基于自然语言处理的智慧政务应用

摘要

本文旨在针对互联网收集的群众问政留言记录以及相关部门对部分群众留言的答复意见进行研究分析，通过自然语言处理和文本挖掘，对群众留言进行分类，挖掘热点问题，并从答复意见的相关性、时效性、完整性和可解释性进行综合评价。

针对问题一，对文本数据进行预处理操作，包括数据清洗、中文分词、去除停用词等等；其次，为后续求得 F-Score 值尽可能达到预期，需要对文本数据进行文本摘要提取、数据增强、特征选择等操作；随后利用基于机器学习文本分类和基于深度学习文本分类实现对每个标签进行分类，通过使用 F-Score 值对分类方法进行评价。

针对问题二，本文基于命名实体识别的方法实现对文本留言主题特定地点和特定人群进行提取，结合余弦相似度和层次聚类实现对特定地点的归类；选取固定时间段对数据集进行划分，建立热度评价指标，计算热度值；建立基于 PSO-BP 神经网络模型实现热点话题的预测，计算热度指数并生成表格。

针对问题三，选取影响答复意见质量的指标，基于自然语言处理和文本挖掘技术对每个指标进行量化处理；利用绩效考核指标分析答复意见的质量进行评价。

关键词：文本分类；命名实体识别；层次聚类；PSO-BP 神经网络预测模型；绩效考核指标

Abstract

The purpose of this paper is to study and analyze the records of the public political message collected by the Internet and the reply opinions of some of the public message by relevant departments. Through natural language processing and text mining, the public message is classified and hot issues are mined, and comprehensive evaluation is made from the relevance, timeliness, integrity and interpretability of the reply opinions.

Aiming to task one, preprocessing the text data, including data cleaning, Chinese word segmentation, removal of stop words, etc. secondly, in order to obtain the F-score value as much as possible to achieve the expectation, text summary extraction, data enhancement, feature selection and other operations need to be carried out for the text data. Then using machine-based learning text classification and depth based learning text classification to achieve each Tags were classified, and the classification method was evaluated by using F-score value.

Aiming at task two, this paper is based on the named entity recognition method to achieve the extraction of specific places and specific groups of text message subject, combined with cosine similarity and hierarchical clustering to achieve the classification of specific places. Selecting a fixed period of time to divide the data set, establishing the heat evaluation index and calculating the heat value is carried out in this paper. And the PSO-BP neural network model is used to realize the prediction of hot topic, calculate heat index and generate tables.

Aiming at task three, the Indicators affecting the quality of reply need to be selected. Based on natural language processing and text mining technology, each index is quantified, and the quality of reply is evaluated by performance evaluation index.

Key words： Text classification; Named entity recognition; Hierarchical clustering; PSO-BP neural network prediction model; Performance evaluation indicators

目录

1 背景与挖掘目标

1.1 问题背景

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，随着大数据的快速发展，新型服务模型智慧政务正成为趋势，智慧政务以大数据分析为核心，不仅是因为相关数据量大、类型多，更重要的是，对海量数据的深度挖掘及多维剖析可以创造出更大的公共价值，有利于更准确地掌握政务动态变化，发现公众新需求，有效提升政务服务能力。尤其是近些年来，随着各类社情民意相关的文本数据量不断攀升，如何利用人工智能等技术实现群众留言的分类以及挖掘热点问题，成为智慧政务建设中的重要课题之一。群众问政留言的智能分类处理对智慧政务有重要意义。

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。

1.2 解决问题

1.2.1 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。

通常使用 F-Score 对分类方法进行评价：

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

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

1.2.2 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，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。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”。

1.2.3 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。

1.3 数据说明

附件 1 提供了一种内容分类三级标签体系，样例如下：

表 1 内容分类三级标签体系

一级分类	二级分类	三级分类
城乡建设	安全生产	事故处理
城乡建设	安全生产	安全生产管理
城乡建设	安全生产	安全隐患
...	...	...

附件 2、附件 3、附件 4 的数据来源于互联网公开渠道，具体表结构如下：

表 2 附件 2 表结构

留言编号	留言用户	留言主题	留言时间	留言详情	一级分类
744	A089211	建议增加 A 小区快递柜	2019/10/18 14:44	我们是 A 小区居民...	交通运输

表 3 附件 3 表结构

留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
744	A089211	建议增加 A 小区快递柜	2019/10/18 14:44	我们是 A 小区居民...	100	2

表 4 附件 4 表结构

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间
744	A089211	建议增加 A 小区快递柜	2019/10/18 14:44	我们是某小区居民...	网民‘A089211’你好...	2019/10/19 8:40

2 问题分析

自然语言处理(NLP)关注的是人类的自然语言与计算机设备之间的相互关系。NLP 是计算机语言学的重要方面之一，它同样也属于计算机科学和人工智能领域。文本挖掘和 NLP 的存在领域类似，它关注的是识别文本数据中有趣并且重要的模式。NLP 和文本挖掘为建立智慧政务系统起到关键作用。

问题给出四个附件数据，附件 1 包括留言的三级分类标签，以及城乡建设等 15 个一级标签，附件 2 给出自互联网公开来源的群众问政留言记录，包括留言编号、留言用户、留言主题、留言时间、留言详情和一级分类共 6 个指标，附件 3 同样给出部分问政留言记录，多出了点赞数和反对数两个指标，附件 4 是部分留言记录的详情以及相关部门对留言的答复意见和答复时间。

问题一要求根据一定的划分体系对群众留言进行分类，建立关于留言内容的一级标签分类模型。根据所给数据的特点，实现文本分类常用步骤是：数据预处理、特征提取、代入机器学习文本分类模型算出 F-Score。过程发现数据存在类别数据量不均衡、特征不明显等现象对文本分类造成巨大的干扰，需要在分类前进一步处理。

问题二要求根据附件 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，给出评价结果，并且挖掘出排名前 5 的热点问题。根据所给数据的特点以及提出的问题，需要提取特点地点或人群，设置正则表达式提取指定地点存在缺陷，理想的做法是用命名实体识别算法。聚类过程采用余弦相似度度量，同时采取 K-Means 聚类、DBSCAN 聚类、层次聚类，经过大量实验证明层次聚类结合聚类性能评估可达到理想类别。定义合理的热度评价指标，预测结果提取排名前 5 的热点问题。

问题三要求从答复的相关性等角度对答复意见的质量综合评价。根据所给数据的特点，定义合理的答复意见质量的指标，采用自然语言处理量化处理，并给出评价方案。

3 数据准备

附件 1~附件 4 总体上无明显异常。按列分析，从群众发表的留言详情内容中有[a]href=[$YingHao$]https://baidu.com/或出现 https://baidu.com/异常字样。按行分析，主要检查群众留言信息是否有重复。

3.1 数据预处理

3.1.1 数据清洗

文本数据中字母、数字、符号无明显作用，故对数据中存在的字母、数字以及符号进行剔除。正则表达式通常用作检索、替换和控制文本。

3.1.2 中文分词

文本处理一定程度上取决于中文分词，现有的分词工具有 Jieba 分词，SnowNLP 分词，PKUSeg 分词等等，如表 5 是部分分词工具的比较。本文选用的分词工具是 Jieba 分词。

表 5 中文分词

分词内容 / 分词工具	A 市经济学院体育学院变相强制实习
Jieba	a 市 经济 学院 体育 学院 变相 强制 实习
SnowNLP	a 市 经济学院 体育 学院 变相 强制 实习
PKUSeg	a 市 经济 学院 体育 学院 变相 强制 实习
THULAC	a 市 经济 学院 体育 学院 变相 强制 实习
HanLP	a 市 经济学院 体育学院 变相 强制 实习
FoolNLTK	a 市 经济 学院 体育 学院 变相 强制 实习
LTP	a 市 经济 学院 体育 学院 变相 强制 实习
CoreNLP	a 市 经济 学院 体育 学院 变相 强制 实习
BaiduLac	a 市经济学院体育学院 变相 强制 实习

3.1.3 去除停用词

文本数据中存在一些停用词对文本研究具有巨大的干扰，需要对停用词进行去除。本文选用通用的停用词表。

3.1.4 去除无用词

单个字符和空格无明显作用，应当去除。

4 数据可视化与文本分析

4.1 数据可视化

数据可视化是关于数据视觉表现形式的科学技术研究。文本数据可视化可以很好理解主旨、组织与分类信息、对比文档信息等等。

4.1.1 类别数据量统计

类别数据量可视化可以直观反映数据是否存在不平衡现象，为后续运用数据增强起到重要作用。从整体上反映和分析类别数量特征，观察样本的数据量，作出正确的判断。

	一级标签	数量
0	城乡建设	2009
1	劳动和社会保障	1969
2	教育文体	1589
3	商贸旅游	1215
4	环境保护	938
5	卫生计生	877
6	交通运输	613

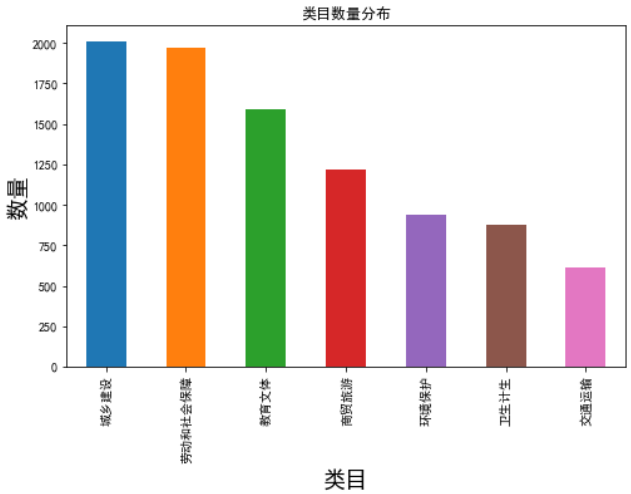

图 1 类别数据量统计

从图 1 可看出，类别数据中存在明显的不平衡现象。增加类别数据量其中

一种手段是对互联网数据进行爬取，本文爬取数据来源是“问政湖南-红网”的对应一级标签类别的数据，采用八爪鱼采集器实现对数据内容的爬取。以下是爬取到的数据量，“城乡建设”数据有 510 条，“环境保护”数据有 475 条，“交通运输”数据有 1315 条，“教育文体”数据有 1475 条，“劳动和社会保障”数据有 1115 条，“商贸旅游”数据有 165 条，“卫生计生”数据有 975 条，并与原数据进行合并。

4.1.2 词云图

词云图是词频数据可视化的表现，呈现字体比例越大，说明词语出现的频次越高。同时可以通过观测数据形式判断潜在的干扰词语。效果如下图：

图 2 词云图

从图 2 可看出，干扰词语有“尊敬”“领导”“西地省”“市”“县”等等，可在数据预处理时进行剔除。

4.2 文本分析

文本分析是指对文本的表示及其特征项的选取；文本分析是文本挖掘、信息检索的一个基本问题，它把从文本中抽取出的特征词进行量化来表示文本信息。

4.2.1 关键词提取

关键词提取就是从文本里面把跟内容意义最相关的一些词语抽取出来。

1） 基于 TF-IDF 算法的关键词提取

TF-IDF 是一个在自然语言处理和信息检索领域常用的一个指标，用来数值化，连续化一个词在一个文档集中的统计特征，为避免分母出现零的情况往往需要做平滑处理。效果如下图 3：

城乡建设：业主 小区 开发商 领导 政府 年月日 公积金 房屋 部门 居民 尊敬 老百姓 建设 住房 相关 规划 房子 交房 解决 办理 您好 请问 房产证 希望 施工 厅长 公司 工程 住户 项目

环境保护：污染 环保局 村民 居民 环保 领导 排放 噪音 生活 投诉 环境 部门 污水 生产 周边 厅长 小区 环评 尊敬 政府 老百姓 影响 环保部门 年月日 您好 希望 垃圾 解决 相关 养猪场

交通运输：高速 厅长 领导 出租车 公路 快递 长沙 尊敬 车辆 高速公路 希望 司机 出行 部门 的士 请问 收费 通车 交通 湖南 您好 年月日 老百姓 刘厅长 建设 道路 湖南省 驾校 政府 高铁

教育文体：学校 学生 教师 教育局 老师 教育 补课 家长 领导 小学 孩子 厅长 中学 尊敬 幼儿园 小孩 招生 考试 工作 您好 希望 教育厅 请问 培训 年月日 收费 工资 学费 班主任 西地省

劳动和社会保障：退休 社保 厅长 工作 工资 职工 人员 单位 领导 西地省 年月日 公务员 医保 尊敬 员工 请问 办理 养老保险 事业单位 养老金 政策 您好 公司 考试 劳动 文件 缴纳 享受 工伤 报销

商贸旅游：电梯 传销 业主 西地省 旅游 部门 小区 景区 领导 年月日 希望 收费 相关 老百姓 公司 游客 投诉 尊敬 物业 您好 开发商 人员 维修 垄断 政府 请问 收取 价格 猪肉 故障

卫生计生：医院 医生 生育 领导 厅长 年月日 乡村 卫生院 政策 执业 尊敬 小孩 医师 请问 您好 西地省 卫生局 患者 卫生室 医疗 手术 独生子女 计生办 国家 病人 计生 再婚 湖南省 希望 办理

领导 厅长 学校 年月日 尊敬 您好 请问 西地省 教师 部门 工作 学生 希望 小区 政府 医院 业主 退休 教育局 人员 相关 工资 解决 老师 老百姓 单位 政策 办理 湖南省 居民 国家 公司 村民 教育 社保 考试 情况 职工 孩子 小孩 湖南 收费 生活 家长 文件 投诉 补课 建设 小学 谢谢 开发商 电梯 医生 一名 污染 回复 医保 时间 工作人员 农村 你好 企业 长沙 公务员 乡镇 管理 高速 费用 请求 恳请 答复 影响 中学 环保局 享受 局长 生育 员工 幼儿园 事业单位 房屋 参加 书记 养老保险 补贴 公积金 缴纳 重视 规划 待遇 发放 通知 环境 群众 百姓 乡村 社会 标准 报销 社区

图 3 基于 TF-IDF 算法的关键词提取

2） 基于 TextRank 算法的关键词提取

TextRank 基于 PageRank 的思想提出，PageRank 认为如果一个网页被链接的次数越多那么这个网页越重要，如果这个网页被一个很重要的网页链接了当然这个网页权值更大，最后经过一系列的证明可以得出多次迭代过后的 PR 值是一定收敛的，也就是说我们对一个网页有且只有一个 PR 值，唯一性由 PR 值计算过程的 Markov 性保证；TextRank 则是取文档中的一个窗口，如果共现则视为有效。本文基于 TextRank 算法提取关键词针对名词与动词。效果如下图 4：

领导 学校 工作 部门 政府 公司 希望 学生 相关 国家 情况 尊敬 政策 人员 医院 教师 解决 请问 小区 单位 生活 教育 老师 厅长 居民 老百姓 业主 建设 孩子 影响 工资 时间 办理 企业 退休 管理 西地省 村民 湖南省 湖南 发展 教育局 考试 职工 文件 小孩 社会 污染 医生 收费 家长 小学 农村 项目 地方 导致 开发商 参加 环境 工作人员 有限公司 生产 长沙 房屋 标准 费用 投诉 规划 服务 享受 建议 重视 请求 百姓 回复 通知 经济 群众 电梯 城市 中学 公路 工程 劳动 中心 专业 公务员 乡镇 关系 电话 员工 申请 旅游 违法 只能 答复 补课 学习 质量 补贴

图 4 基于 TextRank 算法的关键词提取

4.2.2 LDA 主题模型

LDA（Latent Dirichlet Allocation）是一种文档主题生成模型，也称为一个三

层贝叶斯概率模型，包含词、主题和文档三层结构。所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布。

LDA 采用词袋模型。所谓词袋模型，是将一篇文档，我们仅考虑一个词汇是否出现，而不考虑其出现的顺序。

LDA 主题建模效果如下图 5：

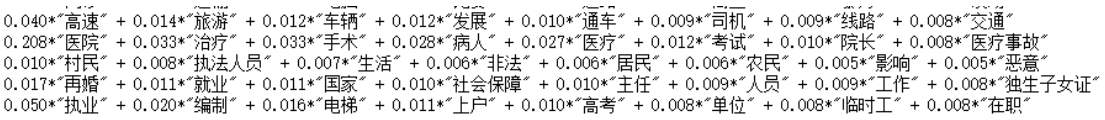

图 5 LDA 主题建模

5 问题一

本文完成问题一的具体流程图如图 6：

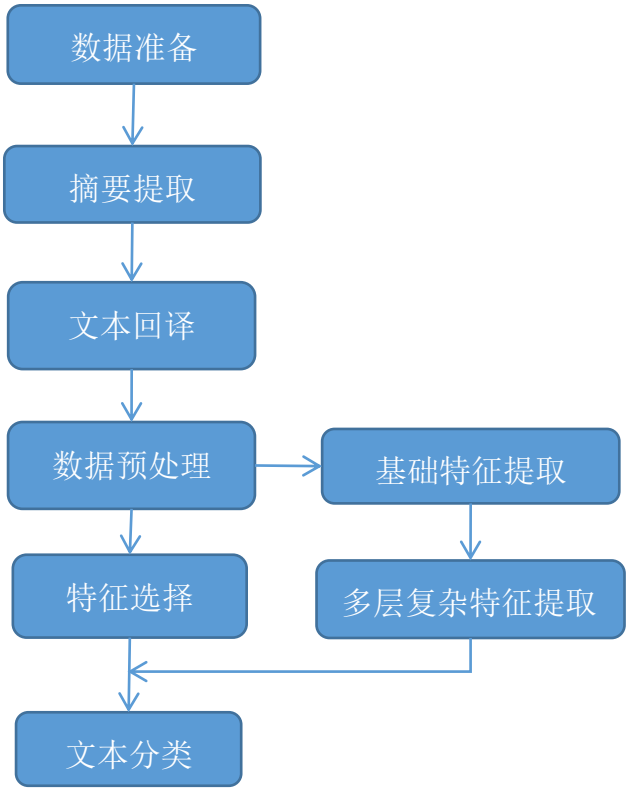

图 6 问题一流程分析

5.1 文本摘要提取

在群众发表留言过程中，较多的群众会讲述事情的起因，经过，结果等等，很多不重要的话语对文本分类造成了较大的干扰。文本摘要提取成为必不可少的过程。有两种办法，一种是基于 word2vec+TextRank 算法生成文章摘要，另一种是 NLPIR 语义分析系统。随机选取附件 2 中群众留言详情进行分析，效果如表 6：

表 6 文本摘要提取

留言主题	A 市高铁站的出租车管理需规范
word2vec+TextRank 算法	作为一名普通市民，我向您反映的这个长沙高铁站出租车管理现状，恳请陈书记百忙之中抽空督促一下，为最广大的普通乘客解决这一难题，也为长沙市的这一窗口树立形象。　　长沙已经是全国文明城市了，高铁站又是长沙的窗口单位，高铁站的这种乱象，给来长沙的游客将会带来深远的负面影响，作为普通的长沙市民，我们也觉得颜面无存。
NLPIR 语义分析系统	作为一名普通市民，我向您反映的这个 A 市高铁站出租车管理现状，恳请陈书记百忙之中抽空督促一下，为最广大的普通乘客解决这一难题，也为 A 市的这一窗口树立形象。

可见 NLPIR 语义分析系统摘要提取比基于 word2vec+TextRank 算法的效果稍好，无明显差别。本文选取 NLPIR 语义分析系统进行摘要提取，根据文本数据长度合理调节最大摘要长度、摘要最大压缩率参数，实现最佳摘要提取。

5.2 文本回译数据增强

在进行文本分类过程中，发现每个类别数据量存在不平衡现象。数据不平衡会严重影响数据分析的结果，在分类问题中的影响尤为巨大。常用的解决办法有过采样、欠采样等等办法，各有优点，同时也存在缺点。较为理想的办法是回译，即用机器翻译把一段中文翻译成另一种语言，然后再翻译回中文。回译的方法不仅有类似同义词替换的能力，它还具有在保持原意的前提下增加或移除单词并重新组织句子的能力。

百度翻译 API 是百度面向开发者推出的免费翻译服务开放接口，任何第三方应用或网站都可以通过使用百度翻译 API 为用户提供实时优质的多语言翻译服务，提升产品体验。利用官方 DEMO 改动部分代码，实现批量数据回译功能。

回译效果见表 7：

表 7 回译过程

	文本数据
回译前	A3 区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。强烈请求文明城市 A 市，尽快整改这个极不文明的路段。
回译后	西湖建设集团燕子山安置房项目施工围挡封闭 A3 地块西侧便道，即无管理处交叉口至加油站路段，人行道包括道路灯杆。这条路上每天都有很多人和车辆，特别是在上下班期间，有很大的安全隐患。强烈要求文明城市 a 尽快对这一极不文明路段进行整治。

可见回译效果良好，为解决数据不平衡问题提供了巨大帮助。

5.3 基于机器学习文本分类

5.3.1 分类前准备

有了前面的知识储备，运行由 Anaconda python3.6 编写的代码，类别数据量以及可视化效果如图 7 所示：

	cat	count
0	教育文体	4000
1	卫生计生	4000
2	城乡建设	4000
3	环境保护	4000
4	劳动和社会保障	4000
5	交通运输	4000
6	商贸旅游	4000

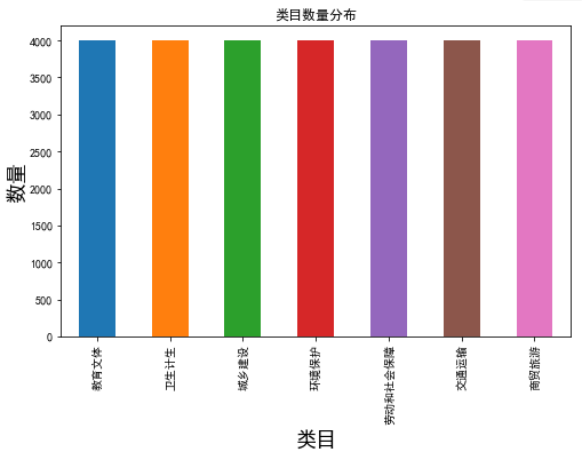

图 7 类别数据量统计

5.3.2 特征选择

在文本分类中，特征提取算法的优劣对于文本分类的结果具有非常大的影响。所以选择效果好的特征提取算法是文本分类前中很重要的步骤。本文运用卡

方检验，这是一个效果很好的特征提取方法。下面是用卡方检验来找出“劳动和社会保障”中关联度最大的词语和词语对的效果图 8：

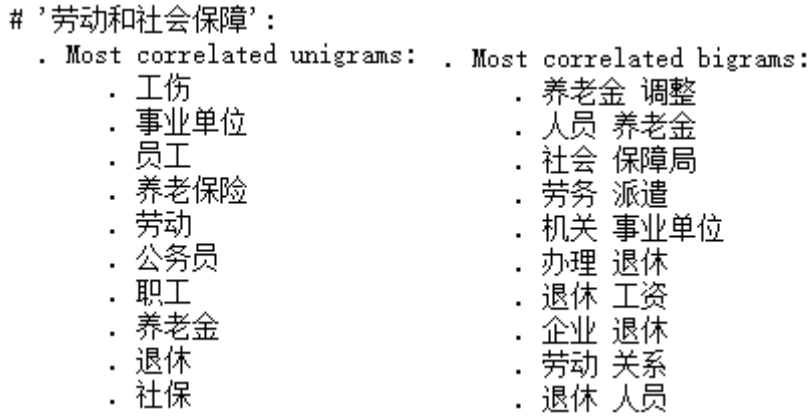

图 8 特征提取

5.3.3 分类模型

大部分机器学习方法都能在文本分类领域有所运用，本文运用了 SVM、朴素贝叶斯、逻辑回归算法、随机森林等等。效果如下表 8：

表 8 不同分类器分类

	Avg precision	Avg recall	Avg f1-score
LogisticRegression	0.90	0.90	0.90
SVM('linear')	0.89	0.89	0.89
SVM('rbf')	0.89	0.89	0.89
MultinomialNB	0.87	0.87	0.87
RandomForestClassifier	0.89	0.89	0.89

5.4 基于深度学习文本分类

基于深度学习文本分类模型已经成为了主流。

5.4.1 Fasttext 文本分类

fastText 是一个快速文本分类算法，与基于神经网络的分类算法相比有两大优点：

①fastText 在保持高精度的情况下加快了训练速度和测试速度

②fastText 不需要预训练好的词向量，fastText 会自己训练词向量

③fastText 两个重要的优化：Hierarchical Softmax、N-gram

生成需要的文本形式，调用 fastText 训练生成模型，求得准确率与召回率为 0.9665

5.4.2 基于神经网络文本分类

本文利用 Tensorflow 深度学习框架，版本型号是 1.14.0。本文运用基于 CNN 的中文文本分类，基于词袋的文本分类以及基于 RNN-GRU 融合的文本分类。设置参数调试运行结果，基于 CNN 的文本分类 Accuracy: 0.865000，基于词袋的文本分类 Accuracy: 0.892000，基于 RNN-GRU 的文本分类 Accuracy: 0.870143。

另一种方式是调用 kashgari 学习框架，kashgari 是一个简单而强大的 NLP 迁移学习框架。将文本数据划分为训练集、测试集、验证集。代入指定参数，效果如图 9 所示：

	precision	recall	f1-score	support
交通运输	0.9051	0.9217	0.9133	600
劳动和社会保障	0.8930	0.9183	0.9055	600
卫生计生	0.9503	0.9233	0.9366	600
商贸旅游	0.8897	0.8333	0.8606	600
城乡建设	0.8625	0.8783	0.8704	600
教育文体	0.9484	0.9183	0.9331	600
环境保护	0.9118	0.9650	0.9377	600
accuracy			0.9083	4200
macro avg	0.9087	0.9083	0.9082	4200
weighted avg	0.9087	0.9083	0.9082	4200

图 9 CNN 文本分类效果图

长文本特征比较明显，语料量也比较大，很容易取得不错的效果。当语料比较少，特征不是很明显时候直接训练可能会导致模型过拟合，泛化能力很差，可以使用预训练的词 Embedding 层来提高模型的泛化能力。例如：word2vec embedding。

6 问题二

本文完成问题二的具体流程图如图 10：

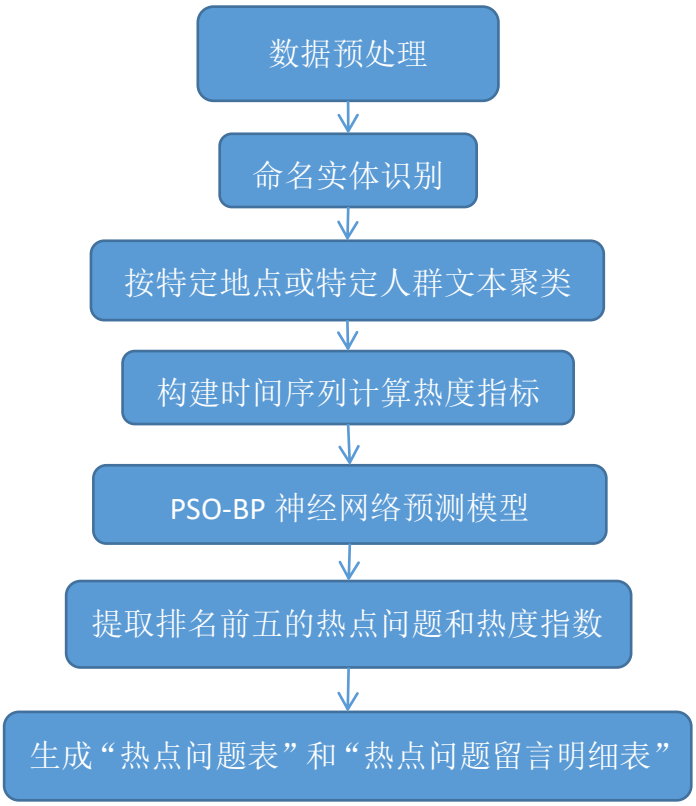

图 10 问题二流程分析

6.1 命名实体识别

NER 又称作专名识别，是自然语言处理中的一项基础任务，应用范围非常广泛。命名实体一般指的是文本中具有特定意义或者指代性强的实体，通常包括人名、地名、组织机构名、日期时间、专有名词等。NER 系统就是从非结构化的输入文本中抽取出上述实体，并且可以按照业务需求识别出更多类别的实体，比如产品名称、型号、价格等。

最先进的命名实体识别有百度 Lac、Standford NER、Spacy NER 等等。以下表 9 是斯坦福大学提供的 Stanza 自然语言处理工具包实现命名实体识别。

表 9 Stanza 自然语言处理效果

Sentence	希望 A 市地铁四号线北延线“同心路站”设在雷峰大道上
Tokenize	希望 A 市 地铁 四 号 线 北 延线 “ 同心 路 站 ” 设 在 雷峰 大道 上
UPOS	希望|VERB A|X 市|PART 地铁|NOUN 四|NUM 号|NOUN 线|PART 北|NOUN 延线|NOUN “|PUNCT 同心|PROPN 路|PART 站|PART ”|

	PUNCT 设|VERB 在|VERB 雷峰|PROPN 大道|NOUN 上|NOUN
XPOS	希望|VV A|FW 市|SFN 地铁|NN 四|CD 号|NN 线|SFN 北|NN 延线|NN “|`` 同心|NNP 路|SFN 站|SFN ”|'' 设|VV 在|VV 雷峰|NNP 大道|NN 上|NN
Name Entity Recognition	Entity: A 市地铁四号线 | Type: FAC Entity: 同心路站 | Type: FAC Entity: 雷峰大道 | Type: FAC

由于 Stanza 工具包的安装会有依赖包，相应对电脑系统和配置会有一定的要求。鉴于这个原因，本文运用了 FoolNLTK 中文处理工具包完成命名实体识别的任务，准确度虽然没有很高，但也能取得良好的效果。

6.2 按特定地点或特定人群文本聚类

聚类分析是一种无监督机器学习算法，它的目标是将相似的对象归到同一个簇中，将不相似的对象归到不同的簇中。

6.2.1 余弦相似度

余弦相似度是用向量空间中两个向量夹角的余弦值作为衡量两个个体差异的大小。相比于欧式距离，余弦相似度更适合计算文本的相似度。余弦值的计算公式如下：

$$\cos\theta = \frac{\sum_1^n (A_i \times B_i)}{\sqrt{\sum_1^n A_i^2} \times \sqrt{\sum_1^n B_i^2}} \tag{6.1}$$

6.2.2 性能度量

为了能让同一个簇内的样本尽可能相似，不同簇的样本尽可能不同，需要建立聚类性能度量指标。

设聚类结果的簇划分 $C = \{C_1, C_2, ..., C_K\}$，定义：

$$avg(C) = \frac{2}{(|C|(|C|-1))} \sum_{1 \le i < j \le |C|} dist(x_i, x_j) \tag{6.2}$$

$$diamC = \max_{1 \le i < j \le |C|} dist(x_i, x_j) \tag{6.3}$$

$$d_{\min(C_i,C_j)} = \min_{x_i \in C_i, x_j \in C_j} dist(x_i, x_j) \quad (6.4)$$

$$d_{cen}(C_i, C_j) = dist(\mu_i, \mu_j) \quad (6.5)$$

其中，μ代表簇C的中心点；$avg(C)$代表簇C内样本的平均距离；$diamC$代表簇C内样本间的最远距离；$d_{\min}(C_i, C_j)$对应于簇C_i和簇C_j最近样本间的距离；$d_{cen}(C_i, C_j)$对应于簇C_i和C_j中心点间的距离。基于以上公式导出两个聚类性能度量内部指标：

• DB 指数（Davies-Bouldin Index，简称 DBI）

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{avg(C_i) + avg(C_j)}{d_{cen}(C_i, C_j)}\right) \quad (6.6)$$

• Dumn 指数（Dumn Index，简称 DI）

$$DI = \min_{1 \le i \le k}\left\{\min_{j \neq i}\left(\frac{d_{\min}(C_i, C_j)}{\max_{1 \le l \le k} diam(C_l)}\right)\right\} \quad (6.7)$$

DB 指数的计算方法是任意两个簇内样本的平均距离之和除以两个簇的中心点距离，并取最大值，DBI 的值越小，意味着簇内距离越小，同时簇间的距离越大；Dumn 指数的计算方法是任意两个簇的最近样本间的距离除以簇内样本的最远距离的最大值，并取最小值，DI 的值越大，意味着簇间距离大而簇内距离小。因此，DBI 的值越小，同时 DI 的值越大，意味着聚类的效果越好。

6.2.3 聚类过程

聚类过程：①分词 ②构建词袋模型 ③权重转换 ④计算余弦相似度
⑤层次聚类算法 ⑥性能度量

层次聚类算法是在不同的层次对数据集进行划分，可以采用“自底向上”的聚类策略，也可以采用“自顶向下”的分拆策略。一般采用“自底向上”的策略，它的思路是先将数据集中的每个样本看作一个初始聚类簇，然后找出两个聚类最近的两个簇进行合并，不断重复该步骤，直到达到预设的聚类个数或某种条件。计算两个簇之间的距离一般有最小距离、最大距离和平均距离。

确定合适的聚类个数，具体思路是：

1） 选定一部分测试样本，对其进行层次聚类分析。

2） 计算性能度量指标 DBI 和 DI 的变化趋势，结合人工校验，得到一个合适的聚类个数和对应的距离阈值。

3） 将此距离阈值作为聚类结束的条件，对所有样本做聚类分析。此时无需再计算 DBI 和 DI 值，计算效率可以大幅提升。

部分数据完成聚类后的效果如图 11：

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
0	360114	A0182491	A市经济学院体育学院变相强制实习	2017-06-08 17:31:20	5公司签了合同，并且公司也要和我们	9	0
0	360110	A110021	济学院寒假过年期间组织学生去工厂	2019-11-22 14:42:14	多难过！虽说不是强制性的，但不去	0	0
0	360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	要求学生必须去学校安排的几个点实习	0	0
0	360113	A3352352	A市经济学院强制学生外出实习	2018-05-17 08:32:04	们都不知道！学校很小但是这几年只	3	0
0	360111	A1204455	市经济学院组织学生外出打工合理吗	2019-11-05 10:31:38	作十个小时以上，（晚班时间是20：3	1	0
1	289408	A0012413	市人才app上申请购房补贴为什么通	2018-11-15 16:07:12	6区也购买了灵活就业人员养老保险和	0	0
2	336608	A0005623	希望省把抗癌药品纳入医保范围	2019-09-08 21:01:59	昂贵。明明有药可治，就是买不起药品	0	0
3	360103	A0012425	劳动东路魅力之城小区临街门面烧烤	2019-09-25 00:31:33	为烧烤夜宵更加扰民，油烟24小时量列	1	0
3	360107	A0283523	路魅力之城小区一楼的夜宵摊严重污	2019-07-21 10:29:36	觉得要维护社会和谐稳定，合法维权	3	0
3	360108	A0283523	之城小区一楼的夜宵摊严重污染附近	2019-08-01 16:20:02	觉得要维护社会和谐稳定，合法维权	6	0
3	360100	A324156	魅力之城小区临街门面油烟直排扰民	2019-09-05 12:29:01	人。一天24小时都是烟。请政府关闭	3	0
3	360101	A324156	A5区劳动东路魅力之城小区油烟扰民	2019-07-28 12:49:18	清洗也没有。每天油烟直排。熏死树	4	0
3	360102	A1234140	劳动东路魅力之城小区底层餐馆油烟	2019-09-10 06:13:27	进屋内，窗户长期不能打开，晚上营	0	0
4	323149	A1241141	请给K3县乡村医生发卫生室执业许可	2019-06-20 20:38:47	新村的是证件下来啊。有些老村医反	0	0
5	343985	A108051	A市能否设立南塘城轨公交站？	2019-10-31 21:19:59	校区，南塘小学，A市一中城南中子	0	0
6	286572	A23525	A市地铁2#线在梅溪湖CBD处增设一	2018-10-27 15:13:26	玩桃花岭和梅溪湖，傍晚坐2#线回家	3	0
7	316619	A235259	请问A市什么时候能普及5G网络？	2019-05-14 11:22:13	全覆盖的城区建成。看消息介绍，5G	0	0

图 11 部分数据完成聚类后的效果

附件 3 有 4326 条数据，运算量相对来说比较大，在计算 DBI 和 DI 性能度量指标运算时间上存在一些困难，本文通过筛选相似度值高的样本归为同一簇。DI 值呈下降趋势，DBI 值呈阶跃上升趋势，根据性能度量的规则（DBI 的值越小越好；DI 的值越大越好），最优值可能出现阶跃点附近，同时结合人工校验，可以确定合理的类别数。

6.3 基于 PSO-BP 神经网络的热点话题预测模型

数据存在不确定性，假设热点问题与时间存在某种特定关系，实现热点话题预测显得尤为重要[10]。

6.3.1 计算热度值

根据附件 3 所给的数据特点，对群众留言信息定义一个八元组 D：

$$D=(id,user\_id,user,topic,time,text,praise\_num,oppose\_num)$$

其中，

$D.id$表示该留言信息的问题 ID

$D.user_id$表示该留言信息的留言编号

$D.user$表示发表留言的用户

$D.topic$表示该留言信息的主题

$D.time$表示该留言信息的发表时间

$D.text$表示该留言信息的内容

$D.praise_num$表示该留言信息的点赞数

$D.oppose_num$表示该留言信息的反对数

特定地点或人群发布的留言信息越多，说明话题的讨论程度越深，从而反映该留言信息的热度，点赞和反对行为表达用户对话题的关注和主观态度，从而也反映出该留言信息的热度。因此，本文选取在时间t内话题h的相关留言数量D_t^h为$tnum$、平均点赞数、平均反对数o_avg作为衡量话题热度t_hot的指标。各指标的计算公式如下：

$$tnum = D_t^h \tag{6.8}$$

$$p\_avg = \frac{\sum d.praise\_num}{tnum}, d \in D_t^h \tag{6.9}$$

$$o\_avg = \frac{\sum d.oppose\_num}{tnum}, d \in D_t^h \tag{6.10}$$

由选取的各项热度指标综合构建出话题热度的度量公式，公式如下：

$$t\_hot = tnum + p\_avg - o\_avg \tag{6.11}$$

通过分析话题的组成形式，以及不同指标的现实意义选取了具有特征性意义的话题热度指标，构建出话题的热度计算公式（6.11），使用公式计算话题的热度值t_hot，并将此值时间间隔设置为t,按时间序列得到每一个时间点上的序列$\{S_1, S_2, S_3, ..., S_l\}$，其中$S_i$表示第$i$个时间间隔内话题的热度$t_hot$，对于时间序列$\{S_1, S_2, S_3, ..., S_l\}$的观测值$S_n$，与前$d$个观测值间是非线性映射关系，其关系可以表示为：

$$S_n = f(S_{n-1}, S_{n-2}, ..., S_{n-d}), n = d+1, d+2, ..., l \tag{6.12}$$

其中，d为维数，$f(S_{n-1}, S_{n-2}, ..., S_{n-d})$为非线性函数。

以$\{S_1, S_2, S_3, ..., S_l\}$为基础构建热度值的训练样本集$\{x_i, t_i\}_i^N$，$x_i = [S_{n-1}, S_{n-2}, ..., S_{n-d}]^T$为热度值的输入样本；$t_i = S_n$为热度值的输出样本；

$N=l-d$ 为训练样本数目。利用该热度值训练样本训练 PSO 优化的 BP 神经网络，可使训练后的网络逼近 $f(S_{n-1},S_{n-2},...,S_{n-d})$，进而实现对 $\{S_1,S_2,S_3,...,S_l\}$ 后续话题热度值数据的预测并计算话题的热度指数。

6.3.2 热点话题预测及结果分析

本文采用 MATLAB2018a[11,12]仿真工具实现预测模型，使用构建好的 PSO 优化的 BP 神经网络模型对话题热度值进行预测分析。

经过解读和大量有效的实验验证，话题热度指数公式可表示为：

$$\alpha=\frac{\sum t\_hot_i}{T},i=1,2,...,l \tag{6.13}$$

其中，α 表示平均热度指数，t_hot_i 表示第 i 个时间间隔内的热度值，T 表示话题的时间范围。平均热度指数可以有效的表示话题在时间范围内热度值表现。

实验结果发现，2020 年 1 月相较其他月份数据量少，运用预测模型实现对该数据进行预测。由于前面 6.1 和 6.2 存在的缺陷而无法聚合，需要人工提取特征明显的热点话题后，剔除部分冷门话题并对热门话题进行二次模型预测，计算热度指数并且生成“热点问题表”和“热点问题留言明细表”。

7 问题三

本文完成问题三的具体流程图如图 12：

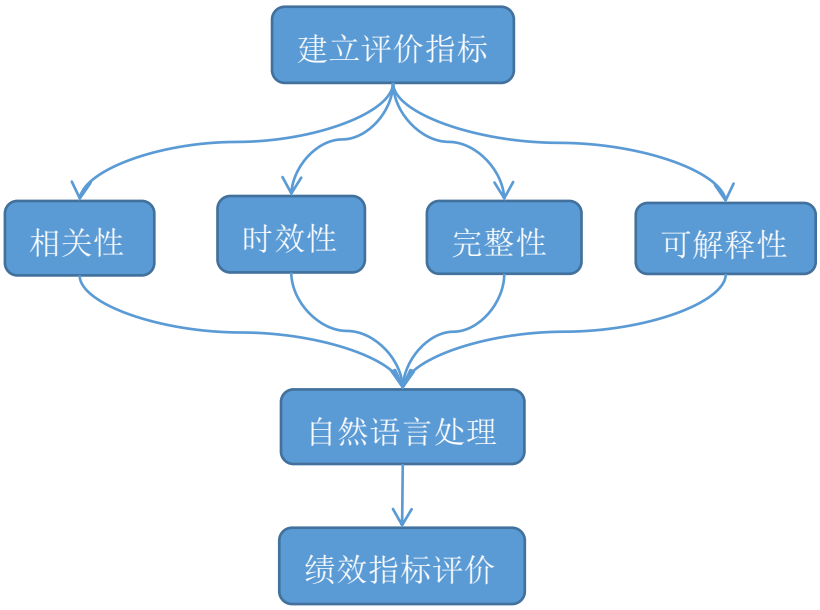

图 12 问题三流程分析

7.1 建立评价指标

根据所给数据的特点，建立合理的评价指标。

相关性:答复意见与群众留言信息的相似程度，用来说明留言集描述的概念和留言回复描述的概念之间的相关程度。好的答复意见应是正确引导群众并解决用户的困难，并不是答非所问。

时效性:答复时间与留言时间的时间差程度。用户提出问题都希望能尽快的解决。

完整性:答复人能否注意言辞礼貌、情绪控制、帮助态度形成答复意见的完整度。好的答复意见能让用户感到舒适。

可解释性:答复意见能否解释用户留言消息，是否存在语义模糊等情况。

7.1.1 相关性

词向量是将词映射到一个语义空间，得到的向量。Word2vec 是借用神经网络的方式实现的，考虑文本的上下文关系，有两种模型 CBOW 和 Skip-gram，这两种模型在训练的过程中类似。Skip-gram 模型是用一个词语作为输入，来预测

它周围的上下文，CBOW 模型是拿一个词语的上下文作为输入，来预测这个词语本身。

Word2vec 主要是将文本语料库转换成词向量。它会先从训练文本数据中构建一个词汇，然后获取向量表示词，由此产生的词向量可以作为某项功能用在许多自然语言处理和机器学习应用中。

Siamese Network[13]是一种神经网络的框架，而不是具体的某种网络，就像 seq2seq 一样，具体实现上可以使用 RNN 也可以使用 CNN。有些场景，字面上看可能不相似，但是从语义上看是相似的，这就需要更复杂的模型来捕捉它的语义信息。将 Siamese Network 架构中的用于表征 X1 与 X2 的 Network 更换为 LSTM 网络，就可以用于判断两个输入文本是否语义上相似。网格的结构如图 13 所示。

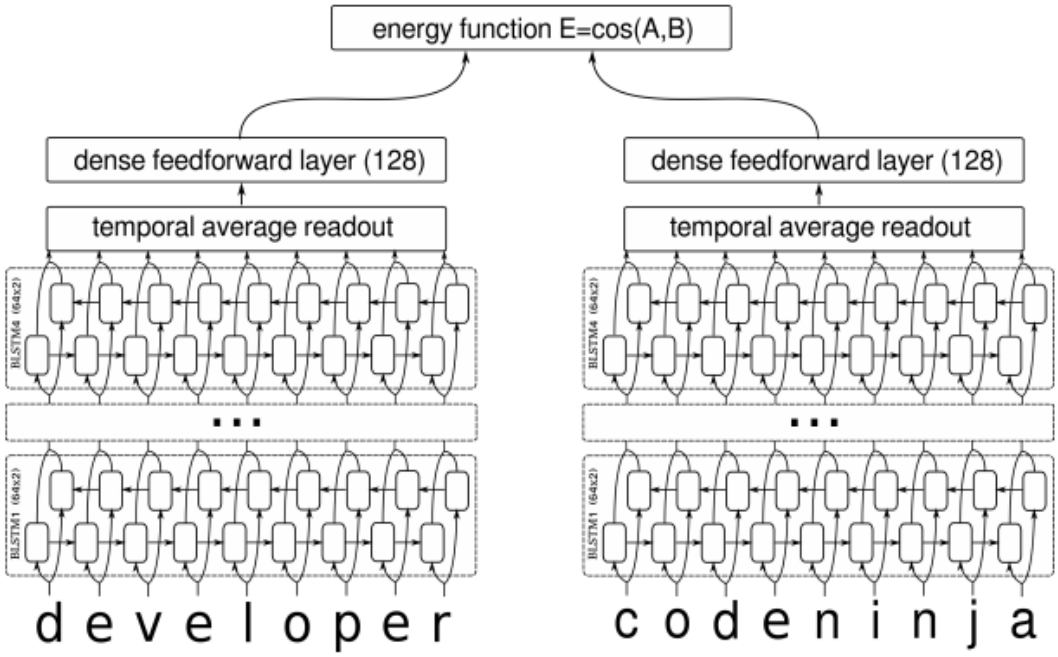

图 13 基于 Siamese Netword 判断文本相似

选择好的相似度是作为重要因素，基于词向量的相似度计算有余弦相似度、曼哈顿距离、欧几里得距离，经过多次试验，发现余弦相似度是作为最佳选择。如果文本相似，那么相似度会尽可能大。

7.1.2 时效性

反映时效性即答复时间与留言时间的时间差。对于附件 4 中的数据，答复时间最短的有 30 分 53 秒，答复时间最长的是 3 年 2 月 5 天 10 小时 28 分 14 秒。答复时间普遍集中在一个月以内。

7.1.3 完整性

答复人能否注意言辞礼貌、帮助态度、情绪控制形成答复意见的完整度。

言辞礼貌：根据数据特点，选取特征性的词语或感谢的句子反映答复礼仪的问题。例如，“网友 xxx 您好”“感谢您对我们工作的支持、理解和监督”。这些词语和句子显著分布在答复意见的开头和结尾，在处理过程中可单独进行分析。

帮助态度：根据数据特点，选取特征性的词语反映答复的态度。例如，“您的留言已收悉”“回复如下”。

情绪控制：文本情感分析，答复信息表达了情感色彩和情感倾向性,如喜、怒、哀、乐和批评、赞扬等。正向情感倾向可以给予加分。

自然语言处理针对言辞礼貌、帮助态度的识别较为简易，通过对句中进行分词，判断是否特征词是否存在于分词列表中即可。

基于深度学习的语义模型是目前较为流行的文本情感分析实现方法，百度飞浆开源了基于海量数据训练好的模型，都取得不错的效果。

7.1.4 可解释性

可解释性是指在用户需要了解或解决一件事情的时候，用户想获得所需要的足够的可以理解的信息。反过来理解，如果在一些情境中用户无法得到相应的足够的信息，那么这些事情对用户来说都是不可解释的。

随机森林是通过组合弱学习器（决策树）来进行分类和回归的有监督学习模型。由于其具有高性能，随机森林成为应用广泛的模型之一。随机森林模型常用作可解释性的研究。

7.2 绩效指标评价

根据所给定的评价指标，将评价指标形成 KPI 考核一份答复意见的质量。设置评价方案如表 10：

表 10 评价方案

考核点	比重	考核详情	得分（分）
相关性	25%	0.8~1 25 分 0.6~0.8 20 分	

		0.4~0.6　10 分 0~0.4　0 分	
时效性	25%	＜10天　25 分 10~20天　20 分 20~30天　10 分 ＞30天　0 分	
完整性	25%	3　25 分 2　20 分 1　10 分 0　0 分	
可解释性	25%	75%~100%　25 分 50%~75%　20 分 25%~50%　10 分 0%~25%　0 分	
总计	100%		

8 参考文献

[1] 周志华.机器学习[M].北京:清华大学出版社,2016

[2] Peter Harrington.机器学习实战[M].人民邮电出版社.2013

[3] 薛峰,胡越,夏帅,许剑东.基于论文标题和摘要的短文本分类研究[J].合肥工业大学学报,2018,41:53-59

[4] 孙明溪,刘春琦.基于 DBSCAN 算法与句间关系的热点话题发现研究[J].图书情报工作,2017,12(61):113-121

[5] 王晓光,王宏宇,黄菡.基于多源数据的专业领域热点探测模型研究[J].图书情报工作,2019,14(63):52-61

[6] 程克非,邓先均,周科,罗昭,陈旭东.基于微博多维度及综合权值的热点话题检测模型[J].重庆邮电大学学报,2019,4:468-475

[7] 李静, 徐路路.基于机器学习算法的研究热点趋势预测模型对比与分析—BP 神经网络、支持向量机与 LSTM 模型[J].现代情报,2019,39(04):24-34

[8] 李法运,陈亮.基于改进 BP 网络的网络论坛热点主题挖掘[J].计算机系统应用,2016,25(3):113-118

[9] 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[N].情报学报,2015,12:1278-1283

[10] 王惠. 微博热点话题分类与热度预测模型研究[D]:[硕士学位论文].天津:中国民航大学,2018

[11] 张采芳,余愿,鲁艳旻.MATLAB 变成及仿真应用[M].武汉:华中科技大学出版社,2014.1-32

[12] 于广艳,吴和静,张尔东,王强.MATLAB 简明实例教程[M].南京:东南大学出版社,2016.65-70

[13] Neculoiu P, Versteegh M, Rotaru M.Learning Text Similarity with Siamese Recurrent Networks[C]. Annual Meeting of the Association for Computational Linguistics(ACL),2016