第 46 卷 第 6 期 计 算 机 工 程 2020 年 6 月
Vol. 46 No. 6 Computer Engineering June 2020

·开发研究与工程应用· 文章编号：1000-3428(2020)06-0288-08 文献标志码：A 中图分类号：TP393

一种结合文本情感分析的微博僵尸粉识别模型

伍 静，詹千熠，刘 渊

（江南大学 数字媒体学院，江苏 无锡 214122）

摘 要： 社交网站中的僵尸粉群体严重威胁社交平台公信力且增加了社交风险。为准确识别僵尸粉，构建一个基于神经网络的僵尸粉识别模型(Zat-NN)。通过分析微博僵尸粉的社交行为得到高级僵尸粉的行为特征，利用累积分布函数研究僵尸粉与正常用户在行为特征上的差异，并结合卷积神经网络与长短时记忆网络加强微博文本情感分析能力，同时增加日均转发微博数、发博工具和微博情感特征 3 个用户新特征提高 Zat-NN 模型识别准确率及鲁棒性。在新浪微博用户数据集上的实验结果表明，Zat-NN 模型能有效识别高级僵尸粉，提升社交网络用户体验。

关键词： 社交网络；僵尸粉；文本情感分析；卷积神经网络；长短时记忆网络

开放科学(资源服务)标志码(OSID)：

中文引用格式： 伍静，詹千熠，刘渊. 一种结合文本情感分析的微博僵尸粉识别模型[J]. 计算机工程，2020，46(6)：288-295.

英文引用格式： WU Jing，ZHAN Qianyi，LIU Yuan. A zombie fans recognition model for microblog combining text sentiment analysis[J]. Computer Engineering，2020，46(6)：288-295.

A Zombie Fans Recognition Model for Microblog Combining Text Sentiment Analysis

WU Jing，ZHAN Qianyi，LIU Yuan

（School of Digital Media，Jiangnan University，Wuxi，Jiangsu 214122，China）

【Abstract】 The social risk caused by prevalence of zombie fans brings significant threat to the credibility of social platforms. To effectively recognize these zombie fans，this paper proposes a zombie fans recognition model，Zat-NN，based on neural network. First，the social behavior of zombie fans on microblog is analyzed to obtain behavior features of high-level zombie fans. Second，the cumulative distribution function is used to study the behavior feature differences between zombie fans and normal users. Then Convolutional Neural Network(CNN) and Long Short Term Memory(LSTM) network are combined to strengthen the sentiment analysis of microblog texts. At the same time，the number of daily forwarded microblogs，blogging tools and microblog emotion features are added as user features to improve the recognition accuracy and robustness of the Zat-NN model. Experimental results on the user dataset of Sina microblog show that the Zat-NN model can effectively recognize high-level zombie fans，improving user experience of social network.

【Key words】 social network；zombie fans；text sentiment analysis；Convolutional Neural Network(CNN)；Long Short Term Memory(LSTM) network

DOI： 10.19678/j.issn.1000-3428.0055232

0 概述

随着信息时代的发展，人们对互联网的依赖性日益增强，越来越多的人在社交网络上发表自己的看法或者记录自己的生活。社交网络改变了人与人之间交流的方式并且拉近了人与人之间的距离。据 2018 年新浪微博第四季度财报数据显示，截至 2018 年11 月，微博月活跃用户已经增至 4.62 亿，与 2017 年同期相比增长 15%，已成为中国最主要的舆论阵地。由于用户在微博上可以随意发布信息，因此微博在成为人们日渐重要的社交平台的同时也成为垃圾用户发布垃圾信息的平台。已有研究表明微博中垃圾用户的异常行为会显著降低用户体验，增加微博社交风险[1-2]。

目前针对垃圾用户已开展了很多研究，但主要都是针对 Twitter 和 Facebook[3-5]，对微博垃圾用户的研究相对而言起步较晚。垃圾用户包含内容垃圾、僵尸垃圾和封号垃圾等，僵尸垃圾又称为僵尸

基金项目： 国家自然科学基金(61672264)。

作者简介： 伍 静(1997—)，女，硕士研究生，主研方向为社交网络、自然语言处理；詹千熠，博士；刘 渊，教授、博士生导师。

收稿日期： 2019-06-18 **修回日期：** 2019-07-22 **E-mail：** 6181611013@stu.jiangnan.edu.cn

计算机工程 www.ecice06.com

粉、社交僵尸,其多数由程序生成并控制,目的是增加用户粉丝的数量,制造出高影响力的假象。除此之外,部分僵尸粉还会传播各种营销信息,严重威胁社交平台的安全和公信力[6]。由于僵尸粉一直在不断地学习和模仿正常用户的行为,导致新浪微博中现存的初代僵尸粉数目已很少,取而代之的是高级僵尸粉。高级僵尸粉不仅拥有正常的昵称和头像,还会发布原创微博,互相给微博点赞和评论。越来越多高级僵尸粉的出现,给微博僵尸粉账号的检测增加了难度。为准确检测僵尸粉,本文基于深度学习方法构造一个基于神经网络的僵尸粉识别模型(Zat-NN)。

1　基础研究与相关工作

1.1　相关基础研究

情感分析也称为意见挖掘[7],是自然语言处理领域的一个重要方向,其是对带有情感色彩的主观性文本进行分析、处理和推理的过程。最初的文本情感分析方法多数是基于情感词词典,但是人工构建词典需要较大代价,且预测准确率较低。如今,情感分析已逐渐引入深度学习方法,例如文献[8-9]。在将深度学习方法应用于文本分类任务时,需将单词转换为高维分布向量以捕获关于单词的形态、句法和语义信息。词向量之间的集合关系对应词与词之间的语义关系[10]。本文对微博文本做情感分析时采用词嵌入方法。

长短时记忆(Long Short Term Memory,LSTM)网络是一种特殊的递归神经网络(Recursive Neural Network,RNN),RNN 是具有循环结构的网络。与传统 RNN 不同的是,LSTM 能使携带信息跨越多个时间步,保留时刻 t 的记忆并允许过去的信息稍后进入,从而解决梯度消失问题。LSTM 单元拥有输入门、遗忘门和输出门,分别用 $\boldsymbol{i}_t$、$\boldsymbol{f}_t$和 $\boldsymbol{o}_t$表示。图 1 是一个标准的 LSTM 单元结构,其中,$\boldsymbol{x}_t$是当前时刻的输入向量,$\boldsymbol{h}_t$是当前时刻单元的输出向量,$\boldsymbol{h}_{t-1}$是前一个时刻单元的输出向量,$\boldsymbol{c}_t$是存储单元,$\boldsymbol{g}_t$是存储单元前一个时刻的状态。

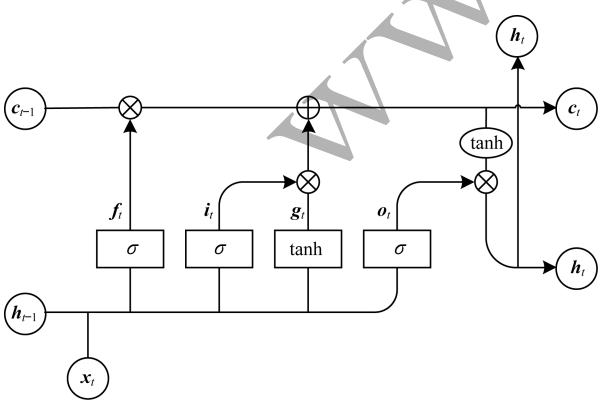

图 1　LSTM 单元结构

Fig. 1　Structure of LSTM unit

1.2　垃圾用户检测研究

水军和僵尸粉为垃圾用户的主要组成部分,两者间有很多相似性,都是通过程序控制微博账号以达到非法传销、盈利等目的,并且对微博水军进行研究可以更好地分析僵尸粉的行为。文献[11]基于贝叶斯模型,通过添加粉丝关注比、平均发布微博数、互相关注数、综合质量评价、收藏数和阳光信用 6 个特征显著提高水军的识别准确率。文献[12]利用 SVM 算法,添加事件参与度、二阶关联性、关系紧密度和引导工具使用率这 4 个新特征,既保证了水军的高识别率,又减少了识别时间。

微博僵尸粉账号检测是一个二分类问题。僵尸粉统称为 spammers,目前国内外已有很多学者对识别僵尸粉进行研究。文献[13]利用用户粉丝数、关注数、博文等特征信息,采用机器学习相关技术从用户中分离出僵尸粉,但其仅是针对 Twitter 上的僵尸粉,对于国内微博僵尸粉的识别有局限性。文献[3]指出合法用户所发的推特通常和个人描述中所写的兴趣爱好相关,并且推特内容具有连续性,通过计算个人描述和推特内容的相似度来区分普通用户和 spammers。文献[14]利用朴素贝叶斯、支持向量机、多层感知器神经网络等分类器对 Twitter 上的用户进行建模和分析,研究发现推特的平均长度、账号年龄、平均发文时长等特征对检测 spammers 有重要作用。文献[15]通过从用户的粉丝中挖掘凝聚子群,结合用户的社会网络关系,提出一种基于用户粉丝聚类现象的僵尸粉检测模型,但该方法需要获取每个用户的全部粉丝信息,实现方法复杂。文献[16]通过构建用户的粉丝数、关注数、微博数、转发数、微博转发情况、微博评论情况、分时段发博数等用户行为特征向量进行僵尸粉识别。文献[17]指出僵尸粉会对传统舆情分析模型造成极大误判,并根据用户个人信息、博文信息实现基于贝叶斯模型的僵尸粉自动判别。

2　问题定义

本节将形式化定义僵尸粉检测,为更好地表达下文内容,给出如表 1 所示的符号定义及其含义。

表 1　符号定义及其含义

Table 1　Definition of symbols and their meanings

符号	含义
C	用户特征集合
U	微博用户集合
n	微博用户数
n'	用户特征数
$\hat{y}$	用户预测标签数
Fans	用户拥有的粉丝数
Follow	用户关注的其他用户数
$\mathrm{FF}_{\mathrm{ratio}}$	用户关注数与粉丝数的比值
$\mathrm{Retweet}_{\mathrm{ratio}}$	用户转发比
$\mathrm{Retweet}_{\mathrm{one}}$	用户日均转发微博数
$\mathrm{Influence}_{\mathrm{w}}$	用户微博影响力
$\mathrm{Original}_{\mathrm{num}}$	用户原创微博数

www.ecice06.com

给定一个社交网络 $G=(V,E)$,其中,G 为一个有向图,$V=\{U\cup C\}$ 表示全部微博用户和用户特征集合,$U=\{u_1,u_2,\cdots,u_n\}$ 表示用户集合,$C=\{c_1,c_2,\cdots,c_n'\}$ 表示用户特征集合,$E\subset\{C\times U\}$ 表示用户和用户特征之间对应关系的集合,每一个用户与其用户特征间的关系为 $e_{ij}=(u_i,c_j)$,表示用户 u_i 具有特征 c_j。本文目标是找到一个预测函数 f 实现全部微博用户的分类:

$$f:(V,E)\rightarrow(Y,Z) \tag{1}$$

其中,Y 和 Z 分别代表正常用户和僵尸粉这两个用户类别集合。

3 微博情感特征

经过研究发现,僵尸粉微博中多数为软文广告和无意义词句等,这与正常用户的微博内容有很大区别。又由于卷积神经网络(Convolutional Neural Network,CNN)可以将长序列转换为由高级特征组成的更短序列,提取有用特征作为 LSTM 网络的输入,因此本文通过添加卷积层来提高情感分析模型的准确率,并基于 CNN-LSTM 神经网络对用户原创微博进行情感分析。

3.1 文本情感分析模型

CNN-LSTM 神经网络模型包含输入层、卷积层、池化层、LSTM 层和分类器层,如图 2 所示。

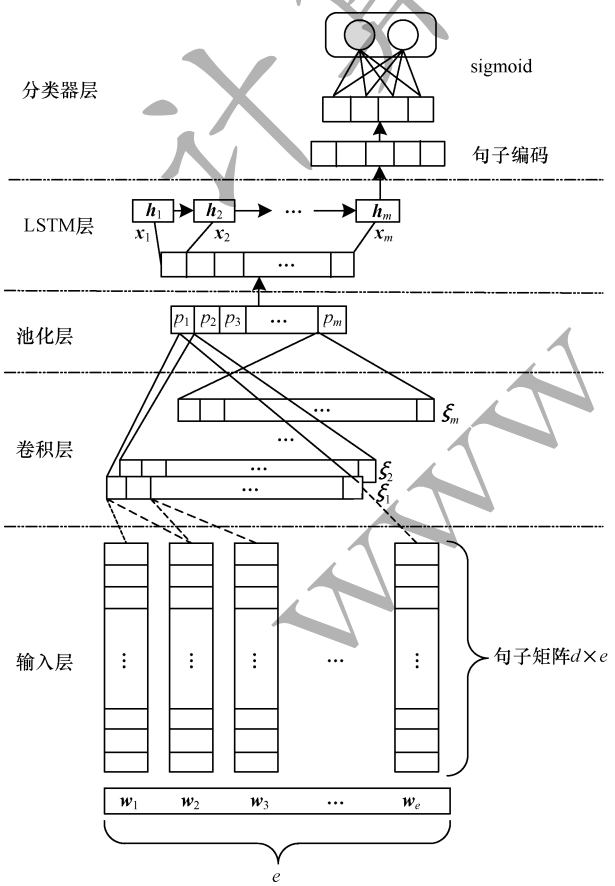

图 2 文本情感分析模型结构

Fig. 2 Structure of text sentiment analysis model

在图 2 中,e 表示输入文本的长度,d 为词向量维数,则 $\boldsymbol{M}\in\mathbb{R}^{e\times d}$ 表示句子矩阵。卷积操作涉及卷积核,通过将卷积核滑动 s 个单位步长进行卷积计算得到局部特征值 $\boldsymbol{\xi}$,当卷积核应用于句子的每个词汇窗口时生成特征图,并在特征图上应用成对最大池操作以捕获最重要的特征,输出特征值 p。池化操作可被视为自然语言处理中的特征选择,$\boldsymbol{x}_1\sim\boldsymbol{x}_m$ 为 LSTM 单元的输入向量,特征值 p 经过 LSTM 层后输出为 $\boldsymbol{h}$ 向量,最终经过 sigmoid 激活函数实现分类。

在将微博文本输入模型前需先对收集到的微博做预处理并删除重复微博。本文用 Jieba 工具去除停用词,通过预训练好的词嵌入模型[18]将文本向量化,在该词嵌入模型中的每个词向量都为 $\mathbb{R}^{300\times1}$ 维。

3.2 模型求解与应用

文本情感分析模型求解的目标是使目标函数最小化,目标函数为:

$$\min\frac{1}{N}\sum_{i=1}^{N}L(f(x;\theta),y)+\lambda J(f) \tag{2}$$

其中,N 为训练样本数目,L 为每个样本的损失函数,$f(x;\theta)$ 为输入 x 时函数的预测输出,y 为目标输出,$J(f)$ 为函数的正则化项。本文采用 Adam[19]优化器来加快该模型收敛速度。此外,为了防止训练时出现过拟合,使用 Dropout[20]方法通过丢弃部分隐藏层神经元来降低过拟合。

采用训练好的文本情感分析模型对每个用户的每条原创微博进行情感预测,将有明显情感特征的微博标记为 1,没有明显情感特征的微博标记为 0,用户原创微博中情感预测标签为 1 的微博所占比例作为用户特征表的第 11 个特征,即 F_{11}。

4 微博非情感用户特征

除了微博情感特征外,本文还提取了 10 个非情感用户特征,具体为粉丝数(F_1)、关注数(F_2)、关注粉丝比(F_3)、用户名特征(F_4)、用户描述(F_5)、微博数(F_6)、转发比(F_7)、日均转发微博数(F_8)、发博工具(F_9)和微博影响力(F_{10}),详见表 2。

表 2 微博僵尸粉识别特征

Table 2 Recognition features of microblog zombie fans

特征编号	特征类别	识别特征	注释
F_1	用户属性	粉丝数[1,21-22]	
F_2	用户属性	关注数[1,13]	
F_3	用户属性	关注粉丝比[1,17]	
F_4	用户行为	用户名特征	简化
F_5	用户行为	用户描述[17]	
F_6	用户行为	微博数[1,21-22]	
F_7	用户行为	转发比[1]	
F_8	用户行为	日均转发微博数	新添加
F_9	用户行为	发博工具	新添加
F_{10}	微博内容	微博影响力	改进
F_{11}	微博内容	微博情感特征	新添加

www.ecocoo.com

由于高级僵尸粉比普通僵尸粉更加难检测,因此本文新添加了日均转发微博数(F_8)和发博工具(F_9)这两个非情感用户特征来提高检测准确率。

4.1　数据获取及特征分析

本节对通过某电商平台购买的 200 个僵尸粉以及微博爬虫爬取的 800 个正常用户进行特征分析。将该 1 000 个用户组成的数据集称为小数据集,并在小数据集上分析并研究用户属性、用户行为和微博内容特征。

4.1.1　用户属性特征

用户属性特征具体如下:

1)粉丝数(F_1)。F_1为某用户的粉丝总数,用 Fans 表示粉丝数,用户 u_j的粉丝数表示为Fans_{u_j},该特征计算公式如下:

$$\text{Fans}_{u_j}=\{\parallel u_i \parallel \mid u_i \rightarrow u_j, u_i \in U, u_j \in U\} \tag{3}$$

其中,$u_i \rightarrow u_j$ 表示用户 u_i关注用户 u_j,Fans_{u_j}为全体 u_i的个数。

2)关注数(F_2)。F_2为某用户关注的其他用户总数,用 Follow 表示关注数,则用户 u_j的关注数表示为Follow_{u_j},该特征计算公式如下:

$$\text{Follow}_{u_j}=\{\parallel u_i \parallel \mid u_j \rightarrow u_i, u_i \in U, u_j \in U\} \tag{4}$$

僵尸粉的关注数远高于正常用户,因为僵尸粉通常通过关注正常用户来获取利益。

3)关注粉丝比(F_3)。F_3用 FF_{ratio}表示关注粉丝比,该特征计算公式如下:

$$\text{FF}_{\text{ratio}}=\frac{\text{Follow}}{\text{Fans}} \tag{5}$$

由于普通僵尸粉关注数很多而粉丝数很少,因此会导致 F_3数值较大。

图 3 为小数据集中的全部用户粉丝数与关注数分布。

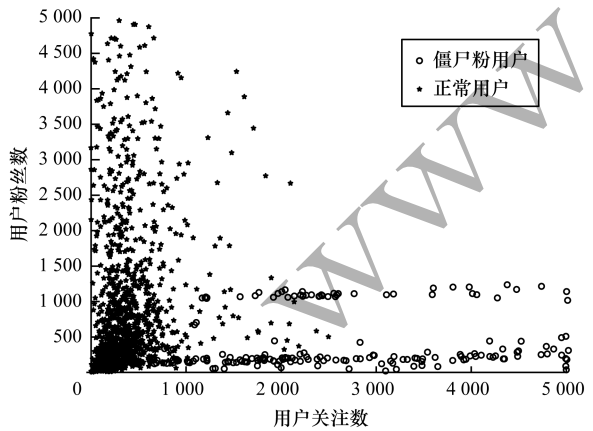

图 3　僵尸粉和正常用户的关注数和粉丝数对比

Fig. 3　Comparison of the number of followers and fans of zombie fans and normal users

可以看出,僵尸粉和正常用户的关注数、粉丝数分布有很大差异。65% 的僵尸粉关注数超过1 000 且粉丝数低于 500 个,属于较低级别的僵尸粉,22.5% 的僵尸粉拥有的粉丝数超过 1 000 个,这与僵尸粉粉丝少的特征相违背。通过对这 22.5% 的僵尸粉研究,发现其粉丝中 90% 以上都是僵尸粉,说明高级僵尸粉通过互相关注的方式来增加双方的粉丝数。僵尸粉将通过这种互相关注行为来大幅降低关注粉丝比值,所以为对这类高级僵尸粉进行准确分类,需要更多有效的用户特征。

4.1.2　用户行为特征

用户行为特征具体如下:

1)用户名特征(F_4)。F_4用于检测某用户的昵称组成结构是否为“用户 + 用户 id 号”,可以快速分类出这类拥有低级昵称的僵尸粉。

2)用户描述(F_5)。F_5为用户注册账号后添加的自我介绍。据统计,小数据集中有约 70% 的僵尸粉没有自我介绍,而正常用户无自我介绍的数量只占 20% ,差异显著。由此可知,F_5可作为分类僵尸粉和正常用户的有效特征。

3)微博数(F_6)。F_6为用户自注册账号以来发布的微博总数。微博总数用 $\text{Weibo}_{\text{num}}$表示,包含原创和转载的微博。

4)转发比(F_7)。F_7用 $\text{Retweet}_{\text{ratio}}$表示,该特征计算公式如下:

$$\text{Retweet}_{\text{ratio}}=\frac{\text{Retweet}_{\text{num}}}{\text{Weibo}_{\text{num}}} \tag{6}$$

其中,$\text{Retweet}_{\text{num}}$为用户转发的微博数,$\text{Weibo}_{\text{num}}$为用户发博总数。转发的微博数衡量了用户与其他用户的互动程度。因为某些僵尸用户只会定期发布原创微博而不会转发他人微博,所以转发比相较于正常用户会比较小。

5)日均转发微博数(F_8)。F_8为本文提出的特征,用 $\text{Retweet}_{\text{one}}$表示用户日均转发微博数,$\text{Retweet}_{\text{day}}$表示用户有过转发行为的天数,则 $\text{Retweet}_{\text{one}}$计算公式如下:

$$\text{Retweet}_{\text{one}}=\frac{\text{Retweet}_{\text{num}}}{\text{Retweet}_{\text{day}}} \tag{7}$$

一些僵尸账号为了制造活跃的假象或者传播营销信息,通常会在较短时间内大量转发指定用户的微博,然而正常用户是有针对性的转发自己感兴趣的微博,所以僵尸粉的日均转发微博数远超过正常用户。图 4 为僵尸粉与正常用户的日均转发微博数的累积分布函数(Cumulative Distribution Function, CDF)对比图,CDF 图不仅能清晰地描述数据的概率分布,而且能直观地体现正常用户和僵尸粉在某个特征上的差异。图 4 中的两条曲线有明显区别,表示日均转发微博数(F_8)在僵尸粉和正常用户之间有较好的区分度。可以看出,几乎所有的正常用户日均转发微博数都很低,而约有 20% 的僵尸粉日均转

www.ecice06.com 计算机工程

发微博数超过 5 条。该特征较 F_7 更具区分度,能更准确地检测出部分僵尸粉。

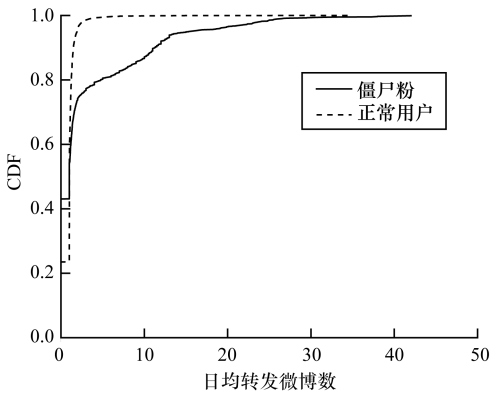

图 4 僵尸粉和正常用户日均转发微博数对比

Fig. 4 Comparison of the number of daily forwarded microblogs of zombie fans and normal users

6)发博工具(F_9)。F_9为本文提出的特征,指用户发布原创微博时的主要工具。若某用户同时用过 weibo. com 和手机客户端发过微博,则计算出每个工具各占的百分比,所占比例最高的工具为用户主要发博工具。对小数据集中的用户发博工具进行统计后发现,98% 的正常用户主要用手机发布微博,只有 2% 的正常用户使用浏览器。而大部分僵尸粉发博工具为浏览器,因为相较于手机微博客户端而言,网页版更容易用程序控制。

4.1.3 微博内容特征

微博内容特征为微博影响力(F_{10}),F_{10}为本文提出的特征,用 Influence_w 表示微博影响力,该特征计算公式如下:

$$\text{Influence}_w = \frac{\text{Like}_{\text{num}} + \text{RR}_{\text{num}} + \text{Com}_{\text{num}}}{\text{Original}_{\text{num}}} \tag{8}$$

其中,Like_{num}、RR_{num}、Com_{num}分别表示用户原创微博所获的点赞总数、转发总数和评论总数,$\text{Original}_{\text{num}}$为用户所发原创微博总数。$F_{10}$简单描述为用户平均每一条原创微博所获点赞、转发、评论数之和。僵尸粉虽然会发布原创微博,但其由于缺乏正常的社交关系,因此所发微博一般不会有用户去评论、点赞,导致 F_{10}值很低。通过将原创微博所获评论数、点赞数以及转发数相结合,可以更有效地区别高级僵尸粉。

5 微博僵尸粉识别模型

本文提出的 Zat-NN 模型基于 BP 神经网络,由输入层、隐藏层和输出层组成。输入层和输出层的神经元个数分别由输入参数及输出参数个数决定,如图 5 所示。

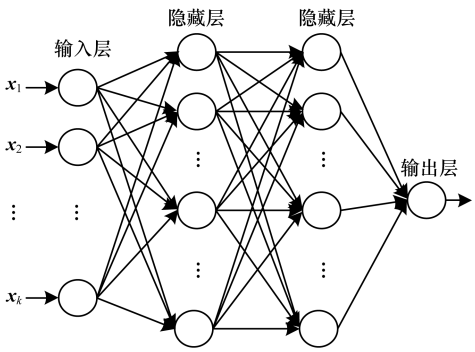

图 5 Zat-NN 模型结构

Fig. 5 Structure of Zat-NN model

Zat-NN 模型结构具体如下:

1)输入层,其中 $\boldsymbol{x}_1 \sim \boldsymbol{x}_k$ 为输入层节点的输入值。在僵尸粉检测中,$\boldsymbol{x}_1 \sim \boldsymbol{x}_k$ 对应于某用户的特征数,若用户有 $F_1 \sim F_{11}$ 全部 11 个特征,则此时 $\boldsymbol{x}_k$ 为 $\boldsymbol{x}_{11}$。

2)隐藏层,Zat-NN 模型共含两层隐藏层,隐藏层节点个数均设为 13 个。

3)输出层,其中 $\hat{y}_i$ 为模型输出结果。由于从微博平台中分离出僵尸粉是一个二分类问题,因此本文使用 sigmoid 函数输出用户分类结果。sigmoid 函数常被用作神经网络的阈值函数,可将变量映射到 0 ~ 1。若用 $S(\boldsymbol{x})$ 表示 $\boldsymbol{x}$ 的输出概率,则:

$$S(\boldsymbol{x}) = \frac{e^{\boldsymbol{x}}}{e^{\boldsymbol{x}} + 1} \tag{9}$$

用户判断是否为僵尸粉的规则如下:

$$\hat{y} = \begin{cases} 1, S(\boldsymbol{x}) > \eta, S(\boldsymbol{x}) \in [0,1] \\ 0, \text{其他} \end{cases} \tag{10}$$

当 $S(\boldsymbol{x})$ 大于 η 时,用户预测标签 $\hat{y}$ 为 1,即将其分类为正常用户;否则分类为僵尸粉。

6 实验结果与分析

6.1 数据集描述

本文设计一个基于 Spider 框架的分布式爬虫,用于爬取新浪微博用户的个人信息及所发微博,微博时间跨度为 2014 年 1 月 1 日——2019 年 4 月 30 日。经过两个星期的爬取,共获得 15 271 个用户信息及 1 535 503 条微博。采用人工方式对爬取到的用户数据进行标注,判别分析得到僵尸粉账号 1 200 个,通过加入网上购买的 800 个纯净僵尸粉账号组成2 000 个僵尸粉样本。为模拟更加真实的微博环境,正常用户与僵尸粉的比例设置为 4∶1。从数据库中提取出 8 000 个正常用户与 2 000 个僵尸粉样本混合组成实验原始数据集。Zat-NN 模型的训练集和测试集用户个数分布如表 3 所示。

计算机工程 www.ecice06.com

表 3　Zat-NN 模型实验数据集

Table 3　Experimental dataset of Zat-NN model

用户数量	训练集	测试集
僵尸粉数量	1 808	192
正常用户数量	7 192	808

CNN-LSTM 情感分析模型的训练集是由没有入选原始实验数据集的用户所发的微博组成。人工对每一条原创微博内容进行情感标注,若微博含有较明显的情感特征,则将其标记为 1,否则标记为 0。最终的微博数据集中共有 10 万条微博,其中有61 424 条标记为 0 的微博,38 576 条标记为 1 的微博。

6.2　数据集预处理

对 Zat-NN 模型实验数据进行预处理,由于将取值范围差异很大的数据输入到神经网络会造成神经网络学习困难并且会使网络过拟合,因此需对除 F_4 和 F_5 这两个特征外的 9 个浮点型类型的特征均做标准化处理,标准化处理公式如下:

$$f_i^* = \frac{f_i - \mu_i}{\sigma_i} \tag{11}$$

其中,f_i^* 为标准化处理后的特征值,$i \in [1,2,3,6,7,8,9,10,11]$,$f_i$ 为未经处理的特征值,μ_i 为该特征值的平均数,σ_i 为该特征值的标准差。处理后的特征平均值为 0,标准差为 1,将 Zat-NN 模型预处理后的实验数据集记为 D_1。需要注意的是训练集和测试集应分开做预处理,对文本情感分析模型微博数据的预处理为删除重复微博无用标签、特殊符号和停用词,然后进行分词。

6.3　实验环境与参数设置

实验运行环境为 Ubuntu 操作系统,2.5 GHz 处理器,8 GB 内存,神经网络模型用 Keras 实现。实验采用十折交叉验证方法来验证模型性能,将数据集的 10% 作为测试集,余下的 90% 随机分成 10 等份互不相交的子集,每次训练都用 9 份子集作为训练集,剩下 1 份子集作为验证集,然后交叉验证重复 10 次。验证准确率取 10 次训练验证结果的平均值。

实验参数设置如下:CNN-LSTM 文本情感分析模型的卷积层卷积核数目为 64,卷积窗口长度为 5,卷积步长设为 1,激活函数采用 ReLU,池化层的最大池化窗口大小为 4,LSTM 层输出空间维度为 70,学习率设为默认值 0.001,dropout 比率设置为 0.2,每次训练 32 个样本;Zat-NN 模型的激活函数采用 ReLU,学习率设为 0.001,dropout 比率设置为 0.2,batch_size 设为 16。

6.4　实验评价指标

微博僵尸账号识别结果的主要评价指标为识别准确性,主要包括正确率(Accuracy)、准确率(Precision)、召回率(Recall)、F1 值、假阳率(False Positive Rate, FPR)、ROC 曲线。本文采用的是 Accuracy、Precision、Recall 和 F1 值这 4 个评价指标。

1)正确率计算公式如下:

$$\text{Accuracy} = \frac{\text{Correct}_{\text{num}}}{n} \tag{12}$$

其中,$\text{Correct}_{\text{num}}$ 为被模型正确分类的用户数,n 为用户总数。正确率越高,模型分类性能越好。

2)准确率计算公式如下:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

其中,TP 为被模型正确分类为僵尸粉的用户数量,FP 为被模型错误分类为僵尸粉的正常用户数量。

3)召回率计算公式如下:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

其中,FN 为被模型错误分类为正常用户的僵尸粉数量。

4)F1 值计算公式如下:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

F1 值为准确率和召回率的加权调和平均,是对准确率和召回率的综合考虑。F1 值越高,模型分类性能越好。

6.5　实验对比模型与结果分析

为证明本文构建 Zat-NN 模型的有效性,将 Zat-NN 模型与文献[16,22]方法进行对比实验,3 种方法的分类结果正确率和召回率如表 4 所示。可以看出,对比文献[16,22]方法,本文 Zat-NN 模型正确率、召回率均最高,证明了本文方法的有效性。另外,分析发现文献[16,22]方法在本文获取到的用户数据集上识别准确率较低的原因为文献[16]方法中添加的用户特征虽然多,但是其中第 9 个至第 12 个用户特征(每天 00:00—06:00 平均发博数、06:00—12:00 平均发博数、12:00—18:00 平均发博数、18:00—24:00 平均发博数)过于冗余,现在的高级僵尸粉不会每天固定时间段发布微博,导致分类效果不明显,而文献[22]方法中添加的用户特征较少,无法全面覆盖现有僵尸粉的特征。

表 4　僵尸粉识别模型对比实验结果

Table 4　Comparative experimental results of zombie fans detection models

模型与方法	准确率	召回率
Zat-NN 模型	0.992	0.991
文献[16]方法	0.936	0.748
文献[22]方法	0.933	0.816

为验证本文提出 3 个新特征的有效性,将 Zat-NN 模型在删除 F_8、F_9 和 F_{11} 特征后的数据集上进行实验,并将该实验结果与添加 F_8、F_9 和 F_{11} 特征后的 Zat-NN 模型实验结果进行对比,如图 6 所示。实验结果表明,模型添加了新特征后正确率、准确率、召回率和 F1 值均有提高,可以看出若使用无新特征的用户集,Zat-NN 模型的各项评价指标均有显著下降,从

计算机工程 www.ecice.com

而说明添加新特征对微博僵尸粉检测的有效性及重要性。

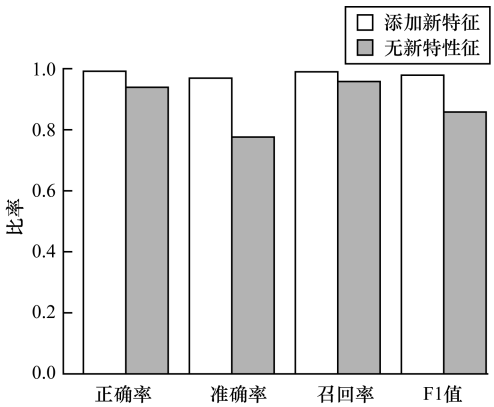

图 6 有无添加新特征的 Zat-NN 模型识别效果对比

Fig. 6 Comparison of the recognition effect of Zat-NN models with or without adding new features

为验证本文提出的微博情感特征(F_{11})的有效性,构建数据集 D_2,数据集 D_2中不包含特征 F_{11}。将 Zat-NN、贝叶斯(Bayes)、支持向量机(Support Vector Machine,SVM)、K 最近邻(K-Nearest Neighbor,KNN)4 种模型分别在数据集 D_1和数据集 D_2上进行实验,实验结果见表 5,其中的 D_1和 D_2分别表示模型在添加了微博情感特征的数据集和在未添加微博情感特征的数据集上进行实验。通过对比发现,在具有同样特征的用户数据集上,Zat-NN 模型比传统机器学习模型识别准确率高,而没有添加微博情感特征的模型评价指标均下降,其中 SVM 模型正确率下降 4%,准确率下降 5.6%,召回率下降 14%,F1 值下降 10.2%,从而验证添加微博情感特征对僵尸粉账号识别的有效性。

表 5 有无添加新特征的 4 种模型识别效果对比

Table 5 Comparison of the recognition effect of four models with or without adding new features

模型	正确率	准确率	召回率	F1 值
Zat-NN-D_1	0.992	0.969	0.991	0.979
Bayes-D_1	0.911	0.965	0.569	0.716
SVM-D_1	0.982	0.968	0.928	0.948
KNN-D_1	0.985	0.966	0.944	0.961
Zat-NN-D_2	0.981	0.940	0.964	0.952
Bayes-D_2	0.841	0.913	0.213	0.345
SVM-D_2	0.942	0.912	0.788	0.846
KNN-D_2	0.960	0.915	0.987	0.896

为证明添加卷积层对情感分析的有效性,将 CNN-LSTM 模型与 Bayes、SVM 和仅含 LSTM 层的模型进行对比。各模型在验证集上的准确率如表 6 所示。由此可知,融合了 CNN 层的 CNN-LSTM 文本情感分析模型准确率最高,比仅含 LSTM 层的模型准确率提高了 4.7%,从而证明添加卷积层对文本情感分析的有效性。

表 6 文本情感分析模型准确率对比

Table 6 Comparison of the precision of text sentiment analysis models

模型	准确率
CNN-LSTM 模型	0.895
Bayes 模型	0.801
SVM 模型	0.788
LSTM 模型	0.848

同时,为验证每个用户特征对模型检测的贡献程度,将神经网络输入层到隐藏层节点的连接权重矩阵的绝对值之和做归一化处理,设 w_i为第 i 个输入层节点权重值,w_{ij}为第 i 个输入层节点的第 j 个隐藏层节点的连接权重值,计算公式如下:

$$\boldsymbol{w}_i = \frac{\sum_{j=1}^{n} |w_{ij}|}{\sum_{i=1}^{m}\sum_{j=1}^{n} |w_{ij}|} \tag{16}$$

其中,m 为输入层节点个数,n 为隐藏层节点个数。在 Zat-NN 模型中,m 为 11,n 为 32。通过计算得到全部特征的权重向量为 $\boldsymbol{W}$ = (0.213,0.060,0.122,0.039,0.070,0.062,0.040,0.052,0.105,0.162,0.074)。各特征的权重分布如图 7 所示,可以看出,本文新添加的微博情感特征、日均转发微博数、发博工具这 3 个特征均具有较大权重,进一步验证了新添加特征对微博僵尸粉检测的重要性。

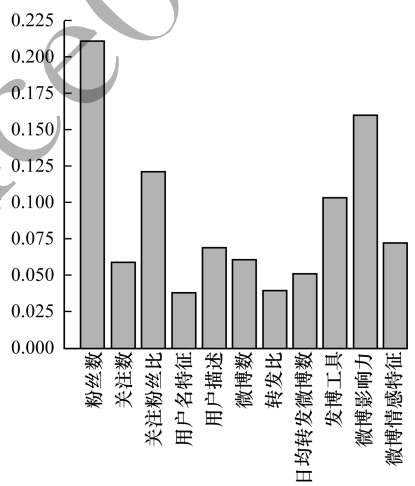

图 7 特征权重分布

Fig. 7 Feature weight distribution

据统计,本文所用的数据集中有 589 个为高级僵尸粉,占全部僵尸粉数量的 29.5%。这些高级僵尸粉均填写了个人资料及个人描述,发博数较多且有转发其他用户微博的行为。随着高级僵尸粉所占比例的提升,僵尸粉检测将会面临更多挑战,因为微博上有一些不活跃用户,他们发博数很少或者发博内容不带有明显个人情绪,微博点赞评论数也很少,识别模型有时会将这类正常用户错误分类为僵尸用户,因此如何准确将这类用户和僵尸粉进行有效区分也是一个难点问题。

www.ecice06.com

7　结束语

微博僵尸粉是微博垃圾用户的主要组成部分。由于僵尸粉的账号大多数由程序生成并控制,其账号安全性能低,容易被攻击者盗用来发送恶意链接、窃取用户的个人隐私,会给个人信息安全和社会公共安全造成严重威胁,因此有效分辨并清除僵尸粉对于提高新浪微博及其他社交网络的用户体验至关重要。本文从用户属性、用户行为、微博内容 3 个方面出发定义 11 个用户特征,并且结合自然语言处理方法对微博文本做情感分析,通过添加微博情感等 3 个新特征增强了模型鲁棒性。实验结果表明,该方法的识别准确率较高,为微博僵尸粉识别提供了一个可行有效的解决方案。下一步将对神经网络识别模型进行多任务学习以捕获更多的僵尸粉用户特征,从而更有效地区分僵尸粉和不活跃用户。

参考文献

[1] ZHANG Yuxiang, SUN Yan, YANG Jiahai, et al. Feature importance analysis for spammer detection in Sina Weibo[J]. Journal on Communications, 2016, 37(8):24-33. (in Chinese)
张宇翔,孙菀,杨家海,等. 新浪微博反垃圾中特征选择的重要性分析[J]. 通信学报,2016,37(8):24-33.

[2] PEI W,XIE Y,TANG G. Spammer detection via combined neural network[C]//Proceedings of International Conference on Machine Learning and Data Mining in Pattern Recognition. Berlin,Germany:Springer,2018:350-364.

[3] ALGHAMDI B,XU Y,WATSON J. A hybrid approach for detecting spammers in online social networks[C]// Proceedings of International Conference on Web Information Systems Engineering. Berlin, Germany: Springer,2018:189-198.

[4] COLLADON A F. Measuring the impact of spammers on E-mail and Twitter networks[J]. International Journal of Information Management,2019,48:254-262.

[5] FAZIL M. A hybrid approach for detecting automated spammers in Twitter[J]. IEEE Transactions on Information Forensics and Security,2018,13(11):2707-2719.

[6] LIU Rong,CHEN Bo,YU Ling,et al. Overview of detection techniques for malicious social bots[J]. Journal on Communications,2017,38(Z2):197-210. (in Chinese)
刘蓉,陈波,于泠,等. 恶意社交机器人检测技术研究[J]. 通信学报,2017,38(Z2):197-210.

[7] ZHAO Yanyan, QIN Bing, LIU Ting, et al. Sentiment analysis[J]. Journal of Software, 2010, 21(8): 1834-1848. (in Chinese)
赵妍妍,秦兵,刘挺. 文本情感分析综述[J]. 软件学报,2010,21(8):1834-1848.

[8] HASSAN A. Deep learning approach for sentiment analysis of short texts[C]//Proceedings of the 3rd International Conferenceon Control, Automation and Robotics. Washington D.C.,USA:IEEE Press,2017:705-710.

[9] RANI S,KUMAR P. Deep learning based sentiment analysis using convolution neural network[J]. Arabian Journal for Science and Engineering,2019,44(4):3305-3314.

[10] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[11] ZHANG Yanmei,HUANG Yingying,GAN Shijie,et al. Weibo spammers' identification algorithm based on Bayesian model[J]. Journal on Communications,2017,38(1):44-53. (in Chinese)
张艳梅,黄莹莹,甘世杰,等. 基于贝叶斯模型的微博网络水军识别算法研究[J]. 通信学报,2017,38(1):44-53.

[12] LI Tao,WANG Yuqiao,XIAO Zhijie. Discovery of features for recognition of social networks spammers[J]. Computer Engineering and Design,2019,40(5):1214-1217,1248. (in Chinese)
李涛,王渔樵,肖智婕. 社交网络水军识别的特征发现[J]. 计算机工程与设计,2019,40(5):1214-1217,1248.

[13] CHU Z,GIANVECCHIO S. Detecting automation of Twitter accounts:are you a human,bot,or cyborg? [J]. Dependable and Secure Computing,2012,9(6):811-824.

[14] HERZALLAH W, FARIS H, ADWAN O. Feature engineering for detecting spammers on Twitter: modelling and analysis[J]. Journal of Information Science,2018,44(2):230-247.

[15] TAO Yongcai, WANG Xiaohui, SHI Lei. Detecting zombies in microblog based on the clustering phenomenon of fans[J]. Journal of Chinese Computer Systems,2015,36(5):1007-1011. (in Chinese)
陶永才,王晓慧,石磊,等. 基于用户粉丝聚类现象的微博僵尸用户检测[J]. 小型微型计算机系统,2015,36(5):1007-1011.

[16] ZHANG Xiying, CHE Xin, TIAN Xianyu. A recognition method of zombie fans on micro-blog user's behavior[J]. Journal of Natural Science of Heilongjiang University, 2014,31(2):250-254. (in Chinese)
张锡英,车鑫,田宪允. 一种基于微博用户行为的僵尸粉识别方法[J]. 黑龙江大学自然科学学报,2014,31(2):250-254.

[17] QIU Xiulian, TIAN Xiaohu, LIAO Wenjian. SEIR microblog public opinion communication model with positive and negative feedbacks[J]. Computer and Modernization,2018(2):44-48. (in Chinese)
邱秀连,田小虎,廖闻剑. 基于正负反馈的 SEIR 微博舆情传播模型[J]. 计算机与现代化,2018(2):44-48.

[18] LI Shen, ZHAO Zhe, HU Renfen, et al. Analogical reasoning on Chinese morphological and semantic relations[EB/OL]. [2019-05-14]. https://arxiv.org/abs/1805.06504? context=cs.

[19] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. [2019-05-14]. https://arxiv.org/abs/1412.6980.

[20] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[EB/OL]. [2019-05-14]. https://arxiv.org/abs/1207.0580v1.

[21] CHEN K,CHEN L,ZHU P D,et al. Unveil the spams in Weibo[C]//Proceedings of 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing. Washington D.C., USA: IEEE Press,2013:916-922.

[22] WANG Yue, ZHANG Jianjin, LIU Fangfang. A multi-feature Weibo zombie powder detection method and implementation[J]. Sciencepaper Online,2014(1):81-86. (in Chinese)
王越,张剑金,刘芳芳. 一种多特征微博僵尸粉检测方法与实现[J]. 中国科技论文,2014(1):81-86.

编辑　陆燕菲

计算机工程 www.ecice06.com