基于 LSTM 与 DBSCAN 的智慧政务系统

摘要

进入 5G 时代，数据化、智能化的办公需求与日俱增。因此自然语言处理技术的改良与应用迫在眉睫。对政府部门而言，基于自然语言处理技术建立智慧政务系统，将成为提高政务工作效率的新趋势。本文构建了基于 LSTM 神经网络与 DBSCAN 算法的自然语言处理模型，主要针对文本信息的提取与文本内容的分类这类问题进行工作。

在数据预处理过程中，我们对文本内容进行数据清洗、去停用词，并通过 word2vec 将各词汇通过优化后的训练模型快速有效地将一个词语表达成向量形式，同时基于 Skip-Gram 模型对其进行特征提取，得到处理后的数据集。

在对留言文本进行一级分类时，为了能够更好得到留言主题与内容之间的准确信息及其相互性，并突显出语义信息，我们将预处理过后的词向量数据放入 LSTM 神经网络中，挖掘出词向量之间的彼此联系，得到最终的词向量表示，并通过 F-SCORE 法对我们构建的一级分类模型进行评估。

对于热点问题的挖掘，考虑到留言内容总量过于庞大，聚类分析时对热点信息的筛选效率会较大程度降低。因此，我们首先按照地点的不同将留言分成 9 类，将不含有地点信息的留言作为无效信息过滤。并基于 DBSCAN 算法进行聚类分析，以此提取出不同地点的留言热点内容，最后根据热度算法对留言热点内容排序，选出前五条留言热点内容。

对用户留言回复内容的评价工作，针对政务文案要求严谨规范的特点，我们基于答复的相关性、完整性、可解释性这三个维度对回复内容进行评价，根据三个维度制定其相匹配的标准进行比较与评价。

关键词：文本提取，文本分类，热点内容挖掘，word2vec，LSTM，DBSCAN

Abstract

In the 5g era, the demand for data and intelligent office is increasing day by day. Therefore, the improvement and application of natural language processing technology is imminent. For government departments, the establishment of intelligent government system based on natural language processing technology will become a new trend to improve the efficiency of government work. In this paper, a natural language processing model based on LSTM neural network and DBSCAN algorithm is constructed, which mainly focuses on text information extraction and text content classification.

In the process of data preprocessing, we clean the text content, remove the stop words, and express each word into vector form quickly and effectively through word2vec through the optimized training model. At the same time, we extract its features based on skip gram model to get the processed data set.

In the first level classification of message text, in order to better get the accurate information and mutual information between the message subject and content, and highlight the semantic information, we put the pre-processed word vector data into the LSTM neural network, mine the relationship between the word vectors, get the final word vector representation, and use the F-score method to build the first level classification The model is evaluated.

For the mining of hot issues, considering that the total amount of message content is too large, the efficiency of hot information screening will be greatly reduced in clustering analysis. Therefore, we first divide messages into 9 categories according to different locations, and filter messages without location information as invalid information. And based on the DBSCAN algorithm, cluster analysis is carried out to extract the hot content of messages in different places. Finally, the hot content of

messages is sorted according to the heat algorithm, and the first five hot content of messages are selected.

According to the characteristics of strict and standard requirements of government documents, we evaluate the content of replies based on the three dimensions of relevance, integrity and interpretability, and make matching standards for comparison and evaluation according to the three dimensions.

Keywords: text extraction,text classification,hotspot information mining, word2vec, LSTM, DBSCAN

目录

一、 简介

1.1 挖掘意义

对于文本信息的提取和分类，目前我们很大程度上仍停留在依靠人工进行划分类别与热点提取的阶段。在大数据时代，大量信息的汇集带给人工处理极大的挑战，而人工根据经验对信息进行提取与分类所存在的工作量大、效率低、差错率高等问题已然成为各种社会工作中的巨大隐患。尤其对于互联网上新出现的电子政务系统而言，仅仅依靠人工对收集到的愈来愈多的社会民意进行处理逐渐开始捉襟见肘，在不计其数的政务留言与民意调查中提取出民生关注的热点内容，对于文本分类工作者而言更是深感应接不暇。对此，我们希望能够开发出一套智能文本处理与分类系统，对于“智慧政务”系统的建立提供帮助。

针对文本内容进行智能阅读、信息提取与分类的工作，机械学习方式仍是目前应用的主流。作为当下最热门的自然语言处理技术之一，提高机械学习体系的语言检索、理解能力，并开发出更高效的处理、分类功能，在目前文本处理并不足够成熟的情况下依然有重要的积极意义。同时，这类技术的改良与创新也能够直接与互联网技术相结合，对文本处理技术进行多方面的运用，仍然具有很广阔的研究前景与很高的研究价值。

1.2 挖掘目标

我们希望构建一个自然语言处理与分类模型。模型将代替现有的人工工作，基于特定的标准帮助信息处理者更高效地完成处理工作。在具体应用层面，本模型将面对庞大的文本数据进行文本内容的信息提取，并基于已学习的信息内容，按照一定标准进行分类。

1.3 挖掘过程

挖掘过程主要包括文本数据预处理、文本分类及热点信息提取三大部分。其中，文本分类与热点信息提取为核心步骤，也是对于目前存在问题的有效解决方式。

文本数据预处理阶段包括数据清洗、分词、去停用词，并将词组向量化。为进一步明确文本信息的关联性，将词向量放入 LSTM 神经系统进行训练并对文本进行一级标签分类；同时，根据文本信息特点（包含地点信息），按照特点制定标准分类，对每一类进行聚类分析，提取出每一类数据的热点留言内容。

二、预处理

2.1 数据清洗

面对数量庞大、可能存在较多重复与错误信息的留言文本库，从中直接提取信息进行分类难度较大。因此，我们的第一步工作是对留言文本库进行数据清洗，包含检测去除异常值、填补空缺值及删除重复项。针对留言数据存在的主要问题在于重复度较高，且经检测，文本数据不存在异常值及空缺值，我们的数据清洗工作主要侧重于去重。

考虑到我们将在后续进程中针对留言文本建立词向量，且词向量训练内容主要基于留言内容，因此我们在对附件二进行文本去重过程中，选择删除文本中的留言内容相同项以提高训练效率；同时聚类分析提取热点内容前，我们会将文本按照留言主题中的地点信息进行分类，并对留言主题和内容共同进行聚类分析，因此，我们删去同一用户留言主题不同而留言内容相同、同一用户留言主题相同而留言内容不同及留言主题中不含地点信息这三类数据。处理结果如下：

附件二原有数据 9210 条，再对相同的留言内容进行去重后，保留数据 9052 条。

附件三原有数据 4326 条，对三类留言进行去重后，保留数据 4231 条。

2.2 分词、构建用户词典、去停用词

在汉语言文本处理中，词是最小的能够独立活动的有意义的语言成分，也是一个合适的语义粒度。由于中文独特的语言体系，导致中文的语义表达中缺少非

常明显的词与词之间的分界，同时逻辑性上远不如其他语言体系精准，所以在文本提取信息的过程中需要进行分词。本文的分词方法采用了 Python 开发的 Jieba 分词，对于数据库中的群众留言进行中文分词。

Jieba 分词的功能基于其自带的 dict.txt 的词典，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG），其中每个词语作为一条有向边。对待分词句子生成 DAG，根据给定的词典进行查词典操作，生成几种可能的句子切分，同时基于统计的方法对各有向边进行权重赋值，并计算得到该句子可形成的最短路径，即为最终的分词结果。

分词完成后，对于留言文本的基本处理完成，我们根据留言文本的分词结果，构建用户词典，以方便后续分类、热点提取及回复评价工作的进行。

表 1：部分用户词典示例

经济学院	劳动东路	在水一方大厦	农村信用合作联社
体育学院	江山帝景	杜鹃文苑小区	锦楚国际星城小区
购房补贴	参保记录	锦楚国际新城	嘉和城润芳园小区
居民小区	明发国际	魅力之城小区	石期市镇老农贸市场
污染空气	冷江东路	西湖建筑集团	……

通常意义上，停用词可概括为自然语言中包含的功能词及使用频率特高的单汉字等，这些功能词极其普遍，但与其他种类词汇相比并不存在实际含义。但这些词的存在无端提高了索引量，却难以帮助缩小搜索范围，同时还会降低搜索的效率。因此在文本处理过程中遇到停用词，则立即停止处理，将其从文本中抽离以增加检索效率，提高检索的效果。

2.3 word2vec 进行文本特征提取

文本进行上述两步处理后，将以词为基本单位存在。下一步我们要将待处理文本输入到神经网络中进行训练。

为了使汉语文本能够转化为计算机语言，从而使其有效进入神经网络中进行训练，我们将每个独立词语单独表示为一个高维向量。但由于汉语词汇量大，表义复杂等特点，在文本处理过程中很容易出现维数灾难、词语相似性、模型泛化

能力以及模型性能等问题。对此，我们需要对当前的原始特征进行降维处理。

Word2vec 可以根据给定的语料库，通过优化后的训练模型快速有效地将一个词语表达成向量形式，是目前广泛适用且能快速有效训练词向量的模型。

Word2Vec 模型中，主要有 Skip-Gram 和 CBOW 两种模型，两种模型的区别在于：Skip-Gram 是给定 input word 来预测上下文。而 CBOW 是给定上下文，来预测 input word。由于要尽可能完整地对文本特征进行提取，同时基于部分词向量预测文本内容，本文采用了 Skip-Gram 模型进行处理

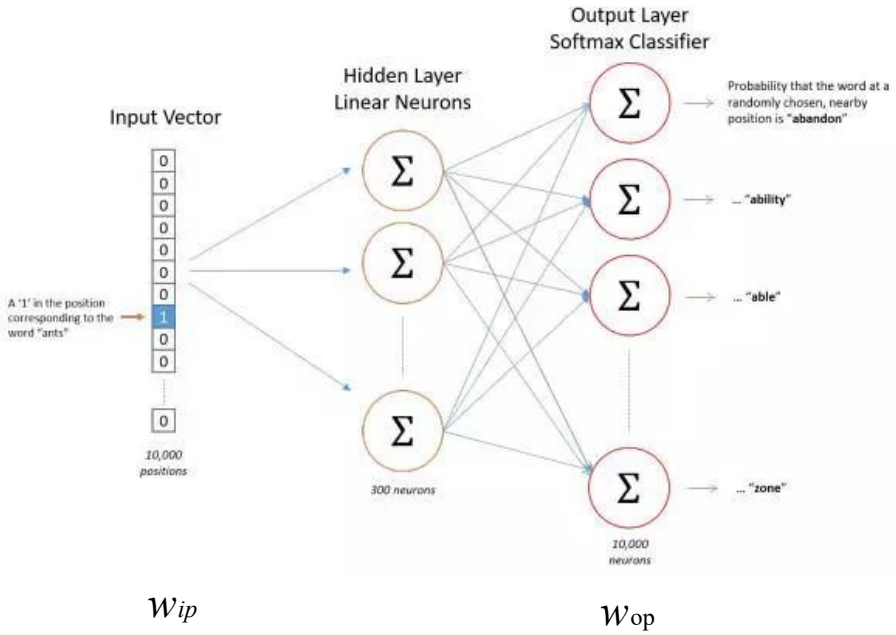

图 1：Skip-Gram 模型示意图

通过 Skip-Gram 模型对文本特征进行提取，为了信息量充足我们对最终输出端的要求下式取得最大值：

$$\frac{1}{T}\sum_{t=1}^{T}\log p(w_{op} \mid w_{ip})$$

其中 T 为训练文本量。

以下数据为我们训练出的高频词向量示例：

表 2：高频词向量示例

原词汇	词向量表示
买	-0.014791357 0.051633313 0.028725415 0.052982926 0.067859836
请	-0.08151165 0.0011495299 -0.008897899 -0.08446391 -0.08991302
建议	-0.093309715 -0.011041587 -0.055616703 -0.024900805 0.061674207
相关政策	-0.003258392 0.044489626 0.0341307 -0.04447144 -0.084455825
有关人员	0.07405699 -0.0035452666 -0.054438308 -0.039418727 0.0440824

三、文本内容分类与提取

3.1 文本一级分类处理

在对文本进行预处理后，我们需要对留言的内容进行一级标签的分类，便于后续细化分类工作的开展。

3.1.1 LSTM 神经网络

长短期记忆网络（LSTM，Long Short-Term Memory）是一种时间循环神经网络，是为了解决一般的 RNN（循环神经网络）存在的长期依赖问题而专门设计出来的，所有的 RNN 都具有一种重复神经网络模块的链式形式，传统的 RNN 神经网络的神经元是将输入运用函数进行计算后进行输出的单元，而 LSTM 将神经元变为记忆单元，每个记忆单元由输入门、遗忘门和输出门构成，其单元结构图如图 1 所示。其中长期状态 c 用于存储长期记忆信使得序列的长期状态可以保存下来，并传递到下一层，同时，遗忘门的设计又使得 c 得到更新，丢弃已经过时的信息。LSTM 的这一设计解决了 RNN 网络存的梯度消失和梯度爆炸问题。因此本文采用 LSTM 神经网络对文本进行一级分类。

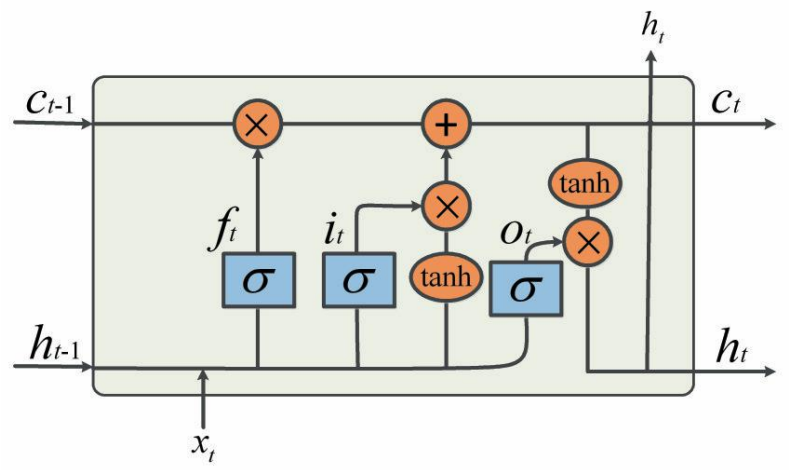

图 2：LSTM 神经元结构

在某一时刻 t 的数据 x_t，与上一时刻该神经元输出的数据 h_{t-1} 一起作为输入数据进入神经元结构内，对 c_{t-1} 进行更新，以此得到长期状态的 c_t。

3.1.2 文本内容一级分类

我们假定，经过文本数据预处理后，某一文本 A 的对应的文本向量空间：

$$A = \{(x_i, y_i)\}_{i=1}^{n}$$

其中 n 表示词的数量，x_i 表示第 i 个词，y_i 表示对应的特征权值。

则经过特征向量提取后，其对应的特征向量为：$A' = \{(x_i, y_i)\}_{i=1}^{w}$

其中 $w \le n$。

该分类的运算方式的运行步骤简述如下：定义输入的文本为 M，其中某一文本经过了预处理及特征向量提取的过程后，可得到特征向量 $A' = \{(x_i, y_i)\}_{i=1}^{w}$ 作为 LSTM 神经网络结构的第一个节点，输出端输出内容为此分类模型对于所有的文本集合 M 做出的分类预测类别集合 C_M。

3.1.3 LSTM 参数设置

模型的第一层为嵌入层(Embedding)，我们使用长度为 100 的向量来表示每

一个词语，并将全部词向量导入 LSTM 神经元层。

LSTM 层包含 100 个记忆单元，输出层为包含 7 个分类的全连接层

训练数据时设置的参数如下：

（1）每次训练包含 5 个训练周期；

（2）指定进行梯度下降时每个 batch 包含的样本数为 64。

3.1.4 文本一级标签分类结果

基于附件 2 给出的七类分类结果，我们经过统计得到如下表所示分类结果：

	城乡建设	174	4	4	0	0	0	0	
	环境保护	20	73	0	1	0	0	2	160
实际结果	交通运输	5	1	43	4	1	4	0	120
	教育文体	5	0	1	128	8	2	1	80
	劳动和社会保障	8	0	0	8	182	1	10	40
	商贸旅游	15	1	1	8	2	84	2	0
	计生卫生	4	3	0	1	7	8	66	
		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	计生卫生	
		预测结果							

图 3：一级分类结果

同时，我们用 F-SCORE 方法对我们的分类方法进行评价，评价标准公式如下：

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

其中， P_i 为第 i 类的查准率， R_i 为第 i 类的查全率

评价结果输出为下表：

表 3：分类评价结果

	Precision	Recall	F1-score	Support
城乡建设	0.75	0.88	0.81	197
环境保护	0.90	0.76	0.82	96
交通运输	0.88	0.74	0.80	58
教育文体	0.84	0.88	0.86	145
劳动和社会保障	0.88	0.87	0.88	209
商贸旅游	0.80	0.74	0.77	113
卫生计生	0.81	0.75	0.78	88
Accuracy	0.83			906
Loss	0.591			

3.2 基于聚类分析的文本热点提取

在对留言内容进行了标签化的分类后，我们可以得到不同主题的留言内容。下一步，我们要对内容进行更细化的提取和分类，提取出留言内容中存在的热点信息。

3.2.1 记录文本词性

在对文本进行预处理过程中，我们已经用 jieba 词库对文本进行分词，但对于词性缺乏认识。因此，我们需要在预处理的基础上，对文本的词性进行记录，从词性角度提取文本的热点内容。

对于热点问题的发掘，我们主要基于留言主题对文本进行聚类分析，同时，绝大部分数据中都包含地点信息，因此，不同于附件二的去重标准，在数据去重过程中，我们删除附件三中同一用户，留言内容相同而主题不同，以及留言内容不同而主题相同的数据，并将没有地点信息的留言判定为无效信息进行过滤。

我们的数据集在处理之后得到了 9052 条短文本，分词得到 14516 个不重复的词汇及其对应的词性，并建立了两者之间的字典联系。

3.2.2 对文本的向量化处理

在记录词性后，我们需要对文本的向量空间进行再分配，基于词性角度对文本进行文本向量化。这里我们选用 TF-IDF 对文本进行处理并赋值，得到文本的 TF-IDF 权重。

TF-IDF 是一种统计方法，用以评估一字词对于某一文本集或某一语料库中的一份文本的重要程度。TF-IDF 实际上是：TF * IDF 的组合，TF 词频(Term Frequency)，IDF 逆向文件频率(Inverse Document Frequency)，其基本原理为：如果某个词或短语在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

TF-IDF 方法可阐释为：TF 表示当前文本 A 中出现的频率。IDF 指：如果包含词条 x 的文档越少，也就是 n 越小，IDF 越大，则说明词条 x 具有很好的类别区分能力。

具体原理如下：

$$\mathrm{tf}_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}\text{（式1）}$$

$$\mathrm{idf}_i=\lg\frac{|D|}{1+|\{j:t_i\in d_j\}|}\text{（式2）}$$

上述两式相乘，得到目标值：$\mathrm{tfidf}_{i,j}=tf_{i,j}\times idf_i$

其中：

式（1）分子表示该词在文本中出现次数，分母表示在文件中所有字词的出现次数之和。

式（2）中分子表示在文件中所有字词的出现次数之和，为避免某一词语不在语料库中导致分母为零的情况发生，分母数值+1

3.2.3 针对词性的词汇权重再分配

在对文本进行预处理、词性分类及文本向量化以后，我们需要根据词性，对词汇的权重赋值进行重新分配，给与其 TF-IDF 权重以不同的乘数，以便突出特

定类型的词汇在反应热点问题时各自的重要性。这一工作将在一定程度上改良聚类分析的效果。

在经过多次调试之后，我们得出以下结论：为了达到最佳的聚类效果，名词类别的 TF-IDF 权重需要乘以 1.2，动词类别需要乘以 1.1，而数词类别则乘以 0.

3.2.4 运用 DBSCAN 算法的聚类分析

基于词汇在词性角度的权重再分配基础上，我们要对当前文本进行聚类分析，目的是提取出最精炼有效的信息，分离文本当中的噪音。本文采用 DBSCAN 算法进行聚类分析。

DBSCAN(Density-Based Spatial Clustering of Applications with Noise)是一个基于密度的聚类算法，其特点在于将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在噪声的空间数据库中发现任意形状的聚类。

该算法的运算流程如下：

确定扫描半径 eps=1 与每个聚类类别中的最小样本数 min_samples=6。任选一个未被访问的点作为起始点，搜索 eps=1 范围内的所有附近点。若附近点的数量≥min_samples，则当前选定的起始点与其附近点形成一个簇，并将该点标记为已访问点。在递归过程中以相同的方法处理所有未被访问的点，将簇进行拓展。若附近点的数量<min_samples，则将该点暂时标记为噪声点。在簇被充分拓展后，可以簇内所有的点都被标记为已访问，其余被标记的噪声点可进行分离。

3.2.5 聚类方法及结果

由于留言主题与内容中地点信息特征明显，我们首先根据文本中的地点信息，对文本进行分类，随后对各个分类内容进行 DBSCAN 聚类，找到热点问题，并将热点问题 top5 作为输出结果示例，如下表：

表 4：热点问题 top5

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	3	100	2019/07/07 至 2019/09/01	A 市伊景园滨河苑	小区车位捆绑销售
2	1	67	2019/11/02 至 2020/01/26	A 市 A2 区丽发新城	小区旁搅拌站有灰尘、噪音污染
3	5	15	2019/07/21 至 2019/12/04	A 市 A5 区劳动东路魅力之城小区	小区一楼商铺有油烟、噪音扰民
4	2	11	2019/06/26 至 2019/10/16	A 市 A7 县经开区泉星公园	公园建设进度缓慢、项目规划不合理
5	4	4	2019/04/26 至 2019/12/15	A 市 A3 区中海国际社区	小区夜间施工噪音扰民、施工进度缓慢

其中，热度指数计算标准为：

热度指数=（相关问题总条数/相关问题持续讨论时间）*100

四、对回复内容的评价

4.1 评价意义

在网络治政理政的大背景下，政务工作人员对于网上留言的及时答复也是网络理政中非常重要的一环。而政务回复文案则被视作政府对问政网民的官方回答。对于政务文案规范性制定评价标准与评价体系，对于网络理政的效率和效果，规范理政流程都将有非常明显的促进作用。

4.2 评价标准制定

针对政务文案要求严谨规范的特点，我们主要从相关性、完整性与可解释性三个维度对答复内容进行评价。

4.2.1 相关性评价标准

对于留言与回复，我们认为两者应当围绕统一主题展开，由此，相对应的留言内容与回复内容将会有语料上的重复。因此，我们首先利用留言内容建立语料库词典，之后用回复内容建立索引与语料库内容进行对比，依此评价回复内容与留言内容间的相关性。具体流程如下：

首先，我们对留言内容进行处理：再读取留言内容后，对要计算的文本进行分词，并将文本整理为完整性评价标准中制定的标准格式。经过拆分再整理后，我们计算各词语频率，对词频较低的词语进行过滤，并通过保留词语建立语料库词典。

然后，我们对回复内容进行读取，将要回复的内容通过 doc2bow 转化为词袋模型，并利用词袋模型建立新的语料库。将新语料库通过 tfidfmodel 进行处理，得到 tfidf，通过 token2id 得到特征数并依此创建稀疏矩阵建立索引。最终相似度结果由稀疏矩阵的相似度体现。我们能规定如下评价标准：

我们定义相似度<30%为不相关，相关度评价为 0；

相似度≥30%为相关，相关性评价为 1；

同时，相关性评价结果直接影响完整性与可解释性评价，若相关性得分为 0，则后续两项不进行评价，记为无效回复；若相关性得分为 1，则继续对后续两项进行评价。

相关性评价结果为：不相关 20 条；相关 2797 条。

4.2.2 完整性评价标准

对于回复内容的完整性，从政务文案的特点这一角度出发进行考虑，我们认为应着重关注其文本的格式化与规范性。因此，我们将回复文本的格式与所制定的标准格式之间的契合度，作为评价回复内容完整性的重要指标。

参考书信格式标准，我们对于回复内容的标准格式制定如下：

标准格式由三部分组成：（1）称谓：您好；（2）正文内容；（3）回复日期。

对于回复内容的检测，我们主要针对成分（1）与成分（3）进行检测，对完整性的评价判断标准如下表所示：

表 5：完整性评价标准表

序号	判断标准	完整性评价
1	格式规范	5
2	缺少成分（1）	3
3	缺少成分（3）	3
4	同时缺少成分（1）与成分（3）	1

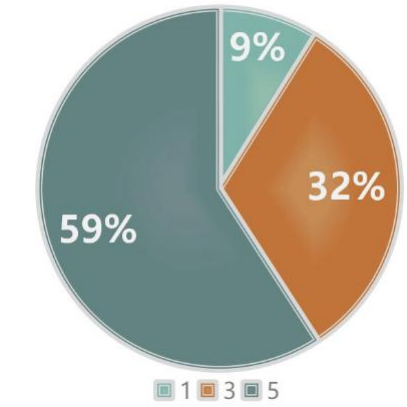

图 4：完整性评价结果

4.2.3 可解释性评价标准

对于回复内容可解释性，我们将其定义为：针对于留言内容问题,是否提出了切实有效的解决方案。

对此，我们基于经过预处理的留言内容与回复内容，使留言内容与回复内容相结合，通过人工对部分回复的可解释性进行示例性打分，随后将示例性打分数据放入 LSTM 模型中进行训练，以预测其他回复内容的打分结果。

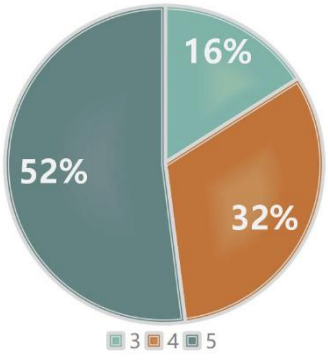

图 5：可解释性评价结果

五、模型评价及优化

5.1 模型评价

5.1.1 模型优点

在分类问题上，LSTM 解决了梯度反传过程由于逐步缩减而产生的 Vanishing Gradient 问题，其在长远的更为复杂的任务上的潜力巨大。它更真实地表征或模拟了人类行为、逻辑发展和神经组织的认知过程。

在热点内容提取方面，DBSCAN 算法聚类速度快且能够有效处理噪声点和发现任意形状的空间聚类；不需要输入要划分的聚类个数；聚类簇的形状不产生偏倚且可以在需要时输入过滤噪声的参数

5.1.1 模型不足

在分类问题上，至今看来以 CNN 为代表的前馈网络依然有着性能的优势。较之于 RNN，其可以处理 100 个量级的序列，而对于 1000 个量级，或者更长的序列则依然会显得很棘手。

对于 DBSCAN 算法的问题，其不能很好处理高维数据，也不能很好反映数据集以变化的密度。同时，如果样本集的密度不均匀、聚类间距差相差很大时，聚类质量较差。

5.2 模型优化

5.2.1 对热点信息提取的细化分类

对于热点信息的提取，由于要降低聚类分析的单次计算量以提高聚类效果，我们基于文本中的地点信息，对文本数据进行分类处理。

同时我们考虑到，不同类型热点信息的提取，对于不同群体的用户影响程度并不相同。因此，我们可以在分类前开展民意普调工作，通过普调结果，建立基于用户群体的分类指标，更精准提取热点信息。

六、参考文献

[1]张超然,裘杭萍,孙毅,王中伟.基于预训练模型的机器阅读理解研究综述[J/OL].计算机工程与应用:1-12[2020-05-07].

[2] 郭艳婕,杨明,侯宇超,孟铭.基于相似性度量的改进 DBSCAN 算法[J].数学的实践与认识,2020,50(06):164-170.

[3] 陈平平,耿笑冉,邹敏,谭定英.基于机器学习的文本情感倾向性分析[J].计算机与现代化,2020(03):77-81+92.

[4]张云翔,饶竹一.基于 LSTM 神经网络的电网文本分类方法[J].现代计算机,2020(02):8-11.