International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Unmasking The Enigma:Exploring Personality Traits Using Social Media Text

S.V Satya Krsihna
Associate professor
*CSE (*AI&ML*) Department*
Siddhartha Institute of Technology and Sciences Hyderabad,Telanagana .

P.Tharun Kumar Reddy
*CSE (*AI&ML*) Department*
Siddhartha Institute of Technology and *Sciences* Hyderabad,Telanagana
pothulatharunkumarreddy@gmail.com .

Pathinavalasa Sai Kumar
*CSE (*AI&ML*) Department*
Siddhartha Institute of Technology and Sciences Hyderabad,Telanagana
Ppathinavalasasaikumar730@gmail.com.

S.Surya Kundan
*CSE (*AI&ML*) Department*
Siddhartha Institute of Technology and Sciences Hyderabad,Telanagana
suryakundans@gmail.com .

MD Khaja Muneeb
*CSE (*AI&ML*) Department*
Siddhartha Institute of Technology and Sciences Hyderabad,Telanagana
mkuddin301@gmail.com .

Abstract-

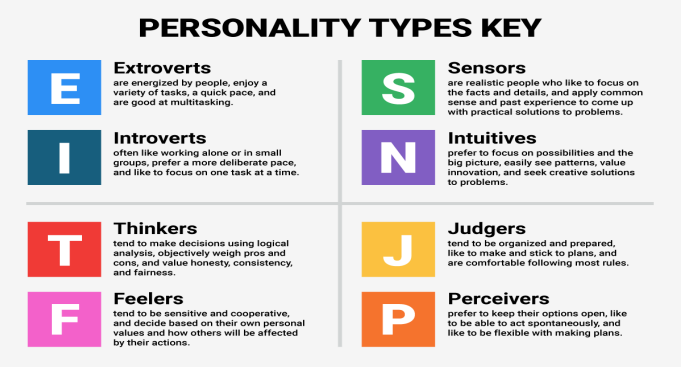
In this paper personality is an important parameter as it differentiates various individuals from another. Predicting personality has many applications in real world. The main objective of this project is to take textual data as input from the user and then run the trained machine learning model on this data to predict his four personality traits which are Introversion vs Extroversion, Sensing vs Intuition, Thinking vs Feeling, Judging vs Perceiving.The main objective is to build an application where users can answer the questions which are processed and analyzed to output his personality traits. The output is a string of four characters where each character determines a personality trait, total of sixteen personality types are possible.The Machine learning model Random Forest Classifier is used to classify the text and output four personality traits. Processing of large textual data is to be done using Natural Language Processing (NLP) techniques with the help of NLTK libraries to process and categorize the data. In order to increase the performance of the model hyper parameter tuning along with cross fold validations is done.

Keywords: **Random Forest**, **Natural Language Processing**, **Text**

I.INTRODUCTION

Personality is what distinguishes the people from one another so it is considered an important parameter. Personality is a key aspect of human life. The study of personality more specifically comes under the branch of psychological study. Personality is constituted of elements like a person's thoughts, feelings, behavior which continuously keep changing over time. The prediction of personality is treated as a classification problem in computer science as the people are classified into different classes of personality types. There are a number of psychological tests that yield different types of personality classes. Popular tests include MBTI, Big Five, DISC. The Myers-Briggs Type Indicator (MBTI) is one of the most famous and widely used personality tests or descriptors. It describes the way people behave and interact with the world around them with four binary categories and 16 total types. They are as follows: Introversion vs Extroversion, Sensing vs Intuition, Thinking vs Feeling, Judging vs Perceiving.

Understanding personality traits can be very useful as it helps users to discover why people behave in certain ways, the areas in which they can improve and also helps users in finding other people with similar personality traits. The main objective of the project is to build an application where users will answer a few questions which will be analyzed and his/her personality traits are outputted. The output is a string of four characters where each character determinesa personality trait so a total of 16 personality types are possible. Each person’s MBTI personality type is defined as the collection of their four types for the four categories, using the bolded identifying letter for each. For example, one who derives their energy mostly from being around other people (E), trusts their gut and uses intuition to interpret information in the world (N), thinks rationally about their decisions (T), and lives life in a carefully planned manner (J) rather than a spontaneous one would have the personality type ENTJ.



II. LITERATURE SURVEY

A Hybrid Deep Learning Technique for Personality Trait Classification From Text” BY HUSSAIN AHMAD1 , MUHAMMAD USAM A ASGHAR 1 , MUHAMMAD ZUBAIR ASGHAR 1 , AURANG ZEB KHAN2 , AND AMIR H. MOSAVI(2021) numerous methods are suggested for exploiting different approaches to resolve the pre diction issue, and techniques are implemented to present personality prediction.

“Detection and Classification of Psychopathic Personality Trait from Social Media Text Using Deep Learning Model” BY Junaid Asghar, 1Saima Akbar,2Muhammad Zubair Asghar, 2 BashAhmad,3Mabro ok S. Al-Rakhami , 4 and Abdu Gumaei(2021)) In their work on id entifying the relationship between personality and online behavior.

III. EXISTING SYSTEM

To predict personality types, existing solutions used Naive Bayes, SVM, and LDA as classifiers, and some of them used a multi-class classification technique. But their accuracy, performance and speed are quite low.

IV. PROPOSED SYSTEM

This Project is implemented using Random Forest Classifier. Random forest is a tree- based algorithm which involves building several trees (decision trees), then combining their output to improve generalization ability of the model. The method of combining trees is known as an ensemble method. Ensembling is nothing but a combination of weak learners (individual trees) to produce a strong learner. This algorithm has a better accuracy compared to the algorithms in the existing system.

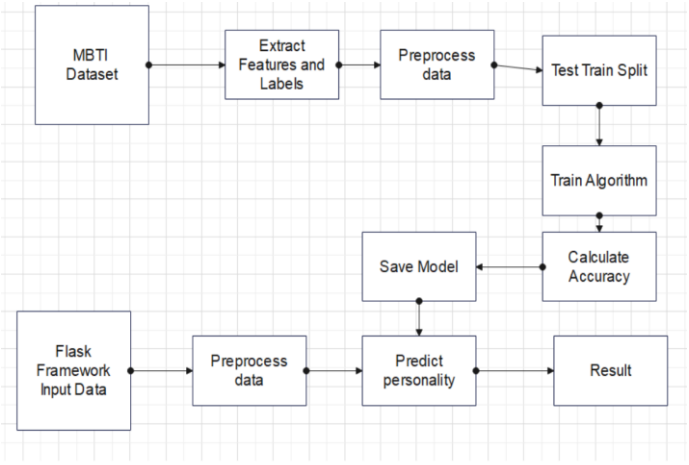
Also, the algorithm's accuracy can be calculated. When the test data is
evaluated against the various machine learning algorithms, following
are the accuracies:

Algorithm	Accuracy
Random Forest Classifier	89.1
Logistic Regression	64.55
Naïve Bayes	60.46
Decision Tree Classifier	54.49

Table : Comparison of accuracies

IV. USING THE TEMPLATE

Two common frameworks for understanding personality are the OCEAN Model (Big Five) and the Myers-Briggs Type Indicator (MBTI). The OCEAN Model categorizes personality into five dimensions: Openness, reflecting a propensity for new experiences and ideas; Conscientiousness, indicating organization, discipline, and goal-oriented behavior; Extraversion, representing sociability, outgoingness, and a preference for social interaction; Agreeableness,denoting cooperativeness, empathy, and trustworthiness; and Neuroticism, highlighting susceptibility to negative emotions and anxiety. On the other hand, the MBTI focuses on preferences influencing personality, categorized into four dichotomies: Extraversion vs. Introversion, Sensing vs. Intuition, Thinking vs. Feeling, and Judging vs. Perceiving. These frameworks offer valuable insights into personality traits, although individuals often exhibit a unique blend of characteristics across these dimensions.



Architecture of Proposed System

A.*Data Collection:* The personality traits involves administering standardized assessments or questionnaires to individuals to gather data on their preferences, behaviors, and emotional tendencies. These assessments are based on established frameworks like the Big Five or MBTI and measure various dimensions of personality. Data can also be collected through interviews, observations, and self-reports. Multiple sources, including self-reports, peer evaluations, and expert ratings, may be used to obtain comprehensive personality data. The goal is to collect accurate and reliable information to facilitate research, assessment, and understanding of human behavior.

B.Data preprocessing: The personality traits involves cleaning and Transforming raw data to prepare it for analysis. This includes handling missing values and outliers, encoding categorical variables, and scaling numerical features. Additionally, feature engineering may involve creating new features or reducing dimensionality. The dataset is typically split into training, validation, and test sets, ensuring balanced distributions of personality traits. Normalization may be applied to ensure proportional contributions from each feature, and techniques like oversampling or under sampling address imbalanced data. Finally, data integration from multiple sources enriches the dataset, and validation ensures the data meets analysis assumptions. These steps ensure the data is accurate and ready for further analysis or modeling.

C. *Feature Extraction:* The personality traits involves identifying key attributes or characteristics from raw data to represent individuals' personalities. This process aims to condense complex personality data into a concise set of informative features.Techniques may include analyzing responses to standardized questionnaires or assessments, identifying patterns or trends in behavior, and deriving numerical representations of personality traits. The extracted features serve as inputs for analysis or modeling tasks, enabling researchers to understand and predict various aspects of human behavior andpersonality.

Random forest is a tree-based algorithm which involves building several trees (decision trees), then combining their output to improve generalization ability of the model. The method of combining trees is known as an ensemble method. Ensembling is nothing but a combination of weak learners (individual trees) to produce a strong learner.

Definition: A random forest is a classifier consisting of collection of tree structured classifiers $h(x, \Theta k)$, $k = 1, ...$ where the Θk are independent identically distributed (i.i.d) random vectors and each tree casts a unit vote for the most popular class at input. Random Forest Algorithm: The following are the basic steps involved in performing the random forest algorithm:

$$\text{Entropy}(S) = -\sum_{i=1}^{c} pi \log 2\ pi$$

Here, c is the total number of classes or attributes and pi is number of examples belonging to the ith class. Information gain is simply the expected reduction in entropy caused by partitioning all our examples according to a given attribute. Mathematically, it is defined as:

$$\text{Gain}(S, A) \equiv \text{Entropy}(S) - \sum |Sv| \text{Entropy}(Sv) / v\text{Values}(A) |S|$$

' S' refers to the entire set of examples that we have. A is the attribute we want to partitionor split. |S| is the number of examples and |Sv| is the num ber of examples for the current value of attribute A. The attribute with the highest information gain sits at the root node, and the tree is first split base d on that attribute.

.

In this model, you can start with some of the software specifications and develop the first version of the software. After the first version if there is a need to change the software, then a new version of the software is created with a new iteration. Every release of the Iterative Model finishes in an exact and fixed periodthat is called iteration.

The Iterative Model allows accessing earlier phases, in which the variations are made respectively. The final output of the project was renewed at the end of the Software Development Life Cycle (SDLC) process.

There are various phases in the iterative model. They are:

1.Requirement gathering & analysis : In this phase, requirements are gathered from customers and checked by an analyst whether requirements will be fulfilled or not. Analyst checks that need will be achieved within budget or not. After all of this, the software team skips to the next phase.

2. Design : In the design phase, the team designs the software by the different diagrams like Data Flow diagram, activity diagram,class diagram, state transition diagram, etc.

3.Implementation: In the implementation, requirements are written in the coding language and transformed into computer programmers which are called Software.

4.Testing: After completing the coding phase, software testing starts using different test methods. There are many test methods, but the most common are white box, black box, and grey box test methods

5.Deployment: After completing all the phases, software is deployed to its work environment.

6.Review: In this phase, after the product deployment, review phase is performed to check the behavior and validity of the developed product. And if there are any errors found then the process starts again from the requirement gathering.

7.Maintenance: In the maintenance phase, after deployment of the software in the working environment there may be some bugs, some errors or new updates are required. Maintenance involves debugging and new addition options.

D. System Design:

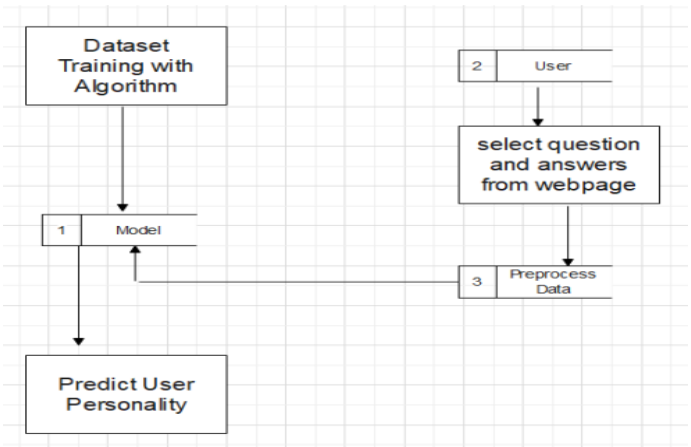
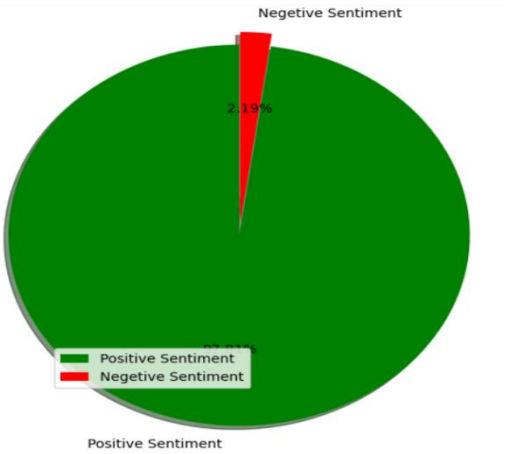


Fig Data Flow Diagram

This Data Flow Diagram (DFD) illustrates the process of predicting a user's personality using a trained model. It begins with a dataset being used to train a machine learning algorithm, creating a model capable of making predictions. The user interacts with the system by selecting questions and providing answers on a webpage. These responses are then preprocessed to clean and format the data. The preprocessed data is fed into the trained model, which analyzes it to predict the user's personality traits. The final output is the predicted personality, generated based on the user's input and the model's analysis.



VI. CONCLUSION

Personality traits play a crucial role in shaping individuals' behaviors, preferences, and interactions. Understanding personality traits allows for better insight into human psychology, behavior, and decision-making processes. Through frameworks like the Big Five and MBTI, researchers can categorize and analyze personality traits, facilitating research, assessment, and practical applications in fields such as psychology, sociology, and human resources. By collecting and preprocessing data effectively, researchers can extract meaningful features to represent personality traits accurately, leading to deeper insights and more informed decision-making. Overall, the study of personality traits offers valuable insights into what makes individuals unique and how they navigate the complexities of life.

REFERENCES

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

[2] Brandon Cui, Calvin Qi, **"**Survey Analysis of Machine Learni ng Methods for Natural Language Processing for MBTI Persona lity Type Prediction**"**, Stanford,2018.

[3] Hernandez, Rayne, Knight, Ian Scott, **"**Predicting Myers-Brig gs Type Indicator with text classification**"**,31st Conference on Ne ural Information Processing System, USA,2017.

[4] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," 2015 International Conference on Data and Software Engineering (ICoDSE), 2015, pp. 170-174, doi: 10.1109/ ICODSE.2015.7436992 .

[5] Shristhi Chaudary, Ritu Singh, Syed Tausif Hasan and Ms. Inderdeep Kaur, A Comparative Study of Different Classifiers
for Myers-Brigg Personality Prediction Model, International Research Journal of Engineering and Technology (IRJET), 2018.
[6] J. Golbeck, C. Robles, M. Edmondson and K. Turner, " Predicting Personality from Twitter," 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 2011, pp. 149-156, doi: 10.1109/PASSAT/ SocialCom.2011.33.

[7] Tommy Tandera, Hendro Derwin Suhartono, Rini Wongso
and Yen Lina Prasetio, "Personality Prediction System from Facebook Users", 2nd International Conference on Computer Science and Computational Intelligence, 13–14 October 2017.