“智慧政务”中的文本挖掘应用

摘要

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。

对于问题 1，通过 Excel 处理，将附件 2 表分成两个表，得到信息不重复的‘附件 2_train’表和‘附件 2_test’表。在附件 2_train 表中，同时采用过采样和欠采样方法，使得各类之间的数据平衡，将其作为训练集，将‘附件 2_test’表作为验证集；再对训练集数据进行数据预处理，对‘留言详情’构建词汇表，使用字符级别表示；将分类目录固定，转换为词典表示，形如{类别: id}。在卷积神经网络 CNN 基础上构建 Text-CNN 模型，调用数据，激活模型对数据进行训练，得出训练集和验证集的精度和损失函数值，保存模型。构建好测试函数，当得到新数据进行测试时，将新数据集作为测试集调用测试函数，得出测试集的分类准确度、损失函数值、模型评价、混淆矩阵。

对于问题 2，通过 jieba 工具，对‘留言详情’分词，预处理；根据文本相似性，利用 Single -Pass 文本聚类算法，对‘留言详情’进行聚类；计数汇总统计，得出类数、每个类的留言数量、中心词等；并返回每个类中的每条留言的元数据行，得出对应的留言信息，并根据类内留言数量、时间、反对数和点赞数，构建热度评价指标，得出热度指数排名，取前五名。

对于问题 3，通过分析相关文本结构，构建评价方案。

关键词：分类 Text-CNN 中文分词 Single－Pass 聚类 TF-IDF 预测 文本相似性 热度

1 挖掘目标

本次挖掘目标是根据网络平台系统上的群众问政留言记录，利用自然语言处理和文本挖掘的方法解决一下问题：

（1）群众留言分类，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；根据附件 3 将某一时段内反映特定地点或特定人群问题留言进行归类定义合理的热度评价指标，并给出评价结果，根据一定的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”；根据一定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。

2 分析方法与过程

2.1 问题 1 分析方法与过程

2.1.1 数据预处理

2.1.1.1 抽取数据

利用 Excel 表格，获得附件 2 中各分类情况，包括 7 个分类标签，每类数据量不相同，最多的是‘城乡建设’类 2009 条留言信息，最少的是‘交通运输’类 613 条，可见数据不平衡。根据各类数据情况，从附件 2 表中先分别抽取 7 类数据自身 80%左右的数据，各类剩余的数据保存在名为“留言分类_验证集.xlsx”的 Excel 表中。

2.1.1.2 数据采样

对抽出的数据同时采用过采样和欠采样方法，在抽出的数据中，超过 1000 条数据的类进行欠采样，不超过 1000 条的类进行过采样，最后形成 7 类数据均为 1000 条留言信息，一共 7000 条的数据集，保存在名为“留言分类_训练集.xlsx”的 Excel 表中。

2.1 构建可运算数据

读取“留言分类_训练集.xlsx”，提取【留言详情】,【分类标签】文本数据。根据【留言详情】的文本内容构建词表并储存，拆分文本每个字符，形成字符级

词汇表，将所有文本构建为同一长度，转换为字典{词：ID}形式；形成分类目录，转换为字典{类别:ID}形式。将词 ID 表示的内容转换为文字；再将将数据集【留言详情】,【分类标签】从文字转换为固定长度的 ID 序列。

2.1.2 构建 TextCNN 模型

2.1.2.1 配置模型参数

根据多次调整，配置以下模型参数

参数表	
名称	参数
词向量维度	embedding_dim = 64
序列长度	seq_length = 600
类别数	num_classes = 7
卷积核数目	num_filters = 128
卷积核尺寸	kernel_size = 5
词汇表大小	vocab_size = 5000
学习率	learning_rate = 1e-3
全连接层神经元	hidden_dim = 128
保留比例	dropout_keep_prob = 0.5
每批训练大小	batch_size = 64
总迭代轮次	num_epochs = 15
每多少轮输出一次结	print_per_batch = 100
每多少轮存入 tensorboard	save_per_batch = 10

表 1：TextCNN 模型参数表

2.1.2.2TextCNN 模型网络结构

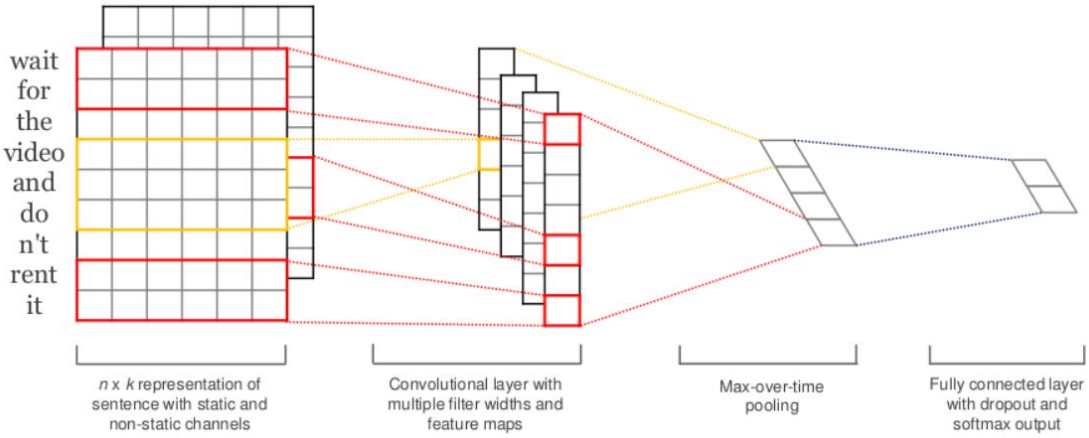

图 1：TextCNN 模型网络结构图

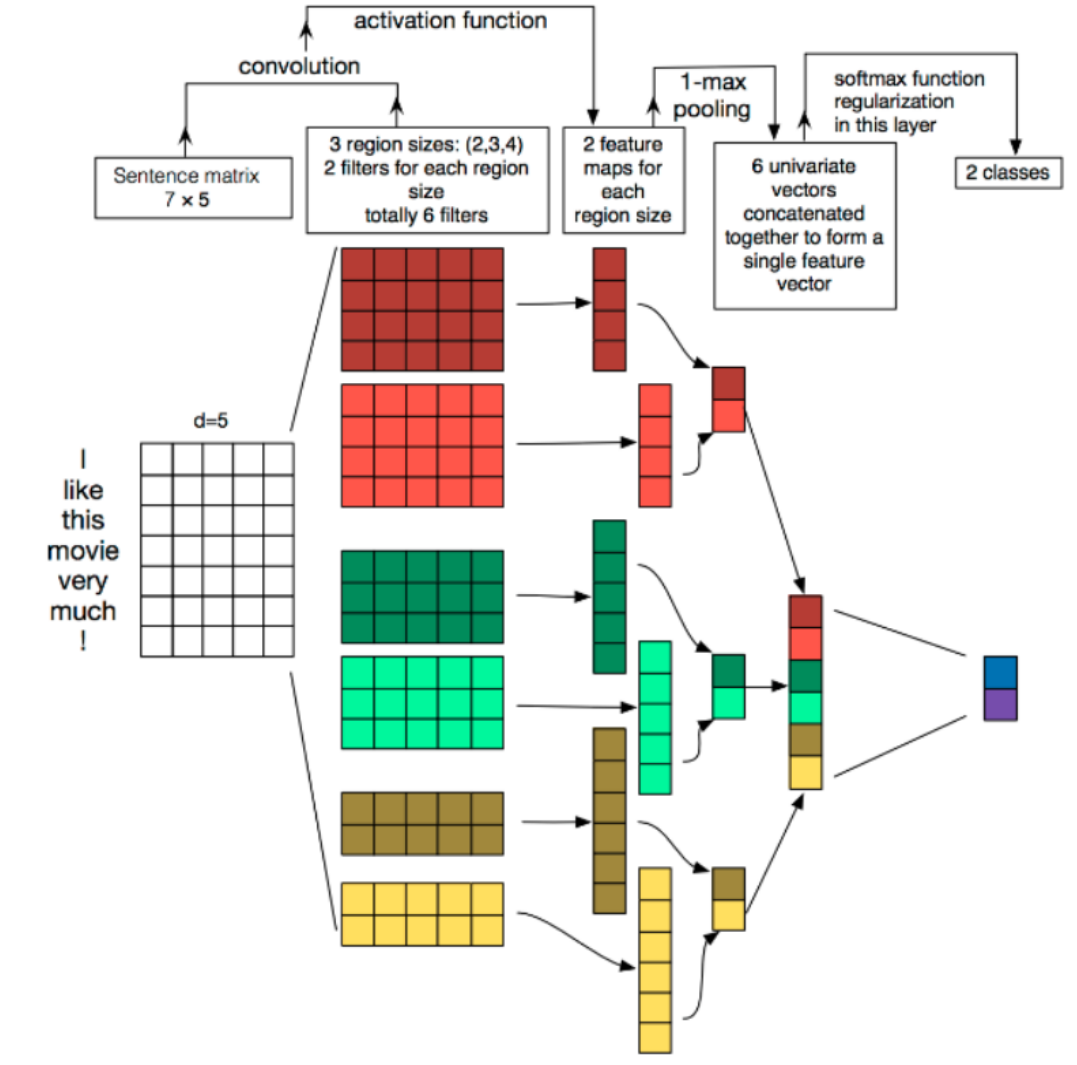

图 2：TextCNN 原理图

2.1.2.3TextCNN 流程

（1）在 embedding 层加载预训练词向量，映射构建向量
（2）卷积
（3）进入池化层，对特征做最大池化。
（4）全连接，进入分类器，分类。

2.2 问题 2 分析方法与过程

2.2.1 数据预处理

将【留言主题】和【留言详情】的内容另外合并为一个新的属性，保存在原 Excel 表中成为第七列【留言主题与详情】。检索表中点赞数和反对数的值，由于点赞数和反对数的值大部分为 0，若检索到为空或者非数值，用 0 填补。最终将处理好的数据保存为名为‘热点问题留言表.xlsx’的 Excel 表。

2.2.2 构建模型

2.2.2.1 模型流程

采用 Single-Pass 算法，根据余弦距离，获得文本相似度，对【留言详情】文本内容进行聚类，从而把反映同一问题的群众留言归为一类，获得类数，每类留言条数，返回留言信息明细等。定义热度系数，再根据留言相关信息获得各类问题热度，得到热点问题前五名。

2.2.2.2Single-Pass 算法流程

Single-Pass 算法处理文本时，以第一篇文档为种子，建立一个新类。当新进入文档，将其与已有类做相似度计算，将该文档加入到与它相似度最大的且大于一定阈值的主题中。若与所有已有的类相似度都小于阈值，则以该文档建立新的类。流程如下：
（1）以第一篇文档为种子，建立一个类；
（2）将文档 A 进行向量化；
（3)将文档 A 与已有的所有类进行相似度计算，采用余弦距离(其他距离也可)；
（4）找出与文档 A 具有最大相似度的类 B；
（5）如果相似度值大于阈值 theta，则把文档 A 加入到最大相似度的类 B 中，转至 7；
（6）如果相似度值小于阈值 theta， 则以文档 A 创建新的类；
（7）聚类结束，下一篇文档进入，直至所有文档归类完成 。

2.2.2.3 定义热度指数

要解决的是找出某一时段内群众集中反映的某一问题，并评价热度。该文本数据集为群众留言，属于社情民意相关的文本数据。评价热度，首先最重要是量，这个量是指这一问题反映的人数、关注这个社会问题的人数等，其次是时间变量。

对于量，可以获得的有同一问题留言数量（可理解为问题反映的人数）P，点赞数 M,P 和 M 越大，说明认同这一问题的人越多，可以使热度提高，可以直观理解为和热度正相关；还有反对数 N，虽然对问题提出反对，可能是对问题的质疑或者否定，但这也是对这一问题的关注，同时说明这一社会问题可能存在争议，为了防止舆论扩大，更加需要相关部门尽快核实，所以反对数 N 与热度也正相关，而且 P、M、N 同等重要。

对于时间，热度是随时间衰减的，时间与热度负相关，从数据中可以获得每类问题中最早留言与最晚留言时间间隔 T，这里以“天”’作为单位变量，根据牛顿冷却定律，时间 T 的衰减速率越来越慢，这里定义$[(\log T)^{-1}]$。如果生活中产生一些社会问题，有时候难以立即发现，民意也不一定会立即反映给相关部门，鉴于一周工作日为 7 天，定义$[(\log_7 T)^{-1}]$。

$$\text{热度指数：} Y = k(P+M+N)\ [(\log_7 T)^{-1}]$$

（注：P 留言数量、M 点赞数、N 反对数，k 为相关系数，这里定义为 1）

2.3 问题 3 的分析

方案构思

导入附件 4，提取【留言主题】与【答复意见】的文本内容，但要注意保留其一一对应的关系。对文本进行分词（可用 jieba 分词等分词工具），标注词性，提取文本实体词。利用 TF-IDF 算法（不限于），基于 LDA 的关键词提取方法等提取留言主题和答复意见的关键词。关键词转化为矩阵的形式，比较相似性，相似性越大，答复与留言相关性越大、完整性越高，评价得分也就越高。

由于个人技能与时间的关系，暂时未能实现并应用。

3 结果分析

1 问题 1 结果分析

对于验证集的精度达到 87%上下，损失函数值在 0.4 左右，结果还是可以接受。在测试集上的结果就还不得而知。

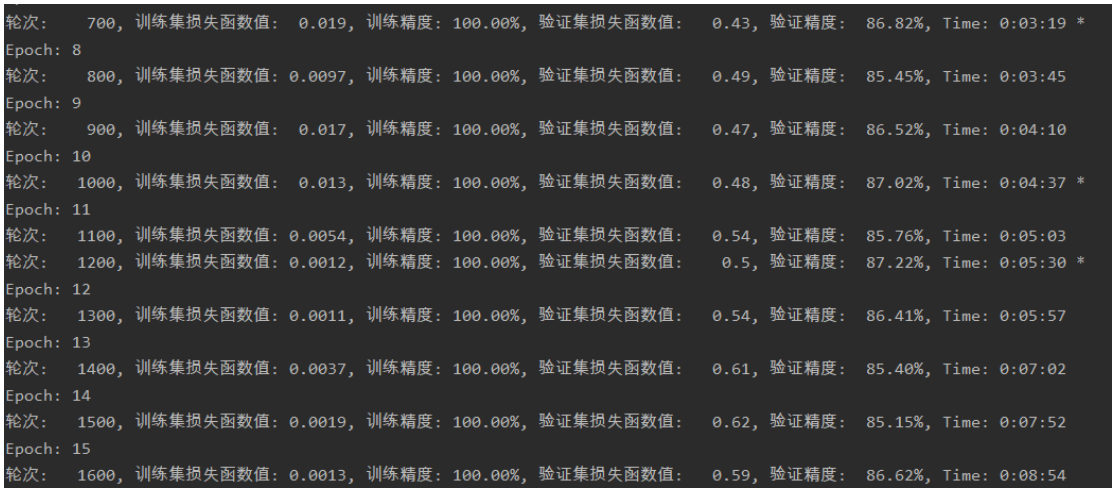

图 3:TextCNN 部分训练结果图

2 问题 2 结果分析

对热点问题留言表进行聚类，得到类数 2948，每类留言数量，从计算热度指数得出的结果看，热度指数前五名都是留言数比较多的，前两名名点赞数和反对数也较多。同一类中时间跨度相差不大，说明这里留言数量 P、点赞数 M、反对数 N,对热度指数影响较大，结果比较符合实际。

4 结论

利用自然语言处理技术建立智慧政务系统,对提升政府的管理水平和施政效率具有极大的推动作用，同时，自然语言处理技术也是一个课题与难题，可以应用在社会上的各行各业。本文主要运用文本分词，CNN 神经网络，Single-Pass 一遍聚类算法，进行文本分类或文本聚类分析，对网络平台的群众留言进行系统性处理，挖掘出热点话题，给出量化相关部门对群众留言的答复意见的方案。

由结果可见，文本分析对留言分类效果不错，这对政府的管理水平和施政效率的提高有推动的作用，甚至可以更细化地对留言进行二、三级分类。对于热点问题挖掘，网络平台每增加一条留言，可以对其进行处理，可以实时更新热点问题，相关部门可以及时了解社情民意，有助于相关部门进行有针对性地处理，提升服务效率。

文本挖掘分析技术也很值得研究，对文本处理技术还可以进一步提升，并应用于社会中的各个行业领域。这对科技进步，社会发展起到积极的作用。

5 参考文献

[1] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.

[2] 涂文博,袁贞明,俞凯.针对文本分类的神经网络模型[J].计算机系统应用,2019,28(07):145-150.
[3] 冯梦莹,李红.文本卷积神经网络模型在短文本多分类中的应用[J].金融科技时代,2020(01):38-42.
[4] 黄建一, 李建江, 王铮, & 方明哲. (2019). 基于上下文相似度矩阵的single -pass短文本聚类. 计算机科学, 46(04), 56-62.
[5] 冯梦莹,李红.文本卷积神经网络模型在短文本多分类中的应用[J].金融科技时代,2020(01):38-42.
[6]周茜. 融合word2vec和Single-Pass的微博话题检测方法研究[D].山东师范大学,2019.
[7]党燕,许志伟,刘利民,王宇,赵思远.基于 Single-Pass 算法的网络舆情文本增量聚类算法研究[J].内蒙古工业大学学报(自然科学版),2017,36(05):364-372.
[8]许峰,张雪芬,忻展红.基于深度神经网络模型的中文分词方案[J].哈尔滨工程大学学报,2019,40(09):1662-1666.
[9]曾小芹.基于 Python 的中文结巴分词技术实现[J].信息与电脑(理论版),2019,31(18):38-39+42.