electronics

Article

User OCEAN Personality Model Construction Method Using a BP Neural Network

Xiaomei Qin [1], Zhixin Liu [2], Yuwei Liu [3], Shan Liu [3], Bo Yang [3], Lirong Yin [4], Mingzhe Liu [5,*] and Wenfeng Zheng [3,*]

[1] College of Translation Studies, Xi'an Fanyi University, Xi'an 710105, China
[2] School of Life Science, Shaoxing University, Shaoxing 312000, China
[3] School of Automation, University of Electronic Science and Technology of China, Chengdu 610054, China
[4] Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA 70803, USA
[5] School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou 325000, China
* Correspondence: liumz@cdut.edu.cn (M.L.); winfirms@uestc.edu.cn (W.Z.)

Highlights:

What are the main findings?

- First, the combination of the methods of machine learning with psychological methods to predict the user's OCEAN personality model could achieve a higher accuracy;
- Second, the model proposed in this paper that is a combination of LDA plus BP neural network is generally superior to the combination model of the same type.

What is the implication of the main finding?

- First, through digital footprints and understanding the rules of user behavior, user behavior can be predicted and targeted recommendations made;
- Second, it was found that predicting the user's OCEAN personality model and then their behavior can provide an effective method for micro-directional recommendations in network communication.

Abstract: In the era of big data, the Internet is enmeshed in people's lives and brings conveniences to their production and lives. The analysis of user preferences and behavioral predictions of user data can provide references for optimizing information structure and improving service accuracy. According to the present research, user's behavior on social networking sites has a great correlation with their personality, and the five characteristics of the OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) personality model can cover all aspects of a user's personality. It is important in identifying a user's OCEAN personality model to analyze their digital footprints left on social networking sites and to extract the rules of users' behavior, and then to make predictions about user behavior. In this paper, the Latent Dirichlet Allocation (LDA) topic model is first used to extract the user's text features. Second, the extracted features are used as sample input for a BP neural network. The results of the user's OCEAN personality model obtained by a questionnaire are used as sample output for a BP neural network. Finally, the neural network is trained. A mapping model between the probability of the user's text topic and their OCEAN personality model is established to predict the latter. The results show that the present approach improves the efficiency and accuracy of such a prediction.

Keywords: OCEAN personality model; digital footprint; LDA topic model; neural network

check for updates

Citation: Qin, X.; Liu, Z.; Liu, Y.; Liu, S.; Yang, B.; Yin, L.; Liu, M.; Zheng, W. User OCEAN Personality Model Construction Method Using a BP Neural Network. *Electronics* **2022**, *11*, 3022. https://doi.org/10.3390/electronics11193022

Academic Editors: Els Lefever and Veronique Hoste

Received: 12 July 2022
Accepted: 16 September 2022
Published: 23 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The Internet has now entered into every corner of people's daily lives. People are getting used to sharing bits and pieces of their life on social networking sites and interacting frequently through the Internet, whereby they leave many digital footprints [1]. There are also significant differences in the social behavior of users with different personality traits on the Internet [2].

In the 1990s, scientists discovered through the method of lexicology that there were about five characteristics that can cover all aspects of personality description. Psychologists developed a model that includes a total of five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. These five factors provide a rich personality structure commonly known as "The Ocean of Personality", that is, the "OCEAN Personality Model" [3,4]. After the development of the OCEAN personality model, scientists have carried out many meaningful studies based on the theory, such as Nasiri et al. [5], who studied the bilateral multi-issue bargaining model based on personality models in e-commerce to predict suitable candidates, and it was found that the compound personality style behavior is suitable for obtaining the best overall utility in terms of the role of buyers and sellers as well as social welfare and market activity. The work of Roussmann et al. [6] confirms that personality is related to the situations people encounter and their interpretations of them, so it is possible to predict how they will act in their daily lives through the OCEAN personality model. Lopez et al. [7], who attempted to automate personality assessments from transliterations of YouTube vlogs using classic and state-of-the-art word embeddings, argue that personality trait studies based on language patterns also need to take into account the knowledge of psychologists and psycholinguists to collect and label data. Holman et al. [8] researched the interaction between the Big-5 personality traits and work characteristics, and it was found that the work characteristics had an impact on the personality traits. Rodriguez et al. [9] performed a social network personality trait analysis based on shared image weak supervised learning and trained an OCEAN personality model based on convolutional neural networks (CNNs) that matched images to people with a given level of features to a high degree.

The research of psychologists has found that the OCEAN theoretical model has a strong connection with the digital footprints of users on social networking sites. The OCEAN personality model can effectively predict the behavior of users on social networks [10]. Bowden et al. [11] reviewed the research on the relationship between an extroverted personality and social media use and divided the results into six areas: (1) content creation, (2) content response, (3) user profile characteristics, (4) usage patterns, (5) perception of social media, and (6) aggression, malicious attacks, and overuse. The results showed that extroverted personalities behave more positively on social media. Gallo et al. [12] predicted user reactions to Twitter feed content based on personality type and social cues and found that predictions based on the OCEAN personality model yielded the best results. Hossain et al. [13] found that people who scored high on responsibility, neuroticism, openness, and low affinity in the OCEAN personality model were prone to cyberbullying and that an analysis of personality could help reduce cyberbullying. Therefore, it is of great significance to analyze the digital footprints of users on social networking sites, to make a relatively accurate evaluation of the needs of each user based on the user's personality and style, and then to provide users with micro-directional content dissemination.

At present, there are two general types of algorithms for predicting user OCEAN personality models:

1.1. An Algorithm That Is Based on Linguistic Features of Users' Texts

First, the text status that a user has published on a social networking site they use was obtained. Then, text features were extracted from the user's text states using natural language analysis tools. The text analysis tools used by the researchers in this algorithm are the emotion analysis classifier, which can identify the emotional polarity of the user's sent state as positive, negative, or neutral; and the text analysis tool, which also includes the natural language word segmentation statistics tool, and which extracts the punctuation marks, verbs, nouns, emotional phrases, and complexity of the text in the user's text state, and then matches these text properties to the OCEAN personality model.

Argamon et al. [14], who first devoted themselves to the field of OCEAN personality model prediction, used the Sequential Minimal Optimization (SMO) algorithm to estimate both the exotropic and neurotic features, and the accuracy of the experimental results was between 57% and 60%. Mairesse et al. [15] used the text analysis tool Linguistic Inquiry and Word Count (LIWC) to predict the user's OCEAN personality model through the support vector machine and M5 model, and the accuracy of the experiment ranged from 54% to 62%. Oberlander and Nowson [16] used users' linguistic features to identify a user's OCEAN personality model through the SMO algorithm and the naive Bayesian algorithm, and they obtained an accuracy rate of 83%~93%. However, when the algorithm is applied to a large data set, it will lead to overfitting and reduce the accuracy to about 55%. Wang et al. [17] proposed a deep learning framework that combines XLNet with the Personality Classification Capsule Network (XLNet-Caps) from text posts and found that personality could be effectively classified, and the average recall rate could reach 77.1%, but the value of F1 was only 68%.

A meta-analysis of studies that predicted Big-5 personality traits from digital footprints on social media was carried out by Azucar D et al. [18], who found that this predictive pathway was feasible and accurate. Al Marouf et al. [19] researched the correlation of conjunction words to predicting personality traits on social media. They found that the percentage of connectives positively correlated with personality traits.

The algorithms above based on the linguistic features of user texts were not particularly accurate, and the results were relatively simple. This is due to the fact that the text sent by the user contains limited information, and the linguistic characteristics of the text are relatively simple according to a dictionary. To improve accuracy, some scholars proposed a second type of algorithm for predicting personality:

1.2. An Algorithms That Integrates Users' Linguistic Features with the Characteristics of Their Social Networks

Researchers who apply this research algorithm analyze user behavior from data by obtaining large-scale user data on social networking sites and create mapping between these behavioral data and the characteristic dimensions of the OCEAN personality model through regression analysis and clustering. Thus, the user's digital footprint on the social networking site can be identified in a timely manner.

Bai et al. [20] used the M4.5 algorithm to identify OCEAN personality model based on the number of friends of the website users and their recently released dynamics, with an accuracy rate of 69% to 72%.

Camacho et al. [21] proposed four dimensions of social network analysis. Gallo et al. [12] adopted a simplified version of the Network Knowledge Base (NKB) model to tackle the problem of predicting basic actions that a user can take given the content of their social media feeds.

However, there are still some disadvantages in the recognition methods for the user OCEAN personality model. The neural network algorithm, which is one of the current methods with a comparative advantage, can be quickly adjusted to fit many direction problems. The purpose of this paper is to improve the identification method of the OCEAN personality model with a neural network algorithm. First, the text feature extraction method based on the Latent Dirichlet Allocation (LDA) topic model [22] was analyzed, and then the prediction method of user OCEAN personality model by the back-propagation neural network (BPNN) was proposed. Finally, the effectiveness of the proposed method was verified through experiments. Compared with the previous research (Table 1), the accuracy, recall rate, and F1 of the method in this paper are high, and the performance is relatively balanced and stable, so it can better predict the user's OCEAN personality model.

Table 1. Comparison of 6 different methods.

Source	Methods	The Best Effect	Year	Bug
This paper	LDA-BPNN	Accuracy: 74% Recall rate: 70% F1: 75%	-	Not found
Argamon et al. [14]	SMO	Accuracy: 60%	2011	Not found
Mairesse et al. [15]	LIWC	Accuracy: 62%	2007	Not found
Oberlander and Nowson [16]	SMO and the naive Bayesian algorithm	Accuracy: 93%	2006	When used at large number of dataset, the accuracy went to 55%
Wang et al. [17]	XLNet-Caps	Recall rate: 83%	2021	The value of F1 was about 68%
Bai et al. [20]	M4.5	Accuracy: 72%	2012	

The main tasks of this paper are as follows:

(1) we obtained experimental data from Sina Weibo users;
(2) we used an LDA model to mine the semantic features of texts;
(3) we built an OCEAN personality model of identification for users based on a BP neural network.

2. The Datasets

Sina Weibo is one of the mainstream social networking sites, with a large number of active users from whom the data is easy to obtain. This study selected Sina Weibo as the main research scenario to explore the relationship between the user's OCEAN personality model and their digital footprints.

There are two parts to the datasets in this paper: the user's OCEAN personality model and the user's behavior data. Finally, we classified all of the data above.

2.1. Users' OCEAN Personality Model

In order to provide accurate experimental data support for the learning model, this study distributed an online personality test scale, inviting people of different ages to participate in the survey, and then established the basic dataset with the results of this survey.

The most recognized OCEAN personality model test scale used in psychology, called the NEO Personality Questionnaire, was selected in this study. The NEO Personality Questionnaire is an OCEAN personality model questionnaire based on the structure of five aspects of the human personality. The questionnaire included questions about both self-assessment and assessment of others. The questionnaire consists of 60 statements, such as "I would rather not decide for a while than plan everything beforehand". Each question tests a feature dimension of the OCEAN personality model, and each feature dimension consists of 12 related questions. About one-third of the problems are negatively correlated, and about two-thirds of the problems are positively correlated. The options were scored on a scale from "strongly disagree" (1 point) to "strongly agree" (5 points) to indicate the user's bias on the topic in the dimension of the measured feature [23].

In addition, before the test users filled in the NEO Personality Questionnaire, a simple survey of their personal account information and the use of their microblog was conducted. The contents were as follows:

(1) personal account information: the user's gender, age, education level;
(2) Weibo usage: the nickname of Sina Weibo, the usage frequency of Weibo.

After screening, 244 users' questionnaires were obtained, and the OCEAN personality models of the 244 users were obtained through mathematical calculation. To ensure the authenticity of the data, the specific screening criteria were as follows:

(1) topic users provide an accurate Weibo account;
(2) the questionnaire answer time was more than 200 s;
(3) the number of microblogs in the corresponding microblog accounts of the topic users was more than 100.

Results were counted for the screened questionnaire and processed to obtain a feature dimension value between 0 and 5. Table 2 gives seven effective user cases, for example.

Table 2. Effective cases of the user OCEAN(Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) personality model.

No.	Openness (O)	Conscientiousness (C)	Extraversion (E)	Agreeableness (A)	Neuroticism (N)
1	3.41667	3.41667	3.66667	2.91667	3.16667
2	2.33333	2.66666	2.5	3.5	3.08333
3	2.66667	3.33333	3.16667	3.58333	2.5
4	3.75	3.75	4.16667	3.5	2.5
5	2.91667	4.08333	2.66667	3.16667	1.75
6	3	2.66667	4	4	2.66667
7	2.5	1.5	3.25	2.5	4.41667

2.2. Users' Behavior Data

Since the public application programming interface (API) of Sina Weibo limits the amount of data and the scope of the content of a web crawler, a web crawler program based on the JAVA language was developed in this paper in order to obtain the basic account information and all microblog information of the Sina Weibo users. First, we logged into the start page of Sina Weibo through simulation, obtaining a series of returned parameters and cookies, and then the website was requested to obtain a response by simulating the browser. After obtaining the response, the extraction of Weibo home page information was first parsed by JSONObject, and then jsoup was used to parse the data and extract the required Weibo home page content.

In this study, the corresponding Sina Weibo information of the 244 topic users was crawled. The information obtained included the user's account information and the user's microblog information.

The basic information of the user's account included the user's name, registration time, number of followers, number of fans, and number of microblogs. In order to eliminate interference from the excessive number of microblogs posted by some users to the post-text experiment, this study crawled the top 300 microblogs of users to be applied as users' text information. The text information obtained from each microblog user was stored separately in a single text.

2.3. Validation Dataset

In the 2013 International AAAI Conference on Weblogs and Social Media (ICWSM) [23] conference proceedings, researchers used a unified Facebook dataset, which was important for measuring the validity of their model. This paper uses the same dataset to verify the validity of our user OCEAN prediction model and to ensure the uniformity of the test dataset.

2.4. Data Classification

After obtaining the data required for the experiment, Weibo users were divided into two groups, A and B, randomly. A was used as a training set, accounting for 80% of the total dataset; B was used as the test set, accounting for 20% of the total dataset.

3. Experimental Methods and Procedures

The prediction of the user OCEAN personality model based on the BP neural network can mainly be divided into three stages: data acquisition and collation, model construction and training, and effect verification and evaluation (Figure 1).

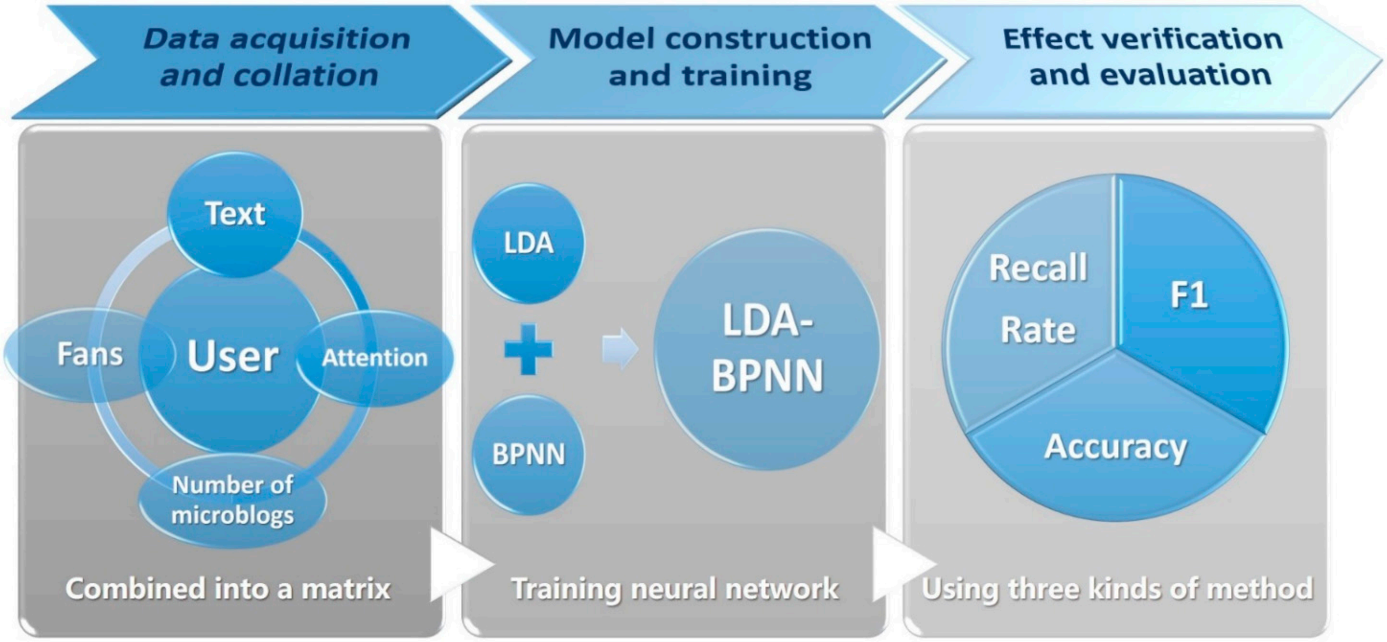
Data acquisition and collation
Model construction and training
Effect verification and evaluation
Text
Fans
User
Attention
Number of microblogs
Combined into a matrix
LDA
BPNN
LDA-BPNN
Training neural network
Recall Rate
F1
Accuracy
Using three kinds of method

Figure 1. Flowchart of user OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) personality model prediction based on BP neural network.

Data acquisition and collation: The data were collected using the methods mentioned in Section 2 of this article, the user's OCEAN personality model and behavior data were obtained, and the obtained data were screened and crawled.

Model building and training: First, for modeling, LDA models and BPNN models were built. Second, for input, content included user text characteristics, user followers, user attention, and the number of microblogs sent by the user. These features were combined into a matrix that served as input to the BP neural network. When the user text feature was taken from the "text-topic probability distribution file" obtained through the LDA topic model in the previous subsection, the file is a matrix, the rows represent a Sina Weibo user, the columns represent a topic, and the element values represent the probability distribution. The above data was used to train a BP network with thirteen inputs (each corresponding to ten topic features plus three user account features). Third, for output, the results of the user's OCEAN model were counted, which were obtained from the questionnaire above. After that, the data were processed, and then the data for each dimension was divided into five categories as sample output. Fourth, for training, the BP trained the test text, established a correspondence between the user's Weibo file and the user's OCEAN model, and identified the OCEAN model of the measured user.

Effect verification and evaluation: In verifying the effectiveness of this study experiment, this paper compares other methods with the same evaluation criteria and dataset to verify the effectiveness of the experimental improvements in this research. Specific evaluation indicators include accuracy, recall rate, and F1 value.

Data collection and collation has been introduced in Section 2, which focuses on model construction and training, as well as effect verification and evaluation.

3.1. Model Construction

There are two main models used in this paper: the LDA topic model was used to extract text features, and the BP neural network model was used for OCEAN personality model prediction.

3.1.1. The LDA Topic Model and Text Features Extraction

(1) The LDA topic model

Short texts feature sparseness, which makes them no longer effective for a method of measuring the similarity between two texts according to repeated words [24]. In the two texts, perhaps no words are the same even though the topics are very similar. Therefore, when judging whether two texts are similar, it is necessary to consider the hidden features of the text [25], such as "I like to use iPhone" or "Steve Jobs is amazing."

These two sentences are relatively similar in topic, but without any shared words. According to the traditional TF-IDF algorithm, these two sentences are judged as having no similarity. However, the LDA topic model can mine the hidden topics of the text. Therefore, this paper uses the LDA model to mine the semantics of the texts.

Latent Dirichlet Allocation (LDA) is a model that finds potentially meaningful topics in multiple texts. Then, to generate a text, the probability of each word appearing is as Equation (1):

$$p(Word|Document) = \sum_{Topic} p(Word|Topic) \times p(Topic|Document) \tag{1}$$

The process generated by the topic is converted into a probability map, as shown in Figure 2:

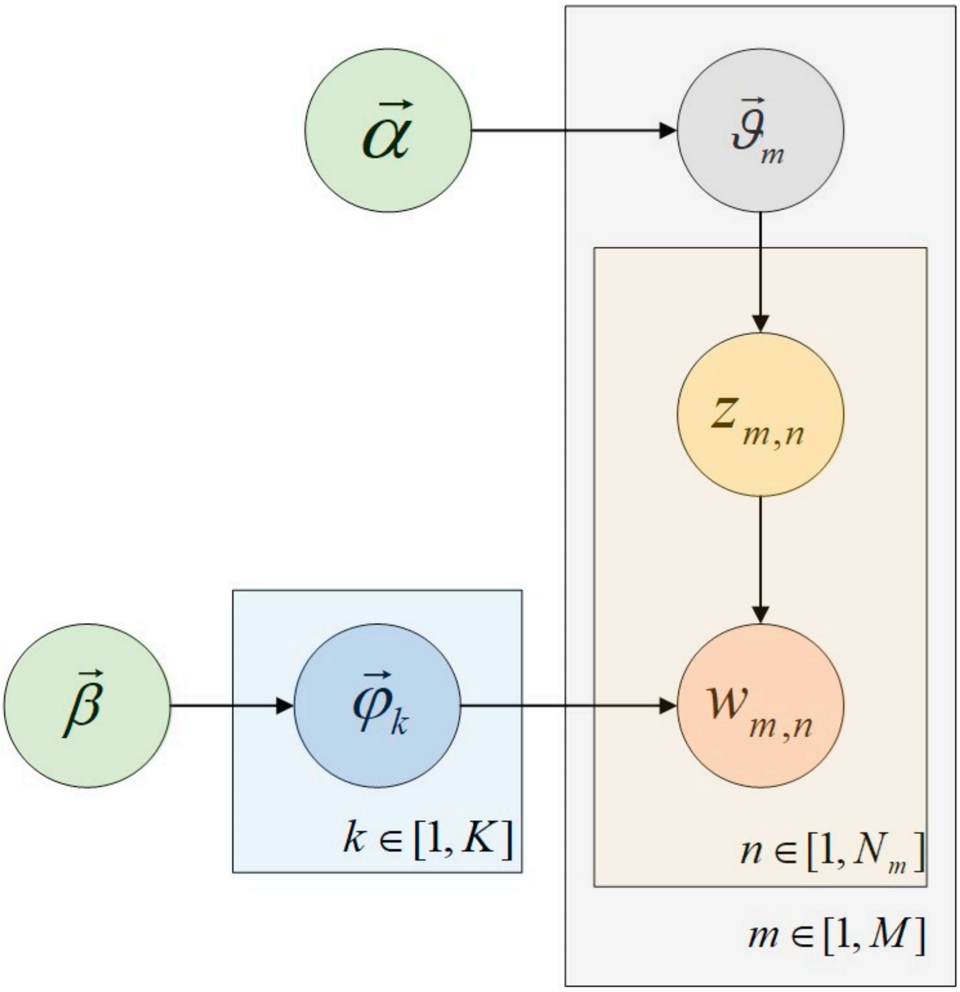

Figure 2. The Latent Dirichlet Allocation (LDA) model diagram.

The symbols in the figure are as in Table 3.

Table 3. Symbol representation of the LDA topic model.

Symbol	Description
M	There are *M* texts
K	Number of topics
N$_m$	The total number of words in the *m*-th text
α	A priori distributed parameters of the topic distribution of each text
β	Parameters of the prior distribution of the word distribution for each topic
$\vec{v}_m$	Indicates the topic distribution of the *m*-th article
$\vec{\phi}_k$	Word distribution representing the *k*-th topic
$z_{m,n}$	The topic of the *n*-th word of the *m*-th article
$w_{m,n}$	The nth word of the *m*-th article

According to the LDA probability map and symbol representation, the text generation process of the LDA topic model is:

Step 1: Choose a $\vec{\vartheta}_m \sim Dir(\vec{\alpha})$
Step 2: For each word $\omega_{m,n}$ to be generated

a. Select a topic $z_{m,n} \sim Multinomial(\vec{\vartheta}_m)$

b. Generate a word $w_{m,n} \sim P(w_{m,n}|z_{m,n}, \vec{\beta})$

where ϑ is a topic vector, the column of the vector represents the probability distribution value of a topic on the text, $Dir(\vec{\alpha})$ is the distribution of the topic *Dirichlet*, and $P(w_{m,n}|z_{m,n}, \vec{\beta})$ represents the probability of the occurrence of the word w when determining the topic z.

(2) Extraction of the user's text features

JGibb LDA (Version 1.0 created by Xuan-Hieu Phan & Cam-Tu Nguyen) [26] is an implementation of a java version of the LDA topic model; the Gibbs Sampling technique is used for parameter estimation and inference. This paper made some modifications to the JAVA source code of JGibb LDA to accommodate the application scenario of the study, including: First, the probability was retained to 10 decimal places to reduce the volume of the file. Second, ik Chinese word segmentation (Version 6.0.0 created by Medcl from INFINI Labs) [27], a Chinese word separation tool, was added on the basis of the source code to make the model applicable to Chinese scenarios. Third, some small improvements were made to the data structure of the source code, improving the operability of the code in this study and the running speed of the model.

Table 4 below shows the input and output of the LDA:

Table 4. Input and outputs of the LDA topic model algorithm.

Input and Output of the LDA Algorithm	
Algorithm input:	A collection of all user text texts, the number of topic *K*, the hyperparameter α and β
Algorithm output:	The output of the LDA topic model consists of the following five files: 1. *params.txt: Information profile 2. *phi.txt: Word-topic probability distribution file 3. *theta.txt: Text-topic probability distribution file 4. *tassign.txt: Text term-topic distribution file 5. *twords.txt: Words-topic inference file

According to the usual experience value [28], combined with the application scenario of this study, set

$$K = 10,\ \alpha = \frac{K}{50},\ \beta = 0.01 \tag{2}$$

The entire output of the five files is not required; according to different research scenarios, different files are chosen to be used. In the scenario of this study, two texts, “words-topic inference file” and “text-topic probability distribution file”, were used.

After 100 iterations, the program outputs the words and probabilities that best represent each topic under 10 topics. Figure 3 shows some of the results of the “words-topic inference file”.

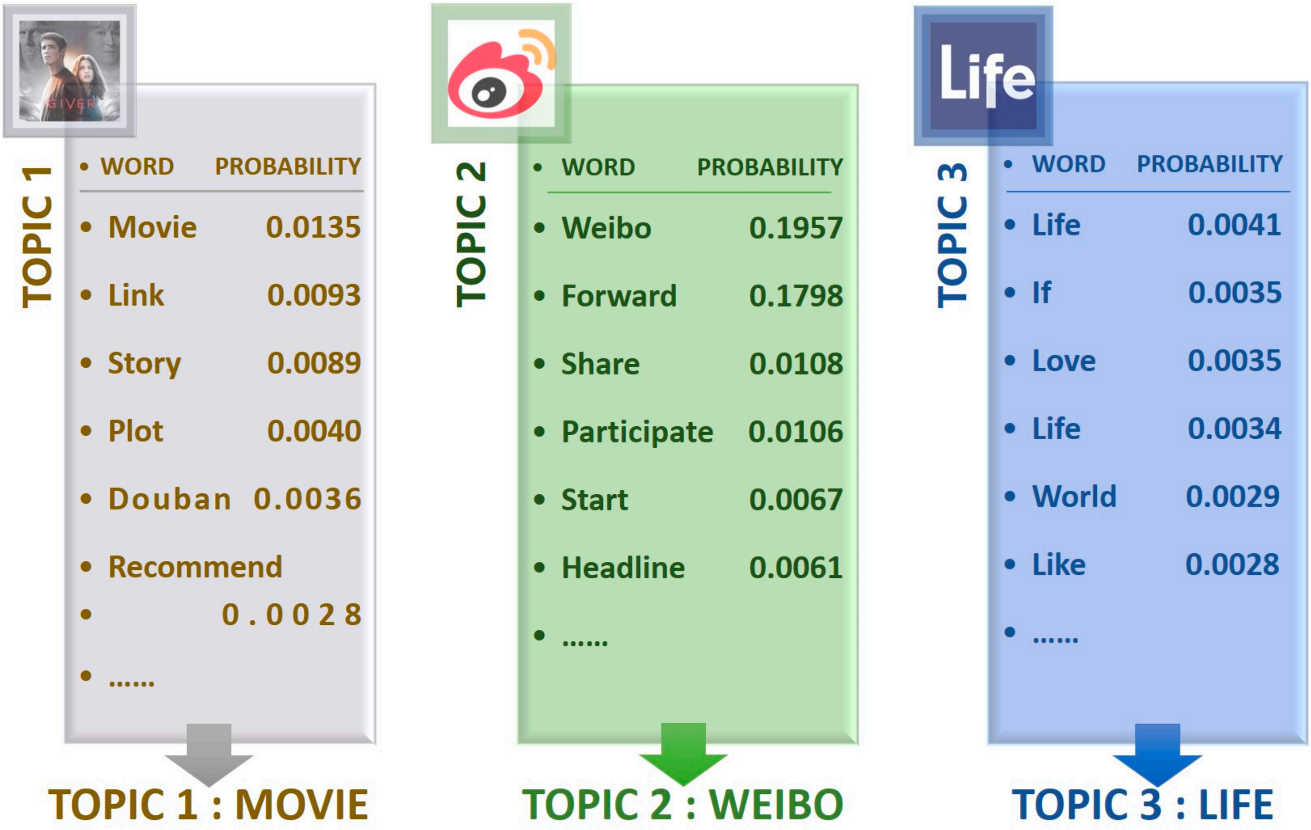

Figure 3. “Words-topic inference file” result example plot.

We can obtain the distribution of topic words analyzed by the LDA theme model in Figure 3. The first topic is related to movies, the second topic is related to activities such as retweeting and sharing on Weibo, and the third topic is related to people’s daily life.

Table 5 shows the partial results of the “text-topic probability distribution file”.

Table 5. Text-topic probability distribution file.

	Topic0	Topic1	Topic2	Topic3	Topic4	...	Topic9
Text 1	0.09108	0.04472	0.04000	0.07381	0.41319	...	0.35473
Text 2	0.32428	0.08246	0.38553	0.03569	0.10418	...	9.41210
Text 3	0.02095	0.52601	0.13106	0.00308	0.28762	...	0.00671

This result represents the probability distribution of 10 topics for the texts of three users tested. Each row represents a user, each column represents a topic, and the element value was the distribution value of the “text-topic probability distribution”.

According to the LDA topic model, the text features of the user texts are extracted by obtaining the topic probability distribution of each user text.

3.1.2. The Multi-Classification Network of User OCEAN Personality Model

The BP neural network is the most representative neural network learning method [29]. This study used a BP neural network to identify a user’s OCEAN personality model. The structure of the model in this study is shown in Figure 4.

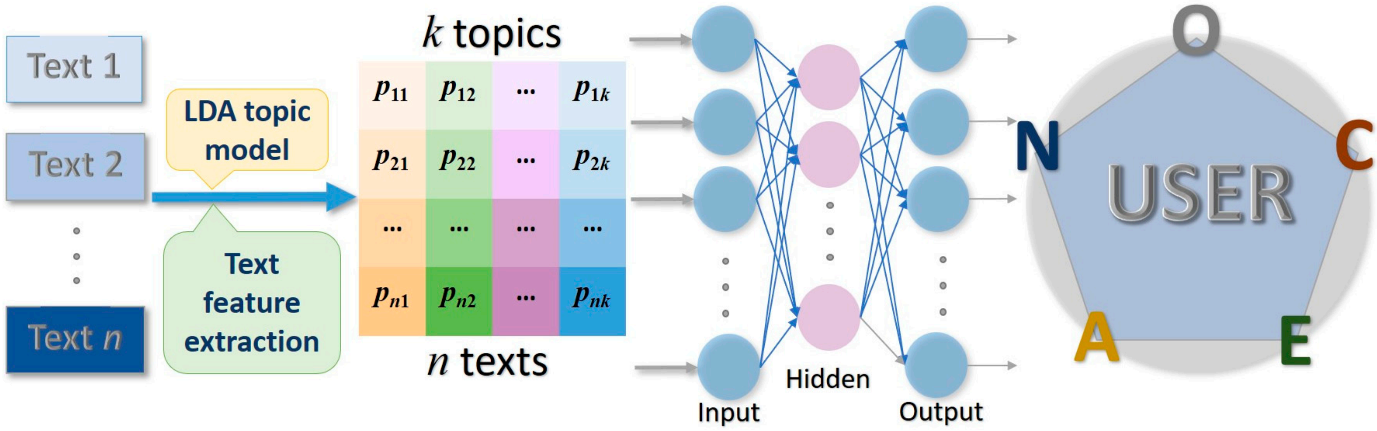

Figure 4. OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) personality model identification structure.

First, the data of each dimension of the OCEAN personality model obtained from the statistics of the questionnaire are divided into five categories. Scores between 0 and 1 are classified into one category, and scores between 1 and 2 are classified into two categories, and so on. The five types of data are numbered 1, 2, 3, 4, and 5. For the input sample, it is necessary to identify which classification belongs to each of the five dimensions of the OCEAN personality model, so each feature dimension of the OCEAN personality model is used as a multi-classification problem based on BP neural network [30].

For example, if the five features of a user's OCEAN model are separate, as in $OCEAN = \{1.750,\ 4.083,\ 4.416,\ 3.250,\ 4.583\}$, then, in the sample output of the model, the sample output matrix of that user can be expressed as Figure 5.

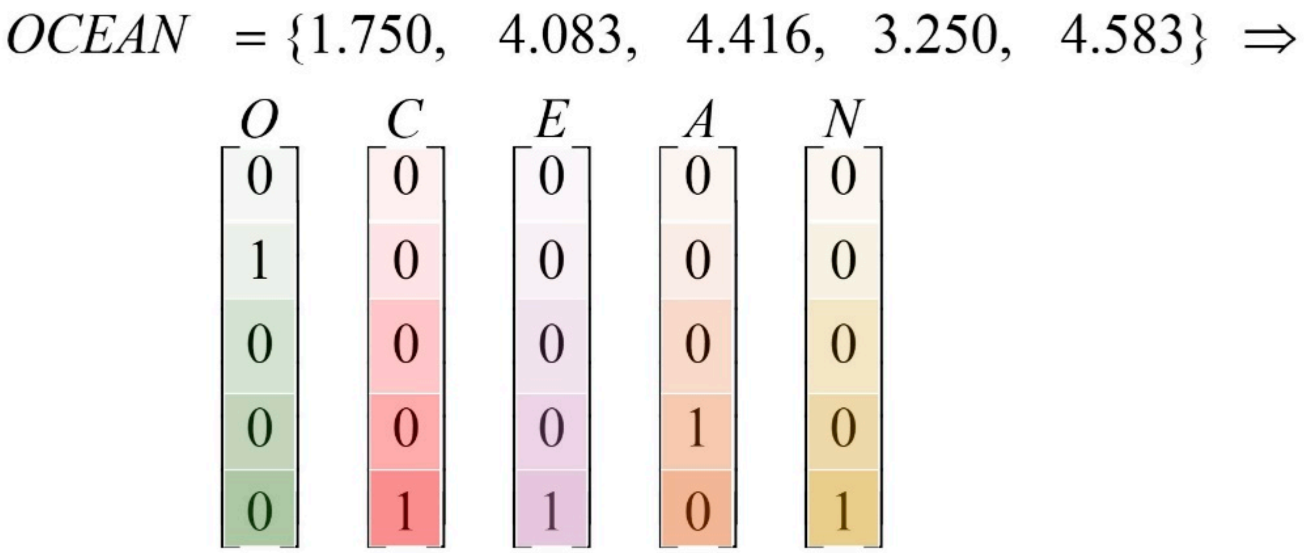
$OCEAN = \{1.750,\ 4.083,\ 4.416,\ 3.250,\ 4.583\} \Rightarrow$
O C E A N
0 0 0 0 0
1 0 0 0 0
0 0 0 0 0
0 0 0 1 0
0 1 1 0 1

Figure 5. Schematic diagram of multi-category output. The letters O, C, E, A, and N represent "Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism", respectively.

In other words, there are five multi-classifiers in the prediction experiment of the OCEAN personality model. For example, the results of sample inputs and outputs for the openness (openness) dimension of the OCEAN personality model can be expressed as Figure 6.

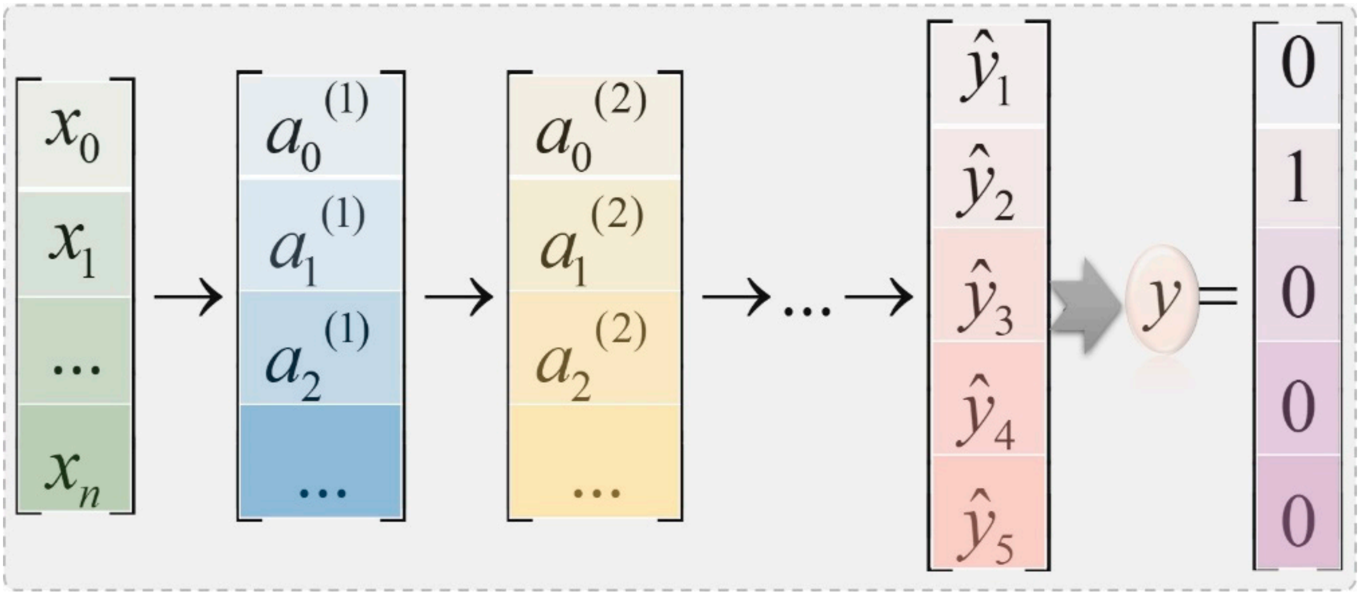

Figure 6. Schematic diagram of multi-category output.

The $x_1 \sim x_m$ are the input signals from a neuron on the previous layer. a_i^j is the i-th activation function of the j-th layer. Each y represents each category of the dimension of openness, corresponding to categories 1 to 5, respectively. A BP neural network is used to estimate the value of each output of $\hat{y}$, i.e., the "likelihood" of each classification to which it belongs. The option with the greatest probability in the classification is selected as the final result and set as "1", and the other is set as "0". The final result of the example in Figure 5 indicates that the user has an openness score of 2.

3.2. Effectiveness Assessment Method

At The International AAAI Conference on Weblogs and Social Media (ICWSM) in 2013, researchers used a unified Facebook dataset in a workshop on personality trait recognition. Accuracy, recall, and F1 values were used to evaluate the experimental results. The dataset includes an OCEAN model of Facebook users, as well as user-posted text material. To ensure the uniformity of the test dataset, the experiment first downloaded the same dataset from the conference's website.

After obtaining the data required for the experiments, the random Weibo users were divided into two groups, A and B. A was used as a training set, accounting for 80% of the total dataset; B was used as a test set, accounting for 20% of the total dataset.

Uniform evaluation criteria and uniform datasets with the ICWSM meeting were used for comparison with the other methods, which can verify the improvement in effectiveness of this experiment. Specific evaluation indicators included accuracy, recall rate, and F1 value. The following details show how to calculate these evaluation indicators.

When the results need to be evaluated, there are four following possibilities, as shown in Table 6.

Table 6. Positive case and negative case definitions of items to be evaluated.

	Actual Positive	**Actual Negative**
Predicted positive	TP (True Positive)	FP (False Positive)
Predicted negative	FN (False Negative)	TN (True Negative)

Based on the definitions of the positive and negative examples in Table 5, the accuracy was defined as in Equation (3):

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Accuracy represents the proportion of the correctly predicted samples to the total predicted positive samples. In general, a higher accuracy means that the prediction was more accurate.

Recall was defined as Equation (4):

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

Recall represents the proportion of samples predicted to be accurate to all true positives. The higher the recall rate, the more accurate the estimate is.

Accuracy and recall do not increase or decrease simultaneously, so another indicator F1 value was introduced to balance the two. The value of $F1$ was defined as Equation (5):

$$F1 = \frac{2 \times Prcision \times Recall}{Percision + Recall} \tag{5}$$

Usually, larger $F1$ values indicate the prediction's higher accuracy.

Based on the multi-classification problem in this paper, when the data above were calculated, each category was then separately defined as "positive", and the other categories were defined as "negative" to calculate the accuracy, recall, and F1 values for each category.

4. Results

The results of this experiment were compared with the results of the ICWSM meeting. "SVM" is an SVM-based meta-learning method proposed by Verhoeven et al. [31] to identify user OCEAN models. "KNN + NB" is a method used by Farnadi et al. [32] to automatically identify the user's OCEAN model from the user's release status using the k-neighbor and naive Bayes methods. "LR + mNB" is a method of Alam et al. [33] that studies logistic regression and polynomial Bayesian methods to identify user OCEAN models. The following three tables compare the accuracy, recall, and F1 values for each method in each dimension of the OCEAN model, with the maximum value of each feature dimension shown in bold. The accuracy, recall, and F1 values for each method in each dimension of the OCEAN personality model are compared in Tables 7–9, where the maximum value for each feature dimension is shown in bold. The letters O, C, E, A, and N represent "Openness, Conscientiousness, Extraversion, Agreeable-ness, and Neuroticism", respectively.

Table 7. Comparison of accuracy.

Method	O	C	E	A	N	Average
Method of this paper	0.79	0.70	0.80	0.70	0.71	0.740
SVM	0.78	0.72	0.79	0.67	0.71	0.734
KNN, NB	0.60	0.55	0.58	0.50	0.54	0.554
LR, mNB	0.60	0.59	0.58	0.59	0.59	0.590

Table 8. Comparison of recall rates.

Method	O	C	E	A	N	Average
Method of this paper	0.76	0.57	0.79	0.72	0.69	0.704
SVM	0.77	0.72	0.79	0.68	0.72	0.736
KNN, NB	0.70	0.54	0.61	0.50	0.53	0.576
LR, mNB	0.60	0.59	0.58	0.59	0.58	0.588

Table 9. Comparison of F1 values.

Method	O	C	E	A	N	Average
Method of this paper	0.77	0.63	0.79	0.71	0.70	0.751
SVM	0.77	0.72	0.79	0.67	0.70	0.748
KNN, NB	0.61	0.54	0.56	0.50	0.52	0.546
LR, mNB	0.60	0.59	0.58	0.58	0.58	0.586

In comparing accuracy, the method in this article had the highest average accuracy of 74%, and four of the five single indicators had the highest accuracy.

In the comparison of recalls, the method in this article was relatively close to the SVM, exceeding 70%.

In the comparison of F1 values, the method in this article was closer to SVM and surpasses SVM on average, at 75.1%.

5. Discussion

Screening information that is useful for oneself amidst the massive amounts of information available on the Internet is a complex project, and the existence of redundant information has become an invisible "time killer". In order to make it easier for people to obtain information that interests them or is useful to them on the Internet, as well as to improve the efficiency and accuracy of searches, it is necessary to screen and recommend content according to a person's preferences and personality characteristics. The development of neural networks has made more accurate recommendations possible. However, the data entered into the neural network still need to be filtered and processed. Therefore, the issue of how to choose the data and which model tools to use to predict people's preference patterns and improve the efficiency and accuracy of predictions is the focus of current research.

If a user's personality characteristics can be accurately predicted through the user's public information on the network and used for a micro-directional push, this will effectively save the user's time and reduce the meaningless transmission of network information. In order to study this problem for the more popular short texts on the network, this paper for the first time used the LDA theme model to extract the user's text features, and the extracted features were used for input to the BPNN. In addition, the user OCEAN model results obtained by the questionnaire were output as samples of the BP neural network, the BPNN was trained, and a mapping model between the probability of the user text theme and the user OCEAN model was established to predict the latter.

On the basis of previous research, this paper has made some attempts at new methods:

(1) to improve the authenticity and accuracy of the original data, we defined the filter rules of the data;

(2) to extract users' text characteristics, we used the LDA topic model, which solves the problem of text information redundancy;

(3) to constructed LDA-BPNN comprehensive models to identify the user's OCEAN personality model that achieved a more accurate and stable recognition effect than the other methods at present.

However, this paper only does the work of predicting the user's OCEAN personality model through the user's textual characteristics, and further research is still needed to combine the results of this study with micro-directional propagation to make the propagation more accurate.

6. Conclusions

In this paper, the definition of the OCEAN personality model was first introduced, and its shortcomings as well as the prediction methods of several models were proposed. Through a process of acquiring user data and considering the characteristics of short texts, the LDA topic model was adopted to mine text topics instead of the traditional text analysis model.

This paper proposes and tests the prediction method of the user OCEAN personality model based on a BP neural network. Finally, the accuracy, recall rate, and F1 value of the method proposed in this paper were compared with those presented at the ICWSM conference.

It can be seen from the experimental results that, compared with some previous methods (SVM, KNN + NB, LR + mNB) for user OCEAN personality model recognition, the method proposed in this paper has made certain improvements in accuracy, recall rate, and evaluation of the F1 value. This proves that the text feature extraction method based on LDA is more suitable for Sina Weibo short texts than other text feature extraction methods and also indicates that the neural network algorithm has better learning effects than other algorithms in some cases.

This paper makes a preliminary exploration into the research of this model. In addition, further research can be produced in the future.

In the future, studies of user OCEAN personality model recognition, in addition to textual feature information, some non-textual feature information, such as pictures, videos, etc., may be added.

Author Contributions: Conceptualization, W.Z. and M.L.; methodology, B.Y. and L.Y.; software, Y.L. and Z.L.; validation, Z.L. and S.L.; formal analysis, Y.L. and L.Y.; investigation, B.Y.; resources, X.Q. and S.L.; data curation, X.Q, and Y.L.; writing—original draft preparation, X.Q., M.L. and L.Y.; writing—review and editing, X.Q., M.L., and W.Z.; visualization, Y.L. and L.Y.; supervision, B.Y.; project administration, W.Z.; funding acquisition, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Sichuan Science and Technology Program, grant number 2021YFQ0003. "The innovation and practice of translation teaching model under 'new liberal art + information technology' (No. 2021ND0605)" and "Reform and practice of applied talent cultivation mode of liberal arts majors in non-government universities under the background of new liberal arts (21BZ080)".

Data Availability Statement: Data available on request due to restrictions, e.g., privacy or ethical. The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the data are not being made publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Anggono, W.; Ni, X.; Yin, L.; Chen, X.; Liu, S.; Yang, B.; Zheng, W. Semantic representation for visual reasoning. *MATEC Web Conf.* **2019**, *277*, 02006. [CrossRef]
2. Durupinar, F.; Pelechano, N.; Allbeck, J.; Güdükbay, U.; Badler, N.I. How the Ocean Personality Model Affects the Perception of Crowds. *IEEE Comput. Graph. Appl.* **2011**, *31*, 22–31. [CrossRef] [PubMed]
3. Plaisant, O.; Guertault, J.; Courtois, R.; Reveillere, C.; Mendelsohn, G.A.; John, O.P. Big Five History: OCEAN of personality factors. Introduction of the French Big Five Inventory or BFI-Fr. *Ann. Med.-Psychol.* **2010**, *168*, 481–486. [CrossRef]
4. O'Keefe, D.F.; Kelloway, E.K.; Francis, R. Introducing the OCEAN.20: A 20-Item Five-Factor Personality Measure Based on the Trait Self-Descriptive Inventory. *Milit. Psychol.* **2012**, *24*, 433–460. [CrossRef]
5. Nassiri-Mofakham, F.; Nematbakhsh, M.A.; Ghasem-Aghaee, N.; Baraani-Dastjerdi, A. A heuristic personality-based bilateral multi-issue bargaining model in electronic commerce. *Int. J. Hum.-Comput. Stud.* **2009**, *67*, 1–35. [CrossRef]
6. Rauthmann, J.F.; Sherman, R.A.; Nave, C.S.; Funder, D.C. Personality-driven situation experience, contact, and construal: How people's personality traits predict characteristics of their situations in daily life. *J. Res. Pers.* **2015**, *55*, 98–111. [CrossRef]

7. Lopez-Pabon, F.O.; Orozco-Arroyave, J.R. Automatic Personality Evaluation from Transliterations of YouTube Vlogs Using Classical and State of the art Word Embeddings. *Ing. Investig.* **2022**, *42*, 13. [CrossRef]
8. Holman, D.J.; Hughes, D.J. Transactions between Big-5 personality traits and job characteristics across 20 years. *J. Occup. Organ. Psychol.* **2021**, *94*, 762–788. [CrossRef]
9. Rodriguez, P.; Velazquez, D.; Cucurull, G.; Gonfaus, J.M.; Roca, F.X.; Ozawa, S.; Gonzalez, J. Personality Trait Analysis in Social Networks Based on Weakly Supervised Learning of Shared Images. *Appl. Sci.* **2020**, *10*, 8170. [CrossRef]
10. Tadesse, M.M.; Lin, H.; Xu, B.; Yang, L. Personality Predictions Based on User Behavior on the Facebook Social Media Platform. *IEEE Access* **2018**, *6*, 61959–61969. [CrossRef]
11. Bowden-Green, T.; Hinds, J.; Joinson, A. How is extraversion related to social media use? A literature review. *Personal. Individ. Differ.* **2020**, *164*, 11. [CrossRef]
12. Gallo, F.R.; Simari, G.I.; Martinez, M.V.; Falappa, M.A. Predicting user reactions to Twitter feed content based on personality type and social cues. *Future Gener. Comput. Syst.* **2020**, *110*, 918–930. [CrossRef]
13. Hossain, M.A.; Quaddus, M.; Warren, M.; Akter, S.; Pappas, I. Are you a cyberbully on social media? Exploring the personality traits using a fuzzy-set configurational approach. *Int. J. Inf. Manag.* **2022**, *66*, 12. [CrossRef]
14. Han, J.; Pei, J.; Tong, H. *Data Mining: Concepts and Techniques*; Morgan kaufmann: Burlington, MA, USA, 2022.
15. Mairesse, F.; Walker, M.A.; Mehl, M.R.; Moore, R.K. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *J. Artif. Intell. Res.* **2007**, *30*, 457–500. [CrossRef]
16. Oberlander, J.; Nowson, S. Whose thumb is it anyway? Classifying author personality from weblog text. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, Australia, 17–18 July 2006; pp. 627–634.
17. Wang, Y.; Zheng, J.Z.; Li, Q.; Wang, C.L.; Zhang, H.Y.; Gong, J.B. XLNet-Caps: Personality Classification from Textual Posts. *Electronics* **2021**, *10*, 1360. [CrossRef]
18. Azucar, D.; Marengo, D.; Settanni, M. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personal. Individ. Differ.* **2018**, *124*, 150–159. [CrossRef]
19. Marouf, A.A.; Hasan, M.K.; Mahmud, H. Secret Life of Conjunctions: Correlation of Conjunction Words on Predicting Personality Traits from Social Media Using User-Generated Contents. In *Advances in Electrical and Computer Technologies*; Sengodan, T., Murugappan, M., Misra, S., Eds.; Springer: Singapore, 2020; pp. 513–525. [CrossRef]
20. Bai, S.; Zhu, T.; Cheng, L. Big-five personality prediction based on user behaviors at social network sites. *arXiv* **2012**, arXiv:1204.4809. [CrossRef]
21. Camacho, D.; Panizo-Lledot, Á.; Bello-Orgaz, G.; Gonzalez-Pardo, A.; Cambria, E. The four dimensions of social network analysis: An overview of research methods, applications, and software tools. *Inf. Fusion* **2020**, *63*, 88–120. [CrossRef]
22. Bhat, M.R.; Kundroo, M.A.; Tarray, T.A.; Agarwal, B. Deep LDA: A new way to topic model. *J. Inf. Optim. Sci.* **2019**, *41*, 823–834. [CrossRef]
23. Moore, K.; McElroy, J.C. The influence of personality on Facebook usage, wall postings, and regret. *Comput. Hum. Behav.* **2012**, *28*, 267–274. [CrossRef]
24. Mehrotra, R.; Sanner, S.; Buntine, W.; Xie, L. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 28 July–1 August 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 889–892.
25. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimed. Tools Appl.* **2018**, *78*, 15169–15211. [CrossRef]
26. Phan, X.H.; Nguyen, C.T. A Java Implementation of Latent Dirichlet Allocation (LDA) Using Gibbs Sampling for Parameter Estimation and Inference (JGibbLDA). Available online: https://sourceforge.net/projects/jgibblda/files/ (accessed on 11 July 2022).
27. Medcl. ik Chinese Word Segmentation. Available online: https://github.com/medcl/elasticsearch-analysis-ik/releases/tag/v6.0.0 (accessed on 11 July 2022).
28. Yang, B.; Liu, C.; Zheng, W.; Liu, S. Motion prediction via online instantaneous frequency estimation for vision-based beating heart tracking. *Inf. Fusion* **2017**, *35*, 58–67. [CrossRef]
29. Jiang, Y.; Zhang, F. Study on BP Neural Network Optimization by Improved Decay Parameter Genetic Algorithm. *J. Phys. Conf. Ser.* **2020**, *1621*, 012054. [CrossRef]
30. Başaran, S.; Ejimogu, O.H. A Neural Network Approach for Predicting Personality from Facebook Data. *SAGE Open* **2021**, *11*, 21582440211032156. [CrossRef]
31. Verhoeven, B.; Daelemans, W.; Smedt, T.D. Ensemble Methods for Personality Recognition. In Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM) in 2013, Cambridge, MA, USA, 8–11 July 2013; Association for the Advancement of Artificial Intelligence: Palo Alto, CA, USA, 2013.
32. Farnadi, G.; Zoghbi, S.; Moens, M.F.; De Cock, M. Recognising Personality Traits Using Facebook Status Updates. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM) in 2013, Cambridge, MA, USA, 8–11 July 2013*; Association for the Advancement of Artificial Intelligence: Palo Alto, CA, USA, 2013.
33. Alam, F.; Stepanov, E.A.; Riccardi, G. Personality Traits Recognition on Social Network—Facebook. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM) in 2013, Cambridge, MA, USA, 8–11 July 2013*; Association for the Advancement of Artificial Intelligence: Palo Alto, CA, USA, 2013.