基于 *ERNIE* 模型的智慧政务文本数据挖掘

摘要

随着“放管服”改革的深入发展，智慧政务理念应运而生，政府管理与公共服务也逐渐实现数字化、社会化和智能化。随之发展而来的网络问政平台的使用，使得民意获取愈加便捷与广泛，然而依此获得的文本数据之庞大仅依靠人工处理无法有效实现，这迫切需要自然语言处理（*NLP*）技术的推动。本文基于大赛提供的数据建立文本库，主要借助 *Python* 软件实现数据的描述和预处理、并构建 *ERNIE* 模型、*Re-Single* 模型以及 *TOPSIS* 评价方法进行相关任务的解决。

针对任务一，首先对附件 2 中的 9210 条留言信息进行数据描述与预处理，将句子最长长度设定确定为 200 字，并将留言详情文本向量化，同时发现此题可不考虑数据倾斜问题。以此为基础，采取 5 折交叉验证（训练集：验证集：测试集的数据量比为 6:2:2），使用 *ERNIE*1.0 模型对留言内容进行分类。分析结果表明该分类器对应 7 个类别预测结果的 *F*1 值全部在 85%以上，且预测的总体平均 *F*1 值为 91.24%，说明 *ERNIE*1.0 模型可以较好的完成对网络问政平台的群众留言主题分类的工作。

针对任务二，首先构建热度评价指标，综合考虑影响力（相关留言单位时间内的点赞数和反对数）和关注度（留言主题单位时间内出现的次数）来计算相关问题的热度指数。其次根据留言主题进行地点分类，并集成句向量间的余弦距离，在每一个地点类的内部使用 *Re-Single* 模型实现文本聚类，基于聚类结果计算每一类的热度指数并提取排名前五的热点问题，分析发现最受市民关注的主题是城乡建设、科技与信息产业和交通运输三个大类，第一类涵盖安全生产、城市建设和市政管理及住房保障与房地产，第二类主要是信息化建设问题，第三类包括规划、管理等交通建设问题。

针对任务三，首先构建留言答复质量评价方案，评价指标有相关性、结构完整性、内容完整性、可解释性、时效性和可读性。随后利用 *TOPSIS* 法综合了各个指标，对留言答复文本进行评分。评分结果采用百分制，其中约 87%留言答复获得 60 分以上，获得 80-90 分的留言答复占比最大，说明政府部门在网络政务平台的给出的留言答复质量较好，但高分段较少低分段较多的现象说明政府部门仍需深入了解市民需求，以进一步实现答复的高质量与高效率。

在文末，小组对本次数据挖掘分析研究进行了总结与后期的改进展望。

关键词：政务文本挖掘 *ERNIE* 模型 集成句向量 *Re-Single* 模型 *TOPSIS* 法

Smart government text data mining based on ERNIE model

Abstract

With the in-depth development of the "decentralization management" reform, the concept of smart government affairs came into being, and government management and public services have gradually been digitized, socialized, and intelligent. The use of the network questioning platform that has evolved has made the acquisition of public opinion more convenient and extensive. However, the huge amount of text data obtained based on this cannot be achieved effectively only by manual processing, which urgently needs the promotion of natural language processing (NLP). This article builds a text library based on the data provided by the contest, mainly using Python software to describe and preprocess the data, and build the ERNIE model, Re-Single model and TOPSIS evaluation method to solve related tasks.

For task one, first of all, the data description and pre-processing of the 9,210 message messages in Attachment 2 are made, the maximum length of the sentence is set to 200 words, and the text of the message details is vectorized. . Based on this, take 5-fold cross-validation (training set: verification set: test set data volume ratio is 6: 2: 2), and use the ERNIE1.0 model to classify the message content. The analysis results show that the F1 values of the prediction results corresponding to the 7 categories of the classifier are all above 85%, and the overall average F1 value of the prediction is 91.24%, indicating that the ERNIE1.0 model can better complete the message to the online questioning platform Topic classification work.

For task two, first build a heat evaluation index, comprehensively considering the influence (the number of likes and objections in the relevant message unit time) and the attention (the number of times the message topic unit time) to calculate the heat index of the relevant question. Secondly, classify the locations according to the subject of the message, and integrate the cosine distance between the sentence vectors, use the Re-Single model to implement text clustering within each location class, calculate the heat index of each class based on the clustering results and extract the top five Analysis, and found that the topics most concerned by the public are urban and rural construction,

technology and information industry, and transportation. The first category covers safety production, urban construction and municipal management, housing security and real estate, and the second category mainly It is an information construction problem. The third category includes planning, management and other transportation construction problems.

For task three, first build a message reply quality evaluation program, the evaluation indicators have relevance, structural integrity, content integrity, interpretability, timeliness and readability. Subsequently, the TOPSIS method was used to synthesize various indicators and score the text of the message reply. The scoring results are based on a 100-point scale. About 87% of the message replies received more than 60 points, and the message replies with 80-90 points accounted for the largest proportion. The phenomenon of low-level and low-level segmentation indicates that government departments still need to understand the needs of citizens in order to further achieve high-quality and high-efficiency responses.

At the end of the article, the team summarized this data mining analysis and research and looked forward to future improvements.

Key words: Government text mining ERNIE model Integrated sentence vector Re-Single model TOPSIS method

目录

一、引言

“互联网+政务服务”是深化“放管服”改革的关键之举，为提升政府服务质量，智慧政务理念应运而生。[1]从“电子政务”到“互联网+政务”，再到“智慧政务”[2]。智慧城市服务集纳了医疗、交通、公安户政、教育、公积金等多种民生服务办事功能，让市民充分享受城市生活的便捷，是“互联网+”在民生服务领域的落地。

智慧政务的兴起与新一代信息技术的成熟密不可分，随着“互联网+”技术在各个领域大规模落的应用，市民与企业的办事愈加便捷。相关部门也逐渐实现政府管理与公共服务的数字化和智能化。微信、微博、市长信箱、阳光热线等网络问政平台也成为政府了解民意、汇聚民智、凝聚民气的重要渠道。然而目前，大部分电子政务系统仍然依靠人工根据经验进行留言划分和热点整理，存在工作量大、效率低，且差错率高等问题，这一现状无法与不断攀升的各类社情民意相关文本数据量相适应。人工智能、大数据、云计算等发展技术便为此类问题提供了解决方案，建立基于自然语言处理（*NLP*）技术的智慧政务系统已成为社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

基于对智慧政务系统理解和认识，本文将立足于以上文本挖掘应用解决下列问题：

（1）文本分类问题，根据题目给出的数据，按照已知的划分体系建立关于留言内容的一级标签分类模型，并使用 *F*1 值对分类方法进行评价，计算模型准确率和召回率；

（2）文本热点话题提取问题，根据题目所给的数据用合适的方式将留言进行归类，并定义合理的热度评价指标，根据评价结果提取出排名前 5 的热点问题；

（3）文本质量评价指标体系构建问题，根据题目所给的数据，从相关性、完整性、时效性、可解释性、可读性等多个角度建立评价指标体系并借助 *python* 实现评价过程，给出评价结果。

二、任务一：群众留言分类

（一）数据描述

在正式进行文本挖掘分析前，先对已有的文本数据进行描述与预处理。基于网络问政平台群众留言文本数据，首先对所有用户留言详情的句长进行判断，得到分布结果。

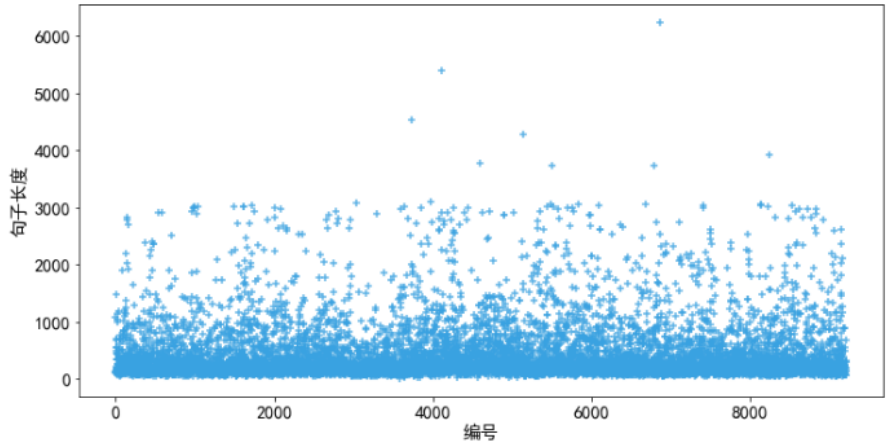

图 1 留言详情长度

由图 1 可知，该网络问政平台群众留言总共有 9210 条，其中群众留言详情句长主要集中在 1000 字以下，其次是 1000-3000 字，另有 8 条留言详情字数达到 3000 字以上。根据上述结果，本文将采取组距式方法对留言详情进行分组，以句子的长度为分组依据，分为 0-100 字，100-200 字，200-300 字，300-500 字，500 字以上 5 组。

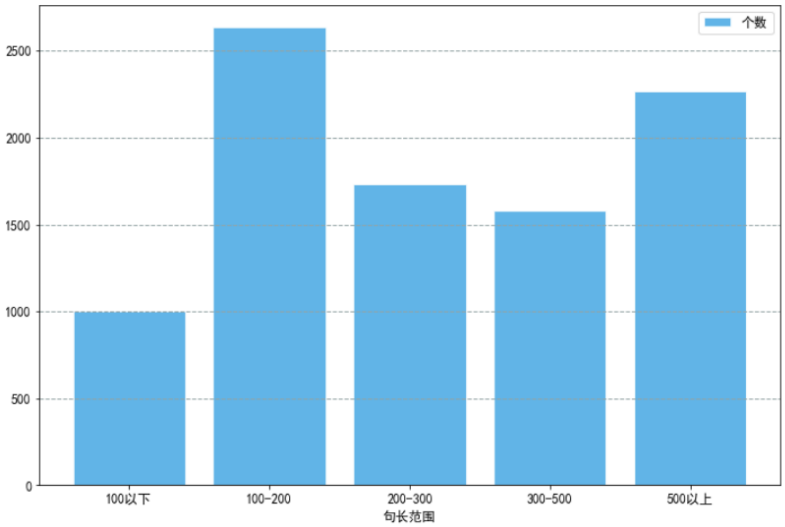

图 2 句子长度分布情况

如图 2 所示，留言详情长度以 100-200 字为主，因此在后续分析中，进行留言文本最长长度限制为 200 字的预处理设定，通过直接选取或截断组合形成分析所用的文本。

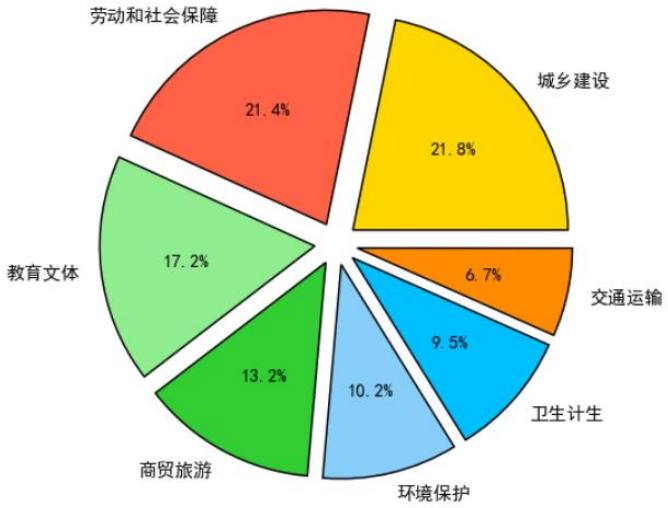

图 3 留言类别分布

附件 2 给出的用户留言数据已经含有一级标签，为避免数据不平衡问题，首先对所有留言一级标签进行统计，结果如图 3 所示。其中城乡建设一级标签留有 2009 条，劳动和社会保障 1969 条，教育文体 1589 条，商贸旅游 1215 条，环境保护 938 条，卫生计生 877 条，交通运输 613 条，分布比例大致为 3:2.5:2:1.5:1.4:1，可以看到留言类别之间数据虽有差异，但是总体成连续的阶段性变化，因此可以不考虑数据倾斜的问题。

（二）基本原理

针对任务一，本文将使用 *ERNIE*1.0 模型对留言内容进行分类。*ERNIE* 模型是由 *BERT* 模型改进发展而来的，是预训练模型大家庭的一员。*BERT* 是一个语言表征模型，通过超大数据、巨大模型和极大的计算开销训练而成，具有深而窄以及同时利用左右两侧的词语的特点，该模型被认为是 *NLP* 的里程碑式改变。

BERT 模型通过堆叠 *Transformer* 子结构来构建基础模型，其输入是文本中各个字/词的原始词向量，该向量既可以随机初始化，也可以利用 *Word2Vector* 等算法进行预训练以作为初始值；输出是文本中各个字/词融合了全文语义信息后的向量表示。模型的核心是 *attention* 机制，对于一个语句，可以同时启用多个聚焦点，而不必局限于从前往后的，或者从后往前的，序列串行处理，可以充分利

用上下文的信息，也摆脱了传统 *RNN* 模型的限制可以进行并行处理。

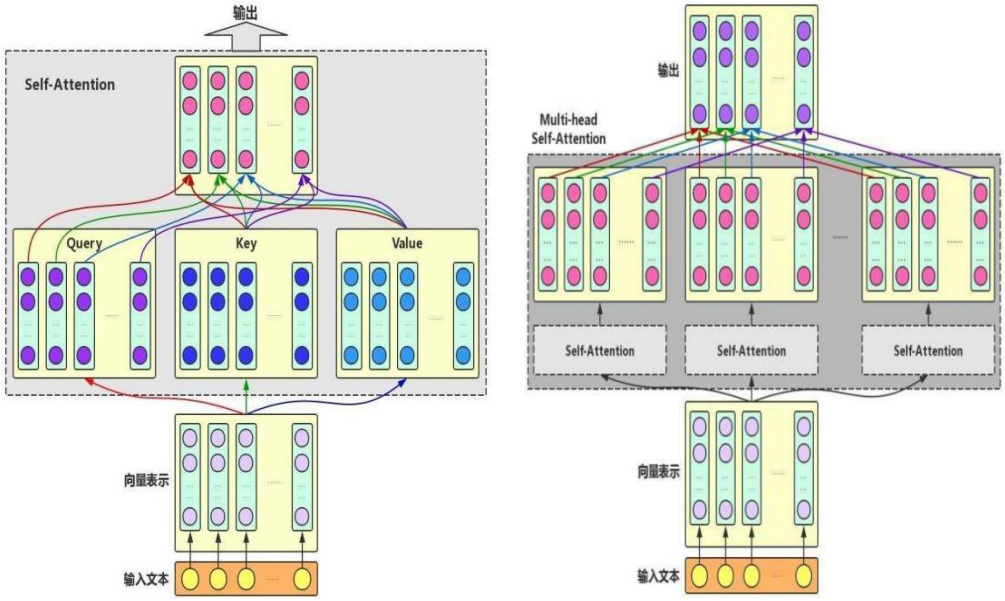

图 4 *Self-attention* 与 *Multi-head Self-Attention* 机制

不仅要正确地选择模型的结构，而且还要正确地训练模型的参数，这样才能保障模型能够准确地理解语句的语义。*BERT* 用了两个步骤，试图去正确地训练模型的参数。第一个步骤是把一篇文章中，15%的词汇遮盖，让模型根据上下文全向地预测被遮盖的词。通过全向预测被遮盖住的词汇，来初步训练 *Transformer* 模型的参数。第二个步骤继续训练模型的参数，两步训练合在一起，称为预训练 *pre-training*，训练结束后的 *Transformer* 模型，包括它的参数，是作者期待的通用的语言表征模型。

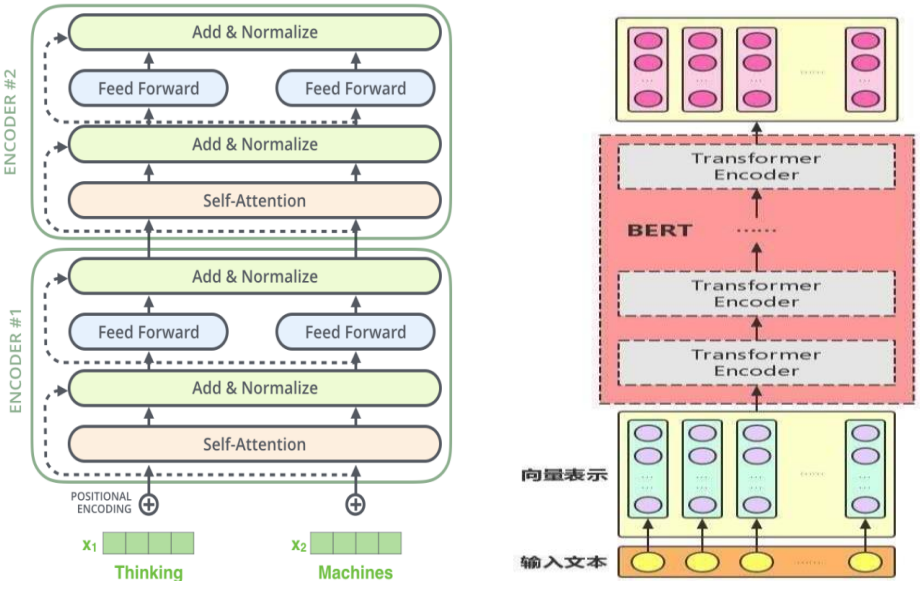

图 5 *Transformer-Encoder* 与 *BERT* 模型

BERT 的出现是建立在前期很多重要工作之上的，包括 *ELMO*, *ULMFiT*, *GPT*, *Transformer* 以及 *Skip-thoughts* 等，是一个集大成者。*BERT* 的出现极大地推动了自然语言处理领域的发展，凡需要构建自然语言处理模型者，均可将这个强大的预训练模型作为现成的组件使用，从而节省了从头开始训练模型所需的时间、精力、知识和资源。然而，*BERT* 模型主要建模原始的语言内部的信号，较少利用语义知识单元建模，导致模型很难学出语义知识单元的完整语义表示。这个问题在中文方面尤为明显。

针对 *BERT* 模型的不足，百度提出基于知识增强的 *ERNIE*1.0 模型，通过引入知识，使预训练模型学习到海量文本中蕴含的潜在知识，进一步提升了预训练语言模型在各个下游任务中的效果。

基于 *BERT*，*ERNIE* 模型提出两种加强的 *masking* 方式，分别是 *entity-level* 和 *phrase-level*。这是一种带有先验知识 *Mask* 机制，一般会包含多个字或词。在训练的过程中，模型可以学习到与 *phrase* 和 *entity* 相关的知识，包括实体间的关系，实体的属性，实体的类别等，使模型更好地泛化[3]。相较于 *BERT* 基于局部词语共现学习的语义表示，*ERNIE*1.0 直接对语义知识单元进行建模，增强了模型的语义表示能力。

表 1 *BERT* 模型和 *ERNIE*1.0 模型对比

句子模型	***BERT* 结果**	***ERNIE*1.0 结果**	**答案**
_____引起高血糖	糖糖内	胰岛素	胰岛素
买菜的市民告诉记者，以往节假日前一元及以下的菜价很少，今年一元左右的菜很多，_____真的很便宜	菜菜	价格	价格

表 1 展示了两个模型的任务完成情况对比。填空任务与 *ERNIE*1.0 引入的知识先验 *Mask LM* 任务十分相似。从表中的比较可知，*ERNIE*1.0 对实体词的建模更加清晰，对实体名词的预测也更准确。如 *BERT* 的答案“糖糖内”不是一个已知实体，且“菜菜”的词边界不完整而 *ERNIE*1.0 的答案能准确命中空缺实体[4]。这是因为在训练数据方面，*ERNIE*1.0 使用包括百科类、新闻资讯类、论坛对话类语料来训练模型，显著提升了模型的语义表示能力。

（三）实验设置

ERNIE 1.0 模型输入的句子长度最大为 512，且数据集中句子的长度绝大部分集中于 200 以下，再综合考虑计算机性能，本文选取的最大输入句子长度为 200。因此需要对数据进行处理，对长度大于 200 的句子进行切割，将一个句子分为 5 个部分，每个部分均选取最前的 40 个字，这样在固定长度的基础上尽可能的保留了全句的信息。由于 *ERNIE 1.0* 模型是基于字级别的预训练模型，因此可以无需进行分词操作，且一般意义上的停词也包含有一定信息，在分类任务中本文也并未去除停词，只去除了一些各个句子中均会出现的会影响语义的特殊字符。

为提升模型的鲁棒性和准确性，本次任务中采取了 5 折交叉验证，训练集：验证集：测试集的数据量比为 6:2:2。*ERNIE* 1.0 模型有一系列超参，其对最后的结果有着很大的影响，本实验在 *pytorch* 框架下进行，具体超参设置如下：

表 2 模型超参表

超参名称	具体数值	超参名称	具体数值
epoch	3	*hidden_dropout_prob*	0.1
batch_size	16	*initializer_range*	0.02
learning_rate	5E-05	*num_attention_heads*	12
pad_size	200	*num_hidden_layers*	12
warmup	0.05	*type_vocab_size*	2
attention_probs_dropout_prob	0.1	*layer_norm_eps*	1E-05
hidden_act	“*relu*”	*vocab_size*	18000

在通过 *ERNIE* 模型得到句向量后，本文运用 *pytorch* 的 *linear* 层，将句向量维度转化为类别维度，并将得分最高的类别作为预测类。

（四）实验结果

1、结果评价指标

混淆矩阵是表示精度评价的一种标准格式，使用准确率、召回率和 *F*1 值来评估模型。查准率（*Precision*）指被正确判定的正例数量占判定正例总数的比例，召回率（*Recall*）指被正确判定的正例数量占样本中正例总数的比例，*F*1 值是准确率和召回率的加权调和平均，其表达式为：

$$F=\frac{\left(\alpha^2+1\right)P\times R}{\alpha^2\left(P+R\right)} \quad (1)$$

当α=1时，即得到 $F1$ 值，为：

$$F1=\frac{2\times P\times R}{P+R} \quad (2)$$

在实际操作中，为正确预测出更多正例，通常会相应增加正例预测数量，而正例预测错误的数量会随之增加，因此准确率与召回率难以兼顾。$F1$ 值综合了准确率和召回率，当 $F1$ 值较高时，说明分类器预测效果较好。

2、预测效果评价

本章将附件 2 用户留言详情数据分为训练集，验证集与测试集，将训练集与验证集代入 *ERNIE*1.0 模型来调整模型参数，并在测试集上进行测试，得到模型预测效果的混淆矩阵如下所示。[1]

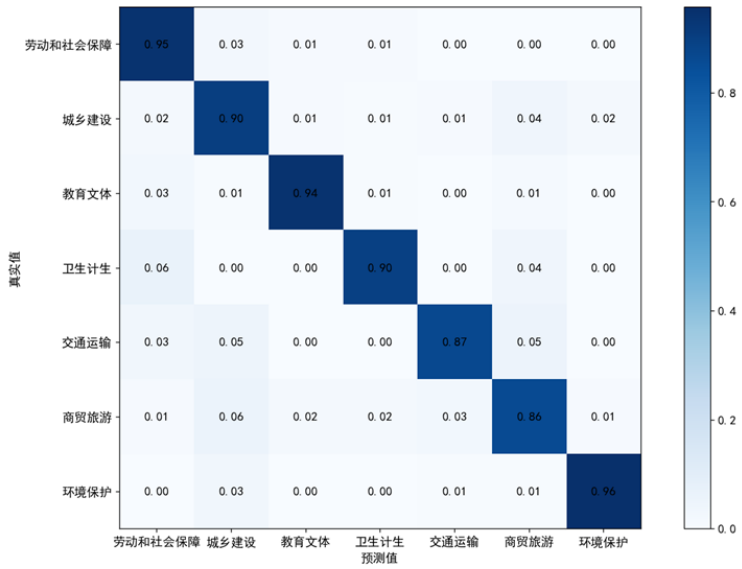

图 6 混淆矩阵

混淆矩阵对角线上的数字表示预测正确地比例，由图 6 可知，预测正确占比最高的是环境保护类留言，达 0.96，最低的商贸旅游类也有 0.86，总体而言，可认为 *ERNIE*1.0 模型分类器对用户留言详情主题分类效果良好。

为进一步了解模型的预测效果，观察表 3 给出的模型的准确率、召回率和 $F1$ 值可知，在经过微调后的模型对于测试集的用户留言详情所有主题分类准确度可以到达 91.45%，且对 7 个分类的预测结果 $F1$ 值均在 85%以上，说明分类器预测

[1] 由于上传文件大小限制，训练好的模型参数已上传至百度云，地址在附件说明中。

效果良好，*ERNIE*1.0 模型可以较好的完成对网络问政平台的群众留言主题分类的工作。

表 3 预测效果

类别标签	准确率	召回率	*F*1 值
劳动和社会保障	0.9158	0.9463	0.9308
城乡建设	0.9043	0.9020	0.9031
教育文体	0.9597	0.9439	0.9517
卫生计生	0.9274	0.8973	0.9121
交通运输	0.8880	0.8672	0.8775
商贸旅游	0.8589	0.8589	0.8589
环境保护	0.9476	0.9577	0.9526
准确度			0.9153
平均值	0.9145	0.9105	0.9124
加权平均	0.9145	0.9153	0.9152

三、任务二：热点问题挖掘

（一）定义热点问题及热度评价指标

对网络问政平台的留言进行挖掘，以发现有价值的热点信息对准确掌握群众的迫切需求具有意义重大[5]。所谓热点问题是指在某一时段内群众集中反映的某一问题。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。为增加确定热点问题的准确率，本文将定义如下热度评价指标：以留言主题单位时间内出现的次数来衡量问题的关注度，以相关留言单位时间内的点赞数和反对数来衡量问题的影响力。综合影响力和关注度计算相关问题的热度指数，计算公式如下：

$$\text{热度指数}=\text{数量}+\text{平均数量}+\text{平均点赞}-\text{平均反对} \quad (3)$$

其中，数量指各类别的留言数量，平均数量指留言数量除以对应留言的时间跨度，点赞数指各类别的留言点赞总数，平均点赞指点赞数除以对应类别内的留言数量，平均反对指点赞数除以对应类别内的留言数量。

热度指数综合考虑了总量指标和相对指标，其中数量因素为总量指标考虑了各个类别留言总数，其他三个为相对指标分别从不同侧面衡量了话题热度，较能全面精准地反映出话题的热度。

（二）提取思路

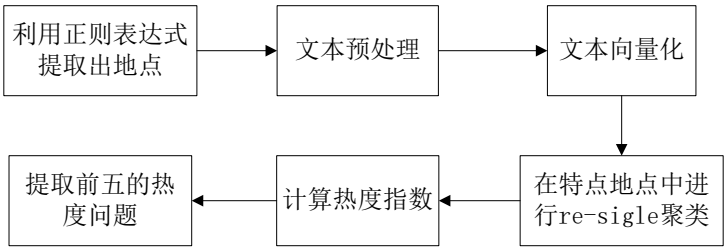

图 7 热点问题提取流程

步骤一：根据附件三的留言主题，使用正则表达式提取每一条留言涉及的地点，并根据地点将留言分为若干类，以简化后续的聚类时间与复杂度，并提高了聚类的精确度。经过我们观察，制定了提取规则，可以提取出留言所属的县市区或者省份，并在此基础上进行聚类；

步骤二：针对留言文本长度进行预处理。将文本不能体现句子特征的词语提出替换为空值，并去除手机号码、网址等词语；

步骤三：使用 *word2vec+TFIDF*、*doc2vec* 和 *ERNIE* 进行文本向量化操作；

步骤四：文本聚类，基于文本句向量，使用基于 *Single-pass* 改进而来的 *Re-Single* 方法进行对留言详情进行聚类，得到若干类集中讨论的热点问题；

步骤五：根据留言数量、留言时间及其点赞与反对数等信息，计算每一类问题的热度指数；

步骤六：利用上述结果进行排序，提取出排名前五的热点问题，并总结分析给出热点问题及其留言明细表。

（三）基本理论

1、文本向量化

文本向量化就是用一个向量的形式表示一个文本，这样处理可以简化两个文本相似度的计算，因此广泛应用于信息检索、机器翻译、文本聚类等自然语言处理相关领域之中。本文将采用以下三种模型进行句向量的提取。

（1）*word2vec+ TFIDF*

计算句子相似度与句子表达有密切联系，常用向量空间模型来表达句子特征，

将文档信息转化为向量。在表达过程中广泛使用词频—逆文档频率（*TFIDF*）权重计算方法衡量某个词汇对于文本的重要程度，若一个词在某一文本中出现频率较高，而在其他文本中出现频率较低，则相应的 *TFIDF* 值越大，认为该词越能表达相应文本的主题，该方法计算简单，且有较高准确率和召回率[6]。

Word2vec 是基于神经网络的词汇稠密向量化表示方法，包含 *CBOW* 和 *Skip-gram* 模两种词训练模型。其中，*CBOW* 模型是根据中心词$W(t)$周围的词来预测中心词，而 *Skip-gram* 模型则是根据中心词$W(t)$来预测周围词。

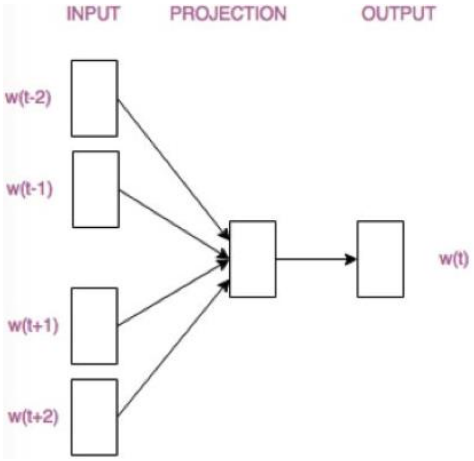

图 8 *CBOW* 模型

CBOW 模型使用 *one-hot* 编码将词进行向量化，模型等价于一个词袋模型的向量乘以一个 *Embedding* 矩阵，从而得到一个连续的 *Embedding* 向量。如上图所示，$W(t-2)$，$W(t-1)$，$W(t+1)$，$W(t+2)$表示当前词$W(t)$前后两个词，然后将这些词所对应向量的累加成和。

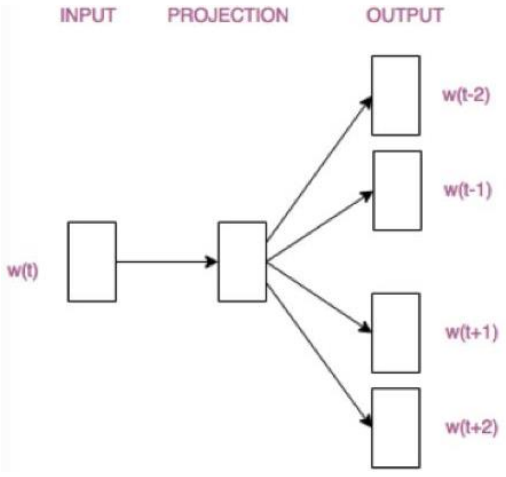

图 9 *Skip-gram* 模型

Skip-gram 模型思路与 *CBOW* 模型呈逆向关系，模型的前向计算过程如公式所示：

$$p\left(w_0 \mid w_i\right)=\frac{e^{U_0 V_i}}{\sum_j e^{U_0 V_i}} \tag{4}$$

其中，v_i 是词 w_i 的输入向量，U_j 是词 w_j 的输出向量。该模型本质是计算输入词的 *input vector* 与目标词的 *output vector* 之间的余弦相似度，并进行 *softmax* 归一化。

在 *word2vec* 基础上加入 *TF-IDF* 值，能够改进对文档向量未考虑单个词产生影响的缺陷，使文本向量化取得更好地效果。

（2）*doc2vec*

在 *word2vec* 模型的基础上，*Mikolov* 等人提出了 *Doc2vec* 方法，可以接受不同长度的句子做训练样本，实现文档或句子的向量化。该模型是一种无监督学习算法，可获得句子、段落和文档的向量化表达，进而计算向量的相似度。在 *Doc2vec* 模型中，每一句话用唯一的向量来表示，用矩阵 D 的某一列来代表。每一个词也用唯一的向量来表示，用矩阵 W 的某一列来表示。与 *word2vec* 模型一样，*Doc2vec* 也有两种训练方式，一种是 *PV-DM* 类似于 *word2vec* 中的 *CBOW* 模型，另一种是 *PV-DBOW* 类似于 *word2vec* 中的 *skip-gram* 模型。对于大多数任务，*PV-DM* 的方法表现很好，但两种方法相结合更值得推荐。

doc2vec 模型的过程，主要有两步：第一步，训练模型，在已知的训练数据中得到词向量，*softmax* 参数以及句向量等；第二步，推断过程，对于新的段落，得到其向量表达。具体地，在矩阵中添加更多的列，在固定的情况下，利用上述方法进行训练，使用梯度下降的方法得到新的 D，从而得到新段落的向量表达[7]。

（3）*ERNIE*

作为一个语言理解模型，*ERNIE* 模型也可以提取词向量实现语义表示。针对静态词向量，模型将文本信息映射到数字空间，变成数字表示的向量，并保留了词语间的距离信息。而针对联系上下文的动态词向量，编码器在进行映射的同时，还保证了词向量的上下文信息。使得词向量不仅保留原始信息，还能保留上下文语义，使预测分类更加在准确[8]。

（4）模型集成

Word2vec 与 *doc2vec* 均存在同样的缺点：1、词和向量是一对一的关系，所以多义词的问题无法解决；2、是一种静态的方式，虽然通用性强，但是无法针对特定任务做动态优化。而预训练 *ERNIE* 模型在存在训练语料的基础上能较好得体现出句子之间的相似度，但本实验仅能根据分类任务得到微调后模型，不能较为精确反映出文本之间的相似度。因此，本文考虑将三种句向量进行相加得到集成的句向量。

2、文本聚类

（1）*Single-pass*

Single-pass 文本聚类是一种流式聚类算法，属于单项遍历算法，在聚类过程中，每个样本只会参与一次聚类，并且对样本的先后顺序有一定的依赖性，使用该算法可以有效地对新闻事件或者垃圾文本进行在线聚类，进而发现热门话题。*Single-pass* 算法具有结果明确，可解释性强的特点，它可以有效防止文本变化的干扰，且对文本输入的顺序相对敏感。

Single-pass 算法顺序处理文本，以第一篇文档为种子，建立一个新的类别。之后再进行新进入文档与已有类别的相似度，将该文档加入到与它相似度最大的且大于一定阈值的主题中。如果与所有已有话题相似度都小于阈值，则以该文档为聚类种子，建立新的主题类别。其算法流程如下：

（1）以第一篇文档为种子，建立一个主题；

（2）将文档 *x* 向量化；

（3）将文档 *x* 与已有的所有话题均做相似度计算，本文采用附加时间因子的余弦相似度；

（4）找出与文档 *x* 具有最大相似度的已有类别；

（5）若相似度值大于阈值 *theta*，则把文档 *x* 加入到有最大相似度的主题中，跳转至（7）；

（6）若相似度值小于阈值 *theta*，则文档 *x* 不属于任一已有主题，需创建新的主题类别，同时将当前文本归属到新创建的主题类别中；

（7）聚类结束，等待下一篇文档进入。

（2）*Re-Single*

热点问题一般是指在某一特定时段公众所集中反映的某一问题，由于给出的问题列表时间分布范围较大，单纯考虑文本向量之间的余弦相似性容易引起混淆将时间跨度很长的问题也归为一类。因此在聚类时，本文采用了附加时间因子的余弦相似度。

$$sim(A,B)=\cos(A,B)\times timef(T_A,T_B) \tag{5}$$

$$\cos(A,B)=\frac{A\times B}{\|A\|\times\|B\|} \tag{6}$$

$$timef(T_A,T_B)=\begin{cases}1, & T_A-T_B<=90\\ 0, & T_A-T_B>90\end{cases} \tag{7}$$

sim 即为附加了时间因素的文本相似度，$\cos(A,B)$ 为文本向量 A 与文本向量 B 的余弦相似度，$timef(T_A,T_B)$ 为时间因子，即认为时间相差在 90 天以内的问题是有联系的，而超出 90 天的问题没有联系。

因为 *Single-pass* 算法较为依赖数据读入的顺序，可能导致分出的类别不能完全贴合实际，本文在 *Single-pass* 聚类的基础上本文进行了改进。在聚类结束之后，每个类别均会产生中心向量，而聚类结束后可能会产生问题数量很少的类别，该类别可能是由于顺序原因为被分到本应该属于的大类中。因此，本文将类中话题数量小于某个值的类别称为小类，大于等于该值的称为大类，对于某一小类衡量该类别中心向量与各个大类中心向量的附加时间因子的余弦相似度，选取出最小值，若该值大于阈值，则将该小类的全部数据放入该大类中。具体操作流程如下：

（1）依据提取出的地点对问题进行分类；

（2）对某一特点的问题进行 *Single-pass* 聚类；

（3）将聚类得到的类别分为大类与小类；

（4）计算与某一小类最相似的大类；

（5）若小类与大类的相似度值大于阈值 *theta*，则把该小类加入到有最大相似度的大类中，跳转至（7）；

（6）若相似度值小于阈值 *theta*，维持原有小类不变；

（7）聚类结束，等待下一小类进入；

（8）重复（4）-（7）步骤，直至所有小类均计算结束，进入下一地点。

（四）模型集成

Re-single 聚类方法最重要的是确定聚类的阈值，经观察得到编号 4313 至 4322 均为同一类数据，所得到的文本集成句向量之间的余弦距离如下所示:

表 4 余弦距离

文本编号	0.00	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00
0.00	1.00	0.87	0.83	0.86	0.91	0.68	0.86	0.84	0.84	0.56
1.00	0.87	1.00	0.90	0.82	0.96	0.77	0.89	0.87	0.87	0.53
2.00	0.83	0.90	1.00	0.90	0.94	0.90	0.91	0.96	0.96	0.67
3.00	0.86	0.82	0.90	1.00	0.90	0.87	0.85	0.88	0.88	0.79
4.00	0.91	0.96	0.94	0.90	1.00	0.83	0.91	0.92	0.92	0.64
5.00	0.68	0.77	0.90	0.87	0.83	1.00	0.81	0.87	0.87	0.78
6.00	0.86	0.89	0.91	0.85	0.91	0.81	1.00	0.93	0.93	0.67
7.00	0.84	0.87	0.96	0.88	0.92	0.87	0.93	1.00	1.00	0.69
8.00	0.84	0.87	0.96	0.88	0.92	0.87	0.93	1.00	1.00	0.69
9.00	0.56	0.53	0.67	0.79	0.64	0.78	0.67	0.69	0.69	1.00

为了尽可能将同一类别事件聚至同一类别中，并控制不让其他类别事件进入该类，综合考虑之后，本文将集成句向量聚类的阈值设置为 0.75，并以同样的方法确定其他方法句向量聚类的阈值。

（五）实验结果

本章使用 *python* 软件，基于 *Re-Single* 模型，运用了四种文本型向量化方法对附件 3 中的留言文本数据进行热点问题归类，结果发现经过集成模型所生成的句向量效果最好，提取排名前五的热点问题，结果如下所示。

表 5 热点问题

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	607	100	2019/6/20 至 2019/9/12	*A* 市五矿万境 *K*9 县	*A* 市五矿万境 *K*9 县存在住房安全问题
2	166	66.98131996	2019/1/14 至 2019/2/21	*A* 市 *A*4 区	*A* 市 58 车货案进展
3	435	64.14263939	2019/1/1 至 2019/6/2	*A* 市	*A* 市房产相关问题

4	507	62.589530 14	2019/5/14	*A* 市	*A* 市什么时候能普及 5G 网络？
5	428	61.440127 34	2019/1/18 至 2019/7/15	*A* 市	*A* 市交通相关问题

由表 5 可知，排名前五的热点问题均发生在 *A* 市，且讨论时间范围不超过半年，其内容主要涉及城乡建设、科技与信息产业和交通运输三个大类，第一类涵盖安全生产、城市建设和市政管理及住房保障与房地产等，第二类主要是信息化建设问题，而第三类包括规划、管理等交通建设问题。根据热度指数可以看出，安全生产问题是市民网友最为关注的问题，结合实际，安全问题的确是最需要保障的重点，因此相关部门必须切实加强城乡建设的安全保障措施。同时随着 5*G* 科技的发展，市民们对科技信息的关注热度也愈加提升。这五个热点问题的明细将在表 6 中展出。

表 6 热点问题明细（部分）

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
607	208636	*A*00077171	*A* 市 *A*5 区汇金路五矿万境 *K*9 县存在一系列问题	2019/08/19 11:34:04	我是 *A* 市 *A*5 区汇金路五矿万境 *K*9 县 24 栋的一名业主，我们小区一开始的定位是一个高端别墅小区……	2097	0
607	215507	*A*0001 03230	*A* 市五矿万境 *K*9 县存在严重的消防安全隐患	2019/09/12 14:48:07	预交房 23 栋没有通往负一楼的楼梯，存在严重的消防安全隐患，开发商处理态度消极不予整改……	1	0
607	234086	*A*00099869	*A* 市五矿万境 *K*9 县房子的墙壁又开裂了	2019/06/20 09:30:44	五矿万境 *K*9 县的房子又出问题了，又是墙壁开裂！令人胆颤心惊！我是五矿万境 *K*9 县 36 栋的业主……	6	0
166	214238	*A*00061787	请问 *A*4 区公安派出所对 58 车货一案办案的进度如何了	2019-01-20 22:28:40	标题：恳请市委胡书记督促 *A*4 区经侦大队和办案警官毛浚发一个 58 车贷恶性退出案件的案情通报哪怕几个字也行胡书记……	1	3
166	220711	*A*00031682	请书记关注 *A* 市 *A*4	2019-02-21 18:45:14	尊敬的胡书记：您好!*A*4 区 *p*2*p* 公司 58 车	821	0

			区58车贷案		贷，非法经营近四年……		
166	272413	A0001 06062	西地省A市58车贷恶性退出，A4区立案已近半年毫无进展	2019-01-14 20:23:57	西地省58车贷邢ze恶性退出，潜逃美国已半年，A市A4区经侦大队和办案警官毛钧懒作为，办案毫无进展，至今没有发过一次案情通报……	2	0
435	261570	A0004 2277	关于A市住房贷款商转公问题的建议	2019-03-22 16:50:16	目前A市的商转公，需要公民自己先借钱还清商业银行贷款，拿到产权证后，再到公……	26	0
435	226996	A0002 2217	A市汽车南站何时能建好？	2019-03-20 09:20:46	A市汽车南站何时建好？按之前向社会公示的工期为2018……	32	0
435	239670	A0008 0329	问问A市经开区东六线以西泉塘昌和商业中心以南的有关规划	2019-01-11 15:46:04	A市经开区东六线以西，泉塘昌和商业中心以南， 新蕾品阁居小区以北，目前闲置的恒天九五重工原厂房和闲置的西地省达源置业有限公司厂房……	41	0
507	219572	A0004 5546	请问A市什么时候能普及5G网络？	2019-05-14 11:19:22	A市A2区之前宣布，5G站址布点1432个，全区5G网络基础建设基本完成……	4	1
507	248712	A0004 5546	请问A市什么时候能普及5G网络？	2019-05-14 11:22:13	A市A2区之前宣布，5G站址布点1432个，全区5G网络基础建设基本完成……	0	0
507	316619	A2352 59	请问A市什么时候能普及5G网络？	2019-05-14 11:22:13	A市A2区之前宣布，5G站址布点1432个，全区5G网络基础建设基本完成……	0	0
428	225849	A0006 1602	反映A市金星北片110kv及以上高压线的现状和规划的几个问题	2019-04-15 16:15:32	胡市长：您好！我是一名在A6区金星北片区普通的投资置业者，惊闻A6区政府准备在月亮岛路上修建110kv……	18	0
428	243808	A0005 3304	强烈建议将地铁7	2019-03-06 14:20:16	尊敬的A市委领导、A市规划局领导：长株潭	31	0

			号线南延至 *A* 市生态动物园		一体化提出已 9 年，暮云作为融城“核心”区域，目前进入……		
428	195915	*A*00018309	建议 *A* 市规划局东延地铁 7 号至泉塘片区，便民出行	2019-02-26 08:22:42	您好！泉塘片区位于星沙，占了 *A* 市经开区 70%的企业，占经开区总面积的三分之二。下辖 14 个社区，入区企业 300 多家……	21	1

注：由于篇幅限制，正文中仅展现每个问题 *ID* 对应的 3 条，共 15 条留言的明细内容

四、任务三：答复意见评价

（一）答复质量评价方案构建

针对相关部门对留言的答复意见，本文将构建以下质量评价方案对答复质量进行评估。

表 7 评价指标

指标	释义
相关性	根据留言答复与留言详情的相似度进行评判
结构完整性	使用三个评判标准来衡量留言答复是否规范，1、答复末尾是否有日期；2、答复开头是否有问候语；3、答复中是否有“答复”、“回复”等提示字
内容完整性	根据留言答复的长度来衡量其内容的完整性
可解释性	可解释性指可以追踪到数据来源，因此本文根据留言答复是否有可靠的来源依据进行衡量，即答复中是否有官方法规或通知等
时效性	依据留言答复和网友留言的时间差进行衡量
可读性	可读性指文本的被理解能力，使用 *ARI* 指数来衡量

针对以上指标，有以下相关补充说明：

1、在相关性中，用余弦相似度[9]计算文本的相关性。其基本原理是，对二维空间的向量 $\vec{A}=(x_1,y_1)$，$\vec{B}=(x_2,y_2)$，有余弦夹角公式：

$$\cos(\vec{A},\vec{B})=\frac{\vec{A}\times\vec{B}}{|\vec{A}||\vec{B}|} \quad (8)$$

通过文本向量化后，留言文本将转化为多维向量，此时上述余弦计算方式同样适用。假定多维向量 $A=(A_1,A_2,\ldots A_n)$，$B=(B_1,B_2,\ldots B_n)$，则有：

$$\cos(A,B)=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}}=\frac{A\times B}{|A|\times|B|} \quad (9)$$

若余弦值越接近 1，则表明夹角越接近 0 度，也就是两个向量越相似。

2、在结构完整性中，将引入虚拟变量进行编码，其中“是”为 1，“否”为 0，最后加总，给每一条答复一个得分。

3、在可解释性中，以后缀、前缀、动词三个标准进行判别，如下表所示，并同样引入虚拟变量进行编码。

表 8 判别标准

标准	标准明细
后缀	《……通知/通令/批复/政策/规定/规划/法规/条例/指南/准则/细则/修订》 《……的决定/请示/公约/方案/协定/意见/纪要/纲要/报告/办法/命令/复函》 《……法/修正案》等
前缀	《中华人民共和国……》 《人民代表大会 /全国人民代表大会……》 《发展改革委/发改委……》 《国务院……》等
动词	按照、依照、参照、依据、根据等

4、在时效性中，将留言答复和网友留言的时间差进行分等级评分，评分赋值依据如下：

表 9 赋值依据

时间差	7 天以内	7-15 天	15-30 天	30-90 天	90-150 天	150 天以上
赋值	5	4	3	2	1	0

5、在可读性中，使用 *ARI* 指数来量化文本的吸引程度，其数值近似等于我们可理解一段文字的最低程度，计算公式为：

$$ARI=4.71\times\left(\frac{\text{总字符数}}{\text{总字数}}\right)+0.5\times\left(\frac{\text{总字数}}{\text{总句数}}\right)-21.43 \quad (10)$$

其中，总字符数指一条回复的所有有效字符的总数，不包括标点符号；总字数指一条回复的所有字数，包括文字与标点符号；总句数指每条回复可分化为完整句子的数量。

（二）基本原理

本文采用 *TOPSIS* 法进行留言答复质量评价指标的综合评价，*TOPSIS* 法的基本原理与逼近理想解基本保持一致：首先构建一个归一化的原始矩阵，在所有方案中找到最优和最劣方案，并用向量表示这两个方案，然后计算目标方案与最优方案的差距，得出评价方案与最优方案的相似度，并以此作为评价的条件。基本步骤如下[10]：

（1）建立评价矩阵。对 n 个需要评价的对象，p 个评价指标，构造 $n\times p$ 的矩阵，若存在逆指标，则进行正向化处理，最终得到的如下矩阵：

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times p} \tag{11}$$

（2）对原始数据归一化：

$$Z=\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}_{n\times p} \tag{12}$$

其中，$z_{ij}=\dfrac{x_{ij}}{\sqrt{\sum_{k=1}^{n}x_{kj}^{\ 2}}}, i=1,2,\ldots,n;\ j=1,2,\ldots,p$

（3）根据各评价指标最优最劣值分别构造最优值向量 Z+和最劣值向量 Z−。

$$Z^{+}=\left(z_{j}^{+},z_{j}^{+},\ldots,z_{p}^{+}\right),\quad Z^{-}=\left(z_{j}^{-},z_{j}^{-},\ldots,z_{p}^{-}\right) \tag{13}$$

其中，$z_{j}^{+}=\left(z_{1j}^{+},z_{2j}^{+},\ldots,z_{pj}^{+}\right), j=1,2,\ldots,p$，$z_{j}^{-}=\left(z_{1j}^{-},z_{2j}^{-},\ldots,z_{pj}^{-}\right), j=1,2,\ldots,p$

特别地，针对本文的任务三，为剔除极端值的影响，将最优最劣值替换为 1%和 99%分位数，进行进一步分析，对于小于 1%分位数的以 1%分位数计算，大于 99%分位数的以 99%分位数计算。

（4）计算各评价单元与 1%和 99%分位数的距离，进行最后评判。

（三）实验结果

评价体系中可解释性、可读性、相关性以及结构完整性均为正向化指标，而时效性为逆向指标，而内容完整性我们认为其为中间型指标，即认为不长不短所包含的内容最为完整，太短容易导致信息缺失太长又易导致信息冗余。通过所建立对政府答复的指标评价体系并对数据进行合理变形之后，本章使用 *TOPSIS* 法对每条留言回复的质量进行评分，得到如下结果。

表 10 留言答复评分（部分）

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	答复评分
4331	*UU*008233 8	质疑 *A* 市新城国际花都的装修价格以及政府发文【2018】53号文件	2018/10/24 10:55:47	过查看在政府备案的该楼盘的装修价格的评估报告分析，我们业主对该楼盘的装修价格以及……	您好！……感谢您的理解和支持！2018年11月2日	2018/11/6 10:21:48	100
5339	*UU*00815	请求 *A* 市执法部门整治井湾子山水庭院小区脏乱差	2018/7/13 10:58:24	未成立业委会，社区对几个热心牵头成立业委会的居民不闻不问，忽悠，认为成立了业委……	您好！您的留言已收悉……感谢您 2018 年 7 月 23 日	2018/8/6 14:12:17	100
79794	*UU*008154 7	反映 *F6* 县教育系统的问题	2019/12/17 16:14:21	育系统每年招聘大量的教师，可偏远地区每年都缺老师……	西地省平台反映 *F6* 县教育……2020 年 1 月 3 日	2019/12/20 17:10:57	99.48593 78
158587	*UU*008224 2	咨询森林植被恢复的收取标准	2015/7/8 22:27:28	在省林业厅办理征占用林地手续时，省林业厅收取森林植被恢复费无……	关于“咨询森林植被恢复的……2015 年 7 月 27 日	2015/7/27 11:34:43	92.94454 45
121137	*UU*008197 8	请求禁止 *L* 市恒大御景湾 *B* 区项目通宵施工	2018/5/16 9:40:59	湾 *B* 区项目通宵施工当前，我们国家正在实行污染防治攻坚战，每个领域都在践行以人民为中心的发展……	网友：您反映的问题已转至市住建局……感谢你 2018 年 7 月 11 日	2018/5/18 10:58:24	91.99234 76
……	……	……	……	……	……	……	……
10924	*UU*008242 0	投诉石长铁路火车夜晚鸣笛扰民	2015/10/30 22:48:11	给您留言，是想投诉石长铁路火车深夜鸣笛对居民的影响……	网友：您好！留言已收悉	2015/11/16 11:18:18	12.11042 93
7500	*UU*0	请求恢复	2018/	个边远山村，原三	您的留言已	2018/6	10.0

0	081539	*F*7 县三联卫生院	3/25 23:20:25	联卫生院，因并乡后，其医务工作者先后退休了……	收悉。关于您反映的问题，已转 *F*7……	/29 16:07:22	715716
12011	*UU*008186	咨询留学生报名省考 *A* 市职位的专业认证问题	2014/9/10:28:39	一名留学生，归国后想参加公务员考试。然而当前公务员考试的专业……	网友：您好！留言已收悉	2014/10/11 17:00:15	9.292545 4
74555	*UU*008839	建议 *F* 市对农村放烟花鞭炮进行一定控制	2019/1/22 12:03:12	者喜庆大多数百姓放烟花鞭炮造成环境……	您的留言已收悉。关于您反映的问题，已转市……	2019/3/8 9:16:10	7.576396 2
30019	*UU*008151	在 *A*6 区准备全款购买二手房事项的咨询	2016/11/3 10:00:17	二手房，房产局资	已收悉	2016/11/22 12:25:56	0.874286 7

由于篇幅限制，以上仅展示了十条留言答复的评分结果，给更好地进行结果分析，以下给出全部留言的评分分布图。

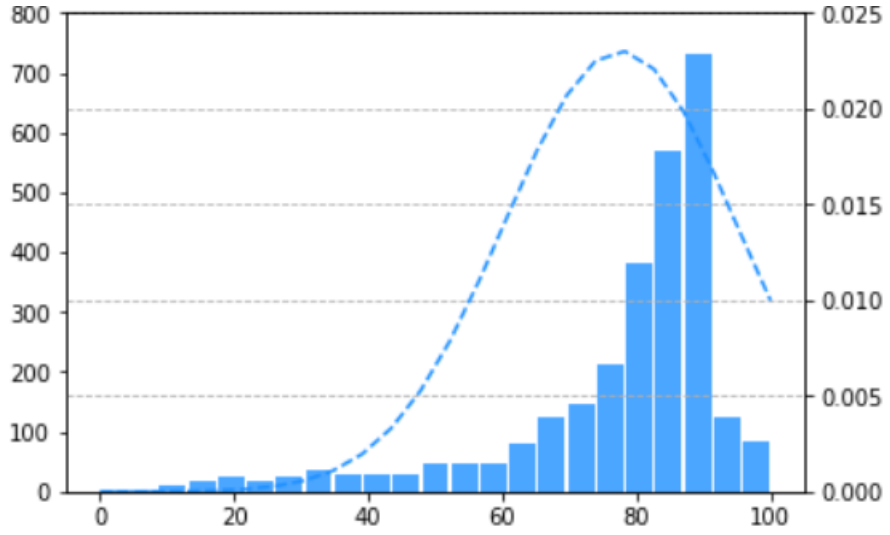

图 10 政府答复评分

如图 10 所示，本次评分结果采用百分制，其中约 87%的留言答复获得 60 分以上，获得 80-90 分的留言答复占比最大。可以看出，政府给出的留言答复在相关性、可解释性、时效性、完整性以及可读性方面，具有较好的质量，但仍有不少低分答复，甚至存在 10 分以下的现象，因此政府还需继续改进，进一步做到对网民留言答复的高质量、高效率的要求。

五、总结与展望

本文基于预训练模型 *ERNIE* 生成了各个问题相关内容的句向量。在第一题中，本文基于提供的数据对模型进行了 *Fine-tune*，使得模型更加贴近政务文本这一具体问题，提取出了留言详情内容的句向量，并以简单 *pythorch* 的 *Linear* 层进行线性变换得到分类结果，取得了 91.3%的较好效果。在第二题中比较了 *word2vec+TFIDF*、*doc2vec*，微调后的 *ERNIE* 模型以及三种句向量集成所生成句向量的聚类效果，发现集成得到句向量得到的聚类效果较好。同时，由于文本类别数量难以确定，因此本文使用了无需确定类别数量的 *Single-pass* 聚类方法，并在此基础上进行了改进，提出了 *Re-Single* 方法在一定程度上克服了 *Single-pass* 方法对输入数据的顺序较为敏感的缺点。为了使热点问题能够符合特定地点某一时间的属性，本文首先对地点进行了提取并在各个地点中进行小范围的聚类，同时在衡量文本相似度时考虑了时间因素。最后通过合理的热度评价指标构建，生成各个热点话题的排名。在第三题中，本文构建了指标体系，从六个方面衡量了回复文本，并利用 *TOPSIS* 法综合了各个指标，对回复文本进行了评分评价。

在本次比赛中，由于时间以及设备的限制，本文没有过多去调整 *ERNIE* 模型的超参，并且在得到句向量后直接进行线性变换得到分类结果，未尝试进行其他操作。针对第二题，本文基于第一题微调后的模型进行句向量提取，并与另外两种方法得到的句向量进行集成，但文本聚类的效果并没有非常完美，容易将第一题的各个分类中的文本聚集在一起，致使类别不纯。若时间充足，可以先标注一部分文本相似度，再用 *ERNIE* 模型进行微调，这样得到的效果将更令人满意。且在第二题中，提取地点是基于规则的提取，容易产生错漏，若对留言数据进行标注之后再进行实体命名提取就可以得到更为精准的地点。同样在第三题中可解释性等指标也均是基于规则的提取，易发生错误和遗漏。同时，对指标体系的建立，仍然存在一定的主观性，缺乏权威性，这在之后的挖掘研究中，可以进行进一步改进。

参考文献

[1] 刘洁.智慧政务 *APP* 应用问题与对策[J].合作经济与科技,2020(08):169-171.

[2] 《打造“智慧政务”离不开哪些“新基建”力量》 http://finance.eastmoney.com/a/202004261467126485.html

[3] 《【预训练语言模型】百度出品 *ERNIE* 合集，问国产预训练语言模型哪家强》 http://mp.weixin.qq.com/s?__biz=MzAxMDk0OTI3Ng==&mid=2247484077&idx=1&sn=fa39b7df8380ea3c49fa5cfd58f446ff&chksm=9b49c55eac3e4c489b3230063700e9414e2e3f5a75018ab3059d2c31a64f9e591da0df927bd9&mpshare=1&scene=1&srcid=&sharer_sharetime=1585903277932&sharer_shareid=21a86c8a927dffbd6d73923f42bdf013#rd

[4] 《*BERT* 和 *ERNIE* 谁更强？这里有一份 4 大场景的细致评测》 https://blog.csdn.net/paddlepaddle/article/details/92717195

[5] 聂文汇,曾承,贾大文.基于热度矩阵的微博热点话题发现[J].计算机工程,2017,43(02):57-62.

[6] 王闻慧.基于谷歌翻译及 *Doc2vec* 的中英句子相似度计算[J].电脑知识与技术,2019,15(15):224-227.

[7] 《*Doc2vec* 学习总结（三）》https://www.cnblogs.com/conan-ai/p/11354926.html

[8] 《【*ERNIE*】深度剖析知识增强语义表示模型——*ERNIE*》 https://cloud.tencent.com/developer/article/1557849

[9] 刘奔奔. 基于集成训练模型的实体链接模型[D].哈尔滨工业大学,2019.

[10] 刘宏亮,罗娇霞,陈国生.基于 *TOPSIS* 法的区域农村一二三产业融合发展综合评价研究——以衡阳市为例[J].福建茶叶,2020,42(04):102-103.