International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

"Personality Prediction Using ML And NLP"

1. Prof. Afsha Akkalkot,2. Abhishek Nagarkar,3.Shrikant Nagawade,4.Rahul Rale,5. Suraj Raut

[1]*Assistant Professor, Department of Computer Engineering, Zeal college of engineering and research narhe pune.*

[2,3,4,5]*Students, Department of Computer Engineering, Zeal college of engineering and research narhe Pune*

Abstract: ***Predicting an individual's personality can help to understand different aspects related to them such as how do they deal with stress, openness to social activities, and if they influence others. One of the key characteristics that affects how people interact with the outside environment showcases their personality. Personality prediction systems have gained significant attention due to their potential applications in various domains, including psychology, marketing, job recruitments, and human resources. Traditional approaches to personality prediction primarily rely on manual evaluation of self-reported questionnaire, which suffers from several limitations. The limitations include subjectivity biases, potential social desirability effects, and time-consuming data collection processes. This project aims to leverage Machine Learning (ML) and Natural Language Processing (NLP) techniques to predict personality traits of an individual using questionnaire responses. Personality prediction system has been implemented based on Myers-Briggs Type Indicator (MBTI) model which is a popular personality prediction framework. The model consists of 16 personality types based on combination of 4-pairs of personality traits.***

Keywords: Personality Prediction, Machine Learning, Natural Language Processing, Myers-Briggs Type Indicator.

I. INTRODUCTION

Personality refers to a unique characteristics and traits of a person based on their thoughts, feelings and behaviours that distinguish a person from others. It's been done for a long time to determine a person's personality based on their nature. Traditional approaches for personality prediction were based on the responses received for a self-reported questionnaire. This way of personality prediction was time-consuming and depends on manual inspection which can lead to human error. In recent years, there has been a significant growth in the field of personality prediction using Machine Learning and Natural Language Processing techniques.

We have used and implemented a popular model for personality prediction which is Myers-Briggs Type Indicator (MBTI), which categorizes individuals into distinct personality types based on their preferences in four key dimensions:

1. Extraversion / Introversion
2. Sensing / Intuition
3. Thinking / Feeling
4. Judging / Perceiving

1.1 PROBLEM STATEMENT

Personality prediction is the task of inferring a person's personality traits from their responses to questionnaire. Personality is a complex trait that is difficult to measure. Traditional approaches to personality prediction were time consuming and cumbersome and might lead to biased predictions. The process was

rigorous and needed to be done manually. The proposed system provides a solution for Personality Prediction of a person using Machine Learning and Natural Language Processing techniques and algorithms. This system will provide the organizations with a better, faster and more accurate prediction of personality without any need of expert person.

II.RELATED WORKS

R. Hegde, S. Hegde, Sanjana, S. Kotian, and S. Shetty, [1] describes, “Personality classification using Data mining approach”. In this paper, the author said that Personality classification refers to the psychological classification of different types of individuals. This project deals with the areas where it determines the characteristics of a person. It can be helpful to classify person using Personality classification using data mining approach. In this paper, we aim to automate the personality prediction of the users by taking a personality test. The system uses classification algorithm i.e., N-closest neighbourhood algorithm (NCN). The analysis is done using vast set of data in data set and has been compared with the user input. This paper mainly focuses on classification algorithm.

F. Q. Annisa, E. Supriyanto and S. Taheri, [2] describes, "Personality Dimensions Classification with EEG Analysis using Support Vector Machine”. In this paper, the author said that Personality is the fundamental thing that forms the behavioural tendencies of each individuality in a situation. A common model used to describe personality is the big five personality that divides personality traits into five dimensions of neuroticism, extraversion, openness, agreeableness, and conscientiousness. Personality assessment through physiological signals offers objectivity and reliability of the test results due to the minimal role of test takers in the examination process. One widely recommended approach is signal-based analysis of electroencephalography (EEG). The EEG signal feature of the ASCERTAIN public database was extracted using discrete wavelet transform (DWT) and was classified using support vector machine (SVM) to determine personality dimensions. The results showed better performance compared to the application of other techniques on the same dataset with 69% and 75.9% accuracy to determine extraversion and neuroticism level, respectively. However, this accuracy still needs to be improved to generate reliable model. Increased data variability can be useful for understanding brain dynamic activity per individual.

D. Al-Hammadi and R. K. Moore, [3] describes, "Using Sampling Techniques and Machine Learning Algorithms to Improve Big Five Personality Traits Recognition from Non-verbal Cues". In this paper, the author said that Automatic personality recognition using the Big Five dimensions (OCEAN: extraversion, agreeableness, conscientiousness, neuroticism and openness) is capturing the attention of researchers. Personality recognition is expected to have encouraging future in Human- computer and Robot Interaction applications. Human speech conveys rich information that can be derived to recognize speaker traits. However, our focus is on the rich content of non-verbal features in human speech. We focus on how humans talk, not what they talk about. The focus in this paper is to experiment with four different machine learning techniques, and their performance in recognizing personality traits, we report our results in this regard. We use the Speaker Personality Corpus provided by the Interspeech 2012 challenge. First, we recognize three issues affecting the system's low performance: dimensionality, judges' agreement, and imbalanced data. Next, we address each issue and provide a solution to improve the system's performance. Finally, we compare our results with the baseline showing better classification results

III.PROPOSED SYSTEM

The system we want to develop is based on MBTI (Myers-Briggs Type Indicator) Model. We are using Machine Learning and Natural Language Processing for the purpose of personality prediction. We have gathered the dataset from Kaggle for personality prediction. We are going to predict the personality through questionnaire which contains objective and subjective questions. We are implementing multiple Machine Learning models and compare them for better accuracy.

The Myers-Briggs Type Indicator (MBTI) is a psychological framework and assessment tool used to categorize individuals into different personality types. It was developed by Katharine Cook Briggs and her daughter Isabel Briggs Myers, based on the theories proposed by Carl Jung.

The MBTI model classifies people into 16 distinct personality types, each represented by a four-letter code. The four dimensions of personality used in the MBTI are:

1. Extraversion (E) vs. Introversion (I):

This dimension reflects how individuals gain energy and focus their attention. Extraverts tend to be outgoing, social, and energized by interacting with others, while introverts prefer solitude, introspection, and gain energy from being alone.

2. Sensing (S) vs. Intuition (N):

This dimension describes how individuals perceive and gather information. Sensors rely on their five senses and prefer concrete, factual information, while intuitive individuals focus on patterns, possibilities, and abstract concepts.

3. Thinking (T) vs. Feeling (F):

This dimension refers to how individuals make decisions and evaluate information. Thinkers tend to rely on logical analysis and objective criteria, while feelers prioritize personal values, empathy, and consider the impact on others.

4. Judging (J) vs. Perceiving (P):

This dimension reflects how individuals prefer to approach the outside world. Judgers prefer structure, organization, and decisiveness, while perceivers are more flexible, adaptable, and open to new experiences.

Combining these four dimensions creates 16 possible personality types, such as ISTJ, ENFP, ENTJ, etc. Each type represents a unique combination of preferences and is associated with different strengths, communication styles, and ways of approaching tasks and relationships.

The MBTI is commonly used in career counselling, team-building exercises, and personal development to gain insights into individuals' preferences, strengths, and potential areas of growth. However, it's important to note that the MBTI has also received criticism for its lack of scientific evidence, limited predictive power, and potential for oversimplification of complex human personalities

IV.METHODOLOGY

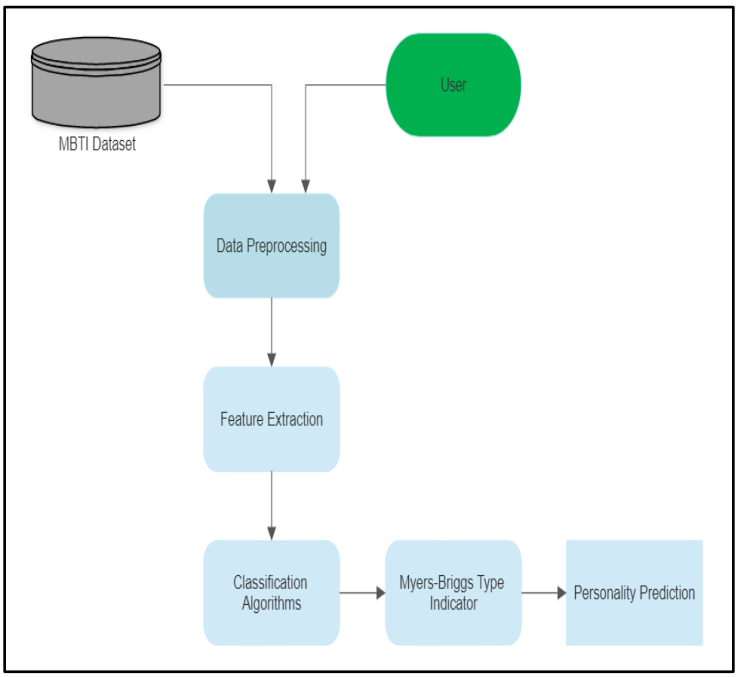


Figure 1 Architecture

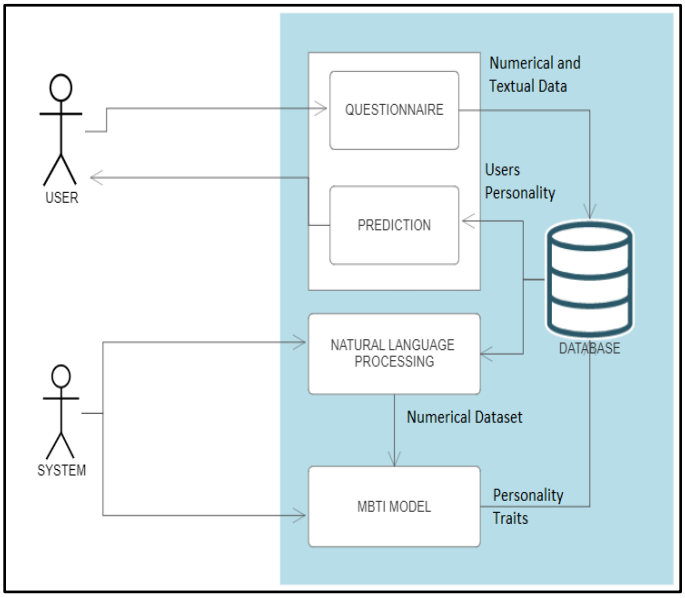


Figure 2 Use case Diagram

1. Data Collection

To start our project we analysed a dataset which includes 32 numeric columns and one textual column. The textual column contained sentences about individuals' personality traits, based on the Myers-Briggs Type Indicator (MBTI) Model. We pre-processed the textual column by removing links and stop words to ensure that the model did not associate any irrelevant information with the personality trait. Stop words were removed as they do not contribute to the meaning of the sentence and can add noise to the data. After pre-processing we used the Term Frequency-Inverse Document Frequency algorithm to extract 32 new features from the textual column. We combined these features with original 32 features from textual column

2. .Pre-processing of data

The dataset is pre-processed using different pre-processing techniques like cleaning values, handling missing values, and or outliers, and converting textual data into numerical data using the concept of vectorization of text.

3. Model Training

Several machine learning models are trained after pre-processing such as Logistic Regression, CatBoost classifier, Multinomial Naïve Bayes, XGBoost Several machine learning models are trained after pre-processing such as Logistic Regression, CatBoost classifier, Multinomial Naïve Bayes, XGBoost Classifier, Support Vector Machine, Random Forest Classifier, Decision Tree Classifier. Each of these models has its own strengths and weaknesses and can be applied to different types of problems depending on the nature of the data. The performance of each model was evaluated using various metrics such as accuracy, precision, recall

and F1 score.Classifier, Support Vector Machine, Random Forest Classifier, Decision Tree Classifier.

4. Model Selection

After evaluating the performance of all the models, we found that the CatBoost model had the highest accuracy score of 90.22%. In addition, the precision, recall, and F1 score metrics of the CatBoost model were also high, indicating that it is a suitable model for personality prediction. The high accuracy score achieved by the CatBoost model is likely due to its ability to handle categorical data and its ability to handle missing values. The CatBoost model's ability to handle categorical data is also important in the context of personality prediction using the MBTI model, as the MBTI model is based on categorical data. Therefore, the CatBoost model's ability to handle categorical data makes it a suitable model for personality prediction using the MBTI model.

5. Web Application Development

The development of a user-friendly and intuitive frontend was a crucial aspect of the project, as it provided an accessible way for users to engage with the machine-learning model. The front end was designed using Streamlit, a popular open-source framework for building web applications and was tailored to meet the needs of the target audience. The frontend was designed to be easy to navigate, with clear instructions on how to use it. Users were prompted to enter their data in a straightforward and intuitive manner, with input fields clearly labelled and easy to understand.

6. Hosting and Deployment

The website is deployed on Streamlit.io cloud platform where the website is accessible globally for anyone to use at any time. The deployment process involved uploading the project on github and creating an account on Streamlit platform. Once this is done, link the github account and add the repository link containing the code for streamlit.

As shown in the above figure, the design of the system is clearly depicted. At first, the required data is collected from the UI created, and further, the pre-processing of the data is carried out for removing the noise in the data and replacing the null value with the proper substituent. The MBTI model carries the process. The four categories in MBTI are intro-version/extraversion, sensing/intuition, thinking/feeling, judging/perceiving.

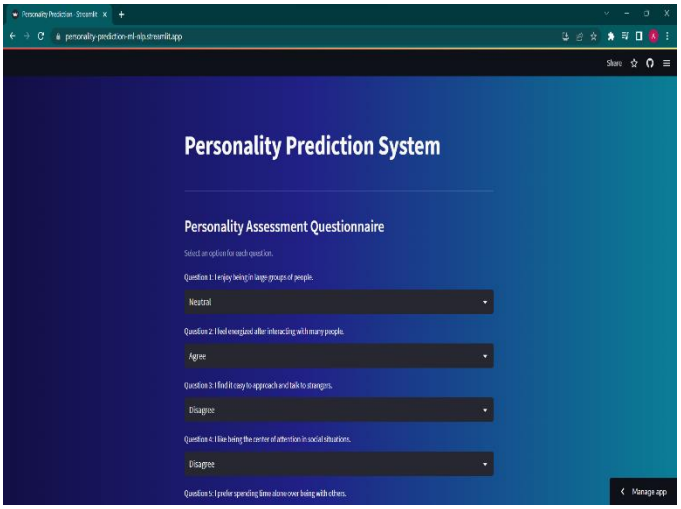
According to the MBTI, each person is said to have one preferred quality from each category, producing 16 unique types. To find out what your MBTI personality type is you need to complete the questionnaire. These questionnaires are answered in multiple-choice as well as descriptive manner. This will help to have a descriptive dataset which will aid in improving the accuracy of finding the exact personality type.

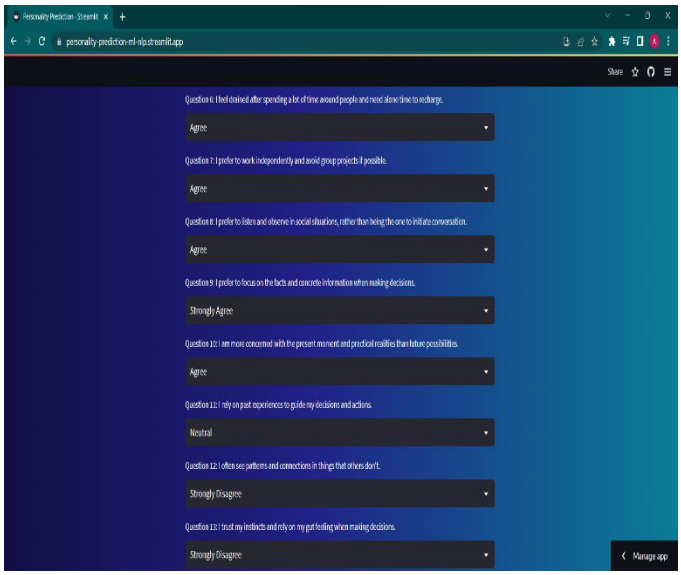
For processing the data using an algorithm, it is necessary to convert textual data into numeric data. This descriptive data is converted into a numeric dataset by using TF-IDF vectorization. The algorithm lists all the unique characters in the dataset and provides a specific number to the descriptive data. The number generated using this algorithm lies between 0 to 1. For the word having maximum occurrence in the data set, the number generated is 0(zero) or nearly equal to 0(zero) and moving towards the unique word with a value near to 1(one).

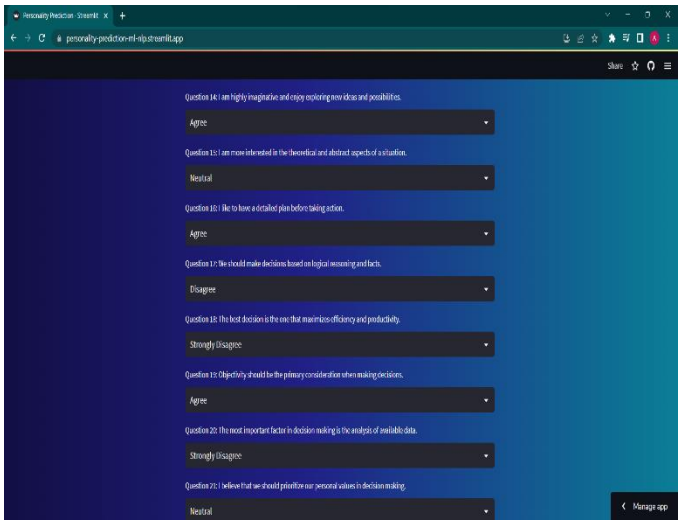
V. RESULTS

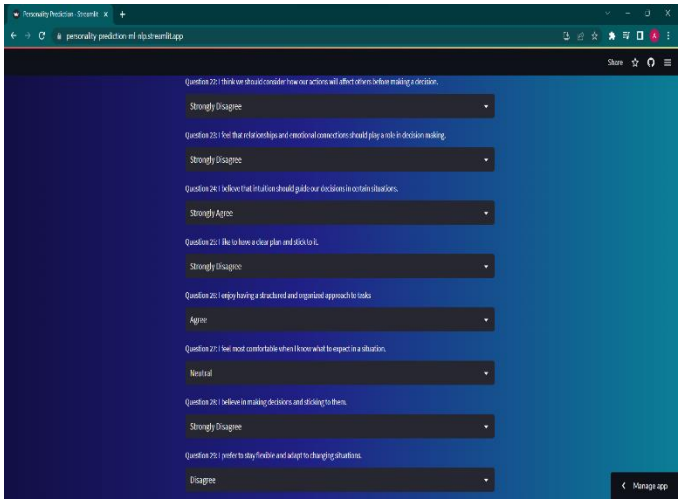
	Models	Test accuracy
0	CatBoost Classifier	98.221325
1	XGBoost Classifier	76.069609
2	logistic regression	74.435150
3	Linear Support Vector classifier	71.762331
4	Multinomial Naive Bayes	56.821459
5	Random Forest Classifier	56.013845
6	Decision Tree classifier	41.121046

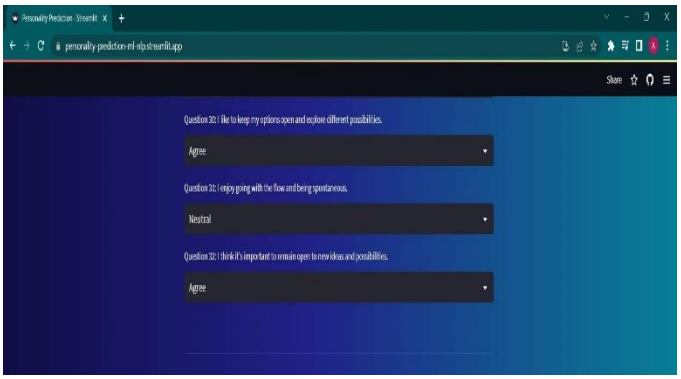
Figure 3 Different Model Accuracies

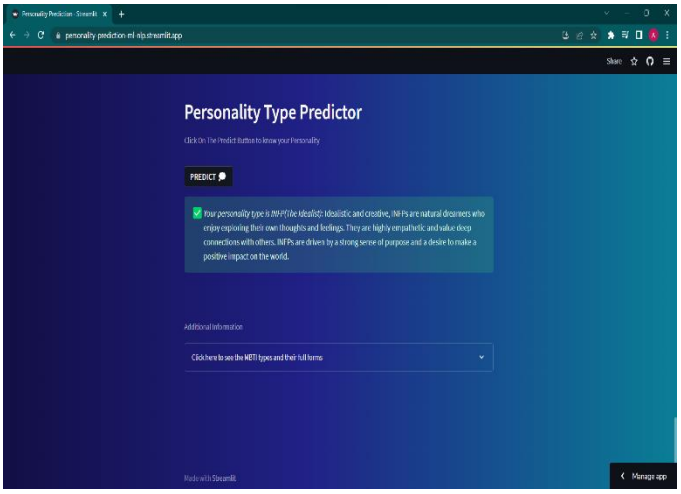












VI. CONCLUSION

The Personality Prediction System using ML and NLP is a beneficial tool for various entities in their field of work. By leveraging ML algorithms and NLP techniques, the system can analyse textual data and extract valuable insights into an individual's personality traits, characteristics, and behaviours.

ML models trained on the dataset of textual and mcq responses have effectively captured the nuances and complexities of personality, allowing for accurate predictions and understanding of individual differences. NLP techniques enabled the analysis of semantic content, contextual information, and even subconscious cues present in text, providing a comprehensive view of personality dynamics.

Overall, Personality Prediction through ML and NLP represents a powerful approach for unravelling the intricacies of human personality. As technology advances and research in this area progresses, we can anticipate further advancements in our understanding of personality traits, leading to more accurate predictions and valuable insights that can benefit diverse fields and contribute to our knowledge of human behaviour.

VI. REFERENCES

[1] R. Hegde, S. Hegde, Sanjana, S. Kotian, and S. Shetty, “Personality classification using Data mining approach,” IJRAR, 2019.

[2] F. Q. Annisa, E. Supriyanto and S. Taheri, "Personality Dimensions Classification with EEG Analysis using Support Vector Machine," 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2020, pp. 79-82, doi: 10.1109/ISRITI51436.2020.9315507.

[3] D. Al-Hammadi and R. K. Moore, "Using Sampling Techniques and Machine Learning Algorithms to Improve Big Five Personality Traits Recognition from Non-verbal Cues," 2021 National Computing Colleges Conference (NCCC), Taif, Saudi Arabia, 2021, pp. 1-6, doi: 10.1109/NCCC49330.2021.9428804.

[4] Alam Sher Khan, Hussain Ahmad, Muhammad Zubair Asghar, Furqan Khan Saddozai, Areeba Arif and Hassan Ali Khalid, “Personality Classification from Online Text using Machine Learning Approach” International Journal of Advanced Computer Science and Applications (IJACSA), 11(3), 2020.

[5] Z. Mushtaq, S. Ashraf and N. Sabahat, "Predicting MBTI Personality type with K-means Clustering and Gradient Boosting," 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, 2020, pp. 1-5, doi: 10.1109/INMIC50486.2020.9318078

[6] Sagar Patel, Mansi Nimje, Akshay Shetty, Sagar Kulkarni, 2021, “Personality Analysis using social media”, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NTASU – 2020 (Volume 09 – Issue 03)

[7] M. N. Sahono et al., "Extrovert and Introvert Classification based on Myers-Briggs Type Indicator (MBTI) using Support Vector Machine (SVM)," 2020 International Seminar on Application for Technology of Information and Communication (semantic), Semarang, Indonesia, 2020, pp. 572-577, doi: 10.1109/iSemantic50169.2020.9234288.

[8] V. dos Santos, I. Paraboni, “Myers-Briggs personality classification from social media text using pre-trained language models”. JUCS - Journal of Universal Computer Science, 28, 4. 378–395, 2022. doi: 10.3897/jucs.70941.

[9] E. J. Zaferani, M. Teshnehlab and M. Vali, "Automatic Personality Perception Using Autoencoder and Hierarchical Fuzzy Classification," 2021 26th International Computer Conference, Computer Society of Iran (CSICC), Tehran, Iran, 2021, pp. 1-7, doi: 10.1109/CSICC52343.2021.9420627.