“智慧政务”中的文本挖掘应用

摘要

针对问题一，关于群众留言分类问题，共分为预处理、特征提取与建向量和分类共三大阶段。预处理阶段，将文本中的特殊符号、重复值及停用词去除，并进行中文分词，得到较为干净的数据。特征提取与建向量阶段，对预处理后的数据建立结巴分词字典，调用腾讯实验室开源的常用 100 万词向量，使用 Gensim 将字典持久化，并结合 TF-IDF 算法，建立 DF 词袋模型，提取出每条留言内容关键字词。在分类阶段，本文使用双向 LSTM 与 LSTM 循环神经网络相结合的算法对文本进行分类，很好地解决了语义依赖问题，同时也避免了神经网络梯度消失和梯度爆炸的问题，模型达到了 80%左右地准确率。在分类模型评估上，本文使用 F1-Score 方法，其结果达到 77%左右；同时，本文通过可视化学习曲线，对模型的性能进行诊断，诊断结果认为本模型的效率是可观的。

针对问题二，热点问题是指某一段时间集中爆发、多人反映的特定地点或特定人群的同一类问题。要找出热点问题，并进行热度评价，待解决的有三点：一是问题的识别；二是问题的归类；三是问题热度的评价。首先，运用文本聚类模型对相似的留言进行归类；然后，运用条件随机场的中文命名实体识别模型识别出特定地点或特定人群，将同一地点人群的问题进行归并；最后建立热度评价指标，使用熵权法确定各指标权重，计算出综合得分，也就是热度指数，根据热度指数进行排名，提取出热度前 5 的热点问题。

针对问题三，综合考虑附件 4 相关部门对留言的答复意见，对答复的相关性、完整性、可解释性、时间性进行定义并综合考虑选取具体特征项：文本语义匹配相似度、答复格式、语义完整、有效字符数、特征词分布路径的平均长度以及答复时间，构建答复意见质量的综合评价指标体系。之后对各项特征项建立相应模型或算法并进行求解，采用层次分析进行评分，最后建立一套答复意见质量的评价方案与模型。

关键词：BiLSTM 模型　TF-IDF 算法　Word2vector 文本聚类 命名实体识别 综合评价体系

目录

一、问题重述

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，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。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是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。

附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。

请利用自然语言处理和文本挖掘的方法解决下面的问题：

1、群众留言分类

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件 1 提供的内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

2、热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点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。请根据附件 3 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，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

3、答复意见的评价

针对附件 4 相关部门对留言的答复意见，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，并尝试实现。

二、问题分析

对于问题一，对于留言分类问题，潜在意思中，要求我们要做的是：在数据作预处理，得到干净的数据的基础上，特征提取与字词向量表示，再输入分类模型训练，最后对训练结果评估，验证模型的可行性。

对于问题二，待解决的问题有三点：一是问题的识别；二是问题的归类；三

是问题热度的评价。首先，对留言主题进行分词、去除停用词等预处理，构建词袋空间，算出 TF-IDF 值，再用 k-means 聚类算法进行聚类。然后，用语料库对条件随机场模型进行训练，利用 jieba 分词对测试语料进行分词和词性标注，并利用训练得到的 CRF 模型进行命名实体的初步识别；通过挖掘未识别实体数据的内部特点和上下文特征，设计了大量的人工规则，进行二次识别，识别出特定地点或人群，将同一地点人群的问题进行归并；最后，综合考虑留言的赞成数、反对数、每个问题留言的数量、每个问题的时间范围进行建立热度评价指标，使用熵权法确定各指标权重，计算出综合得分，也就是热度指数，根据热度指数进行排名，提取出热度前 5 的热点问题。

对于问题三，首先根据附件 4 相关部门对留言的答复意见数据进行分析，对附件 4 进行分类在答复的相关性、完整性、可解释性这三个指标外，同时综合考虑答复的时间长短，引入答复意见时间性的指标。之后对以上四个指标进行量化评价，综合考虑选取文本语义匹配相似度、答复格式、语义完整、有效字符数、特征词分布路径的平均长度以及答复时间指标，进而构建答复意见质量的综合评价指标体系，并综合考虑选取具体特征项：文本语义匹配相似度、答复格式、语义完整、有效字符数、特征词分布路径的平均长度以及答复时间，构建答复意见质量的综合评价指标体系。之后对各项特征项建立相应模型或算法并进行求解，采用层次分析进行评分，最后建立一套答复意见质量的评价方案与模型。

三、模型建立与求解

3. 1 群众留言分类

3.1.1 文本分类概述

文本分类是预先给定的分类体系下，将语料库内容分成某一类或某几类的过程，是一种有监督的学习过程。它首先将文本内容用数学上的符号表示，再用适合的分类算法进行分类。

文本分类主要包括文本预处理、特征选择和特征抽取、构建文本表示形式和训练分类器共五大部分。其中，文本表示通常与特征选择和抽取结合起来，目前最常用的文本表示形式是空间向量表示[1]。

本文通过预处理、特征提取与建向量、分类共三大阶段解决群众留言分类问题。其中预处理阶段调用了官网可靠的停用词字典，特征提取与建向量调用了腾训实验室开源的腾讯词向量，具体流程图如图 1 所示：

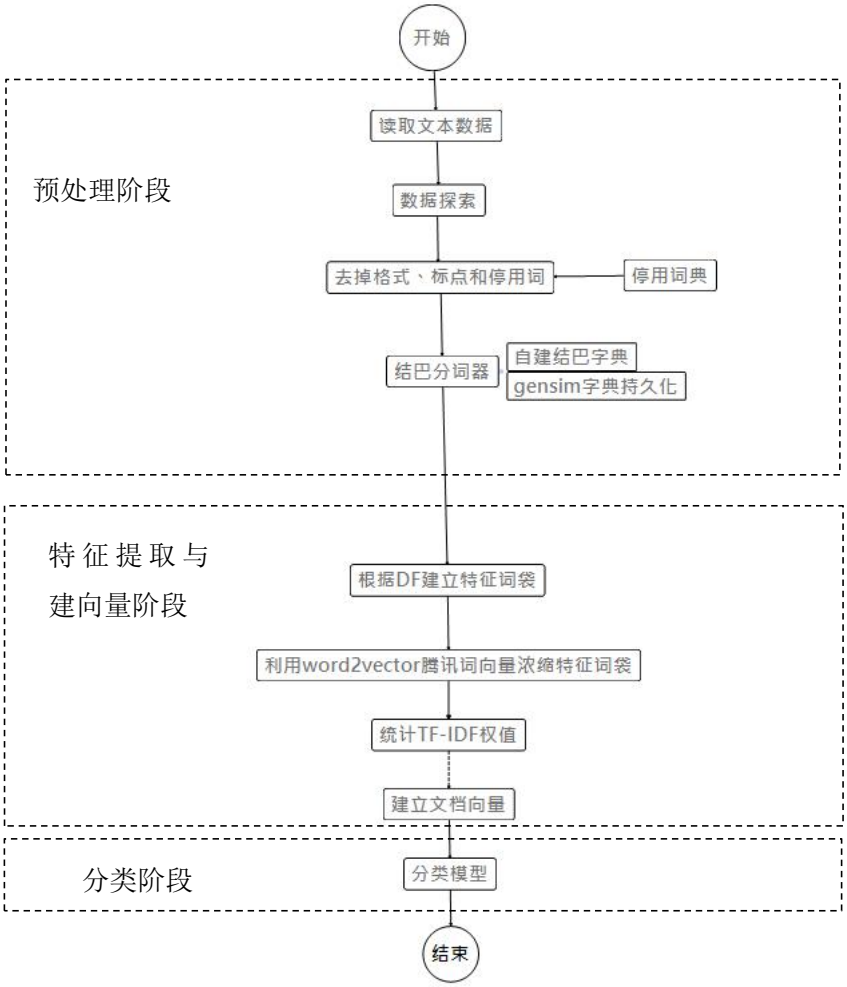


图 1：文本分类流程图

3.1.2 文本预处理

文本预处理是文本分类中最重要的一部分，对后期分类模型有重大影响。对于给定的大量文本数据，可能存在有很多无关的特殊字符、无关主题的字词、语句重复值、拼写错误及编码错误等问题，如果直接利用原始数据建模，不仅造成不必要的开销，同时也会引起模型错误等问题，因而在文本分类中，第一步必先进行预处理。

1.数据探索及文本分词[1]

本附件数据共有 9210 条评论数据，有留言编号、留言时间等共 7 个特征属

性，无空值。本文主要从特殊符号、重复值、分词和停用词四大部分处理。去除重复值后有 9017 条数据。一级标签共分为七类，留言内容，发现有大量无价值的词和特殊符号。其类统计如图 2 所示：

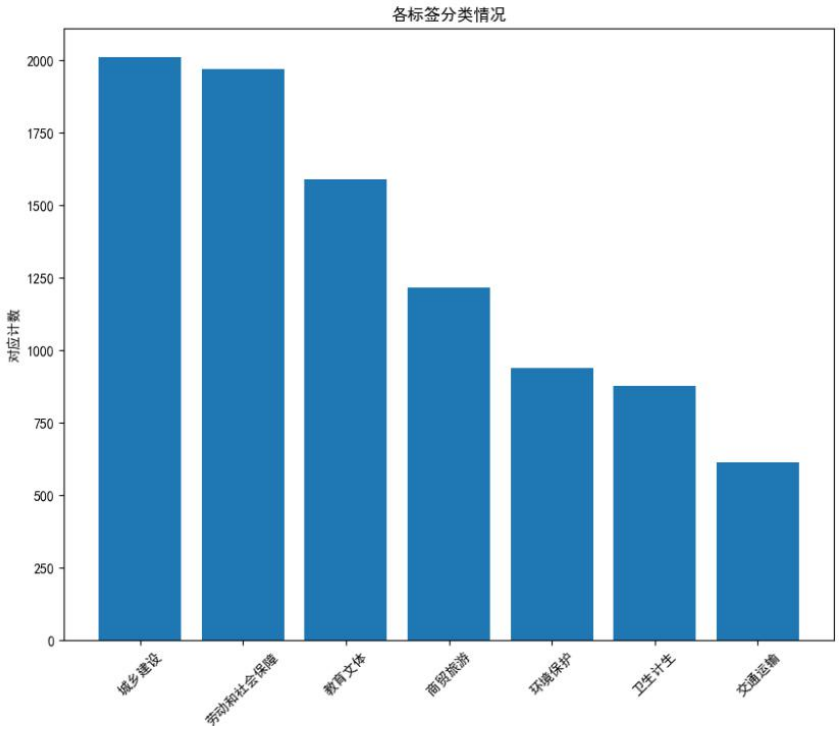


图 2：各标签分类留言数的柱状图

为提高分类的准确性，本文将留言主题合并至留言详情，并重命名为留言内容，将无价值的词和特殊符号去除。

2.文本分词

在中文文本分类问题中，中文分词是必不可少的技术。本文使用 jieba 分词器分词，将文本生成特征基本单位。

3.去停用词

在文本中存在有功能及其普遍，没有实际含义的词语，他们通常是一些数字、符号、高频词，比如“的、我、你、吗”等等。为避免对文本分类造成影响，在预处理阶段将停用词去除。

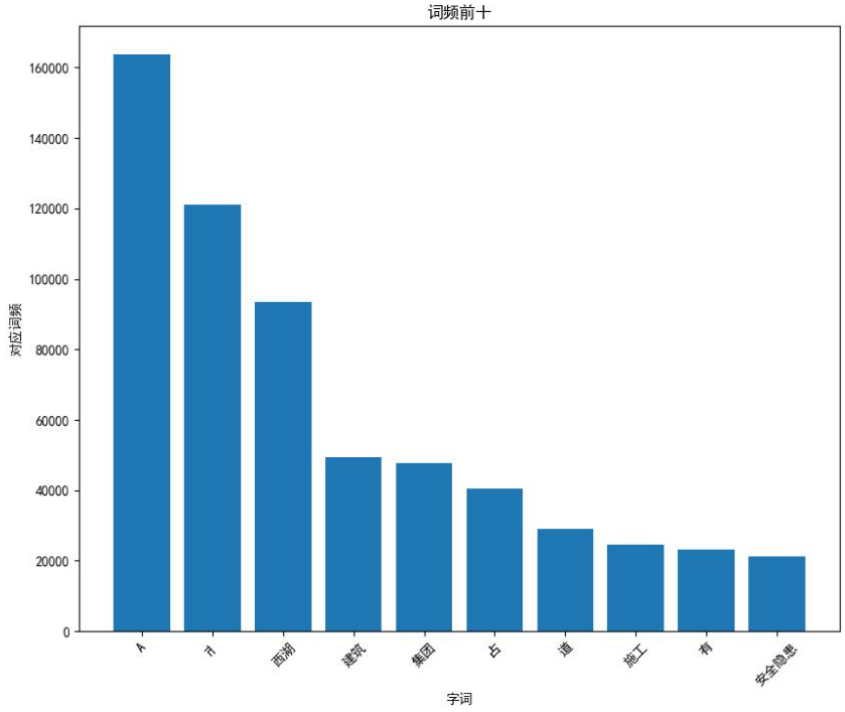


图 3：词频前十情况图 ——去停用词前

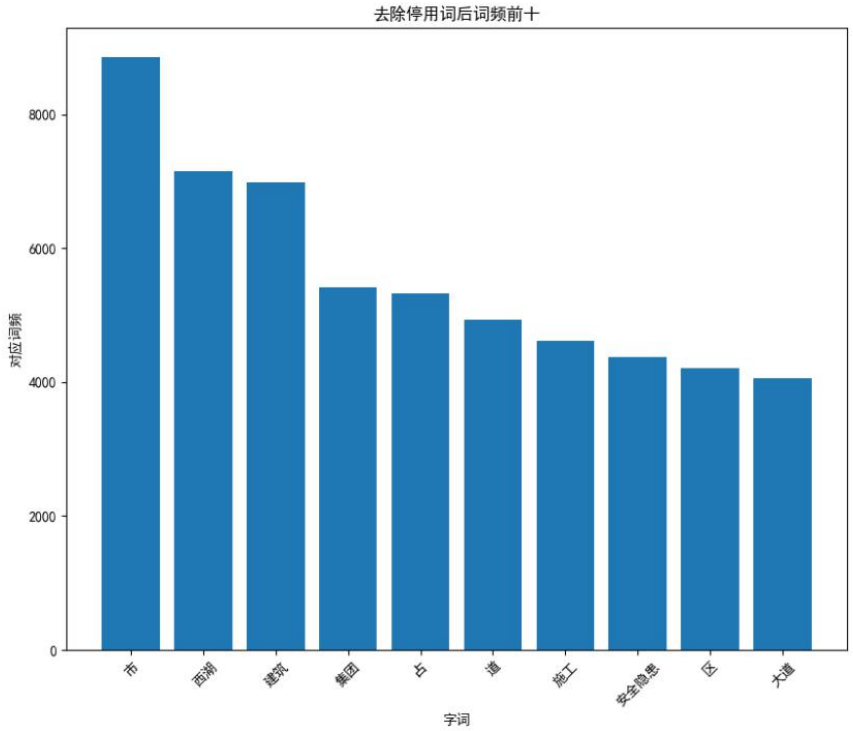


图 4：词频前十情况图 ——去停用词后

以上画出去除停用词前后，词频前十的字词，到此数据预处理结束，已得到了相当干净的数据。

3.1.3 Word2vector 模型原理简介

Word2vector 是 Mikolov 在 2013 年基于神经网络提出的用于快速有效地训练词向量的模型[3]，它通过对词语的上下文及词语与上下文的语义关系建模，将词语映射为 K 维的实数向量。其中，每一维度都代表了词的语义特征，向量之间的距离反映了词语间的语义相似程度，揭示了特征词语之间的相似程度，对于自然语言处理中的文本分类任务有较好的处理效率[4]。Word2vector 模型主要有 CBOW 和 Skip-gram 两种模型，两种模型都包含输入层投影层和输出层。

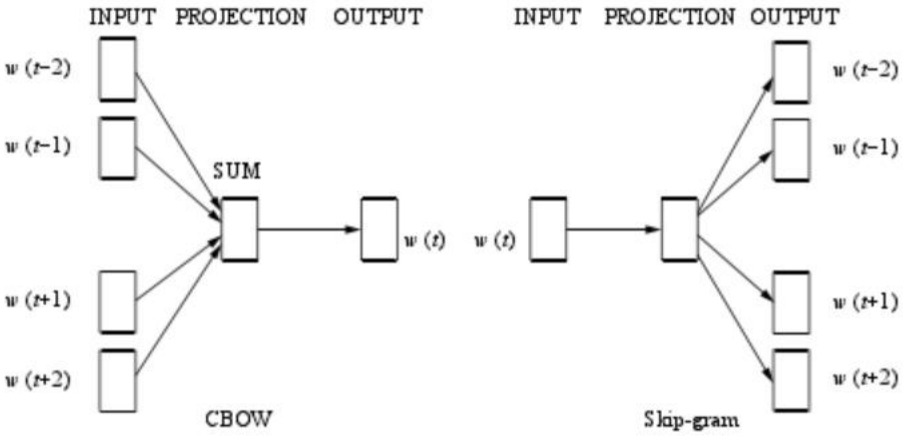


图 5：CBOW 模型与 Skip-gram 模型的原理

CBOW 模型是在已知上下文前提下预测当前词，而 Skip-gram 模型是在已知当前词的情况下预测上下文，即对一个词的预测有 t 次，因而 Skip-gram 模型的效率要比 CBOW 模型高。

$$L_{CBOW} = \sum_{w_t \in C} \log p(w_t \mid w_{ct})$$

$$L_{Skip-gram} = \sum_{w_t \in C} \sum_{-k<c<k} \log p(w_{t+c} \mid w_t)$$

其中，$w_t = w_{t-c},...,w_{t-1},...,w_{t+c}$，$c$ 是给定词 w_t 的前后词语数目，C 代表包含所有词语的语料库，k 代表当前词的前后各 k 个词语。

基于 Skip-gram 模型的高效率，腾讯 AI 实验室改进 Skip-gram 模型，采用自研的 Directional Skip-Gram (DSG)算法，在文本窗口中词对共现关系的基础上，考虑了词对的相对位置，将每个字词用 200 维的向量表示，更加提高了词向量语义表示的准确性。该开源的 800 万词向量涵盖率、准确性与新鲜度都比较高，可供自然语言领域直接使用。

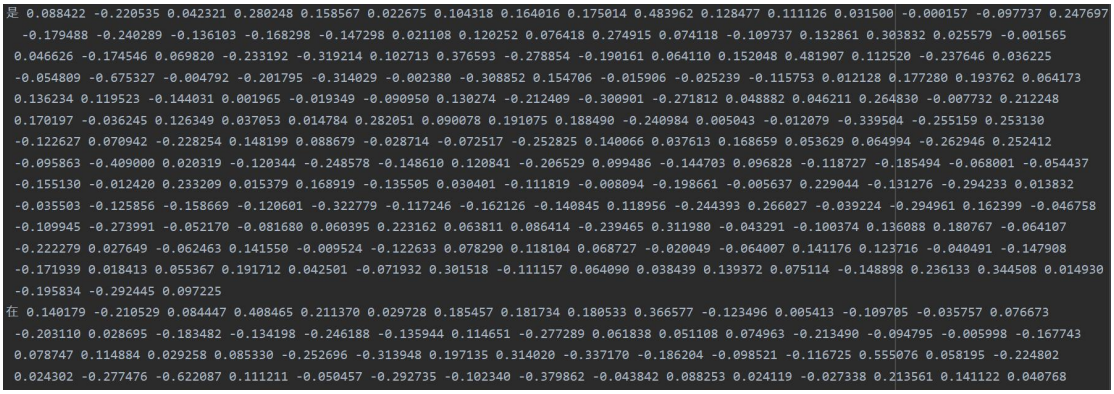


图 6：腾讯词向量示例图

在大量文本数据的任务下，深度学习网络框架中的嵌入层负责文本词汇与向量之间的转换，是一个重要的网络接口。由于腾讯词向量数据庞大，在实践中不易加载，同时实际运用中并不需要如此庞大的向量数据，因而本文使用 Genism 只调用最常用的 100 万腾讯词向量，初始化嵌入层，以减少模型训练代价，得到优性能的深度学习模型。

3.1.4 基于 TF-IDF 的文本向量构建与特征提取

1.TF-IDF 算法原理

TF-IDF(term frequency-inverse document frequency，词频-逆文件频率)是一种统计方法，是文本挖掘的常用加权技术之一，用于评估一个字词对文件或是语料库中的某个文件的重要程度。字词的重要程度与它在文件中出现的次数呈正比关系，与它在语料库中出现的频率呈反比关系。若某个字词在某个文件中出现的频率较高，而在其他文件中出现的频率较低，则认为该字词有较好的区分能力，适合用来分类[1]。

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{k,j}}$$

即：
$$TF_w = \frac{\text{在某一类中词条}w\text{出现的次数}}{\text{该类中所有的词条数目}}$$

其中 TF 表示词频，$n_{i,j}$ 是表示该字词在文件 d_j 中出现的次数，分母表示的是文件 d_j 中所有字词的出现的次数总和。

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

即：
$$IDF = \log(\frac{\text{语料库的文档总数}}{\text{包含词条}w\text{的文档数}+1})$$

其中，*IDF* 表逆向文件频率，$|D|$表示语料库中的文件总数，$|\{j:t_i \in d_j\}|$表示包含词语 ti 的文件数目($n_{i,j} \neq 0$)，为避免该字词不在语料库中的情况，分母以$|\{j:t_i \in d_j\}|+1$ 表示

$$TF-IDF = TF * IDF$$

词频与逆向文件频率作乘得到$tf-idf$ 值，过滤掉文件中常见的词语，提取关键词。

2.文本向量构建与特征提取

在传统的 CountVectorizer 词袋模型向量化中，在统计好词频的基础上，直接将对应的样本对应特征矩阵载入内存，会增加磁盘不必要的负载压力，同时词袋模型中的词向量特征仅仅是以词频表示，无法显示出各词的重要程度。

中文的特征交集大约达到10^5左右的大小[2]，Word2vector 模型能很好地解决传统的文本分类方法，但无法解决高维数据稀疏和语义交叉问题[3]。预处理后的文本数据，特征词往往会有很多，其中包含了很多对分类无价值的数据，如果将预处理后的数据与腾讯词向量结合后，直接输入模型，则可能会因为文本的过度开销，造成数据的维度灾难，增加模型的冗杂度，降低模型的性能[2]。 因而在分类模型前，应先提取文本关键特征。常用的方法有卡方检验、信息增益、互信息等监督方法，文档频词和 TF-IDF 等监督方法。

本文综合考虑了字词权重、计算机磁盘加载负担及模型性能，采用$TF-IDF$算法弥补词袋模型的不足，构建文本向量特征，同时提取各条留言数据的重要字词，以避免语义交叉问题、不必要的维度问题引起模型性能的降低。

本文先是结合腾讯 100 万关键词向量，自建大小为 100 万的结巴词典，再用 Gensim 库将词典持久化，并运用 TF-IDF 算法进行词频统计，依次提取出每句评论数据中权重最高的词语，作为本句评论的关键词。

经过$TF-IDF$算法后，得出每条评论中的关键字词，整理，选前 23 条数据为例，如图 7 所示：

关键词	一级标签
['建筑集团', '西湖', '路段', '西行', '路灯杆', '围墙内', '安置房项目', '未管所']	城乡建设
['在水一方', '烂尾', '护栏', '大厦', '人为', '车辆安全', '电等']	城乡建设
['物业', '车位', '程明', '业主']	城乡建设
['不洗', '水箱', '长年', '楼顶']	城乡建设
['霉味', '霉', '不洗', '健康保障', '自来水龙头']	城乡建设
['停', '盛世', '小区', '电', '供水公司']	城乡建设
['集中供暖', '气候', '近年', '阴冷潮湿', '楼盘']	城乡建设
['停水', '可可', '小城', '桐梓坡', '西路']	城乡建设
['垃圾处理费', '城市']	城乡建设
['小区', '魏家']	城乡建设
['小区', '魏家']	城乡建设
['业委会', '社区居委会', '泰华', '滨江', '非法', '一村', '第四届']	城乡建设
['梅溪湖', '用水', '停水']	城乡建设
['金晖', '鸿涛', '限电', '物业', '延迟交房', '业主']	城乡建设
['来次', '停电', '号线', '三期', '星城', '地铁', '国际', '突然断电']	城乡建设
['高压线塔', '润和', '紫郡', '用电']	城乡建设
['停电', '线路问题', '国际新城', '只有我们']	城乡建设
['融城', '市城区', '金阳大道', '西南']	城乡建设
['滨水', '绿化建设', '城市绿化', '新城', '加大']	城乡建设
['楚府', '停电', '经常停电', '线', '别老', '履行社会责任']	城乡建设
['东安', '建工程', '省建', '集团']	城乡建设
['三单元', '嘉园', '山水', '不同人员', '群租房', '黄谷']	城乡建设

图 7：提取关键字词后的数据示例

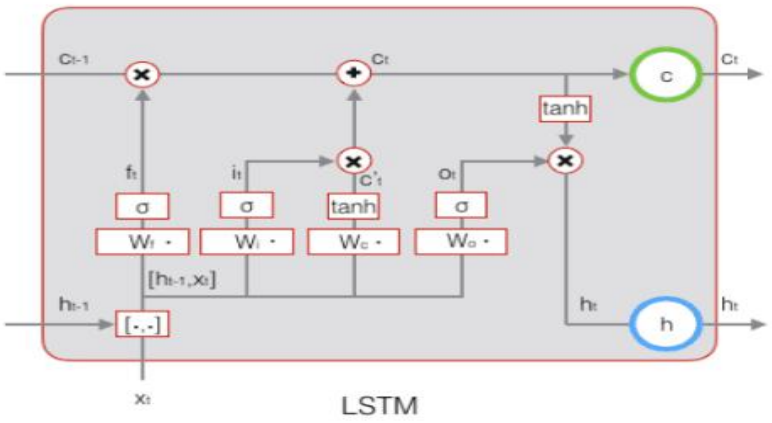
3.1.5 文本分类模型

1.常用分本分类模型概述

LSTM 与 BiLSTM 循环神经网络模型

（1）LSTM 模型原理

LSTM(Long Short Term Memory)是 RNN 模型的一种特殊类型，增加了一个对长序列和长期依赖信息的解决方法。它通过构造一些门，对细胞状态中的信息选择性遗忘和记忆，使得网络记住了很多重要的信息，遗忘无用的信息，避免了 RNN 梯度消失和梯度爆炸的问题[6, 7]。



LSTM

计算遗忘门：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

计算记忆门：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i$$
$$\widetilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

当前时刻细胞状态：

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C}_t$$

计算输出门和当前时刻隐层状态：

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t \tanh(C_t)$$

其中，h_{t-1}表前一时刻的隐层状态，h_t表隐藏层状态， 表当前时刻的输入词，C_t表示当前时刻的细胞状态，f_t是遗忘门的值，i_t是记忆门的值，o_t是输出门的值。

LSTM 有前向和后向两种传播方式，后向传播即是在以上原理基础上从后往前传播。

（2）BiLSTM 模型原理

BiLSTM(Bi-directional Long Short-Term Memory)是由前向的 LSTM 与后向的 LSTM 结合而成的双向 LSTM，它编码了从前往后的信息，同时也编码了从后往前的信息，对不同程度下褒义、中义和贬义的情感词和否定词等相互间的语义依赖有更好的捕捉能力[5]。

BiLSTM 包含输入层、词嵌入层、双向 LSTM 层、聚合层、最大池化层、全连接层和分类层共七层。其中双向 LSTM 层进行特征抽取，聚合层将双向层的向量拼接，再经由最大池化层获取向量中最显著的特征值，最后再经全连接层汇聚为用于情感分类的深度词向量特征。

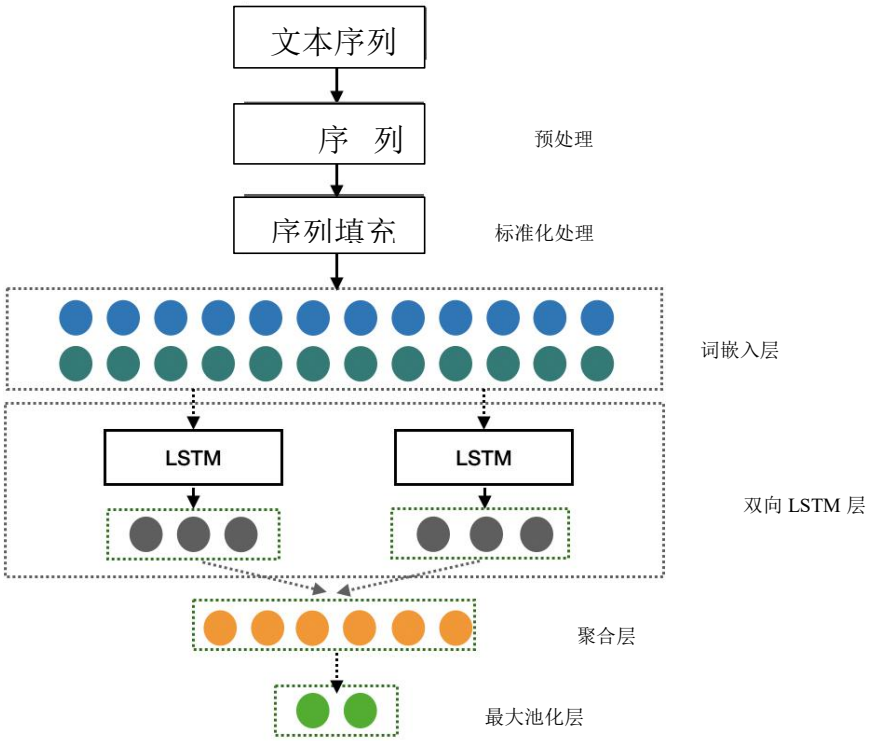


图 8：BiLSTM 模型原理图

对包含有 n 个词语的句子，经输入层和此嵌入层后，再经前向 $LSTM_L$ 层后得 n 个向量 $[h_{L0},h_{L1},...,h_{Ln-1},h_{Ln}]$，经后向 $LSTM_R$ 后，得到 n 个向量 $[h_{R0},h_{R1},...,h_{Rn-1},h_{Rn}]$，前向结合与后向的隐向量结合后得向量 $[(h_{L0},h_{Rn}),(h_{L1},h_{Rn-1}),...,(h_{Ln-1},h_{R1}),(h_{Ln},h_{Rn0})]$，即拼接后取向量聚合层 $[h_0,h_1,...,h_{n-1},h_n]$ 表示句子。

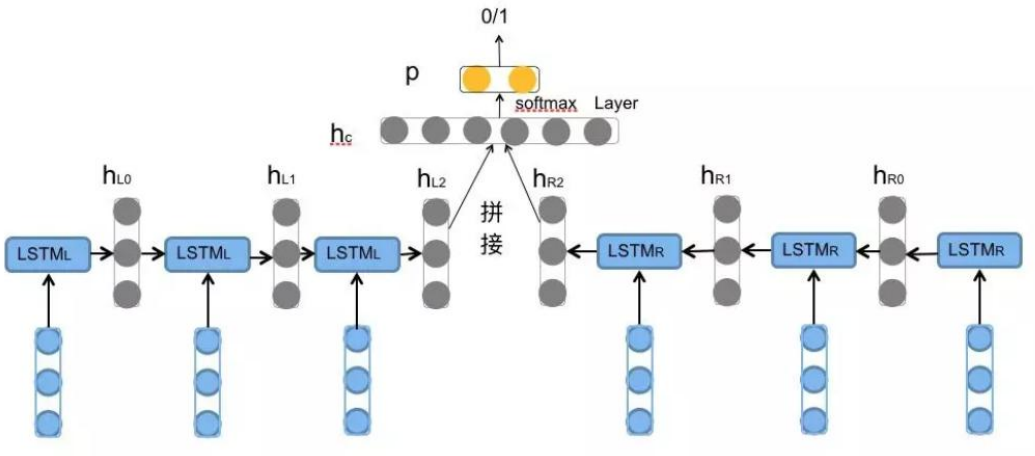


图 9：BiLSTM 模型使用示例图

鉴于BiLSTM的原理，在数据量很大的较细程度的分类下，一般采用BiLSTM模型处理，更好地处理双向语义关系。

2.建立 BiLSTM 留言分类模型

基于 TF-IDF 和 Word2vectors 算法原理，先是 Gensim 加载 Word2vector 下的腾讯词向量，再将所给附件数据中的文本前期预处理后的数据索引化，即把每个评论内容的字词表示为腾讯词向量中对应的索引。接着，运用 TF-IDF 模型提取每条评论数据的关键词序列，将各评论数据索引序列长度结果可视化，得到图10：

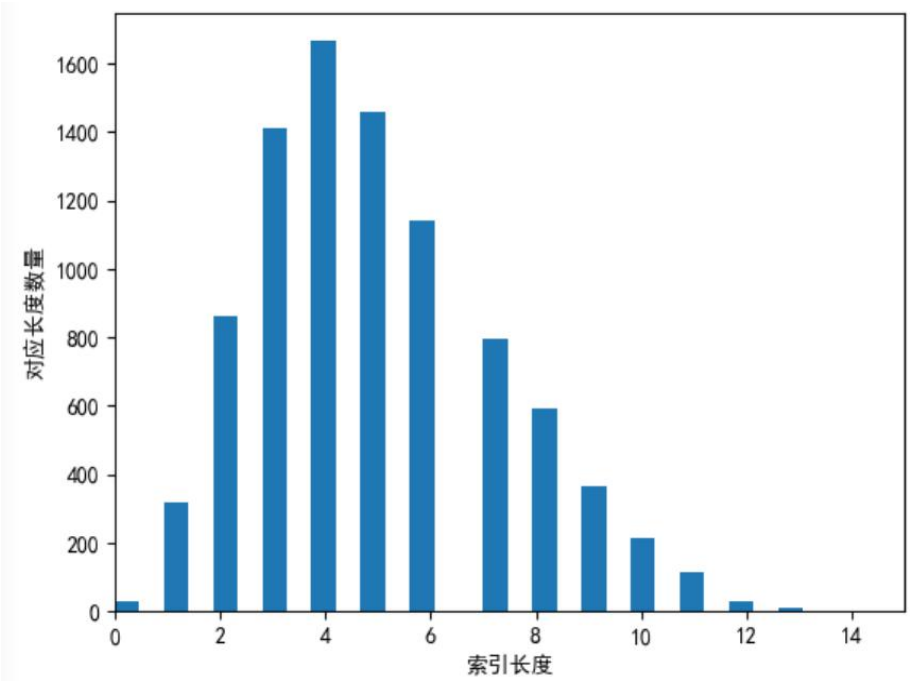


图 10：索引长度分布图

由可视化结果知，经 TF-IDF 算法处理后总体评论数据索引长度不冗长，且各序列长度值分布趋于正态分布。同时，经 TF-IDF 算法处理，各评论数据已为原始评论数据的关键字词，具有很强的区分能力。为避免索引序列较长的评论数据关键词丢失，造成分类影响，将不足长度的评论索引作在前补 0 处理，本文直接将对所有评论数据索引填充至最大索引长度。

基于 LSTM 模型和 BiLSTM 模型原理，将所有评论数据以向量形式输入词嵌入层接口，第一层网络用 32 个单元的 BiLSTM 并返回序列，第二层用 16 个单元的 LSTM 并直接返回最终结果。在全连接层前，加入一个过滤层，设置 Dropout 参数，以避免神经网络模型过度拟合现象。同时，运用早停法，设置在指定轮次的数据遍历后，若训练数据的损失值不再降低，则停止训练，进一步防止模型过拟合现象。

模型训练后得到如下混淆矩阵：

	交通运输	劳动和社	卫生计生	商贸旅游	城乡建设	教育文体	环境保护
交通运输	99	19	0	28	37	1	6
劳动和社	5	475	22	15	19	25	6
卫生计生	1	21	203	11	2	11	2
商贸旅游	15	20	6	239	35	15	19
城乡建设	11	26	3	45	481	17	31
教育文体	0	30	6	32	17	397	3
环境保护	1	12	3	11	34	6	183

模型达到了近 80%的准确性，使用 F1-Score 对分类模型评估，

$$F1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$$

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率

标签	Precision	Recall	F1	Support
交通运输	0.75	0.521053	0.614907	190
动和社会保	0.787728	0.837743	0.811966	567
卫生计生	0.835391	0.808765	0.821862	251
商贸旅游	0.627297	0.684814	0.654795	349
城乡建设	0.7696	0.783388	0.776433	614
教育文体	0.841102	0.818557	0.829676	485
环境保护	0.732	0.732	0.732	250
总体	0.769113	0.767554	0.7665	2706

由计算公式，得出每个标签类别的 F1-Score，总体的 F1-Score 达到了 77%左右，我们认为模型是可行的。

3.模型评估

对模型训练结果作诊断，若模型在训练集上表现很好，在测试集上表现很差，即训练集损失很小，准确性很高，但在测试集上损失很大，准确性很低，则认为模型欠拟合；若模型在训练集和测试集上的效果都不错，损失曲线不断下降并趋于收敛，准确性曲线不断上升并趋于收敛，则认为该模型恰好拟合，有较强的泛化能力；若模型在训练集上的性能很好，但在验证集上的性能达到一个点后开始下降，则认为是过拟合。

考虑到模型的泛化能力，本文通过学习曲线对双向 LSTM 模型的性能进行评估，该学习曲线包括损失曲线和准确性曲线，如图 11-12 所示。

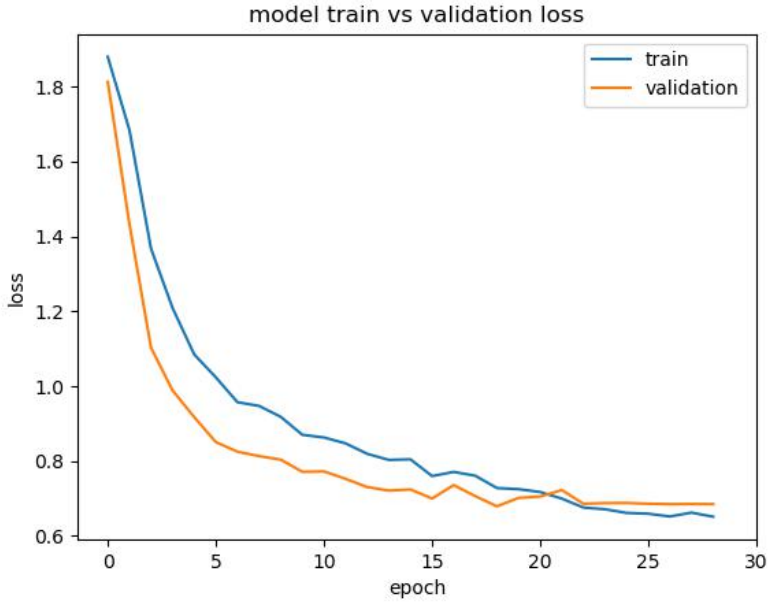


图 11：双向 LSTM 模型损失曲线

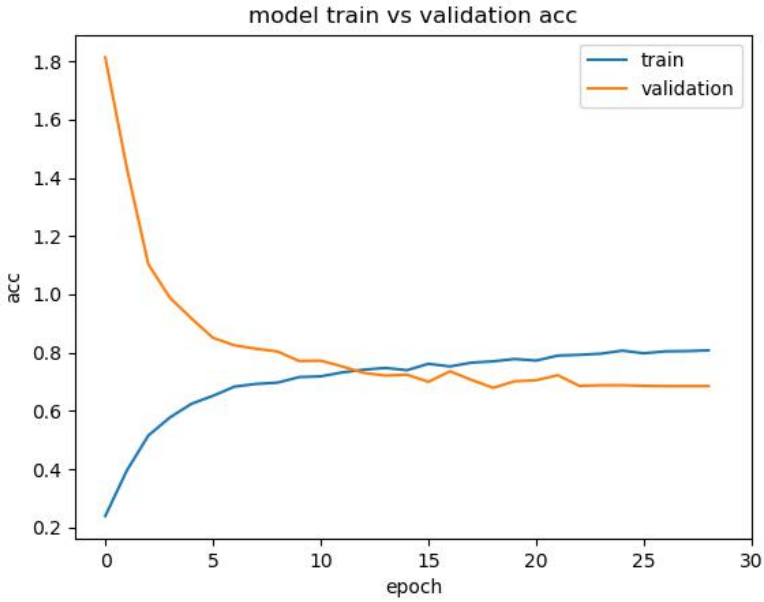


图 12：双向 LSTM 模型准确性曲线

由可视化结果知，模型大概训练至 *epochs* = 28 时，则停止训练。此时，训练集和验证集的损失曲线呈不断下降，并趋于收敛的趋势，效果可观。训练集的准确性趋于上升趋势，而验证准确性集曲线趋于下降趋势，有过拟合现象，但最终训练集和验证集的准确性曲线趋于收敛并达到 70%左右的准确率。

考虑到本文只取了腾讯词向量 100 万常用向量词典，可能忽略了某些重要字词的存在，造成数据缺陷，导致模型了模型误差。若将此模型放至大数据量，且机器允许的情况下，结合所有涉及到的词语向量分析，我们的模型将还是可取的。

3. 2 热点问题挖掘

3.2.1 文本聚类模型

热点问题指某一段时间集中爆发、多人反映的特定地点或特定人群的同一类问题。首先，我们运用文本聚类模型将相似的留言聚为一类，具体步骤如下：

1.分词

利用 jieba 分词工具对附件 3 中的留言主题进行分词

2.去除停用词

下载中文停用词表，过滤掉无意义的词

3.构建词袋空间（VSM）

统计所有词的集合，对于每条数据都构建一个向量，向量的值是对应文本中出现的次数

4.TF-IDF 构建词权重

使用 TF-IDF 来度量每个词的重要程度，统计每个词语的 TF-IDF 权值，获得词在文本中的 TF-IDF 权重

5. k-means 聚类最优 k 值的选取

运用手肘法进行 k-means 聚类最优 k 值的选取，其核心指标是 SSE(误差平方和)，公式如下：

$$SSE = \sum_{i=1}^{k} \sum_{p \in Ci} \left| p - m_i \right|^2$$

其中，Ci 是第 i 个簇，p 是 Ci 中的样本点，mi 是 Ci 的质心（Ci 中所有样本的均值），SSE 是所有样本的聚类误差，代表了聚类效果的好坏。

6.使用 K-means 聚类算法进行聚类[8]

算法过程如下：

（1）从 N 个文本随机选取 K 个文本作为质心

（2）对剩余的每个文本测量其到每个质心的距离，并把它归到最近的质心的类

（3）重新计算已经得到的各个类的质心

（4）迭代 2~3 步直至新的质心与原质心相等或小于指定阈值，算法结束

此处方法中使用余弦夹角来计算文本的相似度，余弦相似性度量为：

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

其中 $similarity(A,B)$ 表示文本 A 和文本 B 之间的相似度，通过计算文本 A 和 B 之间对应特征向量的余弦夹角来判断两文本之间的相似度,余弦值越接近 1,夹角就越接近 0 度,两个向量则越相似。

3.2.2 基于条件随机场的中文命名实体识别模型[9]

命名实体识别（NER）是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等文字。问题二需要从留言主题中提取出特定的地点人群，我们采用条件随机场的中文命名实体识别模型来解决这个问题。

1.条件随机场模型定义

Lafferty 等定义的条件随机场（简称 CRF）为：设无向图 $G=(V,E)$，其中 V 是该无向图中所有顶点的集合，E 是其边的集合。Y 是 G 中的顶点索引，即 $Y=\{Y_v|v\in V\}$。当 Y 的出现条件依赖于 X，且 Y_v 根据图结构的随机变量序列具有马尔可夫特性，即 $p(Y_v|X,Y_v,w\neq v)=p(Y_v|X,Y_v,(w,v)\in E)$，则称$(X,Y)$是一个条件随机域。其中，最常用的是链式条件随机场，结构如图 13 所示。

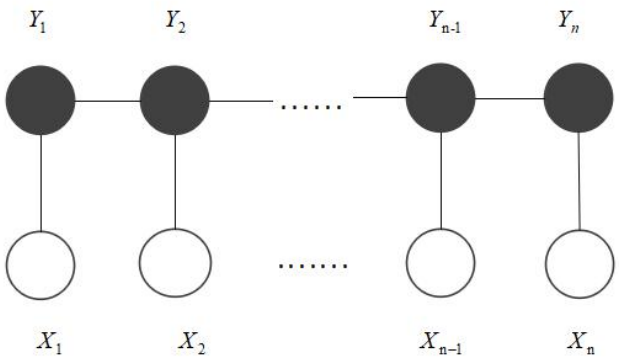


图 13：链式条件随机场

若 $X=(X_1,X_2,\cdots,X_n)$表示被观察序列，$Y=(Y_1,Y_2,\cdots,Yn)$ 表示状态序列，在给定随机变量序列 X 的条件下，其状态序列条件概率：

$$P(y|x)=\frac{1}{Z(x)}\exp(\sum_{i=1}^{n}\sum_{i=1}^{k}\lambda_k f_k(y_{i-1},y_i,x_i)$$

其中 Z 是归一化因子，公式如下：

$$Z（x）=\sum y \exp(\sum_{i=1}^{n}\sum_{k=1}^{k}\lambda_k f_k(y_{i-1},y_i,x_i)$$

2.基于 CRF 的命名实体识别总体流程

第一步 首先将 1998 年 1 月份《人民日报》分为 50000 句训练语句和 50000 句不重叠测试语料; 将训练语料进行标注转换后，利用 CRF 模型对转换后的语料进行训练，最终生成模型参数。

第二步 利用 jieba 分词对测试语料进行分词和词性标注，并利用上一步得到的 CRF 模型进行命名实体的识别。最终将词形和词性标注序列转换为本文定义的标注集序列。

第三步 通过挖掘未识别实体样本的内部特点和上下文特征，设计了大量的人工规则，在上一步识别的基础上，进行二次识别。召回了为识别的地名和组织名。

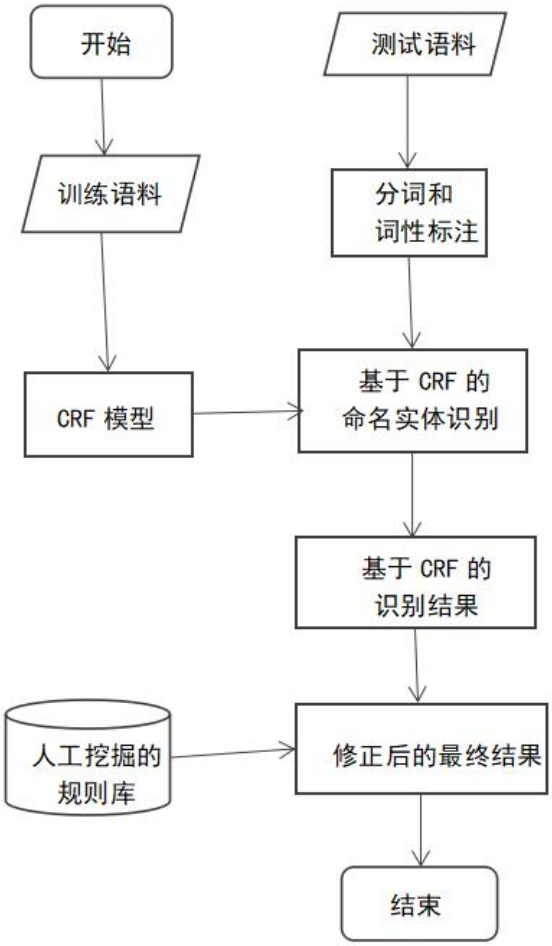


图 14：基于 CRF 的命名实体识别总体流程图

3.命名实体标注集的选择

我们首先定义条件随机场模型的训练所需要的标准集，标注以词尾对象，标

注集中主要定义了命名实体的开始、内部、结尾和非命名实体几种类型，如表 1 所示。

表 1　命名实体内部标注集

标注	含义
B	当前词为地理命名实体的首部
I	当前词为地理命名实体的内部
E	当前词为地理命名实体的尾部
S	当前词是地理命名实体
O	当前词不是地理命名实体或组成部分

4. 特征构造思路

（1）构造词性：直接用 jieba 分词标注分词即可

（2）构造词语边界特征：遍历所有的词，用 len 函数判断这个词的长度，如果长度是 1，那么标记为 W，否则首部标记为 B，内部标记为 I，尾部标记为 E

（3）构造实体指示词特征：在中文中，有些词的出现通常标志着该词周围很可能出现相应的命名实体，这样的词，我们称之为命名实体指示词。命名实体指示词是文本中非常有意义的上下文信息，可以有效的帮助识别命名实体。通常可以分为人名指示词、地名指示词和组织名指示词。遍历每个词，判断这个词是否在实体指示词表里。

（4）构造特征词：特征词就是某一类词具有同一种特征。比如组织名基本都是以特定的词汇结尾，比如公司、学校、小区等等；地名都是以省、市、区、县、等特定的词结尾。

（5）构造常用词

（6）构造标签：对文本进行 jieba 标注分词后，会得到每个词的词性，比如人名标注词性为/nr,组织名是/nt，地名是/ns。

（7）合并以上的特征

5.评测指标

正确率、召回率和 F 值是评测中文命名实体识别系统性能的指标，定义如下：

（1）正确率 P：

$$P = \frac{N_c}{N'}$$

其中 N_c 表示识别正确地点命名实体的总个数，N 表示识别出来的地点命名实体的总个数。

（2）召回率 R：

$$P = \frac{N_c}{N}$$

其中 N_c 表示识别正确地点命名实体的总个数，N 表示测试语料中地点命名实体的总个数。

（3）F 值：

$$F = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 P + R}$$

F 值是综合了争取率和召回率两个值进行评估的办法，同时考虑了 P 和 R 两个值，其中 β 是一个权重，我们取值为 1，即对两个指标都同时重视。

6. 问题的归并

在识别出来的相似留言问题中对同一地点人群的问题进行归并。

3.2.3 热度评价

1.构建热度评价指标

（1）关注热度：附件 3 给的赞成数和反对数，能够反映网上的关注度，我们定义某一热点问题关注热度计算公式为：

$$X_{\mathrm{k}} = \sum_{i=1}^{n} (Yi + Ni)$$

其中 X_{k} 为第 k 个热点问题的关注热度；n 为同一个热点问题的留言数量；Yi 为同一个热点问题第 i 条留言的赞成数；Ni 为同一个热点问题的反对数。

（2）反映度：我们定义反映度为每个热点问题的留言数，其反映了群众对该问题的反映度。

（3）反映持续时间:每个热点问题的 n 条留言中，最早时间与最晚时间的之差即为该问题的反映持续时间。

2 综合热度评价

运用熵权法，对热点问题的 3 个指标进行赋权，从而计算得到该热点问题的热度指数的综合得分，具体步骤如下：

（1）进行标准化处理，$\mathrm{y_{ij}} = \dfrac{x_{ij} - x_{j\min}}{x_{j\max} - x_{j\min}}$，其中 $\mathrm{y_{ij}}$ 为经过无量纲化处理的第 i 个热点问题的第 j 个指标值；x_{ij} 为第 i 个热点问题第 j 个指标数据的原始值。

（2）定义标准化 $Y_{ij}=\dfrac{y_{ij}}{\sum_{i=1}^{m}y_{ij}}$

（3）指标信息熵值 e 和信息效用值 d，第 j 个指标的信息熵为 $e_j=-\dfrac{1}{\ln m}\sum_{i=1}^{m}Y_{ij}\ln Y_{ij}$，信息效用值 $d_{ij}=1-e_j$

（4）评价指标的权重。信息效用值越大，表明指标越重要，对评价的重要性就越大。最后可以得到第 j 个指标的权重为

$$W_j=\frac{d_j}{\sum_{j=1}^{n}d_j}$$

（5）综合得分 $F=\sum W_j y_{ij}$

3.3 答复意见质量的综合评价指标体系与模型

针对问题三，综合考虑附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性、时间性的角度对答复意见的质量建立一套评价方案与模型。

3.3.1 答复意见质量的综合评价指标体系

首先根据附件 4 相关部门对留言的答复意见数据进行分析，对附件 4 进行分类在答复的相关性、完整性、可解释性这三个指标外，同时综合考虑答复的时间长短，引入答复意见时间性的指标。之后对以上四个指标进行量化评价，综合考虑选取文本语义匹配相似度、答复格式、语义完整、有效字符数、特征词分布路径的平均长度以及答复时间指标，进而构建答复意见质量的综合评价指标体系，如表 2 所示。

表 2 答复意见质量评价指标

	一级评价指标	二级评价指标	评分
答复意见质量评价指标	相关性	文本语义匹配相似度	$y=y_1$
	完整性	答复格式	$y=0.5y_1+0.5y_2$
		语义完整	
	可解释性	有效字符数	$y=0.5y_1+0.5y_2$
		特征词分布路径的	

		平均长度	
	时间性	答复时间	$y=y_1$

（1）答复意见的相关性从“文本语义匹配相似度”进行评判，将网友的留言与答复意见进行相似度计算以及分析，得出答复意见与留言的相似度。针对文本相似度判定，余弦相似度和 SimHash 两种算法都可用判定，但余弦相似度适用短文本相似度判定，而 SimHash 算法适合于长文本文本相似度判定，并且该算法能应用于大数据环境中。

（2）答复意见完整性从“答复格式”“语义完整”两方面进行判定。

答复格式应包含“来信人”“问候语”“答复语”“感谢语”以及“时间”即大多数答复意见中“网民 xxx”“您好”“答复如下”“感谢您对我们工作的关心、监督与支持。”“xx 年 xx 月 xx 日”等词语。若这些词语在回信中提及到，便形成较为完整的答复格式。基于语义词素相似度识别算法，利用机器学习，训练自动识别以上答复格式的特定词语，并从文本中识别同义特定内容，对其内容与特定词语内容进行比较，即可判断回复固定格式较完整。

对于语义完整性，目前还没有公认的统一定义。根据实际工作的需要，认为一条语句如果能完整表达出意思，不再需要借助其他的语句，不容易产生歧视，就称它为语义完整。语义完整性分析[10]的主要就是判断一句话是否语义完整。首先需要将数据集转为指定格式传入神经网络，之后基于双层 Bi-LSTM 语义完整性分析模型方法，通过训练数据集进行训练，得到最终的训练模型给预测时使用。

（3）可解释性目前还没有公认的统一定义。根据实际工作的需要，认为文本用语简洁、规范、无冗余，阐述问题、现象、原理，揭示问题本质的深度。根据这一定义，选取主要特征项“有效字符数”“特征词分布路径的平均长度”来判定。

（4）答复时间性对网民反馈的热点问题的及时解答，快速采取相应措施与方案解决起到重要作用，因此答复时间对答复意见质量评价有着较大影响。通过附件 4 获取网民留言时间以及有关部门的回复时间，对留言时间与答复时间的差值的绝对值（即答复速率）进行判定与评价。

3.3.2 评价指标的算法与模型

1.文本相似度 SimHash 算法：

主要步骤如下：

（1）文档特征量化为向量；

（2）计算特征词汇哈希值，并辅以权重进行量化；
（3）针对 f-bit 指纹，按位进行叠加运算；
（4）针对叠加后的指纹，若对应位为正，则标记为 1，否则标记为 0。

如图 15 所示，本文结合问题三解释如下：Doc 表征一篇文本，feature 为该文本经过中文分词后的词汇组合，按列向量组织，weight 为对应词汇在文本中的词频，之后经过某种哈希计算得出哈希值，见图中 1 和 0 的组合，剩余部分不再赘述。

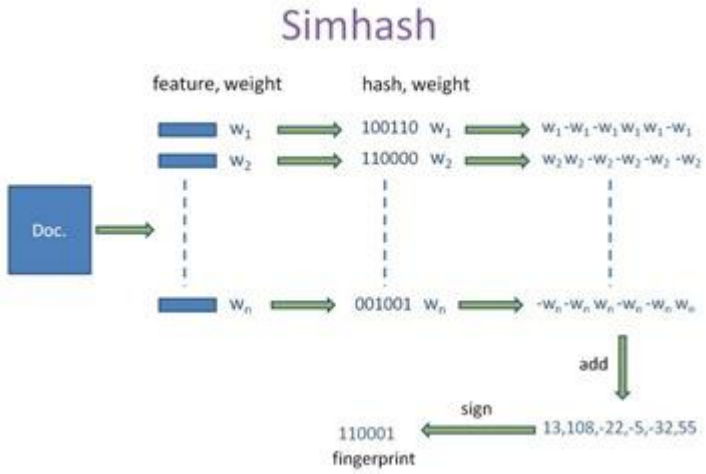


图 15　SimHash 处理过程

2.基于语义词素相似度识别算法

算法主要思想：

首先，建立常用词素的语义词典，对识别词进行切分，在此基础上以词素为单位，以相似性原理为依据，将词素的字面形式转换为语义代码进行相似度判别，在考虑词组的结构关系的前提下进行同义词的识别。其中引入表达度概念，表示词的部分对整体含义所起的作用大小，据此进行加权。

表达度是表示词的部分对整体含义所起的贡献大小，这个词的表达度为 100%，空串对整体的表达度为 0%，词的部分对整体的表达度等于词素表达度之和。

公式成立的条件是以下信息已知：

待匹配词 ctrlword 的信息量总和为 A；

匹配词 keyword 的信息量总和是 B;

两词中表示相同语义的信息量为 C1，C2；

共同部分 C1 对 A 的表达度为 x，C2 对 B 的表达度为 y。

根据这些条件可得：

$$x=\frac{C_1}{A},\ y=\frac{C_2}{B} \qquad (C_1=C_2=C)$$

相似度为：

$$xsd = \frac{2}{\frac{1}{x}+\frac{1}{y}} \qquad (x, y \text{ 不为 } 0)$$

3.基于双层 Bi-LSTM[10] 语义完整性分析模型

采用改进的双层 Bi-LSTM 来进行训练，其中每个层包含多个存储器单元，能够更加准确地学习特征. 第一层 Bi-LSTM 给后一层的 Bi-LSTM 提供序列输出，而不是单个值输出.

①输入层

首先对经过清洗后的数据集进行分词，然后采用四元标注集 T={S, B, M, E}进行标注. 定义 B 表示为一个语义完整句的开头词，M 表示为一个语义完整句的中间词，E 表示为一个语义完整句的结尾词，S 表示为特定符号 (，：、等) 前面和后面最靠近的一个词。

随机欠采样后的词序列因为上下文特征改变，可能会出现欠拟合的现象，为了既对词序列进行采样，又不丢失一个词应有的上下文信息，在随机欠采样前对序列数据进行处理。对于一个有 n 个词的词序列 T(1:n)，用大小为 k 的滑动窗口从首滑动至尾，每次窗口内的子序列作为 Bi-LSTM 的输 入。假设 k 值为 5，序列 T 中下标为 i(下标从 0 开始) 的词生成的子序列表示为 (Ti–2, Ti–1, Ti, Ti+1, Ti+2)，其中 Ti=T[(n+i)%n]。

② 双层 Bi-LSTM

为了方便说明，这里假设滑动窗口的大小为 5 双层 Bi-LSTM 其内部结构如图 16 所示：

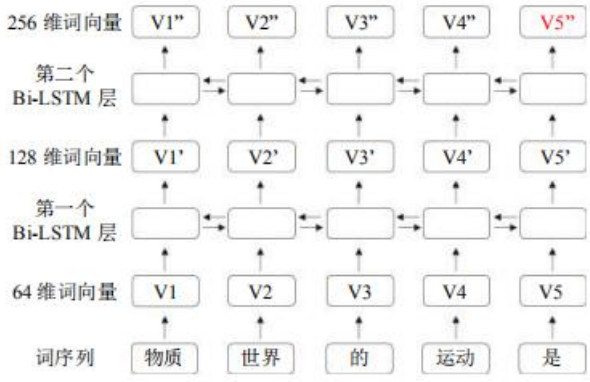


图 16：双层 Bi-LSTM 结构图

在输入层我们已经把输入的词序列转换为维度为 64 的词向量，图 4 中小矩形的数目即序列的长度. 在第一个 Bi-LSTM 中，这里输入为维度 64 的词向量，输出为维度 128 的词向量，由于其不是最后一层 Bi-LSTM，这里会输出 5 个 128 维的词向量 V1'…V5 '。 第二个 Bi-LSTM 输入为 V1'…V5'都为 128 维词向量，经转换后得到 V1"…V5"为 256 维词向量，当前已经是最后一层

Bi-LSTM, 所以这里规定 V5”为窗口中间词即词向量 V3 对应的输出。

③ 输出层

深层神经网络中, 过拟合会使模型泛化性能变差, 为了防止过拟合, 模型中增加 Dropout 层。Drop out 层将在训练过程中每次更新参数时按一定概率随机断开输入神经元, 这就防止了神经元之间过度的协同适应。Dropout 层的输出向量维度与 Bi-LSTM 的输出维度相同, 为了将向量维度转换为与标签类别数一致, 所以增加了一个全连接层, 并采用 elu 激活函数, Dropout 层的输出转换为指定维度的向量. 最后对提取的特征采用 Softma x 激活函数得到输出的概率。

Softmax 的函数定义如下:

$$S_i = \frac{e^{V_i}}{\sum_i^c e^{V_i}}$$

其中, Vi 是全连接层的输出, i 表示类别索引, 总的类别个数为 C, Si 表示的是当前元素的指数与所有元素指数和的比值. 一个含任意实数的 K 维向量, 通过 Softmax 层后, 都会“压缩”到另一个 K 维实向量中, 压缩后的向量每个元素都在[0, 1]范围中, 并且所有元素 的和为 1.

4.有效字符数模型

公式为:

$$\text{read}(\text{TC}) = \begin{cases} 0, & \text{length}(\text{TC}) \leqslant l_0; \\ k, & \text{length}(\text{TC}) \in (l_{k-1}, l_k]; \\ k+1, & \text{length}(\text{TC}) > l_4. \end{cases}$$

式中,lk 可根据实际调整,本文采用的 lk 分别取值为 20,50,100,150,200;其中 k=0,…,4. 函数 length 为获取到的文本字符数,包括文本中出现的标点符号.

5.特征词分布路径的平均长度

答复意见中提及的政策在概念层次上所占的层次越高,则可解释性越强. 因此,根据答复涉及的政策方案最大深度来衡量,公式为:

spec(TP)=max(depth(tpi)),tpi∈CS. (5)

6.答复时间

通过附件 4 获取网民留言时间以及有关部门的回复时间，通过对留言时间与回复时间的差值的绝对值来描述答复时间的长短。

即 $\Delta t = |t_1 - t_0|$

注：t_0 表示留言时间，t_1 表示回复时间

3.3.3 质量评价综合评分

1.评价指标综合评分表

分值	指标			
	相关性	完整性	可解释性	时间性
0	0%	0%	0%	>14day
1	0%～5%	0%～12%	0%～4%	10～14day
2	5%～10%	12%～20%	4%～8%	7～10day
3	10%～20%	20%～30%	8%～15%	5～7day
4	20%～30%	30%～41%	15%～20%	3～5day
5	>30%	>41%	>20%	<3day

2.答复意见的各项指标分值与人工评分均值(部分)

留言用户	公式提取的指标分值				人工评分均值
	相关性	完 整性	可解释性	时间性	
A00045581	3	2	1	0	1.50
A00023583	3	2	3	0	2.00
A00031618	3	3	3	0	2.25
A000110735	3	3	3	0	2.25
A0009233	2	2	2	0	1.50
A00077538	3	2	2	0	1.75
A000100804	3	2	2	0	1.75
UU00812	5	3	3	0	2.75
UU008792	4	2	1	0	2.25
UU008687	2	1	1	0	1.33
...	...	...	...	...	...

对所有指标分值与人工评分均值使用 Spearman 相关系数进行相关性分析.结果表明,相关性、完整性、可解释性、时间性与人工评分均值的相关系数分别为 0.427、0.439、0.482、0.371,相关系数的显著性均为 0.000,小于 0.01,说明以上 4 个指标与答复意见的相关性是显著的.

3.质量评价的回归模型构建

根据上文特征提取的指标分值,进行机器学习训练的基本思路如图 17 所示.已知 X=[xij]mn,Y=[yi]m,其中 m 为数据集的数量,n 为指标数量,即每条答复意见有

4 个特征指标分值,构成 xi. 人工评价分值的平均分作为 yi,得分在 0～5 之间。

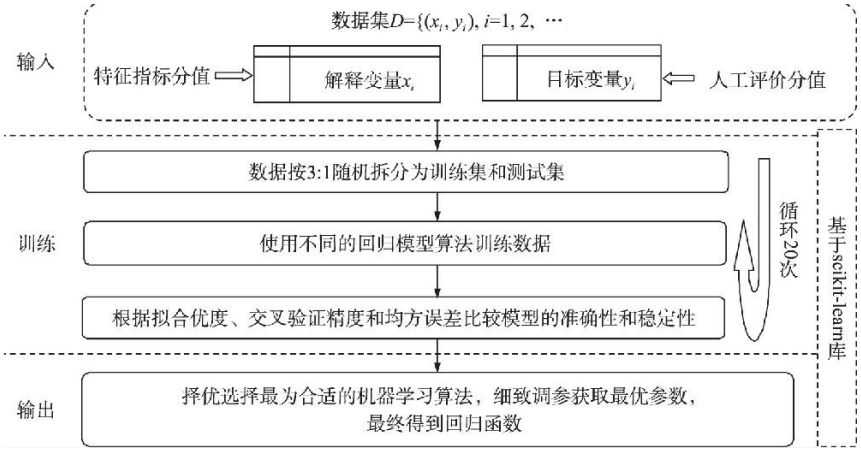


图 17： 质量评价的回归模型构建

四、参考文献

[1]但宇豪,黄继风,杨琳,高海.基于 TF-IDF 与 word2vec 的台词文本分类研究[J].上海师范大学学报(自然科学版),2020,49(01):89-95.

[2]牛雪莹,赵恩莹.基于 Word2Vec 的微博文本分类研究[J].计算机系统应用,2019,28(08):256-261.

[4]汪静,罗浪,王德强.基于 Word2Vec 的中文短文本分类问题研究[J].计算机系统应用,2018,27(05):209-215.

[3]朱磊. 基于 word2vec 词向量的文本分类研究[D].西南大学,2017.

[5]藏润强,左美云,郭鑫鑫.基于 Doc2Vec 和 BiLSTM 的老年患者疾病预测研究[J/OL].计算机工程与科学 :1-8[2020-05-08].http://kns-cnki-net.vpn.vpn.hzu.edu.cn/kcms/detail/43.1258.tp.20200429.1051.006.html.

[6]欧阳红兵,黄亢,闫洪举.基于 LSTM 神经网络的金融时间序列预测[J].中国管理科学,2020,28(04):27-35.

[7]李艳萍,赵晓宇.基于 LSTM 的空气质量预测方法[J].科技与创新,2020(07):7-9.

[8]王俊丰,贾晓霞,李志强.基于 K-means 算法改进的短文本聚类研究与实现[J].信息技术,2019,43(12):76-80.

[9]何炎祥,罗楚威,胡彬尧.基于 CRF 和规则相结合的地理命名实体识别方法[J].计算机应用与软件,2015,32(01):179-185+202.

[10]刘京麦野. 基于循环神经网络的语义完整性分析[D].湘潭大学,2019.

[11]穆亚昆,冯圣威,张静.基于文本与语义相关性分析的图像检索[J].计算机工程与应用,2019,55(01):196-202.

[12]杨涛. 基于标签相关性的文本多标签分类算法的研究[D].北京工业大学,2019.

[13]李雪岩,孙济庆.一个结合信息相关性分析的快速文本分类模型[J].计算机应用与软件,2004(11):12-13+69.

[14]王元波,骆浩楠,汪峥.文本挖掘在主题发现和相关性评估中的应用——以人工智能和机器人领域的专利为例[J].工业控制计算机,2020,33(02):102-103+106.

[15]邹沁含,庞晓阳,黄嘉靖,刘司卓.交互文本质量评价模型的构建与实践——以 cMOOC 论坛文本为例[J].开放学习研究,2020,25(01):22-30.

[16]刘金晶,王丽英.在线学习社区发帖质量评价的回归模型研究[J].南京师范大学学报(工程技术版),2020,20(01):33-41.