INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

SOCIAL MEDIA USER PROFILING PREDICTION USING MACHINE LEARNING ALGORITHMS

[1]M. KHAJA MODDIN, [2]D. RAMMOHAN REDDY

[1]PG Scholar, Dept. of MCA, Newton's Institute of Engineering, Guntur, (A.P)

[2]Associate Professor, Dept. of CSE, Newton's Institute of Engineering, Guntur, (A.P)

Abstract: *Identifying user attributes from their social network activities has been a common research topic nowadays. Age, gender and interest can be common user attributes which can be predicted and are essential for personalization and recommender systems. Most of the researches are based on the textual content created by user, whereas recently multimedia has gained popularity in social networks. This paper tries to propose a gender prediction technique that integrates the sentiments of social media users. This approach is used to predict users' gender attributes, as a way to identify research on the development mechanisms of social media user images. Previous studies on gender prediction have completed little assessment of feelings. This paper has used the knowledge shift idea to investigate customer sentiment and has incorporated sentiment functions into the domain of the dominant system and thus indicates higher performance in terms of accuracy than other techniques. This document specifically uses machine learning techniques to recognize the development of people's profiles on social networks. Gender attributes are studied. First, the feature is extracted for text information of media users. Then, the idea of knowledge transformation is used to analyze the sentiments of users and integrate the sentimental features into the existing system domain. Finally, 2 prediction strategies, namely Random Forest (RF) and Support Vector Machines (SVM) are used to predict the gender of the integrated sentiment.*

Keywords: *profile prediction, machine learning, feature extraction, Online Social Networking sites.*

I. INTRODUCTION

Online Social Networking sites such as Facebook and Twitter are widely used

ISSN: 2057-5688

communication medium, especially among young people. These social sites have become very popular these days and therefore it has become a research topic for studying relationship between users' digital behavior and their demographic attributes such as age, gender, relationship status, etc. Instagram and Pinterst are mainly image based social sites. Images posted by users on online social network may be useful to learn various personal and social attributes of users. We mainly extract the features from the images posted by users using their posting behavior and posted content. We use the images from Pinterest image dataset. There is a difference between male and female preferences. For male users, they are mostly interested in electronics, buildings, men clothes and so on[1]. On the other hand, female users are mainly interested in jewelry, women clothes, gardening and so on. For each user, we extract features like color, texture and shape from their collections of pins in a few different categories, such as art, cars & motorcycles and food & drinks. For posting behaviours, we focus on the users' own labeled distribution of their collections of pins over the limited number of categories provided by Pinterest. Our results suggest that both posting behavior and posted content are beneficial for gender prediction. Our contribution includes predicting the gender of the user based on the type of images posted by him/her and increasing the accuracy of the system. We frame gender classification as a binary classification problem (male and female categories) and evaluate the use of a variety of image-based features[2].

User profiling is a tagging method based on the user's personal attributes, behavior habits and preferences. The construction of user profiling method 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

ISSN: 2057-5688

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, social relations, behavioural 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, 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 user.

II. LITERATURE SURVEY

In 2014, Quanzeng You and JieboLuo and Sumit Bhatia presented a paper "A Picture Tells a Thousand WordsAbout You! User Interest Profiling from User Generated Visual Content," in which they analyze the content of individual images and then aggregate the image-level knowledge to infer user-level interest distribution. They employ image-level similarity to propagate the label information between images, as well as utilize the image category information derived from the user created organization structure to further propagate the categorylevel knowledge for all images. A real social network dataset created from Pinterest is used for evaluation.

In 2013, J. S. Alowibdi, U. A. Buy, and P. Yu presentd a paper "Empirical evaluation of profile characteristics for gender classification on Twitter,"in which they explore profile characteristics for gender classification on Twitter. Unlike existing approaches to gender classification that depend heavily on posted text such as tweets, here they study the relative strengths of different characteristics extracted from Twitter profiles (e.g., first name and background color in a user's

ISSN: 2057-5688

profile page). Their goal is to evaluate profile characteristics with respect to their predictive accuracy and computational complexity. In addition, they provide a novel technique to reduce the number of features of text-based profile characteristics from the order of millions to a few thousands and in some cases, to only 40 features. They prove the validity of their approach by examining different classifiers over a large dataset of Twitter profiles [3].

In 2012, Sridhar, Gowri, presented a paper "Color and Texture Based Image Retrieval," in which they build an interactive image recommendation system, which firstly uses color histogram feature and GCLM texture feature to express image contents, then a kernel-based K-means is utilized to cluster images into multiple classes by their visual features, finally based on a feature vector stored in the database the similar images are retrieved. The HSV color histogram is calculated and the joint histogram is derived based on the combination of the hue and saturation in the hue and saturation histogram. The color feature is extracted from the joint histogram. The chisquare is used to find the similarity between the two images. Thus, global feature is calculated using the joint histogram. The regional feature is extracted using the GCLM technique in which the neighbor pixels is considered into account. The evaluation results demonstrate the accuracy of the retrieval based on the precision and recall false positive and negative ratio. The ROC curve is used to compare the efficiency of the color, texture and the combination of both the color and the texture.

In 2010, Dr., Satyajit Singh presented a paper "Image Retrieval with Shape Features Extracted usingGradient Operators and Slope Magnitude Technique withBTC," in which novel image retrieval methods based on shape features extracted using gradient operators and slope magnitude technique with Block Truncation Coding (BTC). Four variations of proposed „Mask-Shape-

ISSN: 2057-5688

BTC" image retrieval techniques are proposed using gradient masks like Robert, Sobel, Prewitt and Canny. The proposed image retrieval techniques are tested on generic image database with 1000 images spread across 11 categories. In all 55 queries (5 from each category) are fired on the image database. The average precision and recall of all queries are computed and considered for performance analysis. In all the considered gradient operators for shape extraction, „Mask-Shape- BTC" CBIR techniques outperform the „Mask-Shape" CBIR techniques. The performance ranking of the masks for proposed image retrieval methods can be listed as Robert (best performance), Prewitt, Sobel and lastly the Canny.

Michael Fairhurst et al., presented a paper "Using keystroke dynamics for gender identification in social network environment." In which they introduce an approach to addressing risks such as risk of transactions with individuals who deliberately conceal their identity or, importantly, can easily misrepresent their personal characteristics. They use a form of biometric data accessible from routine interaction mechanisms to predict important user characteristics, thereby directly increasing trust and reliability with respect to the claims made tomessage receivers by those who communicate with them [7].

III. PROPOSED WORK

We frame the task of predicting users' gender from their posted images as a binary classification task (fig 1). Given a set of images posted by a user on a social networking site, we predict whether the user is male or female. We suggest that males and females differ in terms of their image posting behavior as well as in the content of posted images. We extract features to capture visual content of images as well as users' posting behavior

Gender prediction of social media profiling

Approach of Learning sentiment feature

ISSN: 2057-5688

The sentimental characteristics of social media users of different genders are different, we need to extract the sentimental characteristics of users. However, the sina Weibo data used does not contain sentimental labels, so the LSTM(Long short-term Memory) neural network 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.

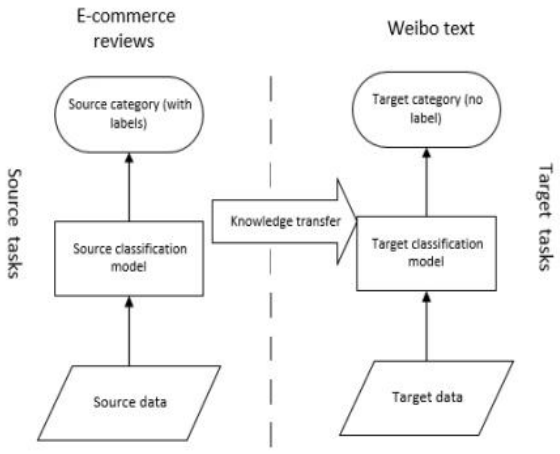

Fig.1 Proposed system architecture

This paper analyzes the sentiments of Weibo users by dividing the texts into a series of words to predict whether they are positive or negative. The input format of LSTM neural network is vector. Word2vec is used to analyse 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

ISSN: 2057-5688

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.

Dataset is been prepared using different categories of image posted by users. These images are collected from Pinterest websites (fig 2). Pinterest: Pinterest is a free website that requires registration to use. Users can upload, save, sort and manage images known as pins and other media contents (e.g., videos and images) through collections known as pinboards. Pinterest acts as a personalized media platform User data: Like Facebook and Twitter, Pinterest now let marketers access the data collected on its users. By granting access to users' data, Pinterest lets marketers investigate how people respond to products. If a product has a high number of repins, this tells the producer of the product that it is liked by many members of the Pinterest community. Now that Pinterest lets marketers access the data, companies can view user comments on the product to learn how people like or dislike it. People use social media sites like Pinterest to direct or guide their choices in products. Sample dataset which is been collected is shown below:

Fig.2 Sample dataset of Pinterest

We are using color, texture and shape feature extraction techniques for extracting the features from the images. We are extracting 102 color features, 4 texture features and 124 shape features. These 230 features are the most important features of an image. Color

ISSN: 2057-5688

Feature: Color is an important feature for image representation which is widely used in image retrieval. This is due to the fact that color is invariance with respect to image scaling, translation, and rotation. The key items in color feature extraction consist of color space, color quantization, and the kind of similarity measurements. Color Feature can be extracted using color moment, color histogram, and Color Coherence Vector (CCV). Color Histogram is commonly based on the intensity of three channels. It represents the number of pixels that have colors in each of a fixed list of color ranges. Color Moment is used to overcome quantization effect in color histogram. It calculates the color similarity by weighted Euclidean distance. Color set is used for fast search over large collection of images. It is based on the selection of color from quantized color space.

Supervised Learning Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

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·

ISSN: 2057-5688

The number of random seeds used.

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

support vector machines (SVM)

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. SVM classification will be used to classify whether the user is male or female. In machine learning, support vector machine (SVM) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis

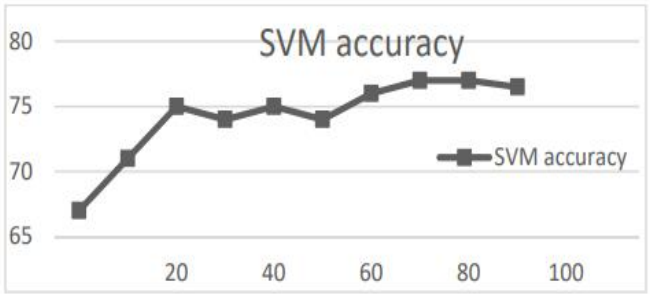

Fig.3 SVM accuracy

IV. RESULTS

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 showed that the accuracy of SVM for gender prediction was the highest, reaching 76.95%, which was

ISSN: 2057-5688

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.

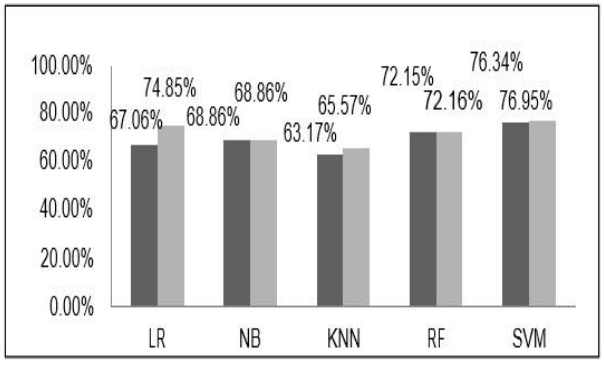

Fig. 4 Comparison of accuracy of gender prediction before and after sentiment fusion

V. CONCLUSION

This paper mainly studies the prediction of social media user profiling. Firstly, starting from the background knowledge, related work and target characteristics of social media user profiling, expounds the purpose and significance of studying social media user profiling. Then the idea of transfer learning is used to analyse 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 profiling 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 analysed.

REFERENCES

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

[2]. Prasadu Peddi (2016), Comparative study on cloud optimized resource and prediction using machine learning algorithm, ISSN: 2455-6300, volume 1, issue 3, pp: 88-94.

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

ISSN: 2057-5688

[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

[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]. Uday Chandrakant Patkar, Sushas Haribabu Patil and Prasad Peddi, "Translation of English to Ahirani Language", *International Research Journal of Engineering and Technology(IRJET)*, vol. 07, no. 06, June 2020.

[10]. N. Srivani, Dr Prasadu Peddi, "Efficient Fr a Geometrical-Model-Based Face Segmentation and Identification in Terms of Identification the Face ", *JFCR*, pp. 1283-1295, Jun. 2022.