

# 基于深度学习的网络政务留言智能分析系统

## 摘要

互联网的蓬勃发展,使得建立基于 NLP 的智慧政务系统成为社会治理创新发展新趋势,及时挖掘出群众对城市建设的反馈信息,对提升政府的管理水平和施政效率有着极大的推动作用,具有现实意义.

针对问题 1,本文对群众留言进行字符粒度处理,并使用词嵌入与 CNN 提取文本特征,最后使用 BP 神经网络进行一级标签分类,建立了基于字符级 CNN 的留言一级标签分类模型.本文将该模型与 GRU、TF-IDF、BOW 等模型在字符与分词两种粒度下进行对比实验,结果表明:字符粒度比分词粒度能更有效地提高模型的性能,CNN 优于其他方法,模型 F-Score 达到 0.904.

针对问题 2,本文结合内部特征与外部特征定义了热度指数对主题进行热点性评价.本文首先利用问题 1 的一级标签分类模型进行预分类,再使用双重 Single-Pass 主题聚类发现主题,接着利用 TextRank 摘要生成算法生成主题描述,并利用基于 Bi-LSTM+CRF 的命名实体识别模型提取主题所反映的特定地点与人群,最后使用热度指数对各主题进行热点性评价,建立了基于双重 Single-Pass 和热点特征的留言热点检测模型.实验结果表明:本文定义的热度指数能够有效反映话题的热度并获得了较好的热度结果;与单次聚类的主题发现相比,使用双重聚类获得了更好的主题识别结果、主题更准确;TextRank 算法能有效提取同一聚类主题中最具代表性的留言标题作为主题描述;本文采用的命名实体识别模型,Precision、Recall 和 F-Score 都略高 0.85,能有效识别留言中的地点、人群实体.

针对问题 3,本文爬取了郴州市市长信箱的群众留言、回复及满意度数据,在问题 1 的基础上提取留言及回复的文本特征,并与相关非文本特征进行特征融合,最后使用 BP 神经网络进行回复满意度预测,建立了基于字符级 CNN 的回复满意度的分类预测模型.与 GRU、TF-IDF、BOW 等模型在字符与分词两种粒度下的对比实验结果表明:该模型相关评价指标优于其他方法,性能有一定幅度的提高.

**关键词:** 字符级 CNN, 留言热点检测, 双重 Single-Pass, 特征融合, 满意度预测

## Abstract

The vigorous development of the Internet has made the establishment of NLP-based smart government systems a new trend in the development of social governance innovation. Mining the public's feedback on urban construction in a timely manner has greatly promoted the improvement of government management and governance efficiency Realistic meaning.

In response to question 1, this article performs character granularity processing on mass messages, uses word embedding and CNN to extract text features, and finally uses BP neural network for first-level label classification, and establishes a first-level label classification model for message based on character-level CNN. This model is compared with GRU, TF-IDF, BOW and other models at two granularities of character and word segmentation. The results show that character granularity can improve the classification performance of the model more effectively than word granularity. CNN is superior to other methods, model F -Score reaches 0.904.

For problem 2, this article combines the internal and external features to define the heat index to evaluate hot topics. This article first uses the first-level label classification model of question 1 to pre-classify, and then uses dual Single-Pass topic clustering to find the topic, and then Use the TextRank summary generation algorithm to generate topic descriptions, and use the named entity recognition model based on Bi-LSTM + CRF to extract the specific locations and crowds reflected by the topics. Finally, use the heat index to conduct hot spot evaluation on each topic, and establish a dual-based Single- Pass and hotspot feature message hotspot detection model. The experimental results show that the heat index defined in this paper can effectively reflect the heat of the topic and obtain good heat results; compared with the topic discovery of single clustering, double clustering is used to obtain Better topic recognition results and more accurate topics; TextRank algorithm can effectively extract the most representative message title in the same cluster topic as the topic description; The named entity recognition model used in this article, Precision, Recall and F-Score are all omitted 0.85 high, can effectively identify the location and crowd entity in the message.

In response to question 3, this article crawls the mass message, reply and satisfaction data of Chenzhou mayor's mailbox, extracts the text features of the message and reply on the basis of question 1, and performs feature fusion with relevant non-text features, and finally uses BP The neural network predicts the satisfaction of replies, and establishes a classification prediction model based on the character-level CNN's response satisfaction. The experimental results of comparison with GRU, TF-IDF, BOW and other models at two granularities of character and word segmentation show that the model is relevant The evaluation index is better than other methods, and the performance has been improved to a certain extent.

**Keywords:** character-level CNN, message hotspot detection, dual Single-Pass, feature fusion, satisfaction prediction

# 目录

摘要.....	1
Abstract.....	2
1 问题重述与分析 .....	1
1.1 问题背景.....	1
1.2 问题分析.....	1
1.3 总体流程.....	1
2 背景知识.....	2
2.1 卷积神经网络(CNN).....	2
2.2 长短时记忆神经网络(LSTM) .....	3
2.3 GRU 神经网络 .....	4
2.4 Single-Pass 聚类 .....	5
3 文本数据预处理 .....	5
3.1 构建词库.....	5
3.2 分词去噪.....	6
3.3 文本序列向量化.....	6
4 基于字符级 CNN 的留言一级标签分类模型.....	6
4.1 特征提取.....	6
4.2 分类器.....	7
5 基于双重 Single-Pass 和热点特征的留言热点检测模型 .....	7
5.1 TF-IDF 特征提取 .....	8
5.2 主题发现.....	8
5.2.1 双重 Single-Pass 主题聚类模型.....	8
5.2.2 TextRank 主题描述生成模型 .....	9
5.3 热点性评价.....	10
5.3.1 内部特征.....	10
5.3.2 外部特征.....	11
5.3.3 热度指数.....	11
5.4 基于 Bi-LSTM+CRF 的地点/人群实体识别模型 .....	11
5.4.1 Bi-LSTM.....	11

5.4.2 CRF .....	13
6 基于字符级 CNN 的回复满意度的分类预测模型 .....	13
6.1 非文本特征 .....	14
6.2 特征融合 .....	15
7 实验与结果 .....	15
7.1 数据来源 .....	15
7.2 实验环境 .....	15
7.3 评价指标 .....	16
7.3.1 F-Score .....	16
7.3.2 轮廓系数 .....	17
7.4 基于字符级 CNN 的留言一级标签分类模型 .....	17
7.4.1 实验数据 .....	17
7.4.2 实验及结果分析 .....	18
7.4.3 对比实验 .....	19
7.5 基于双重 Single-Pass 和热点特征的留言热点检测模型 .....	19
7.5.1 实验数据 .....	19
7.5.2 双重 Single-Pass 主题聚类模型实验及结果分析 .....	20
7.5.3 TextRank 主题描述生成模型实验及结果分析 .....	21
7.5.4 地点、人群实体识别模型实验及结果分析 .....	22
7.6 基于字符级 CNN 的回复满意度的分类预测模型 .....	22
7.6.1 实验数据 .....	22
7.6.2 实验及结果分析 .....	23
7.6.3 对比实验 .....	24
8 总结 .....	24
参考文献 .....	26

## 图目录

图 1-1 整体框架 .....	2
图 3-1 文本数据预处理流程 .....	5
图 4-1 基于字符级 CNN 的留言一级标签分类模型 .....	6
图 5-1 基于双重 Single-Pass 和热点特征的留言热点检测模型 .....	8
图 5-2 双重 Single-Pass 主题聚类模型 .....	9
图 5-3 基于 Bi-LSTM 与 CRF 的地点/人群实体识别模型 .....	12
图 5-4 Bi-LSTM 结构图 .....	12
图 5-5 线性链条件随机场结构 .....	13
图 6-1 基于字符级 CNN 的回复满意度的分类预测模型 .....	14
图 7-1 双重 Single-Pass 轮廓系数对比图 .....	20
图 7-2 双重 Single-Pass 聚类主题数对比图 .....	20

## 表目录

表 6-1 非文本特征描述表 .....	14
表 7-1 附件 2 一级标签分布情况 .....	17
表 7-2 一级标签分类模型混淆矩阵 .....	18
表 7-3 一级标签分类模型分标签性能指标报告 .....	18
表 7-4 一级标签分类模型不同方法性能对比 .....	19
表 7-5 聚类效果较差主题描述生成示例 .....	21
表 7-6 聚类效果较好主题描述生成示例 .....	21
表 7-7 地点、人群实体识别模型性能指标 .....	22
表 7-8 郴州政务留言原满意度分布表 .....	22
表 7-9 郴州政务留言训练满意度分布表 .....	23
表 7-10 满意度预测混淆矩阵 .....	23
表 7-11 满意度预测分满意度性能指标报告 .....	24
表 7-12 满意度预测模型不同方法性能对比 .....	24

# 1 问题重述与分析

## 1.1 问题背景

近年来,随着互联网技术飞速发展,数据库技术的成熟和数据库系统应用的普及,微信、微博、市长信箱、阳光热线等网络问政平台应运而生,逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道.

建立基于自然语言处理技术(Natural Language Processing, NLP)的智慧政务系统已经是社会治理创新发展的新趋势,对于有效利用大规模文本语言在自动划分分类、智能自动回复等具有现实意义,也对提升政府的管理水平和施政效率具有极大的推动作用.

## 1.2 问题分析

本文是利用自然语言处理技术,对政务留言数据进行文本挖掘,要达成的目标如下:

问题 1:根据群众留言数据(附件 2)进行文本挖掘,建立关于群众留言的一级标签分类模型.首先要选择粒度对群众留言进行划分、清洗,其次提取群众留言的文本特征进行一级标签分类.

问题 2:对群众留言数据(附件 3)进行主题聚类,并结合自定义的热点性评价指标建立一个留言热点问题检测模型.首先利用主题聚类对群众留言进行主题识别,并利用摘要生成算法进行描述,其次结合“点赞数”与“反对数”及其他相关特征定义合理的热点性评价指标,对主题进行热点性评价,最后利用命名实体识别模型对热点主题进行地点、人群实体的提取.

问题 3:针对群众留言及其答复意见(附件 4)进行分析处理,建立一个关于回复满意度的分类预测模型.在群众留言及回复文本中提取相关性、完整性、可解释性的相关特征,进行特征融合后进行满意度的分类预测.

## 1.3 总体流程

图 1-1 是本文对问题 1、问题 2、问题 3 分析后,最终确定的整体流程框架,主要包括特征提取、分类预测、热点问题检测三大部分.

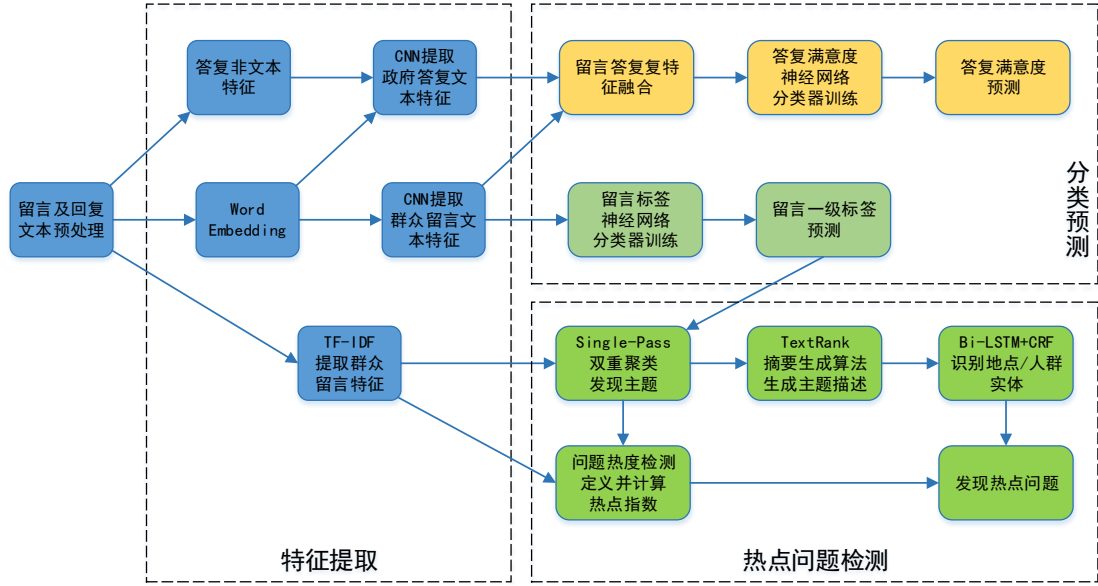


图 1-1 整体框架

## 2 背景知识

### 2.1 卷积神经网络(CNN)

卷积神经网络<sup>[1]</sup> (Convolutional Neural Networks, CNN)一般由若干个卷积层、池化层交替组成。

在卷积层中,通常包含多个可学习的卷积核,上一层输出的特征图与卷积核进行卷积操作,即输入项与卷积核之间进行点积运算,然后将结果送入激活函数,就可以得到输出特征图.每一个输出特征图可能是组合卷积多个输入特征图的值。

卷积层 $l$ 的第 $j$ 单元的输出值 $a_j^l$ 的计算为:

$$a_j^l = f(b_j^l + \sum_{i \in M_j^l} a_i^{l-1} \cdot k_{ij}^l) \quad (2.1)$$

其中,  $M_j^l$ 表示选择的输入特征图的集合, $k$ 表示可学习的卷积核:

$$k_{ij}^l = \begin{pmatrix} X_{11} & X_{12} & X_{13} & \cdots & X_{1j} \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2j} \\ X_{31} & X_{32} & X_{33} & \cdots & X_{3j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & X_{i3} & \cdots & X_{ij} \end{pmatrix}$$

$$a_i^{l-1} = \begin{pmatrix} A_1 & 0 & 0 & \cdots & 0 \\ A_2 & A_1 & 0 & \cdots & 0 \\ \vdots & A_2 & A_1 & \cdots & 0 \\ A & \vdots & A_2 & \cdots & 0 \\ 0 & A & \vdots & \cdots & A_1 \\ \vdots & \vdots & 0 & \cdots & A_2 \\ 0 & 0 & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & A \end{pmatrix}$$

池化层通常出现在卷积层之后,二者相互交替出现,并且每个卷积层都与一个池化层一一对应.池化层 $l$ 中激活值 $a_j^l$ 的计算公式如下

$$a_j^l = f(b_j^l + \beta_j^l \text{down}(a_i^{l-1}, M^l)) \quad (2.2)$$

其中,  $\text{down}(\cdot)$  表示池化函数,常用的池化函数有 Mean-Pooling(均值池化)、Max-Pooling(最大值池化)、Min-Pooling(最小值池化)、Stochastic-Pooling(随机池化)等,  $b_j^l$  为偏置,  $\beta_j^l$  为乘数残差,  $M^l$  表示层所采用的池化框大小为  $M^l \cdot M^l$ .

## 2.2 长短时记忆神经网络(LSTM)

循环神经网络(Recurrent Neural Network, RNN) 因拥有记忆功能使得它广泛应用于 NLP 领域,但是由于 RNN 网络结构存在梯度消失或者梯度爆炸问题,因此在实际应用时 RNN 仅能够回溯利用与它接近时间点上的信息.

长短时神经网络<sup>[2]</sup> (Long short-term memory, LSTM)通过引入记忆与遗忘机制解决 RNN 的长距离依赖问题.LSTM 单元主要由记忆门、遗忘门、输出门三个部分组成.

在 $t$ 时刻,LSTM 单元存在记忆状态 $c_t$ ,输出为 $h_t$ ,输入为 $t$ 时刻的输入的 $x_t$ 与为 $t-1$ 时刻的输出 $h_{t-1}$ 和细胞状态 $c_{t-1}$ .

首先,两个输入信息经过遗忘门 $f_t$ ,遗忘门对输入进行选择性的遗忘:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (2.3)$$

接着,输入信息通过记忆门 $i_t$ ,记忆门对输入进行选择性的记忆:



$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (2.4)$$

$$c'_t = \text{Tanh}(W_c x_t + U_c h_{t-1}) \quad (2.5)$$

其中,  $\sigma$  为 sigmoid 激励函数  $f(x) = \frac{1}{e^{-x} + 1}$

$\text{Tanh}$  为双曲正切函数  $\tanh(x) = \frac{\sin x}{\cos x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

经过记忆门与遗忘门后,对细胞状态进行更新:

$$c_t = f_c c_{t-1} + i_t c'_t \quad (2.6)$$

最后通过输出门  $o_t$ ,是将细胞状态与输入结合,得到输出

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (2.7)$$

$$h_t = o_t \text{Tanh}(c_t) \quad (2.8)$$

## 2.3 GRU 神经网络

GRU<sup>[3]</sup>(Gated Recurrent Unit)是 LSTM 网络的结构更为简单的变体.GRU 模型只有两个门:更新门(update gate)和重置门(reset gate).

首先,上一个传输下来的状态  $h^{t-1}$  和当前节点输入的  $x^t$  来获取两个门控状态.其中  $r$  控制重置门,  $z$  控制更新门.

$$r = \sigma(W^r \cdot \frac{x^t}{h^{t-1}}) \quad (2.9)$$

$$z = \sigma(W^z \cdot \frac{x^t}{h^{t-1}}) \quad (2.10)$$

通过重置门重置后的  $h_{t-1}$  与输入  $x^t$  拼接,再利用  $\tanh$  激活函数得到  $h'$

$$h' = \tanh(W \odot \frac{x^t}{h^{t-1}}) \quad (2.11)$$

其中  $\odot$  是 Hadamard Product,也就是矩阵中对应的元素相乘.

最后进行一个最关键的步骤,对记忆进行更新,更新表达式为

$$h^t = (1 - z) \odot h^{t-1} + z \odot h^t \quad (2.12)$$

## 2.4 Single-Pass 聚类

Single-Pass<sup>[4]</sup>算法由于其简单、高效和低成本而通常用于主题检测和跟踪. Single-Pass 通过计算文本与已有主题间的余弦距离,并与距离阈值进行比较,决定留言归入已有主题或者创建新的主题,根据不同的阈值可以得到不同粒度大小的主题.

Single-Pass 算法聚类过程如下:

- (1) 输入一条新的留言  $d_i$ ;
- (2) 计算留言文本  $d_i$  与已有主题的距离  $\theta$ ,得到距离最小的主题  $Topic_{\min}$  及最小距离  $\theta_{\min}$ ;
- (3) 比较最小距离  $\theta_{\min}$  与距离的阈值  $\varepsilon$
- (4) 若  $\theta_{\min} < \varepsilon$ ,则将留言  $d_i$  归类于主题  $Topic_{\min}$  中,并将  $Topic_{\min}$  的新均值更新为  $Topic_{\min}$  新的质心;否则将留言  $d_i$  作为新主题,创建新的主题;
- (5) 进入下一条留言,直至对所有留言完成聚类.

## 3 文本数据预处理



图 3-1 文本数据预处理流程

由于留言文本具有海量、不完整、模糊、噪声大的特点,所以在挖掘留言文本前需进行文本数据预处理,过程如图 3-1 所示.

### 3.1 构建词库

现有的分词工具的词典,大部分没有涵盖特定领域的特定词汇,一般不能识别特定文本数据中较为特殊的词语,从而导致分词结果较差,对后续的文本处理分析产生较大影响.

本文针对附件所给政务留言文本数据的特点,并结合国家语委现代汉语通用平衡语料库<sup>[5][6][7]</sup>,建立了一个自定义词典(见附件 1),在进行分词时,通过导入自定义词典,提高中文分词的准确率,更准确地对原始文本进行粒度划分。

### 3.2 分词去噪

对于中文文本挖掘,一般不直接使用句子进行处理,而是对所给文本数据进行分词后再进行处理,分词的粒度分为字符粒度与分词粒度两种。

由于文本中有许多没意义的词,如标点符号、html 标签、语气词等,本文引入停用词表对分词数据中这部分无效词进行过滤,达到文本清洗降噪的目的。

### 3.3 文本序列向量化

由于计算机无法直接处理文本数据,因此对文本序列进行分词清洗后,需要对文本序列进行向量化,以便程序对文本序列进行挖掘。

## 4 基于字符级 CNN 的留言一级标签分类模型

对文本分类问题,Yoon Kim<sup>[8]</sup>等人提出了基于预先训练的单词向量的 CNN 模型用于句子分类,Nal Kalchbrenner<sup>[9]</sup>等人提出了用于句子语义建模的动态卷积神经网络,Xiang Zhang<sup>[10]</sup>等人使用字符级卷积网络进行文本分类。

本文针对问题 1,在上述文献基础上建立了基于字符级 CNN 的留言一级标签分类模型,结构如图 4-1。模型采用字符粒度对留言文本进行预处理,使用预处理结果作为输入,用自训练词嵌入模型与 CNN 相结合提取留言文本特征,最后设计了一个 BP 神经网络<sup>[11]</sup>作为分类器进行留言的一级标签分类。

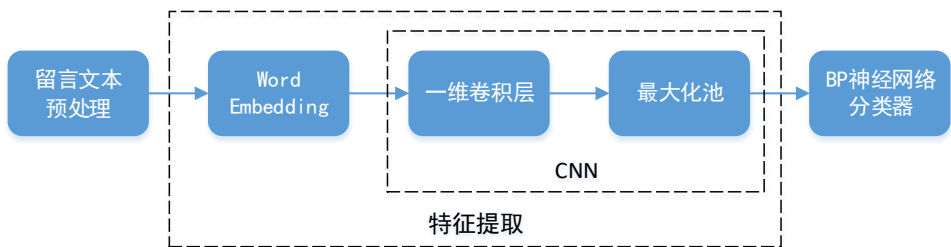


图 4-1 基于字符级 CNN 的留言一级标签分类模型

### 4.1 特征提取

首先将预处理得到的高维稀疏文本序列通过词嵌入模型映射为一个低维稠

密的向量,该向量包含了文本序列语义信息.随后使用 CNN 对文本数据进行特征提取.

CNN 神经网络相比全神经网络的明显优点就是拥有卷积层、池化层.卷积层在处理留言文本时能够更加有效地去除噪声进行特征提取.模型采用的 CNN 由一个一维卷积层与一个最大化池组成.

一维卷积层公式为:

$$a_j^l = f(b_j^l + \sum_{i \in M_j^l} a_i^{l-1} \cdot k_{ij}^l) \quad (4.1)$$

其中,卷积层卷积核为  $k$ .

最大化池公式为:

$$a_j^l = f(b_j^l + \beta_j^l \text{Max}(a_i^{l-1}, M^l)) \quad (4.2)$$

## 4.2 分类器

模型使用三层 BP 神经网络作为分类器,对提取的留言文本特征进行一级标签分类.BP 神经网络使用误差逆向传播算法进行参数学习.

# 5 基于双重 Single-Pass 和热点特征的留言热点检测模型

本文留言热点检测利用热点话题检测与跟踪技术(TDT),主要包括两个任务:一是主题发现,二是对主题进行热点性评价.

对热点话题检测,张辉<sup>[12]</sup>等人先利用 Single-Pass 将新闻聚成微簇,再将微簇与已有事件进行相似性匹配得到模型更新;马永军<sup>[13]</sup>等人综合多种文本要素提出一种多相似度计算组合策略;成卫青<sup>[14]</sup>等人对文本先进行分类,再分别进行聚类分析,得到各个类别的热点簇.

对话题的热度评价,Feng Xu<sup>[15]</sup>等人的研究表明影响微博话题受欢迎程度的评价包含四个因素:微博数量、转发次数、评论数量和赞数;Ma Huifang<sup>[16]</sup>等人根据热点话题的广泛性和突发特性引入了定时窗口,并结合 TF-IDF<sup>[18]</sup>的权重计算提取了热点项;Minjuan Zhong<sup>[17]</sup>等人则结合内部与外部双重特征对聚类得到的主题进行热点评估.

针对问题 2,本文在上述文献的模型及结论的基础上,建立了基于双重 Single-Pass 和热点特征的留言热点检测模型,结构如图 5-1 所示.模型首先使用问题 1 的留言一级标签分类模型对留言文本进行预分类,后使用 TF-IDF 特征进行双重 Single-Pass 主题聚类发现主题,对发现的主题使用摘要生成算法 TextRank<sup>[19]</sup>生成

对应的主题描述,并建立基于 Bi-LSTM+CRF<sup>[20]</sup>的地点人群实体识别模型对主题描述中的地点、人群实体进行提取,最后结合主题的内部特征与外部特征进行热点性评价,得到群众留言的热点问题。

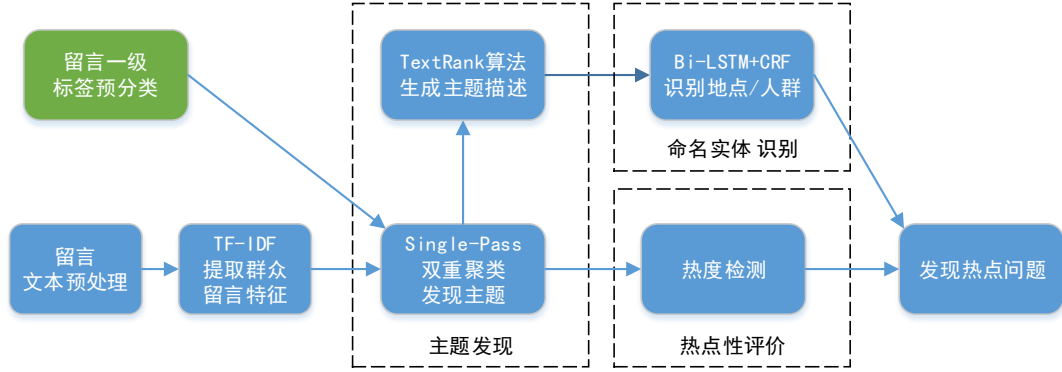


图 5-1 基于双重 Single-Pass 和热点特征的留言热点检测模型

## 5.1 TF-IDF 特征提取

对于采用分词粒度预处理后的留言文本,模型使用 TF-IDF 进行特征提取.TF-IDF 算法是常用于文本挖掘中的加权技术,它是一种统计方法,可以用来评估某一特征项 $t$ 代表的关键问题对于群众留言文本 $d$ 中的重要程度 $w(d,t)$ 。

$$w(d,t) = TF \cdot IDF = tf(d,t) \cdot \log \frac{N}{df_t} \quad (5.1)$$

其中 $tf(d,t)$ 表示特征项 $t$ 在留言文本 $d$ 中出现的频率, $N$ 是留言文本总数, $df_t$ 为特征项在留言文本的频率。

## 5.2 主题发现

主题发现部分由 Single-Pass 双重聚类发现主题与 TextRank 摘要生成算法生成主题描述两部分组成。TextRank 对双重聚类所得结果进行描述。

### 5.2.1 双重 Single-Pass 主题聚类模型

由于大量的留言文本之间容易存在内容语义混淆的问题,直接进行主题聚类容易造成聚类性能下降,且大量文本同时聚类会大大提高模型时空复杂度.为此模型在聚类前先利用问题 1 的一级标签分类模型对留言文本进行标签分类,将留言按照内容进行预分类<sup>[14]</sup>,再进行双重 Single-Pass 主题聚类。

与 K-means、LDA 等聚类算法相比,Single-Pass 以增量的方式进行动态聚类,无需提前确定主题数,能更好地处理留言主题聚类这类主题数动态变化的问题,对不同规模的留言有更强的适应性

模型利用 Single-Pass 算法对群众留言文本进行双重主题聚类,结构如图 5-2 所示.模型首先利用 Single-Pass 分标签进行留言聚类,因为群众留言的问题一般涉及多个部门,所以存在实际上属于一个主题的留言却因分到不同的标签,导致属于不同主题.为解决这个问题,模型借鉴文献[12]的做法,对分标签聚类所得主题进行二次 Single-Pass 聚类,将同属一个主题却分配到不同部门的留言进行融合.

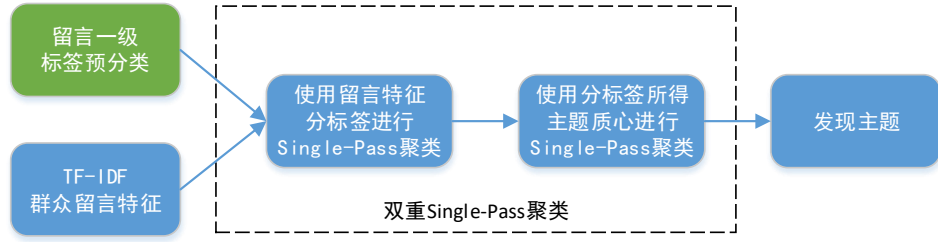


图 5-2 双重 Single-Pass 主题聚类模型

### 5.2.2 TextRank 主题描述生成模型

对聚类后同一主题的所有留言,模型采用抽取式摘要生成算法 TextRank,抽取最具代表性的留言标题作为主题描述.将每个留言文本作为一个节点,建立一个全连接图,节点  $v_i$  与  $v_j$  的边的权值为留言  $i$  与  $j$  之间的相似度  $w_{ij}$ .使用迭代公式计算得到留言的重要性排序后,按需要抽取最重要  $n$  条留言作为主题的描述.

每个留言的重要性迭代公式:

$$S(v_i)^k = (1-d) + d \sum_{(j,i) \in \mathcal{E}} \frac{w_{ji}}{\sum_{v_k \in out(v_j)} w_{jk}} S(v_j)^{k-1} \quad (5.2)$$

整个主题的迭代公式:

$$S(v)^k = (1-d) + d \cdot S(v)^k \cdot A \quad (5.3)$$

$$A = (simi - E) / out(v) \quad (5.4)$$

$$out(v) = \sum_j (simi - E)_{ij} \quad (5.5)$$

其中  $d$  为阻尼系数(一般取 0.85),  $simi$  为留言间的相似度矩阵(模型采用余弦相似度进行度量),  $E$  为单位矩阵.

## 5.3 热点性评价

与普通话题相比,热点话题是群众在某一时段内集中反映的问题.分析发现留言热点话题表现在短时间内群众互动更频繁、留言更多,且持续时间更长.

模型在文献[17]的基础上,引入内部特征与外部特征来定义热度指数,对聚类识别出来的主题进行热点性评价,并根据计算得到的热度指数进行排序,最后得到热点话题排名.

### 5.3.1 内部特征

模型使用群众的直接行为作为内部特征,在政务留言中,群众的直接行为表现为对某些留言进行“点赞”或者“反对”.分析认为,对同一留言,其“点赞数”与“反对数”的区别在于群众对问题的态度,但都体现了群众对该问题的关注程度,且对热度的贡献是一致的.

对同一主题的不同留言,“点赞数”与“反对数”相同,但其与前一个相同主题留言的间隔时间不同,对整个主题的热度贡献也有所区别,且贡献的大小与时间间隔的长短成反比关系,由此定义第 $i$ 个主题的第 $j$ 个留言的时间热度系数:

$$h_{i,j} = \begin{cases} e, j = 1 \\ \frac{T}{\Delta t_{i,j}} + e, j > 1 \end{cases} \quad (5.6)$$

$$\Delta t_{i,j} = \max\{t_{i,j} - t_{i,j-1}, 1\} \quad (5.7)$$

其中: $t_{i,j}$ 为第 $i$ 个主题的第 $j$ 个留言(按时间顺序,下同)的时间; $T$ 为标准时间间隔,取 30 天, $\Delta t_{i,j}$  单位为天.

“点赞数”与“反对数”与热度系数直接相关,使用时间热度系数作为“点赞数”与“反对数”的权值,得到同一第 $i$ 个主题第 $j$ 个留言的内部特征:

$$H_{i,j} = \log(h_{i,j}) \cdot (L_{i,j} + U_{i,j}) \quad (5.8)$$

同一主题的热度内部特征,由该主题的所有留言来共享,使用所有留言的“点赞数”与“反对数”的加权平均作为热度指标的内部特征,定义第 $i$ 个主题的内部特征:

$$Inter_i = \sum_j H_{i,j} \quad (5.9)$$

### 5.3.2 外部特征

模型使用群众的间接行为作为外部特征,在政务留言中,群众参与留言,直接增加了主题的留言数,也间接影响了该主题的起止时间. 同一主题持续时间越长,群众的留言数越多,说明该主题群众反映越集中,热度也越高.同时,由于群众的留言一般同时反映多个问题,对主题的热度贡献,与其相似的程度有关,即与其主题相似度有关.

由此定义第*i*个主题的外部特征为:

$$Exter_i = \frac{t_{i,end} - t_{i,start}}{T} \times S_i \times M_i \quad (5.10)$$

其中, $t_{i,start}$  为第*i*个主题第一条留言的时间;  $t_{i,end}$  为第*i*个主题最后一条留言的时间;  $T$  为标准时间间隔,取 7 天;  $M_i$  为第*i*个主题的留言数;  $S_i$  为第*i*个主题的留言平均主题相似度.

### 5.3.3 热度指数

通过内部特征与外部特征的相关分析,可以得到一个量化的主题热度计算模型,定义第*i*个主题的热度指数为:

$$Hot_i = \lambda_1 \times Inter_i + \lambda_2 \times Exter_i \quad (5.11)$$

其中, $\lambda_1 + \lambda_2 = 1$ .

## 5.4 基于 Bi-LSTM+CRF 的地点/人群实体识别模型

对于 TextRank 算法生成的主题描述,本文利用 Bi-LSTM+CRF 命名实体模型对地点、人群实体进行识别.

目前命名实体识别模型的工作,一般是针对地点、人名、机构组织等实体进行提取,较少针对特定人群进行识别.为此模型专门建立了一个人群命名实体数据集(见附件 2),并对训练的语料进行人群、地点实体的标注,建立了基于 Bi-LSTM 与 CRF 的地点/人群实体识别模型,结构如图 5-3 所示.

### 5.4.1 Bi-LSTM

LSTM 解决了 RNN 的长距离依赖问题,但只能提取文本序列的单向依赖关系,Bi-LSTM<sup>[21]</sup>解决了 LSTM 不能提取文本序列前后依赖的问题,能捕获文本序列



的前后语义关系.

Bi-LSTM 结构如图 5-4 所示,Bi-LSTM 由前向与反向两个 LSTM 组成,将两个 LSTM 的结果  $\vec{h}_t, \bar{h}_t$  进行 Concatenate 融合,将融合结果  $h_t$  作为输出.

Concatenate 公式为:

$$h_t = [\vec{h}_t, \bar{h}_t] \tag{5.12}$$

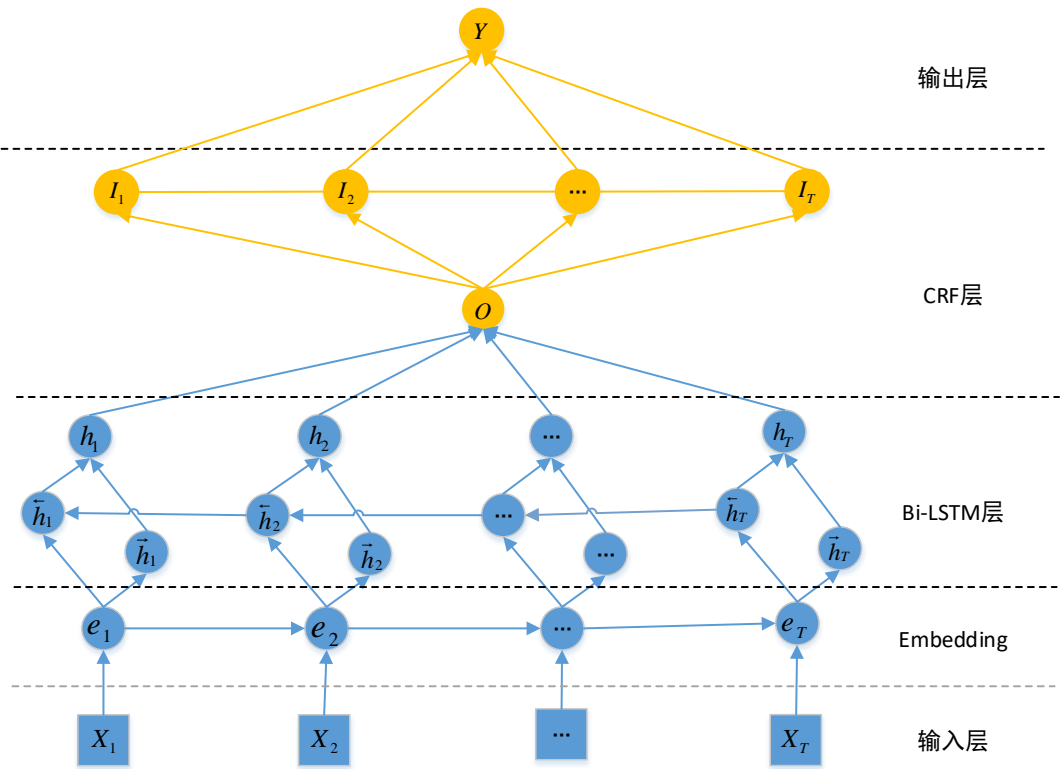


图 5-3 基于 Bi-LSTM 与 CRF 的地点/人群实体识别模型

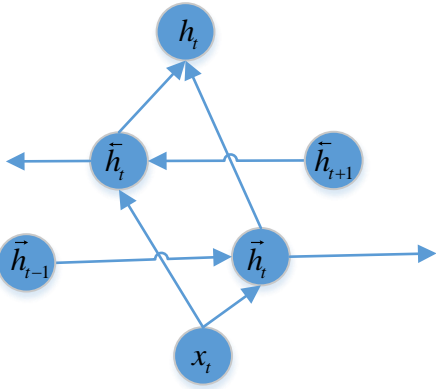


图 5-4 Bi-LSTM 结构图

### 5.4.2 CRF

地点/人群实体识别模型采用线性链结构的条件随机场<sup>[22]</sup>(Linear-chain CRFs)对留言文本序列进行建模,结构如图 5-5 所示.CRF 不仅解决了 HMM 输出独立性假设的问题,还解决了 MEMM 的标注偏置问题,CRF 的概率图为无向图,可以学习到数据在全局的分布.

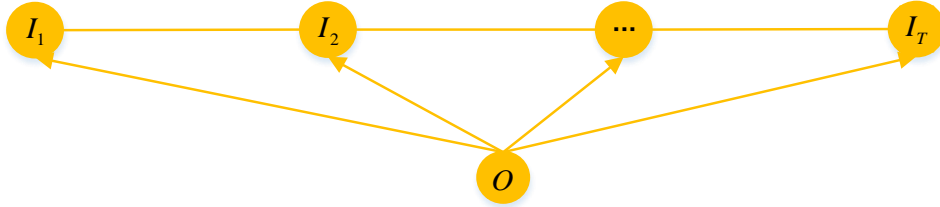


图 5-5 线性链条件随机场结构

对于地点/人群实体识别模型, $O$  表示输入的留言文本序列, $I$  表示对应输出的实体标记序列,有

$$\begin{aligned} P(I|O) &= \frac{1}{Z(O)} \prod_i \psi_i(I_i, O) \\ &= \frac{1}{Z(O)} e^{\sum_i \sum_k \lambda_k f_k(O, I_{i-1}, I_i)} \end{aligned} \quad (5.13)$$

其中,  $\psi_i(I_i, O) = e^{\sum_k \lambda_k f_k(O, I_{i-1}, I_i)}$  为节点  $i$  的势函数;  $Z(x) = \sum_Y \prod_c \psi_c(Y_c | X)$  为归一

化因子;  $f_k(O, I_{i-1}, I_i)$  为第  $k$  个特征函数;  $\lambda_k$  为特征函数  $f_k$  的权重,即模型学习的参数.

## 6 基于字符级 CNN 的回复满意度的分类预测模型

Fabio Calefato<sup>[23]</sup>等人的研究表明只有依赖于较浅的文字特征(如答案长度和句子数量),再加上其他相关特征,才能准确地预测出最佳回复;Roy<sup>[24]</sup>等人将 CQA 回复的文本非文本特征和用户声誉等进行特征融合并使用分类和回归模型来预测回复的有用性.

针对问题 3,本文在上述文献的启发下,在问题 1 的基础上建立了基于字符级 CNN 的回复满意度的分类预测模型,结构如图 6-1.模型采用字符粒度分别对留言文本与回复文本进行预处理,使用预处理结果与相关的非文本特征作为输入,使用词嵌入模型与 CNN 相结合分别提取留言文本特征、回复文本特征,随后将留言文

本特征、回复文本特征与非文本进行特征融合作为分类器输入,最后使用 BP 神经网络作为分类器进行满意度分类预测。

模型对基于字符级 CNN 的留言一级标签分类模型的文本特征提取、分类器部分进行复用。

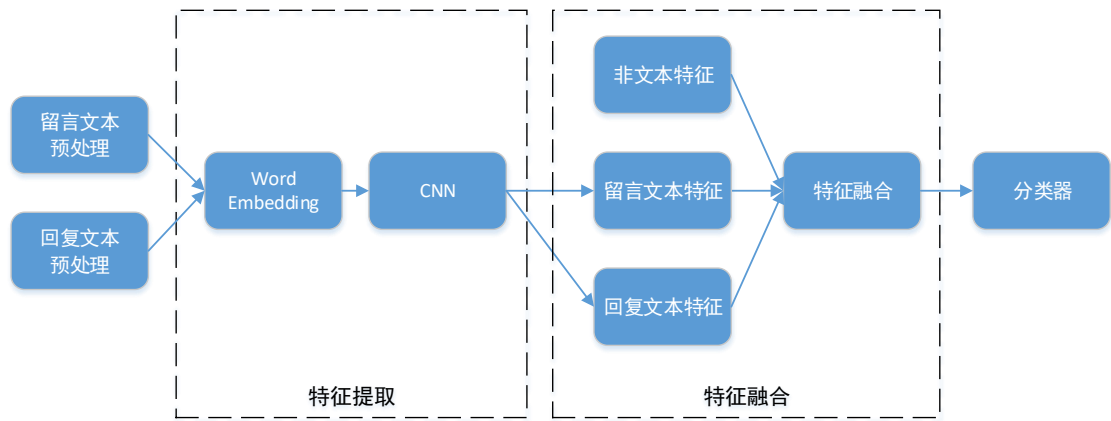


图 6-1 基于字符级 CNN 的回复满意度的分类预测模型

### 6.1 非文本特征

模型提取了与回复满意度相关的 8 个非文本特征,分别是回复总词数、回复名词比例、回复动词比例、回复形容词比例、回复停用词数比例、回复共现词数比例、回复低频词数比例、问答相似度。

非文本特征的具体描述如表 6-1 所示。

表 6-1 非文本特征描述表

特征名称	特征描述
回复总词数	回复文本中的词语的数量(包含停用词)
回复名词比例	回复文本中名词占总词数的比例
回复动词比例	回复文本中动词占总词数的比例
回复形容词比例	回复文本中形容词占总词数的比例
回复停用词数比例	回复文本中停用词占总词数的比例
回复共现词数比例	回复文本中存在于留言中的词语占总词数的比例
回复低频词数比例	回复文本中低频词占总词数的比例
问答相似度	留言文本与回复文本的余弦相似度

非文本特征中,除问答相似度使用 CNN 提取的留言文本特征、回复文本特征进行计算外,其余 7 个非文本特征使用分词粒度对留言文本与回复文本进行分词,经过统计得到.与文本特征使用字符粒度分词相区别。

## 6.2 特征融合

模型分别提取了文本与非文本两类特征后进行 Concatenate 特征融合,公式为

$$v = [v_{text}, v_{other}] \quad (6.1)$$

其中,  $v_{text} \in R^n$  为文本特征,  $v_{other} \in R^m$  为非文本特征,  $v \in R^{n+m}$ .

文本特征包括 CNN 提取的留言文本特征及回复文本特征,对两个文本特征进行 Addition 融合,公式为

$$v_{text} = v_1 + v_2 \quad (6.2)$$

$$v = \{x_i \mid x_i = v_{1i} + v_{2i}, i = 1, \dots, n\} \quad (6.3)$$

其中,  $v_1 \in R^n$  为留言特征,  $v_2 \in R^n$  为回复特征,  $v \in R^n$ .

## 7 实验与结果

### 7.1 数据来源

本论文使用的数据集来源如下:

- A. 比赛组委会提供的 C 题全部数据: <http://www.tipdm.org/bdrace/tzbstysj/20200228/1637.html>
- B. 郴州市政务服务中心市长信箱中的群众来信反馈、政府部门回复、群众满意度等数据:<http://www.app.czs.gov.cn/webapp/czs/email/index.jsp?login=false#top2>
- C. 国家语委现代汉语通用平衡语料库: <http://www.moe.gov.cn/s78/A19/>

除以上数据集外,本文还建立了一个人群实体集与针对附件的自定义词典.

### 7.2 实验环境

硬件环境: Intel64 Family 6 Model 142 Stepping 10 GenuineIntel ~1792 Mhz

Intel64 Family 6 Model 78 Stepping 3 GenuineIntel ~2401 Mhz

操作系统: windows10

编译环境:Pycharm 2019.3

运行环境: python<sup>1</sup> 3.6  
tensorflow<sup>2</sup> 1.14.0  
keras<sup>3</sup> 2.3.1  
scikit-learn<sup>4</sup> 0.22.1  
模型代码相关说明详见附件 3.

## 7.3 评价指标

质量评估是留言文本挖掘中最后一个环节,在本文中,质量评估主要使用 *F-Score*<sup>[25]</sup> 进行分类与实体识别评估,使用轮廓系数<sup>[27]</sup> 指标进行话题聚类进行性能评估<sup>[28]</sup>.

### 7.3.1 F-Score

*F-Score* 计算公式是分类方法常用的评估模型好坏的方法,表示的是 *Precision* 和 *Recall* 的调和平均评估指标,公式为:

$$F_1 = \frac{1}{n} \sum_{i=1}^n \frac{2P_iR_i}{P_i + R_i} \quad (7.1)$$

其中, *Precision*: 准确率,又称为查准率,表示预测结果中,预测为正样本的样本中,正确预测为正样本的概率.

*Recall* 召回率,又称为查全率,表示在原始样本的正样本中,最后被正确预测为正样本的概率;

计算公式如下:

$$Precision = \frac{TP}{TP + FP} \quad (7.2)$$

$$Recall = \frac{TP}{TP + FN} \quad (7.3)$$

其中 *TP*、*TN*、*FP*、*FN* 的含义分别如下:

---

<sup>1</sup> <https://www.python.org/downloads/>

<sup>2</sup> <https://tensorflow.google.cn/install?hl=zh-cn>

<sup>3</sup> <https://pypi.org/project/Keras/>

<sup>4</sup> <https://scikit-learn.org/dev/install.html>

$TP$ : True Positive, 被判定为正样本, 事实上也是正样本;  $FP$ : False Positive, 被判定为正样本, 但事实上是负样本;  $TN$ : True Negative, 被判定为负样本, 事实上也是负样本;  $FN$ : False Negative, 被判定为负样本, 但事实上是正样本.

### 7.3.2 轮廓系数

轮廓系数 (Silhouette Coefficient), 是聚类效果好坏的一种评价方式. 聚类将数据分为了  $k$  个簇后, 对于其中的某一个点  $i$ , 他们的轮廓系数是

$$S_i = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (7.4)$$

其中,  $a(i)$  表示  $i$  向量到同一簇内其他点不相似程度的平均值;  $b(i)$  表示  $i$  向量到其他簇的平均不相似程度的最小值;  $-1 \leq S_i \leq 1$ .

取所有样本的轮廓系数的算数平均值, 作为聚类结果的性能指标. 轮廓系数越趋近于 1, 代表内聚度和分离度越高, 聚类效果越好.

## 7.4 基于字符级 CNN 的留言一级标签分类模型

### 7.4.1 实验数据

表 7-1 附件 2 一级标签分布情况

一级标签 ID	一级分类标签	留言数
A	交通运输	613
B	劳动和社会保障	1969
C	卫生计生	877
D	商贸旅游	1215
E	城乡建设	2009
F	教育文体	1589
G	环境保护	938
总计		9210

模型使用比赛提供的附件 2 留言文本数据进行训练、验证、测试, 附件 2 提供了 7 个留言一级标签组的数据集, 每个标签下的留言数分布如表 7-1 所示, 一共 9210 条留言.

## 7.4.2 实验及结果分析

实验时,随机选择 80%的留言共 7368 条作为训练数据集进行训练,剩余 1841 条留言作为测试数据集进行验证.

gaussic<sup>1</sup>实现了 Xiang Zhang<sup>[10]</sup>的模型,本文在其实现基础上对参数进行适应性调整,确定神经网络的参数为原始序列长度 500,词维数 1300,词嵌入维度 64,学习率 0.001,dropout 比例 0.2,卷积核个数 64,卷积核尺寸为 3,BP 网络隐藏层维数 128,输出类数 7.实验结果见表 7-2、表 7-3:

表 7-2 一级标签分类模型混淆矩阵

真实 标签 ID	预测一级标签 ID						
	A	B	C	D	E	F	G
A	<b>91</b>	1	0	1	16	0	1
B	1	<b>363</b>	7	3	10	7	1
C	0	11	<b>166</b>	6	3	2	1
D	9	4	6	<b>192</b>	20	4	3
E	4	7	1	9	<b>369</b>	4	8
F	0	2	1	2	4	<b>322</b>	1
G	3	0	0	0	13	1	<b>161</b>

表 7-2 是训练模型的一级标签分类模型混淆矩阵结果,可以直观发现大部分结果都落在对角线上,横向看只有商贸旅游(D)被错分到其他标签的个例比较多,其他的标签错分的个例较少;纵向看,只有城乡建设(E)错误预测的个例较多,其他标签的被错分的个例较少.说明基于字符级 CNN 的留言一级标签分类模型总体性能较好.

表 7-3 一级标签分类模型分标签性能指标报告

一级标签	Precision	Recall	F-Score	留言数
交通运输	0.84	0.83	0.83	110
劳动和社会保障	0.94	0.93	0.93	392
卫生计生	0.92	0.88	0.9	189
商贸旅游	0.9	0.81	0.85	238
城乡建设	0.85	0.92	0.88	402
教育文体	0.95	0.97	0.96	332
环境保护	0.91	0.9	0.91	178

<sup>1</sup> <https://github.com/gaussic/text-classification-cnn-rnn>

表 7-3 分类别给出了所有一级标签的留言数以及三个分类性能评价指标 *Precision*、*Recall* 和 *F-Score* 的分数.通过对表 7-3 的观察发现,7 个一级标签的性能指标均高于 0.8,其中大部分高于 0.9.上述结果分析,说明模型的一级标签分类性能总体较好.

### 7.4.3 对比实验

为了进一步的验证基于字符级 CNN 的留言一级标签分类模型的性能,本文将基于字符级 CNN 的分类模型,与基于 RNN(GRU)、基于 TF-IDF、基于 BOW 的三种分类模型进行两种粒度下的对比实验.

在对比实验中,采用相同的训练集和测试集进行实验,实验结果如表 7-4 所示:

表 7-4 一级标签分类模型不同方法性能对比

<i>F-Score</i>	字符级	分词级
CNN	<b>0.904</b>	0.852
RNN(GRU)	0.846	0.789
TF-IDF	0.895	0.863
BOW	0.892	0.852

通过表 7-4 不同方法的对比,可以发现采用分词粒度的分类模型性能显著低于字符粒度的模型,CNN 相对于其他特征方法,对分类性能的提升更显著,在对比的模型中性能最好,*F-Score* 评分略高于 0.9.基于字符级 CNN 的留言一级标签分类模型的评价优于其他方法,性能有一定幅度的提高.

## 7.5 基于双重 Single-Pass 和热点特征的留言热点检测模型

### 7.5.1 实验数据

模型进行主题发现时,使用比赛提供的附件 3 数据进行聚类与生成主题描述.附件 3 提供了群众留言的留言编号、留言用户、留言主题(即留言标题,与前后文所说主题不同,下文称标题)、留言时间、留言详情、点赞数、反对数等信息,共 4326 条留言.

进行地点、人群实体提取时,模型通过建立的人群实体集(标注符号为:nc)与自定义词典,对附件 2、附件 3、附件 4 的留言详情进行地点、人群实体标注,使用标注后的实体集进行地点、人群实体识别模型的训练.



## 7.5.2 双重 Single-Pass 主题聚类模型实验及结果分析

Single-Pass 阈值(threshold)的范围一般位于 0.6-0.9<sup>[26]</sup>,本文在此基础上对不同阈值进行实验.模型使用文献<sup>[14]</sup>的方法作为基线,与双重 Single-pass 进行对比,结果如图 7-1、图 7-2 所示.

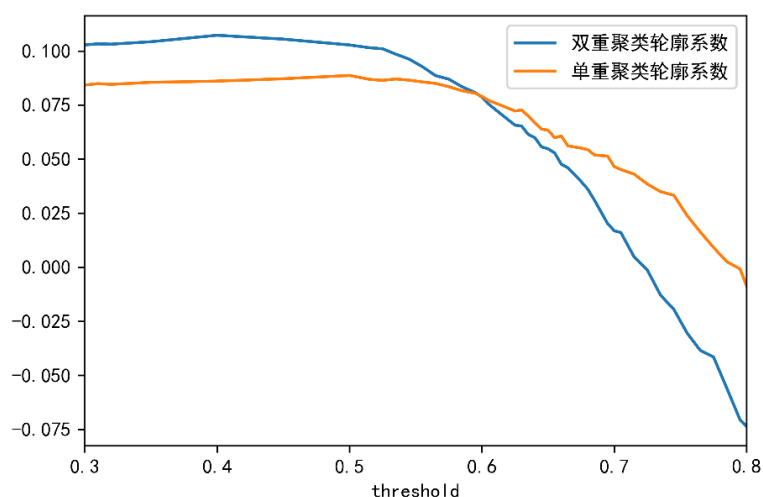


图 7-1 双重 Single-Pass 轮廓系数对比图

由图 7-1 可以发现 threshold 小于 0.61 时,双重 Single-pass 的轮廓系数明显高于单重 Single-pass,两者都随 threshold 值上升而下降, threshold 小于 0.53 时,两种方法的轮廓系数只有轻微波动,此时轮廓系数受 threshold 的影响轻微; threshold 大于 0.53 时,双重 Single-pass 轮廓系数快速下降,至 0.7 左右降为负数,说明此时 threshold 处于双重 Single-pass 的边界左右;而单重 Single-pass 在 0.6 左右才开始缓慢下降,说明双重 Single-pass 的类间界限更明显,聚类效果更好.

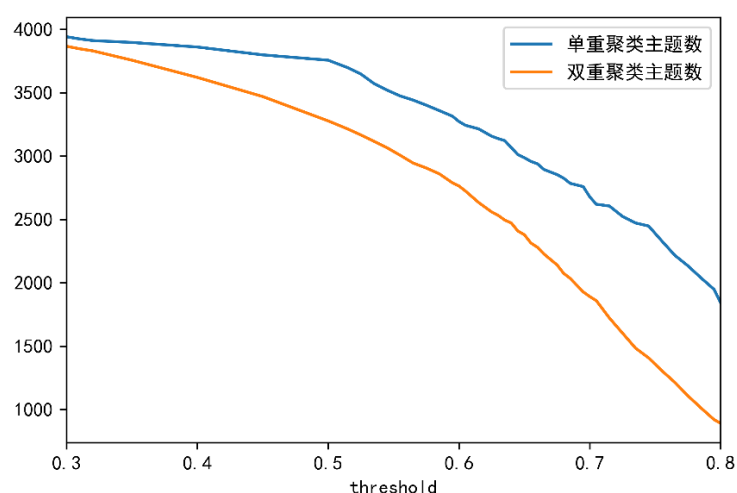


图 7-2 双重 Single-Pass 聚类主题数对比图

图 7-2 给出了不同方法得到的主题数的对比,可以发现双重 Single-pass 的聚

类数明显高于单重 Single-pass,且随着 threshold 的增长差距逐渐,说明双重 Single-pass 能有效解决单重 Single-pass 存在的同一主题不同标签的问题.

综合上述分析,本文确定 threshold 值位于 0.5 至 0.6 之间,对此范围内的 threshold 值聚类情况进行人工评估,最终确定对热点话题聚类性能较好的 threshold 值为 0.5.

7.5.3 TextRank 主题描述生成模型实验及结果分析

表 7-5 聚类效果较差主题描述生成示例

主题描述	A 市经济学院强制学生外出实习
主题留言标题列表	A 市经济学院的体育学院变相强制实习 A 市经济学院强制学生外出实习 A 市经济学院强制学生实习 A 市江山帝景新房有严重安全隐患 咨询移动通信业务问题 A 市经济学院组织学生外出打工合理吗? A 市经济学院寒假过年期间组织学生去工厂工作 对 A 市参保记录的几点疑问

表 7-6 聚类效果较好主题描述生成示例

主题描述	A5 区劳动东路魅力之城小区油烟扰民
主题留言标题列表	A5 区劳动东路魅力之城小区油烟扰民 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 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

模型使用抽取式摘要生成算法抽取主题内最具代表性的留言标题为主题描述,表 7-5、表 7-6 分别展示了聚类效果较差与较好的两种情形的示例.对于聚类效果较差的主题,模型能够识别、提取留言内的最重要的留言标题作为该主题的描述内容;对聚类效果好的主题,其生成的主题描述在留言中具有很强代表性.

通过对比可以得出结论:对于聚类生成的主题,TextRank 算法能有效抽取其中最具代表性的留言标题作为聚类主题的描述.

## 7.5.4 地点、人群实体识别模型实验及结果分析

表 7-7 地点、人群实体识别模型性能指标

评价指标	得分
<i>Precision</i>	<b>0.872</b>
<i>Recall</i>	0.852
<i>F-Score</i>	0.862

buppt<sup>1</sup>实现了基于 Bi-LSTM+CRF 的中文命名实体识别模型,本文在其基础上进行地点、人群的实体修改、训练.表 7-7 展示了基于 Bi-LSTM 与 CRF 的地点、人群实体识别模型的 *Precision*、*Recall* 和 *F-Score* 三个指标得分情况,可以看出三个指标得分都略高于 0.85,这说明模型对文本中实体的识别性能较优,能准确识别政务留言中的地点、人群实体,有较高的实体识别性能.

## 7.6 基于字符级 CNN 的回复满意度的分类预测模型

### 7.6.1 实验数据

表 7-8 郴州政务留言原满意度分布表

满意度	留言数
不满意	3911
非常不满意	81
非常满意	2371
基本满意	33
满意	680
总计	7076

本文使用网络爬虫爬取了郴州市的政务留言数据及满意度评价情况,剔除其中重复与没有满意度评价的留言,剩下的留言共 7076 条,原满意度评价共 5 个类别,满意度分布详情见表 7-8.

通过观察表 7-8 不同满意度的分布,可以发现满意度的分布存在不平衡的现象,“非常不满意”与“基本满意”两个满意度的留言数和比例低于 2%,需要对数据进行平衡处理.

由于群众的满意度的程度比较主观,存在较大的不稳定性,但在满意与否的态

<sup>1</sup> <https://github.com/buppt/ChineseNER>

度这一点上不受其他主观因素的影响,所以在训练时不使用原有的满意度指标,将“不满意”、“非常不满意”两个满意度统一为“不满意”满意度,将“满意”、“非常满意”、“基本满意”三个满意度统一为“满意”满意度,通过将少数类合并到多数类进行数据的平衡处理.

表 7-9 郴州政务留言训练满意度分布表

满意度	留言数
不满意	3738
满意	2878
总计	6616

对爬取的留言观察发现,存在相当一部分的留言与回复,其文本正文以附件形式存在,并不在爬取的留言文本中,留言文本只有极简短的说明,为此,设置了一个词数阈值 10,将总词数不大于 10 的留言进行过滤.处理后的满意度评价分布详情见表 7-9.

### 7.6.2 实验及结果分析

实验时,随机选择 80%的留言共 5292 条为训练数据集进行训练,剩余留言共 1323 条作为测试数据集进行测试.

模型在问题 1 的基础上,进行参数的微调,确定神经网络的参数:非文本特征个数 7,原始序列长度 500,词维数 1300,词嵌入维度 64,学习率 0.001,dropout 比例 0.2,卷积核个数 128,卷积核尺寸为 3,BP 网络隐藏层维数 128,输出类数 2.

实验结果见表 7-10、表 7-11:

表 7-10 满意度预测混淆矩阵

实际满意度	预测满意度	
	不满意	满意
不满意	583	189
满意	236	315

表 7-10 是满意度预测结果的混淆矩阵,评价为“不满意”的留言共 748 条,其中准确预测的有 583 条,错误预测为“满意”的 189 条;评价为“满意”的留言共 575 条,其中准确预测的有 315 条,错误预测为“不满意”的 236 条.结果表明模型对不满意的留言的识别效果更好.

表 7-11 给出了不同满意度的分类预测性能,评价指标为 *Precision*、*Recall* 和 *F-Score*,结果显示“不满意”的指标均高于 0.7,而“满意”的指标只位于 0.6 左右,只有“满意”的 *Recall* 指标小于 0.6,说明模型对不满意的留言的分类预测效果

更好,对满意的留言效果较差.

表 7-11 满意度预测分满意度性能指标报告

一级标签	Precision	Recall	F-Score	留言数
不满意	0.71	0.76	0.73	772
满意	0.62	0.57	0.6	551

### 7.6.3 对比实验

为了进一步的验证基于字符级 CNN 的回复满意度的分类预测模型的性能,本文将基于字符级 CNN 的预测模型,与基于 RNN、基于 TF-IDF、和基于 BOW 的三种分类器在两种粒度下进行对比实验.

在对比实验中,采用相同的训练集和测试集进行对比.结果如表 7-12 所示:

表 7-12 满意度预测模型不同方法性能对比

F-Score	字符级	分词级
CNN	0.679	0.654
RNN(GRU)	0.627	0.586
TF-IDF	0.607	0.631
BOW	0.642	0.644

通过表 7-12 不同方法的对比发现,通过神经网络进行特征提取的模型采用字符级能更有效地捕抓文本的语义信息,基于字符级 CNN 的回复满意度的分类预测模型的评价优于其他方法,性能有小幅度的提升.

## 8 总结

本文的主要工作是利用 NLP 技术对政务留言数据进行处理,包括以下几个方面的工作:

- A. 利用字符级 CNN 模型对留言文本进行分析、处理,并进行一级标签分类;在此基础上融合回复文本特征与相关非文本特征进行回复满意度的分类预测;
- B. 结合热点问题的内部特征与外部特征定义热度指数,对聚类主题进行热点性评价;
- C. 在一级标签分类模型的基础上,对群众留言使用 Single-Pass 算法进行双重主题聚类,并计算相应热度指数进行排名,得到热点问题排名;

D. 使用 TextRank 摘要生成算法生成热点问题描述,并使用基于 Bi-LSTM 与 CRF 的命名实体识别模型提取热点问题反映的地点、人群实体.

实验结果表明,相对于传统方法,本文采用的模型框架,在性能上有一定的提高,具有一定的应用价值及参考意义.

## 参考文献

- [1] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541-551.
- [2] Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8), pp.1735-1780, 1997.
- [3] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *ArXiv abs/1406.1078*, 2014.
- [4] D.G. Bailey and C.T. Johnston, "Single pass connected components analysis", in *Image and Vision Computing New Zealand*, Hamilton, NZ, (2007).
- [5] 肖航,现代汉语通用平衡语料库建设与应用[J],*华文世界(台湾)*,第 106 期,2010.12
- [6] 靳光瑾,肖航,富丽等,现代汉语语料库建设及深加工[J],*语言文字应用*,2005(2):111-120.
- [7] 肖航,语料库词义标注研究[M],云南教育出版社,2016.
- [8] Yoon Kim. Convolutional Neural Networks for Sentence Classification. *ArXiv abs/1408.5882*, 2014.
- [9] Nal Kalchbrenner, Edward Grefenstette, Phil Blunsom. A Convolutional Neural Network for Modelling Sentences. *ArXiv abs/1404.2188*, 2014.
- [10] Zhang Xiang, Junbo Jake Zhao and Yann LeCun. Character-level Convolutional Networks for Text Classification. *ArXiv abs/1509.01626*, 2015.
- [11] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. Learning representations by back-propagating errors. *nature*, 323(6088), p.533, 1986.
- [12] 张辉,李国辉,贾立,孙博良.一种基于 TF·IEF 模型的在线新闻事件探测方法[J].*国防科技大学学报*,2013,35(03):55-60.
- [13] 马永军,刘洋,李亚军,汪睿.基于 Single-Pass 的在线话题检测改进算法[J].*天津科技大学学报*,2017,32(06):73-78.
- [14] 南京邮电大学.一种基于文本分类和聚类分析的网络热点事件发现方法:CN201410432539.4[P].2014-12-24.
- [15] Feng Xu, Jue Liu, Ying He,等. Hot Topic Trend Prediction of Topic Based on Markov Chain and Dynamic Backtracking[C]. *Pacific Rim Conference on Multimedia*. Springer, Cham, 2017.
- [16] Ma, Huifang. Hot topic extraction using time window[C]. *International Conference on Machine Learning & Cybernetics*. IEEE, 2011.
- [17] Minjuan Zhong. Hot Topic Discovery in Online Community using Topic Labels and Hot Features[J]. *Tehnički Vjesnik*,2019,26(4).
- [18] Karen Sparck Jones.Index term weighting.*Information Storage and Retrieval*, Volume 9, Issue 11, Pages 619-633, 1973.
- [19] Mihalcea, Rada & Tarau, Paul. TextRank: Bringing Order into Text. 2014.

- [20] Zhiheng Huang, Wei Xu, Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging. ArXiv abs/1508.01991 ,2015.
- [21] Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. Signal Processing, IEEE Transactions on, 45(11): 2673-2681, 1997.
- [22] Lafferty, J & McCallum, Andrew & Pereira, Fernando & Fields, Conditional. Probabilistic Models for Segmenting and Labeling Sequence Data. 2002.
- [23] Fabio Calefato, Filippo Lanubile, Nicole Novielli. Moving to Stack Overflow:Best-Answer Prediction in Legacy Developer Forums[C]. Acm/ieee International Symposium on Empirical Software Engineering & Measurement. ACM, 2016.
- [24] Roy, P.K., Ahmad, Z., Singh, J.P. et al. Finding and Ranking High-Quality Answers in Community Question Answering Sites. Glob J Flex Syst Manag 19, 53–68 ,2018
- [25] C. J. van Rijsbergen, Information Retrieval, London: Butterworths. <http://www.dcs.gla.ac.uk/Keith/Preface.html>, 1979.
- [26] 邱云飞,程亮.微博突发话题检测方法研究[J].计算机工程,2012,38(09):288-290.
- [27] Peter J.Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20 (1987) 53-65.
- [28] 陈捷. 景区突发事件的话题检测研究[D].北京交通大学,2019.