Personality Traits Classification

20BCE084 Devangi Hadiya, 20BCE096 Aditi Jadeja, 20BCE091 Hetvi Bhansali,
20BCE128 Khushi Jain, 20BCE160 Heni Mehta

***Abstract*—A platform with millions of users is social media. With the growth of social networks, a wide range of methods have been created to categorise users' personalities based on their social interactions and linguistic preferences. Researchers from several domains have recently become interested in the status updates, blog entries, and tweets published by these people. In this study, our main goal is to identify personality traits of users by analysing the many socialising behaviours they share on these platforms. The purpose of this paper is to critically review the literature produced to date, explore the need for personality prediction from social media activity, and describe the different measures adopted to solve the difficulties faced by researchers in this sector.**

***Index Terms*—Classification, Machine Learning Algorithms, Personality Prediction, Big Five Personality Traits, social media**

I. Introduction

The idea is based on determining a person's personality utilising big 5 models and machine learning techniques. The Five-Factor Model (FFM) and OCEAN model are other names for the **Big Five model**. [NRF+13] Such personality-based characteristics are quite successful at raising a person's popularity and attractiveness. When statistical analysis is applied to personality survey data, general and specific words are used to describe the person, and these words can accurately be used to summarise the overall character or personality of the person.

Brief description of these properties is as follow :
• Open to Experience: imagination, sensitivity, attentiveness
• Conscientiousness: carefulness and diligence of the person
• Extraversion: social skills, interaction
• Agreeableness: generosity, ability to adjust with people.
• Neuroticism: mood swings and expressive power.

Some applications of Personality Traits Prediction Model :
• [Ali19]Social networking activities provides an excellent platforms to researchers to study and understand someone's online behaviors, preferences and personality on applications like Facebook, Twitter and Instagram. The statistical data about the user's views as communicated through their social media statuses is one of the key resources for research into the prediction of various human behaviours and personalities.
• This is highly used in dating apps and recommendation systems.
• Personality assessment tools can be used in career planning and employment related decisions.
• It helps recruiters to filter the right candidate and enable the company to choose the right talent.

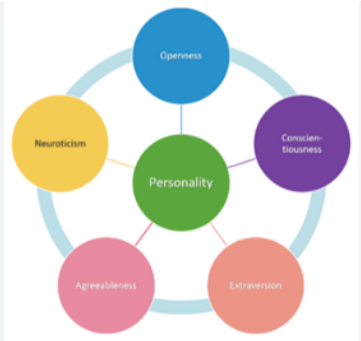

Fig. 1. Big Five Personality Traits

II. Literature Survey

Review of some researchers on the Big Five Personality Traits :

- Positive emotions and a propensity to seek out social interaction are signs of **extraversion**. It symbolises the propensity to be talkative, aggressive, energetic, cheery, and friendly. These people love social situations, appreciate stimulation and excitement, and feel a sense of vigour, fervour, and excitement (Costa and McCrae, 1992; John and Srivastava, 1999).

- The propensity to be dependable, obedient, attentive, kind, and gentle is known as **agreeableness**. Such people have a positive outlook on people. They are compassionate and want to help others, and they also anticipate support from others. Prosocial and having a communal orientation toward others are the main characteristics of agreeable people (Costa and McCrae, 1992; John and Srivastava, 1999)..

- People who exhibit **conscientiousness** are focused and dedicated. They tend to perform obediently, exercise self-control, and strive for success in comparison to a standard or external expectation. Thinking before acting, postponing gratification, abiding by societal norms and laws, as well as organising, planning, and prioritising chores are all examples of task- and goal-directed behaviour that are made possible by conscientiousness (John and Srivastava, 1999).

- The emotional adjustment or stability to emotional maladjustment or **neuroticism** continuum is measured by neuroticism (Costa and McCrae, 1992). The upper

end of neuroticism is exhibited by those who frequently feel dread, anxiety, melancholy, tension, rage, and guilt. Psychologically steady and even-tempered people are those with neuroticism scores in the lower range (Costa and McCrae, 1992; John and Srivastava, 1999).

- **Openness** to experience is a person's propensity for creativity, sensitivity, originality of thought, awareness of inner feelings, appreciation of art, intellectual curiosity, and sensitivity to beauty, according to Costa and McCrae (1992). John and Srivastava (1999). These people are open to innovative ideas and unusual values.

According to **J. Golbeck's** research, they were the first to notice a connection between personality qualities and the statistical information gathered about the social profile. [KS15] Through a Twitter programme they developed, they collected 2000 tweets from 50 different topics. The 45 questions pertaining to the Big Five Personality Traits were among these topics. Two tools were used to process the collected tweets, the first of which was LIWC (Linguistic Inquiry and Word Count), from which a total of 79 features could be extracted. Correlation is given little weight in this study, leaving it open for investigation across bigger datasets. Regression analysis was utilised by the authors to forecast the score of particular personality traits. Ten fold cross validation iterated ten times was used with two algorithms, the Gaussian Process and ZeroR. Scores were predicted by the authors to be between 11% and 18% of their actual values.

A study by **D. Quercia** involved 335 participants who had accounts on both Facebook and Twitter. The relationship between personality and various types of Twitter users was examined, and the personality scores were predicted based on the input of three counts: the number of people the user is following, the number of people who are following them, and the number of people they are listed as following (number of times the user is listed). For each personality trait, the authors used regression analysis with 10-fold cross validation. The authors also calculated Root Mean Square Error, which had a maximum value of 0.88, between anticipated and observed values. This created a solid platform for applying this study to recommender systems, user interface designs, and marketing. However, there have been discussions about how much information individuals reveal on such public profiles.

The Dark Triads of personality, which include narcissism, Machiavellianism, and psychopathy, were added by **C. Sumner** to the research on personality prediction beyond the Big Five Personality traits. To anticipate the dark triads of people's personalities, 2927 Twitter users' language use and profile characteristics were examined. The association between Twitter use and the dark triads of personality was first studied, according to the authors, who made this claim. A maximum of 3200 tweets were gathered, processed using LIWC, and 337 features were chosen for machine prediction use. The authors conducted a comparative analysis of a total of six models, including a polynomial kernel, Random Forest, and Naive Bayes Classifier. The study did uncover some brand-new information about the close ties between language use and antisocial behavior. The study also revealed some drawbacks, including subject selection bias and persistent linguistic usage problems in social media.

The idea of personality trait prediction in text groups was put forth by **Ana C.E.S. Lim**, who also expanded the challenge of personality prediction into a multi-label classification problem. The authors dubbed their model Bayesian Personality Predictor and utilised the Naive Bayes Algorithm to evaluate tweets. Their method was broken down into three steps: preprocessing, transformation, and classification. During the first phase, specific attributes from the tweets were extracted, and the second phase involved mapping multi-label sets into five single-label training sets. With the aid of these training sets, semi-supervised classification is finally carried out in the third phase. Accuracy, Recall, and Precision measures along with k-fold cross validation were used to assess the method. The authors claimed that their method would produce an average accuracy of 84 percent.

Research Paper	Sample Size	Algorithms	Evaluation Metrics
J. Golbeck et al.	279	Gaussian Process, ZeroR	11% to 18% of actual values
D. Quercia et al.	335	M5' Algorithm with 10 fold-cross validation	RMSE maximum 0.88
C. Sumner et al.	2927	SVM using SMO and a polynomial kernel, Random Forest, J48 algorithm, Naïve Bayes Classifier	Top 10% of distribution
Ana C.E.S Lim et al.	30 groups of users	Naïve Bayes Algorithm	84% average accuracy across classifiers

Table: Literature Overview

Fig. 2. Literature Overview

III. Comparative study through classification models

We have computed confusion matrix with the help of which we calculated the accuracy, precision and all the evaluation measures that can best define our model using different classification model like KNN, Gaussian Naïve Bayes, Random Forest, Decision Tree, SVM.

The observations are as following :

Sr. No.	Model Name	Accuracy	Precision	Recall	F1-Score
1	KNN	55.23	0.61	0.55	0.46
2	Gaussian Naïve Bayes	48.88	0.52	0.49	0.49
3	Multinomial Naïve Bayes	84.12	0.84	0.84	0.83
4	Decision Tree	28.88	0.38	0.29	0.31
5	SVM	37.46	0.39	0.37	0.38
6	Linear SVC	82.85	0.85	0.83	0.81
7	Random Forest	14.92	0.44	0.15	0.12

Fig. 3. Comparison of classification models

We observed that KNN because of imbalanced dataset gives "seriuous" for most of the test data. Random forest do have the least accuracy as the predicted data is discrete in nature. While Linear SVC gains highest accuracy because it implements one versus all multiclass reduction and also panelise the intercept.

IV. Methodology

Creating a machine learning model involves seven primary steps such as :

- Defining the problem
- Data collection
- Preparing the data
- Assigning appropriate protocols
- Training the model
- Evaluating success measures
- Parameter Tuning

A. Analysis of dataset

A set of data used to train the model is known as a machine learning dataset. To educate the machine learning algorithm how to make predictions, a dataset is used as an example. Text data and image data are some of the common sorts of data. Data analysis is the process of modifying raw data to produce insightful findings and judgements regarding the model to be implemented.
In our model, we explored and analyzed the train and test datasets thoroughly through manual approach and by using the python library : Pandas *(Built on top of the Python programming language, pandas is an open source data analysis and manipulation tool that is quick, strong, flexible, and simple to use)*
We observed these following points from the dataset :
• It has categorical value for gender.
• It also had some imbalanced set of rows for the classes of classifications.
To solve this, we came up with a solution where in we used **LabelEncoder** *(Character encoding converts data into a machine-readable format and assigns each class of data a unique zero-based number)* which converted categorical data into numerical data.

We concluded from the above analysis that the dataset is 86% pure and 14% imbalanced.

B. Filtration of dataset

Data filtration is the process of examining a dataset to remove, reorganise, or distribute data in accordance with specific criteria. For instance, data filtering can entail calculating the total number of sales for each quarter and excluding entries from the previous month. Data filtering is frequently used by IT professionals to carry out their duties and support others inside their organisation who are performing data analysis. In industries, data filtration has many potential uses such as :

- Process Records
- Modify Values
- Evaluate datasets
- Create new structures from datasets
- Exclude field or values

The major benefits of Data Filtering is :

- Improves efficiency of IT processes
- Allows better data security
- Reduces redundancy and unnecessary data

Feature Scaling is a technique for uniformly distributing independent data features over a predefined range. To work with highly variable quantities, values, or units, this is done as part of data preprocessing.
After detailed analysis, it was determined that the features needed to be scaled for the model to fit the resulting data set. Attributes are widely distributed with some outliers and should be treated accordingly.
We used **StandardScaler** which standardizes the feature by subtracting the mean and then scaling to unit variance.

C. Choosing best fit for the dataset

Model specification is the procedure for determining which independent variables should be included and excluded from a regression equation. An important step in statistics is model selection. Selection of best fit model depends on the type of dataset. If the dataset have linear relationship between input and output, Regression analysis is used for prediction and if the data is categorical, then Classification is used. If you specify incorrectly and choose the wrong model, your results may not be valid. Biases due to specification errors can overestimate, underestimate, or completely hide the existence of key relationships. The best values of independent variables to be included in regression equation can be decided using various metrics and procedures such as :

- Adjusted and Predicted R-Squared
- P-values for independent variables
- Step-wise and Best Subsets Regression

After analyzing and filtering the dataset, now the next step is to find the classification model that best fits our dataset. And thus, we have used **Multinomial Classification Technique**.
Multinomial is used here because some classes were overpowering in dataset and hence other models are prone to underfit and which will lead to decrease in efficiency.

D. Training the model

While training a machine learning model, by feeding data to the ML algorithm, it can recognize and find the best values for all relevant variables. Data sets called learning models are used to train ML algorithms. It includes examples of outputs and related input data sets that affect the outputs. The training model is used to run the input data through an algorithm and compare the processed output with the output of the model. The correlation results are used to tune the model. Model fitting is the term for an iterative process.
There are 2 types of Machine Learning models :

- Supervised learning
- Unsupervised learning

The better the training data is, the better the model performs. For our model, as we have selected multinomial classifier. Thus, we started the optimization of our dataset and trained it through the use of multinomial classification technique and logistic regression.

E. Testing

The run-time behaviour of the software under test is compared to predictions provided by a model in model-based testing. This method is called software testing. A model is a behaviour description of a system. Sequences of inputs, actions, circumstances, outputs, and the movement of data from input to output can all be used to define behaviour. It must provide a precise description of the system being tested in order to be shareable; it should also be practically understandable and reusable. Some of the models which describes the behaviour of the system are following:

- Data and control flow
- Dependency graphs
- Decision tables
- State transmition machines

After building a machine learning model, you need unknown data to test it (using training data). This data, also called test data, can be useful for evaluating the effectiveness and development of training algorithms and for tuning or optimizing algorithms for better results. After training the model, generate hypotheses. This test data set provides models that form hypotheses and predict personality traits. The requirement is to get as few errors as possible. Accuracy, confusion matrix, and all evaluation measures were calculated to better define the model.

Benefits of testing the model includes :

- Cost deduction
- Defect can be detected at an early stage
- Saves a lot of time
- Maintenance is easy

Every firm must put a significant amount of time and attention into MBT deployment. The following are MBT's disadvantages in software engineering.

- Testers must be skilled
- Time of learning curve will be more
- Understanding the model is complex

V. Training Results

- Accuracy : 84.12%
- Precision : 84%
- Recall : 84%
- F1-Score : 83%

VI. Conclusion

Numerous significant life outcomes can be predicted by each of the Big Five Traits of personality. This study used classification algorithms to categorise and assess personalities using the big five model using a particular data set in order to extract the authors' personalities from social networking websites. Personality traits are more accurately predicted by the suggested GUI-based Big Five Personality Traits model utilising Multinomial Classifier Approach. This system produces the highest F1-score of 83 % and accuracy of 84.12% on our reference dataset, averaging performance on most personality approaches implemented. There is still much work to be done, which can only be done by overcoming the limitations imposed by language use and user intent based on their own choices. We hope that this paper will satisfy the desire of the research communities to understand how a person behaves in relation to their categorised traits.

```
[[  7   4   1   0   9]
 [  0  70   1   0   6]
 [  0   3  18   1   2]
 [  2   4   0  26   8]
 [  1   6   1   1 144]]
```

Fig. 4. Confusion matrix

References

[Ali19] Imran Ali. Personality traits, individual innovativeness and satisfaction with life. *Journal of Innovation & Knowledge*, 4(1):38–46, 2019.

[KS15] Amanpreet Kaur Kanupriya Sharma. A review of the existing state of personality prediction of twitter users with machine learning algorithms. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 17(15):1344–1353, 2015.

[NRF+13] Raza Zaidi Nayyar, Abdul Wajid Rana, Batul Zaidi Farheen, Batul Zaidi Ghazala, and Taqi Zaidi Mohammad. The big five personality traits and their relationship with work engagement among public sector university teachers of lahore. *African Journal of Business Management*, 7(15):1344–1353, 2013.