基于自然语言处理与智能语义识别的舆情监测预警模型研究

张君第
（陕西铁路工程职业技术学院，陕西 渭南 714000）

摘要：做好高校舆情分析与预警具有重要的社会意义，针对传统的网络舆情分析方法依靠人工筛选，费时费力、准确度低且无法进行海量数据分析的问题，基于自然语言处理算法，构建了一种舆情监测预警模型。该模型通过TF-IDF算法对文本特征进行提取，使用基于径向量函数的神经网络模型对数据进行训练，实现舆情分析与预警的功能。数值实验测试结果表明，文中所构建算法模型的精确度指标和效率相较其他对比算法均有明显提高，证明了该算法模型可以对网络舆论进行有效的监测和预警。

关键词：舆情监测；TF-IDF算法；RFB神经网络；Scrapy爬虫框架；自然语言处理；深度学习
中图分类号：TN99 **文献标识码：**A **文章编号：**1674-6236(2022)17-0165-05
DOI：10.14022/j.issn1674-6236.2022.17.035

Research on public opinion monitoring and warning model based on natural language processing and intelligent semantic recognition

ZHANG Jundi
(*Shaanxi Railway Institute*, *Weinan* 714000, *China*)

Abstract: It is of great social significance to do a good job of public opinion analysis and early warning in colleges and universities.In view of the traditional network public opinion analysis methods rely on manual screening, which are time-consuming, laborious, low accuracy and unable to analyze massive data, a public opinion monitoring and early warning model is constructed based on natural language processing algorithm.The model extracts text features through TF-IDF algorithm, trains the data using neural network model based on path vector function, and realizes the function of public opinion analysis and early warning.The results of numerical experiments show that the accuracy index and efficiency of the algorithm model constructed in this paper are significantly improved compared with other comparative algorithms, which proves that the algorithm model can effectively monitor and warn network public opinion.

Keywords: public opinion monitoring; TF-IDF algorithm; RFB neural network; Scrapy crawler frame; natural language processing; deep learning

随着互联网技术的发展，用户数量与日俱增。互联网规模增长的一个重要体现就是社交媒体平台的增加，互联网用户通过社交媒体平台发表自身对某新闻的看法已成为常态，而社交媒体也已成为当前最为重要的舆情采集平台。舆情指的是用户对另外

收稿日期：2021-05-20 **稿件编号：**202105136
基金项目：陕西省职业技术教育学会2021年度规划课题(2021SZXGH12)；中国高等教育学会职业技术教育分会2020年度课题(GZYYB202081)
作者简介：张君第(1984—)，女，河南南阳人，硕士，讲师。研究方向：思想政治教育。

的人、事件或者物体所持有的态度、看法和意见[1-2]。

高校学生为互联网用户的主力，学生群体活跃度较高，上网时间也更长。高校舆情数据具有海量性和突发性两大特征，同时，由于部分学生年龄偏小，心智尚未成熟，而不良信息通常会通过极端主义或者道德绑架等形式散播[3]，学生极易被谣言舆情煽动，更有甚者会受到不良意识形态的影响走向歧途，这会对学生的管理和学校的形象造成负面影响。因此高校需建立舆情监测系统和舆情预警系统，及时发现伪舆情，并进行必要的辟谣和疏导，对高校意识形态的建设具有重要作用。

1 网络舆情分析研究

网络舆情的分析是社会各界密切关注的问题之一。网络舆情分析主要是对舆情文本的情感进行分析，分析时需要对舆情数据进行数学计算，通过一定的数值来判断舆情真伪。

目前常见的舆情分析方法有3种：

1）传统方法。传统的网络舆情分析方法依靠人工检测，大部分算法均是主观算法，例如文献[4]中提到的层次分析算法，该算法使用主观权重因子对舆情的真伪进行分辨，费时费力，仅适用于数据量较少的情形。

2）统计学方法。常见的统计算法为意见领袖模型[5-6]，实际为马尔科夫过程模型。其在所有舆情评论中寻找出影响力最高的用户，将其权重调高，再对所有用户分类，从而实现舆情的监测和预警。

3）深度学习方法。随着机器学习的不断发展，互联网的海量数据已经实现了机器自动化训练，而无需人工干预。如文献[7]中构建的SVM模型，使用基于词向量的神经网络模型对Twitter舆情进行分析和判断。

由此看出，传统方法费时费力且准确性较低，统计学方法准确性较前者有所提高，但无法处理目前的海量数据。而深度学习方法可对海量的数据进行训练，更无需人工干预，其准确性高。因此，该文使用深度学习的相关算法进行舆情模型的构建。

2 网络舆情监测预警模型设计

2.1 模型总体框架

该文构建的网络舆情监测预警模型如图1所示。整个模型分为3个模块：数据爬取、数据预处理和数据分析。数据爬取模块使用数据爬虫脚本，对指定网页的内容按照需求进行爬取，然后存储到某文件中供后续使用；随后使用预处理模块对数据进行预处理，预处理部分使用词向量化算法对抓取到的内容进行归一化处理，主要是去重和去噪，以保证计算机可以识别到文本向量；接着将处理好的数据文件传输至模型分析模块，使用语义关联特征算法对文本内容进行分析，并送入至RBF神经网络模型中进行训练，再对舆情的真伪进行判断；最终，输出判断结果并预警。

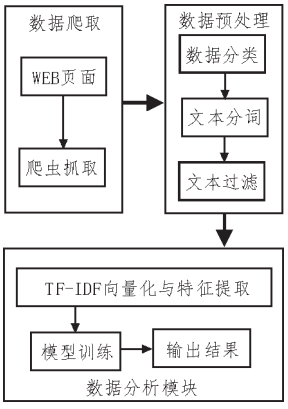

图1 网络舆情监测预警模型

2.2 数据爬取模块

数据爬虫种类繁多，但大部分爬虫的功能是按照一定的规则对互联网的网页信息进行自动探测，高效率的数据爬虫可以有效地采集目标消息。

该文使用的数据爬虫基于Scrapy框架，由该框架搭建的爬虫使用Python语言编写，可以快速地根据用户需求进行网站数据遍历。其与传统爬虫程序不同的是，Scrapy爬虫还可对网站的API数据接口进行爬取，从而大幅提高爬取信息的速度[8-10]。

基于Scrapy框架的爬虫结构包括爬虫脚本主体、爬虫引擎、调度插件、下载模块、爬虫中间件和管道。爬虫脚本主体的目标就是URL地址，爬虫将目标URL地址的内容送入管道中进行存储；爬虫引擎负责内容数据在所有模块中传递；调度插件是将引擎所需的资源请求进行调度；下载模块受爬虫脚本的控制，当爬虫需要下载网页内容时，会调用下载器进行下载。

2.3 数据预处理模块

数据预处理模块分为3个部分，分别为数据分类模块、文本分词模块以及文本过滤模块。

数据分类模块即对采集得来的数据进行标注，

例如负面评论标注a、中性评论标注b、正面评论标注c，这种分类数据作为验证数据集使用；文本分词模块可以使用中文分词脚本，该文使用Jieba第三方分词工具，该工具基于Python语言开发，可以将文本进行准确的切分。此外，Jieba有多种模式，文中使用Jieba.lcut方法，该方法中的cut和HMM参数使用默认值。

2.4　数据分析模块

2.4.1　基于TF-IDF的文本特征提取算法

TF-IDF算法意为词频-逆向文本频率，该算法中的TF为词频，通常用于对某一词语在整个文本出现的频率进行衡量。算法中的IDF为逆文本频率，即在文本中出现次数的倒数。该算法可以表示某一词语在文本中的重要程度[11-12]。TF的计算公式如式(1)所示：

$$\mathrm{TF}_{i,j}=\frac{n_{i,j}}{\sum_{k=1}^{k}n_{k,j}} \tag{1}$$

式中，TF即为词频，$n_{i,j}$为第i个词语在第j个文本中出现的次数，分母为第j个文本中所有词汇的个数。IDF的计算公式如式(2)所示：

$$\mathrm{IDF}_i=\log\frac{n_d}{df\left(d,w_i\right)+1} \tag{2}$$

式中，IDF为逆向文本频率，n_d为所有文本的个数，$df(d,w_i)$为所有文本中包含有特定单词的文本个数。最终的TF-IDF公式如式(3)所示：

$$\mathrm{TF-IDF}_{i,j}=\frac{\mathrm{TF}_{i,j}\times\mathrm{IDF}_i}{\sqrt{\sum w_i\in d_j\left[\mathrm{TF}_{i,j}\times\mathrm{IDF}_i\right]^2}} \tag{3}$$

由式(3)可知，TF-IDF传统算法只考虑了某一特定单词在文本中出现的频率，并未考虑单词所属类别问题，由此会导致在模型训练时对某一冷门类别有贡献的单词丢失。因此还需在TF-IDF算法中加入统计学算法，对单词所属类别问题进行修正。文中加入方差因子，得到改进后的算法如下所示：

$$\gamma_i=\frac{\sqrt{\sum_{j=1}^{N}\left(\frac{df\left(d,w_i\right)}{N}-df\left(d_{c_j},w_i\right)\right)^2}}{N} \tag{4}$$

式中，γ_i为方差因子，N为文本的特征种类数目。可以看到，当某一特殊单词在文本中波动时，γ_i便会发生变化。因此，加入方差因子的TF-IDF算法如下所示：

$$\mathrm{TF-IDF}_{\gamma_i}=\mathrm{TF-IDF}_{ij}\cdot\gamma_i \tag{5}$$

2.4.2　基于径向基函数的神经网络模型

使用神经网络模型可对文本特征数据进行训练。径向基函数也被称为RBF，由该函数组成的神经网络包括输入层、隐藏层以及输出层[13-14]。RBF神经网络模型如图2所示。

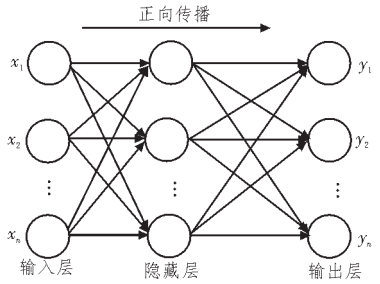

图2　RBF神经网络模型

由图2可知，输入层X为文本数据，数据向量可表示为：

$$\boldsymbol{X}=[x_1,x_2,x_3\cdots,x_n] \tag{6}$$

输出层Y为模型的预测结果，可表示为：

$$\boldsymbol{Y}=[y_1,y_2,y_3\cdots,y_n] \tag{7}$$

隐藏层函数可定义为：

$$R(\boldsymbol{X})=e^{\frac{-\|\boldsymbol{X}-\boldsymbol{C}_i\|^2}{2\delta_i^2}},i=1,2,3,\cdots,m \tag{8}$$

式中，$\boldsymbol{C}_i$为隐藏层中的中心向量；m为隐藏层中神经元的个数；δ_i为隐藏层宽度。

由式(8)可知，输入层神经元和中心向量相隔越远，隐藏层作用函数的值就越低。同时还可以观察出，$\boldsymbol{X}$和$R(\boldsymbol{X})$之间的映射关系属于非线性的。而输出层数据和$R(\boldsymbol{X})$的关系是线性的，则有：

$$y_k=\sum_{i=1}^{m}w_{kp}R(\boldsymbol{X}) \tag{9}$$

式中，w_{kp}为输出向量权重值。按照权重值对输出数据进行排序，即可得到舆情数据的分析结果。

2.5　评价指标

在机器学习领域，常见的模型精度评价指标共有3种，分别为准确率P、召回率R以及F1值[15-16]。准确率是指模型输出结果中正确数据占总数据的比例；召回率是指模型输出结果中正确数据占实际正确数据的比例；而F1值是准确率和召回率的综合计算结果。评价指标的公式如下所示：

$$P=\frac{TP}{TP+FP} \tag{10}$$

$$R=\frac{TP}{TP+FN} \tag{11}$$

$$\mathrm{F1}=\frac{2PR}{P+R} \tag{12}$$

3 实验分析

3.1 数据处理与环境配置

首先使用该文设计的Scrapy爬虫对该校学生在微博、贴吧等社交平台的发言进行爬取，此次共爬取了20 000条学生对于时事热点的发言。其中使用16 000条作为训练样本集，使用4 000条作为测试样本集合。表1为此次测试的数据环境配置。

表1 数据环境配置

参数	项目
训练集/条	16 000
测试集/条	4 000
硬件配置	R5-3600，32 GB

3.2 数据分类

对抓取到的数据进行预警监控，首先需要对数据的主题进行分类。分类后对句子的情感进行判断，筛选出负面消息进行舆情真假判别。

对句子的主题情感进行分类，共筛选出9个与政治相关的舆情话题，按照大类共分为国家安全、政府执政以及社会稳定3个主题。对上述话题按照一定次序排列，如表2所示。

表2 部分数据分类特征

主题	分类编号	内容
国家安全	A1	邪教恐怖主义
	A2	领土完整问题
	A3	外交关系
政府执政	B1	腐败贪污问题
	B2	政策制定
	B3	执政能力
社会稳定	C1	贫富差距
	C2	治安问题
	C3	就业

3.3 算法对比分析

首先对模型的分类能力进行测试，分类数据集合按照表2的主题进行分类。使用训练数据集对模型进行训练，然后对可行性进行验证。

例如，爬虫抓取到的舆情发言为“台湾是中国不可分割的一部分”、“今年就业太难”以及“这项政策对学生是有利的”，将这3句话以编号T1、T2、T3进行指代。模型的分类结果如表3所示。

由表3可知，该文的神经网络模型可以对训练集中的句子进行恰当的内容分类。下面验证舆论情感判断的性能，该文使用其他神经网络模型进行相关指标对比，使用到的对比算法为CNN、KNN和BP神经网络模型。评价指标为准确率、召回率以及F1值。对比测试结果如表4所示。

表3 分类能力验证

编号	分类结果
T1	A2
T2	C3
T3	B2

表4 数据集测试结果

算法	准确率	召回率	F1值
CNN模型	0.70	0.51	0.59
BP神经网络	0.53	0.35	0.421
KNN算法	0.68	0.5	0.629
该文模型	0.75	0.6	0.667

由表4可知，该文模型的准确率、召回率以及F1值三项指标均为最优。在F1值指标中，相较其他算法提高0.077、0.246以及0.038，说明该文算法在舆情敏感话题中有较大优势。

除了对算法准确率进行对比外，还需对算法的运行时间进行分析，进而得到算法的效率。该文以算法训练样本所需时间对算法的效率进行判断，文中训练集合共有16 000条，不同训练样本数量的训练时间如表5所示。

表5 训练时间对比

算法	训练时间/s		
	5 000条	10 000条	15 000条
CNN模型	87.2	159.2	252.3
BP神经网络	142.8	300.2	450
KNN算法	122.2	223.5	345.2
该文模型	77.8	135.2	205.9

由表5可以看出，该文模型在相同样本数量下所需要的训练时间最短，说明该算法同时兼具有高效性。因此，该文模型的综合性能良好，说明所构建的舆情预警模型可以满足设计需求。

4 结束语

高校舆情数据具有海量和突发两大特点，学生极易被谣言舆情所煽动，因此针对高校的舆情管理极为重要。该文针对传统舆情分析方法的不足，基于自然语言技术和深度学习技术设计了高校网络舆情分析预警系统。该系统设计了TF-IDF文本分类算法，同时还使用RBF对数据进行训练。训练测试

结果表明，所设计模型的准确率和效率指标均优于其他对比方法。

参考文献：

[1] 李宗福,李阳,李昂,等.基于Hadoop与机器学习的舆情分析与应用[J].计算机应用研究,2020,37(S1):43-46.

[2] 唐存琛,王椏可.一种结合模型集成的舆情管理模型的研究[J].计算机应用与软件,2019,36(6):31-34,92.

[3] 童玉珍,王应明.基于犹豫模糊集的网络舆情突发事件应急群决策方法[J].计算机系统应用,2019,28(9):9-17.

[4] 覃玉冰,邓春林,杨柳,等.基于决策树的网络舆情类型识别模型研究[J].智能计算机与应用,2018(6):27-32.

[5] 许睿,李艳翠,訾乾龙,等.虚拟学习社区中意见领袖识别模型研究[J].计算机技术与发展,2020,30(5):56-60.

[6] 张媛,陈震,潘尔顺,等.基于自相关观测和隐马尔科夫模型的统计过程监控[J].计算机集成制造系统,2018,24(10):2388-2394.

[7] 甘如饴.基于doc2vec和SVM的舆情情感分析系统的研究与设计[D].北京:北京邮电大学,2017.

[8] Wang D S,Zhang Q B,Hong S X.Research on crawling network information data with scrapy framework[J].International Journal of Network Security,2021,23(2):533-560.

[9] 孙瑜.基于Scrapy框架的网络爬虫系统的设计与实现[D].北京:北京交通大学,2019.

[10]杜鹏辉,仇继扬,彭书涛,等.基于Scrapy的网络爬虫的设计与实现[J].电子设计工程,2019,27(22):120-123,132.

[11]叶雪梅,毛雪岷,夏锦春,等.文本分类TF-IDF算法的改进研究[J].计算机工程与应用,2019,55(2):104-109,161.

[12]Xiao S W,Tong W Q.Prediction of user consumption behavior data based on the combined model of TF-IDF and logistic regression[J].Journal of Physics:Conference Series,2021,175(1):3119-3128.

[13]陈福集,黄亚驹.基于SAPSO_RBF神经网络的网络舆情预测研究[J].武汉理工大学学报(信息与管理工程版),2017,39(4):422-426,438.

[14]Mashor M Y.Modificationos RBF network architecture[J].ASEAN Journal on Science and Technology for Development,2017,17(1):76-83.

[15]金鑫,冯毅,尤雪汐,等.基于机器学习的信息安全设备调配保障技术研究[J].电子科技,2020,33(8):80-86.

[16]张毅,张珉浩.基于机器学习的能力评价与匹配研究[J].计算机工程与科学,2019,41(2):363-369.

（上接第164页）

(12):125-128.

[11]李雄,文开福,钟小明,等.基于深度学习的人脸识别考勤管理系统开发[J].实验室研究与探索,2019,38(7):115-118.

[12]李标俊,姚传涛,杨贵军,等.基于深度学习的变电站轨道自动巡检机器人研究[J].电子设计工程,2020,28(23):68-72.

[13]熊丽华.mjpg-streamer在树莓派上的视频技术应用[J].福建电脑,2020,36(9):101-102.

[14]李濡岩,胡红明,徐建平,等.基于云技术和SSH反向隧道技术的视频监控机器人设计[J].科技与创新,2020(1):54-56.

[15]陈继磊,祁云嵩.基于深度学习的入侵检测方法[J].江苏科技大学学报(自然科学版),2017,31(6):795-800.

[16]吴家俊.基于高速视觉的运动目标检测与跟踪[D].合肥:中国科学技术大学,2019.

(**通信作者：**田会峰，thf830@just.edu.cn)