ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

Investigations on the Use of Machine Learning for the Prediction of Gender in Social Media User Profiling

[1] P. Ramakrishna, [2] P.Ruchitha,

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Article Info

Received: 30-04-2025 Revised: 16-06-2025 Accepted: 28-06-2025

Abstract-

Incorporating the mood of social media users into a gender prediction algorithm is the goal of this work. The goal of this strategy is to facilitate research on the development method of social media user portraits by predicting the gender aspects of users. Little sentiment analysis has been done in previous research on gender prediction. This research demonstrates better accuracy performance than previous techniques by analyzing users' feelings using the concept of transfer learning and integrating emotional data into current machine learning. The major focus of this study is on how to build social media user profiles using machine learning methods. Research on the gender characteristic is ongoing. The text data of media consumers is first processed for feature extraction. Afterwards, the concept of transfer learning is used to assess the user's and include their emotional traits already-existing machine learning. In the conclusion, the gender of the fused feelings is predicted using five different prediction methods: Logistic Regression (LR), Naïve Bayes (NB), k-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machines (1SVM). The findings demonstrate an improvement in the gender prediction effect after sentiment fusion compared to the previous step. as well as an average improvement of 2.1% in accuracy.

Keywords:

ML, gender prediction, user profile, and sentiment

I. INTRODUCTION

Part A: History A technique for labeling users that takes into account their unique characteristics, routines, and preferences is known as user profiling [1]. Building a user profiling approach involves using a number of data mining techniques to categorize individuals into preexisting groups. It may enhance the efficiency of team decision-making and optimize value, allowing users to efficiently acquire information and properly satisfy their own product application demands. User demographics[4], social connections, behavioral patterns, habit preferences, and ideological beliefs are just a few of the factors that may be profiled by user profiling. The popularity of social media platforms like Twitter, Facebook, Sina Weibo, and Instagram has skyrocketed in the last few years due to their quick, widespread, and comprehensive coverage. Features such as a big user base, rapid news transmission, widespread impact, and group effect are present. Departments responsible for social public opinion oversight and advertising media must immediately begin analyzing Weibo users' attributes in order to extract useful and reliable data.B.Work in Progress There was a flurry of activity a few years ago in the fields of social media user profiling, marketing models, and communication mechanisms [6]. An instance of this is Liu Baoqinetal[7]. built a support vector machine classifier that utilize emotion word characteristics to forecast the gender of weibouersin terms of sentiment-related language style aspects. In order to forecast readers' gender from many angles, including the original microblog and the forwarding microblog, Dai Bin et al.[8] used a collaborative training algorithm. Using the interactivity of Sina Weibo posts, Li et al.[9] sought to identify the gender of the users' interactions. Based on features of Chinese microblog names and content, Wang Jingjing et al.[10] were able to estimate the gender of Weibo users. The gender recognition of Weibo user names and verbs were handled by AN Junhui[10] using Naive Bayes, however they neglected to combine the two, resulting in restricted and inaccurate results. Gender prediction using K-nearest neighbor based on tolerance rough set approach was suggested by HUANG Faliang et al.[11], however, this method is vulnerable to tolerance threshold. In order to determine the gender of users, ZHANG Pu et al.[12] used a convolutional neural network model and a manual feature building classifier. To acquire the final prediction results, the XGBoostmodel was used to merge the outputs of the two classifiers. A base classifier was built by learning the original microblog data, microblog user name data, and microblog source data jointly. Then, Bayes was used to infer the gender of users. This was done to address the issue of inactive users on Weibo, as highlighted by CAOYang [13]. Users' attitudes may influence user gender prediction, however the aforementioned studies failed to include these variances.

II. Date set and Analyse

A. DaytimeThis data collection originates from the "micro crowd cup" that took place in the second half of 2016 and was evaluated technically by the National Social Media Processing Professional Committee and the Chinese Information Society. It primarily includes four categories of data. 1) Data from social networks, which includes user IDs; 2) Data about social media users, which includes their source, postings, forwarding number, comment number, and posting time The user's gender, year of birth, and city are part of their usertag, along with the link address information of their social networking username and image. The research made use of 1,534 of these sets of annotated data. Specifically, this uses data culled from the original dataset, which includes the user's label information and their Weibo profile. Section B: Data Pre-processing, Word segmentation is necessary because Chinese social media is composed of a string of words that are generally cohesive [14]. With the article's goals and purposes in mind, we settled on stammer-to-participle word segmentation as our data sampling method.

III. Gender prediction of social media profiling

One example is using social media user mood to predict gender. The thirteenth method of Mastering emotion characteristics

As a result, we need to extract the emotional features of users from social media, which vary by gender [15]. To compensate for the lack of sentimental labels in the used Sina Weibo data, a Long Short-Term Memory (LSTM) neural network[16] is trained on data from an online shopping platform's product evaluations and then applied to the task of predicting the emotional traits of male and female Weibousers. As seen in Figure 1, this paper's sentimental learning procedure makes use of a transfer learning approach.

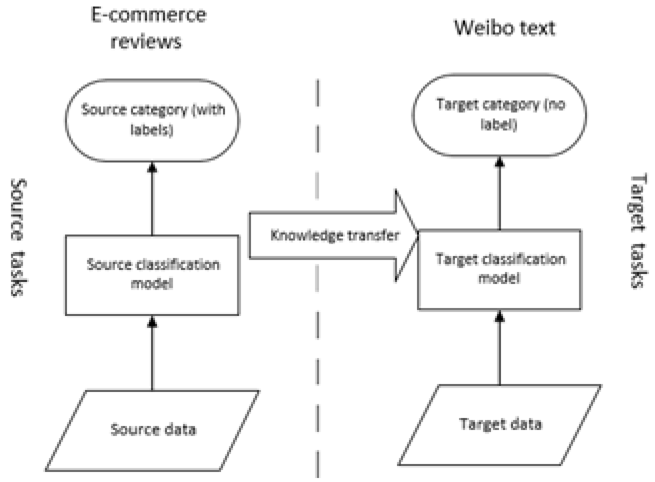
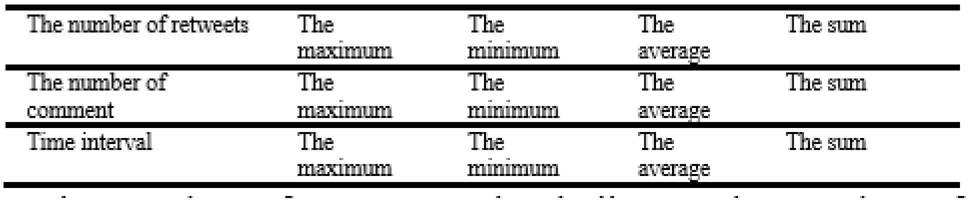

E-commerce reviews
Weibo text
Source category (with labels)
Target category (no label)
Source tasks
Target tasks
Knowledge transfer
Source classification model
Target classification model
Source data
Target data

Figure1LearningmethodofSentimenttransfer

In order to determine whether the attitudes expressed by Weibo users are favorable or negative, this study breaks down the texts into a sequence of phrases [17]. LSTM neural networks use vectors as input. To examine the comment data's word vector, Word2vec[18] is used. It has a 100-dimensional space and four feature vectors. 1)tallyhowmany tweets were sent by various persons. Twitter engagement is the total number of tweets split by the total number of tweets from the same age group. secondly, tally up the proportion of people that write upbeat comments. The output is often an integer between zero and one. A positive ratio is defined as a percentage of positive tweets that is equal to or higher than half. 3) Evaluate each user using the granularity of sentiment analysis on Weibo, specifically looking at the statistics of each post to see whether they are discriminant. Although the discriminant result and the second overlap are both 1, the number of discriminants is zero. Overlapping effects are helpful in identifying the overall accuracy of the results. Regardless of the proportion of positive Weibo, it is still necessary to put the Weibo plus or minus to make it clear that it is a characteristic. 4) evaluate user sentiment according to user granularity; then, assign a value of 0 or 1 to indicate a good or negative sentiment direction, accordingly.23Methods for predicting Gender (a) Normalizing data This article collects information such as the amount of forwarding, the number of comments, and the time interval of Weibo, in addition to the features of text-related phrases on Weibo. The data that was extracted is shown in Table I.

Table IExtractedinformation



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

While the quantity of retweets and comments is comparable, the data value will be much higher when measured in minutes. The data must be normalized. In this work, we use a maximum and minimum value based standardization approach. The feature word vector of related words extracted via word2vec is shown below; each dimension of the vector is less than 1. Since the forwarding number and the number of comments will always equal zero, this study aims to normalize the features acquired on the interval [0,1]. Here in formula (3-1) we can see the result of standardizing the min-max+1 technique.

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

a) Gender-neutral statistics Data imbalance may alter prediction accuracy, hence balancing the data is essential to prevent this problem [19]. Around three to one, there are more males than females in the original dataset. The paper's sample size was reduced using the direct technique, and the gender ratio was set at 1:1. The male-to-female ratio was shown in Figure 2 before correction and again after adjustment.

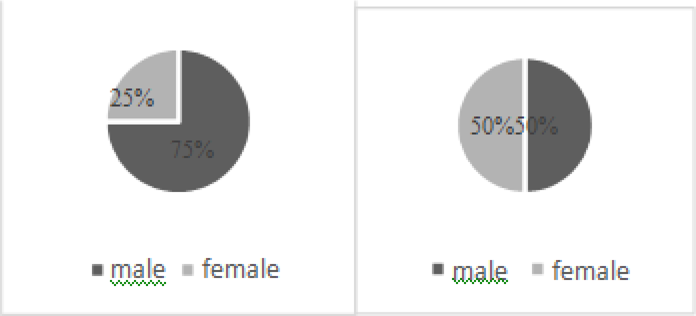


Figure2Samplemale-femaleratiobeforeandafteradjustment

The accuracy of the preliminary prediction before and after data correction, as well as the confusion matrix, are shown in Table 3. This data set is a result of a comparative experiment that used Logistic Regression as its prediction technique. Fifteen hundred thirty-four bits of data are balanced between the sexes. Table II shows the results of a preliminary prediction experiment that was run on the basis of data splitting before and after the gender data were balanced.

TableIIAccuracyandconfusionmatrixratiobeforeandafteradjustment

	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%

Prior to the adjustment of the ratio, the confusion matrix was much worse. After training, the classifier transferred most of the data from the female to the male category. Currently, the test set's accuracy rating of 76.80% is false. Following the adjustment for the gender ratio, the obfuscationmatrix appeared satisfactorily, excluding the need to convert the data that consisted mostly of females into males. While the accuracy fell down to 67.06%, it was still better than before.) Choosing the amount of dataAs for data selection, it's necessary since the supplied data sets aren't perfect. One kind of text prediction is the binary prediction, which includes the gender prediction of Weibo users. We used a set of learning curves to determine how many training and test data points would be most useful for our experiment. The learning curve's x-coordinate represents the total data set size, while the y-coordinate represents the accuracy of the predictions. In most cases, the examination of data prediction accuracy shows that as the amount of data used for training rises, the impact becomes stronger and better, and the curve based on accuracy eventually flattens out. There is enough training data for accurate findings as it starts to flatten out. The data-driven learning curve seen in the image below scales up from tiny to huge steps with a data-to-figure ratio of 10. Figure 3 shows the 10-fold cross-validation drawing that was generated using three methods: LG, NB, and SVM. Illustrations 4. and 5.

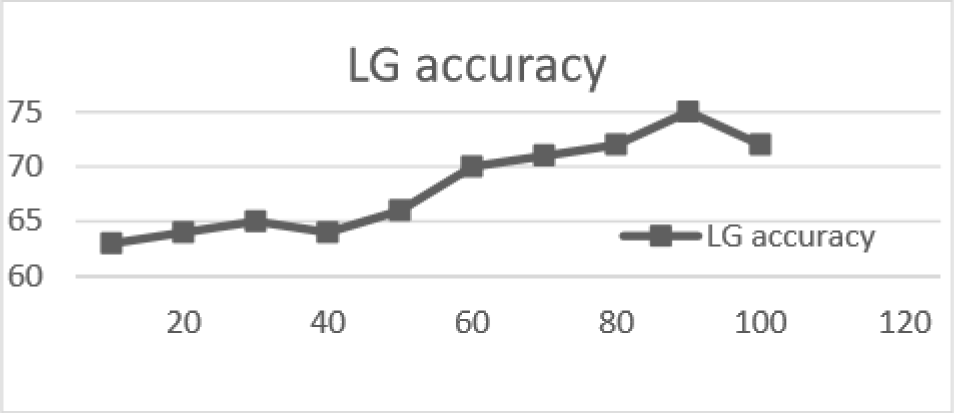


Figure3LearningcurvebasedonLR

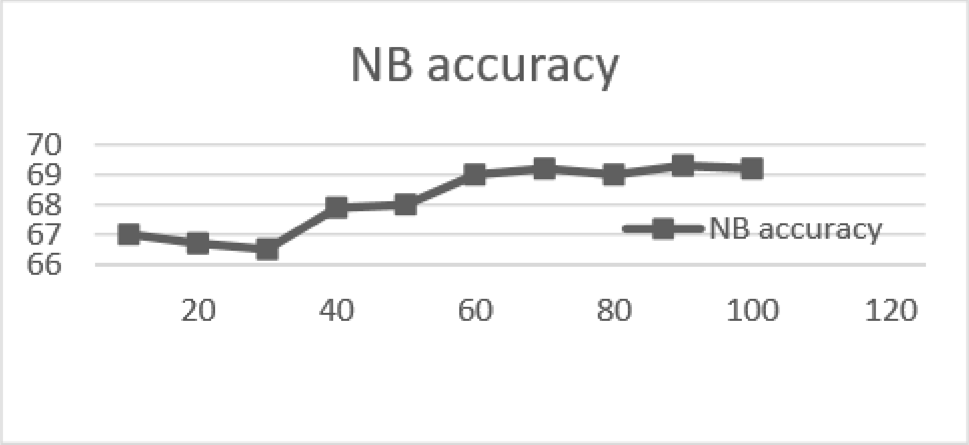


Figure4LearningcurvebasedonNB

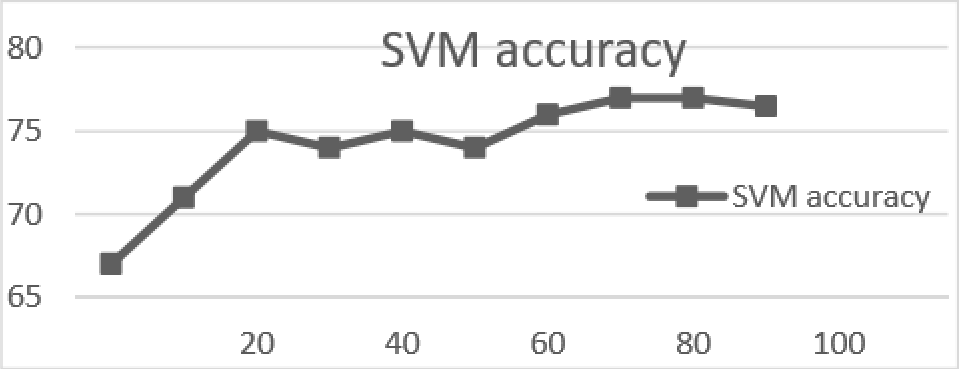


Figure5LearningcurvebasedonSVM

The findings of progressively altering the quantity of training data show that the prediction accuracy curve tends to be steady when the data set accounts for 70% to 80% of the total. Gender data was balanced in 1534 cases. It decided to utilize 1200 data points for training and 334 data points for testing in the gender prediction experiment because it would be simple to split the training data into blocks in the experiment. d) Data Dimension Selection The findings of a series of comparison tests based on logistic regression are shown in Figure 6 and Table III, and they pertain to the choice of data dimensions [20].

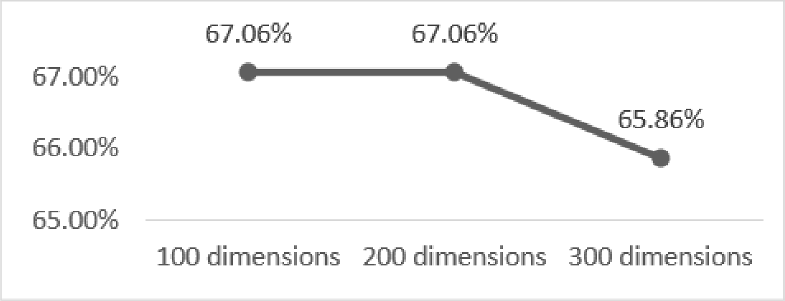
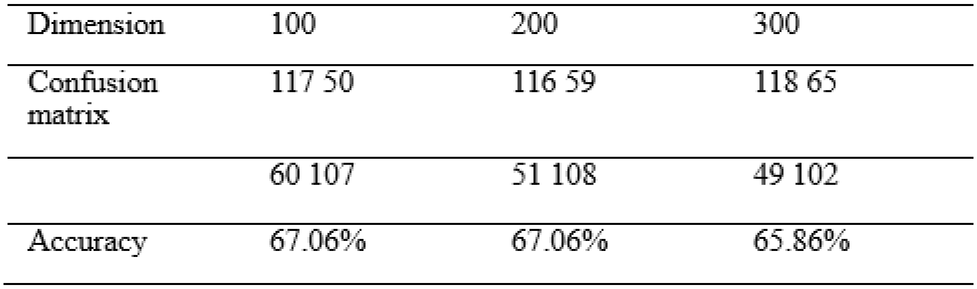


Figure6Comparisonofwordvectorsindifferentdimensions

TableIIIConfusionmatrixandaccuracyofwordvectorsindifferent

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%

The figure clearly shows that the forecast accuracy of 100, 200, and 300 dimensions is comparable, with the 100 and 200 dimensions having the highest accuracy. This is because, as the calculation time increases with each additional dimension, the accuracy of the predictions becomes more important. This work plans to utilize the feature word vector of 200 dimensions to explore because, from the confusion matrix's point of view, the diagonal values in 200 dimensions are more comparable. 33 Gender prediction procedureUsing the gender data processed in the preceding section as a starting point, Figure 7 shows the experimental flow chart for gender prediction using user sentiment on Weibo.

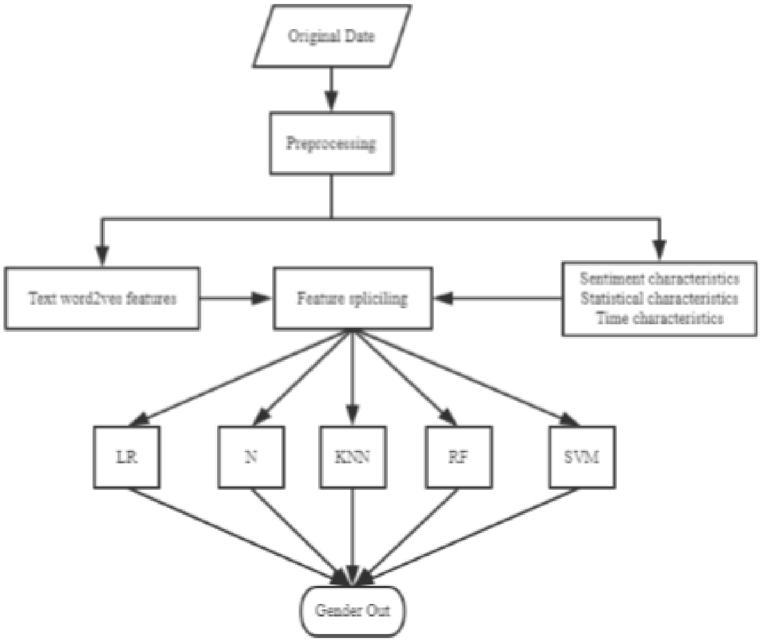


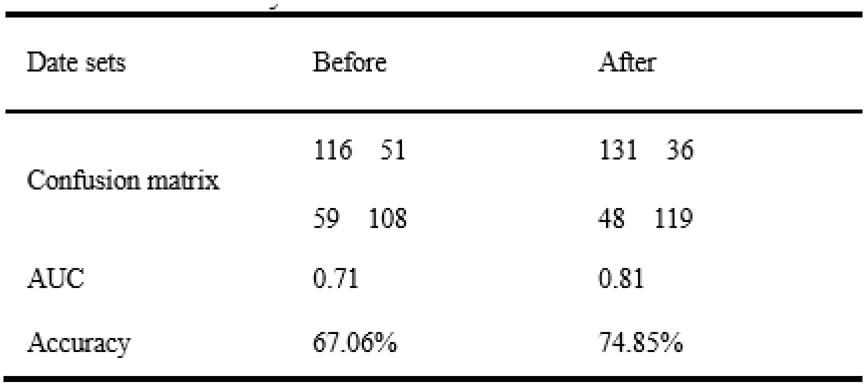
Figure7FlowchartofgenderpredictionofWeibousers

4)Resultsandanalyseofgenderprediction

a)LogisticRegressionprediction

Thedataaftervectorcombinationwasusedinthe experiment[21].The experimentalresultsareshown inTableIV.

TableIV LRAccuracybeforeandafterfusionof sentimentalfeatures



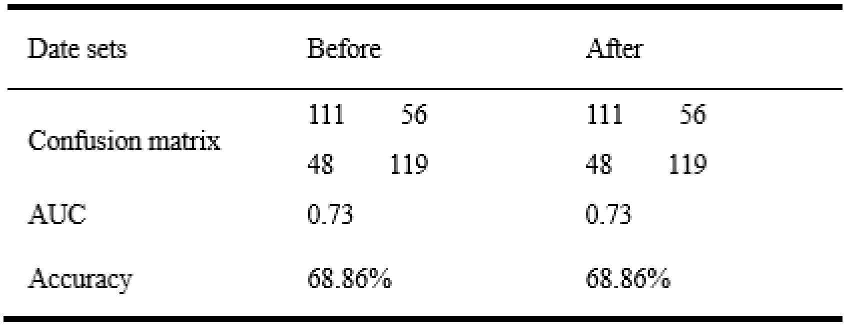
Date sets	Before	After
Confusion matrix	116 51 59 108	131 36 48 119
AUC	0.71	0.81
Accuracy	67.06%	74.85%

It is clear from Table IV that LR prediction greatly enhances the accuracy of users' sentiment fusion data. Following sentiment fusion, the accuracy of gender prediction rises to 74.9 percent, a 7.79 percent improvement from the previous level.

B NaïveBayesprediction

ThespecificresultsareshowninTable V.

Table VGenderpredictionaccuracyofN



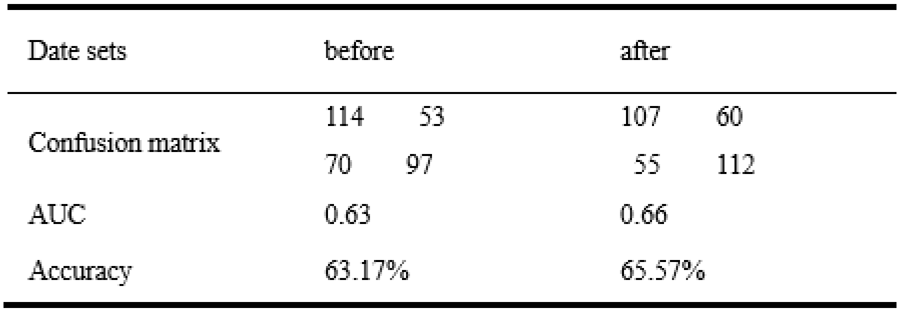
Date sets	Before	After
Confusion matrix	111 56 48 119	111 56 48 119
AUC	0.73	0.73
Accuracy	68.86%	68.86%

The NB approach was used for gender prediction [22], and the prediction accuracy remained at 68.86% both before and after the emotive features were fused.

k-NearestNeighborprediction

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

Table VIGender predictionaccuracyofKNN



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

Table VI shows that there is a 2.4% difference in the accuracy of gender prediction using k-nearest neighbor[23] and 65.57% accuracy before and after emotion fusion. But compared to the previous two approaches, the accuracy of the predictions made by these two is lower. d) Predicting a Forest at RandomOverfitting is successfully prevented by RF since samples are randomly picked from the training set. Based on the results of the RF parameter adjustment test, the following parameters are associated with this dataset in the RF method:

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.

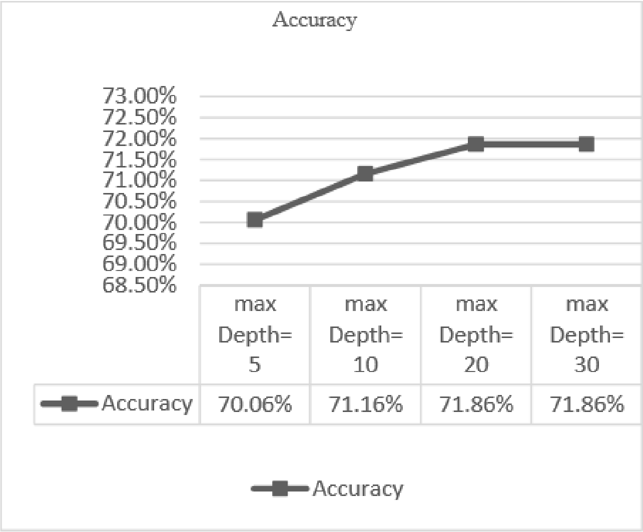
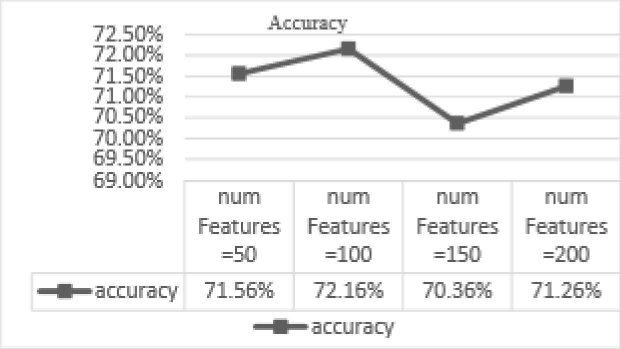


Figure 8 shows that the highest accuracy can be obtained when Max Depth= 10, that is, when the maximum Depth of the numberreaches10,and theaccuracy willnot changewhenMax Depth= 30.

1. WhenMaxDepth=10andseed=1.

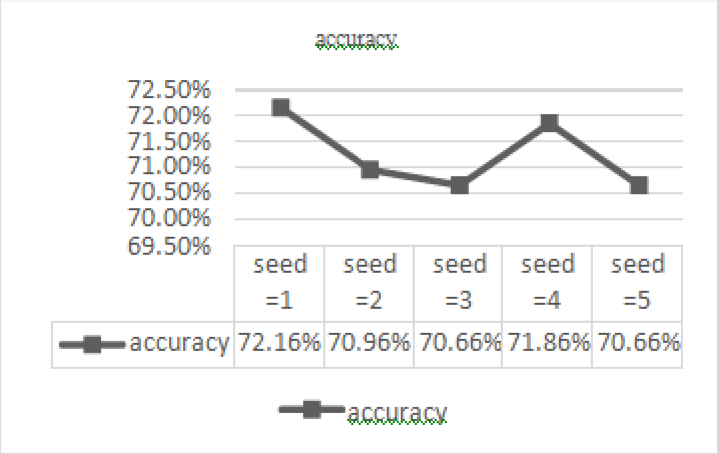


	num Features =50	num Features =100	num Features =150	num Features =200
accuracy	71.56%	72.16%	70.36%	71.26%

Figure9AccuracyofdifferentnumFeaturesvalues

Theexperiment(Figure9)accuracyisthehighest

2.whenMaxDepth=10andnumFeatures=100



	seed =1	seed =2	seed =3	seed =4	seed =5
accuracy	72.16%	70.96%	70.66%	71.86%	70.66%

Figure10Accuracyofdifferentseedvalues

According to Figure 10, increasing the number of seeds to 4 yields the best results in terms of accuracy. According to the findings shown above, the prediction accuracy reaches its peak at 72.16 percent with Max Depth= 10, num Features= 100, and seed = 1.With the same experiment run on pre-fusion data, we found that the maximum accuracy was 72.15 percent with Max Depth= 20, num Features= 100, and seed = 2, which is somewhat lower than the prediction accuracy post-fusion. and support vector machine prediction Choose a radio basis function, a polynomial kernel function, or a linear kernel function to conduct tests comparing the two kinds of C-SVC and nu-SVC. The words "before" and "after" in the graphic stand for the period before and after the merging of emotions, respectively.

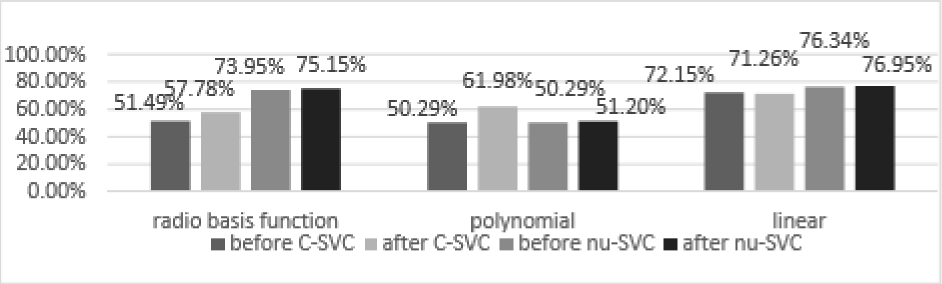


Figure11Accuracyofdifferenttypesofkernelfunctions

Comparing the nu-SVC and C-SVC types, Figure 11 reveals that the former has much better accuracy when applied to the identical kernel function. With the radio basis function as the kernel function, the accuracy gap between the C-SVC type and the nu-SVC type is often very wide. The overall outcome is the prediction of the polynomial kernel function. In most cases, the linear kernel function provides more accurate predictions. Now, after sentiment fusion, the nu-SVC type achieves its maximum accuracy of 76.95%. It was 0.61 percent more than pre-fusion levels.

53-Results Comparison

The results show that k-nearest neighbor is inadequate, and LR's performance is much worse than that of RF and SVM. Overfitting may, in theory, be prevented by RF. When it comes to making predictions, SVM has a decent track record. Figure 12 shows the results of comparing the five prediction systems' accuracy.

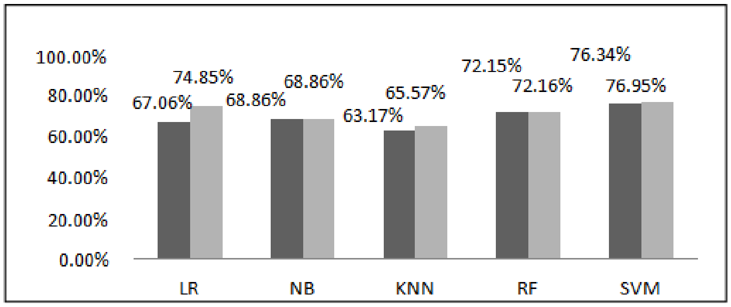


Figure12Comparisonofaccuracyofgenderpredictionbeforeandaftersentimentfusion

Figure 12 demonstrated that SVM achieved the greatest accuracy of 76.95% in gender prediction, surpassing LR, NB, KNN, and RF by 2.1%, 8.09%, 11.38%, and 4.79%, respectively. In most cases, the accuracy of predictions was better before to emotional integration than thereafter.

IV. Conclusion

Most of this study is devoted to researching how to forecast social media user profiles. The first section elaborates on the rationale and relevance of researching social media user profilings by outlining relevant literature, relevant fields of study, and the characteristics of the target audience. After that, we include the sentiment traits into the preexisting machine learning by analyzing the user's sentiment using the concept of transfer learning. On the last step, the gender of fused attitudes was predicted using five different methods: LR, NB, KNN, RF, and SVM.

This article uses gender traits of social media users to forecast how these people will be profiled. A lot of other characteristics of social media users are only baselines, even if some benefits have been accomplished. There are still several careless spots in the whole experiment procedure, and many particular connected elements need additional sorting out and analysis, even if the study concepts and experimental methodologies presented in this work have some referability.

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 MultipleAttribute Decision Making in Social Networks[J]. China Communications,10(12):37-49.

[3].Yan-QuanZ,Ying-FeiH,Hua-CanH.(2007).LearningUserProfileinthePersonalization News Service[C].

[4].KhanA,JamwalS · andSepehriM.(2010).Applying DataMiningto Customer Chum Prediction in an Intemet Service Provider,International Journal of Computer AOplications · 9(7)+8-14´

[5].ThelwallM.(2008).Socialnetworks,gender,andfriending:AnanalysisofMySpacememberprofiles[J].JournaloftheAssociationforInformationScience and Technology, 59(8):1321–1330.

[6].EyharabideV,AmandiA.(2012).Ontology-baseduserprofilelearning[J].

AppliedIntelligence,36(4):857-869.

[7].LIUB,NIUYun.GenderrecognitionofChinesemicroblogusersbasedonemotionalfeatures[J].

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

[9].LiS,WangJ,ZhouG,etal.Interactivegenderinferencewithintegerlinearprogramming[C]//International Joint Conference on Artificial Intelligence,2015:2341-2347.

[10].WANG J, LI S, HUANG L.User gender classification in Chinesemicroblog[J].JournalofChineseInformationProcessing,2014,28 (6):150-155(inChinese).

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

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

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

[14].CAOY.ResearchandApplicationonGenderClassificationofMicroblogUsers.[J] Anhui University,2019.

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