DOI:10.13718/j.cnki.xsxb.2021.08.015

常见中文社交平台中网络欺凌语言的检测分析①

柳致远[1], 范永胜[1], 张万里[1], 冯 骥[1], 李 勇[2], 黄 靖[1]

1. 重庆师范大学 计算机与信息科学学院,重庆 401331;2. 国网四川电力公司天府新区供电公司,成都 610000

摘要:当今中文社交平台中网络欺凌语言十分盛行,而传统的平台管理员人工审核的方式已无法有效地对其进行检测与分析.为解决这一难题,首先,我们提取了十几个典型的中文社交平台中的部分样本进行人工标注,构建了一个训练数据集.然后,我们分别使用朴素贝叶斯、支持向量机、长短期记忆神经网络构建分类模型,对未标注的数据进行分类识别处理.实验表明:选取的分类模型均能有效地识别出网络欺凌语言,其准确率分别是 0.87,0.79,0.88.其中长短期记忆神经网络综合效果最佳.由此得出的结论为:借助大数据手段建立的分类模型,能快速地检测出社交平台上的原始数据中网络欺凌语言的存在.最后,我们对含有网络欺凌语言的评论与用户等级、发表时间等属性上的相关性做了分析,并拟合出高斯分布模型.

关 键 词:自然语言处理;网络欺凌语言;文本分类;中文社交平台;大数据模型

中图分类号:TP391.1 **文献标志码**:A **文章编号**:1000-5471(2021)08-0086-09

网络欺凌语言是指在社交平台上发表的针对个人或群体的攻击性言论,其攻击性多表现为谩骂、诋毁和嘲笑等[1-6].这类语言的提取、甄别工作一般归自然语言处理领域,而自然语言表述的灵活性、无规律性,使得网络欺凌语言常常难以被发现,从而难以被及时处理.在网络信息量呈爆炸式增长的现在,由平台管理员人工审核用户语言的方式已完全无法胜任网络欺凌语言的检测、分析工作,人工智能和机器学习的引入成为解决这一难题的可行且必要的新型途径.

关于网络欺凌语言,国内外的学者们已开展了大量研究.石国亮等[1]对网络欺凌语言的概念、特点进行了总结论述;在对网络欺凌语言的分析方面,刘文宇等[2]侧重从语言学角度对欺凌语言进行分析,朱嘉珺[3]提出了大数据技术对网络侵害防治的探索;在对网络语言的检测方面,强澜[4]从新浪微博搜集了部分数据,并进行了多次迭代的数据处理,然后建立分类模型以达到检测攻击性语言的目的.鲁倪佳[5]构建了一个网络欺凌公开数据集,并引入了卷积神经网络进行分类,同时研究了数据集平衡问题的解决办法.文献[7-9]借助半人工的方式从 twitter 等社交平台上爬取数据并建立数据集,然后对数据集进行分析,最后通过机器学习或深度学习的方法建立分类模型,再用分类模型检测评价数据集,达到检测出网络欺凌语言的目的.

目前学界研究网络欺凌语言时使用的数据集大多来源于英文数据,少量来自其他语种,如 Van Hee C 等[10]研究了荷兰语的网络欺凌语言检测方法.中文研究相对较少,因为中文处理过程中存在一词多义、词

① 收稿日期:2021-01-12
基金项目:国家自然科学基金——青年基金项目(62003065);重庆师范大学(人才引进/博士启动)基金项目(17XCB008);教育部人文社会科学研究项目(18XJC880002);重庆市教育委员会科技项目(KJQN201800539).
作者简介:柳致远,硕士研究生,主要从事 NLP 和大数据的研究.
通信作者:范永胜,副教授,博士.

向量预训练等问题[11]. 为了解决这些问题，赵雅欣等[12]使用哈工大的分词与停用词表，在数据预处理阶段解决了分词问题；龚静等[13]则是研究多语言统一训练分类模型. 由于我国网民数量庞大，社交平台的网络发言具备了大数据特征，欺凌语言也具备了大数据特征. 在这种背景下，要想和谐网络社区氛围、净化评论语言环境，就必须对社交平台上的网络欺凌语言进行有效的检测与分析.

本文首先构建一个经过人工标注了的中文网络欺凌语言数据集，然后使用基于机器学习与深度学习的方法训练分类模型，并对分类结果进行深入分析. 实验表明，基于深度学习的分类模型效果最佳，结果分析能够挖掘出用户在评论字数、用户等级、发言楼层、评论时间等方面的数据特征.

1　建立数据集

1.1　初始数据的获得与清洗

根据艾瑞数据的社交平台使用人数排行报告[14]，本文选取了排名靠前的百度贴吧、知乎、豆瓣、新浪微博等十几个常见的社交平台，采用后羿采集器爬取到 185.87 万条用户评论，构建了初始数据集. 因爬虫软件获取的数据有许多冗余错乱信息，故本文采用 python 编写的程序进行数据清洗. 首先，删除大量异常值如空值、属性缺失数据、重复爬取数据等，得到 115.51 万条评论，作为网络欺凌语言的分析样本集. 随后为了进一步筛选优质数据以便挑选人工标注样本，本文对评论内容进行去重，以及删除过长与过短的评论. 其中，将过长或过短评论定义为：将所有评论按其长度进行排序后，首尾两端共占 20％的评论. 最后得到 86.24 万条评论，可从中抽取样本组成网络欺凌语言的分类训练样本集.

1.2　分类样本集构建

对网络欺凌语言的检测是一种经典的文本分类问题. 在文本分类问题中，正向样本的数量过少时，分类模型的效果将不明显[15]，为了对比含有网络欺凌语言的攻击性评论和不包含网络欺凌语言的正常评论，本文从样本集中随机选取了正向样本和负向样本各 3000 条左右，通过人工标注的方法，建立了网络欺凌语言分类样本集. 部分经过清洗标注后的样本数据集如表 1 所示.

表中“是否攻击性评论”为人工进行的标注. 通过输入大量经过标注的训练样本进行训练，分类模型能够根据学习到的知识来自动化处理无标注的样本.

表 1　部分网络欺凌语言训练样本集表

序号	用户等级	评论楼层	评论内容	评论时间点	是(1)否(0)攻击性评论
1	12	3	我永远喜欢一花	11	0
2	9	1 055	丽江古城夜景很美	8	0
3	11	537	又是当咸鱼的一天	10	0
4	11	38	好温馨	20	0
5	13	10	可口可乐天下第一	6	0
6	6	163	你真的是个纯脑瘫	23	1
7	10	12	网管这种废物都能干的活你不行建议自杀	16	1
8	12	13	宋拉夫的孝子又双标又贱	12	1
9	9	26	傻卵一个	11	1
10	8	116	啥比韩杂自拍照就是全家福	18	1

2　分类模型

网络欺凌语言样本表现为自然语言形式，而分类模型无法直接处理自然语言，因而需要将文字转化为向量形式，即词向量[11]. Word2vec 是单词向量化的重要方法之一，可以根据给定的语料库，通过优化后的训练模型快速有效地将词语表示为矩阵形式，训练方法分别为连续词袋模型 CBOW(Continuous Bag-of-

Words)模式和跳字模型 Skip-gram 模式[16]. CBOW 模式通过原始语句推测目标字词，比较适合小型数据库，而 Skip-gram 模式从目标字词推测原始语句，在大型语料库中表现得更好. 鉴于本文需要对大量的评论词语进行分类，因此我们采用 Skip-gram 模式进行训练. 分类问题可以采用的模型很多，其中朴素贝叶斯与支持向量机是机器学习中经典的算法[17]，而长短期记忆网络是深度学习中针对股票、文本这样的序列数据提出的模型，很适合用来解决文本分类问题[18].

2.1 朴素贝叶斯(Naive Bayes, NB)

朴素贝叶斯[19] 是常见的分类模型之一，适用于文本分类问题. 对于中文自然语言处理领域而言，朴素贝叶斯算法将词向量中每一个元素看作符合独立性假设的一个特征，对训练集所有特征拟合后，即可通过测试文本的特征判断其属性. 例如：对于评论 X，有 $x \in (x_1, x_2, \cdots, x_n)$，其中 x_n 为词向量的特征，而类别为 $y \in (0, 1)$，0 表示正常评论，1 表示分类模型识别出的攻击性评论. 算法的思想为：根据人工标注的语句构建训练集以及学习训练集的特征，再在测试集中，通过其特征计算评论属于分类(0，1) 的概率，取其中较大者作为分类结果.

概率计算公式为

$$P(y_k \mid x) = P(y_k) * \prod_{i=1}^{n} P(x_i \mid y_k) \tag{1}$$

其中，y_k 为输出类别(y_0，y_1)，$P(y_k \mid x)$ 为该评论属于 y_k 分类的概率，$P(x_i \mid y_k)$ 为在 y_k 分类条件下 x_i 的概率.

2.2 支持向量机(Support Vector Machine, SVM)

SVM 是由模式识别中广义肖像算法发展而来的分类器，基于 SVM 算法的分类策略可以将数据集分类成明确的多个集合[20-22]，SVM 通过某种事先选择的非线性映射将输入向量 x 映射到一个高维特征空间 z，在这个空间中构造最优分类超平面，从而使正例和反例样本之间的分离界限达到最大. 构造出的决策函数为

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{N} a_i^* y_i K(x * x_i) + b^*\right) \tag{2}$$

其中，a 与 b 为偏置系数，x_i 与 y_i 为训练数据，K 为自定义的核函数.

SVM 模型中有 2 个重要的参数，C 与 γ. 其中，C 为惩罚系数，即对误差的宽容度，C 设置得过高容易出现过拟合现象，C 设置得过低会出现欠拟合现象，二者均会导致模型泛化能力变差，效果不够理想；γ 为核函数使用高斯函数时其中的重要参数，γ 决定了低维样本到高维的映射，γ 越大，支持向量越少，γ 越小则支持向量越多，因此它影响着模型训练测试的速度. 本文通过网格搜索，在保证 C 与 γ 相互独立的前提下，寻找全局最优解，设置 C 为 13，γ 为 0.8.

2.3 长短期记忆网络(Long Short-term Memory, LSTM)

长短期记忆网络[23] 是基于循环神经网络(Recurrent Neural Network, RNN) 的一种改进网络，广泛应用于各类问题中[24-25]，针对 RNN 对于长期记忆遗忘的问题，LSTM 在细胞中设置了不同的“门”结构，遗忘门结构(公式 3) 决定在传递到下一个细胞时隐层中的信息是保留还是遗忘，更新门(公式 4) 对 c_t 进行了更新，$c_{(t-1)}$ 中的信息借由 f_t 进行有选择的记忆，输出门(公式 5) 在 c_t 已被更新后，再用一个激活函数决定输出的内容，然后通过 tanh 缩放，即完成一个时间序列的输出.

$$f_t = \sigma(W_{if} x_t + b_{if} + W_{hf} h_{(t-1)} + b_{hf}) \tag{3}$$

$$c_t = f_t c_{(t-1)} + i_t g_t \tag{4}$$

$$h_t = o_t \tanh(c_t) \tag{5}$$

上述公式中，f_t，i_t 与 o_t 为各门的神经元，W_{if} 与 b_{if} 为神经网络的权重，g_t 为新信息，c_t 与 $c_{(t-1)}$ 为当前与前一个的细胞状态，x_t 为输入，h_t 与 h_{t-1} 为当前与前一个的隐层状态和输出.

3 算法实现描述

在选定数据源和分类模型之后，数据的采集、处理和分析过程如图 1 所示：

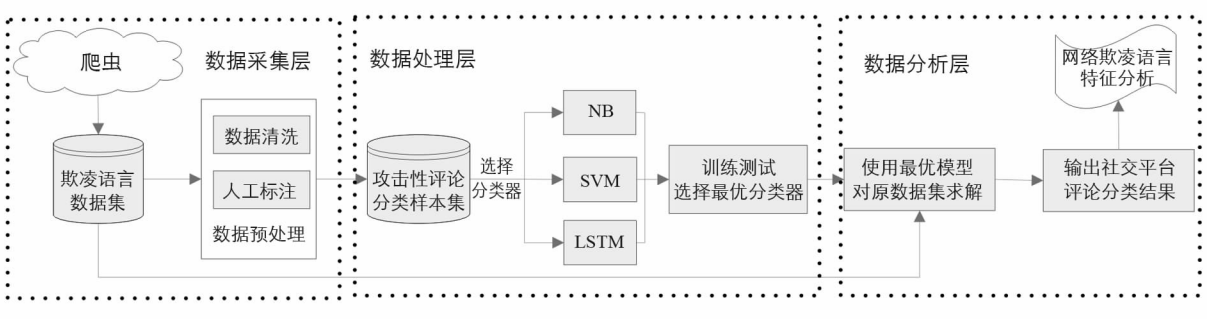

图 1　网络欺凌语言数据流图

1）收集社交平台用户发表的评论数据，构建一个网络欺凌语言数据集；

2）取原数据集中一小部分，进行数据清洗和人工标注，选择使用机器学习中的分类模型进行拟合，建立一个攻击性评论分类样本集；

3）选择 3 种不同的分类算法，分别用分类样本集进行训练，并对比不同的分类模型在测试集上的分类效果，选择其中结果最优的分类模型；

4）使用效果最好的模型，对原数据集中的所有样本求解，得到全部数据的分类结果；

5）基于分类结果对网络欺凌语言的特征进行可视化分析.

表 2 展示了检测社交平台中网络欺凌语言的 3 个算法的伪代码描述，虽然算法的实现有所不同，但基本特点一致：

1）将收集的数据转换成便于处理的形式，比如词向量；

2）分析各自的参数形式与所处理数据属性之间的关系；

3）采用具体算法对数据的各种属性进行处理，形成易于观察的结果；

4）分析结果，研究相应的属性表现出来的特殊性质，如准确度、$F1$ 值等.

表 2　3 个算法流程

朴素贝叶斯算法流程	SVM 算法流程	LSTM 算法流程
1. 对语料库进行分词统计，得到所有的属性 $x=(x_1,x_2,\cdots,x_n)$； 2. 对已标注的数据进行分析，得到所有字词在 (y_0, y_1) 下的概率 $P(x_i\mid y_k)$； 3. 计算测试集中所有语句的 $P(y_k\mid x)$，并取 $\max\{P(y_0\mid x), P(y_1\mid x)\}$ 作为分类结果； 4. 评价分类结果.	1. 用 Word2vec 将输入的文本数据转化为向量形式； 2. 原始问题对偶化为 $\min_a \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}a_i a_j y_i y_j K(x_i, x_j)-\sum_{i=1}^{N}a_i$ $s.t.$ $\sum_{i=1}^{N}a_i y_i=0$；$0\leqslant a_i\leqslant C$，$i=1,2,\cdots,N$； 3. 求解 a^*，b^*，其中 $b^*=y_i-\sum_{i=1}^{N}a_i^* y_i K(x_i\cdot x_j)$； 4. 得到分类决策函数 $f(x)=\mathrm{sign}(\sum_{i=1}^{N}a_i^* y_i K(x*x_i)+b^*)$ 进行分类处理，形成结果.	1. 用 Word2vec 将输入的文本数据转化为向量形式； 2. 数据输入网络进行迭代处理； 3. 基于时间误差反向传播，利用随机梯度下降更新系数； 4. 优化结束，得到网络模型参数，就是处理的结果.

4　实验结果与讨论

在构建好样本集的基础之上，采用上面描述的 3 种方法进行训练，其结果如表 3 所示.

表 3　分类结果对比

分类模型	准确率	召回率	精确率	$F1$
NB	0.87	0.97	0.81	0.88
SVM	0.79	0.87	0.76	0.81
LSTM	0.88	0.87	0.89	0.88

由表 3 可知，长短期记忆网络在构建数据集上的综合效果较好，$F1$ 值达到 88%，同时准确率与精确率达到 88%与 89%. 为此下面的研究将采用长短期记忆网络分类模型来识别检测攻击性语言. 为了进一步检验模型的有效性，我们先选取了百度贴吧中颇具代表性的 bilibili 吧、抗压吧、李毅吧进行验证，统计分类

模型的测试结果，并与人工观察的现象进行对比. 凭借人工观察可知：bilibili吧在人工监督下网友的交流较为友善；抗压吧则相对自由，呈现出较强的攻击性；李毅吧是百度贴吧中用户数量最多的贴吧，其风气呈中性偏多. 而依据构建好的分类模型对这三大贴吧中清洗好的评论进行检测，得到的攻击性语言占比结果如图2所示：bilibili吧与李毅吧的攻击性语言各占16.02%，17.17%，而抗压吧则达到了42.99%. 这一结果与人工观察得出的结论在趋向性上一致，而在精确度上又高于人工观察的结果，从而在一定程度上印证了分类模型的可信度.

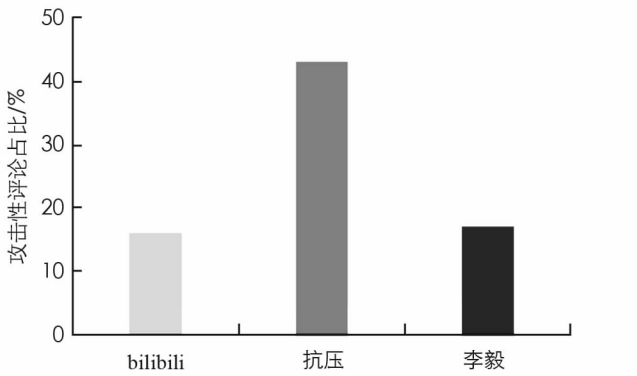

图2 3个贴吧的攻击性评论占比(%)对比图

4.1 评论长度与词性分析

将分类模型分类结果为1的定义为攻击性评论，分类结果为0的定义为正常评论，其评论长度的数量统计如图3所示. 为了公平对比两数据集，我们对攻击性评论的数量进行了一次加权调整，具体为

$$len(\text{sentence})_{new} = len(\text{sentence})_{old} \times num(\text{common}) \div num(\text{aggression})$$

其中，$len(\text{sentence})_{new}$ 为经过加权的语句长度，$len(\text{sentence})_{old}$ 为原语句长度，$num(\text{common})$ 为正常评论的总数量，$num(\text{aggression})$ 为攻击性评论的总数量.

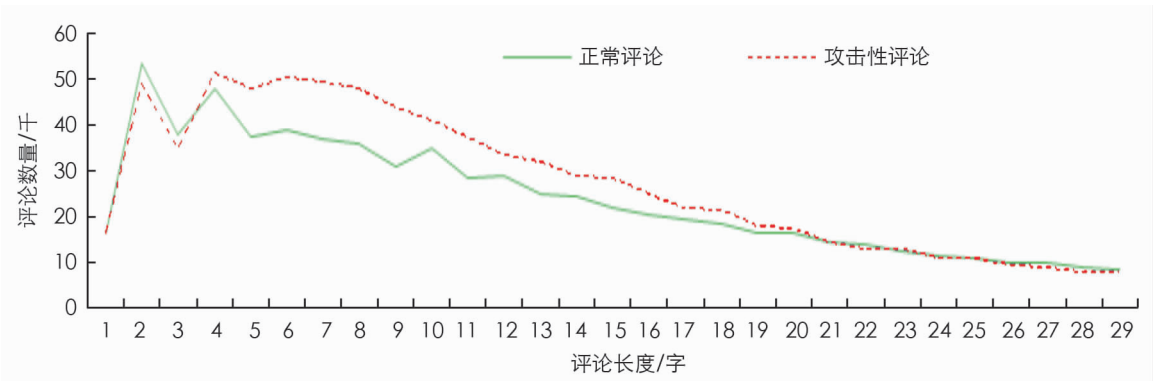

图3 评论长度统计图

由图3可以看出，相对于正常评论，攻击性评论的语句长度在4～20汉字间的数量较多一些. 评论的平均长度与每句评论的不同词性词汇数量统计如表4，表5所示.

表4 评论长度统计表

分类	总长度/字	评论数量/句	平均每句长度(字/句)
正常评论	30 137 560	896 898	33.60
攻击性评论	3 894 467	258 186	15.08

表5 评论不同词性词汇数量统计表 /个

分类	名词	动词	形容词	副词
正常评论	3.59	3.63	0.69	0.02
攻击性评论	1.72	1.73	0.34	0.04

由表4，表5可以看出，攻击性评论平均语句长度只有正常评论的44.88%，同时，每一句话中的名词、动词、形容词数量也要少很多，但副词的使用数量是正常评论的2倍.

4.2　用户等级分析

我们设置用户的等级为横轴，设置参与正常评论/攻击性评论的概率为纵轴，绘制出了柱状分布图. 如图 4 所示.

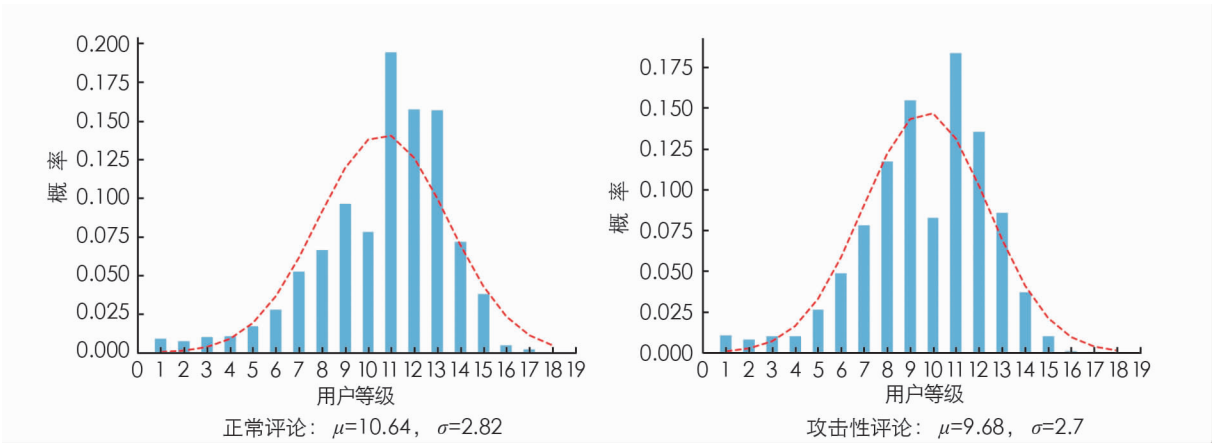

图 4　不同等级用户评论概率分布图

根据图 4 用户等级分布的数据可知，拟合的函数呈现出正态分布特性. 正态分布拟合公式为

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

拟合的各个参数结果如表 6 所示.

表 6　用户等级拟合结果参数表

参数	正常评论	攻击性评论
μ	10.64	9.68
σ	2.82	2.7

由表 6 可以看出，相对于正常评论，攻击性评论的用户等级的分布更为集中(σ 较小)且等级集中的位置较低(μ 较小)，即攻击性评论用户与正常评论用户相比等级集中在较低的位置，同时集中的程度较高.

4.3　评论楼层分析

图 5 为对每一条评论的所在楼层进行统计的分布图.

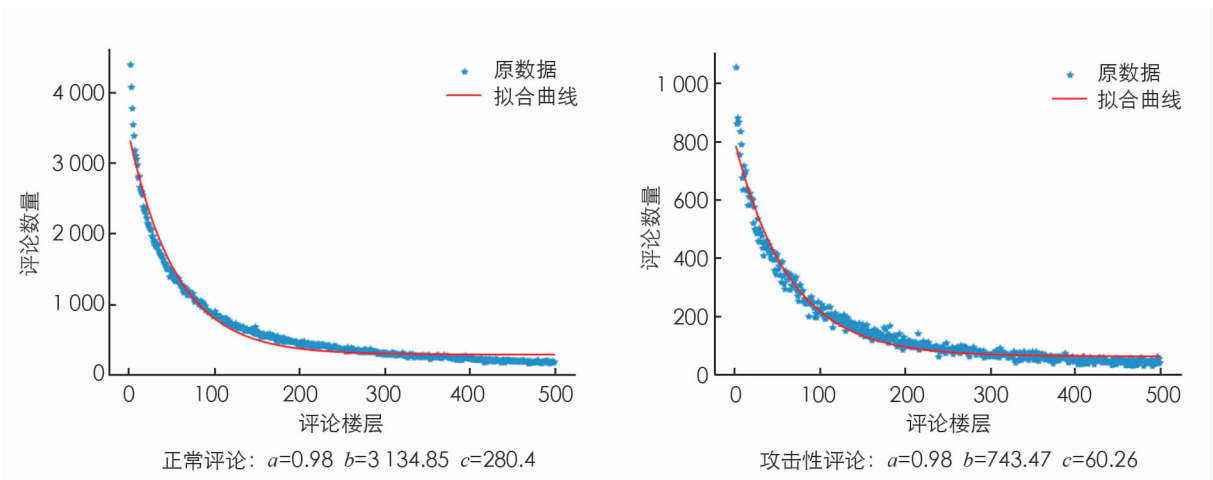

图 5　评论楼层分布图

从图 5 的表现来看，数据呈指数函数曲线形式，其拟合函数为

$$y=b*a^x+c$$

拟合结果如表 7 所示.

表7 评论楼层拟合结果参数表

	a	b	c
正常评论	0.98	3134.85	280.4
攻击性评论	0.98	743.47	60.26

由表7的参数可以看出，攻击性评论相较正常评论而言数量要少一些(c较小)，变化的趋势相差不多(a相近)，但变化的速度较快(b较小)．随着楼层的增加，攻击性评论出现的概率便越小，发表攻击性评论的用户更倾向于在楼层较低时进行攻击．

4.4 评论时间点分析

我们对网友参与评论时间和评论概率的关系进行分析，分别绘制出正常评论和攻击性评论的分布图．如果按0～23时进行观察，很难观察出一定的规律．因此本文将时间属性做了一定的调整，将18～23时的数据调整到0～17时之前，以便更直观地观察出规律．结果如图6所示．

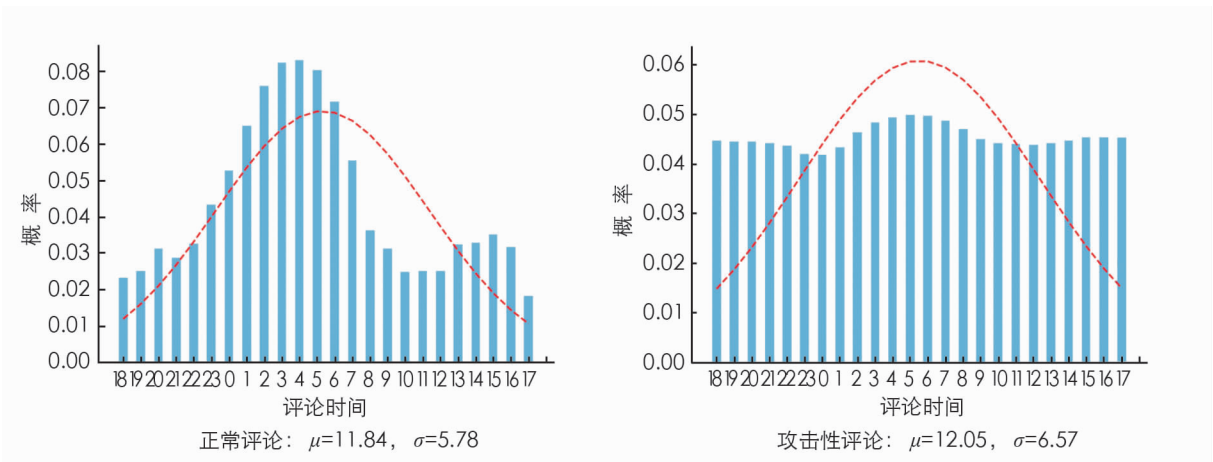

图6 评论时间概率分布图

根据图6，两者的分布图基本呈现正态特性，其正态分布拟合公式为

$$f(x) = 10\ 000 - \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

拟合结果如表8所示．

表8 评论时间拟合结果参数表

参数	正常评论	攻击性评论
μ	11.84	12.05
σ	5.78	6.51

由表8可以看出，相对于正常评论，攻击性评论的时间分布相对离散很多(σ较大)，集中的位置较大(μ较大)．可以解释为攻击性评论呈现出相对不太受时间影响，且攻击性用户熬夜更晚的特点．

5 结 论

本文从常见的社交平台中收集了大量用户评论，清洗后从中选取样本人工标注形成了网络欺凌语言数据集．根据任务特点，选用朴素贝叶斯、支持向量机与长短期记忆网络作为分类模型进行了实验，其中长短期记忆网络综合效果最好．随后使用长短期记忆网络处理未标注的内容，并对结果进行了分析：在3个百度贴吧数据集中，模型分类结果与人工观察结论高度相符，一定程度上验证了模型的可靠性；在全部数据集中，攻击性评论相对于正常评论表现出评论字数较少、用户等级较低、评论时间更离散等分布特征．但是本文仅考虑了传统二分类问题，未对网络欺凌语言的进一步划分作研究，因此下一步考虑使用细粒度

情感分析方法，对网络欺凌现象的成因、发展等因素做深入剖析，从而寻求更有效的检测分析网络欺凌语言的方法.

参考文献：

[1] 石国亮，徐子梁. 网络欺凌的界定及其特点分析 [J]. 中国青年研究，2010(12)：5-8.

[2] 刘文宇，李　珂. 基于批评性话语分析的网络语言暴力研究框架 [J]. 东北师大学报(哲学社会科学版)，2017(1)：119-124.

[3] 朱嘉珺. 大数据视野下的网络侵害防治——一次运用技术解构新型犯罪的探索 [J]. 苏州大学学报(哲学社会科学版)，2019，40(6)：69-76.

[4] 强　澜. 基于社交网络的暴力语言检测研究 [D]. 太原：中北大学，2020.

[5] 鲁倪佳. 面向社交媒体的网络欺凌检测技术研究 [D]. 杭州：杭州电子科技大学，2020.

[6] 俞梅容. 互联网时代的网络语言暴力分析 [J]. 传播与版权，2018(12)：172-173，180.

[7] CHATZAKOU D，LEONTIADIS I，BLACKBURN J，et al. Detecting Cyberbullying and Cyberaggression in Social Media [J]. ACM Transactions on the Web，2019，13(3)：1-51.

[8] AGRAWAL S，AWEKAR A. Deep Learning for Detecting Cyberbullying across Multiple Social Media Platforms [M]// CRESTANI F. Advances in Information Retrieval. Berlin，Germany ：Springer . 2018：141-153.

[9] DADVAR M，ECKERT K. Cyberbullying Detection in Social Networks Using Deep Learning Based Models；A Reproducibility Study [J]. Computation and Language，2018，12：245-255.

[10] VAN HEE C，JACOBS G，EMMERY C，et al. Automatic Detection of Cyberbullying in Social Media Text [J]. PLoS One，2018，13(10)：e0203794.

[11] 李　生. 自然语言处理的研究与发展 [J]. 燕山大学学报，2013，37(5)：377-384.

[12] 赵雅欣，郑明洪，石林鑫，等. 面向电力审计领域的两阶段短文本分类方法研究 [J]. 西南大学学报(自然科学版)，2020，42(10)：1-7.

[13] 龚　静，李英杰，黄欣阳. 基于统计词典和特征加强的多语言文本分类 [J]. 西南师范大学学报(自然科学版)，2018，43(9)：45-50.

[14] 艾瑞数据. PC-Web 社交网络指数 [EB/OL]. (2021-2-13)[2021-2-13]. https：//index. iresearch. com. cn/new/#/pc? cid=3&csid=0.

[15] EMMERY C，VERHOEVEN B，PAUW G，et al. Current Limitations in Cyberbullying Detection：On Evaluation Criteria，Reproducibility，and Data Scarcity [J]. Language Resources and Evaluation，2020，11：1-37.

[16] AYYADEVARA V K. Word2Vec [M]//Pro Machine Learning Algorithms. Berkeley，CA：Apress，2018：167-178.

[17] 杨剑锋，乔佩蕊，李永梅，等. 机器学习分类问题及算法研究综述 [J]. 统计与决策，2019，35(6)：36-40.

[18] 洪　巍，李　敏. 文本情感分析方法研究综述 [J]. 计算机工程与科学，2019，41(4)：750-757.

[19] 李静梅，孙丽华，张巧荣，等. 一种文本处理中的朴素贝叶斯分类器 [J]. 哈尔滨工程大学学报，2003，24(1)：71-74.

[20] CORTES C，VAPNIK V. Support-Vector Networks [J]. Machine Learning，1995，20(3)：273-297.

[21] 杭　立，车　进，宋培源，等. 基于机器学习和图像处理技术的病虫害预测 [J]. 西南大学学报(自然科学版)，2020，42(1)：134-141.

[22] JOACHIMS T. Text Categorization with Support Vector Machines：Learning with many Relevant Features [C]//European Conference on Machine Learning. Heidelberg，Berlin：Springer，2005：137-142.

[23] HOCHREITER S，SCHMIDHUBER J. Long Short-Term Memory [J]. Neural Computation，1997，9(8)：1735-1780.

[24] 王　伟，吴　芳. 基于注意机制和循环卷积神经网络的细粒度图像分类算法 [J]. 西南师范大学学报(自然科学版)，2020，45(1)：48-56.

[25] 林燕榕，张　怡，刘　迪，等. 基于肾病专科电子病历构建肾病医学知识图谱 [J]. 西南大学学报(自然科学版)，2020，42(11)：52-58.

Detection and Analysis of Cybernetics Bullying Language on Common Chinese Social Network Platforms

LIU Zhi-yuan[1], FAN Yong-sheng[1], ZHANG Wan-li[1], FENG Ji[1], LI Yong[2], HUANG Jing[1]

1. School of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China;

2. State Grid Tianfu Electric Power Supply Company, Chengdu 610000, China

Abstract: In order to effectively detect the cyberbullying language on Chinese social platforms, a dozen typical Chinese social platforms are selected, and some samples are extracted from them for manual annotation to construct a training data set. On the basis of the training set, three types of classifiers, i.e. Naive Bayes, support vector machine and long-short-term memory neural network, are used to construct a classification model to classify and recognize unlabeled data. Experiments show that the above selected classifiers can effectively identify cyberbullying language with an accuracy rate of 0.87, 0.79 and 0.88, respectively. Of the three classifiers, the long-short-term memory neural network has the best effect. It is concluded that the classification model established with the help of big data can quickly detect the original data on social platforms and detect the existence of cyberbullying language. Finally, this paper also analyzes the correlation between offensive comment language and user rank, publication time and other attributes, and fits a Gaussian distribution model.

Key words: natural language processing; cyberbullying language; text classification; Chinese social platform; big data model

责任编辑 崔玉洁