# C 题

**摘　要：**随着现今科技的飞速发展，自然语言处理（NLP）技术也在不断发展前进，从开始基于传统机器学习的算法，到现在众多的深度学习算法，这些技术早已渗入了人们的日常生活之中。当下是一个智能的时代，为了提升政府的办公效率，提出和构建了“智慧政务”的概念及平台，而本文主要是对“智慧政务”中的一些文本挖掘任务进行建模，使用 NLP 中的一些技术来对文本分类、热点检测、回复质量评价等任务进行一定程度的挖掘和处理。本赛题主要包括以下三个子任务：（1）文本的多分类任务。对此我们采用了 Doc2vec、PA-II、SVM、FastText、XGBoost 等技术，模型评价采用十折交叉验证的方法，最终分类质量评价指标宏 $F_1$ 值达到 91%；（2）热点话题检测及信息抽取任务。其中热点话题检测基于词嵌入模型，采用增量式聚类 Single-Pass 算法，并在传统算法上进行了几点创新改进：新增向量中心以代替簇中心、新增标题与正文的权重参数、长短留言权重自适应调整等，另外热度指数计算也根据留言环境，如类簇的大小、点赞数、反对数，采用改进的威尔逊区间算法。信息抽取则根据 BIO 标准标注实体，采用双向长短期记忆网络（Bilstm）与条件随机场（CRF）的方法进行命名实体提取继而建立特征向量完成对目标内容的提取。最终效果双双显著；（3）对政府的回复进行评价。由于留言事件内容的复杂性，政府部门回复的多样性，使得问题变得更为复杂。本文将围绕出题方阐述的相关性、完整性、可解释性三个维度进行以下分析：（1）从文本“形”与“意”的角度来判别相关性；（2）对于完整性与可解释性，通过深入的学习，对完整性和可解释性进行定义，将其转换为二分类问题并使用 Sigmoid-Fitting 将分类模型输出转化为概率值，从概率角度看待完整性与可解释性；（3）结合前面三个因素的得分值，最后生成留言回复的评估值。

**关键词 ：**Doc2vec；SVM；FastText；XGBoost；文本多分类；话题检测；Single Pass；事件信息抽取；Bilstm+CRF

# C Question

**Abstract：** With the development of science and technology,natural language processing (NLP) technology is also constantly developing. From algorithms based on traditional machine learning to many deep learning algorithms, these technologies have long penetrated into people's lives.The present is an era of intelligence, in order to promote the government office efficiency, introduced a concept of "political wisdom" and platform, and some of this article is mainly to "e-government" wisdom text mining task modeling, using some of NLP techniques to reply on text categorization, hot spot detection, quality assessment tasks such as mining and processing.This topic mainly includes three sub-tasks :(1) text multi-classification task.We adopted doc2vec, pa-ii, SVM, FastText, XGBoost and other technologies, and adopted the method of 10-fold crossing verification for model evaluation. The macro F1 value of classification quality evaluation index reached 91%;(2) topic detection task and information extraction task.Based on the word embedding (doc2vec) model, we adopted the incremental clustering Single Pass algorithm and improved it. Considering the high real-time performance of the message, the weight of the message title and the body, and the difference in the similarity between the message length and length, we achieved good results in the topic detection.For the text summary, we search for the appropriate message in the cluster as the summary of the cluster.Heat index, we consider the size of class cluster, the number of thumb up, the number of opposition, and finally adopt the improved Wilson interval algorithm as the definition of heat.In terms of information extraction and extraction of population and location, we annotated the entities in the training set according to BIO standard, used the methods of Bilstm and CRF to extract named entities, and then set up feature vectors to extract locations and populations, resulting in remarkable results;(3) evaluate the government's response.The complexity of people's comments, and the diversity of government responses, makes the problem even more complicated.This paper

focuses on the concepts of relevance, completeness and interpretability required by the author :(1) to distinguish relevance from the perspective of "form" and "meaning";(2) as for completeness and interpretability, we defined them through in-depth study, converted them into a dichotomy problem, and used sigmoid-fitting to convert the output of the classification model into a probability value, so as to view completeness and interpretability from a probability perspective;(3) combine the score value of the above three factors, and finally generate the evaluation value of message reply.

**Key words**：Doc2Vec；SVM；FastText；XGBoost；Text Categorization；Topic Detection；Single Pass；Event Information Extraction；Bilstm+CRF

## 1 赛题分析

随着政府对人民生活、意见等方面的日益重视，单单依赖传统的人工进行留言划分和热点整理，其时间、经济上的代价都比较大。同时因为智慧城市的推广，智慧政务又是智慧城市的“基石”，因此用先进的科技建立智慧政务是一项有意义的事。

对于本次赛题，我们将子任务分别定位为：文本分类、话题检测、信息抽取、文本匹配。

文本分类问题一直是 NLP 领域的一个研究热点，本次分类的对象是留言，一般的微博留言属于短文本，但是经过分析发现赛题附件二所给留言经过分词后，文本长度分布情况如下图 1、2，仍有多于 1/4 的留言长度超过了 300，最长者超过了 9000 个字，因此将赛题本文分类归纳为短文本分类问题也不准确。

```
> summary(len)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   15.0   101.0   173.0   303.3   346.8  9164.0
```

图 1 赛题数据分析

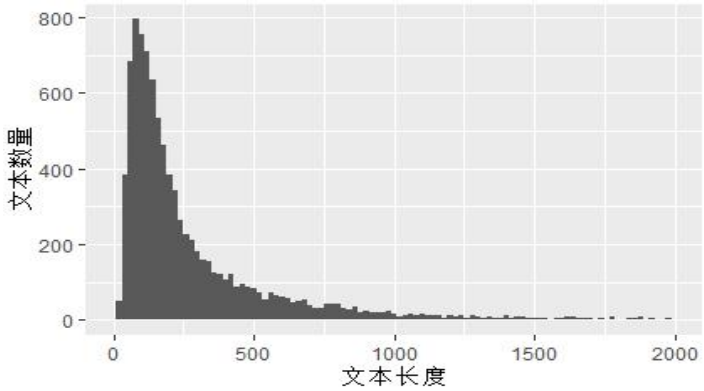


图 2 赛题数据分析

经过分析，我们认为其难点如下：（1）考虑到应用场景是市长信箱，其往往需要较高的实时性，如何在获取较好的分类情况下，同时减少模型复杂度；（2）如何对长、短文本进行文本分类；（3）采用什么文本表示模型；（4）采用哪种文本分类算法。

话题检测（Story Detection）可以帮助人们快速聚焦热点话题、协助企业进行相关决策、帮助国家和政府部门监控社会舆论。结合比赛数据，分析其难点如下：（1）考虑到应用场景需要较高的实时性，因此模型的复杂度应该尽可能低；（2）话题检测主要方法有：基于聚类的方法、基于主题模型的方法，选择哪种方法适合；（3）话题检测后如何获得留言摘要，是自动生成还是进行抽取获得；（4）热度指数的定义，排名算法有很多，往往要结合环境确定合适的算法，考虑点赞数、反对数、聚类条数、时间，最终如何定义热度指数。

信息抽取基于命名实体识别（NER）任务。可以帮助人们从海量信息中提取出对自己有利的信息。在官网的赛题解析中，主讲人认为第二问提取地点/人群是一个 NER 任务，但在我们考察数据后发现，如果将其视为 NER 任务，得到的实体往往不能很好的抓住事件的主题，得到的效果不好。我们分析认为其难点如下：（1）人群的定义很模糊；（2）观察赛题后，发现取而代之具体的真实地点、其赛题中信息都是经过脱敏了的，其操作加大了 NER 难度，选择什么样的工具可以很好的提取出赛题数据中的地点；（3）、提取出这些实体，如何确定事件的地点、事件的人群。

文本匹配任务是要对政府的回复进行评估。我们认为这是一个文本匹配的任务，广义的讲，也就是研究用户提问与政府工作人员回复之间的关系。其难点主要体现在：（1）观察数据后，发现留言内容是较为复杂的，比如：小孩上学问题就有因为证书等导致无法上学、交通不便等，真正相似的问题非常少，建立知识库难度缺乏大量的数据；（2）根据出题方给出的相关性、完整性、可解释性，如何量化它们。

论文组织结构如下：

第 1 部分是赛题的分析；第 2 部分是比赛涉及的相关技术介绍和评价指标说明，主要有文本分类与其评估标准、话题检测与其评估标准、热度公式、信息抽取、文本匹配；第 3 部分是我们所做的实验与实验评估；第 4 部分是总结与展望。

## 2　赛题相关技术与评价指标

我们的工作主要围绕赛题的三个问题展开。

### 2.1　文本分类

第一问定位为文本分类。对于文本分类，主要流程如图 3 所示。

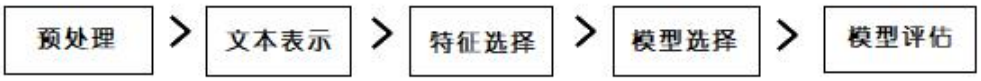


图 3 文本分类任务流程

首先文本表示，常见的文本表示方法有：向量空间模型（VSM）[1.]、布尔模型（Boolean Model）[2.]、概率模型（Probabilistic Model）、词嵌入模型[3.][4.]。对于 VSM，其认为每一篇文档可用多个特征（词）进行表示，通过特征将文本映射到向量空间中，而从将文本向量化，从而可以计算文本之间的相似度等。布尔模型仅仅考虑特征是否出现，对特征重要程度不关心。但是前者们都丢失了大量语义信息，比如：“人控制机器人”，“机器人控制人”，经过切词后，以 VSM 或者布尔模型表示的话，其对于向量是相同，但是明显其语义不同。而词嵌入的方式，在这方面则优于上述两者，但是词嵌入方式往往可解释度不高。

在特征选择方面，如果是采用词嵌入方式，则不需要考虑特征选择问题，因为往往词嵌入转换得到的向量已经不属于高维空间了，而特征选择主要是为了避免（1）文本表示时，向量过于稀疏，计算资源问题；（2）过拟合问题；（3）噪音干扰问题。常见的特征选择方法有：词频（TF）、文档频数（DF）、互信息（MI）、低损降维法（LLDR）、频率差法（RFD）、卡方统计（CHI）、信息增益（IG）、DFCTFS[5.]。我们在复旦大学的中文文本分类语料库[6.]上对这些特征选择方法进行实验对比，分类器采用中心向量算法(Rocchio)，评价指标为宏 $F_1$ 值，发现 DFCTFS 与 CHI 相对稳定，具体结果如图 4 所示。

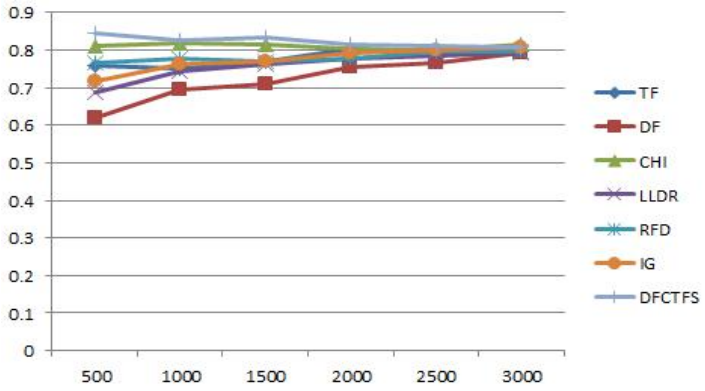


图 4 特征选择算法与其宏 $F_1$ 值（其中横轴是特征维度）

在模型选择上，常见的分类算法如表 1 所示。

表 1 常见分类算法

| | |
|---|---|
| 基于统计方法 | 朴素贝叶斯（NB）、回归模型、支持向量机（SVM）等 |
| 基于连接方法 | 神经网络（FastText、RNN、LSTM 等）等 |
| 基于规则方法 | 决策树（Decision tree）等 |
| 基于集成方法 | Boosting（GBDT、XGBoost）、Bagging 等 |

其中回归模型在数据的分布为线性可分下，效果不错，运算简单，其应用比如垃圾邮件过滤问题上。SVM 是一种基于统计的学习方法[7.]，是 Vapnik 在 1995 年提出，2005 年前还是文本分类上的“一哥”，其主要思想是通过核函数将输入向量映射到一高维空间，在该空间中寻找一个最佳超平面，以区分不同类别样本，该算法考虑了样本点到决策面距离，使得其泛化能力很强。FastText 是 Facebook 于 2016 年开源的一个词向量计算词向量、文本分类工具，其通过训练浅层神经网络，但在文本分类上，其分类效果却可以媲美深度神经网络。循环神经网络（RNN）是为了处理序列数据而生，但面对长序列，则会出现梯度消失和梯度爆炸的问题。相比普通 RNN，长短期记忆网络（LSTM）在不仅仅接收当前的输入，而且接收上一个状态的传递，因此，面对长序列，其效果往往优于普通 RNN。有时候，我们试了很多分类器也不能很好的处理某分类问题，基于集成模型的方法不妨一试，其主要思想是通过训练多个弱分类器，通过策略将其组合，得到最终模型。

在文本分类方面，我们一般使用查准率、查全率、$F_1$ 值来对分类效果进行一个衡量，这里是多分类任务，根据赛题文件，我们使用宏 $F_1$ 值对模型分类效果进行评估，如公式 1 所示。

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \qquad \text{公式 1}$$

其中 $P_i$ 为第 $i$ 类的查准率，$R_i$ 为第 $i$ 类的查全率。

## 2.2 话题检测

话题检测（Story Detection）主要任务是发现数据库当前未知的新话题以及检测其相关报道[8.]。话题识别的研究始于 1996 年，由美国国防高级研究计划委员会提出，话题可以认为是某一个事件。基于聚类的方法其主要思想是将文本进行聚类，从而获取话题，其代表有划分聚类，如：Kmeans 算法等、增量式聚类，如：Single-Pass 算法[9.][10.]等、层次聚类、密度聚类等。而主题模型思想是先对文本构建主题模型，再根据主题模型对文本进行聚类，从而识别话题的类别。代表有隐含狄利克雷分布特征（Latent Dirichlet

Allocation，LDA）的检测方法。

在话题检测方面，我们采用了误检率、查全率来对聚类结果进行一个衡量，如公式 2、3 所示。

$$F = \frac{C}{B + C} \quad \text{公式 2}$$

其中 $F$ 表示误检率，$B$ 表示聚类某一相关话题的留言数量，$C$ 表示聚类某一相关话题内不相关的留言数量。

$$R = \frac{B}{A} \quad \text{公式 3}$$

其中，$R$ 表示查全率（召回率），$A$ 表示实际某一相关话题的留言数量，$B$ 表示聚类某一相关话题的留言数量。

### 2.3 热度公式

热度公式，也就是排名算法，其算法内考虑的因素通常是：点赞数、反对数、时间。这方面可以借鉴许多知名的论坛，比如：Reddit、Stack Overflow、知乎等，有名的排名算法有牛顿冷却定律、威尔逊区间等。其中牛顿冷却定律公式为：$T(t) = T(t_0) \times e^{-k(t-t_0)}$，在热度值的定义上，$T(t_0)$ 为初始得分，$k$ 为热度下降指数，$t - t_0$ 为时间差，可以认为留言热度值随着时间流逝而成反比。对于威尔逊区间方法，其思想是通过计算每个留言好评率（赞成票的比例）的置信区间，以其下限作为得分，进而排名。

### 2.4 信息抽取

信息抽取任务是 NLP 领域一个热点问题，也是难点问题。而根据赛题要求，抽取出地点、人群是属于信息抽取中事件语义信息提取。一般我们通过 5W1H[11.]（When，Where，Who，What，Whom 和 How）来对一个事件进行表达。在一些针对新闻中的语义信息的研究上，Wang W 等人利用新闻标题的信息抽取新闻文本中的主题句，再利用基于规则、SVM 等方法抽取事件的 5W1H 元素[12.]，Wang W 又提出了关键事件识别算法，使用语义角色标注（SRL）抽取中文新闻中 5W1H 元素[13.]，郑立洲[11.]通过观察微博文本的特点，提出了词语聚类与链接的方法，在微博文本上抽取事件语义元素取得不错效果，同时给我们一定的启发。

### 2.5 文本匹配

文本匹配任务，其具体任务是：从大量存储的 doc 中，选取与用户输入 query 最匹配的 doc。通常在智能问题中，doc 也就是检索系统中存放着的问题，query 对应着用户的问题，最匹配意味着用户问题与检索系统中的问题相关度很高，随着深度学习的广泛应用，这方面已经做得很好了[14.]。本赛题第三问与智能问题系统不同，对于智能问答系统，相关度大的 doc 往往有很相关的答案，比如“天为什么是蓝色的？”，该答案是静态的，不受外界因素影响，而本赛题数据不同，里面的留言即便是小孩上学，也可以分为若干小类，其答复也是受时间、地点等因素影响的，今天的政策也许下个月就变化了，A 地政策也不适合与 B 地。同时，出题方提及的相关性、完整性、可解释性，赛题讲解人认为相关性即答复内容是否与问题相关、完整性即回复是否满足某种规范（比如开头是否有称呼，结尾是否结束语）、可解释性即答复意见中内容是否有相关的解释（比如回复是否依据某条法律法规）。通过查阅资料，我们发现对于文本相关性，国内

研究者往往还是从相似性度量入手，对于文本相关性方面研究较少，而完整性、可解释性方面资料就更少了。综合这些因素，使得本小题成为了赛题中的难题。

## 3　实验与评估

### 3.1　任务 1：文本分类

预处理后，我们发现数据脱敏，导致分词精度下降，于是通过正则表达式匹配，添加一些词进入用户词典中。对于停用词，通过向量空间模型（VSM）表示文本，采用多种特征选择方法（卡方独立性检验（CHI）、信息增益（IG）、DFCTFS 等）取交集得到最后对分类影响非常小的词，将这类词添加到停用词中，避免了人工去寻找词作为停用词的不便。经过对数据进行清洗后，我们将要考虑文本的表示问题，我们首先使用的是传统 VSM 的方式，采用 CHI 作为特征选择算法，从而达到去噪音，降维度，提升分类效果。

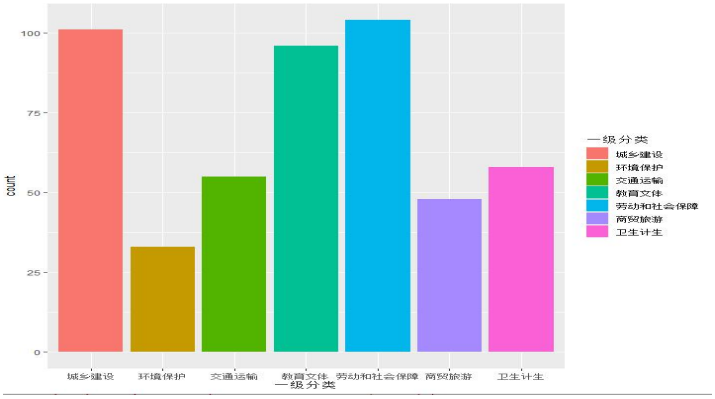


图 5 赛题数据分布图（横轴为类别，纵轴为频次）

由于最初赛题数据分布如图 5 所示，其分布是不均衡的。因此，我们进一步对部分类别的数据进行了增强，颠倒留言语序、机器翻译、删除留言内的信息、增加留言内信息、利用爬虫技术从相应政府部门市长信箱上爬取了一些数据，最后将增强的数据进行整合，人工对数据进行分类。

我们先尝试线性分类器： PA-II 算法，该算法特点是：在线学习、错误驱动、考虑样本点与决策面之间距离，不仅计算简单，而且有一个很好的泛化能力，抗干扰性强[15.]。采用分层抽样，提取出训练集（70%）、验证集（20%）、测试集（10%），利用三折交叉验证确定超参数，在测试集上进行测试，采用 1 对多的方式，在部分类别上 $F_1$ 值可以达到 90%以上，但是在某些类别上就逊色很多，宏 $F_1$ 值（macro $F_1$）达到了 82%。我们猜想数据分布应该不是线性可分的超平面，遂使用 SVM，分别采用多项式核函数、高斯核函数，利用三折交叉验证确定超参数，采用一对一（One-Versus-One，OVO）的多分类方式，在测试集上进行测试，宏 $F_1$ 值达到 90%，分类情况如图 6 所示，其中 1-7 标记分别对应着赛题数据集提供的七个类别：城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生。

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.84 | 0.90 | 0.87 | 378 |
| 2 | 0.94 | 0.96 | 0.95 | 189 |
| 3 | 0.87 | 0.86 | 0.86 | 139 |
| 4 | 0.94 | 0.95 | 0.94 | 288 |
| 5 | 0.94 | 0.93 | 0.94 | 405 |
| 6 | 0.90 | 0.84 | 0.87 | 255 |
| 7 | 0.94 | 0.89 | 0.92 | 188 |
| accuracy | | | 0.91 | 1842 |
| macro avg | 0.91 | 0.90 | 0.91 | 1842 |
| weighted avg | 0.91 | 0.91 | 0.91 | 1842 |

图 6 SVM 算法分类情况

我们猜想会不会留言长短对分类有一定影响，遂剔除留言长度大于留言长度中值的那些留言，再进行分类测试，结果宏 $F_1$ 值几乎没有改变。

尝试了一些传统的机器学习方法后，我们试了试使用词嵌入的文本表示，使用 doc2vec 技术将文本转换成了低维度的向量，再次使用 SVM，利用三折交叉验证确定超参数，采用一对一（One-Versus-One，OVO）的多分类方式，在测试集上进行测试，宏 $F_1$ 值达到 86%。

尝试了基于浅层的神经网络 FastText、基于集成模型的 XGBoost 方法，将文本转换为指定的训练格式，经过交叉验证的方式确定超参数后，在测试集上进行测试，宏 $F_1$ 值分别达到 90%、90%，分类情况分别如图 7、图 8 所示。

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.85 | 0.92 | 0.88 | 378 |
| 2 | 0.92 | 0.94 | 0.93 | 189 |
| 3 | 0.88 | 0.83 | 0.86 | 139 |
| 4 | 0.93 | 0.95 | 0.94 | 288 |
| 5 | 0.93 | 0.94 | 0.94 | 405 |
| 6 | 0.90 | 0.84 | 0.87 | 255 |
| 7 | 0.97 | 0.86 | 0.91 | 188 |
| accuracy | | | 0.91 | 1842 |
| macro avg | 0.91 | 0.90 | 0.90 | 1842 |
| weighted avg | 0.91 | 0.91 | 0.91 | 1842 |

图 7 FastText 算法分类情况

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.84 | 0.91 | 0.87 | 378 |
| 2 | 0.92 | 0.94 | 0.93 | 189 |
| 3 | 0.87 | 0.83 | 0.85 | 139 |
| 4 | 0.96 | 0.97 | 0.96 | 288 |
| 5 | 0.95 | 0.93 | 0.94 | 405 |
| 6 | 0.89 | 0.83 | 0.86 | 255 |
| 7 | 0.94 | 0.89 | 0.92 | 188 |
| accuracy | | | 0.91 | 1842 |
| macro avg | 0.91 | 0.90 | 0.90 | 1842 |
| weighted avg | 0.91 | 0.91 | 0.91 | 1842 |

图 8 XGBoost 算法分类情况

考虑到三个分类器的优缺点，我们决定综合 SVM、FastText、XGBoost 三个分类器得到最终的模型，采用投票的方法、观察学习算法对数据的归纳偏好（我们发现向量空间的维度在 2000 的时候，SVM 对交

通运输类的查准率会高于其他两个分类器）调整集成模型的参数，具体参数如表 2 所示。

表 2 集成模型参数表

| XGBoost | | SVM | | FastText | |
|---|---|---|---|---|---|
| 参数名 | 值 | 参数名 | 值 | 参数名 | 值 |
| max_depth | 6 | Kernel | rbf | Lr | 1 |
| n_estimators | 500 | C | 5 | minCount | 3 |
| colsample_bytree | 0.8 | Gamma | scale | wordNgrams | 2 |
| subsample | 0.8 | decision_function_shape | OVO | Dim | 128 |
| nthread | 4 | | | Ws | 3 |
| learning_rate | 0.1 | | | Epoch | 300 |

集成模型具体分类情况如图 9 所示，可以看到集成模型的宏 $F_1$ 值达到了 91%。

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.88 | 0.90 | 0.89 | 382 |
| 2 | 0.91 | 0.94 | 0.93 | 178 |
| 3 | 0.93 | 0.83 | 0.87 | 122 |
| 4 | 0.95 | 0.96 | 0.96 | 321 |
| 5 | 0.93 | 0.94 | 0.93 | 419 |
| 6 | 0.93 | 0.88 | 0.90 | 258 |
| 7 | 0.89 | 0.90 | 0.89 | 162 |
| accuracy | | | 0.92 | 1842 |
| macro avg | 0.92 | 0.91 | 0.91 | 1842 |
| weighted avg | 0.92 | 0.92 | 0.92 | 1842 |

图 9 集成模型分类情况

### 3.2 任务 2：热点问题挖掘

考虑到市长信箱的实时性比较高，我们希望通过一种简单的模型达到热点问题挖掘。我们首先尝试通过文本聚类的方法来应对话题检测的任务，在预处理方面，我们并不是使用第一问通过特征选择得到的停用词，因为有些词对文本分类影响不大，但是其信息却是我们所需要的，比如各种地点名等，因此，我们重新构建了停用词，主要是一些副词等；在文本表示方面，采用词嵌入（doc2vec）的方式，因为词袋模型会丢失掉大量语义；在算法选择上面，我们根据数据的特点改进了传统的 SinglePass 算法[16.]，具体改进做法：（1）增加了向量中心，以向量中心代替簇中心；（2）根据赛题数据特征，对留言标题和内容设置了不同的权重；（3）权重会根据留言长度自适应调整。

我们随机在出题方提供的附件三中随机抽取了 10 个话题，使用召回率、误检率对话题检测效果进行评估。对于话题检测算法的参数，通过大量的实验，最后设置相似度阈值对 0.6，留言标题的权重为 0.3，话题检测评估效果如表 3、图 10 所示。

表 3 话题检测效果

| 话题语料库 | 聚类相关留言个数 | 聚类不相关留言个数 | 召回率 | 误检率 |
|---|---|---|---|---|
| A 市万科魅力之城商铺无排烟管道，小区内到处油烟味 | 3 | 0 | 1 | 0 |
| A1 区辉煌国际城二期居民楼下商铺违法开饭店，维权近三月没有用 | 5 | 0 | 1 | 0 |
| 咨询 A3 区西湖街道茶场村五组的拆迁规划 | 9 | 0 | 0.9 | 0 |
| 居住证未满一年，小孩不能在 A 市读小学？ | 2 | 0 | 1 | 0 |
| A7 县星沙中贸城欺诈业主、拖欠业主资金不退还 | 3 | 0 | 1 | 0 |

| | | | | |
|---|---|---|---|---|
| 带着情怀提升 A 市规划建设水平，带动经济发展 | 5 | 0 | 1 | 0 |
| A 市能否设立南塘城轨公交站？ | 4 | 0 | 1 | 0 |
| A 市便民服务桥办理居住证一点也不便民，望处理！ | 2 | 0 | 1 | 0 |
| A7 县违建之风盛行谁来管？？ | 3 | 0 | 1 | 0 |
| 坚决反对在 A7 县诺亚山林小区门口设置医院 | 5 | 0 | 1 | 0 |

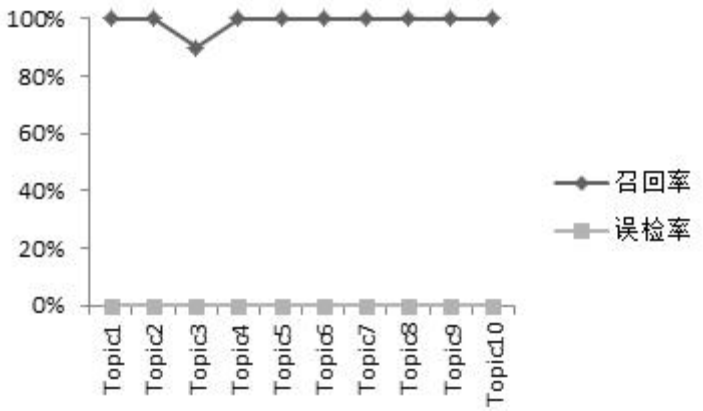


图 10 话题检测效果

在热度值定义问题上，通过观测数据，我们确定如下几个因素：类簇大小、时间、点赞数、反对数。由于任务是市长信箱，我们认为一个留言所反映的问题的影响不应时间的流逝而改变，如果一个留言所反映的问题一直没被解决，那么它的影响就一直存在，因此我们排除了时间的因素，剩下类簇大小、点赞数、反对数三个因素。我们通过对威尔逊区间算法进行调整，得到了适合赛题场景的热度值公式如公式 4 所示。

$$Score_i = \frac{(p_i + \frac{z^2}{2n_i} - \frac{z}{2n_i}\sqrt{4n_i(1-p_i)p_i + z^2})}{1+\frac{z^2}{n_i}} \quad \text{公式 4}$$

其中 $p_i$ 为某类簇 i 的赞同率， $z$ 值取 1.96， $n_i$ 为某类簇中留言数量。在文本摘要方面，只要选择每一个类簇中与类簇中心相似度最高的向量作为文本摘要即可。

在地点、人群提取方面，我们考虑，就地点的问题，方案（1）如果是一个留言，可以先使用命名实体识别（NER）方面的工具提取出地点，最后从若干地点中提取与事件最相关的实体名；方案（2）如果是一个类簇，我们可以通过先对地点聚类以消除中文语言随意性，导致地点描述上的不一致的问题，再根据词频等因素进行筛选。通过观察数据，我们发现，赛题中市长信箱留言并不像微博文本那么随意，语言还是相对正式一些，因此我们选择了方案（1）。那么如何提取这些命名实体便是不可避免的问题，常用的 NER 工具有：Stanford NLP、pyhanlp、FoolNLTK 等。经过测试，我们发现 FoolNLTK 对实体提取，其准确率还不错，但是由于赛题数据已经经过了去敏操作，其命名实体识别的准确率受到一定影响，遂我们通过对赛题数据随机抽取部分留言，采用 BIO 规范，进行人工标注，最后使用双向长短期记忆网络结合条件随机场的方法进行 NER，在地点、人群提取上，效果显著。

对于提取出来的地点、人群，我们需要对其进行筛选，得到与事件相关的实体。在这方面，通过观察数据，我们设计了多级特征的方法，而提取人群的问题与之类似，只是不同在于，人群的定义很模糊，增

加了本题的难度。

具体提取方案如下：

我们对留言中提取出的命名实体赋予两个特征向量，位置特征向量、上下文特征向量。对于某一个实体，其都对应了这两个特征向量，其权值分配如表 4 所示。

表 4 实体位置的特征权值

| | 标题 | 开头 | 内容 | 结尾 |
|---|---|---|---|---|
| 权重 | 4 | 2 | 1 | 2 |

对于上下文特征，这里需要分类讨论，对于地点，其权值分配如表 5 所示。

表 5 地点实体上下文特征权值

| | 在… | 。..或，…或：.. | 来自… | … | 关于… | 是… |
|---|---|---|---|---|---|---|
| 权重 | 2 | 2 | 2 | 4 | 2 | 3 |

对于人群的上下文特征，其权值分配如表 6 所示。

表 6 人群上下文特征权值

| | 系… | 叫… | 我是… | 作为… | 我们是… | 本人… | 代表… | 我为… |
|---|---|---|---|---|---|---|---|---|
| 权重 | 2 | 3 | 4 | 2 | 4 | 3 | 2 | 2 |

举个例子：比如留言标题“魅力之城的路灯坏了！”那么对于地点：魅力之城，其位置特征如表 7 所示。

表 7 “魅力之城的路灯坏了”的位置特征

| | | | |
|---|---|---|---|
| 1 | 0 | 0 | 0 |

其上下文特征如表 8 所示。

表 8 “魅力之城的路灯坏了”的上下文特征

| | | | | | |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 | 0 |

最后计算地点实体魅力之城的总分只要通过向量的内积即可，其得分为 8。

对于人群的计算也是如此，只是相较于地点，人群的定义比较模糊，我们对数据进行观察，发现有的留言出现的是具体的人名，有的是群体，比如“我叫李明”，“我们代表蓝天小区的居民”，对于人名，使用我们训练的 NER 工具可以有效地识别出它，对于群体“居民、员工、学生、司机等”，我们使用分词工具提取即可。最后的处理与地点实体处理类似，就不累述了。

最终效果如下图 11 所示，地点提取的比较详细，如果问题中有人群，将其提取出来了。

| 热点排名 | 问题ID | 热度指数 | 时间范围 | 地点/人群 | 问题描述 |
|---|---|---|---|---|---|
| 1 | 1 | 8 | 2019/01/30至2019/12/26 | A3区旺龙路 | 建议在A3区旺龙路增加人行斑马线 |
| 2 | 2 | 4 | 2019/02/26至2019/06/12 | 畔湖湾小学 | F市建工违规施工使我们无家可归还我房子 |
| 3 | 3 | 3 | 2019/06/18至2019/06/19 | A市地铁3号线星沙大道站/居民 | A市地铁3号线星沙大道站地铁出入口设置极不合理！ |
| 4 | 4 | 3 | 2019/06/20至2020/01/01 | A市生殖医学医院/患者 | 试管做出基因缺陷女婴，不幸家庭雪上加霜 |
| 5 | 5 | 2 | 2019/07/11至2019/12/21 | 黄兴路步行街仕弘教育A市晚报大厦八楼 | 举报A市仕弘教育等培训机构涉嫌欺诈 |
| 6 | 6 | 2 | 2019/07/15至2019/12/03 | A市A6区月亮岛街道 | A6区乾源国际广场停车场违章乱建现象严重 |
| 7 | 7 | 2 | 2019/06/23至2019/08/18 | A市A3区兰亭湾畔小区 | A市A3区兰亭湾畔小区违规开餐饮 |
| 8 | 8 | 2 | 2019/02/21至2019/07/17 | A7县恒基凯旋门万婴格林幼儿园办普高园/业主 | 关于A7县恒基凯旋门万婴格林幼儿园办普高园的咨询 |
| 9 | 9 | 2 | 2019/07/22至2019/11/13 | 丽发新城小区/业主 | 投诉A2区丽发新城附近建筑拌站噪音扰民 |
| 10 | 10 | 2 | 2019/04/16至2019/10/22 | A3区枳溪路丽美幼儿园门口路段/车主 | A3区枳溪路丽美幼儿园门口路段收费停车合法吗 |

图 11 话题挖掘任务效果图（选前 10 个话题）

对于提取出来的地点、人群，随机抽取 20 个留言，人工对机器提取的实体的进行验证，使用正确率作

为评价指标，具体评估结果如表 9 所示。

表 9 事件相关实体提取评估结果

| 地点识别正确率： | 人群识别正确率： |
|---|---|
| 0.8 | 0.85 |

### 3.3 第三问：政府留言评价

对于政府留言评价，我们围绕着相关性、完整性、可解释性。

首先是相关性，我们认为其体现于“形”和“意”上。“形”上，我们使用政府回复与留言主题的共现名词进行衡量，我们认为留言的主题就是留言标题，问题就简化为政府回复与留言标题的共现名词程度的衡量，我们定义如下公式 5 所示。

$$jac01 = \frac{len(A \& B)}{len(A)} \quad \text{公式 5}$$

其中 $A$ 为一个集合，该集合存放着标题内的名词。 $B$ 为一个集合，该集合存放着政府回复中的名词。

由于中文的博大精深，可能使得一些词同义，但是不同形。我们将其归为语义上的问题，我们欲采用词嵌入的方式解决这个问题，最终，从速度与性能综合考虑，我们选择使用 FastText 训练的词向量，将政府回复与留言主题的名词都转换为向量，我们定义如下公式 6 所示。其中 $compare(A,B,\theta)$ 选出的是一个集合，该集合内存放的是：一些标题内名词的子集，子集是通过标题内名词们与政府回复内名词们进行余弦相似度比较且余弦相似度大于阈值 $\theta$ 的集合。

$$jac02 = \frac{len(compare(A,B,\theta))}{len(A)} \quad \text{公式 6}$$

其次是完整性，我们将其定义为三个部分：（1）回复中有对该留言用户的问候和致意；（2）回复中有对该留言相关问题的探讨；（3）回复中有对该留言相关问题的解决方案。同时满足这三个部分的回复必定会包含某些特定关键词，对于这些关键词，我们注重的是有与无的区别，所以我们使用布尔模型（Boolean Model）来表示回复文本，完整性的评判也抽象成了一个二分类问题。基于各种分类算法的结果，我们采用线性核函数的支持向量机算法（SVM）进行分类，使用 Platt 利用 sigmoid-fiitting 将标准 SVM 的输出转换成后验概率的方法，我们将回复完整性的评估分数转换成该回复被判定为完整回复文本的概率，计算公式为公式 7 所示。

$$P(y=1|f) = \frac{1}{1+e^{(Af+B)}} \quad \text{公式 7}$$

其中 $f$ 为标准 SVM 的输出， $A$ 与 $B$ 是待拟合的参数。

因为回复的完整性必定包含对用户留言问题的解答，因此回复的完整性包含着部分的相关性，我们通过给相关性和后验概率权重来使两者结合起来作为一个回复的完整性，我们定义如下公式 8 所示。

$$Integrity = weight * Jac02 + (1 - weight) * P(y=1|f) \quad \text{公式 8}$$

其中 $weight$ 为指定的相关性的权重， $Jac02$ 为之前计算的相关性， $P(y=1|f)$ 为 SVM 分类的后验概率。

接着是可解释性，我们将其定义为三个部分：（1）回复有没有围绕留言的问题进行回复；（2）回复是否有经过调查、是否有根据法律条例、是否有数据来支撑。

因此，回复的可解释性我们也将其抽象为一个二分类问题，采用与回复的完整性相同的方式，计算得到分类的后验概率后，通过设定权重使相关性与后验概率结合来作为回复的可解释性的评估分数，因为可解释性中回复的相关性和回复的可靠性具有相同的重要性，所以我们取两者的均值作为可解释性的评估分数，我们定义如下公式 9 所示。

$$Escore = \frac{Jac02 + P(y=1|f)}{2} \quad \text{公式 9}$$

最后是政府留言的评价，我们认为无论是相关性、完整性还是可解释性都是非常重要的。因此，我们结合相关性、完整性、可解释性，使用调和平均而不是算术均值进行计算。我们从若干留言与回复中随机抽取了 5 个，利用我们的模型对政府回复进行一个粗略的评估，具体情况如下图 12 所示。可以看到第一个留言，对于政府回复的相关性，由于留言信息不完整导致政府回复没有抓住问题要点，其相关性自然就低，得分只有 1 分；对于政府回复的完整性，虽然回复表现出对用户问题的关切，但是由于完整性涉及一定的相关性，其得分也只有 68 分；对于政府回复的可解释性，由于信息不完整导致政府回复中不涉及一些法律法规与相关问题的解释，因此可解释性得分也只有 19 分；对于政府回复的综合得分为 2.8 分。第二个留言，对于政府回复的相关性，我们可以看到政府对留言标题的回复只提到了案件在侦察，可能是由于案件的保密性，这样的回复，无论从“形”还是“意”上，其相关性自然不高；对于政府回复的完整性，其得分为 74 分；对于政府回复的可解释性，由于问题的特殊性，可解释性也只有 52 分；对于政府回复的综合得分为 43 分。

| 留言标题 | 政府回复 | 相关性 | 完整性 | 可解释性 | 综合 |
|---|---|---|---|---|---|
| 举报有公司骗取加盟费 | 网友“UU0081202”您好！您的留言已收悉。现将有关情况回复如下：因您未留下具体联系方式，市工商局无法与您联系获取详细信息及证据，请您直接拨打市工商局合同监管处电话0731-0000-00000000进行反映，感谢您对我们工作的支持、理解与监督！2018年12月17日 | 0.01 | 0.68 | 0.19 | 0.03 |
| 举报A市芒果金融平台涉嫌诈骗 | 网友“UU0081480”您好！您的留言已收悉。现将有关情况回复如下：经查，您所反映的相关警情，已由银盆岭派出所立刑事案件侦查，案件正在侦办中。感谢您对我们工作的支持、理解与监督！2018年12月29日 | 0.27 | 0.74 | 0.52 | 0.43 |
| 建议增开A市261路公交车 | 网友“UU0081227”您好！您的留言已收悉。现将有关情况回复如下：261路公交车全程24公里，配车20台，高峰期发车间距为7-8分钟/趟，平峰为10-15分钟/趟，经查看近期发车时刻表，其发车间隔正常。由于驾驶员工作时间长，劳动强度大，造成车队驾驶员短缺，公司正在积极组织调配人员充实该线路运力，公司人事部正在积极进行驾驶员招聘工作，条件具备后将增加该线路配车。感谢您对我们工作的支持、理解与监督！2019年1月8日 | 0.90 | 0.95 | 0.79 | 0.87 |
| 六线（劳动东路-机场高架）段目 | 网友“UU008694”您好！您的留言已收悉。现将有关情况回复如下：东六线（楚府东路—机场高架）道路工程全长5.8公里，其中楚府路—劳动东路段长度3.8公里，预计2019年4月30日基本具备通车条件；劳动路—机场高架段长度2公里，已完成约800米路基，其余路段因涉及到国有土地（正钢机械厂，星沙机床厂等）征拆问题，暂未启动施工。目前，市城投集团正在积极与A7县进行对接，请求加大拆迁力度，力争2019年年底达到通车条件。感谢您对我们工作的支持、理解与监督！2018年12月28日 | 0.95 | 0.85 | 0.84 | 0.87 |
| 望能在A市怡海星城楼盘建公办学 | 网友“UU0082278”您好！您的留言已收悉。现将有关情况回复如下：经查，“怡海星城”原属A7县审批项目，2008年3月18日通过A7县规划局修建详细规划批复，一期于2009年3月10日、二期于2010年6月25日、三期于2013年12月11日分别通过A7县规划局审批，在1-3期用地范围内分别于2010年7月29日，2011年1月30日通过A7县主管部门批准建设了民办怡雅小学和怡雅中学，2015年1月“怡海星城”通过区划调整由A7县并入A2区后，区教育局就针对“怡海星城”小区内小学生就读公办学校事宜，按照义务制教育阶段学生“就对就近，免试入学”的入学基本原则，已安排小区学生在西湖小学就读小学， | 0.92 | 0.96 | 0.96 | 0.95 |

图 12 政府回复评价

## 4 总结与展望

综上，我们圆满的完成了出题方的三个问题：文本分类、热点问题挖掘、政府回复评估。

对于文本分类，我们的模型在线下测试取得了良好的效果，既考虑了应用场景的实时性，同时宏 $F_1$

值也达到了 91%；对于热点问题挖掘，通过对问题深入的解析，我们针对问题改进了 Single-Pass 算法、改进了热度计算公式、提出了文本中相关人群与地点抽取的方案，同时也对模型热点挖掘的结果进行评估，评估结果显示出我们模型的有效性与可靠性；对于政府回复的评估，我们对出题方所述的相关性、完整性、可解释性进行深入挖掘，给出了一个粗略的评估方案。

当然，我们的方案仍存在一定的缺点：（1）文本分类的效果还存在提升空间，尽管我们尝试了贝叶斯、线性分类器、SVM、KNN、FastText、LSTM、BiLSTM、XGBoost 等模型，最后还做出了我们的集成模型；（2）地点/人群提取上，由于采用多级特征的方式，人工观察得到的特征不具有泛化能力，能否结合更多深度学习的方法，对文本中相关语义信息进行抽取；（3）我们对政府答复的评估，考虑还不够周全，比如完整性的定义，我们考虑了回复中有对该留言用户的问候和致意、回复中有对该留言相关问题的探讨、回复中有对该留言相关问题的解决方案，但是我们却没有深入的理解留言内容与回复内容。又如图 12 的留言 1，虽然三个因素得分都很低，但其原因并不是政府回复不好，而是留言本身的问题。

最后，非常感谢出题方，他们不仅仅提供了既涉及了当下热点领域的热点问题，也非常符合当下背景的赛题，也提供了充分的学习资源、热心的工作人员！在此，我们向各位表示衷心的感谢与真挚的祝福！

**参考文献**

[1.] G.Saltion and M.McGill. Introdution to Modern information Retrieval[J].New York：Mcgraw-Hill，1983.

[2.] Kevin Chen-Chuan Chang， Hector Garcia-Molina， Andreas Paepcke， Boolean Query mapping Across Heterogeneous Information Sources[J].IEEE Transactions on Knowlede and Data Engineering，1996，8(4)：515-512.

[3.] L.Ma and Y.Zhang.Using word2vec to process big text data.2015 IEEE International Conferencce on Big Data，pages 2895-2897，2015.

[4.] Kiros R，Zemel RS，Salakhutdinov R，A multiplicative model for learning distributed text-based attribute representations.Proceedings of Advances in Neural Information Processing Systems.Montreal， Quebec，Canda.2014.2348-2356.

[5.] 赵婧，邵雄凯，刘建舟，王春枝.文本分类中一种特征选择方法研究[J].计算机应用研究，2019，36(8)：2261-2265.

[6.] 复旦大学计算机信息与技术系国际数据库中心自然语言处理小组.复旦大学中文文本分类语料库[EB/OL].

[7.] 白飞云.基于内容的中文垃圾邮件过滤算法研究[D].西安：西安理工大学，2012.

[8.] 王丽颖.增量式聚类的新闻热点话题发现研究[D].广西：广西民族大学，2017.

[9.] 李倩.Single-Pass 聚类算法的改进及其在微博话题检测中的应用研究[D].山东：山东师范大学，2016.

[10.] 周茜.融合 word2vec 和 Single-Pass 的微博话题检测方法研究[D].山东：山东师范大学，2019.

[11.] 郑立洲.短文本信息抽取若干技术研究[D].安徽：中国科学技术大学，2016.

[12.] Wang W， Zhao D， Zou L， et al.Extracting 5W1H event semantic elements from Chinese online news.Proceedings of Web-Age Information Management.Springer，2010：644-655.

[13.] Wang W.Chinese news event 5W1H semantic elements extraction for event ontology population.Proceedings of Proceedings of the 21st international conference companion on World Wide Web.ACM，2012.197-202.

[14.] 郑憧.基于 CNN 语义匹配的自动问答系统构建方法研究[D].哈尔滨：哈尔滨工业大学，2016.

[15.] Koby Crammer， Ofer Dekel， Shai Shalev-Shwartz， Yoram Singer.Online Passive-Aggressive

Algorithms[J].Journal of Machine Learning Research，2006，7：551-585.

[16.] Asharaf S， Murty M N.An adaptive rough fuzzy single pass algorithm for clustering large data sets[J].Pattern Recognition，2013，36(12)：3015-3018.