iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71767

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Survey on Personality Detection on Multilingual Dataset using Machine Learning and Explainable AI

Riya Bhaskar[1], Arunima Jaiswal[2], Harshita Bhakhand[3], Divya Goyal[4], Shreya Ramhans[5]
Department of Computer Science and Engineering, IGDTUW, Delhi

Abstract: ***Personality detection is a key area in AI, significantly impacting psychological profiling, behavioral analysis, and personalized recommendations. It addresses the new frontiers of multilingual personality detection using machine learning (ML) and Explainable AI (XAI). In this review paper, we studied the state-of-art research on deep learning and ensemble methods, while also highlighting the contributions of XAI frameworks like SHAP and LIME in improving model interpretation. We analyzed based metrics on use cases with fair and transparent measures to assess how ML methods generalize across languages and cultural settings through the survey of various multilingual dataset.***

Our findings highlight the significance of XAI in bridging AI predictions with human understanding, making AI-driven personality detection more interpretable and ethically responsible. This review contributes to personality computing by synthesizing advancements in ML-based personality detection, discussing challenges, and identifying future research directions to develop fair, accurate, and explainable AI-driven psychological assessments across languages.

Keywords: Personality detection, machine learning, multilinguality, explainable AI, Myers-Briggs Type Indicator, OCEAN

I. INTRODUCTION

Personality detection is an essential aspect of artificial intelligence and human-computer interaction, with applications ranging from psychological assessments to personalized recommendations and recruitment systems. The ability to accurately predict personality traits from textual data has gained significant attention in recent years, fueled by advancements in machine learning[1] and Explainable Artificial Intelligence [2]. While many existing approaches focus on monolingual datasets, real-world applications require models that can generalize across multiple languages and cultural contexts.

This research explores personality detection on multilingual datasets using machine learning and XAI techniques. By leveraging datasets from diverse linguistic backgrounds, we aim to develop robust models capable of predicting personality traits across different languages while ensuring transparency and interpretability. Traditional black-box machine learning models, despite their high accuracy, often lack explainability, making it challenging to trust and interpret their predictions. To address this limitation, we integrate XAI methods to enhance model transparency, enabling a deeper understanding of the relationship between linguistic features and personality traits.

Our study examines various machine learning algorithms, including deep learning models,[3] ensemble methods, and transformer-based architectures, in combination with XAI frameworks such as SHAP[4] (SHapley Additive Explanations) and LIME[2] (Local Interpretable Model-agnostic Explanations). We evaluate these models on multiple multilingual datasets to assess their generalizability and interpretability. The findings of this research contribute to the growing field of personality computing by providing insights into the effectiveness of ML models across languages and offering interpretable solutions that bridge the gap between AI predictions and human understanding.

Through this research, we aim to advance the state-of-the-art in multilingual personality detection while promoting the use of transparent AI models in psychological and behavioral studies. Our approach ensures that AI-driven personality assessments are not only accurate but also interpretable and ethically responsible.

II. LITERATUREREVIEW

Over a period of time, a wide range of methods have been developed for detecting personality traits using machine learning and deep learning models. This section reviews the key research papers over the years, emphasizing the various technology stacks, models, datasets, and accuracy metrics of each approach.

The following table presents a comprehensive overview of these studies, summarizing their contributions to the field of personality detection while also outlining their challenges and limitations.

Table 1: Key insights from the prominent papers on Personality Detection

Author	Year	Technique	Dataset	Accuracy
Fei Liu et al.[5]	2016	Character-to-word-to-sentence hierarchy	PAN 2015 (Twitter data)	RMSE-based metric
Sun et al.[6]	2018	Bi-LSTM + CNN (2CLSTM), Latent Sentence Group (LSG)	YouTube Comments & Essays Dataset	YouTube: 55.79%, Essays: 60.72%
Khwaja et al.[7]	2019	Machine Learning on Mobile Sensor Data	Mobile sensor data (5 countries)	63%-71%
Siddique et al.[8]	2019	Multilingual embeddings, CNN	PAN 2015 (Twitter data)	F-score: 65-73.4
Leonardi et al.[9]	2020	Transformer-based sentence embeddings	myPersonality (Facebook)	State-of-the-art
Sasidhara et al.[10]	2020	CNN-BiLSTM, Word2Vec	Hinglish Twitter Dataset	83.21%
Khan et al.[11]	2020	XGBoost, KNN, Decision Tree, Random Forest, SVM, MLP	Kaggle MBTI Personality Dataset	99.92%
Hira Shafi et al.[12]	2021	Supervised Machine Learning	MBTI Typology Dataset (2802 instances)	98.4%,
Deilami et al.[13]	2021	CNN, AdaBoost, KNN, Logistic Regression	Essays Dataset (Big Five)	86.2%
Christian et al.[14]	2021	BERT, RoBERTa, XLNet, Model Averaging	Facebook (MyPersonality), Twitter Dataset	Facebook: 86.17%, Twitter: 88.5%
Yanou Ramon et al.[15]	2021	Random Forest, Logistic Regression, Explainable AI	Financial Transaction Records Dataset	Random Forest: 63.98%, Logistic Regression: 58.14%
Savant et al.[16]	2022	Ensemble Learning (KNN, Logistic Regression, Deep Forest, MLP, SVM)	Questionnaire-Based Dataset	Ensemble Classifier: 93.1%
Kerz et al.[17]	2022	Hybrid BERT + BLSTM with Psycholinguistic Feat...	Big Five Essay, MBTI Kaggle dataset	63.5% (Big Five), 77.1% (MBTI)
Chincholkar et al.[18]	2023	Multinomial Logistic Regression, NLP	Various sources (Google Forms, interviews)	Not explicitly mentioned
Tarale et al[19].	2023	SVM, XGBoost, Decision Tree, Psycholinguistic Features, NLP	Kaggle MBTI dataset	SVM: 90.3%, XGBoost: 88.33%
Alsubhi et al.[20]	2023	Logistic Regression, SVM	AraBig5 (Arabic tweets)	67%
Gupta et al.[21]	2023	KNN, CNN, Logistic Regression, Phrase Frequency Algorithm	Recruitment Dataset	Accuracy not provided
Dandash et al.[22]	2024	BERT, Feature Engineering	AraPers (Arabic Twitter users)	74.86% (BERT)

Mehta et al.[23]	2024	Psycholinguistic features + Language Models	Essays dataset, Kaggle MBTI dataset	60.6% (Big 5), 75.9% (MBTI)
Max Murphy[24]	2024	Large Language Models (GPT-3.5, GPT-4)	PersonalityCafe MBTI dataset	73% (GPT-3.5), 76% (GPT-4)
Chraibi et al.[25]	2025	Feature Selection (Filter, Wrapper, Embedded Methods), Regression	MSAPersonality (Modern Standard Arabic texts)	77.34%
Saeteros et al.[26]	2025	BERT, Explainable AI, Logistic Regression	Essays Dataset (Pennebaker & King, 1999)	82.8%

Research Gaps

Several important research challenges persist in the development of personality detection technology that relies on Machine Learning and Artificial Intelligence. Current research mainly investigates one-language detection which creates barriers for model usability across different language-speaking contexts. Large-scale benchmarking together with generalization is limited due to the insufficient availability of well-annotated datasets. Current deep learning systems require better interpretability since their performance falls short of practical needs so alternate explainable AI methods must be developed. The applications used in workplaces alongside healthcare organizations and social media networks impose various constraints which limit the ability to detect personalities across numerous settings.

Research needs to explore how different personality traits extracted from multiple input sources should be merged into a single identity assessment. The need for ethical guidance particularly emphasizes the importance of developing fairness frameworks and regulatory criteria because training datasets create various ethical problems. These models require extensive computational power which limits their practical application for real-time use when implemented in resource-limited areas. The development of more accurate and responsible personality detection systems requires proper solutions for the obstacles encountered.

III. METHODOLOGY

This research analyzed different methods to detect personality through machine learning techniques and deep learning methods and explainable AI strategies. A detailed description of methods, datasets together with systematic research procedures follows in this part. Our study follows the methodology as shown through this flowchart.

A. Flowchart of Method Architecture

The general architecture of machine learning models for personality detection typically follows a pipeline consisting of multiple stages:

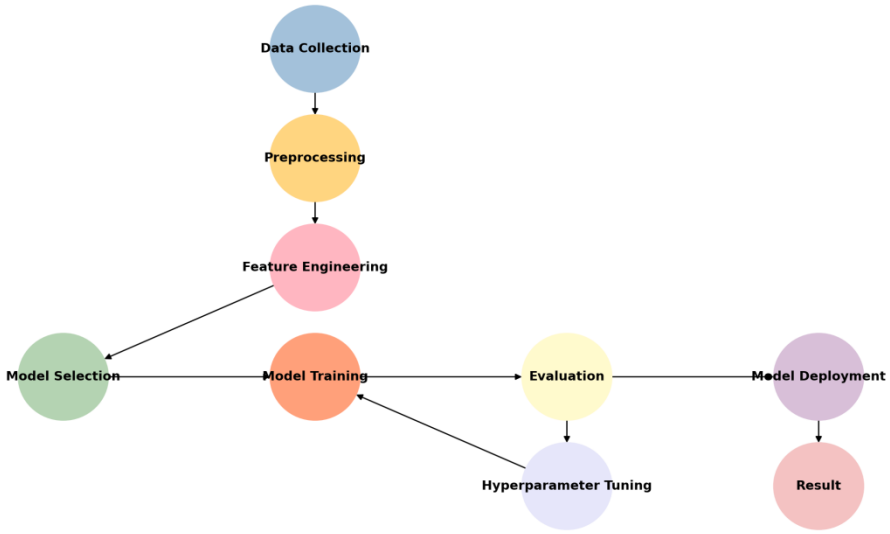


Figure 1: Flowchart on Method Architecture

An organized procedure for personality detection uses multiple stages to create an effective and precise analysis process.

1) Information for analysis originated from a range of different sources that included both linguistic content and social media data and image records and user conduct. The study employed MyPersonality alongside the Big Five Personality Dataset together with social media-based corpora from publicly accessible sources for complete research analysis.

2) Multiple pre-processing actions were applied to the gathered data including clearing noise, turning text into tokens, reducing words to their root forms and dealing with missing information. The processing of structured datasets involved the application of normalization approaches to normalize their data values.

3) The model received increased performance through the implementation of feature extraction techniques during the engineering process. The analysis included TF-IDF alongside Word2Vec and GloVe and BERT and LIWC for linguistic examination of textual content. The image-based personality detection process required feature maps that were generated through convolutional neural networks (CNNs).

4) Model selection occurred through the assessment of the dataset features alongside prediction requirements. The system contains the following classification of models:

- Classifications utilized the supervised learning models consisting of Logistic Regression along with Support Vector Machines (SVM) and Random Forest and Neural Networks.
- The process of Unsupervised Learning used clustering methods including K-Means along with hierarchical clustering for pattern observation.

4.3. The analysis of images for personality detection used Convolutional Neural Networks (CNN) together with Recurrent Neural Networks (RNN) and transformers including BERT for processing sequential text information.

5) Training models took place through the use of labelled data collections. Training prepared loss function definitions followed by gradient descent optimization of parameters together with backpropagation usage for deep learning blocks.

6) Performance evaluation relied on accuracy metrics together with precision and recall along with the F1-score and ROC-AUC metrics. The evaluation used cross-validation (k-fold method) to guarantee performance on data which the algorithm had not learned before.

7) The model performance improved through optimization methods such as Grid Search and Bayesian Optimization during hyperparameter tuning selection process. The prevention of overfitting was achieved through regularizing techniques which included dropout as well as batch normalization.

8) Deployment of the model followed after reaching an optimal performance standard. A continuous monitoring system joined with retraining procedures acted to sustain accuracy throughout time.

B. Data collection

In personality detection it is very important to select the right dataset to ensure improved accuracy of the models. Below is a detailed summary of commonly used benchmark datasets, highlighting their characteristics, significance, and the studies that have leveraged them effectively:

Table : Summary of Datasets Used for Personality Detection

Dataset Name	Description
Bangla Personality Traits Dataset	This dataset consists of 3,000 Bangla informal texts from Facebook, YouTube, and blogs
Myers-Briggs Type Indicator (MBTI) dataset	This dataset is available from Kaggle, and focuses on Myers-Briggs Type Indicator (MBTI) personality classifications. It consists of the big five personality.It consists of 106,067 rows of social media posts
Google Forms Questionnaire Dataset	972 survey responses for Big Five Personality Traits
Financial Transaction Records Dataset	6,408 individuals, 4.5M spending records

Sentiment140 Dataset	This dataset consists of 160,000 pre-processed tweets
Twitter US Airline Sentiment Data	Twitter-based dataset for sentiment classification of US airlines.
GoEmotions dataset	This is a large dataset created by google , annotated with 27 different emotions categories , including anxiety , fear etc.
Pennebaker & King's Essay Dataset	This dataset consists of 2,467 user-written essays for Big Five personality traits
Twitter data andMyPersonality from facebook.	It consists of data from twitter and facebook
Chinese Social Media Personality Dataset	This dataset contains Weibo and PenPen social networks related data
Facial Image Dataset	Used to extract facial features for personality classification
Essays Dataset	This dataset consists of 2,240 undergraduate essays analyzed for personality traits and sex differences
Text-based Personality Dataset	This dataset consists of data from social media posts and digital texts
PersonalityEvd Dataset	A dialogue-based dataset for explainable personality recognition
Healthcare Personality Dataset	This dataset consists of Patient behavioral and psychological data
Data from Facebook and LinkedIn posts.	This dataset consists of data from facebook and linkedin posts.
AraBig5	10,000 Arabic tweets collected from users who took the Arabic personality test and shared their results on Twitter
AraPers	3,250 Twitter users who shared their 16personalities test results and tweet data
MSAPersonality	Modern Standard Arabic (MSA) texts labeled for Big Five personality traits
Hinglish Twitter Dataset	This dataset consists of 12,000 tweets annotated for emotions (Happy, Sad, Anger)
Recruitment Dataset	This is a synthetic or real-world small-scale hiring dataset
PAN 2015(Twitter data)	Twitter data based Dataset
Personality Café MBTI	This dataset consists of social media posts of the users to one of the 16 personality types

C. Techniques used

For personality detection , machine learning and deep learning techniques are applied. These can be broadly classified into traditional machine learning methods, deep learning methods, and hybrid approaches.

Table 3: Techniques used

Technique Name	Description
Logistic Regression	It is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome or event.
XGBoost	XGBoost, short for extreme Gradient Boosting, is an advanced machine learning algorithm designed for efficiency, speed, and high performance.
CNN (Convolutional Neural Network)	CNNs are utilized for automated feature extraction from textual and image data, hence their use is increasing to extract n-gram-like features.
RNN (Recurrent Neural Networks)	RNNs are frequently applied to long form texts like essays, speeches, or conversational data along with LSTM and Bi-RNN.
Bidirectional Encoder Representations from Transformers (BERT)	It is an emerging pre- trained model for extracting word embeddings from text. It indicates bi-directional relationship of words hence making it useful for sentiment analysis, summarizing text and classifying personality traits.
GloVe/Word2Vec	These are word embedding techniques, used to convert words into vector form, which helps to analyze their semantic and syntactic similarity with other words. Helps in clustering, classification and finding similarities in textual data.
Linguistic Inquiry and Word Count (LIWC)	It is used for personality prediction through evaluation of frequency of words. It is combined with SVM or CNN to improve its accuracy to measure psychological traits.
MLP (Multilayer Perceptron)	MLP is a feed-forward neural network, i.e. the information moves forward in a single direction through various layers of nodes. It is used as a classifier after feature extraction from word embeddings.
Bagging Ensemble Methods	In this, the classifiers like SVM are trained on various data subsets and the predictions are then combined, which enhances the stability & accuracy of models by reducing variance. Useful where datasets are imbalanced. Some other ensemble methods are - XGBoost, Random Forest, KBSVE-P
K-Nearest Neighbors (KNN)	It is a simple approach usi in multimodal classification. In personality prediction, KNN helps to segregate individuals on the basis of their proximity to known traits, like word frequency, LIWC features and other features from face recognition. It is also used in Transfer learning.
Deep Belief Networks (DBN)	DBNs are used for unsupervised feature learning processes and reducing dimensionality. It helps in learning high - level data from text and images which is further used for classifications. Hence used to learn to extract hierarchical representation of various types of data.
Naïve Bayes	This technique is commonly used in papers for text based classifications. Highly effective in personality and sentiment prediction from short texts like tweets or messages. .
Active Shape Model (ASM)	It is used in biometrics, facial data recognition, and medical image processing. It learns from different sets of shapes and then fits them into unseen images.

Support Vector Machine(SVM)	It is very frequently used to predict personality traits based on frequency of words, LIWC features and other linguistic data. It is highly efficient in handling high-dimensional data
AdaBoost	Adaptive Boosting is a machine learning technique that combines multiple weak classifiers to create a strong classifier
XLNet	It is an extension of the Transformer-XL model which use an autoregressive method to learn bidirectional contexts
RoBERTa	It is an extension of the BERT model with changes in pretraining procedure.
Explainable AI	Explainable artificial intelligence (XAI) refers to a collection of procedures and techniques that enable machine learning algorithms to produce output and results that are understandable and reliable for human users

IV. TRENDS AND ANALYSIS

The personality detection domain has experienced significant development due to improvements in machine learning and deep learning technologies, alongside multilingual dataset expansion. Research in this field mainly depended on questionnaire-based English datasets such as the Big Five Inventory (BFI) and MBTI until the initial stage. The expansion of research scopes required scientists to add multilingual datasets because this step enhanced model generalization between different linguistic and cultural domains. The PAN 2015 dataset introduced multilingual Twitter data and made it possible to analyze personality across different languages. The expansion of personality detection across multiple languages through datasets such as Kaggle MBTI, recruitment-based corpora and new Arabic-language collections has occurred with additional data collections.

Studies at their early stages used classic machine learning algorithms including Logistic Regression, Decision Trees, Support Vector Machines (SVM) and Naïve Bayes for their research. The employed models succeeded effectively with prepared data input but failed when processing the sophisticated nature of written language. In 2016 scientists started to investigate personality detection through social media texts while deploying logistic regression algorithms together with stochastic gradient descent. These approaches needed manual feature choice thus making their applications restricted to small-scale implementations. Research moves ahead by using Twitter Facebook and YouTube platforms to study real-world personality characteristics during the time period from 2016 through 2018.

Deep learning models consisting of Convolutional Neural Networks (CNNs) and Bi-LSTMs together with ensemble models raised the research accuracy levels across 2018 to 2021. A 2CLSTM model was developed in 2018 to successfully extract text features because it included CNN and Bi-LSTM components. Research scientists between 2022 and 2024 created transformer models including BERT and Sentence-BERT which became dominant because of their enhanced accuracy and operational effectiveness. Research-based investigation merged multitask learning functions with attention mechanisms operating at different levels to enhance the capacity to recognize personalities across texts from various languages.

Time generated changes to the ways features were engineered. The methodologies used by models evolved from traditional text processing to word embedding technologies that incorporate Word2Vec, GLoVe, FastText as well as BERT-based embeddings. Automatic learning of high-dimensional representations now replaces the need for manual feature selection through the advances in newer methods. Explainable AI improvements introduced Integrated Gradients and rule-based feature attribution to explain personality classification systems while protecting human rights.

The field of personality detection investigates fresh research areas that combine financial transaction analysis and studies of facial expressions and video content evaluation through analytical techniques. A scientific research team investigated personality traits by uniting Active Shape Model (ASM) with Deep Belief Networks (DBN) methods for facial characteristic analysis in 2021. A study used Explainable AI models on financial transactions to recognize personality trends defined by customer buying patterns. AI technologies designed to evaluate personality traits form a significant component of recruiting platforms that exist in current corporate applications.

The 2023 Smart-Hire system utilized a ranking process for candidate assessment by predicting personalities through an integration of CNNs and KNN models and logistic regression algorithms. Researchers studied personality traits in Arabic language tweets by using BERT in conjunction with regression-based models throughout 2024 to 2025.

Such categorization systems have experienced steady advancement in their precision rates from the very beginning. Accuracy levels increased significantly when CNNs and Bi-LSTMs were introduced because accuracy rates reached over 80% levels from their initial 50% to 70% rates. BERT transformers achieve benchmark performance between 85-90% and the combination of AdaBoost extended through CNN achieves better predictive accuracy.

A. Efficacy Parameters

In the research area, efficacy parameters are factors used to measure effectiveness of a process which helps us to understand how well the new data is generalized. Table 4 gives an overview of all the efficacy parameters that have been found in this study.

Table 4: Efficacy parameters

Parameter	Description	Significance
Accuracy[56]	Correct predictions out of total predictions	Measures overall performance of model.
Precision[56]	Positive predictions that are actually positive.	Evaluates the model's ability to avoid false positives.
Recall[56]	Actual positive cases that are correctly predicted as positive.	Assesses the ability of model to predict all positives.
F1-score[56]	F1 Score becomes 1 only when precision and recall are both 1	Provides a balanced measure of both precision and recall.
Receiver Operating Characteristic curve[57]	A plot that illustrates the trade-off between true positive rate and false positive rate.	Visualizes the model's performance across different thresholds.
Area under the curve (AUC) [57]	The area under the ROC curve.	Represents the model's overall performance, regardless of threshold.
Computational Efficiency[58]	The time and resources required to train and run the model.	Indicates the model's feasibility for practical applications.

V. CHALLENGES

The advancement of personality detection has occurred significantly though multiple substantial obstacles prevent it from wider implementation and precise measurements. The difficulties in personality detection emerge from inadequate psychological characteristics coverage combined with ambiguous data and limited datasets and diverse linguistic nature of text. Proper research combined with enhanced methodologies and diverse datasets together with advanced models are needed to handle these issues.

1) Focus On Only Big Five Traits (OCEAN): Current research focuses on the Big Five traits (OCEAN) ignoring some traits such as anxiety or confidence which reduces the application of personality detection systems. To solve this issue we can expand research to include more psychological traits by adding new behavioral datasets.

2) Ambiguity Of Input Data: Current Personality detection models struggle with texts that exhibit mixed or ambiguous emotions. Handling such emotional expressions remains a challenge, requiring enhanced research and study in it. Use of techniques like multi-label classification to handle emotional ambiguity can help solving this challenge.

3) Data Limitations: Many models are trained on small, biased, or domain-specific datasets, which limits their applicability to real-world scenarios . We can Collaborate with various research groups to create datasets that include diverse languages, and cultures and hence solve the challenge of data limitation.

4) Multilingual Complexity:The process of understanding personality traits through different languages becomes complex because language and culture affect how people communicate within them. Professional language in various expressions and phrases as well as sentence construction methods differ significantly from one language to another which leads to inconsistent translation results. Research on personae understanding requires models which translate words together with their conceptual meaning more than word-to-word. Advanced NLP techniques along with cross-lingual embeddings that use diverse multilingual datasets enable the improvement of personality detection between multiple languages.

VI. CONCLUSION

Personality detection is the process of identifying individual personality traits and has gained significant attention. The goal of this survey paper is to provide an overview of recent advancements in personality detection methods.

In this survey, we reviewed different research studies on personality detection. The Big five personality traits are the most widely used framework for evaluating personality. We studied different deep learning approaches such as CNN, LSTM models, and K-S sets, comparing their performance across different datasets. A detailed analysis of datasets and their sources is also provided to give a fully understanding of the data used in personality detection.

While the survey provides a broad overview, it lacks in depth analysis of specific methods such as LSTM models. We also recognized the need to include more research on personality detection. To improve this further iterations should consider incorporating additional personality detection methods. Expanding the scope to include approaches beyond deep learning and including hybrid methods could offer a more complete perspective on the current state of personality detection

VII.FUTURE SCOPE

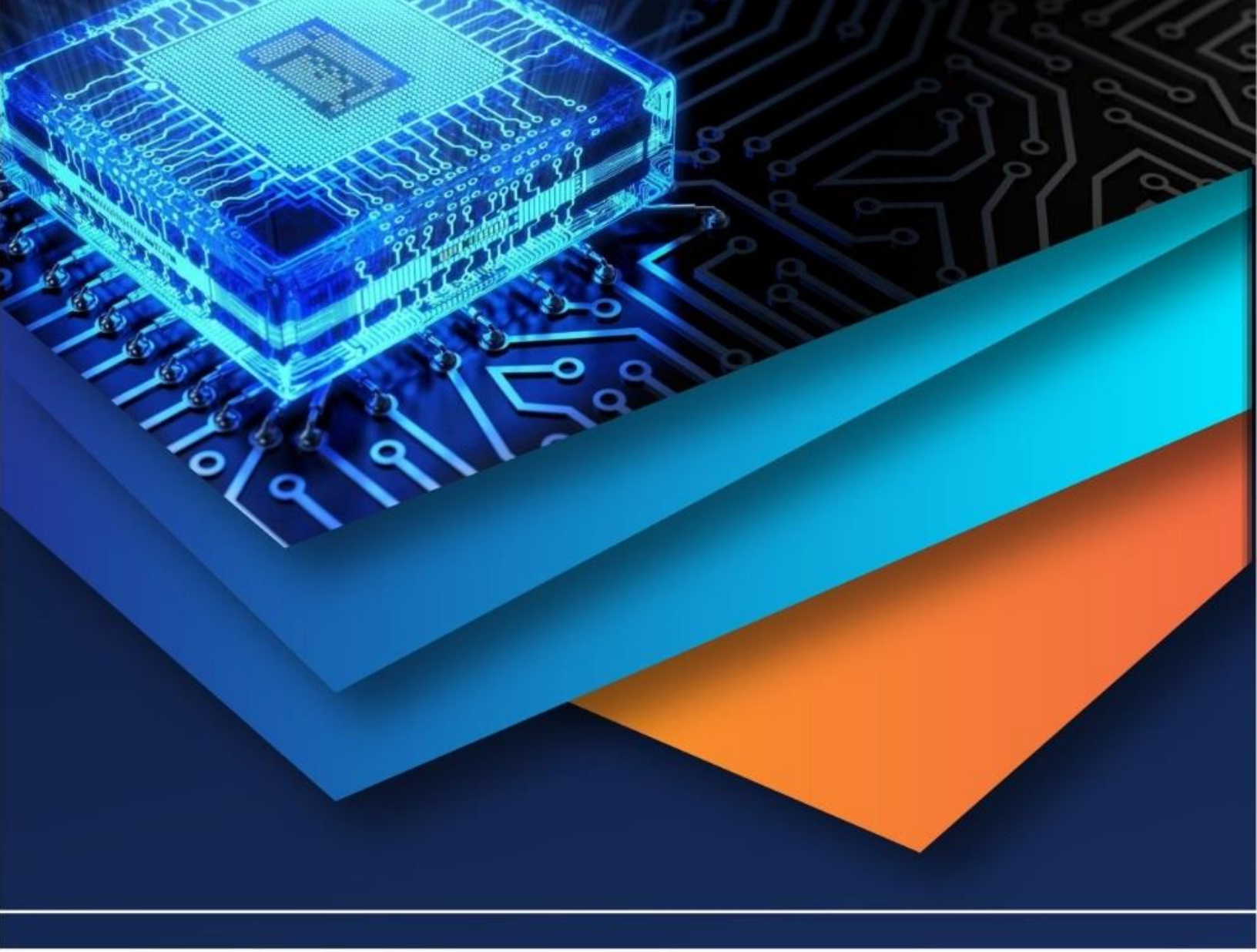
Future research on personality detection needs to expand its focus by adding specific traits such as anxiety and confidence because such additions will make the approach more realistic. Thriving personality detection systems depend on handling the required resolution of ambiguous emotional content in written texts. Contextual analysis presents a possible solution for handling complex emotional situations within personality detection. The performance of multitask learning improves efficiency through simultaneous personality trait predictions. The performance of the system can be enhanced through model optimization that integrates Random Forest along with deep learning and explainable AI methods. A combination of text with audio elements and video data within a detection system will provide better accuracy to personality recognition analyses. The use of diverse language data in datasets helps improve both generalization capabilities and fairness within multicultural contexts of machine learning models.

REFERENCES

[1] “Machine learning,” Wikipedia. Feb. 12, 2025. Accessed: Feb. 16, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=1275382050
[2] “Explainable AI(XAI) Using LIME,” GeeksforGeeks. Accessed: Dec. 03, 2024. [Online]. Available: https://www.geeksforgeeks.org/introduction-to-explainable-aixai-using-lime/
[3] “Introduction to Deep Learning - GeeksforGeeks.” Accessed: Feb. 16, 2025. [Online]. Available: https://www.geeksforgeeks.org/introduction-deep-learning/
[4] “Welcome to the SHAP documentation — SHAP latest documentation.” Accessed: Feb. 05, 2025. [Online]. Available: https://shap.readthedocs.io/en/latest/
[5] F. Liu, J. Perez, and S. Nowson, “A Language-independent and Compositional Model for Personality Trait Recognition from Short Texts,” Oct. 14, 2016, arXiv: arXiv:1610.04345. doi: 10.48550/arXiv.1610.04345.
[6] X. Sun, B. Liu, J. Cao, J. Luo, and X. Shen, “Who Am I? Personality Detection Based on Deep Learning for Texts,” in 2018 IEEE International Conference on Communications (ICC), Kansas City, MO: IEEE, May 2018, pp. 1–6. doi: 10.1109/ICC.2018.8422105.
[7] M. Khwaja, S. S. Vaid, S. Zannone, G. M. Harari, A. A. Faisal, and A. Matic, “Modeling Personality vs. Modeling Personalidad: In-the-wild Mobile Data Analysis in Five Countries Suggests Cultural Impact on Personality Models,” Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 3, no. 3, pp. 1–24, Sep. 2019, doi: 10.1145/3351246.
[8] F. B. Siddique, D. Bertero, and P. Fung, “GlobalTrait: Personality Alignment of Multilingual Word Embeddings,” Nov. 20, 2018, arXiv: arXiv:1811.00240. doi: 10.48550/arXiv.1811.00240.
[9] S. Leonardi, D. Monti, G. Rizzo, and M. Morisio, “Multilingual Transformer-Based Personality Traits Estimation,” Information, vol. 11, no. 4, p. 179, Mar. 2020, doi: 10.3390/info11040179.
[10] T. T. Sasidhar, P. B, and S. K. P, “Emotion Detection in Hinglish(Hindi+English) Code-Mixed Social Media Text,” Procedia Computer Science, vol. 171, pp. 1346–1352, 2020, doi: 10.1016/j.procs.2020.04.144.
[11] A. S. Khan, H. Ahmad, M. Zubair, F. Khan, A. Arif, and H. Ali, “Personality Classification from Online Text using Machine Learning Approach,” IJACSA, vol. 11, no. 3, 2020, doi: 10.14569/IJACSA.2020.0110358.
[12] Hamdard University Karachi, Pakistan et al., “A Machine Learning Approach for Personality Type Identification using MBTI Framework,” JISR-C, vol. 19, no. 2, 2021, doi: 10.31645/JISRC.43.19.2.2.
[13] F. M. Deilami, H. Sadr, and M. Nazari, “Using Machine Learning-Based Models for Personality Recognition”.

[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," J Big Data, vol. 8, no. 1, p. 68, Dec. 2021, doi: 10.1186/s40537-021-00459-1.
[15] Y. Ramon, R. A. Farrokhnia, S. C. Matz, and D. Martens, "Explainable AI for Psychological Profiling from Behavioral Data: An Application to Big Five Personality Predictions from Financial Transaction Records," Information, vol. 12, no. 12, p. 518, Dec. 2021, doi: 10.3390/info12120518.
[16] G. R. Savant, "Personality Classification with Data Mining," vol. 7, no. 5, 2022.
[17] E. Kerz, Y. Qiao, S. Zanwar, and D. Wiechmann, "Pushing on Personality Detection from Verbal Behavior: A Transformer Meets Text Contours of Psycholinguistic Features," Apr. 10, 2022, arXiv: arXiv:2204.04629. doi: 10.48550/arXiv.2204.04629.
[18] Prof. A. Chincholkar, D. Bhosale, S. Adsul, A. Bodkhe, and R. Kadam, "A Comprehensive Survey on Personality Prediction Using Machine Learning Techniques," INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER AND COMMUNICATION ENGINEERING, vol. 12, no. 11, Nov. 2023, doi: 10.17148/IJARCCE.2023.121120.
[19] S. Garg and A. Garg, "Comparison of machine learning algorithms for content based personality resolution of tweets," Social Sciences & Humanities Open, vol. 4, no. 1, p. 100178, 2021, doi: 10.1016/j.ssaho.2021.100178.
[20] S. M. Alsubhi, A. M. Alhothali, and A. A. AlMansour, "AraBig5: The Big Five Personality Traits Prediction Using Machine Learning Algorithm on Arabic Tweets," IEEE Access, vol. 11, pp. 112526–112534, 2023, doi: 10.1109/ACCESS.2023.3297981.
[21] S. Gupta, J. Hingorani, S. Singh, and N. Phadnis, "DESIGNING OF WEB PORTAL FOR TRAINING AND PLACEMENT CELL," vol. 08, no. 05, 2021.
[22] M. Dandash and M. Asadpour, "Personality Analysis for Social Media Users using Arabic language and its Effect on Sentiment Analysis".
[23] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," Artif Intell Rev, vol. 53, no. 4, pp. 2313–2339, Apr. 2020, doi: 10.1007/s10462-019-09770-z.
[24] M. Murphy, "Artificial Intelligence and Personality: Large Language Models' Ability to Predict Personality Type," Emerging Media, p. 27523543241257291, Jun. 2024, doi: 10.1177/27523543241257291.
[25] K. Chraibi, I. Chaker, and A. Zahi, "Predicting personality traits from Arabic text: an investigation of textual and demographic features with feature selection analysis," IJECE, vol. 15, no. 1, p. 970, Feb. 2025, doi: 10.11591/ijece.v15i1.pp970-979.
[26] D. Saeteros, B. Domínguez-Álvarez, D. Gallardo-Pujol, and D. Ortiz-Martínez, "The Written Self: Decoding Personality and Sex Differences Through Explainable AI," Jan. 10, 2025, PsyArXiv. doi: 10.31234/osf.io/eja7r.
[27] U. Rudra, A. N. Chy, and Md. H. Seddiqui, "Personality Traits Detection in Bangla: A Benchmark Dataset with Comparative Performance Analysis of State-of-the-Art Methods," in 2020 23rd International Conference on Computer and Information Technology (ICCIT), DHAKA, Bangladesh: IEEE, Dec. 2020, pp. 1–6. doi: 10.1109/ICCIT51783.2020.9392722.
[28] "Myers–Briggs Type Indicator," Wikipedia. Jan. 31, 2025. Accessed: Feb. 05, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Myers%E2%80%93Briggs_Type_Indicator&oldid=1273106391
[29] "Financial Transactions Dataset." Accessed: Feb. 16, 2025. [Online]. Available: https://www.kaggle.com/datasets/cankatsrc/financial-transactions-dataset
[30] A. Kazemeini, S. S. Roy, R. E. Mercer, and E. Cambria, "Interpretable Representation Learning for Personality Detection," in 2021 International Conference on Data Mining Workshops (ICDMW), Auckland, New Zealand: IEEE, Dec. 2021, pp. 158–165. doi: 10.1109/ICDMW53433.2021.00026.
[31] "Twitter US Airline Sentiment." Accessed: Feb. 16, 2025. [Online]. Available: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment
[32] "GoEmotions: A Dataset for Fine-Grained Emotion Classification." Accessed: Feb. 16, 2025. [Online]. Available: https://research.google/blog/goemotions-a-dataset-for-fine-grained-emotion-classification/
[33] Y. Chai, D. Kakkar, J. Palacios, and S. Zheng, "Twitter Sentiment Geographical Index Dataset," Sci Data, vol. 10, p. 684, Oct. 2023, doi: 10.1038/s41597-023-02572-7.
[34] [M. H. Yimer, Y. Yu, K. Adu, E. Favour, S. M. Liyih, and R. A. Patamia, "Music Genre Classification using Deep Neural Networks," in 2023 35th Chinese Control and Decision Conference (CCDC), May 2023, pp. 2384–2391. doi: 10.1109/CCDC58219.2023.10327367.
[35] "essay." Accessed: Oct. 08, 2024. [Online]. Available: https://paperswithcode.com/dataset/asap
[36] "Logistic regression - Wikipedia." Accessed: Nov. 25, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Logistic_regression
[37] "xgBoost." Accessed: Oct. 09, 2024. [Online]. Available: https://xgboost.readthedocs.io/en/latest/
[38] "CNN vs. RNN: How are they different? | TechTarget." Accessed: Nov. 25, 2024. [Online]. Available: https://www.techtarget.com/searchenterpriseai/feature/CNN-vs-RNN-How-they-differ-and-where-they-overlap
[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, arXiv: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.
[40] "glove." Accessed: Oct. 08, 2024. [Online]. Available: https://nlp.stanford.edu/projects/glove/
[41] "liwc." Accessed: Oct. 08, 2024. [Online]. Available: https://www.liwc.app/
[42] "Multilayer perceptron," Wikipedia. Dec. 29, 2024. Accessed: Feb. 16, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Multilayer_perceptron&oldid=1265916526
[43] "Bootstrap aggregating," Wikipedia. Dec. 27, 2024. Accessed: Feb. 16, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Bootstrap_aggregating&oldid=1265529013
[44] "k-nearest neighbors algorithm," Wikipedia. Feb. 05, 2025. Accessed: Feb. 16, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&oldid=1274106583
[45] "Deep belief network," Wikipedia. Aug. 13, 2024. Accessed: Feb. 16, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Deep_belief_network&oldid=1240122786
[46] "Naive Bayes Classifiers - GeeksforGeeks." Accessed: Nov. 25, 2024. [Online]. Available: https://www.geeksforgeeks.org/naive-bayes-classifiers/
[47] "Active shape model," Wikipedia. Oct. 05, 2023. Accessed: Feb. 16, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Active_shape_model&oldid=1178707117
[48] "Support Vector Machine (SVM) Algorithm - GeeksforGeeks." Accessed: Nov. 25, 2024. [Online]. Available: https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[49] "AdaBoost," Wikipedia. Nov. 23, 2024. Accessed: Feb. 16, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=AdaBoost&oldid=1259173406
[50] "Big Five personality traits," Wikipedia. Feb. 13, 2025. Accessed: Feb. 16, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Big_Five_personality_traits&oldid=1275467842
[51] "Introduction to Multi-Task Learning(MTL) for Deep Learning - GeeksforGeeks." Accessed: Feb. 16, 2025. [Online]. Available: https://www.geeksforgeeks.org/introduction-to-multi-task-learningmtl-for-deep-learning/
[52] "Understanding TF-IDF (Term Frequency-Inverse Document Frequency) - GeeksforGeeks." Accessed: Feb. 05, 2025. [Online]. Available: https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/
[53] "Word2vec - Wikipedia." Accessed: Feb. 05, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Word2vec
[54] "Word Embeddings Using FastText - GeeksforGeeks." Accessed: Feb. 16, 2025. [Online]. Available: https://www.geeksforgeeks.org/word-embeddings-using-fasttext/
[55] "XLNet." Accessed: Feb. 16, 2025. [Online]. Available: https://huggingface.co/docs/transformers/model_doc/xlnet
[56] "Accuracy vs. precision vs. recall in machine learning: what's the difference?" Accessed: Nov. 21, 2024. [Online]. Available: https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall
[57] "roc and auc." Accessed: Oct. 09, 2024. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:text=Receiver%2Doperating%20characteristic%20curve%20(ROC,for%20choosing%20model%20and%20threshold
[58] "comp." Accessed: Oct. 09, 2024. [Online]. Available: https://www.geeksforgeeks.org/idea-of-efficiency-in-computational-thinking/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY