Image analysis considering textual correlations enables accurate user switching tendency prediction*

WANG Jianbin[1,2], SHI Shuyuan[2], WANG Xuna[3], and YU Jiahui[1,4]****

1. Zhejiang University, Hangzhou 310027, China

2. China Telecom Zhejiang Branch, Hangzhou 310014, China

3. School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110159, China

4. School of Computing, University of Portsmouth, Portsmouth PO13HE, UK

(Received 9 March 2023; Revised 28 March 2023)

Predicting likely-to-churn users employing surveys is a challenging task. Individuals with different personalities may make different choices in the same situation, so we introduced social media avatars that reflect the user's psychological state when analyzing their churn tendency. In this paper, we propose a multimodal framework that jointly learns image and text features to establish correlations among users with low net promoter score (NPS) and those likely to churn. We conducted experiments on actual data, and the results show that our proposed method can identify NPS-degraded users in advance, promoting the commercial development of the operator.

Document code: A **Article ID:** 1673-1905(2023)08-0498-8

DOI https://doi.org/10.1007/s11801-023-3043-8

The telecommunications industry is significantly shifting from the long term evolution (LTE) 4G era to the new radio (NR) 5G stage. With the evolution of network systems, users' network usage habits have also transitioned to short vertical video and streaming content viewing. However, user traffic has not grown exponentially, putting pressure on operators to reduce internal costs while investing significantly in 5G. Additionally, product homogenization exacerbates the phenomenon of user churn[1].

The personality characteristics of users must be addressed when predicting user churn trends, as shown in Fig.1. Some studies suggest that introverted individuals prefer low-key, simple avatars, while extroverted individuals may prefer their photos or photos of celebrities. Emotional people may prefer avatars full of emotional expression, and agreeable people may prefer avatars that help establish friendly relationships. User avatars on social platforms can be used to explain individual differences and the reasons for different behaviors exhibited by individuals in the same situation[2]. These avatars contain rich subjective information driven by factors such as personality, which is stable over time and space. Operators can utilize this information to improve customer management and reduce user churn.

Big data analysis and machine learning technologies have important applications in telecommunications, helping operators improve network performance and user experience, improve marketing strategies and security, and provide more intelligent IoT services[3]. Thanks to the rapid development of image processing in recent years, images are used to analyze and predict user characteristics.

(a) Introverted individuals (b) Extroverted individuals

(c) Emotional individuals (d) Agreeable individuals

Fig.1 Social avatars among individuals with different personalities

* This work has been supported by the NSFC-Zhejiang Joint Fund for the Industrialization and Informatization (No.U1809211).

** E-mails: shuyuan_shi90@163.com; jiahui.yu@port.ac.uk

Studies have shown that the accumulation of user transfer trends is not only related to the data performance of operator operational indicators, but also directly influenced by various business data such as package billing, customer service level, and marketing promotion[4]. This textualized information is also the key for trend forecasting. Extracting text information as features and introducing it into an image analysis framework is absent from most existing methods. By learning and analyzing large amounts of user data, neural networks, and machine learning technologies can identify different personality types, such as extroverted, introverted, and emotional. The comprehensive application of these technologies can improve the accuracy and reliability of virtual user avatar personality analysis and provide more precise user portraits and demand analysis for personalized services.

Based on the above analysis, we have developed a flexible and lightweight text feature extraction network, which provides a cost-effective way to enhance image features with text features. By leveraging the complementary nature of text and image data, we can obtain more affluent and informative representations that can improve the performance of downstream tasks such as image analysis. To further improve the effectiveness of our approach, we have also designed a new algorithm for mapping relationships between different features. This algorithm considers the complex and subtle relationships between avatar style, user habits, and transfer tendencies. It builds a more comprehensive model of these relationships than traditional methods. Hence, the proposed framework represents a significant step in developing more efficient and effective deep learning models. By leveraging text data cost-effectively and building more comprehensive models of relationships between features, the proposed framework can improve the accuracy and robustness of deep learning systems across a wide range of applications. Main contributions of the paper are summarized as follows.

We propose a novel multi-modal framework to predict user switching trends, simultaneously considering multiple influencing factors, including images and text, achieving effective adaptive optimization.

We propose to use a dual-channel deep convolutional neural network to extract the style features of images. Combining with the Gram matrix, the convolutional network extracts the color, edge, local shape, and texture information of the image, and finally obtains the style parameters of the image.

We propose novel neural networks to further mine the correlation between user habits and transfer trends without highly demanding data-driven, considering point-to-point and comprehensive correlations.

We conduct a series of experiments to evaluate the proposed method. Especially with the latest collected data, the results show that the proposed method performs better than state-of-the-art methods under the influence of multiple factors.

Currently, the industry mainly focuses on net promoter score (NPS), and there is relatively little exploration of network transfer trends. The identification of NPS detractors is a typical classification problem in machine learning. This problem can be reduced to how to automatically identify NPS detractors through modeling historical user data. Due to this research direction, the telecommunications industry has also made some explorations in the past. HUANG et al[5] modeled the relationship between user behavior and NPS using Bayesian algorithm. LU et al[6] combined frontline practical work to construct an NPS value model related to management field. These NPS prediction studies mainly focus on operational or business fields, lacking a systematic perspective, and the predictability is relatively low. In addition, using statistical methods to predict network transfer trends is also a common research method. SAROHA et al[7] proposed structural equation modeling technique and used AMOS statistical package for structural analysis, where the comparison of each driver factor of customer loyalty is calculated using standardized regression weights obtained from path model. However, the applicability of such research is relatively low in terms of prediction. In addition, the relationship between users' network transfer tendency and their NPS scores needs to be clarified.

In the relevant research on predicting network transfer trends, there has been no research combining personality information from user avatars. Using images to study users' personalities is a research direction that has emerged in recent years. Currently, in the field of computer science, the dominant techniques are to train a deep convolutional neural network model to recognize the style of images. Compared with the traditional classification algorithms, the advantage of the method based on deep convolutional neural networks is that it eliminates the limitations of image features. NICHOLAUS et al[8] automatically predicted personality through Twitter users' personal photos and tweets, where they took users' facial features, expressions, and picture colors from photos. They found through experiments that image features are better at predicting personality than text features or a combination of text and image features, and a single profile picture can reliably predict personality.

In this paper, user social platform avatars are used, which may not contain or make it difficult to identify the user's facial features. Therefore, we only consider obtaining users' personality information through image style. In fact, non-personal images posted by users on online social networks can also be used to distinguish users' personalities. ZHU et al[9] proposed a personality modeling method based on image aesthetic attribute perception graph representation learning, which can use convolutional networks to predict users' image aesthetic attributes, and introduce perceptual graph representation learning to refine images with similar aesthetic attributes.

Finally, personality traits are predicted through multi-layer perceptrons (MLP). SOMAYE et al[10] proposed a method based on deep features and compact convolutional converters to identify image styles. The experimental results show that the method using style has a positive effect compared with personalized image recommendation without style.

In the section, we first detail the dual-stream framework. Next, we technically present the process of image and text feature extraction. Finally, we introduce association analysis and prediction methods. In addition, we also present related modules in detail.

Let $y \in \{0,1\}$ be the user switching tendency classification label, y=0 represents a derogatory user, and y=1 identifies a non-derogatory user. The problem of this study can be expressed as establishing a user derogatory scoring classification model $f()$ according to the image dataset I_u composed of user avatars and the text dataset T_u composed of operation / business domain data.

$$\{I_u, T_u\} \xrightarrow{f(\bullet)} y, y \in \{0,1\}. \tag{1}$$

As shown in Fig.2, this study employs a multimodal approach based on both image and text data. Firstly, anonymized user image and text data were collected and integrated with derogatory categories into a trainable dataset, based on different unique identifiers. Secondly, a dual-channel convolutional neural network was used to extract information from user avatars and obtain image style parameters as the user's personality data. The network was pretrained using Flickr dataset. Concomitantly, correlation mining was conducted on text information. Subsequently, the image style parameters and text data were preprocessed and input into a classification network for training. Finally, the practical logical experience of the model was extracted to guide subsequent iterative analysis. As the analysis is a supervised machine learning problem with binary output, Python programming language was used to implement the neural network model training. The prediction performance of the model was evaluated based on the accuracy P (precision), recall rate R (recall), and their comprehensive indicator F1. A user migration trend analysis mechanism was built after obtaining a satisfactory neural network model performance.

Based on the existing user data from 2021, the information of derogatory and non-derogatory users was integrated into a database. The database consists of user avatars and O/B domain data that has been desensitized, which is a multi-modal database that integrates images and text.

The texture features of images have a significant impact on people's perception and differentiation of image styles[11]. Based on this, we propose a dual-channel convolutional neural network (CNN) with two pathways. The objective channel of this neural network extracts the features used for target recognition tasks in images, such as color, edges, and local shapes. The texture channel incorporates the gram layer to extract the texture features of images.

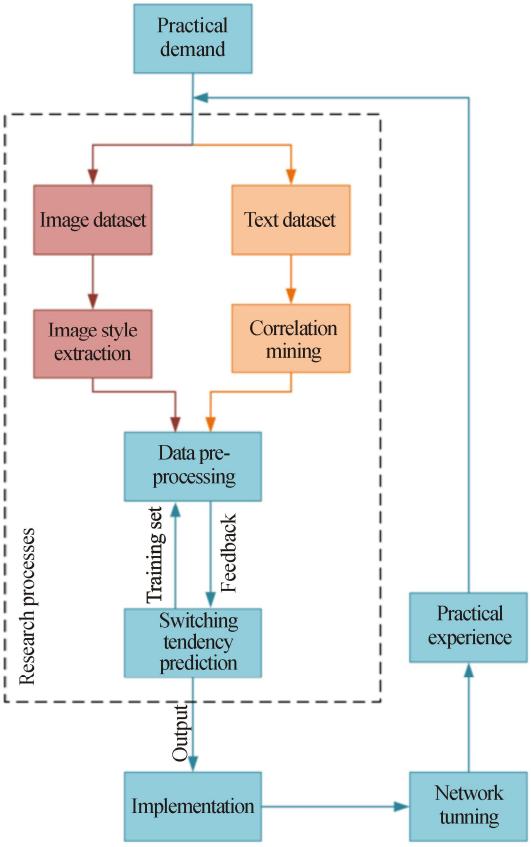


Fig.2 Framework of user switching tendency prediction considering images and text

The network mainly utilizes the gram matrix in the gram layer to enhance the predictive performance of the style model. The advantage of the gram matrix is that it can help preserve the texture features of images, avoiding the loss of texture information caused by smoothing and filtering operations in some traditional image processing methods. Calculating the gram matrix requires using the feature maps in the CNN, reshaping them into matrix form, and calculating the inner product of the matrix to obtain the correlation between each channel. As different feature maps reflect different image features, ideally, the gram matrix can represent the overall feature distribution and the relationships between various features of the image, which are defined as the texture information of the image.

Specifically, the backbone network of the dual-channel deep CNN is Resnet-101[12], which is the target channel. We added a gram layer after each of the four convolution groups, the output is concatenated and then connected to the max-pooling layer, fully connected layer, and softmax layer to form the texture channel. The convolution layers in the network are shared by the target and texture channels, and the model parameters are jointly determined by the training of both channels. The

forward calculation of the gram layer involves multiple stages: (1) down-sampling the feature map, (2) computing the gram matrix, (3) deleting the upper triangle, and (4) connecting to the next layer. Each stage can be implemented as a sub-layer. Different gram layer outputs provide different levels of texture information. The structure of the network is shown in Fig.3.

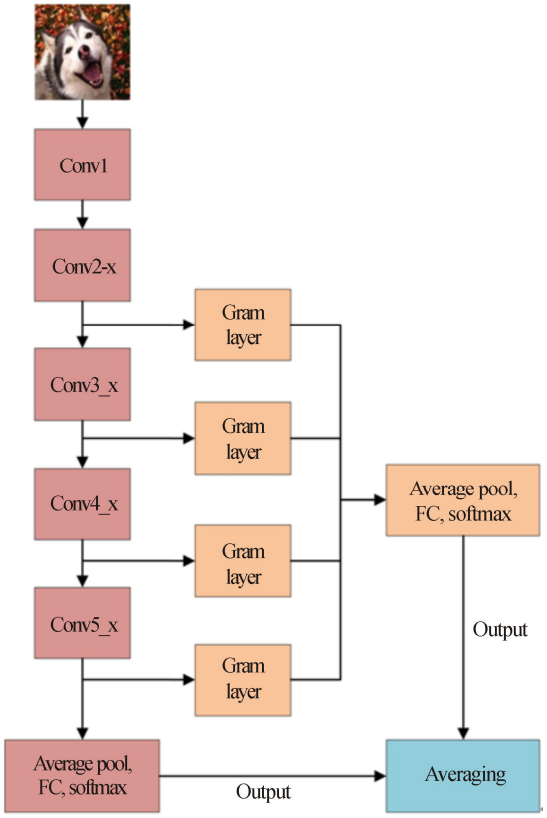
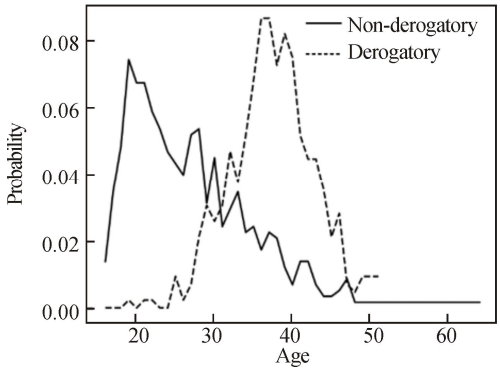


Fig.3 Structure of the dual-channel deep CNN

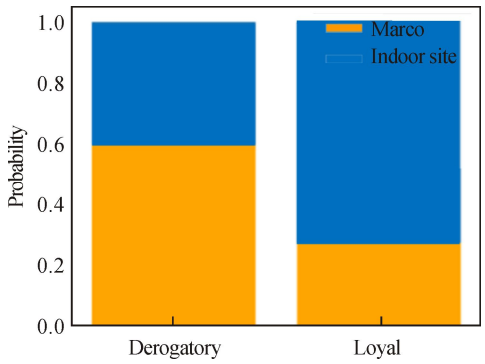
We trained the model on the Flickr dataset[13] to extract abstract style features and predict the average output of the final layer softmax function based on two channels. The Flickr dataset contains 80 000 images with 20 style labels. The labels were grouped into six categories: optical techniques, atmosphere, mood, composition styles, color and genre. We split the dataset into training, validation, and testing sets in a 6: 2: 2 ratio and randomly cropped the input images into four 224×224 images. The model was trained based on pre-trained weights of Resnet-101, and the mean average precision (mAP) of the final testing set was 62.1%.

In terms of the age attribute, as shown in Fig.4(a), there are apparent differences in the age distribution between derogatory and non-derogatory users. Derogatory users are mainly in the range of 35 to 45 years old, while non-derogatory users are primarily in the range of 20 to 25 years old. This shows that the age field has excellent information value in predicting users' NPS scoring.

A user's resident cell type represents whether the resident cell belongs to an outdoor macro site or an indoor microsite. As shown in Fig.4(b), 72.68% of loyal users' resident cells are indoor cells, while for derogatory users, the proportion of users whose resident cells are sub-cells has decreased to 42.42%. It shows that the coverage enhancement brought by the indoor cell significantly impacts the users' network experience.



(a) Age distribution



(b) User resident cell type

Fig.4 Correlation between NPS and O/B domain data

In the context of intensive application of mobile applications, the difference in the use of call duration and data traffic between different users has been proved in many studies[14]. Therefore, user traffic and call usage also have excellent analytical value for user scoring and prediction.

For image data, we utilized the model presented in above to obtain the confidence score for each style. For textual data, we combined the data cleaning approach described above. Then, in order to avoid numerical gradient problems during neural network training, all data was normalized using the Z-score method[15], shown as

$$X' = (x - \mu)/\sigma . \tag{2}$$

After normalization, the total number of samples was 3 000, and they were divided into training and test sets in a ratio of 70% to 30%. The data fields include 1 derogatory index, 20 avatar style data, 11 operation domain data, and 6 business domain data. The derogatory index is the label to be predicted, with 0 representing derogatory users and 1 representing non-derogatory users. The rest of the data are reference data for predicting label, and

detailed information can be found in Tabs.1—3.

Tab.1 Research data dictionary in avatar style field

Attribute	Name	Meaning
Optical techniques	Macro	Confidence score
	Bokeh	Confidence score
	Depth-of-field	Confidence score
	Long exposure	Confidence score
	HDR	Confidence score
Atmosphere	Hazy	Confidence score
	Sunny	Confidence score
Mood	Serene	Confidence score
	Melancholy	Confidence score
	Ethereal	Confidence score
Composition styles	Minimal	Confidence score
	Geometric	Confidence score
	Detailed	Confidence score
	Texture	Confidence score
Color	Pastel	Confidence score
	Bright	Confidence score
Genre	Noir	Confidence score
	Vintage	Confidence score
	Romantic	Confidence score
	Horror	Confidence score

Tab.2 Research data dictionary in business field

Attribute	Name	Meaning
Business data	Age	Age
	Monthly tariff	Tariff this month
	Monthly traffic	Total traffic this month
	Total calling time	Total calling time this month
Customer management data	Customer visits number	The number of times the user handled business in the store in this month

Tab.3 Research data dictionary in operation field

Attribute	Name	Meaning
Wireless network KPI	User weak coverage MRO samples number (monthly)	Number of MRO whose RSRP is less than -105 dBm, represents the network weak-coverage percentage of the user
	User MRO sample number (monthly)	The total number of MR reported by the user within a month, represents the network usage of the user
	User resident cell type	Whether the Top 1 resident cell belongs to a macro station or an indoor micro station, which represents the network structure of the user's resident location
	High-CQI ratio of user resident cell	The High-CQI ratio of the top 10 resident cell in the user month indicates the signal quality of the user resident cell
	VoLTE success setup rate	Represents the user's voice call continuity experience
	User weak coverage MRO samples number (monthly)	Number of MRO whose RSRP is less than -105 dBm, represents the network weak-coverage percentage of the user
	User MRO sample number (monthly)	The total number of MR reported by the user within a month, represents the network usage of the user
	User resident cell type	Whether the Top 1 resident cell belongs to a macro station or an indoor micro station, which represents the network structure of the user's resident location
	User VoLTE average MOS value	User VoLTE average MOS, representing the user's voice call definition experience
	User VoLTE call drop rate	Represents the user's voice call continuity experience
	User wireless drop rate	Represents the user's data service experience
	Downlink rate (kbps)	Represents the user data service experience
	User VoLTE average MOS value	User VoLTE average MOS, representing the user's voice call definition experience
Network experience KPI	Uplink rate (kbps)	Represents user data service experience
	Webpage opening delay (ms)	Represents the user's network browsing experience

Predicting a user's NPS result category is a typical classification problem. Since the sample size involved is not large, it can avoid the computational overload problem

that the neural network may encounter during gradient descent. In addition, back propagation (BP) neural network's adaptive ability and generalization ability have obvious advantages compared with other traditional algorithms. Therefore, the BP neural network algorithm is selected to construct the model. Neural network structure is shown in Fig.5, the total number of research samples is 3 000, and the overall sample is divided into a training set and validation set according to the ratio of 75% and 25%. In terms of the neural model, a total of d information neurons are designed in the hidden layer. The activation function of the neurons uses the rectified linear unit (ReLU) function to activate the neurons, and the output layer uses the Sigmoid function. The Sigmoid function formula is shown in as

$$S(x)=1/(1+e^{(-x)}), \tag{3}$$

whose derivative function $S(x')=S(x)*(1-S(x))$ is beneficial to the computer language implementation in back-propagation[16].

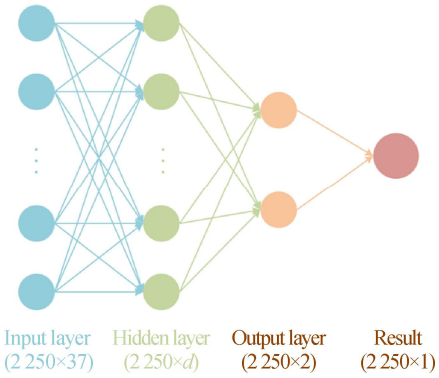


Fig.5 Neural network structure

Under 500 training times, the relationship between the number d of neurons in the hidden layer and the F1 index is shown in Fig.6. When the number of neurons is 96, the model performs optimally on the validation set, with an F1 of 82.42%. When neurons are added subsequently, the model's performance on the validation set tends to decline due to overfitting. Therefore, the number d of neurons in the hidden layer of the model in this study is set to 96.

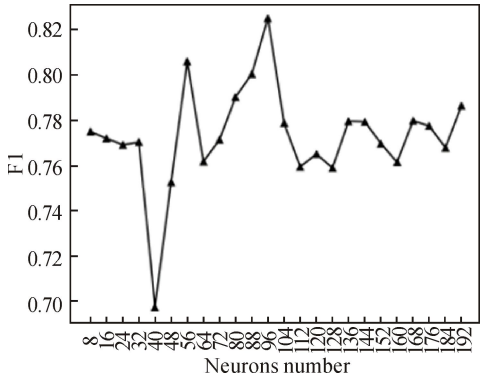


Fig.6 Model F1 performance over different nodes (500 times)

The overall performance of the model in the test set is shown in Tab.4 and Fig.7. The model precision rate P and recall rate 79.04% and 86.11%, respectively. The receiver operating characteristic (ROC) curve is in a convex trend. The area under curve (AUC) of the ROC curve is 0.87.

Tab.4 Confusion matrix of prediction results of the proposed method

Real \ Prediction	Derogatory	Non-derogatory	Performance
Derogatory	279	45	86.11%
Non-derogatory	74	352	82.63%
Performance	79.04%	88.66%	

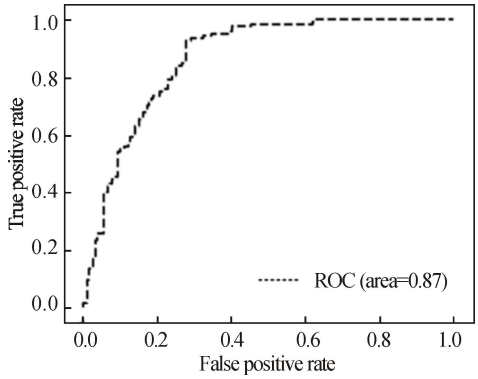


Fig.7 ROC curve

In order to analyze the information value of user avatar data and O/B domain data for user switching tendency prediction, the two types of data were respectively removed and reconstructed based on the model constructed in this paper. These three models were named back propagation neutral network (BPNN), BPNN-text, and BPNN-image.

As shown in Tab.5, the F1 score of BPNN-text constructed only using text data was 90.53% of the original model, and the F1 score of BPNN-image constructed using image data was 65.73% of the original model. The predictive performance of both reconstructed models has decreased, indicating that both user avatar and O/B domain data provide some degree of information input for predicting user switching tendency prediction. It is common practice to use O/B domain data for such problems, and its information content is undoubtedly high. However, the experimental results also show that obtaining user personality information through their avatars has a significant improvement effect on predicting migration trends.

Tab.5 Model performance comparison (0.5 threshold)

	P	R	F1
BPNN	79.04%	86.11%	82.42%
BPNN-text	73.52%	75.76%	74.62%
BPNN-image	56.67%	51.90%	54.18%

As mentioned above, relatively few existing modeling studies exist on NPS users. In addition to building a BPNN model, a support vector machine (SVM)[17] and a random forests (RF) model[18] are also constructed using the same data. As shown in Fig.8, the AUC metric of the BPNN model is the highest among the three. It means that the BPNN model is more sensitive to the experimental data under the same conditions, and the prediction results are more accurate.

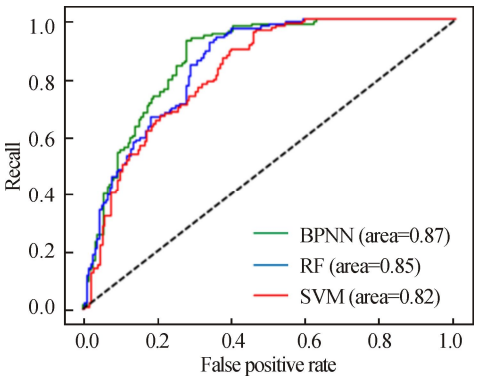


Fig.8 ROC curve over different models

In the telecommunications industry, the main way to obtain users' migration tendencies is through O/B domain data from wireless network operators. However, considering that people with different personalities may make different choices in the same situation, we propose extracting image style parameters from users' social platform avatars as personalized data for this task. This data can serve as a reference for such tasks. Therefore, unlike previous research that only used text information for user tendency prediction, we propose a multimodal prediction framework that uses both image and text data.

As users' avatar styles are related to their personalities, we use a dual-channel convolutional neural network to extract image features and obtain confidence scores for 29 image styles, which are then used as the user's personality parameters. Experimental results show that incorporating user avatar information can improve the model's F1 score by 8%, demonstrating the feasibility of our proposed method.

We overcome the differences in data structures between multimodal tasks by first extracting features from images, normalizing the image features and text information using Z-score, and then inputting the normalized data into the classification model to obtain the prediction results. In terms of images, this is the first study to use users' avatar information for user migration tendency prediction tasks. In terms of text, unlike previous methods that only consider O/B domain data or business domain data, our proposed method considers both types of data.

Wireless network operators have a wealth of data resources, but most of the data collection and storage is scattered across different departments. At this stage, network services are highly homogenous. The use of advanced data mining and machine learning techniques to deeply mine and utilize existing data will be a new direction for wireless operators to pursue business leadership.

In this paper, we propose a multimodal prediction framework that combines image and text data to predict user switching tendencies with consideration for their personality. Thanks to this, we accurately predict the potential intent behind user transfers without requiring a large amount of data-driven training. This is the first time computer vision technology has been used in such studies, setting a precedent for considering user personality as a factor. First, we propose a dual-channel convolutional neural network that combines the Gram matrix for feature extraction from user avatars. Then, we combine O/B domain data with image features by normalizing them using Z-score. Finally, we input the combined user-granular data into a classification network for predicting user migration tendencies. Our proposed framework achieves state-of-the-art results across multiple evaluation metrics, demonstrating its important theoretical research and practical value in effectively learning the underlying correlations in complex interactions of today's factors.

Ethics declarations

Conflicts of interest

The authors declare no conflict of interest.

References

[1] COOPER A B, BLAKE A B, PAULETTI R E, et al. Personality assessment through the situational and behavioral features of instagram photos[J]. European journal of psychological assessment, 2020, 36：959-972.

[2] RIZOMYLIOTIS I, POULIS A, APOSTOLOS G, et al. Applying FCM to predict the behaviour of loyal customers in the mobile telecommunications industry[J]. Journal of strategic marketing, 2020, 28(1)：1-15.

[3] OUYANG Y, WANG L, YANG A, et al. The next decade of telecommunications artificial intelligence[EB/OL]. (2021-01-19) [2022-12-13]. https：//arxiv.org/ftp/arxiv/papers/2101/2101.09163.pdf.

[4] HAPSARI R, HUSSEIN A S, HANDRITO R P. Being fair to customers：a strategy in enhancing customer engagement and loyalty in the Indonesia mobile telecommunication industry[J]. Services marketing quarterly, 2020, 41(1)：49-67.

[5] HUANG Y Y, LIU Y T, YU L M. Research on NPS survey data of telecom companies[J]. Post and telecommunications design technology, 2018, 509(07)：52-56.

[6] LV J. Research on network optimization guided by NPs based on user network net recommendation value[J]. Data communication, 2019, 5：8-17.

[7] SAROHA R, DIWAN S P. Development of an empirical

framework of customer loyalty in the mobile telecommunications sector[J]. Journal of strategic marketing, 2020, 28(8)：659-680.

[8] JEREMY N H, CHRISTIAN G, KAMAL M F, et al. Automatic personality prediction using deep learning based on social media profile picture and posts[C]//2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), December 16-17, 2021, Yogyakarta, Indonesia. New York：IEEE, 2022：21563700.

[9] ZHU H, ZHOU Y, LI Q, et al. Personality modeling from image aesthetic attribute-aware graph representation learning[J]. Journal of visual communication and image representation, 2022, 89：103675.

[10] SOMAYE A, MOHSEN E M. Image recommender system based on compact convolutional transformer image style recognition[J]. Journal of electronic imaging, 2022, 31：043054.

[11] GATYS L A, ECKER A S, BETHGE M. A neural algorithm of artistic style[EB/OL]. (2015-08-26) [2022-12-13]. https：//arxiv.org/abs/1508.06576.

[12] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York：IEEE, 2016：16541111.

[13] SERGEY K, MATTHEW T, HELEN H, et al. Recognizing image style[C]//Proceedings of the British Machine Vision Conference, September 1-5, 2014, Aberdeen, UK. Durham：BMVA Press, 2014：1-11.

[14] DA SILVA D V C, ROCHA A A D A, VELLOSO P B. Mobile vs. non-mobile live-streaming：a comparative analysis of users engagement and interruption using big data from a large CDN perspective[J]. Sensors, 2021, 21：5616.

[15] FEI N, GAO Y, LU Z, et al. Z-score normalization, hubness, and few-shot learning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, October 10-17, 2021, Montreal, QC, Canada. New York：IEEE, 2022.

[16] LI G, ZHANG M, LI J, et al. Efficient densely connected convolutional neural networks[J]. Pattern recognition, 2021, 109：107610.

[17] OTCHERE D A, GANAT T O A, GHOLAMI R, et al. Application of supervised machine learning paradigms in the prediction of petroleum reservoir proper-ties：comparative analysis of ANN and SVM models[J]. Journal of petroleum science and engineering, 2021, 200：108182.

[18] HAN S, KIM H, LEE Y S. Double random forest[J]. Mach learn, 2020, 109：1569-1586.