基于三分支神经网络的多特征微博传播预测模型

尹泽惠 王法玉
(天津理工大学计算机科学与工程学院 天津 300384)

摘 要 针对现如今微博传播预测的模型考虑因素不够全面的问题,提出基于三分支神经网络的多特征微博传播预测模型。该模型以三分支神经网络结构为框架,利用 LDA(Latent Dirichlet Allocation)模型提取微博文本特征,利用改进后的 PageRank 算法分析用户影响力特征,并与微博是否带有图片、链接和视频等其他特征相融合。经实验验证,该模型在微博传播预测准确度上较已有双分支模型有显著提高,且稳定性良好。

关键词 三分支神经网络 微博传播预测 LDA 模型算法

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.11.053

MULTI-FEATURE PREDICTION MODEL OF WEIBO PROPAGATION BASED ON TRIPLET NEURAL NETWORK

Yin Zehui Wang Fayu
(*School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China*)

Abstract Nowadays, there are many models for the prediction of Weibo propagation, but the factors are not completely comprehensive. To solve this problem, this paper proposes a multi-feature prediction model of Weibo propagation based on triplet neural network. The basic framework of this model was a triplet neural network structure. In this model, LDA model was used to extract the text features of micro-blog, and improved PageRank algorithm was used to analyze the characteristics of user influence. The model combined with other features such as whether the microblog had pictures, links and videos. Experimental results show that the proposed model significantly improves the accuracy of Weibo propagation prediction compared with two-branch models, which has good stability.

Keywords Triplet neural network Weibo propagation prediction LDA model algorithm

0 引 言

当今时代,互联网发展突飞猛进,微博作为互联网的产品,其使用率也增长迅速。微博用户日活跃数量猛增,在 2020 年已经达到了 2.41 亿。微博用户可以对微博进行点赞、评论、转发等操作。一条微博的传播在宣传正能量事迹、企业的商品营销等方面都起着重要的作用。例如,戴尔公司对外公布数据称,其微博上宣布各种销售信息已经为他们带来了几百万的盈利。由此可见,研究微博传播预测具有重要的现实意义。

现如今,对微博传播预测进行的研究越来越多。刘超等[1]针对微博传播过程中的用户关注网络问题,提出了一种基于微博关注网络的转发预测模型,但该模型没有注重一条微博的多种特征是否影响转发效果。陈振春等[2]提出基于内容和信任度的舆情扩散预测算法对舆情转发规模和扩散深度进行预测,但该模型只考虑了微博文本内容的特征,没有考虑微博内容中其他的多媒体特征如视频、链接、图片和音乐等特征对微博转发的影响。曾辉等[3]运用 LDA 模型提取微博内容特征结合用户关系网络提取间接关注用户权威度特征等多元特征,构建基于双分支神经网络预测模型预测微博传播行为,但该模型复杂度不够并且在训练时没有过多考虑传播热度低的微博样本对预测准确率的影响。

针对上述微博传播预测模型研究过程中的问题,

收稿日期:2021-07-18。**尹泽惠**,硕士生,主研领域:智能信息处理。**王法玉**,副研究员。

本文提出基于三分支神经网络的多特征微博传播预测模型,以微博内容的文本特征、微博用户的影响力、微博是否带有视频、图片、音乐和链接等其他特征作为影响微博传播因素,利用三分支神经网络模型对微博传播进行预测,对现阶段微博传播预测研究中存在的问题做出改进。

1　微博传播预测模型

在研究影响微博传播因素时,本文利用 LDA 模型对微博的文本内容进行特征的提取,改进 PageRank 算法后应用到用户影响力的分析上,并对是否含有视频、图片和链接等其他特征因素进行分析,利用三分支神经网络模型进行训练并对微博传播进行预测。下面进行详细介绍。

1.1　LDA 主题模型

本文对挖掘到的微博文本内容进行特征提取。首先将 jieba 分词应用到微博的文本内容上,使微博的文本内容转化为一个个的分词状态,之后去掉分词中的停用词,最后利用 LDA 模型提取微博文本的主题。

LDA 模型是一种文档生成模型。模型在一篇文档中随机挑选出不同的主题,再从其中挑选出主题所对应的词汇,形成文档的主题词汇。图 1 为 LDA 模型。

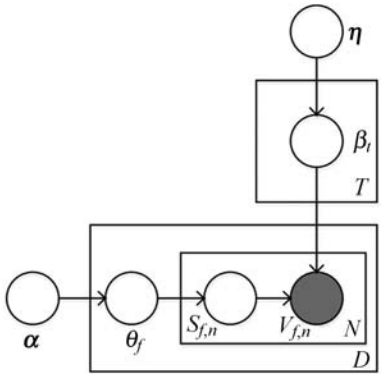

图 1　LDA 模型

首先要了解一篇文档中有多少个主题,将此个数设为 T,其次还需要了解总共有多少篇文档,将总数设为 D。在 LDA 主题模型中,Dirichlet 分布为先验分布。设有一篇文档 f,W 为词汇表中所有词的个数,那么 f 的主题分布为 $\boldsymbol{\theta}_f = Dirichlet(\boldsymbol{\alpha})$,其中,分布的超参数 $\boldsymbol{\alpha}$ 是 T 维的向量。设有一个主题 t,它的词分布为 $\boldsymbol{\beta}_t = \text{Dirichlet}(\boldsymbol{\eta})$,其中,$\boldsymbol{\eta}$ 为 W 维的向量是分布的超参数。令一篇文档 f 中第 n 个词的主题编号 $S_{f,n} = \text{multi}(\boldsymbol{\theta}_f)$,那么该词的概率分布为 $V_{f,n} = \text{multi}(\beta S_{f,n})$。

在 LDA 模型中,$(\boldsymbol{\alpha} \rightarrow \boldsymbol{\theta}_f \rightarrow \boldsymbol{S}_f)$ 组成 Dirichlet-multi 共轭,基于 Dirichlet 分布的文档主题的后验分布可由贝叶斯推理得到。令 $n_f^{(T)}$ 是第 f 个文档中第 T 个主题词的个数,则有 $\boldsymbol{n}_f = (n_f^{(1)}, n_f^{(2)}, \cdots, n_f^{(T)})$,那么 $\boldsymbol{\theta}_f$ 的后验分布为 $\text{Dirichlet}(\boldsymbol{\theta}_f \mid \boldsymbol{\alpha} + \boldsymbol{n}_f)$。同理,令 $n_t^{(w)}$ 是第 t 个主题中第 w 个词的个数,则有 $\boldsymbol{n}_t = (n_t^{(1)}, n_t^{(2)}, \cdots, n_t^{(w)})$,$\boldsymbol{\beta}_t$ 的后验分布为 $\text{Dirichlet}(\boldsymbol{\beta}_t \mid \boldsymbol{\eta} + \boldsymbol{n}_t)$。

本文的 LDA 主题模型使用 Gibbs 采样求解。简化 Dirichlet 表达式:

$$\text{Dirichlet}(\boldsymbol{p} \mid \boldsymbol{\alpha}) = \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{t=1}^{T} p_t^{\alpha_t - 1} \tag{1}$$

式中:$\boldsymbol{\alpha}$ 为 T 维向量,是分布的超参数,$\Delta(\boldsymbol{\alpha})$ 为归一化参数,T 为一篇文档的主题个数,t 是文档的主题。则第 f 篇文档中主题的条件分布为:

$$p(\boldsymbol{s}_f \mid \boldsymbol{\alpha}) = \frac{\Delta(\boldsymbol{n}_f + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \tag{2}$$

式中:f 是一篇文档,$\boldsymbol{n}_f = (n_f^{(1)}, n_f^{(2)}, \cdots, n_f^{(T)})$,$n_f^{(T)}$ 是第 f 个文档中第 T 个主题词的数目。

设在该文档中共有 M 个主题。由此可以计算全部的文档主题的条件概率为:

$$p(\boldsymbol{s} \mid \boldsymbol{\alpha}) = \prod_{f=1}^{M} \frac{\Delta(\boldsymbol{n}_f + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \tag{3}$$

式中:s 是一篇文档中某个词的主题编号,v 是该词的概率分布。则全部主题词的条件分布概率:

$$p(\boldsymbol{v} \mid \boldsymbol{s}, \boldsymbol{\eta}) = \prod_{t=1}^{T} \frac{\Delta(\boldsymbol{n}_t + \boldsymbol{\eta})}{\Delta(\boldsymbol{\eta})} \tag{4}$$

式中:$\boldsymbol{n}_t = (n_t^{(1)}, n_t^{(2)}, \cdots, n_t^{(s)})$,$n_t^{(s)}$ 是第 t 个主题中第 s 个词的个数。

另外,一篇文档的主题分布 $\boldsymbol{\theta}_f$ 以及任一主题的词分布 $\boldsymbol{\beta}_t$ 的后验分布分别为:

$$p(\boldsymbol{\theta}_f \mid \boldsymbol{v}_{\neg i}, \boldsymbol{s}_{\neg i}) = \text{Dirichlet}(\boldsymbol{\theta}_f \mid \boldsymbol{n}_{f,\neg i} + \boldsymbol{\alpha}) \tag{5}$$

$$p(\boldsymbol{\beta}_t \mid \boldsymbol{v}_{\neg i}, \boldsymbol{s}_{\neg i}) = \text{Dirichlet}(\boldsymbol{\beta}_t \mid \boldsymbol{n}_{t,\neg i} + \boldsymbol{\eta}) \tag{6}$$

式中:$\neg i$ 表示的含义是去除下标在 i 之后的词。

经计算,Gibbs 采样的条件概率为:

$$p(s_i = t \mid \boldsymbol{v}, \boldsymbol{s}_{\neg i}) \propto p(s_i = t, v_i = u \mid \boldsymbol{v}_{\neg i}, \boldsymbol{s}_{\neg i}) = E_{\text{Dirichlet}(\theta_f)}(\theta_{ft}) E_{\text{Dirichlet}(\beta_t)}(\beta_{tu}) \tag{7}$$

又有两个期望:

$$E_{\text{Dirichlet}(\theta_f)}(\theta_{ft}) = \frac{n_{f,\neg i}^{t} + \alpha_t}{\sum_{x=1}^{T} n_{f,\neg i}^{x} + \alpha_x} \tag{8}$$

$$E_{\text{Dirichlet}(\beta_t)}(\beta_{tu}) = \frac{n_{t,\neg i}^{u} + \eta_u}{\sum_{y=1}^{W} n_{t,\neg i}^{y} + \eta_y} \tag{9}$$

因此,与主题对应的每个词的条件概率式为:

$$p(s_i = t \mid \boldsymbol{v}, \boldsymbol{s}_{\neg i}) = \frac{n_{f,\neg i}^{t} + \alpha_t}{\sum_{x=1}^{T} n_{f,\neg i}^{x} + \alpha_x} \frac{n_{k,\neg i}^{t} + \eta_t}{\sum_{y=1}^{W} n_{t,\neg i}^{y} + \eta_y} \tag{10}$$

利用全部词和主题的对应关系就能得到不同文档的主题分布 θ_f和词分布 β_t。

1.2 用户影响力模型

1.2.1 PageRank 算法

PageRank 算法[4-5]是针对网页排序问题提出的。这一算法被用来计算和标记一个网页的重要性,衡量一个网站的关注度。一个网站的 PR 值越大则表示该网站的重要性越高,越受欢迎。

一个网页 H 被网页 L 链接,说明网页 H 被网页 L 的所有者判定为比较重要,则网页 L 的一部分重要性就会分给网页 H,该重要性的得分值 S 为:

$$S=\frac{PR(H)}{L(H)} \tag{11}$$

式中:$PR(H)$为网页 H 的 PageRank 值,$L(H)$为网页 H 的出链数。

可以看出 PageRank 算法可以准确方便地计算一个网页的重要性以及该网页的影响力,那么,可以将 PageRank 算法同样地应用到其他领域,比如下面提到的用在计算微博用户影响力领域。但是 PageRank 算法也存在一些问题,比如:PageRank 算法没有进行区分就直接将自身的权值分给链接向它的网址,这些问题在 PageRank 算法应用到用户影响力时需要改进。

1.2.2 改进后的 PageRank 用户影响力算法模型

由于微博用户的网络结构与网页结构有很多相似之处,因此李勇[6]将 PageRank 进行改进,得到了一种新的 MBUInfluence 算法来对用户影响力进行分析。而在本文中,对 MBUInfluence 算法进行了一定的改进:在该算法中计算微博用户影响力[7]部分时,加入了微博用户的会员等级作为其中的一个影响因素,记为 VMBUInfluence 算法。

VMBUInfluence 算法包含两个影响微博用户影响力排名的因素:(1) 用户自身所占的权重,包含用户是否为认证用户、用户的微博总数、用户粉丝数、用户每条微博的转发、点赞和评论数等;(2) 粉丝影响力,主要来源于其粉丝对用户自身所产生的影响。

定义 $SInf(u)$表示由是否为认证用户、用户的微博总数、用户粉丝数、用户每条微博的转发、点赞和评论数等因素决定的用户自身权重。则有:

$$SInf(u)=\left(\frac{a}{T}N_u+\frac{bd}{T}R_u+\frac{be}{T}C_u+\frac{bf}{T}L_u+cV\right)\times M^x \tag{12}$$

式中:a、b、c、d、e 和 f 表示加权系数;T 表示某一时间段;u 表示某一微博用户;N_u 表示用户 u 更新微博的个数;R_u 表示用户 u 转发微博的个数;C_u 表示用户 u 评论微博的个数;L_u 表示用户 u 点赞微博的个数;V 表示微博用户是否被认证;x 为用户的会员等级,M 为常数,本文中的取值范围在 1 到 1.2 之间。

定义 $F(i,u)$表示粉丝用户 i 对微博用户 u 在某一时刻的关注程度,$F(i,u)$的计算表达式为:

$$F(i,u)=\frac{p}{T}R(i,u)+\frac{q}{T}C(i,u)+\frac{r}{T}L(i,u) \tag{13}$$

式中:p、q 和 r 是加权系数;$R(i,u)$表示用户 u 被粉丝 i 转发的微博数量;$C(i,u)$表示用户 u 被粉丝 i 评论的微博数量;$L(i,u)$表示用户 u 被粉丝 i 点赞的微博数量。定义 α_u 表示微博用户 u 的微博被转发的平均概率,则有:

$$\alpha_u=\frac{\sum_{i:(i,u)\in X}F(i,u)}{A_u} \tag{14}$$

式中:X 表示用户 u 的粉丝集合,i 表示 X 中的任一粉丝用户,A_u 表示微博用户 u 的粉丝总数。α_u 与 PageRank 算法中的阻尼系数相类似。

微博用户 u 的影响力值 $VMBUInf$ 计算式为:

$$VMBUInf(u)=SInf(u)+\alpha_u\sum_{i:(i,u)\in X}S(u,i)MBUInf(i)+1-\alpha_u \tag{15}$$

式中:X 表示用户 u 粉丝的集合;i 表示 X 中的任一粉丝用户;$VMBUInf(i)$表示粉丝用户 i 的影响力的值;$S(u,i)$表示粉丝用户 i 将其影响力分给用户 u 的比例,可以有效地解决改进前原 PageRank 算法权重分配不区分的问题,其计算方式为:

$$S(u,i)=\frac{F(i,u)}{\sum_{i:(i,y)\in O}F(y,i)} \tag{16}$$

式中:y 表示粉丝 i 的粉丝;O 表示粉丝 i 的粉丝的集合;$F(y,i)$表示用户 y 对粉丝 i 的关注度。

该算法改进了 PageRank 算法,同时考虑到微博用自身的权重影响以及其粉丝用户对其产生的影响,更加适用于对微博用户影响力进行评估。

1.3 其他特征

除了上述叙述的微博文本特征和微博用户影响力特征外,还有一些其他的微博特征影响着一条微博的传播热度,比如:一条微博的内容是否带有话题、是否带有图片、视频和音乐等多媒体内容、该条微博是否为投票微博、发博时间是否在多数微博用户的活跃时间内、一条微博是否带有其他链接等。本文将这些特征作为微博的其他特征进行研究。

1.4 基于三分支神经网络的多特征微博传播预测模型

本文对微博的传播热度进行预测,预测一条微博在发出后是否会有广泛传播成为热门微博,由此判断该条微博具有多大的影响力。具体的模型如图 2

所示。

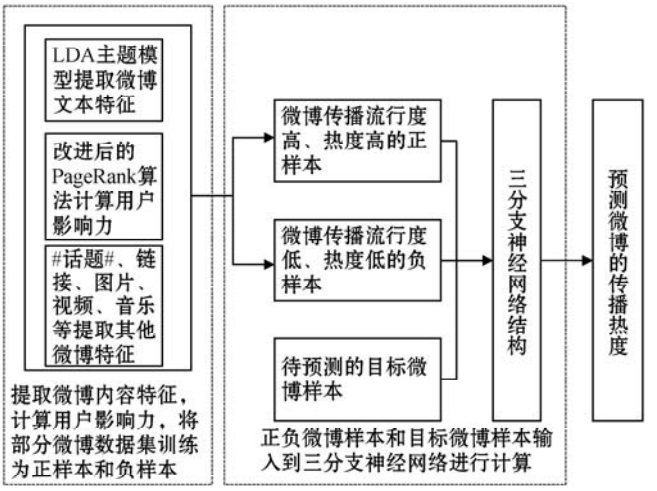

图 2　基于三分支神经网络的多特征微博传播预测模型框架

本文模型对数据集中的微博用户的微博内容利用 LDA 主题模型进行内容特征提取出微博文本内容特征,以及是否带有图片、视频、链接、话题等其他微博特征。同时,利用收集到的微博用户的账号信息对用户的影响力进行评估。融合上述特征分类,传播热度高的数据集作为模型的正样本,传播热度低的数据集作为负样本,正负样本在三分支神经网络模型中进行训练。训练完成之后,将待预测的目标微博样本输入到训练好的三分支神经网络模型当中进行预测,得出结果[8-10]。图 3 所示为三分支神经网络结构模型。

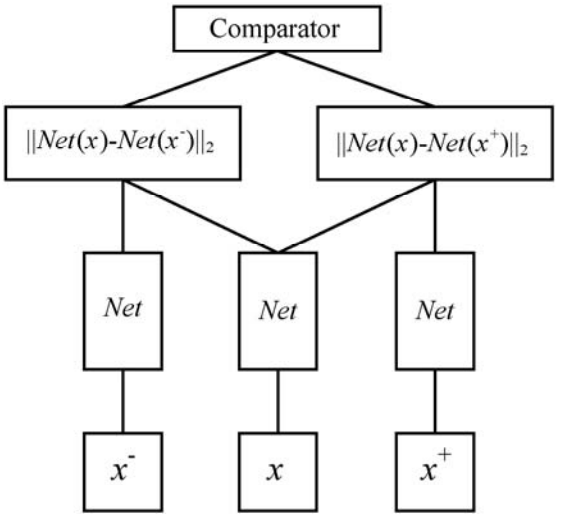

图 3　三分支神经网络模型

可以看出,三分支神经网络每次的输入为三个样本:x 为候选样本,在本文中是待预测传播热度的微博;x^- 为负样本,在本文中是传播热度非常低的微博;x^+ 是正样本,在本文中是传播热度非常高的微博。*Net* 表示三个相同的前馈网络,将三个样本输入进网络之后,网络会输出两个值:候选样本与正样本之间在 embedding 层的距离,候选样本与负样本之间在 embedding 层之间的距离。

$$TrpletNet(x,x^-,x^+)=\begin{bmatrix}\|Net(x)-Net(x^-)\|_2\\ \|Net(x)-Net(x^+)\|_2\end{bmatrix}\in\mathbf{R}_+^2 \tag{17}$$

图 3 中的 Comparator 为比较器。比较器可以对距离向量进行处理,通过比较器中的 loss 函数的训练[11-12],可以让正样本与候选样本之间的距离尽可能的小并且负样本与候选样本之间的距离尽可能大,如图 4 所示。

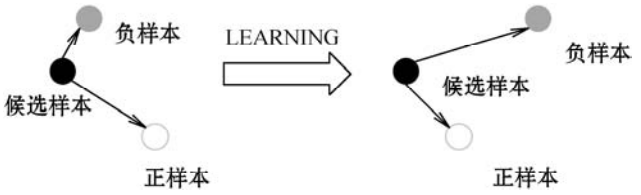

图 4　模型训练示意图

则损失为:

$$Loss(d_+,d_-)=\|d_+,d_--1\|_2^2=c_{onst}\cdot d_+^2 \tag{18}$$

其中:

$$d_+=\frac{e^{\|Net(x)-net(x^+)\|_2}}{e^{\|Net(x)-Net(x^+)\|_2}+e^{\|Net(x)-Net(x^-)\|_2}} \tag{19}$$

并且:

$$d_-=\frac{e^{\|Net(x)-Net(x^-)\|_2}}{e^{\|Net(x)-Net(x^+)\|_2}+e^{\|Net(x)-Net(x^-)\|_2}} \tag{20}$$

式中:d_- 是负样本与候选样本在 embedding 层特征向量的欧氏距离;d_+ 是正样本与候选样本在 embedding 层特征向量的欧氏距离;c_{onst} 为超参数。训练的最终预期是将 d_- 最大化并将 d_+ 最小化,即令 $Loss(d_+,d_-)\to 0$。

2　实　验

2.1　实验数据集

本文采用的实验数据集是由清华大学的科技情报大数据挖掘与服务平台 AMiner 发布的新浪微博社交媒体网络数据集。数据集主要包括用户名、性别、验证状态、城市、所在地、用户粉丝和粉丝的粉丝等用户特征信息和用户的微博内容、微博转发数、微博评论数、微博点赞数等微博信息。数据集中共有 170 万用户,平均每位用户约有 200 位粉丝,收集每位用户的最新的 1 000 条微博,总共 10 亿条微博,采样数据如表 1 所示。

表 1　数据集说明

数据集	用户	粉丝数	源微博	转发
微博	1 776 950	308 489 739	300 000	23 755 810

2.2　评估标准

本文采用准确率(Precision)、召回率(Recall)和

F1 值(F1-Measure)来评价微博传播预测模型的效果。

预测模型的混淆矩阵如表 2 所示。

表 2 预测模型混淆矩阵

真实情况	预测结果	
	正例	反例
正例	T_P(真正例)	F_N(假反例)
反例	F_P(假正例)	T_N(真反例)

用来评价模型准确性的准确率的计算式为:

$$P_{\text{recision}} = \frac{T_P}{T_P + F_P} \tag{21}$$

评价模型是否全面的召回率的计算式为:

$$R_{\text{ecall}} = \frac{T_P}{T_P + F_N} \tag{22}$$

F1 综合度量准确率和召回率性能,其计算式为:

$$F_1 = \frac{2 \times P_{\text{recision}} \times R_{\text{ecall}}}{P_{\text{recision}} + R_{\text{ecall}}} \tag{23}$$

在评价实验模型的好坏时,准确率、召回率和 F1 得分的数值越高,则说明模型的效果越好。

2.3 实验结果分析

本文模型将与其他三个模型进行实验对比:基于双分支融合多特征的微博预测算法通过对原始微博进行分析,运用 LDA 模型算法提取内容特征、构建用户关系网络提取间接关注用户权威度特征等多元特征,构建基于双分支结构神经网络模型预测微博传播行为[3],该算法记为“Double Branch”。基于循环预测网络的转发预测模型通过微博特征、用户特征、微博文本与粉丝兴趣的相似度、转发趋势度与 LSTM 和 DNN 神经网络的优势相结合来建立预测模型[13],该模型记为“SIM-LSTM”。多特征神经网络转发预测模型利用 BP(Back Propagation)神经网络作为预测模型,从发布用户、转发用户、微博文本与用户兴趣相似度 3 个方面切入做特征提取,输入预测模型,得出微博转发概率[14-15],该模型记为“Multi-Feature”。对上述提到的四个模型用相同的数据集进行实验,对微博的传播进行预测。

本文模型利用随机梯度算法数据集进行训练,将初始的学习率设置为 0.5,并将学习率进行指数衰减。令 Momentum(动量梯度下降法)的参数 $\beta = 0.9$。使用 $p = 0.5$ 的 dropout 正则化技术避免过拟合。对数据集进行 500 个 epochs 训练后模型达到了固定的误差。经实验验证,本文模型的激活函数采用 Softmax 函数模型的效果更好。

Multi-Feature 模型的激活函数使用 Sigmoid 函数。初始学习率设置为 0.01,学习率衰减。epoch 次数设置为 300 左右。dropout 率设置为 0.5。预测模型采用梯度下降的策略,即按照负梯度方向对参数进行调整。SIM-LSTM 模型激活函数使用 Softmax 函数,设置初始学习率为 0.1,学习率指数衰减法,衰减系数 0.95。dropout 率设置为 0.5。epoch 次数设置为 300 左右。Double Branch 算法模型激励函数使用 ReLU 函数,优化方法选择 Adam 算法,dropout 率为 0.5,设置学习率 0.01,学习率衰减。epoch 次数设为 1 000。

四种微博传播预测模型的实验结果如表 3 所示。可以看出,本文所提到的基于三分支神经网络的多特征微博传播预测模型的准确率达 80.6%,召回率达 80.7%,F1 值达 80.6%,明显高于其他模型。

表 3 模型准确性对比(%)

模型	Precision	Recall	F1-Measure
Multi-Feature	76.5	76.3	76.4
SIM-LSTM	77.3	77.3	77.3
Double Branch	79.4	79.2	79.2
本文模型	80.6	80.7	80.6

可以看出本文模型的准确率、召回率和 F1 值三个评价指标的数值与其他三个模型的数值相比有所提高。该模型的复杂度最高,引入了三分支神经网络模型,将传播热度高的微博作为正样本,将传播热度低的微博作为负样本,将准备预测的微博作为候选样本进行训练,可以提升预测的准确度。另外,本文选取的微博特征因素较为全面,考虑了微博的文本特征,微博用户影响力特征与微博内容是否含有图片、链接等其他微博特征,增强了模型的性能。

为了更全面地验证本文模型所采用的神经网络框架以及模型所采用的微博特征的效果,本文选取了四类不同的微博特征对四个模型分别进行实验,然后对比四个模型在不同特征值下的 F1 值,结果如图 5 所示。

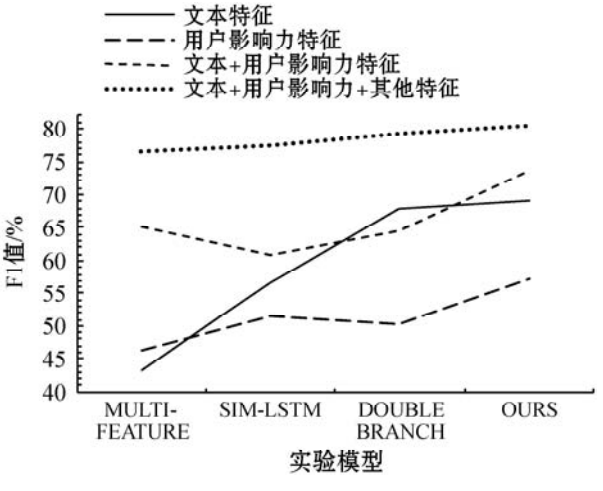

图 5 不同特征值对模型 F1 值影响对比

可以看出不同特征的选取对四个实验模型均产生较大的影响:微博文本内容特征、微博用户影响力特征作为单个特征值进行实验时,四个模型的实验结果均不理想,这是因为造成一条微博的广泛传播的影响因素由多种因素决定,所以在进行微博传播预测时,选取的特征应该尽可能丰富并且全面。图 5 中可以看到选用单一微博的文本特征的实验结果好于选用单一微博用户影响力特征,这说明无论是影响力多大的用户,其微博能够有较高传播的因素中微博本身内容占比较大。另外从图 5 中可以看出在对单一微博文本内容特征进行实验时,Double Branch 模型和本文所用模型的效果相似,这是由于两个模型对于文本内容的处理都用到了 LDA 模型。

由于在不同的时间段里,微博用户的活跃情况不一样,所以实验模型在不同时间段的效果也不同。图 6 为不同时间段用户的活跃情况变化情况统计。

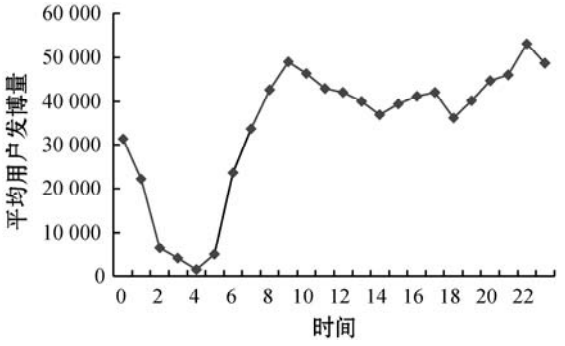

图 6　微博用户不同时刻活跃度统计

可以看出,从 0 时开始,用户活跃度开始下降,4 时过后用户活跃度开始上升,到 9 时用户活跃度到达白日最高峰,接着曲线下降并从 14 时开始有小幅度回升,18 时用户活跃度到达又一极小值,之后开始上升至 22 时到达全天活跃度最高点。4 时、18 时为全天活跃度的极小值点,9 时、22 时为全天活跃度的极大值点。本文选择 4 时、18 时、9 时和 22 时四个时刻对四个模型进行召回率对比,如图 7 所示。

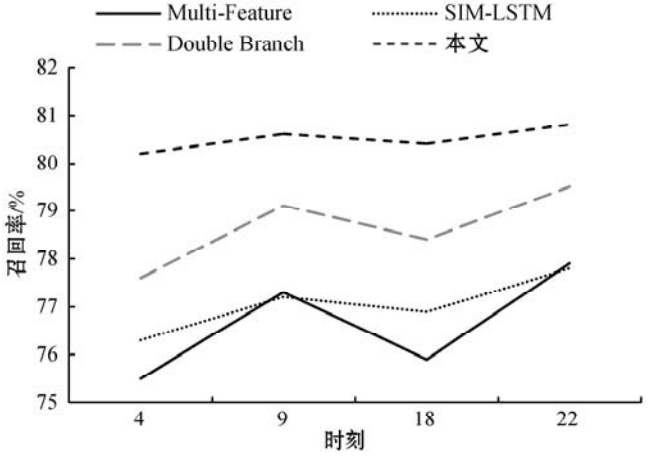

图 7　不同时刻模型召回率对比

可以看出,在 9 时刻、22 时刻,四个模型的召回率均高于各自模型在 4 时刻、18 时刻的召回率。

选取更新频率高的活跃微博用户,对其四个典型时刻发布的微博进行分析计算,对比四个时刻四个模型的召回率。选取的博主有娱乐类博主“当时我就震惊了”、体育类博主“微博体育”和企业类博主“腾讯视频”,三个用户的召回率对比如图 8 – 图 10 所示。

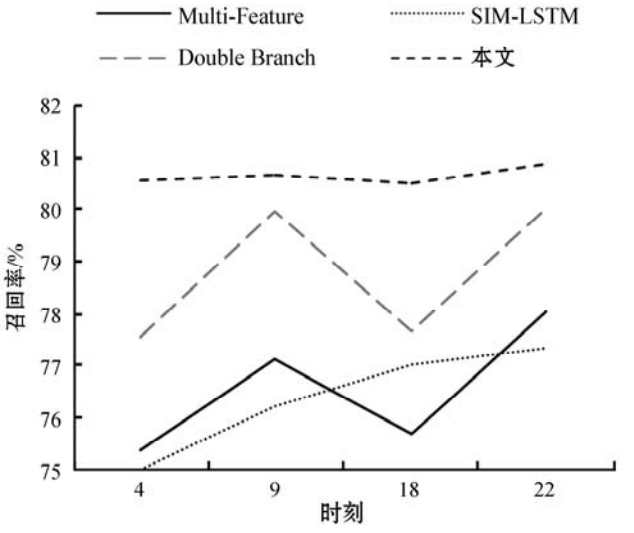

图 8　用户“当时我就震惊了”四时刻召回率对比

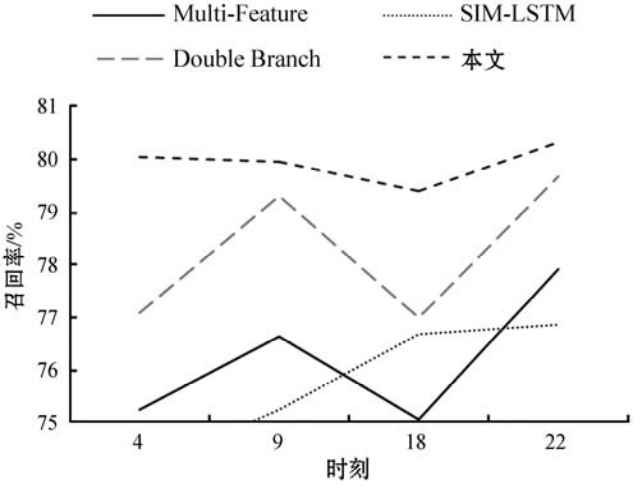

图 9　用户“微博体育”四时刻召回率对比

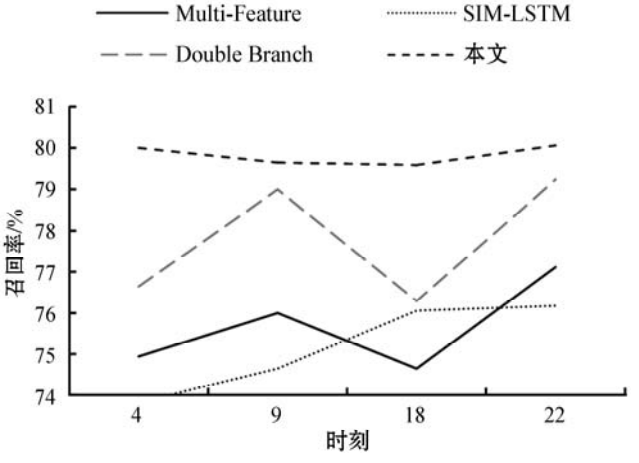

图 10　用户“腾讯视频”四时刻召回率对比

图 7 – 图 10 均说明四个微博传播预测模型在用户活跃度高的时刻以及数据集数量多的时刻效果更好。与其他三个模型相比,本文模型的召回率不会随着时刻的变化出现较大波动,说明本文模型不会随着时间、微博用户活跃情况和数据集的变化而产生较大变化,模型稳定性较好,模型能够在小数量的训练集上获得较好的效果并且模型随着训练集的增大效果随之增加。

单独对指定微博话题进行微博传播预测,可以验证实验模型在指定话题中的预测准确性,衡量模型实用性如何。如图 11 所示,本文随机选定五个微博话题#春晚#、#生日快乐#、#立冬#、#医保新政策#、#期末#,用四个模型对五个话题进行微博传播预测实验,比较四个模型的准确率。

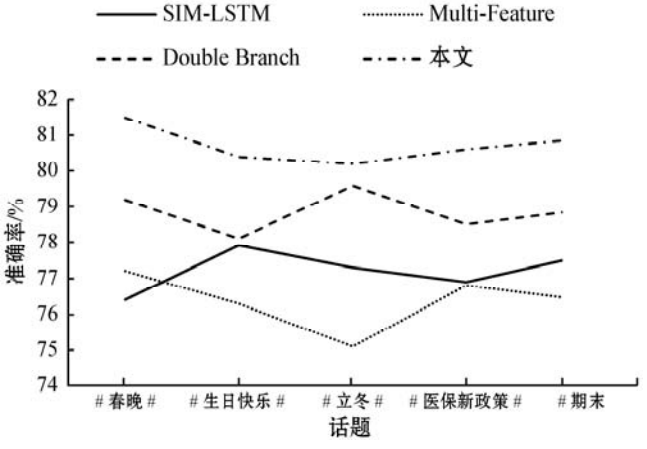

图 11 指定话题模型准确率对比

可以看出,不同模型在不同微博话题中的预测准确率有所差异,但本文所用的基于三分支神经网络的多特征微博传播预测模型在五个不同话题中的预测准确率均高于其他三个模型,说明本文所用模型在单个话题中的预测效果依旧良好,模型稳定且效率高。

通过对整体数据集进行实验,对不同时刻、不同话题的部分数据集进行实验,发现本文模型在准确率、召回率和 F1 值上均高于其他三个模型,模型的稳定性高,性能良好。

3 结 语

本文提出的基于三分支神经网络的多特征微博传播预测模型经过实验验证,其准确率、召回率、F1 值均高于其他模型,该模型具有较高稳定性,在微博传播预测准确度上有了显著提高。但本文在研究过程中没有充分考虑“微博粉丝控评”即某些名人或者微博用户的忠实粉丝会对这些名人和微博用户进行大量重复的转发、评论,从而对微博转发预测的真实效果产生一定影响。今后的研究工作会围绕上述问题展开。

参 考 文 献

[1] 刘超,姚耿,杨宏雨. 基于微博关注网络的转发预测算法研究[J]. 数字技术与应用,2020,38(7):121 – 124.
[2] 陈振春,刘学军,李斌. 基于内容和信任度的舆情扩散研究[J]. 计算机应用与软件,2017,34(10):59 – 65.
[3] 曾辉,淦修修,彭俊,等. 基于双分支结构的融合多特征微博传播行为预测算法[J]. 科学技术与工程,2020,20(26):10822 – 10828.
[4] 孙红,左腾. 基于 PageRank 的微博用户影响力算法研究[J]. 计算机应用研究,2018,35(4):1028 – 1032.
[5] 毛国君,谢松燕,胡殿军. PageRank 模型的改进及微博用户影响力挖掘算法[J]. 计算机应用与软件,2017,34(5):28 – 32,37.
[6] 李勇. 一种改进的微博用户影响力分析算法[J]. 计算机技术与发展,2020,30(8):27 – 33.
[7] 刘玮,贺敏,王丽宏,等. 基于用户行为特征的微博转发预测研究[J]. 计算机学报,2016,39(10):1992 – 2006.
[8] Deng X, Wu W, Wang F. Deep metric learning for text data based on triplet network[J]. IOP Conference Series: Materials Science and Engineering,2020,806:218927660.
[9] He G, Li F, Wang Q, et al. A hierarchical sampling based triplet network for fine-grained image classification[J]. Pattern Recognition,2021,115:107889.
[10] Hoffer E, Ailon N. Deep metric learning using triplet network[C]//International Workshop on Similarity-based Pattern Recognition,2015.
[11] Bhople A R, Prakash S. Learning similarity and dissimilarity in 3D faces with triplet network[J]. Multimedia Tools and Applications,2021,80:35973 – 35991.
[12] Li Y, Chen Y, Wang N, et al. Scale-aware trident networks for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV),2019.
[13] 穆圣坤,张路桥,滕彩峰. 基于循环神经网络的微博转发行为预测[J]. 计算机系统应用,2019,28(8):155 – 161.
[14] 王志峰,冯锡炜,贾强,等. 多特征神经网络微博转发预测[J]. 辽宁石油化工大学学报,2017,37(6):47 – 50.
[15] 王绍卿,李翠平,王征,等. 基于多重信任关系的微博转发行为预测[J]. 清华大学学报(自然科学版),2019,59(4):270 – 275.

(上接第 372 页)

[23] GitHub. BlockBench[EB/OL]. [2021 – 05 – 19]. https://github.com/ooibc88/blockbench.
[24] GitHub. Docker-library[EB/OL]. [2021 – 04 – 27]. https://github.com/docker-library.
[25] Castro M, Liskov B. Practical byzantine fault tolerance and proactive recovery[J]. ACM Transaction on Computer Systems,2002,20(4):398 – 461.