Journal of Business and Economics, ISSN 2155-7950, USA
October 2023, Volume 14, No. 10, pp. 467-480
DOI: 10.15341/jbe(2155-7950)/10.14.2023/003

http://www.academicstar.us

Personality Traits Analysis of Instagram Social Media Community in Terms of Deep Learning Method

Shih Ya-Yueh[1], Chou Han-Ping[2], Li Che-Yu[3]

(1. Department of Management Information Systems, National Chiayi University, Taiwan;

2. Department of Management Information, Chung Hua University, Taiwan;

3. Department of Management Information Systems, National Chiayi University, Taiwan)

Abstract: Nowadays, social media has become the main medium for people to communicate with each other. People begin to care about the quality and effectiveness of making friends on social media. Knowing the personality of users can be used as a basis for judging new friends on social media. Therefore, how to predict the user's personality through the activity records on social media has become an important issue. Due to the long length of traditional personality scale questionnaires, it takes a long time to judge the personality traits of users. In order to improve the efficiency of analyzing the personality traits of users, the purpose of this study is to provide a method for predicting personality traits using user pictures and posts, so as to reduce the time to judge the user's personality. A total of 200 Instagram users' pictures and posts were collected, using the RGB average value of the picture, the percentage of LIWC post vocabulary, and the combination of the RGB average value of the picture and the percentage of vocabulary in the LIWC post as features, and established CNN, SVM, Decision tree and Personality trait prediction models such as Logistic Regression are used to predict the five major personality traits and compare their accuracy. The results of this study confirmed that using the CNN algorithm combined with the image RGB average and the LIWC post vocabulary percentage as features has a high accuracy rate (85%) in predicting personality traits. In terms of predicting individual personality traits, CNN algorithm has a higher accuracy rate of 87.5% for rigorous personality, agreeable personality and neurotic personality; SVM algorithm has higher accuracy rate for extraverted personality and agreeable personality. Both are 87.5%; Decision is better at predicting rigorous personality, with an accuracy rate of 87.5%. This result can be used as a reference for future research.

Key words: Instagram, Big five personality traits, convolutional neural networks, social media, personality trait prediction

JEL codes: O36

1. Introduction

1.1 Research Background and Motivation

With the vigorous development of information technology, the internet has become the primary medium for

Shih Ya-Yueh, Associate Professor, Department of Management Information Systems, National Chiayi University; research areas: data mining electronic commerce. E-mail: moon.shih@gmai.com.

interpersonal communication. According to Hayes' study in 2022, the majority of individuals engage in communication with others through at least one social media platform, and an increasing number of people rely on social media for socializing and making new acquaintances.

Ning and Dhelim et al. (2019) suggested that understanding users' personalities from social media can enhance the judgment features for friend recommendation mechanisms. Hence, a novel approach for detecting personality traits has emerged, utilizing information provided by social media to discern the personality traits of users. This involves employing the Linguistic Inquiry and Word Count (LIWC) tool to calculate the frequency of word categories used in user posts as features, coupled with conventional statistical methods, machine learning, and deep learning algorithms to analyze user personalities.

Howlader et al. (2018) employed the traditional statistical regression analysis to predict user personalities based on Facebook posts, achieving an accuracy rate of 67.12%. In contrast, Michael & Arokia (2018) utilized the Support Vector Machine (SVM) machine learning algorithm to predict user personalities from Facebook posts, attaining an accuracy rate of 71.32%. Many scholars have found that prediction accuracy using Convolutional Neural Network (CNN) exceeds that of other machine learning and deep learning algorithms, as demonstrated by Michael & Arokia (2018), Howlader et al. (2018), and Heci et al. (2020). In these studies, the use of CNN in personality trait prediction consistently yielded the highest accuracy, and in the prediction of extraversion, CNN's accuracy surpassed other algorithms by a considerable margin.

In addition to textual features, some scholars have incorporated images posted by users on social media as features, employing machine learning algorithms and traditional statistical methods to predict user personality traits. For instance, Khorrami & Farhangi (2022) used RGB (Red, Green, Blue) and HSV (Hue, Saturation, Value) mean values as features and applied Random Forest to predict the Big Five personality traits of Instagram users, conducting cross-validation. The study found that RGB values were crucial features in constructing the prediction model.

In summary, there is currently a scarcity of research on personality traits of Instagram users. Therefore, this study will collect posts and images from Instagram users, use LIWC to calculate the percentage of various word categories in the posts, and utilize RGB mean values from images as predictive features. The Chinese version of the Big Five Personality Inventory developed by Deng Jingyi et al. (2011) will be employed to ascertain user personalities. Finally, the study will employ Convolutional Neural Networks (CNN) in deep learning to create a predictive model for the Big Five personality traits. This model will be compared with different machine learning algorithms and traditional statistical methods to identify the most accurate approach in predicting personality traits. This aims to address the issue of traditional questionnaire-based user personality assessment being time-consuming, and consequently enhance its efficiency.

1.2 Research Objectives

As mentioned in the previous section, with the changing habits of knowledge sharing and communication, social media platforms have gained increasing attention from the public. More and more internet users are inclined to upload posts to social media during their leisure time, either to share personal anecdotes or professional expertise. This leads to the accumulation of personalized linguistic features and writing habits in Instagram posts. These characteristics and habits also enable scholars to predict the personalities and traits of users. Given the abundance of posts on social platforms, each poster possesses unique personality traits. Due to the scarcity of studies predicting personality traits based on Instagram posts in the past, this research aims to analyze

and forecast the personality traits of posters on Instagram. Different machine learning algorithms will be utilized for analysis, and their respective accuracy in predicting various personality traits will be compared. Through images and posts, this study seeks to discern the personalities of the posters. The research objectives are summarized as follows:

(1) Utilize the RGB mean values of images from Instagram users as features, and employ machine learning algorithms to predict users' personalities.

(2) Utilize LIWC to calculate the percentage of word categories in users' posts as features, and employ machine learning algorithms to predict users' personalities.

(3) Combine the RGB mean values of images and LIWC post word percentages from Instagram users as features, and employ machine learning algorithms to predict users' personalities.

(4) Compare models established using different features and various algorithms to identify the one with higher accuracy rates.

1.3 Research Contributions

This research encompasses both academic and practical contributions. The academic contributions are outlined as follows:

(1) Providing insights into predicting user personalities based on Instagram posts and images, which can be applied in future research related to Instagram and personality traits.

(2) Utilizing LIWC (Linguistic Inquiry and Word Count) to calculate the percentage of various word categories within posts, and extracting RGB (Red, Green, Blue) mean values from images as features for predicting personality traits of Instagram users. This offers a new research direction in the realm of characterizing user personalities in the context of social media.

(3) Comparing the use of users' posts and images as features with both Convolutional Neural Network (CNN) and multiple machine learning algorithms to identify methods with higher accuracy. This optimization of the process for discerning user personalities aims to enhance the research workflow.

The practical research contributions are as follows:

(1) Employing machine learning and deep learning algorithms to predict personality traits of social media users, and subsequently comparing them to elevate the accuracy of personality prediction. This can serve as a basis for decision-making in adding new friends, enabling users to more accurately find suitable companions and individuals for interaction.

(2) If personality traits can be ascertained with greater precision through users' social media posts, social media platforms can provide customized services and advertisements solely based on posts. Additionally, this knowledge can be integrated into personalized recommendation mechanisms.

2. Related Works

2.1 Personality Traits

In 1943, Raymond Cattell defined personality traits as "thoughts and behaviors that can be predicted in specific situations". In 1967, Eysenck proposed a new personality model based on three major axes: extraversion, neuroticism, and psychoticism. It wasn't until the period between 1980 and 1989 that scholars reached a consensus and introduced the "Big Five" model of personality traits. The Big Five personality trait theory has gained credibility due to multiple validations, leading many scholars to utilize it in personality-related studies

(Nguyen, 2017).

2.1.1 The Big Five Personality Traits

The Big Five is a vector based on five traits used to describe personality, with each trait representing two extremes on a spectrum. To date, the Big Five model has become one of the most widely used personality models in research (Matthews, 2003). These different personality models provide varying perspectives on describing personality traits, offering researchers tools to analyze and assess individual personalities.

The descriptions of the Big Five traits are as follows (Majumder et al., 2017):

(1) Extraversion (EXT): Receptive to external stimuli and energized by social interactions.

(2) Neuroticism (NEU): Signifying emotional instability and a tendency to experience negative emotions.

(3) Agreeableness (AGR): An optimistic and trusting mindset.

(4) Conscientiousness (CON): Organized and committed to fulfilling one's responsibilities.

(5) Openness (OPN): Artistic and possessing an open-minded approach to thinking.

2.1.2 Studies on Personality Traits

Due to its repeated validations, the Big Five personality trait theory has gained a certain level of credibility, leading many scholars to employ the Big Five in personality-related studies. For example:

Eijck & Borgsteede (2005) conducted a survey using a shortened version of the Big Five personality questionnaire, investigating the influence of different personality traits on media preferences and cultural engagement. Nguyen (2017) collected user personality traits based on the Big Five personality questionnaire and utilized them for personalized recommendations. The results demonstrated improved user satisfaction through personality-based recommendations.

Marshall & Lefringhausen (2015) researched which personality types of Facebook users tend to update their posts more frequently. They confirmed subjects' personalities using the Big Five personality inventory. The results indicated that users with higher levels of openness tend to post more knowledge-intensive content, while those with higher self-esteem tend to post content related to children, with narcissistic individuals having the highest posting frequency.

2.2 LIWC (Linguistic Inquiry and Word Count)

2.2.1 Development of LIWC

Pennebaker et al. (1999) developed the Linguistic Inquiry and Word Count (LIWC) system for text analysis of commonly written text data. After calculating word frequencies through programming, the words are categorized, followed by the compilation of a word dictionary. This led to the creation of the Linguistic Inquiry and Word Count Dictionary, abbreviated as the LIWC dictionary.

As the number of words and categories expanded, Pennebaker et al. further developed extended versions of LIWC such as LIWC 2012, LIWC 2015, and the latest LIWC 2022 version. It is currently used to analyze various word categories, and has been widely utilized in analyzing personality traits of social media users with significant reliability and validity (Tandera et al., 2017; Yuan et al., 2018; Howlader et al., 2018; Salsabila Dwi & Setiawan, 2021).

2.2.2 Studies Related to LIWC (Linguistic Inquiry and Word Count)

Since the development of LIWC, a significant body of related research has accumulated. Language, akin to an individual's cognition, emotions, and attitudes, exhibits notable individual differences. Therefore, it is frequently employed in the quantitative analysis of social media text and user feature analysis. Relevant studies

include:

Marengo et al. (2021) collected Facebook users' posts and used LIWC's lexical categories as features to predict users' quality of life. The study found that people's discourse on social media reflects their own psychological condition.

Luhmann (2017) calculated emotional words in Facebook users' posts using LIWC. The results consistently showed that posts on social media contribute to the study and assessment of users' emotional life. The models built also demonstrated moderate to high predictive ability for predicting users' quality of life on Facebook.

2.3 Machine Learning

2.3.1 Definition of Machine Learning

Machine Learning, a branch of Artificial Intelligence, involves the design and analysis of algorithms that enable computers to automatically "learn" from data and make predictions on unknown information (Bishop, 1995). It can be categorized into supervised learning and unsupervised learning. In supervised learning, a function is learned from a given training dataset, allowing predictions to be made for new data. In contrast, unsupervised learning lacks pre-labeled results in the training set (Mitchel, 1997).

2.3.2 Definition of Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is one of the most widely applied architectures in deep neural networks. Proposed by scholars including LeCun in 1989, it gained significant recognition for breakthroughs in image classification and object detection in computer vision. CNN is also one of the few architectures in deep learning that draws inspiration from the human visual system. In 2014, KIM proposed that CNN can also be applied to textual data analysis using one-dimensional convolutional neural networks.

2.3.3 Research in Machine Learning

In previous studies utilizing social media posts as features to analyze user personality, many experiments have employed CNN and other algorithms for personality prediction. These studies demonstrate that CNN excels in predicting personality traits based on post features compared to other algorithms. For example:

Tandera et al. (2017) and Heci et al. (2020) predicted Facebook user personalities using Naive Bayes and CNN algorithms, respectively. Their results showed that CNN achieved higher accuracy in personality prediction, with percentages of 74.17% and 77.3%.

Yuan et al. (2018), Michael & Arokia (2018), and Yuan et al. (2021) utilized the Big Five personality model, using LIWC to extract lexical features. They applied CNN and SVM algorithms to predict user personality traits. The results indicated that CNN outperforms other algorithms in personality prediction, with accuracies of 74.12%, 75.61%, and 78.69% respectively.

Howlader et al. (2018) used Logistic Regression and CNN in their research to predict Facebook users' personality traits, comparing their accuracies. The results demonstrated that CNN achieved a prediction accuracy of 81.26%, surpassing the traditional statistical methods.

In addition to text features, some scholars have used images posted by users on social media as features, employing machine learning algorithms and traditional statistical methods to predict user personality traits. For instance:

Khorrami & Farhangi (2022) used the average values of RGB (Red, Green, Blue) and HSV (Hue, Saturation, Value) as features, applying a random forest algorithm to predict the Big Five personality traits of Instagram users. They conducted cross-validation and found that image features played a crucial role in building prediction models.

All models used demonstrated good predictive performance and generality, with each personality trait corresponding to specific color preferences and image posting styles. Among them, RGB values and HSV color space (Hue, Saturation, Value) were the most important features in modeling.

Furthermore, Kanchana & Zoraida (2020) collected images uploaded by Facebook users, extracting the RGB average values of each image and the tone of the images (cool tone, warm tone, dark tone, bright tone) as features. They used various models (including Extra Tree Regressor, Linear Regression, Elastic Net, and Multioutput Gradient Boosting Regression) to predict the Big Five personality traits of users. The study also confirmed that image color is one of the important features for predicting personality traits. These research findings indicate that, in addition to text features, extracting color features from images posted on social media is significant for predicting user personality traits. These methods expand the scope of personality prediction, providing a more comprehensive and diverse set of features for understanding the personality traits of social media users.

The research process in this study is divided into four steps, as illustrated in Figure 1. The first step involves administering the Big Five Personality Inventory to users through a questionnaire and collecting participants' Instagram posts and images. The second step encompasses data preprocessing, which includes the removal of invalid questionnaires. In the third step, machine learning algorithms from Python packages are employed to construct models for predicting the Big Five personality traits. The fourth step focuses on comparing the predictive accuracy of various algorithmic models established.

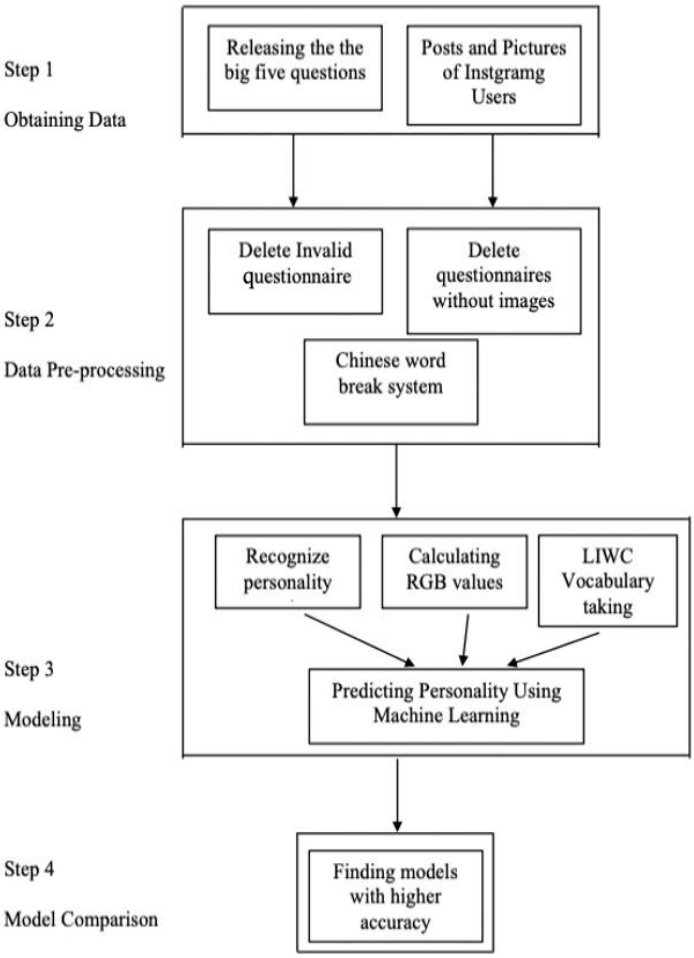

Figure 1 Flowchart of the Study

3. Research Methodology

3.1 Data Acquisition

The data used in this study comprises two main components: scores from the "Big Five Personality Inventory" questionnaire and the posts/images shared by participants on the Instagram social media platform. These data were uniformly collected through Surveycake surveys distributed to users.

The first section of the Surveycake questionnaire includes the "Big Five Personality Inventory" questions. These questions, translated into Chinese by local scholars led by Deng Jingyi in 2011 based on Goldberg's original 1990 inventory, consist of a total of 40 items. These items cover the five major personality traits: extraversion, openness, conscientiousness, neuroticism, and agreeableness. Responses are recorded on a 5-point scale, ranging from 1 (very inaccurate) to 5 (very accurate), including reverse-coded items. Users assign corresponding scores based on their answers to each personality trait item, with the highest score indicating the dominant personality trait.

3.2 Data Preprocessing

The collected questionnaire data was categorized into three parts: "User Posts", "User Images", and the "Big Five Personality Traits Assessment Questions". The preprocessing steps for each part are described as follows:

(1) User Posts: Questionnaires lacking user posts were considered invalid and removed. The five user posts from each valid questionnaire were compiled into a single document, with irrelevant symbols and characters removed.

(2) User Images: Questionnaires lacking image submissions were excluded. The five images provided by each user were grouped into a folder for the calculation of average RGB values.

(3) Big Five Personality Traits Assessment Questions: Answers to the assessment questions were converted from "strongly disagree", "disagree", "neutral", "agree", to "strongly agree", to numerical values ranging from 1 to 5. Reverse-coded questions were handled accordingly.

Before model construction and analysis, invalid questionnaires were removed. Unlike English texts, Chinese text lacks natural word boundaries. Therefore, a Chinese word segmentation system developed by the Academia Sinica was applied to segment each Instagram post and remove irrelevant punctuation marks.

3.3 Model Construction

Upon aggregating and categorizing the validated items from the "Chinese Version of the Big Five Personality Inventory" questionnaire, personality traits were obtained. These traits, along with data derived from the LIWC-extracted vocabulary of user posts, were utilized as features. Using Python programming language packages, algorithms such as CNN, SVM, Logistic Regression, and Decision Tree were employed to predict the personalities of Instagram users.

3.3.1 Confirming User Personality through Weighted Scores

Following the validation of the items in the "Chinese Version of the Big Five Personality Inventory" questionnaire, individual scores for each personality trait were aggregated. The highest score was examined to determine the user's dominant personality trait, serving as the basis for experimental accuracy assessment.

In this study, the "Mini-Markers" proposed by Saucier (1994) were adapted. Together with the updated Chinese translations of word measurements by Teng et al. (2011), a set of personality trait questions for Instagram

users was formulated. Measurement was conducted using a Likert-type five-point scale, where 1 indicated "strongly disagree" and 5 indicated "strongly agree".

3.3.2 Calculating the Average RGB Values of Images

RGB, which stands for Red, Green, and Blue, represents the proportional values of the three primary colors in the RGB color model. This study collected images from Instagram users, grouping them in sets of five for analysis. Python was employed to compute the average RGB values of each set of images, serving as features for predicting user personality traits.

3.3.3 Extracting LIWC Word Percentages

The current Chinese version of LIWC 2022 includes 84 defined word categories, such as WC (total word count), pronoun (pronouns), affect (emotion-related words), and Sad (sadness-related words), among others. In this study, the LIWC program was used to extract post content from Instagram users, utilizing these word categories as variables for predicting the Big Five personality traits.

The collected user posts underwent LIWC (Linguistic Inquiry and Word Count) analysis to calculate the percentages of various word categories. The process involved opening pre-processed documents, sequentially comparing each word with the dictionary. The LIWC program then computed the percentage of words falling into each category relative to the total word count, and provided the output.

3.3.4 Establishing the Algorithmic Prediction Model

In this study, a convolutional neural network (CNN) model was developed using Python libraries including TensorFlow, Keras, NumPy, and Pandas. The research process is outlined as follows:

(1) The percentages of various word categories derived from LIWC analysis of Instagram user posts were compiled into documents.

(2) For each user, the percentages of word categories from the posts and the average RGB values of images were inputted. The input nodes consisted of the percentages of word categories from the posts, while the output nodes represented the user's Big Five personality traits.

(3) TensorFlow was utilized to create functions for convolutional layers and pooling layers, followed by the training of the convolutional neural network.

(4) The Keras open-source neural network library was loaded to support data operations for the convolutional neural network.

(5) Through ParameterGrid, preliminary testing of parameters such as epoch, kernel size, and strides was conducted to establish a model with favorable accuracy. Finally, the disparities between actual values and predicted results were examined to determine the predictive accuracy of the model.

3.4 Model Comparison

This study employed both traditional statistical methods such as Logistic Regression, and machine learning and deep learning algorithms including CNN, SVM, and Decision Tree. Models were established using three different sets of features: image RGB values, LIWC word category percentages, and a combination of image RGB values with LIWC word category percentages. The aim was to identify the method with the highest predictive accuracy.

Previous research on personality trait prediction, as well as literature comparing multiple models, often utilized confusion matrices as evaluation criteria. This includes studies by Michael & Arokia (2018), Howlader et al. (2018), Jun et al. (2021), and others. Evaluation metrics such as Accuracy, Precision, and Recall were also

employed. In this study, the assessment of models will be based on confusion matrices, and calculations will be performed for Accuracy, Precision, and Recall. Accuracy, as the primary indicator for assessing model accuracy, will be utilized. By comparing the accuracy of different models, the study aims to evaluate their differences in overall predictive accuracy. Additionally, Precision and Recall will be employed to assess model performance.

4. Analysis and Discussion

The subjects of this study were users who possessed individual Instagram accounts, and data collection was conducted through an online survey. The survey was distributed on social media platforms and communities on February 25, 2023, employing convenience sampling to obtain responses from users with diverse experiences and age groups. After a three-week period of survey distribution, a total of 20 surveys were collected. After excluding 7 invalid responses, the final valid survey count was 200, resulting in an effective response rate of 96.7%.

4.1 Descriptive Statistics

The basic structure of the collected sample in this study is presented in Table 1. Among the sample, females comprised a higher percentage than males (Males: 44%, Females: 56%). The majority of respondents fell within the age range of 21 to 30 years. Regarding educational attainment, a significant portion held either undergraduate or graduate degrees. Instagram was the most widely used social media platform (67%), with a predominant frequency of 1-3 posts per week (45%). As for personality traits, each of the Big Five personality traits constituted approximately 20% of the sample.

4.1.1 Reliability Analysis

The purpose of reliability analysis is to examine the consistency and stability of measurement results, thus assessing the reliability of the measuring instrument. In this study, SmartPLS 3.0 was employed for analysis. Composite reliability values greater than 0.6 were considered acceptable according to the standard set by Fornell and Larcker (1981). However, some scholars argue that values should exceed 0.7 to be deemed rigorous (Hair et al., 1998). Additionally, Nunnally (1978) suggested that higher Cronbach's α coefficient values indicate greater reliability and stability of the scale. Generally, a Cronbach's α coefficient value should be at least greater than 0.5. In practice, a commonly accepted standard is a value above 0.7.

In this study, the composite reliability values and Cronbach's α coefficient values for all constructs exceeded 0.7. Thus, the valid samples collected in this study exhibit sufficient reliability, instilling confidence in the reliability of the research results. In summary, through factor analysis and reliability analysis, the Big Five personality traits scale in this study demonstrates a robust factor structure and reliability, providing strong support for ensuring the credibility and reliability of the data.

4.1.2 Validity Analysis

Validity refers to the accuracy and truthfulness of measurement results. Good validity is essential for accurately reflecting the characteristics of a construct. Validity primarily consists of two aspects: convergent validity and discriminant validity. Convergent validity is used to measure whether all items within the same construct are highly correlated with the construct. It is assessed using the following two criteria: (1) the factor loadings of all indicator variables must be greater than 0.7; the higher the factor loading, the stronger the correlation between the indicator variable and the construct. (2) The Average Variance Extracted (AVE) for all constructs must be greater than 0.5. In this study, the AVE for all constructs exceeded 0.5. Subsequently, using PLS for validity testing, the factor loadings for each construct were all greater than 0.7. Therefore, the valid

samples collected in this study demonstrate convergent validity.

4.2 Parameter Adjustment

The CNN (Convolutional Neural Network) model established in this study is a one-dimensional neural network built using TensorFlow. The data features include RGB values and LIWC word percentages, both of which are numerical. Through neural network computations, the relationships are identified to predict the Big Five personality traits. To achieve better accuracy in the model and prevent overfitting, the parameters of the CNN (Convolutional Neural Network), specifically the convolutional and pooling layers, are adjusted for optimization to establish a personality trait prediction model.

4.2.1 Data Ratios

In this study, a total of 200 data points were collected and divided into training and testing datasets in three different ratios: 80:20, 90:10, and 70:30. The kernel size and strides parameters were initially set to 1. The dataset sizes for the three ratios were 40, 20, and 60, as shown in Table 4-5. When establishing the CNN model with an 80:20 data ratio, the accuracy was higher (77.5%). For the 90:10 and 70:30 data ratios, the accuracy was 75% and 76.66%, respectively. Precision and recall for all three data ratios exceeded 80%. Based on the above, this study will use an 80:20 data ratio to establish the CNN personality trait prediction model.

4.2.2 Epoch

In this study, the model utilized the Early Stopping technique from the TensorFlow Database to find the optimal number of epochs. During the training process, it monitored both the pre-loss function and accuracy. A lower loss function value indicates that the model's predictions are closer to the true values, demonstrating better predictive capability. Early stopping recorded the results of each training cycle. If accuracy does not improve and the loss function does not decrease over several consecutive training cycles, early stopping terminates the training to prevent overfitting.

In this study, Early Stopping was set and executed with training stopping at 50 epochs, resulting in an accuracy of 85% and a loss function of 2.74%. The model's accuracy plateaued, and the loss function no longer decreased after this number of training cycles.

4.2.3 Kernel Size and Strides

This study utilized ParameterGrid to search for the parameters of kernel size and strides in the CNN convolutional neural network. As mentioned in Section 3.3.4, the kernel size refers to the size of the convolutional kernel used in the convolutional layer (Goodfellow et al., 2016), while strides represent the shift step size of the convolutional kernel when applied to input features (Tseng, 2021). This study tested kernel sizes and strides ranging from 1 to 10, resulting in a total of 100 combinations. Among these combinations, adjusting the kernel size to 3 and strides to 10 yielded the highest accuracy (85%).

4.3 Establishment of the CNN Prediction Model

After parameter adjustments and testing, the CNN model employed an 8:2 data cross-validation method, setting the parameters of kernel size and strides to 3 and 10, respectively. The Early Stopping technique was utilized to automatically determine the optimal epoch. Subsequently, CNN personality trait prediction models were established using RGB values, LIWC word percentages, and a combination of both, as depicted in Table 1. The training dataset comprised 160 samples, while the testing dataset comprised 40 samples. For each of the Big Five personality traits — Openness, Extraversion, Conscientiousness, Agreeableness, and Neuroticism — there were 8 samples each. Notably, utilizing both RGB values and LIWC word percentages as features resulted in a

higher average accuracy (86%).

Table 1 Comparison of CNN Model Accuracy Rates

	RGB value	LIWC vocabulary	RGB value and LIWC vocabulary
Average Accuracy	70%	80%	85%
Openness	70%	80%	87.5%
Extraversion	62.5%	75%	75%
Conscientiousness	75%	87.5%	87.5%
Agreeableness	62.5%	75%	87.5%
Neuroticism	75%	75%	87.5%

4.4 Comparison of Accuracy with SVM, Decision Tree, and Logistic Regression

Using the Python programming language and the scikit-learn package, SVM, Decision Tree, and Logistic Regression models were constructed. Since it was observed from the CNN model that combining RGB values with LIWC word percentages as features resulted in higher accuracy, the same features were utilized for SVM, Decision Tree, and Logistic Regression. The training dataset consisted of 160 samples, while the testing dataset comprised 40 samples. Each of the Big Five personality traits — Openness, Extraversion, Conscientiousness, Agreeableness, and Neuroticism — had 8 samples.

The accuracy of the CNN Convolutional Neural Network was the highest among all models (85%), outperforming SVM, Decision Tree, and Logistic Regression. In terms of individual personality trait prediction accuracy, the CNN excelled over SVM in Openness, Conscientiousness, and Neuroticism. However, the accuracy for Extraversion was slightly lower in CNN compared to SVM. When compared to Decision Tree, CNN exhibited higher accuracy in predicting Openness, Agreeableness, and Neuroticism, while accuracy for Extraversion and Conscientiousness was equivalent at 75% and 87.5%, respectively. In comparison to Logistic Regression, except for Extraversion, where accuracy remained at 75%, CNN demonstrated higher accuracy in predicting Openness, Conscientiousness, Agreeableness, and Neuroticism personality traits.

Table 2 Comparison of Model Accuracy Rates of Algorithms

	CNN	SVM	Decision tree	Logistic Regression
Average Accuracy	85%	80%	72.5%	70%
Openness	87.5%	75%	62.5%	62.5%
Extraversion	75%	87.5%	75%	75%
Conscientiousness	87.5%	75%	87.5%	62.5%
Agreeableness	87.5%	87.5%	75%	75%
Neuroticism	87.5%	75%	62.5%	75%

5. Conclusion

5.1 Research Conclusion

Social media has become an indispensable part of the modern digital era, with a growing user base leading to an unprecedented volume of information. Among the various forms of content, posts stand out as one of the most common ways for users to share information about their lives, emotions, interests, and more. However, posts are not just a means of communication; they can also serve as vital indicators for understanding user personalities.

Therefore, the exploration of methods to understand user personalities through social media posts and the application of machine learning algorithms have become a crucial area of research.

This study utilized the scikit-learn package in the Python programming language to construct comparative models. In the process of establishing the Convolutional Neural Network (CNN) model, different data ratios (9:1, 8:2, and 7:3) were initially employed for testing to evaluate their accuracy. Among these, the 8:2 data ratio yielded superior results. Subsequently, the ParameterGrid package was used to search for optimal convolutional kernel sizes and stride lengths for the model. The Early stopping method was employed to automatically determine the best accuracy while avoiding overfitting.

Before comparing with traditional statistical algorithms and deep learning algorithms, the accuracy of the CNN personality prediction model was evaluated using RGB values of images, LIWC word percentages, and the combination of both as features. The results showed that the model utilizing the combination of RGB values and LIWC word percentages as features performed better in predicting all five major personality traits, achieving accuracies of 70%, 80%, and 85%, respectively. In comparison to models utilizing only LIWC word percentages, the addition of RGB values led to improved prediction accuracy.

When compared to models using only LIWC word percentages, models with the combination of RGB values and LIWC word percentages as features exhibited higher prediction accuracy for Openness, Agreeableness, and Neuroticism personality traits. These results emphasize the importance of combining multiple features in enhancing prediction model capabilities, especially in predicting personality traits. Therefore, considering different feature combinations may lead to better performance when constructing prediction models.

This study established a personality prediction model using the combination of RGB values and LIWC word percentages as features. Different machine learning algorithms, including CNN, SVM, Decision Tree, and Logistic Regression, were employed. In experimental results, the CNN algorithm demonstrated the highest average accuracy of 85%. For the individual prediction of Openness and Neuroticism personality traits, the CNN algorithm outperformed other algorithms. Specifically, in predicting Openness personality, the CNN algorithm achieved the highest accuracy of 87.5% among all algorithms. For predicting Extraversion personality, the SVM algorithm performed the best with an accuracy of 87.5%. In predicting Conscientiousness personality, the Decision Tree algorithm exhibited the highest accuracy, also at 87.5%. For Agreeableness personality traits, both CNN and SVM algorithms achieved the highest prediction accuracy of 87.5%. Similarly, in predicting Neuroticism personality, the CNN algorithm attained the highest accuracy of 87.5%, which was the highest. Therefore, the results indicate that the CNN algorithm performed well in predicting Openness, Conscientiousness, Agreeableness, and Neuroticism personality traits. However, for predicting Extraversion personality, the accuracy was slightly lower than the overall average (75%). The SVM algorithm demonstrated better performance in predicting Extraversion and Agreeableness personality traits. The Decision Tree algorithm showed higher accuracy in predicting Conscientiousness personality.

In summary, the results of this study demonstrate that using the combination of RGB values and LIWC word percentages as features can enhance the prediction accuracy of user personality traits on social media. Furthermore, compared to traditional machine learning algorithms and statistical methods, deep learning algorithms exhibit superior performance in predicting user personality traits. It is worth noting that previous studies have less frequently employed RGB values of images as features, suggesting that the inclusion of relevant image features in the future may further enhance prediction accuracy.

Overall, the information provided by images and posts on social media holds significant value. Through

machine learning algorithms, it is possible to understand user personality traits based on these posts. This study is significant in improving the accuracy of personality trait prediction. Future research could further explore and integrate additional features and algorithms to enhance the accuracy of prediction models. Moreover, these research findings can be applied in practical fields such as social media analysis and personalized recommendations, thereby offering users more precise services and experiences.

5.2 Future Research Directions and Limitations

5.2.1 Future Research Directions

Subsequent research endeavors in this study can incorporate a broader range of diverse visual and textual features. For instance, as highlighted by Kanchana and Zoraida (2020), aspects like the color tones of images and filters applied by users can be considered. Additionally, Huang and Chen (2019) not only utilized image and text features but also integrated behavioral characteristics of social media users, such as likes on posts and the frequency of posting. In future studies, integrating these features and exploring methods to enhance prediction accuracy can be a promising avenue.

Given that this study solely compared the CNN, SVM, Decision Tree, and Logistic Regression algorithms, there exists a multitude of other algorithms suitable for constructing personality trait prediction models. Therefore, future research could involve the utilization of alternative algorithms for model development and subsequent comparison to identify algorithms with even higher prediction accuracy.

5.2.2 Limitations

In terms of the research data, this study collected data from a total of 200 users, predominantly within the age range of 21 to 30 years old, with educational backgrounds primarily consisting of university or postgraduate levels. The data pool may lack representation from other age groups and educational backgrounds, potentially restricting the generalizability of the study's findings. Future efforts could involve extending the data collection period and increasing the sample size to enhance diversity. This could be achieved by collecting data from users spanning a wider range of age groups, educational levels, and cultural backgrounds, ultimately leading to the establishment of a more comprehensive personality trait prediction model. Furthermore, the inclusion of various social media platforms for investigation could be considered. This would enable a comparative analysis of user posts and their associated personality traits across different platforms. Such an approach could further facilitate the exploration of additional effective methods and technologies aimed at refining research outcomes.

References

Bishop C. M. (1995). *Neural Networks For Pattern Recognition*, Oxford University Press.

Cattell R. B. (1943). "The description of personality: Basic traits resolved into clusters", *The Journal of Abnormal and Social Psychology*, Vol. 38, No. 4, pp. 476-481.

Davide Marengo, Cornelia Sindermann, Jon D. Elhai and Christian Montag (2020). "One social media company to rule them all: Associations between use of facebook-owned social media platforms, sociodemographic characteristics, and the big five personality traits", in: *Online Psychology Beyond Addiction and Gaming: A Global Look at Mental Health and Internet-Related Technologies*, *Frontiers in Psychology*, pp. 8-16.

Eijck I. A. J. M. and Borgsteede F. H. M. (2005). "A survey of gastrointestinal pig parasites on free-range, organic and conventional pig farms in The Netherlands", *Veterinary Research Communications*, Vol. 29, No. 5, pp. 407-414.

Howlader P., Pal K. K., Cuzzocrea A. and Kumar S. M. (2018). "Predicting facebook-users' personality based on status and linguistic features via flexible regression analysis techniques", in: *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pp. 339-345.

Heci Peng L., Changhui J., Pengzheng L., Shenke W. and Kejia Z. (2020). "Personality classification based on bert model", in: *2021 IEEE International Conference on Emergency Science and Information Technology*, pp. 150-152.

Khorrami M., Khorrami M. and Farhangi F. (2022). "Evaluation of tree-based ensemble algorithms for predicting the big five personality traits based on social media photos: Evidence from an Iranian sample", *Personality and Individual Differences*, Vol. 188, p. 111479.

Kim Y., Chiu Y. I., Hanaki K., Hegde D. and Petrov S. (2014). "Temporal analysis of language through neural language models", *International Journal of Psychophysiology*, Vol. 65, No. 3, pp. 201-213.

Luhmann M. (2017). "Using big data to study subjective well-being", *Current Opinion in Behavioral Sciences*, Vol. 18, pp. 28-33.

LeCun Y. (1989). "Generalization and network design strategies", *Connectionism in Perspective*, Vol. 19, No. 143-155, pp. 18-24.

Michael Mahesh K. and Arokia Renjit J. (2018). "Evolutionary intelligence for brain tumor recognition from MRI images: A critical study and review", *Evolutionary Intelligence*, Vol. 11, No. 1, pp. 19-30.

Majumder N., Poria S., Gelbukh A. and Cambria E. (2017). "Deep learning-based document modeling for personality detection from text", *IEEE Intelligent Systems*, Vol. 32, No. 2, pp. 74-79.

Mitchell T. M. (1997). *Machine Learning*, Vol. 1, No. 9.

Ning H., Dhelim S. and Aung N. (2019). "PersoNet: Friend recommendation system based on big-five personality traits and hybrid filtering", *IEEE Transactions on Computational Social Systems*, Vol. 6, No. 3, pp. 394-402.

Pennebaker J. W. and King L. A. (1999). "Linguistic styles: language use as an individual difference", *Journal of Personality and Social Psychology*, Vol. 77, No. 6, pp. 129-135.