大数据环境下政务数据的情报价值及其利用研究*

——以海关报关商品归类风险规避为例

王 昊[1,2],邓三鸿[1,2],朱立平[1,2,3],王鑫芸[1,2],范 涛[1,2]

1.南京大学信息管理学院,南京 210023
2.江苏省数据工程与知识服务重点实验室,南京 210023
3.南京海关,南京 210001

摘 要:[目的/意义]为充分挖掘大数据环境下政务数据中的情报价值,文章以海关报关商品自动归类研究为例,探讨经过长时间积累的大规模数据在数据自动处理和分析中的典型应用,从而有效体现数据的情报价值。[方法/过程]提取积累的政务数据中的报关商品及其类目的相关特征信息,进而采用深度学习方法对其进行建模和训练,最终利用机器学习获得情报,实现对未知类目报关商品的自动分类,达到风险规避的目的。[结果/结论]文章首先对不同的深度学习文本分类模型进行比较,在对得到的情报进行分析后,选择构建加入Attention机制的TextRNN模型。实验结果表明,该模型表现最优,能够较好的对海关报关商品进行归类,进而规避风险,并能更加充分挖掘海关报关数据中的情报价值。[局限]实验中对于报关商品特征的讨论有限,选取特征时仅参考了历史研究、专家意见与相关性值,其他有效特征可能被过滤,存在一定的局限性。

关键词:政务数据;风险判别;文本分类;TextRNN;报送商品;情报价值;HS编码

中图分类号:G353 **文献识别码:**A **文章编号:**2096-7144(2020)04-0074-16

DOI:10.19809/j.cnki.kjqbyj.2020.04.007

1 引言

随着电子政务的快速发展,政务数据不断积累,达到大数据量级,并成为不容忽视的重要资源。大数据环境下的政务数据价值密度大,开发价值高,同时也是国家重点发展的战略任务之一[1]。因此,如何利用情报学方法对政务大数据进行充分高效的挖掘,让重要的政务大数据发挥其情报价值,显得极为重要。目前已有研究,利用情报学方法对政务数据进行挖掘,例如,基于区块链技术的政务数据安全共享服

收稿日期:2020-08-12 修回日期:2020-08-28
基金项目:"南京海关税收大数据分析咨询项目";"江苏青年社科英才";南京大学"仲英青年学者(Tang Schloar)"。
作者简介:王昊(ORCID:0000-0002-0131-0823),男,1981年生,博士,教授,博士生导师,主要研究方向:数据计算与分析、本体学习技术及其应用,E-mail:ywhaowang@nju.edu.cn;邓三鸿(ORCID:0000-0002-6910-3935),男,1975年生,博士,教授,博士生导师,主要研究方向:科学计量、信息处理,E-mail:sanhong@nju.edu.cn;朱立平,男,1974年生,博士研究生,主要研究方向:知识图谱构建及应用、智慧海关,E-mail:chemzlp@163.com;王鑫芸(ORCID:0000-0003-3777-7748),女,1997年生,硕士研究生,主要研究方向:科学评价、自然语言处理,E-mail:wang_xy1997@sina.com;范涛(ORCID:0000-0002-6846-2901),男,1995年生,博士研究生,主要研究方向:知识图谱、情感分析,E-mail:fantao@smail.nju.edu.cn。

务[2]、基于社会网络分析法和主题分析法的政务数据分析[1]134和基于文本计量和内容分析的政策文本数据挖掘[1]134。尽管目前已有研究利用情报学方法对政策文本等政务数据进行挖掘，但依旧缺乏对海关报关数据进行相关的情报价值挖掘研究。

海关报关数据作为典型的政务大数据，利用情报学方法发挥数据的情报价值，具有较高的应用前景和实践意义。海关报关单电子化的实施是海关大数据项目中重要的组成部分，报关单是海关监管、征税、统计以及开展稽查和调查的重要依据，是加工贸易进出口货物核销，以及出口退税和外汇管理的重要凭证，还是海关处理走私、违规案件，及税务、外汇管理部门查处骗税和套汇犯罪活动的重要证书[3]。当报关单与实际情况不符时，可能存在瞒报漏报、走私、偷税漏税等危害国家利益的行为，因此对海关商品报关单风险识别工作非常重要。传统的海关报关单的审核需要报关员根据报关单数据，结合内外部其他数据判断报关单的合理性，这一过程受到报关员的知识背景，对商品类目的熟悉程度，以及人为操作等主客观因素的影响[4]。因此，在当前时代背景下，利用数据技术实现报关单的自动风险评估成为了一种趋势[5]，使用历史报关数据构建模型，实现对此类风险的机器判定[6]，也成为了一个亟待解决的问题。

基于此，本文以海关历史报关数据为例，针对海关商品编码与报关单数据，通过深度学习方法构建预测模型，关注海关报关商品中的有效训练特征，对海关中报关商品进行风险自动检测分类，并同相关模型进行对比，从而让海关报关数据发挥情报价值，为实现“智慧海关”提供可行性路径。

2　相关研究

如何利用情报学方法挖掘海关报关数据的情报价值，构建海关报关商品风险自动识别模型，并解决当前处理海关报关数据所面临的问题，将是本文的研究重点。因此，本文将对近期的政务数据研究、海关报关数据的分类文本表示研究以及深度学习建模研究，进行梳理和总结。

2.1　政务大数据研究

宁靓等[7]通过文本分析等方法，对山东政府政务网中的民众办事咨询数据，进行互动信息挖掘。陈平刚等[8]基于区块链技术，构建政务大数据集，并对政府舆情进行预警研究。谭必勇和陈艳[9]利用自然语言处理方法对开放政府数据，进行数据质量评估研究。孙卓林和徐云飞[10]利用文本计量和内容分析方法，对中部六省政策文件进行深入挖掘。赵浚吟[11]利用问卷调查的方式，对政务抖音中的用户信息行为进行研究。

2.2　海关报关数据表示研究

中国目前对商品使用的海关HS编码为每2位一级，不断细分为5级10位编码[12]，这种海关HS编码规则与其他某些领域的编码判别问题较为类似，这些问题多数会根据编码的结构，被转换为多部分分别进行，最终重新组合获得完整编码，具体的研究如：王克海[13]将产品作业逻辑关系分解为零部件代号和工艺过程2部分，用在作业事项号的自动生成研究中；陈东明等[14]根据企业的编码结构特征，提出基于代码自动生成产品结构编码的算法；王昊[15]在研究中提出该类结构编码的判别问题可以采用单层分类法（Flat Classification）和层次分类法（Hierarchical Classification）实现中文数目的分类。其中单层分类法认为编码是独立的，将数据归到置信度最高的类别中；层次分类法则根据类目之间的关系将复杂分类任务简化为较小分类[16-17]，这种方法降低了计算复杂度，适用于大规模的自动分类[18]。从以上研究中可以发现，具有一定结构特征的编码的自动生成与判别归类问题研究，均将研究的关键集中在编码的结构

问题上，如何利用编码的结构特征辅助模型进行编码的生成是研究者关注的重点。除此以外，一般研究者通常将风险的识别研究转化为分类问题解决。如：谢小楚[19]将案件分为一般案件、简单案件和简易程序案件3种，进行海关缉私案件的预测；严俊龙等[20]在对网络安全的研究中，将网络安全性转化为安全与不安全2类，构建网络安全风险评估模型；罗方科等[21]在个人小额贷款信用风险评估的研究中，将用户视为违约与履约2类进行模型的构建。当前海关审单根据风险高低将商品分为不同类别，从而采取不同通关渠道的通关措施，那么在海关商品归类风险判别的建模中是否可以采取这种分类方式，也是本研究需要探讨的问题。

2.3 深度学习建模研究

深度学习是机器学习的一种新的延伸，是相较于之前发展起来的浅层学习而言的[22]。近些年来，由于众多高科技公司在战略层面的高度重视，深度学习的应用越来越深入到人们的日常应用当中[23-26]。在海关进出口商品分类任务中，已有学者做了相关研究。例如，刘昌伟、段景辉[27]使用因子分析方法实现了对海关数据的降维和简化，为进一步挖掘数据间隐含的关系提供了有力的支持；周欣、张弛海[28]采用决策树（Decision Tree，DT）模型，以报关单数据是否为风险记录作为分类依据，建立对海关风险的预测模型；GUO LI等[29]提出了文本图像自适应CNN模型，分别基于报关商品的文本和图像描述提取特征进行分类，再融合得到最终的分类结果。

基于此，本文将利用深度方法构建海关商品风险识别模型，结合真实的海关数据，进行实证研究。

3 方案与算法

3.1 数据与解决方案

本文的数据来自某海关连续4个月的商品报关数据，经过对不同数据表单的清洗、拼接与整理，获得共计240 000余条记录，每条记录对应164个描述字段。其中在审单过程中，出现风险的记录共计7 258条，未出现风险的记录共计234 327条。

根据海关规定，海关贸易以国际公约中规定的统一海关编码进行识别，简称为“协调制度”。

（Harmonized System，HS）[30]，主要采用6位数编码对商品进行区分，且各国家可根据本国实际分出6位后的更多位数码[31]。编码根据商品种类的不同，通常按层级从高到低由每2位码为一级，逐层深入组合得到完整的HS编码。根据中华人民共和国海关进出口税则[32]，我国HS编码有10位，分为5个层次，“HS编码位”“报关商品类目层次”，见表1。根据统计，模型中可能出现的HS编码总计约有4 500余种，参考图书分类中多类目分类法的特点，可以采取层次分类法构建分类模型。本文则将10位HS编码划分为3个层次，如表1“分类层次”所示。层次太少，对应的类目偏多，难以学习；而层次太多，不仅会产生明显的准确率递减效益，而且低层可学习语料明显偏少也会导致学习不充分。为此，本文选择将后6位HS编码笼统记为1层进行建模，进而将HS编码生成问题转化为3个层次的层次分类问题。

表1 HS编码层次分类示例表

序号	HS编码位	报关商品类目层次	分类层次	示例：850110000
1	1-2位	1层	1层	85
2	3-4位	2层	2层	01
3	5-6位	3层	3层	10 00 00
4	7-8位	4层		
5	9-10位	5层		

根据层次HS编码数与样本数据量变化的统计结果，来源数据类目间数据量差异较大，部分类目数据量过少，因此在取编码前2位作为分类标记时，只选择了来源数据中数据量最大的5个类，分别为“39”“73”“84”“85”“90”等，然后以其他类目的数据构成

“其他”类，每一类下抽取约8 000条数据，总计40 052条数据进行实验。各类中参加实验的具体数据量，见表2。

表2 HS编码1–2位实验的数据情况表

序号	1	2	3	4	5	总计
HS	39	73	84	85	90	–
数据量	8 017	8 025	8 017	8 001	7 999	40 052

在4位标记编码的实验中，为保证下位类的充分学习，抽取数据量最大的“85”类进行实验。前2位为“85”的3–4位编码，在该层次的模型构建中分类标记共45种，实验数据的总和为77 214条。在5–10位的实验中按计划应抽取“85”类包括的45种下位类中数据量合适的6位标记编码，经过数据统计检查发现报关数据中细分的后6位类数据量最大的编码包含的类别过多，因此本实验在全部数据中进行统计，采取其中数据量较大、类别较多的“8536”类进行测试，此处的分类标记为HS编码的5–10位共计13种，实验数据的总和为18 582条。

3.2　深度学习算法

深度学习算法部分，本文主要使用了Fasttext、TextCNN和TextRNN等常见的文本分类中常用的深度学习模型。

3.2.1　Fasttext

Fasttext模型是由Facebook在2016年开源的专用于文本分类的深度学习模型[33]，其结构，如图1所示，它将传统的Word2vec模型中基于单词的算法进行了改进，变为基于字符级别来表示一个单词，并使用N–gram表示单词，如对于单词“book”，当N取3时，其Trigram有：“<bo”“boo”“ook”“ok>”，其中“<”和“>”分别为单词的前后缀标识[34]。Fasttext与Word2vec的不同之处主要体现在其输入与输出层上，Fasttext的输入层是整个文本的内容，包括句子、词语与N–gram，其输出为分类的标签，而中间生成的vector不会被保留；Word2vec的输入为规定大小的输入窗口中的词语，而输出则为每个词语的概率矩阵，最终会得到相应的词向量矩阵。总的说来，Fasttext的训练速度快，适用于数据集充足、分类类别大的情况。

3.2.2　TextCNN

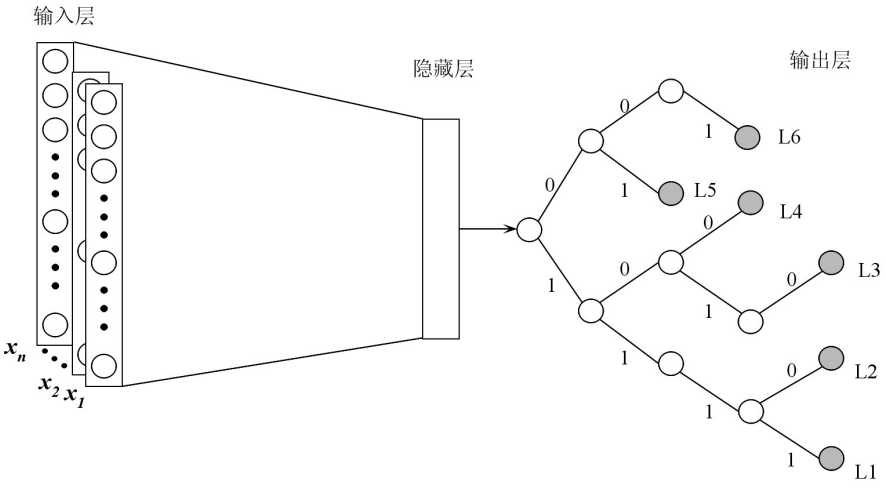

图1 Fasttext文本分类模型图

用n代表输入每条句子的长度，如图2所示，图中500为Word2vec压缩效率与准确率综合最佳的维度，在这里则为TextCNN模型的embedding维度。CNN通常被用在计算机视觉与语音识别领域，其中的卷积有局部特征提取的功能，因此可以类似N–gram地提取出句子中的关键信息。TextCNN的输入

即为word2vec的结果，之后进入卷积层卷积窗口大小即为n×500，卷积后得到若干个n×1的Feature Map，之后进入池化层，在Feature Map中用Maxpooling选取最大值作为最能代表文本的信息，从而得到一个1维向量，之后经过全连接到一个Softmax层获得属于不同类别的概率向量，从而得到了类别标签。

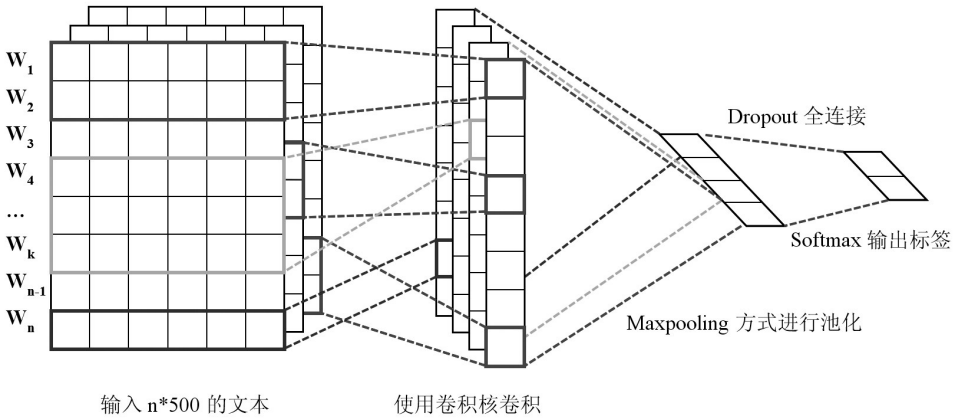

图2 Word2vec+TextCNN的风险识别模型图

3.2.3 TextRNN

本文采用加入Attention机制的TextRNN模型进行风险识别[35]，传统的RNN模型经常被用在序列数据中，较CNN通常被用于图像相比，RNN更多的被使用在文本分类领域，它将每个词看做序列中的一个节点，将词向量作为每个节点的输入，并组合前后词构成双向的特征，计算每个单元的状态与输出特征，一般的TextRNN模型会将序列的特征平均值输入全连接层，可以忽略句子长度的问题，并获得分类标签，但是求特征平均值会使每个词获得同样的参数，相同的参数则会影响模型的表达能力，因此本文使用加入了Attention机制的RNN模型来代替平均的特征，具体模型，如图3所示。

图中可以看出基于深度学习方法的预测实验主要围绕文本特征与参数的调整进行，模型具体参数值，见表3。

实验中调节的参数有：①输入序列长度seq_length，它控制每条文本输入的字符长度，当实际长度长于设定值时会将其截断，短于设定值时会补0以使其符合输入要求。海关的文本通常长度较短，为了获取有效的信息将起始值设定为200，而其实际长度一般很少长于400，因此将其调节范围设定为200~400；②循环层数num_layers指的是多少个RNN结构堆叠在一起，形成一个堆叠RNN结构，后一个RNN接受前一个RNN的输出并计算结果；③隐藏节点数hidden_dim是训练出现过拟合的直接原因，为了避免这种情况，需要在满足精度的情况下尽可能使用紧凑的结果，减少隐藏节点，当其值太小时，网络可能太小而不能训练或性能很差，因此在留出了1 000条左右作为验证集的情况下，训练样本大致为14 000条，训练样本数一般为节点的2~10倍，且为了控制训练效率，经过验证该值设定为128；④keep_prob为Dropout值，过大容易导致训练不充分，过小会导致网络训练成本过高，且容易过拟合，因此选择0.5即删除一半的神经元；⑤learning_rate是学习效率，在梯度下降过程中，该值过小可能长时间无法收敛，过大可能无法收敛，因此学习率选择了常见的经验值，并加入lr_decay作为衰减值，即随着迭代的进行，减小学习的速度，逐渐达到收敛点，grad_clip梯度裁剪阈值同样是避免不收敛的情况；⑥num_epochs迭代次数

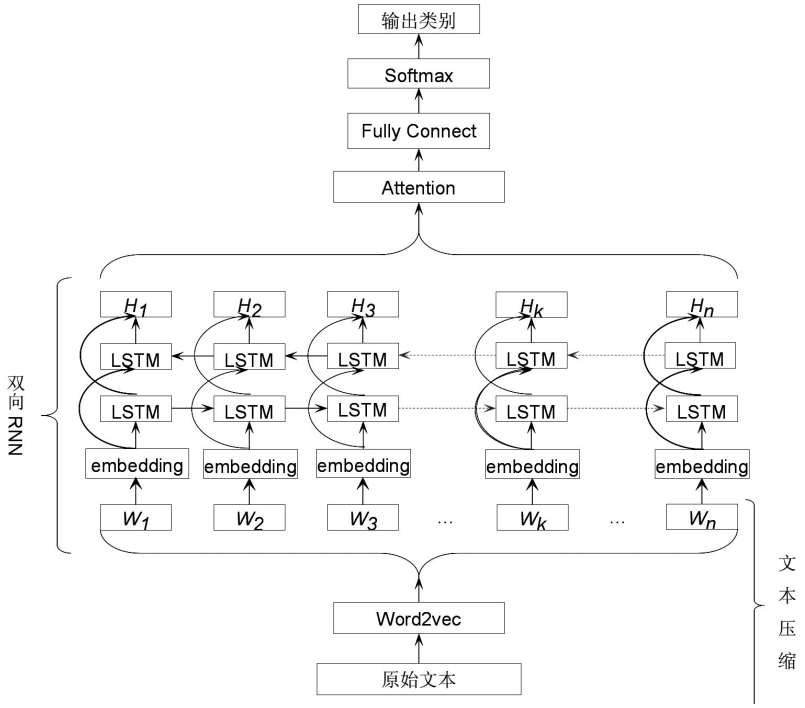

图3 TextRNN+Attention结构的风险识别模型图

表3 模型参数与含义表

序号	参数名称	含义	取值
1	vocab_size	词汇表大小，即Word2vec生成的词汇表	由样本确定
2	num_classes	类别数	由样本确定
3	embedding_dim	Embedding维度，即Word2vec生成的词向量维度	500
4	seq_length	输入序列的长度	200~400
5	num_layers	循环层数	1~2
6	hidden_dim	隐藏层节点数	128
7	keep_prob	dropout保留比例	0.5
8	learning_rate	学习率	$1e^{-3}$
9	lr_decay	学习率衰减率	0.9
10	grad_clip	梯度裁剪阈值	5.0
11	num_epochs	总迭代轮次	10
12	batch_size	每批训练大小	64

是所有的训练数据均进行过1次训练，实验过程中发现迭代至8次以上后，模型基本不再出现很大变化，因此选择迭代10次；⑦batch_size每批训练大小通常设置为2的n次幂，常取值为64，128，256，由于本文构建的网络较大，因此选择64。

3.3 模型影响因素

除了深度学习算法外，以下4个因素也可能会对报关商品的归类模型产生不同程度的影响。

3.3.1 文本表示方法

目前常用的文本表示方法主要有词袋模型和词向量等两种。词袋模型最早由Hinton[36]提出，其经典原模型由Bengio等[37]建立，多采用One-hot[38]和Term Frequency[39]作为权重设置方法。词向量方法的典型代表是Word2vec和BERT，前者是Google在2013年推出的自然语言处理工具，借鉴了神经网络语言模型（Neural Network Language Model，NNLM）的思想，

能够根据使用者给定的语料库，将文本中的词投映到一个低维、稠密的实数向量空间中，其每一维都代表了词的浅层语义特征[40]，从而起到数据降维的作用。

3.3.2 文本粒度

在深度学习的研究中，研究者经常通过将词分为更小粒度的字，提升模型的准确率，因此，在探讨海关商品报关风险的深度学习模型中，也将输入分为词粒度与字粒度分别进行，分词的标准依旧保留所有的字符串，根据深度学习模型的要求，直接输入分词后带空格的文本作为原始文字记录。

3.3.3 结构深度

在参数介绍中提到深度学习的模型深度，即循环层数会影响模型的训练结果，一般来说深度越深，训练效率越低，占用CPU资源越多，效果越好。实际设计神经网络结构的过程中，一般默认有限考虑从1开始进行堆叠，在文本的实验设备条件下，在网络结构层数设置为3时，设备已经无法进行计算，因此仅对比1层与2层网络结构的情况。

3.3.4 序列维度

输入序列维度同样容易影响训练的结果，维度过低时，模型无法获得充分的原始信息；维度过高对构建的网络过大。对于较短的文本，补零会减少实际文本特征的比例，且对设备的性能要求较高。考虑到输入序列需要输入大部分的文本信息，同时要保证一定的设备运行效率，本文将输入序列控制在200~400之间，寻找最佳输入序列值，具体的序列信息与分析见实验结果。

4 实验结果

深度学习模型的输入文件包括训练集与验证集2个文件，其中每行均由标签与文本组成，代表一个样本。通过验证集，模型可以在训练过程中判断迭代的程度，并可以通过及时的验证效果输出判断是否出现了过拟合现象。本文采用P值（分类正确率）作为分类效果的评价指标，而在各类目内部则采用$Micro_P_i$、$Micro_R_i$以及$Micro_F1_i$（其中i为类目标记）作为指定类目的分类效果。

4.1 模型初探

海关的文本数据有以下显著的特点：一是，文本内容是词语的简单组合，词语之间没有十分明显的逻辑上下文关系，且不同方面的描述之间用“|”进行分隔；二是，各记录之间的文本内容、长度等一致性较差，利用Word2vec对其进行学习压缩的效率也是有限的，但总体平均文本长度较短。因此，针对文本描述不规范且分散的特点，本章计划主要使用加入Attention机制的RNN模型针对海关数据进行训练学习，观察训练效果。但是对于基础的文本深度学习模型Fasttext、TextCNN等模型均在2位HS编码层级上进行了实验，以确定3种模型中最适合海关风险识别的模型。

3种模型在2位HS编码上的实验结果，见表4。

观察表4可以发现，Fasttext与普通的TextCNN的训练测试效果非常不理想。这表明模型对海关报关商品无法进行有效的归类，并难以实现风险规避的目标。

Fasttext模型中，虽然其训练速度非常快，但除了在“73”类与“90”类上没有召回任何样本以外，其他类别的准确率或召回率均有显著的缺陷，其总准确率效果也仅有32.07%。分析发现效果较差的原因是Fasttext主要适用场景为超大型数据集，在没有GPU的情况下也可以达到每分钟亿级词汇的训练效率，其支持多语言的表达，并将字符串分为单个字母的级别，且有自己独立的标准，因此对于海关文本这样不符合日常语言逻辑，包含大量字符串的文本，同时训练样本数量仅为万级的情况来说，虽然其训练时

表4 Fasttext、TextCNN与TextRNN实验结果表

模型	类别	Micro_P	Micro_R	Micro_F1	P
Fasttext	其他	21.14%	92.62%	34.42%	32.07%
	39	67.77%	18.10%	28.57%	
	73	–	0.00%	–	
	84	85.48%	41.44%	55.82%	
	85	59.20%	21.17%	31.19%	
	90	–	0.00%	–	
	总计	38.93%	28.89%	25.00%	
Word2vec + TextCNN	其他	69.62%	37.64%	48.86%	54.87%
	39	57.01%	39.51%	46.68%	
	73	38.42%	87.92%	53.47%	
	84	62.58%	64.48%	63.51%	
	85	71.75%	56.94%	63.49%	
	90	52.50%	39.62%	45.16%	
	总计	58.65%	54.35%	56.42%	
Word2vec + TextRNN	其他	95.74%	91.14%	93.38%	93.00%
	39	94.63%	93.38%	94.00%	
	73	95.69%	99.33%	97.48%	
	84	90.78%	91.36%	91.07%	
	85	92.82%	94.31%	93.56%	
	90	87.94%	88.41%	88.17%	
	总计	92.93%	92.99%	92.96%	

间效率非常高，设备资源需求较小，但是很难在“有限”的数据量下达到充分的训练，对数据量较少，或特征较为分散的类别，其训练效果很不理想。

TextCNN模型的训练结果虽然较Fasttext有所好转，但是依旧仅有54.87%的总准确率。训练时间效率上，TextCNN较Fasttext相对较差，但总体来说本实验中万级的训练数据量依旧可以控制在10 min左右，但是实际上海关每条记录本文与通常用来做新闻或情感分类的文本相比并不长，上下记录之间的相关性也不大，这与TextCNN更关注局部词语信息的特点相矛盾；根据之前Word2vec的分析，海关分类之间词语的差别较小，存在一定的迷惑性，即使经过训练TextCNN也难以区分其类目之间的差别，因此训练效果也非常有限。根据查阅该领域模型的相关研究，TextCNN的效果在其他研究人员的测试中也还没有完全超越SVM。

显然加入Attention机制的TextRNN模型训练效果优异，Attention层对特征状态的针对处理，良好解决了海关文本在模型训练中多方面的缺陷，能够对海关报关商品进行有效的归类。本文将继续利用针对性扩大文本有效特征的加入Attention 的TextRNN继续进行实验，之后实验结果中的模型均指代加入Attention机制的TextRNN。

4.2　文本粒度

在文本分词过程中，将文本拆分为字与词，以此探讨深度学习模型在文本粒度上的分类差别。实验结果，见表5。

文本粒度即词特征与字特征的分类粒度结果相近，总体的准确率均较高。

在2位HS编码的实验中，词粒度的实验结果略高于字粒度的特征，在出现数据最多的5个类目中，“85”类与“84”类的描述较为相似，二者容易出现混

表5 文本粒度实验结果表

粒度		词				字			
2位	类别	Micro_P	Micro_R	Micro_F1	P	Micro_P	Micro_R	Micro_F1	P
1	85	92.82%	94.31%	93.56%	93.00%	92.32%	89.86%	91.07%	92.97%
2	84	90.78%	91.36%	91.07%		90.64%	94.56%	92.56%	
3	39	94.63%	93.38%	94.00%		95.26%	97.57%	96.40%	
4	90	87.94%	88.41%	88.17%		90.00%	87.33%	88.65%	
5	73	95.69%	99.33%	97.48%		94.07%	99.33%	96.63%	
6	其他	95.74%	91.14%	93.38%		95.64%	89.11%	92.26%	
–	总计	92.93%	92.99%	92.96%		92.99%	92.96%	92.98%	
4位	类别	Micro_P	Micro_R	Micro_F1		Micro_P	Micro_R	Micro_F1	
1	01	100.00%	95.83%	97.87%	96.46%	96.00%	100.00%	97.96%	96.21%
2	02	–	–	–		–	–	–	
3	03	90.00%	78.26%	83.72%		89.47%	73.91%	80.95%	
4	04	99.32%	96.67%	97.97%		92.95%	96.67%	94.77%	
5	05	80.00%	94.12%	86.49%		100.00%	82.35%	90.32%	
…	…	….	…	…		….	…	…	
45	48	97.22%	89.74%	93.33%		97.14%	87.18%	91.89%	
–	总计	67.07%	63.30%	65.13%		67.83%	63.91%	65.81%	
10位	类别	Micro_P	Micro_R	Micro_F1		Micro_P	Micro_R	Micro_F1	
1	100 000	99.44%	100.00%	99.72%	92.49%	99.44%	100.00%	99.72%	92.92%
2	200 000	97.44%	100.00%	98.70%		100.00%	100.00%	100.00%	
3	300 000	84.30%	98.08%	90.67%		84.21%	92.31%	88.07%	
…	…	…	…	…		…	…	…	
13	909 000	88.80%	8.21%	15.03%		88.83%	4.54%	8.63%	
–	总计	90.02%	77.23%	83.14%		91.83%	76.48%	83.45%	

淆，出现了在词粒度与字粒度的实验中表现相反的现象；总体看文本粒度对实验结果的影响，表现并不显著，但在包含了所有其他类别的“其他”类中，词粒度文本特征的效果显然优于字粒度特征，原因可能是在其他类中，当拆分为更小的字特征时，实际上分散了文本的特性，因此对于本身包含了过多类目的“其他”类来说，文本特征被分散后导致其召回率出现了下滑，在这样的情况中，词粒度特征的效果相对较好；在剔除“其他”类后，词粒度文本与字粒度文本的准确率分别为93.41%与93.81%，这与之前发现的规律相对应。

4位HS编码的表现与2位编码的表现类似，总体上说，类目间的表现与训练样本中的数据量密切相关，数据量相对占比较多，训练充足的类目通常可以获得更好的训练效果，但对于数据占比较少，训练不充分的类目，几乎难以获得有效的召回率；类目之间差距进一步变小后，从文本粒度的角度观察，发现字粒度的文本表现相对词粒度对类目数据量的要求更低，其在数据量占比较小的类目上的表现显著优于词粒度文本构建的模型，可能是由于字粒度特征在类目之间的文本描述更接近时，可以通过模型放大其间的差异，从而有助于分类效果的提升。

10位HS编码的表现验证了4位编码中发现的规律，即字粒度的文本表现相对词粒度对类目数据量的要求更低，且字粒度文本更有助于相似类目之间的判别。当类目数据量有限时，字特征可以有效获得识别目标的特征；当类目数据量较大，包含较多特征时，字特征反而容易出现误判的情况。总得来说，

字粒度文本特征在10位HS编码上的表现更优。

综合3个层次的分类结果,其总体表现,如图4所示。

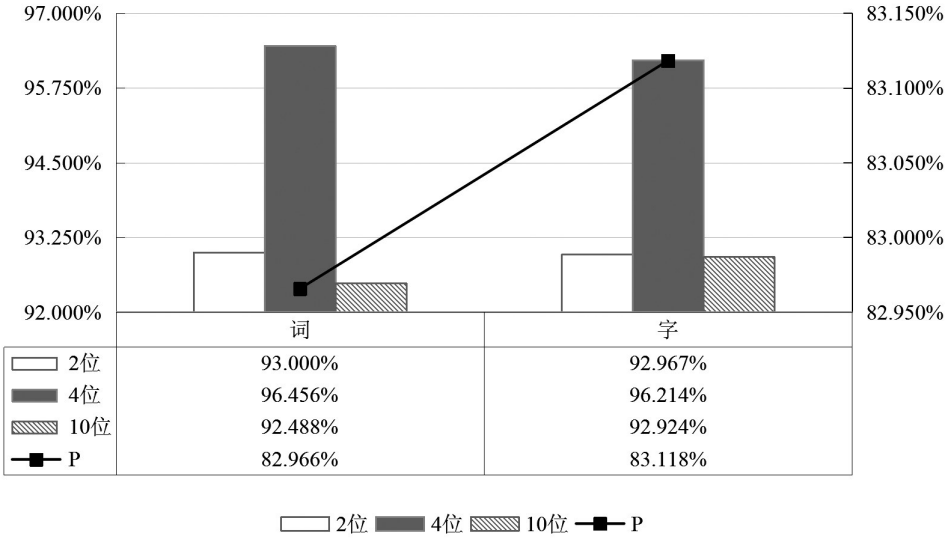

图4 文本粒度的层次实验结果图

总体上看,词粒度的文本与字粒度的文本表现差距很小,在2位与4位的HS编码上词粒度的优势更大,但差距有限;在10位编码的分类问题上,基于字粒度特征的分类优势更为显著,因此叠加后的整体实验里,字特征的总体效果更好,也就是说当每个类数据量有限,或者类之间的相似程度高时,字粒度文本特征的效果优于词粒度文本特征,二者训练时间效率,见表6。

表6 文本粒度训练的时间效率表

特征粒度	词粒度		字粒度	
编码	分支训练速度sec/batch	总训练时间s	分支训练速度sec/batch	总训练时间s
2位	3.11	8 361.86	3.17	7 295.36
4位	3.25	8 731.01	3.24	8 710.02
10位	3.28	6 300.80	3.20	6 140.80

从训练的时间效率上看,训练效率的表现与准确率的表现基本一致,在每个分支上,2位HS编码的词粒度训练更快,但在4位编码与10位编码上,字粒度的速度优势愈发明显;总体上字粒度与词粒度的迭代轮数基本一致,仅在2层编码上词粒度的收敛较慢,因此即使2位编码上词粒度的分支训练速度更快,但由于其收敛迭代轮次更多,因此总训练时间也更长。经过对比字特征的训练准确率与训练效率,较词特征均更优。

4.3 结构深度

经过上一节的实验发现,字粒度文本特征的总体效果优于词粒度文本特征,因此,这里使用字粒度的文本进行实验,受到设备条件限制,构建3层及以上的网络结构已经无法顺利进行训练,因此,这里该值仅为1或2,实验结果,见表7。

在HS编码各层的分类中,不同网络结构深度的表现比较一致,深度越深,效果越差。经过对比具体类目的分类情况,当网络结构变复杂时,不论该类的数据量是否占比更大,其准确率几乎都会显示出一定的下滑,个别类目的准确率出现小幅上升,但实际上,其召回率并没有随之上升。因此,在海关报关商品归类风险的判别中,实验并没有遵循一般模型中深度越深,越能够降低误差,从而提高精度的规律(也有文献认为是出现了过拟合的倾向)。从具体数

表7 结构深度实验结果表

网络结构层数		1				2			
2位	类别	Micro_P	Micro_R	Micro_F1	P	Micro_P	Micro_R	Micro_F1	P
1	85	92.32%	89.86%	91.07%	92.97%	93.03%	90.21%	91.60%	92.93%
2	84	90.64%	94.56%	92.56%		92.53%	91.20%	91.86%	
3	39	95.26%	97.57%	96.40%		92.28%	97.57%	94.85%	
4	90	90.00%	87.33%	88.65%		89.83%	85.71%	87.72%	
5	73	94.07%	99.33%	96.63%		98.00%	98.43%	98.21%	
6	其他	95.64%	89.11%	92.26%		91.74%	94.28%	92.99%	
–	总计	92.99%	92.96%	92.98%		92.90%	92.90%	92.90%	
4位	类别	Micro_P	Micro_R	Micro_F1		Micro_P	Micro_R	Micro_F1	
1	01	96.00%	100.00%	97.96%	96.21%	100.00%	100.00%	100.00%	95.57%
2	02	–	–	–		–	–	–	
3	03	89.47%	73.91%	80.95%		90.00%	78.26%	83.72%	
4	04	92.95%	96.67%	94.77%		97.93%	94.67%	96.27%	
5	05	100.00%	82.35%	90.32%		93.33%	82.35%	87.50%	
…	…	….	…	…		….	…	…	
45	48	97.14%	87.18%	91.89%		97.14%	87.18%	91.89%	
–	总计	67.83%	63.91%	65.81%		62.05%	59.89%	60.95%	
10位	类别	Micro_P	Micro_R	Micro_F1		Micro_P	Micro_R	Micro_F1	
1	100 000	99.44%	100.00%	99.72%	92.92%	98.89%	100.00%	99.44%	91.48%
2	200 000	100.00%	100.00%	100.00%		100.00%	100.00%	100.00%	
3	300 000	84.21%	92.31%	88.07%		83.33%	91.35%	87.16%	
…	…	…	…	…		…	…	…	
13	909 000	88.83%	4.54%	8.63%		85.38%	9.07%	16.40%	
–	总计	91.83%	76.48%	83.45%		90.92%	77.91%	83.92%	

据上分析，本文认为，是出现分类错误的样本从海关报关的文本数据本身已经难以通过模型分辨，而加深模型循环结构深度后，本质上并没有降低网络误差，这样的情况下，应当优先考虑仅有1层结构的模型。

综合HS编码3个层次的分类结果，其总体表现，如图5所示。

该实验的训练时间效率，见表8。

显然由于有2层循环结构的模型在每层HS编码的训练中，效果均差于仅有1层结构的模型，因此整体准确率出现了2%左右的差距。从训练效率上看，显然基于结构深度的理解，循环的增加使原网络需要在原有的输出基础上再增加一个输入输出的结构，其占用的设备资源越多，模型越复杂，使用的时间越长，虽然迭代次数会小于或等于之前的结构，但每个分支的训练时间显著增长，总训练时间大幅增加。因此在海关报关商品的风险判别中，1层的循环结构显然优于2层的结构。

4.4 序列维度

根据之前两节实验的结果，序列维度的实验以字粒度文本以及1层网络结构深度的设置进行，为了确定合理的输入序列维度范围，首先对除去验证集的训练集序列长度进行统计，统计结果，见表9。

可以发现大部分文本的长度在200字符以下，且该长度包含间隔符号，之前默认输入序列维度为200，即当维度为200时，可以包含记录大部分的文本

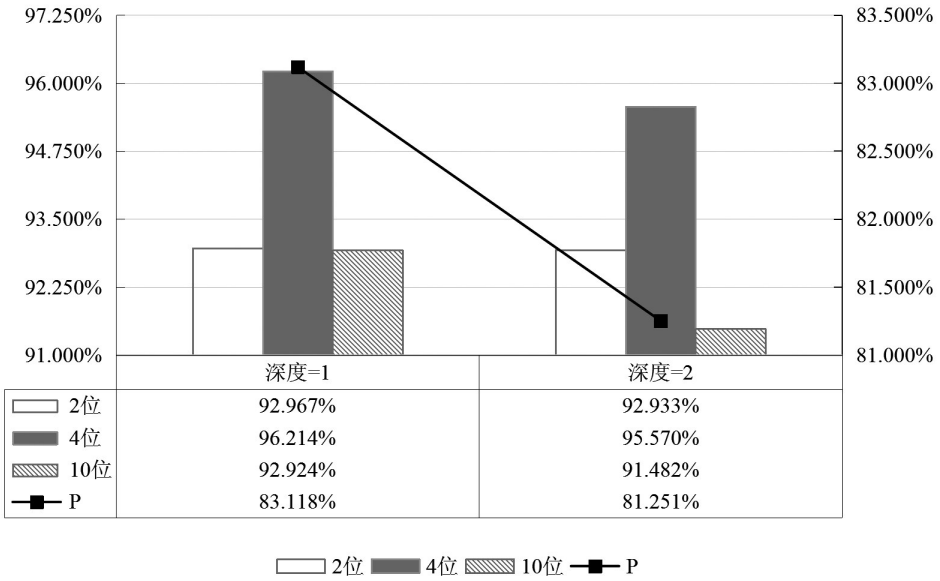

图5 结构深度的层次实验结果图

表8 结构深度训练的时间效率表

结构深度编码	深度=1		深度=2	
	分支训练速度 sec/batch	总训练时间 s	分支训练速度 sec/batch	总训练时间 s
2位	3.17	7 295.36	4.14	7 941.25
4位	3.24	8 710.02	4.26	11 457.41
10位	3.20	6 140.80	3.99	7 155.20

表9 训练集输入序列维度统计表

序列维度范围	2位							4位	10位
	39	73	84	85	90	其他	2位		
1-50	164	347	417	396	166	303	1 793	2 134	964
51-100	1 288	1 575	1 346	1 372	1 404	1 077	8 062	8 283	10 986
101-150	535	353	302	412	478	526	2 606	2 547	2 317
151-200	240	62	67	79	191	198	837	489	267
201-250	100	31	19	16	99	104	369	122	28
251-300	33	3	44	9	54	69	212	55	18
301-350	5	6	11	10	59	21	112	41	4
351-400	1	-	-	-	2	6	9	7	1
序列范围							9~389	11~374	15~366
平均长度	101.66	80.44	80.25	80.75	106.55	105.40	92.73	81.91	81.49
训练样本数	2 366	2 377	2 206	2 294	2 453	2 304	14 000	13 678	14 585
200以下占比							94.99%	98.36%	99.65%

信息。当设置的输入序列长度长于实际值时，模型会补0以保证序列长度的一致性，这可能会对输入信息特征的学习产生影响，且会大大增加训练时间，因此在控制输入长度时，既需要保证能够输入大部分训练数据信息，还需要防止输入过长的情况。根据统计，训练模型中样本的输入长度在9~389维之间，且约95%以上的文本维度均小于200维，当输入维度更小时会损失大量文本信息，因此本节将在200的基础上逐渐增加输入维度，观察输入的序列维度对模型的影响。实验结果，如图6所示。

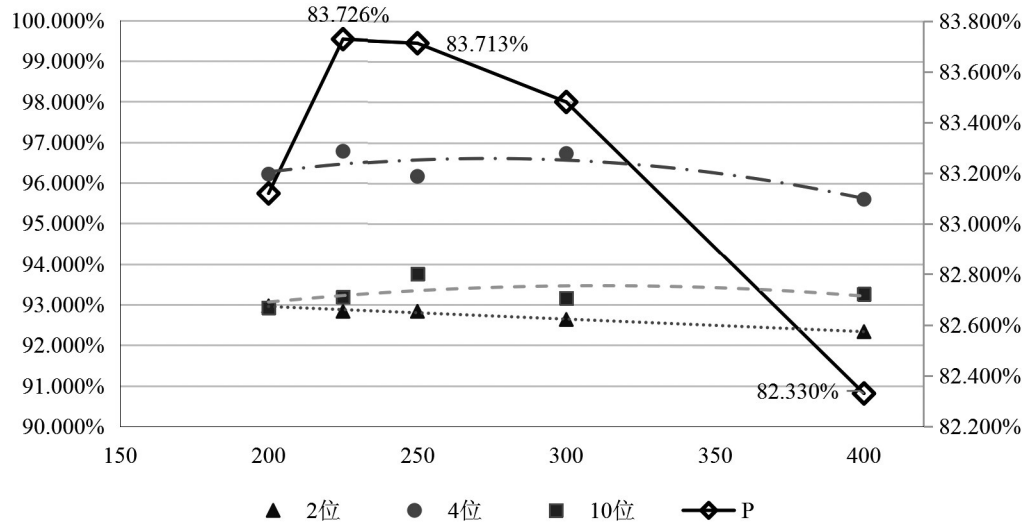

图6 序列维度的层数实验结果图

从实验结果看，每个HS编码层次的实验准确率均维持在相对稳定的水平，上下浮动在1%的范围内，且基本都遵循了识别准确率随着维度的上升先小幅上升，后逐渐下降的总体态势，针对这一现象，这里以2位编码为例观察每个具体类目的准确率变化。可以发现“90”和“39”类的文本平均长度较长，主要分布也向较长的维度偏移，“其他”类包含的商品较为复杂，文本长度平均值也相对较长，相反的“73”“84”和“85”类的文本平均长度相对较短，超过半数的样本长度没有超过100维，其样本的平均序列维度较其他类相差了20维以上。它们在训练中准确率与召回率的调和平均值变化，如图7所示。

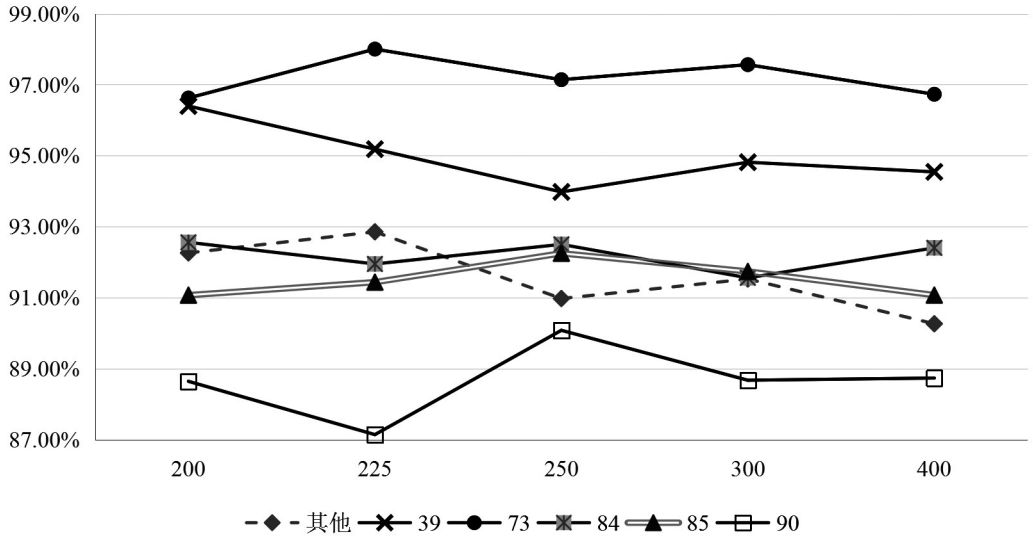

图7 2位HS编码各类目Micro_F值随输入序列维度变化图

从变化趋势看，每个类目的调和平均值随着输入序列维度的变化没有呈现出统一的趋势，均保持着小范围的波动，“39”与“84”类总体出现了小幅的提升，其中“39”类与其他5类之间的差异性较大，且文本长度较短，其准确率随着输入长度的增加发生了下滑；“84”类与“85”类混淆性较大，虽然其总体长度较短，但当输入长度增加时，其效果并没出现有预想中的下滑，而是在一定范围内波动，即文本信息的

截断或不相关信息的混入对其训练效果的影响不显著；与其他几类相比，“90”类的分布最为分散，序列较长的样本也相对较多，因此，随着序列长度的增加，其训练效果出现了小幅的上升。

5　结语

针对当前缺乏对海关报关数据研究的现状，本文利用情报学方法，考虑海关数据特征，构建海关报关商品归类风险识别模型，并结合真实的海关报关商品数据，进行实证研究。同时，本文还将提出的模型同 Fasttext、TextCNN 与 TextRNN 等模型进行对比研究。实验结果表明，本文提出的模型具有一定的优越性，能够较好地对海关报关商品自动分类，并识别风险，从而发挥海关报关商品数据的情报价值。

本文构建海关报关商品归类风险识别模型的思路和实践方法，能够为其他类似的大数据环境下的政务数据处理、利用和发挥其情报价值提供借鉴作用。但本文同时还存在一些不足。首先是数据量有限，本文只获得了 4 个月的海关报关数据，类目与类目间不均衡的分布情况成为模型训练的瓶颈，局限了模型的准确性；其次是对于特征的讨论有限，本文在数值特征的来源中主要参考了历史研究、专家意见与相关性值，未对所有数值特征进行实验。在未来的研究中，本文将持续利用情报学方法，改进模型，从而让海关报关数据更充分的发挥其情报价值。

参考文献：

[1] 段尧清，尚婷，周密．我国政务大数据政策扩散特征与主题分析[J]．图书情报工作，2020，64(13)：133-139.

[2] 苏征，丛凯，陈宏．基于区块链技术在政务大数据中的应用研究[J]．数字通信世界，2020(08)：209-210.

[3] 许留芳，钱华生，夏云青．出口加工区报关单填制和申报的若干问题[J]．对外经贸实务，2014(08)：62-64.

[4] 中华人民共和国海关总署．中华人民共和国海关报关员执业管理办法[EB/OL].[2006-03-20].http://www.gov.cn/gongbao/content/2007/content_588176.htm.

[5] 周欣，张弛海．基于数据挖掘的海关风险分类预测模型研究[J]．海关与经贸研究，2017，38(02)：22-31.

[6] 卢金秋．数据挖掘中的人工神经网络算法及应用研究[D]．杭州：浙江工业大学，2005：42-59.

[7] 宁靓，张卓群，毛万磊．大数据背景下政府网络回应效度研究:以山东政务服务网数据为例[J]．重庆理工大学学报(社会科学)，2019，33(10)：98-109.

[8] 陈平刚，蔡利华，王玲丽．政务大数据支撑的政府舆情预警研究[J]．现代商贸工业，2019，40(34)：141-142.

[9] 谭必勇，陈艳．我国开放政府数据平台数据质量研究：以十省、市为研究对象[J]．情报杂志，2017，36(11)：99-105.

[10] 孙卓林，何云飞．中部省份电子政务文本计量分析：以政策工具为视角[J]．苏州市职业大学学报，2019，30(01)：38-43.

[11] 赵浚吟．大数据视野下抖音政务号用户信息行为研究[J]．江苏科技信息，2019，36(08)：74-77.

[12] 张亦鸣．1996年版《商品名称及编码协调制度》对我国进出口税则的影响[J]．中国海关，1995(02)：27-28.

[13] 王克海．大规模产品生产作业计划作业事项号的自动生成[J]．系统工程理论与实践，1994，14(08)：51-55.

[14] 陈东明，常桂然．基于分段编码自动生成产品结构树的研究[J]．计算机集成制造系统，2005，11(07)：1014-

1018.

[15] WANG J, LEE M C. Reconstructing ddc for interactive classification[C]//Sixteenth ACM conference on Conference on information and knowledge management. New York: ACM, 2007: 137-146.

[16] KOLLER D, SAHAMI M. Hierarchically classifying documents using very few words[C]//Fourteenth international conference on Machine Learning. ICML'97, 1997: 170-178.

[17] ZIMEK A, BUCHWALD F, FRANK E, et al. A study of hierarchical and flat classification of proteins[J]. IEEE/ACM transactions on computational biology & bioinformatics, 2010, 7(03): 563-571.

[18] 王昊, 叶鹏, 邓三鸿. 机器学习在中文期刊论文自动分类研究中的应用[J]. 现代图书情报技术, 2014, 30(03): 80-87.

[19] 谢小楚. 数据挖掘技术在海关缉私系统中的设计与应用[D]. 北京: 北京工业大学, 2007: 1-64.

[20] 严俊龙, 李铁源. 基于SVM的网络安全风险评估模型及应用[J]. 计算机与数字工程, 2012, 40(01): 82-84.

[21] 罗方科, 陈晓红. 基于Logistic回归模型的个人小额贷款信用风险评估及应用[J]. 财经理论与实践, 2017, 38(01): 30-35.

[22] 郭丽丽, 丁世飞. 深度学习研究进展[J]. 计算机科学, 2015, 42(05): 28-33.

[23] 余凯, 贾磊, 陈雨强, 等. 深度学习的昨天、今天和明天[J]. 计算机研究与发展, 2013, 50(09): 1799-1804.

[24] 陈硕. 深度学习神经网络在语音识别中的应用研究[D]. 广州: 华南理工大学, 2013: 1-10.

[25] 卢宏涛, 张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述[J]. 数据采集与处理, 2016, 31(01): 1-17.

[26] 焦李成, 杨淑媛, 刘芳, 等. 神经网络七十年: 回顾与展望[J]. 计算机学报, 2016, 39(08): 1697-1716.

[27] 刘昌伟, 段景辉. 基于因子分析法的海关风险管理评价分析[J]. 海关与经贸研究, 2016, 37(06): 27-42.

[28] 周欣, 张弛海. 基于数据挖掘的海关风险分类预测模型研究[J]. 海关与经贸研究, 2017, 38(02): 22-31.

[29] LI G, LI N. Customs classification for cross-border ecommerce based on text-image adaptive convolutional neural network[J]. Electronic commerce research, 2019, 19(4SI): 779-800.

[30] NODA K, YAMAGUCHI Y, NAKADAI K, et al. Audio-visual speech recognition using deep learning[J]. Applied intelligence, 2015, 42(04): 722-737.

[31] 陆跃平. 《商品名称及编码协调制度》及其公约介绍[J]. 国际贸易, 1992(01): 51-53.

[32] 中华人民共和国海关进出口税则编委会. 中华人民共和国海关进出口税则[M]. 北京: 经济日报出版社, 2012: 2-10.

[33] CHEN Y, LIN Z, ZHAO X, et al. Deep learning-based classification of hyperspectral data[J]. IEEE journal of selected topics in applied earth observations and remote sensing, 2017, 7(06): 2094-2107.

[34] SHEN P, WANG H, MENG Z et al. An improved parallel Bayesian text classification algorithm[J]. Review of computer engineering studies, 2016, 3(01): 6-10.

[35] 陆彦婷, 陆建峰, 杨静宇. 层次分类方法综述[J]. 模式识别与人工智能, 2013, 26(12): 1130-1139.

[36] HINTON G E. Learning distributed representations of concepts[C]//Eighth conference of the cognitive science society.Amherst, Massachusetts:1989.

[37] BENGIO Y,SCHWENK H,SENECAL J S,et al.Neural probabilistic language models[M].Springer Berlin Heidelberg:Innovations in machine learning,2006:137-186.

[38] MATHEW J,RADHAKRISHNAN D.An FIR digital filter using one-hot coded residue representation[C]// 2000 10th European signal processing conference(EUSIPCO).IEEE,2008:1-4.

[39] MIKOLOV T,CHEN K,CORRADO G,et al.Efficient estimation of word representations in vector space [J].Computer ence,2013.

[40] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the association of computational linguistics,2017(05):135-146.

A Study of Intelligence Value and Employment of Political Data in Big Data Environment

——The Risk Avoidance of Customs Declaration Commodities

WANG Hao[1,2],DENG San-hong[1,2],ZHU Li-ping[1,2,3],WANG Xin-yun[1,2],FAN Tao[1,2]

1.School of Information Management,Nanjing University,Nanjing 210023

2.Jiangsu Key Laboratory of Data Engineering and Knowledge Service,Nanjing 210023

3.Nanjing Customs,Nanjing 210001

Abstract: [Purpose / significance]To employ the intelligence value in political big data with Intelligence methods,we take the study of automatic classification of customs declaration commodities as an example to explore the typical applications of large-scale data accumulated over a long period of time in automatic data processing and analysis,so as to effectively reflect the value of intelligence.[Method/process]Extracting the related features of declaration commodity categories,we conduct experiments with deep learning methods. Extracting the relevant feature information of the customs declaration commodities and their categories in the accumulated government affairs data,we use the deep learning method to model and train them, and finally use the intelligence obtained by machine learning to realize the customs declaration commodities of unknown categories automatic classification to achieve the purpose of risk avoidance.[Result/conclusion]We first compare different deep learning text classification models.After analyzing the intelligence,we choose to build a TextRNN model with attention mechanism.The experimental results show that the model has the best performance,can better classify customs declaration commodities to avoid risks,and can more fully explore the intelligence value of customs declaration data.[Limitation]In the experiment,the discussion on risk characteristics is limited.When selecting the characteristics,only historical research,expert opinions and correlation values are referred to.Other effective characteristics may be filtered.

Keywords: political data; risk identification; text classification; TextRNN; declaration commodities; intelligence value; HS code