基于自然语言处理的留言分类算法

李幸阜, 黄伟凯, 桂文涛,何如一*
计算机科学与人工智能学院, 武汉纺织大学, 武汉, 中国
lixingfu1999@163.com, 572511183@qq.com, 3350454488@qq.com, 1171455274@qq.com,

Abstract

随着互联网的普及，越来越多的人开始在各大平台的评论区留言来反映自己的心声，而这些评论具有信息量大、信息质量参差不齐等特点。为了更有效的对这些留言进行处理，特别在今天媒体盛行的时代，其分类效率愈加变得至关重要。本文通过对比现有的分类文本方式，使用 TF-IDF、BOW 模型和 word2vec 模型对文本进行特征提取，构建了支持向量机（SVM）模型、朴素贝叶斯（NaiveBayes）模型以及 GRU 神经网络模型对留言进行对比分类。

关键词：文本分类; SVM; NaiveBayes; GRU; NLP

I. 介绍

社情民意调查是采用科学的调查和统计学方法，对一定时期一定范围内的社会公众进行的对社会现实主观反应的调查，具有反映民意、引导舆论、决策参考、检验政策实效等作用。随着互联网的发展，社情民意调查的途径不断增加，微博，微信，市长信箱，阳光热线成为网络问政平台，成为政府了解民意、汇聚民智、凝聚民气的重要渠道。各类社情民意调查文本数量的不断增加，给留言划分以及热点整理工作带来了挑战[1]，因此，建立基于自然语言处理技术的"智慧政务"系统，提高政府部门管理水平和处理问题的效率成为政府的普遍需求，本文通过实验对比，找到一种相对合理的政务留言文本分类模型。

熊等人进行了一项研究，他们对比了 Support Vector Machines，Logistic Regression，Random Forest 和 Naive Bayes 在评论分类上的准确率和有效性[2]，这项研究的最好结果准确率为84%，该分类效果并不是很好。为了提高评论分类的准确率，本文利用 pkuseg 分词工具对中文评论文本进行分词，使用 TF-IDF、BOW 模型和 word2vec 模型对文本进行向量化处理，并构建支持向量机（SVM）模型、朴素贝叶斯（NaiveBayes）模型以及 GRU 神经网络模型对留言进行分类。实验结果表明，SVM 模型的分类效果最好，准确率达到了 90%以上。本文贡献主要如下：

- 使用 TF-IDF、BOW 模型和 word2vec 模型对文本进行向量化处理。
- 对比了支持向量机（SVM）模型、朴素贝叶斯（NaiveBayes）模型以及 GRU 神经网络模型三种分类模型的分类效果。

本文的组织结构如下：第 1 节介绍了研究背景以及现状，并提出留言分类算法，第 2 节展示了研究的相关工作，第 3 节说明了本研究的预处理方法，评论分类算法，第 4 节介绍了实验结果和评价，第 5 节总结了这项研究。

II. 相关工作

A. 中文分词

中文分词是文本预处理中的关键的一步，分词的准确率往往会对下游任务产生很大的影响，目前常用的中文分词工具有 jieba、THULAC、pkuseg 等分词工具，其中 pkuseg 的分词准确率要优于 jieba、THULAC，pkuseg 基于经典的 CRF 模型，辅以 ADF 训练方法实现更高的测试效果和更好的泛化能力[3]，同时还支持不同领域的分词和多线程分词。

B. SVM

SVM(Support vector machines)早期工作来自于 Vladimir N. Vapnik 和 Alexander Y. Lerner，SVM 是一种监督分类器，广泛用于解决分类和回归问题[4]，它的基本模型是定义在特征空间上的间隔最大的线性分类器，他的优化策略是让支持向量到超平面的间隔最大化。而在实际问题中分类问题一般是多分类，将 SVM 应用于多分类问题有 one vs one 和 one vs all 两种策略，one vs one 策略是对任意两个类别构建一个 SVM 分类器，one vs all 策略是将某个类别划分为正集，而剩下的类别归为负集。假设在数据集中有 K 个类别，one vs one 策略需要构建 K(K-1)/2 个 SVM，而 one vs all 只需要构建 K 个 SVM，因此前者在类别很多的时候开销会比后者大很多。

C. NaiveBayes

贝叶斯分类算法是一种以统计学为基础的分类算法。该算法是对传统贝叶斯分类算法进行朴素的假设，所谓朴素就是假设数据类别之间彼此独立，互不产生任何影响。首先要计算属于某一类的先验概率，然后再利用贝叶斯定理计算其属于此类的后验概率，对各类后验概率的大小进行比较就可进行分类。虽然在现实中不存在这

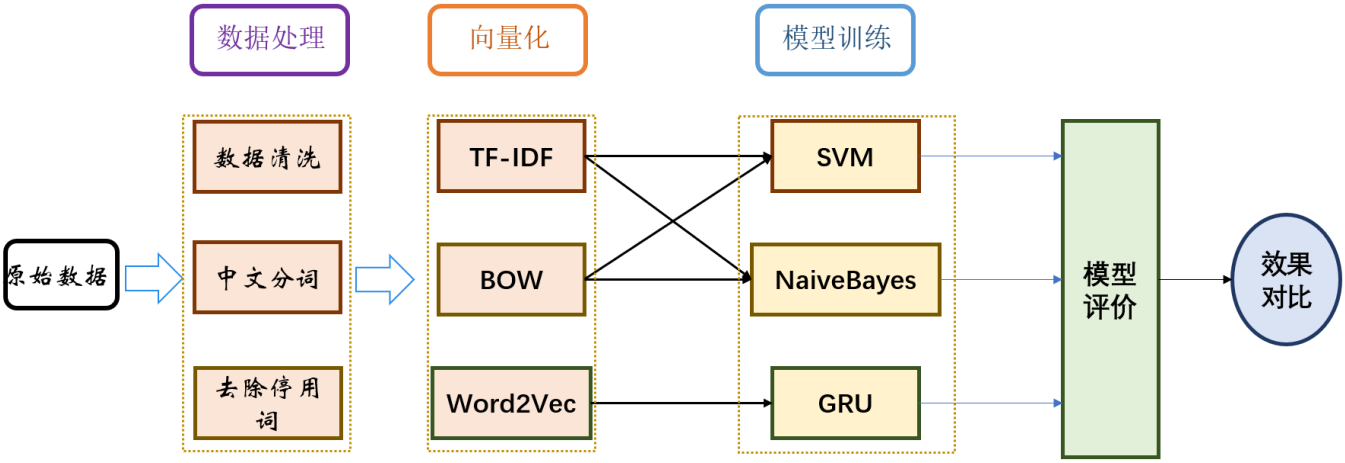

图 1 分类模型介绍

样的情况，但是在实践中对于文本分类问题进行朴素假设可以大大降低贝叶斯分类算法的复杂度。[5]

D. GRU

循环神经网络（Recurrent Neural Network，RNN）是一种用于处理序列数据的神经网络[6]，相比一般神经网络，它能够处理序列变化的数据。长短期记忆（Long short-term memory，LSTM）是一种特殊的 RNN，它是为了解决长序列训练过程中的梯度消失和梯度爆炸问题[7]。LSTM 能够在更长的序列中有更好的表现。在神经网络发展的过程中，所有关于 LSTM 的文章中对于 LSTM 的结构都会做出一些变动，也称为 LSTM 的变体，其中变动较大的是门控循环单元（Gated Recurrent Units，GRU）。GRU 是 2014 年提出的，它是对于 LSTM 结构复杂性的优化[8]。LSTM 能够解决循环神经网络因长期依赖带来的梯度消失和梯度爆炸问题，它有三个不同的门，参数较多，训练起来比较困难，而 GRU 只含有两个门控结构，在超参数全部调优的情况下，二者性能相当，但是 GRU 结构更为简单，训练样本较少，易实现。

III. 方法

A. Overview

首先对数据进行清洗、分词、去除停用词后，使用 TF-IDF、词袋模型、word2vec 工具对文本进行向量化，使用 TF-IDF 矩阵和 BOW 矩阵训练 SVM 模型和朴素贝叶斯模型，使用 word2vec 工具生成的 embedding 训练 GRU 模型，之后对这三种模型进行评价和分类效果对比。具体如图 1 所示。

B. 数据预处理

数据预处理是文本挖掘中关键的一步。首先要对去除重复项的后评论文本数据进行清洗，将链接、空白字符、分隔符标点符号、邮件地址、无意义的数字、日期时间、电话号码、身份证号去除掉，然后利用 pkuseg 分词对每条评论文本进行分词，在分词列表中搜索并剔除停用词。之后使用 TF-IDF 方法对分词列表进行向量化处理。

1) *BOW 模型*

BOW(Beg Of Word)词袋模型首先根据语料库构建字典，字典包含语料库中的每一个词，之后统计所有词的词频生成词袋 B，$B=\{word_i: frequent_i\}$，词袋 B 表示 $word_i$ 的词频为 $frequent_i$。

2) *TF-IDF 算法*

TF-IDF(Term Frequency-Inverse Document Frequency)是一种根据单词在语料库中出现频次判断其重要程度的统计方法。TF(Term Frequency)既词频，他用来描述单词(英文)或词组(中文) t_i 在文档 d_j 中出现的频次，当词项 t_i 的 TF 值越大，也就意味着文档 d_j 与词项 t_i 的关联程度越大。对于某一词项 t_i 在文档 d_j 中的 TF 值计算公式如（1）公式所示。

$$tf_{i,j} = \frac{t_{i,j}}{\Sigma_k t_{k,j}} \tag{1}$$

其中 $t_{i,j}$ 表示 t_i 在文档 d_j 中出现的次数，而 $\Sigma_k t_{(k,j)}$ 表示在 d_j 所有词项出现的次数总和。IDF(Inverse Document Frequency)既逆文档频率，它是用来衡量一个词语的普遍重要性，当包含词项 t_i 的文档个数越多，词项 t_i 的普遍性越高，IDF 值越低，这样可以对常见词进行过滤且突出重要的词。对于某一词项 t_i 的 IDF 值计算公式如（2）所示。

$$idf_{i,j} = ln\frac{1+n_d}{1+df(d,t)} + 1 \tag{2}$$

其中 n_d 表示训练集的文档总数，df(d,t)表示包含词项 ti 的文档个数。文档 d_j 中的词项 t_i 的 TF-IDF 值共可以根据公式（3）计算处来。

$$tf-idf_{i,j} = tf_{i,j} \times idf_{i,j} \quad (3)$$

经过上述处理最终得到一个 M*N 的 TF-IDF 矩阵，其中 M 表示评论个数为 M，N 表示该语料库中有 N 个不同的词。

3) *CBOW*

CBOW 模型的全称为 Continuous Bag-of-Word Model。该模型的作用是根据给定的词w_{input}，预测目标词出现的概率w_t，对应的数学表示为$p(w_t|w_{input})$。如图 2 所示，Input layer 表示给定的词，$h_1, ...,h_N$是这个给定词的词向量（又称输入词向量），Output layer 是这个神经网络的输出层，为了得出在这个输入词下另一个词出现的可能概率，需要对 Output layer 求 softmax。

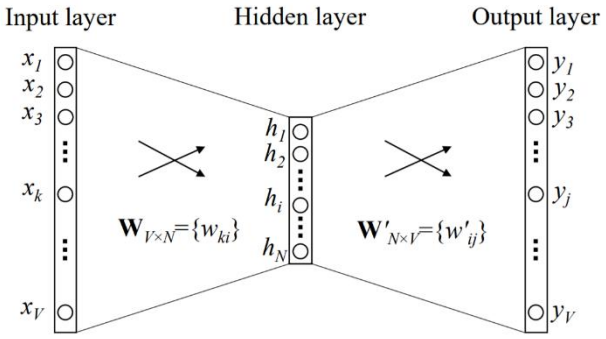

图 2 CBOW 模型[9]

C. 留言分类器

1) *SVM 分类器*

用于分类的 LinearSVC (Linear Support Vector Machine for Classification) 是 SVM 模型中的一种，假设有训练样本{(xi , yi)}，i = 1,2, … , N，训练样本数为 N，分类标签 y ∈ {−1, +1}，x ∈ R^m，表示 m 维特征空间。 则超平面为 g(ω, b) = sign(ω · x + b)

式中 ω 为法向量，决定了超平面的方向；b 为位移项，决定了超平面与原点之间的距离。可以通过求解以下约束优化问题来实现不同类之间的距离最大化，如公式(4)所示。

$$\min_{\omega,b} \frac{1}{2}\|\omega\|^2 \quad (4)$$

对于线性不可分问题，引入松弛变量 g_i 和惩罚参数 c，如公式(5)所示。

$$\min_{\omega,b,g_1,\cdots,g_k} [\frac{1}{2}\|\omega\|^2 + c\sum_{i=1}^{k} g_i] \quad (5)$$

对于非线性问题，将特征向量 x∈R^m映射到高维欧氏空间，并引入核函数：$K(x_i, x_j) = \psi(x_i) \cdot \psi(x_j)$，则超平面可以写成 $g(x) = sign(\sum_i \lambda_i y_i K(x_i, x_j) + b)$，其中，$\lambda_i$ 为拉格朗日乘子。

2) *贝叶斯分类器*

朴素贝叶斯的思想基础为：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率,哪个最大,就认为此待分类项属于哪个类别。朴素贝叶斯分类的正式定义如下:设$x = \{a_1, a_2, \dots, a_m\}$为一个待分类项，而每个 a 为 x 的一个特征属性。有类别集合$C = \{y_1, y_2, \dots, y_n\}$，计算每个 y 在 x 基础的概率分布 $P(y_1|x)$，$P(y_2|x)$，…，$P(y_n|x)$ 如果 $P(y_k|x) = \max\{P(y_1|x), P(y_2|x), \dots, P(y_n|x)\}$，则$x \in y_k$。关键是如何计算每个条件概率。对于此我们可以统计训练集中的条件概率估计，并假设各个属性是条件独立地根据贝叶斯定理有

$$P(y_k|x) = \frac{P(x|y_i)P(y_i)}{P(x)} \quad (6)$$

因为分母对于所有类别为常数所以将分子最大化即可，又因为各特征属性是条件独立的，所以

$$P(x|y_i)P(y_i) = p(y_i)\prod_{j=1}^{m} P(a_j|y_i) \quad (7)$$

3) *GRU 分类器*

RNN 作为一个随着时间的推移而计算的网络，在每个时间节点上执行相同的计算，并依赖于之前的时间状态结果。随着隐藏层的加入，RNN 考虑了单词之间的序列信息，这更适用于NLP任务。长短期记忆(LSTM)是 RNN 的一种变体，主要克服了通过门结构的梯度消失的问题。为了在不影响效果的情况下简化 LSTM 模型，它使用循环块自适应地捕获不同时间尺度上的依赖关系。给定一个序列 S={s0，s1，…，sl}，GRU 模型需要处理之前的单词向量结果，以执行当前的单词向量计算。其计算如(8)、(9)、(10)、(11)[10]所示。

$$r_t = \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{(t-1)} + b_{hr}) \quad (8)$$
$$z_t = \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{(t-1)} + b_{hz}) \quad (9)$$
$$n_t = tanh(W_{in}x_t + b_{in} + r_t * (W_{hn}h_{(t-1)} + b_{hn})) \quad (10)$$
$$h_t = (1 - z_t) * n_t + z_t * h_{(t-1)} \quad (11)$$

它主要由三个门和一个存储单元组成。这三个门分别是输入门、遗忘门和输出门，它们分别控制着输入、存储和输出的数据流。

IV. 实验和分析

A. Experimental Data

本文选用的数据集来自于第八届“泰迪杯”数据挖掘挑战赛 C 题，数据集中的文本数据共 9210 条，类别有城乡建设、环境保护、卫生计生、交通运输、教育文

体、劳动和社会保障、商贸旅游 7 个大类。数量分布图如图 3 所示。

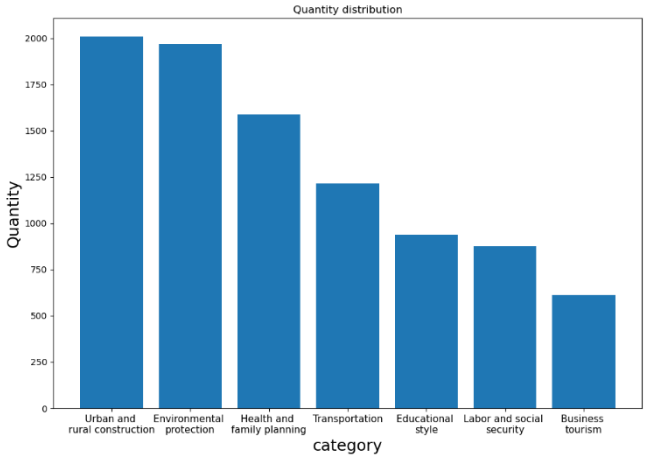

图 3 数量分布图

我们对评论文本数据进行清理、分词、去除停用词操作，分词模型使用 pkuseg 分词。分词后得到的部分数据如表 1 所示。使用 TF-IDF 方法和 BOW 模型将文本中的词对应为向量形式，转化后的文本形式分别为 9120*83658 的 TF-IDF 矩阵和 BOW 矩阵。

表 1 分词结果

评论文本	分词结果
小区在管理上十分落后，尤其是设施设备破旧不堪	校区 管理 落后 设施 设备 破旧 不堪
关于 A8 县回龙铺镇罐子窑地区砖瓦窑污染问题的反映	A 县 回龙铺镇 罐子窑 地区 砖瓦窑 污染
B 市大量出租车司机无证上岗	B 市 出租车 司机 无证 上岗
反映 D 市中学高级职称评定问题	D 市 中学 高级 职称 评定
建议西地省肿瘤医院延长微波炉的使用时间	肿瘤 医院 延长 微波炉 时间

B. 构建分类器

1) *SVM 分类和朴素贝叶斯分类*

分别将上述 TF-IDF 矩阵和 BOW 矩阵按照 7：3 的比例分为训练集和测试集，之后将所以训练集分别作为 SVM 模型和素朴贝叶斯模型的输入数据进行训练，训练完成之后使用测试集对模型进行评估，这里使用 F1-score 查看分类效果。

2) *基于 GRU 模型文本分类*

GRU 神经网络分类模型共分为 Embedding 层，GRU 层，全连接层。首先将每个文本信息的 one-hot 编码作为 Embedding 层的输入，经过 CBOW 模型训练后生成每个词的词向量，one-hot 编码的维度为 83658，因此 Embedding 输入维度为 83658，输出维度设为 512。这将大大降低词向量的维度，方便进行下游任务。之后将这些词向量作为 GRU 层的输入，GRU 的输入维度为 512，这与词向量维度保持一致，输出维度为 128，利用这些时间序列训练 GRU 单元得到隐藏层输出 h_n，然后再将 h_n 作为全连接层的输入，全连接层的输出维度和类型个数相同，因此设置为 7 维，最后利用交叉熵损失函数计算损失，计算出损失值后进行反向传播更新模型参数。模型具体表述如图 4 所示。

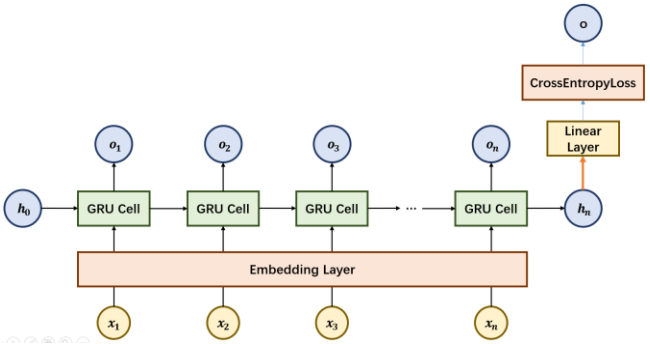

图 4 GRU 分类模型

C. 结果

通过计算并绘制模型混淆矩阵展现评论分类模型预测结果，并采用精准率、召回率、f1-score 对分类模型进行评估。SVM、朴素贝叶斯、GRU 分类模型分类结果的混淆矩阵分别如图 5、图 7 、图 8、图 9 、图 6 所示，X 轴和 Y 轴上的序号{0, 1, 2, ..., 6}分别表示城乡建设、环境保护、卫生计生、交通运输、教育文体、劳动和社会保障、商贸旅游 7 个类别。SVM、朴素贝叶斯、GRU 分类模型分类结果对比分析效果报告如所示。

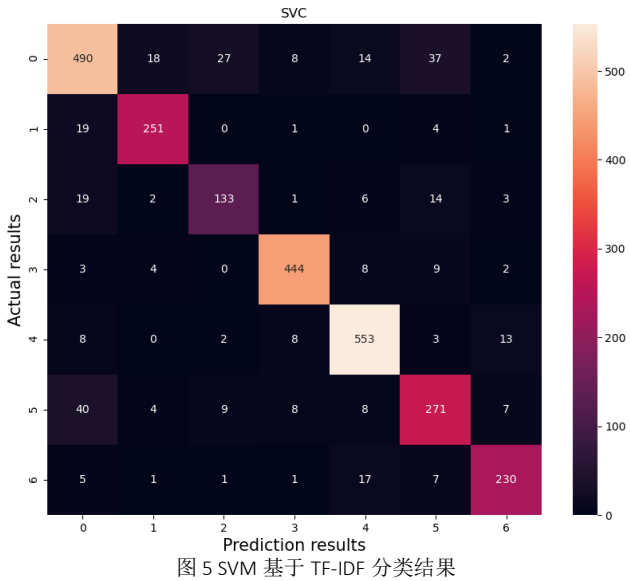

图 5 SVM 基于 TF-IDF 分类结果

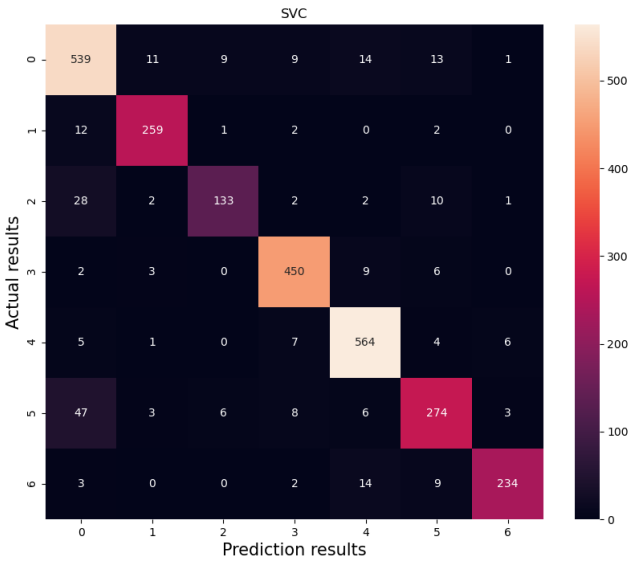

图 7 SVM 基于 BOW 分类结果

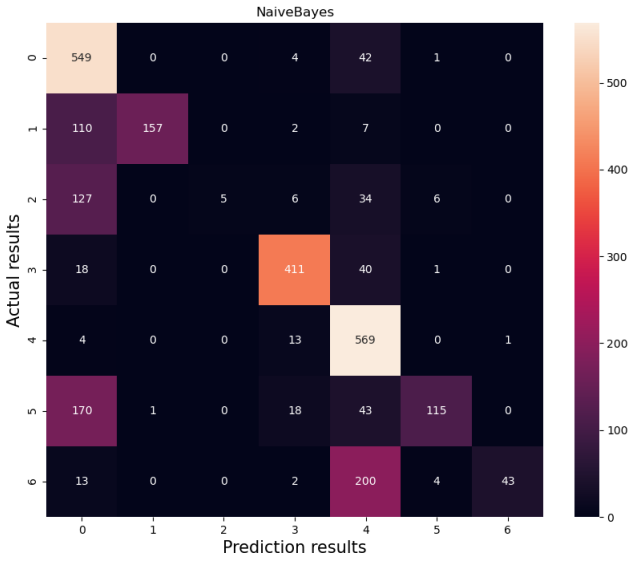

图 8 NaiveBayes 基于 TF-IDF 分类结果

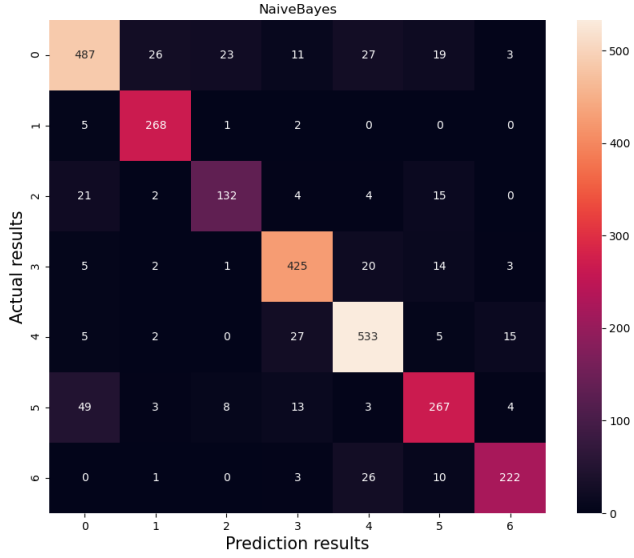

图 9 NaiveBayes 基于 BOW 分类结果

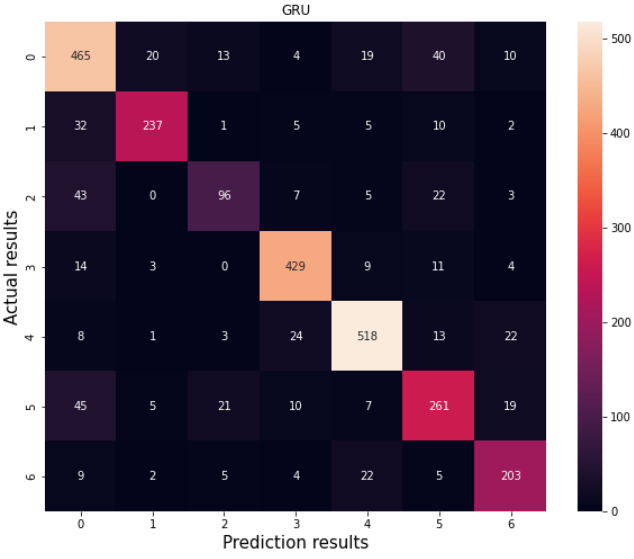

图 6 GRU 模型分类结果混淆矩阵

从混淆矩阵中发现，在使用 TF-IDF 算法进行文本向量化的条件下，朴素贝叶斯对留言的分类结果变得混乱，这使得朴素贝叶斯得效率是所有模型中表现效果最差的。同时还可以看出，无论采取 TF-IDF 算法或者 BOW 向量化文本方式，还是使用 CBOW 进行词嵌入编码，SVM 模型要明显优于其他模型，分类效果表现的很好。

在这些类别中，留言被错误分类到城乡建设类留言最多。但基于 TF-IDF 的朴素贝叶斯模型却对城乡建设的分类正确率最高。而 GRU 神经网络模型对城乡建设类型的留言的分类效果最差，此模型将城乡建设类型的留言分类到商贸旅游类型留言和卫生计生类型留言很多。

最后生成的分类结果报告如表 2 所示

表 2 分类报告

模型 \ measure / embedding		precision	recall	f1-score	ACC
SVM	TF-IDF	0.91	0.89	0.90	0.91
	BOW	0.87	0.87	0.87	0.87
NaiveBayes	TF-IDF	0.79	0.68	0.63	0.68
	BOW	0.86	0.86	0.86	0.86
GRU	CBOW	0.81	0.81	0.81	0.81

从该报告中可以看出该分类模型表现最好的为基于 TF-IDF 的 SVC 分类模型，它的 f1-score 为 0.90，正确率为 0.91。最差的为基于 TF-IDF 的贝叶斯分类模型。它的 f1-score 为 0.63，正确率也仅仅只有为 0.68。其次是

GRU 神经网络模型，他的正确率仅仅达到 0.81，f1-score 有 0.81，但要比基于 TF-IDF 的 NaiveBayes 模型要好的多。在使用 BOW 模型对文本进行向量化时，SVM 分类模型和 NaiveBayes 分类模型效果近似。

V. 结论

本文使用 pkuseg 工具进行中文分词任务，在对文本进行清洗之后，采用了 TF-IDF 算法、BOW 词袋模型和 word2vec 三种工具对文本进行向量化处理。本文构建了支持向量机（SVM）模型、朴素贝叶斯（NaiveBayes）模型以及 GRU 神经网络模型对留言进行分类。其中支持向量机（SVM）模型、朴素贝叶斯（NaiveBayes）模型使用训练好的 TF-IDF 和 BOW 文本矩阵作为输入，而 GRU 神经网络模型使用 word2vec 训练好的词嵌入向量作为输入。

本文对比了三种分类模型在三种词编码方式下的分类效果。实验结果表明，基于 TF-IDF 的 SVM 模型的分类效果最好，准确率达到了 90%以上。

参考文献

[1] 陈曦,″文本挖掘技术在社情民意调查中的应用.″ 中国统计, vol.06, pp.27-29, 2019.

[2] Q. Xiong, L. Li, Y. You, J. Fan, J. Liu, et al., "Experimental Evaluation of Intelligent e-Government System Based on Text Mining," 2020 5th International Conference on Control, Robotics and Cybernetics (CRC), 2020, pp.161-164.

[3] R. Luo, J. J. Xu, Y. Zhang, et al., “PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation,” Arxiv, 2019.

[4] R. Obiedat ,R. Qaddoura, A. Al-Zoubi, et al., "Sentiment Analysis of Customers’ Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," IEEE Access, vol. 10, pp. 22260-22273, 2022.

[5] 许丽, 焦博, 赵章瑞, ″基于 TF-IDF 的加权朴素贝叶斯新闻文本分类算法.″ 网络安全技术与应, vol.11, pp.31-33, 2021.

[6] R. Pascanu, T. Mikolov and Y. Bengio, “On the difficulty of training Recurrent Neural Networks” 2012.

[7] L. Song, A. Wang and J. Ren, "Inverse Dynamic Model using GRU Networks Learning," 2021 International Conference on Advanced Mechatronic Systems (ICAMechS), pp. 52-55, 2021.

[8] N Wang, J Wang and X. Zhang, "Multi-ensemble Bi-GRU Model with Attention Mechanism for Multilingual Emoji Prediction” Proceedings of The 12th International Workshop on Semantic Evaluation“, pp. 459-465, 2018.

[9] X. Rong, "word2vec Parameter Learning Explained." Computer Science, 2014.

[10] K.H. Cho, B.V. Merrienboer, D. Bahdanau and Y. Bengio, “On the properties of neural machine translation: Encoder decoder approaches”, arXiv preprint arXiv:1409.1259, 2014.