2021 IEEE 3rd International Conference on Communications, Information System and Computer Engineering (CISCE 2021)

Research on Gender Prediction for Social Media User Profiling by machine learning method

Chuanbo Liu
School of Mechanical and Electronic Engineering
Wuhan University of Technology
Wuhan, China
e-mail: Lchb72@whut.edu.cn

Fei Li
School of Mechanical and Electronic Engineering
Wuhan University of Technology
Wuhan, China
e-mail: 1203357642@qq.com

Lin Li
School of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
Corresponding author: cathylilin@whut.edu.cn

***Abstract*-This paper aims to propose a gender prediction method that integrates social media users' sentiment. This method is used to predict the gender attributes of users, so as to realize the research on the construction method of social media user portraits. Previous studies on gender prediction have done little analysis of sentiments. This paper has used the idea of transfer learning to analyze users' sentiments and integrated sentimental features into the existing machine learning, and hence, shows superior performance in terms of accuracy as compared to other methods. This paper mainly uses machine learning method to realize the construction of social media user profiling. The gender attribute is studied. Firstly, feature extraction is carried out for text data of media users. Then the idea of transfer learning is used to analyze the user's sentiment and integrate the sentiment characteristics into the existing machine learning. Finally, five prediction methods, i.e. Logistic Regression(LR), Naïve Bayes(NB), k-Nearest Neighbor(KNN), Random Forest(RF) and Support Vector Machines(SVM) is used to predict the gender of the fused sentiments. The results show that The gender prediction effect after sentiment fusion is better than that before. and the accuracy is increased by about 2.1% on average.**

Keywords: gender prediction ; sentiment; user profiling; machine learning

I. INTRODUCTION

A. Background

User profiling[1]is a tagging method based on the user's personal attributes, behavior habits and preferences. The construction of user profiling method[2]is the process of classifying users according to established categories through a series of data mining methods[3]. It can maximize the value and improve the team decision-making efficiency, so that users can improve the efficiency of information acquisition and accurately meet their own needs for product applications. The purpose of user profiling is to make a profiling of different dimensions of users and extract and label information such as users' demographic characteristics[4], social relations, behavioral patterns, habit preferences and ideological views. In recent years, with the rapid, large-scale and full-coverage development of the Internet, sina Weibo(Social media like Twitter, Facebook, sina weibo) has become increasingly popular as a social network platform. It has the characteristics of large user group[5], fast news transmission speed, wide influence and group effect. Advertising media and social public opinion supervision departments urgently need to dig out accurate and usable information through the attribute analysis of weibo users.

B.Related work

A few years ago, there were many researches on the communication mechanism[6], marketing model and future development of social media users' profilings. For example, Liu Baoqin et al[7] constructed a SVM classifier to predict the gender of Weibo users by using the features of sentiment words and language style features related to sentiment. Dai Bin et al[8]used collaborative training algorithm to predict users' gender from different perspectives such as original microblog and forwarding microblog. Li,et al[9] tried to predict the gender of interactive users by using the interactive text between sina weibo users. Wang Jingjing,et al[10]predicted the gender of Weibo users by using the characteristics of Chinese microblog user name and microblog text. AN Junhui[10]used Naive Bayes for gender recognition of Weibo user name and verb respectively, but failed to integrate the two , which was limited and with low accuracy. HUANG Faliang et al[11] proposed to use K-nearest neighbor for gender prediction on the basis of tolerance rough set method, but this method is susceptible to tolerance threshold. ZHANG Pu,et al[12] used two methods, manual feature construction classifier and convolutional neural network model, were used to predict the gender of users. The results of the two classifiers were fused with the XGBoost model to get the final prediction results. CAO Yang[13], for the issue that a single text data cannot accurately predict the gender of inactive users on Weibo, the base classifier was constructed by jointly learning the original microblog data, microblog user name data and microblog source data, and then the base classifier was fused by Bayes to finally infer the gender of users. The above research did not take into account the sentimental differences of users of different genders, and users' sentiments can have an impact on user gender prediction.

2021 International Conference on Communications, Information System and Computer Engineering (CISCE) | 978-1-6654-0352-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/CISCE52179.2021.9445922

II. Date set and Analyse

A.Date set

The data set is from the technical evaluation of the user profiling of the "micro crowd cup" held in the second half of 2016 by the Chinese Information Society of China and the National Social Media processing Professional Committee. It contains four main types of information.1)social network information composed of social media users' ID; 2)the main information of social media user composed of the source of social media, social media posts, forwarding number, comments number and Posting time； 3)the link address information of social media user name and avatar； 4)User tag information consisting of the user's gender, year of birth, and city.

Among them, 1,534 annotated data were used for this study. The data used here selectively from the original data set, including the user's Weibo information and the user's label information.

B.Data pre-processing

Chinese social media consists of a series of relatively coherent Chinese words[14], which needs word segmentation. Combining with the object and purpose of this article, we finally chose to stammer to participle word segmentation is used to sample data.

III. Gender prediction of social media profiling

A.Gender prediction with the sentiment of social media users

1）Approach of Learning sentiment features

The sentimental characteristics of social media users of different genders are different, we need to extract the sentimental characteristics of users[15]. However, the sina Weibo data used does not contain sentimental labels, so the LSTM(Long short-term Memory) neural network[16] is trained by using the data with sentimental labels in the product evaluation of online shopping platform, and then transferred to learn the prediction of sentimental characteristics of Weibo users of different genders. Figure 1 is the transfer learning method used in the sentimental learning process in this paper.

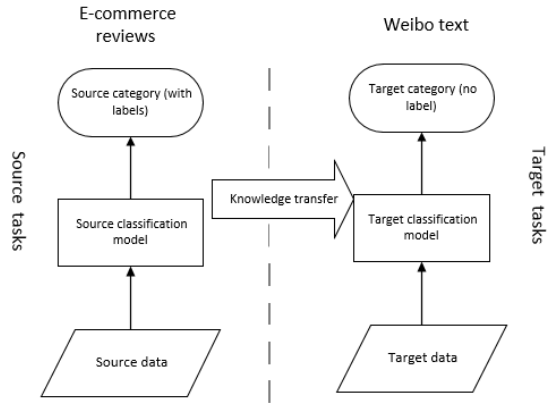

Figure 1 Learning method of Sentiment transfer

This paper analyzes the sentiments of Weibo users by dividing the texts into a series of words[17]to predict whether they are positive or negative. The input format of LSTM neural network is vector. Word2vec[18] is used to analyze the word vector of the comment data. The dimension is 100, and four feature vectors are obtained. 1)count the number of tweets send by users. The number of tweets by users divided by the number of tweets by users of the same sex or age group. 2) count the percentage of users who post messages expressing positive sentiments. The result is usually a number greater than 0 and less than or equal to 1. If the proportion of positive tweets is greater than or equal to 0.5, it is considered positive. 3)judge each user based on Weibo granularity of sentimental tendency, namely the statistics of each post to, plus or minus is discriminant this is 1, the number of discriminant this negative number is zero, although the discriminant result and the second overlap, but often in the process of identifying effect overlay, is conducive to the overall more accurate, even if the front for the proportion of positive Weibo, still need to put the Weibo plus or minus to makes it clear that, as a characteristic. 4)judge user sentiment based on user granularity; Then the positive and negative direction of the sentiment is marked as 0 or 1 respectively.

2）Approach of gender prediction

a)Data normalization

In addition to the characteristics of Weibo text related words, this paper also extracts information such as the number of users' forwarding, the number of comments, and the time interval of Weibo. The extracted information is shown in Table I.

Table I Extracted information

The number of retweets	The maximum	The minimum	The average	The sum
The number of comment	The maximum	The minimum	The average	The sum
Time interval	The maximum	The minimum	The average	The sum

The number of comments is similar to the number of retweets, but the unit of time interval is minutes, the value of the data will be much larger than them. we need to normalize the data.

This paper adopts the method of standardization based on maximum and minimum values. The following text is the feature vector of related words extracted by word2vec, and each dimension of the feature word vector is not greater than 1. This paper hopes to normalize the features obtained on [0,1], because it is inevitable that the forwarding number and the number of comments are equal to 0. Thus, a standardized method of min-max+1 is created, and the formula used is shown in formula (3-1).

$$x^* = \frac{x - \min}{\max - \min + 1} \tag{3-1}$$

b) Balanced gender data

It is necessary to balance the data to avoid the issue of prediction accuracy distortion caused by data imbalance[19]. The ratio of male and female data in the original data set is about 3:1. In this paper, the method of reducing large sample was directly adopted, and the ratio of male and female was adjusted to 1:1. Figure 2 shows the ratio of male and female before and after adjustment respectively.

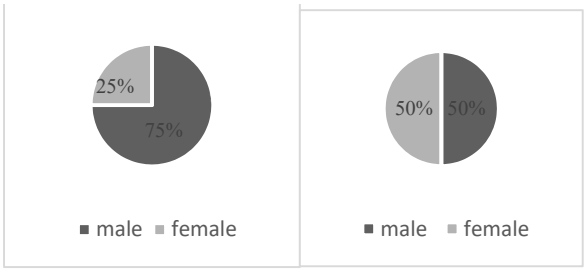

Figure 2 Sample male-female ratio before and after adjustment

The confusion matrix and accuracy of preliminary prediction before and after data adjustment are shown in Table 3. The data is a comparison experiment based on the prediction method of Logistic Regression. There are 1534 pieces of gender-balanced data between male and female. On the basis of data splitting, a preliminary prediction experiment was conducted before and after the gender data were balanced, and the results are shown in Table II.

Table II Accuracy and confusion matrix ratio before and after adjustment

	Before		After	
Date ses	Training sets	Text sets	Training sets	Text sets
Date size	2500 (1878 male)	638 (493 male)	1200 (600 male)	334 (167 male)
Confusion matrix	1796 82 513 109	465 28 120 25	434 166 215 385	116 51 59 108
AUC	0.72	0.65	0.76	0.71
Accuracy	76.20%	76.80%	68.25%	67.06%

The confusion matrix was worse before the ratio was adjusted. The trained classifier divided most of the data belonging to the female label into the male label. At this time, the accuracy rate of 76.80% on the test set was not true. After adjusting for the ratio of male and female, the obfuscation matrix showed up well, at least without dividing the data that were mostly female into male. Although the accuracy was reduced to 67.06%, it was more accurate.

c) Selection of data quantity

As for the selection of data, the data sets actually provided are not ideal, which requires the selection of data sets.

The gender prediction of Weibo users in belongs to the binary prediction of text prediction. For the selection of the number of training and test data, a series of learning curves were drawn to judge the appropriate data in the experiment. The x-coordinate of the learning curve is the size of the whole data set, and the y-coordinate is the prediction accuracy. In general, the prediction accuracy analysis of data is that when the number of data participating in training increases, the effect will become better and better, and the curve based on accuracy will gradually flatten out. The point at which it begins to flatten out is a sufficient amount of training data for reliable results. The figure below is a learning curve drawn on the basis of all the data, which is gradually increased from small to large as the step size of the data ratio is 10. The 10-fold cross-validation drawing was made based on three methods of LG, NB and SVM, respectively, as shown in Figure 3. Figure 4. Figure 5.

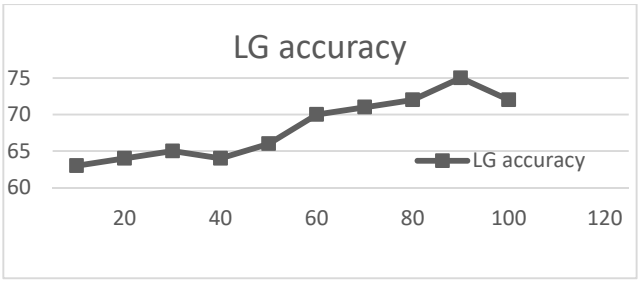

Figure 3 Learning curve based on LR

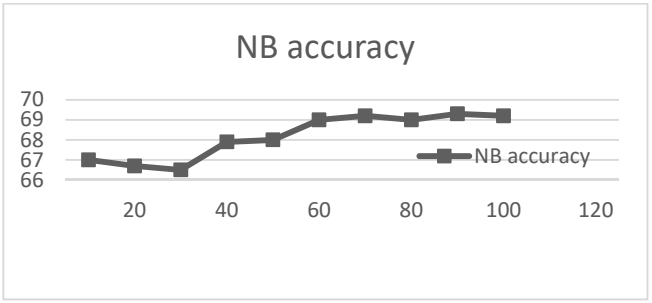

Figure 4 Learning curve based on NB

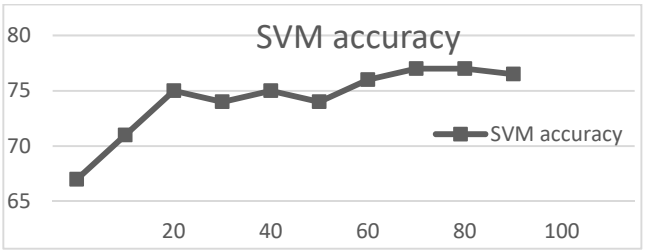

Figure 5 Learning curve based on SVM

It can be seen from the results of gradually changing the proportion of training data that when the data set accounts for 70% to 80% of the total, the prediction accuracy curve tends to be stable. There were 1534 balanced gender data. Considering that it would be convenient to divide the training data into blocks in the experiment, it decided to use 1200 data for training and 334 data for testing in the gender prediction experiment.

d) Selection of data dimension

Regarding the setting of data dimension[20], this paper conducted a group of comparison experiments based on Logistic regression, and the results are shown in Figure 6 and Table III.

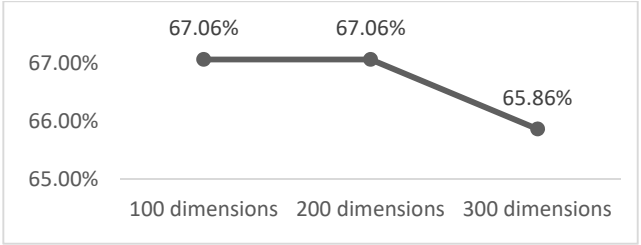

Figure 6 Comparison of word vectors in different dimensions

Table III Confusion matrix and accuracy of word vectors in different dimensions

Dimension	100	200	300
Confusion matrix	117 50	116 59	118 65
	60 107	51 108	49 102
Accuracy	67.06%	67.06%	65.86%

It can be seen from the chart that the prediction accuracy of 100, 200 and 300 dimensions is similar, among which the prediction accuracy of 100 and 200 dimensions is higher, since the higher the dimension is, the more time it takes to calculate. From the perspective of the confusion matrix, the values of the diagonal in 200 dimension are more similar, so this paper intends to use the feature word vector of 200 dimension to study.

3）Process of gender prediction

Based on the preliminary processing of gender data in the previous section, the experimental flow chart of gender prediction that integrates the sentiments of Weibo users is shown in Figure 7.

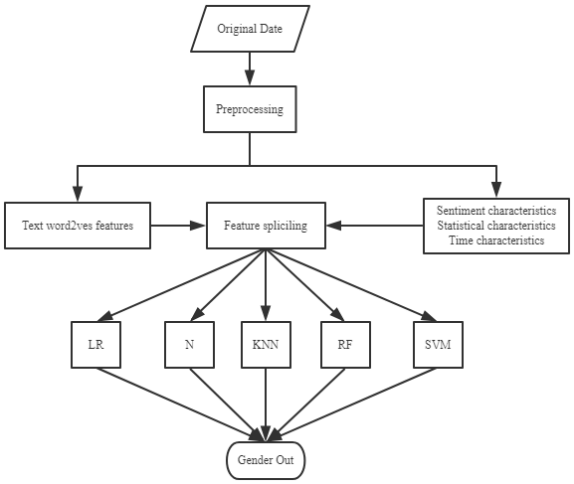

Figure 7 Flow chart of gender prediction of Weibo users

4)Results and analyse of gender prediction

a) Logistic Regression prediction

The data after vector combination was used in the experiment[21]. The experimental results are shown in Table IV.

Table IV LR Accuracy before and after fusion of sentimental features

Date sets	Before	After
Confusion matrix	116 51	131 36
	59 108	48 119
AUC	0.71	0.81
Accuracy	67.06%	74.85%

According to Table IV, it can be seen that the accuracy of users' sentiment fusion data is significantly improved after LR prediction. The gender prediction accuracy after sentiment fusion is 74.85%, which is 7.79% higher than that before.

b) Naïve Bayes prediction

The specific results are shown in Table V.

Table V Gender prediction accuracy of N

Date sets	Before	After
Confusion matrix	111 56	111 56
	48 119	48 119
AUC	0.73	0.73
Accuracy	68.86%	68.86%

Gender prediction was carried out by NB method[22], and the prediction accuracy do not change before and after the fusion of sentimental characteristics, both of which were 68.86%.

c)k-Nearest Neighbor prediction

Table 6 is the gender prediction result based on k-nearest Neighbor method.

Table VI Gender prediction accuracy of KNN

Date sets	before	after
Confusion matrix	114 53	107 60
	70 97	55 112
AUC	0.63	0.66
Accuracy	63.17%	65.57%

According to the results in table VI, the accuracy of gender prediction by k-nearest neighbor[23]is 63.17% and 65.57% before and after sentiment fusion, with a difference of 2.4%. However, the prediction accuracy of the two methods is lower than that of the first two methods.

d)Random Forest prediction

Since the samples are randomly selected from the training set, RF can effectively prevent the problem of overfitting. After the RF parameter adjustment test, it is concluded that the RF method has the following parameters related to this data set:

max Depth： The maximum depth of the tree；

num Features： Number of feature；

seed： The number of random seeds used.

Here is a comparison of the results when setting the different parameters.

1.Assuming num Features = 100, seed = 1.

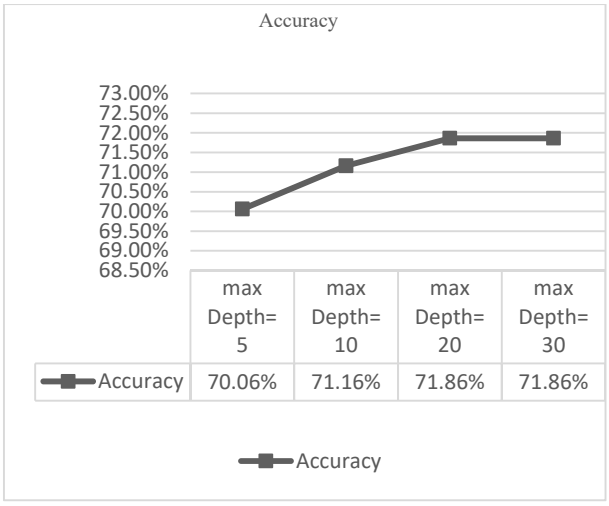

Figure 8 Accuracy of different max Depth values

Figure 8 shows that the highest accuracy can be obtained when Max Depth= 10, that is, when the maximum Depth of the number reaches 10, and the accuracy will not change when Max Depth= 30.

2.When Max Depth = 10 and seed = 1.

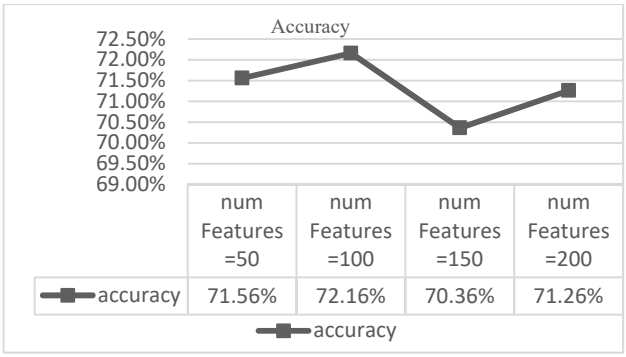

Figure 9 Accuracy of different num Features values

The experiment (Figure 9)accuracy is the highest.

3.when Max Depth= 10 and num Features = 100

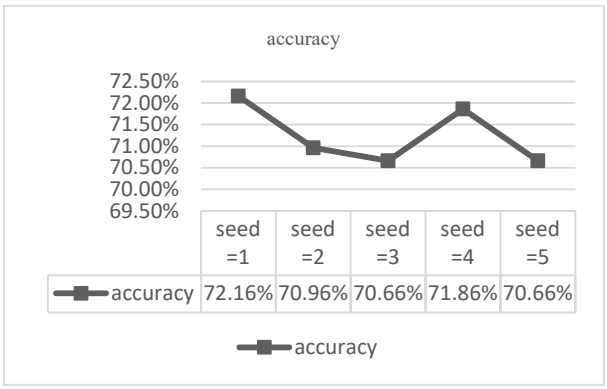

Figure 10 Accuracy of different seed values

Figure 10 shows that the accuracy is highest when the number of seeds is changed to 4.

The above results shows that the accuracy is highest when Max Depth= 10, num Features = 100, and seed = 1, and the prediction accuracy is 72.16%.

The same experiment was conducted on the data before sentiment fusion, and it was concluded that the highest accuracy was 72.15% when Max Depth= 20, num Features =100, and seed =2, slightly lower than the prediction accuracy after sentiment fusion.

e)SVM prediction

For C-SVC and nu-SVC two types, respectively select radio basis function, polynomial and linear three kernel function to do comparative experiments. In the diagram, before and after the fusion of sentiments are respectively represented by "before" and "after".

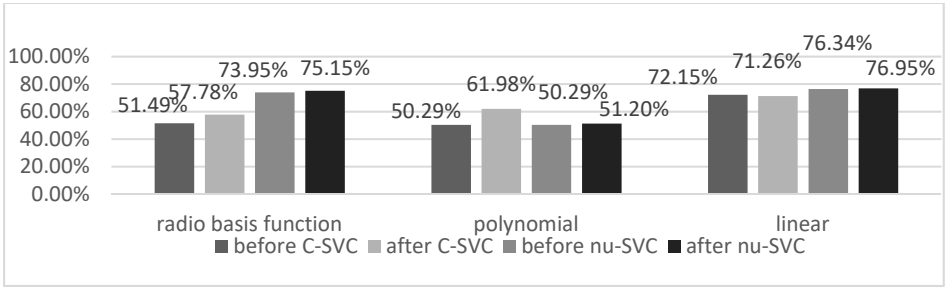

Figure 11 Accuracy of different types of kernel functions

Figure 11 shows that for the same kernel function, the accuracy of nu-SVC type is significantly higher than that of C-SVC type. In general, when the radio basis function is used as the kernel function, the accuracy difference between C-SVC type and nu-SVC type is large. The prediction of the polynomial kernel function is the general effect.The prediction accuracy of Linear kernel function is generally higher. At this point, the accuracy of nu-SVC type is the highest after sentiment fusion, which is 76.95%. It was 0.61% higher than before fusion.

5）Comparison of results

It can be seen from the comparison that LR performance is significantly inferior to RF and SVM, and k-nearest neighbor performance is not good enough. RF can prevent overfitting in principle. The prediction accuracy of SVM is relatively high. The accuracy comparison of the five prediction methods is shown in Figure 12.

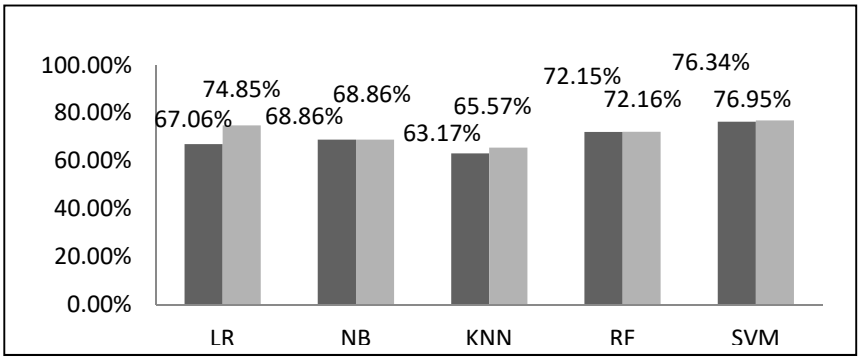

Figure 12 Comparison of accuracy of gender prediction before and after sentiment fusion

Figure 12 showed that the accuracy of SVM for gender prediction was the highest, reaching 76.95%, which was 2.1%, 8.09%, 11.38% and 4.79% higher than LR, NB, KNN and RF, respectively. The prediction accuracy before affective integration was generally higher than that before affective integration.

IV. Conclusion

This paper mainly studies the prediction of social media user profilings. Firstly, starting from the background knowledge, related work and target characteristics of social media user profilings, expounds the purpose and significance of studying social media user profilings. Then the idea of transfer learning is used to analyze the user's sentiment and integrate the sentiment characteristics into the existing machine learning. Finally, five prediction methods, i.e. LR, NB, KNN, RF and SVM, were used to predict the gender of fused sentiments.

The prediction of social media user profilings in this paper is based on the gender attributes of social media users. Although certain effects have been achieved, many other attributes of social media users will only serve as a starting point. Although the research ideas and experimental methods in this paper have certain refer ability, there are still many inconsiderate places in the whole experiment process, and many specific related factors need to be further sorted out and analyzed.

Reference

[1].Ghosh R, Dekhil M. (2008).Mashups for semantic user profiles[C]. Beijing, China: ACM..

[2].ZHAO, Y, DONG, et al.(2013). User Identification Based on Multiple Attribute Decision Making in Social Networks[J]. China Communications, 10(12):37-49.

[3].Yan-Quan Z, Ying-Fei H, Hua-Can H.(2007). Learning User Profile in the Personalization News Service[C].

[4].Khan A,Jamwal S， and Sepehri M.(2010).Applying Data Mining to Customer Chum Prediction in an Intemet Service Provider,International Journal of Computer AOplications，9(7)：8-14．

[5].Thelwall M.(2008). Social networks, gender, and friending: An analysis of MySpace member profiles[J]. Journal of the Association for Information Science and Technology, 59(8):1321–1330.

[6].Eyharabide V,Amandi A.(2012).Ontology-based user profile learning[J]. Applied Intelligence, 36(4):857-869.

[7].LIU B, NIU Yun.Gender recognition of Chinese microblog users based on emotional features[J].

[8].DAI B, LI S, GONG Z, et al.Semi-supervised gender classification with multiple type of text[J].Journal of Shanxi University (Natural Science Edition) , 2017, 40 (1) :14-20 (in Chinese) .

[9].Li S, Wang J, Zhou G, et al.Interactive gender inference with integer linear programming[C]//International Joint Conference on Artificial Intelligence, 2015:2341-2347.

[10].WANG J, LI S, HUANG L.User gender classification in Chinese microblog[J].Journal of Chinese Information Processing, 2014, 28 (6) :150-155 (in Chinese) .

[11].AN J. Research on gender judgment of microblog users based on microblog data [D]. HUAZHONG NORMAL UNIVERSITY, 2015.

[12]. HUANG F, XIONG J, HUANG tianqiang, et al. Gender Recognition of Microblog Users Based on Rough Set [J]. Computer application, 2014, 34(8):2209-2211.

[13].ZHANG P，et al. A Gender Classification Method for Chinese Microblog Users Fused with Two Classifiers [J]. Computer Engineering and Design, 2019,40(01): 268-272.

[14].CAO Y. Research and Application on Gender Classification of Microblog Users.[J] Anhui University,2019.

[15].Guo R, Qiu J, Zhang G. (2015).Web-Based Chinese Term Extraction in the Field of Study[C]// International Conference on Semantics, Knowledge and Grids. IEEE, 133-139.

[16].C. CLAVEL, Z. CALLEJAS.(2016). Sentiment analysis: from opinion mining to human-agent interaction.IEEE Transactions on Affective Computing, 7(1):74-93.

[17].K. S. TAI, R. SOCHER, C.D. MANNING.(2015). Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. Computer Science, 5(1):: 36.

[18].Tomas Mikolov, Kai Chen, Greg Corrado, Jeffery Dean.(2013).Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

[19].Yang Z T, Zheng J.(2016). Research on Chinese text classification based on Word2vec[C]//Computer and Communications (ICCC), 2016 2nd IEEE International Conference on. IEEE,1166-1170.

[20].BATISTA G E A P A , PRATI R C , MONARD M C .(2004). A study of the behavior of several methods for balancing machine learning training data[J]. Acm Sigkdd Explorations Newsletter, 6(1):20-29.

[21].Xu L, Jiang C, Ren Y, et al. (2016).Micro-blog Dimensionality Reduction—A Deep Learning Approach[J]. IEEE Transactions on Knowledge & Data Engineering, 28(7):1779-1789.

[22].T. MIKOLOV, S. W. YIH, G. ZWEIG.(2013). Linguistic Regularities in Continuous Space Word Representations.296-301.29.LEE C H.(2015).A gradient approach for value weighted classification learning in naive Bayes[J]．Knowledge—Based Systems,85(1)：71—79．

[23].LIANG Cong,XIA Shuyin,CHEN Zizhong.(2019).Improvement k—nearest neighbor classification algorithm based on reference points[J].Computer Engineering,45(2)：167—172．