搞成商业门面，噪音扰民严重
A市魅力之城小区底层商铺营业到凌晨，各种噪音好痛苦
A5区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气
A5区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气，急需处理！
万科魅力之城小区底层门店深夜经营，各种噪音扰民

图 3.4-6：留言主题用作问题描述

但是还存在一些留言主题信息极为简略，如“环境污染”、“A 市橘郡开发商去哪了？”等，无法完整表达问题，在从多个留言中提取问题描述时，必须进一步的处理。这里我们使用 TextRank 从留言详情中提取关键句以补充。对于最后的问题描述的给定，由 NER 识别的实体名称、留言数据本身的留言主题以及关键句共同确定。

当然大多数的热点问题不必如此，如一条留言构成的热点事件，其留言主题即可作为问题描述。

综上，给出热点问题的地点/人群要素以及问题描述。

ID	事件热度值	时间范围	事件内容	地点/人群	问题描述

1	0.886	2019/1/2 至 2020/1/7	魅力之城小区餐馆油烟扰民	魅力之城小区/居民	魅力之城小区餐馆油烟扰民
2	0.795	2019/3/28 至 2020/1/8	经济体育学院强制学生外出实习	经济体育学院/学生	经济体育学院强制学生实习
3	0.781	2019/1/8 至 2020/1/6	江山帝景小区存在安全隐患	江山帝景小区/居民	江山帝景小区存在安全隐患
4	0.763	2019/3/28 至 2019/11/18	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	A 市 A5 区汇金路五矿万境 K9 县/居民	A 汇金路五矿万境 K9 县存在系列问题
5	0.724	2019/6/12 至 2019/11/25	A 市保利麓谷林语小区旁 6 号地铁修建扰民	保利麓谷林语小区旁 6 号地铁线/居民	保利麓谷林语小区旁 6 号地铁修建扰民

表 3-3-2：热点事件详细信息

3.5 总结

本章针对问题二，首先对附件 3 数据集进行了统计分析，具体是留言时间分布、留言支持数与反对数分布情况进行了探索，对与可能成为热度评价指标的属性特征有了深入理解；对于评价指标的构建，我们根据舆情传播的影响因素，给出了我们的评价指标，主要包括事件反映持续时间（ERD）、群众留言行为强度（MBS）、当事人或物影响力（IOPO）以及群众关注度（MA），使用层次分析法结合定量与定性的方法完成各个指标权值计算；对于热点事件的确定，我们采用了多层聚类手段，首先第一层对文本向量聚类以发现主题相似的类，最佳簇的数目通过 SSE-k 图发现。在第一层聚类的基础上，对时间维度使用 DBSCAN 进行基于密度的聚类，以发现在时间维度上较为聚集的同类事件，结合数据集特征给出了 MinPts 和 Eps 的最优参数。如此得到热点事件并取 Top5，并获得热点问题明细表（参见**附件热点问题明细表.xlsx**）；进一步地，针对热点问题中 NER 提取，采用了 Bert+BiLSTM+CRF 的模型结构，在数据集上取得较优表现，对于数据脱敏带来的地点前缀失真采用匹配的方式，进而得到完整的命名实体（参加附件 2）。

对于群体的识别，构建联想专家集（参加附件 3），根据 NER 中的机构名称和地点名称进行群体联想。问题描述采用留言主题+详情摘要+NER 结果提供综合参考的方法，得出较为准确的、要素齐全的描述信息（参见**附件热点问题表.xlsx**）。

第四章 答复意见评价方案

本章针对问题三，在附件四的基础上，对留言答复意见进行评价，主要包括相关性、及时性、完整性以及可解释性等。主要包括语义相似度计算衡量相关性、时间差衡量及时性、答复文本长度等衡量完整性，以及一些其他指标。

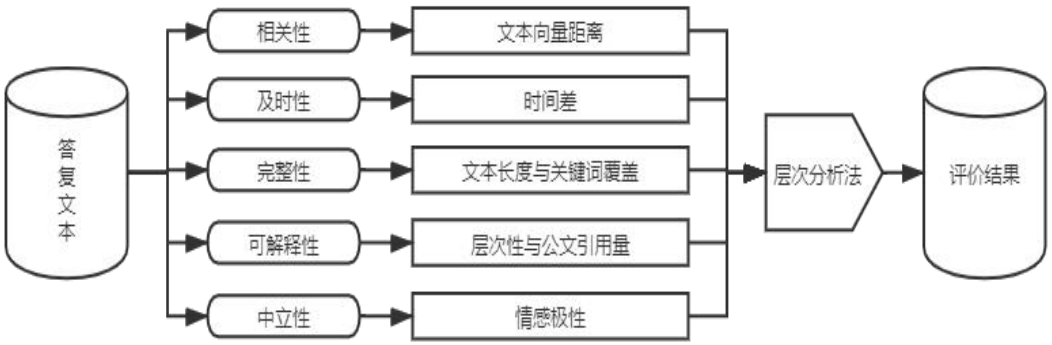


图 4.0-0：本章架构图

4.1 相关性衡量

答复意见的相关性评价是指政府部门回答群众问题时是否主题相关、紧扣群众问题，相关性不强的答复容易对民意造成不良影响。对于留言与文本的相关性，我们很自然地想到使用文本向量距离进行衡量。在问题一留言分类的任务中，我们使用 Bert 模型对文本进行编码取得了很好的效果，这里我们继续沿用这一编码方式。

但是，在政府相关部门的答复中本身会出现一些与主题无关的句子，如“xxx 网友您好，您的留言已收悉。现将有关情况回复如下：……”、“感谢您对我们工作的支持、理解与监督！”等等，当然，群众留言中也存在这类问题，这些非主题句的存在会大大降低问答之间相关性度量的准确性。因此，这里采用 TextRank 提取关键句作为问答文本，进而使用 Bert 模型获取文本向量。

```
网友“UU00812”您好！您的留言已收悉。现将有关情况回复如下：一、关于小区附近幼儿园的问题。经A5区教育局、黎托街道核实，东澜湾幼儿园属于江河东澜湾小区配套幼儿园，由A市江河水利置业投资发展有限公司开发。该园主体建筑于2018年11月建设完工，建筑面积2240平方米，规划9个班，根据《A市A5区城镇小区配套幼儿园专项整治工作方案》文件精神，A5区教育局和A市江河水利置业投资发展有限公司于2018年10月19日签署配套幼儿园移交协议。根据协议规定，该园整体移交给A5区教育局（整体建设完工后办成公立幼儿园或普惠制幼儿园）。由于该园内部功能布局规划方案尚在审批当中，未达到开园要求。区教育局已启动二次装修前期准备及设施设备采购等工作。下一步，街道将督促开发商尽快完成内部功能布局建设，尽快竣工验收投入使用。二、关于东澜湾社区医疗卫生服务的问题。经A5区卫计局核实，2018年6月在黎托街道粟塘社区综合楼开设黎托街道基成社区卫生服务中心（健康热线0731-0000-00000000），距离东澜湾小区4公里左右，中心可为黎托街道及周边辖区居民提供基本公共卫生服务和基本医疗服务。距东澜湾小区1公里左右设有花桥村卫生室、花桥门诊部、沙湾门诊部等医疗机构，可提供医疗服务，建议可前往以上卫生服务机构就医。此外，A5区政府在劳动路与红旗路交汇处西南角规划再建设一所社区卫生中心，目前正在拆迁腾地。感谢您对我们工作的支持、理解与监督！2019年1月15日
摘要:
[{'index': 4, 'sentence': '该园主体建筑于2018年11月建设完工，建筑面积2240平方米，规划9个班，根据《A市A5区城镇小区配套幼儿园专项整治工作方案》文件精神，A5区教育局和A市江河水利置业投资发展有限公司于2018年10月19日签署配套幼儿园移交协议',
'weight': 0.12470705025583519}, {'index': 3, 'sentence': '经A5区教育局、黎托街道核实，东澜湾幼儿园属于江河东澜湾小区配套幼儿园，由A市江河水利置业投资发展有限公司开发', 'weight': 0.10133694260191232}, {'index': 10, 'sentence': '经A5区卫计局核实，2018年6月在黎托街道粟塘社区综合楼开设黎托街道基成社区卫生服务中心（健康热线0731-0000-00000000），距离东澜湾小区4公里左右，中心可为黎托街道及周边辖区居民提供基本公共卫生服务和基本医疗服务', 'weight': 0.09320303883232246}]
```

图 4.1-1：TextRank 提取关键句效果示意

为了说明，问答之间相关性是具有考量意义的，在这里我们对比问答文本向量距离和随机选取留言和答复之间距离的大小，统计结果如下：

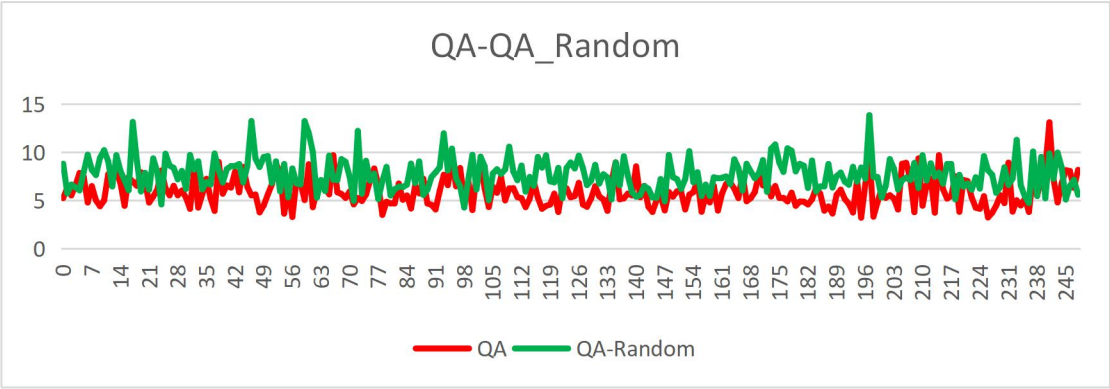


图 4.1-2：对应问答文本向量距离与随机问答文本向量距离对比（局部）

由图 4.1-2 可以看出，对应的问答文本向量之间距离分布曲线（红色）绝大多数情况下低于随机抽取一问一答文本向量，说明对应问答文本向量之间具有值得注意的相关性，充分证明了相关性由文本向量距离衡量的合理性。

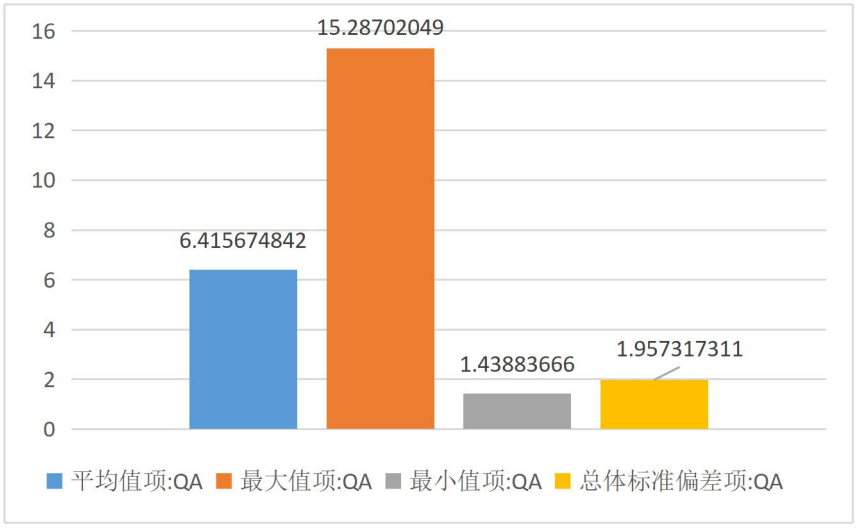


图 4.1-3：问答文本向量距离分布情况

观察较大文本向量距离对应答复情况，发现答复文本多为“反映的问题已转交相关单位调查处置”一类确实与群众留言无关的文本。同时标准差为 1.9 左右，我们采用 4 分位法，对于在[1.43,15.29]之间划分出 4 个档次，以 10 分计，分别打分为 10，8，6，5，出现在最大最小值以外的分别计分 4,10。

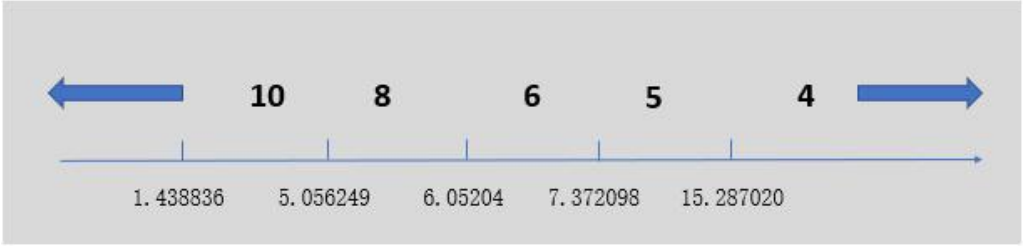


图 4.1-4：相关性计分图示

4.2 及时性衡量

答复及时性评价是指自群众提出问题后，政府相关部门给出的答复的时间差。能在群众可接受时间范围内给出答复是影响群众满意度的重要因素。

附件 4 给出的数据集，绘制答复时效分布图，可以看出大多数答复在群众提出后半年内给出答复，且大多集中在 1 到 3 个月，时间差较大。

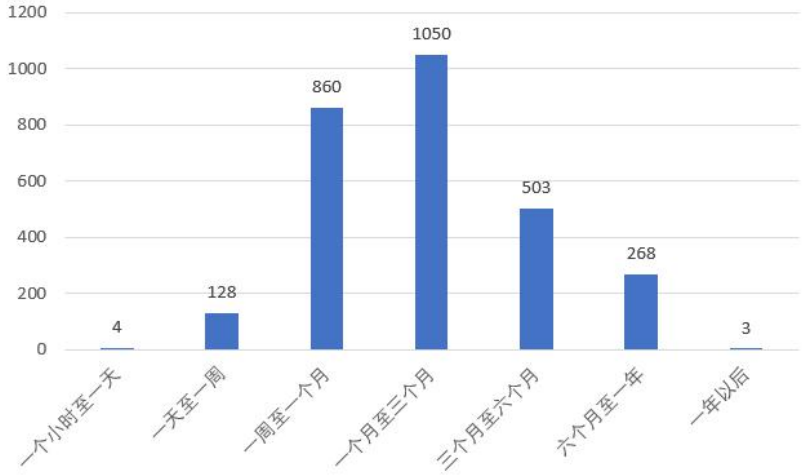


图 4.2-1：答复时间差分布情况

考虑时间离散的分布特点，我们采用分段评分方式，以 10 分计，一天以内回复为满分，一天到一周时间以内答复为 9 分，以此类推，完成对及时性的评价。

4.3 完整性衡量

一般而言，文本完整性体现为结构的完整性和语义的完整性。但是在问答文本中，完整性主要表现在是否完整回答了群众问题。这一点在算法层面很难直接准确判断。但是我们可以从宏观层面把握，比如答复涵盖群众问题中关键词的比例、答复文本长度等。

观察一般对网上群众留言的答复文本，可以看出，答复中一般会概述留言群众反映问题，并以此展开，阐述政府规划、当前政策、目前状况等等。在文本中回答群众问题同样也会不可避免地使用到反映问题的关键词。这种答复模式从语言上具有亲切感，也为我们从关键词覆盖率中评价一个答复质量提供了支撑。

在对关键词的提取上，我们使用 TF-IDF 算法，在第二章数据探索中，我们也使用了相同的手段。TF-IDF 的核心思想是：如果某个字词在一篇文章中出现的频率（使用 TF 值衡量）较高，且在其他文章中很少出现（使用 IDF 衡量），则认为此字词具有很好的类别区分能力，适合作为关键词。TF 与 IDF 的计算公式如下：

$$TF = \frac{\text{在某一类中字词出现的次数}}{\text{该类中所有字词的数目}} \qquad \text{（式 4.3.1）}$$

$$IDF = log\left(\frac{\text{语料库文档总数}}{\text{包含该字词的文档数+1}}\right) \quad \text{（式 4.3.2）}$$

语料库由附件 4 提供的所有留言数据构造，由于留言文本很显然不属于一个类，可以获得质量较高的关键词。

依据覆盖率为答复评分，得到下列结果分布（10 分计）。可以看出以这种方式衡量，答复得分一般在 5 分以下，集中分布于[2,3]。

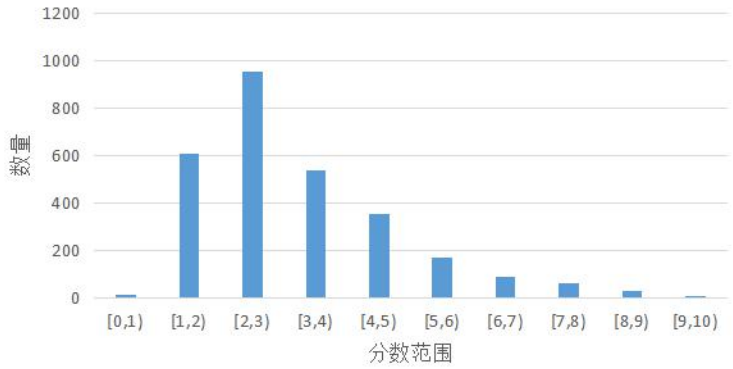


图 4.3-1：评分结果分布图

另一个比较直观的衡量标准是答复文本长度。长度很短的答复通常情况下不认为是一个完整的答复。观察下图：

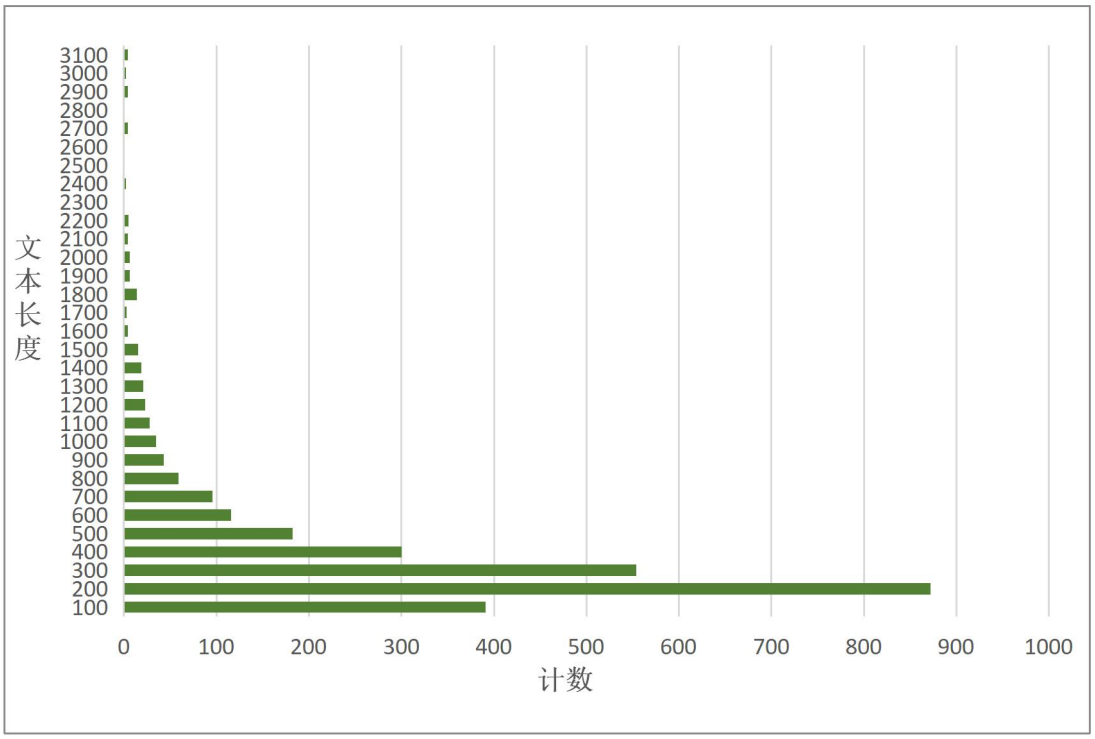


图 4.3-2：附加数据集答复文本长度分布

答复文本长度多集中在 500 字以内，最长可到 3000 余字。过长的答复并不一定代表答复质量较高，可能会大量引用政策、法规等信息，对群众造成一定程度上的困扰。同样的，过短的答复就有可能是涵盖信息不足。如下图示例：

请问领导，农合费用增加了，打狂犬役苗报销比例是多少。盼回音。先谢了！	已收悉

图 4.3-3：答复文本过短而不完整（示例：留言编号 6556）

同时，答复的文本长度在宏观上还取决于留言文本长度，下图展示了附件数据集中答复与留言文本长度之比：

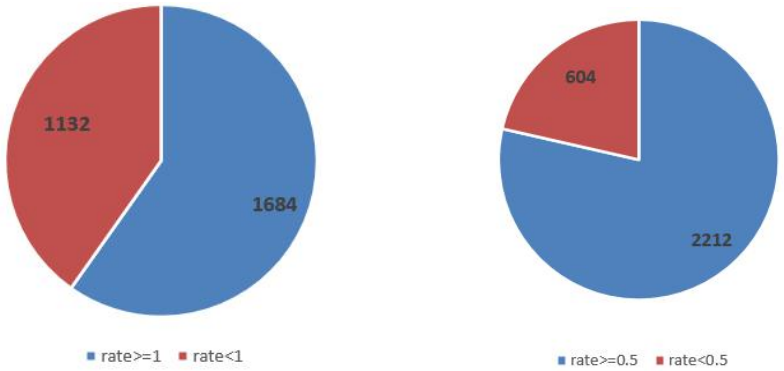


图 4.3-4：答复与留言文本长度比例

（左图以比率 1 分割，右图以比率 0.5 分割）

上图说明，约 60%的答复文本长度要长于对应的留言，78%的答复文本长于留言文本的一半。多次改变 rate 的阈值，发现答复文本满足答复与留言文本长度比例大于 rate 的数量随 rate 呈线性变化，不存在拐点。

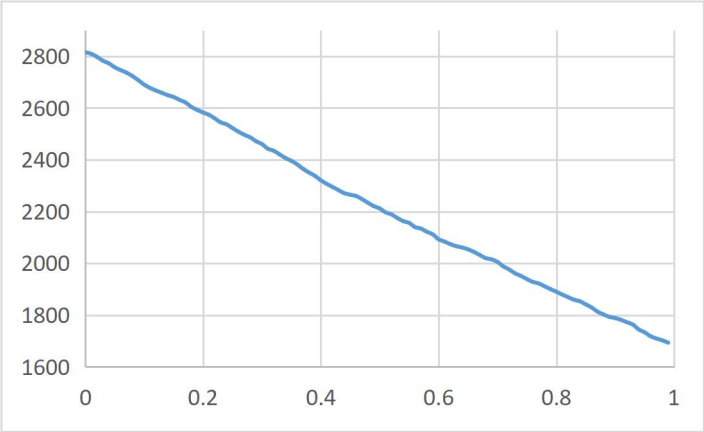


图 4.3-5：答复与留言文本长度比例大于 rate（横轴）的答复文本数量

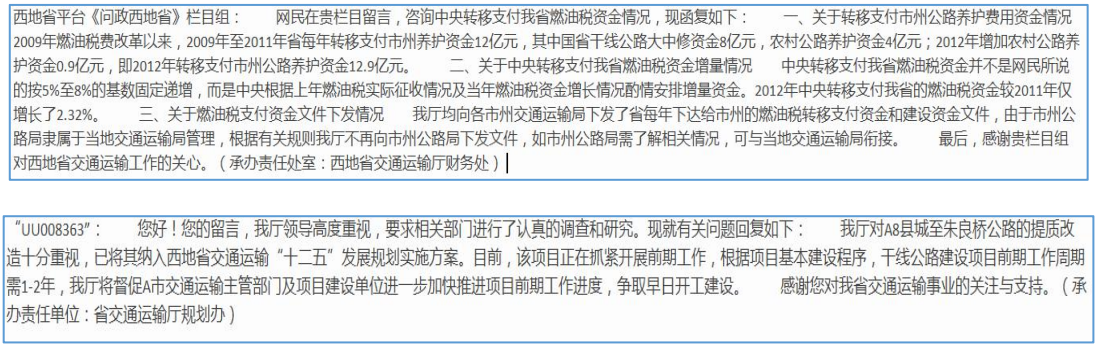
因此在这里我们不存在一个所谓最好的 rate 值来作为评价标准。这时我们给出以下约定：

$$Score = \begin{cases} 9 & rate \geq 10 \\ 10 & 10 > rate \geq 1 \\ 9 & 1 > rate \geq 0.9 \\ \dots & \dots \\ 1 & 0.2 > rate \geq 0.1 \end{cases} \qquad \text{（式 4.3.3）}$$

4.4 可解释性衡量

可解释性指的是答复是否条理明晰、简单易懂。抽象的特性在算法实现层面也很困难。但是我们同样可以进行转换，比如条理清晰的答复体现在要点明确、层次性强；简单易懂体现在政策、法规条文引用少，不“假大空”等。

因此问题转换为对文章层次性衡量和法规政策条文引用识别问题。


西地省平台《问政西地省》栏目组：　网民在贵栏目留言，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，2009年至2011年省每年转移支付市州养护资金12亿元，其中国省干线公路大中修资金8亿元，农村公路养护资金4亿元；2012年增加农村公路养护资金0.9亿元，即2012年转移支付市州公路养护资金12.9亿元。　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　我厅均向各市州交通运输局下发了省每年下达给市州的燃油税转移支付资金和建设资金文件，由于市州公路局隶属于当地交通运输局管理，根据有关规则我厅不再向市州公路局下发文件，如市州公路局需了解相关情况，可与当地交通运输局衔接。　最后，感谢贵栏目组对西地省交通运输工作的关心。（承办责任处室：西地省交通运输厅财务处）

“UU008363”：　您好！您的留言，我厅领导高度重视，要求相关部门进行了认真的调查和研究。现就有关问题回复如下：　我厅对A8县城至朱良桥公路的提质改造十分重视，已将其纳入西地省交通运输“十二五”发展规划实施方案。目前，该项目正在抓紧开展前期工作，根据项目基本建设程序，干线公路建设项目前期工作周期需1-2年，我厅将督促A市交通运输主管部门及项目建设单位进一步加快推进项目前期工作进度，争取早日开工建设。　感谢您对我省交通运输事业的关注与支持。（承办责任单位：省交通运输厅规划办）

图 4.4-1：层次性强弱对比的文本示例

明显可以看出，图 4.4-1 中上图文本的层次性好于下图。上图显示的答复质量也明显高于下图文本示例。在实际实现的过程中，我们将匹配层次编号，如“1.2.3.”、“一、二、三”等作为主要手段，以衡量层次性。

同时，对于答复中对政策文件、法律条文进行识别，得到引用公文数量占比答复长度比例，由一批人工筛选的质量较高的答复得出最佳比例，以此从一个层面衡量答复是否存在晦涩难懂的问题。

由于赛题提供数据集规模局限和人工标注数据集的代价，这里仅给出可解释性度量方案，不做具体实现。

4.5 其他指标与指标综合评价

政府相关部门回答群众问题时必须保持中立，客观的态度，因此对答复的情感倾向进行评价也是必不可少的。常见的情感倾向评价主要有两种方式，一种是基于情感词汇库进行匹配，另一种是使用深度学习手段将之转换为一种分类问题。使用情感词汇库进行匹配的优点是速度快，实现简单，缺点是情感极性识别依赖于情感词汇库的质量；深度学习进行情感极性分类的优点是准确率高，缺点是训练集标注工作量大，尤其在针对新的领域文本时，同时模型训练时间和硬件平台要求较高。

对于以上多个指标，我们可以不做融合直接从各个角度为答复文本质量进行评价。当然，我们也可以采用层次分析法等对指标进行赋权融合，给出一个综合性的指标，如下图：

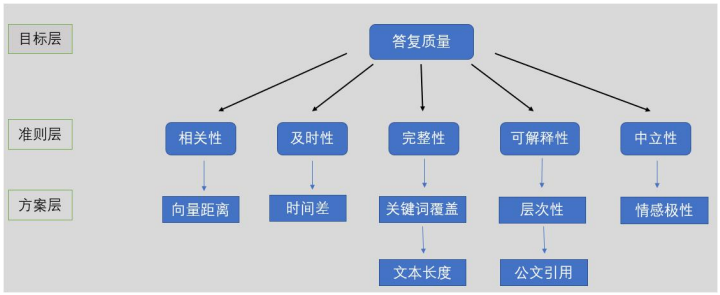


图 4.5-1：层次分析结构

采用定性与定量相结合的手段，我们给出判断矩阵如下：

$$\begin{array}{c} \\ D \\ T \\ C \\ L \\ A \\ Q \\ F \end{array}\begin{bmatrix} 1 & 0.5 & 1 & 1 & 0.5 & 1 & 0.5 \\ 2 & 1 & 2 & 2 & 2 & 2 & 1 \\ 1 & 0.5 & 1 & 0.5 & 1 & 1 & 0.5 \\ 1 & 0.5 & 2 & 1 & 1 & 1 & 0.5 \\ 2 & 0.5 & 1 & 1 & 1 & 1 & 0.5 \\ 1 & 0.5 & 1 & 1 & 1 & 1 & 0.5 \\ 2 & 1 & 2 & 2 & 2 & 2 & 1 \end{bmatrix}$$

符号	含义
D	问答文本向量距离评分
T	问答时间差评分
C	关键词涵盖评分
L	答复文本长度评分
A	答复文本层次性评分
Q	答复文本公文引用量评分
F	答复文本情感极性

表 4-5-1：矩阵符号说明表

```
最大的特征值：7.1002
对应的特征向量为：[-0.2529 -0.5495 -0.2529 -0.3104 -0.3104 -0.2748 -0.5495]
归一化后得到权重向量：[0.1011 0.2198 0.1011 0.1241 0.1241 0.1099 0.2198]
判断矩阵的CI值为0.0167
判断矩阵的RI值为1.32
判断矩阵的CR值为0.0127，通过一致性检验
```

图 4.5-2：判断矩阵通过一致性检验

经过计算，判断矩阵通过一致性检验且各个指标权重为：

$$W = (0.1011,0.2198,0.1011,0.1241,0.1241,0.1099,0.2198)$$

4.6 总结

本章针对问题三，立足于附件中提供的留言答复文本，围绕问答相关性、回答及时性、完整性、文本可解释性等提出了一系列切实可行的评价指标，并根据

实际情况提出了一些其他指标。其中绝大多数评价指标给出了实现方法。各个指标评价数据见附件 4。同时我们使用层次分析法，在构建现有指标判断矩阵的基础上，确定了各个指标的权重，并验证了其合理性。

参考文献

[1]王子牛，姜猛，高建瓴，陈娅先.基于 BERT 的中文命名实体识别方法[J].计算机科学，2019，46（11A）：138-142.

[2]赵舒贞.网络舆情预警指标体系的建构[D].云南：云南师范大学，2013.

[3]邹晴.多分类器融合的文本分类研究[D].湖北：湖北工业大学，2015

[4] zengyy.使用 Bert 生成句向量

[EB/OL].https://gitee.com/zengyy8/bert-utils

[5]于韬，王洪岩.基于 TF-IDF 算法的文本信息提取[J].科技视界，2018,(16):117-119.