Personality prediction system using machine learning approaches: a comparative study

Angad Singh[1], Priti Maheshwary[1], Nitin Kumar Mishra[2], Neerja Dubey[3]
[1]Department of Computer Science and Engineering (Data Science), Oriental College of Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India
[2]School of Computing Science Engineering and Technology, Bennett University, Greater Noida, India
[3]Department of Computer Science and Engineering, Lakshmi Narain College of Technology and Science, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India

Article Info

Article history:

Received Aug 25, 2025
Revised May 25, 2026
Accepted Jun 19, 2026

Keywords:

Analysis
Big Five
Classifiers
Machine learning
Personality prediction systems

ABSTRACT

Identifying personality traits from text offers valuable insights for human resource management, customer service, political campaigning, healthcare, fraud detection, and risk assessment. In psychology, personality prediction from text is an important area with the Big Five model, among the leading frameworks. Popular datasets for this task include Essays. Past works have primarily relied on conventional machine learning (ML) models using linguistic features. This paper evaluates and contrasts ten ML classifiers ‘effectiveness for personality prediction using the Essays dataset. The support vector machine (SVM) classifier yielded the best overall performance, a mean accuracy of 57.68% and means F1-score of 61.16% across all five personality traits, and outperformed logistic regression (LR). Thereby demonstrating its superior predictive capability for this task and significance as an interpretable baseline for future deep learning integration.

Corresponding Author:

Angad Singh
Department of Computer Science and Engineering (Data Science), Oriental College of Technology
Rajiv Gandhi Proudyogiki Vishwavidyalaya
Bhopal, India
Email: angada2007@gmail.com

1. INTRODUCTION

Online communities (social media platforms) i.e., Facebook, WhatsApp, and Twitter are popular among users for sharing information through status updates, tweets, comments, and messages. The utility of social networking sites is constantly growing. It is a context-aware recommendation engine that improves the delivery of customized information on social media networks using advanced natural language processing (NLP) methods [1]. Nowadays, direct communication between people has reduced considerably as people prefer to communicate indirectly using smart devices such as smartphones and tablets. In such a case, it is difficult to identify the personality of users. According to Facebook user statistics, there are 3.065 billion active users on this platform every month. The average time spent per day is 30.8 minutes, with only 1.7% accessing their accounts on their desktop computers, while the majority of users use smartphones and tablets [2]. However, what people write on social platforms helps us get the necessary information as they take time to express their thoughts and feelings. Twitter and Facebook users widely express their emotions and feelings through comments and status updates. This study centers on the linguistic outlook of users' status updates, even though Facebook is also used to share more videos and photos. This study concentrates on the field of psychology that shows the relationship between personality and the meaningful way a person behaves [3], [4]. This relationship can be discovered and demonstrated using NLP. NLP is a technology. It teaches human language in written or spoken form and is a part of artificial intelligence. NLP has its roots in linguistics [5]-[7]. Therefore, the main purpose of this research is to analyze different prediction systems that can predict a user's personality according to his or her behavior on Facebook.

With the widespread development of mobile media such as mobile internet, the influence of the network in users' daily routines is increasing, participation between different platforms is becoming more regular, and cross-stage communication is expanding. For instance, Facebook data can be used to sign in different shopping sites like Amazon, Flipkart, and Sephora, which gives a more extensive application position to recognize clients' personality traits [8]-[10] in view of social media data. However, the correlation between user personality and behavior is the prime focus of most user inclination analyses [11]-[13].

Personality models [14]-[17] such as the Big Five are essential in human psychology research for evaluating personality traits such as extraversion, agreeableness, and emotional stability. A framework for understanding and classifying personality traits [18]-[20] according to dimensions is provided by the Big Five model, such as openness (OPN), conscientiousness (CON), extraversion (EXT), agreeableness (AGR), and neuroticism (NEU), as discussed:

– OPN: is the person innovational and inquisitive vs. adamant and alert?
– CON: is the person competent and ordered vs. awkward and irresponsible?
– EXT: is the person balmy, effusive, and aggressive vs. self-restrained and unsociable?
– AGR: is the person true, candid, benevolent, and prudent vs. dubious, convoluted, sparse, and arrogant?
– NEU: is the person subtle and annoyed vs. protected and positive?

The Myers-briggs type indicator (MBTI) is a well-known personality model that uses four traits, namely introvert/extrovert (IE), thinking/feeling (TF), sensing/intuiting (SN), and judging/perception (JP). Mishra *et al.* [20] presented that core applications of MBTI model such that: i) self awareness, ii) human resource management, iii) career guidance, iv) team building, and v) conflict resolution. In addition, this model was developed by Myers and created MBTI type personality based on preferences. Current studies have shown promising results in predicting MBTI personality types based on machine learning (ML) models from textual data such as Reddit posts, achieving up to 93% accuracy. The bidirectional encoder representations from transformers (BERT) model were created by Google researchers and has been open source since 2018. It is based on advanced technology and is used to classify MBTI personality in a better way than the traditional methods [19]. The results showed that it has better performance in text classification models based on word clustering and fixed word settings. Moreover, in the field of personality computing, the potential of ML models such as logistic regression (LR), random forest (RF), and NLP was used to enhance the MBTI personality type classification on text data. The RF and LR models have been used to predict MBTI personality types based on daily conversation, achieving high accuracy across a variety of personality dimensions.

Classification is a standard task in personality processing. Elmahalawy *et al.* [21] demonstrated that ML classifiers play a core role in personality prediction. Studies have shown the effectiveness of various classifiers in this field. For example, research has used ML algorithms such as LR, RF, XG Boost, k-nearest neighbors (KNN), stochastic gradient descent (SGD), and support vector machine (SVM) to predict personality types which are built on the MBTI design with accuracies ranging from 75.99% to 77.60% [22]. Linguistic Inquiry and Word Count (LIWC) [23] is a standard tool for psycholinguistic feature extraction in personality prediction. Furthermore, the Big Five personality and MBTI models were analyzed using neural network (NN), RF, and SVM methods, with NN reaching the highest accuracy of 76%, followed by RF and SVM [24].

Kamalesh and Bharathi [25] used XG boost approach to predict personality traits with over 99% accuracy and precision for various personality traits, demonstrating its exceptional performance in personality prediction. These research findings in personality processing emphasize the variety of ML classifiers used in personality prediction tasks, each of which exhibits varying levels of accuracy and effectiveness.

The Big Five personality model is well-known trait model that uses five traits such that: i) OPN, ii) CON, iii) EXT, iv) AGR, and v) NEU. It has relationships significantly with Facebook use among Facebook users. Research studies have shown that extroversion is positively associated with social media use, indicating that extroverted persons tend to be more active on platforms such as Facebook [26]. In addition, findings suggest that CON, OPN to experience, and NEU also shape Facebook behavior, with CON and OPN being strongly associated with student participation and time spent preparing for class [27]. In addition, the use of social media, including Facebook, has been linked to dishonest behavior, exclusion anxiety, fear of missing out (FoMO), and abuse, highlighting the influence of personality traits on online behavior and interactions [26].

Serrano-Guerrero *et al.* [28] presented that various ML techniques, such as decision tree (DT), linear discriminant analysis (LDA), Naive Bayes (NB), gradient boosting, KNN, LR, and SVM, have been used to analyze social media data and predict personality based on semantic features and text content. Furthermore, research on the Big Five trait model using social media platforms such as Facebook has attracted considerable attention. Studies have shown that users' online behavior and content can predict personality traits. Moreover, advanced algorithms such as BERT, robustly optimized BERT pretraining approach (RoBERTa), and XLNet for feature extraction have shown promising results in increasing the accuracy of personality prediction models on social media data. These studies focus on the potential of leveraging social media data to accurately understand and classify persons' personalities, providing insights for applications in job satisfaction, relationship success, and targeted advertising.

Bama *et al.* [29] used the transformer-based DL model combined with label attention techniques for MBTI type personality classification from text data. Bhin and Choi [30] used a DL model to extract modality-specific features and combined them using an attention layer to improve personality prediction performance. LIWC [20] is a technique for analyzing text by counting the occurrences of words from different categories in a text. Researchers have found a strong connection between linguistic features and personality traits. Another study by Kolenbrander and Michaels [31] was conducted using LLMs to reproduce human personality traits in text. Singh *et al.* [32] stated that Twitter feed data was used to classify Big Five personality models using DL methods and traditional ML methods. Maharani and Effendy [33] used Twitter data and Big Five model to survey different personality prediction systems. The Essays dataset is the most popular standard dataset for predicting personality from text data. In this research, the Essays [23] dataset is used as a corpus. It contains 2468 essays written by psychology graduate students. In the Essays dataset, every essay is labeled by the Big Five personality trait using binary values (0/1 or absent/present). In this research study, we have also used the Essays dataset.

From the above discussion, it is clear that studies on personality traits identification in recent years shifted toward transformer-based architecture such as BERT, robustly optimized BERT pretraining approach (RoBERTa), and multi-document transformers; however, three major gaps remain:

- These models are data hungry, resource-intensive and in favour of deep learning, interpretability is overlooked.
- Lack of systematic comparison among traditional ML classifiers using unified psycholinguistic feature sets
- Insufficient combination of multiple psycholinguistic features (LIWC, Empath, and medical research council (MRC) Psycholinguistic Database) in a single comparative framework.

The major contributions of this study are enlisted next:

- Comparative evaluation of ten ML classifiers on the Essays dataset.
- Integration of three psycholinguistic features, namely, LIWC, Empath, and MRC to provide three varied perspectives of the feature space.
- Application of two metrics, one primary metric is accuracy, and the other is F1-score metric to calculate the performance of the classification model.

2. METHOD

2.1. Datasets

This study uses the Essays dataset, which is a well-known dataset among scholars in the fields of text computing and psychology. The details of the Essay dataset are given in Table 1.

Table 1. The details of the Essay dataset

S. No.	Binary value	EXT	NEU	AGR	CON	OPN
1.	Absent (0)	1191	1234	1157	1214	1196
2.	Present (1)	1276	1233	1310	1253	1271

2.2. Features

The results and capabilities of this study are compared using various features. The purpose is to determine how well these characteristics work during personality modeling. We employed a closed-vocabulary linguistic feature approach for ML implementation. During this research, we applied linguistic features such as MRC, Empath, and LIWC. Various ML techniques were used independently using the linguistic features of the open vocabulary (OV) approach. OV does not require predefined features. OV approach automatically explores the dataset to find relationships between personality-related words.

In the development of personality models, various linguistic features play an important role. Integration of traditional psycholinguistic features with language model embeddings can significantly enhances personality prediction models. In addition, there is a hybrid model that combines pre-trained language representations with a bidirectional long-short-term memory (BLSTM) network. Further, it uses theory-based psycholinguistic features, and then it has shown promising results in predicting personality traits from text data, outperforming existing models on benchmark datasets such as the Big Five Essay dataset. The details of the feature counts are given in Table 2.

Table 2. Summary of features

S. No.	Source of feature	Number of features
1.	Empath	25
2.	MRC	194
3.	LIWC	102

Each classifier combined all three feature sets (MRC, Empath, and LIWC) into a single feature vector. In this study, different experimental scenarios were not adopted for different feature sets. Feature coefficient analysis of the linear SVM model shows that the three factors that most influence the prediction of EXT and NEU are emotional tone (LIWC), emotional categories (Empath), and imagery/concreteness measure (MRC). This strengthens the idea that personality traits can be significantly predicted by psycholinguistic emotional cues.

Jiang *et al.* [34] demonstrated that language patterns significantly influence personality traits. Differences in personality scores are exhibited by bilingual speakers between languages. Language fluency represents a strong predictor of personality valence. It is noted that fluent speakers report more positive personality profiles. Also, linguistic cues are correlated with personality traits. It leads to the development of models that can extract hidden personality characteristics from text data using deep learning approaches. Naz *et al.* [35] explores the prediction of EXT personality traits (introvert vs. extrovert) using machine and deep learning methods. They applied NLP [36] with TF-IDF, PoS tagging, word embeddings Word2Vec, and GloVe. The key contribution is the use of advanced sentence embeddings, which significantly enhance semantic understanding.

2.3. Classification model

During the classification process, ten well known ML methods were used, as mentioned below. In this experimentation, we include ML methods such as SVM, NB, LR, SGD, LDA, KNN, DT, XG Boost, RF, and NN. In this study, each of the ML algorithms will be distinguished and compared using Essays dataset. We utilized Python libraries for a five-fold cross validation method for model validation. Five-fold cross validation divided the dataset into testing data for 20% and training data for 80%, respectively. To determine the accuracy and F1-score of every one ML algorithm in predicting personality types. Also, we outline the preprocessing pipeline for extracting features from the database of MRC Psycholinguistic, Empath categories, and LIWC lexicon. The tools and libraries used are Empath via the empath Python library, LIWC via the commercial LIWC software and MRC via the NLTK-enhanced dataset, along with the tokenization, filtering, and normalization procedures applied prior to feature extraction.

The preprocessing process begins with collecting raw text data, which is then thoroughly cleaned to remove unnecessary symbols and special characters. After that, the text is tokenized to be divided into meaningful units (words or tokens) and converted to lowercase to ensure uniformity. Lemmatization or stemming is used to normalize words to their original form, while removing stop words that do not provide much semantic value so as to reduce repetition of words. To further improve the text, other measures are taken such as correcting spelling, addressing negative sentences, and removing unnecessary spaces. Finally, the cleaned and standardized text is translated into a structured format suitable for feature extraction (e.g., LIWC, MRC, and Empath), which ensures consistency and reliability for subsequent ML or deep learning models.

Python (version 3.x) was used to conduct tests, as it is easy to use and provides strong support for data analysis and ML. Scikit-learn were used for feature selection, model development, and performance evaluation, while libraries like Numpy and Pandas were used for numerical calculations and data management. The Table 3 exhibits the ML algorithms with the parameters used in the experiment. The data preparation step is shown in Figure 1.

Table 3. ML algorithms with parameters

ML algorithm	Parameters
LR	Regularization: l2 c: 1.0 solver: liblinear
NB	Distribution: gaussian, prior probabilities: none
DT	Criterion: gini, max depth: none
KNN (neighbor-5)	Number of neighbors: 5, distance metric: euclidean
SVM	Kernel: linear, C (regularization): 1.0, gamma: scale
MLP NN (1 hidden layer-40 neurons)	Hidden layers: 1, neurons: 40, activation: ReLU, solver: Adam
SGD	Loss function: log_loss, learning rate: optimal, max iterations: 1000
LDA	Solver: svd, shrinkage: none
RF (estimators-10)	Number of estimators: 10, max depth: none
XG Boost (estimators-10)	Number of estimators: 10 max depths: 6, learning rate: 0.1

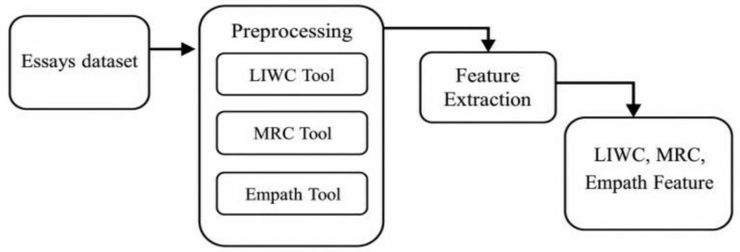

Figure 1. The data preparation step

3. RESULTS AND DISCUSSION

In this section, we describe the experimental results of classification obtained by including well known ML algorithms. Table 3 shows the accuracy and overall mean of each ML algorithms and Table 4 shows the F1-score and overall mean of each ML algorithms. We just report the ML algorithm including the greater accuracy and F1-score in every Big Five personality traits as well as report the ML algorithm with low accuracy and F1-score in every trait.

Table 4. Results (accuracy) of various ML techniques

Classification algorithm	EXT	NEU	AGR	CON	OPN	MEAN
LR	55.79	58.63	54.95	55.71	59.41	56.90
NB	53.86	56.74	47.70	54.86	58.07	54.25
DT	52.25	53.87	51.93	51.39	54.14	52.72
KNN (neighbor-5)	53.26	54.01	51.45	51.09	53.95	52.75
SVM	**56.65**	**59.26**	**56.64**	**56.07**	**59.78**	**57.68**
MLP NN (1 hidden layer-40 neurons)	54.24	55.48	53.91	53.30	54.32	54.25
SGD	51.90	53.13	50.85	53.03	55.44	52.87
LDA	55.65	58.72	54.98	55.57	59.33	56.85
RF (estimators-10)	53.41	54.83	53.32	53.06	56.91	54.31
XG Boost (estimators-10)	56.53	57.53	54.18	54.47	57.56	56.05

3.1. Accuracy

Using Essays dataset, Table 4 exhibits the results (classification) of various ML techniques. Table 4 shows that SVM performs very well on EXT and NEU features, with the highest average accuracy (57.68%). Similar findings were obtained for LDA and LR, suggesting that linear decision boundaries are useful for psycholinguistic features. On the other hand, KNN and DT are less accurate. This may be due to the sparseness of the features and their sensitivity to high-dimensional data.

3.2. F1-score

Using F1-score, Table 5 exhibits the results (classification) of various ML techniques. Table 5 shows that SVM has the highest overall F1-score (61.16%), especially for AGR (64.78%). The instability across traits is evident from NB's strong performance for NEU and weak performance for AGR. The difference between F1-score and accuracy reflects the minor effects of class imbalance.

Table 5. Results (F1-score) of various ML techniques

Classification algorithm	EXT	NEU	AGR	CON	OPN	MEAN
LR	59.09	58.24	60.16	58.08	60.73	59.26
NB	53.61	60.53	57.00	58.34	50.49	55.99
DT	53.72	54.37	54.65	51.79	54.75	53.86
KNN (neighbor-5)	56.46	54.96	55.83	54.32	53.65	55.04
SVM	**61.74**	**59.35**	**64.78**	**59.54**	**60.37**	**61.16**
MLP NN (1 hidden layer-40 neurons)	56.28	54.70	57.24	54.20	55.88	55.66
SGD	47.06	52.95	39.58	44.33	56.37	48.06
LDA	58.88	58.25	60.27	58.00	60.59	59.20
RF (estimators-10)	50.49	49.14	52.55	48.47	53.78	50.89
XG Boost (estimators-10)	58.93	57.64	58.25	55.48	58.82	57.82

Using ML to classify the traits, we experimented with personality prediction according to Big Five Trait models in this study. We utilized ten algorithms for ML. They are SVM, NB, LR, SGD, LDA, KNN, DT, XG Boost, RF, and NN with three elements which are LIWC feature, MRC feature, and Empath feature. The evaluation model uses five-fold cross validation. The experimental scenario based on the Essays dataset demonstrates that the SVM, LR, LDA, and XG Boost algorithm for the entire trait with LIWC feature, MRC feature, and Empath feature yielded the highest accuracy. An experimental result represents that the best results for SVM in all competing algorithms are 56.65, 59.26, 56.64, 56.07, and 58.78 for EXT, NEU, AGR, CON, and OPN. After this, LR, LDA, and XG Boost algorithm represents that the best results in accuracy metric, respectively for all the traits in Big Five model (refer Table 4).

As mentioned above, we are made comparative study between all competing algorithms in Essays dataset with the accuracy and F1-score metric (refer Tables 4 and 5). Table 5 is evident for the highest F1-score of OPN trait is 60.73 by LR; highest F1-score of NUE trait is 60.53 by NB; highest F1-score of EXT trait is 61.74 by SVM; highest F1-score of AGR trait is 64.78 by SVM; and highest F1-score of CON trait is 59.54 by SVM. Also, it can be noted that the lowest F1-score of AGR trait is 05.70 for NB. We also compare the mean value of each personality trait in this experimentation. The highest mean value of SVM is 57.68 for accuracy metric. The highest mean value of SVM is 61.16 for F1-score metric.

The best performance of SVM is due to effectively handling high-dimensional feature spaces with limited training data. It provides a balance between predictive performance and computational efficiency. While deep learning models offer greater representational power, they typically require huge amount of data and higher computational power. SVM provides competitive accuracy with lower training overhead and better generalization. Also, SVM provides a reasonable level of interpretability due to support vectors and margin manipulation. It makes it a practical and effective selection for personality trait prediction.

According to recent research, Transformer-based models perform far better than traditional ML models like SVM, RF, and NB in personality prediction. On the Big Five datasets, methods using BERT, RoBERTa, and DistilBERT demonstrate high accuracy and F1-score, as well as capture deeper contextual information. These results support our findings, which show that deep contextual representations outperform traditional techniques in classifying personality traits.

Error analysis shows that classification of AGR and CON traits is often inaccurate, likely due to overlapping linguistic cues such as duty-related phrases and positive sentiment. In emotionally neutral articles with minimal linguistic cues, NEU predictions appear to be confused. A binary labeling strategy may further increase classification uncertainty.

Comparative studies of social media-based personality prediction [37] have used classifiers such as NB, RF, SVM, and NN. In this study, among the ten ML classifiers evaluated on the Essays dataset, the SVM classifier had the highest average accuracy of 57.68%, demonstrating its effectiveness for personality prediction.

3.3. Comparison with state-of-the-art deep learning models

The best results are obtained in our study by the SVM model, with a mean accuracy of 57.68%; however, this performance is modest compared to recent cutting-edge deep learning models, mainly BERT-based architectures. Model capacity and feature representation are the two main reasons for the difference. Our study uses traditional ML which are interpretable and computation efficient. Further, traditional lexical, psycholinguistic, and statistical features (e.g., MRC, Empath, and LIWC) are utilized. On the other hand, BERT like models is transformer-based models. They are pre-trained on huge datasets and can model context-rich embeddings. However, fine-tuning of these models requires huge computational resources and large annotated datasets. Our work presents a baseline using interpretable, resource-efficient methods, and in the future, we will integrate deep models.

4. CONCLUSION

This study compared 10 traditional ML classifiers for predicting personality traits by integrating three psycholinguistic features. The experimental results show that SVM achieves a superior predictive performance compared to the other ML models. The results reaffirm the effectiveness of SVM. Although recently transformer based models have shown substantially higher performance, basic ML models remain valuable for their interpretability, computational efficiency, and data efficiency. This work therefore provides a strong, interpretable baseline for future research on hybrid or deep learning-based personality prediction systems.

However, there are limitations of the work as follows: i) a small dataset, namely Essays is utilized, may affect performance, ii) personality traits are binary; they may also be fractional numbers representing the degree of traits, iii) only psycholinguistic features were used without contextual embeddings, iv) hyperparameter optimization is done for the basic configuration only, and v) the Essays dataset is a balanced one; the class imbalance across traits may influence performance metrics.

To overcome these limitations, future work will focus on the following directions: i) integration of contextual features with psycholinguistic features, ii) multi-learning regression modeling instead of independent binary classification, iii) application of feature importance for interpretability, iv) hyperparameter tuning using Bayesian optimization, and v) cross-dataset validation for using large imbalanced datasets.

FUNDING INFORMATION

No funding was involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Angad Singh	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Priti Maheshwary	✓	✓	✓		✓				✓	✓		✓	✓	
Nitin Kumar Mishra	✓	✓	✓		✓				✓	✓	✓	✓		
Neerja Dubey	✓	✓	✓	✓	✓				✓	✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing -**O**riginal Draft
E : Writing - Review &**E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Dataset | IEEE DataPort at http://doi.org/10.1037/0022-3514.77.6.1296, reference no. [23].

REFERENCES

[1] Q. Yang, W. Sun, M. Habibi, and I. Albaijan, "An LLM-driven context-aware recommendation system integrating NLP for enhanced social media personalization," *International Journal of Data Science and Analytics*, vol. 22, no. 1, 2026, doi: 10.1007/s41060-025-00964-z.

[2] MixBloom, "Average Time Spent on Social Media (2024)," mixbloom.com, [Online]. Available: https://www.mixbloom.com/resources/average-time-spent-on-social-media-2022. (Accessed: Jul. 15, 2025).

[3] C. S. S. S. Sree, B. Akshayavarthini, K. Sruthi, G. Supraja, and T. Tamilvizhi, "Personality Prediction Using Machine Learning," *International Research Journal on Advanced Engineering Hub (IRJAEH)*, vol. 3, no. 4, pp. 1460–1464, Apr. 2025, doi: 10.47392/IRJAEH.2025.0208.

[4] Y. Yu *et al.*, "Detecting five-pattern personality traits using eye movement features for observing emotional faces," *Frontiers in Psychology*, vol. 15, pp. 1–10, Sep. 2024, doi: 10.3389/fpsyg.2024.1397340.

[5] H. Jiang, X. Zhang, X. Cao, C. Breazeal, D. Roy, and J. Kabbara, "PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits," in *Findings of the Association for Computational Linguistics: NAACL 2024*, Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 3605–3627, doi: 10.18653/v1/2024.findings-naacl.229.

[6] M. A. Akber, T. Ferdousi, R. Ahmed, R. Asfara, R. Rab, and U. Zakia, "Personality and emotion—A comprehensive analysis using contextual text embeddings," *Natural Language Processing Journal*, vol. 9, pp. 1–13, Dec. 2024, doi: 10.1016/j.nlp.2024.100105.

[7] G. Serapio-García *et al.*, "Personality Traits in Large Language Models," *Research Square, preprint*, Aug. 2023, doi: 10.21203/rs.3.rs-3296728/v1.

[8] E. Kafeza, A. Kanavos, C. Makris, G. Pispirigos, and P. Vikatos, "T-PCCE: Twitter Personality based Communicative Communities Extraction System for Big Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1625-1638, 1 Aug. 2020, doi: 10.1109/TKDE.2019.2906197.

[9] M. Gjurković, I. Vukojević, and J. Šnajder, "SIMPA: Statement-to-Item Matching Personality Assessment from text," *Future Generation Computer Systems*, vol. 130, pp. 114–127, May 2022, doi: 10.1016/j.future.2021.12.014.

[10] X. Zhao, Z. Tang, and S. Zhang, "Deep Personality Trait Recognition: A Survey," *Frontiers in Psychology*, vol. 13, pp. 1–18, May 2022, doi: 10.3389/fpsyg.2022.839619.
[11] K. El-Demerdash, R. A. El-Khoribi, M. A. I. Shoman, and S. Abdou, "Deep learning based fusion strategies for personality prediction," *Egyptian Informatics Journal*, vol. 23, no. 1, pp. 47–53, Mar. 2022, doi: 10.1016/j.eij.2021.05.004.
[12] G. Caron and S. Srivastava, "Identifying and Manipulating the Personality Traits of Language Models," *arXiv preprint*, 2022, doi: 10.48550/arXiv.2212.10276.
[13] M. A. Kosan, H. Karacan, and B. A. Urgen, "Predicting personality traits with semantic structures and LSTM-based neural networks," *Alexandria Engineering Journal*, vol. 61, no. 10, pp. 8007–8025, 2022, doi: 10.1016/j.aej.2022.01.050.
[14] H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging," *Journal of Big Data*, vol. 8, no. 1, pp. 1–20, Dec. 2021, doi: 10.1186/s40537-021-00459-1.
[15] 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," *IEEE Access*, vol. 9, pp. 146214–146232, 2021, doi: 10.1109/ACCESS.2021.3121791.
[16] F. Yang, X. Quan, Y. Yang, and J. Yu, "Multi-Document Transformer for Personality Detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, May 2021, pp. 14221–14229, doi: 10.1609/aaai.v35i16.17673.
[17] M. A. Rahman, A. Al Faisal, T. Khanam, M. Amjad, and M. S. Siddik, "Personality Detection from Text using Convolutional Neural Network," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh: IEEE, May 2019, pp. 1–6, doi: 10.1109/ICASERT.2019.8934548.
[18] K. El-Demerdash, R. A. El-Khoribi, M. A. I. Shoman, and S. Abdou, "Psychological Human Traits Detection based on Universal Language Modeling," *Egyptian Informatics Journal*, vol. 22, no. 3, pp. 239–244, Sep. 2021, doi: 10.1016/j.eij.2020.09.001.
[19] H. Saberi and R. Ravanmehr, "Transformer-Based Personality Trait Recognition Enhanced by Contextual Augmentation," *International Journal of Web Research*, vol. 9, no. 1, pp. 45–58, 2026, doi: 10.22133/ijwr.2025.543527.1305.
[20] N. K. Mishra, A. Singh, and P. K. Singh, "Multi-label personality trait identification from text," *Multimedia Tools and Applications*, vol. 81, no. 15, pp. 21503–21519, Jun. 2022, doi: 10.1007/s11042-022-12548-1.
[21] A. R. Elmahalawy, L. Li, X. Wu, X. Tao, and J. Yong, "Computational Personality Analysis with Interpretability Empowered Prediction," in *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, Tianjin, China: IEEE, May 2024, pp. 413–418, doi: 10.1109/CSCWD61410.2024.10580811.
[22] L. Thapa, A. Pandey, D. Gupta, A. Deep, and R. Garg, "A Framework for Personality Prediction for E-Recruitment Using Machine Learning Algorithms," in *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2024, pp. 1-5, doi: 10.1109/Confluence60223.2024.10463354.
[23] J. W. Pennebaker and L. A. King, "Linguistic styles: Language use as an individual difference," *Journal of Personality and Social Psychology*, vol. 77, no. 6, pp. 1296–1312, 1999, doi: 10.1037/0022-3514.77.6.1296.
[24] M. H. A. Rauf, E. J. Johan, and Z. A. Aziz, "A Performance Comparison of NLP Text Pre-Processing Techniques for Analysing Personality Based on Myers Briggs Type Indicator (MBTI)," in *2023 IEEE International Conference on Computing (ICOCO)*, Langkawi, Malaysia, 2023, pp. 172-175, doi: 10.1109/ICOCO59262.2023.10397603.
[25] M. D. Kamalesh and B. Bharathi, "Analysing the Personality traits using machine learning techniques," in *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, Chennai, India: IEEE, Jul. 2022, pp. 1–6, doi: 10.1109/ICSES55317.2022.9914112.
[26] S. Y. Soeharso, C. Moningka, and S. Selviana, "Big-Five Personality: Prediction of Fear of Missing Out, Problematic Social Media Use, and Deceptive Behavior," *Journal An-Nafs: Kajian Penelitian Psikologi*, vol. 8, no. 1, pp. 128–146, Jun. 2023, doi: 10.33367/psi.v8i1.2969.
[27] J. M. Roos, "The Intersection of Personality Traits and Social Media Usage: Large-Scale Representative Samples of Internet Users in Sweden," *Psych*, vol. 5, no. 1, pp. 70–79, Feb. 2023, doi: 10.3390/psych5010008.
[28] J. Serrano-Guerrero, B. Alshouha, M. Bani-Doumi, F. Chiclana, F. P. Romero, and J. A. Olivas, "Combining machine learning algorithms for personality trait prediction," *Egyptian Informatics Journal*, vol. 25, pp. 1–13, Mar. 2024, doi: 10.1016/j.eij.2024.100439.
[29] B. S. Bama, M. S. Hema, S. Esakkirajan, and M. N. Guptha, "A hierarchical transformer network with label attention for personality prediction by MBTI classification," *Applied Soft Computing*, vol. 178, Jun. 2025, doi: 10.1016/j.asoc.2025.113267.
[30] H. Bhin and J. Choi, "Multimodal Personality Recognition Using Self-Attention-Based Fusion of Audio, Visual, and Text Features," *Electronics*, vol. 14, no. 14, pp. 1–20, Jul. 2025, doi: 10.3390/electronics14142837.
[31] J. Kolenbrander and A. J. Michaels, "Personality Emulation Utilizing Large Language Models," *Applied Sciences*, vol. 15, no. 12, pp. 1–24, Jun. 2025, doi: 10.3390/app15126636.
[32] R. Singh *et al.*, "Antisocial Behavior Identification from Twitter Feeds Using Traditional Machine Learning Algorithms and Deep Learning," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 10, no. 4, pp. 1–17, May 2023, doi: 10.4108/eetsis.v10i3.3184.
[33] W. Maharani and V. Effendy, "Big five personality prediction based in Indonesian tweets using machine learning methods," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, pp. 1973–1981, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1973-1981.
[34] G. Jiang, W. Han, M. Xu, C. Zhang, S. C. Zhu, and Y. Zhu, "Evaluating and Inducing Personality in Pre-trained Language Models," in *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, vol. 36, pp. 10622–10643, 2023, doi: 10.5555/3666122.3666588.
[35] A. Naz, H. U. Khan, S. Alesawi, O. I. Abouola, A. Daud, and M. Ramzan, "AI Knows You: Deep Learning Model for Prediction of Extroversion Personality Trait," *IEEE Access*, vol. 12, pp. 159152–159175, 2024, doi: 10.1109/ACCESS.2024.3486578.
[36] A. Koutsoumpis *et al.*, "Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning," *Computers in Human Behavior*, vol. 154, pp. 1–18, May 2024, doi: 10.1016/j.chb.2023.108128.
[37] M. Thahira and A. K. Mubeena, "Comparative study of personality prediction from social media by using machine learning and deep learning method," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 7, 2021, doi: 10.17577/IJERTCONV9IS07024.

BIOGRAPHIES OF AUTHORS

Angad Singh is Ph.D. in Computer Science and Engineering from Rabindranath Tagore University, Bhopal, India. He completed his M.Tech. from Shri Ram Institute of Technology, Jabalpur, India in 2011. He has more than 19 years of teaching experience in reputed institutions like Oriental Group of Institutes, Bhopal. His area of interest is machine learning and deep learning. He can be contacted at email: angada2007@gmail.com.

Priti Maheshwary is Professor in Computer Science and Engineering Department at Rabindranath Tagore University, Bhopal. She had her Doctorate from MANIT, Bhopal in Remote Sensing Image Retrieval. She has completed research project on Climate Change detection and monitoring funded by SAC and Environment Monitoring using Sensor devices funded by AISECT University. Ongoing Project is on Crop Monitoring sponsored by SAC. She is the author of more than 20 publications in different journals of repute. She is the author of book chapter on Software Copyright. Her interests include machine learning, cyber physical systems, mobile networks, WSN, ad hoc networks, data mining, and image processing. Her work experience includes 20 years in teaching computer science and engineering. She is guiding Ph.D. in the field of machine learning, IoT, and networks. She can be contacted at email: pritimaheshwary@gmail.com.

Nitin Kumar Mishra is a Senior Assistant Professor at the School of Computing Science Engineering and Technology, Bennett University, Greater Noida, India. He has a Doctorate in Information Technology from Atal Bihari Vajpayee-Indian Institute of Information Technology and Management (*ABV-IIITM*), *Gwalior*. He has 21 years of academic experience. As a researcher, designing, developing, and optimizing machine learning algorithms for multi-label classification problems is his area. He has presented and published many research papers in SCIE/Scopus-indexed international conferences and journals. He reviews various journals such as Information Sciences, Information Processing & Management, Computers & Electrical Engineering, Expert System with Applications, and Artificial Intelligence Review. He can be contacted at email: nkmishra0701@gmail.com.

Neerja Dubey is an Assistant Professor in Computer Science and Engineering Department at LNCTS, Bhopal, India. She has M.Tech. in Computer Science and Engineering from Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal. She has 8 years of academic experience and her area of interest is artificial intelligence and networking. She can be contacted at email: neerja.dubey2010@gmail.com.