IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Literature Review of Personality Recognition through an Integrated Deep Learning Approach utilizing Convolution Neural Network (CNN) and Recurrent Neural Network (RNN)

Jeel Patel[1], Jayesh M. Mevada[2] , Govind V. Patel[3], Mehul S. Patel[4] , Ankur J. Goswami[5]
[1]Student, [2]Assistant Prof., [3]Assistant Prof., [4]Assistant Prof., [5]Assistant Prof.
[1]Computer Engineering Department
[1] Sankalchand Patel College of Engineering, SPU Visnagar, India

Abstract: Personality prediction is a large area of research. Predicting personality using data from social media is a promising method because it doesn't require users to fill out surveys, which saves time and increases accuracy. Predicting personality has many real-world uses. As social media use continues to grow, a huge amount of text and images are shared online every day. Since most people share a lot of personal information, both knowingly and unknowingly, in their posts, it's possible to figure out personality traits from these texts. Finding individual personality traits from texts opens up many opportunities for different areas, like the forensic field, mental health assessments, and more. Deep learning algorithms are quite effective for text-based personality detection. In this study, we focus on a mixed deep learning approach, which combines different types of neural networks and machine learning techniques.

Keywords: Personality traits, machine learning, deep learning, personality recognition, Big-Five, Text analysis

1. INTRODUCTION

Social media is a place where people show themselves to the world. Social media accounts are private and personal, so they can reflect someone's personal life. Activities on social media, like posting, commenting, and updating statuses, can share personal information. The text users write can be studied to learn more about them, such as their personality.

Personality also influences how people interact with others and their environment. Personality can be used to evaluate things like hiring employees, giving career advice, relationship counseling, and health counseling. To understand their personality, a person usually needs to take different tests. These personality tests can be self-reports, interviews, or observations done by psychologists. However, these traditional methods are expensive and not very practical. [2] focused on identifying users' personality traits, and personality recognition has gained a lot of attention in recent years. First, personality traits, which are clear inner characteristics, are closely connected to the topics users like to talk about. This makes it possible to identify the personality traits of social media users based on content they create (UGC). Second, with the rapid growth of mobile internet, networks are becoming more important in people's daily lives. Different platforms are working together more often, and cross-platform interactions are increasing.

The importance of recognizing personality traits on social media has been shown by the recent interest in developing automatic personality recognition systems. These systems are usually based on well-known personality models, such as the DiSC Assessment[7], the Myers-Briggs Type Indicator (MBTI), and the[4, 5, 8, 11] Big Five Personality Traits Model.

More research is needed to fully improve the effectiveness of automated personality detection by using the ensemble method, as this area is still in its early stages [10]. This review aims to give an overview of the different types of personality models and detection methods.

Computational personality assessment combines psychology with machine learning algorithms. In psychology, a person's behavior and appearance are often influenced by their personality. There are different personality models in psychology, but the most widely accepted one is the "Big Five" personality model [13], which explains human personality through five traits: extraversion, agreeableness, openness, conscientiousness, and neuroticism. According to the research article [13], these five traits can be described with additional related traits as follows.

- **Extraversion:** Friendliness, Sociability, Confidence, Energy Level, Thrill-Seeking, Positivity
- **Agreeableness:** Trust, Ethics, Kindness, Teamwork, Humility, Empathy
- **Openness to Experience:** Creativity, Artistic Interests, Emotional Awareness, Curiosity, Intelligence, Open-mindedness
- **Conscientiousness:** Confidence in Abilities, Organization, Responsibility, Ambition, Self-control, Carefulness
- **Neuroticism:** Worry, Irritability, Sadness, Shyness, Impulsiveness, Sensitivity to Stress

This study aims to provide a more comprehensive and relevant personality assessment than what is typically available from media and polls. [6] The analysis of Twitter sentiments/texts is expanded by sorting tweets into four categories: Dominance, Influence, Submission, and Compliance (DISC). DISC has shown to be effective and consistent with previous research (in Social Sciences and Marketing) using clear and understandable dimensions.

This survey reviews many research articles related to personality prediction. Keywords like "personality prediction from text using deep learning" or "personality recognition/classification using hybrid deep learning approach" were used for searches in Google Scholar. The rest of this paper is organized as follows: Section 2 covers the background, Section 3 discusses related work on personality prediction, and Section 4 presents the conclusion.

Personality Trait	Characteristics
Openness (O)	From cautious/consistent to curious/inventive intellectual, polished, creative, independent, open-minded, imaginative, creative, curious, tolerant
Conscientiousness (C)	From careless/easy-going to organized/efficient reliable, consistent, self-disciplined, organized, hard working, has long-term goals, planner
Extraversion (E)	From solitary/reserved to outgoing/energetic, express positive emotions, excited, satisfied, friendly, seeks stimulation in the company of others, talkative
Agreeableness (A)	From cold/unkind to friendly/compassionate kind, concerned, truthful, good natured, trustful, cooperative, helpful, nurturing, optimistic
Neuroticism (N)	From secure/calm to unconfident/nervous angry, anxious, neurotic, upset, depressed, sensitive, moody

[Overview of the Big Five Personality Traits]

2. BACKGROUND STUDY

2.1 Text mining in social network

Text mining is crucial for extracting useful information and knowledge from unstructured data.

Scholars in [3] noted that social media sites offer a great platform for people to communicate and share their views and opinions. It becomes easier to understand an individual based on their activities. While there has been research on textual or image datasets, the idea of using multimodal datasets (which combine different types of data) has not been thoroughly explored.

2.2 Big Five Model, MBTI (Myers-Briggs Type Indicator), DISC

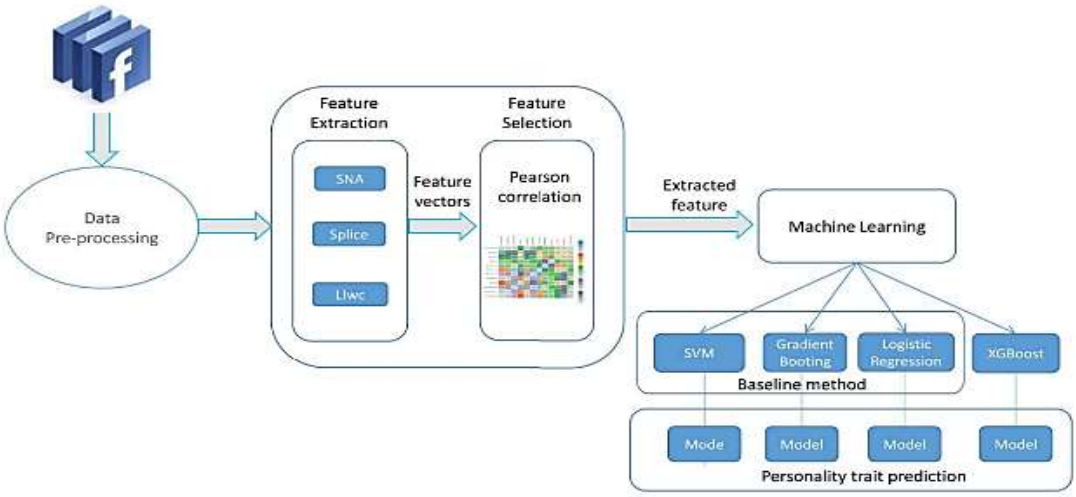
The Big Five Model and MBTI are popular methods for assessing personality. In psychology, the Big Five Factor is used to describe human personality. According to the literature review, the Big Five model is well-known and accurately shows someone's personality traits. This model was first developed in 1990 and is still in use today. The five personality traits are:

The Myers-Briggs Type Indicator (MBTI) has been a well-known and widely accepted personality test around the world for the past 55 years. [6] This test is based on Carl Jung's theory and describes human personality using four basic dimensions: extroversion/introversion, sensing/intuition, thinking/feeling, and judging/perceiving.

DISC is a straightforward, practical, and easy-to-understand personality test that was introduced in 1928. [6] DISC assesses behavior based on four main traits: Dominance, Influence, Steadiness, and Conformity.

3. METHODOLOGY

As using language from social media to predict personality has become more popular [11], there are more methods that combine language and social network features from profiles and status updates to figure out personality traits. The personality prediction framework shown in Fig. 1 includes steps for data cleaning, extracting important features,selecting those features, and then applying machine learning to get prediction results.



[figure 1: Existing System Diagram[11]]

3.1 Data Preprocessing

The dataset from myPersonality was cleaned up before moving on to selecting features and training. To clean the data, they first split each sentence into individual words and combined similar words. Then, they removed URLs, symbols, names, spaces, and changed everything to lowercase. Since many words in the LIWC and SPLICE features have common roots, linking personality to these roots could be problematic.For example, if verbs are changed to their base form, it becomes difficult to tell if they are in the present or past tense [11]. Therefore, in the cleaning process of their experiment, they avoided changing words to their base form and kept all the words as they were.

3.2 Feature Extraction

A user's behavior on social networks is influenced by the presence and actions of other users. These interactions can affect how new information or behaviors spread through groups. In this study, all the data is divided into two main groups. The first group focuses on text features, showing a user's language habits on Facebook, including counts of expressions and topics. To analyze Facebook status texts, they use two dictionaries: LIWC and SPLICE.The second group looks at social interaction behaviors, including network size, density, brokerage, and transitivity. This information shows a user's basic social network behavior on Facebook. LIWC, or the Linguistic Inquiry and Word Count dictionary, is commonly used in psychology studies. In this study, they use it to extract 85 language features from the texts, including five subcategories like standard counts. For text analysis, they chose LIWC2015, which is designed to quickly and efficiently analyze individual or multiple language files. Compared to LIWC 2007 and LIWC 2001, it aims to be clearer and more flexible, allowing users to explore word use in various ways.

SPLICE, or Structured Programming for Linguistic Cue Extraction, is a newer dictionary developed recently. It is still being updated and is expected to be widely used for personality prediction studies [11]. In this study, they use SPLICE to extract 74 language features, including cues related to the speaker's positive or negative self-assessment, as well as complexity and readability scores.

SNA, or Social Network Analysis, is a method used to study the social structure formed by the connections between people in a network. It looks at and measures the patterns of relationships that occur among people who interact with each other. One key idea of this method is that even indirect relationships (like friends of friends) in social groups are important.

3.3 Feature selection

Feature selection is important for building a model for two main reasons. First, it reduces the large number of features in the dataset by removing those that are not needed for training. This helps the model generalize better and speeds up the training process. Second, it helps the model better understand the important features and how they relate to the results.

Additionally, it improves the accuracy of the learning algorithms and lowers the processing needs [11].

To measure how strongly two variables are related and to find important features for predicting personality traits, they used Pearson correlation analysis, Eq. (1), as their main feature selection method. Pearson correlation measures the linear relationship between two variables and is used to predict how personality scores are related to the features extracted. For two variables (x, y), the linear correlation coefficient r is calculated using the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad (1)$$

where $\bar{x}$ and $\bar{y}$ are sample means given by the relations

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (2)$$

In the formula above, n represents the number of samples, and xi and yi refer to individual samples indexed by i. The value of r ranges from -1 to 1, inclusive. If x and y are perfectly correlated, r will be either 1 for a positive correlation or -1 for a negative correlation. If x and y are completely independent, r will be zero [11].

According to the experiment results [11], the XGBoost method performed better than the average baseline for all feature sets. It showed significantly higher accuracy for all personality traits, supporting the idea that XGBoost is a fast and scalable machine learning system. Compared to other gradient boosting methods, XGBoost uses a more controlled model to prevent overfitting and delivers better performance. By using all the features we collected, we can predict personality traits with an average accuracy of 74.2%.

I. PREDICTIVE MODEL BASED ON LANGUAGE

To assess psychological traits from language, there are two main approaches: the open vocabulary approach (which doesn't use predefined features) and the closed vocabulary approach (which uses predefined categories and associated features) for predicting personality on social media. Studies have shown that many papers use features from LIWC (Linguistic Inquiry and Word Count), SPLICE (Structured Programming for Language Cue Extraction), and SNA (Social Network Analysis). SPLICE, MRC, NRC, SentiStrength, and LIWC are tools for linguistic analysis. LIWC measures the percentage of words that fit into each category of its psychology dictionary. The two aspects of linguistic analysis are traditional LIWC dimensions and summary variables.

According to the research paper [11], they provide a method to create and use one type of Social Network Analysis (SNA) features and two types of linguistic features: Linguistic Inquiry and Word Count (LIWC) and Structured Programming for Linguistic Cue Extraction (SPLICE), based on the myPersonality dataset.

4. PREDICTING PERSONALITY

Researchers focused on predicting user personality traits using publicly available information from online social networks. There are also related data sources from social media like blogs, Twitter, Facebook, YouTube, and others in different languages and media. Various works have been done to find the best features and test machine learning algorithms for predicting personality [11].

Authors [4, 12] provided the Essays and MyPersonality datasets as benchmarks for personality recognition. The MyPersonality dataset includes 9,917 status updates from 250 users. It contains features such as status update text, author ID, personality scores, and social network details related to the users. [12] Tommy Tandera, Hendro, and Derwin Suhartono collected data through a Facebook app that used a personality traits test among other psychological tests. Another dataset is the stream-of-consciousness essay collection, which includes 22,468 anonymous users labeled with the Big Five personality traits.

4.1 COMPUTATIONAL PERSONALITY PREDICTION

With the rapid growth of social media, many methods have been developed to predict personality from online social networks (OSN). B. Y. Pratama and R. Sarno [1] used two different datasets and applied classification algorithms to recognize personality traits. The myPersonality dataset, which includes 250 users and 9,917 status updates, was used by M. M. Tadesse, H. Lin, B. Xu, and L. Yang [11].

Yen Lina Prasetio built [12] Support Vector Machine (SVM) and Latent Discriminant Analysis (LDA) models using an open vocabulary approach to recognize Big Five personality traits with the MyPersonality dataset. Their results showed that the LDA models performed better than the SVM model. Additionally, LDA improved computational efficiency by up to 74.17%.

In recent studies, Rhea Mahajan, Remia Mahajan, Eishita Sharma, and Vibhakar Mansotra [10] used two datasets, including MyPersonality, to predict the personality of 100 real-time Twitter users based on the Big Five model. They extracted features from tweets using a combination of CNN (Convolutional Neural Network) and BiLSTM (Bidirectional Long Short-Term Memory). Their findings showed that their model performs better than existing benchmark models, achieving an accuracy of 75.13%.

In this study, A. V. Kunte and S. Panicker [3] took features from a Twitter dataset using the Twitter streaming API and focused on Linear Discriminant Analysis (LDA), Multinomial Naive Bayes, and AdaBoost with the standard Twitter dataset. The results showed that Multinomial Naive Bayes had the highest accuracy of 73.43% for the 'OPN' (Openness) feature. AdaBoost and LDA had almost the same

accuracy, except for the 'OPN' feature.

In a study by [11], the authors examined both social network features and language features related to personality using the MyPersonality dataset. They discovered that using various dictionary-based language features could improve prediction results. They used Pearson correlation analysis to identify important features for predicting personality. They also found that selecting the best features boosts the performance of the machine learning algorithm and shortens training time. Personality predictions using LIWC features were more accurate than those using SPLICE features. Additionally, they showed that using SNA features gave better results than using language features in this study. They found that the XGBoost method outperformed three baseline feature sets. Compared to other gradient boosting methods, XGBoost avoids overfitting and improves both model performance and extraction speed.

4.2 DEEP NEURAL NETWORK

Natural Language Processing (NLP) studies how computers and human languages interact. Language modeling is a fundamental task in artificial intelligence and NLP. Recently, deep learning models and distributed representations have greatly improved tasks like sentence and document modeling, as well as text-based sentiment analysis. Text classification is crucial for many NLP applications. While Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) use different methods to understand NLP, each has its own strengths and weaknesses in text modeling [5].

Although CNNs use different convolution filters to capture higher-level features, they don't keep track of historical and contextual information in long texts. The longer the input sequence, the more convolution and pooling layers are needed. RNNs, which are designed to handle sequences, have a memory that captures long-term relationships, where recent terms are more important than earlier ones. However, RNNs might be less efficient because they need to learn the context of the entire document. To address this issue, the Long Short-Term Memory (LSTM) model is used to improve upon RNNs [5, 9].

To improve how sentences or documents are represented, many modified versions of basic CNN and RNN have been developed. A model using CNN and LSTM was suggested by Tommy Tandera, Hendro, Derwin Suhartono, Rini Wongso, and Yen Lina Prasetio [12]. The authors found that using only CNN or LSTM does not give the best results. This is because CNN struggles to capture long-term sequence information, while LSTM cannot learn high-level features. However, combining CNN and LSTM performs better. This shows that a single LSTM layer cannot fully capture long-term dependencies..

Donghang Pan, Jingling Yuan, Lin Li, and Deming Sheng focused on CNN (Convolutional Neural Network), DNN (Deep Neural Network), LSTM (Long Short-Term Memory), Bi-LSTM, and attention-based LSTM. Using the Chinese Implicit Sentiment Analysis (SMPECISA 2019) dataset, their experiments on public data showed that both the LSTM-based models and the CNN model performed well in sentiment classification. Their results were much better than the DNN model. In this research, the two-category and three-category experiments revealed that the LSTM and CNN models gave good classification results due to their special way of extracting features.

5. CONCLUSIONS

Personality prediction is a growing research area that aims to automatically determine a user's personality traits from publicly shared information on social media. This paper reviews past studies on personality prediction from text using different traditional machine learning and deep learning techniques. Additionally, it examines the language features used in various methods to create prediction systems. Identifying personality traits can be useful in fields like psychology, health assessments, business recommendation systems, and HR management. Future improvements in personality prediction could include better machine learning models, more precise feature selection from social media posts, and improved data processing methods.

6. REFERENCES

[1] B.Y. Pratama and R Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM." In 2015 International Conference on Data and Software Engineering (ICoDSE), pp. 170-174. IEEE, 2015. [2] Jinghua Zhaoa , Dalin Zeng b,∗ , Yujie Xiaoc , Liping Chea , Mengjiao Wang d" User personality prediction based on topic preference and sentiment analysis using LSTM model" In 2020 / Pattern Recognition Letters 138 (2020) 397–402.

[3] A. V. Kunte and S. Panicker, "Using textual data for Personality Prediction:A Machine Learning Approach," 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019, pp. 529-533, doi:10.1109/ISCON47742.2019.9036220.

[4] M. A. Rahman, A. Al Faisal, T. Khanam, M. Amjad and M. S. Siddik, "Personality Detection from Text using Convolutional Neural Network," 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019, pp. 1-6, doi: 10.1109/ICASERT.2019.8934548.

[5] H. Ahmad, M. U. Asghar, M. Z. Asghar, A. Khan and A. H. Mosavi, "A Hybrid Deep Learning Technique for Personality Trait Classification From Text," in IEEE Access, vol. 9, pp. 146214-146232, 2021, doi: 10.1109/ACCESS.2021.3121791.

[6] Nadeem Ahmad, Jawaid Siddique" " Personality Assessment using Twitter Tweets" Procedia Computer Science 112 (2017) 1964–1973

[7] H. Setiawan and A. A. Wafi, "Classification of Personality Type Based on Twitter Data Using Machine Learning Techniques," 2020 3rd International Conference on Information and Communications 5 Technology (ICOIACT), 2020, pp. 94-98, doi: 10.1109/ICOIACT50329.2020.9332152.

[8] D. E. Cahyani and A. F. Faishal, "Classification of Big Five Personality Behavior Tendencies Based On Study Field with Twitter Analysis Using Support Vector Machine," 2020 7th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), 2020, pp. 140-145, doi: 10.1109/ICITACEE50144.2020.9239130.

[9] D. Pan, J. Yuan, L. Li and D. Sheng, "Deep neural network-based classification model for Sentiment Analysis," 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), 2019, pp. 1-4, doi: 10.1109/BESC48373.2019.8963171.

[10] Rhea Mahajan , Remia Mahajan , Eishita Sharma , Vibhakar Mansotra ""Are we tweeting our real selves?" personality prediction of Indian Twitter users using deep learning ensemble model" https://doi.org/10.1016/j.chb.2021.107101

[11] M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Personality Predictions Based on User Behavior on the Facebook Social Media Platform," in IEEE Access, vol. 6, pp. 61959- 61969, 2018, doi: 10.1109/ACCESS.2018.2876502.

[12] Tommy Tandera, Hendro, Derwin Suhartono*, Rini Wongso, and Yen Lina Prasetio" Personality Prediction System from Facebook Users" 2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13-14 October 2017 Procedia Computer Science 116 (2017) 604–611

[13] W. M. K. S. Ilmini, T. G. I. Fernando" Computational Personality Traits Assessment: A Review" 978-1-5386-1676-5/17/2017 IEEE

[14] Zan Mo Mo Aung, Phyu Hninn Myint" Personality Prediction Based on Content of Facebook Users:A Literature Review" 978-1-7281-1651-8/19/2019 IEEE

[15] Hussain Ahmad, Muhammad Zubair Asghar*, Alam Sher Khan, and Anam Habib" A Systematic Literature Review of Personality Trait Classification from Textual Content" t. Sci. 2020; 10:175–193