一种基于卷积神经网络“智慧政务”模型

摘要

政府为了通过各式各样的网络问政平台来了解民意、汇聚民智、凝聚民气，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。而近年来，自然语言处理(NLP)作为人工智能的一个重要领域得到了飞速发展。因此，本文通过不同的方法，构建基于自然语言处理技术的文本分类模型，以解决此类问题。

针对问题一：对群众留言训练集进行数据预处理，对训练集中数据的特征有一个清晰的了解，并对训练集进行去噪处理，防止干扰训练。

针对问题二：将某一时段内反映特定地点或特定人群问题的留言进行归类，根据附件 3 中点赞数和反对数的权值进行比较定义热度评价指标，并给出评价结果。并在该指标的基础之上进行热点评价筛选找出特定的热点问题和对应热点问题的详细信息,取出排名前 5 的热点问题，按照题干中给出的表结构进行存储。

针对问题三：给答复质量设计一套评价方案。考虑到答复相关性、完整性和可解释性的影响，这里需要理清三种影响的关系。以网民想得到的即相关又完整且可解释的答复作为最总质量评价标准。

关键词：**卷积神经网络 自然语言处理 word2vec TF-IDF**

目录

1.引言

随着网络时代的到来，政府可以通过微信、微博、市长信箱、阳光热线等网络问政平台来获取群众的意见。由于文本数据量过于庞大，给以往依靠传统的人工来对留言划分和热点整理的工作带来了极大的挑战。并且人的精力是有限度的，在长时间分类过程下可能也会造成对留言问题区分失误。因此，本文通过不同的方法，构建基于自然语言处理技术的文本分类模型，以解决此类问题。

问题一：对群众留言训练集进行数据预处理，对训练集中数据的特征有一个清晰的了解，并对训练集进行去噪处理，防止干扰训练。

问题二：将某一时段内反映特定地点或特定人群问题的留言进行归类，进行比较定义热度评价指标，并给出评价结果。并在该指标的基础之上进行热点评价筛选找出特定的热点问题和对应热点问题的详细信息,取出排名前 5 的热点问题，按照题干中给出的表结构进行存储。

问题三：给答复质量设计一套评价方案。考虑到答复相关性、完整性和可解释性的影响，这里需要理清三种影响的关系。以网民想得到的即相关又完整且可解释的答复作为最总质量评价标准。

2.分析方法与过程

2.1 问题一的分析方法与过程

2.1.1 数据预处理

分词去停用词

由于中文文本的特点是词与词之间没有明显的界限，从文本中提取词语时需要分词，故采用 Python 开发的第三库--jieba 分词，对留言的每一句话进行中文分词。而且在文本处理时，会有很多的停用词,停用词是指功能极其普遍,与其他词相比没有什么实际含义的词。通常是一些单字，或单字母以及高频的单词。例如中文中的‘我’，

‘你’，‘了’，‘的’，‘吗’等。本文所用的停用词，取自百度停用词表。

word2vec

为了将留言输入神经网络进行训练，采用 Word2vec 将自然语言符号表示成计算机能够理解的数字形式。Word2vec 包含了两种训练模型，分别是 *CBOW* *Skip_gram* 模型，如图 1 所示。中 *CBOW* 模型利用上下文预测当前词，而 *Skip_gram* 模型利用当前词预测其上下文。

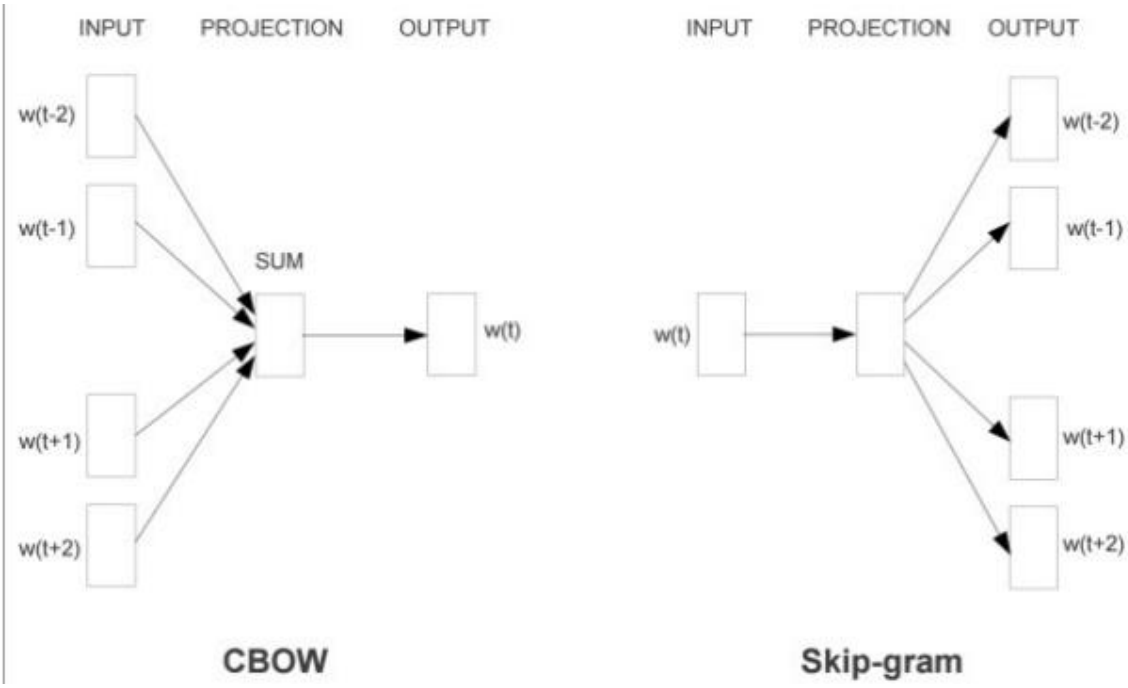


图 1: 两种 word2vec 算法网络示意图

我们使用 *Skip_gram* 模型构建智能阅读系统。Skip-gram 模型的训练目标就是使得下式的值最大：

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

其中，c 是窗口的大小，T 是训练文本的大小。基本的 *Skip_gram* 模型计算条件概率如下式:

$$p\left(w_O\mid w_I\right)=\frac{\exp\left(v_{wo}^{'T}v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left(v_w^{'T}v_{w_I}\right)}$$

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。

2.1.2TextCNN 神经网络的文本分类

卷积神经网络主要由提取特征的卷积层、压缩结果、保留重要特征的池化层以及将文本数据映射到样本标记空间的全连接层组成。应用于文本的卷积神经网络，卷积层输入常是嵌入向量，因此通常会有嵌入矩阵层，将词语索引转换为嵌入向量。卷积神经网络模型架构如图 2 所示。

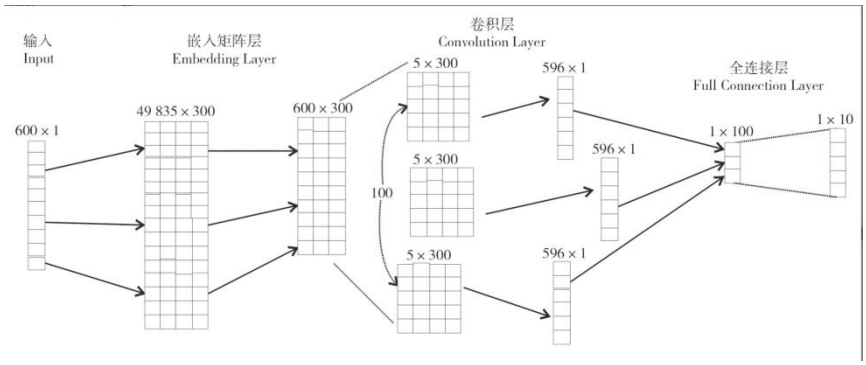


图 2 模型图

(1)输入层：输入层的输入是一个代表文本的矩阵，d 代表词向量的维度，n 代表每个数据所包含的词向量的数量。

(2)嵌入矩阵层：将由嵌入向量堆叠形成的矩阵称为嵌入矩阵，矩阵列数与嵌入向量维度相同。根据嵌入向量是词语或字符级别，将嵌入矩阵分为词嵌入矩阵或字符嵌入矩阵。由于本文工作是基于词语级别的输入，因此后文中嵌入矩阵特指词嵌入矩阵。

(3)卷积层：卷积层涉及卷积核 $w \in R^{hk}$，h 表示卷积窗口大小，k 为卷积维度，等于词向量的维度。一般来说，$W_{i:i+h}$ 表示单词 W_i，$W_{i+1}.....W_{i+h}$。所以生成一个文本特征的表达式为 $C = f\left(w \cdot W_{i:i+h+b}\right)$ 其中 b 为偏置，f 为非线性函数。将此卷积核应用于 $\left(W_{1:h}, W_{2:h+1}.....W_{N-h+1:N}\right)$ 生成一个特征映射 $c = \left(c_1, c_2.....c_{N-h+1}\right)$

(4)全连接层：将池化层输出的多个特征向量进行拼接并输入到全连接层的输入。

2.2 问题 2 分析方法与过程

2.2.1 分析方法

将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果。并在该结果的基础之上进行筛选找出特定的热点问题和对应热点问题的详细信息，按照题干中给出的表结构进行存储。

2.2.2 特定地点或特定人群问题的归类和热度评价指标

对指定问题的人群的问题留言进行分类。首先观察附件 3 中的文本信息可以发现在附件 1 中都能找到对应的分类标准和对应的一级分类、二级分类以及三级分类。根据附件 1 中所给内容对附件 3 中内容进行一一划分。

问题热点评价指标。根据附件 3 所给信息，发现最后两列为点赞数和反对数，于是我们将点赞数的权值和反对数的权值进行相减得到问题的热点评价指数。热点评价指标与点赞数成正比，与反对数成反比。

2.2.3 保存热点问题

根据所给文件附件 3 中，可以发现文件数量非常庞大。按照传统方法非常耗时，且效率低下。Python 中有许多优秀的资源、第三方工具来帮助我们去解决问题。由于题目所给附件为 Excel 表格，这里我们使用 Python 中 panads 进行解析，jieba 进行语义判断。pandas 是基于 NumPy 的一种工具，该工具是为了解决数据分析任务而创建的。Pandas 纳入了大量库和一些标准的数据模型，提供了高效地操

作大型数据集所需的工具。“结巴”中文分词：做最好的 Python 中文分词组件。

第一步 使用 panads 打开附件 3 得到 DataFrame 模型

DataFrame 是一种表格型数据结构，它含有一组有序的列，每列可以是不同的值。DataFrame 既有行索引，也有列索引，它可以看作是由 Series 组成的字典，不过这些 Series 公用一个索引。

第二步 计算出对应每条数据的热点指标

degree = agree - oppose

degree 表示热点度

agree 表示点赞数

oppose 表示反对数

第三步，从留言中提取出地点/人群和问题描述

在附件 3 的 DataFrame 中进行遍历，取出留言主题和留言详情。利用 Python 中 jiaba 库进行文本语义分析。将留言主题和留言详情中语义为名词词汇的词语进行保存和语义为动名词性的词语进行拼接得到留言的问题描述。

将留言主题和留言详情中的词义为名词与词义为地点名词的词汇进行拼接，然后与留言问题描述进行对比，得到其内容相近的短语即为留言中的地点/人群。

2.3 问题 3 分析方法与过程

2.3.1 分析方法

从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。针对这个问题，是在 excel 为源数据的情况下的数据挖掘，主要依赖于自然语言处理技术。

2.3.2 相关性分析过程

判断答复的相关性。在这里提取附件 4.xlsx 里“留言详情”和“答复意见”两列为原始数据进行数据分析。考虑到是对文本的挖掘，每个文本可能会出项的符号会对文本分析产生些许的影响，因词用 python 中替换操作对其进行整理，去除提取文本中的所有符号，只保留有有效文本，称有价值文本。

考虑到问题和答复处于不同的人，语句的结构不可能相同（在这里把概率很小事件，视为不可能事件），加之中文汉字的含义众多。因此，这里采用 python 的中文分词包 jieba 其中的全模式分词模式进行分词。全模式是基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG）。

对分完词后的每对问题与答复的文本使用余弦相似度计算文本相似度。即判断出来答复的相关性。

2.3.3 完整性分析过程

判断答复的完整性。对有价值文本中的答复分析可以看出，答复文本中都是以总结性句子或者不相关性句子开头，而这些句子是没有价值的。并且文本有明显的分隔词，如“答复如下”、“回复如下”。因此以“复如下”为分割点，只保留其后面的部分，从而提高准确度。

答复完整性评价方案，即为对于问题文本提到的关键字，答复文本有涉及到。例如，问题文本中提到“要通过征收小区停车费增加收入”、“路挖得稀烂用围栏围起，一直不怎么动工”。答复中有涉及“停车费”、“路”即为答复的完整性。

考虑到尽可能多的提取关键字以及尽可能少的延伸出无关词汇的所有成词情况，这里选择 jieba 中的精确模式进行分词，将句子最精确地切开。考虑到段落的逻辑性，有很多的铺垫和礼貌用于，在这里需要去掉不相干词性以及不相关词汇，如标点符号、连词、助词、副词、介词、时语素、‘的’、数词、方位词、代词。不相关词汇采用自定义词汇，保存在 stop_word.txt。

经以上操作，仍有大量不相关名词或动词性的词汇存在，使得判

断模型准确度不理想。故进行进一步优化。这里采用关键词提取中的TF-IDF（Term Frequency - Inverse Document Frequency）算法。关键词提取采用无监督学习算法，即先抽取出候选词，然后对各个候选词进行打分，然后输出 topK 个分值最高的候选词作为关键词。打分策略即 TF-IDF。TF-IDF 算法的具体原理如下：

第一步 计算词频，即 TF 权重。

$$TF = \frac{\text{count}(word)}{count(di)}$$

count(word)表示文本 di 中包含词 word 的个数；

count(di)表示文本 di 的词的总数；

第二步 计算逆文档频率，即 IDF 权重。这里的语料库使用 jieba 提供的。

$$IDF = \frac{num(corpus)}{num(word)+1}$$

num(corpus)表示语料库中文档的总数；

num(word)表示语料库 corpus 中包含 word 的文档的数目

第三步，计算 TF-IDF 值。

TF-IDF = TF(词频) × IDF(逆文档频率)

实际分析得出 TF-IDF 值与一个词在文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取文本的关键词。考虑到文本长短不一，进一步优化，提取每个文本长度前 30%的关键词。

对进一步提取的关键字使用余弦相似度计算文本相似度。判断出来答复的完整性。

2.3.3 可解释性分析过程

判断答复的可解释性。考虑到可解释性的概念比较模糊，故需要先假设一个可解释性模型。这里简单的假设网民想要得到的肯定或否定的解释为可解释性。即“要通过征收小区停车费增加收入”，答复

中提到“降低停车费”即为可解释性，否则为不可解释性。

文本预处理。对留言文本进行关键字提取。对答复文本去除留言文本中的关键字和停用词，使提取出来名词以外富含感性色彩的词汇。对答复文本作情感分析。即可从感情的角度来判断答复的执行情况。这也和实际情况相符。对与留言和答复，反映了留言人的“痛点”与事情可行性的难易。且都是口语化的，口语化更易体现出感情色彩。

3.结果分析

3.1 问题 1 结果分析

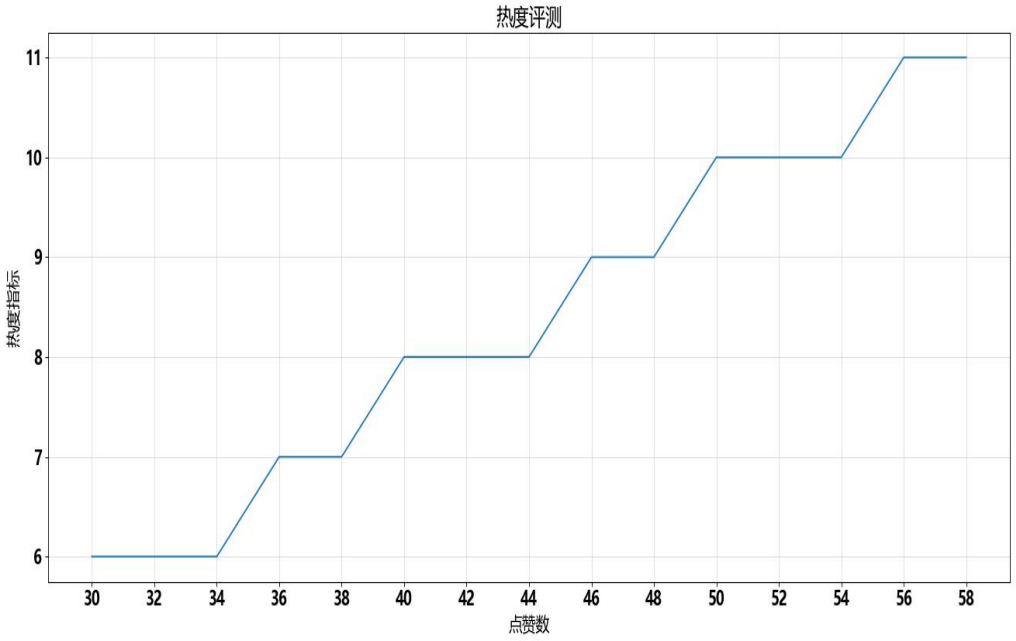
通过表 1 可以看出，基于卷积神经网络的留言分类算法准确率很高，并且稳定。这是因为一方面卷积神经网络模型可以通过增加卷积核来提取到更丰富的分类特征，另一方面还可以通过增加卷积层数来提取相更高层次的分类特征。

分类类型	TextCNN 卷积神经网络分析
城乡建设	0.81
党务政务	0.89
国土资源	0.92
环境保护	0.93
纪检监察	0.76
交通运输	0.83

表 1

3.2 问题 2 结果分析

根据问题附件 3 数据,进行筛选，清洗。有所给出热度评价指标作出评测，可得以下结果：



对于热度问题排名可给出如下结果：

问题ID	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
1	208636	A00077171	A市A5区汇金路五矿万	2019/8/19 11:34:04	我是A市A5区汇金路五矿万境K9县	0	2097
2	223297	A00087522	反映A市金毛湾配套入	2019/4/11 21:02:44	书记先生：您好！我是梅溪湖金毛	5	1762
3	220711	A00031682	请书记关注A市A4区5	2019/2/21 18:45:14	尊敬的胡书记：您好!A4区p2p公司	0	821
4	217032	A00056543	严惩A市58车贷特大集	2019/2/25 9:58:37	胡市长：您好！西地省展星投资有	0	790
5	194343	A000106161	承办A市58车贷案警官	2019/3/1 22:12:30	胡书记：您好！58车贷案发，引发	0	733

3.3 问题 3 结果分析

3.3.1 相关性结果

通过对有价值文本进行全模式分词，利用余弦相似度计算答复的相关性。从得到的可视图中可以看出，答复的相关性基本上介于 0.9 到 1 之间，可以看作相关性是很高。结果与实际相符，因为对于网上问政平台，回答一定是有针对性的。

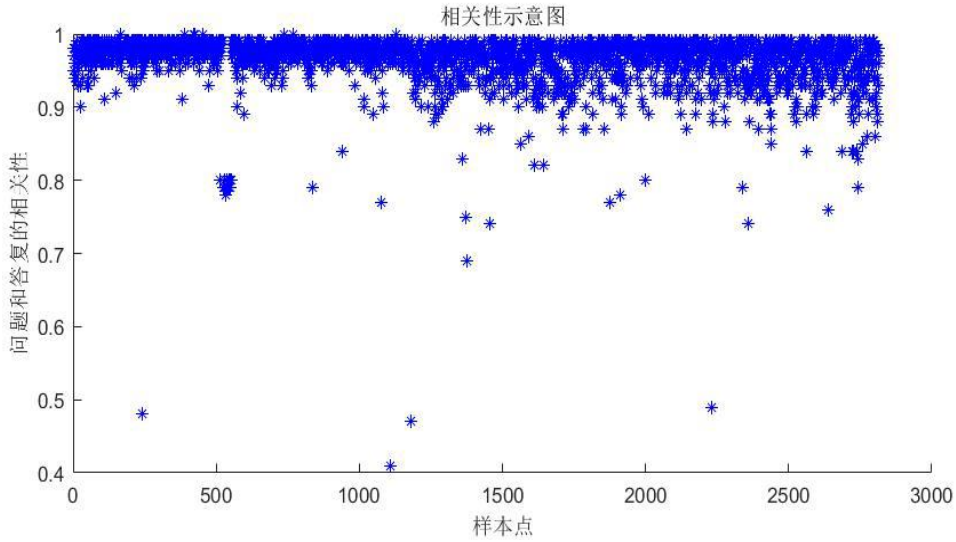


图 相关性示意图

3.3.2 完整性结果

通过对有价值文本进行精确模式分词，去掉停用词和停用词性，再利用 IF-IDF 计算关键词权重并排序，其后抽取文本长度的 30%的关键词得到如下图结果。从该可视图可以看到答复的完整性在 0 到 1 之间都有分布。

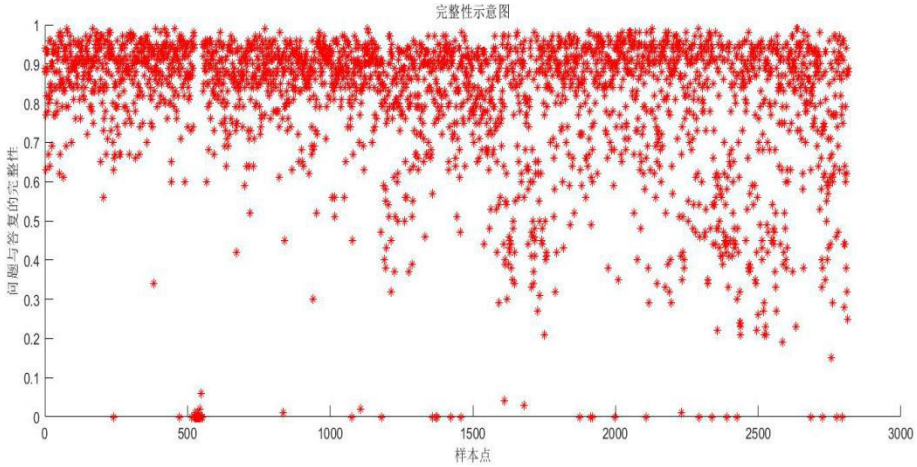
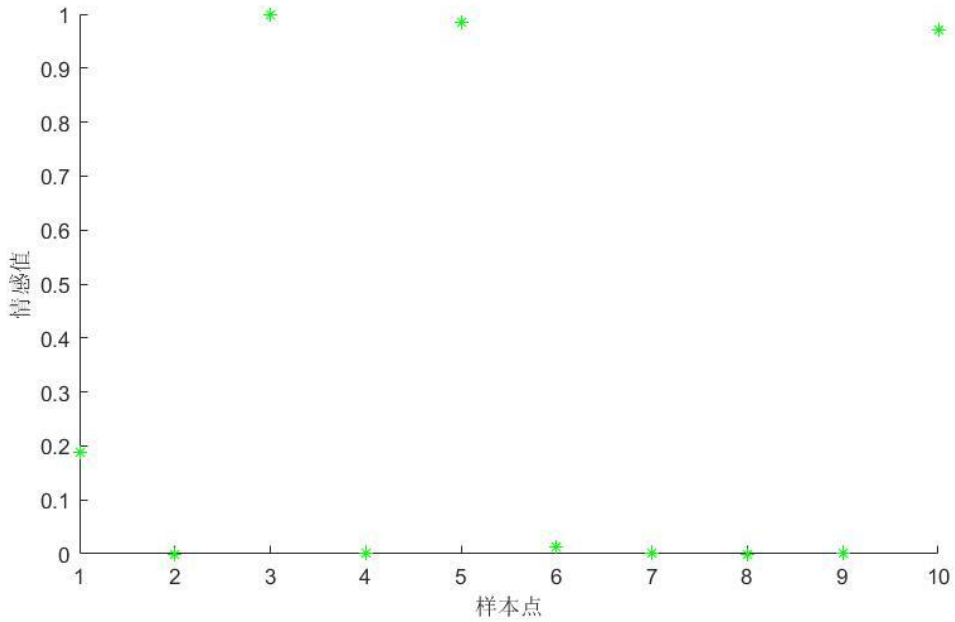


图 完整性示意图

3.3.3 可解释性结果

通过对答复文本剔除留言文本的关键字和停用词，使之保留尽可能多的感情色彩的词汇。建立可解释性方案的模型，以感情色彩来评

价答复的可解释性。得到下图结果。抽取其中部分答复文本测试结果，可以看出由于对可解释性的假定只有可解释性与不可解释性两种划分标准，可解释性呈现是两种极端的。为进一步优化提供数据。



3.3.4 评价方案

针对答复意见的质量，这里从网民的角度考虑。对于答复，在具有相关性的前提下，才有完整性和可解释性之说。相关性、完整性和可解释性的关系如图所示：

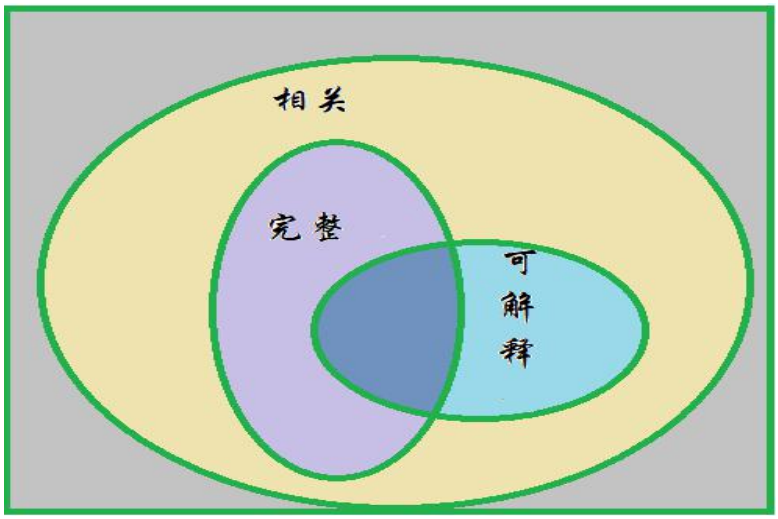


图 关系图

对于网友来说，最想要的答复应该是即相关又完整并且可解释性

强。因此由关系图得出一套答复质量评价方案：

$$best(i) = cor(i) \cap com(i) \cap exp(i),\ i = rang(1, nrows + 1)$$

best(i)：第 i 条答复的最佳答复值

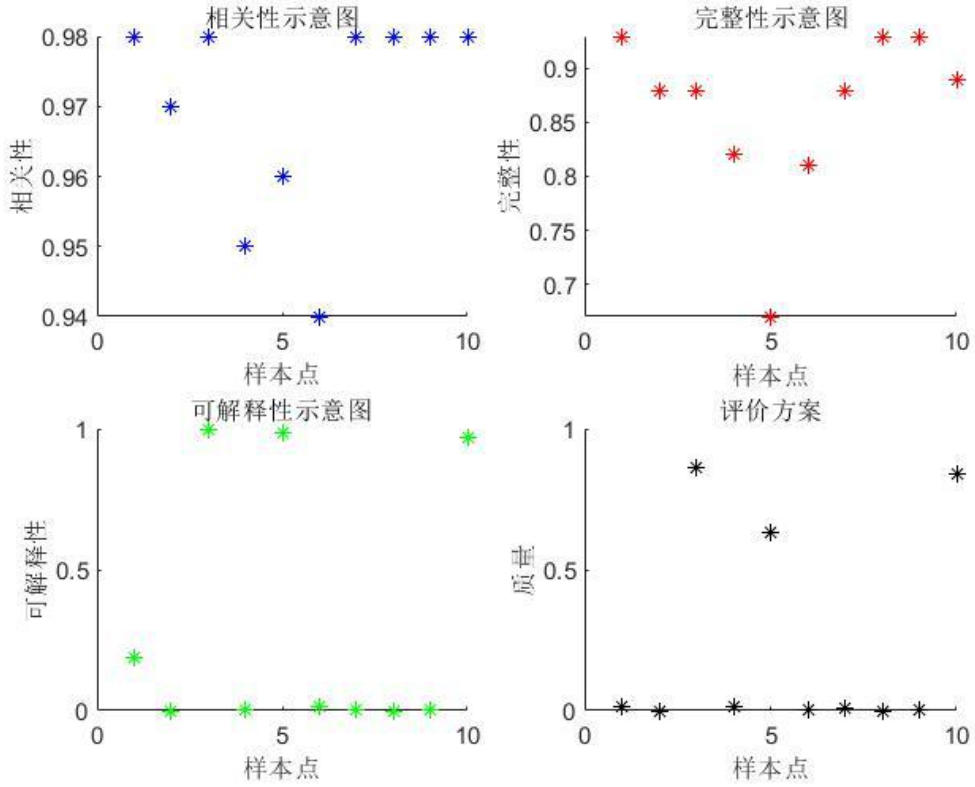
exp(i)：第 i 条答复的可解释性

cos(i)：第 i 条答复的相关性

nrows：excel 行标

com(i)：第 i 条答复的完整性

从最终的答复质量评价方案中，我们可以得出可解释性的影响是非常高的。网民最终想得到的也是最具可解释性的答复。从这个方案中也可以得出，政府部门对于回答网民的提问，应从“对症下药”，下功夫解决网民的问题着手。给网民满意的答复



4. 参考文献

[1] https://github.com/fxsjy/jieba

[2] 张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化.西南科技大学学报.2020-04-16

[3]曾凡锋,李玉珂,肖珂.基于卷积神经网络的语句级新闻分类算法.计算机工程与设计.2020-04-16

[4]李丽华,胡小龙.基于深度学习的文本情感分析.湖北大学学报(自然科学版).2020-03-05

[5]万磊,张立霞,时宏伟。基于 CNN 的多标签文本分类与研究.现代计算机.200-03-15

[6]姚佳奇,徐正国,燕继坤,熊钢,李智翔.基于标签语义相似的动态多标签文本分类算法.计算机工程与应用.2020-03-25

[7]冯梦莹,李红.文本卷积神经网络模型在短文本多分类中的应用.金融科技时代.2020-01-10

[8]江海戬,辛立强.卷积神经网络在博客多标签中的应用.工业控制计算机.2019-12-25

[9]曹鲁慧,邓玉香,陈通,李钊.一种基于深度学习的中文文本特征提取与分类方法.山东科学.2019-12-25

[10]陈巧红,王磊,孙麒,贾宇波.卷积神经网络的短文本分类方法.计算机系统应用.2019-05-15