“智慧政务”中的文本分析

摘要

本文旨在基于互联网收集的群众问政留言信息，及相关部门对部分群众留言的答复意见信息，通过训练词向量、深度学习算法与数据挖掘，对群众反映的问题进行分类。根据相关部门对留言的答复的相关性、完整性、可解释性的分析，可大致对答复意见的质量给出相应的评价模型。

针对问题一，首先利用 jieba 分词对留言内容进行断词处理，再用语料库训练 word2vec 词向量（使用 skip-gram 方法）。然后，建立实验数据中对应的词向量表格，用 CNN 分类器对留言详情进行分类处理，得到数据。最后，利用 F-Score 对该模型进行检测。

针对问题二，首先利用 jieba 分词对留言内容进行断词处理，再用语料库训练 word2vec 词向量（使用 skip-gram 方法）（特定时间，特定地点，发生的问题）。用神经网络模型进行训练，得到排名前 5 的热点问题，作品附件中的热点问题表.xls。排名前 5 的热点问题的详细内容，如热点问题留言明细表.xls。

针对问题三，可用 jieba 分词后，利用 word2vec 比较留言主题和答复意见文本的相似程度，来判断答复的质量。

关键词： jieba 分词；word2vec；卷积神经网络；K-means 聚类

目录

一、 问题分析

人工处理政务存在工作量大、效率低且出错率高等问题，随着微信、微博等网络问政平台逐步成为了解民意、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，建立基于自然语言处理技术的智慧政务系统，有助于提升政府的管理水平和施政效率。

问题所给的任务一要求根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。留言分类方便后续准确地将群众留言分派至相应的职能部门处理。在实际中分类中，要注意文本语义带来的词语交叉，以及数据不平衡带来的分类不准确的问题。

任务二要求根据附件 3 所给的数据，挖掘出某一时段内反应特定地点或特定人群问题的留言，对其热度进行评价，并给出评价结果。

任务三要求根据附件 4 相关部门对留言的答复意见，建立答复的相关性、完整性、可解释性的答复质量的综合评价模型。

二、基本假设

1.假设所给的数据量能够反映一定程度的真实性。

2.假设去停用词后，剩下的词语词语能够反映语句的语义。

三、符号说明

变量	定义
$D=\{\omega_1,\cdots,\omega_T\}$	词语序列
m	上下文窗口大小
N	词表中的词语数量
v_ω	input 向量描述
v'_ω	output 向量描述
s	句矩阵
$\sigma(\cdot)$	非线性激活函数

W	卷积核
c	特征向量
A	精确率
P	查准率
R	查全率
TP	真正例
FP	假正例
TN	真反例
FN	假反例
n	标签类别的个数
P_i	第i类的查准率
R_i	第i类的查全率

四、任务一

4.1 jieba 分词

由于在做文本分析时，为了使机器能够快速、准确识别地文字语言，所以，我们需要先对文本进行分词处理。而 jieba 分词，则是能够精确的将文本切开的一种方式。将附件二中的留言详情进行分词后，得到一个语料库文档。

4.2 Word2vec 训练中文词向量

Word2vec 方法是基于预测的方法，本文采用 skip-gram 模型。该模型使用目标词语来预测它的上下文词语。

$$\omega_1, \omega_2, \cdots, \omega_{t-m}, \cdots, \omega_{t-1}, \omega_t, \omega_{t+1}, \cdots, \omega_{t+m}, \cdots, \omega_{T-1}, \omega_T$$

图 4.1 文本上下文窗口示意图

Skip-gram 方法是在给定目标词语，在一个滑动窗口范围内预测其上下文词语。如图 4.1 所示，给定词语序列 $D=\{\omega_1,\cdots,\omega_T\}$，Skip-gram 方法的目标是最大化公式中的概率函数

$$L(D)=\frac{1}{T}\sum_{t=1}^{T}\sum_{-m\le j\le m,j\neq 0}\log p\left(\omega_{t+j}\mid\omega_t\right)$$

其中，m 是上下文窗口的大小，m 值越大，准确度越高，训练时间也更久。

$p\left(\omega_{t+j}\mid\omega_t\right)$ 是公式中的 softmax 函数定义：

$$p\left(\omega_O\mid\omega_I\right)=\frac{\exp\left({v'_{\omega_O}}^{T}v_{\omega_I}\right)}{\sum_{\omega=1}^{N}\exp\left({v'_{\omega_O}}^{T}v_{\omega_I}\right)}$$

其中 v_ω 和 v'_ω 分别是词语 ω 的“input”和“output”向量描述。N 指词表中的词语数量。

通过该方法可以得到训练好的词向量。

4.3 基于卷积神经网络的留言详情分类模型

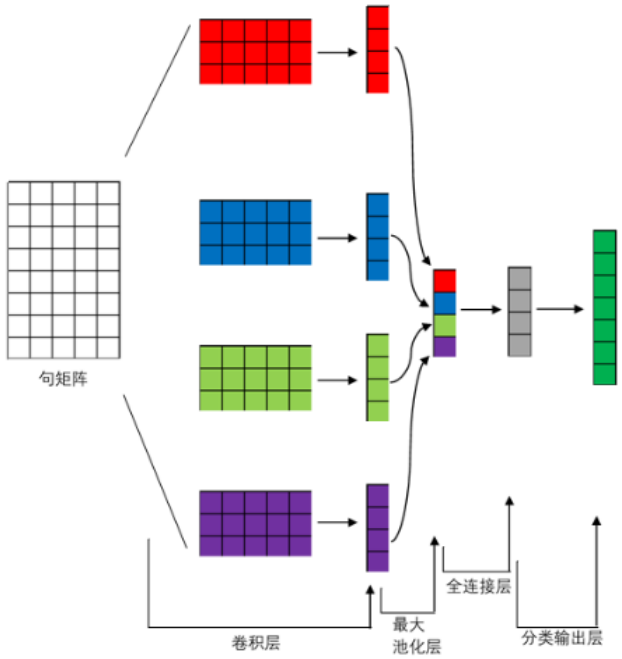

图 4.2 CNN 分类器架构图

卷积神经病网络由卷积层和池化层两大部分。

卷积层中令 $\omega_i \in R^k$ 表示一个句子中的第 i 个词语的 k 维词向量，一个长度为 n 的句子用句矩阵 $s \in R^{n \times k}$ 表示，其中第 i 行对应词向量 ω_i。令 $\omega_{i:i+m}$ 表示由词向量 $\omega_i, \omega_{i+1}, \cdots, \omega_{i+m}$ 连接而成的句矩阵 s 中的一部分。卷积操作则是卷积核 W 被应用在句矩阵 s 中所有大小为 m 的词语窗口 $\{\omega_1, \omega_{2:m+1}, \cdots, \omega_{n-m+1:m}\}$，来从输入矩阵中计算出特征映像 $c = \{c_1, c_2, \cdots, c_{n-m+1}\}, c \in R^{n-m+1}$。该特征 $c_i \in R$ 由以下公式提取出来：

$$c_i = \sigma\left(W \cdot \omega_{i:i+m-1} + a\right)$$

其中，$a \in R$ 是偏移值，$\sigma(\cdot)$ 是非线性激活函数，本文采用 sigmoid 非线性激活函数。操作符·指点积运算。

池化层作用在于依据卷积核 W 计算出的特征向量 c 上，分为最大值池化层和均值池化层。最大值池化层取特征向量 c 中的最大值，均值池化层计算也正向量 c 中所有值的平均值。本文使用的最大值池化层，因为该池化层达到降低计算量的效果，并且保留了最重要的特征，能够提取文本中的局部依赖关系。

之后连接一层全连接层，对最大池化层输出的特征 c 再做进一步的抽象转化处理。最后输入分类输出层，分类输出层神经元的个数由标签分类任务中的标签的个数决定，即任务中有多少个标签，分类输出层就有多少个神经元，每个神经元对应一个标签，并使用 signmoid 函数作为激活函数，来表达标签之间相互独立的关系。本文附件二中一级标签经 Excel 筛选后有七个标签，如图 4.3，所以分类输出层由七个神经元组成。

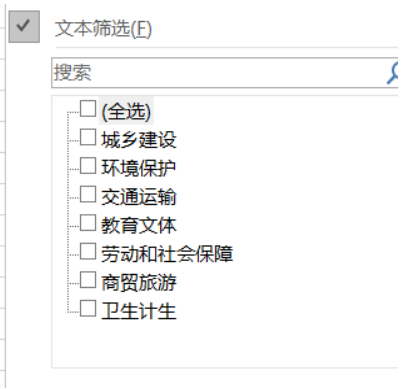

图 4.3 一级标签分类图

由 CNN 分类器对附件二中的进行留言分类，可得到结果作品附件中的热点问题表.xlsx。

4.4 评价模型

根据样本真实的正负分类和训练出的分类器对样本预测的正负分类可将样本划分为如表 1 混淆矩阵中描述的四组，分别为真正例 TP、假正例 FP、真反例 TN、假反例 FN。

表 1 混淆矩阵

真实情况	**预测情况**	
	正例	反例
正例	TP	FN
反例	FP	TN

则样本总数 $=TP+FP+TN+FN$，并产生精确率 A、查准率 P、查全率 R 三个指标，以下为 P、R 的计算公式:

$$P=\frac{TP}{TP+FP},$$

$$R=\frac{TP}{TP+FN},$$

由题中所给的 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 n 为标签类别的个数，P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。该评价指标能扩大样本数较少类别对多标签分类器性能评估的影响。

最后，在预测数据上使用 CNN 分类器预测 F_1 的值，获得最终结果如表 2 所示。

表 2 测试结果

类别	CNN
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	
卫生计生	

由上表得知，CNN 分类器对政务分析中的留言详情分类有很好的效果。

五、任务二

经过分析，可以把某一时间段内出现某一地点或者某一人群出现的频率作为热度评价指标。

采用上述 4.1、4.2 的方法对附件三的留言主题进行分类分词处理和训练词向量。

5.1 K-means 聚类

K-means 算法是以欧氏距离作为相似度测度，它是求对应某一初始聚类中心向量最优分类，使得热度评价指标很高。算法采用误差平方和准则函数作为聚类准则函数。用神经网络模型进行训练，最后得到排名前 5 的热点问题，如作品附件中的热点问题表.xls。排名前 5 的热点问题的详细内容，如热点问题留言明细表.xls。

六、任务三

经过分析可知，问题三可以利用 jieba 分词对附件 4 中的留言主题和答复意见进行分词处理。然后利用 word2vec 来比较留言主题和答复意见文本的相似程度，来判断答复的质量。

七、模型的评价与推广

八、参考文献

[1] 谢凤宏. 基于复杂网络理论的文本聚类和关键词提取方法研究[D].辽宁：辽宁师范大学，2011.

[2] 赵卫东,董亮. 机器学习[M]，人们邮电出版社，2018.8.

[3] 周志华. 2016. 机器学习[M]. 北京：清华大学出版社.

[4] 吴寅恺，陈清萍. 基于文本挖掘和网络爬虫技术度量我国系统性金融风险[J].中国人文社会科学核心期刊，2018，1001（08）：6～25.

[5] 衡宇峰，李俊，彭望龙等 . 基于语义分析的政策法规智能审核研究与实现[J]. 通信技术， 2020，53（04）：937-942.

[6] 陈晓云. 文本挖掘若干关键技术研究[D].上海：复旦大学，2005.