关于文本挖掘在“智慧政务”中的实际应用

摘要

近年来随着互联网的普及，越来越多的人通过网络来表达自己对某些事件、现象、政策的一些看法和意见。而微信、微博、市长信箱、阳光热线等网络问政平台也逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，但是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。此时需要政府部门对群众留言进行分类并提取热点问题，针对分析结果及时解决群众的各项困难。

目前关于文本分类的研究在计算机科学领域有较多的成果。本文以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为数据，对相关问题进行研究。根据文本内容使用 TF-IDF 和词嵌入对进行特征提取，用 BLSTM 模型对留言内容进行标签分类；利用 TextRank 提炼关键词，同时利用 LDA 模型进行相关主题匹配，从而挖掘群众留言中的热点问题；对于网络问政答复质量，从相关性、时效性和完整性的角度设计衡量答复质量指标。这将为相关部门及时了解民生民意提供了可靠依据，对提升政府的管理水平和施政效率具有极大的推动作用。

关键词：群众留言、TF-IDF、BLSTM 模型、TextRank、热点问题、LDA 模型

目录

1. 结合 TF-IDF 的双向长短期神经网络分类模型

本章针对问题一的要求和所提供的数据信息，提出了结合 TF-IDF 的双向长短期神经网络模型用于文本分类。通过对赛题内容及所给数据的分析，确定留言主题和留言详情中的信息可用于群众留言分类模型中。我们采用将传统统计信息和深度学习相结合的方式，设计留言分类模型。首先，利用中文分词工具 Jieba 分词对留言主题和留言详情进行分词处理；其次，利用 Word2vec 将留言主题中的词语转化成为词向量输入到长短期神经网络模型中；再次，利用 TF-IDF 获得留言主题中的传统统计信息，通过特征选择方法 CHI 来减低维度，将得到的信息和长短期神经网络的输出进行拼接；最后，设计全连接层以获得分类输出。

1.1 相关理论技术介绍

1.1.1 文本预处理

处理文本数据前需要对文本数据进行分词、去停用词、关键词提取等处理。中文文本中词与词之间没有明显的分隔符，为有效理解文本，第一步需要进行分词。中文分词算法主要有基于字符串匹配、基于理解和基于统计。分词结果中含有许多对表达句子意思没有实际意义的噪音词，这些噪音词应作为停用词去除，同时提取对句子意思具有决定意义词。

本文分词采用当前广泛使用的基于 Python 的中文分词工具 Jieba，Jieba 中文分词工具内置多个算法，支持多种模式进行分词，能有效解决未登陆词和歧义词。去停用词可以降低句子噪音对句子的理解，减少特征词的数量，从而提高文本分类的准确性。如果英语中 the、is、at、who 等，中文中的在、的、和等副词、量词、介词、叹词、数词等词，这些对理解语句没有实际意义，而且出现频率较高，容易造成噪音，分词后应从分词结果中将这些停用词进行过滤。去停用词只要建立停用词表，然后采用字符匹配的方式扫描分词词典进行删除。

1.1.2 Word2vec

Word2vec 是一种基于统计方法来获得词向量的方法，它是 2013 年由谷歌的

Mikolov 提出的一套新的词嵌入方法。Word2vec 在整个 NLP 里的位置如图 1 所示。

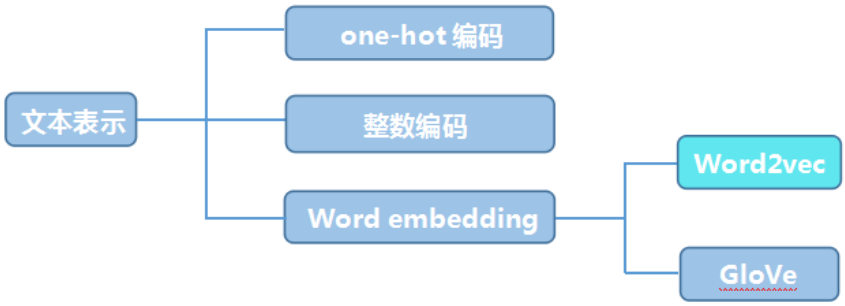

图 1

Word2vec 可以在百万数量级的词表和上亿的数据集上进行高效地训练，而得到的训练结果——词向量，可以很好地度量词与词之间的语义相关性和相似性。Word2vec 包含了两个网络模型：CBOW（Continuous Bag-of-Words）模型和 Skip-Gram 模型，网络结构如图 2 所示。

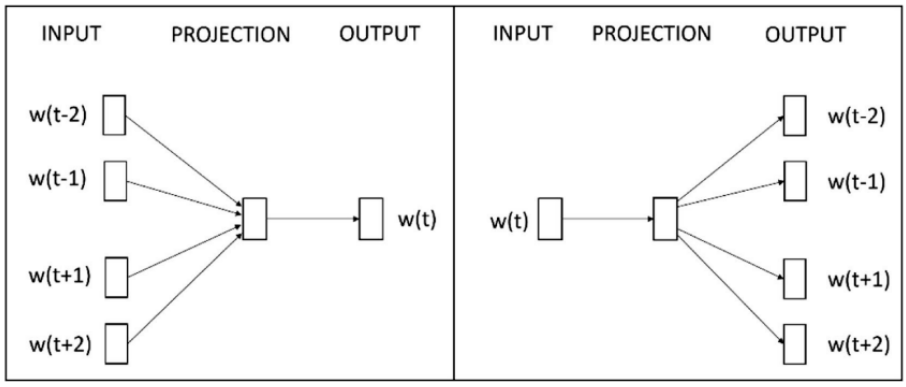

图 2　CBOW 模型（左）和 Skip-Gram 模型（右）

1.1.3 TF-IDF 模型

TF-IDF 由 TF 和 IDF 组成。具体来说，TF（term frequency，词频）指的是某一个给定的词语在该文件中出现的次数。IDF（inverse document frequency，逆向文件频率） IDF 的主要思想是：如果包含词条 t 的文档越少，IDF 越大，词条的类别区分能力越强。TF-IDF 的主要思想是如果某个词在一段文本中出现的频率高，并且在其他文本中出现的次数较少，则认为该词具有较强的类别区分能力，适合成为文本特征表示。具体如下：

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

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

注：这里的分母之所以加 1 是为了避免分母为 0。

$$TF - IDF = TF \times IDF$$

1.1.4 CHI 特征选择方法

CHI 统计方法是用来测量特征项 w 与类别 c 之间的相关性，其关联列表如表 1 所示。

表 1　特征项与类别关系

特征项	类别		
	K	$\overline{K}$	总计
W	A	B	$A+B$
$\overline{W}$	C	D	$C+D$
总计	$A+C$	$B+D$	$A+B+C+D$

其中：属于类别 K 且存在特征项 w 的文本数量，命名为 A；不属于类别 K 但存在特征项 w 的文本数量，命名为 B；属于类别 K 但不包含特征项 w 的文本数量，命名为 C；不属于类别 K 也不包含特征项 w 的文本数量，命名为 D。

则 CHI 值的计算式为：

$$X^2(w,c) = \frac{N(AD-BC)^2}{(A+C)(A+B)(B+C)(C+D)} \tag{1}$$

且有 $N = A+B+C+D$。当 CHI 的值等于 0 时，表示特征项 w 与类别 c 之间没有任何关系；当 CHI 的值越大时，表示特征项 w 与类别 c 的关系性越强。

1.1.5 双向长短期记忆模型

长短期记忆（Long Short-Term Memory，LSTM）是 1997 年由 Sepp Hochreiter 提出的一种神经网络，在循环神经网络的基础上增加了单元状态（cell status）来

保存长期状态，LSTM 允许神经网络可以选择保存更早期的历史信息并选择何时遗忘这些信息。LSTM 的结构示意见图 3。

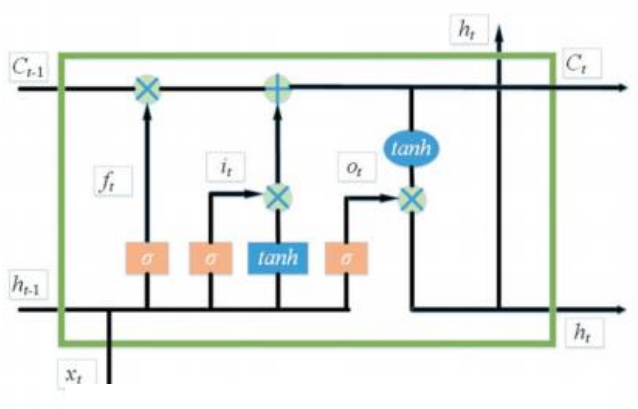

图 3　LSTM 神经元内部结构图

图 3 展示了记忆神经元计算隐藏层状态 h_t 和神经元输出 C_t 的过程。其中，遗忘门通过 h_{t-1} 和 x_t 对神经单元的上一状态信息进行操作。对于来自 C_{t-1} 的信息，当遗忘门 f_t 的值为 1 时，则将 C_{t-1} 的信息保留到 C_t 的计算中；若遗忘门 f_t 的值为 0，则将 C_{t-1} 的信息丢弃。具体的计算方法如公式（2）所示：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$

接下来，同样通过 h_{t-1} 和 x_t 计算出输入门 i_t 和候选状态 $\tilde{C}_t$ 的值，计算公式如下：

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ C_t &= \tan_h(WC \cdot [h_{t-1}, x_t] + b_C) \end{aligned} \tag{3}$$

然后，在神经元上一状态 C_{t-1} 和候选状态 $\tilde{C}_t$ 的基础上，综合输入门 i_t 和遗忘门 f_t 的值，可以获得神经元的当前状态值 C_t，计算方法如公式（4）所示：

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{4}$$

最后，由输出门和神经单元的当前状态值可确定神经单元的输出状态 h_t，计算公式如下：

$$\begin{aligned} o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t \cdot \tanh(C_t) \end{aligned} \tag{5}$$

双向长短时记忆神经网络（BLSTM）本质上是两个独立的 LSTM 网络，一

个网络从头到尾学习数据，另一个网络从尾到头学习数据。

其网络结构如图 4 所示。

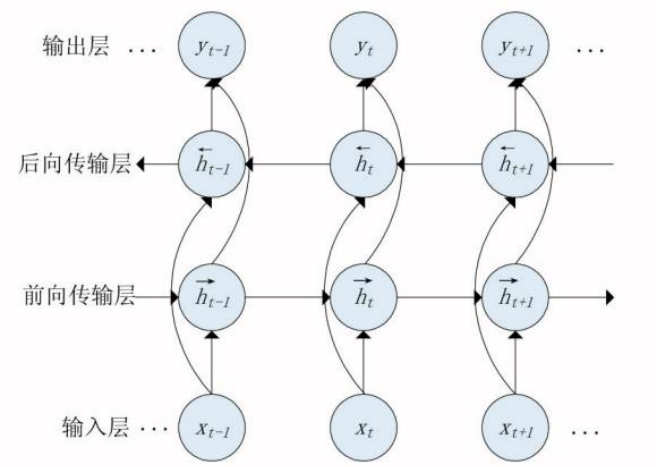

图 4　BLSTM 的网络结构图

图 4 中，BLSTM 网络中两个并行的隐藏层同时从正序和逆序两个方向处理序列信息。它的两个并行神经元分别从前向 $\vec{h}$ 和后向 $\overleftarrow{h}$ 两个方向计算隐藏层状态并合并到输出层。

前向传播隐藏层的计算公式如式（6）所示：

$$\vec{h}_t = \sigma(W^{\overline{hx}} x_t + W^{\overline{hh}} \vec{h}_{t-1} + b_{\vec{h}}) \tag{6}$$

后向传播隐藏层的计算公式如式（7）所示：

$$\overleftarrow{h}_t = \sigma(W^{\overline{hx}} x_t + W^{\overline{hh}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}) \tag{7}$$

然后，将两者合并到输出层可以得到 y_t 可以表示为式（8）：

$$y_t = W^{\overline{hy}} \vec{h}_t + W^{\overline{hy}} \overleftarrow{h}_t + b_y \tag{8}$$

其中，$W^{\overline{hx}}$、$W^{\overline{hh}}$ 和 $W^{\overline{hy}}$ 分别表示输入层到隐藏层的权重矩阵、隐藏层神经元间的权重矩阵和隐藏层到输出层的权重矩阵。$b_{\vec{h}}$、$b_{\overleftarrow{h}}$ 和 b_y 是各层计算是的偏差向量。

1.2 留言分类模型

1.2.1 单词嵌入层

单词嵌入层负责将每个单词映射到一个高维向量空间，该空间可以捕获单词

的语义和句法信息，矩阵的每一列存储相应单词的单词嵌入。令$\{x_1, x_2 \cdots x_T\}$表示输入句子中的单词序列。我们使用 Word2vec 预训练的词向量来获得每个词的固定词嵌入，词嵌入的向量被输入进一个两层的 Highway Network。输出是一个 d 维向量的序列，或者直接说是个矩阵$x \in R^{m \times T}$，其中 m 是一个词向量的维度，T 是句子长度。

1.2.2 BLSTM 层

在本模块中，我们使用 BLSTM 模型来进一步提取留言详情文本信息。在上一层提供的嵌入基础之上，我们使用 BLSTM 网络对句子表示进行建模。具体模型如上文所述，双向长短期神经网络每个方向上的输出为h，我们获得$\mu \in R^{2h}$作为留言详情文本语义的表示结果。

1.2.3 主题特征层

在特征重要性选择方面，传统文本分类方法中通过特征工程提取的一些特征比神经网络提取的特征更具优势。因此，我们还使用一些传统方法来获取一些统计值，这些统计值可以表示为与上一部分获得表示拼接的其他特征。这样，就相当于将传统的特征选择与神经网络的训练过程分开，并最终融合了这两个部分的特征。这种融合方法不仅可以充分发挥传统特征构造方法和神经网络的优势，而且还可以通过人工设置传统特征构造方法中的特征数量来避免文本泛化能力降低的问题，从而实现文本分类。该模型可以更好，更快地为文本选择最有意义的功能，并避免大量冗余功能。我们还可以保留原始的低阶特征以及利用高阶特征。

通过特征选择获得的低阶特征，它仅作用于要分类的当前句子。我们通过 TF-IDF 获得特征权重，使用特征选择的一种方法 CHI 减小维数并提取最相关的特征。经过此步我们可以获得$\mu \in R^{2h+T}$来作为留言主题和留言详情的文本表示。

1.2.4 输出层

输出层专门为任务设置，我们将上文输出的留言主题和留言详情输入该层，得到输出结果：

$$z = soft\max(Wq + b) \tag{9}$$

1.2.5 分类模型

根据赛题要求以及数据特点，我们将留言分类模型设计成了一个多阶段的分类模型。具体流程如下：

第一阶段是词嵌入阶段，我们使用预训练的 Word2Vec 将词映射为对应的词向量表示；

第二阶段是 BLSTM 阶段，使用 BLSTM 来获得留言详情的表示矩阵；

第三阶段是融合主题特征阶段，采用 TF-IDF 提取留言主题特征，并使用 CHI 方法对特征进行降维处理；

第四阶段是全连接层阶段，通过输入留言主题表示和留言详情表示的拼接信息以得到分类结果。

留言分类流程如图 5 所示。

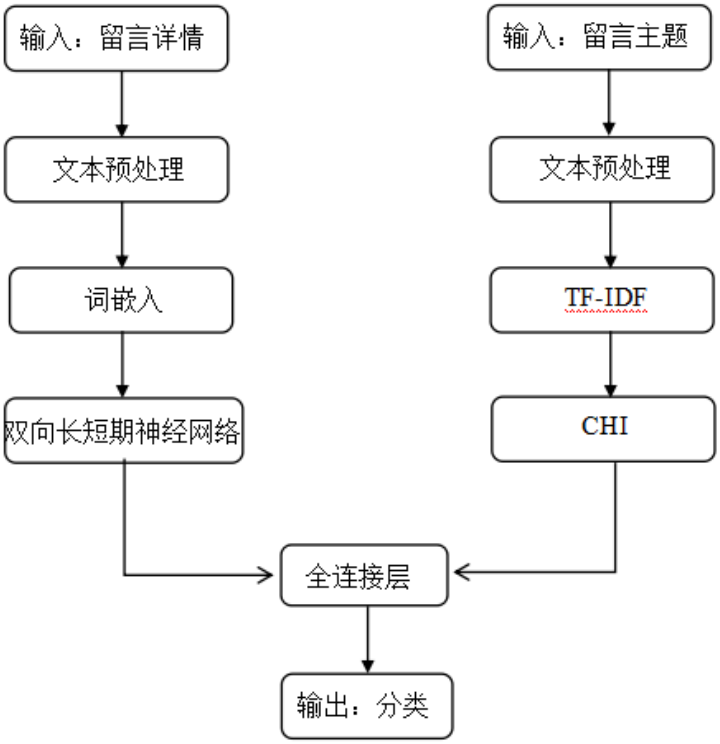

图 5　留言分类流程图

2. 基于 TextRank 的 LDA 主题发现模型

本章针对问题二的赛提要求与信息数据，我们考虑先将留言主题与留言内容留言内容分开进行处理，使用 TextRank 分别进行关键词的抽取，然后将其结果一方面进行 K-means 聚类，获得聚类结果，另一方面过滤提取出的关键词词性，仅保留名词词性的词，作为待定的人群/地点，最后将聚类结果与待定人群/地点通过 LDA 计算余弦相似度来进行话题匹配。

2.1 相关理论技术介绍

2.1.1 利用 TextRank 提取关键关键词

TextRank 算法是将一篇文档转换成一张有向带权的词图模型，它是将文本分割成基本单元，也就是词语。其中，将每个基本单元都看作是一个节点，节点与节点之间的边由词节点之间的共现关系所决定，而节点的重要性又依据相邻节点的指向数量。该算法由 PageRank 算法演变而来，PageRank 最初被用来计算网页的重要性，而 TextRank 的计算公式则在 PageRank 的公式的基础上，引入了新的概念——边的权值，这代表了两个句子的相似度。

TextRank 的模型一般可以表示成一个有向有权图$G=(V,E)$，该图由点集合V和边集合E组成，E是$V\times V$的子集。该有向有权图$G=(V,E)$中，将任两点V_i，V_j之间边的权重记为为w_{ji}，对于一个给定的点V_i，$In(V_i)$为指向该点的点集合，$out(V_i)$为点V_i指向的点集合。点V_i的得分定义见公式（10）。

$$WS(V_i)=(1-d)+d*\sum_{V_j\in I_n(V_i)}\frac{w_{ji}}{\sum_{V_k\in out(V_j)}w_{jk}}WS(V_j) \qquad (10)$$

其中，d 代表阻尼系数，且$d\in[0,1]$，这代表从图中某一特定点指向其他任意点的概率，一般取值为 0.85。

2.1.2 文本聚类算法

在使用 TextRank 对留言详情与留言主题进行初步提取后，对得到的关键词通过使用 TF-IDF 与 K-means 算法进行聚类。

在论文的第一部分中对 TF-IDF 算法进行了简单的介绍, 该算法的原理是如果某个词语在文本中出现的频率较高，并且在其他文章中很少出现，就可以认为该词语具有较好的文本概括性，可以当作主题和关键词。

K-means 算法也叫 K 均值聚类算法，是一种通过迭代求解的聚类分析算法。

步骤如下：

1)首先确定一个 k 值，k 为聚类数目。

2)从数据集中随机选择 k 个数据点作为质心。

3)对数据集中每一个点，计算其与每一个质心的距离（如欧式距离），离哪个质心近，就划分到那个质心所属的集合。

4)把所有数据归好集合后，一共有 k 个集合。然后重新计算每个集合的质心。

5)若新计算出来的质心和原来的质心之间的距离小于某一个设置的阈值（表示重新计算的质心的位置变化不大，趋于稳定，或者说收敛），则可以认为聚类已经达到期望的结果，算法终止。

6)若新质心和原质心距离变化很大，需要迭代步骤 3)- 5)。

该算法中的聚类数目 k 的初始取值为 $k=\left[\frac{n}{\alpha}\right]$，$n$ 为整个数据集的长度，α 是参数，k 值随着随着文本数目的大小而不断变化。

2.1.3 主题匹配

LDA(Latent Dirichlet Allocation)模型是在潜在语义分析（Latent Semantic Analysis, LSA）模型上进行的改进，由于本论文针对的留言详情是篇幅较长的，因此采用对于长文本效果更好的 LDA 模型。

LDA 主题模型是一种典型的词袋模型，其核心思想认为文档是由词构成的一个集合，而文档中的词是没有顺序的独立个体，一篇文档中可以包含多个主题，每一个主题又由词生成。

具体过程如下：对算法每篇需要生成的文档，首先从文档的多项式分布中抽取一个主题，然后根据选定的主题，从对应的词项的多项式分布中抽取一个单词，重复上述过程直到满足文档中的长度为止。结构图见图 6。

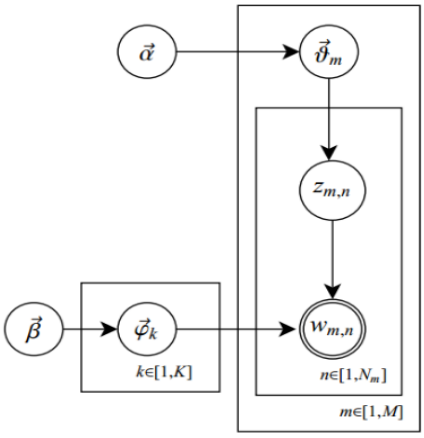

图 6　结构图

2.2 模型构建

关于热点问题挖掘，本文首先对留言主题与留言详情进行预处理，使用 TextRank 算法初步提取特定人群与特定地点，然后对处理好的留言进行聚类，最后通过 LDA 模型中的余弦相似度思想进行话题匹配。该论文拟采用的总技术路线见图 7。

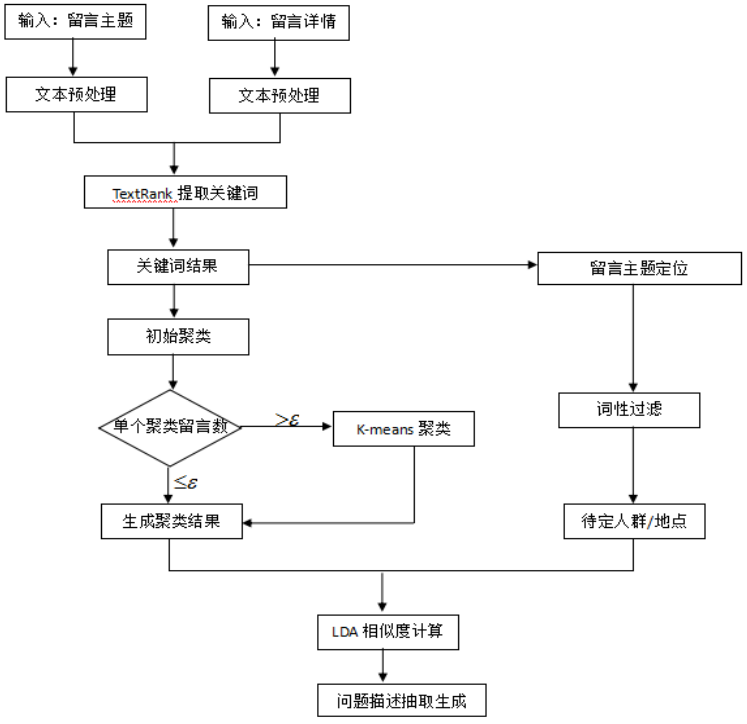

图 7　总技术路线图

2.2.1 文本预处理

本问题旨在提取热点问题，即某一时段内群众集中反映的某一问题，而文本中的数据包括留言主题与留言详情两部分，我们针对目标问题，使用中文分词工具 Jieba 分别对留言主题与留言详情进行预处理，包括去停用词、分词等。

2.2.2 关键词提取

本阶段是在基于预处理后的文本上，使用 TextRank 算法进行关键词提取同时进行词性标注，其中留言主题与留言详情部分分别进行提取。通过借鉴基于 ICTCLAS 的 Ansj 关键词提取技术，该技术主要是依据不同词性词语的初始权重，其中标题中关键词的权重加倍，再结合词在文中出现的位置和频率调整后，得到每个词的权重 score。而在本论文中，由于文本信息包括留言详情与留言主题两个部分，为了结合留言主题和留言详情中的文本信息，因此将留言详情中，留言主题关键词命中的关键词信息权重进行加倍，以此来突出留言主题信息重要性。按照这个想法计算出调整后的留言详情关键词权重，将其进行降序排序，选取排名靠前的关键词作为作为该留言的主题。

2.2.3 文本聚类

利用 k-means 算法对提取的关键词进行初步聚类，即将提取后关键词相似的留言归为一簇。聚类数目 k 需要提前确定，在本问题中我们将 k 的初始取值定为 $k=\left[\frac{n}{\alpha}\right]$，其中 n 为整个数据集的长度，α 是参数，k 值随着随着文本数目的大小而不断变化。为了避免出现每一簇中留言数目过多而导致的聚类冗余的情况，我们设定阈值为 ε，当聚类结果中某一类聚类数目小于该阈值时，将该结果直接作为聚类结果，若聚类数目大于该阈值，则将该类重新分割进行聚类。在此处 TF-IDF 输出的是 $[k,n]$ 数据对，其中 k 是数据集大小，n 为特征词数目。

2.2.4 词性过滤与主题定位

TextRank 算法提取的关键词一方面用来文本聚类，另一方面为接下来的特定人群/特定地点进行关键词定位，考虑到汉语结构，制定过滤定位的关键词之后内容的规则，并可根据该信息追溯到完整的留言信息。

我们的目标数据为热点问题中的特定人群/地点，而在词性上讲，这二者均属于名词，因此我们在利用 TextRank 算法进行关键词抽取的同时，对抽取的关键词进行词性标注，比如名词、动词、形容词等，然后进行过滤，仅保留名词词性的关键词，作为特定人群/地点的待定数据。

2.2.5 话题匹配

在经过 K-means 聚类后我们可以获得一个初步的聚类结果，同时在词性过滤后也可得到待定的人群/地点，通过计算二者之间的相似度来进行匹配。

在该论文我们借鉴 LDA 主题模型的思想计算聚类结果与待定地点/人群的相似度，将使用余弦距离进行计算，计算公式见公式（11）：

$$\cos(a,b)=\frac{\sum_{j=1}^{n} a_j \times b_j}{\sqrt{\sum_{j=1}^{n} a_j^{\,2}} \times \sqrt{\sum_{j=1}^{n} b_j^{\,2}}} \qquad (11)$$

其中，$\cos(a,b)$表示聚类结果与待定人群/地点的相似度，通过计算二者的余弦夹角来判断聚类结果与待定人群/地点的接近程度，余弦值越接近 1，夹角 $\cos(a,b)$就越接近 0°，二者之间就越接近，意味着待定人群/地点与聚类中的留言详情内容相似度越高。

2.2.6 热度评价

舆情热点指标能够帮助政府从海量留言中迅速发现热点民情民意并及时予以应对。考虑到热点问题所包括的时间、人物、内容这几个关键的要素，构造热度评价的思路如下：一方面，某一主题的留言的数目以及该主题所包含的具体留言的支持与反对情况这两个因素对于热度是正相关的，即在其他条件一定的情况下，某一主题的留言数目越多，热度越大，同理，留言数目的点赞或是反对数也

是如此；另一方面，由于仅仅考虑到留言数与点赞/反对数会忽略整体，因分别对其计算频率以衡量总体效果，且由于是二者协同作用，因此在这两个影响因素前面都乘以参数来控制发挥作用；最后，通过指数函数使得不同主题事件的热度值更加明显，易于判断热点事件。

基于以上想法，热度计算公式见公式（12）：

$$temp=\frac{\alpha\times\frac{c_i}{c}+\beta\times\frac{s_i+o_i}{s+o}}{\exp\left(\frac{t-t_0}{t_d}-1\right)} \tag{12}$$

其中，c_i 表示分类中的留言数，c 表示总留言数，s_i 代表分类中的点赞数，o_i 代表分类中的反对数，s 和 o 分别代表总的点赞数和反对数，t_d 为总时间范围 n 内的天数，$t-t_0$ 为某一类舆情的最初发生时间和最新发生时间的时间间隔，α 和 β 均为调整参数。

3. 答复意见评价

网络问政平台上政府部门对于群众留言问政的答复成为群众最能感知相关部门政务工作质量的重要窗口。相关部门答复内容同留言内容的相关性，答复与问政的时间间隔以及答复内容的完整性是影响答复内容质量的重要因素。本文将从答复的相关性、时效性、完整性三方面建立衡量答复质量的指标体系。

3.1 相关性

相关性，是指留言的问题与答复意见的关联程度。相关性强，则回复的质量高，反之，如果相关性低，那么意味着回复的质量低。我们拟采用 LDA 的思想来计算留言详情与答复意见的相似度，以此来衡量相关性的强弱。具体思想与本论文 2.3.2 部分相似，计算公式见公式（13）：

$$R=\cos(q,a)=\frac{\sum_{i=1}^{n}q_i\times a_i}{\sqrt{\sum_{i=1}^{n}q_i^{\,2}}\times\sqrt{\sum_{i=1}^{n}a_i^{\,2}}} \tag{13}$$

其中，$\cos(q,a)$ 表示留言问题与答复意见的相似度，q 为留言详情，a 为答

复意见，余弦值越接近 0，表示答复意见与留言内容越相似，即相关性越大，反之，若余弦值越接近 1，代表答复意见与留言详情内容越不符，即相关性越弱。

3.2 时效性

时效性，是指留言时间与答复时间的间隔大小。若留言时间与答复时间的间隔小，则意味着答复及时，时效性强，但如果留言时间与答复时间的间隔大，则代表答复拖沓，很有可能发生答复问题时该问题已经解决的极端情况，这便是时效性差。

我们用T代表时效性的值，提问时间记作T_q，回复时间记作T_a，则回复提问的时间间隔为$t=T_a$-T_q，则T=$e^{-\frac{t}{\bar{t}}}$，其中$\bar{t}$为平均回复提问的时间间隔。

3.3 完整性

完整性，是指回复的内容针对所提问题是否回答完全。若代表答复意见较为完整的回答了留言问题，则代表完整性强；反之，若留言问题中某些部分并没有在在留言内容中诶诠释，则意味着意味着完整性差。

直观地，我们用留言长度L来个数衡量答复的完整性，为了增强可读性，便于计算，我们将其进行分段处理，见公式（14）。

$$W=\begin{cases}-1, L<15\\ 0, 15\le L\le 40\\ 1, 40\le L\end{cases} \qquad (14)$$

留言长度$L<15$时，回复过短，极有可能出现容易是无效回复的情况，比如在附件 4 中存在答复意见为“已收悉”的情况，故令此种情况下的w取-1；

UU0081320	咨询打狂犬役苗报销比例是多少	2018/3/20 15:19:47	请问领导，农合费用增加了，	已收悉	2018/3/28 16:05:34
UU008151	在A6区准备全款购买二手房事项的咨询	2016/11/3 10:00:17	本人在A6区准备全款购买二手	已收悉	2016/11/22 12:25:56
UU0081119	请求处理K4县城锦豪雍景园小区商品房的费用	2013/6/3 13:05:49	尊敬的县长： 您好！我叫孙玲	已收悉	2013/7/5 16:47:46

留言长度$15\le L\le 40$时，回复的长度在该范围内比较普遍，因此w此时取 0；

留言长度$40\le L$时，回复长度长，一般情况下包含的内容全面，对留言问题的解释性更强，所以w此时取 1。

3.4 评价方案

相关性、时效性与完整性三者协同作用，衡量答复意见的质量，答复意见与留言问题的相关性越高，答复意见质量越好，答复越及时，同样对答复质量有正相关的影响，而有关完整性亦是如此。综合考虑以上因素，我们可列关系式如公式（15）所示。

$$score = (\alpha \times R + \beta \times W) \cdot T = \frac{\alpha \times \cos(q,a) + \beta \times W}{\exp\left(\frac{T_a - T_q}{T_a - T_q}\right)} \qquad (15)$$

其中，α，β 为参数，分别表示相关性与完整性的重要程度。

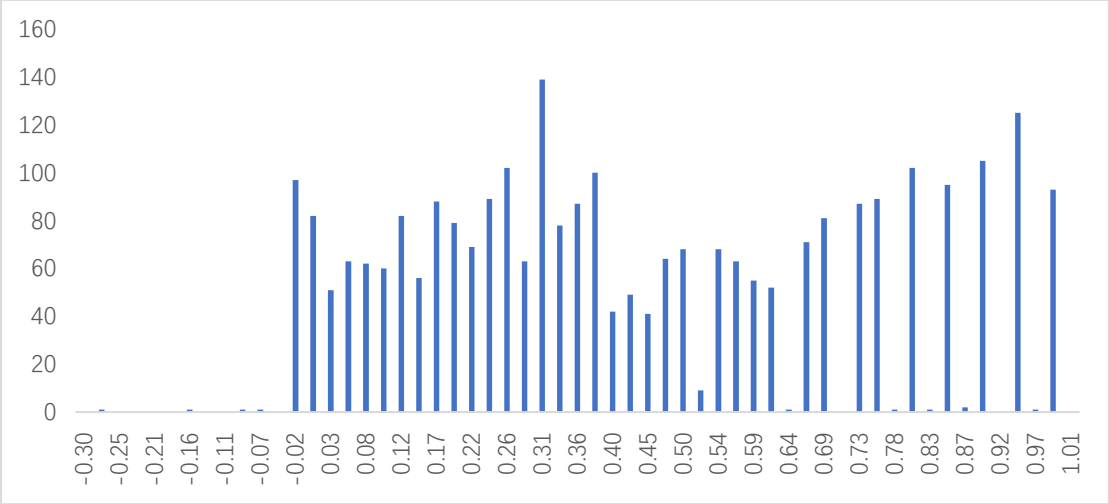

图 8 答复意见指标分布

实验得出答复意见指标分布如上图所示。大部分数据分布于[-0.02,1.01]。质量分数排名靠前的答复意见，比如对于编号 16542 的留言答复。群众网络问政时间为 2018/5/24 8:17:48，答复时间为 2018/5/24 16:12:33，且相关部门对于群众问询的请求道路硬化问题进行了详实完整的回复。而对于诸如留言编号 144850 的“你好！2019 年 6 月 13 日”的答复意见，指标给与了较低的分数。值得注意的是，诸如留言 88899 的答复意见虽然能够详实完整的对群众问询问题进行回复，但因回复时间间隔长达 7 个月，指标也给予了较低的分数。

通过对实验结果的具体分析，我们可以发现，score 值确实能在一定程度上反映答复意见质量。score 值高的词条，一般而言，回复及时、内容完整并且可

以很好的解释留言问题；反之，score 分数低的词条常常会存在一些回复拖沓、内容短小且无效的问题。

参考文献

[1] Dong Dong Xu,Shao Bo Wu. An Improved TFIDF Algorithm in Text Classification[J]. Applied Mechanics and Materials,2014,3512.

[2] 刘春磊,梁瑞斯,邸元浩.基于 TFIDF 和梯度提升决策树的短文本分类研究[J].科技风,2019(24):231.

[3] 牛永洁,田成龙.融合多因素的 TFIDF 关键词提取算法研究[J].计算机技术与发展,2019,29(07):80-83.

[4] 张强强,苏变萍,李敏.基于改进 CHI 的新的短文本混合特征选择方法[J].信息与电脑(理论版),2018(16):34-36.

[5] Jize Yin,Senlin Luo,Zhouting Wu,Limin Pan.Chinese Named Entity Recognition with Character-Level BLSTM and Soft Attention Model[J].Journal of Beijing Institute of Technology,2020,29(01):60-71.

[6] MIHALCEA R,TARAUP. Text Rank: bringing order into texts[Z].Emnlp,2004: 404-411.

[7] Hartigan J，Wong M. Algorithm as 136: a K-means clustering algorithm[J]. Journal of the Royal Statistical Society． Series C(Applied Statistics),1979,28(1) :100-108.

[8] Salton G, McGill M J. Introduction to modem information retrieval[M]. New York: McGraw-Hill Book Company, 1983.

[9] 吴柳,程恺,胡琪.基于文本挖掘的论坛热点问题时变分析[J].软件,2017,38(04):47-51.

[10] Blei D M, NgAY, Jordan M I. Latent Dirichlet Allocation [J]. Journal of machine learning research, 2003, 3:993-1022

[11] 王浩.基于多维度的公安舆情分析模型构建[J].情报探索,2020(03):24-29.

[12] 何跃,蔡博驰.基于因子分析法的微博热度评价模型[J].统计与决策,2016(18):52-54.

[13] 梁昌明,李冬强.基于新浪热门平台的微博热度评价指标体系实证研究[J].情报学报,2015,34(12):1278-1283.

[14] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, Xiaoyong Du, Analogical Reasoning on Chinese Morphological and Semantic Relations, ACL 2018.