Data Mining For Automated Personality Classification

Dr.R N Devendra Kumar[1], P Poovesh [2], N Sethuraman [3], M Tharun Kumar [4]
[1]Associate Professor, Dept of CSE
[2, 3, 4]Dept of CSE
[1, 2, 3, 4] Sri Ramakrishna Institute of Technology

Abstract- *The main applications of personality are in leadership, influence, communication, collaboration, business negotiation, and stress management. One of the key characteristics that affect how people interact with the outside environment is their personality. We can use this initiative more effectively if we have data on individual behavior. This information about personal conduct can be helpful in determining a person's identity based on personality features. The database will already contain the personality traits. The system will determine the user's personality based on the Big Five Personality Traits later when the user enters his personality characteristics, which are examined in the database. One characteristic that affects how people interact with the outside environment is personality. Using personality classification, this information may be useful in determining a person's personality. Based on previous classifications, this learning can now be used to predict user personality. In this project, we suggest a method for evaluating an applicant's personality.*

Keywords- personality, Behaviour, Logistic regression, Decision tree, SVM, Big Five personality Traits

I. INTRODUCTION

A person's personality determines whether he can be a leader or not, have an impact on those around him, be an expert communicator, work collaboratively, be able to negotiate deals, and handle stress. Personality refers to characteristics of persons that determine how they interact with their environment. Based on a person's personality qualities, one can determine that person's personality using personal behavior characteristics. Our model will provide data about the user's personality. The system will compare the personality traits with the data in the database based on the user-provided personality features. The software will automatically categorize the personality of the user and compare the pattern to the patterns in the stored data. The system will review the database's data and compare the user's personality attributes to the information there. The system will then determine the user's personality. The system will offer additional features related to the user's personality based on the user's personality traits. The way a person interacts with their environment and the outside world can vary depending on their personality. Moreover, personality can be used as a factor in the hiring process, career counseling, health counseling, etc. Analyzing a person's actions to predict their personality is an old and ineffective technique.

1.1 Literature Survey:

Novel approaches to automated personality classification:

This article [1] suggests various fresh lines of inquiry into the issue of automated personality classification (APC). First, we look at potential enhancements to the current APC solutions. To do this, we combine several APC corpora, psychological characteristic measures, and learning algorithms. After that, we look at several variations on the APC problem and tasks that are connected to it, such as dynamic APC and text personality discrepancy detection. The environment of social networks and the associated data mining techniques served as the backdrop for the entire research project. The issue of automated personality classification is looked at in this research based on data from the following content: text that the person wrote and metadata about a person that was requested, obtained through social networks, or obtained in other ways. Several studies also examine speech, facial features, gestures, and other behavioral factors, but these are not the focus of our investigation.

A System for Personality and Happiness Detection:

This [2] piece of work suggests a framework for measuring happiness and personality. Based on Eysenck's thesis on human nature, writers want to offer a platform for gathering text messages from social media (WhatsApp) and categorizing them according to various personality traits. Even if there isn't a certain connection between personality traits and happiness, future research may reveal some links. In this paper, we present the platform created and, as a proof-of-concept, we have employed several sources of data to determine whether standard machine learning algorithms can be used to categorize various personality traits and happiness. In this sector, it is acceptable to anticipate that many people will express themselves in writing in a variety of ways, each of which will be unique to the individual. The Sleep scheduling method is created entirely via distributed development.

An Examination of Online Learning:

The idea of online learning was developed by Nurbiha A. Shukora [3], and it has since gained a lot of popularity as a result of technological advancements that have made it possible to participate in conversations even from a distance. The majority of research shows how effective online learning has aided students in enhancing their capacity for learning while also evaluating the learning process. Only by using data mining techniques, where we can access the various experiences of students who filed online based on their log files, is this type of conversation conceivable. To become good online learners, pupils are reportedly advised to work harder.

Using Twitter Content to Predict Psychopathy:

R. Wald [4] has studied human psychology using content from social media sites like Twitter. They claimed that many people use the microblogging website Twitter to discuss their experiences and ideas about their daily lives. A fuller image of the user who is sending the tweets can be created by combining them, despite the fact that academics have frequently abandoned the strategy of predicting personality by studying the tweets because they believe it contains too little content to forecast meaningful information. Twitter has been used to predict psychopathy using Select RUS Boost, a novel form of ensemble learning that makes use of four classification learners and four feature selection methods.

Using Data Mining Techniques to Detect the Personality of Players in an Educational Game:

The objectives of creating systems for simulating student behavior based on data from online interactions, in-class talks, etc., according to Fazel Keshtkar [5]. Yet, techniques like Educational Data Mining (EDM) and Intelligent Tutoring Systems (ITS) employ a person's behavior and personality for analysis. As a result, a system that can be customized by the user and that also analyses pupils' interactional behavior has been created.

social Network Use and Personality:

According to J. Golbeck [6], a social network is a platform where users tend to reveal themselves to the public, providing details about their behavior and letting others into their life. Personality has a significant role in a variety of interpersonal interactions and can be used to forecast success in both romantic and non-romantic relationships as well as in predicting work satisfaction. Up until now, using a survey test, they conducted surveys with a variety of people to precisely anticipate consumer characteristics. Correct personality analysis was a challenge because this was highly impractical while gathering data from social media networks.

1.2 Problem Statement:

Nowadays personality assessment has become the most used test to hire many employees. In this project, we will classify the user’s personality based on the big five personality traits using data mining. We need a strong model that predicts the personality. Because a person’s personality can be predicted by the way she thinks, people's personalities are dependent on their imagination and thoughts. The objective of this research is to create a model that can predict a person's personality. The primary goal of the suggested system is to infer a user's personality from the responses they provide. The project's goal is to create software that will make it easier to determine a person's personality. The idea of machine learning algorithms is utilized. The user's personality will be anticipated and presented after classification. The entire procedure is carried out by a program, which substantially reduces the need for labor and saves a lot of time.

1.3 Scope of the Project:

In this study, we attempted to create a method for predicting personalities using the Big Five personality traits as well as to assess the personality type of the individual. In both his personal and professional life, a person's personality is extremely important. The effectiveness of work is increased since the person is working in what he is competent at and at what he is compelled to perform because many firms now shortlist candidates based on their personalities. Many businesses employ personality assessments as part of the hiring process. These are intended to give staff members a better understanding of each candidate's working methods and abilities. Employers look for candidates that have good personalities in addition to competence. As such, this initiative will benefit those who attend the hiring procedure. People are able to take the personality test and develop the abilities needed by the business.

1.4 Existing System:

In the current system, it is possible for people to make personality judgments about one another based on Facebook profiles or other text data, and some characteristics of Facebook profiles or other social network text data are employed in this process. Therefore, there should not be a perfect alignment between Facebook profile traits that contain the actual personality cues and features that individuals use to make personality assessments. People may disregard or misinterpret some of the real personality indicators while using irrelevant traits as criteria for evaluation. People are prone to prejudices and biases that could skew their perceptions. Moreover, a number of Social network text data, such as Facebook profiles, is challenging for people to comprehend. The number of Facebook friends, for instance, is given on the profile, but it is more challenging for a person to evaluate features like network density.

1.5 Proposed System:

An automated personality categorization system that combines data mining and machine learning algorithms to categorize the personalities of various users is presented as a solution to the issues with the current method. in addition to employing other algorithms, such as the Big Five Personality Model, Logistic Regression, Decision Tree, and Support Vector Machine. Using new methodologies makes it simple to detect the personality by identifying historical data and their patterns, which overrides the current system. The graph in this suggested system will display the Automated Personality System's percentages. These probabilistic values range from 0 to 1. Every applicant or user who registers receives a unique username and password; if not, registration is required prior to doing the survey in an automated personality classification system. Each candidate enters their username and password to log in and complete the survey. The user can complete the survey to learn the Big five personality traits, which are determined by the survey, which has 50 questions for each trait and a 10-question aptitude exam. The user can view the results of his or her personality after completing the survey. After receiving the results, we can view the graph, and individuals whose personalities match receive suggestions, such as a friend suggestion. This is helpful in a variety of fields, including interview recruitment, government sectors, and psychometric testing. Once a user's findings are appropriate, they can enter any organization that hires people for positions based on personality types. In this system, we identify each user's personality and the average of each of the main five personality traits on a scale of 1 to 10, ultimately determining the user's personality type. The kind of personality is predicted based on the user's responses to the personality test. By using that personality type prediction, a user can apply for jobs with ease and learn what personality type is anticipated. Students can also learn about their personalities and take part in competitive exams in the same way.

II. SYSTEM SPECIFICATION

2.1 System Architecture:

High-level knowledge of the system's operation can be obtained from the system architecture. The system's operation starts with the database being queried for data. Now, we choose the attributes and divide the data into training and testing data. The required data is then pre-processed to remove inaccurate and duplicate data. The personality test cannot be taken without the user first logging in. There are 50 questions. 50 questions with 10 questions each must be answered by the user. Following the application of algorithms based on the responses, the model is trained using training data. The top five personality traits are listed below, followed by a description of personality types. By employing test data to test the system, accuracy is evaluated. So, personality is predicted after that.

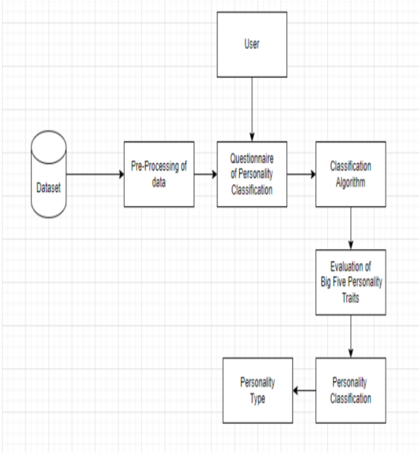
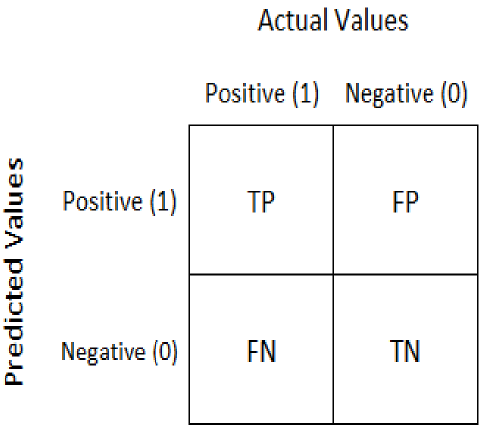

User
Dataset
Pre-Processing of data
Questionnaire of Personality Classification
Classification Algorithm
Evaluation of Big Five Personality Traits
Personality Type
Personality Classification

Fig 2.1(System Architecture)

2.2. Performance Metrics:

A project's evaluation of machine learning algorithms is crucial. In order to anticipate the personality system, we employed machine learning methods including Logistic Regression, Support Vector Machine, and Decision Tree. For predicting the behavior from the test data, we have 7 attributes and 1 attribute that is designated as a personality attribute. This system makes use of a number of characteristics, including Gender, Age, Openness, Agreeability, Extraversion, Neuroticism or Emotional Stability, and Conscientiousness. Five personality traits were used in our system, and performance measures were computed for each trait. For each attribute, we calculated the precision, recall, f1-score, and accuracy. The performance metrics for each categorization algorithm will be obtained.

- **True Positive (TP):** These are cases where the model predicts that the loan will be repaid and the original outcome in the dataset is the same.
- **True Negative (TN):** These are cases where the model predicts that the loan will default and the original outcome in the dataset is the same.
- **False Positive (FP):** These are cases where the model predicts that the loan will be repaid, but the original outcome in the dataset is that it will default.
- **False Negative (FN):** These are cases where the model predicts that the loan will default, but the original outcome in the dataset is that it will not default.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig 2.2(Performance Metrics)

The above figure is a confusion matrix that displays the error types. Using this, other performance metrics like precision, accuracy, true positive rate, and false positive rate are calculated. In the above figure, ‘n’ denotes the total number of cases. There are two possible Predicted and Actual classes: ‘YES’ and ‘NO’. Actual ‘Yes’ means that the loan was originally paid off and Actual ‘NO’ means the loan wasn’t paid off. Predicted ‘YES’ means that the model classifier predicted that the loan would be paid off and Predicted ‘NO’ means that the model classifier predicted that the loan would not be paid off. Thus, the classifier made a total of 165 predictions that were equal to the number of actual outcomes. ‘TN’, ‘FP’, ‘FN’, and ‘TP’ denote True Negatives, False Positives, False Negatives, and True Positives respectively.

The best metric to evaluate the model is the precision of the algorithm to predict whether a customer is going to repay the loan. This is achieved by training the model on the training dataset and then predicting (based on the features) the faithfully paying customers from those that default. The training results in being able to measure the precision, practicality, and realism of the model. The precision, accuracy, true positive rate, and false positive rate of the model are measured as follows:

- Precision= (True Positive/ (True Positive + False Positive))
- Accuracy= (True Positive/ (True Positive+ False Positive+ True Negative + False Negative)
- True Positive Rate: (True Positive/ (True Positive + False Negative))
- False Positive Rate: ((False Positive/ (False Positive + True Positive)))

So, for the confusion matrix in Fig. 4.1, the precision, accuracy, true positive rate and false positive rate are calculated as follows:

1. Accuracy: (TP + TN) / Total
2. Precision: TP/ Predicted YES
3. True Positive Rate: TP/ Actual YES
4. False Positive Rate: FP/ Actual NO

2.3 Algorithms Used:

Logistic regression:

The probability of a target variable is predicted using the supervised learning classification algorithm known as logistic regression. There are only two viable classes since the goal or dependent variable is dichotomous in nature. One of the most often used Machine Learning algorithms, within the category of Supervised Learning, is logistic regression. With a predetermined set of independent factors, it is used to predict the categorical dependent variable. In a categorical dependent variable, the output is predicted via logistic regression. As a result, the result must be a discrete or categorical value. It can be True or False, Yes or No, 0 or 1, etc., but rather than reporting the exact value like 0 or 1, it delivers the probabilistic values that are in the range of 0 and 1. Except for how they are applied, logistic regression and linear regression are very similar. Whereas logistic regression is used to solve classification difficulties, linear regression is used to solve regression problems. Due to its ability to categorize new data using continuous and discrete datasets and to generate probabilities, logistic regression is an important machine learning technique. When classifying observations using various sources of data, logistic regression can be used to quickly identify the factors that will work well.

Decision tree:

The Decision Trees are a type of Supervised Machine Learning where the data is continually divided according to a particular parameter (you explain what the input is and what the related output is in the training data). Decision nodes and leaves are the two components that can be used to explain the tree. The decisions or results are represented by the leaves. The data is divided at the decision nodes. which have results like "fit" or "unfit," respectively. In this instance, the issue was one of binary classification (a yes-no type problem).

Decision trees can be divided into two categories:

1. Trees that classify objects (Yes/No types)
Regression trees
2. (Continuous datatypes)

Support Vector Machine:

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it is largely employed in Machine Learning Classification issues. The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is a name given to this optimal decision boundary. SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors are the name given to these extreme circumstances, and as a result, the technique is known as a Support Vector Machine.

2.4 Requirement specification:

SoftwareRequirements

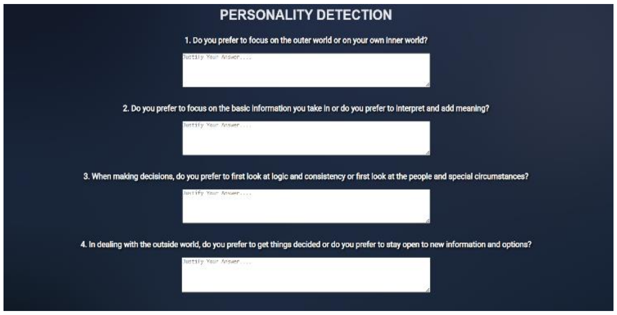
- Operating System: Windows or MacOS
- Platform: IA-32, x86 - 64
- Type: Network Simulator NS 2.34
- Language: Python

Hardware Requirements:
(Minimum requirements):

- CPU type: Intel Pentium 4
- Clock speed: 3.0 GHz
- RAM: 512 MB
- Hard disk capacity: 40 GB

2.5 Results:

The kind of personality is predicted based on the user's responses to the personality test. By using that personality type prediction, a user can apply for jobs with ease and learn what personality type is anticipated. Students can also learn about their personalities and take part in competitive exams in the same way.



PERSONALITY DETECTION

1. Do you prefer to focus on the outer world or on your own inner world?

Justify Your Answer....

2. Do you prefer to focus on the basic information you take in or do you prefer to interpret and add meaning?

Justify Your Answer....

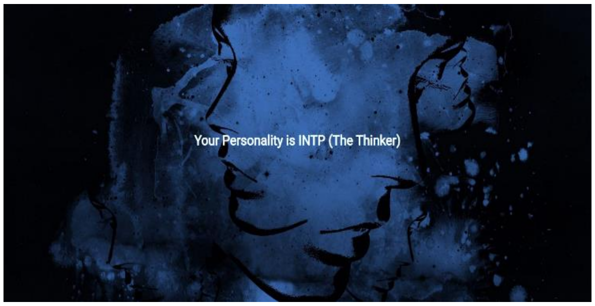
3. When making decisions, do you prefer to first look at logic and consistency or first look at the people and special circumstances?

Justify Your Answer....

4. In dealing with the outside world, do you prefer to get things decided or do you prefer to stay open to new information and options?

Justify Your Answer....

Fig 2.3(Questionaries)



Your Personality is INTP (The Thinker)

Fig 2.4 (Result)

III. CONCLUSION

3.1 Summary:

In this project, we talk about how various classification algorithms can be used to identify personalities. Here, we examine the interaction between the user and the personality system that is utilized on e-commerce sites, in competitive exams, psychometric tests, marriage-related websites, and in government organizations like the army, navy, and the air force. After the user attempts the survey, the system automatically classifies the personality based on the data set provided at the back end. In recent years, personality prediction has increased, so much The data set and algorithms can be improved, which can aid with career recommendations while also increasing accuracy.

3.2 Future Works:

The accuracy of the model may be increased in the future by collecting more real-time data from different users and by focusing more on feature engineering to extract and assess new variables for a more precise prediction of a person’s personality.To increase accuracy, the data collected during the feature engineering step may be taught using a variety of supervised learning algorithms.

REFERENCES

[1] Aleksandar Kartelj, Vladimir Filipović, Veljko Milutinović, Novel approaches to automated personality classification: Ideas and their potentials.

[2] YagoSaez, Carlos Navarro, Asuncion Mochon and Pedro Isasi, A system for personality and happiness detection.

[3] Nurbiha A Shukora ,ZaidatunTasira, Henny Vander Meijden(2015). An Examination of Online Learning Effectiveness using Data Mining, Science.

[4] R. Wald, T. M. Khoshgoftaar, A. Napolitano Using Twitter Content to Predict Psychopath.

[5] Fazel Keshtkar, Candice Burkett, Haiying Li, andArthur C. Graesser, Using Data Mining Techniques to Detect the Personality of Players in an Educational Game.

[6] J. Golbeck, C. Robles, K.Turner, "Predicting

[7] personality with social media," In CHI’11 ExtendedAbstracts on Human Factors in Computing SystemsYagoSaez, Carlos Navarro, Asuncion Mochon and Pedro Isasi, A system for personality and happiness detection.

[8] Durgesh K. Srivastava, LekhaBhambhu, “Data Classification using Support Vector Machine,” Journal of Theoretical and Applied Information Technology.

[9] Yilun Wang, “Understanding Personality through social media,” International Computer Science stand ford University.

[10] Cantandir, I. Fernandez-Tobia] z, A. Belllogin, "Relating personality types with user preferences in multiple entertainment domains," . Cantandir, I. Fernandez-Tobia] z, A. Belllogin, "Relating personality types with user preferences in multiple entertainment domains.

[11] Mahesh Kini, Saroja Devi, Prashant G Desai, Niranjan Chiplunkar,” Text Mining approach to classify Technical Research Document using naive 26 Bayes”, International Journal of Advanced Research in computer and communication engineering.