摘要

近年来，网络问政成为了群众参与政治生活的重要渠道之一，越来越多的群众积极参与政治生活，在网络平台上向政府各部门提出自己的意见或者寻求帮助。这也意味着政府工作人需要处理大量的数据，显然，以往的人工处理方式在新时期遇到了极大的挑战。因此，建立基于自然语言处理技术的智慧政务平台迫在眉睫。

针对问题一，我们对文本进行预处理，并在这之后分别从传统机器学习与深度学习这两个方向进行建模。在传统机器学习方面，我们尝试了高斯贝叶斯、随机森林、SVM、XGBoost、KNN 这五个模型并对 XGBoost 与 KNN 模型进行优化。这几个模型中，F1 值最高的就是 KNN 为 0.836。在深度学习方面，我们构建了 1 维和 2 维的 TextCNN 模型，以及 LSTM 模型，最终将这 3 个模型进行集成学习，得到 F1 值为 0.8942.

针对问题二，我们对文本进行预处理之后，进行命名实体识别，来对特定地点，特定人群进行过滤，接着使用 word2vec 将文字转化为词向量，接着用 k-means 聚类，得到结果。我们热度指标为（点赞数+反对数）*0.5+留言数量*0.4+留言时间*0.1。

针对问题三，我们给出了一个有四个一级指标的 5 分制评价体系，这四个一级指标分别为相关性、完整性、可解释性与及时性，每一个一级指标有不同的得分标准，但总分均为 5 分。对于相关性，我们利用 Python 的 spacy 库计算“留言详情”与“答复意见”之间的文本相似度并将其映射到[0,5]区间内。对于完整性，我们给出了五个得分层次：已经帮助解决问题；尚未解决问题但是明确给出解决问题的建议方案；明确告知来访者已经问题移交相关部门；告知来访者可以解决该问题的特定部门，表明让来访者自行移交问题；并没有给出任何实质性的回答。分别对应 1-5 分。对于可解释性，我们按照其文本的移动程度给分，这一个指标有较大的主观性。对于及时性，我们计算答复与留言的时间差并取整，并划分<=3，<=7，<=15，<=31，>31 这五个时间段分别对应 1-5 分。

关键词：KNN TextCNN LSTM 集成学习 命名实体识别 word2vec k-means

Summary

In recent years, online political inquiry has become one of the important channels for the masses to participate in political life. More and more masses are actively participating in political life, presenting their opinions or asking for help to various government departments on the online platform. This also means that government workers need to process a large amount of data. Obviously, the previous manual processing method has encountered great challenges in the new period. Therefore, it is urgent to establish a smart government platform based on natural language processing technology.

For problem one, we preprocess the text, and then model from the two directions of traditional machine learning and deep learning. In traditional machine learning, we tried the five models of Gauss Bayes, Random Forest, SVM, XGBoost, and KNN and optimized the XGBoost and KNN models. Among these models, the highest F1 value is the KNN of 0.836. In terms of deep learning, we built 1D and 2D TextCNN models, and LSTM models, and finally integrated learning of these 3 models, we got F1 value of 0.8942.

For problem two, after preprocessing the text, we perform named entity recognition to filter specific locations and specific groups of people, then use word2vec to convert the text into word vectors, and then use k-means clustering to get the results. Our popularity index is (number of likes + objections) * 0.5 + number of messages * 0.4 + message time * 0.1.

In response to question three, we gave a 5-point evaluation system with four first-level indicators. These four first-level indicators are relevance, completeness, interpretability and timeliness, each of which is different. The score standard, but the total score is 5 points. For relevance, we use Python's spacy library to calculate the text similarity between "message details" and "reply comments" and map them to the [0,5] interval. For completeness, we give five scoring levels: have helped solve the problem; have not yet solved the problem but clearly given a solution to the problem; clearly inform the visitor that the problem has been transferred to the relevant department; inform the visitor that the problem can be solved The department stated that the visitor

was asked to hand over the question by himself; no substantive answer was given. Corresponding to 1-5 points. For interpretability, we give points according to the degree of movement of its text, this indicator has a greater subjectivity. For timeliness, we calculate and round off the time difference between the reply and the message, and divide the five time periods of <= 3, <= 7, <= 15, <= 31, and> 31 corresponding to 1-5 points respectively.

Keywords: KNN TextCNN LSTM integrated learning named entity recognition word2vec k-means

目录

1 绪论

1.1 挖掘意义

近年来，随着互联网技术的发展以及人民群众参与政治生活的活跃度的提高，越来越多的民众通过微信，微博，市长信箱等网络问政平台发表自己的意见与建议，这使得大量的文本数据涌入政府平台。对于网络问政平台中的民众留言，政府工作者需要根据其信息将其划分到特定的类别并交由特定的部门处理，另外，他们还需要找出热点问题，重点解决。在实际的工作中，工作人员往往并不需要完整地通读所有的民众留言，更多情况下，工作人员希望能够快速的提取出每条民众留言中的时间、地点、人物、事件等关键信息，并根据这些关键信息快速的将民众留言分类并发送给特定的机关单位。同时工作人员也希望依靠这些关键信息方便快捷地找出某一时段某一地区内的热点问题。然而，在面对巨大的数据量时，以往的人工处理的方法显然不再合适，因此，建立基于自然语言处理技术的智慧政务平台迫在眉睫，这对于提高政府的工作效率解决群众问题有极大的推动作用。

将自然语言处理技术与政府工作相结合是近年来较为热门的自然语言处理任务。构建基于自然语言处理技术的智慧政务平台的目的是转变政府工作模式，提高政府工作效率，更精准的解决群众问题。因此，这对于构建和谐的社会政治生活有极大的研究与应用价值。

1.2 挖掘目标

基于群众问政留言记录的内容，我们要构建文本分类模型，模型能准确地识别出留言内容所属的类别，以解决单纯依靠人工根据经验处理所存在的工作量大、效率低，且差错率高等问题。具体来说，在读入大量留言记录后，自动定位到留言内容本身，将此进行预处理后的词向量进入模型进行分类，输出一级标签分类内容。

我们还需构建聚类模型，筛选出某一时段内群众集中反映的某些热点问题，有助于提取针对性内容，使相关部门的工作效率提高，我们将获取大量数据，考虑文本本身的相似度，通过模型将热点内容挖掘出来，输出群众反映的共性问题。

群众关切问题需要相关部门提出答复意见，答复也存在差异之分，答复完整、措施有效都能为部门工作增光添彩，我们希望对此提出一套评价方案，为更好改进政务工作。

1.3 问题分析

1.3.1 群众留言问题分类

该问题要求我们训练处一个合适的模型来对群众的留言进行分类，并用 F1 值来评价该模型。针对这个问题，我们首先进行了初步的属于预处理，对附录一中的数据进行抽样、分词、去停用词、提取关键字；之后，我们分别从深度学习与传统机器学习两个角度来构建模型并优化；最后我们将两种方法的模型结果进行对比，发现深度学习在该问题上有更高的 F1 值，模型性能也更好。该问题的思路图如下：

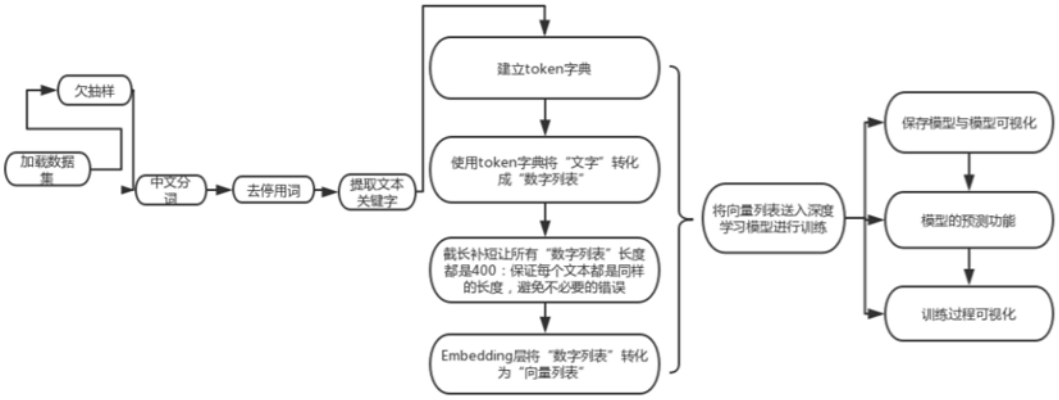


图 1 深度学习建模流程图

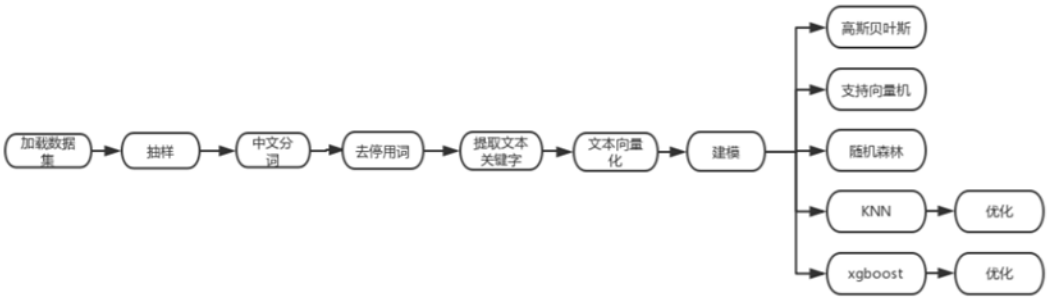


图 2 传统机器学习建模流程图

1.3.2 热点问题挖掘

运用命名实体识别技术以及 k-means 聚类算法将热点问题找出，并根据点赞数加反对数，留言数量，留言时间对热度指数进行评价。

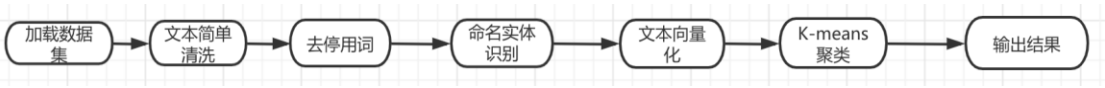


图 3 聚类流程图

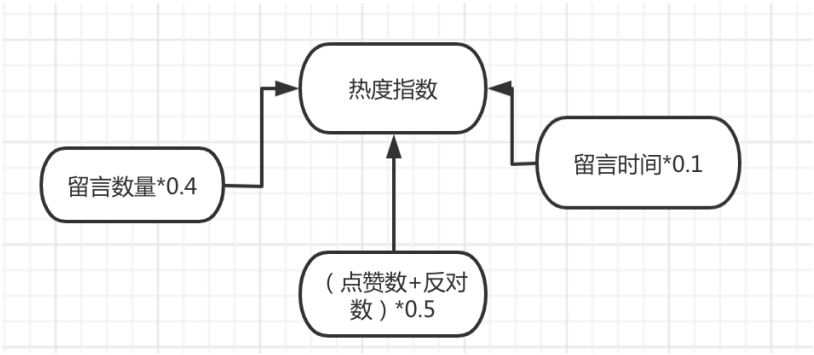


图 4 热度指标

1.3.3 答复性意见评价

问题三要求我们从多个方面给出一套完整的政府答复意见的评价体系并尝试实现。我们仔细探究附录三数据并结合文献资料给出了一个拥有不同权重的有四个一级指标（相关性、可解释性、及时性、完整性）构成的评价体系。该体系采用五分制，除相关性外其他三个指标的得分都分为 1-5 分五个阶段，并给出明确的得分体系。而相关性则是根据数据的“答复意见”与“留言内容”的文本相关度得出的，文本相关度是一个[0,1]区间内的数，我们将其映射到[1,5]区间。这四个指标的加权得分就是该答复意见的最终得分。

2 数据集分析

2.1 附件二：数据集分析

附件二数据集如下图所示：

留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
24	A00074011	建筑集团占道施工有安	2020/1/6 12:09:38	工围墙内。每天尤其上下	城乡建设
37	U0008473	大厦人为烂尾多年，安	2020/1/4 11:17:46	围着，不但占用人行道路	城乡建设
83	A00063999	市A1区苑物业违规收停	2019/12/30 17:06:14	已多次向物业和社区提出	城乡建设
303	U0007137	南路A2区华庭楼顶水箱	2019/12/6 14:40:14	用品，霉是一种强致癌物，	城乡建设
319	U0007137	2区华庭自来水好大一股	2019/12/5 11:17:22	用品，霉是一种强致癌物，	城乡建设
379	A00016773	市盛世耀凯小区物业无证	2019/11/28 9:08:38	物业不是为业主服务的，	城乡建设
382	U0005806	询A市楼盘集中供暖一	2019/11/27 17:14:11	处月亮岛片区近年规划有	城乡建设
445	A00019209	西路可可小城长期停水	2019/11/19 22:39:36	寻求帮助至今没有找到具	城乡建设
476	U0003167	收取城市垃圾处理费不当	2019/11/15 11:44:12	所在的物业公司也未给出	城乡建设

图 5 附件二数据集

分析发现这是一个 9210 行 6 列的数据集，说明该数据集共有 9260 个数据，其中每个数据对用 6 个特征，分别是："留言编号"，"留言用户"，"留言主题"，"留言时间"，"留言详情"，"一级标签"。针对第一题，我主要需要用到的就是"留言详情"和"一级标签"这两个特征。我们统计发现，"一级标签"共有7个大类并且每类的数据量分布并不均衡，其分布情况如下：

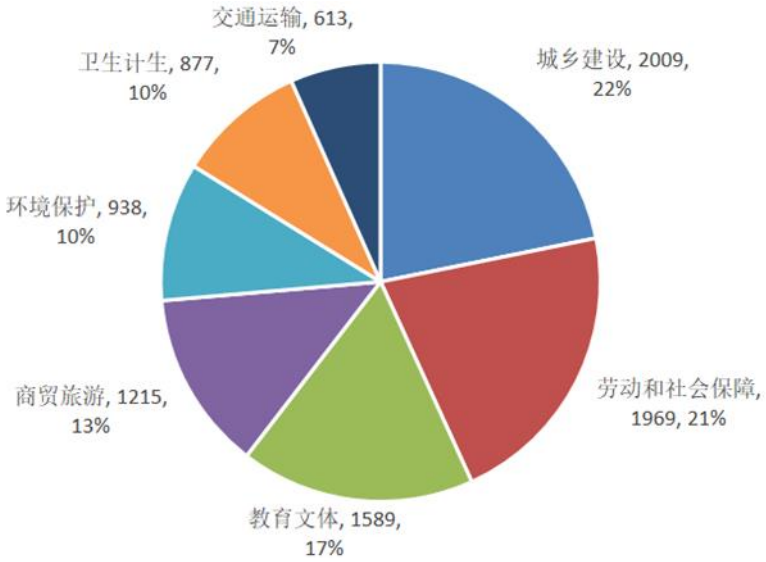


图 6"一级标签"各类数据分布情况图

2.2 附件三：数据集分析

附件三数据集如下图所示：

留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
198766	A000100057	省津楚投资有限公司法人失踪	2019/11/4 16:23:15	投资责任有限公司财务人员吴桐	0	0
238649	A000100057	责任有限公司资金链断裂，	2019/10/29 12:45:40	信任和支持，把多年赞下的血本	0	8
205986	A00010015	期和五期之间的白塞湖路怎	2019/4/29 10:28:46	工一天又停了！为什么？为什	0	0
244529	A000100163	路的华润万家烂尾快7年了，	2019/7/25 15:50:30	三路的华润万家烂尾快7年了，	0	6
227740	A000100203	A市峰之尚物业不给员工买	2019/4/16 18:32:14	时的工作时间，都是在骗员工。	0	0
232520	A000100204	桂园城市花园小区旁违规建	2019/6/26 11:48:25	幼儿园。现如今垃圾站的存在，	0	0
196664	A000100326	学会泄露考生个人信息如此	2019/2/22 14:38:33	的人充当工作人员把关，这里	0	3
246288	A000100333	悦小区1栋二单元一楼105房	2019/7/4 14:30:24	一楼105房，违规敲除外墙，改变	0	5
283304	A000100346	大火，竟成东富地产公司逼	2019/8/6 15:05:35	修并不会影响生活区供电，且生	0	4

图 7 附件三数据集

这是一个包含 4327 个数据，7 个特征的数据集。问题二要求我们找出热点问题并且定义一个合理的热度评价指标。针对第一个要求，我们选择使用“留言主题”这一个特征，并对他聚类；针对第二个要求，我们需要统计总的点击数即“反对数”与“点赞数”，并根据第一个要求所得出的结果统计每一类问题的评论数，再根据“留言时间”计算问题持续时间作为第三个指标。

2.3 附件四：数据集分析

附件四数据集如下图所示：

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
16542	UU0081101	安沙镇花桥村村民请求道路	2018/5/24 8:17:48	约有2公里毛路坑坑洼洼，出行	每年农村公路建设指标有限。建议：1、自为	2018/5/24 16:12:33
23820	UU0081985	县独生子女证领取金的咨	2016/12/1 8:34:10	享受天年，但是享受了独生	，农村户口、现存一个子女或两个女儿、19	2016/12/1 13:08:41
32712	UU0081124	皇图岭镇界联村请求整改	2019/11/20 16:43:58	不畅”专项工作之际，我请求	的留言收悉，已转有关部门调查核处，如	2019/11/20 17:48:59
33742	UU0082128	公园停工一事希望政府能	2018/11/27 8:20:02	违法，股权冻结，我们业主都	求新龙凯公司尽快向经开公安分局报案，依	2018/11/27 21:43:51
32791	UU0081932	的蓝溪谷讨要工资未果，请	2019/10/28 11:44:13	起来有几十万，都没拿到工资	映的情况，建议您向有管辖权的劳动保障行	2019/10/28 15:40:26
33065	UU0082359	市民咨询狂犬疫苗报销一事	2019/8/7 15:18:03	苗，小孩有居民医保可以报	机制，对于学生儿童因突发的、外来的、非	2019/8/7 19:40:26
33132	UU0081841	市研究生人才引进政策的咨	2019/7/16 15:08:12	前听说过有关于B市硕士人才	B市组工网（https://baidu.com/）详细了	2019/7/16 21:07:52
33655	UU0081572	中燃气乱收费现象多次投诉	2018/12/29 16:20:52	燃气管合格，却要求重新更换	质等作出具体规定。我委在2017年11月3日开	2018/12/29 17:35:11
32714	UU0082175	B市足球场新建规划的咨询	2019/11/19 18:17:48	没重点城市，B市有上榜，请问	《问政西地省》的留言收悉，已转有关部门	2019/11/20 17:50:44

图 8 附件四数据集

这是一个拥有 2817 个数据，每个数据拥有 7 个不同特征的数据集。根据我们的答复评价体系，我们需要计算每条数据的“留言详情”与“答复意见”之间的文本相似度，并将答复时间与留言时间之间的差值取整作为我们评价答复及时性的唯一标准，同时我们还需要更具“答复意见”的具体内容，人为的评价其完整性与可解释性。

3 问题一 群众留言分类

本题我们尝试使用传统机器学习方法和深度学习方法对群众留言分类进行建模。

3.1 实验平台

表 1 实验环境配置

CPU	I5-7/4 核
内存	28G
操作系统	Windows10
Python	3.6
GPU	Gtx 1080ti/1 块

3．2 数据预处理

3.2.1 欠抽样

在载入数据后，我们发现该数据集是一个 9210 行，6 列的数据集。数据中各类别数据占比绘制图形如下所示：

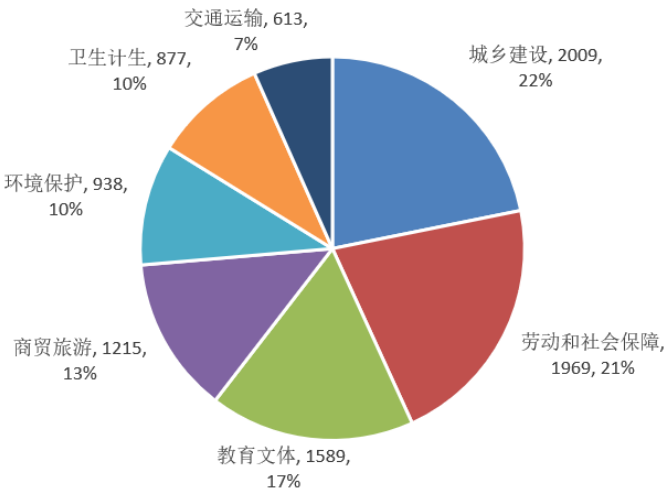


图 9 一级标签各类别数据占比图

由该图可知，我们数据存在类别数量不平衡的问题，其中占比最大的“城乡建设”（22%）是占比最小的“交通运输”（7%）的 3 倍以上，这样可能会导致我们模型偏好占比更大的类别。

为避免这一问题，使得模型泛化能力更强，我们采取欠抽样的方法，以数据量最少的“交通运输”的数量为基准，我们对每一类进行随机抽样，数量为 613 个，共 4293 个数据，并将其打乱顺序组合。其结果如下图所示。

	留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
4818	290809	U0002307	别为了扶贫那点政绩，让全D7县初三学子再受委屈	2018/11/26 8:00:31	\n\n初中公费师范招生不招初中，小学，幼儿园教师，2017年招生50，实际录取22.，20...	教育文体
7235	32561	U0002016	A7县兴隆路金井批发部卖假酒，欺骗老顾客十多年	2017/4/11 2:04:38	\n\t\t\t\t\t\n\t\t\t\t\t 星沙街道兴隆路星沙4区聚鑫大厦附近有家 ...	商贸旅游
8706	116559	U0003299	建议K8县卫计委为保障群众低保民生让路	2015/4/8 14:13:36	\n \n 贫困地区的群众都知道计划生育不合理政策才是真正的“大老虎”，...	卫生计生
6857	301185	U0002079	A市社保卡配送的效率太低了	2019/1/29 11:44:11	\n \n 尊敬的领导：您好！我是一名刚毕业工作了半年的大学生，2018年9...	劳动和社会保障
2186	60611	U000575	投诉D2区友邦化工厂直接向楚江排污	2013/7/13 12:21:54	\n\t\t\t\t\t\n\t\t\t\t\t 友邦化工厂位于D市D2区省道S214线离...	环境保护

图 10 抽样结果

3.2.2 简单文本清洗

利用正则表达式将文本内标点符号，\t,\n 等符号去除

3.2.3 中文文本分词

中文文本分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。在这里我们使用 Python 的 jieba 库进行 jieba 分词，这是一个优秀的中文分词第三方库，他允许我们自己添加新词以提高分词的准确率。其结果如下图 11 所示。

```
4818    [初中, 公费, 师范, 招生, 不招, 初, 中小学, 幼儿园, 教师, 年, 招生, 实...
7235    [星沙, 街道, 兴隆路, 星, 沙区, 聚鑫, 大厦, 附近, 有家, 金井, 批发部, ...
8706    [贫困地区, 的, 群众, 都, 知道, 计划生育, 不合理, 政策, 才, 是, 真正, ...
6857    [尊敬, 的, 领导, 您好, 我, 是, 一名, 刚, 毕业, 工作, 了, 半年, 的,...
2186    [友邦, 化工厂, 位于, 市区, 省道线, 离市, 公里, 处原, 县, 老, 橡胶厂, ...
Name: 留言详情, dtype: object
```

图 11 分词结果

3.2.4 去停用词处理

停用词（Stop Word）是指在信息检索中为了节省存储空间和提高搜索效率，在处理自

然语言数据（或文本）之前或者之后会自动过滤掉的字或词。停用词在检索过程中往往是一个不重要甚至可以忽略的因素，为防止他对我们产生不必要的干扰，我们将其去除。这一步需要用到我们从外部获取的停用词表。去停用词结果如下图 12 所示。

```
4818    [初中, 公费, 师范, 招生, 不招, 初, 中小学, 幼儿园, 教师, 招生, 实际, ...
7235    [星沙, 街道, 兴隆路, 星, 沙区, 聚鑫, 大厦, 附近, 有家, 金井, 批发部, ...
8706    [贫困地区, 群众, 知道, 计划生育, 不合理, 政策, 才, 真正, 大, 老虎, 并且...
6857    [尊敬, 领导, 我, 一名, 刚, 毕业, 工作, 半年, 大学生, 月, 办理, 人才,...
2186    [友邦, 化工厂, 位于, 市区, 省道线, 离市, 公里, 处原, 县, 老, 橡胶厂, ...
Name: 留言详情, dtype: object
```

图 12 去停用词结果

3.3 传统机器学习方法建模

在传统机器学习中，我们将“留言详情”作为主要文本数据

3.3.1 数据进一步处理

3.3.1.1 关键字提取

“留言详情”关键词提取参数 topK 的取值依据如下：

我们以去除停用词后“留言详情”中的单词数为横轴，以数据编号为纵轴，绘制图像如下所示。

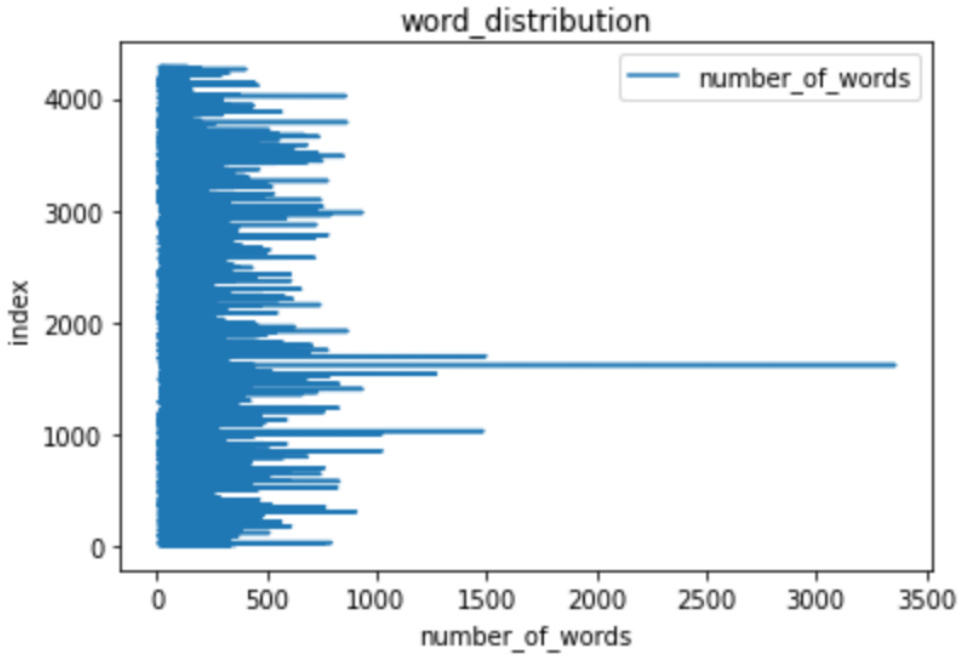


图 13 留言详情单词数分布图

由下图可知，大部分“留言详情”单词数不超过 400，故我们设置参数 topK 为 400

analyse.textrank()方法参数如下：

表 2 留言详情 textrank 方法参数表

3.3.1.2 文本向量化

在本模型中，我们主要采用将自然语言处理的问题转化为传统机器学习的方式来进行学习建模，这需要我们首先将本问题所需的数据集，即民众留言的中文文本进行数字化表示。这里我们选择 TF-IDF 算法来将中文文本转化为词向量。如下图 14 所示。

参数	参数值
Sentence	text
topK	400
withWeight	False
allowPOS	(‘ns’,’n’,’vn’,’v’)

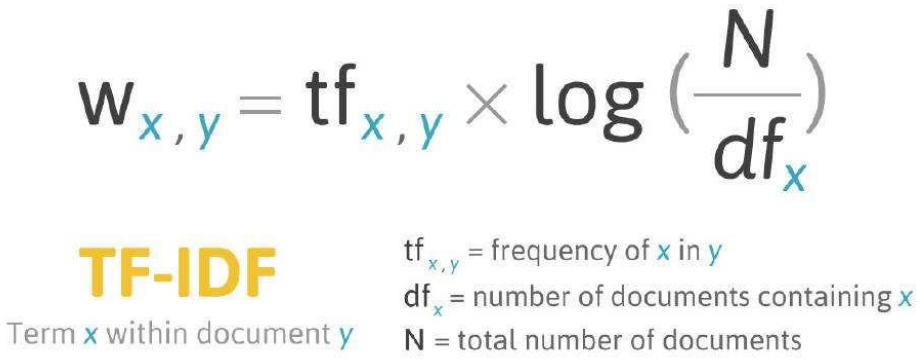


图 14 TF-IDF **算法**

TF-IDF 算法是一种将文本离散化表示的算法，他充分的考虑到了词在文本中出现的频率，并通过分配权重来反映每个词的重要程度。TF-IDF 算法分为两个部分：词频（TF）和逆文档频率（IDF）。其中：词频（TF）表示一个词语与这篇文章的相关性；逆文档率（IDF）表示一个词语的出现的普遍程度。TF-IDF 算法就是将这两部分相乘得到的乘积作为这个词最终的权重。

$$词频(\boldsymbol{TF}) = \frac{某个词在文本中出现的次数}{文本总词数}$$

$$逆文档率(IDF) = log\left(\frac{文章总数}{（包含该词的文章数+1）}\right)$$

计算得出：

$$TF - IDF = TF * IDF$$

其结果如下图 15 所示。

关键词	包含该词的文档数	TF	IDF	TF-IDF
亚洲	62.3亿	0.02	0.603	0.0121
网络	0.484亿	0.02	2.713	0.0543
技术	0.973亿	0.02	2.41	0.0482

图 15 TF-IDF **结果**

该算法实现简单而且相对容易理解，但是也存在着权重分配不全面、精度不够高、忽略词语 X 的位置信息等问题。

由于机器学习模型需要划分训练集与测试集，在这里我们首先调用 Python 的 sklearn 库中的 train_test_split 方法将数据集与标签集分别划分出相应的训练集与测试集。并调用 sklearn 库中的 CountVectorizer,与 TfidfTransformer 分别用于获取权值矩阵以及将训练数据集与测试数据集向量化。最终我们得到训练数据集 data_tr_t，测试数据集 data_te_t，训练标签集 labels_tr，测试数据集 labels_te。

3.3.2 所用模型介绍

3.3.2.1 高斯贝叶斯模型

朴素贝叶斯（Naive Bayes）是一种简单的分类算法，他的理论基础是贝叶斯定理和特征条件独立假设，这里我们所用到的高斯贝叶斯就是类属于朴素贝叶斯算法。高斯朴素贝叶斯假设$P(X_i|Y)$是服从高斯分布（也就是正态分布），然后便可以使用极大似然估计得到高斯分布的参数——均值和方差，来估计每个特征下每个类别上的条件概率。

3.3.2.2 支持向量机模型（SVM）

支持向量机是一种对线性和非线性数据进行分类、回归与异常点检验的有监督的学习方法，其基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。SVM 几乎是一个全能型的机器学习算法，面对不同的问题，他可以通过指定不同的核函数来解决。但是当 SVM 面对拥有大量数据的多分类问题时，他的缺陷也非常明显。因为在面对多个类别时 SVM 为每一个类别训练出一个分类器，每个分类器将训练样本分为K_i类和非K_i类。这将会大大的延长模型的训练时间。

3.3.2.3 随机森林模型

随机森林是一类专门为决策树分类器设计的组合方法，将多棵树集合就是一个“森林”。个体决策树在每个结点使用随机选择的属性决定划分，最终形成一棵完整的数。分类时，每棵树都投票并且返回得票最多的类。

随机森林对于大部分的数据，它的分类效果比较好，模型的训练速度较快，能评估变数的重要性，其分类准确率能与 AdaBoost 相媲美。

3.3.2.4K 邻近算法模型（KNN）

最近邻分类法是基于类比学习的，训练元组用 n 个属性来表示，每个元组代表 n 维空间的一个点，所有元组都存放在 n 维模式空间里。然后，当出现某个未知元组时，KNN 就会搜索原有的模式空间，找出最接近未知元组的 k 个训练元组，那么这 k 个训练元组就是未知元组的 k 个“最近邻”，换句话说就是说的是每个样本都可以用它最接近的 k 个邻居来代表。

基本步骤：

1.计算测试数据与各个训练数据之间的距离

2.选择参数 K

3. 按照距离关系选取距离最小的 K 个点并计算概率

4.将最高概率类别作为测试数据的预测分类

确定 k 值极其重要，如果 k 太小，则最近邻分类器容易由于噪声产生过拟合，如果 k 过大，则可能会误分类测试样例。

3.3.2.5XGBoost

XGBoost 全称是 extreme Gradient Boosting，可翻译为极限梯度提升算法，致力于让提升树突破自身的计算极限，以实现运算快速，性能优秀的工程目标。和以往梯度提升树相比，XGBoost 的优势在于提高了泛化度，提高了精确度与速度，是集大成的机器学习算法。

XGBoost 核心思想为：不断地添加树，不断地进行特征分裂来生长一棵树，每次添加一个树，其实是学习一个新函数 f(x)，去拟合上次预测的残差，逐渐渐形成众多树模型集成的强评估器。

当我们训练完成得到 k 棵树，我们要预测一个样本的分数，其实就是根据这个样本的特征，在每棵树中会落到对应的一个叶子节点，每个叶子节点就对应一个分数，样本在树上的取值，用$f_k(x_i)$表示。

最后只需要将每棵树对应的分数加起来就是该样本的预测值。

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

3.3.3 实验结果分析

这里我们采用规定的 F1_score 作为模型评估指标，这的指标在每次完整运行代码的过程中会有 4%左右的波动。

在实验过程中为了选取最佳的模型我们选用了多种模型，包括高斯贝叶斯、支持向量机、随机森林、XGBoost、KNN。在初次运行时这些模型的表现有较大的差异。其结果如下：

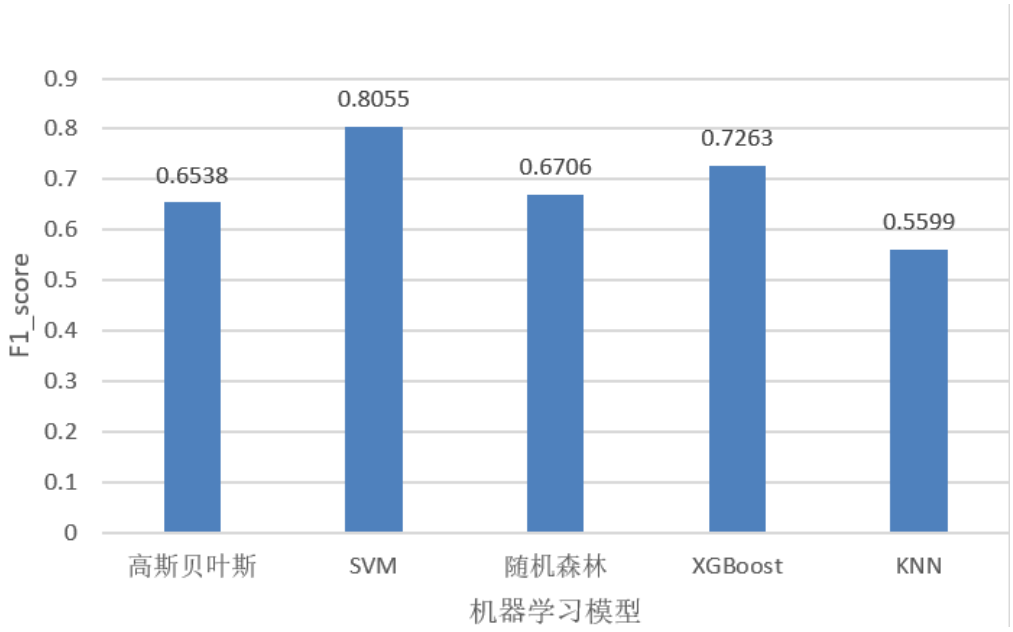


图 16 机器学习模型 F1 值

其中非常适合用来做文本分类的高斯贝叶斯只有 0.6538 的 F1 值，集成算法中随机森林也只有 0.6706 的 F1 值，可见这两个模型在这个问题上的泛化能力并不好。而 KNN 算法只有 0.5599 的 F1 值，泛化能力更差，但是这个算法运行时间较短。最令人期待的 XGBoost 表现较好有 0.7263 的 F1 值。在这些算法中支持向量机（SVM）表现最好，但是，SVM 并不是在左右的核函数上都表现的良好。SVM 共有四个核函数：“liner”、“poly”、“sigmoid”、“rbf”。其中当核函数为“liner”时效果最好，为 0.8055，但是 SVM 在其他三个核函数上表现的很差均只有 0.0322 的 F1 值。

3.3.3.4 模型优化

为了进一步提高模型的泛化能力，我们选择我们认为有较大潜力的 XGBoost 与 KNN 来调参优化。我们调参的方法主要是画出学习曲线来寻找最好的参数值。

3.3.3.4.1XGBoost 调参

XGBoost 库的 XGBClassifier 中有两个重要度参数：

n_estimatores：表示模型总动迭代的次数，即决策树的个数

learning_rate：表示学习率，控制每次迭代更新权重的步长，默认为 0.3

这里我们调节的就是这两个参数。同时考虑到运行效率的问题，我们每次调参范围都选取的较小。

（1）n_estimatores 调参

考虑到模型运行的效率，我们首先选取 90 到 110 内每隔 5 的参数值，其运行结果如下：

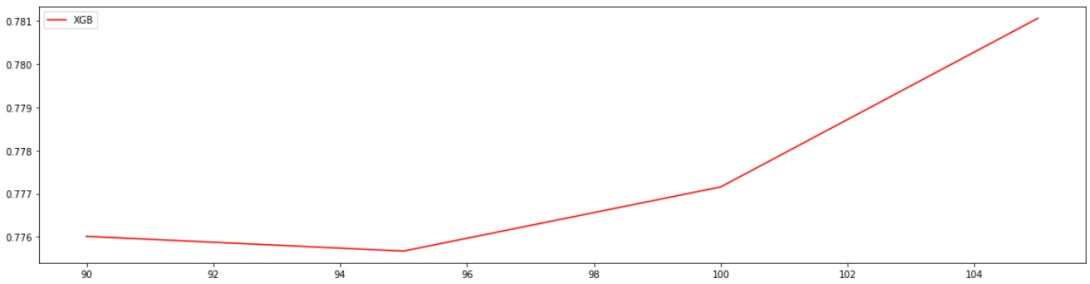


图 17　n_estimatores90-110

可见最高的是参数值为 105 时，此时 F1 值为 0.7810，并且在 105 以后的 F1 值有一个明显的上升趋势，所以我们再次选择 110 到 130 内每隔 5 的参数值调参，每次画出的学习曲线的最高 F1 值对应的参数值以及学习曲线展示的特点，我们继续调适，最终，考虑到模型运行的性能，我们选择参数值为 145，此时有最高的 F1 值为 0.7936

最终我们选择 n_estimatores=145.

（2）learning_rate 调参

在调参过程中我们选择将区间分为 0 到 0.5 与 0.5 到 1 这两个小区间并分别在这两个区间内每隔 0.1 取参数值画学习曲线，当模型的参数 learning_rate 取值为 0.3，即默认值时 F1

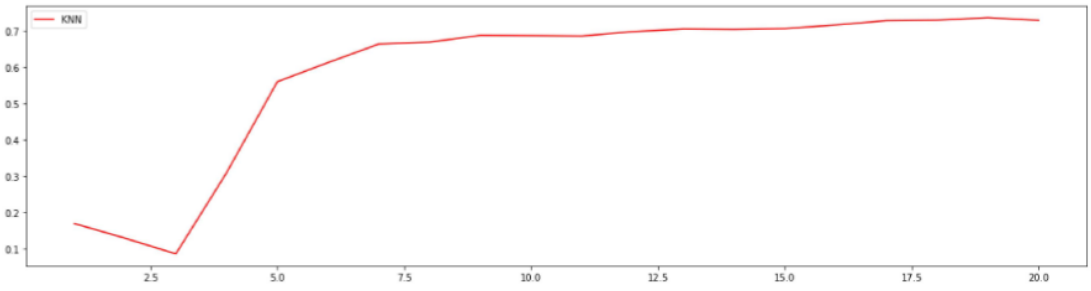
值最高。

综合两个参数，模型的 F1 值从 0.7263 上升到 0.7936，提高了 6.73%。

3.3.3.4.1KNN 调参

KNN 有一个重要的参数 n_neighbors，即 KNN 中的 k 值，K 值较小就意味着整体模型变得复杂，容易发生过拟合；K 值较大就意味着整体模型变得简单，容易发生欠拟合。该值默认为 5，这里我们调节的就是这一个参数。因为 KNN 有较快的模型运行速度，因此在调参时我们选择的区间相对较大。

首先我们选择 0 到 20 内的每一个值作为参数值调参，其运行结果为：



可知当取值为 19 时有最高的 F1 值为 0.7363。显然，在 20 附近 F1 值呈现上升趋势，因此我们继续在 20 右侧区间内画学习曲线，经过多次尝试我们最终确定最优的参数值为 35，此时 F1 值为 0.8369，提高了 27.7%。

两个模型最终的优化结果如下：

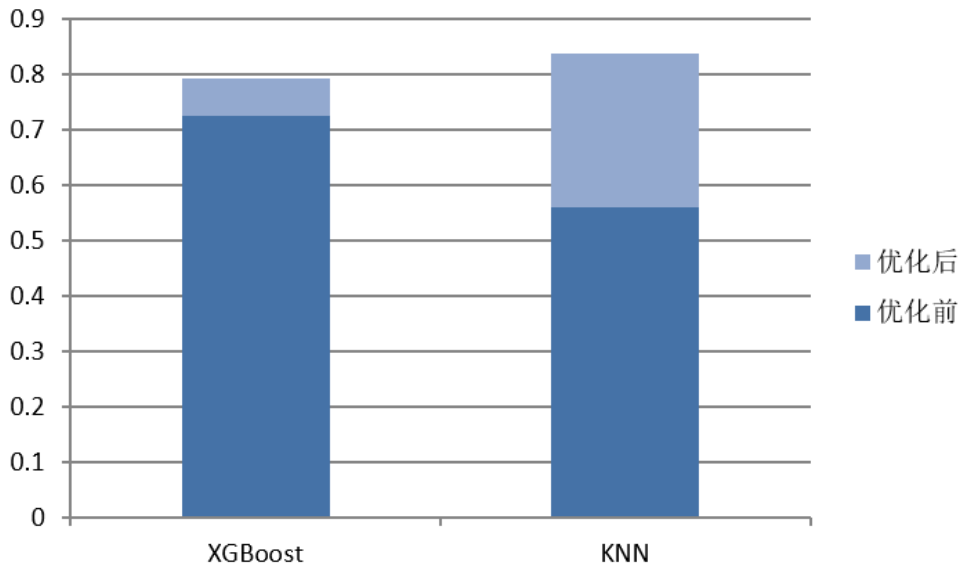


图 18 优化前后模型 F1 值对比图

综上，传统机器学习方法所得最优的 F1 值为 0.8369

3.4 深度学习建模

3.4.1 数据进一步处理

步骤如下：

3.4.1.1 提取关键词

文本关键字提取是从文本中提取与这个文本意义最相关的词语，是一种很好的去除干扰同时降低句子向量维度的方法。在本模型中我们采取 TextRank 算法。TextRank 算法可以脱离语料库背景，对单个文本进行分析以提取该文本的关键字信息。这里调用 jieba 库中的 analyse.textrank()方法进行关键字提取。

该方法共有 4 个参数，sentence, topK, withWeight=False, allowPOS=('ns', 'n', 'vn', 'v'))，我们对其中的关键参数 topK 的取值进行说明，其余均设置为默认值。

考虑到在此问题中“留言详情”和“留言主题”对分类均具有较高价值，故我们对“留言详情” 和“留言主题”均进行关键词提取。

参数	参数值
Sentence	text
topK	10
withWeight	False
allowPOS	(‘ns’,’n’,’vn’,’v’)

留言详情部分在传统机器学习部分已经阐述过，这里不再赘述。

4.4.1．2 “留言主题”关键词提取参数 topK 的取值依据

我们以去除停用词后“留言主题”中的单词数为横轴，以数据编号为纵轴，绘制图像如下所示。

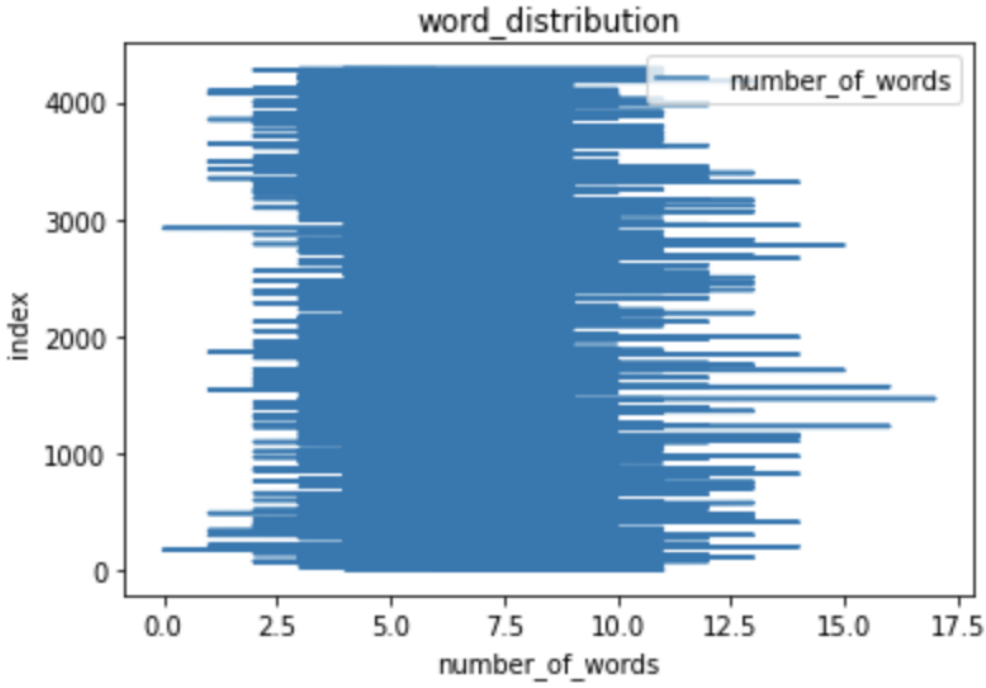


图 19 留言主题单词数分布图

由图可知，大部分“留言主题”单词数不超过 10，故我们设置参数 topK 为 10

analyse.textrank()方法参数如下：

表 3 留言主题 textrank 方法参数表

3.4.1.3 将文本转化为词向量

3.4.1.3.1 建立 token 字典，并使用 token 字典将“文字”转化为“数字列表”

我们使用 keras 库中 `preprocessing.text.Tokenizer` 函数建立文本的 token 字典，并使用 token 字典将“文字”转化为“数字列表”

3.4.1.3.2 对文本数据进行“截长补短”

为了能够正常使用 Embedding 层，我们将对“留言详情”中的每一行文本进行“截长补短”，即使用 keras.preprocessing.sequence.pad_sequences 函数，对于长度不足 maxlen 个词的句子，在前面补 0；对于长度超过 maxlen 个词的句子，从前面截断，只保留 maxlen

个词。

3.4.1.3.3 留言详情“截长补短”的长度参数 maxlen 取值依据

在本节中，我们以“留言详情”中提取关键字后的单词数为横轴，以数据编号为纵轴，绘制图像如下所示。由下图可知，提取关键词后大部分单词数不超过 250。我们确定参数 maxlen 为 250。

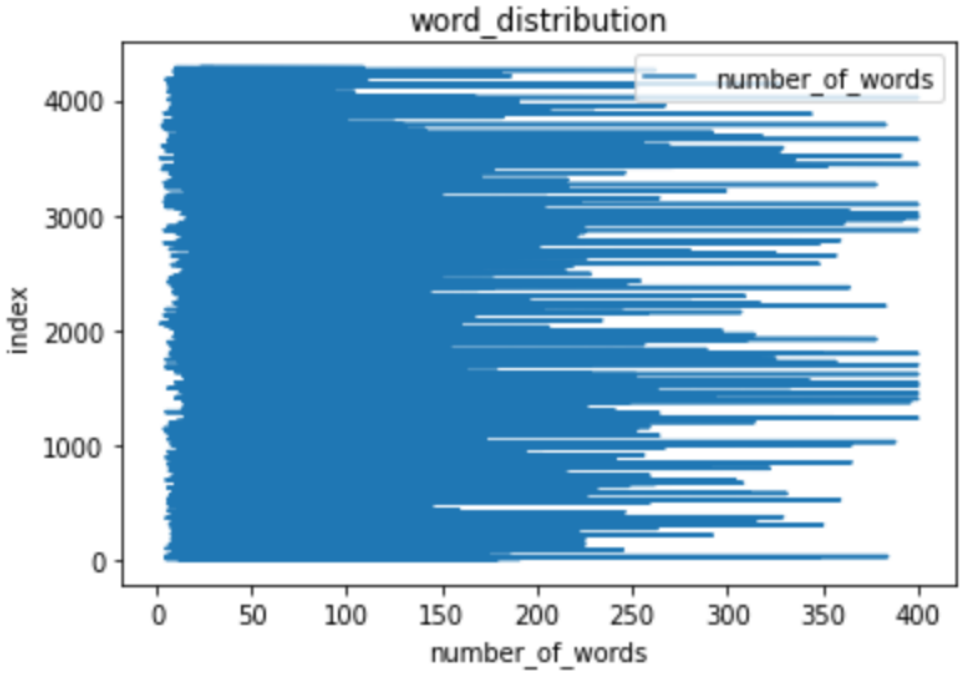


图 20 提取关键词后“留言详情”单词数分布图

3.4.1.4 留言主题“截长补短”的长度参数 maxlen 取值依据

在本节中，我们以“留言详情”中提取关键字后的单词数为横轴，以数据编号为纵轴，绘制图像如下所示。由下图可知，提取关键词后大部分单词数不超过 8。我们确定参数 maxlen 为 8。

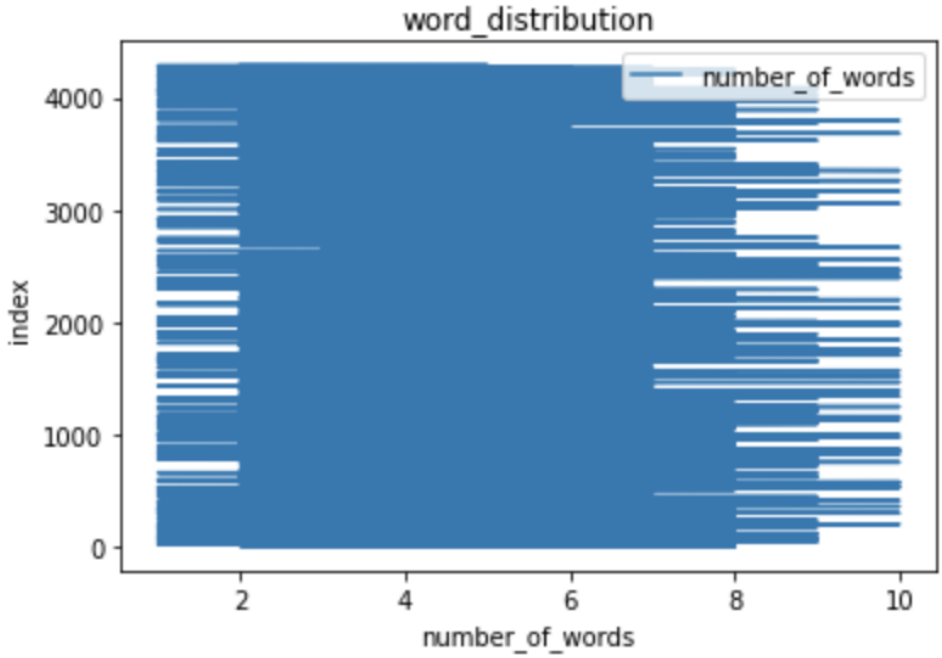


图 21 提取关键词后“留言主题”单词数分布图

3.4.1.5.使用 Embedding 层将“数字列表”转化为"向量列表"

3.4.1.4.1Embedding 层的简介：

使用 Embedding 层可将“数字列表”转化为“向量列表”，利用神经网络做词的向量化，Embedding 可将词的向量训练放入到模型中，形成一个 end-to-end 结构，而这样 Embedding 层训练出来的向量可以更好的适应相应任务。

keras 中的 Embedding 层有两种词嵌入的方式，一个是使用预训练的方式，通过参数指定预训练词向量矩阵。还有一种方式是随机初始化，Embedding 在随机初始化方式下是一个全连接层，而后面得到的词表示是全连接层的权重参数。

3.4.1.4..2 不使用预训练的说明

在本题中，我们不使用预训练的方式，基于以下理由：

从实际结果上看，使用新闻、百度百科、小说数据等数据进行预训练 news_12g_baidubaike_20g_novel_90g_embedding_64.bin 的效果劣于直接用所给数据通过

Embedding 层训练我们的词向量。这可能是使用新闻、百度百科、小说数据等数训练的样本与我们题目的留言样本的差异性较大。同时考虑到使用预训练的词向量会增大时间复杂度和空间复杂度，故本题直接通过所给数据训练词向量。

3.4.2 模型构建：

本模型使用“留言主题”和“留言详情”作为双输入的依据如下：

考虑到“留言主题”部分较为浓缩，可能对“一级分类”难以较为完整的反应；同时考虑到“留言详情”部分较为冗长，可能对“留言类别”难以有集中的反应。

综上两种情况，我们采用以“留言主题”和“留言详情”为双输入，“一级分类”为单输出的模型架构。

同时经过模型实践，仅仅只考虑“留言主题”，或者“留言详情” ，那么最终的结果不如将“留言主题”和“留言详情”做为双输入一同纳入考虑。

3.4.2.1 一种基于 Text CNN 的双输入单输出分类模型

3.4.2.1.1 对 Text CNN 的简介

CNN 是一类包含卷积计算且具有深度结构的前馈神经网络，为深度学习的代表算法之一，TextCNN 模型为 CNN 在文本分类中的应用，利用多个不同 size 的 kernel 来提取句子中的关键信息，其处理过程如下：

为不同尺寸的卷积核建立一个卷积层，因此将有多个 Feature Map。将得到的多个特征融合，连接到池化层，最大池化层只会输出最大值，对输入中的补 0 做过滤。全连接层的每个结点与上一层相连，用来把前面所取特征综合，每个神经元的激励函数一般采用 ReLU 函数。输出层的输入为全连接层的输出，经过 Softmax 层作为输出层，输出每个类别概率。

TextCNN 模型简单，计算量较少，训练速度相对较快，在短文本分类问题上有着不错的效果。故本题使用 TextCNN 作为建模模型。

考虑到本题最终希望建立集成模型，故同时建立 1 维和 2 维的 Text CNN 模型，但考虑到篇幅有限，我们在此介绍 1 维 Text CNN，2 维 Text CNN 则放置于目录中。

3.4.2.1.2 模型架构示意图

考虑到“留言详情”和“留言主题”具有一定的相似性，故我们对每一个输入建立对称的模型。

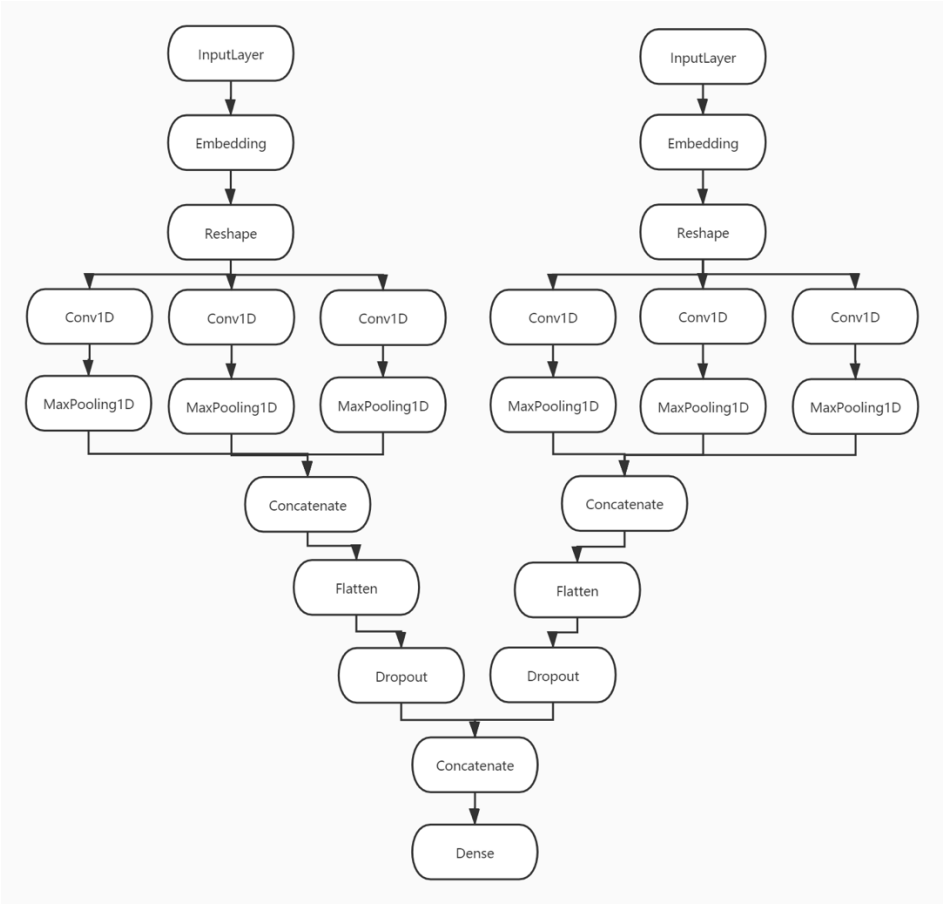


图 221 维 Text CNN 模型架构示意图

3.4.2.1.3 模型参数设置

经过我们对模型参数的反复调整，最终确定模型参数如下时，模型表现较优

Embedding 层以及基础参数设置如下：

参数	参数值
vocabulary_size	5000
sequence_length	400
sequence_length1	10
embedding_dim	256

filter_sizes	[3,4,5]
num_filters	512
drop	0.5
epochs	20
batch_size	30

（1 维 Text CNN 模型参数设置表）

卷积层参数设置如下：

（1 维 Text CNN 池化层参数设置表）

表 4

参数	参数值
Filters	512
kernel_size1	filter_sizes[0]
kernel_size2	filter_sizes[1]
kernel_size3	filter_sizes[2]
padding	‘valid’
kernel_initializer	‘normal’
activation	‘relu’

（1 维 Text CNN 卷积层参数设置表）

池化层参数设置如下：

参数	参数值
pool_size	2
strides	1
padding	‘valid’

3.4.2.1.4 训练与结果呈现：

我们用 sklearn 库中的 model_selection.train_test_split 将预处理后的数据划分为训练集和测试集，其中参数 test_size = 0.2

下图是随着 Epoch 增加，模型训练集和测试集准确率上生的曲线图，最终模型测试集准确率稳定在 0.8556。

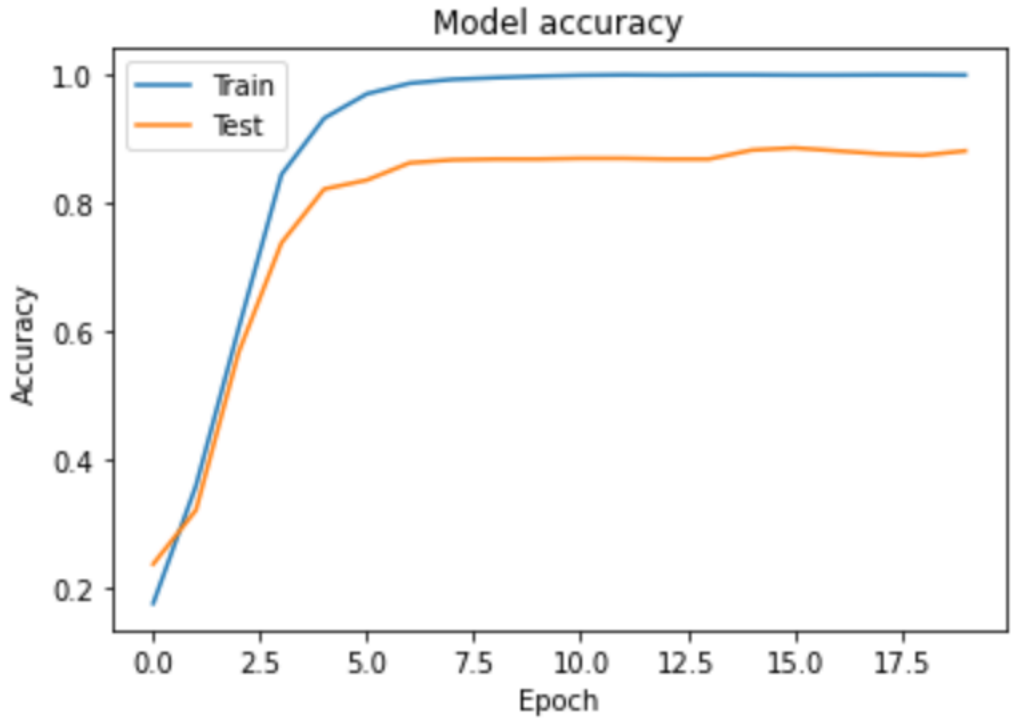


图 231 维 Text CNN 训练和测试准确率随 Epoch 变化图

下图是模型分类结果所对应的混淆矩阵

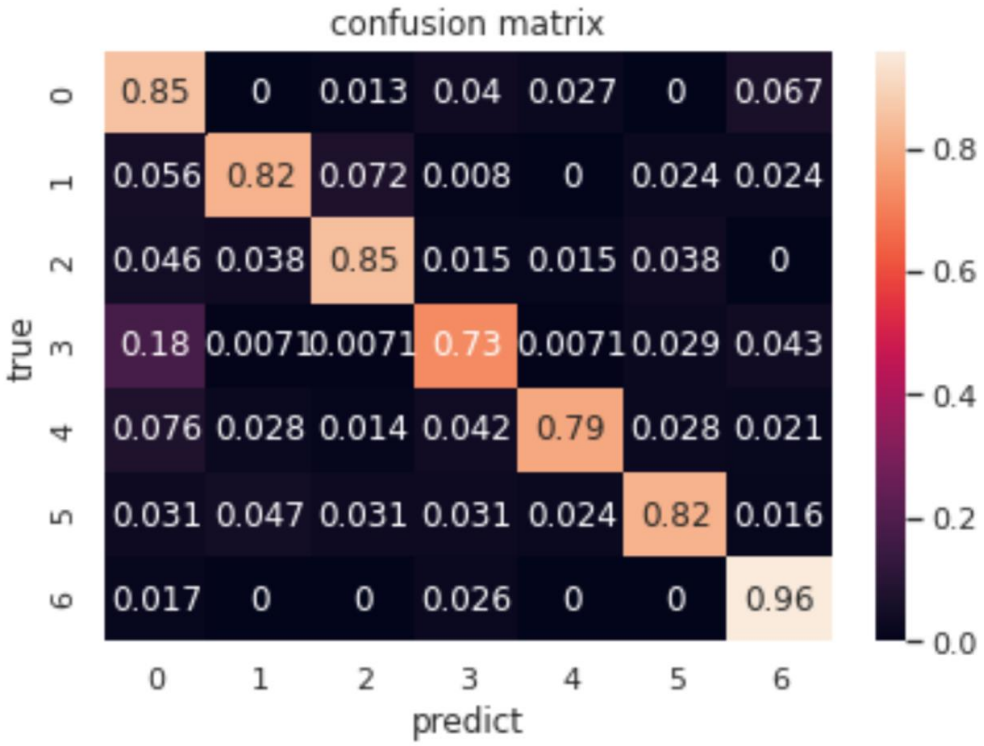


图 241 维 Text CNN 分类结果的混淆矩阵图

此时该混淆矩阵所对应的 F1 值为 0.8554。

为进一步验证模型构建的有效性，我们使用 10 折交叉验证得到该模型的 10 个 F1 值如下表所示：

[0.845518144835201,（变成 4 位）

0.8573182035428955,

0.8563116664213964,

0.8480020011009043,

0.8566989354326572,

0.855897174565607,

0.8558446777292664,

0.8551264751251497,

0.8547384214787459,

0.8524635691845721]

由表可知，F1 数值波动不大，对 10 个 F1 值求平均，我们得到最终的 F1 值：0.8538

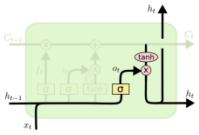
3.4.2. 2 一种基于 LSTM 的双输入单输出分类模型

3. 4. 2. 2.1 LSTM 的介绍

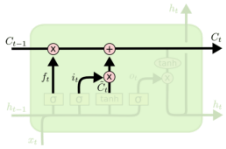
LSTM 是长短时间记忆网络，是一种特殊的 RNN，能有有效解决深度学习中 RNN 梯度消失的问题。其模型由专门的记忆储存单元组成，通过遗忘门、输入门和输出门来控制各记忆储存单位的状态，这些结果实现 LSTM 遗忘或增加信息能力。

遗忘门控制旧信息的遗忘，以备储存新的信息。

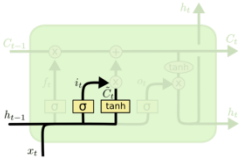
输入门负责处理当前序列位置的输入，判断将何信息保存到细胞中，进行细胞状态更新。



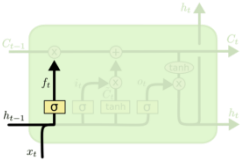
$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$



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



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



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

输出层该最后决定输出什么了，且输出值跟细胞状态有关。

考虑到 TextCNN 和 LSTM 均可以有效地对模型进行分类，同时为了使集成模型具有更好的表现能力，我们应该更多地去考虑模型类别的多样性。考虑到 LSTM 模型的时间复杂度较高，故我们在此建立简单但效果较好的 LSTM 模型。

3.4.2. 2.2 模型架构示意图

同样地我们采取对称式的架构：

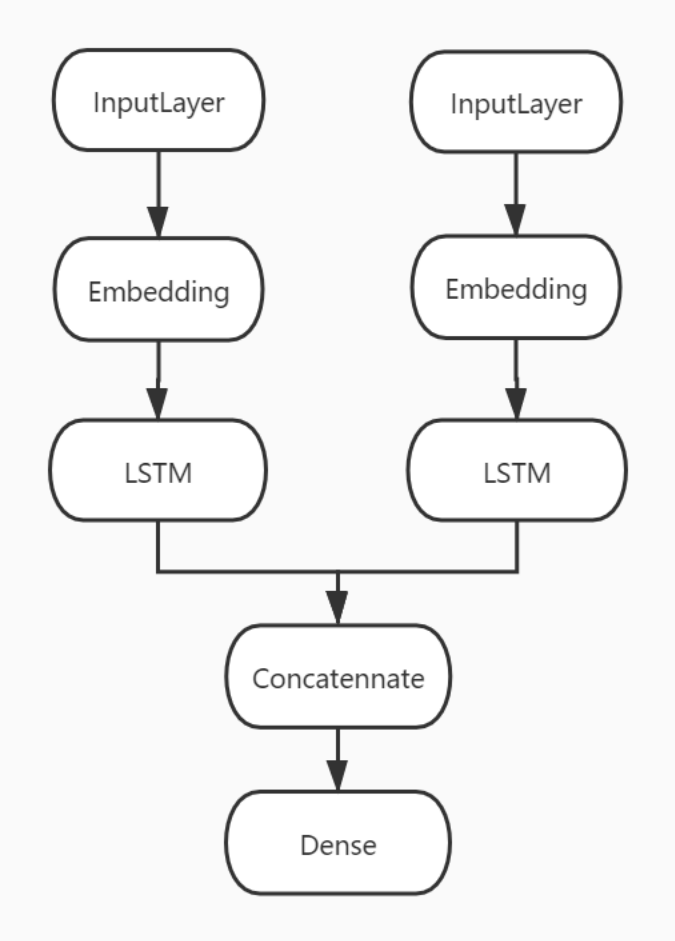


图 25LSTM 模型架构示意图

3.4.2. 2.3.模型参数设置

经过我们对模型参数的反复调整，最终确定模型参数如下时模型表现较优

参数	参数值
input_dim	5000
output_dim	64
units	32
activation	tanh

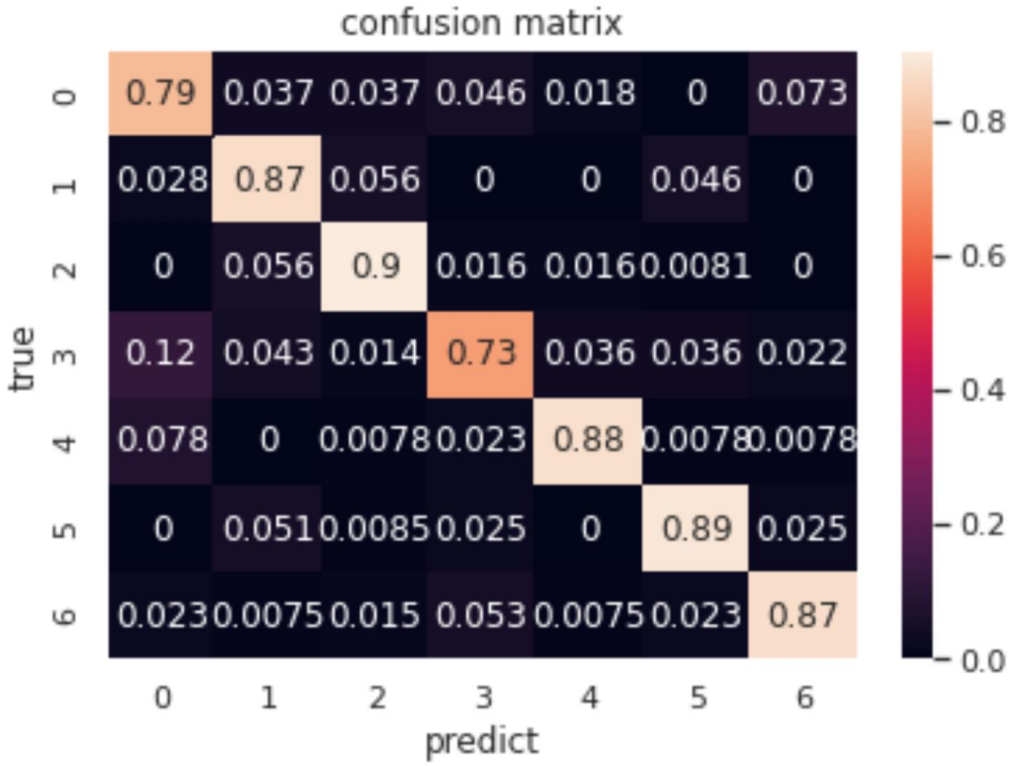
LSTM 模型参数设置

3.4.2.2.4 训练与结果呈现：

我们用 sklearn 库中的 model_selection.train_test_split 将预处理后的数据划分为训练集和测试集，其中参数 test_size = 0.2

下图是随着 Epoch 增加，模型训练集和测试集准确率上生的曲线图，最终模型测试集准确率稳定在 0.8553。

下图是模型分类结果所对应的混淆矩阵



（LSTM 分类对应的混淆矩阵）

此时模型所对应的 F1 值为：0.8561

为进一步验证模型构建的有效性，我们使用 10 折交叉验证得到该模型的 10 个 F1 值如下表所示：

[0.8653338886705303,
0.8660564732535718,
0.84985413809948,
0.8629203836486039,
0.8552726151942864,
0.860010912273954,

0.8526184535909259,

0.8576410876042795,

0.8611688581216179,

0.8594168648252576]

由表可知，F1 数值波动不大，对 10 个 F1 值求平均，我们得到最终的 F1 值：0.8591

3.4.2.3 一种基于 TextCNN 和 LSTM 的集成学习模型

3.4.2.3.1 各学习器的权重比例说明

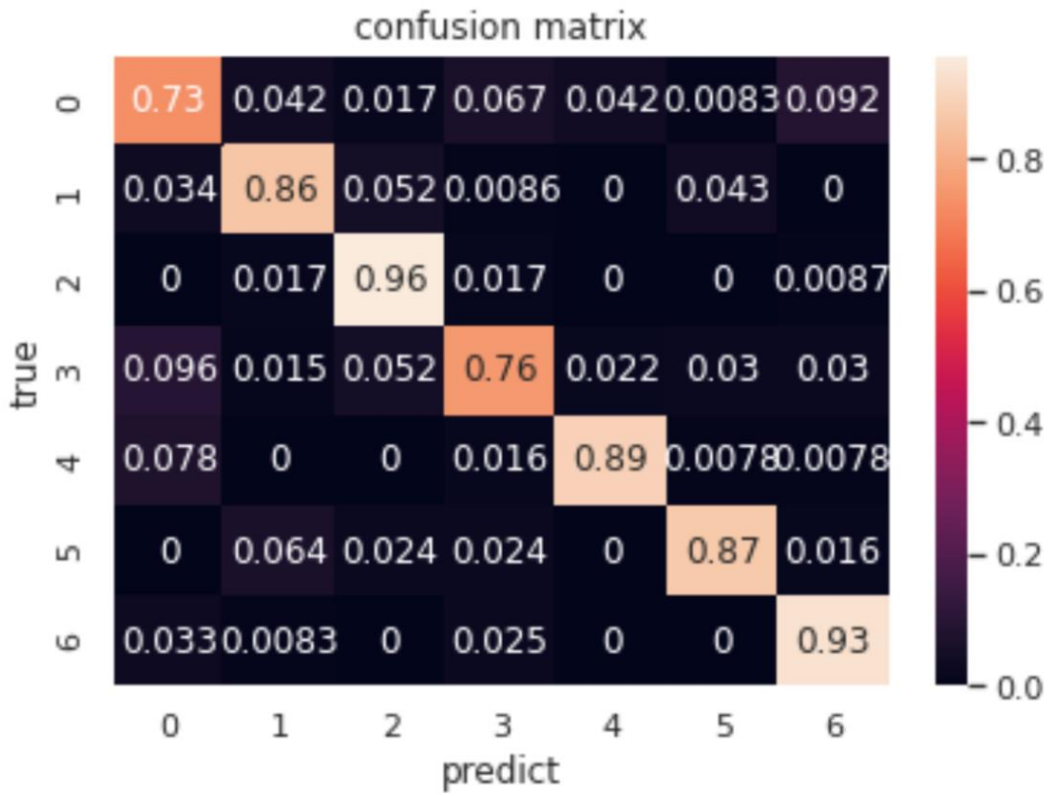
本题集成模型将使用 3 个分类器，分别是 1 维的 Text CNN，2 维的 Text CNN，以及 LSTM。鉴于 3 个模型的表现均较优且 F1 值近似，故我们对这 3 个模型权值进行均分的措施，即设置权重分别为 0.34，0.33，0.33

我们对每一个分类器最后的预测结果，均乘以相对应的权值，再求和，以得到最终的预测结果。

4.4.2.3.2 训练与结果呈现

我们用 sklearn 库中的 model_selection.train_test_split 将预处理后的数据划分为训练集和测试集，其中参数 test_size = 0.2

一次分类的混淆矩阵如下：



（集成模型分类对应的混淆矩阵）

此时 F1 值为 0.8913

为进一步验证模型构建的有效性，我们使用 10 折交叉验证得到该模型的 10 个 F1 值如下表所示：

```
[0.9003387746276573,
 0.9009761788220699,
 0.8926432026317572,
 0.8961608004824038,
 0.8835127539011416,
 0.894303020430469,
 0.8873733422918194,
 0.8910082976143541,
 0.8956452354224241,
 0.9002585874901007]
```

由表可知，F1 数值波动不大，对 10 个 F1 值求平均，我们得到最终的 F1 值：0.8942

从 F1 值可以看出，我们的集成学习模型是有效的，且对比单个分类器，提升了 5%左右

的数值，最终达到一个较好的水平。

3.3 传统机器学习建模与深度学习建模对比

传统机器学习模型的 F1 值最优能够达到 0.8369，深度学习模型最优能够达到 0.8942，较传统机器学习模型增加了大约 6%。可见对于问题一，深度学习模型更加具有优势。

4 问题二 热点问题挖掘

4.1 命名实体识别简介

命名实体识别简称 NER，又称“专名识别”，是自然语言处理的重要技术，是应用领域的重要工具，其主要包括三方面的方法：基于规则的命名实体识别方法、基于统计的命名实体识别方法和基于深度神经网络的命名实体识别方法。

在本题中，我们将使用 jieba 的 analyse. Textrank（）中 allowPOS 参数来对文本进行命名实体识别。

4.2 数据预处理

考虑到目标为挖掘特定人群和特定地点的热点问题，我们将使用对文本内容概括较好的“留言主题”作为文本数据。

4.2.1 文本简单清洗

利用正则表达式将文本中的中英文标点符号去除。

4.2.2 停用词去除

将“的”，“吗”等停用词去除。

4.2.2 命名实体识别处理

我们利用 jieba 的 analyse. Textrank 方法，将参数 allowPOS 设置为('ns','n'),其中'ns'表示地名，'n'表示名词。以此来将文本中的地名与名词筛选出来，有利于后续对"特定人群"和"特定地点"的挖掘。

4.3 利用 Word2Vec 将文本转化为词向量

4. 3.1Word2Vec 简介

对自然语言的建模过程是一个逐渐优化的过程，Word2vec 是深度学习领域能快速高效形成词向量的工具，Word2vec 通过对词语的上下文以及词语与上下文的语义关系进行构建模型，将词语映射到一个抽象的低维空间，生成所需的词向量。

Word2vec 包括 CBOW 和 Skip-gram 两种训练模型，其直接目的就是为了得到高品质的词向量，且简化训练步骤优化合成方式，直接降低运算复杂度，两种模型都包括输入层、投影层、输出层。

本题将采用 CBOW 的训练模型，用于生成词向量。

5． 3.2 将文本转化为词向量

考虑到附件三有 4326 条数据，但是一个热点问题可能所涉及的数据量可能在 10 条左右。所以我们需要利用"非热点问题"所含词语并不高频这一特点对数据进行筛选。

经过反复调参优化，我们最终确定 Word2Vec 中的最小词频为 20。

其他参数如下 size = 300,window=2

通过这一方式，我们过滤了 1738 条数据，最终剩下：2588 条数据

4.4 K-means 聚类算法简介

K-Means 是一种迭代求解的聚类分析算法，其算法实质是通过计算不同样本间的距离来判断他们的相近关系，并将相近部分放到同一类别中。

K-Means 算法取定 k 类与 k 个点作为初始质心，然后不断将各个点指派到最近质心，重新计算各类的均值点并设为新的质点，最终使得各数据点到所属类别质心的距离平方和最小。

最小化损失函数为：

E= ∑_(i=1)^k▒∑▒〖|X-(X_i) ¯|〗

其中 k 为总类别数，(X_i) ¯为质心。

算法流程如下：

1.选定 k 个数据点作为初始聚类质心（c1,c2……ck）

2.对每个数据点 xi 找到最近的质心 ci，并将其分配到所属类

3.重新计算每个簇的质心

4.直至达到稳定，质心不再变化，否则跳转至 2

4.5 热点问题发现

4.5.1 热点问题聚类

我们使用 sklearn 库中的 cluster.KMeans 函数，指定类的个数为 400 个，其他参数如下 max_iter=3000, n_init=40, init='k-means++',n_jobs=-1

建立模型，最终输出聚类结果。

4.5.2 热度评价指标

留言信息的热度可由一时间段内群众对某些问题表达出的热切程度与关注度来表示，针对某一时段内反映特定地点或特定人群问题的留言归类后的信息，我们首先抓取“反对数”和“点赞数”，浏览者在阅读留言后，会对引发共鸣或感到抵触的内容产生一定反应，经探究，我们将反对数和点赞数进行加总，数据呈现四位数与个位数甚至为零的较大差异，因此我们将采用五分制，具体操作为：大于 100 得分 5；50-100 得分 4；15-50 得分 3；5-15 得分 2；1-5 得分 1；0 得分 0，由于该指标能直接反映大范围人群的意见与想法，且表达意见的便捷度较大，我们将其设定 0.5 的权重。

其次，各大类的实际留言数量也是一个量化指标，我们将计算各大类的留言条数，删去重复留言。

删去重复留言是考虑到会有用户针对同一问题进行多次反馈，如留言用户 A00031618，该用户发布了关于“A 市国王陵国家考古遗址公园”的管理，环境问题等 12 条留言。考虑到我们的热点问题是针对于群体，而非个体，所以我们对于在某一个热点问题之内的相同用户

的多条留言只计作 1 条。

通过观察可知，每大类的评论数量大致分布在 1-10，均分到五分制，我们并赋予该指标 0.4 的权重。

最后捕捉“留言时间”，计算聚类后每一类留言的最大时间差，时间差越大反映该问题在该段时间内持续发酵，引起群众的大范围关注，属于重要度较大的热点问题，我们将根据五分制酌情给分，然后由于某些问题能即刻解决或仅带来小段时间的热度，此度量指标存在一定的主观性，我们最终选择赋予 0.1 的权重。

5 问题三 质量指标

5.1 设计思路

5.1.1 指标概念

可解释性可理解我们在对某件事情有了解或解释的需求时，我们可以获得相匹配的信息，并且能清晰地理解与接受它。随着互联网的快速发展与新兴产业的不断出现，大量知识信息被公开，我们会被匹配到各式各样信息，在我们选择性抓取信息时，会因为语言文字涉及的知识体系超出个人理解范围、文本本身通顺度不佳、文字天然接纳度等原因被拒之门外。良好的文字接受度能提升自身质量，给读者愉悦之感。因此相关部门对留言的答复意见也遵循此道理，较高的可解释性能有效提升反馈质量，轻松获得我们所需要的足够可以理解的信息。

及时性牢牢把握时间概念，回馈的不及时有可能会造成不良后果，例如群众因难迟迟无法解决、长时间得不到回馈而持冷漠的态度进而降低参与社会政治生活的积极性、随时间推移加深群众顾虑与担忧等。由此可见，评价回馈质量的高低不仅取决于内容本身，答复的及时性也是不可或缺的一部分，对能够解决的，应及时研究解决；对暂不能解决的，应合理给予反馈，创造条件逐步解决；有悖于法律、政策规定和实际情况的，应向需求者及时说明。

完整性可从广度与深度着手，广度偏向于信息覆盖面的大小，往往多方面多角度的答复更能让群众捕捉到有效信息，且有更多选择权，这无疑提高反馈内容的质量；深度是衡量触及事物本质的程度的指标，相关部门能从更深层次的方面给出答复，有助于群众更好的理解政治信息，有助于社会政治生活的发展。

相关性在这里指每条数据的“答复意见”与“留言详情”之间的相关程度，这一指标是

为了避免出现答非所问的情况。

5.1.2 标准设定

因为可解释性这一指标具有很大的主观性，因此我们评价该指标的标准也具有很强的主观性，但是根据我们的经验，当答复内容包含较多的法律条例等内容时，一般可解释性较低。另外，对于一些答复过于简单或者答复无实质性内容的答复，我们也认为这是一种可解释性较低的答复。因此我们人为的将可解释性分为非常可解释—比较可解释—一般—比较难解释—非常难解释的五级评价标准，分别对应 5-1 分。

对于及时性，浏览答复意见表，我们迅速捕捉到“留言时间”与“答复时间”两个变量，计算时间间隔，可得到在 2817 条有效数据中，3 天内回复的数据共 597 条，反馈迅速，得分 5；3-7 天内回复的数据共 1124 条，反馈较及时，得分 4；7-15 天内回复的数据共 1789 条，反馈及时性一般，得分 3；15-31 天内回复的数据共 2395 条，反馈较慢，得分 2；31 天以上回复的数据共 2817 条，反馈拖延，得分 1。从上述可知，共分 5 阶段来度量及时性指标。

针对完整性，浏览“答复意见”，我们发现所有的意见可以被分为 5 的层次：已经帮助解决问题，得分 5；尚未解决问题但是明确给出解决问题的建议方案，得分 4；明确告知来访者已经问题移交相关部门，得分 3；告知来访者可以解决该问题的特定部门，表明让来访者自行移交问题，得分 2；并没有给出任何实质性的回答，得分 1。

在相关性指标上我们计算出每条数据的“答复意见”与“留言详情”之间的文本相似度，这里我们调用 Python 已有的 spacy 库应选择它自带的中文模型“zh_core_web_sm”来计算两个文本之间的相似度，并将这一相似度乘以 5 映射到[0,5]区间内。同时为了降低评分体系的复杂度，我们将最终得分进行四舍五入，得到一个相应的整数值。

另外，考虑到不同的指标对于评价体系来说有不同的重要程度，显然，对于群众来说答复的完整性是最重要的。另外，答复的及时性对于解决群众问题也是极为重要的。相对来说。答复的相关性和可解释性重要度就比较低，其中，可解释性因为具有较强的主观性以及诸多不可抗力因素使得语言表述不得不变得难懂，因此我们赋予可解释性最低的权重。

最终，我们赋予完整性 0.4 的权重，及时性 0.3 的权重，相关性 0.2 的权重，可解释性 0.1 的权重。

评价指标	权重	评价标准	分数
完整性	0.4	明确已经落实解决群众反映问题，或回答群众疑问	5
		给出多条或单条解决方案	4
		回复群众已将问题移至某部门	3
		告知群众自己将问题移交至相应单位	2
		仅回复已收悉或无回答	1
及时性	0.3	3 天内回复	5
		3-7 天内回复	4
		7-15 天内回复	3
		15-31 天内回复	2
		31 天以上回复	1
相关性	0.2	文本相似度*5	5
			4
			3
			2
			1
可解释性	0.1	非常可解释	5
		比较可解释	4
		一般	3
		比较难解释	2
		非常难解释	1

表 5 答复意见质量的评价指标表

5.2 结果展示

综上，我们部分数据得分情况如下：

留言编号	留言详情	答复意见	时间差	及时性得分	完整性得分	可解释性得分	相关性得分	总得分
16542	2公里毛路坑坑洼洼，	年农村公路建设指标有限。建议：1、自	0	5	4	5	3.2	4.24
32712	畅”专项工作之际，我	的留言收悉，已转有关部门调查核处，如	0	5	3	5	2.6	3.72
47786	也不知道本市有没有具	等职业学校招生工作方案，并在各级教育	0	5	4	4	3	4.1
17715	县级领导重视刑事案件	三是在一般的背街小巷和小区道路，目前	0	5	5	4	4.1	4.72
132270	力非常大，现在下河去	重点项目中包含城北体育场改造项目，目	0	5	5	3	2.1	4.22
93271	体说明是哪天，这样有	网友：您好！留言已收悉	5	4	1	1	2.1	2.12
103744	天了都没反应。还有这	内制作电子化档案资料，转入地收到转出	8	3	5	3	2.3	3.66
132131	两边，曾有一位外地司	、教育，要求个体经营者不得在公路上	243	1	5	4	3.3	3.36
159285	关于低保，残疾人补助，无视实际，落实	尔好。请向当地民政部门询问。2017年8月	1160	1	2	2	1.6	1.62

图 26 答复意见质量数据得分表

6 参考文献

[1] 李宗富，张向先.政务微信公众号服务质量评价指标体系构建及实证研究[D].吉林长春管理学院.2016

[2] https://github.com/fxsjy/jieba

[3]https://github.com/howl-anderson/Chinese_models_for_SpaCy/releases/tag/v2.2.X-0.1.0

[4] 王芳，翟丽娜. 我国地方政府门户网站 G2B 服务能力评价指标体系的构建[D]. 南开大学商学院信息资源管理系.2008